Published in final edited form as:

J Comput Chem. 2011 April 30; 32(6): 1091–1100. doi:10.1002/jcc.21689.

The TIGER2 Empirical Accelerated Sampling Method: Parameter Sensitivity and Extension to a Complex Molecular System

Xianfeng Li and Robert A. Latour^{a)}

Department of Bioengineering, Clemson University, Clemson, South Carolina 29634

Abstract

The recently developed ‘temperature intervals with global exchange of replicas’ algorithm (TIGER2) is an efficient replica-exchange sampling algorithm that provides the freedom to specify the number of replicas and temperature levels independently of the size of the system and temperature range to be spanned, thus making it particularly well suited for sampling molecular systems that are considered to be too large to be sampled using conventional replica exchange methods. Although the TIGER2 method is empirical in nature, when appropriately applied it is able to provide sampling that satisfies the balance condition and closely approximates a Boltzmann-weighted ensemble of states. In this work, we evaluated the influence of factors such as temperature range, temperature spacing, replica number, and sampling cycle design on the accuracy of a TIGER2 simulation based on molecular dynamics simulations of alanine dipeptide in implicit solvent. The influence of these factors is further examined by calculating the properties of a complex system composed of the B1 immunoglobulin-binding domain of streptococcal protein G (protein G) in aqueous solution. The accuracy of a TIGER2 simulation is particularly sensitive to the maximum temperature level selected for the simulation. A method to determine the appropriate maximum temperature level to be used in a TIGER2 simulation is presented.

1. INTRODUCTION

The replica exchange molecular dynamics (REMD) method has been extensively used in sampling the properties of complex systems.^{1–7} By conducting a number of parallel simulations at different temperature levels and periodically attempting to exchange thermally adjacent replicas, the method is able to help a system escape from local energy minima, thus overcoming the kinetic trapping problem experienced in conventional molecular dynamics (MD) simulations. A successful REMD simulation requires sufficient overlap between the potential energy probability distributions for replicas at neighboring temperatures in order to obtain an acceptable degree of replica swapping. Because of the narrowing (relative) width of the energy distribution with increasing system size combined with the linear increase in the mean potential energy with both system size and temperature, an REMD simulation of a large system must use a large number of replicas at closely spaced temperature levels to obtain an acceptable swapping acceptance ratio. The use of a large number of replicas results in not only high demand of computational resources, but also slow diffusion of configurations throughout the temperature space, which increases the time necessary to equilibrate the system.^{8, 9}

To overcome the above problem in the REMD method, we previously developed a new sampling method named “temperature intervals with global energy reassignment” (TIGER),⁸ in which a quenching cycle was introduced to quench all replicas at the elevated

^{a)}latourr@ces.clemson.edu.

temperatures down to the baseline temperature level prior to the application of the exchange criterion. A Metropolis-like exchange procedure is then conducted based on the potential energies of the quenched replicas and the temperature level that they were quenched from, and the replicas are then globally reassigned and reheated to the designated temperature levels. The quenching procedure effectively removes energy differences that are purely thermal, greatly reducing the number of replicas required and eliminating the diffusion problem in REMD, thus substantially improving computational and sampling efficiency. To generate a Boltzmann-weighted ensemble of states at the baseline temperature, the TIGER algorithm relies on the assumptions that the quenching process does not change the general conformational state of the system and the amount of thermal energy removed during quenching is state-independent. However, these assumptions are not easily maintained for complex systems and thus limit the application of the method.

To preserve the advantages of quenching out the excess thermal energy and make the method more effective in modeling complex systems, we recently proposed a revised scheme named ‘temperature intervals with global exchange of replicas’ (TIGER2),⁹ which does not require the limiting assumptions of the TIGER algorithm. The main differences between these two sampling methods are in the exchange and reassignment steps that are carried out at the baseline temperature. In TIGER, replica exchanges are sequentially attempted based on a sampling criteria that is very analogous to that used in the conventional REMD method,^{1–7} whereas in TIGER2, quasi-random sampling of relevant conformational states of a system is performed by randomly selecting one state from among the set of quenched states and then comparing it with the state previously sampled at the baseline temperature using the conventional Metropolis Monte Carlo exchange criterion.¹⁰ In our previous studies,⁹ we empirically demonstrated the ability of TIGER2 to closely approximate Boltzmann statistics at the baseline temperature by comparison with Monte Carlo (MC) and REMD simulations for a set of distinctly different types of molecular systems including butane in vacuum, alanine dipeptide with explicit solvation, and (AAQAA)₃ and chignolin peptides folding in implicit solvent. The TIGER2 results closely matched the MC or REMD results, indicating that, when appropriately applied, the TIGER2 algorithm is able to empirically satisfy the balance condition and is able to generate Boltzmann-like ensembles.^{9, 11}

Because TIGER2 is empirical in nature, however, it is necessary to assess the sensitivity of the algorithm to the protocol applied in a simulation and how to appropriately set simulation parameters in order to provide a close approximation to a Boltzmann-weighted ensemble of states for a given molecular system. To assess these sensitivities and provide guidelines for the application of TIGER2, we have conducted a set of TIGER2 simulations to evaluate the influence of factors such as temperature range, temperature spacing, replica number, and the design of the sampling cycle (i.e., amount of time spent in the elevated temperature sampling and quenching phases of a sampling cycle) on the accuracy of a TIGER2 simulation in comparison with results obtained using the conventional REMD method. Simulations were performed using a simple system of alanine dipeptide and a more complex system consisting of the B1 immunoglobulin-binding domain of streptococcal protein G (protein G) in aqueous solution. The results of these simulations show that TIGER2 sampling is fairly insensitive to all of the above-stated parameters except for the maximum temperature level that is used in the simulation, which was found to have a significant influence on the sampled ensemble of states. Methods to address this issue and to appropriately select the maximum temperature level to be used for a given molecular system are then presented.

2. METHODS AND MATERIALS

2.1. The Replica-Exchange Molecular Dynamics (REMD) Method

Specific details of the REMD method can be found elsewhere.^{1, 2} In brief, an REMD simulation constructs a Boltzmann-weighted ensemble by generating independent trajectories at different temperatures, T_m ($m = 1, \dots, N_r$), with N_r being the number of temperature levels or replicas. An individual replica in this ensemble is represented by $(\mathbf{X}(T_m), \mathbf{P}(T_m))$, where $\mathbf{X}(T_m)$ and $\mathbf{P}(T_m)$ represent the positions and the momenta of all atoms of the m^{th} replica at temperature T_m . The acceptance probability for an attempt to exchange two neighboring replicas $(\mathbf{X}(T_i), \mathbf{P}(T_i))$ and $(\mathbf{X}(T_j), \mathbf{P}(T_j))$ is:

$$\min[1, \exp(-(\beta_j - \beta_i)(E_i - E_j))], \quad (1)$$

where E_i and E_j are the potential energies of replicas i and j , respectively, and β_i and β_j represent $(k_B T_i)^{-1}$ and $(k_B T_j)^{-1}$, respectively, with k_B being Boltzmann's constant. After a successful exchange, the momenta of the swapped replicas are updated by:

$$\mathbf{P}(T_i) = \sqrt{\frac{T_i}{T_j}} \mathbf{P}(T_j), \quad \mathbf{P}(T_j) = \sqrt{\frac{T_j}{T_i}} \mathbf{P}(T_i). \quad (2)$$

2.2. The Temperature Intervals with Global Exchange of Replicas (TIGER2) Method

A TIGER2 simulation is composed of a series of sampling cycles with each cycle containing four stages: (1) rapid heating from a baseline temperature (T_B) to the replica temperature (T_m) by rescaling the momenta of the atoms within the replica by a factor of $\sqrt{T_m/T_B}$ and thermally equilibrating, (2) molecular dynamics sampling at constant temperature (T_m), (3) rapid quenching back down to T_B by rescaling the momenta by a factor of $\sqrt{T_B/T_m}$ followed by thermal equilibration (note that the replica at the baseline temperature does not need to be quenched. The quenching process in TIGER2 consists of just rescaling and equilibrating, and is more efficient than the quenching process in TIGER,⁸ which contains energy minimization, heating and equilibrating.), and (4) global replica reassignment. Stage (4) consists of two substeps: (A) one state from among the set of $(N_r - 1)$ quenched states is randomly selected with a probability of $1/(N_r - 1)$, and the potential energy of this state is then compared with the state from the production run of the baseline replica using the conventional Metropolis sampling criterion,¹⁰

$$\min\{1, \exp[-(E_i - E_B)/k_B T_B]\}, \quad (3)$$

where E_i is the potential energy of the selected quenched state and E_B is the potential energy of the baseline state, and (B) all replicas except the selected baseline replica are then reassigned to higher temperature levels according to their potential energies; i.e., a higher potential energy state is assigned to a higher temperature level. Further details of the TIGER2 method are presented in our initial paper on this method.⁹

2.3. Model Systems and Simulation Setup

Two different molecular systems were considered: (1) alanine dipeptide and (2) protein G in aqueous solution. All simulations were carried out with the CHARMM program^{12, 13} using the CHARMM (c34b2) force field with CMAP.¹⁴ The aqueous solvent was represented with

an implicit solvent model (ISM) based on a screened Coulomb potential (SCP) formulation.^{15, 16} The Langevin dynamics method was used with SCP-ISM to mimic frictional drag and random collision of the solvent on the solute molecule.¹⁷ For simplicity, in all calculations, a constant friction coefficient (expressed as collisions per picosecond) $\zeta_i = 20 \text{ ps}^{-1}$ was used for the heavy atoms in the peptide. All simulations were performed with an atom-based 13 Å cutoff and a shifting function starting at 12 Å. The SHAKE algorithm¹⁸ was used to constrain all bond lengths containing hydrogen, and the leapfrog algorithm¹⁹ was used with a time step of 2 fs to integrate the equations of motion in the dynamics simulations. Both REMD and TIGER2 simulations were conducted for each molecular system and their results were compared. In all simulations, different random seeds were used for different replicas. The TIGER2 simulations for alanine dipeptide and protein G used four and eight replicas, respectively. The number of replicas required for the simulations are determined by short simulations (500 ps) resulting in exchanging rates between 20% and 30% at 300 K baseline temperature.

2.3.1. Alanine Dipeptide with Implicit Solvation—The alanine dipeptide system was used to systematically test the influence of the values of the TIGER2 sampling parameters on the accuracy of a TIGER2 simulation. One REMD simulation and eleven TIGER2 simulations were carried out with each simulation beginning with the same initial conformation of the alanine dipeptide. A 50 ns trajectory was generated in each simulation. A summary of the protocols followed in these simulations is given in Table I, with the different parameter sets designated as TA – TK. The effects of temperature level spacing, the length of sampling time conducted at each elevated temperature level, the length of quenching time, the number of replicas, and the maximum replica temperature were evaluated in comparison with a defined TIGER2 parameter set involving a temperature range of 300 – 600 K using four replicas evenly spaced in 100 K increments with a sampling cycle of 1 ps heating, 1 ps sampling at elevated temperature, and 1.2 ps quenching to the baseline temperature (i. e., TA).

The temperature range used for the TIGER2 simulation needs to be determined before the simulation is conducted. It is noted, that even for standard replica exchange, the value to use for the highest replica temperature is not clearly defined, but may require several preliminary studies to be first conducted before an appropriate maximum temperature level can be identified to enable the relevant phase space of the system of interest to be sampled within a practical timeframe. In this work, we devised a standard method to select the maximum temperature level that should be used with a TIGER2 simulation. Based on the need to be able to sample all relevant microstates in a system within a practical timeframe (i.e., microstates with a potential energy leading to a non-negligible Boltzmann probability of occurrence at the baseline temperature of interest) and the tendency of TIGER2 to over-sample high energy states as the maximum temperature level is increased, it is proposed that the maximum temperature for a TIGER2 simulation should be set to a level that will just enable all microstates of a system to be sampled within a practical timeframe, but not substantially higher, in order to minimize the tendency to over-sample the high energy states of the system. To test this procedure to determine the proper temperature range for the TIGER2 simulations for alanine dipeptide, a separate test simulation was carried out with 12 temperature levels that were evenly distributed from 300 K to 740 K. This simulation was conducted using a 1 ps heating, 1 ps sampling, 1.2 ps quenching sampling cycle as in a typical TIGER2 simulation but without any exchanges being attempted between the replicas. As shown in the results, by monitoring the distribution of the potential energy at each temperature level, this simulation was able to identify the proper temperature range for the TIGER2 simulation of the alanine dipeptide with implicit solvent.

2.3.2. Protein G with Implicit Solvation—At room temperature, the stable structure of the 56-residue protein G contains a packed core of nonpolar residues located between a four-stranded β -sheet and a four-turn α -helix. Its three-dimensional structure has been determined by X-ray crystallography²⁰ and NMR spectroscopy;²¹ these two structures differ primarily in the loop involving residues 46–51 with other minor structural variations being observed in the helix. Computer simulations have been extensively carried out to study the folding mechanism and the properties of this protein in water.^{2, 3, 5, 17, 22 – 24} In this work, the dynamics of the folded protein G was studied by two REMD and eight TIGER2 simulations using different protocols, and the influence of the TIGER2 parameters on the accuracy of the simulation was further examined relative to the REMD results. All simulations for protein G were started from the crystal structure of protein G [Protein Data Bank (PDB) identification number: 1pgb] with all of the crystallographic water molecules being removed. A 50 ns trajectory was generated in each simulation with the objective of the simulations being to assess the ability of each accelerated sampling method to successfully maintain the native-state structure of protein G at the baseline temperature (T_B), which was set at 300 K.

In order to establish the maximum temperature level to be used in the TIGER2 simulations of protein G, we desired to identify the minimum temperature level that would fully unfold the protein within a reasonable timeframe, thus providing access to all relevant microstates in the system while minimizing the tendency to over-sample high energy states at the baseline temperature level of interest (i.e., the ensemble of accepted states at T_B). To determine the maximum temperature level for the TIGER2 simulations for protein G, a set of preliminary simulations were therefore carried out in which the conformation of the protein was sampled at 20 temperature levels that were evenly distributed from 300 K to 680 K. These simulations were conducted using a 1 ps heating, 9 ps sampling, 2 ps quenching sampling cycle as in a typical TIGER2 simulation but without any exchanges being attempted between the replicas. As shown in the results for this set of studies, the longer sampling time at elevated temperature (i.e., 9 ps) was used for this system based on preliminary tests that indicated that the 1 ps production run time used for the alanine dipeptide system did not provide sufficient time for this more complicated structure to unfold at the elevated temperatures. The slightly longer quenching time (i.e., 2.0 ps vs. 1.2 ps used for alanine dipeptide) was selected to provide a bit more time for the removal of thermal energy for this more complex molecular system prior to making exchange decisions based on comparison with the potential energy value of the 300 K baseline replica. By monitoring the distribution of the potential energy and the percent of unfolding of the protein at each temperature level, this set of preliminary simulations was able to be used to identify the minimum temperature that was needed to obtain approximately full unfolding of the protein (i.e., at least 95% unfolding). This temperature was then used as the maximum temperature level in subsequent TIGER2 production-run simulations using a selected subset of these pre-equilibrated replicas, thus minimizing the additional computational cost of performing the preliminary simulations.

Table II summarizes the protocols that were applied for all of our simulations of protein G. To test the effect of the designated maximum temperature level, one REMD (R1) and five TIGER2 simulations were conducted. The TIGER2 simulations are labeled from T1 to T5, with maximum temperature being 520, 552, 580, 610 and 640 K, respectively. The median value of 580 K was used as the maximum temperature used in R1. The temperature range used in T2, from 300 K to 552 K, was used in another set of simulation (T21, T22 and T23 using 1, 6, and 12 ps simulation time at the assigned temperature, respectively) to further examine the influence of cycle simulation time on the accuracy of the simulation. In addition, a second REMD (R2) simulation with same temperature range (i.e., 300 – 552 K) was also performed for comparison sake.

3. RESULTS AND DISCUSSION

3.1. Alanine Dipeptide with Implicit Solvation

For the alanine dipeptide, the main features of the backbone dihedral angle ϕ/ψ distribution are the four regions corresponding to conformations designated C_7^{eq}/C_5 , α_R/β_2 , α_L , and C_7^{ax} . The boundaries between these regions of the ϕ/ψ map are indicated in the Ramachandran plot presented in Fig. 1. The relative energies and positions of the four local minima in the energy surface of alanine dipeptide with SCP-ISM solvation are available from a previously reported analysis (see Table I in Ref. ¹⁷). The sampling probability for each of the four regions was calculated based on the SCP-ISM energies and these results are presented in the first row of Table III. Based on the last 5 ns trajectories obtained from the REMD and TIGER2 simulations designated in Table I, the population ratios for the conformers located in the four regions were calculated to quantitatively contrast the influence of the simulation protocols on the accuracy of the TIGER2 simulations. These results are shown in Table III.

The REMD populations are consistent with the values predicted from SCP-ISM energies, as expected. The overall results of TIGER2 agree qualitatively with the REMD results with relatively minor differences in the sampled density of states. For all simulations except TF, which used the highest maximum temperature level, the population differences between REMD and TIGER2 are less than 5%, even for the most probable conformational state (i.e., C_7^{eq}/C_5). Among the TIGER2 results, the results of TA and TB are very similar, indicating that the spacing of the temperature levels does not strongly influence the simulation results. In the rest of the TIGER2 simulations, the even distribution of temperature levels was used. By comparing the results of TA and TC, and TA and TD, as shown in Table III, it is observed that the variations of the length of sampling time at the elevated temperature or the quenching time do not substantially influence the simulation results. Furthermore, the agreement between the results for the TA and TE simulations indicates that the change in the number of replicas used to span the designated temperature range also has minimal influence on the simulation results.

For alanine dipeptide in SCP-ISM solvent, the only factor that leads to relatively large differences in the sampled populations is the maximum temperature level used for the simulation. As indicated by the comparison of the results of TA, TF, TG, TH, TI, TJ, and TK shown in Table III, the population in C_7^{eq}/C_5 increases by about 12% when the maximum temperature changes from 900 (TF) to 400 (TK), with the results of simulation TJ (450 K maximum temperature) being in closest agreement with the results of the REMD simulation. These results show that an excessively high maximum temperature level leads to under-sampling of the lowest energy state of the system at the baseline temperature, whereas the use of a very low maximum temperature level results in the over-sampling of the lowest energy state. Based on these results, we recommend that users of this method conduct preliminary simulations to establish the proper maximum temperature level to be used before conducting a TIGER2 simulation. We recommend that the maximum temperature level be set to the minimum temperature that will enable all relevant energy barriers in the system to be readily crossed, thus enabling all relevant microstates of the system to be accessed during the simulation, while minimizing the tendency to over-sample the high energy states of the system. An example to show the way of determining the maximum temperature is presented in Fig. 2, which shows the fitted Gaussian distributions of potential energy after quenching each replica to 300 K for TIGER2 cycling without exchanges for a total of 50 ns of sampling using temperature levels from 300 K to 740 K (Fig. 2A). As shown in Fig. 2A, for the simple system of alanine dipeptide, the energy distributions of the quenched replicas greatly overlap with each other, with the distributions shifting towards the sampling of higher energy states as the maximum temperature level is increased. This

shifting of potential energy is quantitatively shown in Fig. 2B, which plots the mean potential energy of the quenched distributions versus the elevated temperature levels that the distributions were quenched from. As clearly shown in this figure, the mean potential energy of each distribution increases with increased maximum elevated temperature level until about 600 K, at which point a distinct transition occurs such that further increases in the maximum temperature level do not lead to substantially higher mean potential energies in the set of sampled states following quenching to 300 K. This relationship demonstrates that once the maximum temperature exceeds 600 K, the energy barriers separating the different states of the system essentially become insignificant compared to the thermal energy of the system, with further increases in temperature leading to negligible changes in the distribution of states that are being sampled. This transition provides a very practical means of identifying the maximum temperature level that should be used in a TIGER2 simulation of this system, with the use of temperature levels above this maximum temperature serving only to cause over-sampling of the high energy states of the system. Fortunately, as indicated by the comparison of the results of TA, and TF through TK shown in Table III, the population in C_7^{eq}/C_5 is still relatively insensitive to the specific value of the maximum temperature level that is selected up to this maximum temperature level, with a difference in sampling probability of only about 5% when the highest temperature changes from 400 (TK) to 600 (TA).

3.2. Protein G with Implicit Solvation

To establish the maximum temperature level for the TIGER2 simulation for protein G, a set of preliminary simulations using TIGER2 cycling but without exchange were first conducted. Starting from the crystal structure of protein G in SCP-ISM solvent, preliminary TIGER2 simulations were performed with 20 independent replicas sampled at temperature levels uniformly distributed from 300 to 680 K in order to identify the maximum temperature required to fully unfold the protein within a practical simulation timeframe (taken as 10 ns for the sake of our simulations). A 9 ps sampling time at the respective elevated temperature level was used in each heating-sampling-quenching cycle to allow sufficient time to cross the local energy barriers of the protein prior to quenching the system to the baseline temperature level. The folded fraction of the protein was calculated using the DSSP program in the CHARMM package.¹³ The results of these preliminary simulations are shown in Fig. 3. As shown in this figure, the protein folding fraction steadily decreased as the temperature level was increased from 300 K to 680 K, with the minimum temperature level for nearly complete protein unfolding being found to occur at a temperature of about 550 K. The fitted Gaussian distributions of the system potential energy and the relationship of the mean energy and the maximum sampling temperature for all quenched replicas are presented in Fig. 4A and 4B, respectively. Similar to the case for alanine dipeptide, the mean potential energy values for the distribution of states quenched to 300 K versus the temperature level that the replicas were quenched from show a distinct transition at about 550 K where the mean potential energy of the set of quenched states increases slowly with increasing sampling temperature. This transition occurs because once the protein is fully unfolded (i.e., all secondary structures disappear), further increases in the elevated temperature followed by quenching will only result in sampling higher-energy conformational states due to random-coil dihedral orientations of the polypeptide chain without additional contributions from the unfolding of the protein's secondary and tertiary structures. Based on these results, the temperature about 550 K was considered to just cover the full range of relevant potential energies after quenching and hence provide full sampling of the relevant phase space of the molecular system without over-sampling of the high energy states. This temperature was thus designated to be the desired maximum temperature level for the subsequent TIGER2 production-run simulations.

Based on the result of these preliminary studies, we then conducted two sets of simulations for protein G (designated Sets A and B) to test the effects of temperature range and sampling time at elevated temperature on the ensemble of states sampled at the baseline temperature of 300 K. Set A contained one REMD simulation (R1) with temperature range of 300 – 580K and five TIGER2 simulations (T1, T2, T3, T4, and T5) with different temperature ranges (see Table II). Set B contained one REMD simulation (R2) and four TIGER2 simulations (T2, T21, T22, and T23) with the same temperature range of trial T2 of Set A, which was 300 – 552 K (see Table II).

For both of the REMD simulations (using 24 replicas for each simulation) and all TIGER2 simulations (using 8 replicas for each simulation), the replica swapping acceptance ratios varied within a range of 0.2 – 0.3, thus demonstrating that a substantial degree of exchange was occurring during the simulations and that ample opportunity was provided for each algorithm to accept structural states of the protein other than its original native state if the algorithm was not following a Boltzmann-weighted sampling process. For each simulation, the folding fraction at each temperature level was calculated. The results for the two sets of simulations (i.e., Set A and B) are plotted in Figs. 5 and 6, respectively. Table IV presents the averaged values of the folding fraction, the root mean squared deviation of C $_{\alpha}$ atoms from the NMR structure of Protein G (C $_{\alpha}$ RMSD), and the corresponding standard deviations at the 300 K baseline temperature for each simulation. The structure of protein G remained stable at 300 K in all of these simulations. The C $_{\alpha}$ RMSDs fluctuate about 0.85 Å for REMD and about 1.0 Å for each TIGER2 simulation and the folding fractions are all predicted to be around 0.8. As shown by the results presented in Table IV, the agreement between REMD and TIGER2 is excellent for each set of conditions, thus showing relatively low sensitivity of the TIGER2 algorithm to the applied changes in its parameter set.

The effect of temperature range is examined with Set A and the result is displayed in Fig. 5. In simulation T1 with 520 K maximum temperature, the protein at the highest temperature was only partially unfolded (with a folding fraction about 0.36 after 50 ns simulation), indicating that this temperature range was not sufficiently high to enable the local energy barriers to be crossed prior to quenching, thus not accessing all folded states of the protein during the sampling process. The melting curves obtained for the rest of the TIGER2 simulations (T2 – T5) are consistent with each other and agree reasonably well with the REMD (R1) results at each temperature level. As also indicated in Fig. 5, the maximum temperature level of 552 K in simulation T2 is shown to be sufficient to provide near full unfolding of protein G for both the TIGER2 and the REMD simulations, as predicted by the preliminary TIGER2 simulations that were conducted to set the desired maximum temperature level for the TIGER2 simulations. The effect of sampling time at the elevated temperatures is clearly shown by the result of Set B presented in Fig 6. It is observed that the length of the sampling phase substantially influences the behavior of the “hot” replicas (sampled at temperatures other than the baseline temperature) if it is below a designated threshold value. In simulation T21 with 1 ps sampling at the elevated temperature levels in each cycle, the protein G replica at the highest temperature (550 K) achieved only partial unfolding (with a folding fraction about 0.56 after 50 ns simulation), thus indicating that this sampling time was not sufficiently long to enable the local energy barriers to be crossed prior to quenching during each sampling cycle. In comparison, when a 9 ps sampling time at the elevated temperatures was used in the sampling cycle (e.g., simulation T2), it is apparent that sufficient time was provided for the local energy barriers to be crossed, with the degree of unfolding for each elevated temperature level being much more closely aligned with the R2 result. In comparison, the results shown in Fig. 6 obtained from T22 with 6 ps and T23 with 12 ps elevated-temperature sampling times show similar folding fractions at each temperature level and closely agree with the R2 results. These combined results indicate that a range of sampling length between 6 – 12 ps is appropriate for complex biomolecular

systems like protein G in implicit solvent. The effect of increased elevated temperature sampling time from 6 ps to 12 ps is also shown in Table IV to cause the C_α RMSD values to increase by about 15%, thus indicating that a longer sampling phase of the sampling cycle does tend to lead to a greater deviation of the folded structure of the protein.

Fig. 7 shows excellent superposition of the structures from simulation R2 and T2 on the crystal structure of protein G. Here the R2 and T2 structures shown represent the lowest energy states picked from the last 10 ns trajectory of each simulation. The C_α RMSD from the crystal structure (red) obtained from the R2 (yellow) and T2 (blue) structures are 0.797 Å and 0.926 Å, respectively. While the agreement between these minimum energy structures and the RMSD and percent folded values for the ensemble of sampled states shown in Table IV may seem like obvious results given that the simulations were all started from the fully folded structures of protein G, it must be recognized that the full range of unfolded states for this protein was represented in the complete set of replicas for both the REMD and TIGER2 simulations and both simulation methods maintained an exchange acceptance ratio of 0.2 – 0.3. Thus there was ample opportunity for the TIGER2 simulation to construct an ensemble of states that was far different from the REMD results. Instead, however, the TIGER2 simulation results are found to be in close agreement with the REMD results, thus demonstrating that the TIGER2 algorithm is again able to satisfy the balance condition and provide a close approximation to the proper Boltzmann-distribution of states for this relatively complex molecular system. To further address this point, we plot the probability density distributions for the sampled potential energies from simulations R2 and T23 in Fig. 8 and present the accompanying statistical analyses in Table V. Comparison of the means (by Student's t-test) and the variances (by F-test) were then conducted to compare each distribution. As shown by Fig. 8, the potential energy probability distributions obtained using TIGER2 (solid line) closely approximate the distributions obtained using REMD (dotted line). Based on 0.05 significance level, there is no significant difference between the variances of the energy distributions from the two methods. However, a difference in the means about 3.3 kcal/mol (<0.06% with respect to the REMD mean) provides a condition of $p < 0.05$. While the means of the energy distributions are shown to be statistically different, the TIGER2 distribution still provides a very close approximation to the REMD distribution.

IV. CONCLUSIONS

In this set of studies, we have investigated the influence of factors such as temperature range, temperature spacing, replica number, sampling time at elevated temperature, and quenching time on the accuracy of TIGER2 simulations in order to evaluate the sensitivity of this empirical accelerated sampling algorithm to these parameters and to provide guidelines for its implementation.

As revealed by the simulations of a relatively simple system like alanine dipeptide in implicit solvent represented by SCP-ISM, the temperature spacing within the designated temperature range, length of the sampling and quenching phases of the TIGER2 sampling cycle, and number of replicas do not significantly influence the simulation results at the baseline temperature. However, the choice of the maximum temperature level used in the simulation was found to have greatest and fairly substantial influence on the distribution of states sampled at the baseline temperature. Similar results were found for the much more complex system represented by protein G with SCP-ISM implicit solvation, but with the additional finding that the use of a 1 ps sampling time at the elevated temperatures for the sampling cycle is not sufficiently long to enable local energy barriers to be efficiently crossed to unfold the protein at the maximum temperature level used in the simulation. There is thus a minimum sampling time necessary at the elevated temperature, which we determine to be in the range of 6 ps. In order to establish the maximum temperature level to

be used in a TIGER2 simulation, we recommend that preliminary equilibration simulations be implemented by applying the TIGER2 simulation cycles without exchange between replicas with replicas spaced about 20 K apart to determine the minimum temperature level and elevated temperature sampling time that will enable all relevant microstates of the system to be accessed within a practical timeframe. Once conditions for complete sampling are established, production-run TIGER2 simulations with exchange between replicas can then be continued using a subset of these pre-equilibrated replicas, which should then provide the ability to obtain a close approximation of a Boltzmann-weighted ensemble of sampled states at the baseline temperature. Because production-run data should not be collected prior to attaining system equilibration for any replica exchange simulation, the need to conduct this preliminary set of simulations actually does not increase the amount of wall-clock time required compared to a simulation that would be performed without this initial step.

The TIGER2 method has now been tested with a number of distinctly different types of molecular systems, such as butane in vacuum, alanine dipeptide with explicit solvation, (AAQAA)₃ and chignolin with implicit solvation in our previous work,⁹ and alanine dipeptide and protein G in implicit solvation in this present work. With proper setup of the simulation parameters, TIGER2 provides sampling distributions that are in agreement with results obtained with the conventional REMD method. These combined results demonstrate that the TIGER2 algorithm is able to provide sampling at the baseline temperature in a manner that closely approximates a Boltzmann-weighted ensemble of states using substantially fewer replicas than is required by the conventional REMD method. Of particular importance, the TIGER2 algorithm allows the number of replicas that are used in a simulation to be set independently of both the temperature range and the number of degrees of freedom in a given molecular system, thus enabling users to select the number of replicas used in a simulation based on the computing resources at hand. This feature is in distinct contrast to the conventional REMD method in which the minimum number of replicas for a given temperature range is determined by system size, with the number of replicas required for a simulation possibly exceeding the number of CPUs available to a user at a given facility. Although shorter exchange times can be used in REMD (e.g., 10 fs)¹ to accelerate the diffusion among replicas and increase the speed of convergence, this does not influence the number of replicas that are needed to perform a simulation for a given system. Therefore, the reduction in the total number of replicas (and thus CPUs) needed to run a given simulation is the primary advantage of TIGER2 over REMD. This unique feature makes TIGER2 a promising accelerated sampling method for modeling complex systems, with its implementation being particularly valuable for sampling systems that are considered too large to be practically handled using the conventional REMD method.

Acknowledgments

This work was supported by 'RESBIO – The National Resource for Polymeric Biomaterials' funded under NIH Grant No. P41 EB001046. Partial support was also provided by NIH under R01 GM074511 and R01 EB006163. We also gratefully acknowledge the computational support provided by the staff and resources of the Palmetto Cluster facility at Clemson University, Clemson, SC.

References

1. Sugita Y, Okamoto Y. Chem Phys Lett. 1999; 314:141–151.
2. García AE, Sanbonmatsu KY. Proteins. 2001; 42:345–354. [PubMed: 11151006]
3. Zhou RH, Berne BJ, Germain R. Proc Natl Acad Sci USA. 2001; 98:14931. [PubMed: 11752441]
4. Liu P, Kim B, Friesner RA, Berne BJ. Proc Natl Acad Sci USA. 2005; 102:13749. [PubMed: 16172406]

5. Andrec M, Felts AK, Gallicchio E, Levy RM. *Proc Natl Acad Sci USA*. 2006; 102:6801. [PubMed: 15800044]
6. Liu P, Huang XH, Zhou RH, Berne BJ. *J Phys Chem B*. 2006; 110:19018. [PubMed: 16986898]
7. Baumketner A, Shea JE. *J Mol Biol*. 2006; 362:567. [PubMed: 16930617]
8. Li XF, O'Brien CP, Collier G, Vellore NA, Wang F, Latour RA, Bruce DA, Stuart SJ. *J Chem Phys*. 2007; 127:164116. [PubMed: 17979328]
9. Li XF, Latour RA, Stuart SJ. *J Chem Phys*. 2009; 130:174106. [PubMed: 19425768]
10. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. *J of Chem Phys*. 1953; 21:1087.
11. Manousiouthakis VI, Deem MW. *J Chem Phys*. 1999; 110:2753.
12. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. *J Comp Chem*. 1983; 4:187.
13. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. *Journal of Physical Chemistry B*. 1998; 102:3586.
14. MacKerell AD Jr, Feig M, Brooks CL III. *J Comput Chem*. 2004; 25:1400. [PubMed: 15185334]
15. Hassan SA, Guarnieri F, Mehler EL. *J Phys Chem B*. 2000; 104:6478.
16. Hassan SA, Mehler EL. *Proteins*. 2002; 47:45. [PubMed: 11870864]
17. Li XF, Hassan SA, Mehler EL. *Proteins*. 2005; 60:464. [PubMed: 15959866]
18. van Gunsteren WF, Berendsen HJC. *J Mol Phys*. 1977; 34:1311.
19. Allen, MP.; Tildesley, DJ. *Computer Simulation of Liquids*. Clarendon Press; Oxford: 1987.
20. Gallagher T, Alexander P, Bryan P, Gilliland GL. *Biochemistry*. 1994; 33:4721. [PubMed: 8161530]
21. Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM. *Science*. 1991; 253:657. [PubMed: 1871600]
22. Dinner AR, Lazaridis T, Karplus M. *Proc Natl Acad Sci USA*. 1999; 96:9068. [PubMed: 10430896]
23. Pande VS, Rokhsar DS. *Proc Natl Acad Sci USA*. 1999; 96:9062. [PubMed: 10430895]
24. Ma B, Nussinov R. *J Mol Biol*. 2000; 296:1091. [PubMed: 10686106]
25. Humphrey W, Dalke A, Schulten K. *J Mol Graphics*. 1996; 14:33.
26. Wessa, P. 2009. http://www.wessa.net/rwasp_autocorrelation.wasp/

Fig. 1.

The conformational distribution of the alanine dipeptide in SCP-ISM solvent generated by 50 ns TA simulation and the designated regions for macro-states $C7_{eq}/C5$ (**A**), α_R/β_2 (**B**), α_L (**C**), and $C7_{ax}$ (**D**).

Fig. 2.

Influence of the sampling temperature on (A) potential energy distribution and (B) average potential energy for alanine dipeptide in 50 ns test simulations without any exchange between replicas.

Fig. 3.
Influence of sampling temperature on the folding fraction for protein G in the test simulation without any exchange between replicas. Error bars represent the standard deviation about the mean values.

Fig. 4. Influence of the sampling temperature on (A) potential energy distribution and (B) average potential energy for protein G in the test simulation without any exchange between replicas.

Fig. 5. Protein G folding fraction at various temperatures obtained from simulations R1, T1, T2, T3, T4 and T5. Error bars represent the standard deviation about the mean values.

Fig. 6. Protein G folding fraction at various temperatures obtained from simulations R2, T2, T21, T22 and T23. Error bars represent the standard deviation about the mean values.

Fig. 7. Superposition of the lowest energy structures from the last 10ns simulations of R2 (Yellow) and T2 (Blue) on the crystal (Red) structure of protein G. The backbone C α RMSDs are 0.797 and 0.926 Å, respectively, with respect to the crystal structure. The plot was created with VMD program.²⁸

Fig. 8. Probability density distributions of potential energy sampling for protein G obtained by using REMD (dotted line) and TIGER2 (solid line) methods, respectively.

TABLE I

Summary of the protocols for the simulations of alanine dipeptide in SCP-ISM solvent.

Simulation	Temperature Levels (K)	Number of Replicas	Length of dynamics phases heating – sampling – quenching
REMD	300, 319, 340, 362, 385, 411, 437, 466, 496, 528, 563, 600	12	1 ps (Sampling only)
TA	300, 400, 500, 600	4	1 ps – 1 ps – 1.2 ps
TB	300, 377, 476, 600	4	1 ps – 1 ps – 1.2 ps
TC	300, 400, 500, 600	4	1 ps – 10 ps – 1.2 ps
TD	300, 400, 500, 600	4	1 ps – 1 ps – 10 ps
TE	300, 343, 386, 429, 472, 515, 558, 600	8	1 ps – 1 ps – 1.2 ps
TF	300, 500, 700, 900	4	1 ps – 1 ps – 1.2 ps
TG	300, 433, 566, 800	4	1 ps – 1 ps – 1.2 ps
TH	300, 466, 633, 700	4	1 ps – 1 ps – 1.2 ps
TI	300, 366, 433, 500	4	1 ps – 1 ps – 1.2 ps
TJ	300, 366, 433, 450	4	1 ps – 1 ps – 1.2 ps
TK	300, 366, 433, 400	4	1 ps – 1 ps – 1.2 ps

TABLE II

Summary of the protocols for the simulations of protein G in SCP-ISM solvent.

Simulation	Temperature Levels (K)	Number of Replicas	Length of dynamics phases heating – sampling – quenching
R1	300 to 580	24	1 ps (Sampling only)
T1	300 to 520	8	1 ps – 9 ps – 2 ps
T2	300 to 552	8	1 ps – 9 ps – 2 ps
T3	300 to 580	8	1 ps – 9 ps – 2 ps
T4	300 to 610	8	1 ps – 9 ps – 2 ps
T5	300 to 640	8	1 ps – 9 ps – 2 ps
R2	300 to 552	24	1 ps (Sampling only)
T21	300 to 552	8	1 ps – 1 ps – 2 ps
T22	300 to 552	8	1 ps – 6 ps – 2 ps
T23	300 to 552	8	1 ps – 12 ps – 2 ps

TABLE III

Population ratios and corresponding standard deviations (SD) calculated over the last 5 ns of REMD and TIGER2 trajectories for the C_7^{eq}/C_5 , α_R/β_2 , C_7^{ax} , and α_L regions.

Method	C_7^{eq}/C_5	SD	α_R/β_2	SD	C_7^{ax}	SD	α_L	SD
SCP-ISM ^a	0.873		0.117		0.001		0.008	
REMD	0.873	8.4E-4	0.121	7.2E-4	0.001	1.7E-4	0.005	2.1E-4
TA	0.848	1.5E-3	0.143	1.3E-3	0.002	6.3E-5	0.007	3.4E-4
TB	0.850	1.5E-3	0.143	1.5E-3	0.001	1.7E-4	0.007	5.5E-4
TC	0.835	1.6E-3	0.155	1.8E-3	0.003	1.6E-4	0.006	1.6E-4
TD	0.881	3.0E-3	0.114	2.6E-3	0.001	2.5E-4	0.004	2.8E-4
TE	0.848	1.4E-3	0.142	1.4E-3	0.003	8.9E-5	0.007	2.3E-4
TF	0.800	1.0E-3	0.179	8.6E-4	0.007	1.2E-4	0.014	3.8E-4
TG	0.817	3.0E-3	0.170	3.0E-3	0.006	3.4E-3	0.013	3.1E-4
TH	0.821	2.0E-3	0.168	2.4E-4	0.003	2.8E-4	0.008	6.3E-4
TI	0.868	4.3E-3	0.126	3.7E-3	0.002	2.6E-4	0.004	3.8E-4
TJ	0.872	2.5E-3	0.122	2.2E-3	0.001	6.0E-5	0.005	4.3E-4
TK	0.894	1.4E-3	0.098	1.4E3	0.0009	2.1E-5	0.007	1.6E-4

^aThe fractions were calculated based on energy minima in SCP-ISM (Table I in Ref. ¹⁷).

TABLE IV

Average C_{α} RMSDs and folding fractions for various simulations of protein G in SCP-ISM solvent at the 300 K baseline temperature.

Property	R1	T1	T2	T3	T4	T5	R2	T21	T22	T23
C_{α} RMSD (Å)	0.849	0.980	0.986	1.118	1.126	1.144	0.849	0.947	0.933	1.071
Standard Dev.	0.192	0.397	0.330	0.411	0.481	0.501	0.192	0.252	0.230	0.488
Folding Fraction	0.809	0.799	0.803	0.806	0.792	0.796	0.809	0.801	0.807	0.799
Standard Dev.	0.071	0.064	0.072	0.068	0.065	0.060	0.071	0.054	0.071	0.072

TABLE V

Statistical Analysis of Potential Energy Probability Distributions for protein G.

Method	τ^d (ps)	N sample	Mean (kcal/mol)	SD	p (t-test) for means	p (F-test) for variance
REMD	12	4170	-5821.5	19.3	0.011	0.053
TIGER2	86	580	-5818.2	22.1		

^a τ is the autocorrelation time that approximately measures the time required to obtain independent samples from the sequentially sampled trajectories. τ is calculated by a web calculator given in Ref ²⁶.