JCTC Journal of Chemical Theory and Computation

Bad Seeds Sprout Perilous Dynamics: Stochastic Thermostat Induced Trajectory Synchronization in Biomolecules

Daniel J. Sindhikara,[†] Seonah Kim,[§] Arthur F. Voter,[||] and Adrian E. Roitberg*,[‡]

Quantum Theory Project and Departments of Physics and Chemistry, University of Florida, Gainesville, Florida 32611, Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, and Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Received December 23, 2008

Abstract: Molecular dynamics simulations starting from different initial conditions are commonly used to mimic the behavior of an experimental ensemble. We show in this article that when a Langevin thermostat is used to maintain constant temperature during such simulations, extreme care must be taken when choosing the random number seeds to prevent statistical correlation among the MD trajectories. While recent studies have shown that stochastically thermostatted trajectories evolving within a single potential basin with identical random number seeds tend to synchronize, we show that there is a synchronization effect even for complex, biologically relevant systems. We demonstrate this effect in simulations of alanine trimer and pentamer and in a simulation of a temperature-jump experiment for peptide folding of a 14-residue peptide. Even in replica-exchange simulations, in which the trajectories are at different temperatures, we find partial synchronization occurring when the same random number seed is employed. We explain this by extending the recent derivation of the synchronization effect for two trajectories in a harmonic well to the case in which the trajectories are at two different temperatures. Our results suggest several ways in which mishandling selection of a pseudorandom number generator initial seed can lead to corruption of simulation data. Simulators can fall into this trap in simple situations such as neglecting to specifically indicate different random seeds in either parallel or sequential restart simulations, utilizing a simulation package with a weak pseudorandom number generator, or using an advanced simulation algorithm that has not been programmed to distribute initial seeds.

1. Introduction

The use of molecular-dynamics (MD) simulations is widespread across various fields.[1] It is often useful to perform MD simulations in the canonical ensemble (NVT) in order to compare with experimental processes. In such circumstances, a thermostat is used to regulate temperature. Many types of thermostats are commonly employed, including Berendsen,[2] Nose-Hoover,[3–5] Andersen,[6] and Langevin.[7] Andersen and Langevin are stochastic in nature -- including random forces to mimic the effect of solvent collisions. Both can be proven[8,9] to give true canonical sampling. The characteristics of simulations using these stochastic thermostats, especially the commonly used Langevin thermostat, are the primary focus of this manuscript. An excellent review of characteristics of various thermostats can be found in ref 9.

Biomolecular simulations run at constant energy or constant temperature using a nonstochastic thermostat, such

* Corresponding author e-mail: roitberg@qtp.ufl.edu.

[†] Quantum Theory Project and Department of Physics, University of Florida.

[‡] Quantum Theory Project and Department of Chemistry, University of Florida.

[§] University of California.

[||] Los Alamos National Laboratory.

10.1021/ct800573m CCC: $40.75
Published on Web 04/28/2009

as Berendsen or Nose-Hoover, are considered to be chaotic and thus extremely sensitive to initial conditions. Braxenthaler et al. found that for a peptide system, root-mean-square deviations between two simulations can grow from only 0.001 Å to roughly 1 Å after only one or two picoseconds.[10] One might then expect that a stochastic thermostat, due to its use of random forces, would increase this divergent behavior. We will show in this article that under certain conditions this assumption is untrue, and failure to recognize this can lead to incorrect simulation results.

Stochastic thermostats use sequences of pseudorandom numbers to mimic the random solvent impacts. Pseudorandom number generators (PRNGs) are deterministic; given an initial 'seed', they always produce the same sequence of numbers. This trait is useful in that it allows for reproducibility of results when needed. Thus, if multiple simulations are run with different initial conditions, $(\vec{x},\vec{v})$, but identical random seeds, their random forces will remain the same for all simulations for their full length. Uberuaga et al. recently showed[11] that for the case of dynamics in a simple harmonic potential basin, Langevin (or Andersen) trajectories with identical seeds are *driven* to synchronize — the difference in both $\vec{x}$ and $\vec{v}$ between two trajectories decay exponentially, ultimately leading to a single trajectory path, no matter how different the initial conditions were. More generally, they argued that for a convex potential basin or even more general confining potentials a similar synchronization effect should occur. This behavior is consistent with rigorous mathematical results stating that under fairly general conditions, when the largest Lyapunov exponent is negative, trajectories starting from any ensemble of initial conditions are attracted to "random sinks".[12] Other groups have also observed this synchronization effect in model systems.[13–17] Cerutti et al. recently described[18] a situation where rapid restarts with the same seed of a single MD simulation in Langevin or Andersen Dynamics would result in a residual (nonzero average) stochastic force. We note, though, that this residual force is not the same as the synchronization effect.

In this paper, we show that synchronization can also occur in the much more complex potential energy landscapes of biomolecular systems. The potential energy surfaces for these systems typically consist of a complex network of many local minima separated by negatively curved saddle regions. Nonetheless, we observe that use of the same random number seed for different trajectories leads to strongly biased behavior due to partial synchronization occurring on the typical simulation time scale. We show this for Langevin dynamics of small peptides (alanine trimer and pentamer) and a simulation of a temperature-jump experiment for peptide folding of a 14-residue peptide.

We also explore the possibility that trajectories with identical seeds at different temperatures can synchronize, extending the harmonic-well derivation of Uberuaga et al. to the case in which the two trajectories have different temperatures. It will be shown that there is a well-defined synchronization of the coordinates, and hence a strong correlation between the trajectories exists. This multiple temperature synchronization has important implications for the method of replica exchange molecular dynamics among temperatures (T-REMD)[19,20] and variants thereof.[21–23] T-REMD is an enhanced sampling algorithm commonly used in the biomolecular simulation community. If the same random number seed is used for all the replicas, correlations among the different trajectories will contaminate the statistics of the study.

We first review the derivation of the driven synchronization for a pair of trajectories in a harmonic oscillator and then extend it to the case of two trajectories at different temperatures. We then present results from various peptide simulations in which the synchronization effect causes a bias in the results, culminating with the case of the T-REMD simulations of alanine trimer. We close with a discussion of the importance of understanding, and avoiding, the statistical contamination that can be caused by this synchronization effect in biomolecular simulations, and we identify and explain several common situations in which a simulator may unknowingly initiate multiple trajectories with the same random number seed, including neglecting to distribute random seeds for simultaneous simulations or sequential restart simulations or using programs that do not enforce distributions of random seeds.

2. Theory

2.1. Single Temperature Langevin Synchronization. In Langevin dynamics, particles are propagated based on the Langevin equation of motion:

$$m_i\ddot{\vec{r}}_i = -\vec{\nabla}V(\vec{r}_i) - \gamma m_i\dot{\vec{r}}_i + \vec{A}(\gamma, T, \upsilon) \qquad (1)$$

Here m_i, $\ddot{\vec{r}}_i$, $\dot{\vec{r}}_i$, and $\vec{r}_i$ are the mass, acceleration, velocity, and position of the i^{th} particle, respectively. $V(\vec{r}_i)$ is the potential energy determined by the force field. Equation 1 is essentially Newton's second law with two extra terms: a solvent drag force represented by $\gamma m_i\dot{\vec{r}}_i$ and a random force, $\vec{A}$, which obeys the fluctuation−dissipation theorem: $\langle A_i(t)A_j(t+\Delta t)\rangle = 2m\gamma k_bT\delta(\Delta t)\delta_{ij}$. Here the average, $\langle\rangle$, is over time, k_b is the Boltzmann constant, and $\delta(\Delta t)$ and δ_{ij} represent the Dirac and Kronecker delta functions, respectively. The magnitude and direction of A are based on a pseudorandom number, υ, and a probability distribution based on the temperature and heat bath coupling strength, γ, also known as the collision frequency, or friction. It has been shown[11] that for two particles in the same harmonic well with the same random number sequence, their trajectories are driven to synchronize. That is, for the i^{th} degree of freedom in a single dimension, $\Delta x_i = x_i^a - x_i^b$, the difference between two trajectories a and b, tends to zero as time increases.

Let us first consider the differences (between trajectories a and b) in the instantaneous accelerations on each degree of freedom in the Langevin regime:

$$\Delta\ddot{x} = -(\partial V/\partial x_a - \partial V/\partial x_b)/m - \gamma\Delta\dot{x} + (A_a - A_b)/m \qquad (2)$$

where $\Delta\ddot{x} = \ddot{x}_a - \ddot{x}_b$ and $\Delta\dot{x} = \dot{x}_a - \dot{x}_b$. If we approximate the local potential region to be a harmonic well of index a or b, $\partial V/\partial x = m\omega_{a,b}^2x$, then the difference in accelerations becomes $\Delta\ddot{x} = -\omega_a^2(x_a - x_b\omega_b^2/\omega_a^2) - \gamma\Delta\dot{x} + (A_a(t) - A_b(t))/m$. If the same pseudorandom number initial seed is used

for both simulations, the difference in random forces becomes zero at every step. What is then left is $\Delta\ddot{x} = -\omega_a^2(x_a - x_b\omega_b^2/\omega_a^2) - \gamma\Delta\dot{x}$. If the basins have identical curvature (or if the two simulations are in the same basin), this reduces to $\Delta\ddot{x} = -\omega^2\Delta x - \gamma\Delta\dot{x}$, which is the equation of a damped harmonic oscillator. For long times, the difference in the coordinates becomes zero; i.e. the trajectories 'synchronize'.

Realistic systems are more complicated than a simple harmonic oscillator. For these systems, synchronization rates are disrupted by a passage of particles through regions of negative curvature.[11] Despite this, any bound system must inevitably exist in some greater basin. Thus, synchronization may eventually occur for almost any simulated system unless a different initial seed for the PRNG is used.

Even before complete synchronization occurs for these many-minima systems, partial synchronization may occur for particles in basins of similar shape (this is where the shift from the identical harmonic well solution is small). As we will show, partial synchronization between trajectories does indeed take place on the time-scale of realistic simulations of peptide systems and has an effect strong enough to corrupt the results.

2.2. Multiple Temperature Langevin Synchronization. Though constant temperature simulations are both useful and commonplace, it is often necessary to use advanced simulation algorithms that utilize multiple temperatures for better sampling. One such enhanced sampling method that employs multiple-temperature simulation is parallel tempering[20] (PT), also known as replica exchange molecular dynamics among temperatures[19] (T-REMD). In this approach, replicas of the same molecule are simulated in parallel at different temperatures, most of which are above physiological temperatures. Periodically, a Metropolis-style Monte Carlo swap is attempted between conformations at different temperatures. A discussion of many generalized ensemble algorithms for enhanced sampling including T-REMD can be found in a review by Okamoto.[24]

The analytical derivation shown in the previous section suggests that single temperature simulations should synchronize. One might expect that the added complication of multiple temperatures might diminish synchronization. We show that even when using multiple temperatures, this effect is present and of consequence. We follow the same scheme and notation as we did in the single temperature derivation with two trajectories a and b, employing the same sequence of random numbers, but at two temperatures, T_a and T_b. Following the fluctuation dissipation theorem, we see that the stochastic forces are related by a simple scaling factor, $A_i^b(t) = cA_i^a(t)$, where $c = \sqrt{T_b/T_a}$. The equation for the difference in the i^{th} degree of freedom between the two runs is then given by

$$\Delta\ddot{x} = -(\partial V/\partial x_a - \partial V/\partial x_b)/m - \gamma\Delta\dot{x} + (1 - c)A_a/m \tag{3}$$

where $\Delta\ddot{x} = \ddot{x}_a - \ddot{x}_b$. As before, for a single harmonic potential with one degree of freedom, $\partial V/\partial x = m\omega^2 x$, and this equation becomes

$$\Delta\ddot{x} = -\omega^2\Delta x - \gamma\Delta\dot{x} + (1 - c)A_a/m \tag{4}$$

When the two temperatures are the same, $c = 1$, and this simplifies to the damped harmonic oscillator equation. When the temperatures differ by a small amount, c is close to unity, and the equation of motion for the difference between the two trajectories, Δx, becomes a Langevin equation with only a small noise term.

We can be more specific about how the two trajectories differ for any two temperatures by using a simple rescaling argument. Equation 3 can be modified to elucidate this behavior:

$$c\ddot{x}_a - \ddot{x}_b = -(c\partial V/\partial x_a - \partial V/\partial x_b)/m - \gamma(c\dot{x}_a - \dot{x}_b) + (c - c)A_a/m \tag{5}$$

The last term (the noise) vanishes, and, for a harmonic oscillator, the linearity allows us to simplify this to

$$c\ddot{x}_a - \ddot{x}_b = -\omega^2(cx_a - x_b) - \gamma(c\dot{x}_a - \dot{x}_b) \tag{6}$$

Defining $y = cx_a - x_b$, this becomes $\ddot{y} = -\omega^2 y - \gamma\dot{y}$, which is again simply a damped harmonic oscillator. Thus the trajectory for cx_a synchronizes to the trajectory for x_b; i.e. the trajectories at the two temperatures are related by a simple rescaling by c. Knowing the trajectory at any temperature is sufficient to specify exactly what the trajectory at any other temperature will be, once they have run long enough to be synchronized. Since the trajectories are now strongly correlated, the effective sampling will be greatly diminished.

3. Methods

3.1. Single Temperature Simulations. Single temperature MD simulations were performed on three peptides: trialanine (Ac-AAA-NH_2, ALA_3), penta-alanine (Ac-AAAAA-NH_2, ALA_5), and a 14-residue peptide (Ac-YGSPEAAA-KAAAA-r-NH_2, where r represents D-Arg). All simulations were performed using the AMBER 9 molecular simulation suite[25] with Langevin Dynamics in generalized-Born implicit solvent. All simulations were performed in AMBER 9 with the AMBER ff99SB force field,[26] and the Generalized Born implicit solvent model $GB^{(OBC)}$ was used to model the water environment in all our calculations.[27] The SHAKE algorithm[28] was used to constrain the bonds connecting hydrogen and heavy atoms in all the simulations. For the polyalanine peptides, a 1 fs integration time step was used, and each calculation was performed in the canonical ensemble (NVT) with a Langevin thermostat, using collision frequencies, γ, of 1 ps^{-1} or 50 ps^{-1} (as specified). For the 14-residue peptide, a 2 fs time step was used with a Langevin collision frequency of 5 ps^{-1}. For the 14-residue peptide, 1200 initial coordinate sets were taken from previously equilibrated run for the DS and SS production runs, which were run at an increased temperature of 372 K in order to simulate a Temperature-jump (T-jump) experiment.

To demonstrate single temperature trajectory synchronization, multiple simulations were run, all with different initial coordinates, using either the same initial random seed (SS) or different seeds (DS). 100 simulations each were run for 1 ns for polyalanines, 1200 simulations for 5 ns each for the 14-residue peptide.

3.2. Multiple Temperature Simulations. Synchronization across multiple temperatures is demonstrated by use of

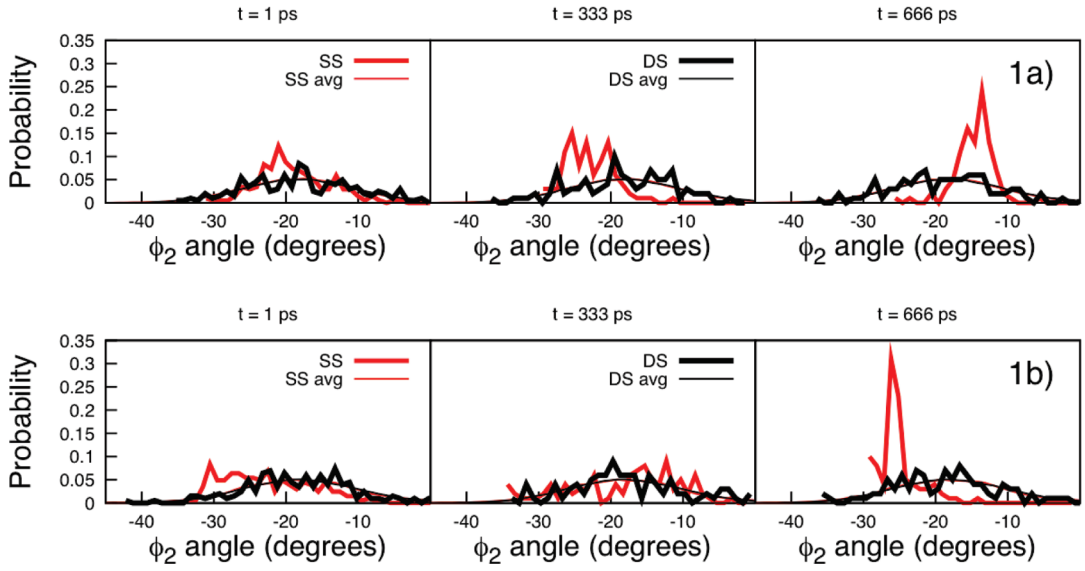

Figure 1. **a,b.** Probability distributions of the dihedral angle φ_2 across sets of 100 simulations for alanine polymer simulations. Data from SS simulations are shown in red, DS in black. Parts Figures a and b show ALA3 with $\gamma = 1$ ps^{-1} and $\gamma = 50$ ps^{-1}, respectively.

T-REMD simulation. Both a DS and SS 100-ns T-REMD simulations were performed using the AMBER9 package with Langevin dynamics in implicit solvent GB model. A Langevin thermostat was used with a collision frequency of 50 ps^{-1}. The SHAKE algorithm was employed allowing use of a 2 fs time step. Both systems utilized 6 replicas and started from the same initial configurations. The replica temperatures were spaced geometrically: 251.8 K, 300.0 K, 357.5 K, 426.0 K, 507.6 K, and 604.8 K. Exchanges were attempted every 500 steps (1 ps). The T-REMD code was altered to keep the random number sequences synchronized for all replicas for the SS simulation. Snapshots were recorded every 25 ps.

4. Results and Discussion

4.1. Single Temperature Synchronization. For both the ALA_3 and ALA_5 simulations, the dihedral angle of the second residue (ϕ_2) was measured versus time as an internal unit. We could have chosen any other set of coordinates to illustrate the synchronization effect.

Figure 1a,b shows probability distributions of ϕ_2 across the sets of 100 simulations for ALA_3 for a collision frequency of 1 ps^{-1} or 50 ps^{-1}, respectively. According to the harmonic theory,[11] in the low collision frequency regime, increasing the frequency, γ, should yield faster synchronization of trajectories. Histograms are shown at arbitrary intervals of 1 ps, 333 ps, and 666 ps into the trajectory. For comparison, probability distributions across the entire trajectories are shown in thin lines though since nearly identical, they are virtually indistinguishable.

Regardless of the random seeds (SS or DS), after a very long time, the distributions of ϕ_2 angles (very thin lines) are the same in each case. If synchronization is not present (as in the case of DS), the distribution of ϕ_2 among the 100 trajectories at a given time should be similar to the longer time average population. Conversely, if the same seeds (SS) are used for all 100 trajectories, the system behaves very differently. For instance, at 666 ps, for parts a and b of Figure 1, a large number of trajectories have very similar values of ϕ_2, as represented by a sharply peaked histogram. Figure 1 clearly shows that even in complex systems, the effect of synchronization is observable. The behavior of a coordinate for a set of SS simulations is similar to a swarm that expands and tightens as if compelled to come together. A movie of the Ramachandran plot (in beta/ppII region) of the first residue of ALA_3 with $\gamma = 50$ ps^{-1} demonstrates this behavior (Supporting Information).

To quantify synchronization among the entire set of 100 simulations, we used a measure of how many of them were similar to each other at any particular time. This was done by histogramming a physical observable (again ϕ_2 in our case) and counting how many of the 100 simulations reside in the histogram bin with maximum population. This is equivalent to the maximum height in Figure 1. If all 100 systems were perfectly synchronized, the highest fractional population (HFP) would be exactly one. Conversely, for completely unsynchronized systems, the HFP should stay relatively constant (and small for small bin sizes). For our simulations, the frame-by-frame ϕ_2 population was binned in narrow 2-degree windows. Figure 2a,b displays the time series of the fractional population of the most popular bin (HFP) for both ALA_3 and ALA_5, versus time. The sideplots show the probability distributions of those HFP time series. Included in the SI are additional HFP time series and probability distributions.

From the figures it is clear that the DS simulations have a small and relatively constant HFP. This is expected since the trajectories evolve independently from each other (there are not many simulations where the ϕ_2 angles are the same). In contrast, the SS simulations HFPs are much larger than for the DS case and in some cases achieve extremely high values. For instance, at 628 ps, a HFP value of 0.78 (Figure 2a) means that 78 of the 100 simulations have the same value of ϕ_2 (to within 2 degrees). The HFP difference between

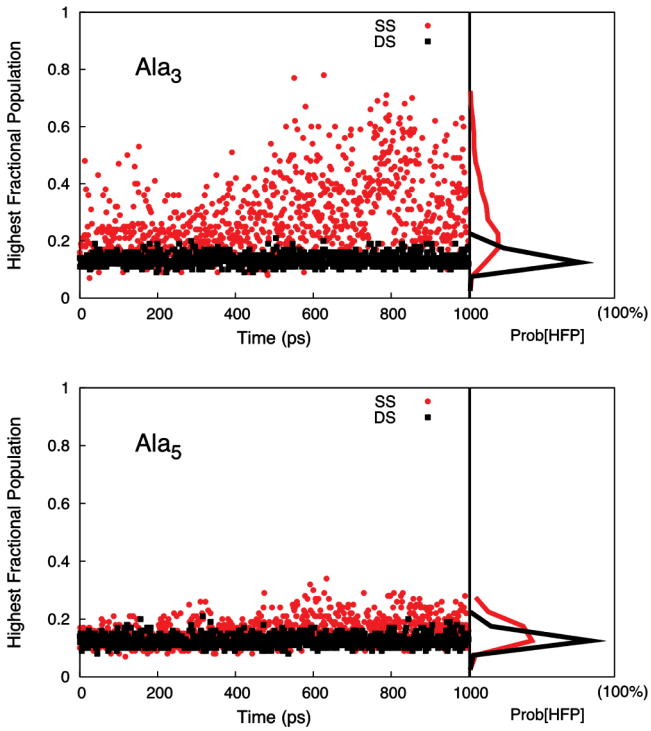

Figure 2. **a,b.** Highest fractional population (HFP) for φ_2 in alanine polymer simulations with a collision frequency, γ, of 50 ps^{-1}. Parts a and b show HFP for ALA_3 and ALA_5, respectively. Sideplots are the appropriate probability distribution of HFPs.

DS and SS is significantly smaller for the ALA_5 possibly because the synchronization is slower for larger systems. However, large homogeneous systems, such as those with explicit solvent, may still synchronize quickly.

To visualize the system as a whole, a time snapshot from the ALA_3 simulations was chosen and shown with all 100 simulation frames superimposed (see Figure 3, created in VMD[29]). In this figure, the red and blue spheres represent atom locations in the SS and DS simulations respectively. A stick representation is shown in gray as a visual aid. There is the same number of red spheres as blue (100 per atom). The figure shows fluctuation among the DS snapshots is much greater than that for SS; since the SS simulations are partially synchronized, the atomic positions are more condensed than they should be otherwise.

When simulating a complex system, it is often useful to utilize many simulations to reduce the error. The average over many simulations of a property, A, at time t,$\langle A\rangle_{sim}(t)$, is likely to be closer to the true average, $\bar{A}$, than the value of a property of a single simulation, $A(t)$ since value of A will fluctuate naturally in time. If the simulations are uncorrelated with each other, it can be shown that the standard deviation over time of these averages over simulations, $\sigma_{time}(\langle A\rangle_{sim})$, is less than the standard deviation over time of a single simulation $\sigma_{time}(A)$ (which is caused by the natural fluctuations):

$$\sigma_{time}(\langle A\rangle_{sim}) = \sigma_{time}(A)/\sqrt{N_{sim}} \tag{7}$$

Here, N_{sim} is the number of simulations. However, if the simulations are correlated with each other, the average over

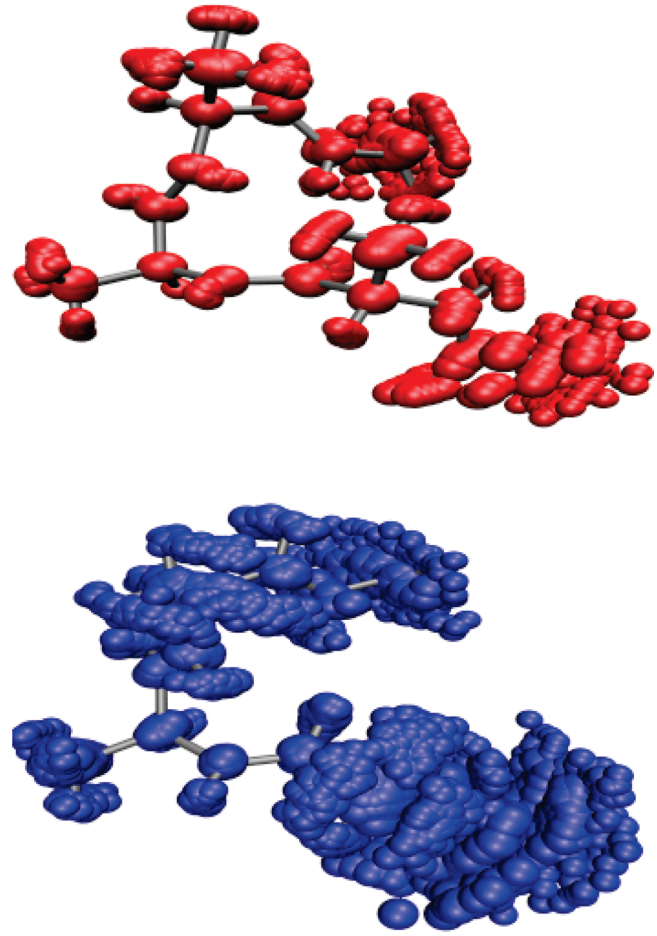

Figure 3. Sphere representations of simultaneous frames of 100 simulations at 836th ps of ALA_3. Red spheres represent atoms in SS simulations; blue spheres represent atoms in DS simulation.

simulations $\langle A\rangle_{sim}$ will fluctuate in time with greater amplitude, similar to that of a single simulation, $A(t)$. This has the same effect as reducing N_{sim}. Thus, according to eq 7, correlated simulations will have a higher standard deviation over time of the average over simulations, $\sigma_{time}(\langle A\rangle_{sim})$. We have presented above some arguments and results showing that many simulations run with the same initial random number generator seed will become somewhat synchronized over time. This effect will cause correlations between different simulations.[30]

We present here a striking example of this effect demonstrated in a simulation of temperature jump folding for a 14-residue peptide. This peptide was chosen since the T-jump kinetics were recently measured experimentally.[31] We have previously published a protocol for the simulation of that type of experiment.[32] In physical T-jump experiments, proteins are heated rapidly by a laser to observe folding events. Typically, a spectroscopic measure such as Trp-fluorescence or IR absorbance is used to follow the subsequent population relaxation. Unfortunately, in simulations, these phenomena are difficult to estimate. The expected CD spectra, rather, can be estimated in simulations based on the structure of the system. Since the CD signal at 222 nm is

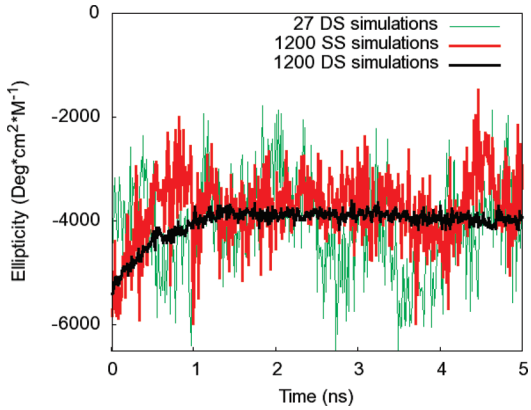

Figure 4. Average ellipticity vs time for a simulated T-jump experiment averaging over 1200 trajectories. SS simulations are shown in red, DS in black. The average of only 27 DS simulations is shown in green.

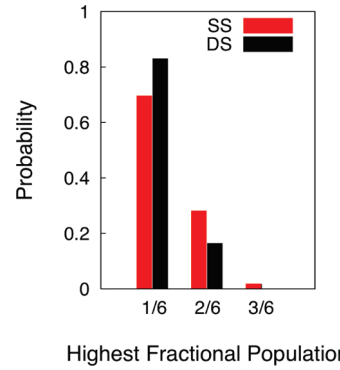

Figure 5. Histogram of highest fractional population for φ_2 angle in 2-degree bins for ALA_3 T-REMD simulation. SS shown in red, DS shown in black.

sometimes used to measure average ellipticity of the molecules, we focused on this measure to observe T-jump kinetics.

We computed ellipticity at 222 nm vs time averaged over 1200 simulations, using the method introduced by Sreerama and Woody.[33] Figure 4 shows the ellipticity vs time averaged over all 1200 simulations for SS (red line) and DS (black line), respectively. As can be seen in the figure, the signal-to-noise ratio is dramatically worse for the SS simulations.

The standard deviation for the last 2.5 ns of the T-jump simulation is 689 and 104 deg*cm^2*dM^{-1} for the SS and DS simulations, respectively. Thus, according to eq 7, the effective number of simulations is 44 ($\sqrt{689/104}$) times smaller for the SS than the DS. According to our preceding explanation, this means that the single seed runs act not like 1200 runs but as if only 27 (1200/44) truly independent runs. Thus, the average over 27 DS simulations (1200/44) should have a similar standard deviation to that of the 1200 SS simulations (thin green line in Figure 4). This effect is clearly shown in Figure 4 as a thin green line. We clarify that the single seed runs are not 'wrong' but that they produce overall fluctuations that are equivalent to a much smaller number of independent runs.

4.2. Multiple Temperature Synchronization. In the T-REMD simulations of alanine trimer, only six simultaneous simulations could be compared — one for each replica (as opposed to the 100 or 1200 simulations from the single temperature simulations above). Figure 5 shows the histogram of the highest fractional population of ϕ_2 bins (2 degree bins) for SS and DS simulations. The DS simulations (black bars) have a higher probability to have an HFP of 1/6, that is, that no two replicas have a ϕ_2 angle within 2 degrees of each other. The SS simulations are more likely to have two or three replicas with the same ϕ_2 angle. This indicates that some synchronization does occur between replicas. Not only such a simulation would be biased but also additional consequences for exchange probabilities may exist. Although T-REMD was used as an example, we expect synchronization to occur for any set of multiple temperature simulations.

4.3. Relevance of Synchronization. As evidenced by our results, thermostat induced trajectory synchronization biases results and should be avoided. Depending on the severity of the synchronization, the bias may or may not be obvious to the researcher. It is thus important to understand the nature of synchronization to be aware of situations where it might occur.

Synchronization occurs when there is an overlap of pseudorandom number sequences, and this is typically caused by using the same initial seed for multiple runs. This can happen inadvertently for many reasons. Some simulation programs use a default random seed. AMBER, for example, uses a default random seed if none is specified. Others may use a time-seeded PRNG, which, depending on the implementation, may give a high risk of giving identical seeds. For example, if the program uses a time seed connected with a clock that is discretized to the nearest second, then if many simulations are initiated simultaneously, there is a high probability that many or all will receive the same seed.

Quite often simulators restart simulations. If one restarts with the same parameters (including initial seed), then the simulations could become self-synchronized. Cerutti et al. recently reported[18] a different negative consequence of repeatedly restarting Langevin or Andersen MD runs with the same initial seed -- trajectory corruption caused by a nonzero average stochastic force. Additionally, coders of new methods, such as T-REMD, which string together MD segments, might unknowingly build a code that uses the same seed. As we have shown, even Langevin MD runs at different temperatures can become synchronized.

Furthermore, since PRNGs have an inherent period, MD runs which call the PRNG more than this amount will naturally repeat the sequence. Although advanced PRNGs such as the Marsaglia algorithm[34] have extremely long periods (2^{144} for AMBERs implementation), older PRNGs have much shorter periods. We highly recommend that simulators take note of the PRNG period of the program they are running.

5. Conclusion

We have shown that identical-noise synchronization effects, previously observed for relatively simple systems under the influence of a stochastic thermostat, can also occur in the much more complex systems typical of

biomolecular simulations. Even in the case of trajectories at different temperatures, harmonic analysis shows a special scaled synchronization will occur. We indeed found evidence of synchronization bias in a replica-exchange simulation. In a simulation study, this synchronization tendency, even if weak, will corrupt the statistical quality of the results and may even lead to incorrect conclusions about the qualitative behavior of the system. Using modern biomolecular simulation programs and methods, many ways exist in which one can inadvertently initiate trajectories with identical seeds. It is possible that many papers have already been published with data that are biased by synchronization. We advise that great care be taken to avoid this situation by meticulous preparation of seeds, and we suggest that authors may wish to state specifically whether different initial seeds have been used when their results are based on multiple trajectories with a stochastic thermostat.

Acknowledgment. The authors acknowledge the University of Florida High-Performance Computing Center for providing computational resources. Computational resources were also provided by Teragrid Grant No. TG-MCA05S010. Work at the University of Florida was funded by the National Science Foundation grant number CHE-0822-935. Work at Los Alamos National Laboratory (LANL) was supported by the United States Department of Energy (U.S. DOE) Office of Basic Energy Sciences, Materials Sciences and Engineering Division. LANL is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. DOE under Contract No. DE-AC52-06NA25396. The authors are grateful to Blas P. Uberuaga, Marian Anghel, Kevin Lin, and John Chodera for helpful discussions.

Supporting Information Available: Additional plots containing statistical data (HFP) and visualization (3 figures, 1 movie). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Tuckerman, M. E.; Martyna, G. J. Understanding modern molecular dynamics: Techniques and applications. *J. Phys. Chem. B* **2000**, *104* (2), 159–178.

(2) Berendsen, H. J. C.; Potsma, J. P. M.; van Gunsteren, W. F.; DiNola, A. D.; Haak, J. R. Molecular Dynamics with Coupling to and External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

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

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

(5) Evans, D. J.; Holian, B. L. The Nose-Hoover Thermostat. *J. Chem. Phys.* **1985**, *83* (8), 4069–4074.

(6) Andersen, H. C. Molecular-Dynamics Simulations at Constant Pressure and-or Temperature. *J. Chem. Phys.* **1980**, *72* (4), 2384–2393.

(7) Zwanzig, R. Nonlinear generalized Langevin equations. *J. Stat. Phys.* **1973**, *9* (3), 215–220.

(8) Andersen, H. C. Molecular Dynamics at Constant Pressure and/or Temperature. *J. Chem. Phys.* **1980**, *72*, 2384.

(9) Hunenberger, P. Thermostat algorithms for molecular dynamics simulations. In *Advanced Computer Simulation Approaches for Soft Matter Sciences I*; Springer: Berlin/Heidelberg, 2005; Vol. 173, pp 105−147.

(10) Braxenthaler, M.; Unger, R.; Auerbach, D.; Given, J. A.; Moult, J. Chaos in protein dynamics. *Proteins: Struct., Funct., Genet.* **1997**, *29* (4), 417–425.

(11) Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of trajectories in canonical molecular-dynamics simulations: Observation, explanation, and exploitation. *J. Chem. Phys.* **2004**, *120* (14), 6363–6374.

(12) Le Jan, Y. Equilibre statistique pour les produits de difféomorphismes aléatoires indépendants. *Annales de l'I.H.P. Probabilités et statistiques* **1987**, *23* (1), 111–120.

(13) Fahy, S.; Hamann, D. R. Transition from Chaotic to Non-chaotic Behavior in Randomly Driven Systems. *Phys. Rev. Lett.* **1992**, *69* (5), 761–764.

(14) Maritan, A.; Banavar, J. R. Chaos, Noise, and Synchronization. *Phys. Rev. Lett.* **1994**, *72* (10), 1451–1454.

(15) Lise, S.; Maritan, A.; Swift, M. R. Langevin equations coupled through correlated noises. *J. Phys. A: Math. Gen.* **1999**, *32* (28), 5251–5260.

(16) Ciesla, M.; Dias, S. P.; Longa, L.; Oliveira, F. A. Synchronization induced by Langevin dynamics. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2001**, *6306* (6), -.

(17) Longa, L.; Curado, E. M. F.; Oliveira, F. A. Roundoff-induced coalescence of chaotic trajectories. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **1996**, *54* (3), R2201−R2204.

(18) Cerutti, D. S.; Duke, R.; Freddolino, P. L.; Fan, H.; Lybrand, T. P. A Vulnerability in Popular Molecular Dynamics Packages Concerning Langevin and Andersen Dynamics. *J. Chem. Theory Comput.* **2008**, *4*, 1669–1680.

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

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

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

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

(23) Rick, S. W. Replica exchange with dynamical scaling. *J. Chem. Phys.* **2007**, *126* (5), 054102.

(24) Okamoto, Y. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graphics Modell.* **2004**, *22* (5), 425–439.

(25) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.

(26) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple amber force fields and development of improved protein backbone

parameters. *Proteins: Struct., Funct., Bioinf.* **2006**, *65* (3), 712–725.

(27) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.* **2004**, *55* (2), 383–394.

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

(29) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics Modell.* **1996**, *14* (1), 33–&.

(30) Friedberg, R.; Cameron, J. E. Test of Monte-Carlo Method - Fast Simulation of a Small Ising Lattice. *J. Chem. Phys.* **1970**, *52* (12), 6049–6058.

(31) Wang, T.; Du, D. G.; Gai, F. Helix-coil kinetics of two 14-residue peptides. *Chem. Phys. Lett.* **2003**, *370* (5−6), 842–848.

(32) Kim, S.; Roitberg, A. E. Simulating temperature jumps for protein folding. *J. Phys. Chem. B* **2008**, *112* (5), 1525−1532.

(33) Sreerama, N.; Woody, R. W., Computation and analysis of protein circular dichroism spectra. *Methods Enzymol* **2004**, *383*, 318−351.

(34) Marsaglia, G.; Narasimhan, B.; Zaman, A. A Random Number Generator for Pcs. *Comput. Phys. Commun.* **1990**, *60* (3), 345–349.

CT800573M