Monitoring Convergence of Molecular Simulations in the Presence of Kinetic Trapping

Kevin Wiehe[†] and Scott C. Schmidler[*,‡]

Department of Bioinformatics and Biostatistics, and Departments of Statistical Science and Computer Science, Program in Computational Biology and Bioinformatics, Program in Structural Biology and Biophysics, Duke University

E-mail: schmidler@stat.duke.edu

Phone: (919) 684-8064. Fax: (919) 684-8594

Abstract

Convergence to equilibrium is an essential requirement for molecular simulation output to be accurate and reproducible. Yet testing for convergence is challenging, especially in instances where kinetic trapping and the consequent plateauing of ensemble quantities can falsely suggest premature equilibrium. Here we introduce a quantitative method for monitoring convergence based on multiple independent simulations started from diverse initial configurations. Our approach addresses common pitfalls of convergence testing, and allows specification of convergence criteria at a hierarchy of resolutions. We demonstrate the utility of this approach in monitoring simulations of several peptide systems of varying sizes, and use it to quantify the efficiency improvements realized by advanced simulation methods.

---

[*]To whom correspondence should be addressed

[†]Department of Bioinformatics and Biostatistics, Duke University, Durham, NC 27708, United States

[‡]Departments of Statistical Science and Computer Science, Program in Computational Biology and Bioinformatics, Program in Structural Biology and Biophysics, Duke University, Durham, NC 27708-0251, United States

1   Introduction

Molecular simulation has become an essential tool for studying biomolecular systems at atomic resolution. Significant progress in simulation methodology and computing power has enabled simulations of increasingly large systems and long time scales.[1] However, the accuracy of these simulations in predicting physical observables remains variable, and relies on both accurate energy functions and effective conformational sampling. Conformational sampling is particularly important because without adequate ensemble averaging, accuracy of potentials cannot be reliably evaluated against experimental data.[2] Adequate conformational sampling is difficult because polypeptides have multidimensional energy surfaces characterized by many local minima separated by high free energy barriers. These rugged energy landscapes can lead to kinetic trapping, with sampling restricted to localized "metastable" regions. To escape these kinetic traps simulations must often run for very long times; a key challenge is to determine when a simulation has run "long enough" to reach equilibrium and control statistical uncertainty.

In the literature, equilibration is often reported based on stabilization or plateauing of one or a few observable quantities over a single simulation run, and if these "summary" quantities (e.g. energy, radius of gyration, RMSD from initial conformation) appear stable, the simulation is deemed equilibrated.[3–5] However, such heuristics can fail in two ways: the summary quantities monitored are one-dimensional projections of a high-dimensional system, and may converge even while other quantities of interest remain unconverged; alternatively *all* quantities may falsely appear converged if the simulation remains stuck in a metastable region with large escape time. Such failures lead to premature termination of the simulation and incorrect results.

Several methods for monitoring equilibration have appeared in the literature, but none adequately address both issues. A natural approach to avoiding reliance on scalar summaries is to partition the conformational space using clustering. The number[6] or better the frequencies[7] of clusters then provide a summary of the entire ensemble, convergence of which can be measured by statistical methods.[8,9] However, this does not address the second problem; comparison of cluster populations within a single simulation[7] remains potentially misleading in the presence of kinetic

trapping. Figure 1a shows the use of a single simulation run to assess convergence of replica exchange molecular dynamics (REMD) simulations of trpzip2, applied to two different simulations initialized in extended and β conformations, respectively. In either case, a comparison of the cluster frequencies between the first and second halves of the simulation would falsely suggest that trpzip2 has converged by 50 ns. However, the frequencies of the respective individual runs are markedly different; only by comparing *between* the two independent simulations can we see that the cluster frequencies differ significantly, indicating that at least one of the runs must be kinetically trapped. In contrast, the diagnostic in this paper (see Methods) clearly identifies the problem (Figure 1b), and shows that convergence even at low resolution takes more than 200ns.
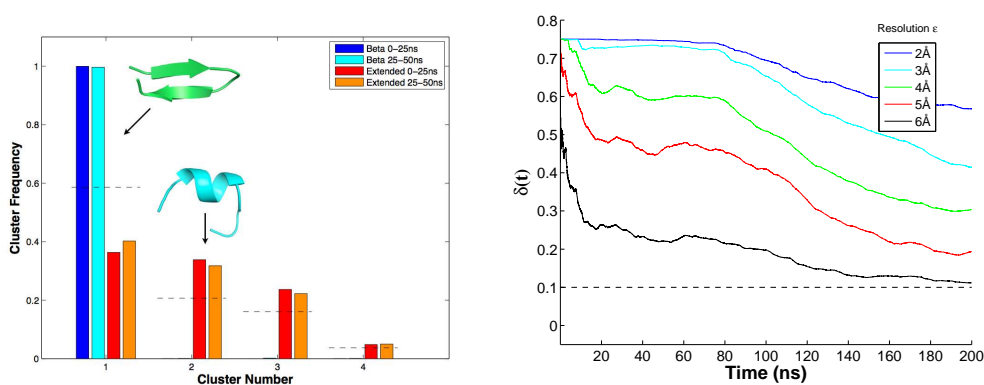

Figure 1: a) Cluster populations for first and second halves of two independent 50 ns trpzip2 simulations initialized from extended (red, orange) and β-hairpin (blue, cyan bars) conformations, respectively. (Four clusters shown cover 99% of ensemble at 5 Å Cα RMSD.) Although first and second halves of each simulation agree, the two simulations differ from each other and from true populations (dashed line, from 4 independent 200 ns simulations). b) Analysis of trpzip2 simulations using the diagnostic proposed here; shown are upper 95% confidence intervals at various structural resolutions. Kinetic trapping is clearly identified, even at low resolutions.

To address the kinetic trapping problem, we have previously adapted techniques from statistics[2,10], employing multiple independent simulations initialized at distinct points dispersed widely throughout the configuration space. Techniques from analysis of variance (ANOVA) are then applied to compare the equilibrium distribution of observables obtained from the different chains; the ratio of the between-chain and within-chain variances approaches one from above as the chains approach equilibrium.[10] In principle this strategy has the power to effectively identify kinetic trap-

ping problems; however, it is limited by the scalar summary quantities selected for monitoring. Figure 2 shows this approach[2] applied to a 50ns REMD simulation of the MPER peptide of HIV gp41 (see Sec. Section 3). Convergence of several commonly monitored quantities is shown. Overall helicity appears to converge at 17 ns, but monitoring energy indicates that the simulation does not converge until 40ns. Replacing overall helicity with helicities of individual residues gives finer detail, and we see that while residue 12 converges in 27 ns, residue 7 has not converged even by 50ns. Monitoring overall helicity, or other one-dimensional projects of the conformation space, is simply too coarse. (The marginal distributions may converge even when the joint has not yet converged.)
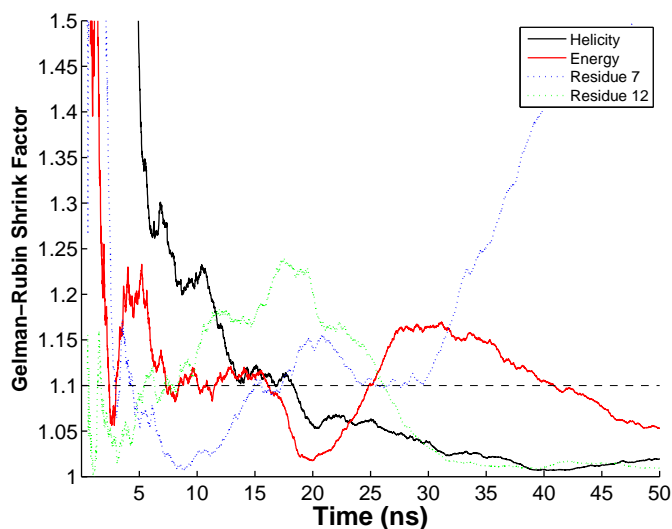

Figure 2: Convergence of scalar summaries in MPER REMD simulation, measured by the shrink factor[2,10] (upper 95% interval). Quantities shown are overall helicity (black), potential energy (red), and residue-specific helicity for positions 7 (blue) and 12 (green). Coarse summaries (helicity, energy) converge at different times and significantly faster than individual residues, making them unsuitable for determining overall convergence.

The approach described in this paper uses ideas closely related to several of these existing methods; in particular our diagnostic combines the advantages of the multiple-chain approach with the benefits of structural clustering, addressing all the pitfalls described above. To our knowledge it is the only method which combines all these key aspects into a single convergence diagnostic.

2   Methods

2.1   Partitioning Conformational Space

Rather than monitor convergence of scalar quantities, which are one-dimensional projections and therefore incompletely characterize conformational space, we wish to monitor the convergence of the entire conformational ensemble directly. However, as conformational space is infinite, we discretize into a finite number of partitions, and represent the ensemble by the partition populations (frequencies). Convergence of the ensemble is then monitored by measuring agreement of these populations across independent simulations.

However, simple uniform volume partitioning would produce an unmanageably large number of elements, most with infinitesimal target frequencies. Instead, we use clustering of the simulated configurations to produce the partition. This provides an effective means of adaptively partitioning the space, achieving finer resolution in precisely those low-energy regions which contribute most to the configuration-space integral as the simulation converges to the target Boltzmann ensemble.

Clustering is performed by C-α RMSD distance (minimal root-mean-square deviation under optimal superposition) on the pooled set of configurations from multiple independent simulations started from over-dispersed initial conformations. Pooling ensures that a region of conformational space sampled by *any* of the simulations will be represented in the ensemble comparison. We cluster the conformations using the following iterative covering method (a similar method has been used previously[7]): First the lowest energy structure is designated as a new cluster center, and all structures within threshold distance ε assigned as members of that cluster. The process is then applied iteratively to all remaining unassigned configurations, until none remain. This approach ensures a constant width of each partition (unlike say hierarchical clustering). The use of energy to define cluster centers means that each cluster has the rough interpretation as a basin in the energy landscape.

2.2   Initialization of multiple independent simulations

A key aspect of our approach is the use of multiple, *independent* simulations started from diverse initial conditions. It is comparison between these independently generated ensembles that allows us to identify kinetic trapping. Since each simulation must have the same equilibrium distribution, any differences between them which exceed those expected by simple statistical fluctuation indicate lack of equilibration.

Choosing the "right" number of runs trades off computational resources (but not time, if multiple processors are available) against the chance of mistakenly initializing all simulations within in the same basin of attraction. If the initial configurations are generated randomly and the probability of the largest basin is p, this has probability approximately p^M for M runs. All of the peptide simulations in the Examples section use $M = 4$ independent runs started from distinct areas of conformational space, typically including α, β, extended, and random configurations.

2.3   Convergence Diagnostic

Once the trajectories have been pooled and partitioned, we quantify the distance between the independent cluster population vectors using the L_1 distance:

$$\delta(t) = \frac{1}{2} \max_j \sum_i^N \left| O_{ij}(t) - \bar{O}_i(t) \right| \tag{1}$$

where $O_{ij}(t)$ is the observed frequency of the ith cluster of the jth simulation at time t, $\bar{O}_i(t) = \frac{1}{M} \sum_j^M O_{ij}(t)$, and M and N are the numbers of simulations and clusters, respectively.

Early in the simulation, the chains will heavily populate regions near their starting configuration and the distance between any one chain to the average cluster populations of all chains will be large. As the chains converge to their equilibrium distribution, the cluster populations will be similar across all chains and the convergence diagnostic $\delta(t)$ will become small. This distance is plotted against time, and the system considered equilibrated when it decreases below a pre-chosen threshold and remains below. We adopt a 0.1 threshold; standard L_1 (total variation) arguments

then imply bounds on errors in ensemble averages.

Because the populations being compared arise from finite sampling, statistical fluctuation will occur. We apply bootstrapping (sampling with replacement from the independent chains) to obtain an approximate upper 95% confidence limit for $\delta(t)$, and it is this quantity which we require to decrease below 0.1.

2.4 Cluster Convergence Structural Resolution

The choice of distance cutoff ε used in partitioning the conformational space affects the convergence time identified for a given simulation. The higher resolution the clustering, the finer the partitioning of conformational space and the longer a simulation may need to run for the cluster frequencies to equilibrate. To ensure monotonicity in convergence times we apply the iterative covering described above at the lowest resolution, and then recursively to sub-partition the clusters to obtain higher resolution clusters. This also enables rough extrapolation of convergence times at low resolution to higher resolutions.

2.5 Preventing Underestimation of Convergence Time

The quantity $\delta(t)$ is best thought of as an *estimated* distance based on the sample history so far. As such, it is subject to fluctuation and does not always decrease monotonically in time. In particular, when one of the independent simulations discovers an area of conformational space that the others have not yet explored, the distance can suddenly increase. Occasionally this can lead to a re-crossing of the designated convergence threshold. Thus, identifying the first threshold crossing as convergence may lead to underestimation of convergence time. To prevent this, we employ the following simple heuristic: if the most recent threshold crossing occurred at time τ, convergence is identified when $\delta(t)$ has remained below the threshold until 2τ. It is important to note that we can never be certain that re-crossing wouldn't occur at longer time scales; essentially we cannot guarantee that *all* chains are not trapped in some common but metastable region. However the use of multiple diverse initial configurations minimize the chances of this, and in practice this heuristic

works well. Figure 3b shows how the application of this heuristic prevents the underestimation of the convergence time for alanine tripeptide.

2.6   Simulation Details

All simulations in this study were run using the AMBER10 molecular dynamics software[11] with the AMBER ff03 force field[12] and generalized Born/surface area (GB/SA) implicit solvent model.[13,14] A time step of 2 fs was used and all bonds involving hydrogen were constrained with SHAKE[15] with a 0.00001 Å tolerance. Temperature was controlled by Langevin dynamics with a collision frequency of 1 ps^{-1}. Non-bonded interactions were calculated within an 8.06 Å cutoff. Electrostatic 1-4 interactions (atom pairs separated by 3 bonds) were scaled by a factor of 1.2 and van der Waals 1-4 interactions were not scaled.

The small system alanine dipeptide was simulated using standard molecular dynamics (MD) at 293K for 1 μs. Additionally, alanine dipeptide was separately simulated with REMD using 4 replicas on a temperature ladder from 293K to 700K for 100 ns with exchanges attempted every 100 fs. Poly-alanine peptides with 2-10 alanine residues were also simulated using MD at 293K for 40 ns. All alanine systems included blocking groups (ACE-A(n)-NME) and were subjected to four independent simulations using four unique starting configurations corresponding to favorable regions in the Ramachandran plot: α, β, poly-proline II, and left-handed helix.

The met-enkephalin peptide (YGGFM) was simulated for 200 ns at 298K using MD. Four independent simulations were run with the same starting configurations as the poly-alanine simulations. The trpzip2 peptide (SWTWENGKWTWK) was simulated for 200 ns using REMD with 10 replicas exponentially spaced between 200K and 400K. Exchange attempts were made every 100fs and four simulations were run in total from four starting configurations of extended, α helix, β hairpin, and random. The 27-residue membrane proximal external region (MPER) of gp41 from HIV-1 (ACE-EQELLELDKWASLWNWFNITNWLWYIK-NHE) was simulated via replica exchange molecular dynamics (REMD) for 60 ns per replica. The MPER peptide is of vital interest for HIV vaccine research due to its role as a conserved epitope in two broadly neutralizing

HIV-1 antibodies.[16] The REMD simulation utilized 16 replicas exponentially spaced on a temperature ladder from 293K to 624K with exchanges attempted every 100fs. Four simulations in total were run from four unique starting configurations: extended, alpha helix, random alpha helix and random.

The MPER peptide was also simulated using the Boltzmann structural reservoir REMD method (srREMD).[17] The reservoir was constructed from 45 ns from the 358K temperature replicas from the four separate REMD simulations listed above. Structures were chosen equidistantly from their trajectories such that the reservoir consisted of 18,000 conformations. Here exchanges were attempted every 1 ps as our previous work has shown that reservoir methods perform poorly at high exchange frequencies (Wiehe et al., in preparation) We used 4 replicas exponentially spaced on a temperature ladder from 293K to 340K with a reservoir temperature of 358K.

3 Examples

We demonstrate the performance of our convergence diagnostic on peptide systems of various sizes.

Alanine Systems   Figure 3a shows the results of applying our convergence diagnostic to MD and REMD simulations of alanine dipeptide. Only 1 Å RMSD (all heavy atoms) clustering is shown. We see that REMD converges (195 ps) more than twice as fast as standard MD (442 ns) for this small system. Figure 3c shows convergence of poly-alanine peptides of increasing length. Growth appears roughly linear ($\rho = .91$).

Met-enkephalin   Convergence of the Met-enkephalin MD simulation occurs at 64 ns using 3 Å heavy-atom clustering Figure 3d. While this convergence time is long for a small pentapeptide system, it is comparable in length to a previous estimate of met-enkephalin convergence at a similar clustering resolution.[7]
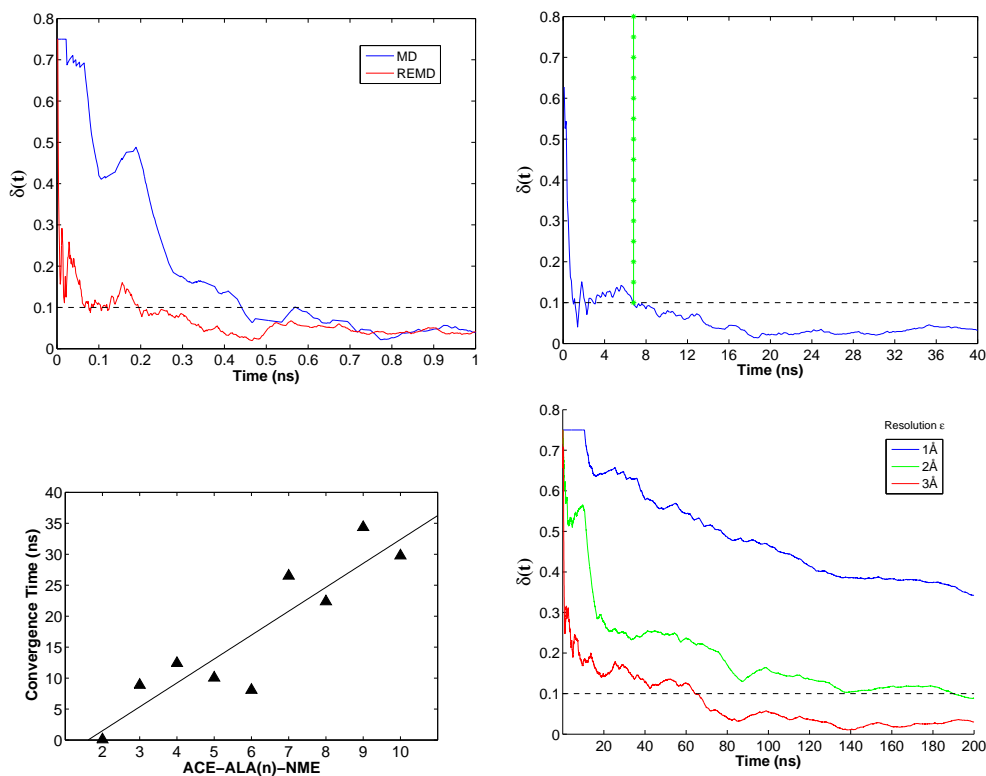
Figure 3: a) Convergence of alanine dipeptide simulations with REMD (red) and MD (blue) at a resolution ε=1 Å heavy-atom RMSD. b) Convergence of alanine tri-peptide MD simulation at a resolution ε=2 Å heavy-atom RMSD demonstrates threshold re-crossing behavior early in the simulation. When this occurs the use of a simple heuristic (starred vertical line) can prevent underestimation of convergence time c) Convergence times with resolution ε=2 Å RMSD (Cα) vs. peptide length for poly-alanine MD simulations, with best linear fit (solid line, ρ=0.91) . D) Convergence of Met-enkephalin at ε=1 Å 2 Å and 3 Å heavy-atom RMSD. Even small peptide systems require long simulations to converge. In panels a, b, and d, the convergence threshold is denoted by a dashed black line and colored solid lines represent the upper limit of the 95% confidence interval of the convergence diagnostic $\delta(t)$.

Trpzip2 Figure 1b shows the convergence of a 200 ns REMD simulation of the 12-residue trpzip2 peptide. We show convergence at multiple resolutions from 2 Å to 6 Å RMSD (Cα) for the lowest temperature (293K) replica. Convergence at even the coarse 6 Å resolution (for which there are only 4 clusters total) does not occur within the 200 ns simulation time. As in Figure 1a, this is due to an overpopulation of β-like configurations when initialized in the β-hairpin conformation. Thus our diagnostic identifies the kinetic trapping missed by other approaches using a single starting configuration (Figure 1a).

HIV gp41 MPER Figure 4a shows results for a 60 ns REMD simulation of the gp41 MPER peptide. Multiple resolutions are shown ranging from 2 Å to 8 Å RMSD (c-α). At 8 Å resolution it has converged by 42 ns, but finer resolutions have not converged by 60 ns. 8 Å is very coarse for a 27-residue peptide, with most of the ensemble represented by only 3 clusters. Comparing to Figure 2 confirms that using a single quantity (e.g. helicity) can grossly underestimate equilibration times.
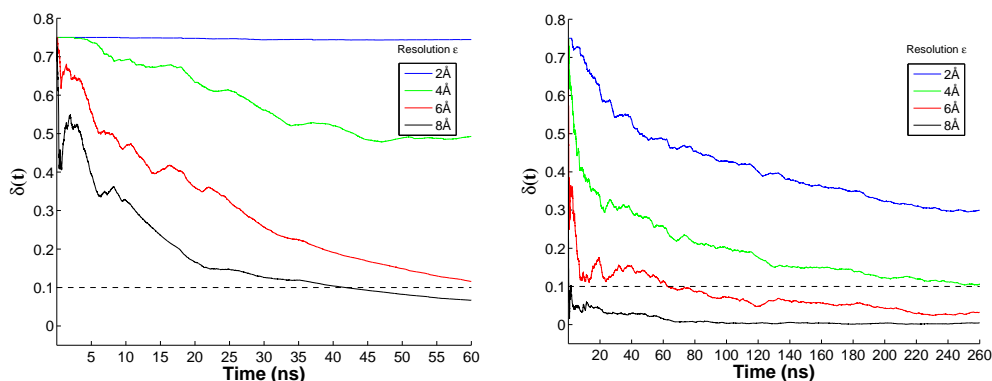

Figure 4: Convergence of the cluster frequencies for the MPER simulations using a) REMD and b) srREMD. Each colored line represents the upper limit of the 95% confidence interval of $\delta(t)$ at the specified resolution ε. srREMD is more efficient than REMD and thus converges faster.

With a reliable method for determining convergence times, we can also compare the efficiency of different simulation algorithms. As the MPER is very slow to converge using REMD, we investigated the efficiency of the recently developed Boltzmann structural reservoir REMD (srREMD) method.[17] At the 8 Å resolution, the srREMD simulation converges in 2 ns (Figure 4b), and by

260 ns has converged down to 5 Å resolution. (Note however that the srREMD is preceded by a 45 ns simulation at high temperature to construct the reservoir. See also Wiehe & Schmidler 2011 for some caveats about the srREMD method.) A structural comparison between two clusters that have not converged at 3 Å but belong to a common cluster that has converged at 5 Å is shown in Figure 5. While there is a broad agreement in the orientation of the backbones, there are substantial differences in the conformations as a whole. Given these structural differences at 3 Å it is likely convergence at this or a higher resolution will be necessary before meaningful conclusions can be made about the properties of the conformational ensemble of the MPER peptide, in contrast to results reported in the literature.[18]

Figure 5: Two MPER peptide conformations (blue and green cartoons) from within a cluster at the converged 5 Å resolution that are in separate clusters at the 3 Å resolution in which the simulation has not yet converged using srREMD. Both a local alignment (left panel) and global alignment (right panel) are shown.

4   Discussion

We have developed a diagnostic for monitoring molecular simulation convergence utilizing multiple independent simulations, which effectively detects kinetic trapping events that can lead other convergence methods to underestimate convergence times. Our method utilizes a efficient partitioning of conformational space and provides convergence at multiple levels of resolution simultaneously. Convergence under this criteria ensures that any ensemble average can be calculated

accurately.

It is apparent from our examples that convergence times for even moderate-sized systems may be much longer than commonly used in molecular simulation applications, as observed previously.[7,9] This highlights the necessity for advancements in sampling algorithms in order to make reliable simulations of larger-scale systems feasible. It is our hope that the methods introduced here will aid both practitioners simulating biomolecular systems, as well as researchers in simulation algorithms and forcefield development, both of which rely implicitly on knowing when simulation output is reliable and reproducible.

5   Acknowledgment

References

(1) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.

(2) Cooke, B.; Schmidler, S. C. *Biophys. J.* **2008**, *95*, 4497–4511.

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

(4) Seibert, M. M.; Patriksson, A.; Hess, B.; van der Spoel, D. *J. Mol. Biol.* **2005**, *354*, 173–183.

(5) Sgourakis, N. G.; Merced-Serrano, M.; Boutsidis, C.; Drineas, P.; Du, Z.; Wang, C.; Garcia, A. E. *J. Mol. Biol.* **2011**, *405*, 570–583.

(6) Smith, L. J.; Daura, X.; van Gunsteren, W. F. *Proteins* **2002**, *48*, 487–496.

(7) Lyman, E.; Zuckerman, D. M. *Biophys. J.* **2006**, *91*, 164–172.

(8) Lyman, E.; Zuckerman, D. M. *J Phys Chem B* **2007**, *111*, 12876–12882.

(9) Grossfield, A.; Feller, S. E.; Pitman, M. C. *Proteins* **2007**, *67*, 31–40.

(10) Gelman, A.; Rubin, D. *Statistical Science* **1992**, *7*, 457–511.

(11) Case, D. et al. AMBER 10. 2008.

(12) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J Comput Chem* **2003**, *24*, 1999–2012.

(13) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Journal of Physical Chemistry* **1996**, *100*, 19824–19839.

(14) Tsui, V.; Case, D. A. *Biopolymers* **2000**, *56*, 275–91.

(15) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *Journal of Computational Physics* **1977**, *23*, 327–341.

(16) Zwick, M. B.; Jensen, R.; Church, S.; Wang, M.; Stiegler, G.; Kunert, R.; Katinger, H.; Burton, D. R. *J. Virol.* **2005**, *79*, 1252–1261.

(17) Okur, A.; Roe, D. R.; Cui, G.; Hornak, V.; Simmerling, C. *Journal of Chemical Theory and Computation* **2007**, *3*, 557–568.

(18) Lapelosa, M.; Gallicchio, E.; Arnold, G. F.; Arnold, E.; Levy, R. M. *Journal of Molecular Biology* **2009**, *385*, 675 – 691.