Reducing the impact of transient effects in rate-constant estimation using the weighted ensemble strategy

Alex J. DeGrave and Lillian T. Chong

Abstract

We present a new method for improving the efficiency of calculating rate constants using the weighted ensemble path sampling strategy. This method reduces the sensitivity of rate-constant estimation to the earliest (and least typical) pathways reaching the target state by incorporating the distribution of event durations (excluding dwell time in the initial stable state) that correspond to pathways captured by the simulation. We demonstrate that the improved method enables accurate estimation of the rate constant for a protein conformational switching process using a fraction of the simulation time required by the original weighted ensemble method. Importantly, our method accounts for systematic error when using data from the entire simulation. Our method is relevant to any simulation strategy that involves unbiased trajectories of similar length to the typical event duration, including weighted ensemble and standard simulations.

1. Introduction

The weighted ensemble (WE) path sampling strategy[1] can be highly efficient at estimating rate constants as well as generating unbiased pathways for rare events (*e.g.* protein folding, protein binding).[2–78] A hallmark of rare events is that the actual transitions between stable states are infrequent, but relatively fast. Furthermore, the barrier crossing time, or event duration t_b, is typically orders of magnitude less than the associated waiting time between events (*i.e.*, the first passage time or the inverse of rate constant k) such that $t_b << k^{-1}$. Rare-events sampling strategies, including the WE strategy, exploit this separation of timescales by focusing the sampling on the *transitions* between stable states rather than the stable states themselves.[9]

Here we present a new method for improving the efficiency of calculating rate constants using the WE strategy. One issue with the original WE strategy is that there is a transient phase prior to relaxation into a steady-state, or equivalently, there is a "ramp up" time between when trajectories are started in the initial state and when the flux is no longer sensitive to the earliest (and least probable) pathways generated for the rare event. Our method significantly reduces this sensitivity by expressing the "ramp up" time in terms of the probability distribution of event durations that correspond to pathways captured by the simulation thereby enabling the calculation of rate constants from shorter trajectories. In addition to the WE strategy, our method is relevant to any simulation strategy that relies on unforced pathways of similar length to the typical event duration, including standard simulations, milestoning, and the construction of Markov state models.

To demonstrate the power of our improved WE strategy for calculating rate constants, we have applied the strategy to estimate the rate constant for protein conformational switching based on simulations of a protein-based Ca^{2+} sensor using a residue-level protein model.[6] This sensor was engineered using the alternate frame folding scheme, which involves fusing together the wild-type protein (in this case, the Ca^{2+}-binding protein, calbindin) and a circular permutant of the protein such that the two proteins partially overlap in sequence.[10] Due to sterics, the two proteins fold in a mutually exclusive manner. The protein conformational switching process therefore involves switching between the two alternate folded states. Importantly, our simulations of the switching process were able to capture the entire distribution of event duration times, making the simulations ideal for a "proof-of-principle" study.

2. Theory

The weighted ensemble (WE) strategy. In the WE strategy,[1] multiple trajectories are started in parallel from the initial state with each trajectory assigned a statistical weight. To control the trajectory distribution, configurational space is divided into bins along a 'progress coordinate' toward the target state. Trajectories are evaluated at fixed time intervals τ for either replication or combination to maintain the same number of trajectories/bin with the goal of generating a sufficiently large ensemble of continuous, successful pathways for

computing rate constants. Rigorous management of trajectory weights ensures that no bias is introduced into the dynamics. To maintain steady-state conditions, trajectories that reach the target state are 'recycled', i.e. terminated followed by the initiation of a new trajectory with the same weight as the terminated trajectory.

Original method for rate-constant estimation calculation. Consider a system in state A at time $t = 0$. Assume that this system has the following properties.

1. While in state A, the system has a constant probability per unit time of entering a successful transition path to state B, denoted k_{AB}.
2. After entering a successful transition path, the transition to B occurs in a nonnegative interval of time. These event durations are randomly distributed according to a probability density function $g_{AB}: [0, \infty) \rightarrow [0, \infty)$ with $\int_0^\infty g_{AB}(t)dt = 1$
3. Upon arriving in state B, the system is immediately "recycled" to state A.

Using the original WE strategy, macroscopic rate constants $k_{AB}$for the switching process involving an initial state A and target state B are computed as follows:

$$k_{AB} = \frac{\langle f_{AB}^{SS} \rangle}{\langle p_A \rangle} = \langle f_{AB}^{SS} \rangle$$

where f_{AB}^{SS} is the steady-state flux of probability carried by trajectories originating in state A and arriving in state B and $p_A$is the fraction of trajectories more recently in A than in B, which is unity due to the recycling conditions. In practice, if a steady state cannot be reached, or if it is unknown whether a simulation has reached steady state, then an analogous naïve estimate may be used:

$$k_{AB} = \langle f_{AB}(t) \rangle_t$$

where $f_{AB}(t)$ is the flux (not necessarily at steady-state) from state A to state B at simulation time t.

Reducing the impact of transient effects in rate-constant estimation. The new method of rate-constant estimation in the present study reduces the impact of transient effects in rate-constant estimation by incorporating information measurable during WE simulation, namely, the distribution of event durations (excluding the dwell time in state A) denoted $g_{AB}(t)$. The flux f_{AB} from the initial state A into the target state B can then be written as a convolution of the initiation of a transition with rate k_{AB} and completion of the event in a time τdistributed according to g_{AB}. Thus, $f_{AB}(t) = \int_0^t k_{AB}\, g_{AB}(\tau)d\tau$, which can be integrated and rearranged to obtain an expression for k_{AB}that depends only upon the cumulative flux $F_{AB}(t_{max}) = \int_0^{t_{max}} f_{AB}(t)dt$ and the cumulative distribution of barrier crossing times $G_{AB}(t) = \int_0^\tau g_{AB}(\tau)d\tau$, both of which are observed from WE simulation: $k_{AB} = F_{AB}(t_{max}) / \int_0^{t_{max}} G_{AB}(t)dt$. Finally, an additional correction is necessary to estimate the distribution $g_{AB}(t)$ of barrier-crossing times from the observed $A \rightarrow B$ events, since during the transient phase we are more likely to observe events with shorter barrier crossing times. From the histogram $\hat{g}_{AB}(t)$, we obtain a corrected estimate $\tilde{g}_{AB}(t)$ by considering the interval of time $[t, t_{max}]$ in which it is *possible* to observe an event of duration t in a simulation of total length t_{max}: $\tilde{g}_{AB}(t) \propto \hat{g}_{AB}(t)/(t_{max} - t)$, where the constant of proportionality is chosen such that $\int_0^\infty \tilde{g}_{AB}(t)dt = 1$. Thus we define an estimate $\tilde{k}_{AB}$ as follows:

$$\tilde{k}_{AB} = \frac{F_{AB}(t_{max})}{\int_0^{t_{max}} \int_0^t \tilde{g}_{AB}(\tau)d\tau dt}$$

While this analysis does not eliminate the need to observe the majority of the distribution of barrier-crossing times, we anticipate that the improved analysis will enable more accurate estimation of rate constants using less total simulation time, since the transient phase can be correctly incorporated into the calculation rather than "thrown away" in favor of later data. Furthermore, we note that this analysis could be especially important for challenging molecular processes, which feature long transient phases.

In cases where it is not possible to sample the entire distribution of event duration times, our method provides a framework for understanding the error that results from not observing longer duration events. Supposing that the maximum trajectory length is t_0, the histogram estimate $\tilde{g}_{AB}(t)$ of $g_{AB}(t)$ will be zero for $t > t_0$and, since $\tilde{g}_{AB}$ is normalized such that $\int_0^\infty \tilde{g}_{AB}(t)dt = 1$, $\tilde{g}_{AB}(t)$ will be artificially inflated for $t < t_0$, *i.e.*:

$$\tilde{g}_{AB}(t) \approx g_{AB}(t) / \int_0^{t_0} g_{AB}(\tau)d\tau \text{ for } t \in [0, t_0]$$

Thus while we desire that $\tilde{k}_{AB} \approx F_{AB}(t_{max}) / \int_0^{t_{max}} \int_0^t g_{AB}(\tau)d\tau dt$, which would be the case if $\tilde{g}_{AB} \approx g_{AB}$, we instead find that $\tilde{k}_{AB} \approx (\int_0^{t_0} g_{AB}(\tau)d\tau) F_{AB}(t_{max}) / \int_0^{t_{max}} \int_0^t g_{AB}(\tau)d\tau dt$. Equivalently, $\tilde{k}_{AB}$tends to underestimate k_{AB} by a factor of $\int_0^{t_0} g_{AB}(\tau)d\tau$, the observed fraction of the event duration distribution. In practice a numerical estimate of this value is prohibited by the lack of knowledge of g_{AB}, but we confirm the intuitive proposition that if, for example, 20% of pathways reaching the target state are of greater duration than the maximum trajectory length, then we tend to underestimate $k_{AB}$by 20%, which is improved relative to the naive estimate $k_{AB} = \langle f_{AB}(t) \rangle_t = F_{AB}(t_{max})/t_{max}$.

3. Methods

WE simulations. WE simulations of the protein conformational switching pathways were carried out as described in DeGrave *et al*.[6] All data represent 10 independent WE simulations of the N′ ⟶N transition of the wild-type E65′Q switch construct. Briefly, the simulations were carried out using the open-source, highly scalable WESTPA software package (https://westpa.github.io/westpa)[11] under steady-state conditions using a Brownian dynamics algorithm with hydrodynamic interactions, as implemented in the UIOWA-BD software.[12,13] All analysis was performed with conformations sampled every 50 ps. A minimal residue-level protein model was employed in which each residue is represented by a single pseudo-atom at the position of its C_α atom. The conformational dynamics of the protein were governed by a Gō-type potential energy function[14,15] that was parameterized to reproduce the experimental folding free energies of the isolated wild-type protein and circular permutant of the protein.[6]

4. Results

As described above, our new method of rate-constant estimation reduces the impact of transient effects by making use of the probability distribution of event duration times that correspond to simulated pathways of the rare event. Given that the transient effects are due to the presence of pathways with non-negligible event durations, our new method $\tilde{k}_{AB}$ quickly converges to the steady-state value when provided with the entire distribution of event durations$\tilde{g}_{AB}$estimated from all simulation data (~10 ns vs ~75 ns for the new method and original WE strategy, respectively) (**Fig. 1a-b**).

To test the real-world utility of our new method, we examined the evolution of $\tilde{k}_{AB}$as a function of the molecular time, where at any given time the estimate $\tilde{g}_{AB}$is based only upon simulation data generated up to and including that time. The new method yields substantially faster convergence of the rate constant k_{AB} for switching between states A and B for the protein-based Ca^{2+} sensor (**Fig. 2**). In particular, as compared to the original WE strategy, a steady-state value of k_{AB} is attained in ~⅓ the molecular time that is required when using the original WE strategy (~25 ns vs ~75 ns for the new method and original WE strategy, respectively); the molecular time is Nτ where N is the number of WE iterations and τ is the fixed time interval of each iteration. Importantly, this comparison suggests that the new method enables accurate estimate of rate constants using only a fraction of the simulation time previously necessary.

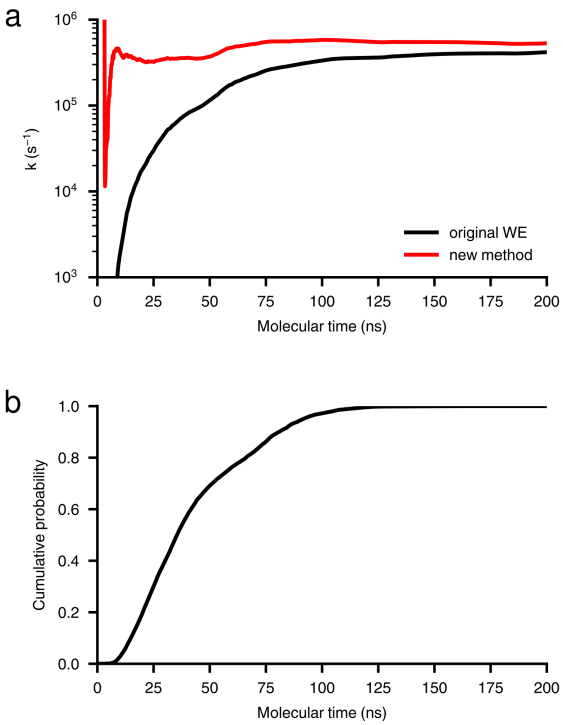

Figure 1 | Effect of barrier crossing times on calculation of the rate constant. a, Comparison of the rate constant $k_{AB}$for switching of the protein-based Ca^{2+} sensor from state A to state B using the original WE strategy and the new method (via the estimate $\tilde{k}_{AB}$) in the present study, as plotted as a function of molecular time, or Nτ where N is the number of WE iterations and τ is the fixed time interval of each iteration. **b,** The cumulative probability distribution $\tilde{G}_{AB}$ of event duration times of switching pathways generated by the same simulation depicted in panel (a).

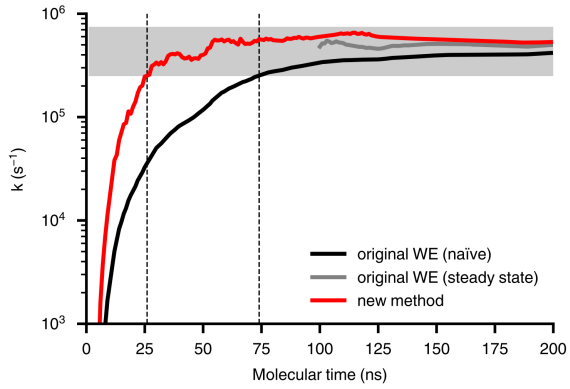

Figure 2 | Comparison of simulation times necessary to achieve converged estimate of the rate constant. The red line indicates $\tilde{k}_{AB}(t)$, where for a given time t the distribution of event durations is estimated using only data up to time t. The black line indicates the naïve estimate $k_{AB} = \langle f_{AB}(t) \rangle_t$, and the gray line indicates $k_{AB} = \langle f_{AB}^{SS} \rangle$, the average flux after the simulation has achieved steady state at 100 ns as determined from the distribution of event durations in fig. 1b. The shaded region highlights the range of converged estimates, defined here as the final value of $\langle f_{AB}^{SS} \rangle$ plus or minus an error margin of 50%.

5. Conclusions

We have developed a new method for calculating rate constants within the framework of the WE strategy that reduces the impact of transient effects on rate-constant estimation. While the method requires that the WE simulation of the rare-event process samples a substantial portion of the event duration distribution (*e.g.*, t

such that $\int_0^t g_{AB}(\tau)d\tau \geq 0.5$), our proof-of-principle test indicates that the method enables accurate estimation of the rate constant using a fraction of the simulation time required by a previous method. Importantly, this method accounts for systematic error when using data from the entire simulation -- even before the molecular time exceeds the maximum event duration time.

Acknowledgements

This work was supported by NIH 1R01GM115805-01 to L.T.C. and D.M.Z., and a University of Pittsburgh Honors College Brackenridge Undergraduate Research Fellowship to A.J.D. Computational resources were provided by NSF CNS-1229064 and the University of Pittsburgh's Center for Research Computing. We thank Daniel Zuckerman (OHSU) for helpful discussions.

References

1. Huber, G. A. & Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys. J.* **70,** 97–110 (1996).
2. Zhang, B. W., Jasnow, D. & Zuckerman, D. M. Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 18043–18048 (2007).
3. Zwier, M. C., Kaus, J. W. & Chong, L. T. Efficient Explicit-Solvent Molecular Dynamics Simulations of Molecular Association Kinetics: Methane/Methane, Na /Cl−, Methane/Benzene, and K /18-Crown-6 Ether. *J. Chem. Theory Comput.* **7,** 1189–1197 (2011).
4. Adelman, J. L. & Grabe, M. Simulating rare events using a weighted ensemble-based string method. *J. Chem. Phys.* **138,** 044105 (2013).
5. Saglam, A. S. & Chong, L. T. Highly Efficient Computation of the Basal kon using Direct Simulation of Protein–Protein Association with Flexible Molecular Models. *J. Phys. Chem. B* **120,** 117–122 (2016).
6. DeGrave, A. J., Ha, J.-H., Loh, S. N. & Chong, L. T. Large enhancement of response times of a protein conformational switch by computational design. *Nat. Commun.* **9,** 1013 (2018).
7. Rojnuckarin, A., Livesay, D. R. & Subramaniam, S. Bimolecular reaction simulation using Weighted Ensemble Brownian dynamics and the University of Houston Brownian Dynamics program. *Biophys. J.* **79,** 686–693 (2000).
8. Zwier, M. C. *et al.* Efficient Atomistic Simulation of Pathways and Calculation of Rate Constants for a Protein-Peptide Binding Process: Application to the MDM2 Protein and an Intrinsically Disordered p53

Peptide. *J. Phys. Chem. Lett.* **7,** 3440–3445 (2016).

9. Chong, L. T., Saglam, A. S. & Zuckerman, D. M. Path-sampling strategies for simulating rare events in biomolecular systems. *Curr. Opin. Struct. Biol.* **43,** 88–94 (2017).
10. Stratton, M. M., Mitrea, D. M. & Loh, S. N. A Ca2 -Sensing Molecular Switch Based on Alternate Frame Protein Folding. *ACS Chem. Biol.* **3,** 723–732 (2008).
11. Zwier, M. C. *et al.* WESTPA: an interoperable, highly scalable software package for weighted ensemble simulation and analysis. *J. Chem. Theory Comput.* **11,** 800–809 (2015).
12. Elcock, A. Molecular simulations of cotranslational protein folding: fragment stabilities, folding cooperativity and trapping in the ribosome tunnel. *PLoS Comput. Biol.* **preprint,** e98 (2005).
13. Frembgen-Kesner, T. & Elcock, A. H. Striking Effects of Hydrodynamic Interactions on the Simulated Diffusion and Folding of Proteins. *J. Chem. Theory Comput.* **5,** 242–256 (2009).
14. Go, N. Theoretical Studies of Protein Folding. *Annu. Rev. Biophys. Bioeng.* **12,** 183–210 (1983).
15. Takada, S. Go-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 11698–11700 (1999).