Overview of Weighted Ensemble Simulation: Path-sampling, Steady States, Equilibrium

Daniel M. Zuckerman, University of Pittsburgh

1 Introduction to Weighted Ensemble Simluation: The very basics

"Weighted ensemble" (WE) simulation is an enhanced sampling method for non-equilibrium and equilibrium processes that would be impractical to observe using straightforward dynamics simulation. Two key examples are (i) activated processes — i.e., overcoming energy barriers and (ii) binding processes — i.e., sampling rare complexation events. The basic goal of such studies is to get from a known initial state (A in Fig. 1) to a target state B, thereby learning pathways and rates for the process. WE uses a multiple-trajectory strategy in which individual trajectories can spawn multiple daughter trajectories upon reaching new regions of configuration space called bins. The daughters are suitably weighted to ensure statistical rigor. WE simulation can yield rigorous estimates for timescales/processes that are much longer than the simulations themselves.

The WE approach has numerous strengths:

- Applicable to folding, binding, conformational change in molecular systems
- Easy to understand and implement
- Statistically rigorous
- Non-equilibrium: Yields rate and path ensemble in a single simulation
- Equilibrium: Can be run in equilibrium mode – states do *not* need to be defined in advance
- Easy to parallelize because of multi-trajectory protocol
- Readily handles multiple pathways and metastable intermediates
- Code is available: `http://chong.chem.pitt.edu/WESTPA/`

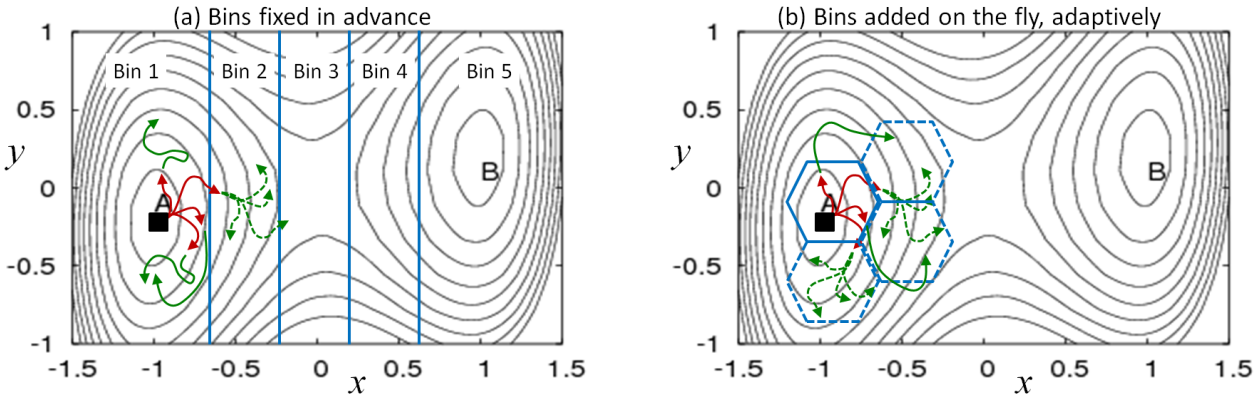

Figure 1: Weighted ensemble (WE) simulation using bins defined in advance or on the fly. Multiple trajectories are initiated from any known configuration (square) and run using standard dynamics simulation, without bias. Each of the four initial trajectories is assigned a weight of 1/4. When a trajectory reaches a new bin, it is replicated with each daughter receiving an equal share of the parent trajectory's weight. Trajectories are checked for their location/bin at fixed time intervals: in the first interval, red trajectories were run, followed by green trajectories in the second interval. The dashed trajectories have each inherited 1/4 of the parent trajectory's weight, resulting in weights of 1/16 for each daughter. (a) If a reasonable progress coordinate is known in advance, such as distance to a target state, then bins can be defined in advance. (b) Bins can also be defined adaptively according to configurations sampled by trajectories, where dashed bins were added after the first interval.

2 Brief history of WE simulation

WE simulation was originally introduced as "weighted ensemble Brownian dynamics" by Huber and Kim to study binding processes[1]. However, the essence of the idea to split and propagate re-weighted trajectories had been introduced early in the Monte Carlo era[2]. Our group showed that WE was applicable to a broad class of stochastic processes, including non-Markovian dynamics[3], and not simply Markovian dynamics as suggested by the initial "Brownian" appellation. We also showed that *steady-state* WE simulations were straightforward to implement and accelerated rate calculations[4]. Because equilibrium is itself a steady state, WE can also be used for equilibrium sampling[4]. Since its introduction, WE has been applied to folding a coarse-grained protein model[5], studying conformational transitions in coarse protein models[6,7] and in a semi-atomistic model[8], as well as for explicit solvent studies of molecular association[9]. In unpublished work, we have used WE to fold small, implicitly solvated all-atom proteins. The Chong group has calculated the on-rate for MDM2 association with a p53 peptide using all-atom implicit-solvent simulation (unpublished).

3 Standard WE protocol for pre-defined bins

The WE procedure is particularly simple to describe when pre-defined bins are used. As with any WE protocol, dynamics can be run using any simulation package. All WE procedures naturally lend themselves to scripting and/or parallel implmentations.

Generalizing Fig. 1(a), let us assume there are N bins, and that M trajectories are initiated from some starting configuration — each with weight $1/M$. Typically, one sets a maximum of M trajectories per bin. Trajectories are examined after each time interval τ, which is also arbitrary. τ will usually consist of many simulation steps, but it is better to have a shorter τ so long as overhead costs for examining trajectories remain small compared to the cost of running dynamics.

After each τ interval, trajectories are examined to determine (i) if there are any occupied bins with fewer than M trajectories, such as newly entered bins and (ii) if there are any occupied bins with more than M trajectories. If a bin has fewer than M trajectories, then one or more of the trajectories in the bin are replicated, according to a statistical procedure where daughters share equally in the parent's weight. Thus, if a bin is occupied by single trajectory, that will be split into M daughters receiving a fraction $1/M$ of the parent's weight. The overall weights continue to sum to 1, maintaining normalization. If a bin is found to contain more than M trajectories, these are pruned ("resampled") down to the limit M. The statistical basis of these procedures has been detailed in our recent paper[3].

WE thus yields the time-evolving probability distribution, as well as the ensemble of trajectories. In simple cases, the rate can be calculated direclty from the evolving distribution, but more generally steady-state simulations are needed: see below.

Different initial conditions can be studied, including from multiple configurations, depending on the problem of interest.

4 The importance of steady-state WE simulations

If metastable intermediate states are present between the initial and target states, standard WE simulation still yields the trajectory ensemble, but the rate cannot easily be calculated. Most systems interesting enough to warrant WE simulation will possess such intermediates. In standard WE simulation, probability/weight will build up very slowly in intermediates, leading to a long transient period before typical transition dynamics are exhibited — thus preventing characterization of long-timescale behavior.

The steady-state WE (WESS) protocol is a straightforward way to sidestep the problems just described[4]. In brief, one uses standard WE simulation until the target state is reached, with suc-

cessfully transitioning trajectories fed back into the initial state. The unbiased trajectories that are always employed in WE simulation are then used to determine the conditional probabilities to hop among bins — i.e., the inter-bin rates k_{ij}. In turn, these rates can be used to estimate the steady-state probabilities of each bin, p_i^{SS} — that would result only after a very long WE simulation.

A new WESS simulation is then begun from the previous state of the system — i.e., with trajectories occupying many or all bins. However, the weights of the trajectories are changed so that the sum of weights in bin i matches p_i^{SS} estimated from the rates. This new WESS simulation is continued to confirm that steady state has been obtained, with unchanging bin probabilities. It is possible to iterate the rate calculations and probability reassignments to accelerate attainment of a steady state under difficult conditions.

5 Calculating the rate

A remarkable feature of WESS simulation is that information on very long timescales (e.g., the first passage time) can be obtained rigorously from a short simulation, even for complex landscapes. Once a steady state has been established using WESS, the transition rate — i.e., the inverse mean-first-passage time — is obtained simply as the total probability/weight entering the target state per unit time — i.e., by the flux[4]. (The related steady-state rate also is obtained directly from WESS[4].) This simplicity is a key strength of WE. No auxiliary calculations are required to obtained the rate.

6 Equilibrium WE simulations

Because equilibrium is a steady state (a special one without any net flows[10]), it can be simulated using the WESS protocol[4]. The only difference is that no feedback from target to initial state should be performed. Operationally, then, a standard WE simulation is run first until its trajectories reach the target state or a desired set of states. Rates are then calculated, and probability is reassigned according to p_i^{SS} — except that the steady state in question is equilibrium. As in WESS, the bin populations must be monitored to ensure they are unchanging. As suggested by Fig. 2, in equilibrium, the amount of probability traveling from one bin to another is exactly balanced by the reverse flow.

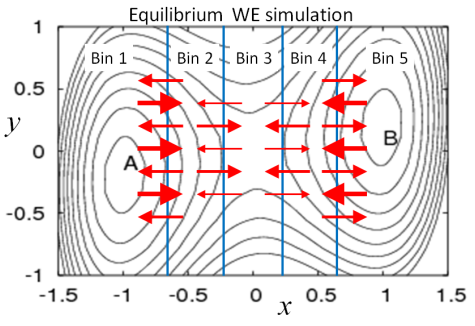

Figure 2: Schematic depiction of equilibrium in WE simulation. Thicker arrows represent higher-weight trajectories. The total probability/weight flowing from one bin to another is exactly balanced, on average, by the reverse flow. Thus, the flow of a given number of low-weight trajectories is matched by a smaller number of higher-weight trajectories moving in the opposite direction. Equilibrium flows of weighted trajectories can be established by estimating and adjusting weights[4].

Note that the equilibrium probability distribution differs from that of a steady state with feedback. The lack of feedback in equilibrium is sufficient to change the rates and p_i^{SS} values so that they correspond to correspond to equilibrium.

Rates can be calculated between arbitrary states – *which do not need to be defined in advance* – based on a WE equilibrium simulation. In brief, the equilibrium ensemble is precisely decomposed into two steady states reflecting forward and reverse events between two states of interest[11]. These two steady states can be used to estimate rates, as we have shown[12].

7 Multiple pathways

As shown in Fig. 3, WE is ideal for identifying different pathways because it employs multiple trajectories. Our group has used WE to sample multiple paths in model systems, small molecules, and a semi-atomistic model of the protein adenylate kinase[4,8,13]. Michael Grabe's group has also used WE to identify different potential pathways in a simplified model of a membrane protein (unpublished). WE handles multiple pathways naturally and requires no auxiliary calculations.

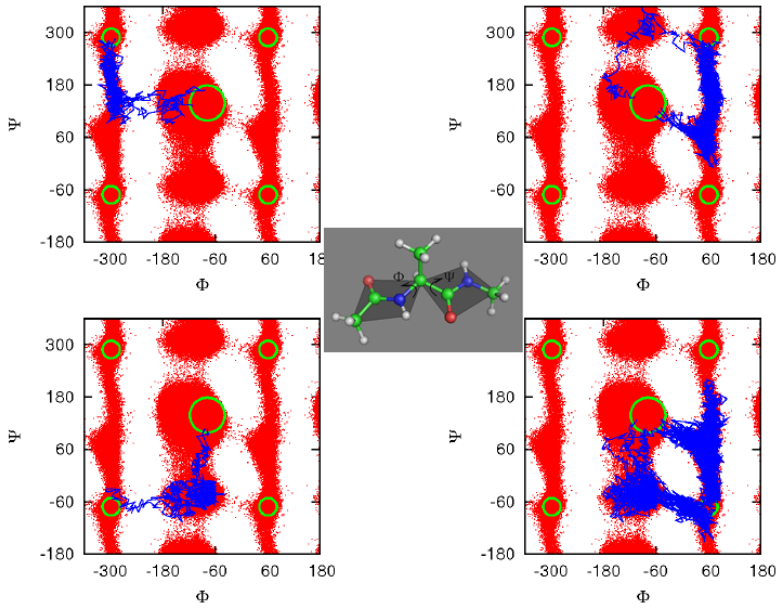

Figure 3: Weighted ensemble (WE) simulation is well suited for finding multiple pathways. A number of different pathways are shown in the alanine dipeptide molecule (center). Transition trajectories start from the c_{7eq} state at the center of each panel and terminate in the c_{7ax} state, which occurs in the four outside corners of each panel due to the cylic nature of torsion angles. WE has also identified distinct pathways in the opening/closing transition of the adenylate kinase protein[8].

8 Parallelization

Because WE examines multiple trajectories (typically hundreds or even thousands) in each τ interval, it is natural to pursue parallelization. The communication overhead tends to be negligible, fortunately, because running dynamics for complex systems is fairly expensive, We have run WE using MPI to start multiple simultaneous Amber simulations (unpublished). Lillian Chong's group has recently reported parallel simulations using GROMACS[9]. Fully functional, documented parallel code developed by the Chong group with input from our group and others is now available: `http://chong.chem.pitt.edu/WESTPA/`.

9 Adaptive binning

The possibility to perform adaptive binning — i.e., to change bins during a simulation — was noted by Huber and Kim[1] and reinforced in our group's work[3]. There is no unique strategy for changing bins. Indeed, this flexibilty is a strength of WE simulation.

As an example, consider a strategy proposed several years ago[6] that we have found useful in folding the Trp-cage mini-protein. A natural coordinate for setting up "radial" bins is the RMSD distance from the experimental folded structure. However, there can be barriers orthogonal to this coordinate. To encourage dispersal of trajectories within the radial bins, one can *adaptively* define sub-bins: the first configuration recorded in a newly entered radial bin can be used as a reference structure to define sub-bins. That is, each inital bin based on RMSD to the target can be further

sub-divided based on local information — i.e., based on RMSD to a configuration in the radial bin. Other adaptive strategies are possible.

10 Summing Up

Weighted ensemble (WE) simulation is a powerful and flexible tool which can handle complex systems in and out of equilibrium. Its straightforward statistical basis makes WE relatively easy to implement and parallelize. The *flexibility* in implementing WE (e.g., adaptive binning strategies) means that the capacities of WE have yet to be fully realized. The use of WE as an equilibrium sampling tool, in particular, is only in its infancy.

References

1. G. A. Huber and S. Kim. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys. J.*, 70:97–110, 1996.

2. Herman Kahn. Use of different Monte Carlo sampling techniques. Technical Report Report P-766, Rand Corporation, 1956.

3. Bin W Zhang, David Jasnow, and Daniel M Zuckerman. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *J Chem Phys*, 132(5):054107, Feb 2010. PMCID: PMC2830257.

4. Divesh Bhatt, Bin W. Zhang, and Daniel M. Zuckerman. Steady state via weighted ensemble path sampling. *Journal of Chemical Physics*, 133:014110, 2010. PMCID: PMC2912933.

5. Atipat Rojnuckarin, Sangtae Kim, and Shankar Subramaniam. Brownian dynamics simulations of protein folding: Access to milliseconds time scale and beyond. *Proceedings of the National Academy of Sciences of the United States of America*, 95(8):4288–4292, 1998.

6. Bin W. Zhang, David Jasnow, and Daniel M. Zuckerman. Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin. *Proceedings of the National Academy of Sciences*, 104(46):18043–18048, 2007.

7. Joshua L. Adelman, Amy L. Dale, Matthew C. Zwier, Divesh Bhatt, Lillian T. Chong, Daniel M. Zuckerman, and Michael Grabe. Simulations of the alternating access mechanism of the sodium symporter mhp1. *Biophys J*, 101:2399–2407, 2011. PMCID: PMC3218348.

8. Divesh Bhatt and Daniel M. Zuckerman. Heterogeneous path ensembles for conformational transitions in semiatomistic models of adenylate kinase. *Journal of Chemical Theory and Computation*, 6(11):3527–3539, November 2010. PMCID: PMC3108504.

9. Matthew C. Zwier, Joseph W. Kaus, and Lillian T. Chong. Efficient explicit-solvent molecular dynamics simulations of molecular association kinetics: Methane/methane, Na+/Cl-, methane/benzene, and K+/18-Crown-6 Ether. *Journal of Chemical Theory and Computation*, 7(4):1189–1197, April 2011.

10. Daniel M. Zuckerman. *Statistical Physics of Biomolecules: An Introduction*. CRC Press, Boca Raton, FL, 2010.

11. Divesh Bhatt and Daniel M. Zuckerman. Beyond microscopic reversibility: Are observable nonequilibrium processes precisely reversible? *Journal of Chemical Theory and Computation*, 7(8):2520–2527, August 2011. PMCID: PMC3159166.

12. Steven Lettieri, Matthew C. Zwier, Carsen A. Stringer, Ernesto Suarez, Lillian T. Chong, and Daniel M. Zuckerman. Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. http://arxiv.org/abs/1210.3094, 2012.

13. Bin W. Zhang, D. Jasnow, and D. M. Zuckerman. Weighted ensemble path sampling for multiple reaction channels. 2010. Preprint available: http://arxiv.org/abs/0902.2772.