3-7-2016

Statistical Biophysics Blog: So you want to do some path sampling …

Daniel M. Zuckerman
Oregon Health & Science University, zuckermd@ohsu.edu

Statistical Biophysics Blog: Statistical mechanics in biology and biocomputation
http://statisticalbiophysicsblog.org/

[Note: Some posts were originally written at the University of Pittsburgh. Corrections and updates are included in this Digital Commons version.]

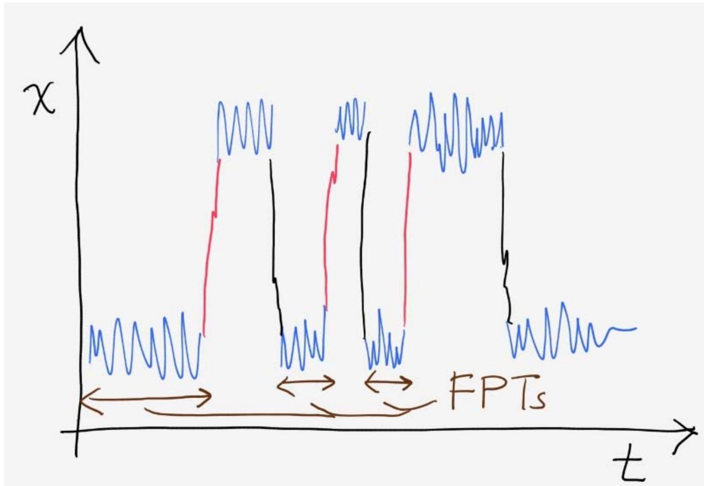
Keywords: molecular dynamics simulation, statistical mechanics, path sampling, trajectory ensemble, rate constant, first passage time

Title: So you want to do some path sampling …
Subtitle: Basic strategies, timescales, and limitations
http://statisticalbiophysicsblog.org/?p=115
Original publication date: March 7, 2016

Key biomolecular events – such as conformational changes, folding, and binding – that are challenging to study using straightforward simulation may be amenable to study using "path sampling" methods. But there are a few things you should think about before getting started on path sampling. *There are fairly generic features and limitations* that govern all the path sampling methods I'm aware of.

Path sampling refers to a large family of methods that, rather than having the goal of generating an ensemble of system configurations, attempt to generate an ensemble of dynamical *trajectories*. Here we are talking about trajectory ensembles that are precisely defined in statistical mechanics. As we have noted in another post, there are different kinds of trajectory ensembles – most importantly, the equilibrium ensemble, non-equilibrium steady states, and the initialized ensemble which will relax to steady state. Typically, one wants to generate trajectories exhibiting events of interest – e.g., binding, folding, conformational change.

A trajectory can be considered a list of configurations (possibly with velocities) for all system coordinates recorded with a fixed time increment. Note that there is indeed a path ensemble even in one dimension: because displacements/velocities will vary along a trajectory, there are an infinite number of trajectories connecting any two points. In principle, trajectory ensemble of transition events could be obtained by collecting transitions of interest from a very long trajectory – for example the red segments below, possibly with their preceding blue segments, which together make up the first-passage times (FPTs) for the system to transition from low to high x values.



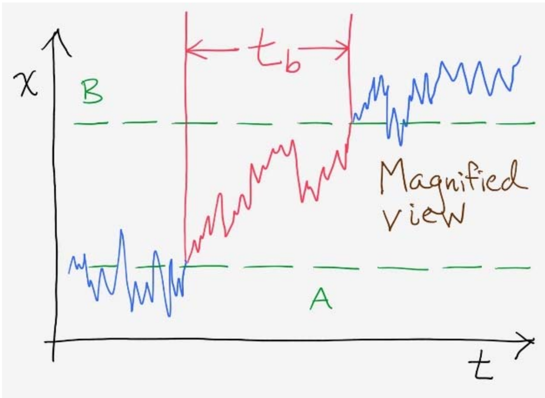
Path sampling methods can focus computing resource on a subset of trajectories (e.g., the red transition events, above) and have been developed using a variety of strategies. We'll mention three that work with continuous trajectories rather than disconnected segments. (1) Lawrence Pratt suggested that, because the probability for a given trajectory to occur could be calculated, one could perform Metropolis Monte Carlo in path (trajectory) space; this was the "transition path sampling" idea later taken up by Chandler and co-workers. The basic idea, however, has its roots in path-integral Monte Carlo. (2) Huber and Kim suggested that an ensemble of trajectories could be orchestrated using replication and pruning steps in a way that could encourage sampling of rare processes. This "weighted ensemble" strategy was really a re-discovery of the "splitting and Russian roulette" strategy published by Los Alamos theorists Herman Kahn & coworkers in the 1950s. (3) "Dynamic importance sampling" was proposed by Woolf, based on prior work by Ottinger, in which trajectories could be biased toward rare events of interest, with reweighting performed after the fact to ensure conformance with statistical principles.

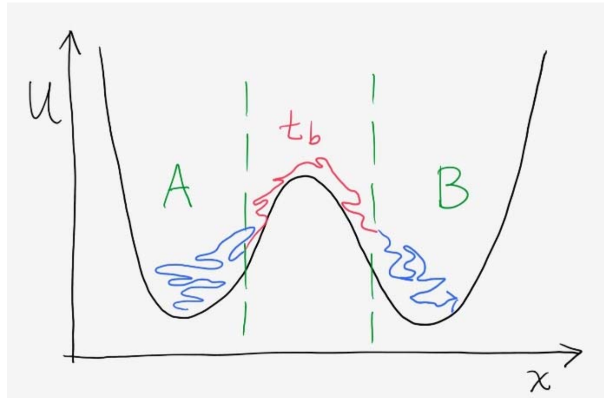
The preceding are three basic approaches that generate ensembles of *continuous* trajectories. It is fair to note that many sophisticated variants and improvements on the basic strategies have been developed, in addition to many approaches using collections of discontinuous segments (see review by Elber noted below); these are quite valuable but a distraction from the main points of this post.

Generic limitations of path sampling

All the continuous-trajectory approaches share two fundamental limitations. One arises from intrinsic system-specific transition timescales, and the other is a consequence of intrinsic sampling limitations.

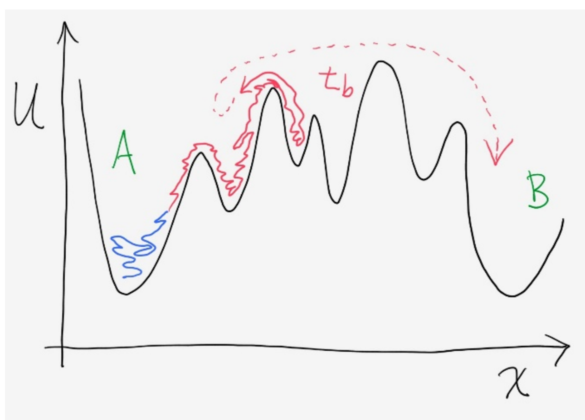
To understand the limitations, let's assume *our goal is to sample 100 statistically independent transition events*. Although every individual trajectory is time-correlated because configurations are generated sequentially, trajectories can be statistically independent – for example, if you started 100 independent simulations in an initial state and simply waited for 100 transitions. Of course, that strategy generally is prohibitive and motivates path sampling in the first place, but truly independent simulations would be a gold standard for independent transition-event trajectories.





System-specific timescales

Generally speaking, there are two kinds of transition events. As shown in our original long one-dimensional trajectory and immediately above in magnified view, in a simple "activated process" characterized by a dominant energy barrier, the duration of a transition t_b will be much shorter than the waiting time (a.k.a. dwell time) in the initial state. The sum of the average dwell and event times is called the mean first-passage time (MFPT). Although t_b may be short and much less than the MFPT, it is still finite. A more challenging scenario is depicted below in the figure with many intermediate states: each intermediate can lead to a separate, possibly lengthy dwell – and don't forget that trajectories can reverse many times leading to more dwells than there are intermediates. In such a case, the transition-event duration t_b may be similar to the overall MFPT.



With this understanding, let's get back to our goal of simulating 100 independent transitions. The minimum cost for doing this with fully continuous trajectories is 100 t_b. If t_b ~ 10 ns, then at least 1 μs is needed for our trajectory ensemble. And there is no guarantee that t_b will be short (compared to timescales that can easily be simulated). So start worrying now … and things only get worse.

Intrinsic limitations of sampling

Path sampling is desirable when system timescales (MFPTs for processes of interest) are too long to simulate. In other words, by definition of the problem, we cannot afford 100*MFPT. Algorithms such as the ones sketched above have the potential to limit computational effort to the transition events themselves. But generating *independent* transition-event trajectories is not a trivial matter! Although starting separate “brute force” (i.e., standard) trajectories is simple to do, if one wants computational effort to be focused on transition events, there are additional costs.

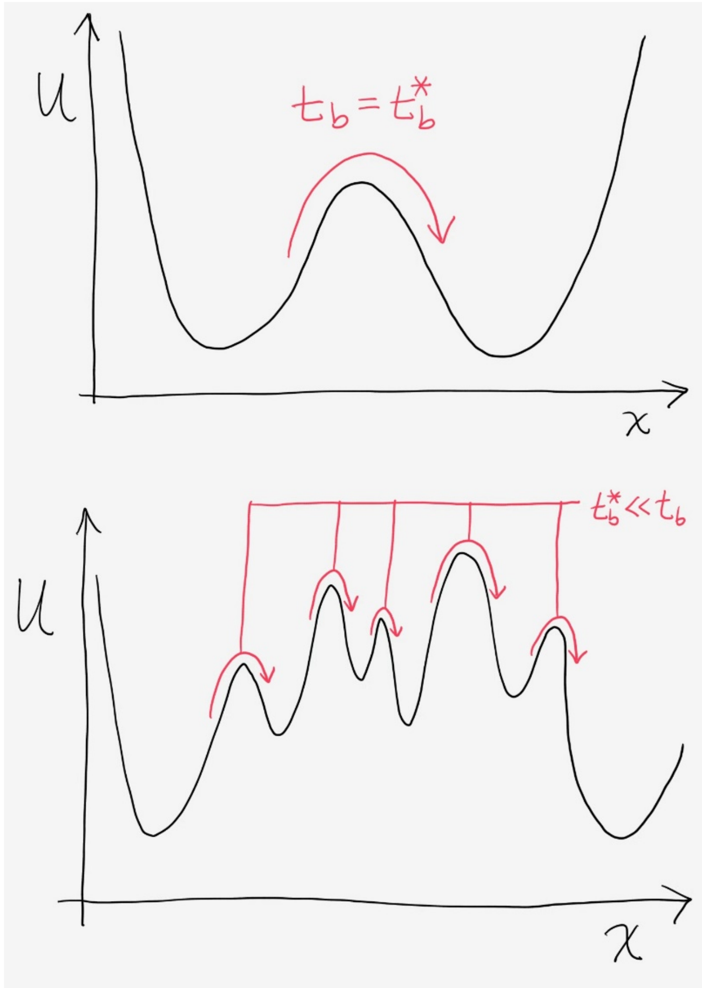
Let’s look in turn at each of the three path-sampling strategies described above.

In *transition path sampling,* Metropolis Monte Carlo in path space requires perturbing the preceding trajectory in a sample to create a trial trajectory (that is correlated and, in fact, typically partially coincident with the prior trajectory) which then is accepted or rejected according to a suitable Metropolis criterion. The sequence of trajectories is significantly correlated, and indeed rejections amount to having the same trajectory twice in the ensemble which is generated. In other words, there is a kind of “correlation number” n_{corr} (akin to a Monte Carlo correlation “time”) measuring the average number of trajectories which must be sampled before a new statistically independent trajectory is sampled. There is no reason why n_{corr} should be small and indeed, just as in a rough energy landscape, one can imagine that the effective landscape for paths is highly corrugated and requires significant sampling “time.” The bottom line is that our 100 independent trajectories will cost a total of $100 * n_{\mathrm{corr}} * t_b$ simulation time: one hopes that this will be less than the “brute force” cost of 100*MFPT! This sounds bad, but other approaches share similar limitations.

Consider the *weighted ensemble* strategy. Trajectories in the ensemble are run independently but occasionally pruned or replicated – and both operations intrinsically reduce information content and hence increase correlations. When a trajectory is pruned, the prior computing effort which generated it now gets wasted (at least partially). When a trajectory is replicated, say midway through the simulation, then the “daughter” or replica trajectories actually were the very same trajectory for half of their existence – and clearly correlated. So once again, there are significant correlations and we can again describe it with an effective correlation number n_{corr}. Whether these correlations are stronger or weaker in the two methods is not our concern here (and indeed would depend on the system and specifics of the implementation of the path-sampling algorithm).

The *dynamic importance sampling* strategy is strictly based on independent trajectories and so does not suffer from correlations … but it has its own challenges. Specifically, trajectories are biased and do not evolve according the correct physical dynamics. Although a probabilistic description of stochastic trajectories enables one to calculate a weight for each of the biased trajectories and thus correct for the bias, these weights degrade the statistical quality of the resulting trajectory ensemble. Specifically, the non-uniformity of weights guarantees that only a fraction of the trajectories (say, $1/n_w$, with $n_w > 1$) will contribute significantly to calculations of any observable, such as a rate. The size of n_w will depend on system and implementation specifics, but it’s clear the approach qualitatively suffers from sampling limitations analogous to the two other strategies we just discussed.

Bottom line: The cost per continuous transition trajectory is $n * t_b$, where $n \gg 1$ is an integer quantifying the efficiency of the path sampling method. Of course, experts in each method strive to reduce n but there are no guarantees for any challenging system.



Meeting the challenge of multiple intermediates

The issue of multiple intermediates (meta-stable on- or off-pathway states) is worth some additional discussion in the context of path sampling. Recall that in such a case, t_b ~ MFPT itself may seem prohibitive – at least, if we insist on having fully continuous trajectories.

The good news is that all three strategies described above can side-step the problem of intermediate states (as can a number of other approaches based on trajectory segments). One example is the non-Markovian post-analysis suggested by Suarez et al., but this post will not go into the details. Qualitatively, it turns out that the limiting timescale for path sampling is not t_b but a quantity we can call t_b^* which represents the sum of all the event durations for transitions among the intermediates – *excluding* the intermediate dwell times. This doesn't solve the problem of trajectory correlations or weights, but at least offers some hope for obtaining useful results.

The papers noted below are only a very small subset of the path sampling literature.

Further reading

Elber, R. "Perspective: Computer simulations of long time dynamics," The Journal of Chemical Physics, AIP Publishing, 2016, 144, 060901

Huber, G. A. & Kim, S. “Weighted-ensemble Brownian dynamics simulations for protein association reactions,” Biophys. J., 1996, 70, 97-110

Pratt, L. R. “A statistical method for identifying transition states in high dimensional problems,” J. Chem. Phys., 1986, 85, 5045-5048

Suárez, E.; Lettieri, S.; Zwier, M. C.; Stringer, C. A.; Subramanian, S. R.; Chong, L. T. & Zuckerman, D. M. “Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories,” J Chem Theory Comput, 2014, 10, 2658-266

Woolf, T. B. “Path corrected functionals of stochastic trajectories: towards relative free energy and reaction coordinate calculations.” Chem. Phys. Lett., 1998, 289, 433-441

Zuckerman, D. M. & Woolf, T. B. “Dynamic reaction paths and rates through importance-sampled stochastic dynamics.” J. Chem. Phys., 1999, 111, 9475-9484

Zuckerman, D. M. & Woolf, T. B. “Transition events in butane simulations: similarities across models.” J. Chem. Phys., 2002, 116, 2586-2591

Zwier, M. C. & Chong, L. T. “Reaching biological timescales with all-atom molecular dynamics simulations,” Curr Opin Pharmacol, Department of Chemistry, University of Pittsburgh, Pittsburgh, PA 15260, USA., 2010, 10, 745-752