Adaptive seeding with short trajectories for protein folding landscapes

Hongbin Wan and Vincent A. Voelz*

Department of Chemistry, Temple University, Philadelphia, PA 19122, USA

E-mail: vvoelz@temple.edu

Abstract

Abstract goes here!

Introduction

In the last decade, new Markov State Model (MSM) methodologies have made possible accurate and efficient estimation of kinetic rates and reactive pathways for slow and complex biomolecular dynamics, including protein folding, ligand binding, and conformational allostery. One of the key advantages touted by MSM methods is the ability to use large ensembles of short-timescale trajectories for sampling events that occur on slow timescales. If enough parallel sampling from short trajectories can be achieved, this can circumvent the need to sample long trajectories.

Many "adaptive" methods have been used for the purpose of accelerating sampling of MSMs. The simplest of these can be called *adaptive seeding*, where one or more new rounds of unbiased simulation are performed by initiating swarms of trajectories "seeded" throughout

*To whom correspondence should be addressed the landscape. The choice of seeds are based on some initial guess for the energy landscape, possibly from non-equilibrium or enhanced-sampling methods.

Another kind of adaptive method is *adaptive sampling*, where successive rounds of seeding are performed in order to better sample states of interest. A popular adaptive sampling strategy is to start successive rounds of simulations from under-sampled states, for instance, from MSM states with the least number of transition counts.[?] The FAST algorithm is an adaptive sampling method designed to encourage sampling of states with desired properties.[?] FAST chooses new seeds based on an objective function that balances under-sampling with some target structural observable. Weighted-ensemble (WE) path sampling algorithms can also be classified as adaptive sampling algorithms. In WE approaches such as WExplore and WESTPA, successive rounds of new trajectories are spawned to better sample a quantity of interest (reactive flux, for example), while the statistical weights of each trajectory are carefully managed so that no bias is introduced.

The above adaptive methods are unbiased, in the sense that each trajectory is sampling from the same, unbiased distribution. But the sampling is *statistically* biased, in the sense that the observed transitions occur with frequencies that do not correspond to the unbiased distribution. This statistical bias has interesting consequences in estimating various quantities, consequences which arise from the trade-off between sampling of transition rates versus equilibrium probabilities. For example, when building MSMs from trajectory data obtained by FAST, it is recommended that row-normalized transition counts be used to estimate the MSM transition matrix,[?] instead of the maximum-likelihood estimator (MLE) that is default in packages like MSMBuilder. This is because the MLE enforces detailed balance; i.e. MLE assumes that the observed counts are sampled from the equilibrium distribution, an assumption which is (purposely) violated by adaptive sampling. One might expect, then, that while quantities like rates and pathways are accurately estimated by adaptive methods, quantities like equilibrium populations may have more uncertainty and/or bias. Indeed, while the weighted ensemble algorithms have recently been used to efficiently sample very slow folding rates[?] ad unbinding rates (residence times > 10 s)[?] , these same algorithms

In this paper, we consider a specific kind of adaptive method for sampling MSMs: adaptive seeding of perturbed folding landscapes. It is often the case that large numbers of expensive simulations in a particular force field are utilized to model the folding landscape of a particular protein sequence. We would like to model the folding of a sequence variant, or perhaps use a different force field potential, without having to perform a heroic amount of simulation. Since there is much prior information from the "wild-type" MSM, it is reasonable to think that adaptive seeding could provide a good picture of how folding rates and populations change with the perturbation. Here, we explore the accuracy of several estimators for obtaining folding rates and populations from adaptive seeding simulations.

Using a 1-D two-state potential as simple model, we explore different estimators and find interesting differences in their relative accuracies in estimating rates versus equilibrium populations from adaptive seeding trajectory data. We also explore the effects of using different trajectory lengths and number of seeds. In general, we find that rates and free energies are more accurately estimated by estimators that incorporate some prior knowledge of the equilibrium populations. We then show how adaptive seeding can be used to model changes in folding rates and populations for GTT WW domain, based on a 1000-state MSM built from ultra-long simulation trajectories.

Results

Adaptive seeding of a 1-D potential energy surface

We consider the following potential energy surface, as used by Stelzl et al. (JCTC 2017): $U(x) = -\frac{2k_BT}{0.596}\ln[e^{-2(x-2)^2-2} + e^{-2(x-5)^2}]$ for $x \in [1.5, 5.5]$, and $k_BT = 0.596$ kcal mol^{-1}. The state space is uniformly divided into 20 bins to calculate discrete-state quantities. Diffusion on the 1-D landscape is approximated by a Markov Chain Monte Carlo (MCMC) procedure in which new moves are translations randomly chosen from $\delta \in [-0.05, +0.05]$ and accepted with probability $\min(1, \exp(\beta[U(x+\delta) - U(x)]))$, i.e. the Metropolis criterion.

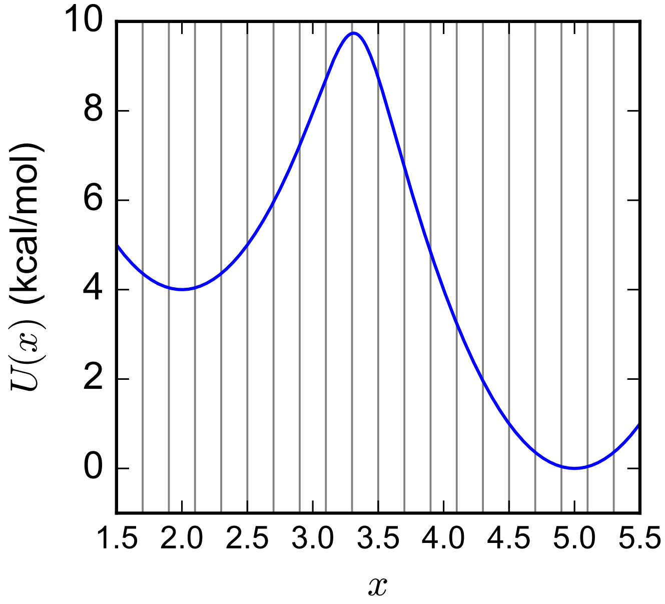

Figure 1: A 1-D two-state potential

.

For all tests with this toy model, we limit the lag time to $\tau = 100$ steps. To estimate the "true" relaxation timescale of the two-state model, we generated long MCMC trajectories of 10^9 steps, sampling from a series of scaled potentials $U^{(\lambda)}(x) = \lambda U(x)$ for $\lambda \in [0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$. For each λ value, 20 trajectories were generated, with half of them starting from $x = 2.0$ and the other half starting from $x = 5.0$, resulting in a total of 120 trajectories. The DTRAM estimator of Wu et al.[?] Wu:2014) was used to estimate the slowest relaxation timescale as 9.66 (± 1.37) $\times 10^6$ steps, using a lag time of 1000 steps.

To emulate adaptive seeding trajectory data, various numbers of trajectories s, each of length $n\tau$ were initiated from the center positions of all twenty bins. The resulting data consists of $20 \times s \times n$ transition counts between states i and j in lag time τ, stored in a 20×20 count matrix of entries c_{ij}. From these counts, estimates of the transition probabilities, p_{ij} can be made.

We explored the accuracy and efficiency of several different transition probability estimators using the adaptive seeding trajectory data as input: (1) a maximum-likelihood estimator (MLE), (2) a MLE estimator where the equilibrium populations π_i of each state are known *a priori*, (3) the MLE estimator where each input trajectory is weighted by the *a priori* equilibrium population of its starting point, and (4) row-normalized transition counts.

1. Maximum-likelihood estimator (MLE). The MLE for a reversible MSM assumes that observed transition counts are independent, and drawn from the equilibrium distribution, so that reversibility (i.e. detailed balance) can be used a constraint. The likelihood of observing a set of given transition counts, $L = \prod_i \prod_j p_{ij}^{c_{ij}}$, when minimized under the constraint that $\pi_i p_{ij} = \pi_j p_{ji}$, for all i, j, yields a self-consistent expression that can be iterated to find the equilibrium populations,[? ? ?]

$$\pi_i = \sum_j \frac{c_{ij} + c_{ji}}{\frac{N_j}{\pi_j} + \frac{N_i}{\pi_i}} \tag{1}$$

where $N_i = \sum_j c_{ij}$. The transition probabilities p_{ij} are given by

$$p_{ij} = \frac{(c_{ij} + c_{ji})\pi_j}{N_j \pi_i + N_i \pi_j} \tag{2}$$

2. Maximum-likelihood estimator (MLE) with known populations π_i Minimization of the likelihood function above, with the additional constraint of fixed populations π_i, yields a similar self-consistent equation that can be used to determine a set of Lagrange multipliers,

$$\lambda_i = \sum_j \frac{(c_{ij} + c_{ji})\pi_j \lambda_i}{\lambda_j \pi_i + \lambda_i \pi_j}, \tag{3}$$

from which the transition probabilities p_{ij} can be obtained as

$$p_{ij} = \frac{(c_{ij} + c_{ji})\pi_j}{\lambda_j \pi_i + \lambda_i \pi_j}. \tag{4}$$

Maximum-likelihood estimator (MLE) with population-weighted trajectory counts. For this estimator, first a modified count matrix c'_{ij} is calculated,

$$c'_{ij} = \sum_k w^{(k)} c_{ji}^{(k)}, \tag{5}$$

where transition counts $c_{ji}^{(k)}$ from trajectory k are weighted in proportion to $w^{(k)} = \pi^{(k)}$, the equilibrium population of the initial state of the trajectory. The idea behind this approach is to counteract the statistical bias from adaptive seeding by scaling the observed transition counts proportional to their equilibrium fluxes. The modified counts are then used as input to the MLE (estimator 1).

Row-normalized counts. For this estimator, the transition probabilities are approximated as

$$p_{ij} = \frac{c_{ij}}{\sum_j c_{ij}}. \tag{6}$$

This approach does not guarantee reversible transition probabilities, which only occurs in the limit of large numbers of reversible transition counts. In practice, however, the largest eigenvectors of the transition probability matrix have very nearly real eigenvalues, such that we can report relevant relaxation timescales and equilibrium populations.

Methods

Conclusions

Acknowledgments

The authors thank the participants of Folding@home, without whom this work would not be possible. We graciously acknowledge D. E. Shaw Research for providing access to the WW domain folding trajectory data. This research was supported in part by the National Science Foundation through major research instrumentation grant number CNS-09-58854 and National Institutes of Health grants 1R01GM123296-01 and NIH Research Resource Computer Cluster Grant S10-OD020095.

Supporting Information

Figures S1-S15. This information is available free of charge via the Internet at `http://pubs.acs.org`

Appendix

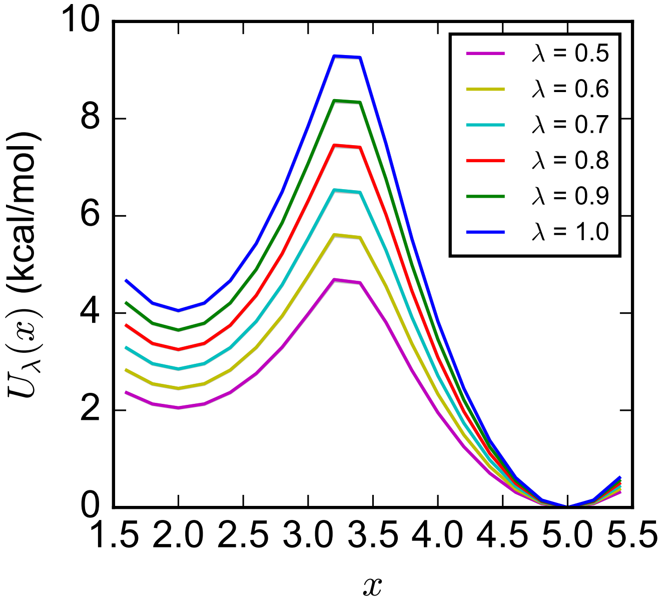

Figure 2: Free energies from dTRAM

.