JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2017, 57, 2729-2740

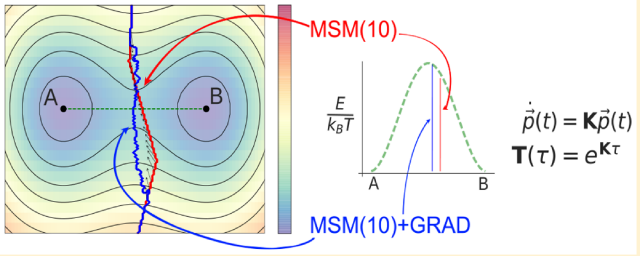
pubs.acs.org/jcim

GRadient Adaptive Decomposition (GRAD) Method: Optimized Refinement Along Macrostate Borders in Markov State Models

P. G. Romano and M. G. Guenza*

Department of Chemistry and Biochemistry, and Institute of Theoretical Science, University of Oregon, Eugene, Oregon 97403

ABSTRACT: Markov state models (MSM) are used to model the kinetics of processes sampled by molecular dynamics (MD) simulations. MSM reduce the high dimensionality inherent to MD simulations as they partition the free energy landscape into discrete states, generating a kinetic model as a series of uncorrelated jumps between states. Here, we detail a new method, called GRadient Adaptive Decomposition, which optimizes coarse-grained MSM by refining borders with respect to the gradient along the free energy surface. The proposed method requires only a small number of initial microstates because it corrects for errors produced by limited sampling. Whereas many methods rely on fuzzy partitions for proper statistics, GRAD retains a crisp decomposition. Two test studies are presented to illustrate the method and assess its accuracy: the first analyzes MSM of idealized model potentials, while the second is a study of the dynamics of unstacking of the deoxyribose adenosine monophosphate dinucleotide.



INTRODUCTION

In recent years Markov state models (MSM) have emerged as a favorite method to extract kinetics information from simulation trajectories. MSM map the dynamics of a complex kinetic process as a series of uncorrelated jumps over a network of discrete states, thus reducing the kinetic process to a random walk in configurational space.[1−4] This assumption simplifies the kinetic formalism into a master equation (ME),[5] which is solved by spectral decomposition, and provides information on equilibrium populations, system time scales, and the dominant pathways between states. The method is simple and mathematically elegant. Thus, the proper solution of the kinetic problem resides in the correct identification of the energetic states that are sampled by a random walk in configurational space.[2]

In general, complex kinetic pathways do not follow a random walk in phase space, except in either the trivial time scale of the simulation step where dynamical events are uncorrelated, or in the infinitely large time scale of diffusion. At all intermediate time scales, kinetic processes are in general correlated as can be easily seen, for the case of protein motion, when considering the fractal nature of the free energy landscape and how it shapes protein conformational dynamics and folding.[6] Dynamical correlation naturally emerges in the kinetic formalism from the process of coarse-graining the microscopic dynamics into a mesoscopic time scale, which formally corresponds to the mathematical operation of applying the projection operator technique. This process leads to an equation for the kinetics, which includes a memory function that, in principle, cannot be discarded.[7,8]

However, if the dynamics of a process can be clearly separated into fast and slow contributions, which occurs when the configurational states are separated by high energy barriers, memory contributions become negligible and can be safely discarded. Then, the system slowly transitions between macrostates, while the faster sampling of local states occurs inside the configurational space of those macrostates. In those cases, the walk between macrostates becomes Markovian and the kinetics of the process is easily analyzed by MSM, if the time step of the sampling is larger than the time step of the fast transition inside the state and smaller than the slow transition between macrostates.

Thus, applying MSM depends on correctly solving the problem of identifying a number of macrostates that are separated by well-defined and large energy barriers.[9] In practice, the MSM procedure starts from the analysis of a long computer simulation trajectory or, equivalently, the analysis of a large number of short simulation trajectories, and identifies a number of possible energetic minima. The procedure then counts how many times the system rapidly transitions between states and groups, through the analysis of the eigenvalues of the transition matrix in the ME, those fast interconverting states into one macrostate. The slow kinetics emerging from the diagonalization of the transition matrix identifies the Markovian kinetic path between macrostates.

For the step of grouping together kinetic states, it is important to define without ambiguity whether a point in the trajectory belongs to one state or its adjacent one. It is sometimes the case that the membership of a configurational point to a state is ambiguous, for example in the regions where the energy landscape is somehow featureless, or in the regions where the statistical errors of the simulations, or lack of sufficient sampling, gives a rough landscape with no well-defined, large, energy barriers. Thus, devising methods that can

Received: May 9, 2017
Published: October 16, 2017

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

be used to study and define with accuracy the border between metastable states can be useful in solving this kinetic problem. The relevance of an accurate determination of energy barriers in configurational space has been stated in a number of papers in the recent literature.[10,11]

In this paper we present a method, called the GRadient Adaptive Decomposition method or GRAD, which provides an accurate definition of the energy barriers and related borders, even when the number of centroids, or seeds, used to define the initial state in MSM is small. The method uses a noise-filtering procedure to smooth the energy landscape, reducing the roughness due to errors, and determines the barriers and related border by directly bringing to the method information about the smoothed slope of the free energy landscape, during the refinement step.

Effectively, the GRAD method finds the separation between states by maximizing the time scales associated with metastable states. It adaptively refines the position of state barriers by randomly sampling microstates along the border wall, a less costly alternative than splitting the full configurational space, and then lumping each microstate in the direction of free energy surface gradient. Newly predicted barriers are accepted on the condition that the system metastability, or probability of simulation data remaining within a state over a given lag time, has increased.

In its present version, the method is developed for two dimensions plus the energy; its generalization to higher dimensions is possible, even if not strictly necessary. For many dynamical problems of interest, simulation trajectories can be conveniently projected onto a low-dimensional space before the MSM procedure is applied. More in general, in the case of protein dynamics, the trajectories can be easily reduced to lower dimensions by applying a principal component analysis (PCA) or the time-lagged independent component analysis (tICA),[12,13] or by using the mode decomposition of the Langevin equation for protein dynamics (LE4PD) that we have recently proposed.[8] Because the GRAD method is integrated with the MSM approach, it shares most advantages and limitations with that method. For example, in principle, there is no limitation on the number of macrostates that can be considered. However, the GRAD method has the potential of improving accuracy in undersampled MSM.

Here, we illustrate the GRAD method and the accuracy of its predictions with a number of test calculations where the energy barriers and their adjacent basins are well-defined. In our examples the barrier can be symmetrical or not symmetrical. We also present the predictions of the method for a small test molecule, the deoxyribose adenosine dinucleotide monophosphate (ApA). For this molecule, we present a simple case of dimensionality reduction as the simulation trajectories of ApA are conveniently mapped onto two coordinates, relevant for the study of breathing fluctuations in DNA. In all cases the GRAD refinement method appears to be useful in improving the accuracy in the calculation of the kinetics.

BRIEF OVERVIEW OF MARKOV STATE MODELS: FROM MICROSTATES TO MACROSTATES

In this section, we briefly describe the current methodology to generate micro and macro states, as well as to calculate characteristic times in kinetics pathways, using Markov state models. We focus on the information necessary to understand the proposed procedure, while for a more detailed discussion of the MSM, we refer to the literature.[14−17] A reader familiar with the MSM procedure should feel free to skip this section without fear of missing some essential information.

Markov state models follow a ME formalism,[18] which describes the molecular kinetics of a process as a Markov-chain of uncorrelated jumps among conformational states. This formalism can be easily described by $\dot{\mathbf{p}}(t) = \mathbf{K}\mathbf{p}(t)$, where $\mathbf{p}(t)$ is the population state vector at time t, $\mathbf{K}$ the kinetic rate matrix, and the dot denotes the differentiation with respect to time. As the system is assumed Markovian, it holds that $\mathbf{p}(n\tau) = [\mathbf{T}(\tau)]^n\mathbf{p}(0)$, where τ is the system lag time to satisfy the Markov condition and T is the matrix of condition probabilities to transition between all states. From these two equations it is simply shown that the transition and rate matrices are related by $\mathbf{T}(\tau) = e^{\mathbf{K}\tau}$.

MSM is based on the Markovian discretization of the configurational landscape into states that need to be kinetically independent. By partitioning continuous energy landscapes into discrete states, and by the projection of the continuum trajectory onto a discrete trajectory along the macrostates, the method introduces discretization errors.[17] To improve the quality of the prediction of the MSM, an automatic decomposition procedure is introduced to carefully tune the number of initial microstates in which the energy surface is initially partitioned.[18] If the number of microstates is too small undersampling can inaccurately identify the position of the barrier between metastable states. If the number is too high, i.e. oversampling, this can be equally detrimental as it can lead to "overfitting", wherein the procedure fits the errors present in the simulated energy landscape instead of the real border.[9]

In order to ensure exhaustive sampling of the energy landscape, either an extensive number of short trajectories or a small number long trajectories are calculated by performing MD simulations. The trajectory is clustered into microstates by performing k-means++.[19] This method improves the standard k-means solution,[20] where centroids minimize center of mass of all nearest neighbors, by carefully choosing seeds.[21] Thus, the multidimensional free energy landscape is first "seeded" by the random generation of the centroids. K-means++ places an additional weight on the acceptance of centroids by the squared distance from the closest centroid. In this way, the procedure tries to sufficiently sample all the regions in the configurational landscape. The precision of the method increases with increasing the number of centroids. The final results partitions the configurational space in a number of states, equal to the number of initial seeds.[18]

In centroid based algorithms, such as k-means++, k-medoids, k-centers, etc., a large number of centroids increases the accuracy of the discretization. It is common to use a few thousands microstates to sample the free energy surface. However, the computational time of MSM increases with the number of sampling centroids, and further increases the computational time for analysis due to the diagonalization of the transition matrix. The latter has dimensions equal to the number of centroids and becomes sparse when the number of centroids is high as the transition between some of the possible states has low probability.

The properties of the MSM are determined by the ME transition matrix, $T_{ij}(\tau)$, which is defined as the conditional probability for a trajectory to enter state j from i over a given time interval τ, also called *lag time*. This is directly estimated from simulation data by counting the observed transition, $C_{ij}(\tau)$. Due to thermal noise or limited sampling, simulations

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

are rarely perfectly reversible, $C_{ij}(\tau) \neq C_{ji}(\tau)$. This can be corrected by enforcing detail balance[9,18] according to

$$\bar{C}_{ij}(\tau) = \frac{C_{ij}(\tau) + C_{ji}(\tau)}{2} \quad (1)$$

The ME transition matrix is then calculated by row normalization

$$T_{ij}(\tau) = \frac{\bar{C}_{ij}(\tau)}{\sum_k \bar{C}_{ik}(\tau)} \quad (2)$$

The procedure to build the transition matrix works well in the limit of infinite sampling, or more practically when the length of the simulations is considerably greater than the predicted time scales of the transition matrix. Commonly, this is not the case and a maximum likelihood estimator as described by Prinz et al. can be employed.[16]

Once the MSM is generated and the ME transition matrix is estimated for a given lag time τ, the model can be evaluated directly by evaluating time scale and metastability according to the properties of $\mathbf{T}$. As the metastability can be thought of how long-lived a state is, this can be quantified as the sum of the diagonal elements of the transition matrix

$$M = \sum_i T_{ii} \quad (3)$$

The time scales of the model can be evaluated according to the Chapman−Kolmogorov[22] (CK) condition

$$\mathbf{T}(n\tau) = [\mathbf{T}(\tau)]^n \quad (4)$$

for integer n intervals of lag time τ. If the trajectory follows a random walk in configurational space, taking n steps with lag time τ is equivalent to taking one step with lag time $n\tau$.

It is reasonable to assume that the system becomes Markovian at large enough τ, as all kinetic events become uncorrelated if they are sampled at times that differ by a interval larger than their correlation time. Fulfilling the CK equation ensures that the relaxation time of a process is independent of the number of uncorrelated steps that are used to model the process. The implied relaxation times for process i is given according to the eigenvalue, λ_i as

$$t_i = \frac{-\tau}{\ln \lambda_i} \quad (5)$$

By analyzing the behavior of t_2 as a function of the lag time τ it is possible to identify the time lag at which the dynamics becomes Markovian.

Coarse-graining is commonly performed by PCCA+ analysis.[4,15−18] However, several coarse-graining schemes exist such as the Bayesian hierarchical[23] and the transition path clustering.[11,24] Accurate coarse-graining still remains an active field of study.[25] The number of macrostates is predicted by the eigenvalues of the transition matrix by identifying "gaps in time" in the list of eigenvalues. PCCA+ employs the structure within the right eigenvectors of this spectral decomposition to assign a fuzzy membership likelihood[26] to each microstate to belong within a set of metastable states, or macrostates. The microstates are clustered as macrostates for which the relevant dynamical observables, for example the metastability M or relaxation time t_2, are measured. These quantities have been used as variational parameters for coarse-graining,[27−29] because they quantify how long-lived is a macrostate in its coarse-grained representation, which correlates to its degree of metastability.

Adopting a fuzzy overlap amounts to assigning each microstate to a number of macrostates with a weighted distribution. As a microstate can belong to a number of different macrostates, this provides a realistic evaluation of the transition time scale. While this is useful from a theoretical standpoint it can result in a large uncertainty along the macrostate borders. Any conformational analysis of transition states can therefore lose chemical insight. For this reason, crisp partitioning is desirable, as it retains molecular information at minima and along the barriers.

The use of a crisp partition is not incompatible with transition path models. In fact, once the paths that describe a transition from one state to another are identified, conformation along the path need to be assigned to either the initial or the final state. If the barrier in the path is crisply defined, the assignment of configurations to either states becomes unambiguous, and the calculation of the transition time by means of the master equation should become more accurate. It is reasonable to think that GRAD applied to transition path models could give consistent answers to the maximum-likelihood propagator-based method, because the identification of the correct position and height of the energy barrier would improve the accuracy in the calculation of the transition time and maximize its value.[11]

In the traditional MSM, "refinement" is performed by generating repeated MSM where the number of initial seeds or microstates is increased. If the separation in macrostate from this new calculation is consistent with the previous one, and the time t_2 converges to the same value, then the procedure is terminated. If agreement is not achieved, the procedure is repeated with an increased number of initial microstates, until the process converges.

Because the number of necessary microstates cannot be known *a priori*, multiple runs of microstate clustering are necessary to adequately define what number of microstates are sufficient to generate a model with low discretization error.[30] This "refinement" procedure is, however, computationally quite expensive. Currently, several thousand to tens of thousands centroids are typically used to sufficiently sample trajectories. However, having more than 4000 centroids requires the use of sparse linear algebra methods to reduce computational complexity, severely affecting the computational time needed to perform the MSM analysis.

METHOD TO REFINE MACROSTATE BORDERS: GRAD ALONG MACROSTATE BORDERS

In the procedure that we propose, i.e. the GRadient Adaptive Decomposition, the number of initial microstates is small, and accuracy is achieved by the iterative refinement of the macrostates along their borders. This is computationally convenient because it limits the number of microstates and localizes the refinement procedure only in a sub area of the free energy space. All analysis codes for GRAD were written in Python[31] using NumPy and SciPy.[32]

In the proposed GRAD procedure, the number of initial microstates is small, and accuracy is achieved by the iterative refinement of the macrostates along their borders. This is computationally convenient because it limits the number of microstates and localizes the refinement procedure only in a sub area of the free energy space.

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

The main goals of our proposed refinement method is to build an MSM with two criteria: (1) to reduce the number of centroids used, (2) while retaining a crisp partitioning of the energy barrier. By refining macrostates to minimize discretization error, in particular reducing error from underfitting the kinetic model, fewer microstates are necessary as initial input. Additionally, as the refinement identifies the barriers between metastable states without resorting to convex hull about microstate centroids, we can ensure an accurate, crisp decomposition. The proposed method refines the macrostate borders iteratively by generating microstates along the border between macrostates. Microstates are assigned to metastable states by lumping in the direction of the gradient along the free energy landscape. In the following section, we introduce and detail the GRAD along macrostate borders refinement method. The procedure maximizes the system metastability while limiting the number of initial centroids necessary in the microstate generation.

In order to retain information from both the free energy landscape as well as the Markov states, a lattice map is made that bins the conformational landscape and assigns a state from the coarse-grained MSM. The free energy surfaces are estimated from the probability along kinetically relevant parameters, X as

$$\Delta G(X) = -k_B T \ln(P(X)) \tag{6}$$

While lattice methods are most commonly developed for a few dimensions, due to the computational costs and extensive memory allocation, the constraint of our method to low dimensions is not a real limitation. The use of dimensionality reduction procedures, such as the principal component analysis (PCA) or the time-lagged independent component analysis (tICA),[12,13] is a convenient strategy to reduce the number of relevant variables before MSM are applied. Identifying either the largest variance or the slowest kinetic processes through PCA and tICA, respectively, has been shown to be a valid method to reduce the dimensionality of a kinetic process in MSM.[30,33] An alternative procedure could be to start from a dynamical mode decomposition as performed by LE4PD.[8] Alternatively, for smaller molecular systems the kinetic process can be described easily by one- or two-dimensional ordered parameters, properly selected, as illustrated in the ApA example here.

Splitting the Metastable Border. The GRAD procedure begins by generating new microstates exclusively along the internal wall between macrostates. We refer to these as *microborders* so as to avoid confusion with the initial microstate seeding. Microborders are generated following a multistep procedure. First, the internal walls of a specific macrostate are padded with a region as shown in Figure 1. We select the shortest axis in the area of the macrostate and define the padding length as a percentage of this axis length. We typically use the small ratios of 0.01, 0.001, and 0.0001 over the course of the refinement to help ensure convergence.

Large width ratios allow for exploration of the surface, whereas small width ratios allow for more detailed refinement. Once the padded region is built, centroids are randomly initialized and microborders are built as a Voronoi cell. Because standard uniform distribution methods are typically plagued by issues of densely packed centroids which create microborders of varying shapes and sizes, centroids are placed using a Poisson disk distribution.[34]

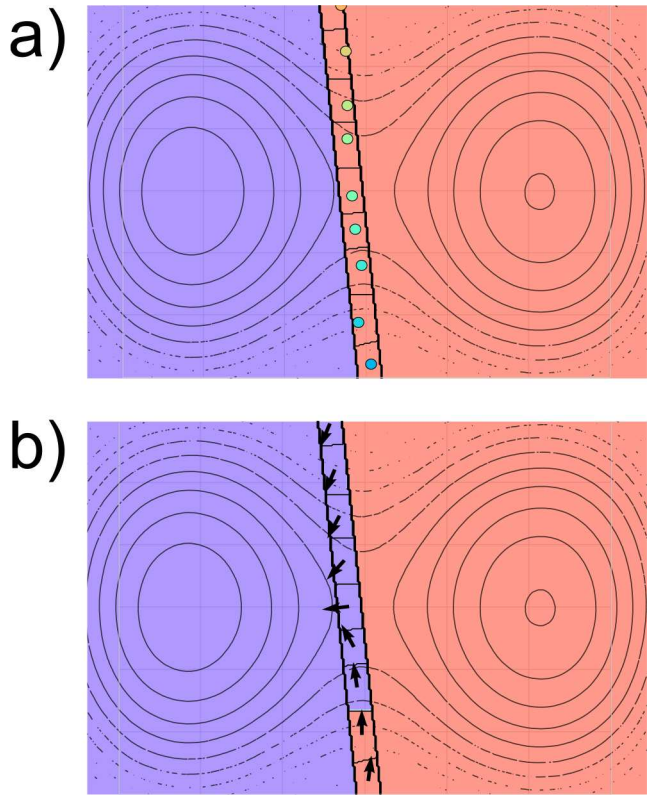


Figure 1. (a) Microborders generated along the wall of an arbitrarily defined macrostate at a fixed padding length. Each microborder is shown with a centroid, as predicted by the poisson disk method. (b) Each microborder clustered to a macrostate (denoted by color) assigned by the direction along the mean gradient within each microborder.

Poisson disk generation uniformly placed centroids with "blue noise" characteristics (avoiding aggregation of the centroids on the surface), where each new centroid maintains a minimum radial separation from all previously accepted centroids.[35] For our method we define the radial separation between centroids as being two times larger than the padding length; we add the extra criteria that all centroids must fall within the padded region. This creates microborders with even, convex shapes and automatically determines the number of microborders necessary to fill the padded region (see Figure 1).

While other centroid generation methods exists, the computational time of this Poisson method scales linearly with the total number of centroids. Disk generation enforces that the microborder centroids have uniform density by evaluating whether any centroids are within a radial disk from a newly proposed centroid. While the centroids are still randomly generated, they have a minimal radial separation between all other centroid positions. Without such a restriction, there would be no guarantee that the contributions from each microborder would refine at the same length scale. Changes in the configurational decomposition would therefore be unevenly weighted, creating nonsmooth, and even nonphysical, division between states.

Assigning Microborders to Macrostates. The discrete states generated are stored along a lattice map at a fixed number of grid-points, allowing for trivial one-to-one mapping of surface properties such as the free energy surface (FES). As such, microborders can be clustered to corresponding macrostates by using the gradient along the FES to identify barrier

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

maxima. The mean gradient along the FES within each newly generated microborder is therefore used to determine in what direction the microstate should be regrouped across the border (see Figure 1). The method *appends* to the standard MSM workflow. While the traditional MSM repeats the procedure while increasing the number of centroids until the predicted time scales converges, our proposed method selects a small number of centroid and then refines the borders between macrostates until the metastability is maximized.[18]

Each microborder is assigned to a neighboring macrostate by lumping the microborder in the direction of the gradient along the free energy surface. This pushes the border between macrostates closer to the barrier and reduces the discretization error from the initial MSM. Within each microborder, the mean gradient along the energy surface is computed by a numerical gradient.[36] Because these numerical methods can have error at low numbers of grid-points, and MD simulations often produce sparse regions of poor sampling, a 2D Savitzky−Golay filter[37] is applied to the energy surface prior to calculating the gradient. This filter reduces statistical noise, improves the signal-to-noise ratio, and ensures that the gradients are accurately calculated even in the presence of sparse or noisy sampling (see Figure 2).

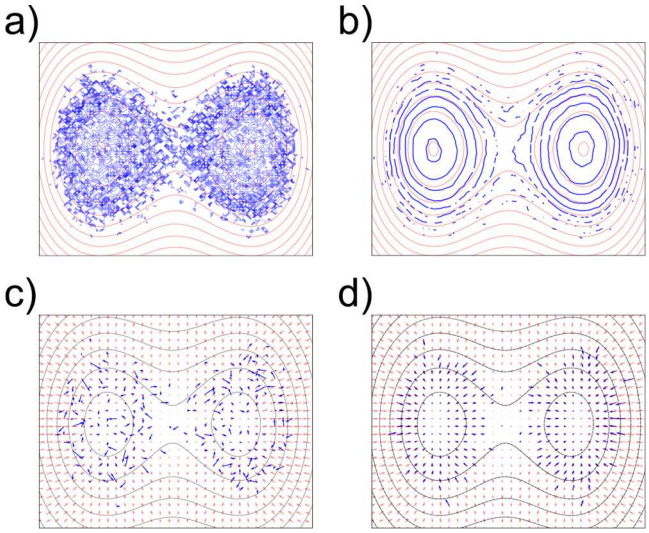


Figure 2. Free energy calculated from a single diffusion simulation along a symmetric two-well potential. (left to right) Smoothing process by 2D Savitzky−Golay filter on the energy calculated from the simulation trajectory (top panels), and as well on its gradient (bottom panels). Red lines and vectors are calculated by an analytical function (noise free), while blue lines and vectors are from simulated data.

The Savitzky−Golay filter has two critical parameters, the first is the size of the window used to fit the data with a polynomial function, and the second is the degree of the polynomial that fits the data. As the surface has a fixed number of points along the lattice, the window size is kept fixed as 10% of the total number of grid-point.The order of the polynomial is determined by direct inspection of a number of critical regions in the energy map. The goal is to obtain a smoothed surface without changing the characteristic shape of the FES, while improving continuity in the gradients (see Figure 2). It is noteworthy that the time required to calculate the filtered FES with our code is orders of magnitude faster than other computational steps in the MSM+GRAD workflow. It follows that defining the parameters by direct inspection in practice does not significantly increase the computational time of the MSM+GRAD procedure. The method smooths the structure in the data and cannot correct for missing features along the conformational topology. However, we have found that gradient filtering with Savitzky−Golay works well with sets of at least 10 000 data-points so long as transitions are representatively sampled (data not shown).

As stated earlier, the border between macrostates is moved in the direction of the mean gradient. The direction is defined from the median centroid of a microborder along the state space lattice using the Bresenham algorithm.[38] Because the method maps along a grid, the line drawn needs to be restricted to positions along the lattice. The Bresenham algorithm "rasterizes" a line along a lattice (i.e., represents a line by the shortest corresponding path on the lattice) by moving stepwise to discrete points that minimize the error away from the line. The rasterized line is drawn until it reaches a point outside the microborder, and the microborder is clustered into either the macrostate from which it originated or into a neighboring macrostate (see for example Figure 1).

Once all microborders within a macrostate have been clustered, the metastability is computed. If the metastability increases with respect to the previous iteration and within an established threshold, the new arrangement is accepted; otherwise, it is rejected, and a new macrostate is selected. A single iteration in the refinement scheme ends when all macrostates, selected in random order, have been refined.

The procedure terminates when the change in metastability for a complete iteration becomes smaller than a pre-established threshold, e.g. for step i we test that

$$0 \leq [M_{i+1}(\tau) - M_i(\tau)]/M_i(\tau) \leq 10^{-6} \tag{7}$$

This ensures that all macrostates are refined and that convergence is met not for individual, but for all states.

Because the GRAD method is based on the calculation of the free energy landscape from simulation trajectories, which have usually noisy and sparse sampling at the barrier, it is important to use an iterative procedure that minimizes the error to find where each microstate belongs.

TEST OF GRAD ALONG MACROSTATE BORDERS WITH MODEL POTENTIALS

In order to illustrate the capabilities of our refinement method, we performed a number of test simulations where a particle was free to diffuse on a free energy landscape. We selected symmetric and asymmetric double and triple-well potentials. The tridimensional potential surface is described by a sum of N elliptical Gaussian functions

$$V = -10k_BT \sum_{i}^{N} \exp[-a_i(x - x_{0,i})^2 - 2b_i(x - x_{0,i})(y - y_{0,i}) + c_i(y - y_{0,i})^2] \tag{8}$$

where N is also the number of minima in the potential, k_B is the Boltzmann constant, and T is the temperature. The coefficients of the potential are defined as

$$a_i = \cos^2(\theta_i)/(2\sigma_{x,i}^2) + \sin^2(\theta_i)/(2\sigma_{y,i}^2)$$
$$b_i = -\sin(2\theta_i)/(4\sigma_{x,i}^2) + \sin(2\theta_i)/(4\sigma_{y,i}^2)$$
$$c_i = \sin^2(\theta_i)/(2\sigma_{x,i}^2) + \cos^2(\theta_i)/(2\sigma_{y,i}^2) \tag{9}$$

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

In eq 9 σ_x and σ_y represent standard deviation along the x and y axis respectively; and the parameter θ is the angle by which the axes are rotated with respect to the x axis.

Each diffusion model defines a test case, where the number of macrostates is well-defined as the number of minima in the free energy surface. By easily tuning the complexity of the potential landscape, we were able to directly evaluate the accuracy of the MSM generated with, and without, our refinement procedure.

The single particle diffusion was modeled by a Langevin equation

$$\vec{r}(t + \Delta t) = \vec{r}(t) + \frac{\Delta t}{\gamma}\left(-\frac{\partial V}{\partial \vec{r}}\right) + \vec{r}_{\text{random}} \tag{10}$$

where $\vec{r} = \{x, y\}$, with γ as the friction coefficient, and $\vec{r}_{\text{random}}$ as a random displacement obeying the white-noise fluctuation−dissipation condition. For simplicity, we reduced the energy scale such that the simulated particle had thermal energy, $k_B T = 1$. All the simulations were performed for one million time steps, and were repeated to allow the initialization from all possible minima. The potentials that were studied are (i) a symmetric two well, (ii) an asymmetric two well, (iii) a symmetric three well, and (iv) an asymmetric three well.

For all diffusion simulations, initially a MSM analysis was generated using 10 microstates using the KMeans package in Scikit-Learn,[39] and the trajectories were clustered by PCCA+ into macrostates from PYEMMA,[40] determined by the number of well minima. This 10-centroid MSM analysis is refined following two different procedures. In the first, we used an MSM performed with an increased number of centroids. In the second, we keep the 10-centroid MSM as the starting system, and we refine this model using the GRAD method. The purpose is to assess how the shape and symmetry of the potential affects the accuracy of the GRAD procedure compared to MSM with increased sampling.

Using the conventional MSM method, we first performed calculations with microstates generated using k-means++ with 10, 100, and 1000 centroids for the four potentials presented in Figure 3. As a shorthand notation we identify each MSM analysis of the simulation trajectory as MSM_n^m, where n is the number of centroids in the microstate generation, and m the number of macrostates used to coarse-grain the model. Then, to test the GRAD method, we started from the MSM_n^m and refined the borders following the procedure described in the previous section. Refinement for all diffusion potentials were carried out until reaching the convergence of the metastability parameter, eq 7, while the calculations were performed using the trajectories from the diffusive simulations for the four potentials presented in Figure 3.

In order to evaluate how closely the proposed refinement method corrects under-sampling of the MSM, we compared in Figure 4 the refined MSM_{10}^2+GRAD (panel *d*), with the MSM_{10}^2 (panel *a*) as well as with the MSM_{100}^2 (panel *b*) and with the MSM_{1000}^2 (panel *c*) models. Each panel in the figure displays how the MSM partitions the energy surface into macrostates, given a fixed number of initial centroids, or microstates.

Panels a−c in Figure 4 show the conventional MSM microstates grouped by PCCA+ into macrostates, after convergence to the Markovian statistics for the symmetric two-well potential. The microstates are delimited by black lines, while the macrostates are shown as the blue and the purple areas to illustrate their MSM border. At the lag time τ selected for this figure, the macrostate are optimized and do not show further modification of their areas.

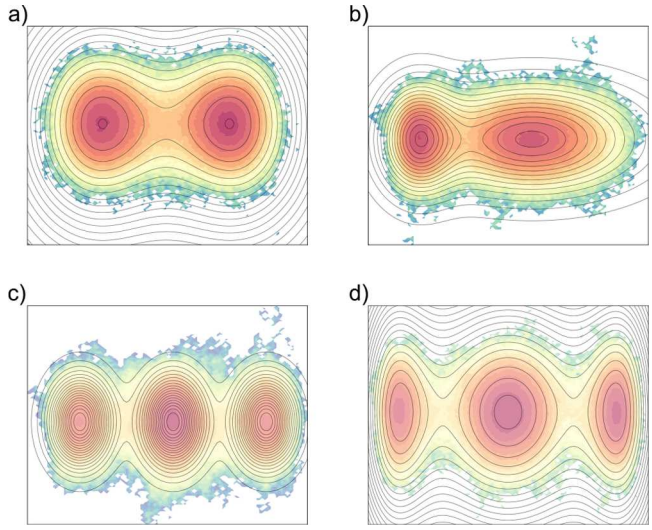


Figure 3. Free energy surface of the four model potentials: (a) symmetric two well, (b) asymmetric two well, (c) symmetric three well, and (d) asymmetric three well. The free energy surface calculated from the analytical equation is shown as smooth contour lines, while the free energy sampled by the diffusive simulation trajectory is shown as filled contour surfaces. As the energy scale increases from red to blue, the figure indicates that the diffusive simulations preferentially sample the states with lowest energy.

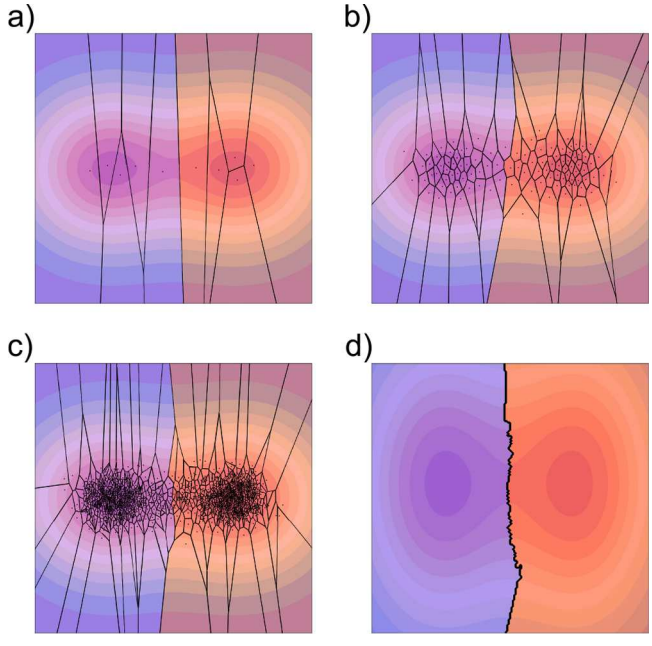


Figure 4. Panels display the process of identifying the energy barrier for a symmetric two-well potential. The vertical barrier is located at the center between the two minima. MSM for (a) MSM_{10}^2, (b) MSM_{100}^2, and (c) MSM_{1000}^2 centroids where black lines represent the crisp borders of microstates and each fill color (purple and red) denotes assignment in macrostates. (d) Macrostate MSM initialized by 10 centroids and refined with GRAD. The barrier predicted by MSM_{10}^2 is clearly on the right of the correct position. This is slightly improved in the MSM_{100}^2 panel where, however, the straight barrier is approximated by a fragmented, irregular pattern. The agreement is improved in the MSM_{1000}^2 sample and even more so in the MSM initialized by 10 centroids and refined with GRAD calculations.

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

The initial number of centroids in which the free energy surface is partitioned is MSM^2_{10} in the top left panel, MSM^2_{100} in the top right panel, and MSM^2_{1000} in the bottom left panel. Because of the analytical structure of the potential, the exact border is well-defined and it is given by a straight vertical line exactly positioned at equal distance between the minima in the two wells. The figure shows that by increasing the number of microstates the resolution of the energy border between macrostates improves. The last, bottom-right, panel shows the two macrostates obtained from MSM initialized with 10 microstates (MSM model of panel a) and refined with the GRAD along macrostate borders procedure. Even in the under sampled limit of our MSM^2_{10}+GRAD model, the refinement of the border leads to a precise definition of the border between macrostates. Similar results are obtained for all the three other potential shapes.

The comparison between the MSM^m_{10}+GRAD and the MSM^m_{1000} is shown, for all potentials, in Figure 5. Specifically,

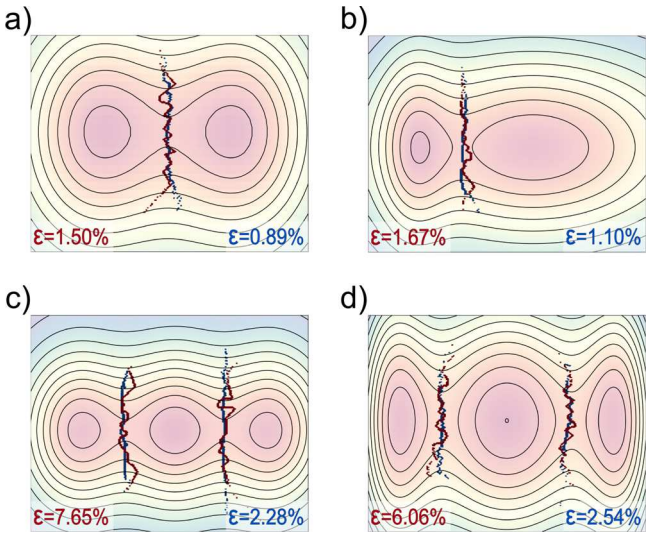


Figure 5. Illustration of the refined decomposition of the conformational space into macrostates for diffusion models (a) symmetric m = 2, (b) asymmetric m = 2, (c) symmetric m = 3, and (d) asymmetric m = 3. Lines represent the crisp partition between metastable states predicted from 1000 centroids (red) and refined with GRAD from 10 centroids (blue). The error, ϵ, reported is the mean squared error predicted via harsh boolean metric against the analytical barrier, for MSM^m_{1000} (red, bottom left) and MSM^m_{10}+GRAD (blue, bottom right). The predictions of the MSM^m_{10}+GRAD method are more accurate than those of MSM^m_{1000}.

the figure shows, for each potential, how the free energy surface is decomposed in m macrostates for the two refinement procedure (MSM^m_{1000} and MSM^m_{10}+GRAD). The border between macrostates is depicted in blue for MSM^m_{10}+GRAD refinement method and in red for the MSM^m_{1000}. The demarcation line is crisper for the GRAD refinement method with low centroid number for both symmetric and asymmetric potentials, with two or three wells.

To further test the precision of the GRAD refinement method versus the MSM with sufficient microstate sampling we evaluated how accurately the barriers between macrostates were reproduced according to the analytical potential for which the true gradients along the barrier is easily calculated. First we evaluated the mean squared error between MSM_{1000} and the analytical barrier of the potential, using a harsh boolean metric, where all matching points along the lattice receive a score of 0, and all differences a penalty of 1. The error is reported, as the mean squared error over all sampled positions, at the bottom of each panel in Figure 5.

Figure 6 reports the error per accepted iteration of the MSM_{10}+GRAD method, and, as a horizontal red line, the error

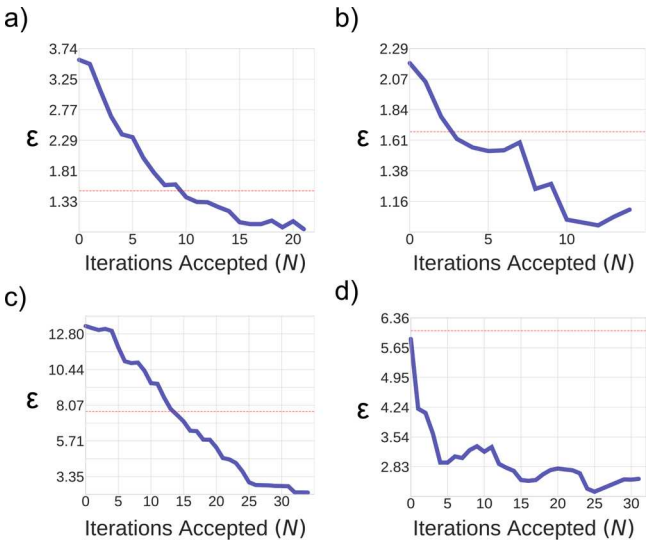


Figure 6. Mean squared error predicted via harsh boolean metric between analytical barrier and MSM^m_{10}+GRAD refinement method for all the accepted iterations (blue) in (a) symmetric m = 2, (b) asymmetric m = 2, (c) symmetric m = 3, and (d) asymmetric m = 3. The error is additionally shown for MSM^m_{1000} model (dashed red line). While at a small number of iterations MSM^m_{10} is less precise than the 1000 centroids MSM, with increasing number of iteration MSM^m_{10}+GRAD method converges to a smaller error.

of the MSM_{1000} calculation. For three of the potentials, initially the MSM_{10} is evidently incorrect when compared with the better sampled MSM_{1000}. However, as the GRAD refinement procedure proceeds, the error is reduced, and the GRAD method rapidly finds the correct energy barrier decomposition. The final predicted error in MSM_{10}+GRAD is less than the error in the well sampled MSM_{1000} calculation. These results are significant as they demonstrate the ability of the GRAD methods to reduce discretization error even below those predicted by well sampled centroid models.

Figure 6 illustrates an issue with overfitting in MSM. For the asymmetric three-well potential, MSM^2_{10} performs better than MSM^2_{1000} even in the absence of refinement. It appears that the proposed refined method can correct for both under sampling as well as for the discretization error of oversampling. The reason for this is that the GRAD method is directly informed by the free energy landscape at the barrier, while in the MSM approach the energy landscape enters only through the sampling performed by the centroids. In this way the quality of the method used to sample the free-energy-landscape determines both the computational efficiency of the method and the precision of the results. One could object that the error in the free-energy-landscape, i.e. the roughness of the surface and its associated noise, could affect the local slope at the border. However, this error is accounted for by the procedure that smooths the energy surface.

In addition refining the crisp partitioning of the conformational landscape, a key aspect of the GRAD method is its ability

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

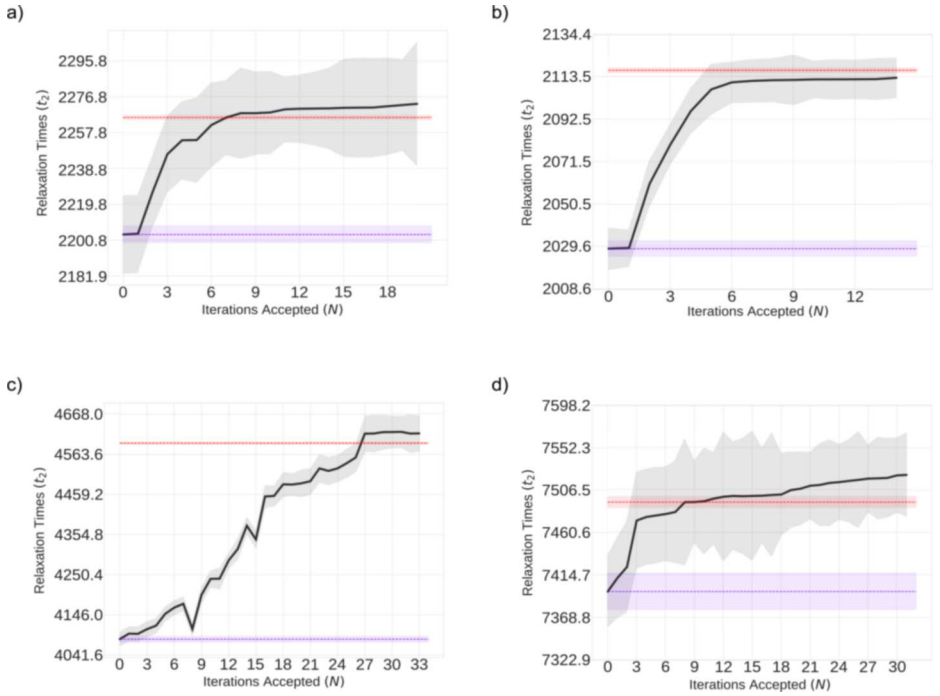


Figure 7. Calculations for the four diffusion potentials: (a) symmetric $m = 2$, (b) asymmetric $m = 2$, (c) symmetric $m = 3$, and (d) asymmetric $m = 3$. The t_2 relaxation times of the MSM+GRAD refinement approach are reported as black lines and show how t_2 evolves per accepted step in the refinement procedure. The predicted t_2 for MSM_{10} and MSM_{1000} are shown as purple and red lines, respectively. Errors are displayed as shaded regions of the same corresponding color, where statistical uncertainty is calculated by the reversible transition matrix sampling algorithm.[41]

to successfully recover time scales predicted by the kinetic model.

In Figure 7, the predicted long time scales (t_2) of all refined cases, MSM_{10}^m+GRAD, are compared to the predicted times from low-sampled MSM_{10}^m and well-sampled MSM_{1000}^m models. For each MSM_n, the centroids were clustered in macrostates using PCCA+ at a lag time, τ. The lag time, defined in eq 5, was calculated by finding the time at which the slowest relaxation time, t_2, converged and the CK condition was fulfilled. For all four model potentials we observe that the MSM converged to Markovian statistics within the time of the simulation run.

In MSM generation, the larger the number of centroids initialized during the microstate clustering, the smaller the discretization error. In the case of the simple diffusion models, we find that MSM_{100}^m gives an accurate enough decomposition of the macrostates in most cases, as comparing models MSM_{100}^m to MSM_{1000}^m. In its predictions of the time scale for the slowest kinetic process, t_2, the GRAD method with 10 centroids is comparable in accuracy to the MSM with a high number of centroids. Statistical errors were calculated by the reversible transition matrix sampling algorithm.[41] In all cases, convergence via GRAD refinement produce significant t_2 values.

The diffusion models illustrate the benefit of adopting the MSM+GRAD refinement approach because with iterative refinement the method converges to a decomposition of the molecular landscape, successfully improving the accuracy of both under and oversampled models. Additionally, as the refinement method is not centroid dependent, it more accurately decomposes the macrostates along the barriers because it can accurately represent nonconvex shapes, while Voronoi tessellation only does so in the limit of a large number of centroids.

The results presented in this section can be explained considering that the conventional MSM method assumes that the barriers can be defined as the midpoint between centroids, which allows for the use of Voronoi cells in the procedure of energy-surface decomposition. However, if the shape of the barrier is asymmetric, the Voronoi cells procedure can introduce errors in the calculation of the kinetic. More specifically, if the number of starting centroids is small, i.e. the system is undersampled, the use of Voronoi cells introduces discretization errors because of possible underfitting or overfitting of the MSM. However, if the energy surface is sufficiently sampled, barrier shape can be well-represented by MSM, at the expense of large computational times. Thus, in the case of undersampled systems, our method can be useful as it provides an accurate prediction of border decomposition independent of the shape of the barrier. When compared with highly sampled MSM calculations, our method can still be convenient because it is less computationally expensive.

DEOXYRIBOSE ADENOSINE DINUCLEOTIDE MONOPHOSPHATE

As an additional example, we studied deoxyribose adenosine dinucleotide monophosphate (ApA). For this system the free energy landscape has complex features including several minima. More challenging to capture, the roughness of the landscape varies significantly, making it difficult to calculate surface properties robustly. The combination of the complexity and roughness of the landscape, provides further tests to evaluate the capabilities of the GRAD method to define, with accuracy, the border between macrostates.

ApA is a structurally small molecule, which allows for the exhaustive sampling of its configurational free-energy space by standard atomistic molecular dynamics simulations. All-atom equilibrium molecular dynamics (MD) simulations for ApA were performed using GROMACS[42] with the Amber99+-parmbsc0[43] force-field, in explicit TIP3P water. Sodium ions

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

were added to concentration such that charges along the phosphate backbone were neutralized.[44] Structures were prepared[45−47] by energy minimization using a steepest descent algorithm for 5000 steps, heated to T = 300 K under equilibration as an NVT ensemble for 100 ps and, then, followed by a secondary 100 ps equilibration in the NPT ensemble using the Parrinello−Rahman barostat.[48] MD production runs were performed in the NPT ensemble with velocity rescaling thermostat and Parrinello−Rahman barostat, evolving atomic coordinates every 2 fs with the Verlet integrator under LINCS constraints.[49] In total, 10 independent, 1 μs simulations were performed. Each independent run used the same methodology but had different initial configurations, starting from energy minima identified from a preliminary MSM analysis of the first simulation trajectory. In total, the simulation data gave a cumulative sampling of 10 μs.

While the ApA dinucleotide is a small molecule, the configurational topology has several degrees of freedom due to nonrestricted dihedral rotations.[50,51] To analyze dinucleotide base stacking, we adopted a two site per nucleotide description, which defines vectors within the plane of each nucleotide. The first site was positioned as the midpoint between the C_4 and C_5 atoms, and the position of the second site was given by a 1 Å displacement oriented by an in-plane 90° rotation from the bond between the C_4 and C_5 atoms (see Figure 8). The

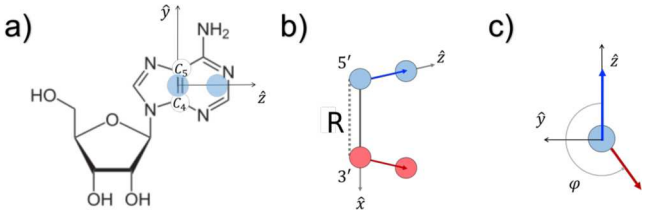


Figure 8. Depiction of the conformational model wherein (a) the fictitious sites are placed within the plane of the base. The independent order parameters are (b) the radial separation between C_4 and C_5 midpoints within each adenine monomer. (c) Aerial view, shown 5′ → 3′ into the page, of the dihedral between the in-plane vectors.

orientation in this four bead model captures the distance between nucleotides and their related torsion angle. Figure 8 illustrates the two site per nucleotide model for ApA and the relative distance and orientation of the two in-plane vectors. The distance, R, between nucleotides is calculated as the distance between the C_4 and the C_5 midpoints of each base. The stacking torsion, φ, is defined by the dihedral between the in-plane vectors. This reference frame is consistent with the convention that left and right handed helices are classified by $\varphi < 0$ and $\varphi > 0$ respectively. These coordinates have been selected because they are order parameters for the transition between stacking and unstacking of the adenosine rings in the ApA dinucleotide and are of use in the calculation of the circular dichroism signal.[52,53]

From the free energy surface of ApA shown in Figure 9 along the coordinates previously defined, the landscape appears complex, with several minima characterized by a variety of base-stacking, which can not be trivially separated into a two-state model, based on the transition between "stacked" and "unstacked" configurations. For this complex landscape an MSM analysis of the complex transitions between states is appropriate. We generated an MSM for the adenosine dinucleotide using 100 starting microstates and repeated the

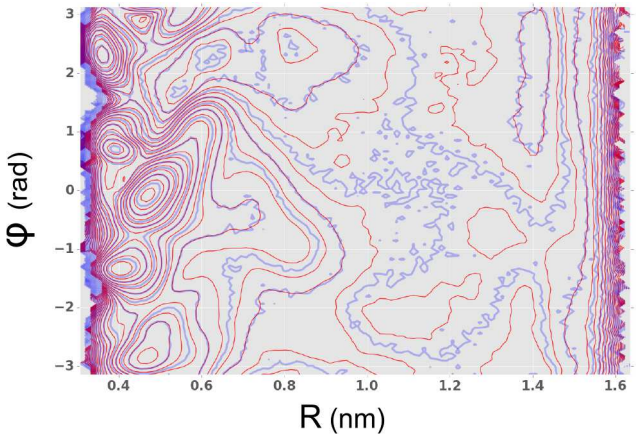


Figure 9. Free energy landscape calculated from simulations of ApA smoothed via 2D Savitzky−Golay filter. Blue contours are calculated directly from simulation data, whereas red contours have noise reduced via filtering.

calculations with higher resolution using 10 000 centroids. We selected a lag time of 100 ps, where the system converges to Markov statistics. The free energy landscape is coarse-grained with PCCA+ to 5 macrostates for both the MSM_{100} and MSM_{10000} (see Figure 10). The resulting times, t_2, for both MSM calculations are reported in Table 1.

The predicted MSM, while informative, required the use of 10 000 centroids which is computationally costly and memory exhaustive, both in the under sampled k-means++ seeding and in the diagonalization of large matrices, which involves sparse linear algebra. Thus, our refinement method could present an opportunity to reduce the number of microstates needed and return an accurate kinetic description from an under sampled MSM. A kinetic model was generated starting from MSM_{100} and analyzed at the lag-time τ = 500 ps. The MSM_{100}+GRAD refinement was performed iteratively until convergence. The predicted slow time, t_2, for MSM_{100}+GRAD is reported in Table 1, and it is found to be consistent with MSM_{10000}.

When exploring complex energy landscapes, convergence can be slowed down due to complications in gradient minimization as the barrier line can become locally trapped and not further explore neighboring maxima. As the full configuration of all macrostate decompositions is too large to sample ergodically, overcoming this problem requires an intelligent exploration to find the "true" division between metastable states. This is addressed within the GRAD method by defining a padding length, which is refined using coarse sizes to extend beyond local peaks, and then finishing the refinement at finer padding lengths. The procedure is repeated for consistency: initially the padding length is set to be constant until convergence is reached for the metastability. Then the procedure is repeated with a padding length decreased by an order of magnitude. This allows for exploration of local minima, and then after several iterations, fine-scale refinement to ensure the largest increase in metastability.

For the ApA system, a key evaluation of the refinement was to recover the landscape decomposition of a system generated from significantly more centroids, in this case MSM_{100}+GRAD method was compared with MSM_{10000}. As this system is nontrivially defined (as opposed to the diffusion potentials), the landscape decomposition predicted via MSM is far more challenging and as such more prone to error. While the number of ApA simulations provide enough statistics to reduce sparse

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

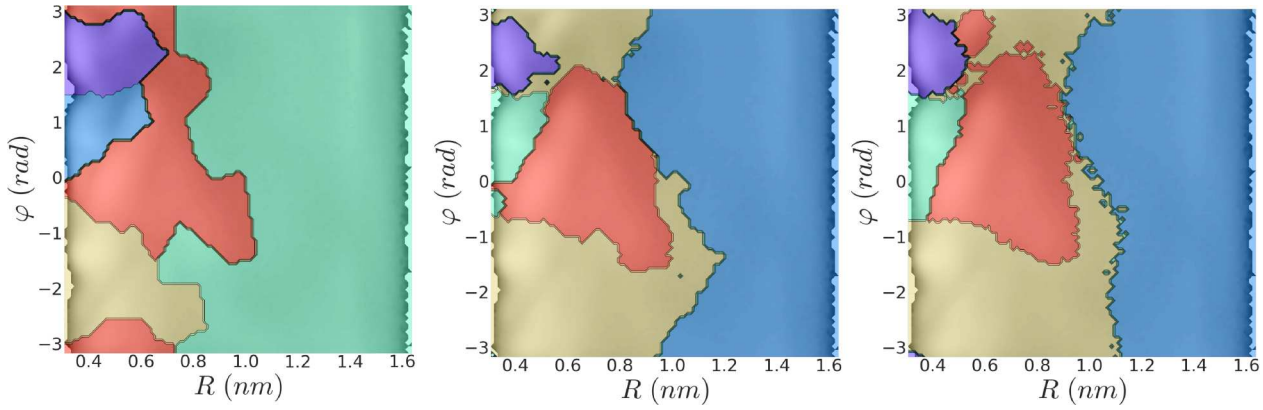


Figure 10. Decomposition of the free energy surface for ApA, as predicted by (a) MSM with 100 centroids and no refinement, (b) MSM+GRAD with 100 centroids and refinement, and (c) MSM with 10 000 centroids and no refinement.

Table 1. Time Scale for the Slowest Kinetic Process, t_2, in the Dynamics of the Deoxyribose Adenosine Dinucleotide Monophosphate

model	t_2 (ps)
MSM^{S}_{100}	2484.09
$MSM^{S}_{10,000}$	3686.19
MSM^{S}_{100}+GRAD	3851.34

sampling, several regions of the free energy surface are still not sufficiently well sampled by simulations to minimize noise due to numerical error. Therefore, the energy surface is first analyzed after applying Savitzky−Golay filtering, which smoothes the surface and reduces error during refinement. Figure 10 shows a comparison between the two refinement methods. While the method does converge to a slightly different decomposition, it is evident that the refinement is able to largely correct the decomposition predicted by MSM_{100} and produce a macrostate model that is similar to that MSM_{10000} model. The mean squared error between the two refinement methods, shown in Figure 11, demonstrates the improvement capabilities of the proposed method.

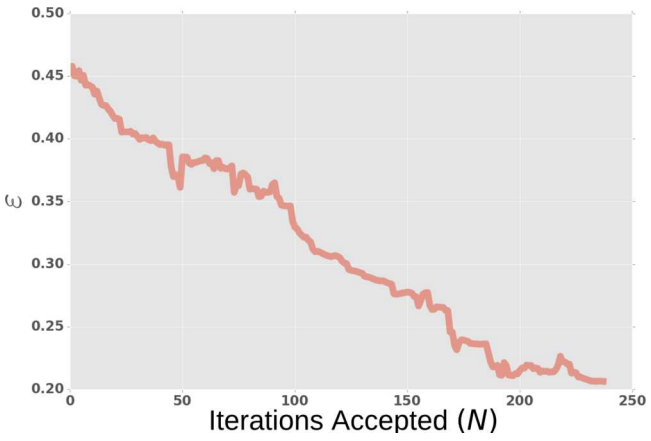


Figure 11. Mean squared error predicted via harsh boolean metric between the MSM_{10000} decomposition and the MSM_{100}+GRAD decomposition, as a function of the accepted iterations. The error of the MSM_{100}, with respect to the standard of MSM_{1000}, is given by the point at zero iterations accepted and decreases with the implementation of the GRAD refinement procedure.

DISCUSSION

MSM are widely employed to evaluate the kinetics of transition in systems that have a complex energy landscape, by analyzing simulation trajectories of the time evolution of the system. The goal of the MSM is to represent kinetic pathways in the time evolution of the system as uncorrelated transitions between states with no memory of the process history. Because complex systems are rarely Markovian, the goal of the method is to find a number of macrostates, or metastable states, that are kinetically independent and as such are connected by memory-less Markovian transitions. Thus, the method shifts the challenge from evaluating the complex kinetic pathway leading from the initial to the final states of a transition, to finding the proper states in the pathway that are sampled by the system when following a random walk along the reaction pathway.

Given that the pathways are inherently non-Markovian at short time, the MSM finds the pathway that minimizes the departure from Markov statistics by first identifying the long-time scale at which the process becomes Markovian and then, at that time, the macrostates that are sampled by the system. Those are metastable states that the system "visits" during its dynamical evolution. Following this procedure, the pathway is represented in a simpler way as a random walk among states in configurational space, where the ME formalism applies.

To calculate the kinetics of the process, one has to construct the transition matrix in the ME formalism, and to do so, one has to precisely count, during the simulation trajectory, the number of fast transitions that keep the system still inside each macrostate, and separate them from the slow transitions that occur between macrostates; the latter forming the Markovian pathway. To precisely allocate each transition inside or outside a macrostate, it is important to perform the MSM analysis with a precise definition of the location of the borders between macrostates. This partition between macrostates has to be crisp to ensure a precise count of intra and interstate transition.

In the traditional MSM the search of the macrostate borders is performed by progressively increasing the number of seeds, or centroids, used to build the transition matrix, until the slowest kinetic time converges. This process is precise but computationally costly, both in the seeding procedure and in the numerical diagonalization of the sparse transition matrix of the ME formalism.

In this paper we propose an alternative method to refine macrostates borders that we call GRAD. The new method,

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

which is easily integrated in the traditional MSM workflow, starts from a MSM performed by using a minimal number of centroids, larger than or equal to the number of minima in the free energy landscape. A region along the border is then decomposed into microstate borders, or microborders, which are subsequently assigned to the proper confining macrostates where they belong, using as information the slope of the energy landscape in the center of the microborder. The decomposition and recomposition of the borders is methodically performed on all the macrostates and iteratively performed while the method checks that the overall metastability of the sum of the macrostates increases and then converges. The metastability is calculated as the number of transitions in the simulation trajectory that keep the system in the same metastable state. The use of metastability is not novel; where our method is different is in its effort to maximize the metastability of the system by crisply refining the separation between macrostates and minimizing the uncertainty of the coarse-grained model.

Because the number of initial centroids is small and because the calculation of the metastability requires an optimization of the diagonal element of the transition matrix, but not its diagonalization, the GRAD method is relatively computationally inexpensive. While the MSM improves accuracy by scaling up the number of discrete states, the GRAD method improves accuracy while remaining at the low limit of the number of microstate centroids.

The GRAD refinement method is a novel protocol that ensures the accurate decomposition of conformational space on discrete metastable states. It works well to accurately and crisply decompose the conformational landscape from under sampled MSM. The method is robust to landscape complexity with respect to sparsity as well as noise inherent in the simulation data. This is predominantly due to the implementation of the Savitzky−Golay filtering, which is used for data smoothing to reduce error from noisy as well as sparse sampling. Implementing information from surface data can prove particularly useful in reducing the effect of limited sampling of energy barriers in simulation trajectories.

AUTHOR INFORMATION

Corresponding Author

*E-mail: mguenza@uoregon.edu.

ORCID

M. G. Guenza: 0000-0002-1151-4766

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank M. Dinpajooh and E. R. Beyerle for carefully reading the manuscript and providing helpful feedback. This work was mostly supported by NIH training grant T32 GM007759 (to P.G.R.). Partial support of this work was given by the National Science Foundation (NSF) Grant No. CHE-1362500. CPU time was provided by NSF Grant No. ACI-1053575 through Extreme Science and Engineering Discovery Environment (XSEDE) resources.

REFERENCES

(1) Noé, F.; Fischer, S. Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.

(2) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011−19016.

(3) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. Markov State Models Based on Milestoning. *J. Chem. Phys.* **2011**, *134*, 204105.

(4) Lane, T.; Bowman, G.; Beauchamp, K.; Voelz, V.; Pande, V. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. *J. Am. Chem. Soc.* **2011**, *133*, 18413.

(5) Reichl, L. E. *A Modern Course in Statistical Physics*; John Wiley & Sons, 2016.

(6) Wales, D.; Miller, M.; Walsh, T. Archetypal Energy Landscapes. *Nature* **1998**, *394*, 758−760.

(7) Lyubimov, I.; Guenza, M. A First Principle Approach to Rescale the Dynamics of Simulated Coarse-Grained Macromolecular Liquids. *Phys. Rev. E* **2011**, *84*, 031801.

(8) Copperman, J.; Guenza, M. G. Coarse-Grained Langevin Equation for Protein Dynamics: Global Anisotropy and a Mode Approach to Local Complexity. *J. Phys. Chem. B* **2015**, *119*, 9195−9211.

(9) Bowman, G.; Beauchamp, K.; Boxer, G.; Pande, V. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131*, 124101.

(10) Jain, A.; Stock, G. Identifying Metastable States of Folding Proteins. *J. Chem. Theory Comput.* **2012**, *8*, 3810−3819.

(11) Buchete, N.-V.; Hummer, G. Coarse Master Equations for Peptide Folding Dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.

(12) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, 015102.

(13) Noé, F.; Wu, H.; Prinz, J.; Plattner, N. Projected and Hidden Markov Models for Calculating Kinetics and Metastable States of Complex Molecules. *J. Chem. Phys.* **2013**, *139*, 184114.

(14) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States. *J. Chem. Phys.* **2007**, *126*, 155102.

(15) Bowman, G.; Huang, X.; Pande, V. Using Generalized Ensemble Simulations and Markov State Models to Identify Conformational States. *Methods* **2009**, *49*, 197−201.

(16) Prinz, J.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(17) Bowman, G.; Pande, V.; Noé, F. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Springer Science & Business Media, 2013; Vol. *797*.

(18) Chodera, J.; Singhal, N.; Pande, V.; Dill, K.; Swope, W. Automatic Discovery of Metastable States for the Construction of Markov mModels of Macromolecular Conformational Dynamics. *J. Chem. Phys.* **2007**, *126*, 155101.

(19) Arthur, D.; Vassilvitskii, S. K-means++: The Advantages of Careful Seeding. In Proc. of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. *SODA '07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*; 2007; pp 1027−1035.

(20) Lloyd, S. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129−137.

(21) Arthur, D.; Vassilvitskii, S. How Slow is the K-means Method? *SCG '06* **2006**, 144−153.

(22) van Kampen, N. *Stochastic Processes in Physics and Chemistry*; Elsevier, 1995.

(23) Bowman, G. Improved Coarse-Graining of Markov State Models Via Explicit Consideration of Statistical Uncertainty. *J. Chem. Phys.* **2012**, *137*, 134111.

(24) Leahy, C. T.; Murphy, R. D.; Hummer, G.; Rosta, E.; Buchete, N.-V. Coarse Master Equations for Binding Kinetics of Amyloid Peptide Dimers. *J. Phys. Chem. Lett.* **2016**, *7*, 2676−2682.

(25) Li, Y.; Dong, Z. Effect of Clustering Algorithm on Establishing Markov State Model for Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2016**, *56*, 1205−1215.

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740

(26) Deuflhard, P.; Weber, M. Robust Perron Cluster Analysis in Conformation Dynamics. *Linear Algebra and its Applications* **2005**, *398*, 161−184.
(27) Noé, F.; Nuske, F. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Model. Simul.* **2013**, *11*, 635−655.
(28) Hummer, G.; Szabo, A. Optimal Dimensionality Reduction of Multistate Kinetic and Markov-State Models. *J. Phys. Chem. B* **2015**, *119*, 9029−9037.
(29) Martini, L.; Kells, A.; Hummer, G.; Buchete, N.-V.; Rosta, E. Identification and Analysis of Transition and Metastable Markov States. *Phys. Rev. X* **2017**, 31060.
(30) McGibbon, R.; Pande, V. Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics. *J. Chem. Phys.* **2015**, *142*, 124105.
(31) Oliphant, T. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, *9*, 10.
(32) van der Walt, S.; Colbert, S.; Varoquaux, G. The NumPy array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22−30.
(33) Schütte, C.; Sarich, M. A Critical Appraisal of Markov State Models. *Eur. Phys. J.: Spec. Top.* **2015**, *224*, 2445−2462.
(34) Dunbar, D.; Humphreys, G. A Spatial Data Structure for Fast Poisson Disk Sample Generation. *ACM TOG* **2006**, *25*, 503−508.
(35) Bridson, R. Fast Poisson Disk Sampling in Arbitrary Dimensions. *ACM SIGGRAPH 2007 sketches* **2007**, 22.
(36) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*; University of Cambridge: New York, 1992.
(37) Savitzky, A.; Golay, M. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627−1639.
(38) Bresenham, J. A Linear Algorithm for Incremental Digital Display of Circular Arcs. *Commun. ACM* **1977**, *20*, 100−106.
(39) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(40) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.
(41) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. Estimation and Uncertainty of Reversible Markov Models. *J. Chem. Phys.* **2015**, *143*, 174101.
(42) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations Through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1*, 19−25.
(43) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.* **2007**, *92*, 3817−3829.
(44) Bergonzo, C.; Galindo-Murillo, R.; Cheatham, T. E. *Curr. Prot. Nucl. Acid Chem.* **2001**, DOI: 10.1002/0471142700.
(45) Cheatham, T. E., III; Kollman, P. A. Molecular Dynamics Simulation of Nucleic Acids. *Annu. Rev. Phys. Chem.* **2000**, *51*, 435−471.
(46) Cheatham, T. E. *Curr. Prot. Nucl. Acid Chem.* **2001**, DOI: 10.1002/0471142700.
(47) Bergonzo, C.; Galindo-Murillo, R.; Cheatham, T. E. *Curr. Prot. Nucl. Acid Chem.* **2001**, 1 DOI: 10.1002/0471142700.
(48) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(49) Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(50) Bloomfield, V.; Crothers, D.; Tinoco, I. *Physical Chemistry of Nucleic Acids*; Harper Collins Publishers, 1974.
(51) Cantor, C.; Schimmel, P. *Biophysical Chemistry. P. 3, The Behavior of Biological Macromolecules*; Freeman, 1980.
(52) Bush, C.; Tinoco, I. Calculation of the Optical Rotatory Dispersion of Dinucleoside Phosphates. *J. Mol. Biol.* **1967**, *23*, 601−614.
(53) Nordén, B. *Circular dichroism and linear dichroism*; Oxford University, Press, USA, 1997; Vol. *1*.

DOI: 10.1021/acs.jcim.7b00261
J. Chem. Inf. Model. 2017, 57, 2729−2740