Estimating rates from reduced representations of dynamics in an Ising model of ion channel voltage gating

Daniel Sigg,[1] Vincenzo Carnevale,[2] and Vincent A. Voelz[2, a]

[1] *dPET, Spokane, WA, USA*

[2] *Department of Chemistry, Temple University, Philadelphia, PA 19122, USA*

Extracting transport rates from atomically-detailed simulations of high-dimensional systems such as voltage gated ion channels remains a challenging prospect. One approach to obtaining this kinetic information is to calculate the potential of mean force over a 1D reaction coordinate, which, along with position-dependent diffusion coefficients, can be used to estimate rates. To test various approaches for doing this, we use as a model system a 2D Ising model that captures the essential behavior of ion channel voltage gating. From simulated kinetic Monte Carlo trajectories we estimate the phenomenological rate of gating, and compare it to mean first passage times predicted by a Smoluchowski equation and simulated Langevin dynamics in 1D the gating current coordinate. We also test an alternative approach in which time-lagged independent component analysis (tICA) is used to project the dynamics to a low-dimensional subspace enabling microstate clustering and subsequent construction of a Markov State Model (MSM) of the dynamics. While all methods tested overestimate the phenomenological rate, the MSM approach performs the best, underscoring the importance of preserving mesoscopic detail when inferring kinetics from low-dimensional projections of simulation data.

INTRODUCTION

In this workshop, we?ll perform Monte Carlo simulations of a 20 x 20 kinetic Ising lattice model. The macroscopic behavior of the lattice resembles that of a 2-state gating particle. The purpose of this exercise is demonstrate, using a simple-to-understand model, how one goes about coarse-graining a ?large? dynamical system with an astronomical number of states into a simpler system with just two states. We?ll look at the thermodynamic and kinetic aspects of the problem. Participants will employ two methods of coarse-graining: 1) Diffusion landscape approach: we calculate a potential of mean force (W) and diffusion coefficient (D), from which a coarse-grained partition function (Z) and rate constants can be derived for the 2-state macroscopic scheme. 2) Markov state model (MSM) approach: we construct a large-dimensional Markov chain, then apply successive coarse-graining that eventually yields the macroscopic behavior.

Lattice model.

The traditional Ising model comprises a grid of interacting cells that flip between resting and activated states. Here we consider a 20 x 20 Ising grid containing 400 interacting cells. Each cell flip moves a small charge, and so the system is voltage-dependent. The cellular flip rates are determined by continuous-time rate constants. Determining the system?s thermodynamic and kinetic properties as a function of temperature (T) and voltage (V) is a non-trivial problem, currently solvable only through ?kinetic? Monte Carlo sampling of the enormous space of 2^{400} states.

Thermodynamics. : The system energy E for any of the 2^{400} configurations is given by

$$E = \sum_{i,j}(e_i + e_j - 2e_ie_j)\delta\varepsilon - \sum_i e_i \delta q V \qquad (1)$$

The index i ranges over all 400 cells, whereas index j covers the nearest-neighbor cells (north, south, east, and west of cell i). Cells can take on values ei = 0 (resting) or 1 (activated). In the first term (interaction term), adjacent cells with the same orientation do not interact. If they are oppositely aligned, there is an energy penalty of $\delta\varepsilon$. The second sum is the field term, driven by the voltage V. As cells flip from resting to activated they move a microscopic gating charge q. Thus, the total gating charge $q = \sum_i e_i \delta q$ of the system ranges from 0 to $400\delta q$.

There are a large number of microscopic states containing the same E and q. We denote this degeneracy by the traditional symbol $\Omega(E, q)$. The reaction coordinate, traditionally known as the “order parameter”, is the gating charge q, which is very close to being a continuous variable, so we?ll treat it as one. The Helmholtz free energy $A(T, q)$ is logarithmically related to a Laplace-like transform of $\Omega(E, q)$:

$$\begin{aligned} A(T, q) &= -kT \ln \int \Omega(E, q) \exp(-E/kT) dE \qquad (2)\\ &= -kT \ln \int \exp(-(E - TS)/kT) dE \qquad (3)\\ &= -kT \ln Z_A(T, q) \qquad (4)\end{aligned}$$

Here $S(E, q) = k \ln \Omega(E, q)$ is the microcanonical entropy, and $Z_A(T, q)$ is the partition function specific to A. The potential of main interest to us, which we?ll designate as $\Phi(T, V)$, is a function of the two control variables T and V. It has no official name so we?ll call it the “gatin” potential. Just as $A(T, q)$ is proportional to the

[a] Electronic mail: voelz@temple.edu

logarithm of the Laplace transform of $\Omega(E,q)$, $\Phi(T,V)$ is related to the Laplace transform of $Z_A(T,q)$:

$$A(T,q) = -kT\ln\int Z_A(T,q)\exp(qV/kT)dE \quad (5)$$

$$= -kT\ln\int \exp(-(A(T,q)-qV)/kT)dE \quad (6)$$

$$= -kT\ln Z_\Phi(T,V) \quad (7)$$

The potential of mean force (PMF), which we designate by the symbol W, is given by

$$W(T,V,q) = A(T,q) - qV \quad (8)$$

When we speak of the ?energy landscape? of our gating particle, we are talking about $W(T,V,q)$. The PMF is related to the partition function $Z_\Phi(T,V)$ through:

$$Z_\Phi(T,V) = \int \exp(-W(T,V,q))/kT)dq \quad (9)$$

All thermodynamic parameters can be obtained from $Z_\Phi(T,V)$, and therefore from W(T, V, q). For example, the mean gating charge $\langle q(T,V)\rangle$, is given by

$$\langle q(T,V)\rangle = -\frac{\partial\Phi(T,V)}{\partial V} = kT\frac{\partial \ln Z_\Phi(T,V)}{\partial V} \quad (10)$$

One of the aims of this exercise is to compute W(T, V, q) through Monte Carlo simulation.

Microscopic kinetics: There are 5 activating and 5 deactivating rate constants ai for each cell i, consistent with 10 possible local configurations. The rate constants are a function of the energy needed to activate a cell from its resting state. This activation energy is equal to:

$$\Delta\varepsilon_n = 2(2-n)\delta\varepsilon - \delta qV \quad (11)$$

where n is the number of nearest-neighbor cells (0-4). Note that for n = 2, there is no penalty for activation/deactivation, provided V = 0. The formula for the value of the 5 activating or “forward” rate constants α_n is given by:

$$\alpha_n = \nu\exp(-x\Delta\varepsilon_n/kT), \quad (12)$$

where ν is a pre-exponential factor, and x (the Brønsted or LFER slope) is any number between 0 and 1 (usually assigned the value 0.5). There is no temperature-dependence ascribed to ν (implying the microscopic transition is a ?bottleneck? or purely entropy-driven event) so that any temperature sensitivity of the macroscopic rate can be attributed to cell-cell interactions. The corresponding 5 deactivating or ?backward? rate constants β_n are:

$$\beta_n = \nu\exp((1-x)\Delta\varepsilon_n/kT), \quad (13)$$

These formulas satisfy detailed balance, as evidenced by:

$$\frac{\alpha_n}{\beta_n} = \exp(-\Delta\varepsilon_n/kT), \quad (14)$$

Monte Carlo simulation: The method outlined here is different from the conventional ?Metropolis? or ?Glauber? Monte Carlo algorithms, which utilize a fixed time step. Instead, we draw random numbers that yield real-valued dwell times between microscopic transitions. The method is as follows:

(1) Choose starting configuration (time t = 0). The program randomly assigns individual cells the value of 0 or 1 according to a predetermined probability (we will use 0.5, placing the configuration at the ?top? of the macroscopic energy barrier). The system is allowed to briefly equilibrate from the starting configuration before commencing data recording. If starting near the barrier peak, the system rapidly falls into one of two macroscopic states: mostly resting cells (state R) or mostly activated cells (state A). (2) Assign to each of the 400 cells one of the 10 permissible rate constants ai. If there are NA active cells, then there will be (1NA) forward rate constants, and NA backward rate constants. (3) Sum all the previously assigned rate constants ai together to obtain the total rate constant a of leaving the current configuration. Then pick a uniform random number r1, and calculate the exponentially distributed dwell time using the formula

$$\tau = \frac{-\ln r_1}{a}. \quad (15)$$

(4) Pick a second random number r2 in order to determine which cell to flip. Each cell has probability ai/a of being chosen. (5) Advance the simulation time t by ?, and flip the randomly chosen cell. (6) Repeat steps 2-5 until the simulation time exceeds a certain value (400 ms). The result is a q-trajectory that randomly flips between two stable states (filtering accounts for the low noise):

Biased sampling: The above 400 ms trajectory took an entire hour to simulate on a laptop. If we want to speed up subsequent simulations, we can adjust the rate constants in step (2) by dividing the ai by the mean rate a(q) (determined by time-averaging a in the unbiased trajectory), and multiplying by a fixed rate, say ?. This effectively ?flattens? the PMF so that a fairly uniform sampling occurs rapidly. This

Despite the loss of kinetic information in the biased runs, we can still perform thermodynamic averaging in order to construct PMF and diffusion coefficient landscapes. Coarse-graining: The goal is to arrive at the correct 2-state macroscopic description of gating. The thermodynamics of the 2-state gating particle is determined by the coarse-grained partition function Z(T, V) = 1 + K(T, V), where the equilibrium constant K has the following empirical form:

$$K = \exp(-(\Delta A - \Delta q V)/kT) \tag{16}$$
$$= K_0 \exp(\Delta q V/kT) \tag{17}$$

The two parameters $\Delta A = A_A = A_R$ and $q = q_A?q_R$ are the free energy and charge differences between the two macroscopic states R and A. Thus, from a coarse-grained thermodynamic point of view, the activation landscape is reduced to just two coordinate pairs: (qR, AR) and (qA, AA), which are, as always, functions of T and V. The plot below demonstrates the location of the two state coordinates (blue crosses) superimposed on a PMF with zero applied voltage. A third coordinate pair (corresponding to the barrier) is described later.

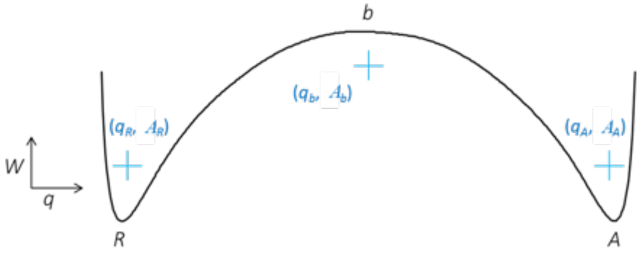

FIG. 1. Caption goes here.

A similar coarse-graining can be applied to a kinetic description of gating. Standard empirical formulas for forward and backward rate constants are as follows:

$$\alpha = \alpha_0 \exp(\Delta q_\alpha V/kT) \tag{18}$$
$$\beta = \beta_0 \exp(\Delta q_\beta V/kT) \tag{19}$$

The connections to the thermodynamic parameters are $K_0 = \alpha_0/\beta_0$ and $\Delta q = \Delta q_\alpha - \Delta q_\beta$. An alternative to these equations is the LFER formulism:

$$\alpha = \nu K^x \tag{20}$$
$$\beta = \nu K^{x-1} \tag{21}$$

where the LFER slope is $x = \Delta q_\alpha/\Delta q$. and the pre-exponential factor is $\nu = \alpha_0^{1-x}\beta_0^x$.

Participants will employ two methods for determining these kinetic parameters. The first is through an intermediate PMF, from which dynamics can be obtained by adding a diffusion coefficient and solving the diffusion equation. The second is to construct a high-dimensional Markov state model (MSM) from a long Monte Carlo trajectory, and apply coarse-graining to decrease the dimensionality.

Kinetic coarse-graining through an intermediate PMF. In this method, we construct and solve a diffusion equation based on the PMF and an additional diffusion constant D. The PMF will be exact allowing for some numerical error, but a significant assumption will be attached to D. The assumption is that D possesses no memory of past configurations (Markovian). The Markov property requires rapid vertical thermalization compared to horizontal outflow for every value of q. If the Markov assumption is valid, we can equate D with the diffusion coefficient for a Brownian motion random walk, namely:

$$D(q) = \frac{\langle \alpha(q) \rangle + \langle \beta(q) \rangle}{2} \delta q^2 \tag{22}$$

Here, the quantities $\langle \alpha(q) \rangle$ and $\langle \beta(q) \rangle$ are the time-averaged forward and backward rates obtained from our biased Monte Carlo simulations. Keep in mind these average rates are T- and V-dependent, and so all quantities derived from them are also T- and V-dependent, though we don?t always write this explicitly. In fact all thermodynamic as well as kinetic quantities of interest can be derived from $\langle \alpha(q) \rangle$ and $\langle \beta(q) \rangle$, since by invoking detailed balance, we can construct the PMF:

$$\frac{\langle \alpha(q) \rangle}{\langle \beta(q+\delta q) \rangle} = \exp\left(\frac{-W(q+\delta q) - W(q)}{kT}\right) \tag{23}$$

The appropriate diffusion equation for Brownian motion with position-dependent W and D is the Smoluchowski equation, which can be written as a sum of drift and diffusion terms:

$$\frac{\partial}{\partial t} p(q,t) = -\frac{\partial}{\partial q}\left[\frac{-W'(q)}{R(q)} p - D(q)\frac{\partial}{\partial q} p\right]. \tag{24}$$

Note that D(q) and the resistance R(q) are related through the Einstein equation D(q)R(q) = kT. There are many methods for extracting macroscopic rate constants from the Smoluchowski equation. We?ll examine two of them: (1) mean first passage time (MFPT) and (2) Langevin simulation. These are not necessarily the best methods from the standpoint of numerical accuracy, but are nevertheless instructive. Other methods include Monte Carlo simulation and eigenvalue decomposition.

MFPT. This approach offers a rational approach for kinetic coarse-graining. The mean first passage time in the forward direction (from qR to qA) is:

$$\text{MFPT} = \int_{q_R}^{q_A} dq D(q)^{-1} \exp\left(\frac{W(q)}{kT}\right) \int_{q_{(-)}}^{q} dq' \exp\left(\frac{-W(q')}{kT}\right), \tag{25}$$

While the MFPT can easily be calculated in this way, in order to properly equate the MFPT with the inverse rate constant α^{-1}, there must be a transition barrier sufficiently high (¿ 4-5 kT) that the peak is rarely visited and so the leak of probability over the barrier is very slow. We can pull out a constant value Do from the integral and, and after absorbing the diffusion coefficient term into the positive exponent, we have:

$$\alpha^{-1} = D_0^{-1} \int_{q_R}^{q_A} dq \exp\left(\frac{A_D(q) - qV}{kT}\right) \int_{q_{(-)}}^{q} dq' \exp\left(\frac{-(A(q') - q'V)}{kT}\right), \tag{26}$$

where $A_D(q) = A(q) - kT\ln(D(q)/D_0)$ is a modified Helmholtz energy containing the "spurious drift" term $-kT\ln(D(q)/D_0)$.

The above equation is a double integral. We can express it as the product of two conventional integrals provided there is minimal overlap between the two integrands, both of which have the form of a probability distribution. We can then pull the inner integral out of the outer integral. This is a satisfactory approximation when a large barrier separates two well-defined states. Pulling out the inner integral, we have:

$$\alpha^{-1} \approx D_0^{-1} \left[\int_{\{R\}} \exp\left(\frac{A(q) - qV}{kT}\right)\right] \left[\int_{\{b\}} dq' \exp\left(\frac{A_D(q) - qV}{kT}\right)\right], \tag{27}$$

We can write this more simply as:

$$\alpha \approx \frac{D_0}{Z_R \hat{Z}_b}, \tag{28}$$

where Z_R is the local partition function for the resting state, equal to $\exp(-\Phi_R/kT)$, and $\hat{Z}_b$ is a "special" partition function applied to the barrier region. Just as we did for the resting and activated states, we can derive discrete values of the barrier charge (q_b) and barrier free energy (Φ_b) values from $\hat{Z}_b$, thus completing the coarse-graining of barrier landscape. We obtain the traditional forward rate parameters from $\Delta q_\alpha = q_b - q_R$ and $\alpha_0 = D_0 \exp(-(\Phi_b - \Phi_R)/kT)$. A similar process is used to derive reverse rates, leading to $\Delta q_\beta = q_b - q_A$ and $\beta_0 = D_0 \exp(-(\Phi_b - \Phi_A)/kT)$.

Langevin equation

We can recast the Smoluchowski equation as a stochastic differential (Langevin) equation:

$$\frac{dq}{dt} = \frac{-W(q,V)}{R(q)} + \xi(q,t) \tag{29}$$

where $\xi(q,t)$ is a rapidly fluctuating force with zero mean and variance $2D(q)$. Simulating trajectories with the Langevin equation is fraught with difficulties when, as in this case, the diffusion coefficient depends on q, since there multiple interpretations possible, depending on whether one evaluates the noise term before or after the integration step. The appropriate interpretation for physical systems is that of Stratonovich, which attaches equal weight to contributions before and after the step. The simplest algorithm that converges to the Stratonovich interpretation is the Heun method:

$$q_{n+1} = q_n + \frac{1}{2}(g_1(q_n) + g_1(\hat{q}))\Delta t + \frac{1}{2}(g_2(q_n) + g_2(\hat{q}))\Delta\Gamma, \tag{30}$$

where the Weiner increment $\Delta\Gamma = N(0,1)\Delta t^{1/2}$ is derived from the standard normal distribution, and we use the following functions: $g_1(q) = -W'(q,V)/R(q)$ and $g_2(q) = \sqrt{2D(q)}$.

The intermediate (Euler) increment is given by:

$$\hat{q} = q_n + g_1(q_n)\Delta t + g_2(q_n)\Delta\Gamma. \tag{31}$$

If we start each simulation at the resting state, the mean of many simulations relaxes to a final position with time constant $\tau = (\alpha + \beta)^{-1}$. With zero applied force, we have $\alpha = \beta$, and the expression simplifies to $\tau = \alpha/2$.

KINETIC COARSE-GRAINING THROUGH MARKOV STATE MODEL APPROACHES

Markov State Models (MSMs) are the name given to a particular class of models–now widely used in the biomolecular simulation community–to model rates and pathways of molecular conformational dynamics using kinetic network models.[1,2] Starting from molecular simulation trajectory data, the goal is to (1) identify a finite set of N metastable states, (2) estimate transition rates between these states, and (3) infer information about rates and pathways. In many cases, this kinetic network can be coarse-grained to a small number of macrostates to obtain human-understandable mechanistic information.

MSMs have some key advantages that make them very attractive to the simulation community. For one, the kinetic network can be used to solve the chemical master equation, which (ostensibly) is a complete description of how the state populations evolve over time. Thus, while the transition rates may be estimated from ensembles of short trajectories, information about dynamics over much longer timescales can be obtained.

Another key advantage of MSM approaches is that they do not (strongly) rely on reaction coordinates; instead, the prevailing approach has been to avoid projecting the trajectory data to a pre-defined reaction coordinate, and instead use fine-grained conformational clustering over molecular (atomic) coordinates to define a network of states whose connectivity represents the important degrees of freedom. The representation can always later be projected to reaction coordinates for visualization and understanding. That said, this is a very data-driven approach which inevitably breaks down at

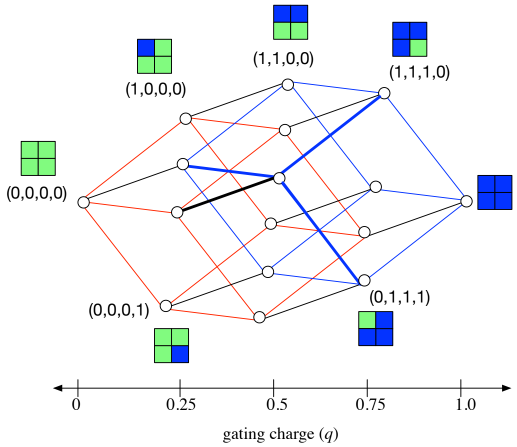

FIG. 2. A map of all possible microstates and pathways in a four-state Ising model.

some point because that we have a limited amount of very high-dimensional data. Finite-sampling error limits the resolution and accuracy of the MSM, and some projection of the data to a lower-dimensional space tends to be very useful for conformational clustering.

The Ising model examined here, an analogy for ion channel gating, represents a very interesting and challenging application case for MSM methods. While MSMs are typically useful in discovering reaction coordinates without preconceptions, here we do not need our usual caution. We already know the reaction coordinate that best corresponds to the slowest motions: q, the gating charge. As we will see, data-driven MSM approaches easily recover this fact.

But having a single reaction coordinate is perhaps not all it's cracked up to be. One approach to predicting kinetics (of channel gating, transport, etc.) is to use enhanced thermodynamic sampling methods to compute a potential of mean force (PMF) along a reaction coordinate, and then use this profile to model rates by considering diffusion along the surface. As Daniel Sigg has shown (see his workshop notes), when this procedure is done on the Ising model, the calculated rate of activation is too fast compared to the "real" answer extracted from simulated time traces, $q(t)$. Apparently, the assumption of rapid thermalization at each value of q is not valid.

We can see this more clearly in a four-state system where we can see every microstate and how they are connected (Figure **??**). The thermodynamic profile for this system has a maximum density of states at $q = 0.5$, with neighboring densities in a 4:6:4 ratio. But if we look at the connectivity of the microstates, individual pathways crossing the transition state (see bolded paths in the figure) have state densities of 2:1:2, resulting in a tighter kinetic bottleneck.

Could an MSM do a better job at estimating crossing rates compared to 1D diffusion models? We attempt to answer this question by applying current MSM methods for biomolecular simulation to the Ising problem. We will employ the following protocol:

1. Simulate a large number of microstate trajectories according to the kinetic Monte Carlo scheme outlined above.
2. Convert the kMC trajectories to discrete-time trajectories suitable for MSM analysis
3. Use the time-lagged independent component analysis (tICA) method to find a low-dimensional projection of the trajectory data that preserves the slow kinetics
4. Perform conformational clustering in the reduced tICA space to define metastable states
5. Estimate transition rates Tij between metastable states i and j from numbers of observed transitions
6. Estimate the macroscopic activation rate from the slowest implied timescales

It should be noted that the success of this approach is not guaranteed. There are several technical issues that we need to overcome, due to (1) the difficulties of sampling such an enormous space of 2^{400} microstates, and (2) the loss of information due to conformational clustering/coarse-graining.

The time-lagged independent component analysis (tICA) method

One of the problems with conformational clustering of simulation data in Cartesian space is the large numbers of degrees of freedom. Slow, correlated motions may occur only over a few degrees of freedom, with the remaining degrees of freedom contributing noise which makes it difficult to infer metastable states directly.

Consider some N-dimensional trajectory data $\mathbf{X}(t)$. The tICA approach[3?] is similar to principal component analysis (PCA), which works by diagonalizing the covariance matrix $\mathbf{C} = \langle \mathbf{X}(t)\mathbf{X}^T(t)\rangle$. In the case of PCA, the principal components (PCs) are the eigenvectors of $\mathbf{C}$, and the ones with the largest corresponding eigenvalues represent the degrees of freedom over which the largest amount of covariation in the data occurs. In contrast, tICA considers the time-lagged correlation matrix $\mathbf{C}_{TL}(\tau) = \langle \mathbf{X}(t)\mathbf{X}^T(t+\tau)\rangle$, where τ is the tICA lag time. $\mathbf{C}_{TL}(\tau)$ is a matrix containing the time-correlation of all pairs of coordinates in the system. The time-lagged independent components (tICs) are found as the eigenvectors in the solution of the generalized eigenvalue problem

$$\mathbf{C}_{TL}(\tau)\mathbf{V} = \mathbf{C}\mathbf{V}\mathbf{\Lambda} \qquad (32)$$

where $\mathbf{V}$ is the matrix of of eigenvectors. The tICA eigenvectors represent the degrees of freedom over which the largest amount of time correlation in the data occurs. $\mathbf{\Lambda}$ is a diagonal matrix of tICA eigenvalues.

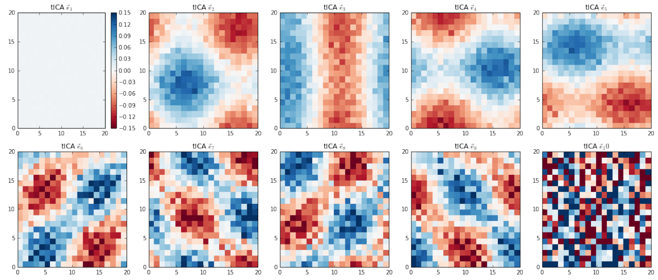

FIG. 3. The ten largest tICs calculated for the 32-trajectory data.

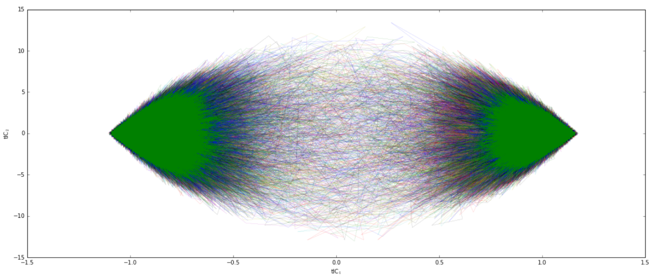

FIG. 5. Projection of trajectory data to the first two tICs.

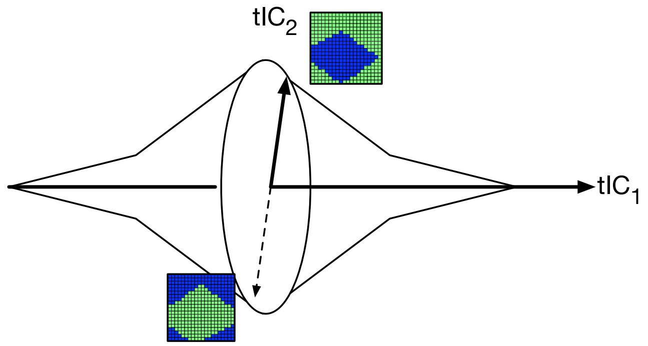

FIG. 4. The higher-indexed tICs are associated with low-energy pathways. The cartoon represents the density of states along the gating charge coordinate, q (which is also tIC1); the oval is supposed to represent a cross-section. Excursions from tIC1 must occur along higher-indexed tICs.

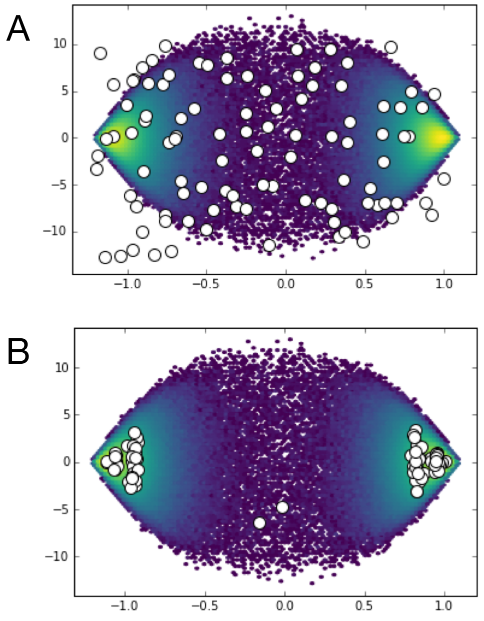

FIG. 6. Projection of trajectory data to the first two tICs.

a. Applying tICA to the Ising model. tICA was applied to a data set of 32 kMC trajectories sampled at 298K, each of length ∼280 ms. The kMC trajectories were subsampled using a discrete time step of 1 μs (yielding about 83,000 samples per trajectory), for use with MSM analysis. The tICA lag time was also taken to be one step (1 μs).

Figure 2 shows the ten largest tICs computed from the data. The 2D color maps represent each 400-element eigenvector, with red denoting negative values and blue denoting positive values. The largest tICA component, tIC1, is a nearly uniform vector of slightly positive numbers. This makes sense, given what we already know to be the slowest degree of freedom: the vector connecting the state $(0, 0, 0, ..., 0)$ to $(1, 1, 1, ..., 1)$. This is just a unit vector.

The next largest tICs reflect the next-slowest motions in the system, which are associated with a diversity of low-energv transition pathways. Dynamics cannot proceed along tIC1 alone, because individual cells must be activated (i.e. there is no $(0.5, 0.5, ..., 0.5)$ state!). Excursions from tIC1, which arise from the nature of the realized pathways (granular and minimally-frustrated) are associated with the higher-order tICs (Figure 3).

A plot of the trajectory data onto tIC_1 and tIC_2 (Figure 5) is strikingly similar the four-state Ising model in Figure 2. The horizontal axis corresponds to the gating charge reaction coordinate, q, while the vertical axis captures pathway diversity.

b. Conformational clustering. We use the k-means algorithm to perform conformational clustering. In the k-means algorithm, we first predetermine the number of clusters we want and initialize a tentative set of cluster centers (also called generators). The distances of data points to each cluster center are computed, and each point is assigned to the closest generator. After this, the algorithm iterates the two steps of recomputing the cluster centers (as the mean of all the assigned data points), and reassigns the data, iterating until all the assignments converge.

As an example, we take our 32-trajectory data set and perform k-means using 100 clusters, over 10 tICs. We start with a completely random guess for the cluster centers. After k-means, the cluster centers have moved significantly!

Apparently, there is a lot of variation in tICA space near the two main basins, and most of the cluster centers have migrated there to best cover the space. This makes sense, because the system spends so much time there and can sample many many states. We can visualize this further by plotting higher-order pairs of tICs:

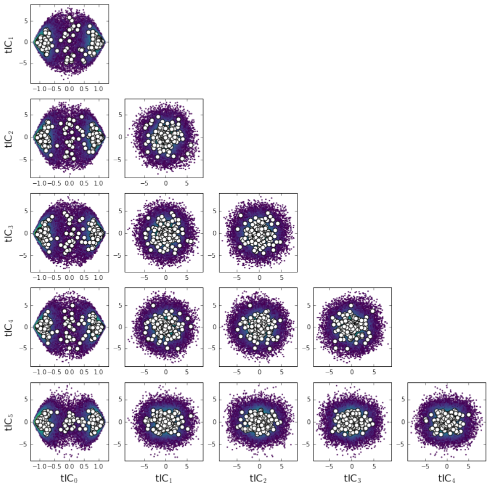

FIG. 7. Projection of trajectory data to the first two tICs.

c. MSM construction and rate estimation. Next, we build a Markov State Model from these 100 states by estimating the probabilities $T_{ij}(\tau)$ of transitioning from state i to state j in some time interval τ, which we call the MSM lag time. In practice, we go through all the trajectory data and literally count how many times we observe the system in state i at some time, and state j some time τ later. We store these counts C_{ij} in a matrix. While is it tempting to simply calculate the transition probabilities as $T_{ij} = C_{ij}/\sum_j C_{ij}$, this results in an estimate that does not obey detailed balance. Instead, a maximum-likelihood procedure can be used to infer T_{ij} that do obey detailed balance.

Once we have an estimate of the transition matrix $\mathbf{T}$ with elements T_{ji}, we have everything we need to know! The equilibrium state probabilities π are given by the stationary eigenvector $\mathbf{T}\pi = \pi$, and the complete dynamics of state populations $\mathbf{p}(t)$ is given by the rate matrix $\mathbf{K}$, through $d\mathbf{p}/dt = \mathbf{K}\mathbf{p}$, where $\mathbf{K}$ is related to $\mathbf{T}$ by matrix exponentiation, $\mathbf{T} = \exp(\mathbf{K}\tau)$. The quantities we are most interested in are the relaxation rates, which are given by the eigenvalues of $\mathbf{K}$, λ_i. The values of λ_i are real (ensured by detailed balance) and are all negative, such that each relaxation eigenmode decays over time in proportion to $\exp(\lambda_i t)$, reaching equilibrium as $t \to \infty$. The λ_i can be ordered such that $|\lambda_1| < |\lambda_2| ... < |\lambda_n|$, making $-\lambda_1$ the slowest relaxation rate.

Now, here we should point out that for a two-state system connected by forward (k_f) and backward (k_b) rates, the observed relaxation rate k_{obs} from any out-of-equilibrium starting state is $k_{\text{obs}} = k_f + k_b$. Therefore, when we build our MSM for the Ising model (with no bias potential), its slowest relaxation rate will be twice the activation rate (because in this case $k_f = k_b$).

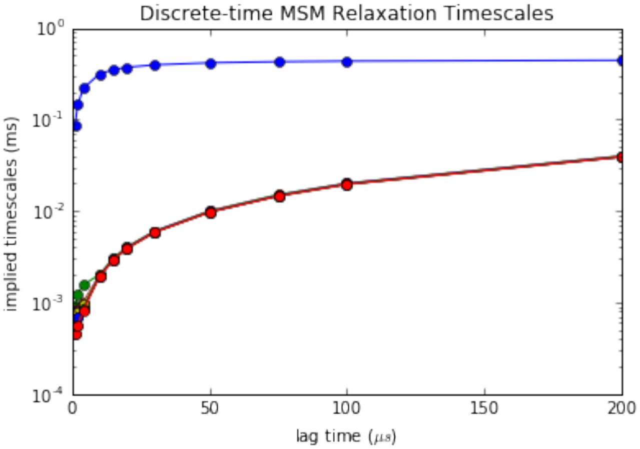

FIG. 8. Implied timescales vs. lag time.

Finally, we note that we don't actually have to calculate $\mathbf{K}$ to get this information. Because $\mathbf{T} = \exp(\mathbf{K}\tau)$, the eigenvalues of $\mathbf{T}$ are related by $\mu_i = \exp(\lambda_i \tau)$. Thus, we can diagonalize $\mathbf{T}$ to get μ_i, and get the relaxation rates as $-\lambda_i = -(\ln \mu_i)/\tau$. Equivalently, we can take the inverse of these relaxation rates to get a series of implied timescales, τ_i:

$$\tau_i = (-\lambda_i)^{-1} = \frac{\tau}{-\ln \mu_i} \tag{33}$$

d. Implied timescales. The implied timescales depend on estimates of $\mathbf{T}$, which in turn depend on a choice of MSM lag time, τ. The choice of lag time is very important, because if it is too short, the assumption of Markovian dynamics breaks down, and the implied timescales will be underestimated.

We choose an appropriate MSM lag time by examining how the implied timescales τ_i change with lag time. If the lag time is sufficiently long, the implied timescales should level off and be robustly independent of the lag time.

Figure 5 shows the slowest ten implied timescales vs. lag time for our 100-state model. One can clearly see that the implied timescales level off after about 50 ns. The blue line shows the slowest implied timescale. It is separated by over two orders of magnitude from the others (all in red, and overlapping).

To be prudent, we choose a lag time of 100 μs, for which the longest implied timescale is τ_1= 5.620 ms, which means that the MSM estimate of the activation rate is $(1/(5.620 \text{ ms}))/2 = 0.0889$ kHz. This compare very well to the "actual" answer calculated by Daniel Sigg from distributions of dwell times culled from hundreds of ∼400 ms time traces.

DISCUSSION

blah

CONCLUSION

This result suggests that MSM approaches may be increasingly useful for analyzing simulations of ion channels and other systems traditionally reliant on thermodynamic profile analysis.

ACKNOWLEDGMENTS

V.A.V was supported in part by the National Science Foundation through major research instrumentation grant number CNS-09-58854 and MCB-1412508.

REFERENCES

[1] J. D. Chodera and F. Noé, Current opinion in structural biology **25**, 135 (2014).

[2] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, The Journal of Chemical Physics **134**, 174105 (2011).

[3] C. R. Schwantes and V. S. Pande, Journal of Chemical Theory and Computation **9**, 2000 (2013).

SUPPORTING INFORMATION

Supporting Text

Preparation of initial p53-MDM2 configurations for simulation. Four non-native structures of the p53 fragment were chosen from 500 ns of replica exchange molecular dynamics (REMD) simulations of the p53 fragment (residues 17-26) taken from PDB structure 1YCR. The Amber ff96 force field with the OBC GBSA implicit solvent model were used to simulate 16 exponentially-spaced temperature replicas between 300 and 450 K. Conformational clustering was performed on the lowest-temperarture replica using a backbone-RMSD distance metric, from which representative structures

Supporting Tables

TABLE S1. Estimated rates for flux analysis.

Parameter	Value (Tmatrix)	Value (TPT)
k_{on}^{w}	1.8×10^9 M^{-1} s^{-1}	1.2×10^9 M^{-1} s^{-1}
k_{off}^{w}	2.9×10^6 s^{-1}	1.8×10^6 s^{-1}
k_{wt}	1.1×10^4 s^{-1}	1.7×10^4 s^{-1}
k_{tw}	1.0×10^7 s^{-1}	9.0×10^5 s^{-1}
k_{wt}^{MDM2}	9.6×10^5 s^{-1}	1.1×10^6 s^{-1}
k_{tw}^{MDM2}	3.9×10^6 s^{-1}	4.4×10^6 s^{-1}
k_{on}^{t}	4.0×10^9 M^{-1} s^{-1}	8.7×10^7 M^{-1} s^{-1}
k_{off}^{t}	2.9×10^4 s^{-1}	1.8×10^4 s^{-1}

TABLE S2. Corrected off-rates for the four-state binding mechanism model

Parameter	Corrected value
k_{off}^{w}	6.0 s^{-1}
k_{off}^{t}	0.06 s^{-1}

TABLE S3. Inferred rates for high p53 helicity

Parameter	Inferred rates
k_{wt} $(h\% = 28\%)$	3.2×10^5 s^{-1}
k_{tw} $(h\% = 28\%)$	4.7×10^4 s^{-1}
k_{wt} $(h\% = 64\%)$	6.9×10^5 s^{-1}
k_{tw} $(h\% = 64\%)$	2.2×10^4 s^{-1}