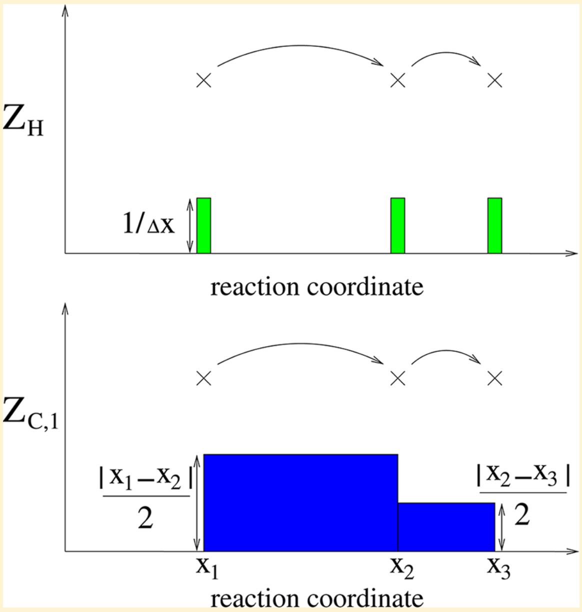
JCTC Journal of Chemical Theory and Computation
Article
pubs.acs.org/JCTC

On Reaction Coordinate Optimality

Sergei V. Krivov*

School of Molecular and Cellular Biology, Leeds University, Leeds, United Kingdom

ABSTRACT: The following question is addressed: how to establish that a constructed reaction coordinate is optimal, i.e., that it provides an accurate description of dynamics. It is shown that the reaction coordinate is optimal if its cut free energy profile, determined using length-weighted transitions, is constant, i.e., it is position and sampling interval independent. The observation leads to a number of interesting results. In particular, the equilibrium flux between two boundary states can be computed exactly as diffusion on a free energy profile associated with the coordinate. The mean square displacement, for the trajectory projected onto the coordinate, grows linear with time. That for the same trajectory projected onto a suboptimal coordinate grows slower than linear with time. The results are illustrated on a number of model systems, Sierpinski gasket, FIP35 protein, and beta3s peptide.



1. INTRODUCTION

The description of complex multidimensional dynamics can be simplified by projecting it onto a reaction coordinate. The analysis of the protein folding dynamics is the prominent example where such dimensionality reduction is often employed. The dynamics of protein folding is then described as diffusion on a low-dimensional free energy surface as a function of the reaction coordinates with a position dependent diffusion coefficient. The surface and the diffusion coefficient can be determined from the reaction coordinate time series. During such a dimensionality reduction, some information is inevitably lost. It is desirable to choose the coordinates in such a way that the information of interest is preserved. Determination of a single or few coordinates, which accurately describe the dynamics of the entire system, is challenging. A number of approaches to computing such coordinates have been suggested. Simple choices are the root-mean-square distance from the native structure, the number of native contacts, the radius of gyration, principal components, and their generalizations as dihedral PCA[1] and bond PCA.[2,3] More sophisticated approaches include, for example, diffusion maps[4] and their variation[5] and the optimization of the so-called cut profiles.[6,7]

An important dynamical quantity in the protein folding dynamics is the folding (splitting or committor) probability—the probability to fold before unfolding, starting from a given structure.[8] The folding probability is an optimal reaction coordinate, the projection on which preserves a number of important properties of the dynamics; e.g., the mean first passage time can be computed exactly in milestoning if the isocommittor surfaces are taken as milestones.[9] While the coordinate is optimal, its determination for realistic multidimensional systems is difficult. A number of approaches have been suggested to determine the p_{fold} reaction coordinate from a dynamical trajectory.[10−12]

The existence of multiple approaches, naturally, poses the question of how different reaction coordinates can be compared. Which reaction coordinate provides a better description of the dynamical process when different coordinates lead to different descriptions? A recent example is the analysis of the long equilibrium atomistic simulation of protein folding obtained by Shaw et al.[13] Two approaches applied to the analysis of the same trajectory lead to completely different results.[7,13] A related question is whether one can establish that a putative optimal reaction coordinate is indeed the optimal one. A possible approach could be to compare properties of the dynamics (e.g., the mean first passage time, the p_{fold} values, the transition path times) computed directly from the trajectories with those computed by assuming diffusive dynamics on the determined free energy landscape,[7,14] or to check whether the dynamics is Markovian[15] or diffusive.[7] Here, we suggest another criterion, which, we believe, is more fundamental and is easier to apply, namely, that the reaction coordinate is optimal if its

Received: September 25, 2012

ACS Publications © XXXX American Chemical Society

cut free energy profile $F_{C,1}$ (determined using length-weighted transitions) is constant: it does not depend on the position and on the sampling interval used to compute the profile.

A brief outline of the paper is as follows. The cut based free energy profiles are introduced. The equation for the optimal coordinate between two given boundary states with the highest cut profile is derived. The cut profiles along the optimal coordinate are shown to be constant. The constancy of the profile is exploited to obtain (in a simple way) different equations for its value. In particular, the value is trivially computed in the limit of an infinitely large sampling interval and is related exactly to the equilibrium flux between the boundary states. It is shown that the mean square displacement (MSD) computed for the trajectories projected on the optimal reaction coordinate grows linear with time, whereas that projected onto a suboptimal coordinate grows slower than linear with time, indicative of subdiffusion. The theoretical results are illustrated on a number of model systems. In particular, the MSD of a random walk on a Sierpinski gasket (a popular example of a system with anomalous transport) grows linear with time if analyzed along an optimal reaction coordinate. The Appendix contains detailed step-by-step computation of the cut profile and the MSD for a simple system.

2. THEORY

Consider the trajectory of a multidimensional equilibrium stochastic process $\mathbf{X}(i\Delta t)$ sampled with interval Δt. Reaction coordinate $x = R(\mathbf{X})$ defines a projection of the full configuration space $\mathbf{X}$ onto a coordinate x. By applying the projection to the trajectory, one obtains the reaction coordinate time series $x(i\Delta t) = R(\mathbf{X}(i\Delta t))$. The cut free energy profile $F_{C,1}(x) = -kT \ln Z_{C,1}(x)$ is a function of the reaction coordinate, or the reaction coordinate time series. Its partition function $Z_{C,1}(x)$ equals (at point x) half the sum of distances of those trajectory steps that go through the point x.[12] More precisely,

$$Z_{C,1}(x) = 1/2 \sum_i{}' |x(i\Delta t) - x((i+1)\Delta t)| \tag{1}$$

where the prime sign over the sum indicates here (and analogously below) that the sum is taken over all such i that x is between $x(i\Delta t)$ and $x((i+1)\Delta t)$. Figure 1 illustrates how the partition functions of the conventional (histogram-based) Z_H and cut-based $Z_{C,1}$ free energy profiles are computed from a reaction coordinate time-series. Here, we show that for the optimal reaction coordinate, $Z_{C,1}$ is constant. It is position (x) independent and does not change when the sampling interval (Δt) is changed.

In the theoretical analysis below, it is assumed, for simplicity, that the process is described by a Markov state model (MSM or network). In applications, however, an explicit construction of such a model is not required. The approach uses only the properties of the cut profile $Z_{C,1}$, which can be constructed directly from the reaction coordinate time series without constructing the model.

Consider a Markov process $p_i(t + \Delta t) = \sum_j P_{ij} p_j(t)$, where $p_i(t)$ is the probability to be at state i at time t, and P_{ij} is the transition probability from state j to state i. The equilibrium probability is defined as $p_i^{\mathrm{eq}} = \sum_j P_{ij} p_j^{\mathrm{eq}}$, and $n_i = p_i^{\mathrm{eq}} T/\Delta t$ is the equilibrium number of times node i is visited by a trajectory of length T. The equilibrium number of transitions between nodes i and j is $n_{ij} = P_{ij} n_j = P_{ji} n_i = n_{ji}$ (the detailed balance), with $n_i = \sum_j n_{ij}$ and $P_{ij} = n_{ij}/n_j$.

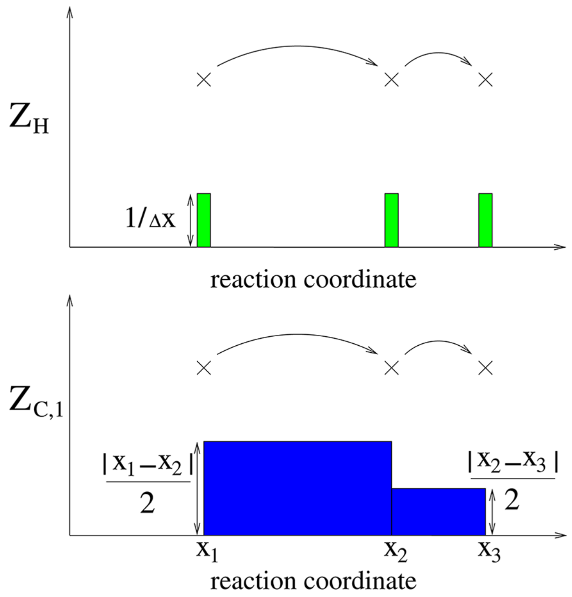


Figure 1. The partition functions of the conventional (histogram) Z_H and cut-based $Z_{C,1}$ cut profiles computed from the first three points of a trajectory x_1, x_2, and x_3 (shown as crosses). For $Z_H(x)$, when the trajectory visits a bin, the bin count is increased by $1/\Delta x$, where Δx is the bin width. For $Z_{C,1}(x)$, when the trajectory makes a transition from one point to another, the counts of all the bins in between are increased by half the length of the transition. While for the histogram there is a tradeoff between accuracy and resolution (too narrow bins have too few points), for cut profiles, one can make the bins as narrow as desired.

An optimal reaction coordinate between two given states (A and B) of the Markov state model is constructed as follows. Let the states of the model be projected onto a reaction coordinate $x_i = R(i)$. The cut profile equals $Z_{C,1}(x) = 1/2\sum'_{i,j} n_{ij} |x_i - x_j|$ for such i and j that x is between x_i and x_j. The optimal coordinate is determined by optimizing (minimizing) $I = \int Z_{C,1}(x)\, dx$ under constraints that $x_A = 0$ and $x_B = 1$. The cut profile due to transitions just between two nodes i and j is a rectangular pulse (Figure 2), which equals $|x_i - x_j|(n_{ij} + n_{ji})/2$ for x between x_i and x_j and 0 otherwise. The integral for the profile equals $\int Z_{C,1}(x)\, dx = (x_i - x_j)^2(n_{ij} + n_{ji})/2$. Since the construction of the cut profiles and integration are linear operations, the integral for the profile due to transitions between all nodes equals the sum $I = 1/2\sum_{ij} n_{ij}(x_i - x_j)^2$. It attains minimum value when $\partial I/\partial x_i = 0$, which leads to the system of linear equations

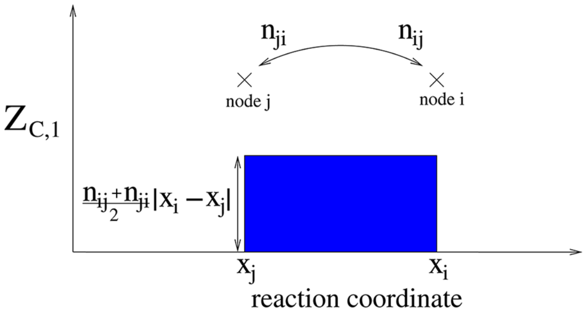


Figure 2. A rectangular pulse $Z_{C,1}$ due to transitions just between two nodes i and j of a Markov state model.

$$\sum_j n_{ij}(x_i - x_j) = 0 \tag{2}$$

with boundary conditions $x_A = 0$ and $x_B = 1$. The equation can be transformed to $x_i = \sum_j n_{ij} x_j / \sum_j n_{ij} = \sum_j n_{ij} x_j / n_i = \sum_j P_{ji} x_j$. The solution is known as the committor function, folding or splitting probability, $x_i = p_{\text{fold}}(i)$, i.e., the probability to end up in state B rather than in state A, when starting from state i.[8,12]

***$Z_{C,1}(x)$* along *p_{fold}* Is Constant.** To show this, note that $Z_{C,1}(x)$ can change only when x advances through a node, since the contribution from any pair of nodes (i,j) is constant $(n_{ij}|x_i - x_j|)$ for all x in between (Figure 3). To compute the change in $Z_{C,1}$ around node a, when x changes from $x < x_a$ to $x > x_a$, we subtract the contribution from all the nodes on the left and add the contribution from all the nodes on the right (Figure 3):

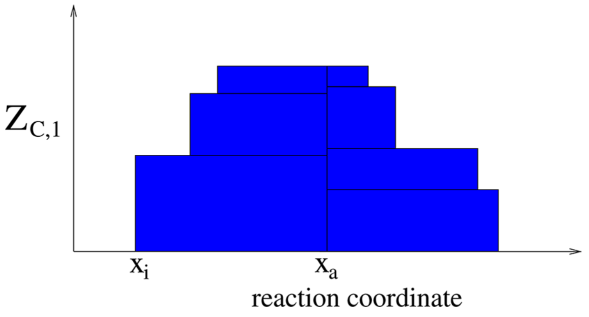


Figure 3. Change of $Z_{C,1}$ around node a. The profile shows only the contribution due to transitions from or to node a.

$$\begin{aligned}\Delta Z_{C,1} &= -\sum_{x_i < x_a} n_{ia}|x_i - x_a| + \sum_{x_i > x_a} n_{ia}|x_i - x_a|\\ &= \sum_{x_i < x_a} n_{ia}(x_i - x_a) + \sum_{x_i > x_a} n_{ia}(x_i - x_a)\\ &= \sum_i n_{ia}(x_i - x_a)\\ &= 0\end{aligned}$$

Any reaction coordinate (including suboptimal) can be rescaled $(x \to p_{\text{fold}}(x))$ so that $Z_{C,1}(x)$ (computed at a particular sampling interval) is constant. A transition network n_{ij} can be constructed by discretizing the coordinate and counting the transitions between the nodes. New positions for the nodes are found by solving eq 2. For diffusive dynamics, the transformation can be written analytically

$$\begin{aligned}p_{\text{fold}}(x) &= \int_A^x \frac{dx}{e^{-F(x)/kT}D(x)} \Bigg/ \int_A^B \frac{dx}{e^{-F(x)/kT}D(x)}\\ &= \int_A^x \frac{dx}{Z_{C,1}(x)} \Bigg/ \int_A^B \frac{dx}{Z_{C,1}(x)}\end{aligned}$$

Such transformation, obviously, does not change the optimality of the coordinate; however, it makes the optimization functional $\int Z_{C,1}(x)\, dx$ more optimal. Conversely, the optimal reaction coordinate can be locally rescaled so that $Z_{C,1}(x)$ is not constant. Thus, position independence of the $Z_{C,1}$ (computed at particular sampling interval) is neither a sufficient nor a necessary condition for reaction coordinate optimality.

***$Z_{C,1}(p_{\text{fold}})$* Is Independent of *Δt*.** Consider the cut profile $Z_{C,1}$ computed from the same trajectory (or transition network) but using longer sampling interval Δt. Unlike the conventional histogram free energy profile $F_H(x)$, which does not depend on the sampling interval (up to an overall reference constant), the cut profile does depend on the sampling interval. To compute the cut profile $Z_{C,1}$ with different sampling interval $k\Delta t$, every kth snapshot of the reaction coordinate time series is taken

$$Z_{C,1}(x) = 1/2 \sideset{}{'}\sum_i |x(ik\Delta t) - x((i+1)k\Delta t)| \tag{3}$$

for such i that x is between $x(ik\Delta t)$ and $x((i+1)k\Delta t)$. The statistics can be increased by averaging over the starting snapshot, so that one obtains

$$Z_{C,1}(x) = 1/(2k) \sideset{}{'}\sum_i |x(i\Delta t) - x((i+k)\Delta t)| \tag{4}$$

for such i that x is between $x(i\Delta t)$ and $x((i+k)\Delta t)$. For the network, it corresponds to

$$Z_{C,1}(x) = 1/(2k) \sideset{}{'}\sum_{ij} n(k\Delta t)_{ij}|x_i - x_j| \tag{5}$$

for such i and j that x is between x_i and x_j, where $n(k\Delta t)_{ij}$ represents the transition network obtained by computing the number of transitions from node j to node i with a time interval of $k\Delta t$.

If the dynamics is Markovian, i.e., $\mathbf{P}(k\Delta t) = \mathbf{P}^k(\Delta t)$, then from $\sum_j n_{ij}(x_i - x_j) = 0$ it follows that $\sum_j n(k\Delta t)_{ij}(x_i - x_j) = 0$ (for all nodes but those at the boundaries); i.e., the optimal reaction coordinate for short and long sampling intervals is the same. For example, $\sum_j n(2\Delta t)_{ij}(x_i - x_j) = \sum_j P_{ij}^2 n_j(x_i - x_j) = \sum_{jm} P_{im} P_{mj} n_j (x_i - x_m + x_m - x_j) = \sum_m P_{im} n_m (x_i - x_m) + \sum_m P_{im} \sum_j n_{mj}(x_m - x_j) = \sum_m P_{im} n_m (x_i - x_m) = 0$. Where we used $\sum_j n_{mj}(x_m - x_j) = 0$, which is valid for every node (m) except the boundary nodes. p_{fold} values computed with every step and every second step differ around boundary nodes due to the following. At a longer sampling interval, some of the events, when the system visits the boundary node and leaves it, are undetected. A simple modification is described below that rectifies the issue.

The difference between (constant) $Z_{C,1}(x)$ computed with different sampling intervals equals the difference between $\int Z_{C,1}(x)\, dx$, which is simpler to compute.

$$
\begin{aligned}
Z_{C,1}(2\Delta t) &= \int Z_{C,1}(x, 2\Delta t)\,\mathrm{d}x \\
&= 1/4 \sum_{ij} n(2\Delta t)_{ij}(x_i - x_j)^2 \\
&= 1/4 \sum_{ijm} P_{im}P_{mj}n_j(x_i - x_m + x_m - x_j)^2 \\
&= 1/4[\sum_{mij} P_{jm}P_{mi}n_i(x_i - x_m)^2 \\
&\quad + \sum_{mij} P_{jm}P_{mi}n_i(x_m - x_j)^2 \\
&\quad + 2\sum_{mij} P_{jm}P_{mi}n_i(x_i - x_m)(x_m - x_j)] \\
&= 1/4[\sum_{mi} P_{mi}n_i(x_i - x_m)^2 \\
&\quad + \sum_{mj} P_{jm}n_m(x_m - x_j)^2 + 2\sum_{m}\sum_{ij} P_{jm}P_{mi}n_i \\
&\quad (x_i - x_m)(x_m - x_j)] \\
&= Z_{C,1}(\Delta t) + 1/2 \sum_{m} \langle (x_i - x_m)(x_m - x_j)\rangle_{ij}
\end{aligned} \tag{6}
$$

where we used $\sum_j P_{jm} = 1$, $\sum_i P_{mi}n_i = n_m$. The correlation term $\langle (x_i - x_j)(x_m - x_j)\rangle_{ij} = \sum_{ij} P_{jm}P_{mi}n_i(x_i - x_m)(x_m - x_j) = \sum_j P_{jm}(x_m - x_j)\sum_i P_{mi}n_i(x_i - x_m)$ is zero for every node (since $\sum_i n_{mi}(x_i - x_m) = 0$), but the two boundary nodes, i.e., when either m = A or m = B, where it is strictly negative since x_A is smaller (x_B is greater) than x_i and x_j. Alternatively, with a longer sampling interval, some of the transitions from a node to the boundary nodes and back are no longer detected, which makes $Z_{C,1}$ smaller. Thus, while increasing the sampling interval does not change the cut profile in the middle, it decreases the profile at the boundaries.

There is a simple way to rectify the issue by modifying the way the profile with a longer sampling interval is computed. It is equivalent to considering the transition paths rather then the trajectory itself. More precisely, the trajectory is partitioned into segments (transition paths) between nodes A and B. Pathways returning to the same node ($A \rightarrow A$) are included, which means that each state is sampled with the equilibrium probability and, in particular, that the detailed balance is satisfied $n_{ij} = n_{ji}$. Note that, conventionally, the transition paths include only the reactive paths (i.e., A → B and B → A),[16] which results in nonequilibrium sampling of the configuration space. Each segment, say $x_1 = x_\mathrm{A}, x_2, ..., x_{n-1}$, $x_n = x_\mathrm{A}$ is continued into $-\infty$ and $+\infty$ by repeating the first and the last nodes as $...x_\mathrm{A}, ..., x_1 = x_\mathrm{A}, x_2, ..., x_{n-1}$, $x_n = x_\mathrm{A}, ..., x_\mathrm{A}, ...$ In fact, for every finite Δt, it is sufficient to elongate the segments just by $\pm\Delta t$. The cut profile is computed for the ensemble of the transition path segments. The correlation term for the boundary nodes is zero because either the preceding or the following node is the same boundary node and $\Delta x = 0$. Hence, $Z_{C,1}$ does not depend on the Δt. An illustrative computation of the cut profile using the transition path segments for a simple three state system is presented in the Appendix.

The value of $Z_{C,1}$ can be found by considering the limit of very large Δt, when most of the transitions are just between the boundary nodes. The cut profile tends to the limiting value, which is independent of the reaction coordinate choice. It is assumed that while reaction coordinates may provide inaccurate description of the region between A and B, they all separate the boundary nodes equally well. The limiting value (from eq 3) equals $Z_{C,1} = 1/2(N_\mathrm{AA}|x_\mathrm{A} - x_\mathrm{A}| + N_\mathrm{AB}|x_\mathrm{A} - x_\mathrm{B}| + N_\mathrm{BA}| x_\mathrm{B} - x_\mathrm{A}| + N_\mathrm{BB}| x_\mathrm{B} - x_\mathrm{B}|) = N_\mathrm{AB}$, i.e., the total number of transitions from node A to node B (or that from B to A). To summarize, the optimal reaction coordinate between two nodes (A and B) is the one with the highest cut profile $F_{C,1}$, which does not depend on the position or on the sampling interval, and $Z_{C,1}$ equals the number of transitions from one node to the other.

Kinetics. The cut profile along the optimal reaction coordinate (p_fold) equals $Z_{C,1}(p_\mathrm{fold}) = N_\mathrm{AB} = N_\mathrm{BA}$. Other expressions for N_AB can be derived as well. Integrating constant $Z_{C,1}(p_\mathrm{fold})$, one obtains

$$N_\mathrm{AB} = \int_0^1 Z_{C,1}(p_\mathrm{fold})\,\mathrm{d}p_\mathrm{fold} = 1/2 \sum n_{ij}(p_i^\mathrm{fold} - p_j^\mathrm{fold})^2$$

or

$$N_\mathrm{AB}^{-1} = \int_0^1 \mathrm{d}p_\mathrm{fold}/Z_{C,1}(p_\mathrm{fold}) = \int_{x_\mathrm{A}}^{x_\mathrm{B}} \mathrm{d}x/Z_{C,1}(x) \tag{7}$$

where x is any coordinate obtained by a monotonic transformation $x = x(p_\mathrm{fold})$, and it is assumed that the average displacement distance during Δt is sufficiently small so that

$$
\begin{aligned}
Z_{C,1}(x) &= 1/2 \sum_i{}' \left| x(i\Delta t) - x((i+1)\Delta t) \right| \\
&= 1/2 \sum_i{}' \frac{\mathrm{d}x}{\mathrm{d}p_\mathrm{fold}} \left| p_\mathrm{fold}(i\Delta t) - p_\mathrm{fold}((i+1)\Delta t) \right| \\
&= \frac{\mathrm{d}x}{\mathrm{d}p_\mathrm{fold}} Z_{C,1}(p_\mathrm{fold})
\end{aligned}
$$

Assuming that Z_H is approximately constant on the average displacement distance during Δt, one can show[12] that for diffusive dynamics $Z_{C,1}(x) = \Delta t Z_H(x)D(x)$ and $Z_{C,-1}(x) = Z_H(x)/2$, where $Z_{C,-1}(x) = 1/2\sum'_{ij} n_{ij}(x_i - x_j)^{-1}$ and $\sum'_{ij}$ denotes the sum over such i and j that x is between x_i and x_j. The assumption is certainly valid for the conventional diffusive dynamics on a continuous configuration space when sufficiently fine discretization at a sufficiently small sampling interval is employed. For the general case of an arbitrary MSM, we define

$$
\begin{aligned}
Z_H(x) &= 2Z_{C,-1}(x) \\
D(x) &= Z_{C,1}(x)/(\Delta t Z_H(x))
\end{aligned} \tag{8}
$$

The definition is self-consistent because

$$D(x) = \frac{\sum{}' n_{ij}(x_i - x_j)^{-1}(x_i - x_j)^2}{2\Delta t \sum{}' n_{ij}(x_i - x_j)^{-1}} = \langle \Delta x^2 \rangle/(2\Delta t)$$

can be considered the average of $(x_i - x_j)^2$ with weights $n_{ij}(x_i - x_j)^{-1}$, which contribute to the partition function $Z_H(x)$. The total number of transitions between the points A and B can be accurately computed as

$$N_\mathrm{AB}^{-1} = \frac{1}{\Delta t}\int_{x_\mathrm{A}}^{x_\mathrm{B}} \frac{\mathrm{d}x}{Z_H(x)D(x)} = \frac{1}{\Delta t}\int_{x_\mathrm{A}}^{x_\mathrm{B}} \frac{dx}{\mathrm{e}^{-F(x)/kT}D(x)}$$

i.e., assuming diffusive dynamics on free energy profile $F(x) = -kT \ln Z_H(x)$ with position dependent diffusion coefficient $D(x)$ defined by eq 8.

The equilibrium flux from one boundary node to the other (the number of transitions per unit time) equals $J_{\mathrm{AB}} = N_{\mathrm{AB}}/T$, where the total time equals $T = \Delta t \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} Z_H(x)\ \mathrm{d}x$. Thus, one obtains

$$\begin{aligned} J_{\mathrm{AB}}^{-1} &= \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} Z_H(x)\ \mathrm{d}x \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} \frac{\mathrm{d}x}{Z_H(x)D(x)} \\ &= \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} \mathrm{e}^{-F(x)/kT}\ \mathrm{d}x \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} \frac{\mathrm{d}x}{\mathrm{e}^{-F(x)/kT} D(x)} \end{aligned} \tag{9}$$

The Kramers equation[17] estimates the mean first passage time in the high friction limit as

$$\langle t_{\mathrm{AB}} \rangle = \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} \frac{\mathrm{d}x}{\mathrm{e}^{-F(x)/kT} D(x)} \int_{x_{\mathrm{A}}}^{x} \mathrm{e}^{-F(y)/kT} \mathrm{d}y$$

For "the full trip" mean time, one obtains

$$\langle t_{\mathrm{AB}} + t_{\mathrm{BA}} \rangle = \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} \frac{\mathrm{d}x}{\mathrm{e}^{-F(x)/kT} D(x)} \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} \mathrm{e}^{-F(y)/kT}\ \mathrm{d}y = J_{\mathrm{AB}}^{-1}$$

Note that eq 9 assumes that the x coordinate is (possibly rescaled) p_{fold}. The equation can be written (without this assumption) as

$$\begin{aligned} J_{\mathrm{AB}}^{-1} &= \max \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} \mathrm{e}^{-F(x)/kT}\ \mathrm{d}x \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} \frac{\mathrm{d}x}{\mathrm{e}^{-F(x)/kT} D(x)} \\ &= Z \max \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} \frac{\mathrm{d}x}{\mathrm{e}^{-F(x)/kT} D(x)} \end{aligned} \tag{10}$$

where the maximum is taken over all possible reaction coordinates $x = R(\mathbf{X})$ between the two states ($R(\mathrm{A}) = x_{\mathrm{A}}$, $R(\mathrm{B}) = x_{\mathrm{B}}$) and Z is the total partition function. Analogously,

$$\begin{aligned} J_{\mathrm{AB}} &= Z_{C,1}(p)/T \\ &= 1/T \int_0^1 Z_{C,1}(p)\ \mathrm{d}p \\ &= \Delta t/T \int_0^1 Z_H(p)D(p)\ \mathrm{d}p \\ &= 1/Z \int_0^1 Z_H(p)D(p)\ \mathrm{d}p \\ &= 1/Z \min \int_0^1 Z_H(x)D(x)\ \mathrm{d}x \end{aligned} \tag{11}$$

Thus, for any complex equilibrium stochastic process, which can be described by a Markov state model, it is possible to construct an optimal reaction coordinate between any two chosen boundary states by optimizing the cut free energy profile $F_{C,1}$. The associated free energy profile and the position dependent diffusion coefficient are determined using eq 8 from the reaction coordinate time series or from an MSM. The equilibrium flux between the boundary states is found exactly by assuming diffusive dynamics on the free energy profile with the diffusion coefficient, e.g., by using eq 9. Equation 8 can be used to determine the free energy profile and diffusion coefficient along any putative reaction coordinate. The equilibrium flux computed from them by assuming diffusive dynamics (e.g., eq 9) attains its minimum when the putative reaction coordinate is the optimal reaction coordinate, and the value of the minimum is exactly the equilibrium flux (eqs 10 and 11). Note that no assumptions are made regarding the complexity of the free energy landscape or the separation of time scales. In particular, for the protein folding, defined as stochastic dynamics between the unfolded and native states, such a coordinate, in principle, can always be constructed. The coordinate should, in general, depend on solvent degrees of freedom. If, however, the protein degrees of freedom contain all the essential information about the solvent degrees of freedom, the optimal reaction coordinate can be constructed using only the former. An interesting open question is which other properties, besides the equilibrium flux, can be computed exactly by using such one-dimensional representation.

Similar variational principles have been derived before. Berezhkovskii and Szabo have analyzed diffusive barrier crossing for a multidimensional system with a high barrier, harmonic transition state, and anisotropic diffusion coefficient.[18] They computed the rate by using the one-dimensional Kramers' equation for the potential of mean force along a putative reaction coordinate and found that it attains the minimum when the putative coordinate is perpendicular to the stochastic separatrix ($p_{\mathrm{fold}} = 0.5$). The minimal value is the exact rate given by multidimensional Langer equation. E and Vanden-Eijnden considered metastable diffusive dynamics with a constant diffusion coefficient.[19] They showed that the p_{fold} reaction coordinate can be obtain by optimizing either $I_1 = \int_\Omega |\nabla q|^2\ \mathrm{e}^{-\beta V}\ \mathrm{d}\mathbf{X}$ or $I_2 = \int_0^1 [\int_\Omega |\nabla q|^2 \mathrm{e}^{-\beta V} \delta(q(\mathbf{X}) - q)\ \mathrm{d}\mathbf{X}]^{-1}\ \mathrm{d}q$. Integral $\int_\Omega |\nabla q|^2\ \mathrm{e}^{-\beta V} \delta(q(\mathbf{X}) - q)\ \mathrm{d}\mathbf{X}$ is proportional to $Z_{C,1}(q)$ (if the diffusion coefficient is constant) so the two optimization functionals are equivalent to $\int_0^1 Z_{C,1}(x)\ \mathrm{d}x$ and $\int_0^1 \mathrm{d}x/Z_{C,1}(x)$, correspondingly. In fact, since $Z_{C,1}(x)$ is constant at the optimum, one can optimize any functional of the type $\int_0^1 A[Z_{C,1}(x)]\ \mathrm{d}x$, where A is an arbitrary monotonic function.[12] The variational principle established in the present work is more general: it is valid for any Markov state model. It has also been shown that the value of the equilibrium flux between the two boundary states computed by assuming diffusive dynamics on the determined one-dimensional free energy profile with the determined diffusion coefficient is exact. In particular, no metastability (two state process) or separation of time scales is required. From a practical point of view, the proposed variational principle is much simpler because it does not require a detailed knowledge of the systems equations of motion, e.g., the potential energy surface and the coordinate dependent diffusion coefficient or a MSM. For example, one can analyze Newtonians dynamics at a sufficiently large sampling interval, when the dynamics loses memory and the diffusive description (in the overdamped regime) becomes appropriate. It is sufficient to know only the reaction coordinate time series $x(t)$. The optimal coordinate can be found by optimizing the cut profile $F_{C,1}(x)$.

The equilibrium flux can be also computed as

$$\begin{aligned} J_{\mathrm{AB}} &= N_{\mathrm{AB}}/T = Z_{C,1}(p)/T \\ &= 1/2\sum_{ij}^{\prime} \frac{n_{ij}}{T}|p_i^{\mathrm{fold}} - p_j^{\mathrm{fold}}| \\ &= 1/2\sum_{ij}^{\prime} \frac{P_{ji}}{\Delta t}\frac{n_i}{T/\Delta t}|p_i^{\mathrm{fold}} - p_j^{\mathrm{fold}}| \\ &= 1/2\sum_{ij}^{\prime} \frac{P_{ji}}{\Delta t}p_i^{eq}|p_i^{\mathrm{fold}} - p_j^{\mathrm{fold}}| \\ &= \sum_{ij}^{\prime\prime} \frac{P_{ji}}{\Delta t}p_i^{eq}(p_i^{\mathrm{fold}} - p_j^{\mathrm{fold}}) \end{aligned}$$

where $\sum_{ij}'$ denotes the sum over such i and j that p is between p_i^{fold} and p_j^{fold} and $\sum_{ij}''$ is that for $p_i^{\mathrm{fold}} > p > p_j^{\mathrm{fold}}$. If Δt is sufficiently small that off-diagonal elements of the transition matrix are P_{ij}=$\Delta t K_{ij}$, where K_{ij} is the reaction rate matrix, then $J_{\mathrm{AB}} = \sum_{ij}''K_{ji}p_i^{\mathrm{eq}}(p_i^{\mathrm{fold}} - p_j^{\mathrm{fold}})$. The equation was derived previously.[20,21]

On Diffusion and Subdiffusion. The optimal reaction coordinate has the highest free energy profile (the lowest partition function) by definition. All other coordinates have lower profiles. Consider such a suboptimal coordinate. As the sampling interval increases, the profile $F_{C,1}$ increases as well ($Z_{C,1}$ decreases) until it reaches the limiting optimal value. Assume that the decrease in $Z_{C,1}$ is monotonic. The difference between the profiles equals the correlation term: $Z_{C,1}(2\Delta t) - Z_{C,1}(\Delta t) = 1/2\sum_m\langle(x_i - x_m)(x_m - x_j)\rangle_{ij} < 0$, meaning that (on average) the consecutive displacements are anticorrelated. By considering the difference between the cut profiles integrated locally (and neglecting the boundary effects) $\int_{p_0}^{p_1}Z_{C,1}(x,2\Delta t)\ \mathrm{d}x - \int_{p_0}^{p_1}Z_{C,1}(x,\Delta t)\ \mathrm{d}x \approx 1/2\sum_{p_0<x_m<p_1}\langle(x_i - x_m)(x_m - x_j)\rangle_{ij} < 0$, one discerns that the consecutive displacements are anticorrelated locally. The negative correlation means that the dynamics is subdiffusive. For the optimal reaction coordinate, the correlation between the consecutive displacements is zero, i.e., the dynamics is diffusive. Thus, the equilibrium Markovian dynamics projected on a reaction coordinate is never superdiffusive, diffusive only on optimal coordinates and subdiffisive on the rest. A precise definition of what is meant by the diffusive dynamics in this context is given below.

In practice, the sampling of the configuration space is always finite. In such cases, if the reaction coordinate functional form has many parameters, it might be possible to overfit the trajectory during optimization, so that $Z_{C,1}$ is lower than the correct one.[7,12] Thus $Z_{C,1}$ increases as the sampling interval increases, which means that the consecutive displacements are positively correlated and dynamics is superdiffusive. This superdiffusivity of the dynamics can be used as an indicator of overfitting.[7,12] The optimal reaction coordinate with the $Z_{C,1}$ profile constant for all x and Δt is free of overfitting.

One may give an alternative interpretation of the results in terms of the mean square displacement (MSD). From $\int_0^1 Z_{C,1}(x,k\Delta t)\ \mathrm{d}x = 1/2k\sum_{ij}n(k\Delta t)_{ij}(x_i - x_j)^2 = \Delta x^2(k\Delta t)Z/2k$, one obtains the closed form expression for the MSD computed over the transition path segments

$$\langle\Delta x^2(k\Delta t)\rangle = 2kZ_{C,1}(k\Delta t)/Z \tag{12}$$

where $Z = \sum_{ij}n_{ij}$ is the total partition function. Note that such computed MSD grows infinitely even for a system with finite configuration space. Since the result is somewhat counterintuitive, a simple three state system is analyzed in detail in the Appendix. The MSD computed in the conventional way, using trajectory, is bounded by some equilibrium value for a system with finite configuration space. The two ways to compute the MSD are in agreement at small times, when the boundary effects can be neglected. If $Z_{C,1}(k\Delta t)$ is constant with respect to k, then the MSD grows linear with time, i.e., the dynamics is diffusive. If $Z_{C,1}(k\Delta t)$ decreases with k increasing, the MSD grows slower than linear with time, i.e., the dynamics is subdiffusive. And conversely, if $Z_{C,1}(k\Delta t)$ increases with k increasing, the MSD grows faster than linear with time and the dynamics is superdiffusive. Note, that the MSD is computed by taking the average over all the states as starting points (with equilibrium probabilities). Thus, when the dynamics is said to be diffusive or subdiffusive it is referred to the dynamics of an equilibrium ensemble of trajectories starting from an equilibrium set of states, whereas conventionally one starts with a single source. One can integrate the profile over some local region instead of the entire coordinate $\int_{p_0}^{p_1}Z_{C,1}(x,k\Delta t)\ \mathrm{d}x/\int_{p_1}^{p_1}Z_H(x)\ \mathrm{d}x \approx \langle\Delta x^2(k\Delta t)\rangle/2k$ and discern that the MSD in the region grows linear with time, but it would still involve averaging over all the states in the region. Obviously, when all the states are equivalent, the MSD averaged over the states reduces to the conventional MSD for a single source.

To find the optimal coordinate, along which the MSD grows in the fastest way (linear with time), one optimizes the cut profile, which is equivalent to minimizing the MSD $1/2\sum_{ij}n_{ij}(x_i - x_j)^2$. It may sound counterintuitive, but the reason is as follows. The optimal coordinate has the smallest MSD for every sampling interval. At very large intervals ($\Delta t_{\mathrm{large}}$), when the transitions are, essentially, just between the boundary nodes, the MSDs for all coordinates are the same. Then if at some small sampling interval ($\Delta t_{\mathrm{small}}$) a coordinate had the MSD higher than the optimal coordinate, then somewhere in between ($\Delta t_{\mathrm{small}} < \Delta t < \Delta t_{\mathrm{large}}$) its MSD should grow slower than linear so that the MSDs of the two coordinates are in agreement at $\Delta t_{\mathrm{large}}$. It means that the dynamics should be anticorrelated or subdiffusive. In other words, to grow steadily as fast as possible, one starts with the lowest value.

The Relation between $Z_{C,1}$ and Z_C. If the dynamics along the optimal reaction coordinate is diffusive, then such a coordinate can be determined by optimizing either $Z_{C,1}$ or Z_C cut profiles.[12] $\int Z_{C,1}(x)\ \mathrm{d}x$ attains the optimum when $\sum_j n_{ij}(x_i - x_j) = 0$, i.e., x_i equals the mean of the connected nodes. $\int Z_C(x)\ \mathrm{d}x$ attains the optimum when $\sum_{j,x_j < x_i}n_{ij} = \sum_{j,x_j > x_i}n_{ij}$ or $\sum_j n_{ij}\mathrm{sign}(x_i - x_j) = 0$, i.e., x_i equals the median of the connected nodes. For diffusive dynamics with Gaussian distribution of transitions between connected nodes $n_{ij} \sim \exp[-(x_i - x_j)^2/4D\Delta t]$, the median and the mean coincide, and the two optimization functionals are equivalent. As the sampling interval increases, the distribution of transitions may start to deviate from the Gaussian. In particular, it happens when the steps are so large that they reach the boundaries. In this case, while optimization of $Z_{C,1}$ still produces the optimal p_{fold} reaction coordinate, optimization of Z_C may not. It is easier to optimize Z_C than $Z_{C,1}$ because the former is invariant to monotonic transformations of the reaction coordinate. To attain the optimum, the reaction coordinate has only to reproduce the relative order of the nodes x_i, not their exact value $x_i = p_{\mathrm{fold}}(i)$, as when the optimum of $Z_{C,1}$ is attained. The optimization functional $\int\ \mathrm{d}x/Z_{C,1}(x)$ is invariant to reaction coordinate transformation, but only when the sampling interval is sufficiently small. Both $Z_{C,1}$ and Z_C can be used to inspect whether dynamics is diffusive or subdiffusive. For the former, the coordinate should be rescaled so that $Z_{C,1}$ is constant; thus the entire coordinate is either optimal or not. The latter can be used to test the local optimality; however it is necessary that Z_H be approximately constant on the average displacement distance, which limits its applicability to transition state regions and small sampling intervals. Finally, it is much easier to work analytically with $Z_{C,1}$.

3. ILLUSTRATIVE EXAMPLES

1D Model System. The 1D model system is mainly used to demonstrate the properties of the $Z_{C,1}$ cut profiles. The system has a potential energy profile of $U(x) = \cos(2\pi x - \pi)$ for $x \in [0,1]$. The segment is discretized as $x_i = i/100$ for $i = 0,1,...,100$. The transition network is $n_{ij} = \exp[-(U(x_i) + U(x_j))/2]$ if $|x_i - x_j| \leq 0.03$. The distribution of steps is taken to be step-like rather than Gaussian to make the dynamics a bit more general. The trajectory is generated by simulating MC dynamics with transition matrix $p_{ij} = n_{ij}/\sum_i n_{ij}$ for 10^7 steps. Figure 4 shows $F_{C,1}$ for different reaction coordinates. The $F_{C,1}$ as a function of the x coordinate equals $U(x)$ up to an unimportant constant (and boundary effects). The $F_{C,1}$ as a function of the optimal reaction coordinate p_{fold} is position independent and on average higher than $F_{C,1}$ as function of x. The p_{fold} coordinate is more optimal in the sense that it gives a smaller value to the optimization functional $\int Z_{C,1}(x)\ dx$. For the 1D system, the p_{fold} coordinate is equivalent to the x coordinate and to any other coordinate obtained by monotonic transformation from x. Figure 4 shows that the $F_{C,1}$ profiles, computed using the partitioning of the trajectory into transition path segments, are constant with respect to the position and the sampling interval. As the sampling interval increases, the profiles computed without the partitioning deviate more and more from the constant value around the boundaries.

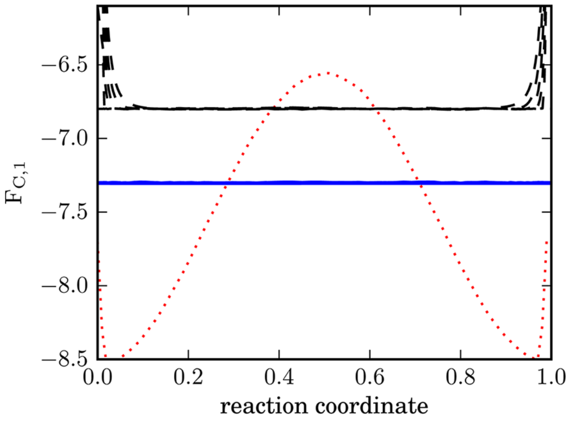


Figure 4. $F_{C,1}$ profiles along the x coordinate (red dotted line) and along the optimal reaction coordinate (p_{fold}) computed at different sampling intervals Δt with and without partitioning into transition path segments. The former are shown with a blue solid line for $\Delta t = 1$, 2, 4, 1024, 4096. The latter are shown with a black dashed line for $\Delta t = 1$, 2, 4, 8, 16, 32 and are shifted up by 0.5 for visual clarity.

System with Two Parallel Pathways. A system with two parallel pathways is considered to illustrate the difference in the behavior of the cut profiles $F_{C,1}$ as functions of optimal and suboptimal reaction coordinates. The system contains two straight parallel 1D pathways $(0 \leq x \leq 1, y = 1)$ and $(0 \leq x \leq 1, y = 2)$. The corresponding terminal nodes of the pathways are considered to be identical: $(x = 0, y = 1) = (x = 0, y = 2) = A$ and $(x = 0, y = 1) = (x = 0, y = 2) = B$. Each pathway has a barrier described by the potential $U(x,y) = 2 \exp[-9(3x - y)^2]$. The pathways are discretized as $x_i = i/100$ for $i = 0, 1, ..., 100$. The transition network is $n_{ij} = \exp[-(U(x_i,y_i) + U(x_j,y_j))/2]$ if $|x_i - x_j| \leq 0.03$ and $y_i = y_j$. The trajectory was generated by simulating MC dynamics for 4×10^7 steps.

Figure 5a shows the $Z_{C,1}$ cut profiles along the optimal reaction coordinate ($p_{\text{fold}}(x_i,y_i)$ between nodes A and B) computed at different sampling intervals. The profiles are constant with respect to the position and the sampling interval; i.e., the reaction coordinate is optimal. To mimic a suboptimal choice of reaction coordinate, the configuration space was initially projected onto the x coordinate by removing the y values from the trajectory. That x is a suboptimal reaction coordinate is evident from the fact that the barriers of the two

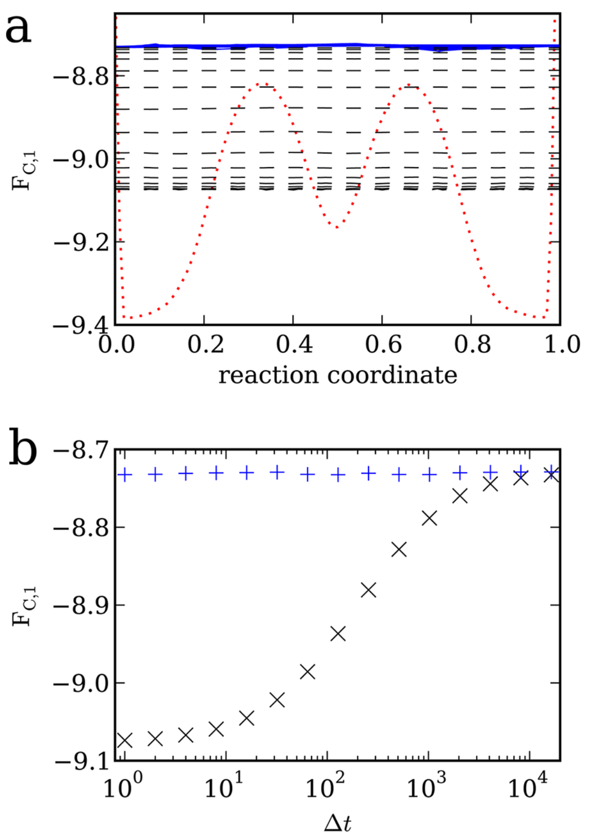


Figure 5. $F_{C,1}$ along different reaction coordinates for the model system with two one-dimensional pathways (x_i,y_i). (a) $F_{C,1}(p(x,y))$ as a function of the optimal reaction coordinate $p = p_{\text{fold}}(x,y)$ for different sampling intervals ($\Delta t = 1, 2, ..., 2^{16}$) are shown by (blue) solid lines. They are constant with respect to position and the sampling intervals, i.e., the reaction coordinate is optimal. $F_{C,1}(x)$ (at $\Delta t = 1$) as a function of the suboptimal reaction coordinate x is shown with a (red) dotted line. It is not constant with respect to the position x. $F_{C,1}$ as a function of an improved but still suboptimal reaction coordinate, $p = p_{\text{fold}}(x)$, is shown with dashed (black) lines. The profiles are position independent (the transformation to $p_{\text{fold}}(x)$ is applied to every sampling interval) but increase as the sampling interval increases, from −9.07 for $\Delta t = 1$ to the correct value of −8.72 at $\Delta t = 2^{16}$. (b) Dependence of $F_{C,1}$ on the sampling interval Δt for the optimal and suboptimal coordinates.

pathways (centered around $x = 1/3$ and $x = 2/3$) do not align well when projected on x. As a result, their height is decreased after projection from 2.0 to 0.6. Figure 5a shows that $Z_{C,1}$ profile along x is not constant and lower than that along the optimal reaction coordinate $p_{\text{fold}}(x,y)$. The (suboptimal) x reaction coordinate can be "improved" (to give smaller $\int Z_{C,1}(x)\ dx$) by computing p_{fold} as function of x_i. $Z_{C,1}$ along the $p_{\text{fold}}(x)$ is constant and higher than that along x. However, it is still lower than that along the optimal $p_{\text{fold}}(x,y)$. Note that to make the profiles along the (suboptimal) $p_{\text{fold}}(x)$ coordinate position independent for different sampling intervals Δt, the $p_{\text{fold}}(x)$ coordinate had to be recomputed for every Δt. Since the dynamics projected on the (suboptimal) $p_{\text{fold}}(x)$ coordinate is not Markovian, the $p_{\text{fold}}(x)$ coordinates for different sampling intervals are different. Hence, the profile determined with particular Δt along the $p_{\text{fold}}(x)$ constructed for different Δt is not position independent.

Figure 5b shows how $F_{C,1}$ changes with the sampling interval Δt. $F_{C,1}$ for the optimal coordinate is constant. $F_{C,1}$ for the suboptimal coordinate is constant for small Δt, increases with Δt in the intermediate region, and reaches the limiting (optimal) value for Δt larger than the mean first passage time between the basins of ~3000 steps. The limiting value is independent of the choice of reaction coordinate and equals half the total number of transitions between nodes A and B. The plot illustrates that the suboptimal coordinate is suboptimal for all time scales. It is impossible to select a time scale after which the coordinate becomes optimal and dynamics along it becomes diffusive. The relatively constant value of the $F_{C,1}$ for small Δt is due to the fact that a deterministic contribution due to the potential is proportional to Δt and can be neglected at small Δt compared to the stochastic contribution, proportional to $(\Delta t)^{1/2}$. In other words, all the coordinates may seem optimal at very small Δt, and to detect suboptimality one needs to explore all the time scales. Alternatively, one may compare the value of $Z_{C,1}$ with the long time limit of N_{AB}.

The Sierpinski Gasket. The Sierpinski gasket (triangle) is considered as the next model system. The gasket is a popular example of a disordered system used to study the anomalous transport.[22] It can be constructed according to the following iterative procedure (Figure 6). We start with a network consisting of three pairwise joined nodes (a triangle). At the next iteration, each triangle is partitioned into four triangles by placing the nodes in the middle of each edge and joining them. The central triangle (out of these four) is then discarded from further consideration.

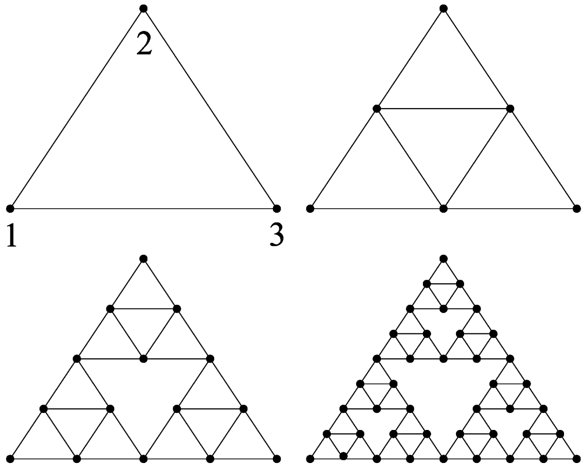


Figure 6. The iterative construction of the Sierpinski triangle.

Random walk in such systems with fractal, self-similar configuration space shows characteristic subdiffusive behavior.[22] The MSD grows slower than linear with time. In particular, for the Sierpinski gasket, it grows as $\langle R^2 \rangle \sim t^{2/d_w}$, where $d_w = \ln 5/\ln 2 = 2.322$ is the anomalous diffusion exponent.[22] The fractal structure of the configuration space was suggested to be the reason why the protein dynamics shows subdiffusive behavior.[23] The subdiffusive dynamics in proteins has been observed in many experimental and theoretical works and is a subject of intense research.[23−36]

We apply the framework to analyze the dynamics of the random walk on the gasket. The transition network equals 1 for directly connected nodes and 0 otherwise. The iterative process was continued for six generation and resulted in a network with 1095 nodes, on which MC dynamics was simulated for 4×10^7 steps.

The optimal reaction coordinate (p_{fold}) between boundary nodes 1 and 3 is determined by solving eq 2 with $x_1 = 0$ and $x_3 = 1$. Figure 7 shows that the $F_{C,1}$ profile along the coordinate is constant with respect to position and sampling interval ($F_{C,1} \sim -6.4$). It indicates that dynamics, when projected on the optimal reaction coordinate, is diffusive. As a suboptimal reaction coordinate, we take the x coordinate. The cut profile along the coordinate is nonconstant and lower than that along the optimal coordinate. Reoptimization of the coordinate ($p_{\text{fold}}(x)$) leads to a constant profile that is a bit higher $F_{C,1} \sim -7.4$ but still lower than the optimal one. The profile gets higher as the sampling interval increases, until it reaches the optimal value. The elevation of the profile is an indication that dynamics along the suboptimal reaction coordinate is subdiffusive. To confirm it, Figure 8 shows the (conventional) MSD vs time along the optimal and suboptimal coordinates. The MSD was computed for trajectories starting between x = 0.4 and x = 0.6. For the optimal coordinate, the MSD grows linearly with time, indicating diffusive dynamics. For the suboptimal coordinate, the MSD grows subdiffusively with the

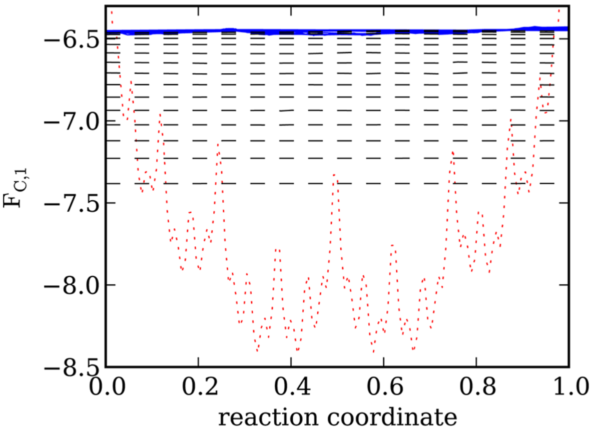


Figure 7. $F_{C,1}$ for the Sierpinski gasket along different reaction coordinates between two boundary nodes 1 and 3 for different sampling intervals (Δt = 1, 2, ..., 2^{16}). $F_{C,1}$ as a function of the optimal reaction coordinate are shown with (blue) solid lines. $F_{C,1}(x)$ (at Δt = 1) as a function of the suboptimal reaction coordinate x is shown with a (red) dotted line. $F_{C,1}$ as a function of improved but still suboptimal reaction coordinate, $p = p_{\text{fold}}(x)$ are shown with dashed (black) lines.

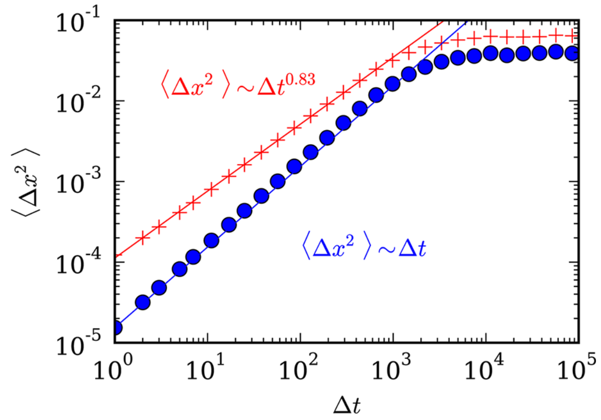


Figure 8. The MSD along the optimal (blue circles) and suboptimal (red crosses) reaction coordinates for the Sierpinski gasket. The solid lines show diffusive (blue) and subdiffusive (red) slopes and are to guide the eye.

exponent close to the theoretical value (residual discrepancy is, likely, due to finite system size).

The second optimal reaction coordinate which describes the dynamics between nodes 1 and 2 can be constructed by solving the equation $\sum_j n_{ij}(y_i - y_j) = 0$ with constraints $y_1 = 0$ and $y_2 = 1$. Since the equations are linear, both reaction coordinates $\mathbf{r} = (x,y)$ can be determined by solving $\sum_j n_{ij}(\mathbf{r}_i - \mathbf{r}_j) = 0$ with $\mathbf{r}_1 = (0,0)$, $\mathbf{r}_2 = (0,1)$, and $\mathbf{r}_3 = (1,0)$. The equation

$$\sum_j n_{ij}(\mathbf{r}_i - \mathbf{r}_j) = 0 \tag{13}$$

generalizes the p_{fold} equation (eq 2) to many dimensions. The random walk on the Sierpinski gasket embedded in such a way into 2D is diffusive, because $\Delta R^2(\Delta t) = \Delta x^2(\Delta t) + \Delta y^2(\Delta t) \sim \Delta t$. Since linear transformations of the coordinates do not change the diffusivity of the dynamics, the position of the top node can be also taken as $\mathbf{r}_2 = (1/2, \sqrt{3/2})$ so that the resulting embedding in 2D is symmetric (Figure 9). Thus, if the positions of the nodes of the Sierpinski transition network can be varied, the network can be embedded into 2D in such a way that dynamics is diffusive, i.e., the MSD grows linearly with time.

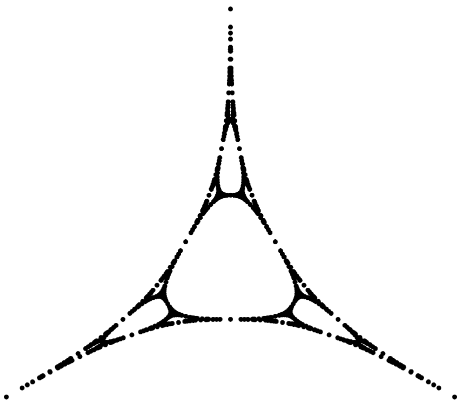


Figure 9. The network of the Sierpinski gasket embedded into 2D in such a way that the dynamics is diffusive.

The analysis suggests that dynamics described by a MSM is not diffusive or subdiffusive *per se*.[35] When analyzed along the optimal reaction coordinate, the dynamics is diffusive, while the same dynamics is subdiffusive when analyzed along the suboptimal reaction coordinate. The root MSD can be considered as the optimal reaction coordinate for spaces with simple topology. Random walk in such spaces, projected on the root MSD, shows diffusive behavior. For more complex systems, however, the Cartesian coordinates or the root MSD are not good reaction coordinates and thus exhibit subdiffusive dynamics.

Figure 10 shows the free energy profile along the optimal reaction coordinate (p_{fold}) between nodes 1 and 3. The coordinate is rescaled so that the diffusion coefficient is constant and the free energy profile provides a complete description of the dynamics.[6] The profile is self-similar and not smooth. Thus, while the optimal coordinate provides a diffusive description of the dynamics, using the free energy profile as a function of the coordinate to simplify the description of the dynamics as diffusion makes little sense.

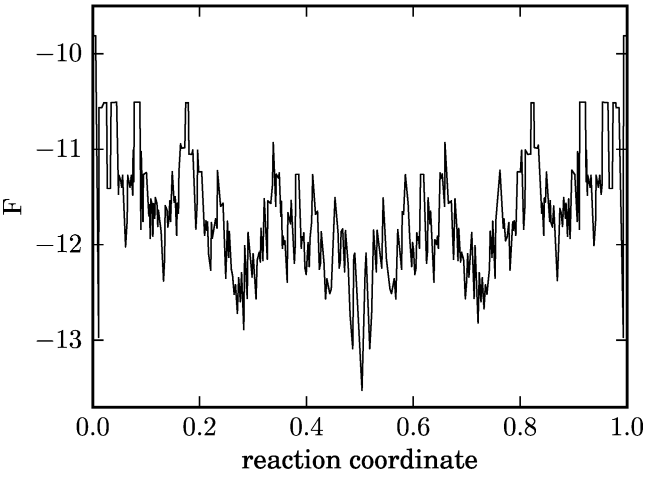


Figure 10. Free energy profile of the Sierpinski gasket along the optimal reaction coordinate, rescaled so that the diffusion coefficient is constant.

System with Multiple Parallel 1D Pathways. The sawtooth shape of the profile shown in Figure 10 is likely to be a peculiarity of the Sierpinski gasket, its self-similarity, discreteness, and low dimensionality. A realistic multidimensional system with multiple pathways is likely to have a free energy profile which is inherently smooth. The existence of multiple pathways is assumed, for example, in the "funnel" picture of protein folding.[40] To illustrate this point, ideally, a model high-dimensional system should be analyzed. However, the construction of the exact optimal reaction coordinate (p_{fold}) for such a system is computationally very demanding. A simple model system, which retains the essential property of the high dimensional systems—the existence of multiple pathways—is considered instead. It is a generalization of the two pathway system considered before (Figure 5) to multiple parallel pathways with different potential energy profiles. The potential energy profile along each path is taken as $U(x) = \cos(2\pi x - \pi) + \Delta U(x)$ for $x \in [0,1]$, where $\Delta U(x)$ denotes the random part of the profile $\Delta U(x) = 0.25\sum_{i=5}^{10}\xi_i \cos[i(2\pi x - \pi)] + \eta_i \sin[i(2\pi x - \pi)]$, where ξ_i and η_i are random variables, uniformly distributed between −1 and +1, different for every pathway. The random, fluctuating part of the profiles mimics the potential landscape ruggedness due to, e.g., the dihedral angle barriers. For every pathway, the optimal coordinate is trivially constructed, and all of the pathways are projected onto the coordinate.

Figure 11 shows how the "total" free energy profile along the optimal reaction coordinate depends on the number of pathways considered. For a single pathway, where $F(x) = U(x)$, the fluctuations are present. As the number of pathways increases, the fluctuations become "averaged out," so that the free energy profile becomes inherently smooth. The p_{fold} reaction coordinate here is rescaled so that the diffusion coefficient is constant and the dynamics is described completely by the free energy profile alone. To summarize, the single optimal reaction coordinate with a smooth free energy profile provides a quantitatively accurate description of the stochastic dynamics between the two boundary nodes along all the pathways with a "rugged" potential energy landscape. Note, that in simulation, where the sampling is often very limited, it is possible that the determined free energy landscape will show some "residual" ruggedness due to the limited number of pathways sampled.

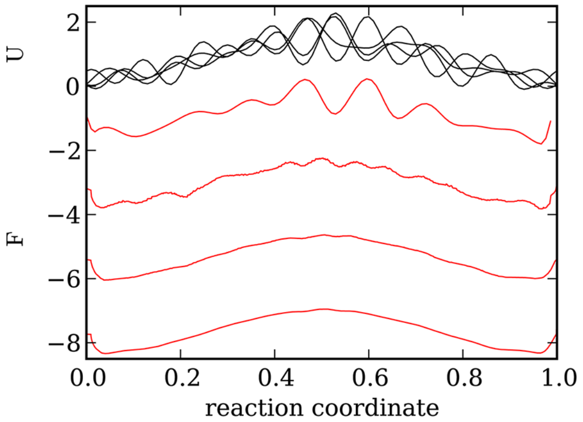


Figure 11. The free energy profile F along the optimal reaction coordinate for the system with multiple pathways. Each pathway has random small scale fluctuations on top of the $U(x) = \cos(2\pi x - \pi)$ profile. Black lines show potential energy profiles along some of the pathways. Red lines show F, determined using ensembles of 1, 10, 100, and 1000 pathways (from top to bottom). The optimal reaction coordinate is rescaled so that the diffusion coefficient is constant. As the number of pathways increases, the free energy profile becomes inherently smooth, since the fluctuations get averaged out.

FIP35. Finally, the developed approach is applied to test the optimality of various reaction coordinates constructed to analyze the folding dynamics of the FIP35 protein simulated by Shaw and co-workers.[13] The following coordinates are considered: the p_{fold} coordinate determined by optimizing $\int dx/Z_{C,1}(x)$;[12] the coordinate employed by Shaw et al.,[13] determined by the variational approach of Best and Hummer;[10] and the RMSD from the native structure.

Figure 12a shows the free energy profiles along the coordinates between the native and denatured states. As the native and denatured states (the nodes A and B), the configurations to the left (right) from the bottom of the native (denatured) basins were taken, respectively. Figure 12b shows the dependence of $F_{C,1}$ on the sampling interval Δt for the coordinates. The (approximate) p_{fold} coordinate,[12] while being much better than the other coordinates, is still slightly suboptimal. The dynamics along other coordinates become approximately diffusive on the time scale of 500−1000 steps (100−200 ns). While the RMSD coordinate has a profile which is slightly more informative than that of the coordinate used by Shaw et al. (e.g., it shows an intermediate state), it describes the dynamics significantly worse. The limiting long time value of $F_{C,1}/kT$ for the RMSD coordinate is different from the exact number of −ln(7.5) because the coordinate does not properly separate the native and denatured states and overestimated the number of folding-unfolding events.

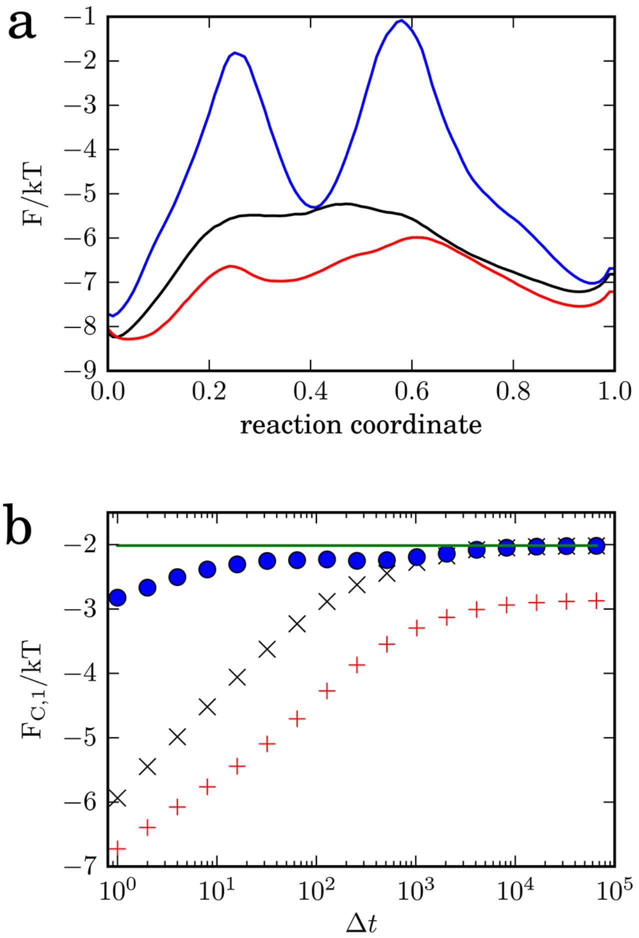


Figure 12. (a) The free energy profile of the FIP35 folding simulation along different reaction coordinates: the p_{fold} coordinate determined by optimizing $\int dx/Z_{C,1}(x)$[12] (blue line); the coordinate employed by Shaw et al. and determined by the variational approach of Best and Hummer[10] (black line); the RMSD from the native structure (red line). (b) Δt dependence of $F_{C,1}$ for the corresponding coordinates is shown with blue circles, black crosses, and red pluses, respectively. The green line shows the theoretical optimal value of −ln(7.5).

MSM Validation. If a MSM is known, the optimal reaction coordinate between any two nodes (p_{fold}) can be constructed analytically. In this case, the reaction coordinate optimality test is actually a test for the MSM, i.e., whether the constructed model is indeed Markovian. It can be used in addition to other tests for MSM.[15] To illustrate this, we consider three MSMs for the beta3s peptide,[41] constructed by clustering a long equilibrium MD trajectory in different ways: all-atom RMSD clustering with a 2.0 Å threshold, all-atom RMSD clustering with a 2.5 Å threshold, and secondary structure clustering.[6,42] Figure 13 shows that the MSM obtained with 2.5 Å clustering provides a Markovian description of the dynamics. The optimal reaction coordinates for the MSM obtained with 2.0 Å clustering "overfits" the data. Due to the small threshold, some of the links in the transition network (MSM) are missing.

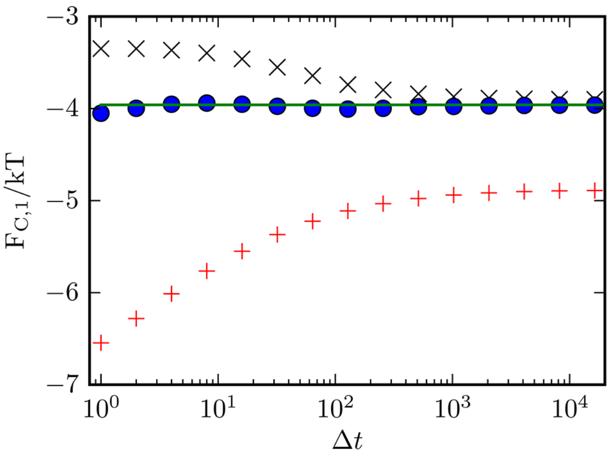


Figure 13. Δt dependence of $F_{C,1}$ along the optimal reaction coordinate for MSM of bets3s peptide obtained with different clustering procedures: 2.0 Å all-atom RMSD clustering (black crosses), 2.5 Å all-atom RMSD clustering (blue circles), and secondary structure clustering (red pluses). The green line shows the theoretical optimal value.

As a result, the number of folding−unfolding transitions is underestimated. In contrast, the MSM obtained with secondary structure clustering has many spurious links between kinetically different regions of the configuration space.[6,42] As a result, it does not provide a Markovian description of the dynamics.[6,42]

4. CONCLUSION

It has been shown that for equilibrium dynamics described by a Markov model, it is always possible to construct the optimal reaction coordinate that quantitatively describes the dynamics between any two states. The optimal coordinate has the highest $F_{C,1}$ profile (the smallest $Z_{C,1}$). $Z_{C,1}$ as a function of the optimal reaction coordinate is position and sampling interval independent and equals the total number of transitions from one state to the other. This criterion can be used to test the quality of reaction coordinates or Markov state models employed for the analysis of long equilibrium MD simulations or other types of complex dynamics. The criterion generalizes the criterion based on F_C profiles.[6,35] The latter was derived assuming that the free energy profile is relatively flat, which limits its applicability to transition state regions and/or small sampling intervals.

For a system, described by a MSM, the dynamics is not diffusive or subdiffusive *per se*. It is diffusive when projected on the optimal reaction coordinate. The same dynamics is subdiffusive when projected on a suboptimal reaction coordinate. This result demonstrated before using F_C profiles is applicable only to relatively flat regions of free energy profile.[35] Usage of the $F_{C,1}$ profiles allowed us to extend the result to arbitrary landscapes and sampling intervals. The result can be also interpreted as follows. Employing the Mori−Zwanzig formalism,[37,38] one can derive generalized Langevin equations which describe system dynamics projected on a coordinate. The generalized Langevin equation contains a memory kernel which leads to non-Markovian dynamics. To completely specify dynamics in this case one has to compute the memory kernel, which is not trivial.[39] The proposed approach suggests an alternative strategy. Instead of employing conventional, simple coordinates (e.g., the Cartesian coordinates), one seeks more complex, optimal coordinates, which minimize the non-Markovian contributions from the kernel, so that the stochastic dynamics is accurately described by just the free energy profile and diffusion coefficient.

The free energy profile for a realistic multidimensional system with multiple pathways is inherently smooth. The dynamics between two states along all the pathways on a rugged potential energy surface is quantitatively described as diffusion along a single reaction coordinate with a smooth free energy profile. In particular, the equilibrium flux from one boundary state (A) to the other (B) equals $J_{AB} = Z_{C,1}(p_{\text{fold}})/Z$ or

$$J_{AB}^{-1} = Z\max \int_A^B \frac{dx}{e^{-F(x)/kT}D(x)}$$

where the maximum is taken over all reaction coordinates connecting the states and Z is the total partition function. The equation is valid for any equilibrium Markovian dynamics and generalizes the results obtained before by employing the harmonic approximation for the transition state[18] or assuming constant diffusion coefficient.[19]

APPENDIX

A System with Three States

Here, we illustrate in detail how the cut profiles and the MSD are computed by using the transition path segments. Consider a system with three states A, B, and C with the following transition probabilities: $P_{AB} = P_{CB} = 1/2$, $P_{BA} = P_{BC} = 1$. The reaction coordinate is constructed between nodes A and C, so that $x_A = 0$, $x_B = 0.5$, and $x_C = 1$. The following transition path segments are possible (with equal probabilities): ABA, ABC, CBC, and CBA. The equilibrium number of segments of each type is $T/(8\Delta t)$ for a trajectory of length T with sampling interval of Δt.

We start with the MSD, which is more familiar to the reader than the cut profiles. Segment ABA is extended as ...AAAABAAAA... For this segment, one obtains $\Delta x^2(\Delta t) = \sum_i[x(i\Delta t) - x(i\Delta t + \Delta t)]^2 = (x_A - x_B)^2 + (x_B - x_A)^2$; $\Delta x^2(2\Delta t) = \sum_i[x(i\Delta t) - x(i\Delta t + 2\Delta t)]^2 = (x_A - x_B)^2 + (x_B - x_A)^2 + (x_A - x_A)^2$; $\Delta x^2(k\Delta t) = \sum_i[x(i\Delta t) - x(i\Delta t + k\Delta t)]^2 = (x_A - x_B)^2 + (x_B - x_A)^2$. Segment ABC is extended as ...AAAABCCCC... For this segment, one obtains $\Delta x^2(\Delta t) = \sum_i[x(i\Delta t) - x(i\Delta t + \Delta t)]^2 = (x_A - x_B)^2 + (x_B - x_C)^2$; $\Delta x^2(2\Delta t) = \sum_i[x(i\Delta t) - x(i\Delta t + 2\Delta t)]^2 = (x_A - x_B)^2 + (x_B - x_C)^2 + (x_A - x_C)^2$; $\Delta x^2(k\Delta t) = \sum_i[x(i\Delta t) - x(i\Delta t + k\Delta t)]^2 = (x_A - x_B)^2 + (x_B - x_C)^2 + (k - 1)(x_A - x_C)^2$. For segment CBA, one obtains the same results as for ABC (since it is the same sequence just reversed). Segment CBC is extended as ...CCCCBCCCC... For this segment, one obtains $\Delta x^2(\Delta t) = \sum_i[x(i\Delta t) - x(i\Delta t + \Delta t)]^2 = (x_C - x_B)^2 + (x_B - x_C)^2$; $\Delta x^2(2\Delta t) = \sum_i[x(i\Delta t) - x(i\Delta t + 2\Delta t)]^2 = (x_C - x_B)^2 + (x_B - x_C)^2 + (x_C - x_C)^2$; $\Delta x^2(k\Delta t) = \sum_i[x(i\Delta t) - x(i\Delta t + k\Delta t)]^2 = (x_C - x_B)^2 + (x_B - x_C)^2$. The total sum over all (equiprobable) segments is $\Delta x^2(k\Delta t) = T/(8\Delta t)[4(x_A - x_B)^2 + 4(x_B - x_C)^2 + 2(k - 1)(x_A - x_C)^2] = Tk/(4\Delta t)$, and the MSD is $\Delta x^2(k\Delta t)/(T/\Delta t) = k/4$. The MSD grows linear with time because the sum is computed over the segments growing in size but is divided by the original trajectory length, which is constant.

The cut profile $Z_{C,1}$ is computed as follows. Segment ABA (extended as ...AAAABAAAA...), contributes to the profile with $x_A < x < x_B$: $Z_{C,1}(x,\Delta t) = |x_A - x_B|$; $Z_{C,1}(x,2\Delta t) = 1/2|x_A - x_B| + 1/4|x_A - x_A|$; $Z_{C,1}(x,k\Delta t) = 1/k|x_A - x_B|$. Segment CBC (extended as ...CCCCBCCCC...) contributes to the profile with

$x_B < x < x_C$: $Z_{C,1}(x,\Delta t) = |x_B - x_C|$; $Z_{C,1}(x,2\Delta t) = 1/2|x_B - x_C| +1/4|x_C - x_C|$; $Z_{C,1}(x,k\Delta t) = 1/k|x_B - x_C|$. The contribution of ABC segment (extended as ...AAAABCCCC...) consists of three parts: for transitions from A to B, $x_A < x < x_B$: $Z_{C,1}(x,k\Delta t) = 1/2k|x_A - x_B|$; for transitions from B to C, $x_B < x < x_C$: $Z_{C,1}(x,k\Delta t) = 1/2k|x_B - x_C|$; for transitions from A to C, $x_A < x < x_C$ and $k > 1$: $Z_{C,1}(x,k\Delta t) = (k - 1)/2k|x_A - x_C|$. Segment CBA is equivalent to ABC. By summing up all the segments with a weight of $T/(8\Delta t)$, one obtains for the total cut profile for $x_A < x < x_B$ (and the same for $x_B < x < x_C$): $Z_{C,1}(x,k\Delta t) = T/(8\Delta t)[1/k|x_A - x_B|+2/2k|x_A - x_B|+2(k - 1)/2k|x_A - x_C|] = T/(8\Delta t)$. The equilibrium flux is $J_{AC} = Z_{C,1}(x,k\Delta t)/T = 1/(8\Delta t)$ and is equal to the reciprocal of the sum of the mean first passage times between nodes A and C $\langle t_{AC}\rangle = \langle t_{CA}\rangle = 4\Delta t$. For the MSD, one obtains (eq 12) $\langle \Delta x^2(k\Delta t)\rangle = 2kZ_{C,1}/Z = 2kT/(8\Delta t)/(T/\Delta t) = k/4$ in agreement with the above.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: s.krivov@leeds.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work has been supported in part by an RCUK fellowship and BBSRC grant (BB/J016055/1). The author is grateful to Attila Szabo for critical comments.

■ REFERENCES

(1) Mu, Y.; Nguyen, P. H.; Stock, G. *Proteins* **2005**, *58*, 45−52.

(2) Allen, L. R.; Krivov, S. V.; Paci, E. *PLOS Comput. Biol.* **2009**, *5*, e1000428.

(3) Hori, N.; Chikenji, G.; Berry, R. S.; Takada, S. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 73−78.

(4) Ferguson, A. L.; Panagiotopoulos, A. Z.; Kevrekidis, I. G.; Debenedetti, P. G. *Chem. Phys. Lett.* **2011**, *509*, 1−11.

(5) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. *J. Chem. Phys.* **2011**, *134*, 124111.

(6) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 13841−13846.

(7) Krivov, S. V. *J. Phys. Chem. B* **2011**, *115*, 12315−12324.

(8) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334−350.

(9) Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. *J. Chem. Phys.* **2008**, *129*, 174102.

(10) Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6732−6737.

(11) Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.

(12) Krivov, S. V. *J. Phys. Chem. B* **2011**, *115*, 11382−11388.

(13) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341−346.

(14) Chodera, J. D.; Pande, V. S. *Phys. Rev. Lett.* **2011**, *107*, 098102.

(15) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.

(16) E, W.; Vanden-Eijnden, E. *J. Stat. Phys.* **2006**, *123*, 503−523.

(17) Kramers, H. A. *Physica* **1940**, *7*, 284−304.

(18) Berezhkovskii, A.; Szabo, A. *J. Chem. Phys.* **2005**, *122*, 14503.

(19) E, W.; Vanden-Eijnden, E. In *Multiscale Modelling and Simulation*, 1st ed.; Attinger, S., Koumoutsakos, P., Eds.; Springer: New York, 2004; p 277.

(20) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Chem. Phys.* **2009**, *130*, 205102.

(21) Metzner, P.; Schuette, C.; Vanden-Eijnden, E. *Multiscale Model. Simul.* **2009**, *7*, 1192−1219.

(22) Havlin, S.; Ben-Avraham, D. *Adv. Phys.* **1987**, *36*, 695−798.

(23) Neusius, T.; Daidone, I.; Sokolov, I. M.; Smith, J. C. *Phys. Rev. Lett.* **2008**, *100*, 188103.

(24) García, A. E.; Blumenfeld, R.; Hummer, G.; Krumhansl, J. A. *Phys. D (Amsterdam, Neth.)* **1997**, *107*, 225−239.

(25) Kneller, G. R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2641−2655.

(26) Min, W.; Luo, G.; Cherayil, B. J.; Kou, S. C.; Xie, X. S. *Phys. Rev. Lett.* **2005**, *94*, 198302.

(27) Michalet, X.; Weiss, S.; Jager, M. *Chem. Rev.* **2006**, *106*, 1785−1813.

(28) Luo, G.; Andricioaei, I.; Xie, X. S.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 9363−9367.

(29) Matsunaga, Y.; Li, C.-B.; Komatsuzaki, T. *Phys. Rev. Lett.* **2007**, *99*, 238103.

(30) Senet, P.; Maisuradze, G. G.; Foulie, C.; Delarue, P.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 19708−19713.

(31) Li, C.-B.; Yang, H.; Komatsuzaki, T. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, 536−541.

(32) Granek, R.; Klafter, J. *Phys. Rev. Lett.* **2005**, *95*, 098106.

(33) Magdziarz, M.; Weron, A.; Burnecki, K.; Klafter, J. *Phys. Rev. Lett.* **2009**, *103*, 180602.

(34) Sangha, A. K.; Keyes, T. *J. Phys. Chem. B* **2009**, *113*, 15886−15894.

(35) Krivov, S. V. *PLOS Comput. Biol.* **2010**, *6*, e1000921.

(36) Cote, Y.; Senet, P.; Delarue, P.; Maisuradze, G. G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 10346−10351.

(37) Mori, H. *Prog. Theor. Phys.* **1965**, *33*, 423−455.

(38) Zwanzig, R. *Nonequilibrium Statistical Mechanics*; Oxford University Press: New York, 2001.

(39) Darve, E.; Solomon, J.; Kia, A. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 10884−10889.

(40) Onuchic, J. N.; Socci, N. D.; Luthey-Schulten, Z.; Wolynes, P. G. *Folding Des.* **1996**, *1*, 441−450.

(41) Ferrara, P.; Caflisch, A. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 10780−10785.

(42) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701−8714.