THE JOURNAL OF PHYSICAL CHEMISTRY B

ARTICLE

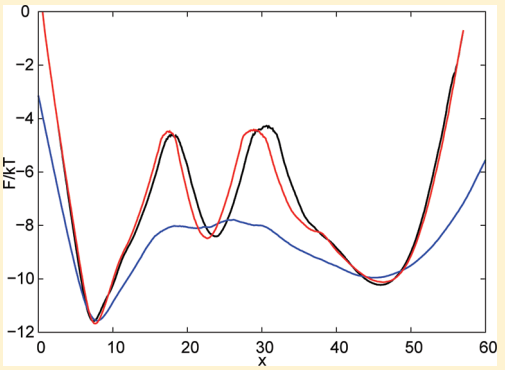
pubs.acs.org/JPCB

Numerical Construction of the p_{fold} (Committor) Reaction Coordinate for a Markov Process

Sergei V. Krivov*

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

ABSTRACT: To simplify the description of a complex multidimensional dynamical process, one often projects it onto a single reaction coordinate. In protein folding studies, the folding probability p_{fold} is an optimal reaction coordinate which preserves many important properties of the dynamics. The construction of the coordinate is difficult. Here, an efficient numerical approach to construct the p_{fold} reaction coordinate for a Markov process (satisfying the detailed balance) is described. The coordinate is obtained by optimizing parameters of a chosen functional form to make a generalized cut-based free energy profile the highest. The approach is illustrated by constructing the p_{fold} reaction coordinate for the equilibrium folding simulation of FIP35 protein reported by Shaw et al. (*Science* **2010**, *330*, 341−346).



INTRODUCTION

Reaction coordinates are often introduced to simplify the description and analysis of complex multidimensional dynamical process. In particular, in studies of protein folding dynamics, the reaction coordinates and associated free energy landscapes are used to describe the complex dynamics in a clear and intuitive way as diffusion on the free energy landscape.[1,2] The construction of an optimal reaction coordinate, which provides accurate quantitative description of dynamics, is challenging. The coordinate should absorb all the important dynamical information contained in the other degrees of freedom, so they can be neglected. The conventional reaction coordinates, constructed based on physical intuition (e.g., the number of native contacts, the root-mean-square distance from the native state, the radius of gyration), or obtained via generic dimensionality reduction techniques (e.g., the principal component analysis) do not take the dynamics into account and thus often fail to accurately describe the dynamics.[3,4] In particular, the dynamics when projected on such suboptimal coordinates is non-Markovian and subdiffusive.[5]

An approach to construct reaction coordinates based on the system dynamics has been described recently.[5] It defines the optimal coordinate as the one with the highest cut-based free energy profile (cFEP).[6,7] The coordinate is constructed by numerically optimizing parameters of a chosen functional form, so that the cFEP (computed from simulated trajectories) is the highest. The cFEP is complementary and superior to the conventional (histogram) free energy profile as it is invariant to reaction coordinate rescaling, insensitive to statistical noise, and capable of detecting subdiffusion. Together they determine the coordinate-dependent diffusion coefficient $D(x)$ and thus completely specify diffusive dynamics.[7] The approach, originally developed for the analysis of protein folding dynamics, was successfully applied to construct the optimal reaction coordinate for the game of chess. The dynamics of the chess game is quantitatively described as diffusion on the associated free energy landscape.[8] The successful application suggests that the approach and the free energy landscape framework, in general, can be used for the analysis of many important complex dynamical phenomena, for example, disease dynamics and economics.

The folding probability (P_{fold}) or committor is another reaction coordinate that takes dynamics into account. It was first mentioned as the splitting probability by Onsager[9] and is often used in protein folding studies.[10−13] For the description of protein folding dynamics it equals the probability for the trajectory to reach the native state before reaching the denatured state, starting from a given configuration. It has been shown that this coordinate provides optimal dividing surfaces for the milestoning algorithm[14] and that diffusive dynamics on the free energy surface projected on this coordinate has the same mean first passage time as the original one.[14,15] In describing the dynamics of the chess game, the p_{fold} coordinate denotes the probability of winning the game from a current position, arguably, the most important quantitative information about the game.[8]

While a number of approaches have been suggested to construct the p_{fold} coordinate, they all have limitations. The p_{fold} coordinate can be found from the transition matrix of a Markov state model.[6,16] However, an accurate Markov state model is difficult to construct. The coordinate can be found by the string method, assuming that a transition pathway can be closely approximated by a narrow transition tube.[17] The approximation is not always valid.[18] The coordinate around the $p_{fold} = 0.5$ can be found by optimizing the probability of being on a transition path $p(TP|r)$.[11] Assuming that $p(TP|r) = p_0(1 - \tanh[r]^2)$ the

Received: June 3, 2011
Revised: July 15, 2011
Published: August 23, 2011

coordinate can be constructed from transition path trajectories by maximum likelihood.[19] The last two approaches assume that the transition state is defined by $p_{fold} = 0.5$. The assumption, reasonable for systems with a single dominant barrier, might not be valid for systems with an intermediate state, as shown by the illustrative example.

Here I describe a new, computationally efficient approach to construct the p_{fold} reaction coordinate for Markov process, satisfying the detailed balance. The p_{fold} coordinate is constructed by numerical optimization of the parameters of a chosen functional form, so that a generalized cut free energy profile is maximal. The approach is similar to the one that constructs thet reaction coordinate by optimizing the cut free energy profile;[5] they differ only by the employed cut profiles. The approach is illustrated by constructing reaction coordinate and associated free energy landscape for the all-atom equilibrium folding simulation of FIP35 protein reported by Shaw et al.[20]

GENERALIZED CUT FREE ENERGY PROFILE

The reaction coordinate time series is computed as $x(i\Delta t) = R(\vec{X}(i\Delta t))$, where $\vec{X}(i\Delta t)$ is a multidimensional trajectory recorded with time interval Δt and $x = R(\vec{X})$ defines the reaction coordinate. The partition function of the conventional histogram-based free energy profile in a bin $[x_i, x_i + \Delta x]$ equals to the density of points in the bin

$$Z_H(x) = N/\Delta x \quad \text{for} \quad x \in [x_i, x_i + \Delta x] \tag{1}$$

where N is the number of time-series points in the bin and Δx is the size of the bin. The partition function of the cut-based free energy profile equals half the total number of transition trough point x:

$$Z_C(x) = 1/2 \sum_i \Theta\{(x(i\Delta t) - x)(x - x(i\Delta t + \Delta t)\} \tag{2}$$

where $\Theta\{x\}$ is the Heaviside step function. The corresponding free energies are $F_H(x) = -kT \ln Z_H(x)$ and $F_C(x) = -kT \ln Z_C(x)$. For diffusive dynamics (with Gaussian increments Δx)

$$Z_C(x) = \sqrt{D(x)\Delta t/\pi} Z_H(x) \tag{3}$$

which can be used to estimated the coordinate-dependent diffusion coefficient $D(x)$.[7] To transform reaction coordinate x with variable diffusion coefficient $D(x)$ to coordinate y with constant diffusion coefficient $D(y)$, one numerically integrates $dy = Z_H(x)/Z_C(x)dx$. In the latter case $F_C(y)$ differs from $F_H(y)$ by an unimportant constant.

$Z_{C,r}(x)$ generalizes $Z_C(x)$ by summing up the transitions through point x with weights equal to the length of the transition to the power r:

$$Z_{C,r}(x) = 1/2 \sum_i |x(i\Delta t + \Delta t) - x(i\Delta t)|^r \Theta\{(x(i\Delta t) - x)(x - x(i\Delta t + \Delta t)\} \tag{4}$$

For transitions in the positive direction, from $y < x$ to $y + \Delta x > x$, one has

$$Z^+_{C,r}(x) = \int_{y=-\infty}^{x} Z_H(y)\, dy \int_{y+\Delta x=x}^{y+\Delta x=\infty} P(\Delta x)\, d\Delta x |\Delta x|^r = \int_{\Delta x=0}^{\Delta x=\infty} P(\Delta x)\, d\Delta x |\Delta x|^r \int_{x-\Delta x}^{x} Z_H(y)\, dy$$

Assuming that $Z_H(y)$ is approximately constant on the distance of the mean absolute displacement $\langle|\Delta x(\Delta t)|\rangle$ the second integral can be taken

$$Z^+_{C,r}(x) = \int_{\Delta x=0}^{\Delta x=\infty} P(\Delta x)\, d\Delta x |\Delta x|^{r+1} Z_H(x)$$

Combining with the corresponding value for transitions in the negative direction

$$Z^-_{C,r}(x) = \int_{\Delta x=-\infty}^{\Delta x=0} P(\Delta x)\, d\Delta x |\Delta x|^{r+1} Z_H(x)$$

one obtains

$$Z_{C,r}(x) = 1/2(Z^+_{C,r}(x) + Z^-_{C,r}(x)) = Z_H(x)\langle|\Delta x(\Delta t)|^{1+r}\rangle/2 \tag{5}$$

Two cases are of particular interest: with $r = -1$

$$Z_{C,-1}(x) = Z_H(x)/2 \tag{6}$$

and $r = 1$

$$Z_{C,1}(x) = Z_H(x)\langle\Delta x^2(\Delta t)\rangle/2 = \Delta t D(x) Z_H(x) \tag{7}$$

Equation 6 computes the conventional (histogram) partition function $Z_H(x)$. In contrast to eq 1, $Z_H(x)$ is defined for every point, not per bin. Equation 7 allows one to determine the diffusion coefficient, which can be determined by eq 3 only for a special, though important, case, when the distribution of jumps $P(\Delta x)$ is Gaussian.

If the mean square displacement scales with time as $\langle\Delta x^2(\Delta t)\rangle \sim \Delta t^{2\alpha}$, then $Z_{C,1}$ scales with the time interval Δt as $Z_{C,1}(x,\Delta t) \sim t^{2\alpha-1}$ (the profiles are computed by varying the Δt, while the length of the trajectory is fixed, that is, $Z_H \sim 1/\Delta t$). For diffusive dynamics, $\alpha = 0.5$, and $Z_{C,1}(x)$ does not change with varying Δt. α can be estimated from $Z_{C,1}(x)$ computed with two different time intervals

$$\alpha(x) = 1/2 + \frac{\ln Z_{C,1}(x, \Delta t_2) - \ln Z_{C,1}(x, \Delta t_1)}{2(\ln \Delta t_2 - \ln \Delta t_1)}$$

Reaction Coordinates for the Markov Process. Consider a Markov process $p_i(t + \Delta t) = \Sigma_j P_{ij} p_j(t)$, 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 by equation $p_i^{eq} = \Sigma_j P_{ij} p_j^{eq}$ with shorthand notation $n_i = p_i^{eq}$. The equilibrium number of transitions between nodes i and j is $n_{ij} = P_{ij} p_j^{eq} = P_{ji} p_i^{eq} = n_{ji}$ (the detailed balance), with $n_i = \Sigma_j n_{ij}$ and $P_{ij} = n_{ij}/n_j$. n_{ij} (accordingly n_i and P_{ij}) can be estimated from a trajectory (a realization of the stochastic process) by counting the number of transition from node j to node i. The folding probability can be found from

$$p_i^{fold} = \sum_j P_{ji} p_j^{fold} \tag{8}$$

with boundary conditions $p_A^{fold} = 0$ and $p_B^{fold} = 1$.[21] For rescaled probability vector $\psi_i = p_i/(n_i)^{1/2}$ the stochastic dynamics is described by $\psi_i(t + \Delta t) = \Sigma_j K_{ij} \psi_j(t)$ with symmetric matrix $K_{ij} = n_{ij}/(n_i n_j)^{1/2}$. The eigenvalues and eigenvectors of the symmetric matrix are defined by

$$\lambda v_i = \sum_j K_{ij} v_j \tag{9}$$

The largest eigenvalue equals $\lambda_1 = 1$ with the corresponding eigenvector describing equilibrium probability distribution $v_i^1 = (n_i)^{1/2}$. If there is a gap in the spectrum $1 - \lambda_2 \ll 1 - \lambda_3 < 1 - \lambda_4$, and so forth, for example, as in the two state folding process, the second largest eigenvalue λ_2 together with associated eigenvector v^2 describe the long time asymptotics of dynamics and can be used to distinguish the two states by the sign of v_i^2. Berezhkovskii and Szabo have shown that $p_i^{\text{fold}} \approx a(v_i^2/v_i^1 + b)$ around the transition state region.[22] Using the second eigenvector instead of p_{fold} might be more convenient since one does not have to pick up the (boundary) nodes A and B which should be on different sides of the transition state.

If the equilibrium n_{ij} can be computed, then the p_{fold} reaction coordinate can be obtained from eq 8.[6] However, the construction of an accurate Markov state model from a simulation (e.g., protein folding trajectories) is challenging.[23] For example, the requirements on the size of the states are contradictory. The states should be taken as small as possible to obtain Markovian dynamics. At the same time they should be as large as possible (particularly, in high-dimensional space) to have reasonable statistics to estimate n_{ij}. I now show how the p_{fold} reaction coordinate can be found by optimizing functional $\int A(Z_{C,1}(x))dx$, without constructing the Markov state model.

Construction of the p_{fold} Reaction Coordinate for a Markov Process. Let x_i be the position of node i on the reaction coordinate. The value of the cut profile $Z_{C,1}$ at point x equals to the sum $1/2|x_i - x_j|n_{ij}$ over such i and j that x_i and x_j are on opposite sides of x. Consider first the simple functional $\int Z_{C,1}(x)dx$. $Z_{C,1}(x)$ for a single transition from point x_i to point x_j is a rectangular pulse, which equals $|x_i - x_j|/2$ for x between x_i and x_j and 0 otherwise. The functional for the single transition equals $\int Z_{C,1}(x)dx = (x_i - x_j)^2/2$. Since the construction of the cut profiles and integration are linear operations, the functional for all transitions equals the sum $\int Z_{C,1}(x)dx = (1/2)\Sigma_{i,j}(x_i - x_j)^2 n_{ij}$.

The derivative of the functional $\int A(Z_{C,1}(x))dx$ with respect to the change of the position x_i of node i is given by

$$\frac{\partial}{\partial x_i}\int A(Z_{C,1}(x))\,dx = \int A'(Z_{C,1}(x))\frac{\partial Z_{C,1}(x)}{\partial x_i}dx$$

$$= \int A'(Z_{C,1}(x))\frac{\partial Z^i_{C,1}(x)}{\partial x_i}dx = \int_{I_{x_i}} A'(Z_{C,1}(x))\frac{\partial Z^i_{C,1}(x)}{\partial x_i}dx$$

where $A' = dA(Z)/dZ$, $Z^i_{C,1}(x)$ denotes the part of the cut profile due to transitions from or to node i and is zero for x not in I_{x_i}, where I_{x_i} is the segment $[\min(x_j), \max(x_j)]$ which contains all of the nodes j connected with node i. Assume that $A'(Z_{C,1}(x))$ is approximately constant for x in I_{x_i}, so that it can be taken out of the integral sign. The size of I_{x_i}, which is about $|x_i - x_j| \sim (D\Delta t)^{1/2}$, can be made sufficiently small by selecting an appropriate Δt, so that the assumption is valid. In regions where $Z_{C,1}(x)$ does not change much, that is, around local minima of the free energy profile and transition states, the assumption is valid at larger Δt. Then

$$A'(Z_{C,1}(x_i))\int_{I_{x_i}}\frac{\partial Z^i_{C,1}(x)}{\partial x_i}dx = A'(Z_{C,1}(x_i))\frac{\partial}{\partial x_i}\int_{I_{x_i}} Z^i_{C,1}(x)\,dx$$

Remembering that

$$\int_{I_{x_i}} Z^i_{C,1}(x)\,dx = \frac{1}{2}\sum_j (x_i - x_j)^2(n_{ij} + n_{ji})$$

one finds

$$\frac{\partial}{\partial x_i}\int A(Z_{C,1}(x))\,dx = A'(Z_{C,1}(x_i))\frac{1}{2}\frac{\partial}{\partial x_i}\sum_j (x_i - x_j)^2(n_{ij} + n_{ji}) \quad (10)$$

The functional attains extremum $(\partial/\partial x_i \int A(Z_{C,1}(x))dx = 0)$ at

$$x_i = \frac{\sum_j (n_{ij} + n_{ji})x_j}{\sum_j (n_{ij} + n_{ji})} = \sum_j P_{ji}x_j \quad (11)$$

which, supplemented by boundary conditions $x_A = 0$ and $x_B = 1$, defines the folding probability $x_i = p_i^{\text{fold}}$; see eq 8.

Let the reaction coordinate be defined by some functional form $R(\vec{X},\vec{\alpha})$, where $\vec{\alpha}$ is a vector of the parameters. The position of state (i) of the Markov process on the reaction coordinate is $x_i = R(\vec{X}_i,\vec{\alpha})$. The cut profile computed from the network n_{ij} equals to the cut profile computed from the process time series, assuming sufficient sampling. This eliminates the cumbersome procedure of constructing the network. The functional, computed from the time series, attains extremum (with respect to the parameter vector) when the reaction coordinate equals the folding probability $R(\vec{X},\vec{\alpha}^*) = p_{\text{fold}}(\vec{X})$.

Function $A(x)$ is chosen such that its optimization makes the free energy profile higher ($Z_{C,1}(x)$ smaller), for example, min $\int Z_{C,1}(x)dx$, max $\int -\ln Z_{C,1}(x)dx$, or max $\int 1/Z_{C,1}(x)dx$. If the functional form of reaction coordinate $R(\vec{X},\vec{\alpha})$ is very flexible, so that it can accurately approximate $p_{\text{fold}}(\vec{X})$, the exact choice of the optimization function A is not important. Otherwise, when not all of the parts of the profile are optimal, function A defines the relative importance (weight) of different parts of the profile. In particular, min $\int Z_{C,1}(x)dx$ optimizes mainly regions with low free energy (large $Z_{C,1}$), that is, free energy basins, while max $\int 1/Z_{C,1}(x)dx$ optimizes mainly high free energy regions (small $Z_{C,1}$), that is, transition states.

The functional $\int 1/Z_{C,1}(x)dx$ has a useful property. It is invariant with respect to reaction coordinate rescaling by an arbitrary monotonic function $R(x) \to g(R(x))$.[12] In that case, $R(x)$ should only reproduce the order of $p_{\text{fold}}(x_i)$ which is a much simpler task than reproducing the numerical values of $p_{\text{fold}}(x_i)$. However, the property is likely to be less helpful when one projects more than one pathway onto a single reaction coordinate. In that case the corresponding p_{fold} values for different pathways need to be aligned.

If the reaction coordinate is a weighted sum of basis functions $R(\vec{X},\vec{\alpha}) = \Sigma_{k=1,N}\alpha_k r_k(\vec{X})$, then the optimal weights $(\alpha_1, ..., \alpha_N)$ that deliver extremum to $I = 2\int Z_{C,1}(x)dx = \Sigma_{ij}(x_i - x_j)^2 n_{ij} = \Sigma_i[x(i\Delta t + \Delta t) - x(i\Delta t)]^2$ can be found analytically. The functional equals $I = T\Sigma_{ij}\alpha_i\alpha_j\langle\Delta r_i\Delta r_j\rangle$, where $\Delta r_i = r_i(\vec{X}(i\Delta t + \Delta t)) - r_i(\vec{X}(i\Delta t))$, $\langle\rangle$ denotes the trajectory average, and T is the trajectory length. A minimum of I/T with boundary conditions $x_{i0} = \Sigma_{k=1,N}\alpha_k r_k(\vec{X}(i_0\Delta t)) = 0$ and $x_{i1} = \Sigma_{k=1,N}\alpha_k r_k(\vec{X}(i_1\Delta t)) = 1$ is the solution of the system of linear equations $Ax = b$, where

$$\begin{aligned} x &= \{\alpha_1, ..., \alpha_N, \lambda_0, \lambda_1\} \\ A_{ij} &= \langle\Delta r_i\Delta r_j\rangle, 1 \le i,j \le N \\ A_{i,N+1} &= A_{N+1,i} = r_i(\vec{X}(i_0\Delta t)), 1 \le i \le N \\ A_{i,N+2} &= A_{N+2,i} = r_i(\vec{X}(i_1\Delta t)), 1 \le i \le N \\ b_i &= \delta_{i,N+2} \end{aligned} \quad (12)$$

Since the functional favors optimization of the free energy minima, the solution is less useful for optimizing the transition states, the most important parts of the free energy landscape for kinetics. Optimization can be focused at the transition state region by an appropriate reweighting of the trajectory.

The analysis can be repeated for optimal reaction coordinate defined as the one with the highest $F_C(x)$ (lowest $Z_C(x)$).[5,7] The extremum of $2\int Z_C(x)dx = \Sigma_{ij}|x_i - x_j|n_{ij} = \Sigma_i|x(i\Delta t + \Delta t) - x(i\Delta t)|$ is attained when $\Sigma_{x_j<x_i} n_{ij} = \Sigma_{x_j>x_i} n_{ij}$, that is, x_i is the median of the distribution of steps. For Gaussian distribution of steps $n_{ij} \sim e^{-(x_0-x_j)^2/4D\Delta t}$ the median (x_0) coincides with the mean (eq 9), and the two definitions of the optimal reaction coordinate are equivalent. It is convenient to optimize $\int A(Z_C(x))Z_H(x)dx$ which is invariant to trivial rescaling transformation of the reaction coordinate.[5]

A similar variational principle with functional $\int \exp(F(q))/\langle|\nabla q|^2\rangle dq$ was established for Langevin dynamics with a constant diffusion coefficient.[12,17] The present analysis differs in the following. The cut profile $Z_{C,1}$ is a function of reaction coordinate time series only, which is much easier to compute than, for example, $\langle|\nabla q|^2\rangle$. Precise knowledge about the dynamics in the original configuration space is not required. For example, the optimal reaction coordinate that describes the dynamics of the chess game was constructed by optimizing Z_C without knowing "the equations of motion".[8] One may construct the reaction coordinate using a "proxy" space of feature variables, for example, collective degrees of freedom. The variational principle is valid for any Markov process satisfying the detailed balance, with Langevin dynamics being a particular case.

The extremum of $\Sigma_{ij}(x_i - x_j)^2 n_{ij}$ under constraint $\Sigma\, n_i x_i^2 = 1$ is attained when $(1-\lambda)x_i = \Sigma_j n_{ij}/n_i x_j$. The equation can be rewritten as

$$(1-\lambda)(x_i\sqrt{n_i}) = \sum_j \frac{n_{ij}}{\sqrt{n_i n_j}}(x_j\sqrt{n_j}) \qquad (13)$$

and is equivalent to eq 9. Thus, in principle, one can find eigenvectors by optimizing $I = 2\int Z_{C,1}(x)dx = \Sigma_i[x(i\Delta t + \Delta t) - x(i\Delta t)]^2$ under the constraint $\Sigma_i x^2(i\Delta t) = 1$. The lowest eigenvalue $\lambda_1 = 0$ corresponds to the (trivial) equilibrium eigenvector $x_i^1 = 1$ ($v_i^1 = (n_i)^{1/2}$). The second lowest eigenvector equals $x_i = v_i^2/(n_i)^{1/2} = v_i^2/v_i^1$, where v_i^2 is the second eigenvector of eq 9. As suggested by Berezhkovskii and Szabo,[22] it approximates the p_{fold} reaction coordinate.

If reaction coordinate is a weighted sum of basis functions $x = R(\vec{X},\vec{\alpha}) = \Sigma_{k=1,N}\alpha_k r_k(\vec{X})$ then the optimal weights $(\alpha_1, ..., \alpha_N)$ can be found analytically by minimizing $I/T = \Sigma_{ij}\alpha_i\alpha_j\langle\Delta r_i \Delta r_j\rangle$ under the constraint $\Sigma_{ij}\alpha_i\alpha_j\langle r_i r_j\rangle = 1$. They are the solution of the generalized eigenvalue problem

$$\begin{aligned} A\,\vec{u} &= \lambda B\,\vec{u} \\ A_{ij} &= \langle\Delta r_i \Delta r_j\rangle \\ B_{ij} &= \langle r_i r_j\rangle \end{aligned} \qquad (14)$$

The (approximate) p_{fold} reaction coordinate is $R(\vec{X}) = \Sigma_{k=1,N}\alpha_k r_k(\vec{X})$, where $\alpha_i = u_i^2$ is the second eigenvector. The reaction coordinate time series is computed as $x_i = \Sigma_{k=1,N}\alpha_k r_k(\vec{X}_i)$.

It is instructive to compare the approach with two similar dimensionality reduction techniques: the Laplacian eigenmaps[24] and the diffusion maps.[25,26] The three approaches determine the low dimensional embedding (x) by minimizing $\Sigma_{ij}K_{ij}(x_i - x_j)^2$. The Laplacian eigenmaps approach defines K_{ij} as the heat (diffusion) kernel $K_{ij} = \exp(-\|\vec{X}_i - \vec{X}_j\|^2/t)$ if $\|\vec{X}_i - \vec{X}_j\| < \varepsilon$ or 0 otherwise. The diffusion maps approach defines the kernel as $K_{ij} = \exp(-\|\vec{X}_i - \vec{X}_j\|^2/\varepsilon)/(p(\vec{X}_i)p(\vec{X}_j))^{1/2}$ for all pairs of i and j, where $p(\vec{X}_i)$ is the probability density at $\vec{X}_i$. In the present approach K_{ij} equals n_{ij}, the equilibrium number of transitions between points $\vec{X}_i$ and $\vec{X}_j$. Thus, while the present approach uses the actual dynamics to define K_{ij}, the Laplacian eigenmaps and diffusion maps introduce a model of dynamics, namely, diffusion in the configuration space. Using a model of dynamics has the advantage of knowing all n_{ij} without extensive (and computationally expensive) sampling. However, it also means that the constructed reaction coordinate is accurate for the model dynamics rather than the actual one. As a consequence, the result depends on the choice of coordinate set $\vec{X}$ (observables) to define the model dynamics. A different set of coordinates $\vec{Y}$ with a different set of distances $\|\vec{X}_i - \vec{X}_j\| \neq \|\vec{Y}_i - \vec{Y}_j\|$, for example, obtained by rescaling of some coordinates, produces a different embedding (reaction coordinate). The presented approach is manifestly invariant with respect to the selection of any set of coordinates for the description of the process, provided they contain all of the important information about the dynamics. It is formulated in terms of a Markov state model, where each state is defined by index only, without reference to coordinates.

It can be said, loosely, that the Laplacian eigenmaps and diffusion maps perform dimensionality reduction with a focus on preserving the properties (proximity) of a given configuration space. The dynamical information contained in the temporal sequence of the configurations (trajectory) is ignored. The proximity in configuration space, in general, does not necessarily mean that regions are well-connected dynamically; they can be separated by a high free energy barrier. For example, configurations of a relatively large biomolecule that differ only locally, e.g., in a few consecutive dihedral angles or in a side chain orientation, are likely to have close distance in configuration space. However, it is possible that the transition between the configurations occurs only by overcoming a large free energy barrier (e.g., the trans−cis isomerization). At the same time motion along the low energy normal modes (i.e., low barriers) is generally associated with large conformational changes. As another example, consider the construction of optimal reaction coordinate to describe the dynamics of the game of chess.[8] For every chess position there are many nearby positions (small $\|\vec{X}_i - \vec{X}_j\|$) which cannot be reached by a legal move, for example, positions obtained by moving a knight to a nearby square. Moreover, not every legal move is selected by a skillful player. By considering the actual n_{ij}, the presented approach performs dimensionality reduction while preserving the dynamics.

AN ILLUSTRATIVE EXAMPLE: THE P_{FOLD} REACTION COORDINATE FOR FIP35

Recently, Shaw et al. reported a "brute-force" 200 μs equilibrium folding simulation of FIP35 protein in explicit water that contains 15 folding-unfolding events with the folding rate and the native structure in agreement with the experiment.[20] The simulation is the longest-ever realistic simulation of protein folding. The folding dynamics was analyzed by using the p_{fold} reaction coordinate constructed by variational approach of Best and Hummer.[11]

The folding dynamics has been also analyzed by using an alternative optimal reaction coordinate, constructed by optimizing the cut-based free energy profile F_C.[27] The analysis produced results markedly different from those reported by Shaw et al. In particular, FIP35 is not an incipient downhill folder, but it folds

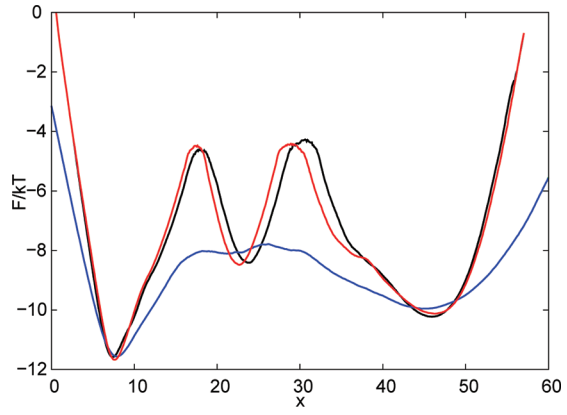


Figure 1. Free energy profiles of FIP35 along different reaction coordinates. The free energy profile for the p_{fold} coordinate constructed by maximizing $\int 1/Z_{C,1}(x)dx$ is shown by a black line. That for the optimal reaction coordinate with the highest F_C, constructed by maximizing $\int Z_H(x)/Z_C(x)^2 dx$ is shown by a red line. That for the p_{fold} coordinate of Shaw et al.,[20] constructed by the approach of Best and Hummer,[11] is shown by a blue line. The coordinates are rescaled, so that the diffusion coefficient is constant; $D(x) = 2.1$ for the coordinate of Shaw et al. and $D(x) = 1$ for the other two coordinates.

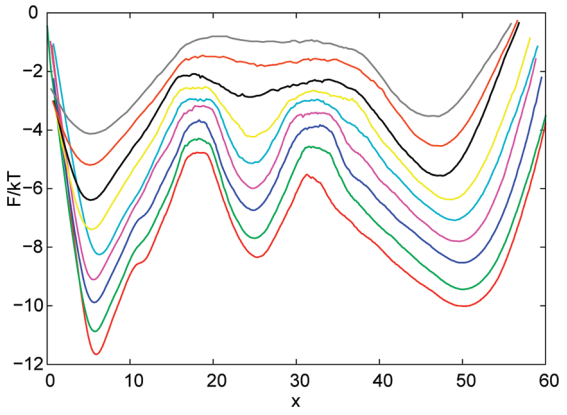


Figure 2. Free energy profiles along coordinates computed during iterative optimization with different Δt = 128, 64, 32, 16, 8, 4, 3, 2, 1. The profiles are rescaled along x and shifted along F/kT for visual clarity. The profile with $\Delta t = 1$ is the lowest (red), while that with $\Delta t = 128$ is the highest (gray).

via a populated on-pathway intermediate separated by high free energy barriers; the high free energy barriers, rather than landscape roughness, are a major determinant of the rates for conformational transitions. The preexponential factor of folding kinetics is $k_0^{-1} \sim 10$ ns rather than 1 μs.

In principle, the difference in the analysis can be attributed to the different definitions of optimal reaction coordinate, that is, p_{fold} versus the one with the highest F_C. To investigate the possibility, the p_{fold} reaction coordinate is constructed by maximizing $\int 1/Z_{C,1}(x)dx$. A detailed description of the optimization procedure is given below. The free energy profile along such constructed p_{fold} reaction coordinates is very similar to that along the coordinate constructed by optimizing F_C (Figure 1). The two coordinates provide a very similar description of the folding dynamics, which is very different from that given by coordinates of Shaw et al. It confirms that the difference in the results is due to the suboptimal reaction coordinate employed by Shaw et al. rather than the different definitions of the optimal reaction coordinate. A detailed analysis of the folding dynamics and comparison with the analysis of Shaw et al. is presented elsewhere.[27]

The approach of Best and Hummer[11] constructs the p_{fold} reaction coordinate by optimizing it just around $p_{fold}\sim 0.5$. Regions before or after $p_{fold}\sim 0.5$ are not optimized.[28] $p_{fold}\sim 0.5$ is a reasonable definition of the transition state for a two-state system (see, however ref 29). In FIP35, the $p_{fold}\sim 0.5$ defines the intermediate, rather than the transition state, due to the presence of two similar free energy barriers (Figure 1). The transition states have $p_{fold}\sim 0.75$ and $p_{fold}\sim 0.2$[27] and hence are not optimized by the approach of Best and Hummer. The presented approach optimizes the whole p_{fold} coordinate.

The p_{fold} reaction coordinate constructed by the approach, as well as the coordinate constructed by optimizing F_C,[27] are optimal only around the transition states, the most important parts of the free energy profile to describe the kinetics of folding.[27] This is due to the very simplistic form of the reaction coordinate $R(\vec{X},\vec{\alpha})$, namely, the (smoothed) number of contacts. More sophisticated functional forms, for example, neural networks,[30] are likely to be useful to construct the reaction coordinate which accurately describes larger regions of the free energy profile.

Optimization Procedure. The optimization procedure is identical to the one used in ref 27. The optimal reaction coordinate $R(x)$ was constructed by numerically optimizing the parameters $\{\alpha_i\}$ of reaction coordinate functional form $R(x,\alpha)$ to make the cut free energy profile $F_{C,1}(R)$ the highest. The whole coordinate was optimized by maximizing $I_{AB} = \int_A^B 1/Z_{C,1}(x)dx$ where A and B are the positions of local minima in states A and B. Initially, the coordinate was optimized with functional $I_{ND} = \int_N^D 1/Z_{C,1}(x)dx$, where N and D are the positions of local minima of the native and denatured states, respectively (Figure 2). It was found that the free energy profile has two transition states. However, one of the transition states (e.g., that around $x \approx 35$ in Figure 2) is suboptimal, as indicated by subdiffusion exponent $\alpha < 0.5$.[5,27] To obtain uniform optimization of both the transition states (to refine the profile) it was useful to compute the functional for each transition state and optimize their product $I_{NI}I_{ID}$, where N, I, and D are the native, intermediate, and denatured states, respectively. Note that a single reaction coordinate is used to describe both of the transition states. Such a constructed reaction coordinate is used to compute the free energy profile shown in Figure 1.

The putative reaction coordinate is taken as (smoothed) number of contacts $R(x,\alpha) = \Sigma_{ij}\alpha_{ij}h(\alpha_{ij}^r - r_{ij})$, where α_{ij} is either 1 or -1 and α_{ij}^r is a threshold when a contact is considered to be formed, r_{ij} is the distance between atoms i and j; $h(x) = \min(1,x)$ if $x > 0$ and zero otherwise. Distances between backbone HN and O atoms were considered.

The numerical optimization was performed by randomly modifying the parameters α_{ij}, recomputing the reaction coordinate, the profiles, and the functional and accepting the new parameter value if $\Delta I/I < 0.01$, where ΔI and I are the change and the current value of the optimization functional. The profiles and the integral were computed by discretizing the coordinate with $\Delta x = (rc_{max} - rc_{min})/10\,000$, where rc_{max} and rc_{min} are the current maximum and minimum of the reaction coordinate time series. Such a small Δx was chosen to make any potential approximation errors negligible. Increasing Δx does not speed

up the algorithm because the time-limiting step is the computation of the reaction coordinate time series, which is proportional to trajectory length (10^6 steps here). For efficient computation of the cut profiles from a time series, one computes $dZ_{C,1}(x)/dx$ rather than $Z_{C,1}(x)$ directly. Transition from x_i to x_j increments the array corresponding to $dZ_{C,1}(x)/dx$ by $|x_i - x_j|/2$ at point $\min(x_i,x_j)$ and decrements by the same amount at point $\max(x_i,x_j)$. To compute $Z_{C,1}(x)$, one needs to modify all of the points in between.

When sampling is limited and a reaction coordinate has many parameters, it is possible to overfit the data by constructing a reaction coordinate with a free energy profile higher than the correct one. In this case the dynamics projected on the coordinate is superdiffusive,[5,27] which can be used as an indicator of overfitting. A penalty term is introduced in the optimization to avoid overfitting: $\int p(x)Z_H(x)/Z_C(x)dx$, where $p(x) = k^{10}(x)$ if $k(x) > 1$ and zero otherwise, and $k(x) = Z_C(x,\Delta t_2)/Z_C(x,\Delta t_1)(\Delta t_2/\Delta t_1)^{1/2}$, with $\Delta t_1 = 1$ and $\Delta t_2 = 8$ trajectory steps of 0.2 ns.

The reaction coordinate optimization problem was complicated by having many competing local minima, where the numerical optimization can stuck. The following procedure was found to be useful to find the putative global minimum. The large time scale description of the dynamics should be simpler since many metastable states are absent. The reaction coordinate optimization problem at the large time scale (large sampling interval Δ) should be simpler as well. The optimization was performed in an iterative manner starting with large sampling interval $\Delta t = 128$ and halving it at the next iteration until $\Delta t = 1$. Optimization was initialized with a seed reaction coordinate (the rmsd from the native structure here), which contribution was slowly decreased to zero.

Figure 2 illustrates the performance of the iterative algorithm at the initial stage of the reaction coordinate construction, when no information about the intermediate state is available and the optimization functional is $I_{ND} = \int_N^D 1/Z_{C,1}(x)dx$. At $\Delta t = 128$ the free energy profile has just one broad transition state. At higher time resolution, $\Delta t \leq 16$, the intermediate state appears. At $\Delta t = 1$ the profile shows the intermediate and the two transition states. The transition state around $x \approx 35$ is about 1 kT lower than that in Figure 1. Thus even without the information about the intermediate, the approach has been able to find the intermediate and the two-transition states. Comparing this profile with that obtained with the method of Best and Hummer[11] is "more fair", in the sense that the latter had not used the information about the intermediate. Figure 2 establishes the general tendency and the robustness of the approach with respect to the choice of the sampling interval (Δt).

To illustrate that the approach works with very few actual transitions, the extreme case is considered, when the reaction coordinate is constructed for a segment of the trajectory that contains single (first) folding event. The segment consists of 90 000 snapshots with total length of 18 μs. The coordinate was optimized by the iterative procedure using $I_{ND} = \int_N^D 1/Z_{C,1}(x)dx$ functional; that is, no information about the intermediate was used. The FEP along the constructed coordinate (Figure 3) is in good agreement with that for the whole trajectory, as in Figure 1. The barriers for the first transition state ($x \approx 35$) from the denatured ($x \approx 45$) and intermediate states ($x \approx 25$) are around 6 kT and 4 kT, respectively, in agreement with that in Figure 1. The agreement is explained by the fact that, while the segment of the trajectory contains a single total folding event, it contains five transitions between the denatured and the intermediate states, which are enough for a rough estimation of the free energies. The barriers for the second transition state ($x \approx 15$) from the intermediate ($x \approx 25$) and native states ($x \approx 5$) as well as the free energy of the native state are different from that in Figure 1. The discrepancies are due to the very limited sampling; evidently a single transition is not enough to accurately estimate the equilibrium free energies. The width of the both transition states is in reasonable agreement with that in Figure 1, meaning that the diffusion coefficients for both profiles agree. It is because the trajectory makes a sufficient number of forward−backward moves along the coordinate (even during single transition) that a reasonable estimation of the diffusion coefficient is possible.

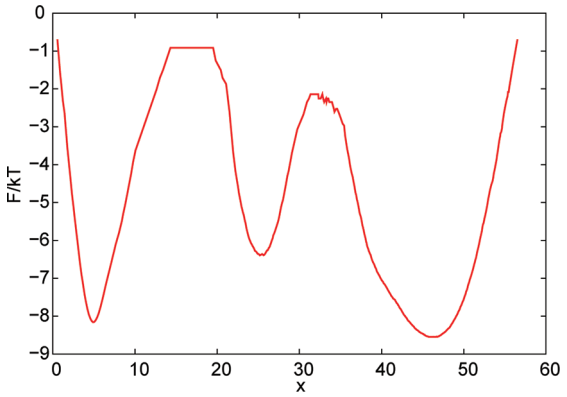


Figure 3. Free energy profile along the optimal coordinate constructed for a segment of the trajectory that contains a single folding event.

CONCLUSION

The approach to construct the p_{fold} reaction coordinate for a Markov process has been presented. The coordinate is constructed by optimizing the generalized cut free energy profiles $F_{C,1}$ computed from the process trajectories. Compared to other approaches, the approach makes no assumptions regarding the shape of the free energy profile,[19] the shape of the folding pathway,[18] or the position of the transition state.[11] In comparison to the popular dimensionality reduction techniques, namely, the Laplacian eigenmaps[24] and the diffusion maps,[25,26] the approach performs dimensionality reduction based on the actual dynamics, rather than the model dynamics. As a consequence the proposed approach is invariant with respect to the choice of the observables. Provided that the process is Markovian in some configuration space and that the available coordinates contain all of the important information about the process, the approach can be used to analyze dynamics when the full configuration space is not accessible (e.g., dynamics of a disease) and "the equations of motion" are not known (e.g., the game of chess).[8] The approach has been illustrated by constructing the p_{fold} reaction coordinate for analyzing the folding dynamics of FIP35 protein. It is believed that the approach shall be useful for the analysis of protein folding dynamics and many other complex dynamical phenomena.[8]

AUTHOR INFORMATION

Corresponding Author

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

■ ACKNOWLEDGMENT

I am grateful to David Shaw and his coworkers for making the folding trajectories available. The work was supported by an RCUK fellowship.

■ REFERENCES

(1) Onuchic, J. N.; Socci, N. D.; Luthey-Schulten, Z.; Wolynes, P. G. *Folding Design* **1996**, *1*, 441–50.
(2) Dobson, C. M.; Sali, A.; Karplus, M. *Angew. Chem., Int. Ed.* **1998**, *37*, 868–893.
(3) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.
(4) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751–758.
(5) Krivov, S. V. *PLoS Comput. Biol.* **2010**, *6*, e1000921.
(6) Krivov, S.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689–12698.
(7) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13841–6.
(8) Krivov, S. V. *Phys. Rev. E.* **2011**, *84*, 011135.
(9) Onsager, L. *Phys. Rev.* **1938**, *54*, 554–557.
(10) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334–350.
(11) Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6732–7.
(12) E, W.; Vanden-Eijnden, E. In *Multiscale Modelling and Simulation*, 1st ed.; Attinger, S., Koumoutsakos, P., Eds.; Springer: New York, 2004; p 277.
(13) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877–5882.
(14) Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. *J. Chem. Phys.* **2008**, *129*, 174102.
(15) Berezhkovskii, A.; Szabo, A. *J. Chem. Phys.* **2005**, *122*, 14503.
(16) Wales, D. J. *J. Chem. Phys.* **2009**, *130*, 204111–7.
(17) E, W.; Ren, W.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2005**, *413*, 242–247.
(18) Ovchinnikov, V.; Karplus, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134*, 085103.
(19) Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.
(20) 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.
(21) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Chem. Phys.* **2009**, *130*, 205102.
(22) Berezhkovskii, A.; Szabo, A. *J. Chem. Phys.* **2004**, *121*, 9186–9187.
(23) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.
(24) Belkin, M.; Niyogi, P. *Neural Comput.* **2003**, *15*, 1373–1396.
(25) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7426–7431.
(26) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 13597–13602.
(27) Krivov, S. V. Submitted for publication, **2011**.
(28) Peters, B. *Chem. Phys. Lett.* **2010**, *494*, 100–103.
(29) Muff, S.; Caflisch, A. *J. Chem. Phys.* **2009**, *130*, 125104.
(30) Ma, A.; Dinner, A. R. *J. Phys. Chem.* **2005**, *109*, 6769–79.