HHS Public Access

Author manuscript

J Phys Chem Lett. Author manuscript; available in PMC 2024 September 16.

Published in final edited form as:

J Phys Chem Lett. 2022 October 13; 13(40): 9263–9271. doi:10.1021/acs.jpclett.2c02529.

Committor-Consistent Variational String Method

Ziwei He[†], **Christophe Chipot**[‡], **Benoît Roux**[¶]

[†]Department of Chemistry, The University of Chicago, 5735 S Ellis Ave, Chicago, IL 60637, Chicago, IL 60637, USA

[‡]Laboratoire International Associé Centre National de la Recherche Scientifique et University of Illinois at Urbana-Champaign, Unité Mixte de Recherche n°7019, Université de Lorraine, B.P. 70239, 54506 Vandæuvre-lès-Nancy cedex, France

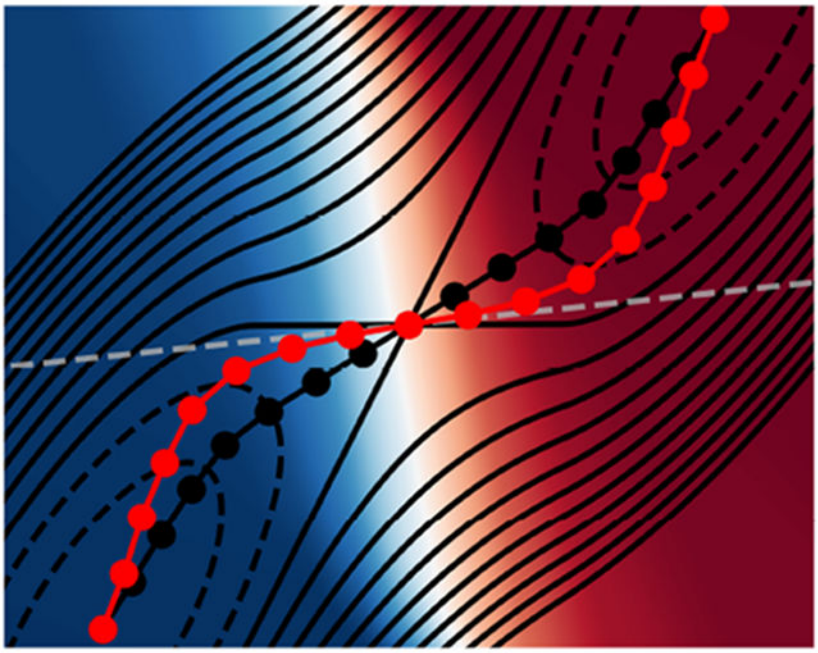
[¶]Department of Biochemistry and Molecular Biology, The University of Chicago, 929 E. 57th Street W225, Chicago, Illinois 60637, USA

Abstract

The treatment of slow and rare transitions in the simulation of complex biomolecular systems poses a great computational challenge. A powerful approach to tackle this challenge is the string method, which represents the transition path as a one-dimensional curve in a multi-dimensional space of collective variables. Commonly used strategies to optimize the pathway have been to align the tangent of the string to the local mean force, yielding the minimum free energy path, or to the mean drift determined from swarms of short trajectories. Here, a novel strategy is proposed, allowing the string to be optimized on the basis of a variational principle involving the unidirectional reactive flux expressed in terms of the time-correlation function of the committor. The method is illustrated with model systems and probed with the alanine dipeptide as well as a coarse-grained model of the barstar-barnase protein complex. It is shown that successive iterations variationally refine the trial string toward an optimal transition pathway following the gradient of the committor between two metastable states.

Graphical Abstract

roux@uchicago.edu .



A wide range of computational methods and strategies have been developed to treat slow and rare transitions in complex molecular systems[1-13]. Many key ideas and concepts can be established by considering a prototypical model system with two metastable states A and B in the context of transition path theory (TPT)[14]. In particular, one critically important quantity is the committor probability that a trajectory initiated at some configuration will ultimately reach state B before ever reaching state A[15,16]. The committor probability is often thought to represent an optimal one-dimensional (1D) reaction coordinate (RC) representing the progress of the slow A-B transitions[17-19]. A number of approaches have been proposed to determine the committor probability by shooting trajectories to identify RC in complex systems[6,8,13,20-25]. Alternative strategies to identify the slowest degrees of freedom and potentially discover relevant slow coordinates from a set of user-defined collective variables (CVs) $\mathbf{z}$ also include the time-lagged independent component analysis (TICA)[26-28], and the spectral gap optimization of order parameters (SGOOP)[27,28]. More recently, machine learning (ML) methodologies appear to provide a promising route to enable the discovery of complex RCs that are nonlinear functions of the CVs[29-35].

This progress notwithstanding, the lack of sampling can undermine the ability of ML methodologies to correctly learn arbitrarily complex nonlinear RCs from high-dimensional MD data[36,37]. To may help circumvent the issue, we sought a way to reduce the dimensionality of the problem while preserving its nonlinear character. To this end, a powerful framework inspired by the string method is to represent the RC as a 1D curve, or chain-of-states, embedded in the high-dimensional space of the CVs[7,12,38]. According to TPT, the “reaction tube” embracing the string is expected to support most of the reactive probability current governing the long time relaxation from A to B, and this current is largely determined by variations of the committor[14]. However, present algorithms to optimize a string between two metastable states either seek to find the minimum free energy path

(MFEP)[7], or a path following the local probability current determined from local mean drifts calculated from swarms of trajectories[12,39]. While useful in their own rights, these prescriptions to construct the string do not directly provide information about the committor probability, and hence, the reactive probability current. Our goal here is to address this issue by introducing a novel algorithm to variationally determine an optimal "committor-consistent" curvilinear string transition pathway in the space of the CVs.

Theoretical Development

Effective dynamical propagator

To reduce the complexity of the problem, one typically seeks to determine the optimal pathway between two metastable states within the subspace of reduced dimension spanned by a subset of collective variables (CVs), $\tilde{\mathbf{z}}(\mathbf{x}) = (\tilde{z}_1(\mathbf{x}), \ldots, \tilde{z}_N(\mathbf{x}))$, i.e., a vector-valued function that maps every atomic configuration $\mathbf{x}$ of the system on a set of values $\tilde{\mathbf{z}}(\mathbf{x})$[39]. In this context, the string represents the pathway linking states A and B as a discrete "chain of states", i.e., a collection of M images located at positions $\{\mathbf{z}^1, \ldots, \mathbf{z}^M\}$ in the subspace of the CVs[7,12]. Knowledge of the long-time dynamics of the system within the subspace of the CVs is critical to characterize the string pathway[39]. The probability density of the system at time t is expressed as $\rho(\mathbf{z}; t)$. The forward propagation step $(\mathbf{z} \rightarrow \mathbf{z}')$ for the probability density from the time t to the time $t + \tau$ is,

$$\rho(\mathbf{z}'; t + \tau) = \int d\mathbf{z}\, \mathscr{P}_\tau(\mathbf{z}' \mid \mathbf{z})\, \rho(\mathbf{z}; t), \tag{1}$$

where $\mathscr{P}_\tau(\mathbf{z}' \mid \mathbf{z})$ is the propagator (also called transfer operator). The dynamics within the reduced subspace of the CVs is assumed to be Markovian with a finite lag-time τ, and that $\mathscr{P}_\tau$ obeys the Chapman-Kolmogorov equation, $\boldsymbol{\rho}(t + n\tau) = \mathscr{P}_{n\tau} \cdot \boldsymbol{\rho}(t)$, with $\mathscr{P}_{n\tau} = (\mathscr{P}_\tau)^n$. The system is also assumed to be in thermodynamic equilibrium and that detailed balance is satisfied, $\mathscr{P}_\tau(\mathbf{z}' \mid \mathbf{z})\, \rho_{\mathrm{eq}}(\mathbf{z}) = \mathscr{P}_\tau(\mathbf{z} \mid \mathbf{z}')\, \rho_{\mathrm{eq}}(\mathbf{z}')$. Under these conditions, the effective transfer propagator $\mathscr{P}_\tau(\mathbf{z}' \mid \mathbf{z})$ yields a self-consistent representation of the dynamics of the system within this subspace (closure of the dynamical propagation).

Committor probabilities for two metastable states

Assuming two metastable states A and B, the forward committor $q(\mathbf{z})$ is the sum of the probability over all paths starting at $\mathbf{z}$ that ultimately reach the state B before ever reaching the state A (see Supporting Information). The probability of each of these paths is expressed as a product of discrete propagation steps $\mathscr{P}_{n\tau} \cdots \mathscr{P}_{n\tau}$ with lag-time τ, under the restriction that the intermediate states resulting from all these steps are neither $\in A$ or B. Summing over all possible paths, it follows that $q(\mathbf{z})$ can be written as,

$$q(\mathbf{z}) = \int d\mathbf{z}'\, q(\mathbf{z}')\, \mathscr{P}_\tau(\mathbf{z}' \mid \mathbf{z}), \tag{2}$$

with the constraints $q(\mathbf{z}) = 0$ if $\mathbf{z} \in A$, and $q(\mathbf{z}) = 1$ if $\mathbf{z} \in B$, see Eq. (S2) for details. By construction, $0 \leq q(\mathbf{z}) \leq 1$. While the equations for the committor probabilities involve only the propagator $\mathscr{P}_\tau(\mathbf{z}' \mid \mathbf{z})$ for the lag-time τ, the fundamental validity of these equations is predicated upon the necessity to satisfy Markovianity of the dynamics as expressed by the Chapman-Kolmogorov equation, $\mathscr{P}_{n\tau} \equiv (\mathscr{P}_\tau)^n$.

Net forward flux from reactive pathways

In the context of TPT, one can express the net forward reactive flux from A to B as[14,23,39-42]

$$J_{AB} = \frac{1}{2\tau}\int d\mathbf{z} \int d\mathbf{z}' \, (q(\mathbf{z}') - q(\mathbf{z}))^2 \mathscr{P}_\tau(\mathbf{z}' \mid \mathbf{z}) \, \rho_{\text{eq}}(\mathbf{z}) \, . \tag{3}$$

Equivalently, J_{AB} can also be expressed as a time-correlation function,

$$\begin{aligned} J_{AB} &= \frac{1}{2\tau}\langle (q(\tau) - q(0))^2 \rangle \\ &= \frac{1}{\tau}(\langle q(0)q(0)\rangle - \langle q(\tau)q(0)\rangle) \, . \end{aligned} \tag{4}$$

In the following sections, we refer to $\langle q(0)q(0)\rangle - \langle q(\tau)q(0)\rangle = C(\tau)$ as the the committor time-correlation function. While $q(\mathbf{z})$ is a probability, the committor time-correlation function in Eq. (4) can be understood by recalling that the CVs are vector-valued functions that maps every microscopic configuration $\mathbf{x}(t)$ of the system onto a set of values $\tilde{\mathbf{z}}[\mathbf{x}(t)]$ along a dynamical trajectory. Assuming that the function $q(\mathbf{z})$ is known, we write $q(\tilde{\mathbf{z}}[\mathbf{x}(t)])$ as $q(t)$ for the sake of simplicity. See Supporting Information for additional details.

Basis set expansion of the committor

The quadratic expression for the reactive flux J_{AB} from Eqs. (3), or equivalently Eq. (4), can serve as a robust variational principle to optimize a trial committor $q(\mathbf{z}')$. Minimizing the quantity J_{AB} with respect to a trial function $q(\mathbf{z})$, $\delta J_{AB} / \delta q = 0$ recovers Eq. (2) that formally defines the committor probability Minimization of the steady-state flux J_{AB} defined by Eq. (4) for a trial function q with the constraints $q(\mathbf{z}) = 0$ if $\mathbf{z} \in A$, and $q(\mathbf{z}) = 1$ if $\mathbf{z} \in B$ yields the correct committor $q(\mathbf{z})$, as defined by Eq. (2). See Eq. (S13) in Supporting Information for additional details.

We seek to express the trial committor $q(\mathbf{z})$ in terms of a basis set expansion. However, the construction of this trial function requires special care to handle the constraints imposed by the boundary states[43,44]. For our set of basis functions $f_i(\mathbf{z})$, we choose Voronoi cells supported by a set of M centroids corresponding to the images of a string connecting boundary states A and B. We write the committor as

$$q(\tilde{\mathbf{z}}(\mathbf{x})) = h_A(\tilde{\mathbf{z}}(\mathbf{x}))\, q_A + h_I(\tilde{\mathbf{z}}(\mathbf{x}))\left(\sum_{i=1}^{M} b_i\, f_i(\tilde{\mathbf{z}}(\mathbf{x}))\right) + h_B(\tilde{\mathbf{z}}(\mathbf{x}))\, q_B, \tag{5}$$

where $q_A = 0$ and $q_B = 1$, and h_A, h_B, and h_I are indicator functions correspondingly equal to 1 when the system is in the A, B, or the intermediate region, respectively, and zero otherwise. By construction, $h_A + h_I + h_B = 1$, and all cross products of indicator functions are identically zero ($h_A h_I = h_A h_B = h_I h_B = 0$) because there is no overlap between the three different regions. Using the trial function in Eq. (5) we can express the committor time-correlation function as,

$$\langle q(\tau)\, q(0)\rangle = \left\langle \left(h_I(\tau)\left(\sum_{i=1}^{M} b_i\, f_i(\tau)\right) + h_B(\tau)\right) \left(h_I(0)\left(\sum_{j=1}^{M} b_j\, f_j(0)\right) + h_B(0)\right)\right\rangle. \tag{6}$$

The committor time-correlation function can be expanded as,

$$C(\tau) = \frac{1}{2}\mathbf{b}^t(\mathbf{D}(0) - \mathbf{D}(\tau))\mathbf{b} + (\mathbf{g}(0) - \mathbf{g}(\tau))\cdot\mathbf{b} + \langle h_B(0)\, h_B(0)\rangle - \langle h_B(0)\, h_B(\tau)\rangle, \tag{7}$$

where

$$D_{ij}(\tau) = \langle h_I(0)\, h_I(\tau)\, f_i(0)\, f_j(\tau)\rangle + \langle h_I(0)\, h_I(\tau)\, f_j(0)\, f_i(\tau)\rangle, \tag{8}$$

and

$$g_i(\tau) = \langle h_I(\tau)\, h_B(0)\, f_i(\tau)\rangle + \langle h_I(0)\, h_B(\tau)\, f_i(0)\rangle. \tag{9}$$

Then, taking the derivative with respect to the basis set coefficients yields the linear system of equations,

$$(\mathbf{D}(0) - \mathbf{D}(\tau))\mathbf{b} + (\mathbf{g}(0) - \mathbf{g}(\tau)) = 0, \tag{10}$$

with the simple solution,

$$\mathbf{b} = -(\mathbf{D}(0) - \mathbf{D}(\tau))^{-1}(\mathbf{g}(0) - \mathbf{g}(\tau)). \tag{11}$$

Substituting the M basis set coefficients $\{b_i\}$ determined from Eq. (11) into Eq. (7) yields the optimized committor time-correlation function $C(\tau)$ for a given basis set. Because of the indicator functions $h_I(0)$ and $h_I(\tau)$ in **D**, the basis functions must be chosen such that they either fully or at least partially reside in the intermediate region to avoid singularities with the matrix inversion. Eq. (11) is a central result of this analysis and enables the optimization of a committor-consistent string pathway.

Global optimization of the committor-consistent string

Using the variational principle with the committor time-correlation function, $C(\tau)$ it is possible to optimize the position of the M images of a string in a committor-consistent manner. Adopting a Voronoi basis set expansion, $C(\tau)$ depends on the M basis set coefficients $\{b_1, \ldots, b_M\}$, and depends also–in a non-linear fashion–on the position of the M centroids $\{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$. Thus, we have the correlation function $C(\tau; \{b_1, \ldots, b_M; \mathbf{z}_1, \ldots, \mathbf{z}_M\})$. To optimize the string according to the variational principle, we must minimize the committor time-correlation function with respect to the position of the M images. To this end, we have adopted an iterative Monte Carlo procedure, whereby we first determine the basis set coefficient via Eq. (11), and then introduce random changes in the position of the images that are accepted or rejected on the basis of $C(\tau; \{b_1, \ldots, b_M; \mathbf{z}_1, \ldots, \mathbf{z}_M\})$. To achieve a complete optimization of the pathway and obtain a committor-consistent string, these two operations must be repeated iteratively until convergence, i.e., until one cannot further minimize $C(\tau; \{b_1, \ldots, b_M; \mathbf{z}_1, \ldots, \mathbf{z}_M\})$. The optimization process can be initiated with a string constructed either from the mean force[7], or the mean drift calculated from swarms of trajectories[12]. Further analysis indicates that the resulting pathway is a one-dimensional (1D) line going from A to B that follows the committor gradient $\nabla q(\mathbf{z})$ in the space of the CVs **z**.

Equilibrium average and enhanced sampling

The committor time-correlation function $C(\tau)$ should be averaged over equilibrium initial conditions. In practice, it is likely that enhanced sampling techniques would be necessary to have an accurate result in the presence of large free energy barriers. An effective strategy to improve the conformational sampling relevant to a slow A-B transition is to introduce a biasing potential along a progress path collective variable (PCV) $\tilde{s}[\mathbf{z}]$ function, expressed as a differentiable function of the CVs **z**[38]. Once this enhanced sampling along $\tilde{s}[\mathbf{z}]$ is achieved, we can generate unbiased trajectories initiated from specific regions at $t = 0$ according to the value of $\tilde{s}[\mathbf{z}]$. The unbiased committor time-correlation function $C(\tau)$ can then be expressed as,

$$C(\tau) = \int_0^1 ds\, \rho_{\mathrm{eq}}(s) \langle q(\tau)\, q(0) \rangle_{(s)},$$

(12)

where $\langle q(\tau)q(0)\rangle_{(s)}$ is the time-correlation function calculated from an unbiased trajectory initiated with $\tilde{s}[\mathbf{z}(t)] = s$ at $t = 0$, and $\rho_{eq}(s)$ is the equilibrium probability starting of starting at s. See Eqs. (S24) and (S25) for details.

RESULTS AND DISCUSSION

Illustration with one-dimensional double-well potential

We first illustrate the variational framework in the case of a simple one-dimensional double well. A simple set of discrete one-hot indicator functions was used. The widths of the one-hot indicators were automatically assumed by even spacing across the intermediate region. The result in Figure 1 shows that the exact committor (black dashed line) is essentially reproduced by solving Eq. (11) for the basis functions. An example with Gaussian functions is given in Supp Info (Figure S1) to illustrate how basis set and hyperparameter choices affect the committor probability.

Illustration with two-dimensional potential

We now illustrate the usefulness of the proposed framework in finding a string that follows the dominant reaction pathway in a multi-dimensional space. As a prototypical two-state system, we used the two-dimensional (2D) potential shown in Figure 2 that was previously studied by Berezhkovskii and Szabo to examine the effect of anisotropic diffusion on the reaction rate[17]. This 2D potential, which we will refer to as the 2D-BS potential, was previously used by Tiwary and Berne to illustrate the SGOOP method[45] and by Roux to illustrate the committor variational principle[39]. Here, the string is assumed to be a straight line with one-hot indicator functions evenly spacing across the intermediate region. The amplitude of the basis set coefficients is determined by solving Eq. (11).

To demonstrate the effect of anisotropic diffusion, we considered three conditions: $D_y / D_x = \delta = 0.1$, 1, and 10. In each case, a different optimal reaction direction about the saddle region was discovered (depicted by the dashed lines in Figure 2, top left). For the analysis and computation of the committors, a lag time τ of 1 time step was selected. The boundary conditions for where $q = 0$ or 1 were defined by ellipses at the well minima such that the trajectory spends approximately 50% of the time within the ellipses and the other 50% of the time outside of the region.

Considering first a straight pathway, the committor time-correlation function is computed by Eq. (7) as a function of the angle of the path with respect to the x-axis (Figure 2 A). The minimum correlation function value is taken to be the best angle for the reaction pathway through the saddle point (Figure 2 C). From a Voronoi tessellation of the 2D space built from the images of the straight pathway, the estimated committor for $\delta = 1$ is determined (Figure 2 B). Considering now a curvilinear pathway, we determined the optimized string for the three diffusion cases (Figure 3). The positions of the images of the curvilinear string were optimized by variationally minimizing the committor correlation function using Monte Carlo sampling, solving Eq. (11) and evaluating Eq. (7) at every iteration. The

curvilinear strings cross the transition state region differently depending on the parameter δ. The minimum free energy path (MFEP), used as initial string, is shown in black. This path is valid only in the case of isotropic diffusion with $\delta = 1$. When $\delta \neq 1$, the committor-consistent optimized string departs from the MFEP.

The string method aims to determine a curvilinear RC linking state A and B. However, the criterion to determine the most relevant curve is not immediately clear. Should the best choice of RC be the 1D coordinate that yields the correct transition rate? Or should it be the 1D coordinate that captures the most representative events at the kinetic bottleneck for a productive transition? In fact, identifying the most relevant and useful RC is not straightforward, even with knowledge of the committor $q(\mathbf{z})$ and the reactive forward flux, J_{AB} [17], or extensive information about the transition path ensemble[6,8,13,14,20-25,46]. The 2D-BS potential provides a great opportunity to illustrate these fundamental issues. Depending on the assumptions about the local tangent of the string at $\mathbf{z}$, different curvilinear pathways can be constructed to link states A and B. For example, the tangent to the string may follow the mean force (the gradient of the PMF), $-\nabla W(\mathbf{z})$. This prescription generates the MFEP[7]. Alternatively, the tangent to the string could follow the gradient of the committor, $\nabla q(\mathbf{z})$, the reactive flux density $\mathbf{J}_{AB}(\mathbf{z}) = \mathbf{D}\nabla q(\mathbf{z})$, or the mean drift from swarms of trajectories $\langle \Delta z(\tau) \rangle$[12].

To examine the different pathways across the saddle point, we express the potential near the top of the barrier in terms of the Hessian matrix, $\mathbf{V}$ (see Supplementary Information for more details). Fig. 4 illustrates the principal directions computed for these different cases. The two principal axes of the 2D potential defining the tangent of the MFEP are determined from the right-eigenvector of the matrix $\mathbf{V}$. The gradient of the committor and the reactive flux $\mathbf{J}_{AB}$ follows the right-eigenvector of the matrices $\mathbf{DV}$ and $\mathbf{VD}$, respectively[47]. The tangent of committor-consistent optimized string points in the direction of $\nabla q(\mathbf{z})$. This is a natural outcome of constructing Voronoi cells basis functions from the images along the string used as centroids because it yields isocommitor surfaces that are essentially orthogonal to the local tangent. Interestingly, the mean drift from the swarms of trajectories $\langle \Delta z(\tau) \rangle$ (blue arrow) and the reactive flux $\mathbf{J}_{AB} = \mathbf{D}\nabla q(\mathbf{z})$ (purple arrow) both point in the direction of the right-eigenvector of the matrix $\mathbf{DV}$ at the saddle point.

While these different pathways provide different information about the mechanism of the reaction, they do not yield an effective 1D RC. For instance, in the case of a multidimensional activated process controlled by diffusion, using a 1D coordinate parallel to the gradient of the committor at the saddle point yields a result that is consistent with the multidimensional Kramers-Langer rate theory[17,48]. Projection onto any other 1D coordinate leads to an incorrect result. Furthermore, it was shown recently that $\nabla q(\mathbf{z}^{\dagger})$ indeed represents the best choice to determine the direction of an effective 1D reaction coordinate[19]. An alternative strategy to determine the committor is to monitor the outcome of multiple shooting trajectories initiated from different positions near the transition state[8,13,21,22]. See Supplementary Information for more details.

Alanine dipeptide and enhanced sampling

The simulation used for the study of the 2D-BS potential was sufficiently long to provide an equilibrium sampling. However, this may not always be feasible due to the long-lived timescales of many systems of interest and compounded by the large computational overhead in order to simulate the rare events. Here we apply the committor-consistent string method to the N-acetyl-N-methyl-L-alanylamide molecule, commonly known as the alanine dipeptide (or dialanine), to study the conformational transition between C_{7eq} and C_{7ax} in vacuum at room temperature.

To initiate the pathway optimization, enhanced sampling was performed with an adaptive biasing force acting along an initial string, used to define the progress collective variable (PCV) s[38]. Two different strings were considered to define the PCV in the $\phi - \psi$ plane: the MFEP from a 5 μs unbiased trajectory and the mean drifts from swarms of trajectories[12]. Nevertheless, these two strings are fairly similar qualitatively, and also yield similar values of ΔG (2.3 kcal/mol for the black string and 1.9 kcal/mol for the red string shown in Supp. Fig. S5). Then, defining the variable s between the strings from the above two methods, a reaction tube of starting points was chosen from the biased trajectories to run additional unbiased simulations for the computation of q (Supp. Fig. S6). The optimized string after 20 iterations is shown in Fig. 5.

Because of the isotropic nature of the ϕ and ψ dihedral angles, the committor-consistent string remained close to the initial MFEP string. The committor $q(\mathbf{z})$ calculated along the optimized string from Eqs. (7) and (11) is very similar to that calculated from the PMF along the PCV. In particular, the halfway crossing at $q(\mathbf{z}) = 0.5$ agrees relatively well, although with a somewhat steeper slope. Compared to the committor calculated from the PCV by the string method with swarms of trajectories, the top of the free energy barrier is slightly shifted, which is expected since the initial strings were different.

Coarse-grained model of barstar-barnase binding

The committor-consistent string method was used to examine the kinetics of protein–protein association and dissociation in the context of the barstar-barnase complex. The simulations are based on a coarse-grained (CG) model of the complex developed previously[49]. Briefly, the CG representation used in a previous study maps each amino acid residue as a single bead with its mass and position corresponding to the Cα carbon atom. Attractive Lennard-Jones 6-12 potentials are used to represent four pairwise contacts of the native complex to simulate the protein-protein association. Using a long Langevin dynamics simulation generated previously[49], the string was optimized within a subspace of two order parameters, namely, the center-of-mass (COM) distance and the root-mean-square deviation (RMSD) of the contact residues. The optimization was initiated by starting from a straight path as shown in Fig. 6. Monte Carlo with random moves over the two order parameters (0.05 × 0.05) was carried out for a total of 10 iterations until convergence. Supp. Table S1 gives the rates of association and dissociation computed for different lag times τ. The rate constants k_{on} and k_{off} estimated from the steady-state reactive flux with the committor-consistent string at a lag time τ of 15 ns are $2.41 \times 10^{13} \text{Å}^3\ \text{s}^{-1}$ and $3.47 \times 10^{7}\ \text{s}^{-1}$, respectively, which are close to

the rates of 2.31×10^{13} Å^3 s^{-1} and 2.70×10^7 s^{-1} from the MSM and Perron-cluster cluster analysis (PCCA) analysis with the six order parameters at the optimal MSM lag time of 12 ns[49]. The k_{on} and k_{off} values here are also close the MSM rates of 2.36×10^{13} Å^3 s^{-1} and 2.74×10^7 s^{-1}, respectively, with two order parameters at an optimal lag time of 12 ns.

CONCLUSION

An extension of the string method that produces an optimal reaction pathway following the gradient of the committor by variationally minimizing a committor time-correlation function was proposed. By representing the RC as a 1D curvilinear path embedded in the space of the CVs, the string provides a natural framework that simultaneously reduces the high dimensionality of the problem while retaining nonlinearity. By virtue of the Voronoi tessellation, the tangent of the optimized path from the committor-consistent string method follows the gradient of the committor ∇q. The committor probability is broadly viewed as an 'ideal' RC[18]. Furthermore, transition path analysis of two-state systems showed that the ideal 1D RC in a reduced subspace of coordinates is directed along the gradient of the committor ∇q[19], a choice that is also consistent with the multidimensional Kramers-Langer theory[17,48]. This illustrates the clear advantage of a committor-consistent string over a more conventional MFEP-based string following the local mean force[7]. The images of the committor-consistent curvilinear string can be optimized by a Monte Carlo annealing method. In this algorithm, the string images are moved randomly in the CVs space and accepted or rejected based on the difference in likelihood. The present framework bears some similarities with previous methods[8,20-22,50,51]. It is closest in spirit with the nonlinear RC analysis proposed by Bolhuis and coworkers whereby a string pathway is optimized via a maximum likelihood criterion to model the committor data obtained from a path sampling simulation[13]. It might be possible to combine the two approaches in a unifying framework to determine an optimal committor-based pathway. Moreover, recent developments in ML techniques may help discovering the optimal multidimensional space of CVs[29-35]. Additional developments shall also consider the treatment of systems with more than two metastable states. Efforts in this direction would address the question as to whether we can predict the correct order of states through which the string should pass, and whether we can discover more than one important reaction pathway.

Methods

Brownian dynamics simulations were performed for the 1-dimensional double well and the Berezhkovskii-Szabo potential using in-house scripts. The NAMD software[52] was used to run Langevin dynamics and perform PCV in vacuum for alanine dipeptide and CG barstar-barnase. The CHARMM force field[53] was used in both cases. For alanine dipeptide, an unbiased Langevin dyanmics simulation was performed for a simulation time 5 μs with a time step of 0.5 fs to generate an initial MFEP path. Then, 4000 trajectories were generated using PCV at a time step of 0.5 fs for 100 ps. For the CG barstar-barnase, 25 independent trajectories were generated of 1 μ each with a time step of 1 fs. The PyEMMA software[54] was used to build the 2-dimensional MSM for comparison with our string method. The

nearest-neighbor regression algorithm in Scikit-learn[55] was used to produce a continuous range of q's on the potential surface of Figs. 3 and 4.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the National Science Foundation (NSF), through Grant No. MCB-1517221. The authors wish to acknowledge useful discussions and feedback from Attila Szabo, Sergei Krivov, and Aaron Dinner.

References

(1). Elber R; Karplus M A Method for Determining Reaction Paths in Large Molecules: Application to Myoglobin. Chem. Phys. Lett 1987, 139, 375–380.

(2). Jónsson H; Jacobsen KW In Nudged Elastic Band Method for Finding Minimum Energy Paths of Transitions; Berne BJ, Ciccotti G, Coker DF, Eds.; Classical and Quantum Dynamics in Condensed Phase Simulations; World Scientific: Singapore, 1998; Chapter 16, p 385.

(3). Bolhuis PG; Dellago C; Chandler D; Geissler P Transition Path Sampling: Throwing ropes over mountain passes, in the dark. Annu. Rev. of Phys. Chem 2002, 59, 291.

(4). E W; Ren W; Vanden-Eijnden E String method for the study of rare events. Phys. Rev. B 2002, 66, 052301.

(5). van Erp TS; Moroni D; Bolhuis PG A novel path sampling method for the calculation of rate constants. J. Chem. Phys 2003, 118, 7762–7774.

(6). Best R; Hummer G Reaction coordinates and rates from transition paths. Proc. Natl. Acad. Sci. U.S.A 2005, 102, 6732–6737. [PubMed: 15814618]

(7). Maragliano L; Fischer A; Vanden-Eijnden E; Ciccotti G String method in collective variables: minimum free energy paths and isocommittor surfaces. J. Chem. Phys 2006,125, 24106. [PubMed: 16848576]

(8). Peters B; Trout BL Obtaining reaction coordinates by likelihood maximization. J. Chem. Phys 2006,125, 054108. [PubMed: 16942204]

(9). Branduardi D; Gervasio FL; Parrinello M From A to B in free energy space. J. Chem. Phys 2007,126, 054103. [PubMed: 17302470]

(10). Yang H; Wu H; Li D; Han L; Huo S Temperature-Dependent Probabilistic Roadmap Algorithm for Calculating Variationally Optimized Conformational Transition Pathways. J. Chem. Theory and Comp 2007, 3, 17.

(11). Miller TF; Predescu C Sampling diffusive transition paths. J. Chem. Phys 2007, 126, 144102. [PubMed: 17444696]

(12). Pan A; Sezer D; Roux B Finding transition pathways using the string method with swarms of trajectories. J. Phys. Chem. B 2008,112, 3432–3440. [PubMed: 18290641]

(13). Lechner W; Rogal J; Juraszek J; Ensing B; Bolhuis PG Nonlinear reaction coordinate analysis in the reweighted path ensemble. J. Chem. Phys 2010,133, 174110. [PubMed: 21054009]

(14). E W; Vanden-Eijnden E Transition-path theory and path-finding algorithms for the study of rare events. Annu. Rev. Phys. Chem 2010, 61, 391–420. [PubMed: 18999998]

(15). Onsager L Initial recombination of ions. Phys. Rev 1938, 54, 554–557.

(16). Bolhuis PG; Dellago C; Chandler D Reaction coordinates of biomolecular isomerization. Proc. Natl. Acad. Sci. U. S. A 2000, 97, 5877–5882. [PubMed: 10801977]

(17). Berezhkovskii A; Szabo A One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions. J. Chem. Phys 2005,122, 14503. [PubMed: 15638670]

(18). Li W; Ma A Recent developments in methods for identifying reaction coordinates. Mol. Sim 2014, 40, 784–793.

(19). Roux B Transition rate theory, spectral analysis, and reactive paths. J. Chem. Phys 2022, 156, 134111. [PubMed: 35395906]

(20). Ma A; Dinner AR Automatic method for identifying reaction coordinates in complex systems. J. Phys. Chem. B 2005,109, 6769–6779. [PubMed: 16851762]

(21). Peters B; Bolhuis PG; Mullen RG; Shea J-E Reaction coordinates, one-dimensional Smoluchowski equations, and a test for dynamical self-consistency. J. Chem. Phys 2013, 138.

(22). Peters B. Reaction Coordinates and Mechanistic Hypothesis Tests. Annu. Rev. Phys. Chem 2016, 67, 669–690. [PubMed: 27090846]

(23). Krivov SV On Reaction Coordinate Optimality. J. Chem. Theory Comp 2013, 9, 135–146.

(24). Banushkina PV; Krivov SV Nonparametric variational optimization of reaction coordinates. J. Chem. Phys 2015,143.

(25). Okazaki KI; Wöhlert D; Warnau J; Jung H; Yildiz Ö; Kühlbrandt W; Hummer G Mechanism of the electroneutral sodium/proton antiporter PaNhaP from transitionpath shooting. Nat Commun 2019,10, 1742. [PubMed: 30988359]

(26). Pérez-Hernández G; Paul F; Giorgino T; De Fabritiis G; Noé F Identification of slow molecular order parameters for Markov model construction. The J. Chem. Phys 2013, 139, 07B604_1.

(27). Tiwary P; Berne BJ Spectral gap optimization of order parameters for sampling complex molecular systems. Proc. Natl. Acad. Sci. U.S.A 2016,113, 2839–2844. [PubMed: 26929365]

(28). Smith Z; Pramanik D; Tsai ST; Tiwary P Multi-dimensional spectral gap optimization of order parameters (SGOOP) through conditional probability factorization. J. Chem. Phys 2018,149, 234105. [PubMed: 30579304]

(29). Chen W; Ferguson AL Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. J Comput Chem 2018, 39, 2079–2102. [PubMed: 30368832]

(30). Jung H; Covino R; Hummer G Artificial intelligence assists discovery of reaction coordinates and mechanisms from molecular dynamics simulations. arXiv 2019, 1901.04595.

(31). Bonati L; Zhang YY; Parrinello M Neural networks-based variationally enhanced sampling. Proc. Natl. Acad. Sci. U.S.A 2019,116, 17641–17647. [PubMed: 31416918]

(32). Sidky H; Chen W; Ferguson AL Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. Molecular Physics 2020, 118, e1737742.

(33). Tsai ST; Kuo EJ; Tiwary P Learning molecular dynamics with simple language model built upon long short-term memory neural network. Nat. Comm 2020, 11, 5115.

(34). Bonati L; Piccini G; Parrinello M Deep learning the slow modes for rare events sampling. Proc. Natl. Acad. Sci. U.S.A 2021, 118, e2113533118. [PubMed: 34706940]

(35). Trizio E; Parrinello M From Enhanced Sampling to Reaction Profiles. J. Phys. Chem. Lett 2021, 12, 8621–8626. [PubMed: 34469175]

(36). Pant S; Smith Z; Wang Y; Tajkhorshid E; Tiwary P Confronting pitfalls of AI-augmented molecular dynamics using statistical physics. J. Chem. Phys 2020, 153, 234118. [PubMed: 33353347]

(37). Tsai ST; Smith Z; Tiwary P SGOOP-d: Estimating Kinetic Distances and Reaction Coordinate Dimensionality for Rare Event Systems from Biased/Unbiased Simulations. J. Chem. Theory Comp 2021, 17, 6757–6765.

(38). Branduardi D; Gervasio FL; Parrinello M From A to B in Free Energy Space. J. Chem. Phys 2007, 126, 054103. [PubMed: 17302470]

(39). Roux B String Method with Swarms-of-Trajectories, Mean Drifts, Lag Time, and Committor. J Phys Chem A 2021, 125, 7558–7571. [PubMed: 34406010]

(40). Vanden-Eijnden E In Transition path theory; Ferrario M, Ciccotti G, Binder K, Eds.; Computer Simulations in Condensed Matter: From Materials to Chemical Biology; Springer, 2006; Vol. 2; p 439.

(41). Berezhkovskii A; Hummer G; Szabo A Reactive flux and folding pathways in network models of coarse-grained protein dynamics. J. Chem. Phys 2009, 130, 205102. [PubMed: 19485483]

(42). Berezhkovskii AM; Szabo A Committors, first-passage times, fluxes, Markov states, milestones, and all that. J. Chem. Phys 2019, 150, 054106. [PubMed: 30736684]

(43). Thiede EH; Giannakis D; Dinner AR; Weare J Galerkin approximation of dynamical quantities using trajectory data. J. Chem. Phys 2019, 150, 244111. [PubMed: 31255053]

(44). Strahan J; Antoszewski A; Lorpaiboon C; Vani BP; Weare J; Dinner AR Long-time-scale predictions from short-trajectory data: A benchmark analysis of the trp-cage miniprotein. J. Chem. Theory Comp 2021, 17, 2948–2963.

(45). Tiwary P; Berne BJ Predicting reaction coordinates in energy landscapes with diffusion anisotropy. J. Chem. Phys 2017, 147, 152701. [PubMed: 29055314]

(46). Li W Optimizing reaction coordinate by flux maximization in the transition path ensemble. J. Chem. Phys 2022, 156, 054117. [PubMed: 35135266]

(47). Berezhkovskii AM; Szabo A Diffusion along the splitting/commitment probability reaction coordinate. J Phys Chem B 2013, 117, 13115–13119. [PubMed: 23777371]

(48). Langer J. Statistical theory of the decay of metastable states. Annu. Phys 1969, 54, 258–275.

(49). He Z; Paul F; Roux B A critical perspective on Markov state model treatments of protein-protein association using coarse-grained simulations. J Chem Phys 2021, 154, 084101. [PubMed: 33639768]

(50). Ma A; Nag A; Dinner AR Dynamic coupling between coordinates in a model for biomolecular isomerization. J. Chem. Phys 2006, 124, 144911. [PubMed: 16626249]

(51). Hu J; Ma A; Dinner AR A two-step nucleotide-flipping mechanism enables kinetic discrimination of DNA lesions by AGT. Proceedings of the National Academy of Sciences 2008, 105, 4615–4620.

(52). Phillips JC; Hardy DJ; Maia JD; Stone JE; Ribeiro JV; Bernardi RC; Buch R; Fiorin G; Hénin J; Jiang W et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys 2020, 153, 044130. [PubMed: 32752662]

(53). Brooks BR; Brooks CL III; Mackerell AD Jr.; Nilsson L; Petrella RJ; Roux B; Won Y; Archontis G; Bartels C; Boresch S et al. CHARMM: The Biomolecular Simulation Program. J. Comp. Chem 2009, 30, 1545–1614. [PubMed: 19444816]

(54). Scherer MK; Trendelkamp-Schroer B; Paul F; Pérez-Hernández G; Hoffmann M; Plattner N; Wehmeyer C; Prinz J-H; Noé F PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. J. Chem. Theory Comp 2015, 11, 5525–5542.

(55). Pedregosa F; Varoquaux G; Gramfort A; Michel V; Thirion B; Grisel O; Blondel M; Prettenhofer P; Weiss R; Dubourg V et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011, 12, 2825–2830.

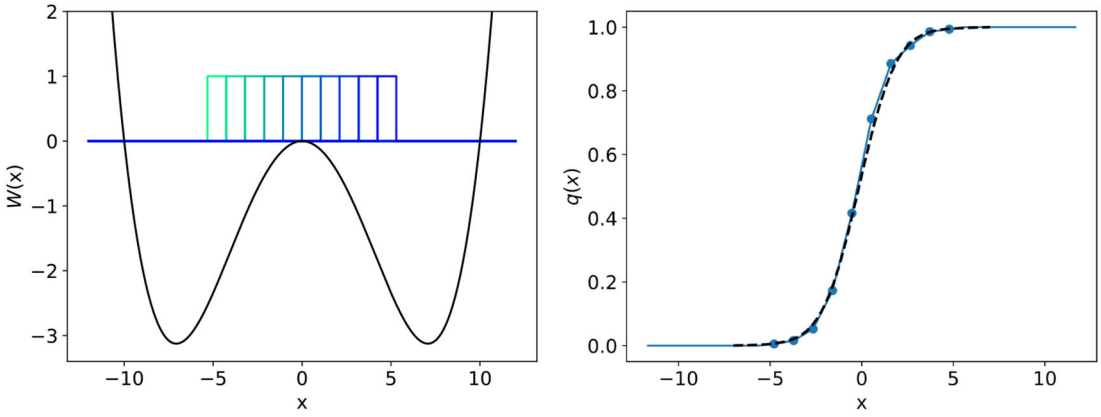


Figure 1:

Illustration of a one-dimensional double well using using ten one-hot indicator functions and the resulting committor probability. The boundaries near the well minima where $q = 0$ or 1 were defined to be within 1 k_BT of the well depth, or at $x = -5.31$ and $x = 5.31$. The black dashed line is the exact committor calculated numerically from the double well potential.

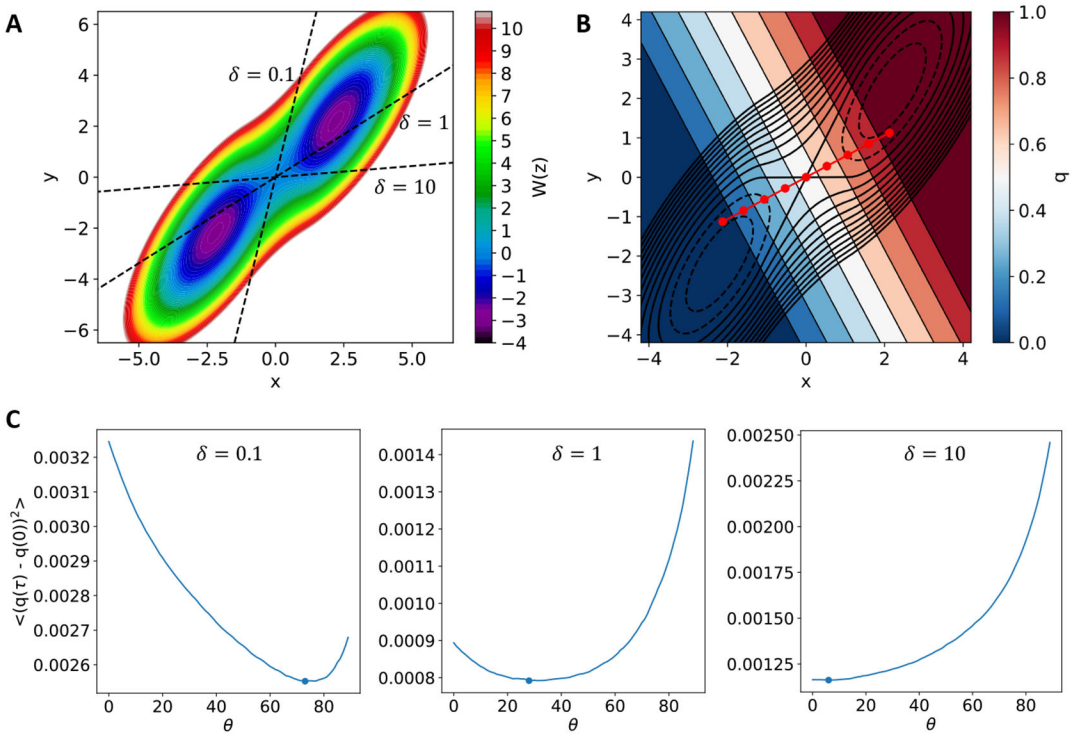


Figure 2:

Berezhkovskii-Szabo potential and correlation functions. (A) The 2D-BS potential surface with dashed lines showing the angle of the optimal reaction pathway for the diffusion cases obtained using a basis set of 9 Voronoi in the intermediate region: $\theta = 6^{\circ}$ for $\delta = 10$, $\theta = 28^{\circ}$ for $\delta = 1$, and $\theta = 73^{\circ}$ for $\delta = 0.1$. (B) Voronoi tessellation of the straight pathway for $\delta = 1$. The Voronoi cells are colored by the value of the committor, sequentially increasing from $q = 0$ (blue, bottom left) to $q = 0$ (white, middle region) to $q = 1$ (red, top right). (C) Dependence of the time-correlation function on the angle θ for $\delta = 10$ (left), $\delta = 1$ (middle), $\delta = 0.1$ (right). The minimum of the correlation function of each diffusion case is indicated by the blue dot and taken to be the angle of the optimal reaction pathway at the saddle point.

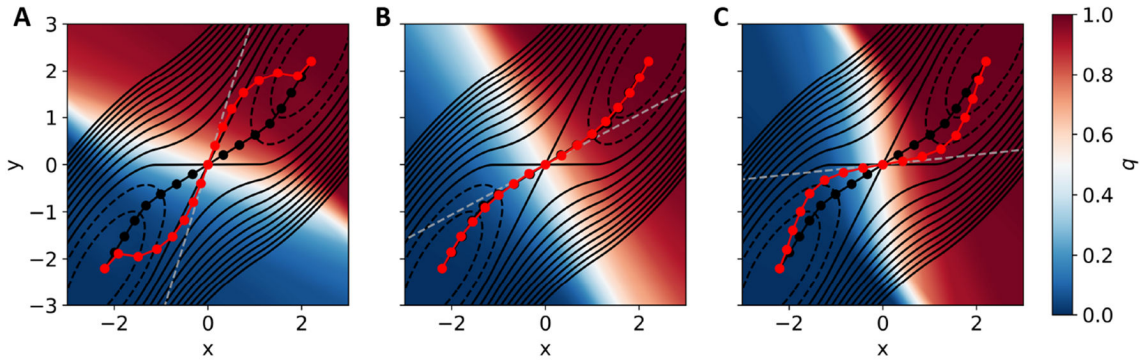


Figure 3:

String obtained by variationally minimizing the correlation function with Monte Carlo sampling and the resulting committor probability for three diffusion conditions: (A) for $\delta = 0.1$, (B) for $\delta = 1$, and (C) for $\delta = 10$. For all three cases, optimization was performed by starting from the MFEP string shown as a black curve; a box of size 0.05×0.05 was used to sample around each image position in the 2D plane. One Monte Carlo iteration cycle comprises moving the image, and then computing the committor-correlation function. Once a move is accepted for a new image position that results in a lower committor-correlation function value, then the Monte Carlo sampling for that image stops and proceeds to the next image. If after 1000 moves none were accepted, then that image in the iteration cycle is assumed to be converged and we proceed to sample for the next image. For $\delta = 0.1$, 1, and 10, the Monte Carlo procedure required 56, 25, and 42 iterations, respectively, to reach convergence.

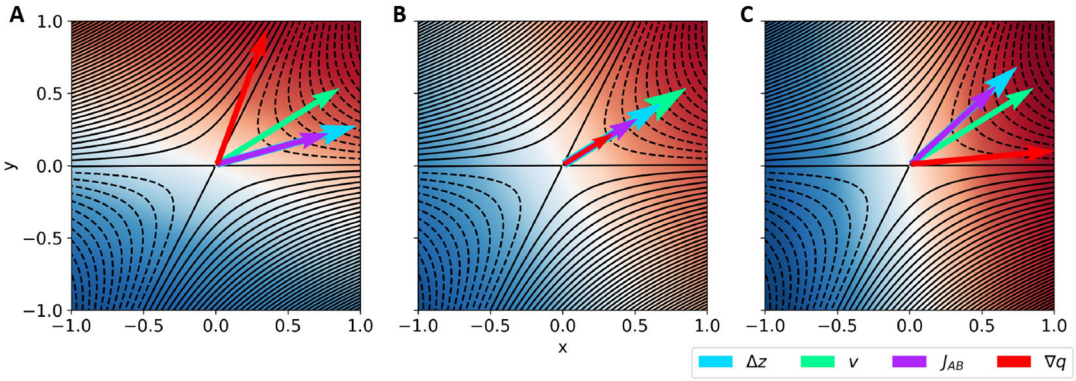


Figure 4:

Comparison of selected eigenvectors computed from the diffusion matrix and Hessian. (A) $\delta = 0.1$. (B) $\delta = 1$. (C) $\delta = 10$. For each diffusion condition, we show the directions of the principal component of the Hessian matrix $\mathbf{V}$ (green), the overall reactive flux $\mathbf{J}_{AB}$ (purple), the mean drift $\langle \Delta z(\tau) \rangle$ (blue), and $\nabla q(\mathbf{z})$ (red).

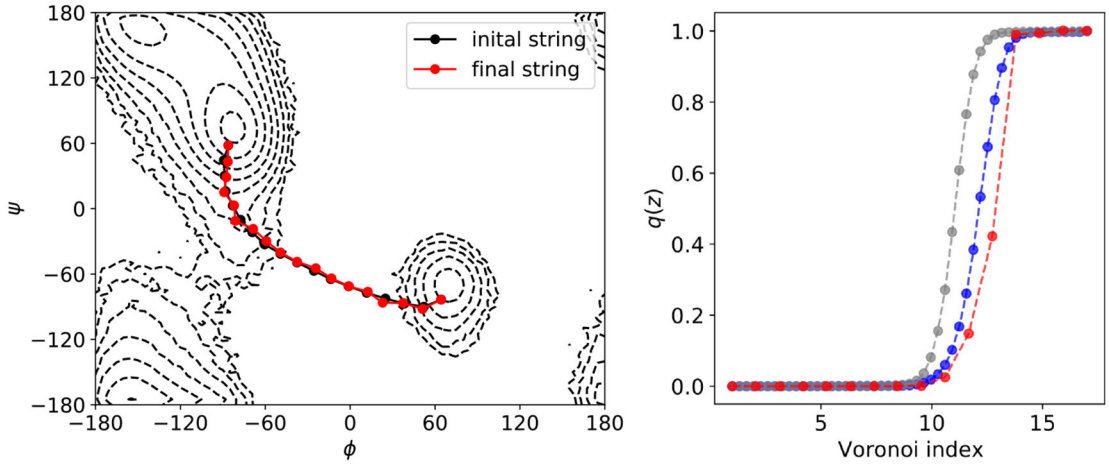


Figure 5:

Variational string method implemented for the dialanine. The initial string (black) consists of straight paths connecting the local minima while the final string (red) is obtained by variationally minimizing the committor-correlation function. The committor probability of this final string (red) increases from 0 to 1 with a steep curve at the transition barrier. The position of the images was optimized by minimizing the committor time-correlation function using the iterative Monte Carlo procedure with random moves over a box 1.0° × 1.0°. The Monte Carlo optimized string and q is compared with results from PCV using a string following the mean force (blue) and string method with swarms of trajectories (gray).

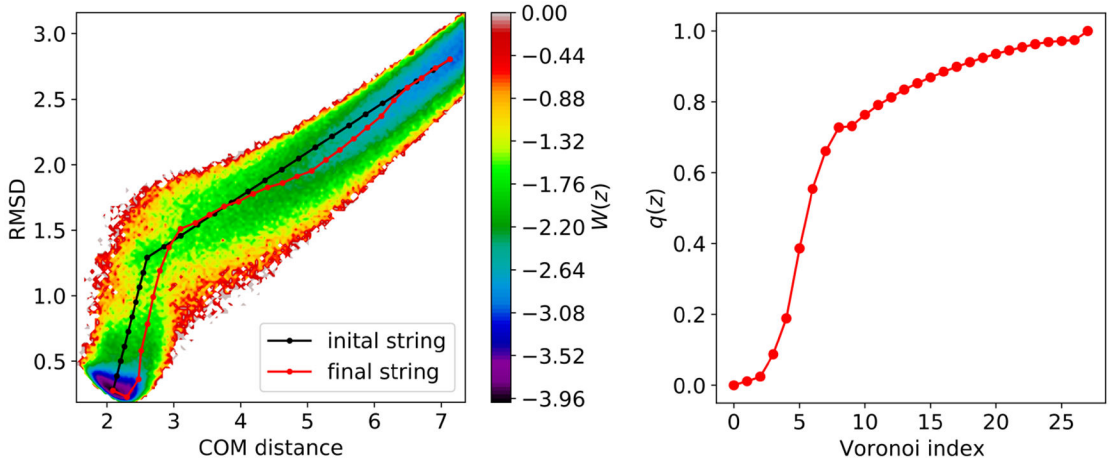


Figure 6:

Variational string method implemented for the barnase-barstar complex. Left: The initial string (black) consists of straight paths connecting the local minima while the final string (red) is obtained by variationally minimizing the committor-correlation function. Right: The committor probability of the string increases from 0 to 1 relatively smoothly.