Determination of reaction coordinates via locally scaled diffusion map

Mary A. Rohrdanz,[1] Wenwei Zheng,[1] Mauro Maggioni,[2] and Cecilia Clementi[1,a)]
[1]*Rice University, Department of Chemistry, Houston, Texas 77005, USA*
[2]*Duke University, Department of Mathematics, Durham, North Carolina 27708, USA*

(Received 14 October 2010; accepted 3 March 2011; published online 30 March 2011)

We present a multiscale method for the determination of collective reaction coordinates for macromolecular dynamics based on two recently developed mathematical techniques: diffusion map and the determination of local intrinsic dimensionality of large datasets. Our method accounts for the local variation of molecular configuration space, and the resulting global coordinates are correlated with the time scales of the molecular motion. To illustrate the approach, we present results for two model systems: all-atom alanine dipeptide and coarse-grained src homology 3 protein domain. We provide clear physical interpretation for the emerging coordinates and use them to calculate transition rates. The technique is general enough to be applied to any system for which a Boltzmann-sampled set of molecular configurations is available. © *2011 American Institute of Physics*. [doi:10.1063/1.3569857]

I. INTRODUCTION

The pursuit of collective reaction coordinates is of great interest to researchers that deal with macromolecular dynamics, as such coordinates are crucial for extracting meaningful information from the large volume of data routinely produced by molecular dynamics (MD) simulations. This search is based on the premise that while the dimensionality of molecular configuration space is high, oftentimes the distribution of physically relevant states is highly clustered around a set of much lower dimensionality. This working assumption has been empirically verified for a number of different systems, e.g.,[1–4] and motivates the definition of reaction coordinates with which to study the system's collective dynamics and identify (meta-)stable states.

Methods for the determination of coordinates capable of describing this low-dimensional space have been developed by a number of researchers. For example, isocommittor surfaces can be computed, giving the probability of a given configuration to transition to a reactant or product free-energy minimum.[5] Genetic neural network algorithms[6] and Bayesian analysis methods[7] have been developed to assess and select the best reaction coordinates from a set of prospective ones. In addition, methods for finding the minimum free-energy path, such as the string method,[8,9] transition path sampling,[10] and milestoning,[11] provide a reaction coordinate as the collective variable mapping the resulting path. For all of these methods, some initial specification of collective variables and/or reactant and product states are required.

Geometrical dimensionality reduction techniques have also been applied to molecular systems, including linear principal component analysis[12] and its nonlinear variants,[13,14] local linear embedding,[15] and Isomap.[16,17] These techniques do not require any input information concerning potential reaction coordinates or reactant/product states; however, they are limited in that they consider the number of effective dimensions *only* as a global property and do not account for the local heterogeneity of MD simulation data. We (and others[18]) have found these variations to be important, and in the currently proposed method such differences are used in the construction of overall coordinates.

The approach we propose builds upon two recently developed mathematical techniques: diffusion maps[19–23] (which have been applied in various context such as machine learning tasks[24,25] and manifold parametrization[26]), and estimation of the intrinsic dimensionality of noisy datasets.[27] We make two new contributions: (1) extending the diffusion map method to include a locally determined variable length scale and (2) applying the method to molecular dynamics simulation data. A collection of Boltzmann-distributed molecular configurations plays the role of "noisy dataset," and our algorithm determines both the number of effective dimensions at each configuration, and the length scale within which this intrinsic dimensionality persists. These position-dependent local length scales are input to a diffusion map calculation, yielding a few global coordinates that correlate with different time scales in the system. Since our work is a combination and extension of diffusion map and local scale analysis, we refer to the method as "locally scaled diffusion map" (LSDMap). Our method does not require any *a priori* knowledge about the system (such as prospective reaction coordinates and/or the definition of reactant and product states), and the local heterogeneity of the MD data is accounted for in the construction of global coordinates.

We apply the LSDMap framework to characterize the dynamics of two very different and well-understood test systems: an all-atom model of alanine dipeptide and a coarse-grained model of the src homology 3 domain protein (SH3). Through an analysis of the LSDMap, we find insights into the nature of the free-energy minima, transition regions, and overall free-energy landscape. We verify the mathematical assertion that the diffusion coordinates (DCs) are good reaction coordinates through calculation of the diffusion rate between free-energy minima. We find that the diffusion

a)Electronic mail: cecilia@rice.edu.

coordinates provide rates closer to the simulation rate than those of competing empirical coordinates. By comparing with the empirical coordinates and probability of contact formation, we obtain a straightforward physical interpretation of the diffusion coordinates.

The remainder of this paper is organized as follows. In Sec. II we outline the mathematical underpinnings of the diffusion map and local scale determination. We detail our procedure for obtaining an LSDMap from molecular dynamics simulation data in Sec. III. In Sec. IV we provide results for the two test systems; in Sec. V we give concluding remarks.

II. MATHEMATICAL BACKGROUND: LSDMap

Below we present some of the relevant mathematical background on diffusion maps in the current context of MD simulations. We refer the reader to the original literature[19,23] for the full details.

For a system with N atoms, with a given potential energy function $E(\boldsymbol{x})$, at constant temperature T, and in the limit of high friction, the Fokker–Planck equation governs the temporal evolution of the probability distribution $p(\boldsymbol{x}, t)$ at any configuration $\boldsymbol{x} \in \mathbb{R}^{3N}$ of the system,

$$\frac{\partial p}{\partial t} = -\sum_{i}^{3N} \frac{\partial}{\partial x_i}\left(\frac{1}{\beta}\frac{\partial}{\partial x_i} + \frac{\partial E}{\partial x_i}\right) p = -\mathbf{H}_{\mathrm{FP}}\, p, \tag{1}$$

where $\beta = 1/(k_B T)$, k_B is Boltzmann's constant, and t is the time variable. Under rather general conditions, the operator $\mathbf{H}_{\mathrm{FP}}$, which acts on an infinite-dimensional space of probability distributions, has a discrete eigenspectrum of non-negative eigenvalues λ_i, with $\lambda_0 = 0 < \lambda_1 \leq \lambda_2 \leq \ldots$, and corresponding eigenfunctions $\phi_i(\boldsymbol{x})$. Formally (and rigorously in an appropriate metric that depends on various assumptions about $\mathbf{H}_{\mathrm{FP}}$), the general solution of the Fokker–Planck equation is

$$p(\boldsymbol{x}, t) = \phi_0(\boldsymbol{x}) + \sum_{i=1}^{\infty} c_i \phi_i(\boldsymbol{x}) e^{-\lambda_i t}, \tag{2}$$

where the coefficients c_i are determined by the initial distribution $p(\boldsymbol{x}, t = 0)$. The eigenfunction $\phi_0(\boldsymbol{x})$ is the Boltzmann distribution, approached by any initial distribution when $t \gg 1/\lambda_1$.

For systems with one (or a few) slow process(es) dominating the dynamics (such as the crossing of a free-energy barrier), the eigenspectrum will present a gap; i.e., $\lambda_{k+1} \gg \lambda_k$ for some k, and the evolution of the probability distribution toward equilibrium may be approximated as the first k terms of the general solution,

$$p(\boldsymbol{x}, t) = \phi_0(\boldsymbol{x}) + \sum_{i=1}^{k} c_i \phi_i(\boldsymbol{x}) e^{-\lambda_i t}, \tag{3}$$

at least at time scales $t \gg 1/\lambda_{k+1}$. In these situations it has been shown that $\phi_i(\boldsymbol{x})/\phi_0(\boldsymbol{x})$, which are eigenfunctions of the backward Fokker–Planck operator,[28] serve as collective coordinates in the sense that their time evolution is approximately Markovian and independent of the remaining degrees of freedom. These are the diffusion coordinates, and the diffusion map is the nonlinear mapping from the space of molecular configurations to the diffusion coordinate space.

An efficient numerical method to approximate these first few eigenfunctions and associated eigenvalues using samples of the equilibrium distribution has been recently proposed.[23] The approach involves defining a weighted graph on the simulation data and determining the first few eigenvalues and eigenvectors of a random walk on the graph. The weights are related to the transition probability between configurations and will be larger for configurations that are similar in structure. Here we measure similarity by the root mean square deviation (RMSD) between structures (as opposed to Euclidean distance used previously[23]) in order to quotient out irrelevant translational and rotational degrees of freedom. The transition probability between any two structures is based on the kernel,

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||^2}{2\varepsilon_i \varepsilon_j}\right), \tag{4}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent two molecular configurations, and $||\boldsymbol{x}_i - \boldsymbol{x}_j||$ is their RMSD. An appropriate renormalization of K, described in Sec. III B, leads to a random walk on $\{\boldsymbol{x}_i\}$ whose eigenfunctions approximate those in Eq. (3).

This method is predicated on the idea that the high-dimensional space of molecular configurations can be approximated by a lower dimensional set $\mathcal{M}$, and that the intrinsic dimensionality of $\mathcal{M}$ is location dependent. The local scale parameter ε_i can be interpreted as the distance around $\boldsymbol{x}_i$ within which $\mathcal{M}$ can be well approximated by a low-dimensional hyperplane tangent to $\mathcal{M}$ at $\boldsymbol{x}_i$. This is the region around $\boldsymbol{x}_i$ in which $\mathcal{M}$ is approximately linear (i.e., "locally flat"). In previous applications of diffusion maps, ε_i has always been chosen equal to a constant value ε independently of $\boldsymbol{x}_i$.[19,23]

Little is known about the choice of this crucial local scale parameter, with theoretical results providing only some guidance in the asymptotic regime when the number of configurations is very large (at least exponential in the intrinsic dimension of the effective configuration space), and often *ad hoc* techniques are used in practice. If the data sample is dense and lies on a smooth, non-noisy, low-dimensional manifold, the choice of ε is not critical to the numerical estimation of the Fokker–Planck eigenfunctions—using a constant value yields meaningful results (as the number of samples grows, the estimated generator of the diffusion converges to the true generator of the Fokker–Planck equation).

However a dataset of macromolecular configurations from MD simulations has highly variable density (due to the properties of the Boltzmann distribution), it is very noisy (with the characteristics of the noise changing with the region of configurational space), and it is not infinitely dense. In such a situation, if the parameter ε is selected too small, e.g., comparable to the scale of the noise, the results will be corrupted because the "locally flat" region will correspond to that of the noise rather than that of the actual data. On the other hand if ε is too large, regions of the system will be considered artificially flat, again corrupting the results. We have found that in molecular dynamics applications a uniform value of ε yields a

Fokker–Planck eigenspectrum strongly dependent on the selected value of ε, and no straightforward interpretation of the results is possible. Examples of the application of diffusion map with constant ε to both systems considered here are discussed in the supplementary material[29] Sec. I.

Inspired by the results of Little *et al.*,[27] we define below an algorithm for determining the intrinsic dimension and local scale associated with each configuration in a set of MD data. As the local scale parameter ε_i indicates the (unknown) length scale around $\boldsymbol{x}_i$ at which $\mathcal{M}$ can be well approximated by its (unknown) tangent hyperplane at $\boldsymbol{x}_i$, we obtain an estimate of ε_i by performing multidimensional scaling (MDS), a linear dimensionality reduction technique, over increasingly large neighborhoods of $\boldsymbol{x}_i$. Under very general assumptions on the geometry of $\mathcal{M}$, the density of points, and the noise, such a technique leads to robust identification of the local scale and intrinsic dimension around any point on $\mathcal{M}$.[27] Moreover, this technique requires a number of samples $\{\boldsymbol{x}_i\}$ linear in the intrinsic dimension of $\mathcal{M}$ and independent of the large ambient dimensionality.[27]

III. ALGORITHM: LSDMap

The first step in the LSDMap calculation is to acquire a Boltzmann-distributed set of molecular configurations. These may be obtained either as one long run or many short runs of molecular dynamics simulation. Then the local scale and the diffusion coordinates can be obtained through the algorithm we detail below.

A. Determination of local scale and dimension

The estimation of the local scale ε_i for each configuration $\boldsymbol{x}_i$ in the dataset is as follows. For each $\boldsymbol{x}_i$ we order the remaining configurations according to their RMSD to $\boldsymbol{x}_i$. We perform MDS on increasingly larger neighborhoods (ε-balls) around $\boldsymbol{x}_i$. This yields an MDS singular value spectrum as a function of the RMSD radius of the ε-balls. We divide these singular values by the square root of the number of configurations within the ε-ball. These are the normalized MDS spectra.

An analysis of the gaps between the singular values as a function of neighborhood size provides information about both the intrinsic local dimensionality and the local scale. First of all, this analysis allows for a separation of the relevant degrees of freedom ("data") from the irrelevant ones ("noise"). Singular values corresponding to noise will decrease in value and clump together at larger length scales, while the singular values that correspond to the actual data continue to increase (see, for example, Fig. 1 of Little *et al.*[27]). A conservative estimate of the local scale ε_i around a point $\boldsymbol{x}_i$ is obtained by considering the size of the ε-balls at which the noise spectra begin to decrease and separate from the data. We have found that the intrinsic dimensionality of MD data changes rapidly from region to region (see Sec. IV), and we have modified the algorithm proposed by Little *et al.*[27] to take it into account; the full details are explained in Appendix A.

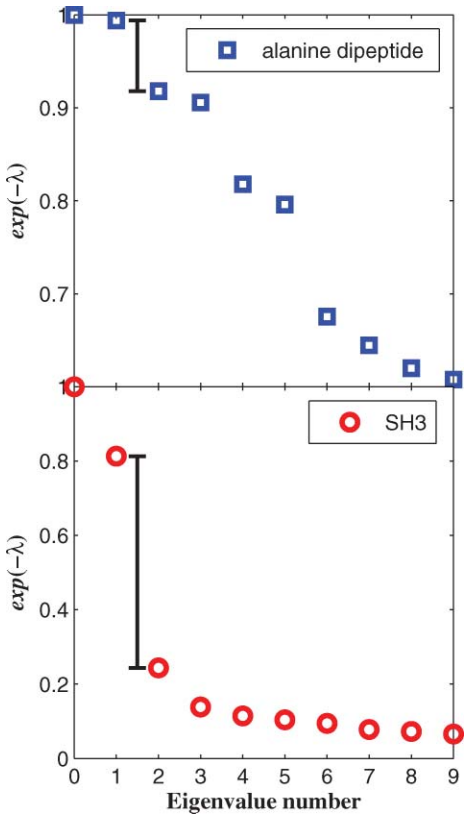

FIG. 1. The negative exponential of the eigenvalues of the Fokker-Planck operator are plotted as a function of eigenvalue number. For both systems the zeroth eigenvalue $\lambda_0 = 0$, $\exp\{-\lambda_0\} = 1$ corresponds to the Boltzmann distribution. The spectral gap between the first and second eigenvalues (denoted by the black bars) suggests that there is a single slow time scale dominating the dynamics. This time scale corresponds to the C_5, $P_{\|} \rightarrow \alpha_P, \alpha_R$ isomerization process in alanine dipeptide and the folding/unfolding transition for SH3.

B. Diffusion map with local scale

Once the local scale of each $\boldsymbol{x}_i$ is determined, the first few eigenfunctions of the backward Fokker–Planck operator, i.e., the diffusion coordinates, are calculated as follows. The algorithm below closely parallels that of Coifman *et al.*[23]; the differences being the use of the RMSD as the distance measure (rather than the Euclidean distance) and set of ε_i values $\{\varepsilon\}$ (rather than a uniform ε). Here the RMSD distance is more appropriate because molecular systems are invariant under rigid rotations and translations of the system; the locally determined ε_i is required due to the very high variability of the molecular data, as discussed above. In practice, for a dataset with $\mathcal{N}$ configurations:

1. Construct the $\mathcal{N} \times \mathcal{N}$ matrix, the transition probability kernel K, as

$$K_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\varepsilon_i \varepsilon_j}\right), \tag{5}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent two molecular configurations, ε_i and ε_j are their respective local scales, and $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is the RMSD distance between them.

2. For each $\boldsymbol{x}_i$, sum the corresponding row of K to compute

$$P_i = \sum_{j=1}^{\mathcal{N}} K_{ij}, \tag{6}$$

which is proportional to a density estimation around $\boldsymbol{x}_i$.

3. Normalize the kernel as

$$\tilde{K}_{ij} = \frac{K_{ij}}{\sqrt{P_i P_j}}. \tag{7}$$

4. Define the diagonal matrix D as $D_i = \sum_{j=1}^{\mathcal{N}} \tilde{K}_{ij}$, and construct a Markov matrix $M = D^{-1}\tilde{K}$,

$$M_{ij} = \frac{\tilde{K}_{ij}}{D_i}. \tag{8}$$

5. Compute the first few largest eigenvalues and the corresponding right eigenvectors of M.

These eigenvectors serve as the diffusion coordinates. Coifman and Lafon[19] have shown that for data points $\boldsymbol{x}_i$ randomly sampled from a Boltzmann distribution, as $\mathcal{N} \to \infty$ and as $\varepsilon \to 0$ (for uniform ε and at an appropriate rate in $\mathcal{N}$, depending on unknown quantities, such as the intrinsic dimension of the data, and possibly the size of the noise), the right eigenvectors of M converge (in probability) to the eigenfunctions of the backward Fokker–Planck operator. This result enables approximation of the eigenfunctions of the Fokker–Plank operator from simulated trajectories even for high-dimensional systems where standard discretization methods are not feasible.

The matrix M is adjoint to a symmetric matrix $M_s = D^{-1/2}\tilde{K}D^{-1/2}$, and the numerical computation of the first few eigenvalues and eigenvectors is in practice performed on M_s. The complexity of the above algorithm, including the local scale determination, is $O(k\mathcal{N}^2 N)$ for $\mathcal{N}$ configurations in $\mathbb{R}^{3N}$ and k eigenvectors. In practice, one may not construct the full matrix K but rather a sparse version where entries below a certain threshold are set to 0. If the cost of identifying the nonzero entries, i.e., finding the ε-neighbors of each point, is less than $O(\mathcal{N}^2)$ and the resulting matrix is sparse, substantial computational savings may be achieved.

IV. RESULTS AND DISCUSSION

We apply the LSDMap approach to two test systems: all-atom alanine dipeptide in implicit water and coarse-grained SH3. To verify the mathematical assertion that the diffusion coordinates function as good reaction coordinates, we calculate transition rates between free energy minima. We determine the free-energy profile along the diffusion coordinates and various competing empirical coordinates, then calculate transition rates using Kramers' expression for the escape rate of a system moving over a barrier.[30] The Kramers' escape rate is given by

$$\text{rate} = \left(\int_{\text{barrier}} \frac{e^{\beta F(x)}}{D(x)} dx \int_{\text{well}} e^{-\beta F(x')} dx' \right)^{-1}, \tag{9}$$

with $\beta = (k_B T)^{-1}$, free-energy $F(x)$, and diffusion coefficient $D(x)$. In the evaluation of the integrals above, the barrier region is defined as the segment of the coordinate between the free-energy minima; the well region is defined as the half of the configurational space containing the free-energy minimum corresponding to the "reactant" state and delimited by the top of the barrier. In practice only configurations at the top of the barrier or bottom of a minimum will significantly contribute to these integrals, and the resulting rates are very robust upon changes of the integration limits around the ones so defined.

The Kramers' rate calculated from a free-energy profile is strongly dependent on the choice of coordinates used in defining the free-energy.[31] A poor reaction coordinate tends to convolute motion directed over the top of a free-energy barrier with motion perpendicular to the barrier, and therefore underestimates the barrier height and overestimate the rate. An optimal reaction coordinate is perpendicular to the separatrix defining the transition state,[4] and the rate evaluated via the Kramers' expression along such a coordinate should provide a good estimate of the actual rate.

In order to calculate the Kramers' rate through Eq. (9), the coordinate-dependent diffusion coefficient $D(x)$ along the reaction coordinates is required. These were obtained through Bayesian analysis,[32] which allows for an estimation of $D(x)$ from MD simulation data. We used these techniques as originally proposed in Ref. 32; the choice of the parameters and a brief description of the methods are given in Appendix B.

A. Alanine dipeptide

Alanine dipeptide is a typical testbed for collective dynamics studies. Although the molecule consists of 22 atoms, multiple steric constraints effectively reduce the configuration space to two dimensions under standard conditions. The two dimensions of choice are the dihedral angles Φ and Ψ. As the two angles are *a priori* known, this system represents an ideal case to test our approach.

The MD data are obtained with AMBER (Ref. 33) from a 300 K simulation with the AMBER99 force field in implicit water. Configurations collected every 0.1 ps during a 20 000 ps simulation are used as input to the local scale determination and diffusion map calculation. The hydrogen atoms were removed before the local scale determination, since they do not contribute to important conformational changes of the molecule.

It is important to emphasize a few points about the trajectory data. A much smaller dataset can be used in the LSDMap approach; using only 10 000 configurations yields a free-energy landscape as a function of diffusion coordinates that is indistinguishable from that of the 200 000 configuration result and can be calculated in a day of computer time on a single-core workstation. The reason for using such a large dataset here is twofold: to test the robustness of the small sample results and to provide adequate sampling for the Bayesian analysis used to calculate the diffusion coefficients[32] from a single long trajectory. Alternatively, if a smaller data sample is used, the position-dependent diffusion coefficients

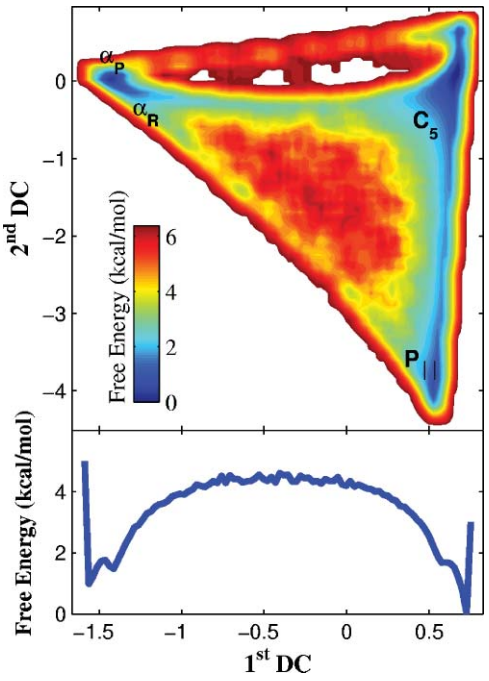

FIG. 2. (Top) Free energy of alanine dipeptide as a function of the first and second diffusion coordinates. (Bottom) Free energy profile along the first diffusion coordinate (DC). The Kramers' rate along the first DC is shown in Table I.

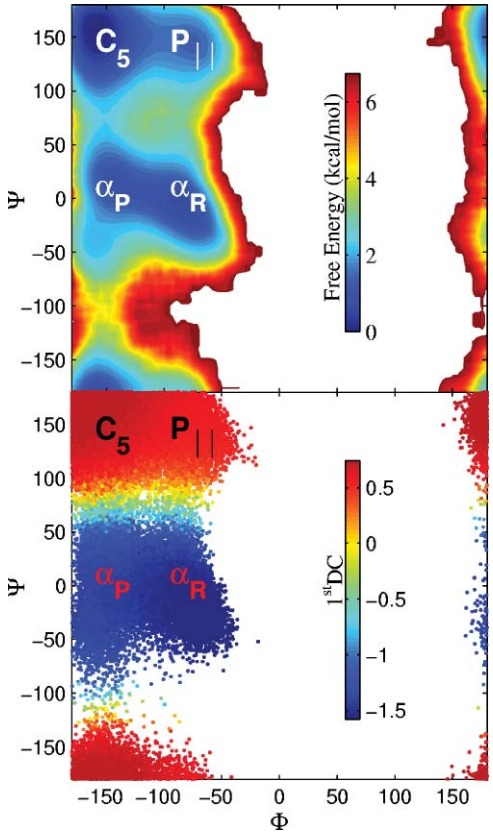

FIG. 3. Comparison of the first DC with empirical coordinates Φ and Ψ. (Top) Free energy (kcal/mol) as a function of dihedral angles Φ and Ψ; displayed to show the locations of the free energy minima in Φ−Ψ space. (Bottom) Raw molecular configuration data plotted according to Φ and Ψ, and colored according to first DC. The smooth color change between the pairs of minima C_5–$P_{||}$ and α_R–α_P shows that the first DC corresponds to a transition between these pairs, and that the first DC correlates well with Ψ. Analogous figures for the second and third DCs are available in the supplementary material.

could be estimated by using many independent short simulations.[32]

The Fokker–Planck eigenvalue spectrum calculated with the LSDMap is displayed in the top panel of Fig. 1. The $\lambda_0 = 0$ eigenvalue corresponds to the Boltzmann distribution; λ_1 to the first DC; λ_2 to the second DC, etc. The gap between $\exp(-\lambda_1)$ and $\exp(-\lambda_2)$, denoted by the vertical bar, shows that there is a separation of time scales between the collective motion corresponding to the first DC and that of the second DC.

Figure 2 shows the free-energy as a function of the first and second DCs (top panel), from which it is clear that the first DC corresponds to a transition between two pairs of minima: C_5–$P_{\|}$ and α_R–α_P. This can be corroborated through Fig. 3, the top panel of which shows the free-energy as a function of the dihedral angles Φ and Ψ and is displayed for reference here in order to locate the free-energy minima. In the bottom panel the first DC is seen to change smoothly along the path between the pairs of minima, C_5–$P_{\|}$ and α_P–α_R, and corresponds to a transition between these two pairs. In addition, the first DC is well correlated with the empirical coordinate Ψ.

In Fig. 1, the fact that the gap between $\exp(-\lambda_2)$ and $\exp(-\lambda_3)$ is small compared to the gap between $\exp(-\lambda_1)$ and $\exp(-\lambda_2)$ shows that the second and third DCs describe motions on similar time scales. We analyze these motions in the supplementary material[29] and find that the second DC corresponds to diffusion from the $P_{\|}$ minimum to the C_5 minimum, while third DC to transitions between the α_P and α_R minima. Supplementary material[29] Fig. S5 shows the free-energy as a function of the first and third DCs; Supplementary material[29] Fig. S6 shows the figures analogous to Fig. 3 for the first (left) and third DCs (right).

Figure 4 displays the results of the local scale analysis for a representative configuration near a transition barrier (top) and free-energy minimum (bottom). For the configuration near a transition barrier, there are a few MDS singular values that are large and well separated from the remaining "noise" MDS singular values, while for the configuration near the minimum, the "data" and "noise" singular values are more closely spaced. This figure can be compared with Fig. 1 of Little *et al.*[27] to get a sense of the difference between MD datasets and high-dimensional noisy datasets generated by the addition of Gaussian white noise to lower dimensional datasets.

For the configuration near the minimum, the local intrinsic dimensionality of $\mathcal{M}$ is larger, and the locally linear region is smaller, than that for the configuration near the transition region; representative examples of this are shown in Fig. 4. This trend is apparent throughout the dataset as evidenced in Fig. 5, which displays the results of the local scale analysis for all of the data points. Figure 5 plots the local scale (top), and the intrinsic dimensionality of $\mathcal{M}$ (bottom) at each configuration in the dataset as a function of

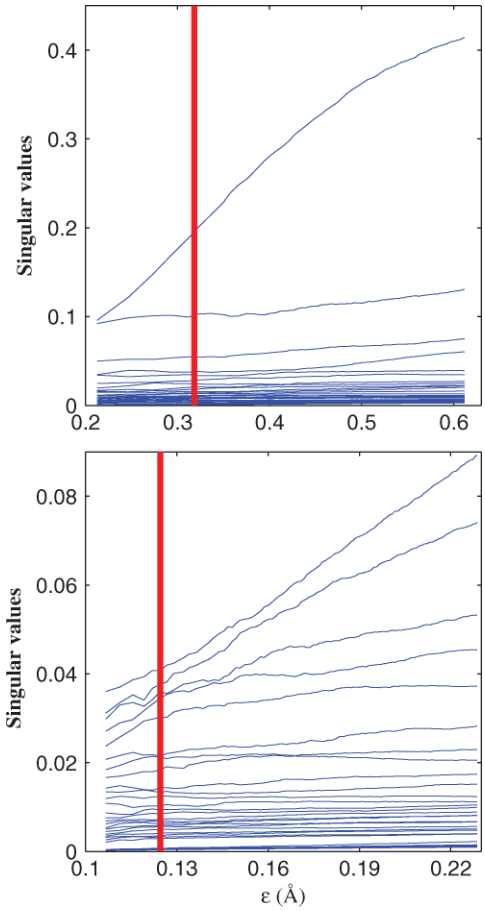

FIG. 4. MDS singular value spectra for configurations inside ε-balls around a configuration near a transition region (top), and near a free-energy minimum (bottom) for alanine dipeptide. The horizontal axis is the RMSD radius of each ε-ball in Å. For the top (bottom) panel the intrinsic dimension determined by our algorithm is 2 (8), and the red vertical line denotes the value of the estimated local scale ε_i. Note the differences in the scales of the axes between the two figures.

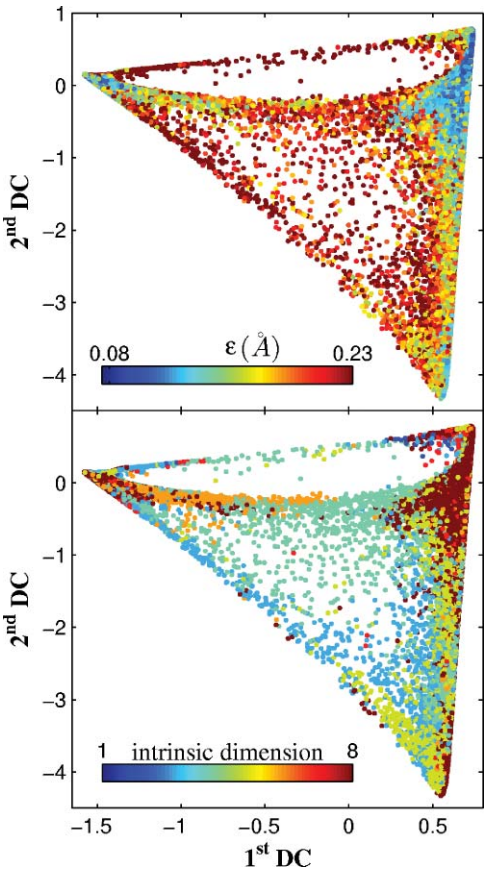

FIG. 5. Alanine dipeptide local scale analysis. Raw molecular configuration data plotted as a function of the first and second DCs, and colored by the local scale ε_i in Å (top), and the local intrinsic dimension (bottom). For visual clarity the following cutoffs have been imposed on the color bars: the maximum value of ε is 0.23 Å, and the maximum value of the intrinsic dimension is set at 8. There are a few outliers near transition regions that have local scales high above this cutoff. Most of the configurations in the minima have an intrinsic dimensionality in the range 8–24.

the first and second DCs. A comparison with the free-energy (Fig. 2) demonstrates that the "locally flat" region of $\mathcal{M}$ is smaller in length and of a higher intrinsic dimension near the free-energy minima compared to transition regions. This result is in accord with chemical intuition: the classic definition of a transition state is a state in which the energy is a maximum along one degree of freedom and a minimum along all other orthogonal degrees of freedom. Following the minimum energy path of a reaction, we expect the intrinsic dimension in such a state to be close to one.

We quantify the assertion that the first DC captures the essential dynamics of the isomerization process between C_5–$P_\parallel$ and α_R–α_P by calculating the Kramers' rate from Eq. (9) along both the first DC and Ψ. A comparison with the rate obtained directly from the simulation is reported in Table I. Both the rate along the first DC and Ψ are in excellent agreement with the rate extracted directly from simulation, suggesting that both the first DC and Ψ are good reaction coordinates for this transition. It is expected that these coordinates perform similarly, as they are strongly correlated.

We have performed these same calculations using a constant value of ε, and find that the diffusion coordinates emerging from such calculations depend strongly on the ε value chosen. Moreover it is *a priori* unknown which value provides meaningful results. These results are detailed in Sec. I of the supplementary material.[29]

B. SH3

The folding dynamics of the 57-residue protein domain SH3 have been well characterized both by simulation studies[34–36] and wet-lab experiments.[37] This protein is known to fold in a two-state manner, that is, only the unfolded or folded states are significantly populated near the folding transition temperature T_f. The folding/unfolding

TABLE I. Alanine dipeptide isomerization rates (ps^{-1})[a].

Coordinate	$C_5, P_\parallel \rightarrow \alpha_R\ \alpha_P$	$\alpha_R\ \alpha_P \rightarrow C_5, P_\parallel$
Direct simulation[b]	0.023	0.047
1st DC	0.023 ± 0.001	0.048 ± 0.003
Ψ	0.020 ± 0.001	0.040 ± 0.003

[a]See Appendix B for details on the error analysis.
[b]Standard deviation for simulation rates are of order 10^{-4} ps^{-1}.

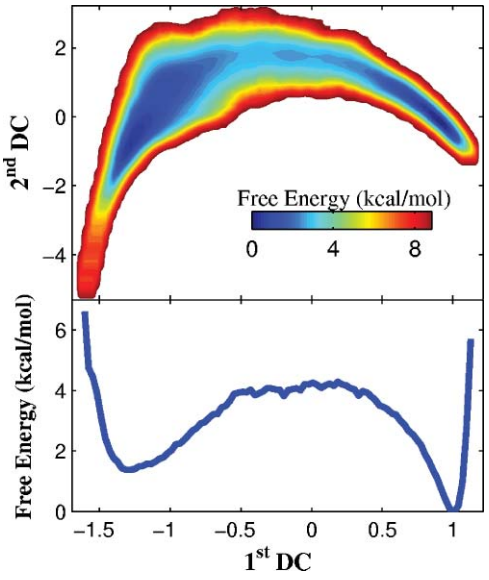

FIG. 6. (Top) Free energy of the coarse-grained SH3 model as a function of the first and second DCs. (Bottom) Free energy profile along the first DC. The Kramers' rate along the first DC is shown in Table II.

process is known to be the longest time scale and corresponds to diffusion over the free-energy barrier separating the folded and unfolded states. Previous studies have shown that the free-energy landscape of SH3 can be well approximated by a few global coordinates, either empirically defined[34,38] or obtained through nonlinear dimension reduction techniques.[17,39] As with alanine dipeptide, the previous work on this system makes it an ideal test case for our approach.

We apply our method to simulation data obtained with the coarse-grained DMC model of SH3.[40] MD simulations were performed in GROMACS (Ref. 41) near T_f, and configurations collected every 5 ps during a 500 000 ps run. As with alanine, the trajectory used here is at least ten times longer than needed to obtain reliable results; a similar free-energy landscape is obtained with only 10 000 configurations.

The large spectral gap in the Fokker–Planck eigenspectrum (Fig. 1) between $\exp(-\lambda_1)$ and $\exp(-\lambda_2)$, denoted by the vertical bar, shows that the time scale between the collective motion corresponding to that of the first DC is much longer than that of the second DC. Figure 6 displays the free-energy as a function of the first and second DCs in the top panel and the profile along the first DC in the bottom panel. The minimum on the right corresponds to the folded state, and the minimum on the left is the minimum of the unfolded state. From this we see that the first DC separates the folded and unfolded minima, while the second DC seems to be related to the motion toward the transition region.

In order to relate the diffusion coordinates to quantities that have a direct physical interpretation, in Fig. 7 we compare the first DC with two empirical coordinates: RMSD to the native structure and the fraction of native contacts, Q. The top panel shows the free energy in terms of these coordinates, presented for reference to locate the free-energy minima. The lower right minimum is the folded minimum (with a high fraction of native contacts and a small RMSD to the native structure), and the upper left minimum is the unfolded minimum.

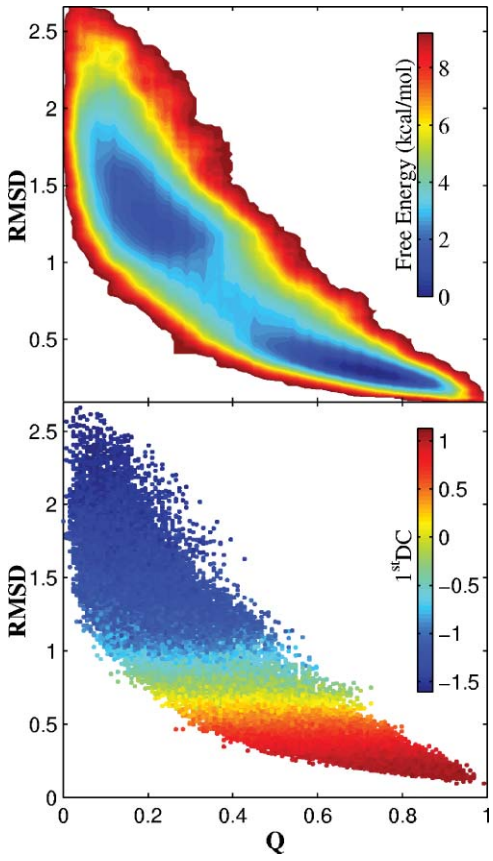

FIG. 7. (Top) Free energy (kcal/mol) as a function of the empirical coordinates RMSD to the native structure (Å) and fraction of native contacts Q. (Bottom) Raw molecular configuration data points, plotted in terms of the same empirical coordinates, and colored according to the first DC. The smooth color transition along the first DC shows qualitatively that the first DC and the RMSD to the native structure are correlated.

The bottom panel displays the first DC as a function of these same two coordinates, and a comparison between the panels shows that the first DC corresponds to the folding/unfolding transition and correlates well with the RMSD to the native structure.

For protein systems, it is also possible (and illuminating when there are no empirical coordinates available) to examine the relationship between the diffusion coordinates and contact probabilities. We calculate the probability of contact formation along the first and second DCs for all nonbonded contacts of SH3, and from that calculate the correlation between the contact probabilities and first DC (second DC) in the lower (upper) triangle of Fig. 9. Both the Spearman rank correlation ρ (which measures the monotonicity of the correlation: $\rho = 1$ for a perfect monotonic relationship) and the Pearson correlation coefficient r (which measures the linearity of the relationship) are considered. Only contacts with both the Spearman and Pearson correlations greater than 0.8, and a probability of formation greater than 0.1 are included in the figure.

The contacts shown in blue tend to form as the first DC increases. These are mostly the native contacts, confirming that the motion along the first DC corresponds to the folding process. The contacts in red tend to form as the second DC increases. Interestingly, these contacts include the set of *non-native* contacts involved in the formation of a nonspecific

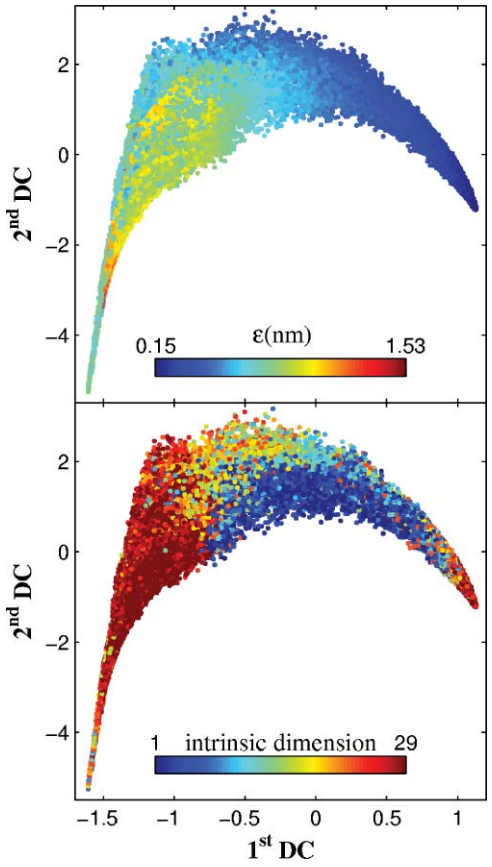

FIG. 8. SH3 local scale analysis. Raw molecular configuration data plotted as a function of the first and second DCs, colored according to the local scale ε_i in nm (top) and the local intrinsic dimension (bottom).

hydrophobic nucleus (circled in red in the figure). This is the only set of non-native contacts with both the Spearman rank and Pearson correlation coefficients larger than 0.9. Both experimental[42] and simulation (see Fig. 4 of Das *et al.*[40]) results suggest these contacts to be important in the folding mechanism of SH3.

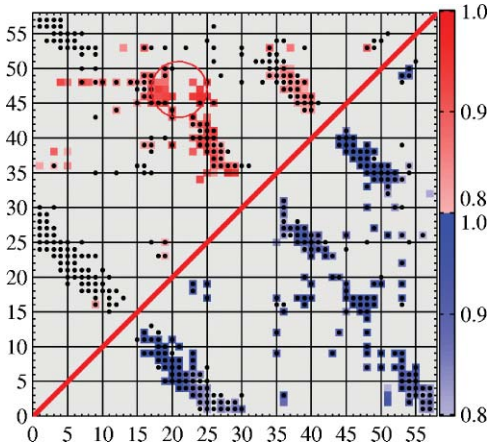

FIG. 9. Correlation of SH3 probability of contact formation with the first (lower right) and second DCs (upper left). Native contacts are marked by a black dot. Different shades of red or blue indicate different values of the Pearson correlation coefficient, as indicated in the colorscale on the right.

TABLE II. SH3 folding rates (ps^{-1})[a].

Coordinate	Folded → Unfolded	Unfolded → Folded
Direct simulation	$(4.4 \pm 0.4) \times 10^{-5}$	$(6.4 \pm 0.4) \times 10^{-5}$
1st DC	$(6.9 \pm 0.8) \times 10^{-5}$	$(8.0 \pm 0.9) \times 10^{-5}$
RMSD[b]	$(9 \pm 1) \times 10^{-5}$	$(12 \pm 1) \times 10^{-5}$
Q	$(3.1 \pm 0.5) \times 10^{-4}$	$(3.7 \pm 0.7) \times 10^{-4}$

[a]See Appendix B for details on error analysis.
[b]RMSD with respect to the native structure.

The analog of Fig. 7 for the second DC shows that it has an extremum at the transition state and decreases when moving to either of the minima (Fig. S7 of the supplementary material[29]). Taken together with the results of the correlation analysis, these results indicate that the second slowest time scale in the folding of SH3 corresponds to the formation of a folding nucleus involving this set of non-native contacts, which are formed at the transition state, but not formed in the unfolded and folded states.

The results of the local scale analysis are displayed in Fig. 8. It is interesting to compare and contrast these results with those of alanine dipeptide. For SH3 the local scale is small and the local intrinsic dimension of $\mathcal{M}$ is larger around configurations in the folded minimum in comparison with configurations around the transition region—a result similar to that of alanine dipeptide (see Fig. 5). However for configurations in the unfolded minimum, the local intrinsic dimension is high, while the local scale is large when compared to configurations in the transition region. While for alanine dipeptide configurations in both free-energy minima have small local scale, for SH3 there is a clear difference between the folded, *potential* energy minimum and the *entropic* minimum of the unfolded states. This difference is expected by considering that the structures in the unfolded minimum are far apart from one another in terms of RMSD distance, while in a potential energy minima the structures are more similar to one another.

The rates obtained by using Kramers' escape rate expression[30] along a reaction coordinate are presented in Table II; the rates along three different coordinates: the first DC, the RMSD to the native structure, and the fraction of native contacts Q, are compared with the rate obtained directly from simulation. The rate estimate along the first DC is more accurate than the rate along both RMSD and Q, confirming that the first DC better describes the overall folding/unfolding motion of this SH3 model than empirically defined reaction coordinates.

V. CONCLUSIONS

We present a multiscale, mathematically justified approach for extracting collective coordinates from a configurational sample of macromolecular motion. This method provides not only global reaction coordinates and a free-energy landscape but also information about the geometrical structure of the configuration space. In addition, no prior estimation of prospective reaction coordinates and/or definition of reactant/product states is required. The approach is based on the determination of the length scale at which the

dynamics can be considered locally linear at each point in the configuration space; this position-dependent length scale is then used to locally "renormalize" the kernel of the transition probability between each pair of configurations. A diffusion map is then constructed on the global diffusion process.

For systems with a separation of time scales in which the slowest time is associated with the diffusion over a free-energy barrier, the first diffusion coordinate is a good reaction coordinate. Reaction rates computed by using Kramers' rate expression along the first diffusion coordinate are in remarkable agreement with the rates measured directly from simulation data.

The analysis of the correlation of the first few diffusion coordinates with collective variables such as empirical reaction coordinates (if available), and/or the probability of contact formation allows for a physical understanding of the collective motions corresponding to the diffusion coordinates. Through such an analysis of coarse-grained SH3, we find the slowest time scale corresponds to the folding/unfolding transition, and the second slowest time scale corresponds to the formation of a set of non-native contacts at the core of the protein.

The local scale analysis at the base of the LSDMap approach provides insight into the local intrinsic dimensionality of the molecular configurational space, which can be used to approximately locate free-energy minima and transition regions, and gain some understanding of the nature of these regions.

At present, the method is only applicable to systems for which it is possible to obtain Boltzmann-sampled data. We are currently working to extend the method to non-Boltzmann-distributed data, such as that from biased molecular dynamics runs.

To the best of our knowledge, this is the first time mathematical techniques in multiscale geometric theory have been extended to the analysis of macromolecular dynamics data; this is a first step in the direction of quantifying and exploiting geometric properties of trajectories arising from MD simulations. We believe the approach presented provides a powerful tool to understand the collective processes in complex diffusion reactions over a spectrum of different time and length scales.

APPENDIX A: LOCAL SCALE DETERMINATION

To obtain information about the local geometry around a configuration $\boldsymbol{x}_i$, we perform multidimensional scaling on increasingly larger neighborhoods (ε-balls) around $\boldsymbol{x}_i$. We normalize these singular values by the square root of the number of configurations within the corresponding ε-ball. In order to numerically distinguish between the non-noise and noise MDS singular values, we calculate the gap between each pair of consecutive singular values at three locations along the range of values of ε considered: 3/7, 1/2, and 4/7 of the largest value. The reason for performing the analysis at three distinct points is to ensure robustness of the results. To analyze the gaps between the singular values, we construct a "status vector" as follows. The first entry corresponds to the gap between the largest and second largest singular values, the second entry to the gap between the second and third largest, etc. The entry for each pair of consecutive singular values is "1" if the value of the gap for that pair is greater than twice the value of each of the following five gaps at any of the three fractions considered; the entry is "0" otherwise. The separation between the non-noise and noise singular values is defined to be between the first pair whose entry is "1" and with the following three entries equal to "0". This analysis is similar in spirit to the one proposed in Little *et al.*,[27] and provides robust results.

The next step is to determine the local scale ε_i. A conservative estimate is to define ε_i as the length scale at which the noise singular values begin to decrease and clearly separate from the non-noise ones. To numerically find these lengths, we fit the noise spectra to a low-order polynomial (decreasing the order of the polynomial if the Vandermonde matrix is ill conditioned) and calculate the derivatives of the noise spectra as a function of the neighborhood size ε from the fit. We then scan from smallest to largest ε, and define ε_i as the value at which the first derivative of *each* noise singular value is less than a given cutoff (0.03 for alanine dipeptide, 0.04 for SH3—the results are robust against variations of these parameters). If no such value of ε is found, we define the noise more conservatively by considering the highest noise singular value to belong with the data and repeat the scan.

APPENDIX B: BAYESIAN DETERMINATION OF DIFFUSION COEFFICIENTS AND KRAMERS RATE ERROR ANALYSIS

We determine the diffusion coefficients along the various coordinates (first DC and Ψ for alanine dipeptide; first DC, RMSD, and Q for SH3) using Bayesian analysis.[32] The method is based on the fact that through Bayes inference theorem, the probability distribution of position-dependent diffusion coefficients giving rise to a trajectory is proportional to the probability of observing the same trajectory for given values of the diffusion coefficients. In order to obtain a likelihood function associated with the MD data, for a given choice of the collective coordinate X, the range of values spanned by X is discretized in m cells X_i, $i = 1, \ldots, m$. The information associated with a given MD trajectory (or a set of many short MD trajectories) is then translated into the number of transitions N_{ij} between cells i and j observed in a time t_α. Assuming Markovian dynamics, the position-dependent diffusion coefficient $D_{i+1/2} = D(\frac{1}{2}(X_i + X_{i+1}))$ at the boundary between two consecutive cells i and $i+1$ can be defined in terms of the rate matrix R as

$$D_{i+1/2} = |X_{i+1} - X_i|^2 \sqrt{R_{i,i+1} R_{i+1,i}}. \tag{B1}$$

The likelihood function L associated with the observed N_{ij} for a given rate matrix R is

$$\ln L = \sum_{i=1}^{m} \sum_{j=1}^{m} N_{ij} \ln (e^{t_\alpha R})_{ij}. \tag{B2}$$

The rate matrix R (and the corresponding diffusion coefficients) is then determined by performing a Metropolis Monte Carlo simulation in the space of the matrix elements

R_{ij} in which the negative log-likelihood is used as an "energy function."[43] The resulting distribution for R_{ij} is sharply peaked around the most probable values of the matrix elements, which are then used to determine the diffusion coefficients. For each system we use a long MD trajectory to obtain the N_{ij} matrix elements for the likelihood function.

In order to ensure smoothness in the diffusion coefficients we use the prior in the form,

$$\prod_i \exp\left\{\frac{-[D(X_i) - D(X_{i+1})]^2}{2\gamma^2}\right\}, \tag{B3}$$

as proposed in Eq. (14) of Hummer's work.[32] The values of the γ parameter used for alanine dipeptide are 0.1 for both the first DC and Ψ; and for SH3 are 0.0001 for the first DC, and 0.00005 for both RMSD and Q. Some caution must be used in choosing the values of γ. Too large a value may produce large spikes in the diffusion coefficients in the slightly less sampled regions; a value too small may artificially flatten the diffusion coefficient profile along the coordinate. For all the coordinates considered, a range of values of γ consistently reproducing the same diffusion coefficients can be defined. Our results are robust against variations of the parameter γ around the values reported above.

The other parameters to be chosen for the Bayesian analysis are the observation times t_α and the number of cells m along a collective coordinate. We used several sets of these parameters and calculated the Kramers' integrals from each set of diffusion coefficients. The final rate values in Tables I and II of the main text are a result of averaging over the following sets of t_α and numbers of cells. For alanine dipeptide, along the first DC: $t_\alpha = 0.5$, 0.6, 0.7, and 0.8 ps, with 20, 24, 28, and 32 cells; along Ψ: $t_\alpha = 0.5$ and 0.6 ps, with 24, 28, 32, and 36 cells. For SH3, along the first DC: $t_\alpha = 60$ and 70 ps, with 16, 24, 36, and 48 cells; along RMSD: $t_\alpha = 60$ and 70 ps, with 16, 24, 36, and 48 cells; along Q: $t_\alpha = 60$ and 70 ps, with 16, 24, 36, and 48 cells. These choices for t_α and the numbers of cells m were motivated by the need to obtain an N_{ij} matrix with transitions between each pair of neighboring cells. Very long t_α values result in poor sampling of the transition between cells along the top of the barrier; very short t_α values result in the observation of too few transitions from the cells near free-energy minima. In the supplementary material,[29] Figs. S8 and S9 show the rates of the various transitions for each of the choices of t_α and m.

The errors reported for each rate in Tables I and II are calculated as follows. The Bayesian analysis produces diffusion coefficients evaluated on the cell edges as well as their standard deviations. These standard deviations are the largest errors in the calculation and are propagated through the numerical evaluation of the Kramers' integral to yield an estimate of the error for the Kramers' rate calculated for each pair of parameters t_α and m. These are the error bars shown in Figs. S7 and S8 of the supplementary material.[29] The errors reported in Tables I and II are the average of the errors for each pair of t_α and m that were used in determining the Kramers' rate.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation (NSF) (CDI-type I Grant No. 0835824 to C.C. and Grant No. 0835712 to M.M., NSF CAREER Award No. CHE-0349303 to C.C., and NSF CAREER Award No. DMS-0650413 to M.M.), the Welch Foundation (C-1570 to C.C.), and the Sloan Foundation (to M.M.). Simulations and other computations were performed on the following shared resources at Rice University: the Rice Computational Research Clusters funded by NSF under Grant No. CNS-0421109 and in partnership between Rice University, AMD and Cray; the Cyberinfrastructure for Computational Research funded by NSF under Grant No. CNS-0821727; the Shared University Grid at Rice University funded by NSF under Grant No. EIA-0216467 and in partnership between Rice University, Sun Microsystems, and Sigma Solutions, Inc.; and a 2010 IBM Shared University Research (SUR) Award on IBM's Power7 high performance cluster (BlueBioU) to Rice University as part of IBM's Smarter Planet Initiatives in Life Science/Healthcare and in collaboration with the Texas Medical Center partners, with additional contributions from IBM, CISCO, Qlogic, and Adaptive Computing. We thank Paul Ledbetter for useful discussions.

[1] B. Qi, S. Muff, A. Caflisch, and A. R. Dinner, J. Phys. Chem. B **114**, 6979 (2010).
[2] R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **107**, 1088 (2010).
[3] S. Yang, J. N. Onuchic, and H. Levine, J. Chem. Phys. **125**, 054910 (2006).
[4] A. Berezhkovskii and A. Szabo, J. Chem. Phys. **122**, 014503 (2005).
[5] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[6] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).
[7] R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **102**, 6732 (2005).
[8] W. E, W. Ren and E. Vanden-Eijnden, J. Chem. Phys. **126**, 164103 (2007).
[9] L. Maragliano and E. Vanden-Eijnden, Chem. Phys. Lett. **446**, 182 (2007).
[10] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
[11] A. K. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).
[12] I. T. Jolliffe, *Principal Components Analysis* (Springer, Berlin, 1986).
[13] P. H. Nguyen, Proteins **65**, 898 (2006).
[14] Y. Mu, P. H. Nguyen, and G. Stock, Proteins **58**, 45 (2005).
[15] S. T. Roweis and L. K. Saul, Science **290**, 2323 (2000).
[16] J. B. Tenenbaum, V. de Silva, and J. C. Langford, Science **290**, 2319 (2000).
[17] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **103**, 9885 (2006).
[18] Y. Yao, J. Sun, X. Huang, G. R. Bowman, G. Singh, M. Lesnick, L. J. Guibas, V. S. Pande, G. Carlsson, J. Chem. Phys. **130**, 144115 (2009).
[19] R. R. Coifman and S. Lafon, Appl. Comput. Harmon. Anal. **21**, 5 (2006).
[20] R. R. Coifman and M. Maggioni, Appl. Comput. Harmon. Anal. **21**, 53 (2006).
[21] R. R. Coifman *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 7426 (2005).
[22] R. R. Coifman *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 7432 (2005).
[23] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Multiscale Model. Simul. **7**, 842 (2008).
[24] A. D. Szlam, M. Maggioni, and R. R. Coifman, J. Mach. Learn. Res. **9**, 1711 (2008).
[25] S. Mahadevan and M. Maggioni, J. Mach. Learn. Res. **8**, 2169 (2007).
[26] P. W. Jones, M. Maggioni, and R. Schul, Proc. Nat. Acad. Sci. U.S.A. **105**, 1803 (2008).
[27] A. Little, Y.-M. Jung, and M. Maggioni, *Multiscale Estimation of Intrinsic Dimensionality of Data Sets*, AAAI Fall Symposium Series (November 5-7, 2009, Arlington, Virginia, USA). Available online: http://aaai.org/ocs/index.php/FSS/FSS09/paper/view/950.

[28]The functions $\phi_i(\boldsymbol{x})/\phi_0(\boldsymbol{x})$ are eigenfunctions of the backward Fokker–Plank operator. The forward and backward Fokker–Planck operators are adjoint to one another, share the same set of eigenvalues, and their eigenfunctions differ by a factor of $\phi_0(\boldsymbol{x})$.
[29]See supplementary material at http://dx.doi.org/10.1063/1.3569857 for more details.
[30]H. A. Kramers, Physica **7**, 284 (1940).
[31]L. Huang and D. E. Makarov, J. Chem. Phys. **128**, 114903 (2008).
[32]G. Hummer, New J. Phys. **7**, 34 (2005).
[33]D. A. Case, T. A. Darden, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, D. A. Pearlman, M. Crowley, R. C. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D. H. Mathews, C. Schafmeister, W. S. Ross, and P. A. Kollman, AMBER9 University of California, San Francisco, 2006.
[34]C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).
[35]C. Clementi and S. S. Plotkin, Protein Sci. **13**, 1750 (2004).
[36]L. Li, L. A. Mirny, and E. I. Shakhnovich, Nat. Struct. Biol. **7**, 336 (2000).
[37]V. P. Grantcharova, D. S. Riddle, J. V. Santiago, and D. Baker, Nat. Struct. Biol. **5**, 714 (1998).
[38]S. S. Cho, Y. Levy, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **103**, 586 (2006).
[39]E. Plaku, H. Stamati, C. Clementi, and L. E. Kavraki, Proteins **67**, 897 (2007).
[40]P. Das, S. Matysiak, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **102**, 10141 (2005).
[41]D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comput. Chem. **26**, 1701 (2005).
[42]A. R. Viguera, C. Vega, and L. Serrano, Proc. Natl. Acad. Sci. U.S.A. **99**, 5349 (2002).
[43]As in Hummer's work, we also perform Metropolis Monte Carlo on P_i, the probability of being in the i^{th} cell. The elements of R and P are related through detailed balance: $R_{i+1,i}/R_{i,i+1} = P_{i+1}/P_i$. While the Bayesian analysis method provides estimates of the free-energy, we use the more finely sampled free-energy directly from our simulation for the evaluation of the Kramers' integrals.