Discovering Mountain Passes via Torchlight: Methods for the Definition of Reaction Coordinates and Pathways in Complex Macromolecular Reactions

Mary A. Rohrdanz, Wenwei Zheng, and Cecilia Clementi

Department of Chemistry, Rice University, Houston, Texas 77005; email: cecilia@rice.edu

Annu. Rev. Phys. Chem. 2013. 64:295–316

First published online as a Review in Advance on January 4, 2013

The *Annual Review of Physical Chemistry* is online at physchem.annualreviews.org

This article's doi: 10.1146/annurev-physchem-040412-110006

Keywords

path sampling, reaction rate, machine learning, molecular dynamics

Abstract

The long-timescale dynamics of macromolecular systems can be oftentimes viewed as a reaction connecting metastable states of the system. In the past decade, various approaches have been developed to discover the collective motions associated with this dynamics. The corresponding collective variables are used in many applications, e.g., to understand the reaction mechanism, to quantify the system's free energy landscape, to enhance the sampling of the reaction path, and to determine the reaction rate. In this review we focus on a number of key developments in this field, providing an overview of several methods along with their relative regimes of applicability.

1. INTRODUCTION

Reaction coordinate: collective variable used in the quantitative understanding of macromolecular motion

Order parameter: collective variable used in the qualitative understanding of macromolecular motion

MD: molecular dynamics

Understanding complex molecular reactions is paramount in chemistry. To quantify the progress of a reaction process from the reactant to the product state, one often invokes reaction coordinates. The idea is to find global coordinates that can aid in understanding how the system proceeds from one state to the other; ideally these coordinates have a physicochemical interpretation, and their evolution in time can shed light on the reaction mechanism. In this review, we focus our attention on the characterization of reactions involving configurational rearrangements of large macromolecules (e.g., proteins) that can be described in terms of atomic motions.

Many of the key ideas for determining and using reaction coordinates were initially formulated several decades ago (see, e.g., 1–4). However, with the developments in computer hardware and software of the past decade, a great amount of work has been recently done to define theoretically robust and computationally practical methods for extracting and using reaction coordinates in simulations.

Reaction coordinates have a variety of uses, e.g., to obtain a qualitative model of atomic motion during the course of a reaction or to provide the basis for a quantitative calculation of reaction rates. The performance of a reaction coordinate depends on the function being queried. A coordinate that is good for visualizing the progress of a reaction might not be best for using, e.g., classical transition-state theory to calculate the rate. In this context, the distinction is sometimes made between an order parameter and a reaction coordinate. An order parameter is usually thought of as a rough variable used to gauge the progress of a reaction and is often based on physical or chemical intuition. It is sometimes needed as input for more quantitative methods, and is discussed further below. Conversely, the term reaction coordinate is often reserved for a collective coordinate that can be used for quantitative rate calculations. As such, it has meaning only when coupled with a corresponding reaction-rate theory. The dynamics of large macromolecular systems is often modeled as an overdamped diffusion, and in this regime Kramers' reaction-rate theory can be used to extract a reaction rate and mechanism when free energy as a function of the appropriate reaction coordinates is available. We maintain the distinction between order parameters and reaction coordinates throughout this review; when we refer to properties shared by both, we use the terms collective variables or collective coordinates.

Traditionally, collective variables have been used to interpret molecular simulations once an exhaustive sampling of the system of interest has been obtained. More recently, such coordinates have started to play an additional and increasingly important role: They are often used in molecular dynamics (MD) simulations as a guide to sample the configurational space relevant to the reaction more efficiently. In principle, the exploration of a complex free energy landscape directly in terms of the appropriate reaction coordinates (possibly partitioned in terms of the timescales and length scales involved) could present significant advantages with respect to the application of standard MD. Let us consider, for instance, a system with a clear separation of timescales, in which one rare event (such as the crossing of a large free energy barrier) dominates the long-timescale dynamics. Let us assume that the optimal reaction coordinate associated with this process is known a priori. Then this information could be used to properly sample along the reaction coordinate, and a free energy map could be calculated efficiently. Many methods discussed below could be straightforwardly used for this purpose. However, the definition of the optimal reaction coordinate(s) is oftentimes highly nontrivial for complex macromolecular systems, and little can be inferred a priori. However, methods have been developed to define reaction coordinates and test their validity a posteriori, once an equilibrium sampling of the system is available. The methods discussed in this review present different approaches to this chicken-and-egg problem. To the best of our knowledge, no method has been proposed that solves the problem entirely. Different

approaches are better suited for different applications, and we try to highlight their advantages and limitations.

We discuss methods designed to use predefined sets of plausible collective coordinates to guide the sampling, methods to define the reaction coordinates starting from a sampling, and approaches in which the definition and calculation of reaction coordinates and reaction rates become intimately entangled with the sampling itself. A main advantage of the coupling of sampling and analysis is that it often allows one to simulate a large collection of short trajectories instead of one or a few extremely long trajectories visiting all the relevant space. In practice, methods based on these ideas require some initial coarse sampling to define at least an approximate partitioning of the configurational space. Information can then be extracted from the local sampling and used to refine or further direct the search of the conformational space. The generation of the initial sample itself is sometimes the main limitation of such methods.

RMSD: root-mean-square deviation

As a starting point, we have divided the methods into a few broad and somewhat overlapping categories. We first briefly discuss methods based on the physical and chemical intuition on the system, and the idea of the isocommittor, which is often underlining other approaches. We then review methods based on the definition of pathways connecting a priori known reactant and product states; next we discuss methods that invoke the partitioning of the system's phase space into multiple substates. We then examine methods based on machine-learning approaches for the definition of reaction coordinates. Ideally, machine-learning methods do not require any a priori knowledge on the properties of the system under investigation; however, in their present implementation, they require an existing sampling of the configurational space. Finally, we briefly discuss a few techniques that invoke collective coordinates to enhance sampling.

We stress that what follows is not an all-encompassing cataloging of existing methods, as we have chosen to highlight a subset of approaches based on our particular field of expertise and interest. In particular, we focus this review on methods developed on and applicable to biomolecular systems. Some of these methods are more generally applicable, and different applications are mentioned.

2. INTUITION-BASED COORDINATES AND THE ISOCOMMITTOR

For simple systems, reaction coordinates can be chosen based on physical intuition. For example, the various isomerizations of the alanine dipeptide can be understood in terms of the dihedral angles ϕ and ψ, and ψ can be used in rate calculations (5). For minimally frustrated protein systems, coordinates such as Q (the fraction of native contacts) and the root-mean-square deviation (RMSD) from the native state can function as good reaction coordinates (6, 7). Such coordinates have the advantage of being straightforwardly connected to physically meaningful quantities. However, for complex systems, it can be challenging to find a reasonable coordinate, and even for deceivingly simple systems, a seemingly good intuitive coordinate may lead to inaccurate rate calculations (8, 9).

Another popular coordinate used for the characterization of macromolecular reactions is the isocommittor coordinate (10), the notion of which dates back to at least the 1930s (11, 12). For molecular systems with two metastable minima, A (reactant) and B (product), the isocommittor surfaces, or isosurfaces, are hypersurfaces in configuration space that span the region between A and B, do not intersect one another, and have the property that trajectories initiated on a given hypersurface have the same probability of reaching state B before state A. The value of the isocommittor coordinate ranges continuously from 0, on the hypersurface bounding region A, to 1, on the hypersurface bounding region B. The transition-state ensemble is identified with the hypersurface in which the value of the isocommittor is 0.5 (12). The isocommittor coordinate is variably referred to as the transmission coefficient (10) [not to be confused with the transmission

coefficient from transition-state theory (13)], the committor (14), and P_{fold} (15) in protein-folding studies.

TPS: transition-path sampling

TIS: transition interface sampling

PPTIS: partial path transition interface sampling

It has been stated that the isocommittor is the best reaction coordinate for a system with only two free energy minima (16), as it groups together configurations of the system that have the same probability of proceeding toward the product state. Conversely, it has also been argued that its usefulness is limited because of heavy computational requirement (17), applicability only if no intermediate states are present between reactants and products (although it is always possible to define an isocommittor between each relevant pair of states), sensitivity to temperature, and the lack of direct relationships to experimental observables (15).

The committor function is ubiquitous in the methods described in this review, both as actual steps in various algorithms and as checks of the resulting reaction coordinates for reactions limited to two metastable states. Additional examples of approaches that invoke the isocommittor include methods to discover reaction coordinates based on neural networks (18) and methods that examine the reactive flux through a network (19).

3. PATH-BASED METHODS

The path-based approaches discussed in this section are focused on providing information on the dynamics of the system through various types of enhanced sampling of transitions between the system's metastable states. For example, transition-path sampling (TPS) (Section 3.1) begins with a single reactive trajectory connecting the reactant and product, from which it generates a large transition-path ensemble. Transition interface sampling (TIS) and partial path TIS (PPTIS) build on this foundation to improve computational efficiency and are more applicable to high-friction or highly diffusive systems.

3.1. Transition-Path Sampling

Starting from a single reactive transition path, TPS performs a random walk in trajectory space to collect a set of reactive paths (20). From this transition-path ensemble, one can estimate information about the reaction mechanism, transition states, and rate constants (14). Some of the first attempts to statistically analyze collections of reactive paths were proposed in the 1970s and 1980s (12, 21), and a precursor of the modern formulation of TPS (22, 23), based on stochastic path integrals, was presented in the mid-1990s (24). TPS has been applied to a large range of systems, from the study of the rearrangement of Lennard-Jones clusters (25) to chemical reactions (26). Methods that use TPS data to rank putative reaction coordinates have also been developed (see the sidebar Reaction Coordinates from Path Sampling).

REACTION COORDINATES FROM PATH SAMPLING

Taking advantage of the data collected by TPS, Best & Hummer (124) applied Bayesian interference to identify transition regions as those regions in configuration space such that the trajectories passing through them are most likely reactive. Similar ideas are used to identify the best reaction coordinate from a linear combination of candidate ones. Peters & Trout (125) developed a related approach that uses likelihood maximization to find the best reaction coordinate from a set of collective variables. Both methods quantify the quality of reaction coordinates for a specific purpose; a similar goal is pursued in the work of Ma & Dinner (18), who used the isocommittor, instead of the configuration's probability of being on the transition path, to test putative reaction coordinates.

Let us consider a system with two metastable states, a reactant state, A, and product state, B, and imagine one set of molecular configurations along a dynamical path between A and B. From this initial trajectory, new ones can be generated in several ways. For example, the coordinates of one point along the trajectory can be shifted by a small amount, and a new trajectory can be obtained by running MD both forward toward the B state and backward toward the A state. Based on a Monte Carlo criterion, the new path is then either accepted as a member of the transition-path ensemble or rejected. This procedure is repeated—the new path becoming the old one—to generate an ensemble of transition paths. With the help of an order parameter λ, one can statistically analyze the transition paths to calculate the transition rate from A to B (see 14 for details).

The key drawback for the application of this method to biological systems is that the time for a trajectory to move from one state to another in such systems can be quite long because of high friction, high barriers, and a very rough free energy landscape. As such, the simulation times can become prohibitively expensive (14). In addition, the method is designed for systems with only two minima separated by a high free energy barrier, whereas biological systems often have several metastable free energy minima with complex pathways between them. This limitation can sometimes be circumvented by analyzing each pair of minima separately (27).

Several methods have been developed, based on TPS, to overcome these limitations, including TIS and PPTIS, which are discussed in the next section. In addition, a transition-path theory has been developed (28) that makes a connection to string methods (see Section 4.1).

3.2. Transition Interface Sampling and Partial Path Transition Interface Sampling

TIS grew out of TPS with the goal of improved numerical convergence and computational efficiency. In TIS, an order parameter λ is used to divide the space between reactants A and products B. The rate is expressed in terms of the effective positive flux (i.e., recrossings are ignored) through hyperplanes defined by the values of the order parameter. This division of the region between A and B has the effect of avoiding the need for the simulation of full reactive paths, thereby making the method more applicable to biological systems.

Reference 29, figure 3, presents a schematic of the hypersurfaces along an order parameter λ. The reaction-rate constant can be expressed as

$$k_{AB} = \frac{\langle \Phi_{A,\lambda_1} \rangle}{\langle h_A \rangle} \prod_{i=1}^{n-1} \mathcal{P}(\lambda_{i+1}|\lambda_i) \mathcal{P}(\lambda_B|\lambda_n). \quad (1)$$

Here h_A represents all the configurations that were in state A more recently then they were in state B (i.e., all the configurations in state A, and all the phase-space points that, if integrated backward in time, would reach state A before state B). The steady-state flux of trajectories $\langle \Phi_{A,\lambda_1} \rangle$ from state A to λ_1 (that is, the fraction of trajectories from A that reach the surface associated with the order parameter value λ_1) can be estimated via MD simulation. The conditional probabilities $\mathcal{P}(\lambda_i|\lambda_j)$ represent the probability of a trajectory reaching interface λ_i given that it came from state A and passed through λ_j. In the evaluation of the various $\mathcal{P}(\lambda_i|\lambda_j)$, trajectories are propagated both backward in time to A and forward in time to reach either λ_j (these trajectories contribute to the forward flux) or A (these trajectories do not contribute to the forward flux).

For dynamics with high friction or diffusivity, which includes most macromolecular systems, the sampling of these trajectories can be inefficient. For this reason, PPTIS was developed (30). In this method, memory loss between interfaces is assumed, and simulations need only be run

between sets of three consecutive interfaces. This method can also provide the free energy along the order parameter at no extra cost (31).

ZTS: zero-temperature string method

A number of improvements have been made to these methods in the past few years. For systems with multiple paths, TIS and PPTIS may have difficulty obtaining a good sampling of the various pathways. Path-swapping techniques, modifications to the paths' generation algorithm, and the use of time as an order parameter have been proposed to address this issue (32). The inclusion of multiple metastable states with TPS and TIS has also been proposed (33). A recent publication, with applications to small proteins, provides practical information on the use of these methods and makes connections to isocommittor surfaces (34).

TPS-like methods have recently been applied to characterize a wide variety of rare-event processes, including reaction networks (35), simulations of Trp-cage folding transitions (36) (which compare favorably to experiment), homopolymer folding dynamics (37), and nucleation of hard spheres (38) (for which application/comparison to experiment is still in progress). In addition, techniques for nonequilibrium systems have been developed (see the sidebar Forward-Flux Sampling), and recent advances allow the application to nonstationary systems (39).

4. STATE-BASED METHODS

The methods discussed in this section place emphasis on the partitioning of the relevant configurational space into discrete states, rather than on reaction paths. Instead of relying on the statistical analysis of long trajectories, these methods combine information from shorter simulations between various locations in phase space to obtain global information such as reaction paths and rates. As such, they may be more advantageous in systems with many metastable minima and multiple paths.

4.1. String Methods

String methods seek to find reaction paths and rates by constructing a string of configurations along a path connecting reactants and products, and then moving those configurations toward the minimum energy path (40–43). A recent review of these techniques, including the corresponding theory of transition paths (44), presents the details of our sketch below (28).

String methods can be developed in terms of isocommittor surfaces (see Section 2). Let us consider, for example, a set of isocommittor surfaces between the reactants A and products B. The zero-temperature string (ZTS) method obtains the minimum energy path from A to B by finding the curve through the isosurfaces with the maximal flux of transition current. The

FORWARD-FLUX SAMPLING

Forward-flux sampling (FFS) was developed for nonequilibrium systems (126–128), for which methods such as TPS and TIS are not applicable (for a recent review, see 129). The fundamental setup for FFS is the same as TIS, and the equation for the rate (Equation 1) is identical. The difference is in how the conditional probabilities $\mathcal{P}(\lambda_{i+1}|\lambda_i)$ are calculated. In TIS, the trajectories initiated on each interface are integrated both forward and backward in time. However, in FFS, the trajectories initiated at each interface are integrated only forward in time. The final point of each trajectory initiated at λ_i that reaches λ_{i+1} is stored, and sets of new trajectories are initiated from these successful configurations. Without the need to run trajectories backward to the initial state, the computational demand is less than that of TIS. However, the end result is susceptible to errors in the initial sampling of the reactant distribution and a poor choice of order parameter.

finite-temperature string (FTS) method, alternatively, finds the curve corresponding to the average flux through each isosurface, which corresponds to the principal curve (45) from A to B.

The ZTS method starts from a series of configurations that lie along a path connecting A and B. A string γ is then parameterized along that path. An example parameterization is according to the arc length $\gamma : \gamma = \{\phi(s) : s \in [0, L]\}$, where L is the total length of the string from A to B. The set of configurations corresponds to a discrete set of points ϕ_i equally spaced along the string and is moved toward the minimum energy path in a two-step iterative process. First, it is propagated forward in time according to the underlying potential energy function $U(\phi)$ and then redistributed along the string to regain equal distribution. The details of the algorithm can be found in Reference 41.

FTS: finite-temperature string method

MSM: Markov state model

The ZTS method is most useful for systems with smooth energy surfaces but is perhaps less appropriate for systems with a rough energy surface (see, e.g., Reference 46, figure 1). In addition, as explained in Reference 28, the minimum energy paths might not coincide with a reaction path, even for smooth energy surfaces. The FTS method is designed to locate these types of reactive paths, by finding the principal curve instead of the minimum energy path. As with the ZTS method, FTS starts with an initial string of discretized configurations ϕ_i on a path between A and B (see 47). In the latest version of the algorithm (46), Voronoi cells are used to sample the region of configuration space around each ϕ_i, and the ϕ_i are moved toward the geometric center of each Voronoi cell. This process is iterated until convergence, at which point the set of ϕ_i approximates the principal curve between A and B. The free energy in terms of the Voronoi cells is also obtained.

String methods have been applied to a number of systems: the discovery of a zipper mechanism for nanotube fusion (48), which includes experimental results; wetting transitions on curved substrates (49); and the ability of DNA-repair proteins to recognize drug-DNA complexes (50). The ZTS method is related to chain-of-states approaches, such as the elastic band (51) and nudged elastic band (52) methods, which also find minimum energy paths.

4.2. Markov State Models

Another recent approach rapidly gaining in popularity is the use of Markov state models (MSMs). The method divides molecular configurational space into many metastable states (or microstates) such that a separation of timescales exists between the fast intrastate motion and slower interstate motion. A network of stochastic transitions among the microstates is built, and information about reaction mechanisms and rates can be obtained by considering the transition probabilities between the states. A recent review (53) provides a nontechnical introduction, with references to the explicit details of the method. Much of what is discussed here has been developed by Pande's group; however, similar approaches have been proposed, e.g., by Noé and colleagues (54), that use a more directly kinetic method to define the metastable states. A more recent publication lays out the latest advances and includes detailed error analysis (55).

Below we outline the basic ideas; fuller details on the algorithm are presented in Reference 56. MSM construction begins with a large set of molecular configurations that adequately cover important regions of configuration space. A suitable metric (e.g., RMSD) is used to make an initial clustering of configurations into microstates. As MSMs are examples of discrete-time Markov processes, if $\mathbf{p}(t)$ represents the vector of probabilities of being in each state at time $t = n\tau$, then the evolution of the system is determined by $\mathbf{p}(n\tau) = \mathbf{T}(\tau)^n\mathbf{p}(0)$. Here $\mathbf{T}(\tau)$ is a matrix of transition probabilities, with $T_{ji}(\tau)$ the probability of finding the system in state j at time t, given that it was in state i at time $t - \tau$. To be Markovian, one must choose the time τ such that the probability of being in state j at time t is independent of the previous state of the system. In other words, τ must be longer than the correlation time within each state.

Because no process that occurs on a timescale shorter than τ can be studied with a given MSM, iterative algorithms have been developed to construct a discretization of configuration space that attempts to minimize τ. Short simulations are run from the various microstates to construct a transition matrix between them. This resulting network must then be checked for convergence and Markovianity and can then be used to predict properties of the system. The details of these procedures are presented in References 55 and 56.

Recent applications of MSMs include the simulation of temperature-jump experiments on the trpzip2 peptide and calculation of the corresponding time-resolved infrared and two-dimensional infrared signals, which has been compared to experimental results (57). There is an ongoing dialogue between MSM calculations and experiments concerning the folding of the λ_{6-85} fragment (58, 59).

4.3. Milestoning

Milestoning has similar goals to the other methods presented in this review. As with MSMs (Section 4.2), the general tactic is to collect statistical data from short simulations between marker regions in configuration space (called milestones) rather than long reactive trajectories. The ideas of milestoning were developed in a series of papers beginning in 2004 (60). The initial milestoning algorithm required the use of an a priori reaction coordinate to define the milestones and was therefore limited to processes that could be described by a single degree of freedom. In addition, the algorithm included the approximation of using an equilibrium distribution of configurations at each milestone. This was later shown to be inappropriate (61), and a milestoning method based on the Voronoi tessellations of the configurational space was proposed (62). The use of such tessellations inspired the development of the latest version of the algorithm, directional milestoning (63), which is briefly outlined below and discussed in detail in a recent paper (64).

Directional milestoning begins with a set of configurations, termed anchors, that reasonably cover configurational space. A set of collective variables Q_l is also required, and the distance between two configurations $\mathbf{x}_i$ and $\mathbf{x}_j$ is calculated in terms of these variables: $d(\mathbf{x}_i, \mathbf{x}_j) \equiv \left[\sum_l \{Q_l(\mathbf{x}_i) - Q_l(\mathbf{x}_j)\}^2\right]^{1/2}$.

The directional milestone from anchor $\mathbf{x}_a$ to $\mathbf{x}_b$, denoted by $M(a \to b)$, is the set of points $\mathbf{x}$ defined through

$$M(a \to b) \equiv \{\mathbf{x} | d(\mathbf{x}, \mathbf{x}_a)^2 = d(\mathbf{x}, \mathbf{x}_b)^2 + \Delta_a^2, \quad \text{and } \forall \text{ anchors } \mathbf{x}_{\{c \neq a,b\}}, \quad d(\mathbf{x}, \mathbf{x}_b) \leq d(\mathbf{x}, \mathbf{x}_c)\}. \tag{2}$$

In other words, the directional milestone is the collection of molecular configurations such that the distance squared to anchor $\mathbf{x}_a$ is equal to the sum of the distance squared to anchor $\mathbf{x}_b$ and a term $\Delta_a = \min_{c \neq a} d(\mathbf{x}_c, \mathbf{x}_a)$. In addition, there is the extra requirement that the milestone $M(a \to b)$ configurations be closer to anchor $\mathbf{x}_b$ than any other anchor. The directional milestones partition off regions around each anchor, and the term $\Delta_a = \min_{c \neq a} d(\mathbf{x}_c, \mathbf{x}_a)$ creates a displacement between $M(a \to b)$ and $M(b \to a)$. This is shown schematically in Reference 64, figure 1. This asymmetry was introduced to increase the applicability of one of the major assumptions in milestoning: memory loss between milestones.

Once the directional milestones have been determined, a first hitting point distribution is approximated at each milestone. This distribution is such that, if the trajectory from each configuration in the distribution is run backward, the first milestone encountered will be different from the starting milestone. Short trajectories run from the first hitting point distribution at each

milestone are used to estimate a transition probability matrix **K**. Expressions for quantities such as the mean first passage time are written in terms of this transition matrix (see 64 for calculation details).

Recent work has used the directional milestoning approach in conjunction with experiments to determine how changes in HIV reverse transcriptase structure affect substrate selection (65) and to examine the transport of a blocked tryptophan molecule through a 1,2-dioleoyl-*sn*-glycero-3-phosphocholine membrane (66).

PCA: principal component analysis

5. MACHINE-LEARNING METHODS

For many systems, it may be possible and recommendable to use physical intuition and any a priori data available; however, when investigating a new system, such information may simply not yet exist. In these situations, even the definition of reactants and products may not be possible, making methods such as those based on TPS inapplicable. For these reasons, numerous groups have begun to adapt machine-learning methods for use in analyzing biomolecular systems. Many of these methods can be seen as dimensionality-reduction algorithms. That is, they take as input data from a high-dimensional space (such as molecular configuration space) and output a set of coordinates in a much lower dimensional space (such as a few collective variables). These methods rely on the assumption that physically relevant molecular configurations exist on a low-dimensional manifold that is embedded in a much higher dimensional space. Most of these methods preserve a metric of some sort, as discussed below.

5.1. Principal Component Analysis

The oldest and perhaps best-known technique for the definition of collective variables from large data samples is principal component analysis (PCA). It was initially proposed in 1901 by Pearson (67) as a way to optimally fit a large group of points in space and was independently discovered by Hotelling (68) in 1933. The basic idea of PCA is to find the linear transformation of variables that best captures the variance of the data; the implementation details and modern developments are discussed in Reference 69.

In the context of macromolecular dynamics, PCA was first introduced to estimate protein configurational entropy by characterizing the anharmonicity of collective motions (70–72) and was successfully applied to a number of different systems, with slight changes to the initial preprocessing of the input atomic data (73, 74). Numerous applications have been proposed in the past two decades to the analysis of protein folding and allosteric dynamics (75–79).

Despite its popularity, PCA presents some nontrivial problems when applied to macromolecular systems. Primarily, PCA is a linear method, whereas large macromolecular motions such as protein folding are usually highly nonlinear. As such, the usefulness of PCA is limited to small regions of configurational space. A nonlinear invariant of PCA has been proposed to overcome this problem (see the sidebar Kernel Principal Component Analysis).

Another problem results from the need for the molecular configurations to be optimally aligned to a common reference structure to remove trivial rotation and translation from the system motion. The results depend both on the particular choice for the reference configuration and on which subset of the molecule's atoms is used in the alignment. Reference 80 presents an example of this problem with a model system with two independently rotating methyl groups attached to the same rigid backbone.

Van Aalten et al. (81) proposed the use of dihedral angles as the input to PCA for macromolecular systems and found the results to be less noisy and to yield fewer important collective motions. Problems in dealing with angle periodicity were addressed by Mu et al. (82) by using the cos and

KERNEL PRINCIPAL COMPONENT ANALYSIS

To apply PCA to highly nonlinear data, investigators have developed advanced versions that use as input nonlinear functions of the system coordinates rather than the coordinates themselves. In particular, kernel PCA methods involve the diagonalization of a more general kernel $k(x_i, x_j)$ instead of the covariance matrix (130). The standard PCA algorithm can be seen as a particular form of kernel PCA with the kernel $k(x_i, x_j) = x_i \cdot x_j$. Polynomial kernels $(x_i \cdot x_j)^d$, sigmoidal kernels $tanh(\kappa x_i \cdot x_j)$, and exponential kernels $exp(-|x_i - x_j|^2/2\sigma^2)$ are commonly used to capture nonlinear effects in the system's collective dynamics. As a example, a polynomial kernel has recently been used to characterize the transition-state ensemble of the enzymatic reaction catalyzed by lactate dehydrogenase (131). In general, the choice of a particular form for the kernel is based on intuition on the system, or is a posteriori validated; to our knowledge, there is no theoretical justification for the use of different kernels for applications to macromolecular systems.

sin of the angles. The main disadvantage of these approaches is that large conformational changes might not result in correspondingly large variances in dihedral angle space, causing dihedral angle PCA to miss some important collective motions.

In protein-folding studies, the use of PCA on the contact map space instead of the Cartesian coordinate or dihedral angle space has been proposed (83–85). Working in the contact map space bypasses the optimal alignment problem. Additionally, a contact map of a minimally frustrated protein usually correlates with the protein energy. The formation of contacts usually involves relatively large energy changes but may correspond to a small change in the configuration space, and therefore may be difficult to capture by using geometry coordinates such as Cartesian coordinates or dihedral angles.

The problematic optimal alignment in PCA can also be bypassed by another popular dimensionality-reduction method, multidimensional scaling (MDS) (86, 87), which defines a set of low-dimensional variables best preserving pairwise distances in the high-dimensional space (e.g., the RMSD between molecular configurations). When the high-dimensional space is Euclidean, the results of MDS are equivalent to PCA; however, because of the optimal alignment step in the RMSD calculation, the molecular configurational space is not Euclidean. Details on the MDS algorithm are available in Reference 86.

5.2. Isomap

Isomap can be seen as a variant of MDS in which the geodesic distance is preserved in going from the high-dimensional to low-dimensional space (88). The geodesic distance is the distance between two points on a manifold. For example, the geodesic distance between Houston and Shanghai can be defined as the shortest flight between these two cities, which is much longer than the Euclidian distance measured through Earth. Given that we are constrained to move on Earth's surface (i.e., on a manifold), the geodesic distance is more meaningful here than the Euclidean distance. Isomap makes this same argument for molecular configurations.

MDS: multidimensional scaling

The Isomap algorithm approximates the geodesic distances between all pairs of molecular configurations by finding the shortest path between the configurations through a nearest-neighbor network. To preserve the accuracy of this approximation, a very dense sampling of configurations is often required, so dense that it becomes computationally demanding. A variant of the Isomap

algorithm, Scalable Isomap (ScIMAP), attempts to overcome this difficulty by randomly choosing landmark configurations and approximating only the geodesic distances from these to the remaining configurations (89, 90). ScIMAP has been successfully applied to analyze the folding/unfolding of coarse-grained SH3 (90), providing results in agreement with P_{fold} analysis on this system.

Despite significant improvement over PCA, Isomap and ScIMAP may present difficulties in application to molecular systems. First, MD data are nonuniformly distributed. Because of the Boltzmann distribution, macromolecular data are usually only sparsely sampled near a transition region, whereas there are many configurations near minima. Heterogeneous data create problems in the approximation of the geodesic distances on the nearest-neighbor network (91). Second, MD data are noisy. Although we view the configurations as lying on a manifold, this is of course only an approximation, and configurations should be considered as lying around the assumed manifold. This noisiness makes it difficult to choose the needed parameters for the nearest-neighbor searches (5, 92).

5.3. Sketch-Map

The sketch-map is a recently developed method that extracts coordinates from high-dimensional simulation data by preserving the middle-range RMSD distances between configurations. The argument is that simulation data fail to satisfy the main assumptions underlying dimensionality-reduction algorithms, and therefore the resulting coordinates can reveal only a sketch of the underlying manifold, rather than providing collective variables that correspond to reaction coordinates.

The motivation of the sketch-map method stems from the comparison between the RMSD distributions of all pairs of configurations collected during an MD simulation of a peptide (alanine-12) and those of two model data sets. Ceriotti et al. (93) found that the RMSD distribution of the MD data approximated that of a 24-dimensional, isotropic Gaussian distribution data set in the short range (RMSD $<$ 2 Å) and a 24-dimensional uniform distribution data set in the long range (RMSD $>$ 8 Å). This result suggests that the most interesting (less trivial) behavior is in the intermediate range of RMSD (between 2 and 8 Å, in the case of the peptide), and the low-dimensional space best preserving the RMSDs in this range is used as a set of collective coordinates to describe the dynamics of the MD data set. The sketch-map approach can also be used to enhance sampling in the metadynamics scheme (see Section 6.1); an application to reconstruct the free energy landscape of alanine-12 has been recently proposed (94).

Some of the problems discussed in the previous sections are also present in the sketch-map scheme. Additionally, it is not clear for a general system how to determine the distance range to preserve. In the case of the alanine-12 considered by Ceriotti et al. (93), the RMSD distribution of the MD data can be compared to two 24-dimension model systems, as its intrinsic dimensionality should be close to the number of dihedral angles. For larger, more complex systems, the intrinsic dimensionality might not be easily guessed a priori.

5.4. Diffusion Map

The diffusion map method extracts a low-dimensional coordinate system from MD data by preserving the diffusion distance between configurations. Roughly speaking, this distance represents the ease with which one configuration can diffuse into another. The coordinates that emerge from this technique also have the property of being good reaction coordinates for the system, as they in principle correspond to eigenfunctions of the Fokker-Planck (FP) operator for the system (95–97). **Figure 1** compares the various distance metrics discussed in previous sections on a toy system.

The dynamics of a macromolecular system in the high-friction limit can be modeled as a diffusion process obeying an FP equation, the solution of which is the probability density at position

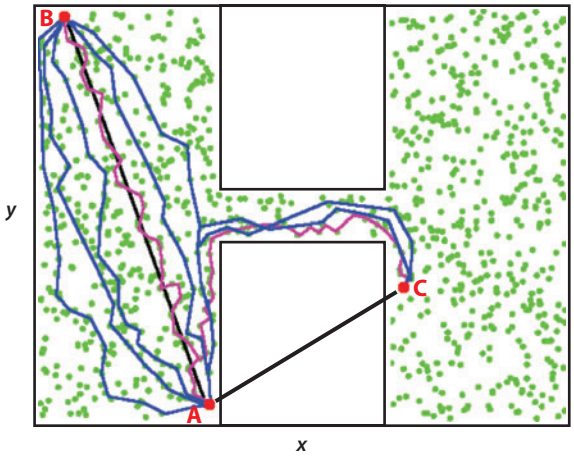

Figure 1

The difference between the multidimensional scaling (MDS), Isomap, and diffusion map approaches in the case of a simple diffusion in a two-dimensional box with a purely entropic barrier in the middle. All three can be considered metric-preserving dimensionality-reduction methods: MDS preserves Euclidean distances (*black lines*) between points, Isomap preserves geodesic distances approximated on the nearest-neighbor network (*purple lines*), and diffusion map preserves the diffusion distance, which can be considered as the ease of diffusion between the two points (visualized by the different diffusion trajectories connecting the points, in *blue*). It is clear that the diffusion between A and B is much easier than that between A and C because of the presence of the entropic barrier. The Euclidean distance between A and B is much larger than that between A and C, and it does not capture the shape of the manifold associated with the configurational space. The geodesic distance between A and B is comparable with the geodesic distance between A and C; although it captures the shape of the manifold more accurately than the Euclidean distance, the geodesic distance still does not correctly describe the diffusion process.

x at time t, $p = p(x, t)$, and it can be expressed by the eigenfunctions of the FP operator. For systems with a separation of timescales between the m slow collective motions and the remaining faster ones, the solution for the long timescale can be approximated by the first $m + 1$ eigenfunctions. When normalized by the eigenfunction with an eigenvalue equal to 0 (corresponding to the Boltzmann distribution), the other eigenfunctions possess qualities of good reaction coordinates (see 97), and only the first m collective coordinates are needed to describe the slow motions of the system (98). For a discrete data set, the first few eigenfunctions of the FP operator can be numerically approximated by building the kernel

$$K(x_i, x_j) = \exp[-||x_i - x_j||^2/(2\varepsilon^2)], \tag{3}$$

where ε is a measure of the length scale within which the transition between two configurations x_i and x_j is meaningful. By diagonalizing a properly normalized version of this kernel, one can define diffusion coordinates (DCs) to describe the system collective motions. The zeroth DC is trivial, the first DC corresponds to the slowest motion of the system, and so on.

In principle, the kernel (Equation 3) connects all pairs of configurations. However, in practice, only the pairs of points closer than the length scale ε are considered connected, as points further apart play an exponentially smaller role in the diffusion map calculation. This assumption is usually correct if a suitable value for ε is picked. This choice can be highly nontrivial, and it is discussed in detail in the next section. Through this assumption, the diffusion map method usually considers a larger connectivity network (and consequently a denser Markov matrix) than an MSM (see Section 4.2), in which only nearby configurations in the data set are considered connected.

The diffusion map method preserves the transition probabilities between all pairs of configurations, which are calculated on all the possible paths between these configurations. Therefore, it provides a more accurate description of the diffusion process, and it is more robust to noise than other metric-preserving methods.

LSDMap: locally scaled diffusion map

The diffusion map approach has been applied to characterize the (meta)stable states of *n*-alkane chains in water (99), and the conformational landscape of the antimicrobial peptide Microcin J25 (100). A weighted version of the diffusion map approach can also be used to deal with non-Boltzmann distributed data sets obtained from enhanced sampling techniques such as umbrella sampling (101) and to discover candidate physical variables corresponding to the slow modes of the system to better enhance the sampling. An iterative scheme can be found in recent work (102).

Although diffusion map is directly connected to the Fokker-Planck equation, which is used to characterize a classical diffusion process, it can also be considered as a more general dimensionality-reduction tool to analyze high-dimensional data from different types of simulations. For example, Virshup et al. (103) have applied the diffusion map approach to describe the excited-state reaction dynamics from ab initio molecular dynamics simulations.

5.5. Locally Scaled Diffusion Map

As discussed in the previous section, the local scale parameter ε entering the definition of the kernel (Equation 3) of the diffusion map plays a crucial role in determining the quality of the low-dimensional embedding, as it defines the scale within which the transition probability between two configurations contributes to the Markov matrix. In this section, we discuss the extension of the original diffusion map algorithm to the locally scaled diffusion map (LSDMap) (5), which introduces a configuration-specific local scale into the kernel.

If the data sample lies on a non-noisy low-dimensional manifold and is infinitely dense, the choice of ε is not too important to the numerical approximation of the FP equation, and using an increasingly small constant value of ε would generate a meaningful low-dimensional embedding. However, the data associated with the Boltzmann sampling of a macromolecular systems are not uniformly distributed: They present large density in free energy minima and are very sparse on top of barriers. Moreover, the data set is very noisy, and there is evidence that the noise changes with the region of the configuration space (5). For such a complex data set, if the scale ε is chosen to be too small and is comparable with the scale of the noise, the low-dimensional embedding will be corrupted. Conversely, if the local scale is selected to be too large, the manifold is artificially flattened, and again the results are corrupted. For this reason, a position-dependent local scale is necessary when applying the diffusion map to the characterization of large macromolecular systems.

A position-dependent length scale ε_i can be determined locally for each configuration x_i in the data set by considering the behavior of the MDS eigenspectrum as a function of an increasingly large RMSD radius ball centered at x_i. An analysis of the gaps between eigenvalues in the eigenspectrum provides information of the local intrinsic dimensionality and separates the eigenvalues corresponding to the collective local modes from the ones corresponding to the noise (104). The details of the algorithm and evidence of heterogeneous intrinsic dimensionality and local scale in different regions of the configuration space can be found in Reference 5.

As an example, **Figure 2** shows the results from the application of LSDMap to characterize the dynamics of a peptide. The dynamics of a polyalanine system, alanine-12, is simulated at 400 K in vacuum. As the conformational landscape of the peptide spans both helical and hairpin structures, it provides a good test system. It is clear that the first DC captures the folding/unfolding motion of the helical state, whereas the second DC corresponds to the formation of the hairpin from misfolded states. The folding/unfolding pathway can be characterized by considering

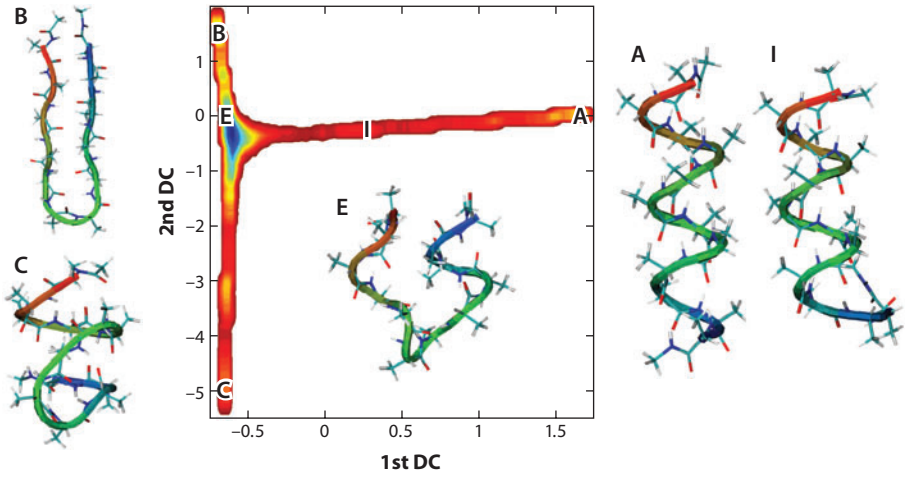

Figure 2

The application of the locally scaled diffusion map method to the characterization of the alanine-12 peptide dynamics. The free energy of this system is plotted as a function of the first two diffusion coordinates (DCs). Typical configurations corresponding to the major metastable states are shown.

intermediate structures along the first DC and shows that the lower part of the helix unfolds before the upper one. The application of LSDMap to another peptide, β-sheet miniprotein Beta3s, can be found in Reference 105.

LSDMap has also been recently tested on an all-atom model of alanine dipeptide, a coarse-grained SH3 model protein (5), and has been applied to characterize the dynamics of polymer reversal inside a nanopore (9). In these cases, the first DC correctly captures the slowest motion of the system, whereas the higher-order DCs describe faster motions (e.g., the formation of a nonspecific hydrophobic core in SH3 and partial misfolding in Beta3s).

Two main issues are present in the current implementation of the LSDMap approach. First, it is computationally intensive: The local-scale determination scales linearly with the number of points in the data set, and the eigenvalue decomposition of the normalized Markov matrix scales quadratically with the number of points. A preprocessing of the data set—i.e., selecting landmarks from clustering of points and assigning a weight to each landmark (similarly to that proposed in 93)—might be a way to decrease the computational cost. A more important problem is that the motions described by the DCs are not immediately and directly connected to the variation of a physical observable; the DCs obtained from LSDMap need to be correlated a posteriori to a set of relevant variables to physically characterize the slow motions of the system. This same problem is present in all the nonlinear dimensionality-reduction techniques discussed above. A possible solution may be the application of a genetic neural network (18, 106, 107) to correlate the motions defined by the DCs with a set of candidate physical variables, as discussed in Reference 105.

6. COLLECTIVE COORDINATES FOR ENHANCED SAMPLING

In addition to allowing for the determination of reaction mechanisms and rates, reaction coordinates are also useful in various enhanced sampling methods. Below we outline a few such techniques, which, rather than yielding a reaction coordinate as an output, take putative reaction coordinates as input for use in increasing the sampling of configuration space.

6.1. Metadynamics

The main idea behind the popular metadynamics approach (108), related techniques (109, 110), and recent extensions (111, 112) is to use a handful of collective coordinates as input and, during the dynamics, modify the underlying potential along those coordinates by the addition of small Gaussian functions to the visited regions in the reaction coordinates' space. These history-dependent additions to the potential encourage the system to leave free energy minima and explore other regions.

One difficulty with the method, as described above, is determining when to end a metadynamics run: In the long-time limit, the recovered free energy surface fluctuates around the actual free energy, and the magnitude of the fluctuations is controlled by the rate at which the small Gaussian functions are added to the potential energy (113, 114). In addition, long runs may result in the system becoming trapped in a physically irrelevant configuration. To overcome these problems, investigators developed the well-tempered metadynamics method (111), which allows for some control over regions of the free energy surface that are explored and offers as limiting cases ordinary MD and ordinary metadynamics. In practice, both metadynamics and well-tempered metadynamics have difficulties dealing with large numbers of collective variables. The recently developed reconnaissance metadynamics method (112) was designed to deal with large numbers of collective variables. Metadynamics has found applications to a wide variety of different classes of systems: structural changes in crystals (115), chemical reactions (116), protein-ligand docking (117), and analysis of Lennard-Jones and water clusters (118).

6.2. Adiabatic Free Energy Calculations

Adiabatic free energy calculation methods are designed to efficiently obtain the free energy associated with a molecular system by elevating both the temperature and mass of key collective degrees of freedom in MD simulations. Two similar methods have been proposed independently: TAMD (temperature-accelerated molecular dynamics) (120) and d-AFED (driven adiabatic free energy dynamics) (121), and its antecedent adiabatic free energy dynamics (119), allow for determination of the free energy in terms of collective variables.

These methods require a set of collective variables q_i as input, where the number of such variables n is much less than the number of particles in the system. Along these collective coordinates, the temperature is artificially elevated to allow easy crossing of high free energy barriers. Two thermostats are required: one for the high temperature of the collective variables T_s and another for the temperature of the regular system T. To keep the T_s coordinates from polluting the dynamics in the rest of the system, the mass along the collective coordinates is increased enough to achieve adiabatic decoupling between the q_i variables and the remaining degrees of freedom.

The free energy in terms of the q_i at the ordinary system temperature T can be calculated directly from the probability distribution P_{adb} obtained from such a simulation: $F(q_1, \ldots, q_n) = -k_B T_s \ln\{P_{adb}(q_1, \ldots, q_n)\}$. The derivation is presented in Reference 121. Studies of molecular crystal polymorphism (122), insulin receptor kinase dynamics (123), and short-range hydrogen diffusion in Na_3AlH_6 have used the above-mentioned methods.

7. CONCLUSIONS

Above we present and discuss recent methods for the definition and application of reaction coordinates and reaction paths on the characterization of complex macromolecular reactions. Different techniques are more appropriate for application to different systems, and we discuss the advantages and problems for each. It was our intention to offer a general overview of the main ideas and philosophy underlying existing approaches, and we refer the interested reader to the technical literature cited above for the practical implementation of the discussed methodologies.

Although some of the basic ideas behind the methods presented here may date back several decades, their implementations for the characterization of complex macromolecular reactions are fairly recent. Usually, new techniques are initially tested and presented on simple models, such as the alanine dipeptide. The application to these well-characterized test systems, in which reaction paths are known and rates can be measured by direct simulation of transitions between reactants and products, allows one to evaluate the performance of a method and the reaction coordinates it defines, independently of the other approximations involved in the simulation, such as the choice of a force field. Ultimately, however, the power of a theoretical method needs to be assessed by considering its ability to make predictions that are experimentally verifiable and to interpret experimental results. Most techniques discussed above have been recently proposed. Some are still in the initial testing phase, and some are just beginning to be applied to systems that are experimentally testable. The jury is still out on the ability of most of these reaction coordinates methods to make quantitative predictions on genuine systems of relevance. We expect that the work done on the theoretical definition of these approaches will soon be followed by applications to the characterization of complex molecular reactions that cannot be studied by more classical methods.

SUMMARY POINTS

1. The definition and application of collective variables are critical for the characterization of the long-timescale dynamics of macromolecular systems and the associated free energy landscape. These coordinates can be used to understand reaction mechanisms and determine reaction rates.
2. Path-based techniques have been developed to understand reaction mechanisms and compute reaction rates by analyzing a large set of reactive paths connecting metastable states.
3. State-based techniques have been developed to determine reaction mechanisms and perform rate calculations by dividing the configurational space of the system into different regions and performing most calculations locally in the different regions.
4. Machine-learning methods have been developed to extract the main collective motions and associated collective coordinates from a large sampling of MD data, with minimal a priori knowledge of the system required.
5. If the collective variables associated with the long-timescale dynamics of a system are known, they can be used to bias the MD simulation to sample rare and slow motions efficiently.

FUTURE ISSUES

1. No method can presently extract reaction coordinates on the fly during MD simulations and at the same time use them to enhance the sampling of the configurational space. We expect that the techniques discussed here will be extended and combined to explore this possibility in the near future.
2. Most methods for the determination of reaction coordinates and pathways are tested on simple systems. The definition of a set of benchmarks would be very useful to evaluate new methods.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are indebted to Peter Wolynes and Mauro Maggioni for insightful discussions. The work of Clementi's group is supported by the NSF (CDI-type I grant CHE-0835824 and grant CHE-1152344) and the Welch Foundation (C-1570 to C.C.).

LITERATURE CITED

1. McCammon JA, Karplus M. 1979. Dynamics of activated processes in globular proteins. *Proc. Natl. Acad. Sci. USA* 76:3585–89
2. Rothman MJ, Lohr LL Jr. 1980. Analysis of an energy minimization method for locating transition states on potential energy hypersurfaces. *Chem. Phys. Lett.* 70:405–9
3. Williams IH, Maggiora GM. 1982. Use and abuse of the distinguished-coordinate method for transition-state structure searching. *Chem. Phys. Lett.* 89:365–78
4. Bell S, Crighton JS. 1984. Locating transition states. *J. Chem. Phys.* 80:2464–75
5. **Rohrdanz MA, Zheng W, Maggioni M, Clementi C. 2011. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.* 134:124116**
6. Nymeyer H, Socci N, Onuchic JN. 2000. Landscape approaches for determining the ensemble of folding transition states: Success and failure hinge on the degree of frustration. *Proc. Natl. Acad. Sci. USA* 97:634–39
7. Clementi C, Nymeyer H, Onuchic JN. 2000. Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–53
8. Huang L, Makarov DE. 2008. The rate constant of polymer reversal inside a pore. *J. Chem. Phys.* 128:114903
9. Zheng W, Rohrdanz MA, Maggioni M, Clementi C. 2011. Polymer reversal rate calculated via locally scaled diffusion map. *J. Chem. Phys.* 134:144109
10. Du R, Pande VS, Grosberg A, Tanaka T, Shakhnovich E. 1998. On the transition coordinate for protein folding. *J. Chem. Phys.* 108:334–50
11. Onsager L. 1938. Initial recombination of ions. *Phys. Rev.* 54:554–57
12. Pratt LR. 1986. A statistical method for identifying transition states in high dimensional problems. *J. Chem. Phys.* 85:5045–48
13. Eyring H. 1962. Transmission coefficient in reaction rate. *Rev. Mod. Phys.* 34:616–19
14. **Bolhuis PG, Chandler D, Dellago C, Geissler P. 2002. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* 53:291–318**
15. Cho S, Levy Y, Wolynes PG. 2006. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl. Acad. Sci. USA* 103:586–91
16. E W, Ren W, Vanden-Eijnden E. 2005. Transition pathways in complex systems: reaction coordinates, isocommittor surfaces, and transition tubes. *Chem. Phys. Lett.* 413:242–47
17. Clementi C, Jennings P, Onuchic JN. 2001. Prediction of folding mechanism for circular-permuted proteins. *J. Mol. Biol.* 311:879–90
18. Ma A, Dinner AR. 2005. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B* 109:6769–79
19. Berezhkovskii A, Hummer G, Szabo A. 2009. Reactive flux and folding pathways in network models of coarse-grained protein dynamics. *J. Chem. Phys.* 130:205102
20. Dellago C, Bolhuis PG, Geissler PL. 2002. Transition path sampling. *Adv. Chem. Phys.* 123:1–78

5. Presents the locally scaled diffusion map approach, designed for the noisy and nonuniformly distributed data from macromolecular systems.

14. Discusses transition-path-sampling techniques in detail.

21. Anderson JB. 1973. Statistical theories of chemical reactions: distributions in the transition region. *J. Chem. Phys.* 58:4684–92
22. Dellago C, Bolhuis PG, Csajka F, Chandler D. 1998. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* 108:1964–77
23. Dellago C, Bolhuis PG, Chandler D. 1999. On the calculation of reaction rate constants in the transition path ensemble. *J. Chem. Phys.* 110:6617–25
24. Olender R, Elber R. 1996. Calculations of classical trajectories with a very large time step: formalism and numerical examples. *J. Chem. Phys.* 105:9299–315
25. Dellago C, Bolhuis PG, Chandler D. 1998. Efficient transition path sampling: application to Lennard-Jones cluster rearrangements. *J. Chem. Phys.* 108:9236–45
26. Ensing B, Baerends E. 2002. Reaction path sampling of the reaction between iron(II) and hydrogen peroxide in aqueous solution. *J. Phys. Chem. A* 106:7902–10
27. Bolhuis PG. 2003. Transition-path sampling of β-hairpin folding. *Proc. Natl. Acad. Sci. USA* 100:12129–34

28. Details the theory and application of string methods and path-finding algorithms.

28. E W, Vanden-Eijnden E. 2010. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.* 61:391–420

29. van Erp TS, Moroni D, Bolhuis PG. 2003. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.* 118:7762–74
30. Moroni D, Bolhuis PG, van Erp TS. 2004. Rate constants for diffusive processes by partial path sampling. *J. Chem. Phys.* 120:4055–65
31. Moroni D, van Erp TS, Bolhuis PG. 2005. Simultaneous computation of free energies and kinetics of rare events. *Phys. Rev. E* 71:056709
32. van Erp TS, Bolhuis PG. 2005. Elaborating transition interface sampling methods. *Proc. Natl. Acad. Sci. USA* 205:157–81
33. Rogal J, Bolhuis PG. 2008. Multiple state transition path sampling. *J. Chem. Phys.* 129:224107
34. Juraszek J, Vreede J, Bolhuis PG. 2012. Transition path sampling of protein conformational changes. *Chem. Phys.* 396:30–44
35. Morelli MJ, Allen RJ, Tanase-Nicola S, ten Wolde PR. 2008. Eliminating fast reactions in stochastic simulations of biochemical networks: a bistable genetic switch. *J. Chem. Phys.* 128:045105
36. Velez-Vega C, Borrero EE, Escobedo FA. 2010. Kinetics and mechanism of the unfolding native-to-loop transition of Trp-cage in explicit solvent via optimized forward flux sampling simulations. *J. Chem. Phys.* 133:105103
37. Růžička Š, Quigley D, Allen MP. 2012. Folding kinetics of a polymer. *Phys. Chem. Chem. Phys.* 14:6044–53
38. Filion L, Ni R, Frenkel D, Dijkstra M. 2011. Simulation of nucleation in almost hard-sphere colloids: The discrepancy between experiment and simulation persists. *J. Chem. Phys.* 134:134901
39. Becker NB, Allen RJ, ten Wolde PR. 2012. Non-stationary forward flux sampling. *J. Chem. Phys.* 136:174118
40. E W, Ren W, Vanden-Eijnden E. 2002. String method for the study of rare events. *Phys. Rev. B* 66:052301
41. E W, Ren W, Vanden-Eijnden E. 2007. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.* 126:164103
42. Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G. 2006. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* 125:024106
43. Maragliano L, Vanden-Eijnden E. 2007. On-the-fly string method for minimum free energy paths calculation. *Chem. Phys. Lett.* 446:182–90
44. E W, Vanden-Eijnden E. 2006. Towards a theory of transition paths. *J. Stat. Phys.* 123:503–23
45. Hastie T, Tibshirani R, Friedman JH. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer
46. Vanden-Eijnden E, Venturoli M. 2009. Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J. Chem. Phys.* 130:194103
47. E W, Ren W, Vanden-Eijnden E. 2005. Finite temperature string method for the study of rare events. *J. Phys. Chem. B* 109:6688–93
48. Yoon M, Han S, Kim G, Lee S, Berber S, et al. 2004. Zipper mechanism of nanotube fusion: theory and experiment. *Phys. Rev. Lett.* 92:075504

49. Qiu C, Qian T. 2009. Nucleation of wetting films on cylindrical and spherical substrates: a numerical study by the string method. *J. Chem. Phys.* 131:124708
50. Elder RM, Jayaraman A. 2012. Sequence-specific recognition of cancer drug-DNA adducts by HMGB1a repair protein. *Biophys. J.* 102:2331–38
51. Gillilan R, Lilien R. 2004. Optimization and dynamics of protein-protein complexes using B-splines. *J. Comput. Chem.* 25:1630–46
52. Sheppard D, Terrell R, Henkelman G. 2008. Optimization methods for finding minimum energy paths. *J. Chem. Phys.* 128:134106
53. Pande VS, Beauchamp K, Bowman GR. 2010. Everything you wanted to know about Markov state models but were afraid to ask. *Methods* 52:99–105
54. Noé F, Horenko I, Schutte C, Smith JC. 2007. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J. Chem. Phys.* 126:155102
55. Prinz JH, Wu H, Sarich M, Keller B, Senne M, et al. 2011. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* 134:174105
56. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC. 2007. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* 126:155101
57. Zhuang W, Cui RZ, Silva DA, Huang X. 2011. Simulating the T-jump-triggered unfolding dynamics of trpzip2 peptide and its time-resolved IR and two-dimensional IR signals using the Markov state model approach. *J. Phys. Chem. B* 115:5415–24
58. Bowman GR, Voelz VA, Pande VS. 2010. Atomistic folding simulations of the five-helix bundle protein 685. *J. Am. Chem. Soc.* 133:664–67
59. Prigozhin MB, Gruebele M. 2011. The fast and the slow: folding and trapping of λ_{6-85}. *J. Am. Chem. Soc.* 133:19338–41
60. Faradjian A, Elber R. 2004. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* 120:10880–89
61. Vanden-Eijnden E, Venturoli M, Ciccotti G, Elber R. 2008. On the assumptions underlying milestoning. *J. Chem. Phys.* 129:174102
62. Vanden-Eijnden E, Venturoli M. 2009. Markovian milestoning with Voronoi tessellations. *J. Chem. Phys.* 130:194101
63. Majek P, Elber R. 2010. Milestoning without a reaction coordinate. *J. Chem. Theory Comput.* 6:1805–17
64. Kirmizialtin S, Elber R. 2011. Revisiting and computing reaction coordinates with directional milestoning. *J. Phys. Chem. A* 115:6137–48
65. Kirmizialtin S, Nguyen V, Johnson KA, Elber R. 2012. How conformational dynamics of DNA polymerase select correct substrates: experiments and simulations. *Structure* 20:618–27
66. Cardenas AE, Jas GS, DeLeon KY, Hegefeld WA, Kuczera K, Elber R. 2012. Unassisted transport of *N*-acetyl-L-tryptophanamide through membrane: experiment and simulation of kinetics. *J. Phys. Chem. B* 116:2739–50
67. Pearson K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 2:559–72
68. Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24:417–41
69. Jolliffe IT. 2002. *Principal Component Analysis*. Berlin: Springer-Verlag
70. Karplus M, Kushick JN. 1981. Method for estimating the configurational entropy of macromolecules. *Macromolecules* 14:325–32
71. Levy RM, Karplus M, Kushick J, Perahia D. 1984. Evaluation of the configurational entropy for proteins: application to molecular dynamics simulations of an α-helix. *Macromolecules* 17:1370–74
72. Levy RM, Srinivasan AR, Olson WK, McCammon JA. 1984. Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers* 23:1099–112
73. Ichiye T, Karplus M. 1991. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins* 11:205–17

53. Reviews the theory and application of Markov state models.

64. Explains in detail the calculation of reaction coordinates by means of directional milestoning.

74. Kitao A, Hirata F, Go N. 1991. The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem. Phys.* 158:447–72
75. García AE. 1992. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* 68:2696–99
76. Amadei A, Linssen A, Berendsen HJC. 1993. Essential dynamics of proteins. *Proteins* 17:412–25
77. Nolde SB, Arseniev AS, Orekhov VY, Billeter M. 2002. Essential domain motions in barnase revealed by MD simulations. *Proteins* 46:250–58
78. Levy Y, Caflisch A. 2003. Flexibility of monomeric and dimeric HIV-1 protease. *J. Phys. Chem. B* 107:3068–79
79. Teodoro ML, Phillips GN Jr, Kavraki LE. 2003. Understanding protein flexibility through dimensionality reduction. *J. Comput. Biol.* 10:617–34
80. Prompers JJ, Brüschweiler R. 2002. Dynamic and structural analysis of isotropically distributed molecular ensembles. *Proteins* 46:177–89
81. Van Aalten DMF, De Groot BL, Findlay JBC, Berendsen HJC, Amadei A. 1997. A comparison of techniques for calculating protein essential dynamics. *J. Comput. Chem.* 18:169–81
82. Mu Y, Nguyen PH, Stock G. 2005. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins* 58:45–52
83. Shen T, Zong C, Hamelberg D, McCammon JA, Wolynes PG. 2005. The folding energy landscape and phosphorylation: modeling the conformational switch of the NFAT regulatory domain. *FASEB J.* 19:1389–95
84. Lätzer J, Eastwood MP, Wolynes PG. 2006. Simulation studies of the fidelity of biomolecular structure ensemble recreation. *J. Chem. Phys.* 125:214905
85. Lätzer J, Shen T, Wolynes PG. 2008. Conformational switching upon phosphorylation: a predictive framework based on energy landscape principles. *Biochemistry* 47:2110–22
86. Härdle W, Simar L. 2007. *Applied Multivariate Statistical Analysis*. Berlin: Springer-Verlag
87. Troyer JM, Cohen FE. 1995. Protein conformational landscapes: energy minimization and clustering of a long molecular dynamics trajectory. *Proteins* 23:97–110
88. Tenenbaum JB, De Silva V, Langford JC. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319–23
89. Silva V, Tenenbaum JB. 2003. Global versus local methods in nonlinear dimensionality reduction. *Adv. Neural Inf. Process. Syst.* 15:705–12
90. Das P, Moll M, Stamati H, Kavraki LE, Clementi C. 2006. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. USA* 103:9885–90
91. Donoho DL, Grimes C. 2003. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA* 100:5591–96
92. Balasubramanian M, Schwartz EL. 2002. The isomap algorithm and topological stability. *Science* 295:7
93. Ceriotti M, Tribello GA, Parrinello M. 2011. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. USA* 108:13023–28
94. Tribello GA, Ceriotti M, Parrinello M. 2012. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* 109:5196–201
95. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, et al. 2005. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. USA* 102:7426–31
96. Coifman RR, Lafon S. 2006. Diffusion maps. *Appl. Comput. Harmon. Anal.* 21:5–30
97. Coifman RR, Kevrekidis IG, Lafon S, Maggioni M, Nadler B. 2008. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Model. Simul.* 7:842–64
98. Jones PW, Maggioni M, Schul R. 2008. Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels. *Proc. Natl. Acad. Sci. USA* 105:1803–8
99. Ferguson AL, Panagiotopoulos AZ, Debenedetti PG, Kevrekidis IG. 2010. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc. Natl. Acad. Sci. USA* 107:13597–602
100. Ferguson AL, Panagiotopoulos AZ, Kevrekidis IG, Debenedetti PG. 2011. Nonlinear dimensionality reduction in molecular simulation: the diffusion map approach. *Chem. Phys. Lett.* 509:184903

90. Presents an application of a scalable version of Isomap to macromolecular systems best preserving geodesic distances.

93. Provides the algorithm of the sketch-map approach, preserving medium-range distances between pairs of configurations.

97. Presents the theoretical definition of a diffusion map to characterize high-dimensional diffusion processes.

101. Torrie GM, Valleau JP. 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation-umbrella sampling. *J. Comput. Phys.* 23:187–99
102. Ferguson AL, Panagiotopoulos AZ, Debenedetti PG, Kevrekidis IG. 2011. Integrating diffusion maps with umbrella sampling: application to alanine dipeptide. *J. Chem. Phys.* 134:135103
103. Virshup AM, Chen J, Martínez TJ. 2012. Nonlinear dimensionality reduction for nonadiabatic dynamics: the influence of conical intersection topography on population transfer rates. *J. Chem. Phys.* 137:22A519
104. Little AV, Jung YM, Maggioni M. 2009. Multiscale estimation of intrinsic dimensionality of data sets. In *AAAI Fall Symposium*, pp. 26–33. Palo Alto, CA: Assoc. Adv. Artif. Intell.
105. Zheng W, Qi B, Rohrdanz MA, Caflisch A, Dinner AR, Clementi C. 2011. Delineation of folding pathways of a β-sheet mini-protein. *J. Phys. Chem. B* 115:13065–74
106. So SS, Karplus M. 1996. Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural networks. *J. Med. Chem.* 39:1521–30
107. So SS, Karplus M. 1996. Genetic neural networks for quantitative structure-activity relationships: improvements and application of benzodiazepine affinity for benzodiazepine/GABAa receptors. *J. Med. Chem.* 39:5246–56
108. Laio A, Parrinello M. 2002. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* 99:12562–66
109. Huber T, Torda AE, Gunsteren WF. 1994. Local elevation: a method for improving the searching properties of molecular dynamics simulation. *J. Comput.-Aided Mol. Des.* 8:695–708
110. Wang F, Landau DP. 2001. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* 86:2050–53
111. Barducci A, Bussi G, Parrinello M. 2008. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* 100:020603
112. Tribello G, Ceriotti M, Parrinello M. 2010. A self-learning algorithm for biased molecular dynamics. *Proc. Natl. Acad. Sci. USA* 107:17509–14
113. Bussi G, Laio A, Parrinello M. 2006. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.* 96:090601
114. Laio A, Rodriguez-Fortea A, Gervasio FL, Ceccarelli M, Parrinello M. 2005. Assessing the accuracy of metadynamics. *J. Phys. Chem. B* 109:6714–21
115. Martoňák R, Laio A, Bernasconi M, Ceriani C, Raiteri P, et al. 2005. Simulation of structural phase transitions by metadynamics. *Z. Kristallogr.* 220:489–98
116. Ensing B, De Vivo M, Liu Z, Moore P, Klein ML. 2006. Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Acc. Chem. Res.* 39:73–81
117. Gervasio FL, Laio A, Parrinello M. 2005. Flexible docking in solution using metadynamics. *J. Am. Chem. Soc.* 127:2600–7
118. Tribello GA, Cuny J, Eshet H, Parrinello M. 2011. Exploring the free energy surfaces of clusters using reconnaissance metadynamics. *J. Chem. Phys.* 135:114109
119. Rosso L, Miináry P, Zhu Z, Tuckerman ME. 2002. On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles. *J. Chem. Phys.* 116:4389–402
120. Maragliano L, Vanden-Eijnden E. 2006. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* 426:168–75
121. Abrams JB, Tuckerman ME. 2008. Efficient and direct generation of multidimensional free energy surfaces via adiabatic dynamics without coordinate transformations. *J. Phys. Chem. B* 112:15742–57
122. Yu TQ, Tuckerman ME. 2011. Temperature-accelerated method for exploring polymorphism in molecular crystals based on free energy. *Phys. Rev. Lett.* 107:015701
123. Vashisth H, Maragliano L, Abrams CF. 2012. "DFG-Flip" in the insulin receptor kinase is facilitated by a helical intermediate state of the activation loop. *Biophys. J.* 102:1979–87
124. Best RB, Hummer G. 2005. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. USA* 102:6732–37
125. Peters B, Trout BL. 2006. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.* 125:054108
126. Allen R, Warren P, ten Wolde P. 2005. Sampling rare switching events in biochemical networks. *Phys. Rev. Lett.* 94:018104

127. Allen R, Frenkel D, ten Wolde P. 2006. Simulating rare events in equilibrium or nonequilibrium stochastic systems. *J. Chem. Phys.* 124:024102
128. Borrero EE, Escobedo FA. 2007. Reaction coordinates and transition pathways of rare events via forward flux sampling. *J. Chem. Phys.* 127:164101
129. Allen RJ, Valeriani C, ten Wolde PR. 2009. Forward flux sampling for rare event simulations. *J. Phys. Condens. Matter* 21:463102
130. Schölkopf B, Smola A, Müller KR. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* 10:1299–319
131. Antoniou D, Schwartz SD. 2011. Toward identification of the reaction coordinate directly from the transition state ensemble using the kernel PCA method. *J. Phys. Chem. B* 115:2465–69

129. Reviews forward flux sampling and includes a brief discussion of equilibrium methods.

RELATED RESOURCES

MSMBuilder: software to build Markov state models. **https://simtk.org/home/msmbuilder/**

The String Method Page: information and Matlab code for running the string method on the Mueller potential. **http://www.cims.nyu.edu/~eve2/string.htm**

PLUMED: plugin for free energy calculations, with an emphasis on biological applications. **http://www.plumed-code.org/**

LSDMap code: software to extract locally scaled diffusion map coordinates from simulation data. **http://sourceforge.net/projects/lsdmap/**

Annual Review of
Physical Chemistry
Volume 64, 2013

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Physical Chemistry* articles may be found at http://physchem.annualreviews.org/errata.shtml