Transition Path Sampling and Forward Flux Sampling. Applications to Biological Systems

Fernando A Escobedo,* Ernesto E Borrero, and Juan C Araque

School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853 USA

Abstract

The last decade has seen a rapid growth in the number of simulation methods and applications dealing with the sampling of transition pathways of rare nanoscale events. Such studies are crucial, for example, to understand the mechanism and kinetics of conformational transitions and enzymatic events associated with the function of biomolecules. In this review, a broad account of transition path sampling approaches is provided, starting from the general concepts, progressing to the specific principles that underlie some of the most important methods, and eventually singling out the so-called forward flux sampling method for a more detailed description. This is done because forward flux sampling, despite its appealing simplicity and potential efficiency, has thus far received limited attention by practitioners. While path sampling methods have a widespread application to many types of rare transitional events, here only recent applications involving biomolecules are reviewed, including isomerizations, protein folding, and enzyme catalysis.

* Author to whom correspondence should be addressed at fe13@cornell.edu

Table of Contents

List of Main Abbreviations

ANOVA	Analysis of variance
BD	Brownian dynamics
BG	Branch Growth
DFFS	Direct forward flux sampling
FFS	Forward Flux Sampling
FTS	finite temperature formulation of the String method
GNN	Genetic neural network
KMC	Kinetics Monte Carlo
LSE	Least square estimation
MC	Monte Carlo
MD	Molecular dynamics
MEP	Minimum energy pathway
MFEP	Mean free energy path
ML	Maximum likelihood
p_B	Committor probability to basin B
PPTIS	Partial path transition interface sampling
RxC	Reaction coordinate
TIS	Transition interface sampling
TS	Transition state
TSE	Transition state ensemble
TPE	Transition path ensemble
US	Umbrella sampling
WE	Weighted ensemble

1. Introduction

Proteins, DNA, and RNA are examples of biomolecules that can be seen as programmable heteropolymers which self-assemble into well-defined structures. To enact their functions, these molecules form complexes with other molecules (e.g., for protein-protein signaling, protein-antigen binding, DNA-enzyme interactions during replication, etc.). Any one of these biological events, namely, those involving the folding of a biomolecule into its "native" state, the binding of two biomolecules, and any specific enzymatic action, can all be considered "rare" events on the time scale of the long pre- and post-stages of the process. In the following, we constrain ourselves to elaborate on the protein folding problem, the paradigm of such rare events, to motivate the importance of molecular simulation approaches that target such events.

Proteins are biochemical molecules known to be involved in almost every biological process since their function ranges from catalysis of chemical reactions to maintenance and structural support of the cell. Their primary structure is synthesized by the ribosome as linear sequences of amino acids; these then assume secondary structures (i.e., α-helix and β-sheet) and tertiary structures based on a variety of chemical interactions that form between amino acid residues. The transformation of genetic information into unique three-dimensional native protein structures capable of complex biological activity depends on the accuracy and efficiency of the folding process. However, protein folding can be difficult given that it competes with other cellular events such as misfolding and aggregation. The understanding of how a protein folds successfully into its functional configuration is important for the separation, purification and formulation of therapeutic proteins, the aging and storage of proteins drugs, and the analysis of neurodegenerative human diseases (such as Alzheimer's, Huntington's, and Parkinson's) associated with protein misfolding and aggregation.[1, 2] These are active questions that numerous simulation studies have aimed to understand since the 1980s.[3] However, because the characteristic waiting time for the transition to the folded state is typically orders of magnitude longer than the time for the folding event itself, conventional "brute-force" approaches are not indicated for the study of this process.[4, 5] Furthermore, from "brute-force" simulations is difficult to obtain the

transition path ensemble (TPE), the transition state ensemble (TSE), as well as the mechanistic details contained in the folding pathways. It is here where path sampling methods are starting to contribute to our understanding of folding events.[6-8]

Taking advantage that a rare event itself is not slow, just very infrequent, path sampling algorithms are effective because they concentrate the computational effort to the transitional events only. These algorithms thus aim to collect characteristics pathways (i.e., the TPE) that describe the system's transition between stable states (e.g., unfolded and folded state in a protein). The TPE is then used to characterize the system's dynamics at a macroscopic level by computing rate constants. The TPE could also be analyzed to understand transitions at a microscopic level by extracting information about the TSE. Assuming that the pathways in the TPE harvested by a path sampling algorithm include examples of the transition state (TS) intermediates, a properly weighted TSE could be collected by screening configurations along reactive pathways. The TSE can then be used to characterize the mechanism by which the transition in complex systems occurs by identifying the sequence of key events and bottlenecks involved in a transition.

For high-dimensional complex systems, the committor probability to the basin B (p_B) can serve as guide to identify the TSE. This function is the probability that a system with a given configuration will commit to state B before reaching the state A; p_B is essentially only a function of the system's configuration and quantifies the tendency of a configuration to relax to a particular basin of attraction under the system's intrinsic dynamics. The location of the TSE can be determined by screening configurations with p_B=1/2.

In general, path sampling methods can generate transition paths by using an order parameter; i.e., an initial approximation of the reaction coordinate (RxC), that allows distinguishing the stable states of the system and monitoring the progress along the reaction pathways without the need to know the exact RxC of the system. However, the TPE harvested from the path sampling simulation can be used to estimate a good RxC by relating p_B data of all states on the TPE to a set of "collective" variables

(i.e., physically meaningful properties that condense many atomistic degrees of freedom).[9, 10] The knowledge of the RxC provides a more detailed understanding of the dynamics of the rare event.[11]

While several reviews on transition path sampling methods and applications have recently appeared in the literature (e.g., Refs. [7,12]), in this review a stronger emphasis is placed on interface-based methods and the so-called "forward flux sampling" in particular, and on applications entailing model biological systems. Our selection of methods and applications is not comprehensive and necessarily reflects our own experience in the field. In the following section, we introduce some of the most common path sampling methods and how the collection of the TPE can be used to obtain information of the RxC, the TSE, and mechanistic details of the transition. Afterwards, we review some applications to biological systems, ending with a discussion of the isomerization of alanine dipeptide and the folding of the Trp-cage mini-protein, cases for which several path sampling methods have been employed.

2. Path Sampling Methods

The central idea of path sampling methods is to generate trajectories for rare events that constitute the transition path ensemble (TPE); for example, for transitions between two well-defined stable states A and B separated by a high free energy barrier (compared to the thermal energy of the system).[13-15] In this section, we will introduce briefly the TPS formalism, the archetype of such computational techniques,[13-15] to set the stage for the description of other path sampling methods. A detailed description of the formalism and applications of TPS can be found in several TPS review articles.[16, 7, 15]

2.1. Transition Path Sampling (TPS) scheme

In the TPS method, the trajectory space is sampled using a Monte Carlo (MC) procedure, where a new trajectory $x^{(n)}(\Gamma)$ is generated from an old one $x^{(0)}(\Gamma)$. To this end, each path is represented by an ordered sequence of space points (or time slices):

$$x(\Gamma) \equiv \{x_0, x_{\Delta t}, x_{2\Delta t}, ..., x_\Gamma\} \qquad (1)$$

where x denotes the microscopic state of the system depending on the dynamics (e.g., $x = \{r, p\}$; position (r) and momenta (p) of particles) and Γ specifies the length of the trajectory. Hence, a trajectory results from $\Gamma / \Delta t$ steps that are separated by a time increment Δt and ultimately connecting state A to B. Regions A and B are defined in terms of an order parameter $\lambda(r)$, where r denotes the coordinates of the phase space. The parameter $\lambda(r)$ can be chosen such that the system has values $\lambda(r) \leq \lambda_A(r)$ in region A and $\lambda(r) \geq \lambda_B(r)$ in region B.

While there are various schemes to generate these new trajectories, the shooting algorithm has been proven to be particularly efficient to sample trajectory space.[16] In this procedure a new $x^{(n)}(\Gamma)$ pathway is generated from a randomly selected time slice $x_{t'}^{(0)}$ of the old path (see Fig. 1). For deterministic dynamics, the $x_{t'}^{(0)}$ point is modified, for instance, by adding a small perturbation to the momenta (no perturbation is required for stochastic dynamics), yielding $x_{t'}^{(n)}$ from which a new trajectory is constructed by following forward and backward the intrinsic dynamics of the system (e.g., integrating the equations of motion). This newly generated path is then accepted or rejected depending on the Metropolis acceptance probability:

$$P_{acc}\left[x^{(0)}(\Gamma) \to x^{(n)}(\Gamma)\right] = h_A\left[x_0^{(n)}\right] h_B\left[x_\Gamma^{(n)}\right] \min\left[1, \frac{\rho\left(x_{t'}^{(n)}\right)}{\rho\left(x_{t'}^{(0)}\right)}\right] \quad (2)$$

where $\rho(x)$ is the stationary distribution evaluated at x and $h_A[\xi]$ is a indicator function for state A such that it is unity if the system is in the initial basin at ξ and is zero otherwise.[13-15] Analogously, $h_B[\xi]$ is unity if the system is in the final basin at ξ and zero otherwise.

To increase the efficiency of the TPS simulations several alternative optimization techniques have been proposed including methods to control the average acceptance probability in Eq. (2) {e.g., by changing the magnitude of the momentum perturbation (for deterministic trajectories),[17, 14]}, combination of TPS with parallel tempering simulations at the path level,[18, 19] and complementing shooting moves with shifting moves and path reversal moves.[15] Other researchers have developed

techniques to enhance the effectiveness of the shooting approach by guiding the selection of configurations for shooting moves along the paths.[20, 10] For example, Chopra et al.[20] proposed an algorithm to enhance the local sampling of transition states by using committor probabilities to bias the selection of configurations for shooting moves and performing simulations in parallel.

The transition rate constant between two stable states A and B can be determined by a correlation function of state populations in time, $C(t)$.[15] If the rare transitions between the stable states A and B are separated by a single dynamical bottleneck, $C(t)$ can be defined as the conditional probability to find the system in final region B at time t provided it started in A at time t=0, i.e.,

$$C(t) = \frac{\langle h_A(x_0) h_B(x_t) \rangle}{\langle h_A(x_0) \rangle} \quad . \quad (3)$$

where $\langle ... \rangle$ denotes equilibrium ensemble averages and x_t is the set of coordinates specifying the state of the system at time t. Once $C(t)$ is determined for times greater than the characteristic time (τ_{mol}) required to cross the dynamical bottleneck and forget how it got from A to B, $k_{A\to B}$ can be extracted by taking the derivate of the correlation function $C(t)$:[14, 15]

$$k_{A\to B} \approx \frac{dC(t)}{dt} = \frac{d\langle h_A(x_0) h_B(x_t) \rangle / dt}{\langle h_A(x_0) h_B(x_{t'}) \rangle} \times C(t') \qquad (4)$$

which implies that the time-dependent derivate of $C(t)$ reaches a plateau for $\tau_{mol} < t << \tau_{rxn}$ [where $\tau_{rxn} = (k_{A\to B} + k_{B\to A})^{-1}$]. Hence, the average transition rate constant with TPS can be computed from $C(t')$ for a time t' that can be much smaller than t . $C(t)$ in the time interval $0 < t < \Gamma$ is determined by: (i) a single path sampling simulation to calculate $\langle h_A(x_0) h_B(x_t) \rangle$, and (ii) an umbrella simulation to get $C(t')$ for $t' << t$:

$$C(t) = \int_{\lambda^B_{\min}}^{\lambda^B_{\max}} d\lambda P(\lambda, t) \quad . \qquad (5)$$

where $P(\lambda,t)$ is the probability to find a time slice x_t at an order parameter λ, which is usually determined by dividing the phase space into a sequence of N+1 overlapping regions.

TPS has several shortcomings that led to the development of new path sampling methods. TPS requires knowledge of the initial state phase density, which means that the system must be in equilibrium. The initial trajectory is usually obtained by letting the transition occur spontaneously in a long brute force simulation.[15] However, this may not be practical for some complex systems and other alternative approaches have been used. For example, the initial path can be obtained by driving the system from A to B artificially from a brute force simulation at high temperature.[21, 22] In this case, the initial path is not representative of the TPE at the conditions of interest and TPS may lead to the collection of non-reactive trajectories.[16] Moreover, the rule described in Eq. (2) provides a method to sample only pathways of a fixed Γ length which limits the efficient implementation of TPS.[16] To overcome these limitations, improved TPS variants have been developed such as an approach where trajectories are sampled using flexible path lengths[8], and methods that employ a series of interfaces in phase space to facilitate the generation of transition paths.[23-25]

2.2. Interface-based Transition Path Sampling schemes

Transition interface sampling (TIS) was developed to estimate the kinetics of rare events with flexible path length by a TPS-like procedure.[24, 25] This method employs a series of nonintersecting interfaces (n+1) between the initial (A) and final (B) regions $\{\lambda_{0,}\ldots\ldots,\lambda_n\}$ by means of a suitable order parameter λ that increases monotonically as the interfaces $\lambda_{0,}\ldots,\lambda_n$ are crossed such that $\lambda \geq \lambda_0$, $\lambda_n = \lambda_B$, and $\lambda_i > \lambda_{i\text{-}1}$. The transition rate constant $k_{\mathrm{A}\to\mathrm{B}}$ is thus given by the product of two terms:

$$k_{A\to B} = \overline{\Phi}_{A,0} P\left(\lambda_{n=B} \mid \lambda_0\right), \qquad (6)$$

where $\overline{\Phi}_{A,0}$ is the effective flux of trajectories that leave the basin of attraction A through the first interface λ_0, and $P(\lambda_{n=B} \mid \lambda_0)$ is the probability that a trajectory that leaves A and crosses λ_0 will reach B before returning to A. Hence, Eq. (6) defines the rate constant of the process as an average rate of transitions from A to B using an "effective positive flux" expression.[5, 24, 25] $P(\lambda_{n=B} \mid \lambda_0)$ can be decomposed as a product of conditional probabilities:

$$P(\lambda_{n=B} \mid \lambda_0) = \prod_{i=0}^{n-1} P(\lambda_{i+1} \mid \lambda_i) \quad (7)$$

where $P(\lambda_{i+1} \mid \lambda_i)$ is the probability that a trajectory that visits A and crosses λ_i for the first time will subsequently reach λ_{i+1} without returning to the initial region A. Each of the factors in Eq. (7) is then computed in separate path sampling simulations for each pair of adjacent interfaces. For example, for the interface λ_i, the TIS approach consists of choosing a random slice on an existing trajectory that leaves A and crosses λ_i, reaches λ_{i+1} or returns to A. This time slice is modified (only for deterministic trajectories) and then used to generate a new partial trajectory using the shooting algorithm[15] (see Sec. 2.1). This new partial path is accepted in the TIS path ensemble belonging to interfaces *i* and *i*+1, if it leaves A, crosses λ_i, and either reaches λ_{i+1} or returns to A. Figure 2 shows a schematic illustration of the TIS procedure.

For systems involving diffusive barrier (and long trajectories), the partial path TIS (PPTIS) method[23] holds an advantage over TIS because the partial paths generated are much shorter, increasing the sampling efficiency. In contrast to the TIS method, PPTIS samples partial paths that only span one or two interfaces using the TIS framework. However, this approach assumes Markovian "memory loss" over subsequent interfaces to justify the shorter paths.[23] The crossing probabilities in Eq. (7) are thus calculated by a recursive relation in terms of single interface crossing probabilities which are defined depending in their starting and ending interfaces. An advantage of PPTIS method is that the forward and backward rate constants can be determined from a single path sampling simulation.

TIS and PPTIS have also been combined with replica exchange methods to enhance dramatically the path sampling efficiency.[26, 27] Rogal and Bolhuis[28] extended the formulation of the TIS algorithm to harvests trajectories connecting multiple stable states and obtain expressions for the rate constants of all possible transitions. Moreover, the TIS and PPTIS algorithm can be also used to calculate simultaneously the reaction rate constant and the free energy profile along a chosen order parameter.[29]

In the jargon of Monte Carlo (MC) methods, TPS-like methods (including TIS and PPTIS) are "dynamic" MC techniques because a new path is generated by proposing a "change" to an older one. It is also possible to formulate "static" MC methods where new pathways are generated from scratch; the method described next belongs to this category.

2.3. Forward Flux Sampling (FFS) schemes

FFS algorithms were developed to encompass stochastic nonequilibrium systems in which a-priori knowledge of the phase-space density is not required.[5, 30] FFS-type simulations are not limited to nonequilibrium systems and can use different types of stochastic dynamics, including Molecular Dynamics by incorporating a stochastic component in the trajectory.[31] Like TIS and PPTIS, FFS schemes allow the computation of both rate constants and TPE by dividing the phase space between the initial and final region into a series of interfaces. Distinctively, FFS does not require backward pathways hence eliminating the need for the backward shooting part of TPS or TIS; this is the feature that allows FFS to be applicable to non-equilibrium systems. The rare paths are generated such that any trajectory from A to B passes through each interface in turn. The transitions between interfaces are free to follow any possible path between A and B, including paths crossing previous interfaces several times, as illustrated in Fig.3. Although a series of interfaces in the phase-space is used as in other sampling techniques, a long decorrelation time does not present an issue because FFS is no treated as a MC Markov chain.[7] Hence, trajectories are generated without assuming a steady-state distribution (or "memory loss" during the transition) at each interface.[29]

The rate constant for the system's transition between stable states is also computed using the "effective positive flux" expression in Eq. (6). The main difference between FFS and (PP)TIS is the way that the crossing probabilities are computed. At the present, three path sampling schemes have been proposed to generate the TPE: the Direct Forward Flux Sampling (DFFS), Branched Growth method (BG), and Rosenbluth method (RB).[5] The DFFS and BG schemes, the simpler and arguably more efficient ones,[5] will be discussed in the following subsections. Allen et al.[4, 5] have proposed several extensions of the FFS formalism to improve the efficiency of all three methods. For instance,

the computational expense of simulating "failed" trial from λ_i all way back to A can be reduced by pruning trial paths which go toward A (i.e., trials paths reaching λ_{i-1} from λ_i are terminated with certain probability). Extensions of the FFS algorithm have also been reported to describe pathways connecting the two stable states through multiple intermediates states.[32]

The FFS method can be also used for a simultaneous calculation of the kinetic data and the underlying stationary probability distribution of the system. For example, Valeriani et al.[33] computed stationary distributions $P(\lambda)$ (i.e., the "free energy" profile for equilibrium systems or the steady-state probability distribution for nonequilibrium systems) along an order parameter λ by performing two FFS simulations to obtain the rate constants for the forward and backward transitions. These rates are then used to reweigh contributions to $P(\lambda)$ from trajectories originating from both region A and region B. A complementary method[32] that is restricted to equilibrium systems (denoted FFS-US) entails an umbrella sampling (US) simulation following an original FFS protocol (pre-optimized for order parameter[34] and staging[35]) to sample the regions inside each window until the partial path ensemble loses any "memory" of where it originated.

Compared to other methods to obtain the TPE and rate constants, the main advantages of FFS are arguably its simplicity and its ability to describe not only equilibrium but also non-equilibrium systems. On the other hand, in applications to complex systems the efficiency and accuracy of FFS can be sensitive to:[4, 5, 7,34, 35] (i) the choice of order parameter, and (ii) the positions of the interfaces and the extent of sampling of the interface ensembles, in particular for the first interface. Such a sensitivity of FFS is partially due to a tendency of propagating sampling errors that is larger than that of other interface-based path sampling methods. These points are discussed in more detail in Sec. 2.5.

2.3.1. Direct Forward Flux Sampling (DFFS)

The DFFS scheme is a two-stage algorithm illustrated in Fig. 3. The first stage entails the evaluation of the flux $\overline{\Phi}_{A,0}$ in Eq. (6); this is done (as with TIS) by a simulation in basin A and

computing the ratio of the total number N_0 of crossing configurations at λ_0 to the total length Γ of this run. Each such state at λ_0 (whose phase space coordinates are saved) corresponds to a configuration that crosses λ_0 in a trajectory coming from A; i.e., the trajectory has to return to A between consecutive stored points at λ_0.

In the second stage of the algorithm, consecutive path sampling simulations are performed for each interface λ_i to get conditional probabilities, $P(\lambda_{i+1} \mid \lambda_i)$, of reaching λ_{i+1} from λ_i [see Eq. (7)]. Partial paths are thus generated from the collection of stored points at λ_0 (N_0) by firing M_0 trials runs, which are continued until either reaching λ_1 or returning to the initial region A. Each trial run starts from a random point selected from the N_0 points at λ_0. If no successful trial runs were generated at λ_1 (i.e., $N_S^{(0)} = 0$), the procedure is stopped and $P(\lambda_{n=B} \mid \lambda_0) = 0$. Otherwise, all end point configurations $N_S^{(0)}$ that reached λ_1 are stored and used as starting points for M_1 trial runs toward λ_2 (or back to A). If $N_S^{(1)}>0$ of the trial runs reach λ_2, the partial paths are continued by initiating M_2 trials runs to λ_3, randomly chosen from the $N_S^{(1)}$ successful configurations. This procedure is repeated until either the final region $\lambda_n=\lambda_B$ is reached or no successful trials were generated at some intermediate interfaces. The conditional probabilities in Eq. (7) for each interface are estimated from:

$$P(\lambda_{i+1} \mid \lambda_i) = N_S^{(i)} / M_i \qquad (8)$$

The correctly weighted TPE can be extracted from the phase-space coordinates of the interfacial points of the system along all trial runs which successfully reach λ_{i+1} from λ_i, and the information on the connectivity of the partial paths. The characteristic transition paths are thus obtained beginning with the collection of trials which arrive at $\lambda_B=\lambda_n$ from λ_{n-1} and tracing back the sequence of connected partial paths which link them to region A. As illustrated in Fig. 3, the TPE results in a "branching tree" of transition paths, in which partial paths between interfaces close to A may be shared by many members of the TPE.

2.3.2. The Branched Growth (BG) Sampling Method

The BG method is illustrated schematically in Figure 4, where branched transition paths are generated one by one. In the first stage of the algorithm, the flux $\overline{\Phi}_{A,0}$ in Eq. (6) is obtained as described for the DFFS method in the previous section. In the second stage of the algorithm, branched paths are generated from the stored configurations at λ_0 [gray circles at λ_0 in Fig. 4] and the conditional probabilities $P(\lambda_{i+1} \mid \lambda_i)$ are estimated. To do so, the BG method samples k_i trial runs per stored point at λ_i, rather that sampling M_i randomly selected points at λ_i as with DFFS. The branched path is started by selecting randomly a state at λ_0 from which k_0 trial runs are fired and continued until either reaching λ_1 or returning to the initial region. Each end point configuration $N_S^{(0)}$ resulting from a reactive partial path (i.e., reaching λ_1) is stored and used as starting point for k_1 trial runs. If $N_S^{(1)}>0$ of those trial runs reach λ_2, the branching tree path is continued by initiating k_2 trials runs to λ_3 from each of the $N_S^{(1)}$ successful configurations. This procedure is repeated until either the final region $\lambda_n=\lambda_B$ is reached or because no successful trials were generated at some intermediate interfaces. An estimate of the conditional probabilities to jump one interface is given by:

$$P(\lambda_{i+1} \mid \lambda_i) = \frac{N_S^{(i)}}{k_i N_S^{(i-1)}} \quad . \qquad (9)$$

The same procedure described before is used to create a new branching tree starting from a different randomly chosen point at λ_0. After many such branched paths have been generated, final estimates for $P(\lambda_{i+1} \mid \lambda_i)$ and $P(\lambda_{n=B} \mid \lambda_0)$ [via Eq. (7)] are obtained using N_s data from all the paths. The TPE is obtained just as was described before for the DFFS. Note that for the BG, M_i , the total number of trial runs fired at λ_i (from each starting point at λ_0) is given by:

$$M_i = k_i \prod_{j=0}^{i-1} k_j P(\lambda_{j+1} \mid \lambda_j) \qquad (10)$$

2.3.3. Computational efficiency of FFS-type schemes

Allen et al. [4] estimated the computational efficiency (ε) for FFS-type simulations as the inverse of the product between the computational cost C and the relative statistical error ν in the estimated value $k_{A\to B}$ of the rate constant per starting point at λ_0:

$$\varepsilon = 1/[C\nu] \qquad (11)$$

Following Eq. (6), the variance, V, in the estimate of $k_{A\to B}$ depends on the relative variance of both $\overline{\Phi}_{A,0}$ and $P(\lambda_{n=B} \mid \lambda_0)$. Assuming that the error in $\overline{\Phi}_{A,0}$ could be ignored, the V in $k_{A\to B}$ is approximated to be:

$$V[k_{A\to B}] \propto V[P(\lambda_{n=B} \mid \lambda_o)]. \qquad (12)$$

Assuming that successful trial runs at λ_i ($N_S^{(i)}$) are uncorrelated, the statistical error ν is

$$\nu \approx N_0 \frac{V[P(\lambda_n \mid \lambda_0)]}{P(\lambda_n \mid \lambda_0)^2} \qquad (13)$$

where N_0 is the number of starting points at λ_0. If $V[P(\lambda_{i+1} \mid \lambda_i)]$ is the variance in the estimates of $P(\lambda_{i+1} \mid \lambda_i)$, then propagating errors from Eq. (7) leads to

$$V[P(\lambda_n \mid \lambda_0)] = P(\lambda_n \mid \lambda_0)^2 \sum_{i=0}^{n-1} \frac{V[P(\lambda_{i+1} \mid \lambda_i)]}{P(\lambda_{i+1} \mid \lambda_i)^2} \qquad (14)$$

Allen et al.[4] defined the computational cost [factor C in Eq. (11)] as the average number of simulation steps required by a particular FFS-type sampling scheme per starting point at λ_0. Hence, ignoring any other contributions to the CPU time, the average cost is approximated by:

$$C = R + \frac{1}{N_0} \sum_{i=1}^{n-1} M_i C_i \qquad (15)$$

where R is the average cost of generating one starting point at λ_0 and C_i stands for the cost of firing one trial run from interface λ_i, which is approximated by:

$$C_i = S\{P(\lambda_{i+1} \mid \lambda_i)[\lambda_{i+1} - \lambda_i] + (1 - P(\lambda_{i+1} \mid \lambda_i))[\lambda_i - \lambda_A]\} \qquad (16)$$

where the average length of a partial trajectory from λ_i to λ_j was assumed to be linearly proportional to $\left|\lambda_j - \lambda_i\right|$, with a proportionality constant *S*.[4]

For the DFFS method, $P(\lambda_{i+1} \mid \lambda_i)$ is found from Eq. (8), where $N_s^{(i)}$ can be modeled using the binomial distribution so that Eq. (14) becomes:[4]

$$V^{DFFS}\left[P(\lambda_n \mid \lambda_0)\right] = P(\lambda_n \mid \lambda_0)^2 \sum_{i=0}^{n-1} \frac{1 - P(\lambda_{i+1} \mid \lambda_i)}{M_i P(\lambda_{i+1} \mid \lambda_i)} \qquad (17)$$

The computational cost for DFFS simulation C^{DFFS} is approximated by substituting equation (16) into Eq.(15), and taking into account the possibility that the cost is reduced by failing to reach later interfaces (i.e., as in the case that M_i is small):

$$C^{DFFS} = R + \frac{1}{N_0}\left\{ M_0 C_0 + \sum_{i=1}^{n-1}\left[M_i C_i \prod_{j=0}^{i-1}\left[1 - \left(1 - P(\lambda_{j+1} \mid \lambda_j)\right)^{M_j}\right]\right]\right\}. \qquad (18)$$

Finally, equations (13), (17) and (18) can be substituted into Eq. (11) to give the complete expression for the computational efficiency of DFFS simulations.

For the BG method, the variance V^{BG} in the estimated value of $P(\lambda_{n=B} \mid \lambda_0)$ is now: [4]

$$V^{BG}\left[P(\lambda_n \mid \lambda_0)\right] = \frac{P(\lambda_n \mid \lambda_0)^2}{N_0} \sum_{i=0}^{n-1} \frac{1 - P(\lambda_{i+1} \mid \lambda_i)}{\prod_{j=0}^{i} k_j P(\lambda_{j+1} \mid \lambda_j)} \qquad (19)$$

The computational cost for a BG simulation C^{BG} per starting point at λ_0 is estimated by substituting Eqs. (10) and (16) into Eq. (15):

$$C^{BG} = R + k_0 C_0 + \sum_{i=0}^{n-1}\left[k_i C_i \prod_{j=0}^{i-1} P(\lambda_{j+1} \mid \lambda_j) k_j \right]. \qquad (20)$$

Finally, equations (13), (18) and (20) can be substituted into Eq. (11) to get the computational efficiency of BG simulations.

2.4. Other Related Interface-based Methods

2.4.1. Weighted-Ensemble Method.

Motivated by the difficulty of estimating reaction rate constants for systems with diffusion-controlled transitions, Huber and Kim[36] developed a path sampling approach that attempts to maintain a measurable steady-state flux along a "progress" coordinate by dynamically controlling the distribution, in configurational space, of a weighted ensemble (WE) of trajectories. Each resulting unbiased transition trajectory represents a juxtaposition of multiple pathways that either divided or merged, according to their weights, when progressing through sequential reaction regions (slabs); the statistical path ensemble thus obtained yields asymptotically the correct TSE and reaction rate constants. Because of the subdivision of the reaction progress into slabs in configuration space, it has been noted[37] that the WE method could also be considered the precursor of other interface-based transition pathway formulations, e.g., of the branched growth FFS reviewed above.

In the original formulation,[36] the WE was used with Brownian dynamics (BD) to address the time-scale problem of rare diffusional events by applying a pruning-enriching procedure to an ensemble of independent Brownian pseudoparticles. These pseudoparticles, in configurational space, represent different partial trajectories of the system which must be pruned (merged) or enriched (cloned) according to their progress from the reactant to the product basins. The individual state of each pseudoparticle is monitored by subdividing a coordinate that measures the reaction progress into several slabs (bins), and the number of pseudoparticles distributed throughout each bin, n, is maintained approximately constant by dynamically pruning and enriching them and their weights. This assures that slabs of the configurational space from which the pseudoparticles diffuse rapidly (i.e., in the neighbourhood of transition states) are sampled uniformly with respect to those near the basins (where pseudoparticle diffusion relaxes slowly), and that the statistics of probability fluxes across their interfaces are measured accordingly. A schematic illustration, shown in Fig. 5, depicts the WE dynamics for a simple two-state system.

The WE method represents an improvement of the flux-over-population approach of Farkas,[38, 39] and also over the method of Northrup-Allison-McCammon,[40] for problems with long diffusive transitions. For instance, the WE calculation of the reaction rate for the binding of a monoclonal antibody (NC6.8) to a hapten (a case mediated by a large barrier) was shown to be about eight times more efficient than standard BD under the formulation of Northrup-Allison-McCammon.[41, 42] For systems without significant barriers, as that of the homodimerization of CuZn superoxide dismutase, the use of WE led to negligible computational gains but yielded results in agreement with experiment and with standard BD.[42] Similarly, the WE method has been used to study the conformational transition of a "double-native" model of calmodulin's N-terminal domain,[43] from a calcium-bound to an unbound state, and found that it yielded increasingly larger efficiency gains, when compared to brute-force simulations, as the energetic barrier is increased (i.e., lowering the temperature). For this case, a decrease of 20% in temperature led to an increase in efficiency from 1 to 3 orders of magnitude. Even larger gains in terms of CPU time, of about 5 orders of magnitude compared to brute-force BD, were predicted in relation to the folding time of a four-helix bundle protein within the framework of the diffusion–collision model.[44] The WE has also allowed accessing the long time-scales of the electrostatically steered homodimerization of hemoglobin and an exhaustive scrutiny of the role of polar residues in the association mechanism.[45-47, 42]

2.4.2. Milestoning

The milestoning method[48, 49] attempts to capture the complete dynamics of long time scale phenomena occurring over a reactive pathway by first subdividing it into transition interfaces or milestones (i.e., coarse graining), and then piecing together the contributions of the local microscopic dynamics at each slice. This is shown schematically in Fig. 6a. The milestones are hypersurfaces H_i that split the phase space along a reaction pathway, i.e., between the reactant (A) and product (B) basins. If this partitioning is possible, then the statistical properties of the sequence of milestones $\{H_i\}_{i=1,2,..,M}$ can be obtained by starting short MD trajectories at initial configurations with phase space coordinates

$X \in H_i$, where X is distributed according to either the equilibrium[48, 49] or the first hitting point probability densities.[50] Each MD trajectory integrated for a time τ, contributes to the local, forward $K_i^+(\tau)$ or backward $K_i^-(\tau)$, distributions of pausing times according with the corresponding ending interface, H_{i+1} or H_{i-1} respectively; where τ is the time spent (incubation) in the neighbourhood of H_i before first hitting any of the adjacent hypersurfaces.

The evolution of the system is computed following the Montroll-Weiss continuous-time random walk [51], in which $Q_i(t)$, the probability that the random walk reaches H_i at time *t*, is calculated by a probability balance considering the earlier arrival to the neighbouring hypersurfaces H_{i+1} and H_{i-1}, and the subsequent transition from these to H_i, therefore

$$Q_i(t) = \eta_i \delta(t-0) + \int_0^t \left[K_{i+1}^-(t-t')Q_{i+1}(t') + K_{i-1}^+(t-t')Q_{i-1}(t') \right] dt', \qquad (21)$$

where η_i is the initial milestone probability distribution and t' is the arrival time to the adjacent hypersurfaces. Equation (21) allows to compute the time evolution of the global probability $P_i(t)$ of having crossed H_i and remaining between H_{i+1} and H_{i-1} at time t, through the integral equation

$$P_i(t) = \int_0^t \left(1 - \int_0^{t-t'} K_i(\tau) d\tau \right) Q_i(t') dt', \quad (22)$$

where $K_i(\tau) = K_i^+(\tau) + K_i^-(\tau)$; note that the term in the parenthesis is the probability of waiting in the neighbourhood of H_i and not leaving before t', expressed in terms of the probability distribution of pausing times $K_i(\tau)$.

Within the milestoning approximation, the problem of estimating the transition rate constant between two stable states can be reduced to an initial value problem with absorbing boundary at H_{B}, i.e., a first-passage time problem.[52] The mean first-passage time can be estimated from the first moment of the first-passage time distribution, $\varphi(\tau) = dP_f(t)/dt_{t=\tau}$, as

$$\tau_{MFPT} = \int_0^\infty t\varphi(t)\mathrm{d}t, \qquad (23)$$

where the transition is from H_i at $t = 0$ to the absorbing boundary milestone H_f whose $K_f^{\pm}(\tau) = 0$. Then, the transition rate constant is simply $k_{\mathrm{AB}} = 1/\tau_{MFPT}$.[39]

For biologically systems, milestoning has been applied to the α-helix to β-sheet transition pathway of the alanine dipeptide in aqueous solution,[49] and to the allosteric transition of the *Scapharca* hemoglobin (HbI).[53] The transition minimum energy path required to define the sequence of milestones $\{H_i\}_{i=1,2,\ldots,M}$ was derived without explicit calculations for the alanine dipeptide, whereas for the HbI a preliminary estimation was required. In both cases, significant computational gains were obtained with respect to direct MD simulations of about 1 and 3 orders of magnitude, respectively.

2.4.3. String Method

This approach, in its finite temperature formulation (FTS),[54] shares a fundamental similarity with the previously discussed interface-based methods, in that it is based on the idea of studying the transition kinetics as a stochastic process across hypersurfaces orthogonally defined along the reaction pathway; although it differs from milestoning, for example, in that it is effectively a path sampling algorithm defined within the framework of the transition state theory (TST).[55, 56] In its limiting formulation, the zero temperature string method (ZTS),[57] the approach is reminiscent of the nudge elastic band method[58, 59] because it searches the minimum energy pathway (MEP) and saddle point along a smooth curve φ^* (*string*) connecting two potential minima; the hypersurfaces in this limit become points or images on the *string* which dynamically evolve to satisfy the condition of null potential gradient $(\nabla V)^{\perp}$ normal to φ^*.[57] This latter limiting case will not be discussed further as it does not considers the TPE.

In the FTS approach, schematically illustrated in Fig. 6b, the MEP after N iterations is no longer described by a single pathway but by the mean transition path φ_N of a *transition tube* containing the most probable transition trajectories.[54, 60] The process of finding such trajectories is directly

linked to the problem of finding the hypersurfaces that rigorously define the reaction coordinate (RxC), i.e., isoprobability surfaces with constant commitor probability ($p_{\mathrm{B}}(x) = z \in [0,1]$). Transition tubes are therefore defined by the hitting probability distribution of reactive pathways crossing the isocommitor surfaces, which in turn are stochastically characterized, within the framework of a Markov process, by the backward Kolmogorov equation.[61] From the Markovian property and statistical time reversibility, it follows that the distribution of points where the trajectories hit the isocommitor surfaces (i.e., the transition tube in configuration space), is equivalent to the equilibrium distribution of trajectories weighted on the isocommitor surface. This result has important methodological consequences because it allows to simplify the high dimensional space of the backward Kolmogorov equation within a variational formulation of $p_{\mathrm{B}}(x)$. Under this approximation, the resulting isocommitor surfaces $\Gamma_z = \{x : p_{\mathrm{B}}(x) = z\}$, shown in Fig. 6b,are reduced to the family of hyperplanes $\{P_\alpha, \alpha \in [0,1]\}$ orthogonal to the mean string path $\varphi(\alpha)$ of the transition tube.[62]

Since the transition pathways are defined in configuration space instead of physical time (as in strictly TPS-based methods), this formulation has an essential numerical advantage over others based on TPS: the isocommitor surfaces can be specified initially without prior knowledge or integration of the transition trajectories.[60, 62] A variational formulation is also possible for a large set of collective degrees of freedom, where the FTS method yields the isocommitor surfaces that minimize the mean free energy path (MFEP)[63, 64].

The FTS method has been tested extensively on the alanine dipeptide problem, both in vacuum and in aqueous solution.[9-11] The isocommitor surfaces obtained by FTS at the transition state region between the metastable conformers C_{7eq} and C_{7ax}, over the (ϕ, ψ)[9, 11] and (ϕ, ψ, θ, ζ)[10] dihedral angles, were estimated in accord with the commitor distribution computed directly on the hyperplanes; this is obtained with a lower computational cost than other TPS-based methods.[12, 13] The FTS in collective variables has been also implemented[14] to characterize the collapse mechanism of a hydrophobic chain in water;[15] a process with significant relevance to understand the assembly of

multiple protein fragments and folding of single proteins in aqueous solutions. The MFEP computed with the string method, in agreement with more expensive MD calculations, confirmed the collapse pathway previously proposed for this process using a coarse-grained description of the solvent.[15]

2.5. Selecting and sampling the order parameter λ

Path sampling methods usually have several drawbacks that are worth mentioning. For instance, the efficiency and completeness of the TPE sampling depends on the quality of the order parameter λ. In this sense, FFS-type methods are more dependent on a good order parameter than methods such as TIS and TPS (which only uses λ to distinguish the stable states of the system).[7] In the case of FFS-type simulations, the rate constant estimate and the sampled pathways will depend strongly on the quality of the ensemble of starting points at the first interface λ_0. If this ensemble is under-sampled, errors will propagate through successive interfaces. This does not happen in TIS where each individual interface ensemble eventually converges.[7] By its "static" nature, FFS does not allow pathways to relax the portion before the current interface. Moreover, even though a good order parameter could be determined and used to partition the phase-space, the efficiency of the sampling is still sensitive to the number and position of the interfaces and to how extensively different interfaces are sampled.[4, 5, 34, 35] Hence, in general, the two main challenges to address in interfaces-based path sampling simulations are: (i) the determination of a good order parameter and (ii) the optimization of interface sampling.[34, 35]

An adequate order parameter λ should be a variable that quantifies progress along the reaction pathways and permits to discriminate between the stable states. A clear distinction should be made between the concepts of reaction coordinate (RxC) and a good order parameter. The former is able to tell how far a "reaction" has proceeded and should be able to identify whether a particular configuration belongs to the transition state ensemble (TS). While a good order parameter should be able to tell whether a reaction has just started or is about to be completed, the RxC is a dynamical variable that measures the complete progress of the reaction from start to finish. Therefore, a good RxC (i.e., one that

approximates well the true RxC) can serve also as a good order parameter, but the inverse is not necessarily true. As stated in the introduction, the RxC is closely related to the probability of a configuration x to commit to the final state B; i.e., the "committor probability" $p_B(x)$, which quantifies the tendency of configuration x to relax to the basin of attraction B under the system's intrinsic dynamics.[10] A schematic view of the procedure for the determination of committors is given in Fig.7. Configurations in the initial basin A have $p_B = 0$, those in basin B have $p_B = 1$, and those at the TS have $p_B = ½$. Hence, p_B can be seen as a perfect RxC in the sense that it provides a quantitative description of the dynamic behavior of every state along a trajectory. Indeed, any good RxC should parametrize the committor such that its distribution $P(p_B)$ for configurations with the same RxC value should be sharply peaked around a characteristic p_B value.[9] On the other hand, a poor RxC leads to non unimodal $P(p_B)$ as configurations with the same value of the RxC can have very different p_B values.

While the committor p_B can be taken itself as the perfect order parameter for path sampling simulations, to be of practical use, p_B (i.e. $\lambda(\boldsymbol{q}(x))= p_B$) should be related to a few collective variables (themselves functions of the configurations) that encapsulate many atomistic details into physically important properties.[9] Here, $\boldsymbol{q}(x)=q_1, q_2, ..., q_m$, refers to a set of m collectives variables, which are considered potentially useful descriptors for the isocommittor surface. Since a model for the RxC corresponds to the p_B surface response, the TSE and hence the mechanistic details of the process can be readily obtained by only analyzing characteristics of the collective variables at the p_B contour of $½ \pm \sigma$ (where σ is the desired level of statistical accuracy).[10] In high-dimensional complex systems, however, it is not a trivial task to find a good RxC. Several committor-based analysis methods have been proposed, including conventional committor analysis,[15, 10] Bayesian path statistics,[65] genetic neural networks,[9] string method,[63, 66] likelihood maximization,[67, 10] and FFS- least square estimation.[34]

As illustrated in Fig. 7, in the conventional committor analysis,[15, 10] a minimum number of fleeting trajectories, N_{min}, are initiated from a starting configuration along one of the paths belonging to

the TPE. The committor probability p_B is therefore estimated from the fraction of paths that end in state B. A statistical analysis carried out by Peters and Trout[10] indicates that good statistics require hundreds of estimates for p_B histograms and analysis of ≥ 100 trajectories in the TPE. Unfortunately, each p_B estimate requires on the order of N_{min} =10 fleeting trajectories, each half as long as a reactive trajectory.[10] Hence, the difficulties and computational cost of the committor analysis have motivated recent attempts to systematize the search for RxCs. For example, Hummer[68, 65] used Bayesian path statistics to introduce a new criterion for the TS as those points in configurational space with high probability p(TP|x) that equilibrium trajectories passing through them are reactive (i.e., connect stables states). Hence, this definition of p(TP|x) gives higher values to those configurations (x) in the TPE which are common to many transition paths but are rarely visited in equilibrium. The projection of p(TP|x) onto a good RxC should thus give a sharply peaked distribution, which can be used to choose among different candidate order parameters,[9] but it requires costly estimation of a p(TP|x) histogram for each iterative improvement of RxC model. Ma and Dinner[9] proposed a method based on neural networks (denoted GNN-method) to determine the functional dependence of p_B on a set of coordinates, and a genetic algorithm that selects the combination of inputs that yields the best fit via the estimation of p_B histograms.[69-71] The collection of configurations for the database is taken from transitions pathways, for instance harvested with TPS. Although the GNN-method provides the advantage that a very large pool of possible RxCs can be efficiently searched without the need to redo the path sampling (with an improved order parameter), the computational cost for the calculation of the committor values for a sufficiently large database is very expensive. Maragliano et al.[63] combined a string method with TPS to determine the MFEP. Their approach presumes that transitions are most likely to occur around the MFEP and thus isocommittor surfaces are determined therein. However, this approach requires many iterations of the mean force and variable entanglement calculations.[72, 66] The majority of these committor-based methods use expensive histograms calculations, and mainly focus on improving the trial and error aspects of the p_B calculation.[7, 10] This has led to the development of statistical

approaches capable to determine commitment probabilities "on-the-fly" as the pathways are generated.[34, 67, 10]

2.5.1. Likelihood Maximization

Peters and Trout[10] proposed the Maximum Likelihood (ML) method which, like the GNN method, screens a set of candidate collective variables for a good RxC estimate that depends on a few relevant variables but it does not require expensive calculation of commitment probabilities. A simple model for the RxC; e.g., a lineal combination of the collective variables, is assumed and used to calculate the likelihood of the model given the shooting data. The data are built on information about the accepted and rejected shooting moves accumulated during a TPS simulation. This is achieved by using a method denoted "aimless shooting" to harvest independent realizations of p(TP|x) or p_B. This aimless shooting algorithm differs from standard shooting algorithm (illustrated in Fig. 1) in: (i) the new momenta at configuration x (where x is the randomly selected state from which the shooting move is attempted) need to be drawn from the Maxwell-Boltzmann distribution rather that obtained by small perturbation of the old momenta, and (ii) the shooting points are selected from a small region around the previous shooting points rather than from the whole path. This procedure leads to an ensemble of shooting points that form a normal distribution peaked near the TS. From the TPS simulation, the configurations of the shooting points are stored together with the information on whether the trajectories were reactive (connecting A to B) or not. Then, M collective variables, $\boldsymbol{q}(\mathrm{x}) = q_1,\ldots, q_M$, are calculated for each of the collected shooting points and then used to calculate the likelihood of an assumed RxC model (q) in term of these physical properties:

$$q(x) = \alpha_0 + \sum_{k=1}^{M} \alpha_k q_k + q^T \mathbf{A} q \qquad (24)$$

Here, $\alpha_j, j=0,1,\ldots,m$, are the fitting coefficients and absorb the units from the collective variables, so that $\boldsymbol{q}(x)$ is dimensionless. Interactions between collective variables are included by the cross quadratic term in Eq. (24) where **A** is a matrix of adjustable parameters.

The outcome of each shooting move is viewed as a particular realization of the process whose statistics is described by $p(\mathrm{TP}|q)$: the probability to be on a transition path given a particular value of the RxC. Therefore, $p(\mathrm{TP}|q)=p_B(x)$ is the PDF model for a good RxC, which depends on the M collective variables, $\boldsymbol{q}(x)$;[10] a general form for the model that is peaked at the value of q corresponding to the TS and decays to zero away from the peak [65] is:

$$p(TP \mid q) = p_0[1 - \tanh^2(q)], \qquad (25)$$

where p_0 is an adjustable parameter. The Bayesian information criterion is then used to determine significant variables for the RxC. To this end, a likelihood function which quantifies the probability of the observed data as function of the model parameters is constructed:

$$L(\alpha) = \prod_{x \in acc} p(TP \mid q(x)) \prod_{x \in rej} [1 - p(TP \mid q(x))]. \qquad (26)$$

The products extend over all accepted (*acc*) and rejected (*rej*) shooting points. The log likelihood [in Eq. (26)] is then maximized to obtain the optimal parameters (α) that yield the best RxC model.

Because the likelihood function in Eq. (26) uses information over the full range of $p(\mathrm{TP}|q)$ values, the RxC applies at every $p(\mathrm{TP}|q)$ along the trajectories. Once the shooting data are obtained from the TPS-like procedure, extension of the RxC to include more collective variable does not require any substantial additional computational effort. Peters et al.[67] have reported further improvements of the original ML method.

2.5.2. FFS- least square estimation (FFS-LSE) method

The FFS-LSE approach is related to the ML method [10] for obtaining RxC models. However, the two main differences are that FFS-LSE uses a different method for sampling p_B (FFS rather than TPS) and for finding the model (LSE rather than ML). The FFS-LSE formalism starts by harvesting an ensemble of typical trajectories for the transition from a FFS-type simulation (usually the BG scheme) using an initial guess for the order parameter. An approximate value of p_B for each of the points crossing the interfaces is extracted from their path connectivity and then used to fit a model for the RxC

in terms of several collective variables, $\boldsymbol{q}(\boldsymbol{x})=q_1, q_2, ...,q_m$. Standard least-squares estimation (LSE) is used to find the coefficients of the model and an analysis of variance (ANOVA) is used to determine the significant terms in the model.

Figure 8(a) illustrates schematically the procedure to obtain p_B history from a BG simulation. During the BG run, the phase-space coordinates for all points along all the trial runs which successfully reach λ_{i+1} from λ_i are stored together with information on the connectivity of the partial paths. From these data, the p_B values of any stored point j at λ_i, $p_{B_j}^{\ i}$, can be estimated from the $p_{B_j}^{\ i+1}$ values of all connecting points at λ_{i+1} using recursively:

$$p_{B_j}^{\ i} = (1/k_i)\sum_{m=1}^{N_j^i} p_{B_m}^{\ i+1} \qquad (27)$$

Hence, $p_{B_j}^{\ i}$ values are obtained by following the trials that reached B (where $p_{B_j}^{\ n} = 1$) back to λ_{n-1}, then following their connected partial paths back to λ_{n-2}, and so on back to A. In this way, the FFS-LSE method obtains "on-the-fly" estimates for the p_B history from an FFS simulation with an initial guess for the order parameter. In a second stage, an improved order parameter is obtained from the collected p_B data as follows. At each stored point where the p_B value was estimated, m candidate collective variables are evaluated and a simple model for the RxC is assumed:

$$\lambda(q) = p_B(q) \approx \sum_{k=1}^{m} \alpha_k q_k + q^T \mathbf{A} q + \alpha_0 \qquad (28)$$

where the regression parameters in α_j and matrix **A** have a similar meaning as that in Eq. (24). These coefficients are determined by standard LSE. An ANOVA is then performed to check the adequacy of the model fit by determining if there is a statistically significant correlation between the response variable p_B and a subset of the q_k collective variables, and identifying the variables whose coefficients are significant. These variables are used to construct a simpler RxC model which is subjected again to LSE and ANOVA analysis.

The FFS-LSE protocol can be iterated, so that the current best model is used as the λ parameter in a new FFS run to generate additional p_B data to be LSE-fitted to the model of Eq. (28) and get an even better estimate of the RxC, and so on. Moreover, the DFFS scheme (rather than BG) could also be used to get p_B estimates if one only considers points for which a minimum number of trials runs have been fired.[34] Figure 8(b) shows results from the application of the FFS-LSE to a particle moving (via BD) on a rugged 2D potential energy surface. This surface is formed by the superposition of 109 Gaussians functions, [20] leading to two well defined global minima (basins A and B, respectively) and three local minima (metastable states). After iterating the FFS-LSE method,[32] it led to a model of the form (28) with several significant high-order terms; i.e., $p_B \approx 0.2\text{-}1.68x\text{-}1.78y\text{-}0.06xy+6.06x^2+6.25y^2\text{-}4.1x^3\text{-}4.17y^3$. Such high order terms were needed in this case to capture the curvature and complex topology of the p_B iso-surfaces which, as shown in Fig. 8(b) exhibit several disconnected domains.

2.5.3. Sampling optimization for FFS

Even when a good order parameter is used to partition the phase-space in interface-based methods, the efficiency of the sampling is still sensitive to the number and position of the interfaces and to how extensively different interfaces are sampled.[4, 35] Adaptive algorithms have been proposed to optimize the λ sampling for either the number and position of the interfaces (i.e., optimized λ phase staging), and/or the number of fired trial runs per interface.[35] Please refer to Figs. 3, 4, and 8 for basic definitions.

i. Approach 1: Optimizing the {λ} set (staging).

Optimizing the position of the starting interface.

The suitable positioning of the first interface λ_0 is crucial: if λ_0 is too close to the initial basin then the crossing points (which serve as starting points of all trajectories) are abundant but highly correlated, while if λ_0 is too far, then crossing points are well uncorrelated but too costly to generate. If too few uncorrelated starting points are used; i.e., if the ensemble of states at λ_0 is under-sampled, errors will

propagate through the next interfaces and lead to erroneous transition rate constants. Ideally then, the ensemble of stored configurations should be uncorrelated and distributed over all the phase space sampled by the characteristic pathways; this can be seen as minimizing the cost term "*R*" in Eq. (15). For this purpose, an observable property *y* is identified, whose values can be taken as providing a measure of phase space change that is nearly "orthogonal" to that provided by λ. The correlation of a set of *N* measurements of *y* for states at λ_0, can be estimated from an autocorrelation function; e.g.:

$$\mathrm{ACF(lag)} = \sum_{i=1}^{N-lag} \frac{\left(y_i - \overline{y}\right)\left(y_{i+\mathrm{lag}} - \overline{y}\right)}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2} \quad . \quad (29)$$

where $\overline{y}$ is the average for the complete run in basin A and the *lag* is the separation between stored states (in units of number of consecutive states at λ_0). Because this ACF should decay exponentially, i.e., $\mathrm{ACF(lag)} \propto \exp(-\mathrm{lag}/\tau_{\lambda_0})$, the constant τ_{λ_0} provides a measure of the autocorrelation time at λ_0. The simulation time required to obtain an uncorrelated state at λ_0 is approximated by $\tau^* \propto \tau_{\lambda_0} \times \Delta t_{\lambda_0}$ where $\Delta t_{\lambda_0} = 1/\overline{\Phi}_{A,0}$ is the average simulation time required to reach consecutive points at λ_0 [$\overline{\Phi}_{A,0}$ is was defined in Eq. (6)]. By minimizing τ^*, one reduces the simulation time (in basin A) needed to obtain a preset number of uncorrelated points at λ_0. The optimum location of λ_0 can then determined by the minimum of the τ^* versus λ_0 curve (this curve can be constructed from a *single* run in basin A).[73]

Optimizing the position of subsequent interfaces

Assuming that λ_A=λ_0 and λ_B are fixed, the variance in $P(\lambda_{n=B} \mid \lambda_0)$ can be reduced by minimizing Eq. (17) and (19) for the DFFS and BG scheme, respectively; with the constraint that $P(\lambda_{n=B} \mid \lambda_0) = \prod_{i=0}^{n-1} P(\lambda_{i+1} \mid \lambda_i)$ [i.e., Eq. (7)] must remain constant. This leads to: [35]

$$M_i P(\lambda_{i+1} \mid \lambda_i) = N_s^{(i)} = N_s = \text{constant.} \qquad (30)$$

i.e., a constant flux of reactive trajectories between interfaces is desirable. Equation (30) is applicable to both DFFS and BG schemes even though the objects to minimize are different. Also, for

$M_i P(\lambda_{i+1} \mid \lambda_i)$ to remain constant in a BG simulation [where M_i is given by Eq. (10)], k_i has to be chosen such that $k_0 = N_s / P(\lambda_1 \mid \lambda_0)$ and $k_i = 1/P(\lambda_{i+1} \mid \lambda_i)$ for $0<i<n$ where N_s is the desired number of partial paths between interfaces.

Since Eq. (30) does not fully constrain the values of $P(\lambda_{i+1} \mid \lambda_i)$ (as M values can be adjusted), one has some freedom in choosing them; e.g., for the particular case that it is desired to have a uniform distribution with $P(\lambda_{i+1}|\lambda_i)$ = constant for i=1,2,..,n-1, then

$$P(\lambda_{i+1} \mid \lambda_i) = \left[P(\lambda_n \mid \lambda_0)\right]^{1/n}. \quad (31)$$

To get a new $\{\lambda'\}$ staging, a special function f of $P(\lambda_{i+1} \mid \lambda_i)$ is constructed [with the current $P(\lambda_{i+1} \mid \lambda_i)$ vs. λ data] to interpolate the new $\{\lambda'\}$ set corresponding to the desired distribution of $P(\lambda_{i+1} \mid \lambda_i)$ values. This f function is not unique but should provide a one-to-one correspondence between an f value and a λ value; one such choice is,[35]

$$f(\lambda_i) = \frac{\ln[k(\lambda_A \rightarrow \lambda_i)]}{\ln[k(\lambda_A \rightarrow \lambda_B)]} = \frac{\sum_{j=0}^{i-1} \ln P(\lambda_{j+1} \mid \lambda_j)}{\sum_{j=0}^{n-1} \ln P(\lambda_{j+1} \mid \lambda_j)}, \qquad i = 1,...,n \quad (32)$$

where the denominator is simply a constant. The function in Eq. (32) monotonically increases with λ, going from $f(\lambda_0)$=0 to $f(\lambda_n=\lambda_B)$=1. Therefore, the algorithm consists of (1) running FFS to get P's for given λ's and construct the f curve, and (2) using the f curve to obtain new λ values for the desired set of $\{P(\lambda_1|\lambda_0)$, $P(\lambda_2|\lambda_1)$, …$\}$ values (one iteration provides suitable convergence of the λ staging). [35] For the choice of Eq. (31), Eq. (32) reduces to $f(\lambda_i) = i/n$, for $1<i<n$ (with $\lambda_0'=\lambda_0$ and $\lambda_n'=\lambda_n$ remaining fixed); i.e., the intermediate λ interfaces are to be distributed in such a way that $\Delta f = f(\lambda_{i+1}) - f(\lambda_i)=1/n$ is constant. This case is illustrated in Fig. 9, showing that the method identifies the "bottleneck" of the FFS simulation wherein sampling is automatically concentrated. Note that rather than fixing n (number of interfaces) in Eq. (31) and computing $P(\lambda_{i+1} \mid \lambda_i)$, the latter could be specified and the value of n calculated.

Individual interfacial points at the corresponding λ_i will have similar p_B value when the order parameter is a good estimate or the true RxC of the system. Based on this, the $\{\lambda_i'\}$ set could be optimized by distributing the $P(\lambda_{i+1} \mid \lambda_i)$ values trying to target prescribed values of the average $\langle p_B \rangle_\lambda$ for a given interface obtained from: [35]

$$\langle p_B \rangle_{\lambda_i} = \prod_{k=i}^{n-1} P(\lambda_{k+1} \mid \lambda_k) \qquad (33)$$

such that the FFS sampling is concentrated in the desired region (i.e., near the TSE). More generally, both λ-sampling and order parameter optimization (via FFS-LSE) could be combined such that in each iteration the λ staging of the current order parameter is optimized and used to obtain a new estimate for the RxC until a satisfactory convergence is attained (e.g., until the TS isosurfacce, p_B =0.5, coincides with the $\langle p_B \rangle_\lambda$ =0.5 interface from the staging optimization). Indeed, the optimization of λ staging can significantly reduce the computational effort of the FFS-LSE method in screening suitable RxC models.[35]

The strategy to optimize the staging discussed here could be extended to any other interface-based path sampling methods [e.g., like milestoning methods, TIS, or PPTIS].[35] Interestingly, the WE method (Sec. 2.4.1) was formulated to maintain a constant flow of partial reactive trajectories in each window, consistent with the idea underlying Eq. (30).

ii. Approach 2: Optimizing the trial runs $\{M_i\}$ set

Assuming that the λ staging has already been fixed, the statistical ν error in the estimate of $k_{A\to B}$, or equivalently the variance in the probability $P(\lambda_{n=B} \mid \lambda_0)$, can be minimized by optimizing the number of trial runs M_i or k_i at each interface for a fixed computational cost. For the DFFS [BG] scheme, the optimized $\{M_i'\}$[$\{k_i'\}$] set is then found by choosing the M_i [k_i] values such that Eq. (17)[(19)] is minimized with the constraint that the cost given by Eq. (19) [(20)] remains constant

[terms *R* and *S* in Eq. (15) are assumed constant]. For example, for the DFFS scheme, varying the distribution of M_i values, and assuming that $\left[1-P(\lambda_{i+1}\mid\lambda_i)\right]^{M_i}\approx 0$, this procedure leads to:

$$M_i \propto P(\lambda_n\mid\lambda_0)\left(\frac{1-P(\lambda_{i+1}\mid\lambda_i)}{P(\lambda_{i+1}\mid\lambda_i)}\right)^{1/2}\left(P(\lambda_{i+1}\mid\lambda_i)\left[\lambda_{i+1}-\lambda_i\right]+\left(1-P(\lambda_{i+1}\mid\lambda_i)\right)\left[\lambda_i-\lambda_A\right]\right)^{-1/2}. \quad (34)$$

To implement Eq. (34), one of the M_i values is fixed to a desired value M (i.e., $M_0 = M$) and set the computational cost of the DFFS simulation. Additional expressions for optimizing $\{M_i'\}$ for DFFS and for optimizing $\{k_i'\}$ for BG are given in [35]. To achieve a larger combined optimization effect, one can optimize first the staging (approach 1) for a specific prescribed set of $P(\lambda_{i+1}\mid\lambda_i)$ values, and then optimize for the $\{M_i\}$ set (approach 2). [35]

3. Applications

3.1. General biological applications

The appearance of rare event processes in biological systems has long been recognized to be a fundamental difficulty for computational studies of their dynamic and equilibrium behaviour. The use of path sampling methods is thus becoming an almost indispensable tool to properly account for the ubiquitous separation of time and length scales, typically spanning for several decades, caused by pervasive entropic or energetic barriers; the application of TPS-based methods has indeed allowed to unravel fundamental transition mechanisms not predicted before by other computer simulation methods.[9, 74] A recent account of the status of the field concerning biological problems has been presented before by Dellago and Bolhuis,[7] where a broad range of applications was already identified, including: biomolecular isomerization,[75, 9, 63] protein folding,[6, 76-78, 8, 74] DNA base pair unbinding,[79] enzyme catalysis,[80-83] lipid bilayers,[84, 85] and biochemical network switches.[5, 30] Here an update to that previous review is provided, where both old and some of the newer methods described above were used.

3.1.1. Protein folding

The kinetics of the transition pathways for the folding mechanism of a single chain protein in open space and in confined spaces was evaluated using the FFS framework with coarse-grained lattice models.[86] This work showed how the rapid initial formation of a critical core of amino acids affects the global properties and the folding mechanism of a single chain protein. The critical core is formed by those residues that have a higher chance of being in contact in the transition state;[87, 88] these residues can vary depending on the confinement conditions of the protein. Their observation regarding the importance of the formation of key transition state intermediates is consistent with the nucleation folding mechanism in proteins. The results from this work were used by Contreras et al. [89] in a computational study of the reassembly process of split proteins (i.e., how two fragments of a protein reconstitute the original folded structure). They found that the way in which the critical nucleus is fragmented plays a key role in the reassembly kinetics and mechanism; e.g., the two fragments reassembled more efficiently if the critical nucleus was fragmented in roughly equal parts which then acted as a catalytic "glue". The FFS-LSE approach was also applied to estimate a good RxC for the folding transition of model lattice proteins. [75] It was found that p_B could be well approximated via a model linear on conformational energy and number of native contacts.

The application of path sampling methodologies to the problem of two-state kinetics in all-atoms protein folding, recently reviewed by Bolhuis,[90] has yielded relevant insights of the (un)folding mechanisms and reaction rates of the Trp-cage in aqueous solution.[8, 31] This Trp-cage is a model mini-protein designed to yield fast folding rates accessible to computer simulations, but containing enough complexity to be a good test for path sampling algorithms. A more detailed account of this system and the different TPS-based methods used to study it are presented later.

The folding pathway of a 54-residue polyglutamine chain into a β-helical structure, in both explicit and implicit solvent, was studied by Chopra *et al.* [91] using constant-path TPS Monte Carlo simulations;[15] this system, unlike the one studied by Juraszek and Bolhuis[31], lacks higher order

(tertiary) structure formation. The pathway of β -helical structure formation was characterized by the formation of bridging hydrogen bonds between the coils, which nucleate primarily at the turns and drastically break the solvation hydrogen bonds; the same pattern of structure formation was observed in the implicit solvent case. Analysis of the TSE in both directions, folding and unfolding, revealed that contacts located at the corner of the helical coils control both reactive pathways; this was seen in both implicit and explicit solvent. Remarkably, the TPS analysis of this study found that at least 36 residues are necessary to stabilize the β-helical structure of polyglutamine sequences, which is consistent with the experimental evidence that correlates the presence of fibrils of expanded polyglutamines (i.e., those having more than ~36 residues) in samples of neuronal cells affected with neurodegenerative diseases.

3.1.2. DNA melting and base pair stability

Transitions between double- and single-stranded states of DNA were investigated by Sambriski *et al.* [92] by means of TPS simulations on a model where DNA nucleotides are coarse-grained to three sites (phosphate, sugar, and nucleobase). The dynamics between duplex melting and its reverse process, i.e., renaturation, involves not only a complex interplay of backbone conformation and strong short-ranged directional pairing and stacking interactions, but also spans across large time scales. The TSE was identified, using the committor probabilities, for two short oligonucleotide fragments (14 and 15 base-pairs) having different degree of sequence heterogeneity and guanine-cytosine base-pair content (repetitive and random, respectively). For the random sequence, the TSE occurs at a low value of the extent of reaction with a narrow (specific) probability distribution, whereas the repetitive sequence exhibits a largely broad (nonspecific) distribution. This behaviour was shown to be correlated with the location of the nucleation base pairs for conformations belonging to the TSE. The association pathway of the random sequence was shown to use fewer but localized base pairs, whereas that of the repetitive involves a large number of possible shifted base pairs; nevertheless, the central base-pairs associate preferentially in both cases.

Another problem of interest, involving base pair binding and unbinding transitions, is that of enzymatic repair of DNA lesions to maintain the stability of DNA replication and transcription processes.[81] A fundamental step in the repair of a localized lesion along double-helical DNA involves the opening of the base pair containing the lesion by flipping the damaged base. The process of lesion recognition by the O^6-alkylguanine-DNA alkyltransferase (AGT) and the forces that later promote flipping of a methylated guanine were studied by Hu *et al.* [93] using TPS simulations with bias annealing.[22] Path sampling allows capturing the correct dynamics of the rare event of base flipping, otherwise inaccessible to direct MD simulations, which occurs scarcely on the order of milliseconds even in the presence of the protein. Relevant RxCs for this process were found with a genetic network approach in terms of the committor probabilities;[9] a complex RxC was identified relating specific atomic distances in the active site of the DNA-enzyme complex. Comparing the features of the TSE for AGT-induced flipping of guanine (intact base) or methylguanine (alkyl lesion), allowed the identification of a kinetic "gate-keeping" strategy for lesion discrimination via a two-state process in which the methylguanine exhibits a faster rate of flipping to the active site and a slower rate of flipping back to the base-paired state; this mechanism reformulates a previously proposed one based on energetic stabilization of the flipped lesion.

A related process involving the bypass of an oxidative lesion during DNA polymerase β (pol β) replication was addressed by Wang and Schlick[94] with TPS simulations. In this bypass strategy, pol β has been shown to prefer the insertion of a correct nucleotide (d**C**TP) instead of an incorrect one (d**A**TP) when replicating the complement of an 8-Oxoguanine (8-oxo**G**) damaged base; this selection is controlled with large precision even when (d**A**TP) and (8-oxo**G**) form a stable base-pair much like a correct Watson-Crick **G**:**C** base pair. Different order parameters, in terms of molecular coordinates, were considered in the TPS simulations for the pairing-unpairing transition of both d**A**TP:8-oxo**G** and d**A**TP:8-oxo**G** base pairs. The identification of relevant molecular conformations of the transition states and transition pathways, as well as estimates of the reaction rates constants, allowed to recognize the

unfavourable interactions leading to a lower insertion efficiency for d**A**TP compared to that of d**C**TP; further kinetic and energetic effects favouring the pairing of dCTP were also found.

3.1.3. Biomolecular isomerization

The classical example of alanine dipeptide isomerization will be discussed in a later subsection. In this section, recent studies on other prominent problems involving conformational transformations of biomolecules will be reviewed. Such is the case of the isomerisation in vacuum of the methyl β-D-maltoside (disaccharide) between two of its four stable conformations, in terms of its ϕ and ψ dihedral angles, studied by Dimelow *et al.*[95]. In this work, the results from a TPS simulation were compared with those from a less computationally expensive potential of mean force (PMF) calculation. Although the two approaches were shown to be complementary, the former generated a more complete view of the relevant reaction mechanisms that included and additional reaction channel not predicted by the PMF simulations. This study also highlights the ability of TPS to discriminate the specific intramolecular interaction leading to different transition pathways. By using committor probabilities, the TSE was identified and found to be consistent with the free-energy landscape barriers along the (ϕ, ψ) coordinates.

The two-state allosteric conformational transition of the nitrogen regulatory protein C (NtrC) has been studied independently by Pan *et al.* [96] with a modified implementation of the FTS in collective variables[63] and by Khalili and Wales[97] using the discrete path sampling method.[98] The biological activity of NtrC, involved in bacterial signal transduction, is controlled by the phosphorylation of an aspartate residue which modulates the population of the active conformation from 2 to 10% in the unphosphorylated form, to 99% in the phosphorylated one. In the work by Pan *et al.*[96], a coarse-grained elastic two-state network was used to constraint the dynamics to transitions between the inactive and active states. Simulations of this model using the modified FTS method allowed to precisely determine a transition state having a $p_B \approx 0.5$ isocommittor pathway, corresponding to the free-energy barrier; for this purpose a complex multidimensional space accounting

for more than 550 collective variables (inter-residue distances) was considered. The discrete path sampling study by Khalili and Wales[97] on this same system at the atomistic level in implicit solvent, also identified kinetically relevant pathways by which this large scale conformational transition is achieved. This method, although not guided by the help of the isocommittor probabilities, allows selection of a suitable set of order parameters using disconnectivity graphs of the potential energy landscape for both active and inactive conformations. Then, an enumeration algorithm was used to locate the activation pathway (having the dominant contribution to the reaction rate) containing specific displacements, rotations and (de)stabilization interactions of residues that were largely correlated to experimental observations.

The allosteric activation mechanism of another signalling protein, the *E. coli* chemotaxis Y protein (CheY), was studied using TPS on an all-atom model in explicit solvent by Ma and Cui;[99] this protein is also activated by phosphorylation. In this case, the phosphorylated residue is displaced by the formation of a hydrogen bond, which in turn makes sterically possible the isomerization of a neighbouring residue from a solvent exposed configuration to a buried rotameric state. A large set of reactive trajectories was generated from path sampling simulations that used an initial state with p_B close to the transition state; the analysis of the TPS simulations was complemented with free-energy landscapes generated with umbrella sampling. This led to the identification of an alternative pathway that is kinetically competitive with respect to the one mentioned above; this new pathway is remarkable in that it does not require the formation of the hydrogen bond in the phosphorylated residue. The combination of free-energy and TPS results was also crucial to identify an alternative dynamics for a "loop gating" mechanism better correlated to experimental evidence than a previously proposed one based purely on biased MD simulations.

3.1.4. Enzymatic catalysis

The previous TPS study by Basner and Schwartz[80] on the enzymatic reaction catalyzed by the lactate dehydrogenase (LDH) was reassessed recently on the basis of committor distribution

analysis by Quaytman and Schwartz.[100] The former work identified reactive paths where residues outside the active site exhibit significant motion that seemed to provide additional stabilization to the donor-acceptor atoms; this hypothesis however was not proven to be correlated with the RxC. The latter study confirmed that indeed these motions make a significant contribution to the catalytic activity of LDH. This finding was validated by appropriately selecting a RxC (dependent on the displacement external residues mentioned above) which exhibited a peaked committor distribution at $p_B = 0.5$. Similarly, Schwartz et al.[101] studied the mechanistic dynamics of another enzymatic reaction, the phosphorolysis of guanosine catalyzed by the human purine nucleoside phosphorylase (PNP), using TPS simulations and committor probabilities. Similar to the behaviour observed in the LDH study,[80, 100] the kinetic details of the reactive trajectories obtained in this study also showed that the motions of protein residues outside the active site are directly coupled to the TSE and make significant contribution to the RxC. In this case, even fast motions were seen to be a critical part of the transition state, which shows that the time scale of these motions need not be the same as that of the enzymatic reaction to contribute to the rate constant.

In a similar study, the enzymatic conversion of chorismate into prephenate, catalyzed by the chorismate mutase, was characterized by Crehuet and Field[102] in terms of the TSE using TPS and committor probabilities. This particular reaction is a test model for computational studies but the interpretation of results from previous investigations has yielded several mechanistic contradictions. In addition to provide new insights on the reactive transition pathways, the path sampling analysis also revealed that, similarly to LDH and PNP, chorismate mutase exhibits motions of protein residues that cause conformational compression during the reaction and play an important role in TSE. In this work, the capabilities of TPS simulations to capture the correct dynamics of reactive trajectories were found adequate to address a problem that includes a large and complex space of collective variables; although the committor analysis failed to provide a relevant RxC. Accurate predictions of reaction rates, however,

seem to be out of the reach of TPS simulations for such cases, as that would require exceedingly expensive computations.

3.1.5. Genetic Switches

FFS is particularly advantageous to study oscillatory biochemical networks due to the nonequilibrium nature of the rare switch flipping events. Two models of bistable genetic switches have been investigated recently by Valeriani *et al.* [33] and Morelli *et al.* [103] using the FFS method.[5] The two models studied describe the action of two genes that repress mutually and encode two transcription factors (proteins); the production of both transcription factors (A and B) is controlled by their binding in homodimeric form (A_2 and B_2) to a regulatory DNA sequence (O). In one model, the exclusive switch, the two factors mutually exclude each other's binding, whereas in another, the general switch model, both factors can bind simultaneously; a network of biochemical reactions describe the kinetics behaviour of the switches. The switching pathways obtained for both models from FFS in combination with committor distribution analysis, yielded a complex dependence on the transcription factor homodimerization and their DNA binding and unbinding reactions. For the exclusive switch, the TSE revealed a strong dependence of on the latter but not on the former, whereas the switching pathways of the general switch are largely independent to fluctuations in the rate constants of both reactions.

More recently, Morelli et al. [104] have studied via FFS the bistable gene regulatory switch controlling the transition from lysogeny (dormant state) to lysis (lytic state) in bacteriophage lambda. The model adopted encompassed several hundred reactions that included DNA looping, the detailed dynamics of binding of transcription factors to the promoters, and gene depletion of by non-specific binding. By taking into account the stochastic character of the chemical reactions, these authors were able to reproduce the bistability of the switch and to find evidence that DNA looping provides a likely explanation for the puzzling stability and robustness (experimentally observed) of the lysogenic state to perturbations of the transcriptional regulatory interactions.

3.1.6. Biopolymer Translocation and Motion through Pores

The transport of a biopolymer through a nanopore is a phenomenon that occurs in several important biological processes such as in the passage of proteins through the ribosomal tunnel (during protein synthesis), the protein entrance to and departure from chaperone cavities, the motion of oligomeric species through cell transmembrane protein pores, and the insertion of genetic material by viruses. Several nascent biotechnological techniques also rely on biopolymer translocation, including microfluidic devices designed to separate or to sequence DNA. These processes are typically unidirectional and rely on the aid of an external driving force, such as flow, a motor protein, or an electric field. The non-equilibrium nature of the process makes FFS again an ideal tool for kinetic studies. This was recognized very early by Allen et al. [4, 5] who demonstrated the use of FFS to study a simple model of polymer translocation through a pore. In that study, the chain was modeled by mutually attracting beads, chain motion was evolved via Langevin dynamics, and the pulling force was assumed to be intermittent so that in the "on" state all monomers inside the pore experienced a forwardly-directed force, and in the "off" state no external force was exerted (such an action could be seen as a simplistic analog to that of a motor protein). Despite this auspicious beginning and the multitude of interesting related processes, applications of FFS to biopolymer motion through pores are still scant (two examples are discussed below); most often, such problems have been studied via conventional MD, BD, and MC methods.

In a recent paper, Huang and Macarov [105] studied the dependence of the time scale of polymer reversal on the pore size and on the polymer length. The system consisted of an unstructured, flexible chain inside a neutral, infinitely long cylindrical pore. They compared the predictions of simple one-dimensional theories (like TS theory) and exact FFS-BD simulation results in describing the reversal rate constant. More recently, Hernandez-Ortiz and de Pablo [106] used FFS to quantify the effect of hydrodynamics on the flow-induced translocation rate of DNA molecules through a narrow pore. The setup and dimensions of this system are depicted in Fig. 10; long DNA molecules (ranging

from 10 to 420 μm) were modelled as flexible bead-spring chains with repulsive beads and their motion described by a BD scheme that incorporates hydrodynamic effects under confinement. The λ interfaces were defined as parallel planes perpendicular to the flow direction and located at specific distances along the pore; the chain was assumed to reach an interface when its chain centre of mass crossed it. This study found that hydrodynamic forces can either accelerate or hinder translocation (depending on DNA molecular weight) by many orders of magnitude, thus emphasizing the limitations of the free-draining approximation (which ignores hydrodynamic interactions) in the modelling of such processes and in the interpretation of experimental data.

3.2. Illustrative Cases

3.2.1. Isomerization of Alanine dipeptide

Alanine dipeptide is a small molecule that in aqueous environment displays short-timescale transitions between conformations typical of α helix and β strand motifs in proteins.[107] Fig. 11 shows a schematic representation of the main conformers involved in the transitions of this peptide projected onto ψ and ϕ dihedral angles in vacuum and explicit solvent. In vacuum, the free energy landscape shows three distinct stable basins corresponding to states C_{7ax}, C_{7eq} and C5. In explicit solvent, the free energy landscape shows several minima in which the $\beta_2/\alpha_R \Leftrightarrow C5/C_{7eq}$ transition is one of the most studied. Various researchers have estimated transition rate constant values for the forward [107, 108, 49] and reverse transitions in vacuum and explicit solvent environment,[75, 107] as well as the collective variables that are important for the description of the transitions.[75, 9, 49]

Fig. 12 compares some of the rate constant estimates found in the literature. For example, in vacuum, Chun et al.[109] and Vedell and Wu[110] estimated the kinetic $C_{7eq} \Rightarrow$ C5 transition time to be around 3.0 ps. The former study used both atomistic MD simulations and a rigid body MBO(N)D[108] method using CHARMM force field, while the latter used a multiple shooting algorithm with a MOIL based force field. More recently, Velez-Vega et al. [73] estimated a 4.0 ps transition rate time for the

same reaction using both atomistic FFS- MC and MD simulations with CHARMM force field. The results from both FFS-type simulations were similar and consistent with those found in the literature.[75, 107, 108, 49] These authors employed a FFS-MD scheme by incorporating a stochastic component (associated with an Andersen-type thermostat[111]) into the MD simulation to achieve distinct pathways between the stable states.

The fast $\beta_2/\alpha_R \Rightarrow C5/C7_{eq}$ reaction in explicit solvent has been simulated in several studies. For example, Chekmarev et al.[107] used BD and an Analytic Generalized Born with Nonpolar Interactions implicit solvent model with OPLS force field to obtain a mean first passage time of 27 ps. In other study, Oliveira et al.[108] estimated a transition time of 80 ps using accelerated MD simulations with AMBER force field in explicit solvent. More recently, West et al.[49] and Velez-Vega et al.[73] determined transition rate for the same transition using path sampling methods. The former found a mean first passage time of 64 ps via the Milestoning method using the MOIL package, whereas the latter obtained a 31 ± 6 ps transition time via FFS-MD simulations using the CHARMM force field. For the slow $C5/C_{7eq} \Rightarrow \beta_2/\alpha_R$ reaction, Chekmarev et al.[107] calculated a mean first passage time of 249 ps using BD and an Analytic Generalized Born with Nonpolar Interactions (AGBNP) implicit solvent model with OPLS-AA force field. However, Bolhuis et al.[75] estimated a rate constant of 100 ps using TPS with AMBER 94 force field, while Velez-Vega et al. [73] estimated a rate constant of 328 ± 62 ps using FFS-MD simulations with the CHARMM force field. A general conclusion from the above works is that path sampling simulations provide an efficient way to estimate rate constant for the transitions compared to conventional methods, but the results typically have large errorbars and are highly sensitive to the force field details.

Although the peptide ψ and ϕ dihedral angles satisfactorily describe the system's distinct stable states (as seen in Figs. 12 and 13), this does not imply that they are accurate descriptors for the dynamics of the transition. [75, 9, 49] Thus, other variables and/or interaction terms between variables may also be important in the RxC model. For example, Bolhuis et al.[75] use a conventional committor

analysis[15] and found that for the $C_{7eq} \Rightarrow C_{7ax}$ isomerization in vacuum, other variables besides the ψ and ϕ angles are necessary, suggesting that adding the θ (O-C-N-C_α) angle could provide a reasonable description of the RxC. The string simulations performed by Maragliano et al.[63] suggested that including the ζ (C_α-C-N-H) angle in addition to ψ ,ϕ and θ is required to make the committor distribution peaked at the correct value. Ma and Dinner [9] also estimated the collective variables that are important for the description of the $C7_{eq} \Rightarrow$C5 isomerization reaction in vacuum using their genetic neural network method. Their results confirm that a RxC in term of three of the main dihedrals angle correlates strongly with the committor probability distribution. More recently, Velez-Vega et al.[73] used the FFS-LSE algorithm and found that the ψ and ϕ order parameters are sufficient for predicting the dynamic pathways of the $C7_{eq} \Rightarrow$C5 transition; however, an interaction term between these variables (i.e., $\psi\phi$) and quadratic terms (i.e., ψ^2 and ϕ^2) are also necessary for a complete description of the isocommittor surface curvature. Isocommittor lines for this transition are shown in Fig. 13.

For the $\beta_2/\alpha_R \Leftrightarrow C5/C_{7eq}$ transition in explicit water, Bolhuis et al.[75] used conventional committor analysis[15] to find that the solvent degrees of freedom may play a role in this transition and suggested their incorporation in the RxC model of the process. Using a large p_B database and a genetic neural network method, Ma and Dinner[9] determined that the best RxC model is composed of three descriptors: the ψ angle, the distance between atoms 2H and $2C_\beta$ ($\left|r_{\mathbf{2H-C_\beta}}\right|$) , and the electrostatic torque around bond 1C-2N from solvent forces on atom 3H ($\Gamma^{3H}_{1C\text{-}2N}$). Velez-Vega et al.[73] used the FFS-LSE algorithm to determine that a simpler linear model involving the dihedral angles ψ and ϕ , and the quantities $\left|r_{|2H-C_\beta}\right|$ and $\Gamma_{1C-2N3H}$) provided a reasonably complete description of the system's dynamics. However, for the reverse $C5/C_{7eq} \Rightarrow \beta_2/\alpha_R$ transition, it was found that in addition to the linear terms in those variables, nonlinear interaction terms between them were needed for a better description of the isocommitor surfaces.[73] A general consensus of these studies is that although the ψ angle seems to be

a key component of the RxC, other variables, including some associated with solvent degrees of freedom, play a non-negligible role in the TS for isomerizations in water.

3.2.2. Trp-cage mini-protein (un)folding

The Trp-cage protein is a 20-residues polypeptide chain (NLYIQWLKDGGPSSGRPPPS) designed by Neidigh et al.[112] to fold via a two-state process. Its native structure contains both secondary and tertiary structures having a hydrophobic core in which Trp-6 is buried. Although the folding events of Trp-cage have been studied using all-atom implicit-solvent MD simulations, coarse-grained models, and replica exchange MD simulations;[113-116], it is the all-atom solvent-explicit TPS simulations performed by Juraszek and Bolhuis[8] that have fully elucidated the mechanism through which this protein folds. These authors [8, 31] employed the OPLSAA force field, and an the SPC water model to obtain the TPE in a simulation time of several μs. They found that this mini-protein folds through two distinct mechanisms [as depicted in Fig. 14]: (i) a diffusion-collision route where secondary structure elements fold before the tertiary structure and (ii) a nucleation-condensation mechanism where a tertiary-structure nucleation event precedes the formation and stabilization of the secondary structures. The latter mechanism (i.e., the N-L-U route) was found to be the predominant folding route (i.e., 80% of the paths follow this route). On the basis of committor calculations, piece-wise RxCs were found between the basins and the transition states on each route. Although a first study demonstrated that TPS is indeed capable of capturing entirely different paths in a complex system,[8] the second study from the same authors[31] compared rate constants for the folding and unfolding of the Trp-cage protein in explicit solvent using TIS and FFS. In particular, they studied the major (N-L-U) unfolding-folding route which contributes most to the rate constant. The root mean-square deviation of the α-helical (2-8) residues from an ideal helix ($rmsd_{hx}$) was used as order parameter to partition the phase space. From the TIS simulations, they estimated the rates for folding and unfolding to be k_{LN}= $(0.4\ \mu s)^{-1}$ and k_{NL}= $(1.2\ \mu s)^{-1}$, respectively, which are about one order-of-magnitude higher than the experimental values. In contrast, they found that the rate constant for the unfolding transition obtained

from FFS simulations [k_{NL}= (100 μs)$^{-1}$] was a factor-80 smaller than the one obtained with TIS. In this case, the paths sampled by FFS bunched up along a non-representative route resulting in the overestimation of the free energy barrier and hence underestimation of the rate constant. The authors argued that, compared to TIS, FFS is likely more sensitive to the choice of the order parameter and has more difficulty relaxing the pathways in directions orthogonal to the imposed order parameter. This highlights the importance of thoroughly optimizing the implementation of FFS (something that was arguably limited in Ref. [31]), making use of various strategies such as those described in Sections 2.5.2 and 2.5.3. In particular, the use of a better order parameter (closer to the RxC) to define the λ space, and of an optimized positioning of the first interface are likely to vastly improve the performance of FFS for this system.

Juraszek and Bolhuis[31] also applied the ML method for the L-N TPE obtained from the TPS simulations[8] to obtain a better estimate for the RxC. They found that a combination of the $rmsd_{hx}$ and the root mean square deviation from the native α-carbons ($rmsd_{ca}$) provided a good description of the local dynamics. On the other hand, for the L-N TPE a single RxC model in terms of $rmsd_{ca}$ was enough to describe the transition.

4. Conclusions

The last few years have witnessed a rapid increase in the number of path sampling methods and a wider range of applications to biological and physico-chemical systems. This trend is only expected to accelerate in the years to come as these methods become more refined, computers get faster and cheaper, and more experience (and wisdom) is gained on how to avoid pitfalls and make the most out of a given method. While many of these advances lie outside the scope of this article, this review has attempted to convey some of this collective experience as it pertains to a subset of path sampling methods and applications to biomolecular transitions. By understanding the core concepts underlying some of the competing methods, similarities and differences among them can be better appreciated,

which in turn can lead to further methodological improvements, cross-fertilization or to a clearer view of instances when two different methods become complementary.

Expectedly, our review of recent applications contains still many more examples of the use of the older, more established methods (like TPS), rather than of the use of the newer or less known methods we had tried to highlight here (for which often only proof-of-principle calculations exist). Altogether though, it is clear that path sampling methods are becoming more sophisticated and are already tackling challenging systems and unraveling new mechanistic details of important biomolecular processes. On the other hand, methods employing accelerated dynamics and other time-saving strategies (like coarse graining) will be needed to study the countless biological systems that not only entail very many degrees of freedom, but also exhibit a multiplicity of intermediates and transition channels, and span very broad time scales. It is important to keep in mind, however, that the successful implementation of novel path sampling methods (and of forward flux sampling in particular) crucially depends on the careful selection of the method's parameters; this selection should not be based only on experience or "art" but also on systematic optimization strategies such as those discussed in this review.

Acknowledgements

This publication is based on work supported in part by Award No. KUS-C1-018-02, made by King Abdullah University of Science and Technology (KAUST). Additional support from the National Science Foundation Award 0553719 is also gratefully acknowledged. The authors are also grateful to J. Hernandez-Ortiz and P. Bolhuis for allowing us to modify their picture files.

References

[1] Sethuraman A, Vedantham G, Imoto T, Przybycien T and Belfort G 2004 Protein unfolding at interfaces: Slow dynamics of alpha-helix to beta-sheet transition *Proteins: Struct. Funct. Bioinform.* **56** 669-78

[2] Tsumoto K, Ejima D, Kumagai I and Arakawa T 2003 Practical considerations in refolding proteins from inclusion bodies *Protein Expression Purif.* **28** 1-8

[3] Snow C D, Sorin E J, Rhee Y M and Pande V S 2005 How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.* **34** 43-69

[4] Allen R J, Frenkel D and ten Wolde P R 2006 Forward flux sampling-type schemes for simulating rare events: Efficiency analysis *J. Chem. Phys.* **124** 194111

[5] Allen R J, Frenkel D and ten Wolde P R 2006 Simulating rare events in equilibrium or nonequilibrium stochastic systems *J. Chem. Phys.* **124** 024102

[6] Bolhuis P G 2003 Transition-path sampling of beta-hairpin folding *Proc. Natl. Acad. Sci. USA* **100** 12129-34

[7] Dellago C and Bolhuis P G 2007 *Atomistic Approaches in Modern Biology: from Quantum Chemistry to Molecular Simulations,* (Berlin: Springer-Verlag Berlin) pp 291-317

[8] Juraszek J and Bolhuis P G 2006 Sampling the multiple folding mechanisms of Trp-cage in explicit solvent *Proc. Natl. Acad. Sci. USA* **103** 15859-64

[9] Ma A and Dinner A R 2005 Automatic method for identifying reaction coordinates in complex systems *J. Phys. Chem. B* **109** 6769-79

[10] Peters B and Trout B L 2006 Obtaining reaction coordinates by likelihood maximization *J. Chem. Phys.* **125** 054108

[11] van Erp T S 2006 Efficiency analysis of reaction rate calculation methods using analytical models I: The two-dimensional sharp barrier *J. Chem. Phys.* **125** 174106

[12] Dellago C and Bolhuis P G 2008 *Advanced Computer Simulation Approaches for Soft Matter Sciences III,* (Springer: Eds. C. Holmer, K. Kremer) pp 167-233

[13] Dellago C, Bolhuis P G and Chandler D 1999 On the calculation of reaction rate constants in the transition path ensemble *J. Chem. Phys.* **110** 6617-25

[14] Dellago C, Bolhuis P G, Csajka F S and Chandler D 1998 Transition path sampling and the calculation of rate constants *J. Chem. Phys.* **108** 1964-77

[15] Dellago C, Bolhuis P G and Geissler P L 2002 Transition path sampling *Adv. Chem. Phys.* **123** 1-78

[16] Bolhuis P G, Chandler D, Dellago C and Geissler P L 2002 Transition path sampling: Throwing ropes over rough mountain passes, in the dark *Annu. Rev. Phys. Chem.* **53** 291-318

[17] Bolhuis P G 2003 Transition path sampling on diffusive barriers *J. Phys.: Condens. Matter* **15** S113-S20

[18] Geyer C J and Thompson E A 1995 Annealing Markov-Chain Monte-Carlo with Applications to Ancestral Inference *J. Amer. Statistical Assoc.* **90** 909-20

[19] Vlugt T J H and Smit B 2001 On the efficient sampling of pathways in the transition path ensemble *PhysChemComm* **2** 1

[20] Chopra M, Malshe R, Reddy A S and de Pablo J J 2008 Improved transition path sampling methods for simulation of rare events *J. Chem. Phys.* **128** 144104

[21] Grunwald M, Rabani E and Dellago C 2006 Mechanisms of the wurtzite to rocksalt transformation in CdSe nanocrystals *Phys. Rev. Lett.* **96** 255701

[22] Hu J, Ma A and Dinner A R 2006 Bias annealing: A method for obtaining transition paths de novo *J. Chem. Phys.* **125** 114101

[23] Moroni D, van Erp T S and Bolhuis P G 2004 Investigating rare events by transition interface sampling *Physica a-Statistical Mechanics and Its Applications* **340** 395-401

[24] van Erp T S and Bolhuis P G 2005 Elaborating transition interface sampling methods *J. Comput. Phys.* **205** 157-81

[25] van Erp T S, Moroni D and Bolhuis P G 2003 A novel path sampling method for the calculation of rate constants *J. Chem. Phys.* **118** 7762-74
[26] Bolhuis P G 2008 Rare events via multiple reaction channels sampled by path replica exchange *J. Chem. Phys.* **129** 114108
[27] van Erp T S 2007 Reaction rate calculation by parallel path swapping *Phys. Rev. Lett.* **98** 268301
[28] Rogal J and Bolhuis P G 2008 Multiple state transition path sampling *J. Chem. Phys.* **129** 224107
[29] Moroni D, van Erp T S and Bolhuis P G 2005 Simultaneous computation of free energies and kinetics of rare events *Phys. Rev. E* **71** 056709
[30] Allen R J, Warren P B and ten Wolde P R 2005 Sampling rare switching events in biochemical networks *Phys. Rev. Lett.* **94** 018104
[31] Juraszek J and Bolhuis P G 2008 Rate Constant and Reaction Coordinate of Trp-Cage Folding in Explicit Water *Biophys. J.* **95** 4246-57
[32] Borrero E E and Escobedo F A 2009 Simulating the kinetics and thermodynamics of transitions via forward flux/umbrella sampling *J. Phys. Chem. B* **113** 6434-45
[33] Valeriani C, Allen R J, Morelli M J, Frenkel D and ten Wolde P R 2007 Computing stationary distributions in equilibrium and nonequilibrium systems with forward flux sampling *J. Chem. Phys.* **127** 114109
[34] Borrero E E and Escobedo F A 2007 Reaction coordinates and transition pathways of rare events via forward flux sampling *J. Chem. Phys.* **127** 164101-17
[35] Borrero E E and Escobedo F A 2008 Optimizing the sampling and staging for simulations of rare events via forward flux sampling schemes *J. Chem. Phys.* **129** 024115-16
[36] Huber G A and Kim S 1996 Weighted-ensemble Brownian dynamics simulations for protein association reactions *Biophys. J.* **70** 97-110
[37] Dickson A, Warmflash A and Dinner A R 2009 Nonequilibrium umbrella sampling in spaces of many order parameters *J. Chem. Phys.* **130** 074104
[38] Farkas L 1927 The speed of germinitive formation in over saturated vapours *Z. Phys. Chem.* **125** 236-42
[39] Hänggi P, Talkner P and Borkovec M 1990 Reaction-rate theory: fifty years after Kramers *Reviews of Modern Physics* **62** 251
[40] Northrup S H, Allison S A and McCammon J A 1984 Brownian Dynamics Simulation Of Diffusion-Influenced Bimolecular Reactions *J. Chem. Phys.* **80** 1517-26
[41] Gabdoulline R R and Wade R C 2002 Biomolecular diffusional association *Curr. Opin. Struct. Biol.* **12** 204-13
[42] Rojnuckarin A, Livesay D R and Subramaniam S 2000 Bimolecular reaction simulation using Weighted Ensemble Brownian dynamics and the University of Houston Brownian Dynamics program *Biophys. J.* **79** 686-93
[43] Zhang B W, Jasnow D and Zuckerman D M 2007 Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin *Proc. Natl. Acad. Sci. USA* **104** 18043-8
[44] Rojnuckarin A, Kim S and Subramaniam S 1998 Brownian dynamics simulations of protein folding: Access to milliseconds time scale and beyond *Proc. Natl. Acad. Sci. USA* **95** 4288-92
[45] Fisher E W, Rojnuckarin A and Kim S 2001 Kinetic effects of mutations of charged residues on the surface of a dimeric hemoglobin: insights from Brownian dynamics simulations *J. Mol. Struct.* **549** 47-54
[46] Fisher E W, Rojnuckarin A and Kim S 2002 Effects of local repositioning of charged surface residues on the kinetics of protein dimerization probed by Brownian dynamics simulations *J. Mol. Struct.* **592** 37-45

[47] Fisher E W, Rojnuckarin A and Kim S 2002 Exhaustive enumeration of the effects of point charge mutations on the electrostatically driven association of hemoglobin subunits, using weighted-ensemble Brownian dynamics simulations *Struct. Chem.* **13** 193-202
[48] Faradjian A K and Elber R 2004 Computing time scales from reaction coordinates by milestoning *J. Chem. Phys.* **120** 10880-9
[49] West A M A, Elber R and Shalloway D 2007 Extending molecular dynamics time scales with milestoning: Example of complex kinetics in a solvated peptide *J. Chem. Phys.* **126** 145104
[50] Vanden-Eijnden E, Venturoli M, Ciccotti G and Elber R 2008 On the assumptions underlying milestoning *J. Chem. Phys.* **129** 174102
[51] Montroll E W and Weiss G H 1965 Random Walks on Lattices. 2 *J. Math. Phys.* **6** 167
[52] Kampen N G v 1992 *Stochastic processes in physics and chemistry* (Amsterdam: North-Holland)
[53] Elber R 2007 A milestoning study of the kinetics of an allosteric transition: Atomically detailed simulations of deoxy Scapharca hemoglobin *Biophys. J.* **92** L85-L7
[54] E W, Ren W Q and Vanden-Eijnden E 2005 Finite temperature string method for the study of rare events *J. Phys. Chem. B* **109** 6688-93
[55] E W and Vanden-Eijnden E 2006 Towards a theory of transition paths *J. Stat. Phys.* **123** 503-23
[56] Vanden-Eijnden E and Tal F A 2005 Transition state theory: Variational formulation, dynamical corrections, and error estimates *J. Chem. Phys.* **123** 184103
[57] E W, Ren W Q and Vanden-Eijnden E 2002 String method for the study of rare events *Phys. Rev. B* **66** 052301
[58] Henkelman G and Jonsson H 2000 Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points *J. Chem. Phys.* **113** 9978
[59] Henkelman G, Uberuaga B P and Jonsson H 2000 A climbing image nudged elastic band method for finding saddle points and minimum energy paths *J. Chem. Phys.* **113** 9901
[60] E W, Ren W Q and Vanden-Eijnden E 2005 Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes *Chem. Phys. Lett.* **413** 242-7
[61] Gardiner C W 2004 *Handbook of stochastic methods: for physics, chemistry, and the natural sciences* (Berlin; New York: Springer)
[62] Ren W, Vanden-Eijnden E, Maragakis P and E W 2005 Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide *J. Chem. Phys.* **123** 134109
[63] Maragliano L, Fischer A, Vanden-Eijnden E and Ciccotti G 2006 String method in collective variables: Minimum free energy paths and isocommittor surfaces *J. Chem. Phys.* **125** 024106
[64] Maragliano L and Vanden-Eijnden E 2007 On-the-fly string method for minimum free energy paths calculation *Chem. Phys. Lett.* **446** 182-90
[65] Hummer G 2004 From transition paths to transition states and rate coefficients *J. Chem. Phys.* **120** 516-23
[66] Weinan E, Ren W Q and Vanden-Eijnden E 2005 Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes *Chem. Phys. Lett.* **413** 242-7
[67] Peters B, Beckham G T and Trout B L 2007 Extensions to the likelihood maximization approach for finding reaction coordinates *J. Chem. Phys.* **127** 034109
[68] Best R B and Hummer G 2005 Reaction coordinates and rates from transition paths *Proc. Natl. Acad. Sci. USA* **102** 6732-7
[69] Dinner A R, So S S and Karplus M 1998 Use of quantitative structure-property relationships to predict the folding ability of model proteins *Proteins: Struct. Funct. Bioinform.* **33** 177-203
[70] Dinner A R, So S S and Karplus M 2002 *Computational Methods for Protein Folding,* pp 1-34
[71] So S S and Karplus M 1997 Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations *J. Med. Chem.* **40** 4347-59

[72] Metzner P, Schutte C and Vanden-Eijnden E 2006 Illustration of transition path theory on a collection of simple examples *J. Chem. Phys.* **125** 084110
[73] Velez-Vega C, Borrero E E and Escobedo F A 2009 Kinetics and reaction coordinate for the isomerization of alanine dipeptide by a forward flux sampling protocol *J. Chem. Phys.* **(in press)**
[74] ten Wolde P R and Chandler D 2002 Drying-induced hydrophobic polymer collapse *Proc. Natl. Acad. Sci. USA* **99** 6539-43
[75] Bolhuis P G, Dellago C and Chandler D 2000 Reaction coordinates of biomolecular isomerization *Proc. Natl. Acad. Sci. USA* **97** 5877-82
[76] Bolhuis P G 2005 Examining the folding of small two-state proteins in explicit water using path sampling techniques *Biophys. J.* **88** 182A-A
[77] Bolhuis P G 2005 Kinetic pathways of beta-hairpin (Un)folding in explicit solvent *Biophys. J.* **88** 50-61
[78] Evans D A and Wales D J 2004 Folding of the GB1 hairpin peptide from discrete path sampling *J. Chem. Phys.* **121** 1080-90
[79] Hagan M F, Dinner A R, Chandler D and Chakraborty A K 2003 Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA *Proc. Natl. Acad. Sci. USA* **100** 13922-7
[80] Basner J E and Schwartz S D 2005 How enzyme dynamics helps catalyze a reaction in atomic detail: A transition path sampling study *J. Am. Chem. Soc.* **127** 13822-31
[81] Radhakrishnan R and Schlick T 2004 Orchestration of cooperative events in DNA synthesis and repair mechanism unraveled by transition path sampling of DNA polymerase beta's closing *Proc. Natl. Acad. Sci. USA* **101** 5970-5
[82] Radhakrishnan R and Schlick T 2005 Fidelity discrimination in DNA polymerase beta: Differing closing profiles for a mismatched (G: A) versus matched (G: C) base pair *J. Am. Chem. Soc.* **127** 13245-52
[83] Radhakrishnan R, Yang L J, Arora K and Schlick T 2004 Exploring DNA polymerase beta mechanisms by advanced sampling techniques *Biophys. J.* **86** 34A-A
[84] Marti J 2004 A molecular dynamics transition path sampling study of model lipid bilayer membranes in aqueous environments *J. Phys.: Condens. Matter* **16** 5669-78
[85] Marti J and Csajka F S 2004 Transition path sampling study of flip-flop transitions in model lipid bilayer membranes *Physical Review E* **69** 061918
[86] Borrero E E and Escobedo F A 2006 Folding kinetics of a lattice protein via a forward flux sampling approach *J. Chem. Phys.* **125** 164904-14
[87] Fersht A R 1997 Nucleation mechanisms in protein folding *Curr. Opin. Struct. Biol.* **7** 3-9
[88] Vendurscolo M, Paci E, Dobson C M and Karplus M 2001 Three key residues form a critical network in a protein folding transition state *Nature* **409** 641-5
[89] Martinez L M C, Quintana E E B, Escobedo F A and DeLisa M P 2008 In silico protein fragmentation reveals the importance of critical nuclei on domain reassembly *Biophys. J.* **94** 1575-88
[90] Bolhuis P G 2009 Two-state protein folding kinetics through all-atom molecular dynamics based sampling *Frontiers In Bioscience* **14** 2801-28
[91] Chopra M, Reddy A S, Abbott N L and de Pablo J J 2008 Folding of polyglutamine chains *J. Chem. Phys.* **129** 135102
[92] Sambriski E J, Ortiz V and de Pablo J J 2009 Sequence effects in the melting and renaturation of short DNA oligonucleotides: structure and mechanistic pathways *J. Phys.: Condens. Matter* **21** 034105
[93] Hu J, Ma A and Dinner A R 2008 A two-step nucleotide-flipping mechanism enables kinetic discrimination of DNA lesions by AGT *Proc. Natl. Acad. Sci. USA* **105** 4615-20
[94] Wang Y L and Schlick T 2007 Distinct energetics and closing pathways for DNA polymerase beta with 8-oxoG template and different incoming nucleotides *BMC Struct. Biol.* **7** 7

[95] Dimelow R J, Bryce R A, Masters A J, Hillier I H and Burton N A 2006 Exploring reaction pathways with transition path and umbrella sampling: Application to methyl maltoside *J. Chem. Phys.* **124** 114113
[96] Pan A C, Sezer D and Roux B 2008 Finding transition pathways using the string method with swarms of trajectories *J. Phys. Chem. B* **112** 3432-40
[97] Khalili M and Wales D J 2008 Pathways for conformational change in nitrogen regulatory protein C from discrete path sampling *J. Phys. Chem. B* **112** 2456-65
[98] Wales D J 2002 Discrete path sampling *Mol. Phys.* **100** 3285-305
[99] Ma L and Cui Q 2007 Activation mechanism of a signaling protein at atomic resolution from advanced computations *J. Am. Chem. Soc.* **129** 10261-8
[100] Quaytman S L and Schwartz S D 2007 Reaction coordinate of an enzymatic reaction revealed by transition path sampling *Proc. Natl. Acad. Sci. USA* **104** 12253-8
[101] Saen-oon S, Quaytman-Machleder S, Schramm V L and Schwartz S D 2008 Atomic detail of chemical transformation at the transition state of an enzymatic reaction *Proc. Natl. Acad. Sci. USA* **105** 16543-8
[102] Crehuet R and Field M J 2007 A transition path sampling study of the reaction catalyzed by the enzyme chorismate mutase *J. Phys. Chem. B* **111** 5708-18
[103] Morelli M J, Tanase-Nicola S, Allen R J and ten Wolde P R 2008 Reaction coordinates for the flipping of genetic switches *Biophys. J.* **94** 3413-23
[104] Morelli M J, ten Wolde P R, and Allen R J 2009 DNA looping provides stability and robustness to the bacteriophase lambda switch *Proc. Natl. Acad. Sci. USA* **106** 18101-6
[105] Huang L and Makarov D E 2008 The rate constant of polymer reversal inside a pore *J. Chem. Phys.* **128** 114903
[106] Hernandez-Ortiz J P and de Pablo J J 2008 Hydrodynamic Effects on the Translocation Rate of a Polymer Through a Narrow Pore *Phys. Rev. Lett.* **(submitted)**
[107] Chekmarev D S, Ishida T and Levy R M 2004 Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete-state kinetic models *J. Phys. Chem. B* **108** 19487-95
[108] de Oliveira C A F, Hamelberg D and McCammon J A 2007 Estimating kinetic rates from accelerated molecular dynamics simulations: Alanine dipeptide in explicit solvent as a case study *J. Chem. Phys.* **127** 175105-8
[109] Chun H M, Padilla C E, Chin D N, Watanabe M, Karlov V I, Alper H E, Soosaar K, Blair K B, Becker O M, Caves L S D, Nagle R, Haney D N and Farmer B L 2000 MBO(N)D: A multibody method for long-time molecular dynamics simulations *J. Comput. Chem.* **21** 159-84
[110] Vedell P and Wu Z 2008 The solution of the boundary-value problems for the simulation of transition of protein conformation *International Journal of Numerical Analysis and Modeling* **(submitted)**
[111] Frenkel D and Smit B 2002 *Understanding Molecular Simulation: From Algorithms to Applications* (Boston: Academic)
[112] Neidigh J W, Fesinmeyer R M and Andersen N H 2002 Designing a 20-residue protein *Nat. Struct. Biol.* **9** 425-30
[113] Linhananta A, Boer J and MacKay I 2005 The equilibrium properties and folding kinetics of an all-atom Go model of the Trp-cage *J. Chem. Phys.* **122** 114901
[114] Ota M, Ikeguchi M and Kidera A 2004 Phylogeny of protein-folding trajectories reveals a unique pathway to native structure *Proc. Natl. Acad. Sci. USA* **101** 17658-63
[115] Simmerling C, Strockbine B and Roitberg A E 2002 All-atom structure prediction and folding simulations of a stable protein *J. Am. Chem. Soc.* **124** 11258-9
[116] Snow C D, Zagrovic B and Pande V S 2002 The Trp cage: Folding kinetics and unfolded state topology via molecular dynamics simulations *J. Am. Chem. Soc.* **124** 14548-9

Figure captions

Figure 1. Schematic illustration of the shooting algorithm in TPS simulations. A new trajectory path (blue) is generated from an old one (red) by choosing randomly one state on the old trajectory. For deterministic trajectories, this shooting point is modified by adding a small perturbation δp to the particle momenta. From this modified point a new path is constructed by following forward and backward the intrinsic system dynamics. For stochastic dynamics no perturbation of momenta is required. The color scheme of the background free-energy landscape changes from highest (white) to lowest (black) elevations; this convention also applies to Figs. 2-4 and 7-9.

Figure 2. In the TIS method, the phase space between two well defined stable states A and B is partitioned via a series of non-intercepting interfaces. The average rate constant is estimated by measuring the effective positive flux through these interfaces. For each pair of adjacent interfaces (λ_i, λ_{i+1}), a path ensemble simulation is performed where a new path belonging to this ensemble is generated if it starts in A, crosses λ_i and then either crosses λ_{i+1} or returns to region A (red). Trajectories that do not cross λ_i are not part of this ensemble (blue).

Figure 3. In the DFFS, a branched path (red thick lines) is constructed by generating partial paths between consecutives interfaces λ_i and λ_{i+1} for $0 \le i \le n$-1. The first stage entails a simulation run in the A basin shown by a green thick line. Starting points for the subsequent generation of branched paths are labeled with a gray circle at λ_0. The second stage corresponds to the trial runs (M_i) fired from λ_i ; those that reached the next λ_{i+1} interface are shown by a red thick line and those which failed to reach λ_{i+1} are shown by a blue dotted line.

Figure 4. Generation of a branched path (red thick lines) in the BG method. The first stage involves a simulation run in the A basin shown by the green line. Starting points for the subsequent generation of branched paths are marked with a gray circle at λ_0. The second stage corresponds to the trial runs (k_i)

fired from λ_i ; those that reached the next λ_{i+1} interface are shown by a red thick line and those that failed to reach λ_{i+1} are shown by a blue dotted line.

Figure 5. Schematic illustration of pseudoparticle dynamics under the WE method. The partial trajectories (or pseudoparticles, represented as partially filled circles) move along a progress coordinate (dashed line), subdivided into 3 bins (slabs in configurational space), from basin A to B. After every period of time τ of dynamic evolution, a pruning-enriching process maintains the balance of the weights and number of pseudoparticles such that each bin contains 2 reactive trajectories (adapted from Ref. [43]).

Figure 6. Schematic illustration of: (a) three short MD trajectories started from milestone H_i , with initial positions distributed according to $X \in H_i$, and reaching neighboring milestones H_{i+1} or H_{i-1}; where the basin A to B represent the reactant and product states, respectively (adapted from Ref. [49]); and (b) the initial (iteration 0) and final (iteration N) states of a path sampling simulation using the string method, highlighting the iterative redistribution of the isocommitor hypersurfaces Γ_z and the development of a defined reaction tube delimited by the dashed lines; the basin A to B are defined as in (a) (adapted from Ref. [62]).

Figure 7. Sketch of the conventional committor analysis procedure. The committor for a state along a trajectory (thick solid line) is estimated from the fraction of fleeting trial trajectories started therein that reach region B. Partial paths reaching B are show by a red line and those which reached A before than B are shown by a red line. $p_B \approx 0$ for states close to the A basin and $p_B \approx 1$ for states close to the B basin. The ensemble of configurations for which $p_B \approx 1/2$ is called transition state ensemble (TSE).

Figure 8. (a) A schematic illustration of the generation of committor probabilities, $p_B(\lambda)$, via the FFS-LSE method. In this example, $k_i=4$, $k_{i+1}=3$, and $k_{i+2}=2$. $p_{B_j}^{\,n} = 1$ for all points collected at $\lambda_{n=B}$. $p_{B_j}^{\,i+1}$ for the points j at λ_{i+1} are estimated from Eq. (27) as follows: $p_{B1}^{\,i+1} = 1/3 \times \left[2/2 + 1/2\right]$, $p_{B2}^{\,i+1} = 1/3 \times \left[1/2 + 2/2\right]$, and $p_{B3}^{\,i+1} = 1/3 \times \left[1/2 + 1/2 + 2/2\right]$. The $p_{B1}^{\,i}$ value for the point 1 at λ_i

is then obtained from: $p_{B_1}^{\ i} = 1/4 \times \left[p_{B_1}^{\ i+1} + p_{B_2}^{\ i+1} + p_{B_3}^{\ i+1} \right] = 1/4 \times \left[1/2 + 1/2 + 2/3 \right]$. The TSE is enclosed by the white ellipse. (b) Predicted p_B model derived from the FFS-LSE method for the motion of a particle on a 2D potential energy surface. [32] The estimated RxC iso-lines are shown as dashed (black) lines and commitor values as labels. The 2D potential energy surface is from Ref. [20] and the corresponding free energy landscape is shown as a colored contour plot where the color scheme changes from red (highest) to blue (lowest) elevations. The initial and final regions are shown by the squares labeled *A* and *B*, respectively. Adapted from Ref. [32].

Figure 9. Initial and optimum λ staging for the system of a particle moving in a two dimensional energy potential. The color scheme changes from highest (black) to lowest (white) elevations. The initial and final regions are shown by the circles labeled A and B, respectively. The initial λ staging for the FFS-type simulation is shown by dotted lines (white). The optimized $\{\lambda_i'\}$ set was obtained for the prescribed values of $P(\lambda_{i+1} \mid \lambda_i) = \left[P(\lambda_n \mid \lambda_0) \right]^{1/n}$, where n=9 is the number of interfaces. The thick line (red) shows the optimized $\{\lambda_i'\}$ staging. Note that interfaces are concentrated in the region preceding the transition state (TS), i.e., in the "bottleneck" of the simulation. In the free-energy surface contour plot, the TS is visually identifiable at $\lambda = x = 0$ where $p_B = 0.5$ (x is the x-coordinate). Since $P(\lambda_{n=B} \mid \lambda_{n-1}) = \left\langle p_B \right\rangle_{\lambda_{n-1}} \approx 0.44$, λ_{n-1} is located a little before $\lambda = x = 0$ (the TSE region).

Figure 10. Frontal view of the system used to simulate the flow-induced translocation of DNA through a rectangular pore [106], depicting three snapshots of a 420μm DNA molecule at different stages during the translocation.

Figure 11. Stable states defining the reactions for the bimolecular isomerization of alanine: $C7_{eq} \Leftrightarrow C5$ and $C7_{eq} \Leftrightarrow C_{7ax}$ transitions in vacuum (dotted lines) and $C5/C7_{eq} \Leftrightarrow \beta_2/\alpha_R$ transitions in explicit solvent (solid lines).

Figure 12. Comparison of the rate constant estimates for the isomerization of alanine dipeptide in vacuum (black bars) and explicit solvent (gray bars). The *x*-axis corresponds to the reference number.

Figure 13. Results from a FFS-MD simulations for the $C7_{eq} \Rightarrow C5$ transition of alanine dipeptide in vacuum at 300 K. over the ψ-ϕ plane. The color scheme for the underlying free energy landscape changes from highest (gray/light yellow) to lowest (black/dark red) elevations. The solid (black) lines correspond to particular interfaces of the initial order parameter $\lambda = \psi$: 80 (state A upper limit) and 150 (state B lower limit). The dashed lines correspond to the $\lambda = p_B$ isocommittor surface as computed from the FFS-LSE method (adapted from Ref. [73]).

Figure 14. Illustration of the two major routes between the native and unfolded states for the (un) folding of Trp-cage mini-protein. This sketch is based on Fig. 1 in Ref. [31].

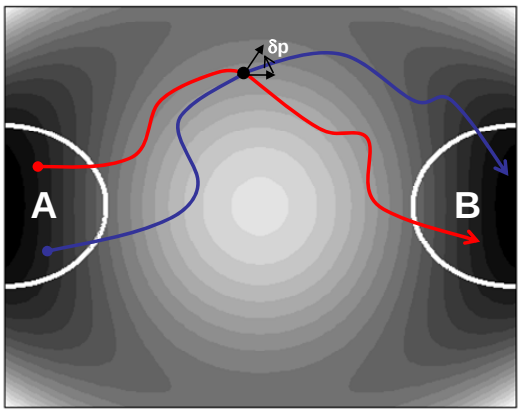


Figure 1

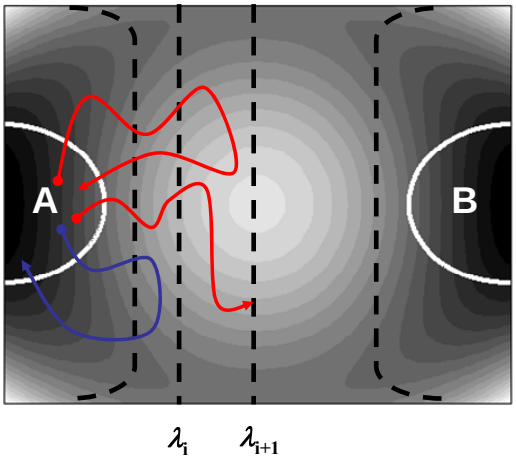


Figure 2

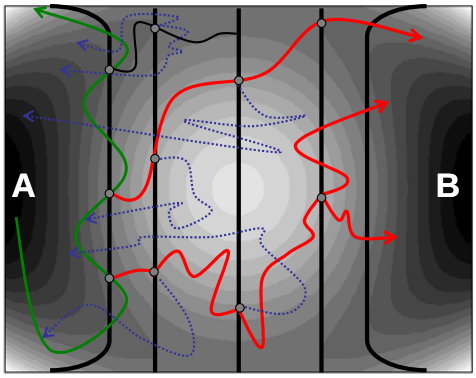

A
B
$\lambda_A=\lambda_0$
λ_1
λ_2
λ_{n-1}
$\lambda_n=\lambda_B$
$M_0=10$
$M_1=7$
$M_2=4$
$M_{n-1}=3$

Figure 3

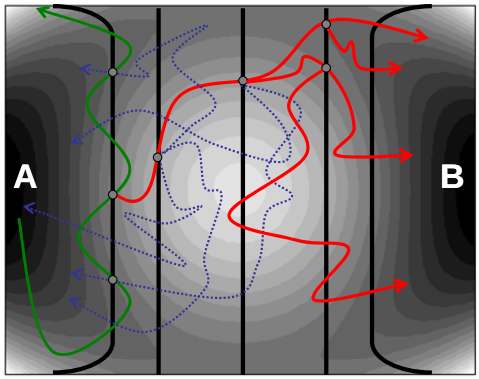

A
B
$\lambda_A=\lambda_0$
λ_1
λ_2
λ_{n-1}
$\lambda_n=\lambda_B$
$k_0=10$
$k_1=4$
$k_2=4$
$k_{n-1}=2$

Figure 4

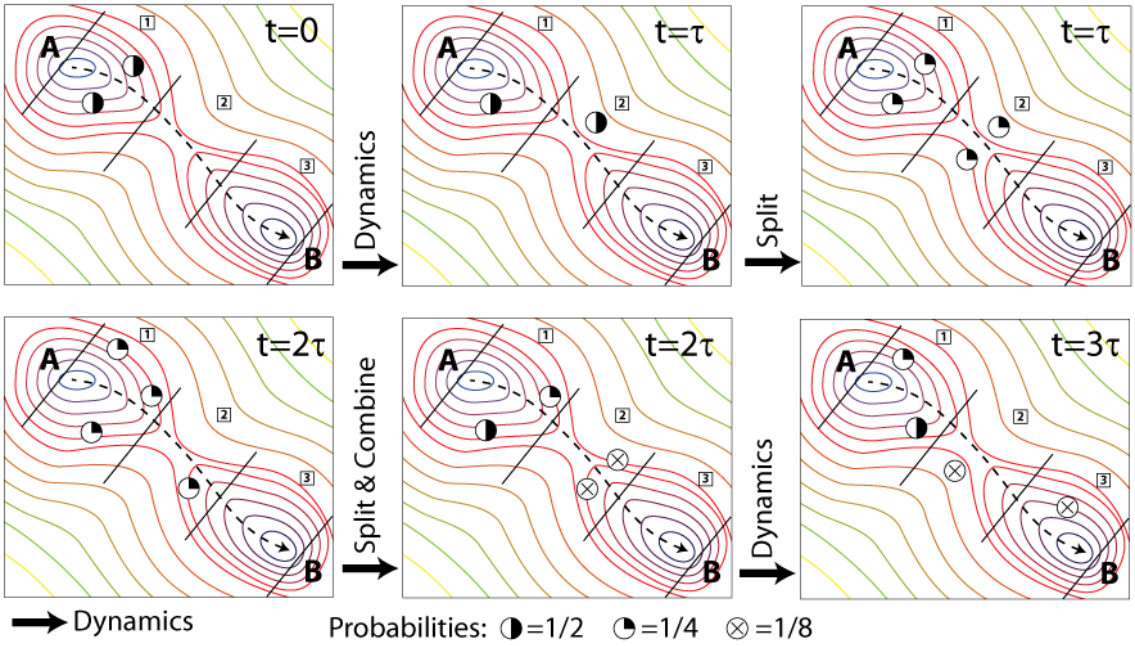

t=0
t=τ
t=τ
t=2τ
t=2τ
t=3τ
A
B
1
2
3
Dynamics
Split
Split & Combine
Dynamics
Dynamics
Probabilities: ◑=1/2 ◔=1/4 ⊗=1/8

Figure 5

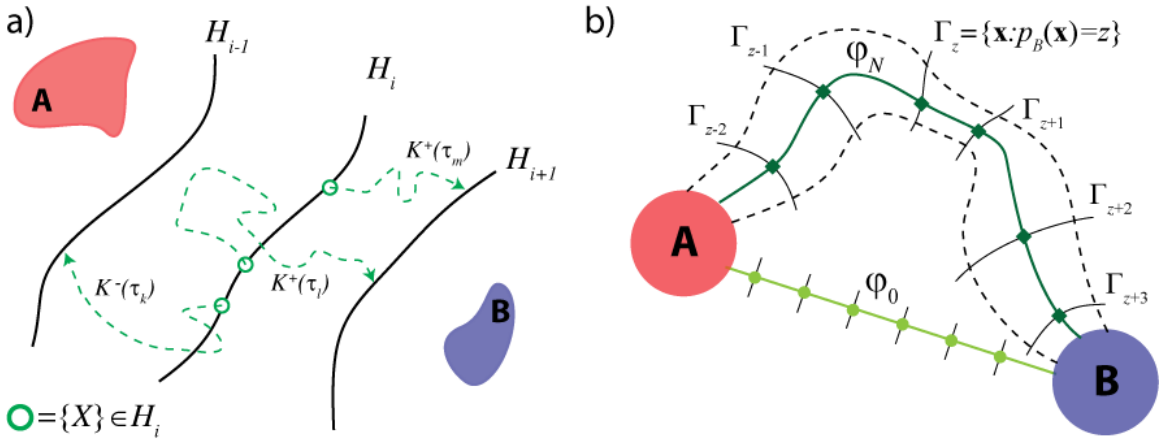

a)
A
B
H_{i-1}
H_i
H_{i+1}
$K^+(\tau_m)$
$K^-(\tau_k)$
$K^+(\tau_l)$
$O=\{X\}\in H_i$
b)
A
B
Γ_{z-1}
$\Gamma_z=\{\mathbf{x}:p_B(\mathbf{x})=z\}$
Γ_{z-2}
Γ_{z+1}
Γ_{z+2}
Γ_{z+3}
φ_N
φ_0

Figure 6

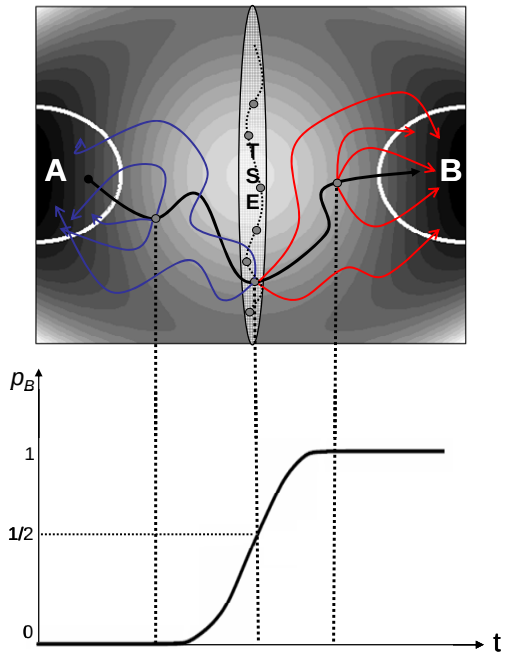
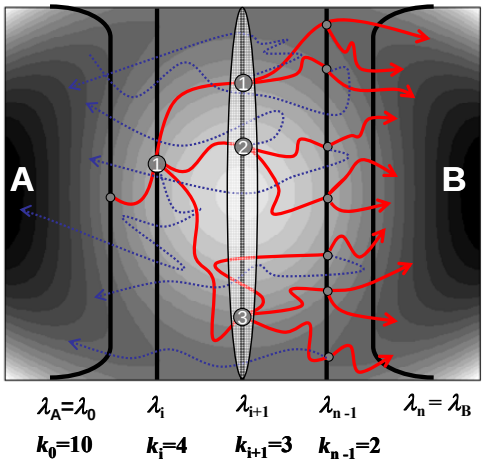


Figure 7

(a)


A
B
1
2
3
$\lambda_A=\lambda_0$
λ_i
λ_{i+1}
λ_{n-1}
$\lambda_n=\lambda_B$
$k_0=10$
$k_i=4$
$k_{i+1}=3$
$k_{n-1}=2$

(b)

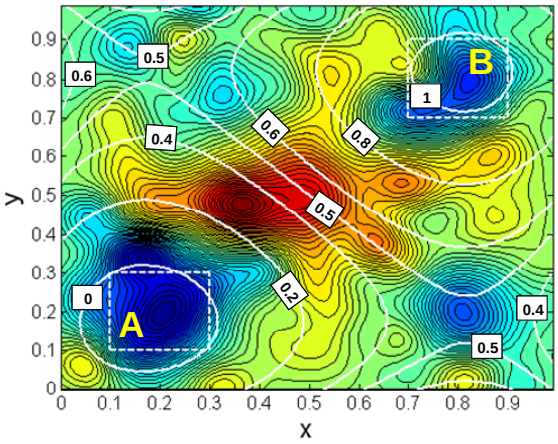

A
B
0
0.2
0.4
0.5
0.6
0.8
1
y
x
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Figure 8

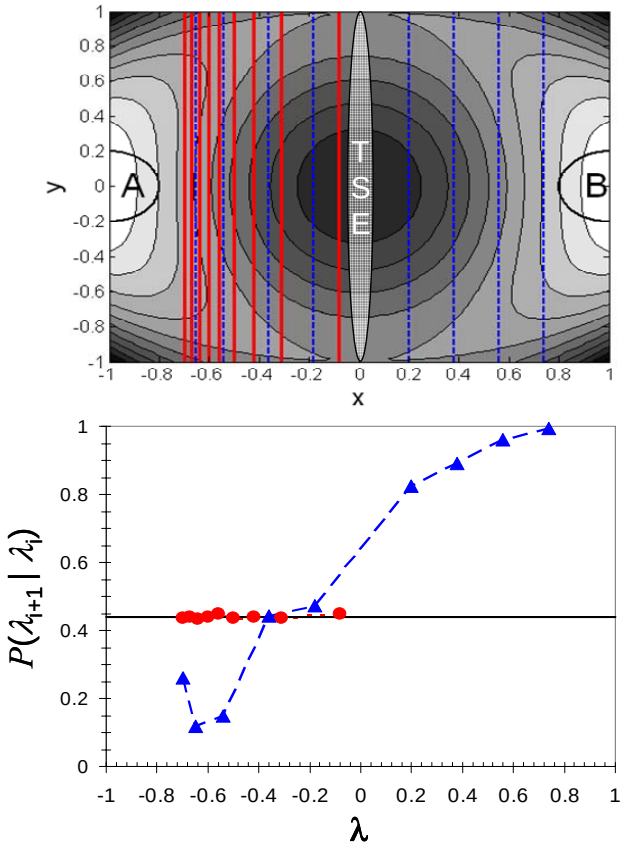


Figure 9

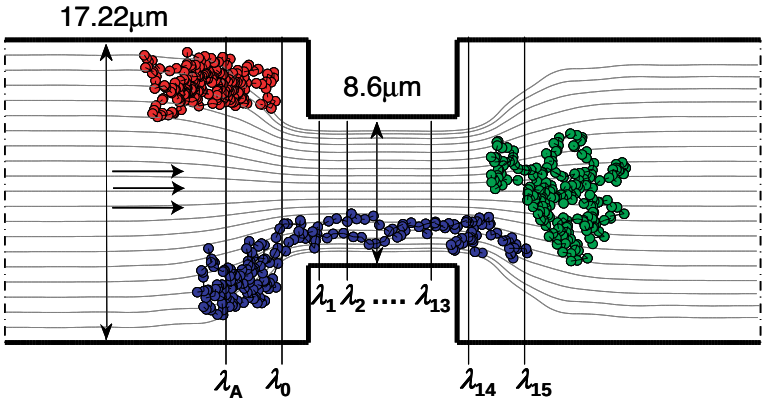

17.22μm
8.6μm
λ_1 λ_2 .... λ_{13}
λ_A
λ_0
λ_{14}
λ_{15}

Figure 10

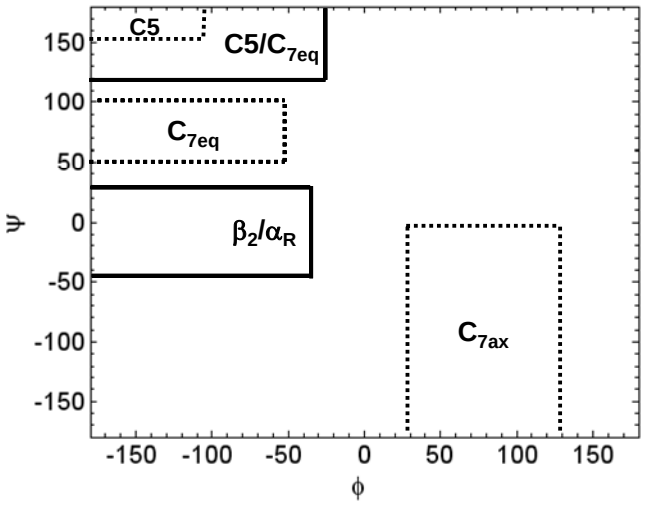

C5
C5/C7eq
C7eq
β2/αR
C7ax
150
100
50
0
-50
-100
-150
ψ
-150 -100 -50 0 50 100 150
ϕ

Figure 11

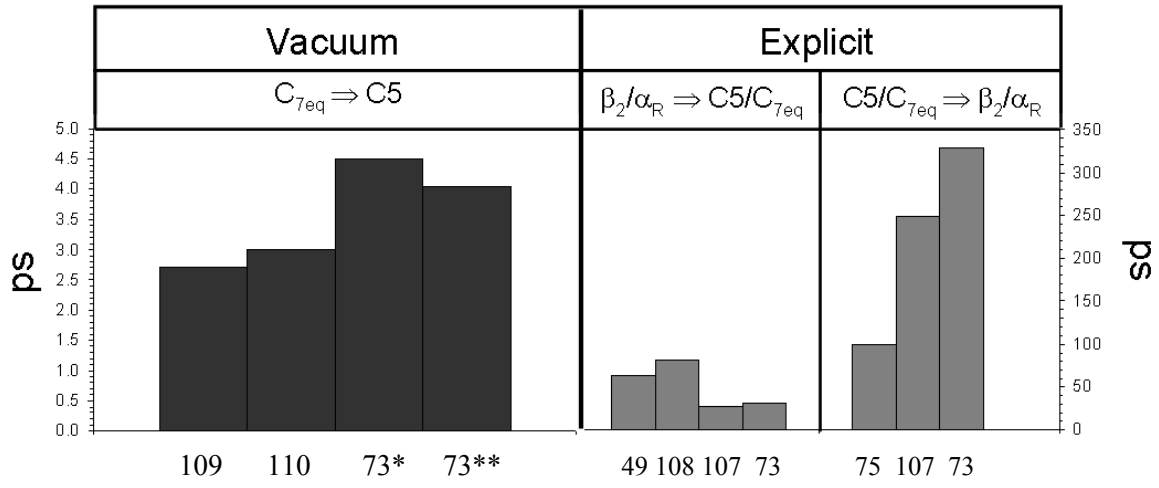

Vacuum
Explicit
$C_{7eq} \Rightarrow C5$
$\beta_2/\alpha_R \Rightarrow C5/C_{7eq}$
$C5/C_{7eq} \Rightarrow \beta_2/\alpha_R$
ps
sd
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
350
300
250
200
150
100
50
0
109 110 73* 73**
49 108 107 73
75 107 73

Figure 12

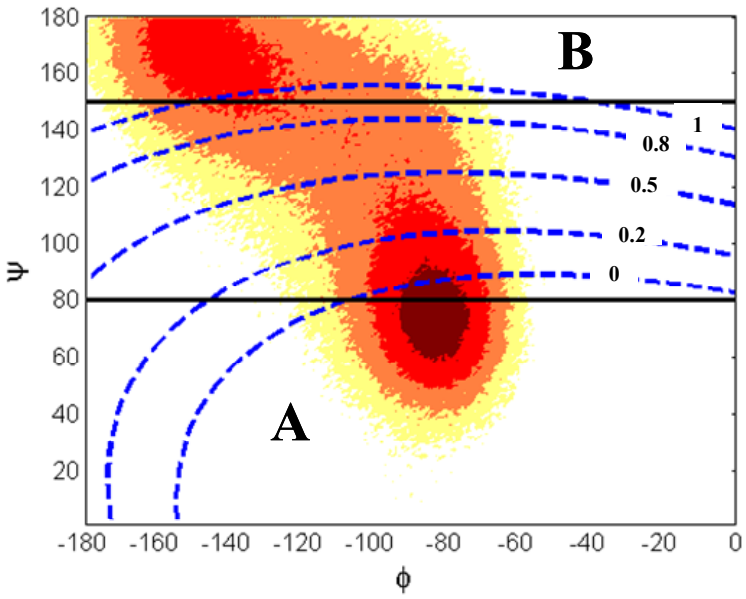

B
A
1
0.8
0.5
0.2
0
ψ
ϕ
180
160
140
120
100
80
60
40
20
-180
-160
-140
-120
-100
-80
-60
-40
-20
0

Figure 13

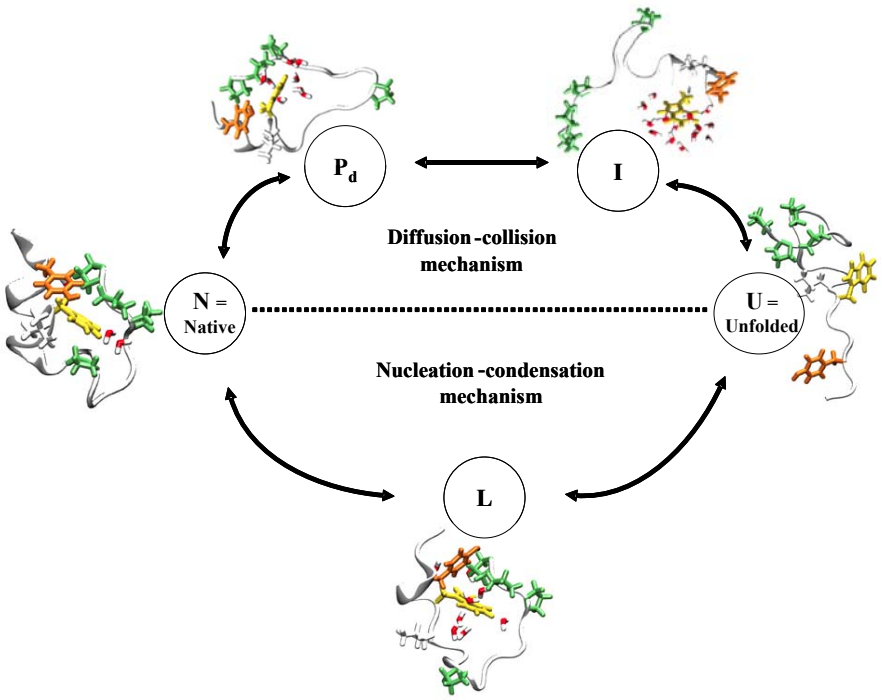

P_d
I
Diffusion-collision mechanism
N = Native
U = Unfolded
Nucleation-condensation mechanism
L

Figure 14