PCCP
Dynamic Article Links

Cite this: *Phys. Chem. Chem. Phys.*, 2011, **13**, 16912–16927

www.rsc.org/pccp

PERSPECTIVE

Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables†

Jan-Hendrik Prinz, Bettina Keller and Frank Noé

Received 20th April 2011, Accepted 26th July 2011
DOI: 10.1039/c1cp21258c

Markov (state) models (MSMs) have attracted a lot of interest recently as they (1) can probe long-term molecular kinetics based on short-time simulations, (2) offer a way to analyze great amounts of simulation data with relatively little subjectivity of the analyst, (3) provide insight into microscopic quantities such as the ensemble of transition pathways, and (4) allow simulation data to be reconciled with measurement data in a rigorous and explicit way. Here we sketch our current perspective of Markov models and explain in short their theoretical basis and assumptions. We describe transition path theory which allows the entire ensemble of protein folding pathways to be investigated and that combines naturally with Markov models. Experimental observations can be naturally linked to Markov models with the dynamical fingerprint theory, by which experimentally observable timescales can be equipped with an understanding of the structural rearrangement processes that take place at these timescales. The concepts of this paper are illustrated by a simple kinetic model of protein folding.

1 Introduction

Folding of proteins and other macromolecules depends on their ability to undergo conformational transitions between substates. A hallmark of protein dynamics is that these substates are often metastable sets of structures, *i.e.* the protein will typically fluctuate within a set of structures for a long time before enough thermal energy is accumulated to leave this set and transition to another metastable set. Such dynamics has been conceptualized as a walk on a complex energy landscape with basins corresponding to metastable sets of structures and energy barriers separating these basins.[27] It is the interest of chemical physicists and biophysicists to identify the essential metastable states, quantify their free energies or probabilities, the kinetics arising from the transitions between them, and the structural mechanisms involved.

The study of protein folding has been a particular driving force in investigating protein dynamics *via* both experiment and simulation. The fact that protein folding is an intramolecular process involving changes to almost the entire structure makes a variety of physical or chemical probes and measurement techniques available that report on aspects of the folding process. For small and moderately sized proteins, simulations are now feasible that can access experimentally resolvable timescales.[64,78,87]

There is a large body of experimental research indicating that protein folding is characterized by single-exponential kinetics.[7,38,56] This suggests that protein folding is a two-state transition, where the two states are believed to correspond to an enthalpically stabilized native state and an entropically stabilized denatured state.[11,19,20,65] It has been noted, however, that it is unclear whether such a two-state interpretation is meaningful as the as-equilibration of the denatured state may not occur on a timescale faster than folding.[9,21] In current research, there is increased attention to complexity in the kinetics that was difficult to detect in earlier studies. Moreover, for some formerly apparent two-state folders, additional relaxation timescales have been found using measurement methods with increased resolution.[57,63,71] A more disturbing fact is that although kinetic experiments will in principle measure all relaxation timescales, any given combination of measurements and observables will only be sensitive to a few—typically one or two.[63] It is thus conceivable that protein folding has significantly more complex kinetics than apparent in individual experiments. This feeling seems to be supported by careful analyses of single-molecule experiments that have reported on the existence of multiple metastable states.[25,29,30,52,55,73,92] Also ensemble experiments can, with appropriate design, probe conformational heterogeneity, hidden intermediates and the existence of parallel pathways.[32,46–48,80]

In order to overcome the limitation of indirect observability of experiments, molecular dynamics (MD) simulations are becoming increasingly accepted as a tool to investigate structural details of molecular processes and relate them to experimentally resolved features.[64,75,85] In the simulation community, there is also a tendency to move towards more sophisticated analysis methods. Previous projections of the

FU Berlin, Arnimallee 6, 14195 Berlin.
E-mail: jan-hendrik.prinz@fu-berlin.de, bettina.keller@fu-berlin.de, frank.noe@fu-berlin.de

† This article was submitted as part of a themed collection on the Physical Foundations of Protein Folding.

simulation data onto simple one- or two-dimensional observables usually suggest simplicity in the kinetics, this however owing often to a disguise of the true and often complex nature of the kinetics by creating overlaps between kinetically distinct structures.[44,54,58] In the past few years, there has been a rapid increase of studies that first partition the simulated structural data into relatively small substates and then study the kinetics that emerges from a transition network between these substates.[12,17,37,40,54,58,66,70,76,89,90] Early work has focused on energy-landscape models that use rate theories to generate transition networks between local potential energy minima.[60,61,89] More recently, approaches that are based on directly counting transitions in MD simulations were established.[14,44,54,59,66,76,83] The resulting models are often called transition networks, Master equation models or Markov (state) models (MSMs), where "Markovianity" means that the kinetics are modeled by a memoryless jump process between states.[14,15,18,59,64,67,77,79,83,84]

MSMs provide access to the complexity of the essential kinetics without discarding information by projection onto order parameters. The essential kinetics can be understood by studying the metastable sets arising from the model[59,77,90] or graph-based visualization tools.[70] The ensemble of folding pathways can be calculated and quantified from MSMs with Transition Path Theory.[64] Kinetic experimental measurements can be calculated from MSMs directly and the experimentally-detectable kinetic features can be linked to structural changes.[63] In the present review, we will explain the essentials of MSM theory and these analysis methods.

2 Markov models

In this section we outline the basic theory of Markov models, explain where their limitations are and how well they can approximate original kinetics, and give approaches for constructing Markov models from simulation data.

2.1 Basics

We briefly sketch the main mathematical ideas underlying Markov models of molecular kinetics. The dynamics of a molecular system can be understood as a long trajectory $\boldsymbol{x}(t)$ describing the positions and momenta of all atoms considered. The state space of positions and momenta is huge—6N dimensions when a molecular system with N atoms are considered. However, we know that in most macromolecular systems only a very small subset of this state space is actually populated—namely the region which contains conformations of relatively low energies. It is therefore reasonable to ask whether we can computationally characterize this region by subdividing it into sets, each of which comprising a group of similar molecular structures. We then aim at approximately describing the dynamics in terms of the transition probabilities between these sets. This state-space discretization and the corresponding transition probabilities will be our Markov model.

More formally, it is assumed that the molecular system studied lives in a continuous state space Ω consisting of positions and momenta, its time evolution $\boldsymbol{x}(t)$ obeys the following properties:

(1) $\boldsymbol{x}(t)$ is a Markov process in the full state space Ω, *i.e.* the instantaneous change in $\boldsymbol{x}$ only depends on the current value of $\boldsymbol{x}$ and not its history.

(2) $\boldsymbol{x}(t)$ is ergodic, *i.e.* all states of Ω could be reached by an infinitely long trajectory and are visited with a frequency given by the Boltzmann distribution:

$$\mu(\boldsymbol{x}) = \mathrm{Z}(\beta)^{-1} \exp(-\beta H(\boldsymbol{x})). \quad (1)$$

(3) $\boldsymbol{x}(t)$ is reversible, *i.e.*, the probability density of going from state $\boldsymbol{x}$ to state $\boldsymbol{y}$ in time τ, $p(\boldsymbol{x},\boldsymbol{y};\tau)$, fulfills the condition of *detailed balance*:

$$\mu(\boldsymbol{x})p(\boldsymbol{x},\boldsymbol{y};\tau) = \mu(\boldsymbol{y})p(\boldsymbol{y},\boldsymbol{x};\tau). \quad (2)$$

These conditions are fulfilled by many dynamical models frequently used to simulate molecular dynamics, such as Anderson-thermostatted dynamics or Hybrid Monte Carlo. We can then perform the following, at this stage purely formal, trick, and describe the evolution of the dynamics in terms of an ensemble distribution $p_t(\boldsymbol{x})$:

$$p_{t+\tau}(\boldsymbol{x}) = \mathcal{Q}(\tau)\cdot p_t(\boldsymbol{x}). \quad (3)$$

This means when $p_t(\boldsymbol{x})$ is the probability distribution of molecules in the ensemble at time t, then $p_{t+\tau}(\boldsymbol{x})$ is the probability distribution of the ensemble at a later time $t + \tau$. The evolution of probability density is described by the operator $\mathcal{Q}(\tau)$. The important fact of this equation is that the same operator $\mathcal{Q}(\tau)$ holds at all times t and that it is a linear operator, *i.e.* a mathematically simple object, which allows us to propagate to arbitrarily long times by repeated usage:

$$p_{t+2\tau}(\boldsymbol{x}) = \mathcal{Q}(\tau)\cdot(\mathcal{Q}(\tau)\cdot p_t(\boldsymbol{x})) = \mathcal{Q}(2\tau)\cdot p_t(\boldsymbol{x}). \quad (4)$$

and so on. The whole purpose of Markov models is to discretize state space Ω such that $\mathcal{Q}(\tau)$ can be approximated by a matrix and $p_t(\boldsymbol{x})$ can be approximated by a vector, such that the equations above are well approximated. We will explain below that this is indeed possible even for complex molecular systems, and that then all interesting long-time dynamical quantities can be calculated from the discrete version of $\mathcal{Q}(\tau)$ despite that only short trajectories of length τ are needed. Note that τ can be orders of magnitude shorter than the longest timescales of the system.

Before going into the discrete representation, we shall first illustrate how $\mathcal{Q}(\tau)$ operates on a one-dimensional example. Fig. 1a shows a potential energy landscape with associated Boltzmann density $\mu(\boldsymbol{x})$. Fig. 1b is an illustration of the operator $\mathcal{Q}(\tau)$: the horizontal and vertical axes correspond to the coordinate $\boldsymbol{x}$ and the color coding quantifies how much probability density is transported between two points $\boldsymbol{x}$ in a time τ. The dark colour blocks near the diagonal correspond to the fact that there is a high probability to move around within an energy basin, while the white colors in off-diagonal regions correspond to the fact that there is a small probability to jump between the basins.

Fig. 1c–e show a spectral decomposition of the operator $\mathcal{Q}(\tau)$ which will be discussed below. The eigenvalues shown in Fig. 1c and eigenfunctions shown in Fig. 1d fulfill the equations

$$\mathcal{Q}(\tau)\cdot\phi(\boldsymbol{x}) = \lambda_i\phi(\boldsymbol{x}).$$

whose relevance will also be explained below.

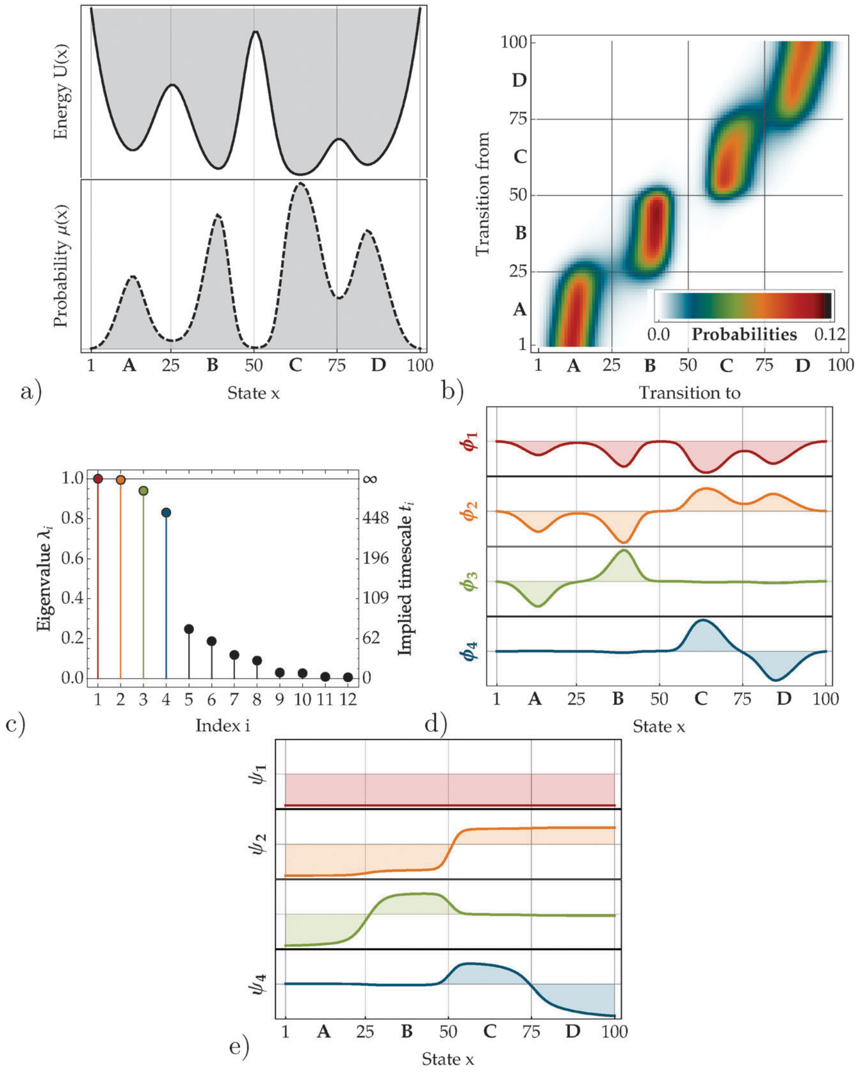

Fig. 1 (a) Potential energy function with four metastable states and corresponding stationary density $\mu(\boldsymbol{x})$, (b) density plot of the transfer operator for a simple diffusion-in-potential dynamics defined on the range $\Omega = [0,100]$, black and red indicates high transition probability, white zero transition probability. Of particular interest is the nearly block-diagonal structure, where the transition density is large within blocks allowing rapid transitions within metastable basins, and small or nearly zero for jumps between different metastable basins. (c) Eigenvalues of the transfer operator, the gap between the four metastable processes ($\lambda_i \approx 1$) and the fast processes is clearly visible. (d) The four dominant eigenfunctions of the operator $\mathcal{Q}(\tau)$, $\phi_1,\ldots,\phi_4$, which indicate the associated dynamical processes. The first eigenfunction is associated to the stationary process, the second to a transition between $A + B \leftrightarrow C + D$ and the third and fourth eigenfunction to transitions between $A \leftrightarrow B$ and $C \leftrightarrow D$, respectively. (e) The eigenfunctions weighted with the $\mu(\boldsymbol{x})^{-1}$.

Imagine now that the coordinate $\boldsymbol{x}$ is discretized into sets $\{S_1,\ldots,S_n\}$. It is obvious that when these sets are many and small enough, we can approximate $\mathcal{Q}(\tau)$ by discrete transition probabilities between sets. $T_{ij}(\tau)$ represents the time-stationary probability to find the system in state j at time $t + \tau$ given that it was in state i at time t:

$$T_{ij}(\tau) = \mathbb{P}[\boldsymbol{x}(t+\tau) \in S_j | \boldsymbol{x}(t) \in S_i],$$

defining a transition matrix $\boldsymbol{T}(\tau) \in \mathbb{R}^{n\times n}$. The transition matrix can also be written in terms of correlation functions:[83]

$$T_{ij}(\tau) = \frac{c_{ij}^{\mathrm{corr}}(\tau)}{\pi_i} \qquad (5)$$

where π_i is the stationary probability to be in set S_i:

$$\pi_i = \int_{\boldsymbol{x}\in S_i} \mathrm{d}\boldsymbol{x}\mu(\boldsymbol{x}),$$

and the unconditional transition probability $c_{ij}^{\mathrm{corr}}(\tau) = \pi_i T_{ij}(\tau)$ is an equilibrium time correlation function which is normalized such that $\sum_{i,j} c_{ij}^{\mathrm{corr}}(\tau) = 1$. Since we assume the dynamics to fulfill detailed balance, the correlation matrix is symmetric ($c_{ij}^{\mathrm{corr}}(\tau) = c_{ji}^{\mathrm{corr}}(\tau)$). If we would manage to generate a very long trajectory $\boldsymbol{x}(t)$ and simply count transitions in time steps τ, we would obtain a count matrix $c_{ij}(\tau)$ that is proportional to $c_{ij}^{\mathrm{corr}}(\tau)$.

Suppose that $\boldsymbol{p}(t) \in \mathbb{R}^n$ is a column vector whose elements denote the probability, or population, to be within a set

$j \in \{1,\ldots,n\}$ at time t. After time τ, the probabilities will have changed according to:

$$p_j(t+\tau) = \sum_{i=1}^{n} p_i(t) T_{ij}(\tau), \tag{6}$$

or in matrix form:

$$\boldsymbol{p}^T(t + \tau) = \boldsymbol{p}^T(t)\boldsymbol{T}(\tau) \tag{7}$$

Note that an alternative convention often used in the literature is to write $\boldsymbol{T}(\tau)$ as a column-stochastic matrix, obtained by taking the transpose of the row-stochastic transition matrix defined here.

The stationary probabilities of discrete states, π_i, yield the unique discrete stationary distribution of $\boldsymbol{T}$:

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \boldsymbol{T}(\tau). \tag{8}$$

2.2 Estimation and statistics

In practice, the transition probabilities cannot be directly calculated. Instead, we have a microscopic model of the molecular system which permits us to calculate energies and forces at every state $\boldsymbol{x}$ and a dynamical model (*e.g.* integrator + thermostat) which propagate these dynamics with short timesteps (typically femtoseconds). Suppose we have this machinery to generate trajectories of our molecular system. Then we can use these trajectories to estimate the transition probability between any pair of discrete sets S_i and S_j. Of course, such an estimation will involve an estimation error resulting from finite sampling, and this error will become smaller the more trajectory data are generated.

More formally, consider one trajectory generated under equilibrium conditions with N configurations stored at a fixed time interval Δt:

$$X = [\boldsymbol{x}(t = 0), \boldsymbol{x}(t = \Delta t),\ldots,\boldsymbol{x}(t = (N-1)\Delta t)] \tag{9}$$

$$= [\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N] \tag{10}$$

and consider that a state space discretization has been defined such that each structure can be assigned to one discrete state $\boldsymbol{x}_k \in S_i \rightarrow s_k = i$, and the trajectory information can be simply stored as the sequence $s_1,\ldots,s_N$ of discrete states.

We also assume that $\boldsymbol{x}_1$ was drawn from the equilibrium density pertaining to state s_1, $\mu_{s_1}(\boldsymbol{x})$ (see discussion in ref. 64, 69 and 72). We can now define the discrete state count matrix $C^{\text{obs}}(\tau) = [c_{ij}^{\text{obs}}(\tau)]$ at lag time τ, where τ now needs to be an integer multiple of the available data resolution Δt:

$$c_{ij}(\tau) = c_{ij}(l\Delta t) = |\{s_k = i, s_{k+1} = j\}\ |k = 1\ldots N-l|. \tag{11}$$

which provides an estimator of the correlation matrix defined in eqn (5) by:

$$\hat{c}_{ij}^{\text{corr}}(\tau) = \frac{c_{ij}(\tau)}{N-l}. \tag{12}$$

$\boldsymbol{C}(\tau)$ simply counts the number of observed transitions between discrete states, *i.e.* c_{ij} is the number of times the trajectory was observed in state i at time t and in state j at time $t + \tau$, summed over all times t. If multiple trajectories are available, then the count matrices of these trajectories are simply added up. It can be shown[69] that based on $\boldsymbol{C}(\tau)$, the transition matrix can be estimated with maximum likelihood by:

$$\hat{T}_{ij} = \frac{c_{ij}}{c_i}, \tag{13}$$

where c_i are the row sums of $\boldsymbol{C}$:

$$c_i := \sum_{k=1}^{n} c_{ik}, \tag{14}$$

which are equal to the total number of times the trajectory was found in state i. This estimator is asymptotically unbiased, *i.e.* for a long enough trajectory it will converge to the correct transition matrix:

$$\lim_{N\to\infty} \hat{T}_{ij} = T_{ij}. \tag{15}$$

Since simulation data are finite, all validation procedures (either consistency checks or comparisons to experimental data) need to account for statistical uncertainties. For these, standard deviations or confidence intervals induced by the posterior distribution of transition matrices are of interest. It follows from well-known properties of the distribution of transition matrices[2] that the expectation value for transition matrices is

$$\bar{T}_{ij} = \mathbb{E}[\hat{T}_{ij}] = \frac{c_{ij}+1}{c_i+n}, \tag{16}$$

and the variance is given by

$$\text{Var}[\hat{T}_{ij}] = \frac{(c_{ij}+1)((c_i+n)-(c_{ij}+1))}{(c_i+n)^2((c_i+n)+1)} = \frac{\bar{T}_{ij}(1-\bar{T}_{ij})}{c_i+n+1}. \tag{17}$$

Also everything calculated from $\boldsymbol{T}(\tau)$ will have statistical error associated. These errors can be rigorously evaluated.[16,35,62,68]

It is important to note that $\hat{T}_{ij}$ as given by eqn (13) does not necessarily fulfill the detailed balance equations: $\pi_i T_{ij} = \pi_j T_{ji}$, but generally $\pi_i \hat{T}_{ij} \neq \pi_j \hat{T}_{ji}$. This is a result of limited statistics and can be avoided by using a maximum likelihood estimator that makes sure that the detailed balance equations are fulfilled.[69]

2.3 Predicting long-term kinetics from short simulations and the systematic error done by this

Markov models are an approximation of molecular kinetics in two ways: as discussed above, a Markov model is estimated from a finite number of trajectories and thus involves statistical error. However, there is a systematic source of error, which is addressed here: the fact that we discretize state space into sets $(S_1,\ldots,S_n)$ erases the information where exactly the continuous process $\boldsymbol{x}(t)$ was. As a result, the jump process on $(S_1,\ldots,S_n)$ is no longer Markovian even if $\boldsymbol{x}(t)$ is, nevertheless we approximate it by a Markov chain. This apparent contradiction is what has raised criticism against Markov models.

The purpose of this section is to make clear that this criticism is equally justified or unjustified as in any other area of numerics. Consider the numerical evaluation of the area under a curve by approximating it with a finite number of step functions and adding up their areas. Despite the fact that the curve of interest may not nearly be step-like, we trust numerical integrators, because we know they can deliver the desired result to arbitrary precision by making the discretization finer—and that "fine enough" is practically feasible. In other words, we have some way to control the error. For Markov models we can get a similar result, although Markov model numerics is not yet as well developed as other areas of numerics. We can make useful theoretical statements of the systematic error introduced by the discretization. What may currently be more important is that a practical test is available to validate that a Markov model built is quantitatively acceptable.

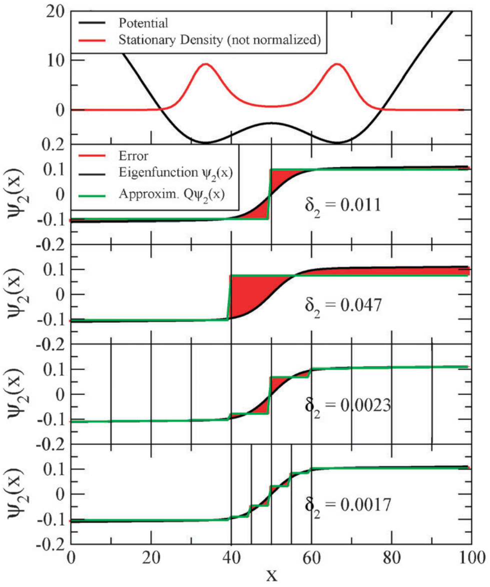

Fig. 2 Illustration of the eigenfunction approximation error on the slow transition in the diffusion in a double well (top, black line). The slowest eigenfunction is shown in the lower four panels (black), along with the step approximations (green) of the partitions (vertical black lines) at $x = 50$; $x = 40$; $x = 10, 20,\ldots,80, 90$; and $x = 40,45,50,55,60$. The eigenfunction approximation error δ_2 is shown as red area and its norm is printed. Figure adapted from ref. 69.

The following two quantities are obtained from Markov models *without* systematic error:

(1) The propagation of transition probabilities by one step τ, $\boldsymbol{p}^T(t + \tau) = \boldsymbol{p}^T(t)\boldsymbol{T}(\tau)$.

(2) Stationary properties, such as the stationary distribution $\boldsymbol{p}$ and associated expectation of state functions $\mathbb{E}_\pi(a) = \langle \boldsymbol{\pi},\boldsymbol{a}\rangle$.

However, state space discretization introduces systematic error is in the reproduction of long-time kinetics, *i.e.* the prediction:

$$\boldsymbol{p}^T(t + k\tau) \approx \boldsymbol{p}^T(t)\boldsymbol{T}^k(\tau), \tag{18}$$

is only approximately true. However, good approximation of this equation is essential, because it represents one of the main advantages of Markov models, namely to predict long-time kinetics by using short trajectories of length order τ. Based on rigorous theoretical results of ref. 22, 69 and 74, the following statements are true:

(1) The error of eqn (18) decreases with increasingly small discretization states. Quantitatively, what matters is how well the discretization can approximate the slow eigenfunctions (weighted with the stationary density, see Fig. 1e).

(2) The error of the approximation of timescales t_i decreases when the discretization better approximates the corresponding eigenfunction.

(3) For a given discretization, the error both of eqn (18) and of the approximation of timescales t_i decreases with increasing lag time τ.

These results are illustrated in Fig. 2 and 3. A diffusion on the two-well potential shown in Fig. 2a has a sigmodial-shaped eigenfunction (when weighted by the stationary density) shown in Fig. 2b. When only using two states to discretize the state space, the separatrix is best placed on the transition state (Fig. 2b), or otherwise may generate a very large error (Fig. 2c). However, the error of the Markov model decreases

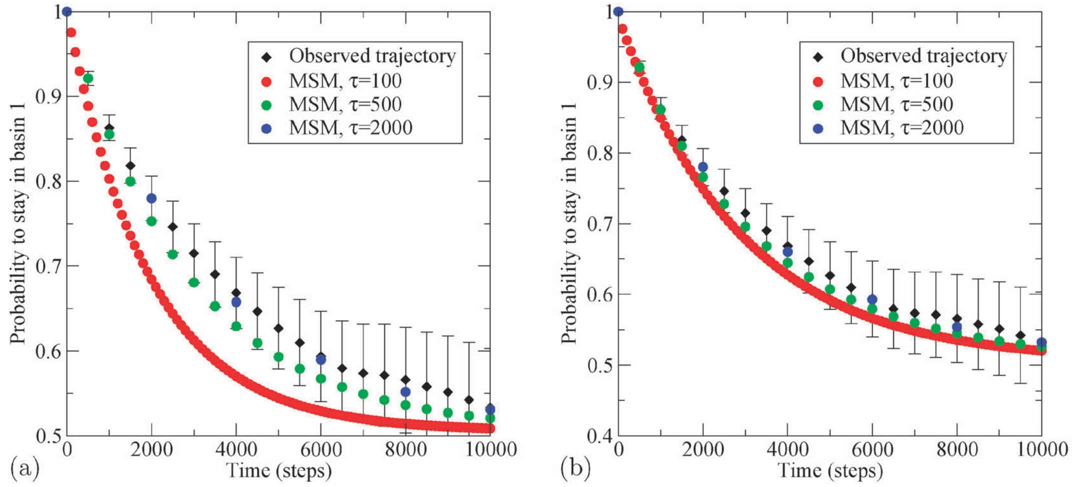

Fig. 3 The so-called Chapman–Kolmogorov test. This corresponds here to compare between MSM and original dynamics how the probability of being in the left minimum relaxes when starting in the left basin. The test was done for the two-well potential using a trajectory of length 10^6 steps. Tested are Markov models that use lag times $\tau = 100$, 500, 2000 and (a) 2-state discretization (split at $x = 50$), (b) 6-state discretization (split at $x = 40, 45, 50, 55, 60$). Figure adapted from ref. 69.

when more than two states are used (Fig. 2d and e). Thus, in contrast to previous assumptions, it is not the most metastable partition of state space that produces the best Markov model. Fig. 3 compares the propagation of probability by the Markov model and the true dynamics, shows the result of using the two- and the six-state partitions on the error in eqn (18) over time. The six-state partition clearly outperforms the two-state partition.

2.4 Spectral properties

At the end of our theoretical investigation of Markov models we come to the spectral properties of the operator $\mathcal{Q}(\tau)$ and the associated transition matrix $\boldsymbol{T}(\tau)$. Although this point is somewhat difficult to understand at first, it is essential in order to see what metastable states are, why some Markov models work better than others, and eventually also how kinetics experiments work. At this point a comparison to another approach that is more commonly used in the Chemical Physics community may be useful: consider the principal component analysis method,[1] where the relative distances of a set of data points (*e.g.* molecular structures) are captured by a covariance matrix. When performing an eigenvalue decomposition one obtains eigenvectors and eigenvalues. The eigenvectors with the largest eigenvalues are called "principal components" and describe the directions along which the data set has the greatest spatial extent. The corresponding eigenvalues capture the variance of the data set along these principal directions. Analogously, a transition matrix $\boldsymbol{T}(\tau)$ can also be decomposed into eigenvectors and eigenvalues. The eigenvectors also represent "principal modes", but since the transition matrix contains probabilities these modes are vectors that contain changes of the probability for each discrete state S_i. The principal modes with the largest eigenvalues are indeed the main modes of probability flow between the system's substates. The corresponding eigenvalues have magnitude expressing how slow or fast the corresponding probability flow occurs. Thus, the eigenvalue decomposition of a transition matrix may be understood as a principal component analysis of the dynamics.

More formally, transition matrices can, as any diagonalizable matrix, be written as a linear combination of their left eigenvectors, their eigenvalues and their right eigenvectors. For the here assumed case of matrices fulfilling detailed balance, the right eigenvalues can be replaced by the left eigenvalues (and *vice versa*), leading to the decomposition:

$$\boldsymbol{T}(\tau) = \boldsymbol{\Pi}^{-1}\sum_{i=1}^{n}\lambda_i(\tau)\boldsymbol{l}_i\boldsymbol{l}_i^{\top}. \tag{19}$$

with the diagonal matrix $\boldsymbol{\Pi}^{-1} = \mathrm{diag}(\pi_1^{-1},\ldots,\pi_n^{-1})$. Thus, for longer timescales:

$$\boldsymbol{T}^k(\tau) = \boldsymbol{\Pi}^{-1}\sum_{i=1}^{n}\lambda_i^k(\tau)\boldsymbol{l}_i\boldsymbol{l}_i^{\top}. \tag{20}$$

The transition matrix $\boldsymbol{T}(k\tau) = \boldsymbol{T}^k(\tau)$ which transports an initial probability k time steps forward is again a linear combination of the eigenvectors and eigenvalues. These linear combinations (eqn (19) and (20)) are known as *spectral decomposition* of the transition matrix. They are very useful for connecting the dynamics of the molecule to experimentally-measured signals, which is described in Section 6.

Eqn (20) is the key for understanding how the transition matrix transforms a probability vector. The complete process consists of n subprocesses $\boldsymbol{l}_i\boldsymbol{l}_i^{\top}$, each of which is weighted by the eigenvalue λ_i raised to the power k. Because the transition matrix is a row-stochastic matrix, it always has one eigenvalue which is equal to one $\lambda_1 = 1$.[18] Raising this eigenvalue to the power k does not change the weight of the corresponding subprocess $\boldsymbol{l}_1\boldsymbol{l}_1^{\top}$: $1^k = 1$. $\boldsymbol{l}_1\boldsymbol{l}_1^{\top}$ is the stationary process, which we postulated in eqn (8), and $\boldsymbol{l}_1 = \pi$. All other eigenvalues of the transition matrix are guaranteed to be smaller than the one in the absolute value.[18]

The weights of the processes hence decay exponentially with the implied timescale t_i of the decay process

$$t_i = -\frac{\tau}{\ln\lambda_i}. \tag{21}$$

Since the relaxation timescales t_i are physical properties of the dynamics, they should be invariant under change of the lag time τ used to parametrize the transition matrix.[83] For large enough τ, t_i should converge to their true value (assuming sufficient statistics). Therefore, the convergence of t_i with increasing τ has often been employed as an indicator for selecting τ.[14,59,69,83] For the two-well potential diffusion dynamics in Fig. 2, the τ-convergence of the slowest timescale t_2 is shown in Fig. 4. These curves illustrate that discretizations that allow for better approximation of the eigenfunctions also provide nearly-correct timescales at shorter lag times τ.

The smaller the eigenvalue λ_i, the smaller the implied timescale t_i, the faster the corresponding process decays. To understand the interplay of multiple relevant eigenvalues and eigenvectors let us review again Fig. 1 which shows the diffusion dynamics on an energy landscape with four basins (A, B, C, D) and high intervening energy barriers. Fig. 1d shows the 15 largest eigenvalues of the transition matrix in Fig. 1b. There is one eigenvalue, λ_1, which is equal to one,

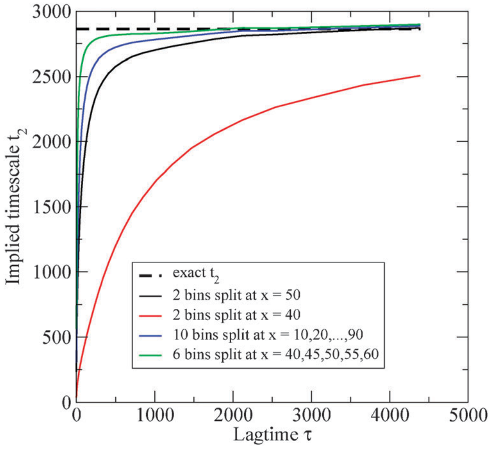

Fig. 4 Convergence of the slowest implied timescale $t_2 = -\tau/\ln\lambda_2(\tau)$ of the diffusion in a double-well potential depending on the MSM discretization. The metastable partition (black, solid) has greater error than non-metastable partitions (blue, green) with more states that better trace the change of the slow eigenfunction near the transition state. Figure adapted from ref. 69.

followed by three eigenvalues, λ_2 to λ_4, which are close to one. These four *dominant eigenvalues* are separated by a gap from the remaining eigenvalues. Hence, the transition matrix consists of a stationary process, three slow processes and many processes which decay quickly. After a few time steps, only the four dominant processes contribute to the evolution of the probability vector. How these processes alter this vector is determined by the shape of the corresponding eigenvectors.

Fig. 1c shows the four dominant right eigenvectors. The first eigenvector corresponds to the stationary process and is, therefore, constant. The second eigenvector corresponds to the slowest process and has positive signs in regions A and B and negative signs in regions C and D. This shape effectively moves probability density across the largest barrier in the energy surface. Since the eigenvector is approximately constant within the combined regions (A, B) and (C, D) left and right of the barrier, it does not alter the relative probability distribution within these regions. The third eigenvector, analogously, moves density between A and B, the fourth moves density between C and D.

3 Illustrative protein folding model

We use a simple protein folding model throughout this study in order to illustrate the concepts described in this paper. We consider three structural elements called *a*, *b* and *c* that form independent of each other. A simple energy model has been designed in which the folding of each structure element contributes a loss in potential energy and also a loss of entropy (Table 1).

The entropic part is chosen that the formation of a structural element decreases the accessible conformational space by a factor *a* → 2, *b* → 3 and *c* → 5 favouring the unfolded state for high temperatures. A small additive number (0.5) is added to the conformation space volumes in order to break the perfect independence of structure elements. In addition, for each formed structural element the potential energy is lowered so as to favor the folding at low temperatures.

Table 1 Energy model of the simple protein folding model. Shown is the potential energy ΔU and the entropy ΔS depending on the folding state. The potential energy drops with the number of structural elements formed, while the entropic part mimics a reduction of conformational space when one of the elements forms (by a factor of *a* → 2, *b* → 3 and *c* → 5)

	U	S
Unfolded	0	103 804 = log(60 + 120 − 0.5)
a	−1.5	6.76878 = log(30 − 0.5)
b	−1.5	5.94083 = log(20 − 0.5)
c	−1.5	4.88469 = log(12 − 0.5)
a/b	−3.75	4.50258 = log(10 − 0.5)
a/c	−3.75	3.4095 = log(6 − 0.5)
b/c	−3.75	2.50553 = log(4 − 0.5)
$a/b/c$	−4.5	0.81093 = log(2 − 0.5)

Thus, at any given temperature T, the free energy $F_i = U_i - TS_i$ for each of the eight possible foldamers $\{0,a,b,c,ab,ac,bc,abc\}$ can be calculated and also the associated stationary distribution

$$\pi_i = \frac{\exp(-F_i/k_\mathrm{B}T)}{\sum_j \exp(-F_j/k_\mathrm{B}T)}.$$

Assuming furthermore that the model protein can jump between states by forming or breaking one structure element with transition probabilities

$$T_{ij} = \exp\left(-\frac{\Delta + \max(0, F_j - F_i)}{k_\mathrm{B}T}\right)$$

with minimum barrier height $\Delta = 4$, we have a consistent dynamical model that can be used for analysis. Fig. 5 illustrates this model at low, intermediate and high temperatures, showing that the folded state is stable at low temperatures and the unfolded state is stable at high temperatures.

Fig. 6 shows the eigenvectors in the protein folding model. For the low-temperature situation, the folding process is interestingly not the slowest, but the third-slowest process, which exchanges probability between unfolded-*a*–*b*–*c* and states *ab*–*ac*–*bc*–*abc*. The slowest process corresponds to the formation of *a*, while the second-slowest process is a more complex transition involving the exchange of unfolded, *c* and *ac* with the rest.

In the intermediate-temperature situation, the slowest process is the one that most closely resembles folding—it mostly exchanges probability between unfolded-*c* and *ab*–*abc*. The second- and third-slowest processes correspond to the formation of *c* and *b*, respectively.

In the high-temperature situation, the slowest process is a folding process which exchanges probability between unfolded and the rest. It is therefore a different kind of folding process than the third-slowest process in the low-temperature case. One might say that the transition state has shifted towards the unfolded side. The second- and third-slowest processes again correspond to the formation of *c* and *b*, respectively.

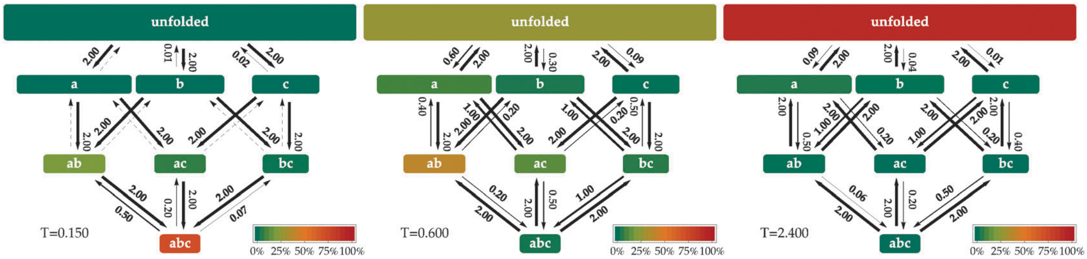

Fig. 5 Illustrative protein folding model for low, intermediate and high temperatures. The colours indicate the stationary probability of states, while the thickness of the arrows and the numbers next to them quantify transition probabilities (within some fixed but arbitrary timescale).

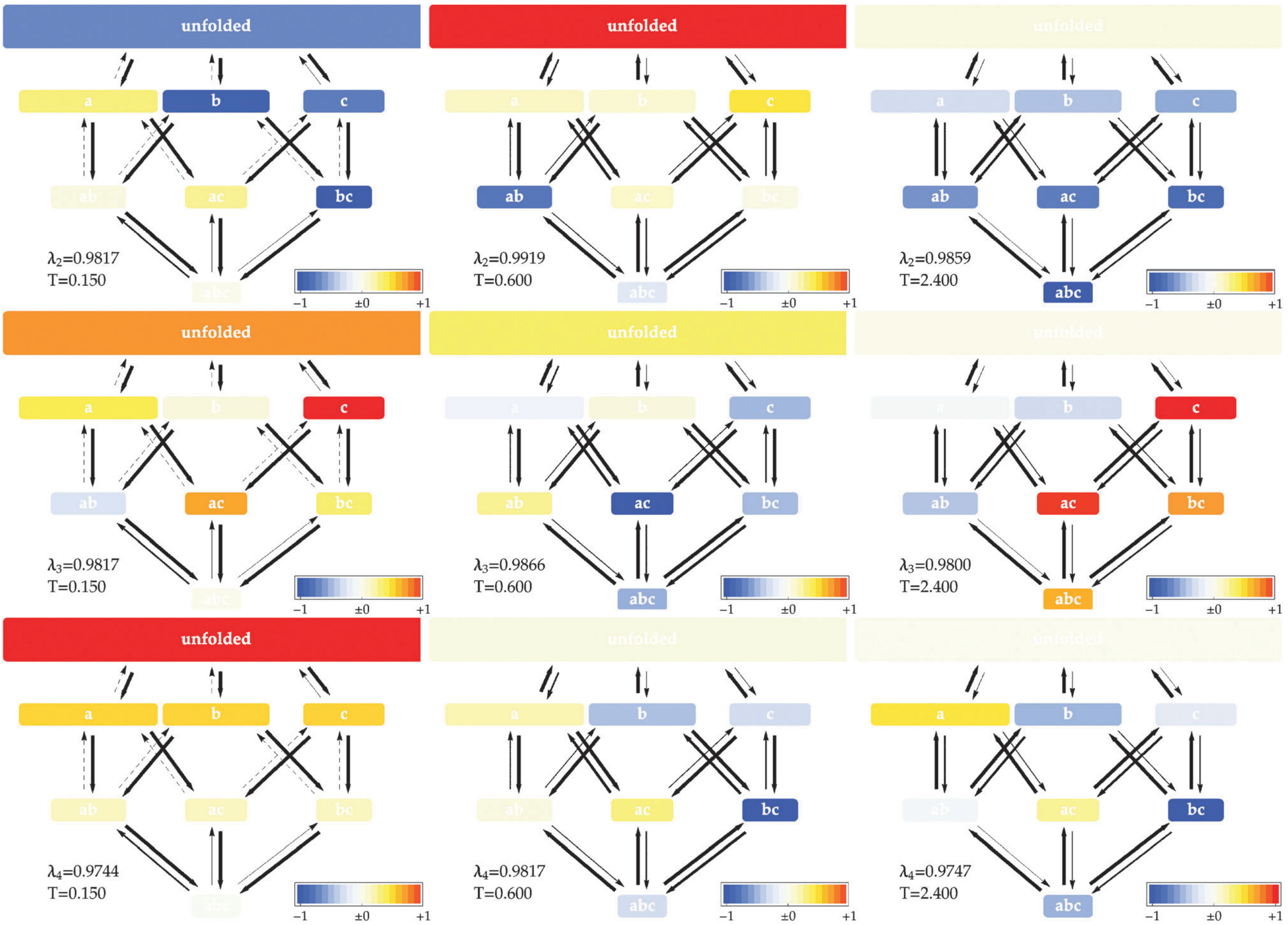

Fig. 6 Dominant eigenvectors and eigenvalues of the protein folding model.

4 Metastable states

The protein folding model used here for illustration consists of only 8 states and is thus easy to comprehend. When building Markov models from clustered molecular dynamics data one often requires several thousands of states in order to approximate the system kinetics well. Network approaches have been developed to visualize the network of transitions arising from such a model,[70] but especially when the network is dense, this is not straightforward. It is thus desirable to find an effective representation that communicates the essential properties of the kinetics. In this section we describe a way to cluster the large discrete state space into a few metastable sets that have the property that they capture the dynamics for long times before jumping to another set. Let us stress that the purpose of finding these sets is purely illustrative (*e.g.* for lumping fluxes, see Section 5). For quantitatively calculating kinetic properties, the full Markov model should be used, as the approximation of the system's kinetics will generally deteriorate when using a lumped Markov model.[45,69,74]

Let us consider the coarse partition of state space $\Omega = \{C_1, C_2, \ldots, C_n\}$ where each cluster C_i consists of a set of states S_j. We are interested in finding a clustering that is maximally metastable. In other words, each cluster C_i should represent a set of structures that the dynamics remains in for a long time before jumping to another cluster C_j. Thus, each cluster C_i can be associated with a free energy basin.

As shown above (see Fig. 1 and Section 2.4), we can understand the slow kinetics in terms of probability transport by the dominant eigenvectors of the transition matrix. Consequently, these dominant eigenvectors can also be used in order to decompose the system into metastable sets.[77,90] Consider the eigenvector corresponding to the slowest process in Fig. 1 (yellow line): this eigenvector is almost a step function which changes from negative to positive values at the saddle point. When we take the value of this eigenvector in each state and plot it along one axis, we obtain Fig. 7a. Partitioning this line in the middle dissects state space into the two most metastable states of the system (Fig. 7b). The two most metastable states exchange at a timescale given by the slowest timescale t_2. If we are interested in differentiating between smaller substates, we may ask for the partition into the three most metastable states. In this case we consider two eigenvectors simultaneously, $\boldsymbol{r}_2$ and $\boldsymbol{r}_3$. Plotting the coordinates in these eigenvalues for each state yields the triangle shown in Fig. 7c whose corners represent the kinetic centers of metastable states. Assigning each state to the nearest corner partitions state space into the three most metastable states (Fig. 7d) that exchange at timescales of $\boldsymbol{r}_3$ or slower. The same partition can be done using three eigenvectors, $\boldsymbol{r}_2$, $\boldsymbol{r}_3$ and $\boldsymbol{r}_4$, yielding four metastable states exchanging at timescales t_4 and slower,

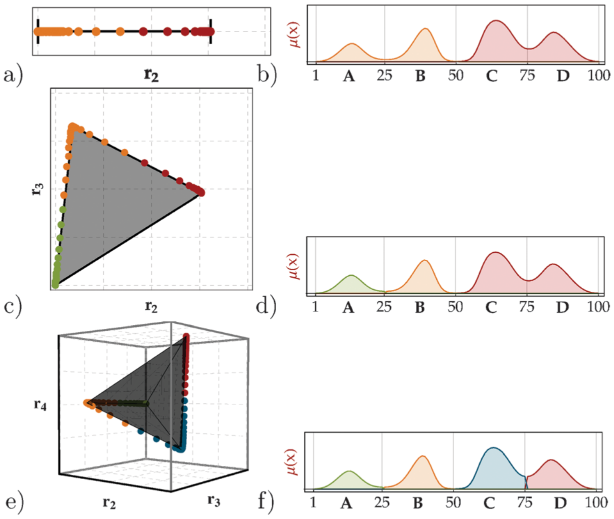

Fig. 7 Metastable states of the one-dimensional dynamics (see Fig. 1a) identified by PCCA+. (a), (c), (e) Plot of the eigenvector elements of one, two, and three eigenvectors. The colors indicate groups of elements (and thus conformational states) that are clustered together. (b), (d), (f) Clustering of conformation space into two, three, and four clusters, respectively.

and so on (Fig. 7e and f). Generally, it can be shown that when n eigenvectors are considered, their coordinates lie in an n-dimensional simplex with $n + 1$ corners called *vertices* which allow the dynamics to be partitioned into $n + 1$ metastable sets.[59,90]

Each of these partitionings is a valid selection in a hierarchy of possible decompositions of the system dynamics. Moving down this hierarchy means that more states are being distinguished, revealing more structural details and smaller timescales. For the system shown in Fig. 1, two to four states are especially interesting to distinguish. After four states there is a gap in the timescales ($t_5 \ll t_4$) induced by a gap after the fourth eigenvalue (Fig. 1c). Thus, for a qualitative understanding of the system kinetics, it is not very interesting to distinguish more than four states. However, note that for quantitatively modeling the system kinetics, it is essential to maintain a fine discretization as the MSM discretization error will increase when states are lumped (see Section 2.3).

Fig. 8 shows the metastable states of the protein folding model. Interestingly, there is no simple partition that splits unfolded and folded states. In the intermediate temperature case this is most closely the case as the unfolded state is a metastable state and separated from all other states with a partial structure. The remaining space and the conformation space at other temperatures are clustered in a non-obvious manner. Sometimes these clusters are defined by the presence of particular structural elements (*e.g.* red cluster in the high-temperature case is characterized by having c formed).

5 Transition pathways

Understanding the folding mechanism of macromolecules, and proteins in particular, is one of the grand challenges in biophysics. The field was driven by questions such as:[21] how does an ensemble of denatured molecules find the same native structure, starting from different conformations? Is folding hierarchical?[4,5] Which forms first: secondary or tertiary structure?[31,94] Does the protein collapse to compact structures before structure formation, or concurrently?[3,36,80] Are there folding nuclei?[39] Is there a particular sequence in which secondary structure elements are formed?

Heterogeneity in folding pathways has been found in a number of experimental studies. For example, using time-resolved FRET with four different intramolecular distances, it was found in barstar[81] that there are multiple folding routes, and that different routes dominate under different folding conditions. Moreover, changing the denaturant can change the dominant pathway.[46] Extensive mutational analysis of the seven ankyrin sequence repeats of the Notch ankyrin repeat domain has revealed its funnel landscape.[10,48,82] Some folding is sequential, as in FynSH3,[43] cytochrome,[26] T4 lysozyme,[13] and Im7,[28] and some folding is parallel, as in cytochrome C[32] and HEW lysozyme.[47]

Formally, the question about folding pathways boils down to the following: let A and B be two subsets of state space, defined so as to specify the transition process one wants to investigate. For example, A may correspond to the strongly denatured set of sets while B is the metastable set around the known crystal structure.[64] All remaining states are unassigned "intermediate" states I. What is the probability distribution of the trajectories leaving A and continuing on to B? *I.e.*, what is the typical sequence of I states used along the transition pathways?

When an MSM is already available, the information of transition pathways is easily accessible *via* transition path theory,[50,64,91] which is explained below. Transition path theory is related to transition path sampling (TPS) in the sense that both are trying to generate statistical information about the ensemble of $A \rightarrow B$ pathways. TPS is a direct approach to sampling pathways directly[8] and could in principle be used to sample folding pathways. However, in TPS the sampled trajectories are in practice of limited length and it is thus

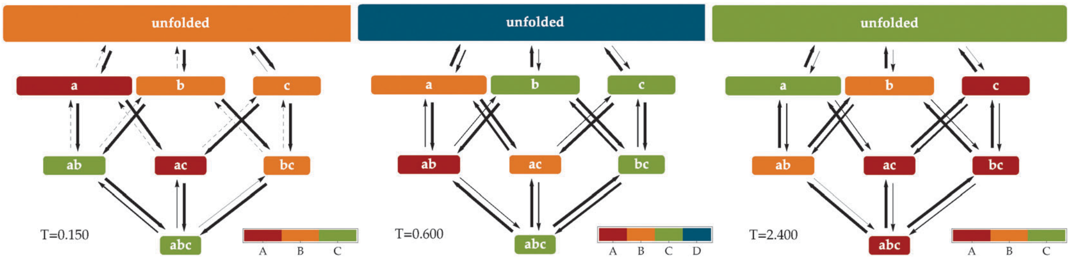

Fig. 8 Metastable sets of the folding model.

unpractical to use TPS when the intermediate states I contain metastabilities. One can run multiple TPS-samplings between pairs of metastable states after having identified them.[88]

5.1 Transition path theory

The essential ingredient required to compute the statistics of transition pathways is the committor probability q_i^+. q_i^+ is the probability when being at state i, the system will reach the set B next rather than A.[8,24,86] In protein folding contexts, it is the probability of folding.[24] By definition, all states in A have $q_i^+ = 0$ while all states in B have $q_i^+ = 1$. For all intermediate states, the committor gradually increases from A to B (see Fig. 9), and its value can be calculated by solving the following system of equations:

$$\sum_{k\in I} T_{ik} q_k^+ = -\sum_{k\in B} T_{ik}$$

(see ref. 64 for derivation). Fig. 9 shows the committor (color-coding) for the protein folding model: at low temperatures, the committor changes rapidly after leaving the unfolded state and forming the first structure elements. At high temperatures, it changes rapidly when entering the full-structured native state. At both temperatures, the folding process has thus essentially two-state character, although with different definitions of the two states. At intermediate temperatures, the committor increases gradually from the unfolded to the native state, indicating that it is important to consider the intermediate states in the folding process.

We further need the backward-committor probability, q_i^-. q_i^- the probability, when being at state i, that the system was in set A previously rather than in B. For dynamics obeying detailed balance (which is assumed here) this is simply

$$q^- = 1 - q^+.$$

Consider the probability flux between two states i and j, given by $\pi_i T_{ij}$ (absolute probability of finding the system at this transition). We are only interested in trajectories that successfully move from A to B without recurring to A beforehand. The flux pertaining to these *reactive* trajectories only is given by multiplying the flux by the probability to come from A and to move on to B:

$$f_{ij} = \pi_i q_i^- T_{ij} q_j^+.$$

This flux is the quantity that could be obtained directly from a converged TPS sampling by counting transitions of the reactive path ensemble. However, we further want to remove contributions that come from recrossings or detours. For example, a trajectory that would jump on its way from A to B multiple times between two substates i and j would produce an increase in the flux $i \rightarrow j$ and the backward flux $j \rightarrow i$. However, we only want to consider a single transition per pathway and thus define the net flux, given by:

$$f_{ij}^+ = \max\{0, f_{ij} - f_{ji}\}.$$

Considering detailed balance dynamics and when ordering states along the reaction coordinate q_i^+ such that $q_i^+ \leq q_j^+$, an equivalent expression is:[6]

$$f_{ij}^+ = \pi_i T_{ij}(q_j^+ - q_i^+).$$

f_{ij}^+ defines the net flux and is a network of fluxes leaving states A and entering states B (see Fig. 9). This network is flux-conserving, *i.e.* for every intermediate state i, the input flux equals the output flux (see ref. 50 and 64 for proof). There the only set in the network that produces flux is A and the only set that consumes flux is B. Due to flux conservation, these amounts of flux are identical and are called total flux F of the transition $A \rightarrow B$:

$$F = \sum_{i\in A}\sum_{j\notin A} \pi_i T_{ij} q_j^+ = \sum_{i\notin B}\sum_{j\in B} \pi_i T_{ij}(1 - q_i^+).$$

The value of F gives the expected number of observed $A \rightarrow B$ transitions per time unit τ that an infinitely long trajectory would produce. Of special interest is the reaction rate constant k_{AB} (see ref. 64 for derivation):

$$k_{AB} = F \Big/ \left(\tau \sum_{i=1}^{m} \pi_i q_i^-\right).$$

Note that all states that trap the trajectory for some time will reduce k_{AB}. The effect of these traps is properly accounted for in the folding flux, even if they do not contribute to productive pathways.

5.2 Transition paths between macrostates

Since the number of n conformational states used to construct a Markov model is often very large, it is convenient for illustration purposes to compute the net flux of $A \rightarrow B$ trajectories amongst only a few coarse sets of conformations. We consider a coarse partition of state space $S = \{C_1, C_2, \ldots, C_n\}$ which may be based on a decomposition into metastable states as described in Section 4, or another partition that the user defines *e.g.* based on order parameters of interest. We make the restriction, however, that this

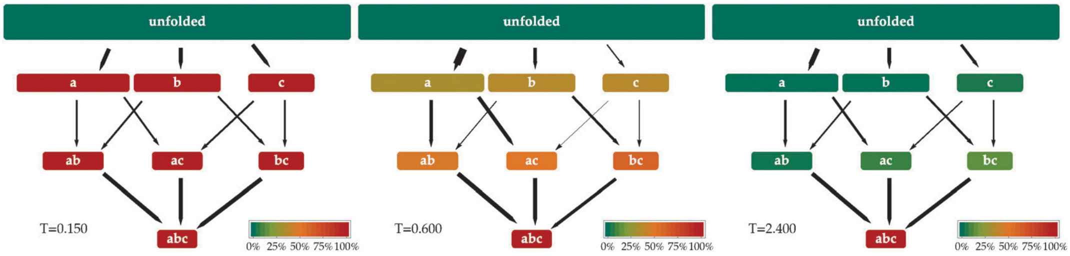

Fig. 9 Committor and net flux from unfolded to folded state.

decomposition preserves the boundaries of sets A, B and I, *i.e.* A and B are either identical to individual C_i, or to a collection of multiple C_i.

The coarse-grained flux between two sets is then given by:

$$F_{ij} = \sum_{k \in C_i, l \in C_j} f_{kl},$$

and the net flux by

$$F_{ij}^{+} = \max\{0, F_{ij} - F_{ji}\}.$$

We note a technicality here: the second step of again removing backfluxes to obtain a coarse-grained net flux is necessary only if the clusters used do not partition state space along the isocommittor surfaces. Thus it may be desirable to use a partition that only groups states with similar committor values.

Fig. 10 shows the coarse-grained flux from the unfolded to the folded states where the coarse-graining has been done according to metastable states. At low and intermediate temperatures, the topology of the folding network is equal, but the flux becomes smaller and the *ab* intermediate is used less. At higher temperatures, the topology of the folding network changes due to a change in the boundaries of metastable states and the unfolded state first splits into three intermediate states before converging to *abc*.

Coarse-graining generates a simplified but correct view on the folding flux. The actual dynamics, represented by the Markov model $\boldsymbol{T}(\tau)$ cannot easily be coarse-grained without loosing information, and no statement is made here about the transition probability between two coarse sets C_i and C_j.

5.3 Pathway decomposition

The flux network can be decomposed into pathways from $A \to B$. When the dynamics are reversible, then the flux can be completely decomposed into such $A \to B$ pathways and no cycles will remain. Consider a pathway consisting of k nodes

$$P = (i_1 \in A \to i_2 \to \cdots \to i_{k-1} \to i_k \in B)$$

Along each of its edges, say $i_l \to i_{l+1}$, the flux network can carry a flux of up to $f_{i_l i_{l+1}}^{+}$. Thus, the capacity or flux of the pathway is given by the minimum of these fluxes:

$$f(P) = \min\{f_{i_l i_{l+1}}^{+} | l = 1 \ldots k\}$$

A pathway decomposition consists of choosing a pathway P_1, and then removing its flux $f(P_1)$ from the flux along all the edges of P_1. This may be repeated until the total flux F has been subtracted and the network is thus free of $A \to B$ pathways. Note that while the flux network is unique, such a decomposition is not unique, because one may choose different strategies to select pathways. Nevertheless pathway decompositions are useful in at least the following aspects:

(1) The strongest pathway, *i.e.* the pathway whose minimum flux $f(P)$ is the largest of all pathways, is of special interest. Especially so, if $f(P)$ is not much smaller than the total flux F.

(2) One reasonable way to perform a pathways decomposition is to first remove the strongest pathway, then remove the strongest pathway of the remaining network, and so on.[51] This decomposition is useful to estimate how many $A \to B$ are necessary to obtain a certain percentage of the flux.[64]

(3) Any pathway decomposition, even a decomposition in which pathways are chosen randomly, gives the same answer when calculating the probability of certain events. Let us consider the probability that, in the protein folding model, one of the three structural elements, a, b, and c, is formed before the other ones in the intermediate-temperature case. The network can, *e.g.* be decomposed into the pathways with corresponding fluxes:

unfolded $\to a \to ab \to abc$	0.000241655
unfolded $\to a \to ac \to abc$	0.000276008
unfolded $\to b \to ab \to abc$	0.000782191
unfolded $\to b \to bc \to abc$	0.000175341
unfolded $\to c \to ac \to abc$	0.000306848
unfolded $\to c \to bc \to abc$	0.000592429

and the probability of forming a, b or c first is given by the flux fraction of pathways where this occurs:

$$\mathbb{P}(a\,\text{first}) = \frac{1}{F}\sum_i f(P_i)\chi_i(a\,\text{first}) = 60.11\%$$

$$\mathbb{P}(b\,\text{first}) = \frac{1}{F}\sum_i f(P_i)\chi_i(b\,\text{first}) = 29.44\%$$

$$\mathbb{P}(c\,\text{first}) = \frac{1}{F}\sum_i f(P_i)\chi_i(c\,\text{first}) = 10.44\%$$

Where χ_i is 1 if $a/b/c$ forms first in pathway P_i, respectively, and 0 otherwise.

The pathway decomposition is usually done on the original flux network. It can also be done on a coarse-grained flux

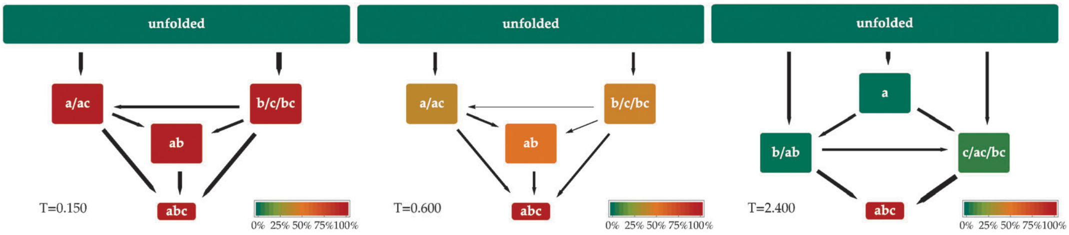

Fig. 10 Coarse-grained folding fluxes.

network, provided that the coarse-graining does not lump states which need to be distinguished in order to calculate the probabilities of the events investigated.

6 Experimental observables/dynamical fingerprints

In experimental studies of protein folding, the conformational dynamics is mapped onto an observable a which is measured. a could be a fluorescence or transfer efficiency in a fluorescence experiment, the chemical shift in an NMR experiment, the intensity of a given spectral peak in an IR experiment, the distance in a pulling experiment, and so forth. In the following we assume that a has a scalar value for every state S_i, *i.e.* there is a mapping $S_i \rightarrow a_i$, where a_i is the mean values of a over the state S_i. We note that vector- or function-valued observables (such as entire spectra in IR or NMR data) could be treated in a similar way, although this is not done here. Given the observable vector, various experimental measurements can be expressed as derived in ref. 41 and 63.

In equilibrium experiments, the observed molecule is in equilibrium with the current conditions of the surroundings (temperature, applied forces, salt concentration *etc.*), and the mean value of an observable a, $\mathbb{E}_\pi[a]$, is recorded. This may be either done by measuring $\mathbb{E}_\pi[a]$ directly from an unperturbed ensemble of molecules, or by recording sufficiently many and long single molecule traces $a(t)$ and averaging over them. The expected measured signal is

$$\mathbb{E}_\pi[a] = \sum_{i=1}^{n} a_i \pi_i = \langle \boldsymbol{a}, \boldsymbol{\pi} \rangle. \tag{22}$$

where $\mathbb{E}[x]$ denotes the expectation value of an observable $x(t)$ and $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ denotes the scalar product between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Since $\boldsymbol{\pi}$ is the eigenvector to eigenvalue 1 of the transition matrix $\boldsymbol{T}(\tau)$, it can easily be calculated from the MSM. $\mathbb{E}_\pi[a]$ does not depend on time and therefore bears no kinetic information.

Kinetic information is available through time-correlation experiments. These may be realized by taking trajectories from time-resolved single molecule experiments, such as single molecule fluorescence or pulling experiments, and computing time correlations from these trajectories. Given a partition into states S_i, the autocorrelation of a for time $k\tau$ can be expressed as:

$$\begin{aligned} \mathbb{E}[a(t)a(t+k\tau)] \ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \mathbb{P}(s_t = S_i) \cdot a_j \\ &\quad \mathbb{P}(s_{t+k\tau} = S_j | s_t = S_i). \end{aligned} \tag{23}$$

The terms under the summation signs contain the product the signal in state i and the signal in state j, $a_i a_j$, where a_i is weighted by the probability of finding the system in state S_i, and a_j is weighted by the conditional probability of finding the system in state j given that it has been in state i at k timesteps τ earlier. In equilibrium, the former probability is given by the equilibrium probability π. Assuming that the process is Markovian, the latter probability is given by the transition matrix element of the corresponding transition matrix. Eqn (23) can be rewritten as a matrix equation in which $\boldsymbol{T}(\tau)$ appears explicitly

$$\begin{aligned} \mathbb{E}[a(t)a(t+k\tau)] &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \pi_i \cdot a_j [\boldsymbol{T}^k(\tau)]_{ij} \\ &= \boldsymbol{a}^\top \Pi \boldsymbol{T}^k(\tau) \boldsymbol{a}. \end{aligned} \tag{24}$$

Replacing $\boldsymbol{T}^k(\tau)$ by its spectral decomposition (eqn (20)), one obtains

$$\begin{aligned} \mathbb{E}[a(t)a(t+k\tau)] \ &= \boldsymbol{a}^\top \left[\sum_{i=1}^{n} \exp\left(-\frac{k\tau}{t_i}\right) \boldsymbol{l}_i \boldsymbol{l}_i^\top \right] \boldsymbol{a} \\ &= \langle \boldsymbol{a}, \pi \rangle^2 + \sum_{i=2}^{n} \exp\left(-\frac{k\tau}{t_i}\right) \langle \boldsymbol{a}, \boldsymbol{l}_i \rangle^2. \end{aligned} \tag{25}$$

Likewise, cross-correlation functions can be computed as

$$\begin{aligned} \mathbb{E}[a(t)b(t+k\tau)] &= \langle \boldsymbol{a}, \pi \rangle \langle \boldsymbol{b}, \pi \rangle \\ &\quad + \sum_{i=2}^{n} \exp\left(-\frac{k\tau}{t_i}\right) \langle \boldsymbol{a}, \boldsymbol{l}_i \rangle \langle \boldsymbol{b}, \boldsymbol{l}_i \rangle. \end{aligned} \tag{26}$$

Eqn (25) and (26) have the form of a multiexponential decay function

$$f(t) = \gamma_1^{\text{corr}} + \sum_{i=2} \gamma_i^{\text{corr}} \exp\left(-\frac{t}{t_i}\right), \tag{27}$$

with amplitudes

$$\gamma_i^{\text{corr}} \ = \ \langle \boldsymbol{a}, \boldsymbol{l}_i \rangle \langle \boldsymbol{b}, \boldsymbol{l}_i \rangle. \tag{28}$$

Each of the amplitudes is associated with an eigenvector of the transition matrix and the decay constant t_i is the implied time scale of this eigenvector, $t_i = -\tau/\ln \lambda_i$.

Alternatively, relaxation experiments can be used to probe the molecules' kinetics. In these experiments, the system is allowed to relax from a nonequilibrium starting state with probability distribution $\boldsymbol{p}(0)$. Examples are temperature-jump, pressure-jump, or pH-jump experiments, rapid mixing experiments, or experiments where measurement at $t = 0$ starts from a synchronized starting state, such as in processes that are started by an external trigger like a photoflash. After time $t = 0$ the conditions are governed by a transition matrix $\boldsymbol{T}(\tau)$ with stationary distribution $\boldsymbol{\pi} \neq \boldsymbol{p}(0)$. The ensemble average $\mathbb{E}_{\boldsymbol{p}(0)}[a(t)]$ is recorded while the system relaxes from the initial distribution $\boldsymbol{p}(0)$ to the new equilibrium distribution π. The expectation value of the signal at time $t = k\tau$ depends on the current probability distribution $\boldsymbol{p}(k\tau)$ and is given by

$$\mathbb{E}_{\boldsymbol{p}(0)}[a(k\tau)] = \sum_{i=1}^{n} a_i p_i(k\tau) = \langle \boldsymbol{a}, \boldsymbol{p}(k\tau) \rangle. \tag{29}$$

Eqn (29) is analogous to eqn (26). $\boldsymbol{p}(k\tau)$ evolves under the influence of the transition matrix $T(\tau)$ (eqn (18)). Using the spectral decomposition of $\boldsymbol{T}(\tau)$ (eqn (20)) and expressing λ_i^k *via* implied timescales t_i, we obtain

$$\begin{aligned} \mathbb{E}_{\boldsymbol{p}(0)}[a(k\tau)] \ &= \langle \boldsymbol{p}'(0), \pi \rangle \langle \boldsymbol{a}, \pi \rangle \\ &\quad + \sum_{i=2}^{n} \exp\left(-\frac{k\tau}{t_i}\right) \langle \boldsymbol{p}'(0), \boldsymbol{l}_i \rangle \langle \boldsymbol{a}, \boldsymbol{l}_i \rangle \end{aligned} \tag{30}$$

Table 2 Overview of the expressions for the amplitudes in correlation experiments

	Equilibrium correlation experiment	Relaxation experiment
Relaxation experiment	—	$\gamma_i^{\text{relax}} = \langle \boldsymbol{a}, \boldsymbol{l}_i \rangle \langle \boldsymbol{p}'^{\top}(0), \boldsymbol{l}_i \rangle$
Autocorrelation	$\gamma_i^{\text{eq,auto-cor}} = \langle \boldsymbol{a}, \boldsymbol{l}_i \rangle^2$	$\gamma_i^{\text{jump,auto-cor}} = \langle \boldsymbol{a}, \boldsymbol{P}'(0), \boldsymbol{l}_i \rangle \langle \boldsymbol{a}, \boldsymbol{l}_i \rangle$
Cross-correlation	$\gamma_i^{\text{eq,cross-cor}} = \langle \boldsymbol{a}, \boldsymbol{l}_i \rangle \langle \boldsymbol{b}, \boldsymbol{l}_i \rangle$	$\gamma_i^{\text{jump, auto-cor}} = \langle \boldsymbol{a}, \boldsymbol{P}'(0), \boldsymbol{l}_i \rangle \langle \boldsymbol{b}, \boldsymbol{l}_i \rangle$

		Obs A	Obs B	Obs C
T-Jump 0.15 → 0.20	γ_2	0.71	0.19	0.13
	γ_3	0.29	0.81	0.87
T-Jump 0.60 → 0.65	γ_2	0.94	0.89	0.17
	γ_3	0.06	0.11	0.83
T-Jump 2.40 → 2.45	γ_2	0.98	0.95	0.89
	γ_3	0.02	0.05	0.11

Fig. 11 Normalized amplitudes of the slowest and second-slowest processes of simulated temperature-jump experiments of the folding model.

where $\boldsymbol{p}'(0)$ is the *excess probability distribution* $\boldsymbol{p}'(0) = \Pi^{-1}\boldsymbol{p}(0)$. $\mathbb{E}_{\boldsymbol{p}(0)}[a(k\tau)]$ is again a multiexponential decay function with amplitudes

$$\gamma_i^{\text{relax}} = \langle \boldsymbol{p}'(0), \boldsymbol{l}_i \rangle \langle \boldsymbol{a}, \boldsymbol{l}_i \rangle. \tag{31}$$

A summary of the amplitudes of various types of experiments is given in Table 2.

These equations are useful to calculate based on simulations which processes a given experiment will be sensitive to. To illustrate this, consider again the protein folding model and let us consider three different observables. In observable A, we measure the formation of structure element a, *i.e.* $a = 1$ for states in which a is formed while $a = 0$ for states in which a is not formed. Likewise observables B and C measure the formation of structure elements b and c. This can be realized *e.g.* with a fluorophor and a specific quencher at appropriate positions.[23] We also consider three ways of measuring each of these three constructs, namely temperature jump experiments at three different temperatures from 0.15 to 0.2, from 0.6 to 0.65, and from 2.4 to 2.45. We calculate the amplitude that is in the slowest and second-slowest processes and report the normalized results in Fig. 11.

It is apparent that the processes that can be measured drastically depend on the way the measurement is done and the observable used. For example, at high temperatures, all observables yield nearly single-exponential kinetics with the timescale of moving between the unfolded state and the partially structured state. At low temperature, the kinetics may appear biexponential, provided that measurement noise is sufficiently small, with the main amplitude being in the formation of $a(\gamma_2)$ and $c(\gamma_3)$.

The combination of Markov models and the spectral theory given is useful to compare simulations and experiments *via* the *dynamical fingerprint* representation of the system kinetics.[63] Furthermore, this approach permits us to design experiments that are optimal to probe individual relaxations.[63]

7 Conclusions and perspectives

The combination of Markov models with analysis methods such as transition path theory and dynamical fingerprinting provides a theoretically solid and computationally feasible approach to obtain deep insights into the microscopic complexity of protein folding and relate molecular simulation data (or protein folding models) to experiments that probe the kinetics of the molecular system in reality.

In contrast to projections on few pre-defined order parameters, a sufficiently fine clustering in the MSM will retain the relevant details of the complex energy landscape, specifically the information which states are kinetically connected and which aren't. This allows relatively detailed analyses such as using transition path theory in order to calculate the ensemble of pathways that lead from the unfolded to the folded state. Based on the resulting path ensemble, mechanistic questions such as "with what probability does structure element a form before the others" can be answered.

With regard to connecting to experiments, the main advantage of the MSM approach over traditional MD analyses is that the processes that occur at given timescales are unambiguously given by the theory. In the Markov model, this assignment is present by the one-to-one association of transition matrix eigenvalues (that correspond to measurable relaxation timescales) and eigenvectors (that describe structural changes). When the experimentally-measured relaxation data are further subjected to a spectral analysis, experiment and simulation can be reconciled on the basis of dynamical fingerprints, *i.e.* by matching peaks of the timescale density. A comment is in order on the fact that in all cases, the slow relaxations in kinetic measurements are found to have the form of a sum of single exponential terms, each term corresponding to an eigenvalue/eigenvector pair in our analysis. This is a general result which can also be obtained by performing the analysis in full continuous state space (as opposed to our discrete-state treatment here). The only assumptions that are made to arrive at this result are the following: (1) the dynamics of the system is Markovian in full continuous state space, (2) the state space is ergodic, *i.e.* all states of the system can interconvert, (3) the relaxations are measured under equilibrium conditions. These assumptions can be assumed to be fulfilled for most protein folding measurements. However, even in such a situation, apparent nonexponentiality has been found over significantly long timescales, such as stretched exponentials[42,49] or power laws.[52] Note that this is not a contradiction because such apparent nonexponentialities can be easily explained by sums of a few single exponential relaxations with particular spacings of timescales and amplitudes[33,63,93]—and thus also correspond to dynamical fingerprints with multiple peaks (see ref. 63).

The methodology introduced here is generally applicable to all dynamical processes that possess a stationary distribution, and especially those which are in equilibrium (*i.e.* fulfill detailed balance). Markovian dynamics and transition path theory have *e.g.* been used to calculate ligand binding pathways.[34] Markov models have been used to characterize different native substates in conformational changes.[53] Applications worthwhile exploring include Physics-based models of matter, such as Ising models. Moreover, chemical processes treated by *ab initio* dynamics are an interesting and challenging field of application, because here direct simulation of sufficiently long trajectories is unfeasible.

References

1 A. Amadei, A. B. Linssen and H. J. C. Berendsen, Essential dynamics of proteins, *Proteins: Struct., Funct., Genet.*, 1993, **17**, 412–425.

2 T. W. Anderson and L. A. Goodman, Statistical Inference about Markov Chains, *Ann. Math. Stat.*, 1957, **28**, 89–110.

3 A. Bachmann and T. Kiefhaber, Apparent two-state tendamistat folding is a sequential process along a defined route, *J. Mol. Biol.*, 2001, **306**(2), 375–386.

4 R. L. Baldwin and G. D. Rose, Is protein folding hierarchic? I. Local structure and peptide folding, *Trends Biochem. Sci.*, 1999, **24**(1), 26–33.

5 R. L. Baldwin and G. D. Rose, Is protein folding hierarchic? II. Folding intermediates and transition states, *Trends Biochem. Sci.*, 1999, **24**(2), 77–83.

6 A. Berezhkovskii, G. Hummer and A. Szabo, Reactive flux and folding pathways in network models of coarse-grained protein dynamics, *J. Chem. Phys.*, 2009, **130**(20), 205102.

7 O. Bieri, J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello and T. Kiefhaber, The speed limit for protein folding measured by triplet–triplet energy transfer, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**(17), 9597–9601.

8 P. G. Bolhuis, D. Chandler, C. Dellago and P. L. Geissler, Transition path sampling: throwing ropes over rough mountain passes, in the dark, *Annu. Rev. Phys. Chem.*, 2002, **53**(1), 291–318.

9 G. R. Bowman and V. S. Pande, Protein folded states are kinetic hubs, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**(24), 10890–10895.

10 C. Bradley and D. Barrick, The Notch Ankyrin Domain Folds *via* a Discrete, Centralized Pathway, *Structure (London)*, 2006, **14**(8), 1303–1312.

11 J. D. Bryngelson and P. G. Wolynes, Spin Glasses and the Statistical Mechanics of Protein Folding, *Proc. Natl. Acad. Sci. U. S. A.*, 1987, **84**, 7524–7528.

12 N. V. Buchete and G. Hummer, Coarse Master Equations for Peptide Folding Dynamics, *J. Phys. Chem. B*, 2008, **112**, 6057–6069.

13 J. Cellitti, R. Bernstein and S. Marqusee, Exploring subdomain cooperativity in T4 lysozyme II: Uncovering the C-terminal subdomain as a hidden intermediate in the kinetic folding pathway, *Protein Sci.*, 2007, **16**(5), 852–862.

14 J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope and J. W. Pitera, Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics, *J. Chem. Phys.*, 2007, **126**, 155101.

15 J. D. Chodera, W. C. Swope, J. W. Pitera and K. A. Dill, Long-time protein folding dynamics from short-time molecular dynamics simulations, *Multiscale Model. Simul.*, 2006, **5**, 1214–1226.

16 J. D. Chodera and F. Noé, Probability distributions of molecular observables computed from Markov models. ii: Uncertainties in observables and their time-evolution, *J. Chem. Phys.*, 2010, **133**, 105102.

17 B. de Groot, X. Daura, A. Mark and H. Grubmüller, Essential Dynamics of Reversible Peptide Folding: Memory-free Conformational Dynamics Governed by Internal Hydrogen Bonds, *J. Mol. Biol.*, 2001, **301**, 299–313.

18 P. Deuflhard and M. Weber, Robust Perron cluster analysis in conformation dynamics, *ZIB Report*, 03–09, 2003.

19 K. A. Dill, Polymer principles and protein folding, *Protein Sci.*, 1999, **8**(6), 1166–1180.

20 K. A. Dill and H. S. Chan, From Levinthal to pathways to funnels, *Nat. Struct. Biol.*, 1997, **4**(1), 10–19.

21 K. A. Dill, S. Banu Ozkan, M. Scott Shell and T. R. Weikl, The Protein Folding Problem, *Annu. Rev. Biophys.*, 2008, **37**(1), 289–316.

22 N. Djurdjevac, M. Sarich and C. Schütte, Estimating theeigenvalue error of Markov State Models, *Multiscale Model. Simul.*, 2010, submitted.

23 S. Doose, H. Neuweiler and M. Sauer, Fluorescence Quenching by Photoinduced Electron Transfer: A Reporter for Conformational Dynamics of Macromolecules, *ChemPhysChem*, 2009, **10**(9–10), 1389–1398.

24 R. Du, V. S. Pande, A. Yu, T. Tanaka and E. S. Shakhnovich, On the transition coordinate for protein folding, *J. Chem. Phys.*, 1998, **108**(1), 334–350.

25 E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay and D. Kern, Intrinsic dynamics of an enzyme underlies catalysis, *Nature*, 2005, **438**(7064), 117–121.

26 H. Feng, Z. Zhou and Y. Bai, A protein folding pathway with multiple folding intermediates at atomic resolution, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**(14), 5026–5031.

27 H. Frauenfelder, S. G. Sligar and P. G. Wolynes, The energy landscapes and motions of proteins, *Science*, 1991, **254**, 1598–1603.

28 C. T. Friel, G. S. Beddard and S. E. Radford, Switching two-state to three-state kinetics in the helical protein Im9 *via* the optimisation of stabilising non-native interactions by design, *J. Mol. Biol.*, 2004, **342**, 261–273.

29 A. Gansen, A. Valeri, F. Hauger, S. Felekyan, S. Kalinin, K. Táth, J. Langowski and C. A. M. Seidel, Nucleosome disassembly intermediates characterized by single-molecule FRET, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**(36), 15308–15313.

30 J. C. Gebhardt, T. Bornschlögl and M. Rief, Full distance-resolved folding energy landscape of one single protein molecule, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**(5), 2013–2018.

31 R. Gilmanshin, S. Williams, R. H. Callender, W. H. Woodruff and R. B. Dyer, Fast events in protein folding: Relaxation dynamics of secondary and tertiary structure in native apomyoglobin, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 3709–3713.

32 R. A. Goldbeck, Y. G. Thomas, E. Chen, R. M. Esquerra and D. S. Kliger, Multiple pathways on a protein-folding energy landscape: kinetic evidence, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**(6), 2782–2787.

33 S. J. Hagen and W. A. Eaton, Nonexponential structural relaxations in proteins, *J. Chem. Phys.*, 1996, **104**(9), 3395–3398.

34 M. Held, P. Metzner and F. Noé, Mechanisms of protein–ligand association and its modulation by protein mutations, *Biophys. J.*, 2011, **100**, 701–710.

35 N. S. Hinrichs and V. S. Pande, Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics, *J. Chem. Phys.*, 2007, **126**, 244101.

36 L. Hoang, S. Bédard, M. M. G. Krishna, Y. Lin and S. Walter Englander, Cytochrome c folding pathway: Kinetic native-state hydrogen exchange, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**(19), 12173–12178.

37 I. A. Hubner, E. J. Deeds and E. I. Shakhnovich, Understanding ensemble protein folding at atomic detail, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**(47), 17747–17752.

38 M. Jäger, H. Nguyen, J. C. Crane, J. W. Kelly and M. Gruebele, The folding mechanism of a beta-sheet: the WW domain, *J. Mol. Biol.*, 2001, **311**(2), 373–393.

39 D. H. Jane, P. E. E. Wright and H. A. A. Scheraga, The role of hydrophobic interactions in initiation and propagation of protein folding, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**(35), 13057–13061.

40 M. E. Karpen, D. J. Tobias and C. L. Brooks, Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV, *Biochemistry*, 1993, **32**(2), 412–420.

41 B. Keller, J.-H. Prinz and F. Noé, Markov models and dynamical fingerprints: unraveling the complexity of molecular kinetics, *Chem. Phys.*, 2011.

42 J. Klafter and M. F. Shlesinger, On the relationship among three theories of relaxation in disordered systems, *Proc. Natl. Acad. Sci. U. S. A.*, 1986, **83**(4), 848–851.

43 D. M. Korzhnev, X. Salvatella, M. Vendruscolo, A. A. Di Nardo, A. R. Davidson, C. M. Dobson and L. E. Kay, Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR, *Nature*, 2004, **430**(6999), 586–590.

44 S. V. Krivov and M. Karplus, Hidden complexity of free energy surfaces for peptide (protein) folding, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 14766–14770.

45 S. Kube and M. Weber, A coarse graining method for the identification of transition rates between molecular conformations, *J. Chem. Phys.*, 2007, **126**(2), 024103–+.

46 M. O. Lindberg and M. Oliveberg, Malleability of protein folding pathways: a simple reason for complex behaviour, *Curr. Opin. Struct. Biol.*, 2007, **17**(1), 21–29.

47 A. Matagne, S. E. Radford and C. M. Dobson, Fast and slow tracks in lysozyme folding: insight into the role of domains in the folding process, *J. Mol. Biol.*, 1997, **267**(5), 1068–1074.

48 C. C. Mello and D. Barrick, An experimentally determined protein folding energy landscape, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**(39), 14102–14107.

49 R. Metzler, J. Klafter, J. Jortner and M. Volk, Multiple time scales for dispersive kinetics in early events of peptide folding, *Chem. Phys. Lett.*, 1998, **293**, 477–484.

50 P. Metzner, C. Schütte and E. Vanden Eijnden, Transition Path Theory for Markov Jump Processes, *Multiscale Model. Simul.*, 2009, **7**, 1192–1219.

51 P. Metzner, C. Schütte and E. Vanden-Eijnden, Illustration of transition path theory on a collection of simple examples, *J. Chem. Phys.*, 2006, **125**(8), 084110.

52 W. Min, G. Luo, B. J. Cherayil, S. C. Kou and X. S. Xie, Observation of a Power-Law Memory Kernel for Fluctuations within a Single Protein Molecule, *Phys. Rev. Lett.*, 2005, **94**, 198302–+.

53 F. Morcos, S. Chatterjee, C. L. McClendon, P. R. Brenner, R. López-Rendón, J. Zintsmaster, M. Ercsey-Ravasz, C. R. Sweet, M. P. Jacobson, J. W. Peng and J. A. Izaguirre, Modeling Conformational Ensembles of Slow Functional Motions in Pin1-WW, *PLoS Comput. Biol.*, 2010, **6**(12), e1001015–+.

54 S. Muff and A. Caflisch, Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of α-sheet mini-protein, *Proteins: Struct., Funct., Bioinf.*, 2007, **70**, 1185–1195.

55 H. Neubauer, N. Gaiko, S. Berger, J. Schaffer, C. Eggeling, J. Tuma, L. Verdier, C. A. Seidel, C. Griesinger and A. Volkmer, Orientational and dynamical heterogeneity of rhodamine 6G terminally attached to a DNA helix revealed by NMR and single-molecule fluorescence spectroscopy, *J. Am. Chem. Soc.*, 2007, **129**(42), 12746–12755.

56 H. Neuweiler, M. Löllmann, S. Doose and M. Sauer, Dynamics of Unfolded Polypeptide Chains in Crowded Environment Studied by Fluorescence Correlation Spectroscopy, *J. Mol. Biol.*, 2007, **365**, 856–869.

57 H. Nguyen, M. Jäger, A. Moretto, M. Gruebele and J. W. Kelly, Tuning the Free-Energy Landscape of a WW Domainby Temperature, Mutation, and Truncation, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**(7), 3948–3953.

58 F. Noé and S. Fischer, Transition networks for modeling the kinetics of conformational transitions in macromolecules, *Curr. Opin. Struct. Biol.*, 2008, **18**, 154–162.

59 F. Noé, I. Horenko, C. Schütte and J. C. Smith, Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States, *J. Chem. Phys.*, 2007, **126**, 155102.

60 F. Noé, D. Krachtus, J. C. Smith and S. Fischer, Transition Networks for the Comprehensive Characterization of Complex Conformational Change in Proteins, *J. Chem. Theory Comput.*, 2006, **2**, 840–857.

61 F. Noé, M. Oswald, G. Reinelt, S. Fischer and J. C. Smith, Computing Best Transition Pathways in High-Dimensional Dynamical Systems: Application to the alpha$_L$–beta–alpha$_R$ Transitions in Octaalanine, *Multiscale Model. Simul.*, 2006, **5**, 393–419.

62 F. Noé, Probability Distributions of Molecular Observables computed from Markov Models, *J. Chem. Phys.*, 2008, **128**, 244103.

63 F. Noé, S. Doose, I. Daidone, M. Löllmann, J. D. ChoderaJ, M. Sauer and J. C. Smith, Dynamical fingerprints: Understanding biomolecular processes in microscopic detail by combination of spectroscopy, simulation and theory, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 4822–4827.

64 F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich and T. R. Weikl, Constructing the full ensemble of folding pathways from short off-equilibrium simulations, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 19011–19016.

65 J. N. Onuchic, P. G. Wolynes, L. Z. Schulten and N. D. Socci, Toward an outline of the topography of a realistic protein-folding funnel, *Proc. Natl. Acad. Sci. U. S. A.*, 1995, **92**, 3626–3630.

66 A. C. Pan and B. Roux, Building Markov state models along pathways to determine free energies and rates of transitions, *J. Chem. Phys.*, 2008, **129**(6), 064107–+.

67 S. Park and V. S. Pande, Validation of Markov state models using Shannon's entropy, *J. Chem. Phys.*, 2006, **124**, 054118.

68 J.-H. Prinz, M. Held, J. C. Smith and F. Noé, Efficient computation of committor probabilities and transition state ensembles, *SIAM Multiscale Model. Simul. 9, 545*, 2010.

69 J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Fischbach, M. Held, J. D. Chodera, C. Schütte and F. Noé, Markov models of molecular kinetics: Generation and validation, *J. Chem. Phys.*, 2011, **134**, 174105.

70 F. Rao and A. Caflisch, The Protein Folding Network, *J. Mol. Biol.*, 2004, **342**, 299–306.

71 A. Reiner, P. Henklein and T. Kiefhaber, An unlocking/relocking barrier in conformational fluctuations of villin headpiece subdomain, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **107**, 4955–4960.

72 S. Röblitz, *Statistical Error Estimation and Grid-free Hierarchical Refinement in Conformation Dynamics*, *PhD thesis*, 2009.

73 Y. Santoso, C. M. Joyce, O. Potapova, L. Le Reste, J. Hohlbein, J. P. Torella, N. D. F. Grindley and A. N. Kapanidis, Conformational transitions in DNA polymerase I revealed by single-molecule FRET, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**(2), 715–720.

74 M. Sarich, F. Noé and C. Schütte, On the approximation error of Markov state models, *Multiscale Model. Simul.*, 2010, **8**, 1154–1177.

75 D. D. Schaeffer, A. Fersht and V. Daggett, Combining experiment and simulation in protein folding: closing the gap for small model systems, *Curr. Opin. Struct. Biol.*, 2008, **18**(1), 4–9.

76 V. Schultheis, T. Hirschberger, H. Carstens and P. Tavan, Extracting Markov Models of Peptide Conformational Dynamics from Simulation Data, *J. Chem. Theory Comput.*, 2005, **1**, 515–526.

77 C. Schütte, A. Fischer, W. Huisinga and P. Deuflhard, A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo, *J. Comput. Phys.*, 1999, **151**, 146–168.

78 D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan and W. Wriggers, Atomic-Level Characterization of the Structural Dynamics of Proteins, *Science*, 2010, **330**(6002), 341–346.

79 N. Singhal, C. Snow and V. S. Pande, Path sampling to build better roadmaps: predicting the folding rate and mechanism of a Trp Zipper beta hairpin, *J. Chem. Phys.*, 2004, **121**, 415–425.

80 K. Sridevi, The slow folding reaction of barstar: the core tryptophan region attains tight packing before substantial secondary and tertiary structure formation and final compaction of the polypeptide chain, *J. Mol. Biol.*, 2000, **302**(2), 479–495.

81 K. Sridevi, G. S. Lakshmikanth, G. Krishnamoorthy and J. B. Udgaonkar, Increasing stability reduces conformational heterogeneity in a protein folding intermediate ensemble, *J. Mol. Biol.*, 2004, **337**(3), 699–711.

82 T. O. Street, C. M. Bradley and D. Barrick, Predicting coupling limits from an experimentally determined energy landscape, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**(12), 4907–4912.

83 W. C. Swope, J. W. Pitera and F. Suits, Describing protein folding kinetics by molecular dynamics simulations: 1. Theory, *J. Phys. Chem. B*, 2004, **108**, 6571–6581.

84 W. C. Swope, J. W. Pitera, F. Suits, M. Pitman and M. Eleftheriou, Describing protein folding kinetics by molecular dynamics simulations: 2. Example applications to alanine dipeptide and beta-hairpin peptide, *J. Phys. Chem. B*, 2004, **108**, 6582–6594.

85 W. van Gunsteren, J. Dolenc and A. Mark, Molecular simulation as an aid to experimentalists, *Curr. Opin. Struct. Biol.*, 2008, **18**(2), 149–153.

86 E. Vanden-Eijnden, Transition Path Theory, in *Computer Simulations in Condensed Matter: From Materials to Chemical Biology, Vol. 2, Lecture Notes in Phys. 703*, ed. M. Ferrario, G. Ciccotti, and K. Binder, Springer-Verlag, Berlin, 2006, 439–478.

87 V. A. Voelz, G. R. Bowman, K. Beauchamp and V. S. Pande, Molecular Simulation of *ab initio* Protein Folding for a Millisecond Folder NTL9, *J. Am. Chem. Soc.*, 2010, **132**(5), 1526–1528.

88 J. Vreede, J. Juraszek and P. G. Bolhuis, Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 2397–2402.

89 D. J. Wales, *Energy Landscapes*, Cambridge University Press, Cambridge, 2003.

90 M. Weber, Improved Perron cluster analysis, *ZIB Report*, 03–04, 2003.

91 E. Weinan and E. vanden-Eijnden, Towards a Theory of Transition Paths, *J. Stat. Phys.*, 2006, **123**(3), 503–523.

92 B. G. Wensley, S. Batey, F. A. C. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia and J. Clarke, Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family, *Nature*, 2010, **463**(7281), 685–688.

93 J. B. Witkoskie and J. Cao, Single molecule kinetics. I. Theoretical analysis of indicators, *J. Chem. Phys.*, 2004, **121**(13), 6361–6372.

94 S. R. Yeh and D. L. Rousseau, Hierarchical folding of cytochrome c, *Nat. Struct. Biol.*, 2000, **7**(6), 443–445.