cluster incorporates also the cost for right location of the substrate.

These theoretical results are quite consistent with the experimental indications about the reduced catalytic activity of the metal-depleted enzyme (Scolnick et al. 1997).

Thus, arginase enzyme presents a higher catalytic activity in alkaline medium. An intact binuclear manganese cluster and a ligand field are required for optimal catalysis. In fact, MnA cation favors the correct binding and orientation of the substrate, as the higher activation energy obtained for the MnA-depleted cluster confirms.

Cross-References

▶ Quantum Mechanical Simulations of Biopolymer Vibrational Spectra

References

Bewley MC, Jeffrey PD, Patchett ML, Kanyo ZF, Baker EN. Crystal structures of *Bacillus caldovelox* arginase in complex with substrate and inhibitors reveal new insights into activation, inhibition and catalysis in the arginase superfamily. Structure. 1999;7:435–48.

Cavalli RC, Burke CJ, Kawamoto S, Robert Soprano D, Ash DE. Mutagenesis of rat liver arginase expressed in *Escherichia coli*: role of conserved histidines. Biochemistry. 1994;33:10652–7.

Hellerman LA, Perkins ME. Activation of enzymes III. The role of metal ions in the activation of arginase. The hydrolysis of arginine induced by certain metal ions with urease. J Biol Chem. 1935;112:175–94.

Ivanov I, Klein ML. First principles computational study of the active site of arginase. Proteins. 2004;54:1–7.

Khangulov SV, Sossong Jr TM, Ash DE, Dismukes GC. L-arginine binding to liver arginase requires proton transfer to gateway residue His141 and coordination of the guanidinium group to the dimanganese(II, II) center. Biochemistry. 1998;37:8539–50. and references there in.

Krebs HA, Henseleit K. Studies on urea formation in the animal organism. Hoppe-Seyler's Z Physiol Chem. 1932;210:33–66.

Kuhn NJ, Talbot J, Ward S. pH-sensitive control of arginase Mn (II) ions at submicromolar concentrations. Arch Biochem Biophys. 1991;286:217–21.

Leopoldini M, Russo N, Toscano M. The determination of the catalytic pathway of manganese arginase enzyme throughout density functional investigation. Chem Eur J. 2009;15:8026–36.

Scolnick LR, Kanyo ZF, Cavalli RC, Ash DE, Christianson DW. Altering the binuclear manganese cluster of arginase diminishes thermostability and catalytic function. Biochemistry. 1997;36:10558–65.

Xie K, Fidler IJ. Therapy of cancer metastasis by activation of the inducible nitric oxide synthase. Cancer Metastasis Rev. 1998;17:55–75.

Marcus Theory

▶ Electron Transfer Theory

Markov Model

▶ Hidden Markov Modeling in Single-Molecule Biophysics

Markov Models of Molecular Kinetics

Frank Noé
Institute of Mathematics, FU Berlin, Berlin, Germany

Introduction

Markov (state) models (MSMs) are an approach to understand conformational dynamics of molecules using computer simulations. An MSM consists of (1) a subdivision of the state space into a discrete set of *microstates*, often using some clustering method, and (2) a Markovian model to describe the transition dynamics amongst these microstates, usually a transition probability matrix or rate matrix.

MSMs are especially useful when studying complex macromolecular changes, such as folding, native-state transitions, and binding. Such systems are often metastable, that is, the protein(s) 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 (Frauenfelder et al. 1991; de Groot et al. 2001). 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.

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. MSMs are a systematic framework for analyzing and also for driving molecular dynamics simulations. Compared to standard analyses of molecular dynamics simulations, MSMs have a number of useful features:

1. Long-term molecular kinetics may be predicted from short-time simulations.
2. Great amounts of simulation data can be analyzed with relatively little subjectivity of the analyst.
3. Stationary and kinetic quantities can be calculated, such as conformational free energy differences, metastable states, and the ensemble of transition pathways.
4. Simulation data and measurement data can be reconciled in a rigorous and explicit way.
5. Statistical information contained in MSMs can be used to allocate new simulations adaptively.

Due to the advent of large-scale distributed computing frameworks and the recent performance increase of computer clusters, large numbers of short trajectories are becoming more and more easy to generate (Voelz et al. 2010; Noé et al. 2009; Shaw et al. 2010). MSMs and other methods based on trajectory ensembles are thus increasingly useful and important in the process of investigating conformational dynamics with simulations. There are currently two relatively complete software packages for building and analyzing MSMs: MSMbuilder (Beauchamp et al. 2011) and EMMA (Senne et al. 2012).

Markov Model Theory

Basics

The dynamics of the molecular system considered is given by trajectories of a stochastic process $\mathbf{x}(t)$ in the continuous state space consisting of positions and momenta. The stochasticity of $\mathbf{x}(t)$ comes through coupling the system to a thermostat. Following properties are assumed for $\mathbf{x}(t)$: (1) $\mathbf{x}(t)$ is Markovian in full state space, (2) $\mathbf{x}(t)$ is ergodic and states are visited with a frequency given by the Boltzmann distribution $\mu(x) = Z(\beta)^{-1} \exp(-\beta H(\mathbf{x}))$ (3) the dynamics are in thermal equilibrium and thus $\mathbf{x}(t)$ fulfills microscopic detailed balance. See Prinz et al. (2011) for a more extensive description. These conditions are fulfilled by not all, but many dynamical models frequently used to simulate molecular dynamics. Even for setups violating these conditions, MSMs are often useful, although they are then not justified by a solid theory.

Let the state space with coordinates $\mathbf{x}$ be discretized into "microstates" $\{S_1,\ldots,S_n\}$. $T_{ij}(\tau)$ 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:

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

defining a transition matrix $\mathbf{T}(\tau) \in \mathbb{R}^{n \times n}$. Note that τ can be orders of magnitude shorter than the longest timescales of the system. The transition matrix can also be written in terms of correlation functions (Swope et al. 2004):

$$T_{ij}(\tau) = \frac{c_{ij}^{\mathrm{corr}}(\tau)}{\pi_i}, \tag{1}$$

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

$$\pi_i = \mathbb{P}\left[\mathbf{x}(t) \in S_i\right]$$

and $c_{ij}^{\mathrm{corr}}(\tau) = \pi_i T_{ij}(\tau)$ is an unconditional transition probability. Suppose that $\mathbf{p}(t) \in \mathbb{R}^n$ is a column vector whose elements denote the probability to be within a set $j \in \{1, \ldots, n\}$ at time t. After time τ, the probabilities will have changed according to:

$$\mathbf{p}^T(t+\tau) = \mathbf{p}^T(t)\mathbf{T}(\tau) \tag{2}$$

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

$$\pi^T = \pi^T \mathrm{T}(\tau) \tag{3}$$

Clustering, Estimation and Statistics

The transition probabilities T_{ij} are usually estimated from molecular dynamics simulations. Suppose a trajectory $\mathbf{x}(t)$ is given. The simulation data is first discretized onto a microstate discretization $(S_1,\ldots, S_n)$. This is usually done with clustering methods such as density-based clustering, k-medoids or k-means, using RMSD, Euclidean positions, or internal coordinates as metric (Voelz et al. 2010; Chodera et al. 2007; Prinz et al. 2011).

The count matrix $\mathbf{C}(\tau)$ is then defined by counting the number of transitions between discrete sets along the trajectory:

$$c_{ij}(\tau) = |\{\mathbf{x}(t) \in S_i,\ \mathbf{x}(t+\tau) \in S_j, \}|. \quad (4)$$

If multiple trajectories are available, then the count matrices of these trajectories are simply added up. Based on $\mathbf{C}(\tau)$, the transition matrix can be estimated with maximum likelihood by Prinz et al. (2011):

$$\hat{T}_{ij} = \frac{c_{ij}}{\Sigma_{k=1}^{n} c_{ik}}, \quad (5)$$

Provided that the trajectories $\mathbf{x}(t)$ are started from a local equilibrium within the set S_i that contains the starting structure $\mathbf{x}(0)$ (Prinz et al. 2011), the estimator Eq. 5 is asymptotically unbiased, that is, for a long enough trajectory, $\hat{\mathbf{T}}(\tau)$ will converge to the correct transition matrix $\mathbf{T}(\tau)$. It is important to note that $\hat{T}_{ij}$ as given by Eq. 5 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 is usually accounted for by using a maximum likelihood estimator that makes sure that the detailed balance equations are fulfilled (Prinz et al. 2011).

Since simulation data is finite, all validation procedures (either consistency checks or comparisons to experimental data) need to account for statistical uncertainties. Standard deviations or confidence intervals of the transition matrix elements and of properties computed from the transition matrix can be calculated from the count matrix $\mathbf{C}(\tau)$. See Singhal and Pande (2005), Noé (2008) for details (Fig. 1).

Predicting Long-Term Kinetics from Short Simulations and the Systematic Error Caused by This

Markov models are an approximation of molecular kinetics. The discretization of state space into sets (S_1,..., S_n) erases the information where exactly the continuous process $\mathbf{x}(t)$ was. As a result, the jump process on (S_1,..., S_n) is no longer Markovian even if $\mathbf{x}(t)$ was; nevertheless, it is approximated by a Markov chain. What are the consequences of this approximation? The following two quantities are obtained from Markov models *without* systematic error:

1. The propagation of transition probabilities by one step τ, $\mathbf{p}^T(t+\tau) = \mathbf{p}^T(t)\mathbf{T}(\tau)$.
2. Stationary properties, such as the stationary distribution $\boldsymbol{\pi}$ and associated expectation of state functions $\mathbb{E}_\pi(a) = \langle \pi, \mathbf{a} \rangle$.

However, state space discretization introduces systematic error in the reproduction of long-time kinetics, that is, the prediction:

$$\mathbf{p}^T(t+k\tau) \approx \mathbf{p}^T(t)\mathrm{T}^k(\tau), \quad (6)$$

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 (Sarich et al. 2010; Prinz et al. 2011), it is now known that the error of Eq. 6 decreases toward zero with increasingly fine discretization and increasingly long lagtime τ.

A practical way to test the quality of a specific state space discretization is the Chapman-Kolmogorow Test (Prinz et al. 2011), which tests the approximate validity of Eq. 6. Figure 2 shows the results of such a test for a two- and a six-state partition of a diffusion in a double well. The six-state partition clearly outperforms the two-state partition.

Eigenvalues and Eigenvectors

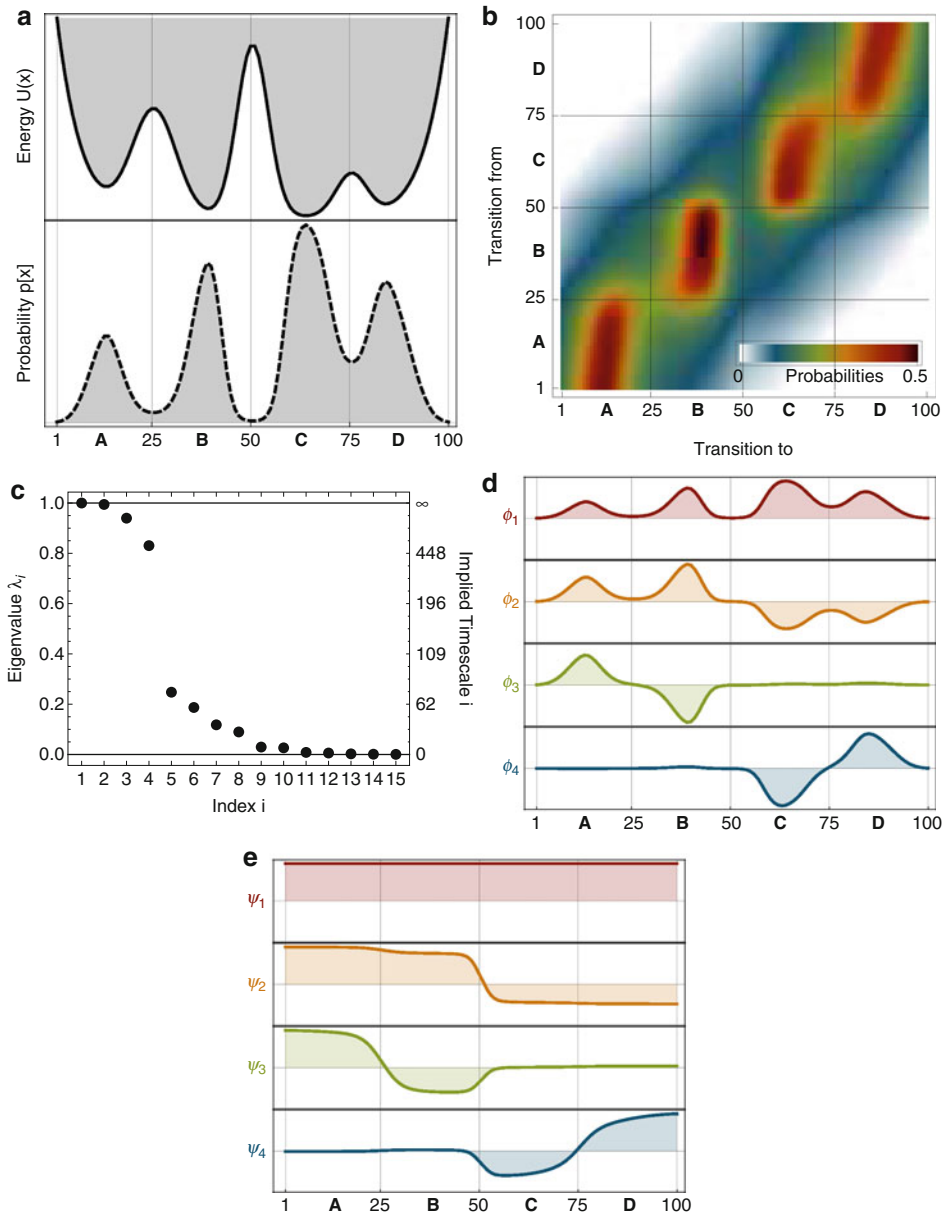
The transition matrix $\mathbf{T}(\tau)$ can be written as a linear combination of its eigenvalues λ_i and left eigenvectors l_i:

$$\mathbf{p}^T(k\tau) = \pi^T + \sum_{i=2}^{n} \lambda_i^k(\tau)\alpha_i \mathbf{l}_i^{\mathrm{T}}. \quad (7)$$

with coefficients α_i that depend on the initial distribution $\mathbf{p}(0)$. The first eigenvector is equal to the stationary distribution $\mathbf{l}_1 = \pi$ and has the eigenvalue $\lambda_1 = 1$. All other eigenvalues are smaller than one, hence $\lim_{k\tau\to\infty} \mathbf{p}^T(k\tau) = \boldsymbol{\pi}$. The terms with $i \geq 2$ indicate exponential relaxation processes with a timescale implied by the eigenvalues:

$$t_i = -\frac{\tau}{\ln \lambda_i} \quad (8)$$

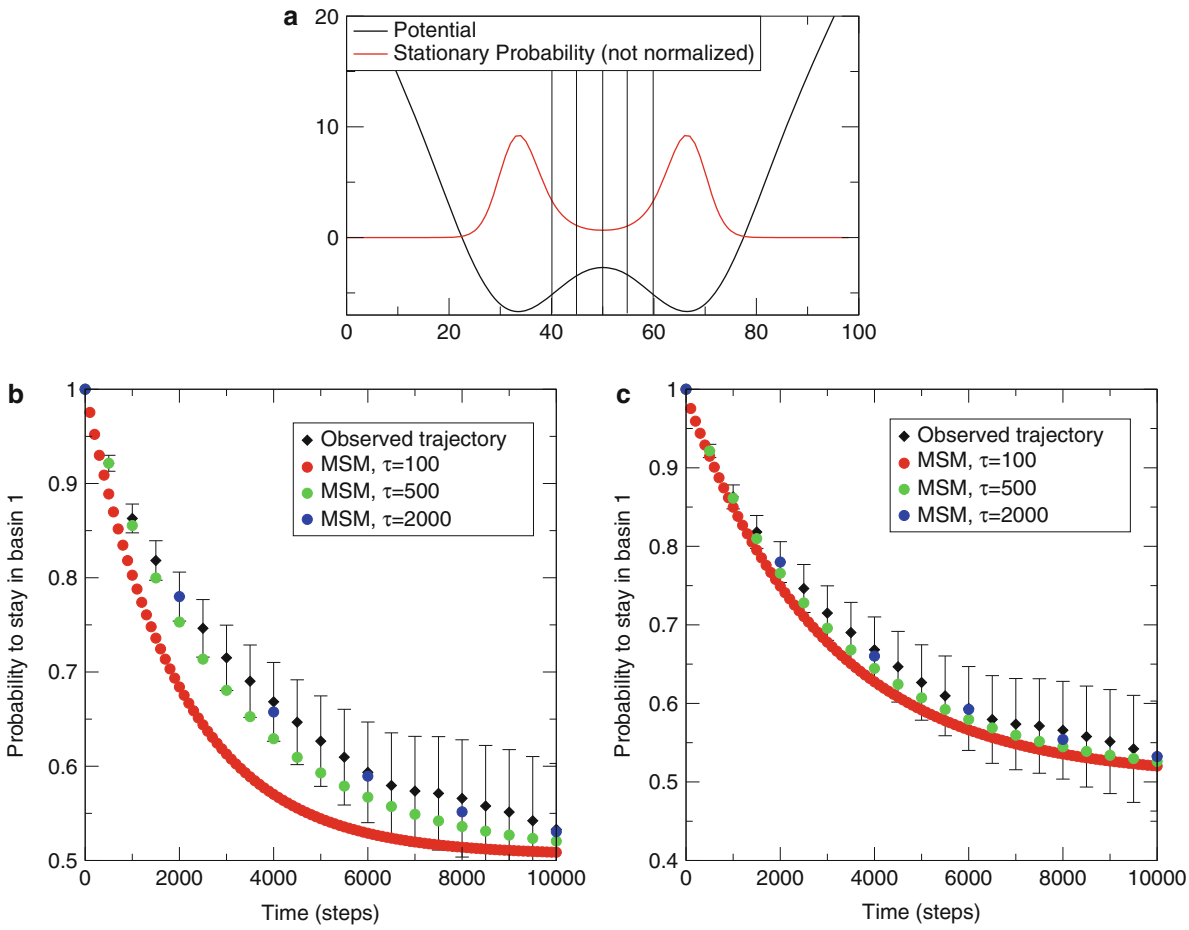
Since the relaxation timescales t_i are physical properties of the dynamics, they should be invariant under change of the lag time τ used to parametrize the transition matrix (Swope et al. 2004). For large enough τ, t_i should converge to their true value



Markov Models of Molecular Kinetics, Fig. 1 (**a**) Potential energy function with 100 microstates and four metastable sets and corresponding stationary probabilities π_i. (**b**) Density plot of the transition matrix for a simple diffusion in the potential. The matrix is nearly block-diagonal, with the transition probability being large within blocks allowing rapid transitions within metastable basins, and small or nearly zero for transitions between different metastable basins. (**c**) Eigenvalues of the transition matrix. The gap between the four slow processes ($\lambda_i \approx 1$) and the fast processes is clearly visible. (**d**) The four dominant eigenvectors, $\mathbf{r}_1, \ldots \mathbf{r}_4$, which indicate the associated dynamical processes. The first eigenvector 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 left eigenvectors $\mathbf{l}_1$, $\ldots$, $\mathbf{l}_4$ (Figure adapted from Prinz et al. (2011))

(assuming sufficient statistics). Therefore, the convergence of t_i with increasing τ has often been employed as an indicator for selecting τ (Swope et al. 2004; Chodera et al. 2007; Prinz et al. 2011) (see Fig. 3).

The relevance of the eigenvectors is illustrated in Fig. 1d, showing the four dominant eigenvectors for the diffusion in a four-well potential. The first eigenvector corresponds to the stationary distribution. The



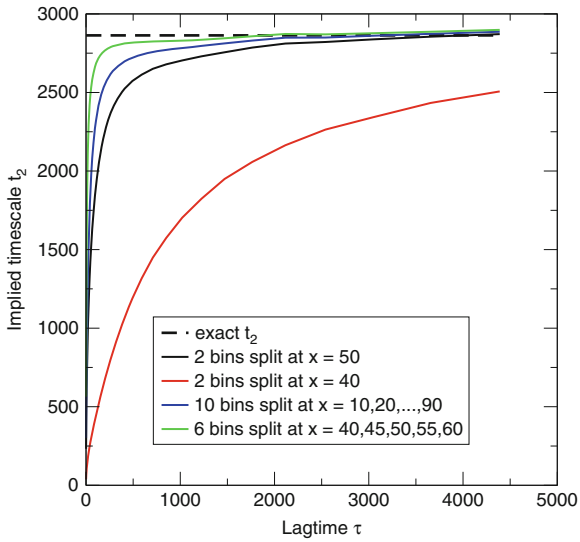
Markov Models of Molecular Kinetics, Fig. 2 Chapman-Kolmogorov-Test for MSMs of a diffusion in a double-well potential (**a**). (**b**, **c**) compare the probability of being in the left minimum over time, given that the dynamics starts in the left basin. The test was done for the two-well potential using a trajectory of length 10^6 steps. Tested are Markov models that use lag times $\tau = 100$, 500, 2000 and (**b**) 2-state discretization (split at $x = 50$), (**c**) 6-state *discretization (split at* $x = 40, 45, 50, 55, 60$) (Figure adapted from Prinz et al. (2011))

second eigenvector corresponds to the slowest process and has positive signs in regions A and B and negative signs in regions C and D, thus corresponding to the transition between (A,B) and (C,D). The third eigenvector corresponds to the transition between A and B, while the fourth corresponds to the transition between C and D.

Metastable States

Markov models from clustered molecular dynamics data often require thousands of microstates. It is thus desirable to find a simplified representation that communicates the essential properties of the kinetics. Let us consider the coarse partition of state space $\Omega = \{C_1, C_2, \ldots, C_n\}$ where each cluster C_i contains multiple microstates 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 remain 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.

Schütte and coworkers proposed that the metastable states could be identified by grouping microstates according to the signs in the dominant eigenvectors $\mathbf{l}_2, \mathbf{l}_3$ etc. (Schütte et al. 1999; Weber 2003). Based on that, Weber (2003) developed PCCA+, an optimal method for identifying metastable sets. When plotting the values each microstate has in its $m - 1$ right



Markov Models of Molecular Kinetics, Fig. 3 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 the finer partitions (*blue*, *green*) (Figure adapted from Prinz et al. (2011))

eigenvectors $\mathbf{r}_2$, $\mathbf{r}_3$, …, $\mathbf{r}_m$, these values lie in a simplex whose vertices correspond to metastable states and the most metastable partition is found by assigning microstates to their closes vertices. See Fig. 4 for an illustration, and Weber (2003) for a more detailed description.

Note that metastable states are very useful for illustrative purposes. If the dynamics are very metastable, they may even serve as MSM microstates because then the dynamics loses memory before exiting to another metastable state, yielding an effectively Markovian partition de Groot et al. (2001). In general, this is not the case, and for quantitatively modeling the system kinetics, it is thus recommeded to maintain a fine discretization as the MSM discretization error will increase when states are lumped (see section "Long-Term Kinetics from Short Simulations").

Figure 5 shows metastable states of the folding dynamics of Pin WW Noé et al. (2009).

Transition Pathways

Understanding the mechanisms of conformational transitions, such as protein folding, RNA folding, native conformational transitions in proteins, or binding of ligands to proteins, is one of the grand challenges in biophysics. Let A and B be two subsets of state space (e.g., denatured and folded state), and let all remaining states be "intermediate" states I. What is the probability distribution of the trajectories leaving A and continuing on to B? That is, what is the typical sequence of states I used along the transition pathways? When an MSM is available, these questions can be answered by Transition Path Theory (TPT) (Weinan and vanden-Eijnden 2006; Metzner et al. 2009; Noé et al. 2009).

The TPT equations are derived for rate matrices in Metzner et al. (2009) and for transition matrices in Noé et al. (2009). The essential ingredient required to compute the statistics of transition pathways is the committor probability. $q_i^+ \cdot q_i^+$ is the probability when being at state i, the system will reach the set B next rather than A (Du et al. 1998). The committor can be calculated from the equations:

$$\begin{aligned} q_i^+ &= 0 & i &\in A \\ q_i^+ &= 1 & i &\in B \\ \Sigma_{k\in I} T_{ik} q_k^+ &= -\Sigma_{k\in B} T_{ik} & i &\in I \end{aligned}$$

The backward-committor probability, q_i^-, is the probability, when being at state i, that the system was in set A previously rather than in B. For dynamics obeying detailed balance:

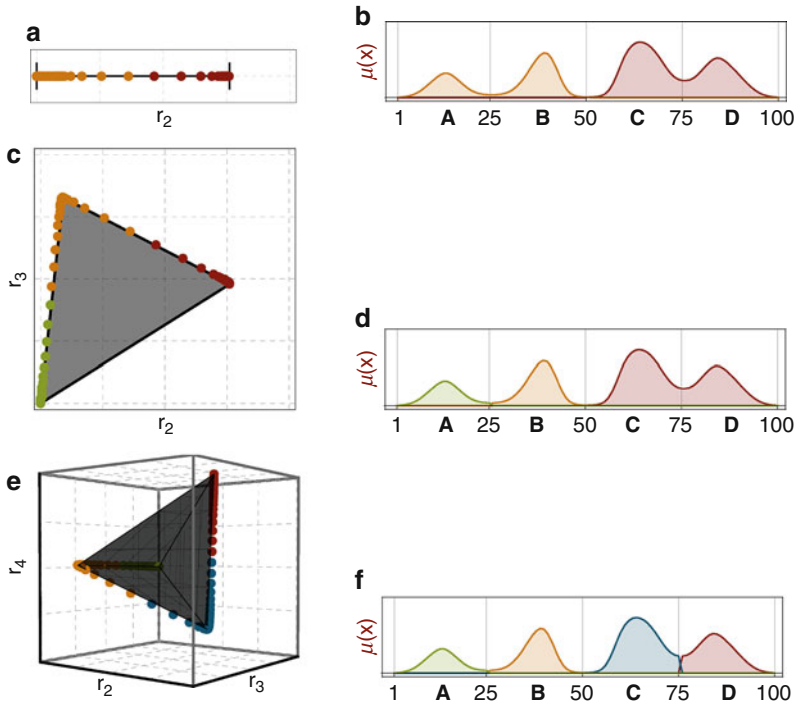
$$q_i^- = 1 - q_i^+.$$

Consider the probability flux between two states i and j, given by $\pi_i T_{ij}$. TPT only considers 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:

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

Furthermore, contributions from recrossings or detours are removed. Thus, the net flux is defined 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 $q_i^+ \le q_j^+$, an equivalent expression is (Berezhkovskii et al. 2009):

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



Markov Models of Molecular Kinetics, Fig. 4 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. Each of these partitions 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 (Figure adapted from Prinz et al., Phys. Chem. Chem. Phys. 13, p 16912 (2011))

f_{ij}^+ defines a network of fluxes leaving states A and entering states B. It is similar to an electric network where the "voltage" $(q_j^+ - q_i^+)$ across a "conductivity" $\pi_i T_{ij}$ gives rise to a "current" f_{ij}^+. The total flux created in A and consumed in B is:

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

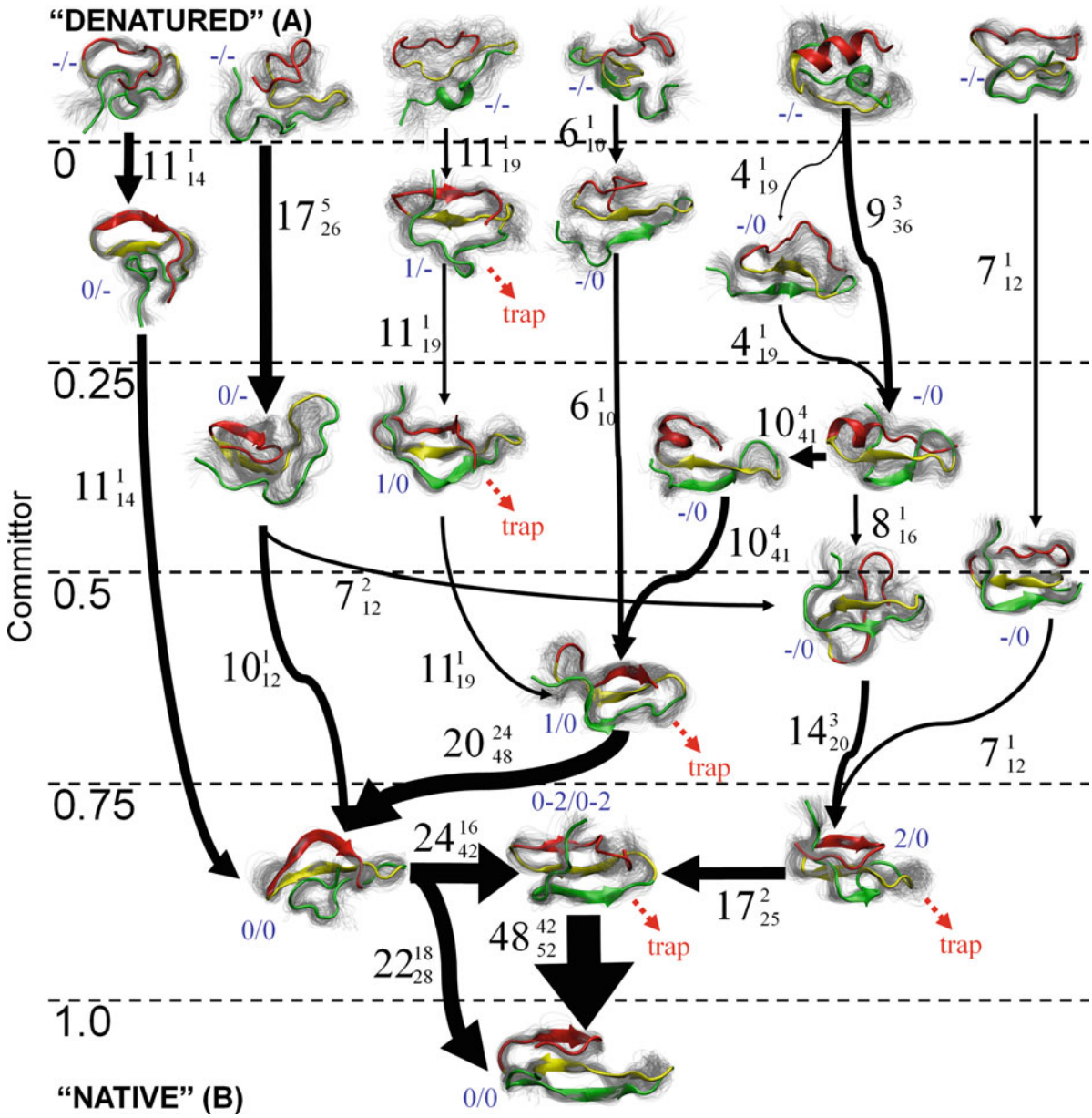
Of special interest is the reaction rate constant Noé et al. (2009):

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

The fluxes f^+ can be coarse grained by summing fluxes crossing the boundaries of metastable states (see Fig. 5). This yields a simplified view on the transition investigated. The flux f^+ can be decomposed into individual pathways (Metzner et al. 2009) and their relative contribution to the $A \to B$ process can be evaluated (Noé et al. 2009). As an example, the folding pathways for the Pin WW protein are shown in Fig. 5.

Experimental Observables/Dynamical Fingerprints

Biophysical experiments measure one or multiple observables $a(\mathbf{x})$ which are functions of the high-dimensional macromolecular coordinates. a could be a fluorescence or transfer efficiency in a fluorescence experiment, an NMR chemical shift, the intensity of an IR spectral peak, the distance in a pulling experiment, etc. Let a_i be the mean value of observable a over the state S_i. Given the observable vector $\mathbf{a} = [a_i]$, various experimental measurements can be expressed in terms



Markov Models of Molecular Kinetics, Fig. 5 Flux of the folding transitions among the metastable states of the PinWW protein (Figure adapted from Noé et al. (2009))

of the transition matrix $\mathbf{T}$ and linked to its eigenvalues and eigenvectors (Noé et al. 2011).

In stationary equilibrium experiments, the mean value of an observable a, $\mathbb{E}_\pi[a]$, is recorded. This may be either done my 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:

$$\mathbb{E}_\pi[a] = \sum_{i=1}^{n} a_i \pi_i = \langle \mathbf{a}, \pi \rangle. \tag{9}$$

where $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the scalar product between two vectors $\mathbf{x}$ and $\mathbf{y}$.

Kinetic information is available from time-correlation experiments. These may be realized by

computing time-correlation functions from single molecule trajectories (e.g., fluorescence correlation spectroscopy) or by scattering techniques (e.g., inelastic neutron scattering). The time cross-correlation of two observables a and b can be computed as:

$$\mathbb{E}\left[a(t)\, b(t+k\tau)\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \pi_i a_j \left[\mathbf{T}^k(\tau)\right]_{ij} = \langle \mathbf{a}, \pi\rangle\langle \mathbf{b}, \pi\rangle + \sum_{i=2}^{n} \exp\left(-\frac{k\tau}{t_i}\right)\langle \mathbf{a}, \mathbf{l}_i\rangle\langle \mathbf{b}, \mathbf{l}_i\rangle. \tag{10}$$

For the autocorrelation of a:

$$\mathbb{E}\left[a(t)\, a(t+k\tau)\right] = \langle \mathbf{a}, \pi\rangle^2 + \sum_{i=2}^{n} \exp\left(-\frac{k\tau}{t_i}\right)\langle \mathbf{a}, \mathbf{l}_i\rangle^2$$

In relaxation experiments, the system is allowed to relax from a nonequilibrium starting state with probability distribution $\mathbf{p}(0)$. 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 $\mathbf{T}(\tau)$ with stationary distribution $\boldsymbol{\pi} \neq \mathbf{p}(0)$. The ensemble average $\mathbb{E}_{\mathrm{p}(0)}$[a (t)] is recorded while the system relaxes from the initial distribution $\mathbf{p}(0)$ to the new equilibrium distribution π:

$$\mathbb{E}_{\mathbf{p}(0)}\left[a(k\tau)\right] = \sum_{i=1}^{n} a_i p_i(k\tau) = \langle \mathbf{a}, \pi\rangle + \sum_{i=2}^{n} \exp\left(-\frac{k\tau}{t_i}\right)\langle \mathbf{p}'(0), \mathbf{l}_i\rangle\langle \mathbf{a}, \mathbf{l}_i\rangle. \tag{11}$$

where $p_i'(0) = p_i(0)/\pi_i$. Both Eqs. 10 and 11 have the form of a multiexponential decay function with implied timescales of the transition matrix. Each timescale enters the observation with an amplitude that depends on the overlap between the corresponding eigenvector Eq. 9 and the observable (s), and in relaxation experiments also on the initial conditions of the experiment. For any given experimental observable, many amplitudes will be near zero; thus, even complicated kinetics may have the signature of two- or three-state systems in a single given kinetic experiment.

The ability to link experimentally measurable relaxation timescales to individual eigenvalue/eigenvector pairs allows structural processes to be assigned to these timescales via the eigenvector (see "Interpretation of Eigenvectors" above). This reconciliation of simulations and experiments is described in detail via the concept of *dynamical fingerprints*. Furthermore, this approach permits to design experiments that are optimal to probe individual relaxations (Noé et al. 2011).

Summary

Markov modeling is a theoretical framework suitable for analyzing molecular dynamics or any other stochastic process that is ergodic and Markovian in full state space. Markov (state) models (MSMs) approximate the complex original dynamics by transition probabilities between discrete subsets of the possibly high-dimensional state space. In molecular dynamics, these subsets may correspond to molecular conformations, rotamers, foldamers, or binding states. 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 are not. This allows relatively detailed analyses such as using transition path theory to calculate the ensemble of transition pathways between two subsets of state space, or the assignment of structural processes to the kinetic features of experimental observables.

It has been intensively debated whether it is generally feasible to approximate the high-dimensional continuous dynamics of macromolecules by a discrete Markov process on relatively few (typically 10^2–10^5) discrete states. A number of theoretical developments between 2000 and 2010 have shown that this is indeed feasible if the system has relatively few slow relaxation processes, typically arising from the transitions between metastable states. This makes MSMs especially interesting to biological macromolecular processes, such as conformational changes, folding, binding, and oligomerization of peptides, proteins, and nucleic acids. Whether MSMs can also be practically useful to investigate processes with combinatorially

exploding state spaces, such as spin systems, remains a subject of ongoing research.

Current challenges of Markov model methodology lie especially in the development of robust adaptive methods for discretization and sampling: (1) The basic mathematical relation between a state space discretization and the quality of an MSMs is now understood. The translation of the mathematical insight of discretization quality into a robust adaptive discretization algorithm is an important step toward efficient construction of MSMs for complex systems. (2) Enhanced sampling methods to explore the state space such as metadynamics and multi-ensemble methods are complementary to MSM modeling of the equilibrium dynamics. Consistently integrating these approaches is an important step toward efficient simulation (Sriraman et al. 2005). (3) It has been demonstrated on simulation models that the statistical uncertainties of the MSM transition matrix and quantities calculated from it can be used to allocate new simulations so as to speed up the convergence (Singhal and Pande 2005). These approaches need further development and are likely to significantly influence the molecular dynamics field.

Cross-References

► Molecular Dynamics Simulations of Lipids

References

Beauchamp KA, Bowman GR, Lane TJ, Maibaum L, Haque IS, Pande VS. MSMBuilder2: modeling conformational dynamics at the picosecond to millisecond scale. J Chem Theory Comput. 2011;7(10):3412–9. doi:10.1021/ct200463m. ISSN: 1549-9626. URL: http://dx.doi.org/10.1021/ct200463m.

Berezhkovskii A, Hummer G, Szabo A. Reactive flux and folding pathways in network models of coarse-grained protein dynamics. J Chem Phys. 2009;130(20):205102. doi:10.1063/1.3139063. ISSN: 1089-7690. URL: http://dx.doi.org/10.1063Z1.3139063.

Chodera JD, Dill KA, Singhal N, Pande VS, Swope WC, Pitera JW. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. J Chem Phys. 2007;126:155101.

de Groot B, Daura X, Mark A, Grubmüller H. Essential dynamics of reversible peptide folding: memory-free conformational dynamics governed by internal hydrogen bonds. J Mol Biol. 2001;301:299–313.

Du R, Pande VS, Yu A, Tanaka T, Shakhnovich ES. On the transition coordinate for protein folding. J Chem Phys. 1998;108(1):334–50. doi:10.1063/1.475393. URL: http://dx.doi.org/10.1063/1.475393.

Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science. 1991;254:1598–603.

Metzner P, Schütte C, Vanden Eijnden E. Transition path theory for Markov jump processes. Multiscale Model Simul. 2009;7:1192–219.

Noé F. Probability distributions of molecular observables computed from Markov models. J Chem Phys. 2008;128:244103.

Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR. Constructing the full ensemble of folding pathways from short off-equilibrium simulations. Proc Natl Acad Sci USA. 2009;106:19011–6.

Noé F, Doose S, Daidone I, Löllmann M, Chodera JD, Sauer M, Smith JC. Dynamical fingerprints: understanding biomolecular processes in microscopic detail by combination of spectroscopy, simulation and theory. Proc Natl Acad Sci USA. 2011;108:4822–7.

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

Sarich M, Noé F, Schütte C. On the approximation error of Markov state models. SIAM Multiscale Model Simul. 2010;8:1154–77.

Schütte C, Fischer A, Huisinga W, Deuflhard P. A direct approach to conformational dynamics based on hybrid Monte Carlo. J Comput Phys. 1999;151:146–68.

Senne M, Trendelkamp-Schroer B, Mey ASJS, Schütte C, Noé F. Emma – a software package for Markov model building and analysis. J Chem Theory Comput (in revision). 2012.

Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W. Atomic-level characterization of the structural dynamics of proteins. Science. 2010;330(6002):341–6. doi:10.1126/science.1187409. ISSN: 1095-9203. URL: http://dx.doi.org/10.1126/science.1187409.

Singhal N, Pande VS. Error analysis and efficient sampling in Markovian state models for molecular dynamics. J Chem Phys. 2005;123:204909.

Sriraman S, Kevrekidis IG, Hummer G. Coarse master equation from Bayesian analysis of replica molecular dynamics simulations. J Phys Chem B. 2005;109:6479–84.

Swope WC, Pitera JW, Suits F. Describing protein folding kinetics by molecular dynamics simulations: 1. Theory. J Phys Chem B. 2004;108:6571–81.

Voelz VA, Bowman GR, Beauchamp K, Pande VS. Molecular simulation of ab initio protein folding for a millisecond folder NTL9. J Am Chem Soc. 2010;132(5):1526–8. doi:10.1021/ja9090353. URL: http://dx.doi.org/10.1021/ja9090353.

Weber M. Improved Perron cluster analysis, ZIB Report, 03-04. Berlin-Dahlem: ZIB; 2003.

Weinan E, vanden-Eijnden E. Towards a theory of transition paths. J Stat Phys. 2006;123(3):503–23. doi:10.1007/s10955-005-9003-9. ISSN: 0022-4715. URL: http://dx.doi.org/10.1007/s10955-005-9003-9.