Published in final edited form as:
Curr Opin Struct Biol. 2014 April ; 25: 135–144. doi:10.1016/j.sbi.2014.04.002.

Markov state models of biomolecular conformational dynamics

John D. Chodera[a] and **Frank Noé**[b]

John D. Chodera: choderaj@mskcc.org; Frank Noé: frank.noe@fu-berlin.de

[a]Computational Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[b]Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany

Abstract

It has recently become practical to construct *Markov state models* (MSMs) that reproduce the long-time statistical conformational dynamics of biomolecules using data from molecular dynamics simulations. MSMs can predict both stationary and kinetic quantities on long timescales (e.g. milliseconds) using a set of atomistic molecular dynamics simulations that are individually much shorter, thus addressing the well-known sampling problem in molecular dynamics simulation. In addition to providing predictive quantitative models, MSMs greatly facilitate both the extraction of insight into biomolecular mechanism (such as folding and functional dynamics) and quantitative comparison with single-molecule and ensemble kinetics experiments. A variety of methodological advances and software packages now bring the construction of these models closer to routine practice. Here, we review recent progress in this field, considering theoretical and methodological advances, new software tools, and recent applications of these approaches in several domains of biochemistry and biophysics, commenting on remaining challenges.

Keywords

Markov state models; molecular kinetics; molecular dynamics; ligand binding; protein folding; conformational change

1. Introduction

The study of biomolecular systems by molecular dynamics simulations is by no means straightforward. Aside from a myriad of issues related to the accurate treatment of intra- and intermolecular interactions and appropriate modeling of the chemical environment, the timescales relevant to biomolecular folding and function (often microseconds to seconds) are enormously long compared to the timesteps required for stable integration (generally femtoseconds) [1]. Even with expensive special-purpose hardware such as Anton [2],

simulation trajectories can barely reach biomolecular timescales of typical interest, let alone *exceed* to permit their statistical characterization beyond simple anecdotal observation.

While various solutions to the timescale problem have been explored, many practitioners have now adopted a practice of extracting stochastic kinetic information from multiple simulations that are *shorter* than the timescales of interest to build a dicrete-state stochastic model capable of describing long-time statistical dynamics. These *Markov state models* (MSMs) describe the stochastic dynamics of a biomolecular system using two objects: (1) a *discretization* of the high-dimensional molecular state space, usually into *n* disjoint conformational sets $S_1, \ldots, S_n$, and (2) a model of the stochastic trasitions between these discrete states, usually described by a matrix of conditional transition probabilities estimated from the simulation trajectories x_t, and termed the *transition matrix* $\mathbf{P} \equiv (p_{ij})$:

$$P_{ij}(\tau) = \mathrm{Prob}(x_{t+\tau} \in S_j | x_t \in S_i). \quad (1)$$

Here, τ is the *lag time* or *observation interval* for which the transition matrix is constructed. As recent literature (reviewed below) highlights, this lag time τ turns out to be an important parameter in determining the approximation quality and utility of the MSM, with larger τ providing models of higher fidelity but coarser termporal resolution.

A transition matrix **P** gives rise to a stationary distribution π by virtue of the simple eigenvalue problem:

$$\pi^{\mathrm{T}}\mathbf{P} = \pi^{\mathrm{T}}. \quad (2)$$

This is a key property of MSMs: While the matrix $\mathbf{P} \equiv (p_{ij})$ only contains conditional transition probabilities (which can be computed from short trajectories of length $\sim \tau$, usually orders of magnitude shorter than the longest relaxation timescales), the global stationary distribution π can still be computed from **P**. The MSM correctly recovers the equilibrium thermodynamic and kinetic properties of the system, even if the short trajectories used to construct it were not initiated from equilibrium. Additionally, while identifying a suitable state space discretization is by no means trivial, MSMs offer the advantage over many other methods addressing sampling problems that slow order parameters do not need to be defined *a priori*.

Another important feature of MSMs is that many quantities of interest can be easily calculated from them. A conceptually simple approach is to use the transition matrix **P** to generate discrete trajectories with time resolution τ, sampling a molecular configuration from the configurations associated with each discrete state. Alternatively, any molecular observable amenable to such a treatment can also be computed by algebraic equations involving **P** without the need to resort to such sampling schemes. The latter approach avoids trajectory sampling, and can often yield additional insight into the dominant contributions of the observed temporal behavior of a given experiment. Key ingredients are the eigenvalues λ_i and eigenvectors $\mathbf{r}_i$ of the transition matrix:

$$\mathbf{P}\mathbf{r}_i = \mathbf{r}_i\lambda_i, \quad (3)$$

where the eigenvalues translate to molecular relaxation timescales, $t_i = -\tau/\ln|\lambda_i(\tau)|$, and the eigenvectors indicate which structural changes occur at this timescale (see Figure). This duality is crucial for the interpretration of kinetic molecular experiments via MSMs.

Recently, two books have been produced on MSM theory and use. The first is a reasonably comprehensive survey of the current theory and practice of Markov state model construction [4], while the second focuses on advanced mathematical and theoretical aspects [5].

While a number of literature reviews and overview articles cover the fundamentals of Markov state models [6, 7], the present review focuses on theoretical advances and applications that have been published since these reviews were written.

2. Recent theoretical and methodological advances

Paradigm shift: From maximizing metastability to approximating eigenspaces

For many years, MSM construction techniques were driven by the goal of attempting to construct a state space discretization that was *maximally metastable*, based on the intuition that the discrete state dynamics should be approximately Markovian (memoryless). To achieve this, most schemes attempted to maximize quantities related to the *lifetimes* of the projected discrete states, ensuring that subsequent state transitions were maximally decorrelated from their previous transition history.

Recent theoretical work has shown it is more useful to instead consider the MSM as a *discrete approximation* to the dynamics of the Markov operator (transfer operator or dynamical propagator) in the full state space of positions and velocities [8, 7, 5]. As a result, the fundamental goal of state discretization has shifted from maximizing lifetimes to minimizing approximation error of the statistical long-time dynamics.

As an example, consider a double-well potential. Maximizing the lifetimes would lead one to construct a two-state model, with a single partition placed at the transition state between the two wells. Introducing additional partitions near the transition state will reduce the discrete state lifetimes but actually *increase* the approximation accuracy of the model by allowing it to better approximate the slow eigenspace of the Markov operator [7]. In addition, selected dynamical processes of interest can be approximated to arbitrary accuracy [9].

Eigenvalues and eigenvectors

A key finding is that good MSMs are able to accurately approximate long-time molecular kinetics because their eigenvectors closely approximate the corresponding eigenfunctions of the Markov operator associated with large eigenvalues [8, 7, 5]. These eigenvalues $\lambda_i(\tau)$ for a Markov operator of observation time τ are related to intrinsic molecular relaxation timescales, while the eigenfunctions describe the associated structural reconfigurations.

This perspective has also facilitated the exposure of fundamental connections between MSMs and related approaches. For instance, the construction of diffusion maps—which aim to approximate these eigenfunctions for overdamped Langevin dynamics—have been

extended to model biomolecular dynamics, despite being originally introduced as a general data analysis tool [10]. Diffusion maps have been used for adaptive exploration of the conformation space along the slow-process eigenfunctions [11].

Given the above insights, it is not surprising that the *systematic estimation error* of MSM-derived relaxation timescales (or rates) can be bounded in terms of how well the MSM discretization approximates the Markov operator eigenfunctions [12]. In Ref. [13], it was found that virtually all rate theories trying to extract transition rates from trajectory data—either from molecular dynamics simulations or biophysical experiments—can be cast in a similar manner in which the error intrinsic to many classical rate theories could be computed in terms of this eigenfunction approximation error.

Variational approach

Viewing MSMs as a method to approximate the eigenfunction of a Markov operator invokes parallels to quantum chemistry, where the goal is to approximate the eigenfunctions and eigenvalues of the Hamiltonian. It was recently discovered that the Rayleigh-Ritz variational principle—a fundamental concept in quantum chemistry—has an analog in molecular conformation dynamics, and that the eigenfunction approximation problem that appears in constructing MSMs can be cast as a generalized eigenproblem [14]. This formulation implies that MSM-derived relaxation timescales are always underestimated except when linear combinations of the true eigenfunctions of the Markov operator are used as basis functions. Standard “crisp paritioning” MSMs based on clustering simulation data were shown to be a special case in which the basis set used in variational optimization is chosen to be a set of functions that are constant on the discrete MSM states [14].

This insight has far-reaching consequences. Just as better basis sets led to better computational models in quantum chemistry, alternative basis sets for MSM construction could lead to better and more informative models of molecular kinetics. Ref. [15] discusses differences and similarities to quantum-mechanical approaches that may be exploited and applies the variational principle to the approximation of the peptide dynamics via Gaussian basis sets. The variational approach is the keystone for searching for more efficient and informative basis sets than Voronoi partitions of high-dimensional coordinate space.

Coordinate spaces for discretization

A main difficulty in constructing accurate MSMs from biomolecular simulation data is the need to balance statistical error and systematic approximation error: while a fine partitioning will minimize approximation error, the limited quantity of trajectory data means that fine partitionings will increase statistical error. An important question in the field has been what combination of distance metric and clustering method would provide a reproducibly good approximation of the dominant eigenfunctions of the Markov operator. Earlier work employed torsion angles, Cartesian coordinates (potentially following translation/rotation onto a reference structure), or principal components of any of these coordinates. Another popular approach has been the use of pairwise minimal root-mean-square deviation (RMSD). Solvent degrees of freedom that play a critical role in defining kinetically distinct states are also notoriously difficult to deal with [16].

A challenge in all these approaches is that they propose a distance metric *a priori*. Because trajectory data is always limited in quantity by practical simulation times, it is essential that configurations that are highly similar by this metric are actually kinetically related. Fixed distance metrics suffer from the drawback that very small neighborhoods are generally required in order to guarantee this kinetic relatedness, which could then require an unrealistically large quantity of data to build a high-quality MSM. For this reason, the community sought out data-adapted distance metrics that were able to enlarge these neighborhoods of similarity by learning about kinetic similarity from the trajectories. For example, a scheme using *kinetic discriminatory metric learning* was proposed in which a generalized weighted squared Euclidean metric was optimized [17]. In a recent study [18], an alternative approach to the discretization of the diffusion operator relevant for transport of small molecules (e.g. ions) in biological cells was proposed.

A major breakthrough followed in the exploitation of time-delayed correlation analysis, also called time-lagged or time-structure-based independent component analysis (tICA or TICA) [19]. This approach, independently rediscovered by two groups [20, 21], linearly transforms the input coordinates (e.g. torsions, distances, contacts) into collective coordinates sorted by "slowness", thus providing an excellent dimension reduction method for MSM construction. This can be understood as a generalization of principal component analysis, as it solves a generalized eigenvalue problem with the instantaneous and time-lagged correlation matrices of the input coordinates [20, 21]. This approach can also be shown to solve the variational problem for the specific choice of basis functions identical to the input coordinates [21].

Full partition versus cores

Several recent contributions have addressed the question of how best to project from high-dimensional trajectory data to a sequence of discrete state labels, from which the MSM transition matrix is subsequently estimated. Most of the recent MSM studies construct a Voronoi tesselation to use in assigning configurations to discrete states. An alternative approach is the *core*-based projection method, long familiar to the transition path sampling community but first introduced for MSM construction in [22]. Here, one defines an incomplete partition of space into highly metastable *cores* located within distinct metastable basins. The continuous trajectory is discretized by counting changes in state assignment only upon encountering a different core, maintaining its association with the previous core in the intermediate region between cores. This core discretization scheme was subsequently shown to be equivalent to *milestoning*, with the cores being milestones [23]. In the MSM framework, this core projection method effectively uses the core committor functions as basis functions, which often provides a superior approximation to the eigenfunctions of the transfer operator when the number of cores is small [23]. However, a general procedure for identifying cores in high-dimensional configuration space has not yet been proposed.

Estimation of MSMs

Once the high-dimensional trajectory has been projected onto a discrete sequence of states, an MSM is obtained by estimating a kinetic model of the discrete dynamics between states. In most cases, a matrix of conditional transition probabilities is estimated using some lag time τ from the trajectory data [7]. Early approaches estimated transition probabilities as the

number of transition events from *i* to *j*, divided by the number of frames in *i*, while recent MSM methods have used higher degrees of sophistication. Commonly, the transition matrix is only estimated over the largest connected subset of states [24, 25] and proceeds via an iterative maximum likelihood estimator that respects detailed balance [26, 7].

To assess the statistical uncertainty of the MSM model and its predictions, it is important to consider not only the most likely transition matrix, but *all* such matrices statistically consistent with the data. This can now be efficiently achieved (including the use of detailed balance constraints) through Markov chain Monte Carlo (MCMC) techniques [27]. A recent publication has shown how knowledge of the stationary distribution (from, say, a long parallel tempering simulation) can be used to further constrain the space of transition matrices [28].

MSMs can also be constructed directly from parallel tempering simulations using the short trajectory segments generated between exchange attempts, provided the discretization is sufficiently good to permit the use of very short lag times [29]. This approach was subsequently improved upon [24] by the introduction of a dynamical reweighting method [30] that permits use of data from all temperatures, which also allows the resulting MSM to capture the continuous dependence on temperature.

Coarse-graining of MSMs

An MSM that closely approximates the statistical dynamics of a biomolecular system may have hundreds or thousands of discrete states. To obtain an interpretable model from this, some form of coarse-graining is often necessary. Such a lumping operation—unless made in a specific mathematical form—should only serve the better understanding of an MSM, e.g., for the sake of a visualization of structures in kinetically distinct states. While useful for illustration purposes, re-estimating the transition matrix on a coarser state space is not trivial, as lumping of states will degrade the approximation quality of the MSM, and potentially lead to vastly underestimated timescales.

The foundational work on coarse-graining MSMs was the development of PCCA [31] and PCCA+ [32]. These spectral clustering methods use the sign structure of the leading eigenvectors of the MSM transition matrix to relate conformational transitions among metastable states with dynamical relaxation processes and kinetic experimental observables.

However, when used for simple lumping of states, PCCA(+) operates on only a single estimate of the transition matrix, resulting in choices of set boundaries that can be dominated by statistically uncertain features. Recent methods such as the Bayesian agglomerative clustering (BACE) method [33], a hierarchical Nystrom method [34], flux PCCA+ [35] and HMM-based coarse-graining [3] have proposed clustering methods to address this statistical reliability issue. The performance of several such methods have been compared [36], although we note that the result of such a comparison depends on the choice of benchmark. Ref. [37] proposed a network-based approach to lumping MSM states into metastable sets, and observed near-exponential exit time distributions when applied to peptide dynamics.

Projected Markov models and Hidden Markov models

The fundamental approximation of MSMs is that the statistical dynamics on discrete sets can be accurately approximated by a Markov chain. Many of the methods described above aim at reducing the discretization errors due to this decision, by good choices of the input coordinates and the discretization metric and method used. However, the discretization error made by the Markov approximation can be large. While it is known that the molecular relaxation timescales are better estimated at large lag times τ, Ref. [13] shows that the error behaves as τ^{-1}, explaining the slow convergence observed in implied timescale plots of recent papers.

Ref. [3] proposes a different approach, the Projected Markov model (PMM), in which discretized molecular dynamics trajectories can be *exactly* described by a simple mathematical structure consisting of two matrices—an eigenvalue matrix (corresponding to the relaxation timescales) and a projection matrix (corresponding to eigenfunctions that have been projected onto the cluster states, and are therefore different from MSM eigenvectors). It was shown that PMMs can be efficiently estimated by Hidden Markov model techniques (HMMs), and that the resulting HMM is as easy to interpret as an MSM. Future research will have to address the question how PMMs can be generally estimated, without making the HMM approximation. Ref. [13] gives an optimal estimation method for two-state PMMs, which is also expected to be useful for the analysis of single-molecule experimental data.

Software

A number of software packages have been developed to aid in the construction, validation, and interpretation of Markov state models, and have now reached relative maturity. Both the **EMMA** [38] [https://simtk.org/home/emma] and **MSMBuilder** [25] [http://msmbuilder.org/] software packages facilitate the construction and validation of Markov state models from molecular simulation data in various trajectory formats. The **MSMExplorer** software package [https://simtk.org/home/msmexplorer] allows for the visualization of MSMs using graph and network diagrams, scatterplots of state properties, and interactive structure visualization [39].

3. Applications

MSM methodologies have now been applied to a wide variety of problems in biomolecular dynamics to study the folding of proteins, the dynamics of intrinsically disordered peptides and proteins, ligand binding processes, native-state and functional dynamics, and the connection between molecular dynamical processes and their experimentally-resolved single-molecule and ensemble kinetics. Below, we review a few notable recent examples.

Protein folding

The study of how unfolded or disordered proteins reach their native states has long been one in which simulation, with its ability to resolve dynamics in atomistic detail, could provide critical insight difficult to extract from bulk or single-molecule experiments, which are inherently limited in either time or structural resolution. However, numerous challenges have traditionally stood in the way of simulations reaching experimentally relevant

timescales and gathering sufficient statistics to make definitive statements about protein folding mechanism [1]. While computational advances have ushered in a new era of molecular simulation software accelerated by graphics processing hardware [40, 41, 42] or even custom ASICs [2], raw advances in computational power are insufficient to produce more than anecdotal observations of protein folding or unfolding events [43]. By contrast, the MSM approach has led to the construction of extremely detailed statistical models of folding mechanism and unfolded state dynamics for some small peptides and proteins (e.g. [44, 45]).

Applications of MSM methodology to model systems in protein folding have generally revealed a surprising complexity to the apparently simple dynamics observed in experiments, in defiance of the “Occam’s Razor” interpretation (but still consistent with the experimental observations). For example, a combined experimental and computational study of Acyl-CoA binding protein (ACBP) examined the MSMs constructed from over 30 ms of aggregate molecular simulation data to probe the unfolded state dynamics, intermediate states, and native state formation rates, concluding that a previously-characterized fast kinetic phase did not correspond to population of a specific intermediate structure (as had previously been assumed), but was instead due to heterogeneous structure acquisition within the unfolded state [46, 47], a phenomenon also observed in a fast-folding protein [48].

Even more surprisingly, an examination of fourteen massive protein simulation datasets suggested that there are significant, experimentally detectable deviations from the two-state behavior generally believed to describe the folding of many simple model protein systems, suggesting a previously unappreciated complexity [35]. A folding study of ubiquitin has revealed the existence of several intermediates and misfolded states [49]. The dominant experimentally observable slow processess appear to also be robust to perturbations in intrinsic rates that would be expected from variations in experimental conditions, suggesting this behavior should be relatively universally observable [50]. In fact, an examination of the well-studied fast-folding HP35 domain by simulation identifies an intermediate state whose existence was only recently revealed by triplet-triplet energy transfer experiments [51].

MSMs have also been used to examine the role of glassy dynamics in protein folding, finding that true glass transitions appear to be absent in atomistic models of solvated protein dynamics [52]. In fact, sidechains of residues within the cores of folded proteins appear to show some degree of liquid-like behavior [53].

Protein-ligand binding

Markov state models also have been used to illuminate the mechanism of small molecule binding and unbinding from protein targets, often with the aim of discovering new putative binding sites that could be exploited for the design of novel inhibitors. While direct simulation can achieve timescales sufficient to observe ligand binding events [54], the bound half-lives of typical drugs is often on the hours timescale, far out of reach of drug binding studies. Additionally, by aggregating statistical information from many trajectories, a more detailed (and less anecdotal) mechanistic understanding of binding can be obtained.

The viability of using MSMs to study small molecule ligand binding in atomistic detail was established in an initial study of the binding kinetics of the trypsin inhibitor benzamidine was described by an MSM constructed from a collection of short trajectories, revealing the existence of long-lived binding intermediates [55]. A related network-based method using a scheme for enforcing detailed balance was able to describe the multiple unbinding pathways of ligands dissociating from FKBP [56]. A study that used the simplifying assumption of fixed receptor geometry used the framework transition path theory to extract descriptions of unbinding pathways and how these changed in response to receptor point mutations [57]. More recently, the often significant effects of *rebinding* on modulating effective dissociation kinetics of multivalent ligands has been illuminated through the use of MSMs [58].

Intrinsically disordered proteins

Studies of intrinsically disordered proteins have recently benefitted from a number of technical advances to better deal with conformationally and dynamically heterogeneous dynamics, such as the aforementioned tICA scheme. This proved critical in enabling the construction of MSMs for the intrinsically disordered KID peptide, which was found to adopt a conformation that may precede its KIX-bound geometry even in the absence of its binding partner [21]. Studies of the unfolded dynamics of the intrinsically disordered hIAPP peptide (found in 95% of type II diabetes patients) also found a surprising quantity of structure in these intrinsically unfolded states, suggesting these metastable conformations may seed aggregation processes [59]. A similar study on Aβ peptides implicated in Alzheimer's disease found similar propensities for disease-promoting truncations or mutations to populate structures that likely promote aggregation [60].

Native state conformation changes

With increasing simulation power, the possibility to study conformational changes of native proteins associated with function—which often have timescales in the microsecond to millisecond range—have become possible through the use of MSMs. The intrinsic fluctuations of β-lactamase in its native state reveals a multitude of potential allosteric binding sites that could potentially be exploited in the design of allosteric modulators of activity [61]. Recent MSM studies of the activation pathways of kinases [62] or GPCRs [63] has similarly revealed the potential for identifying putative allosteric binding sites or distinguishing between agonists and antagonists using structural information along putative functional pathways.

Other MSMs studies have revealed putative mechanisms for the autocatalytic step in HIV maturation of HIV protease [64] and the release of phosphate from bacterial RNA polymerase II during transcription elongation [65]. A combined experimental-computational study of dynamin tetramers was able to propose feasible oligomeric structures using MSM techniques [66].

4. Connecting simulation with experiment

The ability of MSMs to describe the statistical dynamics of a single biomolecule or the deterministic time-evolution of an ensemble of noninteracting biomolecules prepared in a

nonequilibrium state affords enormous power in the ability to connect models derived from atomistic simulation with a multitude of biophysical experiments. By coupling the atomistic resolution within conformational states with a spectroscopic model of the appropriate experimental technique, experimental observables can be compared directly with their computed counterparts, rather than resorting to interpretation through some intermediate-scale model that might suffer from a loss of potentially critical information.

Reconciling MSMs and kinetic experiments

While temperature-jump IR/fluorescence and fluorescence correlation measurements have seen widespread use in probing multiple kinetic timescales in biomolecular systems, it is not always straightforward to identify the structural relaxation processes associated with these timescales.

More recently, the framework of *dynamical fingerprints* has been developed as a principle way of separating both experimental and MSM-derived kinetic data into relaxation processes with distinct amplitudes and timescales, and associating specific observed structural relaxation processes with each [67, 68]. It was found that individual experimental observables can often pick up only few (1–2) relaxation processes even if many are present, suggesting a resolution to the apparent contradiction between simple experimental and complex simulated kinetics. However, it was suggested that MSMs and dynamical fingerprints could be used to design kinetic experiments so as to optimally probe selected relaxation timescales [67].

A number of concrete successes in connecting MSMs derived from atomistic simulations to biophysical experiments have been reported. NMR relaxation and order parameters are inherently dynamical quantities, and their computation from MSMs can be made straightforward through the use of the model-free framework approach of Lipari and Szabo, as recently demonstrated for HIV-1 protease [69]. Similar work has been carried out in developing a theory for the reconstruction of inelastic neutron scattering spectra from MSMs derived from atomistic simulations [70, 71]. An elegant framework for the simulation of spin-labeled continuous-wave electron spin resonance (cw-ESR) experiments from molecular simulations using MSMs has also been developed, as described recently in chapter 10 of [4].

MSMs have also found use in interpreting temperature-jump data using a variety of spectroscopic probes. Using exciton methods, MSMs were able to illuminate the processes giving rise to temperature-jump infrared (IR) and 2DIR data, providing physical insight into the observed spectroscopic signatures of distinct states in the trpzip2 peptide [72] and an α/β fragment of NTL9 [73]. Dynamical fingerprints were used to reconcile fluorescence correlation spectroscopy (FCS) measurements in glycine-serine peptides with molecular dynamics simulations [67]. MSMs were also used to interpret temperature-jump fluorescence measurements, once again in a spectroscopic study of the trpzip2 peptide [74].

Markov models from single-molecule experiments

While there is a long history of the use of *hidden* Markov models to identify kinetically metastable states in single-molecule biophysical experiments, only recently have sophisticated analysis techniques developed to the point of reliably resolving many states within experimentally observed traces, such that some of the theoretical pathway analysis tools described earlier (such as transition path theory) can be applied to models derived solely from experiment. A detailed six-state model of the folding/unfolding kinetics of adeylate kinase was constructed from a large quantity of single-molecule FRET data, permitting the application of transition path theory for the analysis of folding pathways [75]. Optical force spectroscopy experiments of calmodulin extrapolated to zero force also yielded a multistate MSM describing exchange among on- and off-pathway intermediates [76]. The complex MSM obtained from single-molecule FRET studies of the Diels-Alderase ribozyme permitted the use of transition path theory and transition matrix eigenvector analysis for describing the Mg^{2+}-dependent folding pathway [77].

5. Challenges and potential solutions

The MSM field has virtually exploded in the last few years, and has started to make significant impact in basic problems of biomolecular modeling and simulation. Despite advances in computational power, the sampling problem is still a major bottleneck in molecular simulation. It has been suggested that MSMs have the potential to solve this problem by using the statistical information encoded in the model in order to direct the simulation effort where new conformational states can be found with high probability or where statistics are still poor. However, an automatic and unsupervised realization of such an adaptive sampling procedure that works reliably for complex biomolecular systems remains elusive.

A key ingredient of such an adaptive approach is that the construction and estimation of MSMs must be fully automated. This will require additional theoretical advances to provide insight into the selection of approrpiate system- and data-appropriate parameters for the construction of MSMs. A particularly important aspect of this is balancing the statistical and the systematic error of MSMs—in the simplest case, the selection of an appropriate number of states that is not too small to cause major discretization errors and not too large to incur enormous statistical uncertainties or biases. It is likely that unsupervised MSM construction will also only be solved by an approach that is itself adaptive.

Both of the above aspects (adaptive sampling and fully automated MSM construction) must also be accompanied by software development that make these methods available to public use. We have already seen the rapid progross in this field that has been enabled by the open software tools already available.

A third aspect is the use of MSMs in force field parametrization. Force field development has steadily moved to include thermodynamic data, such as free energies of transfer or hydration [78]. With the computational power available to generate well-converged molecular dynamics simulations of sizable protein systems, thermodynamic and kinetic data on larger systems could be systematically included in the next generation of force field

development. MSMs provide a systematic approach to calculate these quantities in a way that is largely independent of subjective decisions of the modeler.

Acknowledgments

FN acknowledges support from ERC starting grant “pc-Cell”. JDC gratefully acknowledges support from the Memorial Sloan Kettering Cancer Center. The authors thank Antonia S. J. S. Mey for helpful comments on this manuscript.

References

1. Lane TJ, Shukla D, Beauchamp KA, Pande VS. To milliseconds and beyond: challenges in the simulation of protein folding. Curr Opin Struct Biol. 2013; 23:58–65. [PubMed: 23237705]
2. Shaw DE, Dror RO, Salmon JK, Grossman J, Mackenzie KM, Bank JA, Young C, Deneroff MM, Batson B, Bowers KJ, Chow E, Eastwood MP, Ierardi DJ, Klepeis JL, Kuskin JS, Larson RH, Lindorff-Larsen K, Maragakis P, Moraes MA, Piana S, Shan Y, Towles B. Millisecond-scale molecular dynamics simulations on Anton. Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis (SC09). 2009
3. Noé F, Wu H, Prinz JH, Plattner N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. J Chem Phys (submitted). 2013; 139:184114.
4••. Bowman, GR.; Pande, VS.; Noé, F., editors. Experimental Medicine and Biology. Springer; 2014. An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation, Vol. 797 of Advances. A comprehensive collection of recent reviews of various aspects of MSM construction.
5•. Schütte, C.; Sarich, M. in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches. American Mathematical Society; 2013. Metastability and Markov State Models. Vol. 24 of Courant Lecture NotesA comprehensive theoretical and mathematical treatment of advanced topics in MSM construction.
6. Pande VS, Beauchamp K, Bowman GR. Everything you wanted to know about Markov State Models but were afraid to ask. Methods. 2010; 52:99–105. [PubMed: 20570730]
7. Prinz JH, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera J, Schütte C, Noé F. Markov models of molecular kinetics: Generation and validation. J Chem Phys. 2011; 134:174105. [PubMed: 21548671]
8. Sarich M, Noé F, Schütte C. On the approximation error of markov state models, SIAM Multiscale Model. Simul. 2010; 8:1154–1177.
9. Sarich M, Schütte C. Approximating selected non-dominant timescales by Markov state models, Comm. Math Sci. 2012; 10:1001.
10. Rohrdanz MA, Zheng W, Maggioni M, Clementi C. Determination of reaction coordinates via locally scaled diffusion map. J Chem Phys. 134:124116. [PubMed: 21456654]
11. Zheng W, Rohrdanz MA, Clementi C. Rapid exploration of configuration space with diffusion-map-directed molecular dynamics. J Phys Chem B. 2013; 117:12769–12776. [PubMed: 23865517]
12. Djurdjevac N, Sarich M, Schütte C. Estimating the eigenvalue error of Markov State Models, Multiscale Model. Simul. 2012; 10:61–81.
13. Prinz J-H, Chodera J, No F. Spectral rate theory for two-state kinetics. Phys Rev X. 2014; 4:011020.
14•. Noé F, Nüske F. A variational approach to modeling slow processes in stochastic dynamical systems, SIAM Multiscale Model. Simul. 2013; 11:635–655. Derives a variational principle for molecular kinetics. Lays the mathematical groundwork for approximating molecular kinetics with basis sets different from direct cluster partitions and establishes a link between molecular kinetics and quantum mechanics.
15. Nueske F, Keller B, Mey A, Perez-Hernandez G, Noé F. Variational approach to molecular kinetics. J Chem Theory Comput. 2014 in press.
16. Gu C, Chang HW, Maibaum L, Pande VS, Carlsson GE, Guibas LJ. Building Markov state models with solvent dynamics. BMC Bioinformatics. 2013; 14(Suppl 2):S8. [PubMed: 23368418]

17. McGibbon RT, Pande VS. Learning kinetic distance metrics for Markov state models of protein conformational dynamics. J Chem Theor Comput. 2013; 9:2900–2906.

18. Teo I, Schulten K. A computational kinetic model of diffusion for molecular systems. J Chem Phys. 2013; 139:121929. [PubMed: 24089741]

19. Molgedey L, Schuster HG. Separation of a mixture of independent signals using time delayed correlations. Phys Rev Lett. 1994; 72:3634–3637. [PubMed: 10056251]

20•. Schwantes C, Pande V. Improvements in markov state model construction reveal many non-native interactions in the folding of ntl9. J Chem Theory Comput. 2013; 9:2000–2009. Introduces the time-structured independent component analysis as a generalization to PCA and a way to approximate the slow subspace of a set of input coordinates. Provides an excellent dimension reduction technique to precede MSM construction. See also Perez-Hernandez et al, 2013. [PubMed: 23750122]

21•. Perez-Hernandez G, Paul F, Giorgino T, de Fabritiis G, Noé F. Identification of slow molecular order parameters for Markov model construction. J Chem Phys. 2013; 139:015102. Introduces the time-lagged independent component analysis as an optimal way to approximate the slow subspace of a set of input coordinates. Provides an excellent dimension reduction technique to precede MSM construction. See also Schwantes et al, 2013. [PubMed: 23822324]

22. Buchete N, Hummer G. Coarse Master Equations for Peptide Folding Dynamics. J Phys Chem B. 2008; 112:6057–6069. [PubMed: 18232681]

23. Schütte C, Noé F, Lu J, Sarich M, Vanden-Eijnden E. Markov state models based on milestoning. J Chem Phys. 2011; 134:204105. [PubMed: 21639422]

24. Prinz JH, Chodera J, Pande V, Swope W, Smith J, Noé F. Optimal use of data in parallel tempering simulations for the construction of discrete-state markov models of biomolecular dynamics. J Chem Phys. 2011; 134:244108. [PubMed: 21721613]

25. Beauchamp K, Bowman G, Lane T, Maibaum L, Haque I, Pande V. MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. J Chem Theo Comput. 2011; 7(10):3412–3419.10.1021/ct200463m

26. Bowman G, Beauchamp K, Boxer G, Pande V. Progress and challenges in the automated construction of Markov state models for full protein systems. J Chem Phys. 2009; 131(12): 124101+.10.1063/1.3216567 [PubMed: 19791846]

27. Noé F. Probability Distributions of Molecular Observables computed from Markov Models. J Chem Phys. 2008; 128:244103. [PubMed: 18601313]

28. Trendelkamp-Schroer B, Noé F. Efficient Bayesian estimation of Markov model transition matrices with given stationary distribution. J Phys Chem. 2013; 138:164113.

29. Buchete N, Hummer G. Peptide folding kinetics from replica exchange molecular dynamics. Phys Rev E. 2008; 77:030902.

30. Chodera JD, Swope WC, Noé F, Prinz JH, Pande VS. Dynamical reweighting: Improved estimates of dynamical properties from simulations at multiple temperatures. J Phys Chem. 2011; 134:244107.

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

32. Deuflhard, P.; Weber, M. Robust perron cluster analysis in conformation dynamics. In: Dellnitz, M.; Kirkland, S.; Neumann, M.; Schütte, C., editors. Linear Algebra Appl. Vol. 398C. Vol. 2005. Elsevier; New York: 2005. p. 161-184.

33. Bowman G. Improved coarse-graining of markov state models via explicit consideration of statistical uncertainty. J Chem Phys. 2012; 137:134111. [PubMed: 23039589]

34. Yao Y, Cui RZ, Bowman GR, Silva DA, Sun J, Huang X. Hierarchical nyström methods for constructing Markov state models for conformational dynamics. J Chem Phys. 2013; 138:174106. [PubMed: 23656113]

35. Beauchamp KA, McGibbon R, Lin YS, Pande VS. Simple few-state models reveal hidden complexity in protein folding. Proc Natl Acad Sci USA. 2012; 109:17807–17813. [PubMed: 22778442]

36. Bowman GR, Meng L, Huang X. Quantitative comparison of alternative methods for coarse-graining biological networks. J Chem Phys. 2013; 139:121905. [PubMed: 24089717]

37. Jain A, Stock G. Identifying metastable states of folding proteins. J Chem Theory Comput. 2012; 8:3810.
38. Senne M, Trendelkamp-Schroer B, Mey A, Schütte C, Noé F. EMMA – A software package for Markov model building and analysis. J Chem Theory Comput. 2012; 8:2223–2238.
39. Cronkite-Ratcliff B, Pande V. MSMExplorer: visualizing Markov state models for biomolecule folding simulations. Bioinformatics. 2013; 29:950–952. [PubMed: 23365411]
40. Harvey MJ, Giupponi G, De Fabritiis G. ACEMD: Accelerated molecular dynamics simulations in the microseconds timescale. J Chem Theory Comput. 2009; 5:1632.
41. Eastman P, Friedrichs MS, Chodera JD, Radmer RJ, Bruns CM, Ku JP, Beauchamp KA, Lane TJ, Wang LP, Shukla D, Tye T, Houston M, Stich T, Klein C, Shirts MR, Pande VS. OpenMM 4: A reusable, extensible, hardware independent library for high performance molecular simulation. J Chem Theory Comput. 2013; 9:461. [PubMed: 23316124]
42. Salomon-Ferrer R, Götz AW, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. J Chem Theory Comput. 2013; 9:3878.
43. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. Science. 2011; 334:517. [PubMed: 22034434]
44. De Sancho D, Mittal J, Best RB. Folding kinetics and unfolded state dynamics of the GB1 hairpin from molecular simulation. J Chem Theory Comput. 2013; 9:1743.
45. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl T. Constructing the full ensemble of folding pathways from short off-equilibrium simulations. Proc Natl Acad Sci USA. 2009; 106:19011–19016. [PubMed: 19887634]
46. Voelz VA, Jäger M, Zhu L, Yao S, Bakajin O, Weiss S, Lapidus LJ, Pande VS. Markov state models of millisecond folder ACBP reveals new views of the folding reaction. Biophys J. 2011; 100(3):515a.10.1016/j.bpj.2010.12.3015
47. Voelz VA, Jäger M, Yao S, Chen Y, Zhu L, Waldauer SA, Bowman GR, Friedrichs M, Bakajin O, Lapidus LJ, Weiss S, Pande VS. Slow unfolded-state structuring of acyl-CoA binding protein revealed by simulation and experiment. J Am Chem Soc. 2012; 134:12565–12577. [PubMed: 22747188]
48. jie Deng N, Dai W, Levy RM. How kinetics within the unfolded state affects protein folding: An analysis based on Markov state models and an ultra-long MD trajectory. J Phys Chem B. 2013; 117:12787–12799. [PubMed: 23705683]
49. Piana S, Lindorff-Larsen K, Shaw DE. Atomic-level description of ubiquitin folding. Proc Natl Acad Sci USA. 2013; 110:59155920.
50. Weber JK, Pande VS. Protein folding is mechanically robust. Biophys J. 2012; 102:859–867. [PubMed: 22385857]
51. Jain A, Stock G. Hierarchical folding free energy landscape of HP35 revealed by most probably path clustering. J Phys Chem B. 2014 in press. 10.1021/jp410398a
52. Weber JK, Jack RL, Pande VS. Emergence of glass-like behavior in Markov state models of protein folding dynamics. J Am Chem Soc. 2013; 135:5501–5504. [PubMed: 23540906]
53. Bowman GR, Geissler PL. Extensive conformational heterogeneity within protein cores. J Phys Chem B. 2014 in press. 10.1021/jp4105823
54. Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE. How does a drug molecule find its binding site? J Am Chem Soc. 2011; 133:9181. [PubMed: 21545110]
55. Buch I, Giorgino T, de Fabritiis G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proc Natl Acad Sci USA. 2011; 108(25):10184–10189.10.1073/pnas.1103547108 [PubMed: 21646537]
56. Huang D, Caflisch A. The Free Energy Landscape of Small Molecule Unbinding. PLoS Comput Biol. 2011; 7(2):e1002002+.10.1371/journal.pcbi.1002002 [PubMed: 21390201]
57. Held M, Metzner P, Prinz JH, Noé F. Mechanisms of protein-ligand association and its modulation by protein mutations. Biophys J. 2010; 100:701. [PubMed: 21281585]
58. Weber M, Bujotzek A, Haag R. Quantifying the rebinding effect in multivalent chemical ligand-receptor systems. J Chem Phys. 2012; 137:054111. [PubMed: 22894336]

59. Qiao Q, Bowman GR, Huang X. Dynamics of an intrinsically disordered protein reveal metastable conformations that potentially seed aggregation. J Am Chem Soc. 2013; 135:16092. [PubMed: 24021023]

60. Lin YS, Bowman GR, B KA. Investigating how peptide length and a pathogenic mutation modify the structural ensemble of amyloid beta monomer. Biophy J. 2012; 102:315–324.

61. Bowman G, Geissler P. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. Proc Natl Acad Sci USA. 2012; 109:11681–11686. [PubMed: 22753506]

62•. Shukla D, Meng Y, Roux B, Pande VS. Activation pathway of Src kinase revealse intermediate states as targets for drug design. Nature Communications. 2014; 5:3397. A recent demonstration of the utility of Markov state models in studying functional native-state biomolecular dynamics, illustrating the potential for revealing allosterically targetable binding sites.

63•. Kohlhoff KJ, Shukla D, Lawrenz M, Bowman GR, Konerding DE, Belov D, Altman RB, Pande VS. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. Nature Chemistry. 2014; 5:3397. Two milliseconds of molecular dynamics simulations of a major drug-target G-protein-coupled receptor were aggregated using Markov state models providing an atomistic description of GPCR ligand-modulated activation pathways.

64•. Sadiq S, Noé F, de Fabritiis G. Kinetic characterization of the critical step in HIV-1 protease maturation. Proc Natl Acad Sci USA. 2012; 109:20449–20454. 0.5 milliseconds of molecular dynamics simulations of the HIV protease dimer were aggregated using Markov state models providing an atomistic description of the self-association pathway preceding the polyprotein cleavage required for HIV maturation. [PubMed: 23184967]

65. Da LT, Avila FP, Wang D, Huang X. A two-state model for the dynamics of pyrophosphate ion release in bacterial RNA polymerase. PLoS Comput Biol. 2013; 9:e1003020. [PubMed: 23592966]

66. Faelber K, Posor Y, Gao S, Held M, Roske Y, Schulze D, Haucke V, Noé F, Daumke O. Crystal structure of nucleotide-free dynamin. Nature. 2011; 477:556–560. [PubMed: 21927000]

67. Noé F, Doose S, Daidone I, Löllmann M, Chodera J, Sauer M, Smith J. Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. Proc Natl Acad Sci USA. 2011; 108:4822–4827. [PubMed: 21368203]

68. Keller BG, Prinz JH, Noé F. Markov models and dynamical fingerprints: Unraveling the complexity of molecular kinetics. Chem Phys. 2012; 396:92.

69. Xia J, jie Deng N, Levy RM. NMR relaxation in proteins with fast internal motions and slow conformational exchange: Model-free framework and Markov state simulations. J Phys Chem B. 2013; 117:6625–6634. [PubMed: 23638941]

70. Lindner B, Yi Z, Prinz JH, Smith J, Noé F. Dynamic Neutron Scattering from Conformational Dynamics I: Theory and Markov models. J Chem Phys. 2013; 139:175101. [PubMed: 24206334]

71. Zheng Y, Lindner B, Prinz JH, Noé F, Smith J. Dynamic Neutron Scattering from Conformational Dynamics II: Application using Molecular Dynamics Simulation and Markov modeling. J Chem Phys. 2013; 139:175102. [PubMed: 24206335]

72. Zhuang W, Cui RZ, Silva DA, Huang X. Simulating the T-jump-triggered unfolding dynamics of trpzip2 peptide and its time-resolved IR and two-dimensional IR signals using the Markov state model approach. J Phys Chem B. 2011; 115(18):5415–5424. [PubMed: 21388153]

73. Teo I, Schulten K. A molecular interpretation of 2D IR protein folding experiments with Markpv state models. Biophys J. 2014; 106 in press.

74. Song J, Gao F, Cui RZ, Shuang F, Liang W, Huang X, Zhuang W. Investigating the structural origin of trpzip2 temperature dependent unfolding fluorescence lineshape based on a markov state model simulation. J Phys Chem B. 2012; 116:12669–12676. [PubMed: 22994891]

75. Pirchi M, Ziv G, Riven I, Cohen S, Zohar N, Barak Y, Haran G. Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. Nature Comms. 2011; 2:493.

76. Stigler J, Ziegler F, Gieseke A, Gebhardt J, Rief M. The complex folding network of single calmodulin molecules. Science. 2011; 334:512–516. [PubMed: 22034433]

77. Keller BG, Kobitski A, Jäschke A, Nienhaus GU, Noé F. Complex RNA folding kinetics revealed by single molecule FRET and hidden Markov models. J Am Chem Soc. 2014; 136:4534–4543. [PubMed: 24568646]

78. Jämbeck JPM, Lyubartsev AP. Update to the General Amber Force Field for small solutes with an emphasis on free energies of hydration. J Phys Chem B. 2014 in press.

Highlights

* Markov state models (MSMs) are now widely used to study the long-time statistical dynamics of biomolecules

* Recent theoretical advances emphasize MSMs can closely approximate the true statistical dynamics

* MSM construction has been greatly simplified by new software packages

* Numerous applications have demonstrated the utility of MSMs for extracting insight and connecting with experiment

* Remaining challenges include fully automated adaptive MSM construction and balancing of statistical and systematic error

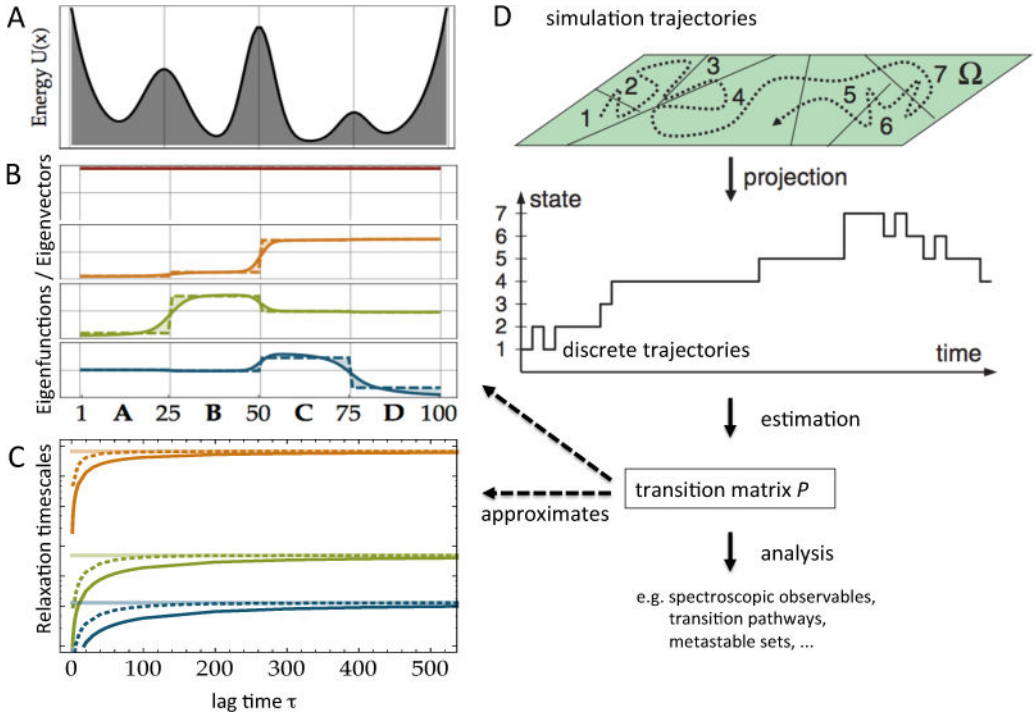

Figure 1.

Schematic illustrating important properties of MSMs and their construction. (A) A metastable four-well potential. (B) Eigenfunctions of the Markov operator (solid) and eigenfunctions approximated by the eigenvectors of a MSM using four states separating the four basins. (C) Implied timescales plot. The MSM timescales (solid) converge to the true relaxation timescales with increasing observation interval, or *lag time*, τ, but may do so only slowly. PMM timescales [3] (dashed) converge rapidly. (D) MSM construction. Using some state space discretization, the simulation trajectories are mapped to discrete trajectories. From these, the transition matrix P is estimated whose largest eigenvalues/eigenvectors approximate the true largest eigenvalues/eigenfunctions. P is then analyzed for molecular observables of interest.