Introduction and Overview of This Book

1

Gregory R. Bowman, Vijay S. Pande, and Frank Noé

Computer simulations are a powerful way of understanding molecular systems, especially those that are difficult to probe experimentally. However, to fully realize their potential, we need methods that can provide understanding, make a quantitative connection with experiment, and drive efficient simulations.

The main purpose of this book is to introduce Markov state models (MSMs) and demonstrate that they meet all three of these requirements. In short, MSMs are network models that provide a map of the free energy landscape that ultimately determines a molecule's structure and dynamics. These maps can be used to understand a system, predict experiments, or decide where to run new simulations to refine the map. Protein folding and function will often be used to illustrate the principles in this book as these problems have largely driven the development of MSMs; however, the methods are equally applicable to other molecular systems and possibly entirely different problems. Whether you are an experimentalist interested in understanding a bit of theory and how it could complement your work or a theorist seeking to understand the details of these methods, we hope this book will be useful to you.

This introduction provides a brief overview of the background leading to the development of MSMs, what MSMs are, and the contents of this book.

1.1 Background

Molecular systems are exquisitely sensitive to atomistic details—for example, a single point mutation can have dramatic effects on protein folding or function—a complete understanding would require atomically detailed models that capture both the thermodynamics and kinetics of the system of interest. There are many powerful experimental methods for probing the structure and dynamics of molecular systems but, currently, none can provide a complete understanding of a system.

Structural biologists have developed a range of methods for building atomically detailed models of proteins and other molecules; however, we are far more limited when it comes to dynamics. For example, when monitoring the relaxation of an ensemble of unfolded proteins back to the native state, one typically sees simple behavior that can be fit well by a single or double exponential. By Occam's razor, it is difficult to justify explaining such data with anything more complicated than a two- or three-state model. To push beyond these extremely coarse models, one has to begin making perturbations like mutations or trying

G.R. Bowman (✉)
Departments of Molecular & Cell Biology and Chemistry, University of California, Berkeley, CA 94720, USA
e-mail: gregoryrbowman@gmail.com

V.S. Pande
Department of Chemistry, Stanford University, Stanford, CA 94305, USA

F. Noé
Institut für Mathematik II, Freie Universität Berlin, Arnimallee 2-6, 14195 Berlin, Germany

G.R. Bowman et al. (eds.), *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Advances in Experimental Medicine and Biology 797,
DOI 10.1007/978-94-007-7606-7_1,

to incorporate other experimental data. However, the sensitivity of many molecular processes to atomistic changes makes interpreting the effects of perturbations difficult and combining different types of experimental data is also nontrivial—for example, how does one weight the relative contributions of two different types of data to a model? As a result, while there are certainly many opportunities in these directions, there is currently no clear path to building atomically detailed models for the entirety of a system from experimental data alone.

An alternative is to develop computer models that can complement experiment by providing an unambiguous description of a system's atomic motions. Ideally, these models could be validated by comparison to existing experimental data. One could then delve into the rich structural and kinetic information the model would provide to explain the origins of experimental results and generate hypotheses to guide the design of new experiments.

Atomistic molecular dynamics simulations are one powerful tool for achieving this vision. In these simulations, one iteratively evaluates the force each atom experiences due to the other atoms in the system, calculates where each atom will be some small timestep in the future, and finally updates their positions.

Unfortunately, it is extremely challenging to reach biologically relevant timescales in a molecular dynamics simulation, much less to obtain sufficient statistics to accurately characterize a system's behavior. The large forces and small length scales involved in such simulations necessitate a very small timestep—typically on the order of a femtosecond, or 10^{-15} seconds. One must then build up, about one femtosecond at a time, to the microseconds, milliseconds, and seconds timescales where many of the molecular processes of interest typically occur. Simulating a single millisecond on a typical desktop computer could easily take hundreds of years and is still essentially intractable with large computer clusters, though some progress has been made with distributed computing and specialized hardware.

Many advanced methods have been developed to overcome this gap between biological and simulation timescales but none is a magic bullet. For example, generalized ensemble methods—like replica exchange—allow a simulation to perform a random walk in temperature space. The hope is that at low temperatures the simulation will slowly explore the landscape of interest but that at high temperatures the system can easily jump to new regions of conformational space. Such methods are extremely powerful for small systems where energetic barriers dominate but can actually perform worse than conventional molecular dynamics for more complicated systems where entropic barriers dominate because these will become even more insurmountable at high temperatures. Coarse-graining can also provide reasonable speedups by reducing the number of pairwise interactions that must be calculated. However, there is always the danger that the degrees of freedom one coarse-grains out are actually important, in which case the coarse-grained simulation is of no value.

Even if these advanced methods could access arbitrarily long timescales, the issue of how to extract understanding from them would still remain. One cannot simply report what happened in a simulation because molecular processes like protein folding are inherently stochastic, so the exact sequence of events in one simulation is extremely unlikely to appear in a second trajectory.

One common analysis method is to project the free energy landscape onto order parameters but, once again, this is not a general solution. Projections of the free energy surface are really only valid if the order parameters chosen are truly reaction coordinates for the process of interest—i.e. they accurately reflect progression from reactants to products. In a very few cases, it is clear what the reaction coordinates are. For example, the alanine dipeptide only has two degrees of freedom, so it is perfectly legitimate to project the system's free energy landscape onto these order parameters. However, for processes like protein folding that occur in extremely high-dimensional spaces, finding a reaction coordinate is not so simple. Researchers often project free energy surfaces for proteins onto popular order parameters, like the number of native contacts or the RMSD to a known crystal structure, but one can find drastically different landscapes by choosing different

order parameters Therefore, these methods often do not provide clear and consistent models of molecular processes.

Clustering the conformations sampled with a set of simulations based on some geometric criterion—like the RMSD between conformations—is a less biased approach but is still not completely satisfactory. One major advantage of clustering is that it is less biased than projections since no reaction coordinate has to be assumed a priori. Furthermore, once the data has been clustered, many analyses can be performed easily. For example, comparison of the relative amounts of time spent in different clusters gives information about their relative free energies. One can also attempt to estimate the transition rates between clusters from the number of transitions observed between them and then begin looking at the most probable pathways between arbitrary start and end points. However, many important questions remain. For example, how many clusters are necessary and where, exactly, should the boundaries between them lie? Given two different clusterings, which one is better? Does a given clustering contain useful information? As will be discussed in more detail later, many problems can also arise when trying to estimate kinetic parameters from these models.

1.2 Markov State Models

A Markov model consists of a network of conformational states and a transition probability matrix describing the chances of jumping from one state to another in some small time interval. Many readers will recognize them as discrete time master equation models. Importantly, the states in an MSM are defined based on kinetic criteria rather than geometric criteria. Therefore, it is possible to accurately identify the boundaries between free energy basins and model dynamic processes like the relaxation to equilibrium.

A Markov model is a coarse-graining of a system's dynamics that reflects the underlying free energy landscape that determine's the system's structure and dynamics. Intuitively, it is often useful to think of the states in a Markov model as corresponding to free energy minima. However, as discussed in the next few chapters, this is not always necessarily true. Nonetheless, Markov models can provide important insights into a molecule because we have a much better intuition for states and rates (or, equivalently, transition probabilities) then we do for the large numbers of three dimensional structures generated by MD simulations.

The states and rates picture also provides a natural means to make a quantitative connection with experiments. For example, it is often possible to calculate an experimental observable (like the distance between two probes) for each state. A set of initial conditions can then be prepared by populating a subset of states and the relaxation to equilibrium can be modeled using the transition probabilities between states. This dynamics can be projected onto the experimental observable and the resulting signal can be compared to experiment.

Finally, adaptive sampling methods leverage Markov models to direct efficient simulations. In adaptive sampling, one iteratively run simulations, builds a Markov model, and then uses the current model to decide where to spawn new simulations to improve the model. Such methods can lead to tremendous improvements in computational efficiency compared to simply running one long simulation and waiting for it to gather statistics on the entirety of conformational space.

1.3 Outline of This Book

The remainder of this book can be divided into two sections. The first section, which includes Chaps. 2 through 7, presents the theoretical foundations of Markov state models. The second section, which includes Chaps. 8 through 10, focuses on a number of exciting applications of Markov models that serve to demonstrate the value of this approach. Below, we briefly review the contents of each chapter.

Chapter 2 provides a more thorough overview of Markov state models and how they are constructed. This discussion includes a description of the key steps for building an MSM and some of

the options available for each stage of the model building process. An important theme is that there is no single right way to perform many of these steps. Therefore, it is valuable to have some understanding of the tradeoffs between the available options.

Chapter 3 lays the theoretical foundation of MSMs. As indicated by the name, Markov models assume that the current discrete state of the system is sufficient to know the probabilities of jumping to any other state in the next time interval, without having to know the previous history. While the Markov assumption may be correct for the dynamics in the full-dimensional phase space, it cannot be *exactly* correct for the discrete partition of state space used for the MSM. The associated error, i.e. the difference of the MSM kinetics from the exact kinetics is a *discretization error*. Fortunately, we do not depend on a leap of faith when constructing MSMs. As a result of thorough mathematical work, especially during the last couple of years, the MSM *discretization error* is now well understood and can even be quantitatively bounded. Chapter 3 describes the nature of this error in the absence of additional statistical error, derives properties that a "good" partition of state space must fulfill, and suggests advanced approaches for MSM construction that go beyond simple state decomposition by clustering.

In practice, constructing MSMs progresses by defining a partitioning of conformational space into states and subsequently testing and possibly refining it. In order to do so, the MD trajectory data must be mapped on the discrete state space partitioning, and the MSM transition matrix must be estimated. Chapter 4 describes this step in detail and derives statistically optimal estimators for the transition matrix given a dataset and a state space partitioning. Subsequently, practical tests are described to assess the quality of the estimated MSM. It is these tests that will report on success or failure of the MSM to be a consistent kinetic model, and appropriate steps can be taken, e.g. by refining the state space partitioning used.

Since an MSM is estimated from a finite amount of MD trajectory data, the associated transition matrix and all properties computed from it will involve statistical uncertainty. Clearly, this is an issue for any model of the equilibrium or kinetic properties that is built from computer simulations—not just MSMs. Fortunately, for MSMs we now have a very complete theory that allows us to quantify these statistical errors from the number of transitions observed between the discrete sub-states. Chapter 5 attempts to give an overview of these methods and then goes into detail with Bayesian methods to sample the statistical uncertainties of transition matrices, and any quantity computed from them. Importantly, one can use estimates of uncertainties from an existing MSM to decide where to run new simulations in order to refine the model as efficiently as possible. Such methods are called adaptive sampling.

Chapter 6 gives an overview of some of the most useful analyses that can be performed with a valid Markov model. Three aspects are discussed and illustrated using a toy model of protein folding. First, we describe the significance of eigenvalues and eigenvectors of MSM transition matrices. Eigenvalues are related to the relaxation timescales of kinetic processes, and eigenvectors indicate the associated structural changes. Consequently, the eigenvectors associated with the slowest relaxation timescales can be used to find the metastable states of the molecular system studied. Secondly, the ability to associate relaxation timescales with structural changes via the eigenvalue-eigenvector duality is arguably one of the main advantages of MSMs over many other approaches to analyze MD simulation data. It permits one to uniquely assign structural transition events to experimentally measurable timescales, which makes MSMs a very valuable tool for quantitatively comparing simulation and experiment. We can go further and quantitatively predict the relaxation or correlation functions measured by kinetic experiments using three ingredients: the MSM eigenvalues, eigenvectors and the mean value of the spectroscopic observable for each discrete state. Chapter 6 describes the associated theory.

Finally, MSMs allow us to compute complex kinetic quantities that may not be directly experimentally accessible. One example is the ensemble of transition pathways and the transition state

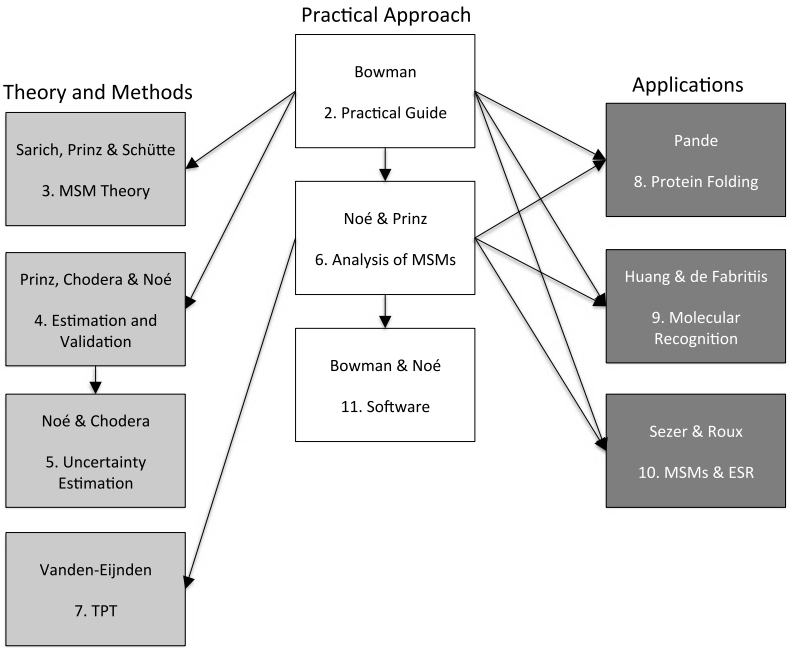

Fig. 1.1 Overview of the chapters in this book

ensemble. Given an MSM, both can be easily computed with transition path theory. Chapter 6 gives an introduction to transition path theory and illustrates it on the folding of the Pin WW peptide.

Chapter 7 is a theoretical chapter that goes into more detail on transition path theory. While transition path theory was originally derived for continuous Markov processes, this chapter focuses on its use in conjunction with MSMs and illustrates it using a simple example—a random walk in a two-dimensional maze. The basic mathematical quantities needed for computing transition pathways are defined and the equations for computing them from transition matrices are given. Furthermore, an approach to efficiently generate samples of reactive trajectories is introduced.

MSMs meet all three of the requirements laid out at the beginning of this chapter: providing understanding, making a quantitative connection with experiment, and driving efficient simulations. The subsequent application chapters will show that MSMs have already proven consistent with existing experimental data for a variety of molecular processes, allowed researchers to better understand—and sometimes even reinterpret—existing data, and led to new hypotheses that have been borne out in subsequent experiments.

Chapter 8 describes the application of Markov models to the protein folding problem and the new insights this has provided. This problem has two major components. First, how can we predict the structure of a protein from its sequence? And, second what is the sequence of events that allows a protein to fold? Besides showing how Markov models address both of these issues, this chapter will discuss how MSMs have allowed researchers to study much larger and slower systems than would otherwise be possible.

Chapter 9 summarizes recent work on using Markov models to understand how proteins bind small molecules. This application has important implications for drug design and our understanding of signaling within cells. It also presents an interesting methodological challenge because it is non-trivial to move from studying single-body problems (like protein folding, where all the atoms in the system of interest are covalently linked together) to multi-body problems.

Chapter 10 discusses how Markov models can be used to connect with new experimental techniques, like electron spin resonance. An impor-

tant emphasis is the impressive degree of agreement between simulation and experiment that one can achieve.

Since the construction, validation and analysis of MSMs is a nontrivial task, the existence of software that can support the user in these tasks is crucial. Chapter 11 provides an overview of existing MSM software packages and their current capabilities. Clearly, these packages are rapidly evolving and thus this chapter is just meant as a starting point. Therefore, links are provided to the manuals and tutorials of the software packages described.

Figure 1.1 provides an overview of the chapters of this book. For readers who decide not to follow the sequence of chapters in the book, we indicate the dependencies between chapters from the viewpoint of a practically oriented reader that is unfamiliar with MSMs. Theoretically inclined readers may start with the theory sections. Readers familiar with MSMs may read the book chapters in any sequence.