

# Network models for molecular kinetics and their initial applications to human health

Gregory R Bowman<sup>1</sup>, Xuhui Huang<sup>2,3</sup>, Vijay S Pande<sup>1,4</sup>

<sup>1</sup>Biophysics Program, Stanford University, Stanford, CA 94305, USA; <sup>2</sup>Department of Chemistry, The Hong Kong University of Science and Technology, Kowloon, Hong Kong; <sup>3</sup>Department of Bioengineering, Stanford University, Stanford, CA 94305, USA; <sup>4</sup>Department of Chemistry, Stanford University, Stanford, CA 94305, USA

Molecular kinetics underlies all biological phenomena and, like many other biological processes, may best be understood in terms of networks. These networks, called Markov state models (MSMs), are typically built from physical simulations. Thus, they are capable of quantitative prediction of experiments and can also provide an intuition for complex conformational changes. Their primary application has been to protein folding; however, these technologies and the insights they yield are transferable. For example, MSMs have already proved useful in understanding human diseases, such as protein misfolding and aggregation in Alzheimer's disease.

**Keywords:** Markov state models, molecular dynamics, simulations, protein folding, conformational change, Alzheimer's disease

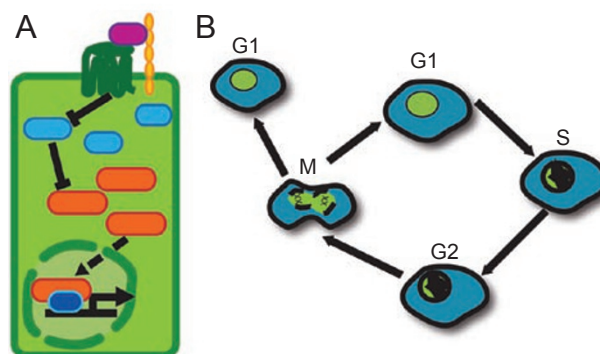
*Cell Research* (2010) 20:622-630. doi:10.1038/cr.2010.57; published online 27 April 2010

## Introduction

Much of today's biological research is motivated by the desire to understand higher-order processes more fully. For example, our desire to understand human development and disease helps us to drive research on cell biology. In turn, our desire to understand cells has motivated us to understand signaling pathways and gene networks. At each of these levels, networks — entities connected by arrows based on their relationships to one another — have proven to be a valuable way of representing knowledge (Figure 1).

Now our desire to understand the molecular underpinnings of biology and disease is motivating research into molecular kinetics. For example, it has recently been discovered that small oligomers of just a few A $\beta$  peptides may be the toxic elements in Alzheimer's disease [1]. Determining their structures could aid in designing drugs to prevent their formation, however, the structural heterogeneity of these oligomers makes accurate structural characterization with conventional methods difficult. Fortunately, computational modeling can capture both the dominant structures and dynamics of these molecules

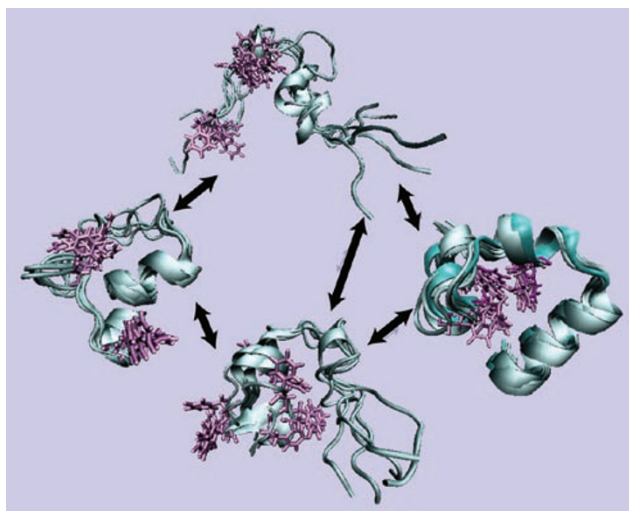
[2]. Another example is proteomics. Now that genomics has given us the sequence of the human genome, there is a push for high throughput structure prediction to obtain the structures of all the proteins encoded therein. Information-based methods have proved useful for small globular proteins, but physical models (which capture both thermodynamics and kinetics) are likely required to push to systems like membrane proteins for which less



**Figure 1** Example networks. **(A)** A signaling pathway with a cell (large green rectangle), proteins (colored ovals), a nucleus (dashed green circle), repression as blunted arrows, movement into the nucleus as a dashed arrow, and transcription as a solid black arrow. **(B)** The cell cycle with stages G1, S, G2, and M.

structural data is available. Since the majority of drugs on the market target membrane proteins, such information would again be valuable for designing therapeutics.

As in many other fields, networks are a valuable framework for representing knowledge of molecular kinetics. In particular, networks called Markov state models (MSMs) are proving to be a powerful means of understanding processes like protein folding and conformational changes [3-12]. The power of MSMs derives from the fact that they are essentially maps of the conformational space accessible to a system. That is, like a road map with roads labeled with speed limits and cities labeled with populations, MSMs give the probability that a protein or other molecule will be in a certain set of conformations (called a state or node) and describe where it can go next and how quickly. Figure 2 shows a portion of an MSM as an example. Examining these maps can give tremendous insight into processes like protein folding and could even suggest how to manipulate these processes with small molecules, mutations or other perturbations. Models with sufficient resolution can also yield quantitative agreement with, or even prediction of, experimental observables like folding rates and structures [4-6].



**Figure 2** An example MSM for the villin headpiece. Shown here are four clusters of conformations automatically identified by MSMBUILDER. Each cluster represents a state of the villin protein. Arrows indicate transitions between states, also identified by MSMBUILDER. The group or cluster representing the native state (right-most) was accurately identified. Its members match the crystal structure (shown in darker blue and magenta) with an average root mean square deviation (RMSD) of 1.8 Å. *Courtesy of Joy Ku and Gregory R. Bowman, reproduced with permission from Biomedical Computation Review.*

MSMs are typically constructed from simulation trajectories (i.e. series of conformations that were visited one after another in a physical simulation of a system, like a protein) [4, 5, 10, 12-14]. Because of the temporal relationship between conformations in a trajectory, it is possible to group conformations that can interconvert rapidly into states and then determine the connectivity between states by counting the number of times a simulation went from one state to another. By employing these kinetic definitions, one ensures that the system's dynamics can be modeled reasonably well by assuming stochastic transitions between states [3-7, 10, 11, 13, 15]. Thus, it is possible to perform analyses, such as identifying the most probable conformations at equilibrium. In addition, one can naturally vary the temporal and spatial resolution of an MSM by changing the definition of what it means to interconvert rapidly or slowly [4, 5, 15-17]. By choosing a long timescale cutoff, one can obtain humanly comprehensible models with just a few metastable (or long-lived) states that capture large conformational changes, like folding. Such coarse-grained models are useful for gaining an intuition for a system. With a short timescale cutoff, on the other hand, one can obtain a model with many states. By using such high-resolution models, one sacrifices ease of comprehension for more quantitative agreement with experiments [4, 5, 18].

To date, MSMs have mostly been used to understand phenomena like peptide and protein folding [4, 5, 13, 17, 19-23], RNA folding [6, 16], and conformational changes [8, 9, 24]. Having been validated by these studies, they are now being applied to important topics in human health. Examples include protein aggregation in Alzheimer's disease [2] and vesicle fusion [25], an important step in influenza infection.

The remainder of this review will be divided into four major sections. First, we review the application of MSMs to biomolecular folding, one of the driving problems behind the development of this technology. The next section focuses on the application of MSMs to human health, particularly protein misfolding diseases and influenza. This is followed by a review of MSM methodology and a discussion of efficiently capturing long timescales with short simulations and MSMs in the last two sections.

## Protein and RNA folding

Protein folding is one of the biological problems that has driven the development of MSMs. From a biophysical point of view, it is simply amazing how proteins collapse to specific structures so quickly given the astronomical number of possible conformations they can

adopt. From a biomedical point of view, understanding folding is an important first step towards understanding and preventing misfolding, as in Alzheimer's and Huntington's diseases [26]. Better models for protein folding could also allow more efficient structure prediction [27, 28], aiding high-throughput proteomics initiatives.

Unfortunately, understanding protein folding is notoriously difficult. Much experimental work has been done to characterize protein folding, but no experiment can capture the dynamics of an entire system with atomic resolution in order to provide a complete picture of the process. Computer simulation has the potential to fill this void, but reaching biologically relevant timescales with sufficient (atomic) resolution is extremely challenging [29], and even with an infinitely long simulation one would still need a way to extract meaningful information from it. MSMs can overcome these issues by performing a kinetic clustering of conformations to provide maps of a system's conformational space and extracting long timescale dynamics from many short simulations [4, 5, 10, 13, 30], similar to the way relay runners can cover large distances despite the fact that each runner only goes a little way. As explained in the section 'Reaching biologically relevant timescales with MSMs', MSMs may even be used to direct simulations to where they are needed most (via adaptive sampling) to efficiently explore the conformations that a molecule can adopt [6, 31-33].

Many of the initial applications of MSMs have been to small peptides due to the challenges of simulating larger proteins [12, 13, 17, 20-22, 34, 35]. For example, some of the fundamental methods were laid out by Singhal *et al.* [36] in an early work on a tryptophan zipper  $\beta$  hairpin. In another noteworthy example, Chodera *et al.* [13] developed and applied one of the first automated methods for building MSMs to a number of small systems. Buchete and Hummer [21] have also demonstrated how to tune MSMs to obtain models with varying degrees of coarse graining.

In one of the first forays into full protein systems, Jayachandran *et al.* [37] constructed an MSM for the villin headpiece. Villin was selected because it has a hydrophobic core and tertiary contacts, which make it a real protein rather than just a peptide, but is still extremely small and fast folding [38, 39]. Villin only has  $\sim 35$  residues and folds on the microsecond timescale, whereas most proteins have a hundred or more residues and fold on millisecond to second timescales; thus, villin is more amenable to computer simulations, which are generally no longer than tens of microseconds.

Villin was also the first target to be studied by automated methods [4], mainly the MSMBUILDER package [15], which is now freely available on the web ([https://](https://simtk.org/home/msmbuilder/)

[simtk.org/home/msmbuilder/](https://simtk.org/home/msmbuilder/)). This new MSM, a portion of which is depicted in Figure 2, was validated by showing that it is capable of reproducing the raw simulation data. Moreover, it also yields reasonable agreement with experimental measurements. For example, based on free energies calculated from the model, it was possible to predict the native structure to within 2 Å of the X-ray crystal structure (Figure 2). This was a significant advance since the native state of a system is determined by the lowest free energy state, yet most structure prediction algorithms are still based on identifying low-energy conformations due to the difficulties inherent in calculating the entropic component of free energies [40, 41]. Making predictions based on free energies will be especially important for more complex systems, like inherently disordered proteins, where the lowest energy state may not be the lowest free energy state. The longest timescales in our villin MSM were also consistent with experimentally measured folding rates [4]. Capturing these rates is critical because most experimental observables can be derived given complete knowledge of the relevant rates and a few conformations from each state.

Similar methods have also been used to study the Pin-WW domain [5]. This small protein consists of a three-stranded  $\beta$ -sheet and is another common model system for protein folding studies [42]. One important contribution of this work was a method for probing the equilibrium ensemble of folding pathways, which is based on transition path theory [43] and related methods [44]. This study also revealed numerous parallel folding pathways, as well as misregistered trapped states which slow the folding process.

MSMs have also been applied to the related problem of RNA folding [6, 16]. This research has been driven by recent developments demonstrating that RNA does more than convey information from DNA to proteins and, like proteins, often must fold into specific three-dimensional shapes to accomplish these roles [45]. The ribosome is a noteworthy example; however, it is far too large to sample exhaustively. Thus, simulations have tended to focus on small motifs, like RNA hairpins [46, 47]. RNA hairpins are composed of a double helix stem capped by a short loop. Despite their small size, they play many important roles. They are ubiquitous in known RNA structures, may serve as nucleation sites for the folding of larger molecules and play important roles in RNA-RNA and RNA-protein interactions [46, 48]. In the past, generalized ensemble simulations have been used to try to understand the thermodynamics of RNA hairpins [46, 49, 50]. More recent work has begun to apply MSMs to capture both their thermodynamics and kinetics, showing that current force fields are at least sufficient to identify a

hairpin as the native state and get the order of magnitude of the folding rate correct [6, 16].

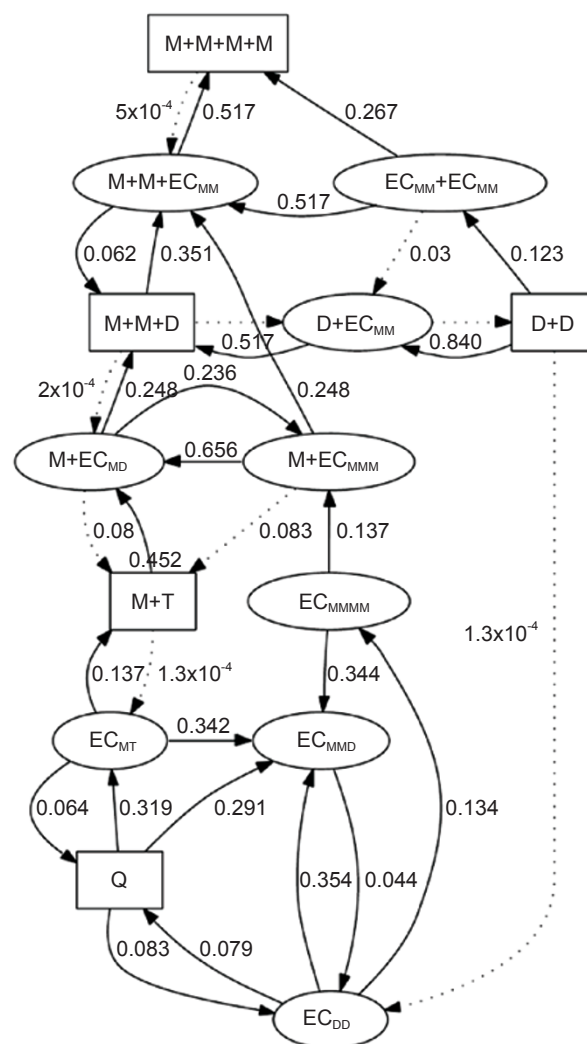
Other studies have also used MSMs to try to understand phenomena like downhill folding [51], conformational changes in membrane proteins [9], the effects of mutations [52], proton pumping [53] and human disease (to be discussed in the next section). However, most of the work on protein and RNA folding has focused on methods development and validation. In the coming years, it will be exciting to identify common features in protein free energy landscapes, explain exactly how proteins are able to fold so quickly and elucidate the origins of experimental observables. Capturing longer timescales (i.e. milliseconds to seconds and beyond) will be crucial for these future developments. An important step in this direction is a recent work that was able to capture the folding of NTL9 [54], an  $\alpha/\beta$  protein that folds on the millisecond timescale [55]. Further application to other biomolecules should also prove fruitful.

### MSMs and human health

Protein folding is not the only process that is difficult to characterize experimentally. Inherent disorder in processes like aggregation and vesicle fusion also makes them good candidates for study with MSMs. Again, experimental studies of these systems have limited resolution. Computer simulations can provide atomic-level descriptions of their dynamics and, with MSMs, reach the relevant timescales.

One particularly noteworthy example of applying MSMs to human disease is a recent work on A $\beta$  aggregation, which plays an important role in Alzheimer's disease [1, 2]. A $\beta$  is extremely difficult to work with experimentally; however, with computer simulations and MSMs, Kelley *et al.* [2] were able to obtain models with atomic resolution capable of capturing dynamics on tens of seconds timescales. To accomplish this feat, they first ran atomistic simulations of encounter complexes with varying numbers of monomers. An MSM was then constructed from this data using state definitions based on physical intuition (e.g. based on the number of monomers, dimers, trimers and tetramers). To achieve experimentally relevant concentrations of A $\beta$ , they then added a diffuse state and rates of transitioning between this state and encounter complexes were calculated using an analytic diffusion theory. This model, a schematic of which is shown in Figure 3, was found to give reasonable agreement with experimentally measured rates, a tremendous achievement given that standard simulations are about six orders of magnitude shorter than the relevant timescales. The MSM also gave important mechanistic

insight into A $\beta$  aggregation. In particular, the authors identified a reasonably populated C-terminal  $\beta$ -hairpin that was the main source of interactions between ag-



**Figure 3** Markovian model for A $\beta$  oligomerization. Our model was built using the different aggregation states as the Markov states; in a system with four chains, there are five such states: four monomers (MMMM), two monomers and one dimer (MMD), two dimers (DD), one monomer and one trimer (MT), and finally, one tetramer (Q). In addition, to include the effects of low concentration found experimentally, we discriminate EC states (in which states are close) from separated states. The rate-limiting steps in the aggregation process are shown as dotted lines. The numbers associated with the transitions are transition probabilities. The significant figures were determined from the uncertainties in the transition probabilities. Some transitions with very low probability have not been shown for the sake of clarity. Reprinted with permission from Kelley NW, Vishal V, Krafft GA, & Pande VS, *J Chem Phys*, 129, 214707, 2009. Copyright 2009, American Institute of Physics.

gregated peptides, leaving the N-termini exposed to the solvent. The exposure of the N-termini to solvent makes them more accessible for binding with other molecules, explaining why antibodies targeting the N-terminus tend to have higher binding affinities than those targeting the C-terminus.

In another example, MSMs were used to study vesicle fusion [25]. Vesicle fusion is important for processes like infection by enveloped viruses (e.g. in influenza); however, their large size and slow timescales make them inaccessible to standard computer simulations. Typical simulations have ~10 000 atoms and reach tens of nanoseconds timescales, but understanding vesicle fusion requires reaching hundreds of microseconds timescales for systems with over 500 000 atoms. Once again, a combination of simulations with MSMs was able to overcome these limitations, leading to important mechanistic insights. First, vesicles were found to fuse via two pathways: a fast pathway wherein two vesicles become connected by a short stalk and then rapidly fuse, and a slow pathway in which the stalk state transitions to a strongly metastable hemi-fused state before finally fusing. In a later work, the authors were even able to probe the dependence of the relative probabilities and rates of these pathways on the lipid composition of the vesicles involved [56].

Other human diseases are also excellent candidates for study with MSMs. For example, thousands of short simulations have already been performed for the Huntingtin protein [57] and the influenza hemagglutinin fusion peptide [58]. Thus, these systems are prime for study with adaptive methods [6, 31, 32], which allow one to build an initial MSM with available simulation data and then refine it by running new simulations from each state.

## MSM methodology

The dynamics of proteins and other molecules are governed by the system's underlying free energy landscape. Much like hikers on a natural landscape, proteins prefer to stay in the valleys of their landscape and only rarely (and generally slowly) cross over the barriers and peaks between valleys. MSMs are essentially maps of such landscapes [3-7, 10, 13, 15, 19]. However, whereas road maps have cities connected by roads labeled with speed-limits, MSMs have conformational states (sets of conformations in the same valley) connected by edges labeled by the probabilities of transitioning between them.

While MSMs may be visualized as networks, as in Figure 2, computers represent them as transition probability matrices ( $P$ ) [3, 7], where the entry in row  $i$  and column  $j$  ( $P_{ij}$ ) gives the probability of transitioning from

state  $i$  to state  $j$  in an interval called the lag time of the model ( $\tau$ ). The probability of being in any state at a particular time can be represented as a vector,  $v(t)$ . The time evolution of  $v(t)$  can then be calculated using

$$v(t + \tau) = v(t)P(\tau) \quad (1)$$

where  $P(\tau)$  is a transition matrix with lag time  $\tau$  and each multiplication advances the model through time by one lag time.

The eigenvalue/eigenvector spectrum of a transition probability matrix gives important thermodynamic and kinetic information [3, 7]. For example, the first eigenvalue always has a value of one and the corresponding eigenvector gives the equilibrium probabilities of all the states. Subsequent eigenvector/eigenvalue pairs give information about sets of states that interconvert on the same timescale and the rates at which these transitions occur. For example, a rate can be calculated from an eigenvalue using

$$k = \frac{-\tau}{\ln(\mu)} \quad (2)$$

where  $\mu$  is an eigenvalue,  $\tau$  is the lag time, and  $k$  is a rate. This equation comes from the equivalence between discrete time MSMs and continuous time master equations (see Refs [10] and [13] for details). Using these properties and a few representative conformations from each state, it is possible to compute experimental observables [4, 5]. One can even put error bars on these properties using Bayesian statistics [59, 60].

MSMs are straightforward to work with once you have a valid state definition (one yielding Markovian behavior). One simply has to assign conformations to these states and represent trajectories as a series of state assignments rather than as a series of conformations. One can then simply count the number of transitions between pairs of states and store these values as a transition count matrix ( $C$ ), where the entry in row  $i$  and column  $j$  ( $C_{ij}$ ) gives the number of observed transitions from state  $i$  to state  $j$ . One can then obtain a transition probability matrix ( $P$ ) by normalizing each row of the transition count matrix. From  $P$ , one can then calculate all the eigenvalues, eigenvectors and observables they desire.

The most challenging part of building MSMs is identifying a good state decomposition. A good state decomposition should group conformations that can interconvert rapidly into the same state because this implies that they are not separated by a significant free energy barrier. Conformations that cannot interconvert rapidly, however, should be separated into different states because they are

likely separated by a significant free energy barrier.

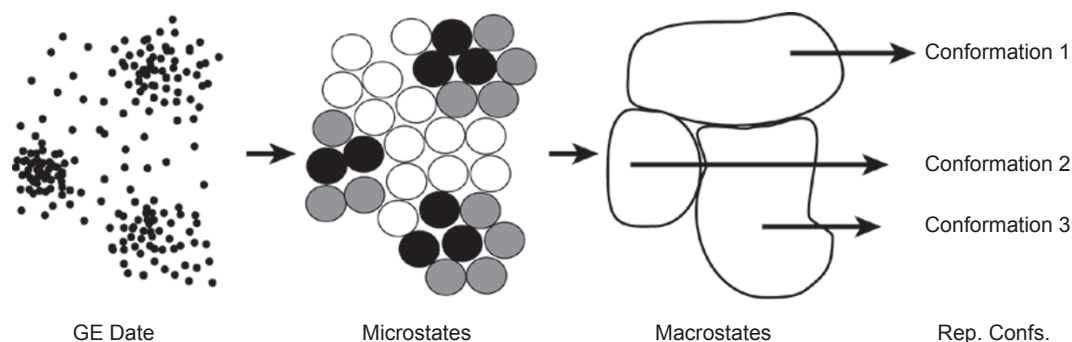
Practically speaking, it is neither possible nor desirable to determine the rate (or equivalently the probability) of transitioning between two conformations. Rather, one must consider the rates of transitioning between sets of conformations. To understand this, one can imagine trying to measure how long it takes to get home from work. You could measure the time it takes to get from the front doorway of your office to the front doorway of your home. However, it would likely be more appropriate to measure the time it took you to get from anywhere in your work place to any point in your home, regardless of whether you left work through the stairwell door and entered your home through the front door or left work through the main door and entered your home through the garage door. In the same spirit, measuring the transition rates between sets of kinetically equivalent conformations is more meaningful than measuring the transition rates between individual conformations.

In the case of MSMs, one often obtains initial sets of conformations by clustering them into microstates based on geometric criteria (Figure 4), with the objective of having conformations within a given microstate be so similar that their geometric similarity implies a kinetic similarity [10, 13, 15]. This initial decomposition may be ideal for making quantitative comparisons with experiments [4, 5, 18]. To gain an intuition for the system, kinetically related microstates can then be lumped into macrostates to ensure a direct connection to the underlying free energy landscape (e.g. identify various valleys and how quickly one can get from one to another) [10, 13, 15]. For example, in an earlier automated algorithm,

one first generated a set of microstates, lumped them into macrostates using a method called PCCA [61], and then iteratively broke the macrostates up into new microstates and re-lumped them until convergence [13]. A more recent automated algorithm (MSMBuilder), which is now freely available at <https://simtk.org/home/msmbuilder/>, avoids any iteration by using a different clustering algorithm to obtain finer clusters and then doing a single round of lumping using either PCCA or an improved method, called PCCA+ [62, 63]. This procedure is outlined schematically in Figure 4.

Regardless of how an MSM is built, it is necessary to choose a lag time and validate the final model. Checking that the implied timescales of the model level off as the lag time is increased is a first indication that a given state decomposition is reasonable and allows one to choose a lag time (the lag time at which the implied timescales first level off should be chosen) [64]. Intuitively, this equates to checking that the slowest rates (often on the order of hundreds of nanoseconds or greater) are invariant with respect to the interval at which you count transitions (often on the order of ten nanoseconds or less). Unfortunately, evaluating where the implied timescales of a model level off is extremely subjective and new methods are needed to replace this criterion. Some new methods employing information theory and Bayesian statistics point to how this may be done [59, 65]. Once a lag time has been selected, the transition probability matrix can be calculated and used to further validate the MSM by comparing its dynamics to the raw simulation data (the Chapman-Kolmogorov test) [4, 5, 13].

Another important question is how many macrostates



**Figure 4** Schematic representation of the steps required for building an MSM and obtaining representative conformations for each state. First, Generalized Ensemble (GE) data (or other data for that matter) represented by points are grouped into microstates represented by circles, with darker circles for more highly populated microstates. Kinetically related microstates are then lumped together into macrostates or metastable states, represented by amorphous shapes. Finally, representative conformations are obtained by extracting the most probable conformation from each macrostate. *Reprinted from Methods, 49, Bowman, GR, Huang, X, Pande, VS, Using generalized ensemble simulations and Markov state models to identify conformational states, 197-201, Copyright (2009), with permission from Elsevier.*

to construct when building a model to gain some intuition for a system. In the past, researchers have chosen a number of macrostates based on gaps in the implied timescales [10, 13, 15]. Doing so imposes a separation of timescales – fast intrastate transitions, slow interstate transitions – that should give rise to Markovian behavior, i.e., the next state a trajectory visits should be independent of its history because it should be able to enter a state and quickly lose memory of where it came from before transitioning to a new state. However, recent work has shown that there is often a continuum of timescales without any obvious gap. Thus, the number of macrostates has come to be seen as a tunable parameter that can be adjusted based on one's objectives [4, 5, 18]. For example, to gain an intuition for a system one may desire to lump a dataset into as few states as possible, while still preserving the Markov property and reasonable agreement with the raw simulation data.

### Reaching biologically relevant timescales with MSMs

An important advantage of MSMs is their ability to capture long timescale dynamics from many short simulations [4, 5, 9, 15, 21, 30]. Traditional approaches to molecular simulation require global equilibration, i.e., each simulation must be orders of magnitude longer than the slowest relaxation process so that every possible transition can be observed multiple times [66]. In contrast, MSMs only require local equilibration [6], i.e., each simulation only needs to be long enough to equilibrate within a subset of all the possible states. One can then obtain a global model by statistically stitching together many short, parallel simulations covering different parts of the conformational space, much like creating a quilt from many small patches or relay runners covering long distances.

Adaptive sampling algorithms take this reductionist approach a step further [6, 31, 32]. In adaptive sampling, one first obtains an initial model and then uses Bayesian statistics to calculate the contributions of each state to statistical uncertainty in some observable of interest, like the rate of the slowest process. A new round of simulations is then started from these states in order to increase the model precision as efficiently as possible. Thus, one avoids running more simulations where they are unnecessary while gathering more data where it can be of most use. Work with toy models has shown that performing multiple rounds of adaptive sampling can quickly refine a model [31]. Indeed, it has recently been shown that adaptive sampling can reduce the wall-clock time necessary to achieve a given model quality by a factor of  $N$ , where

$N$  is the number of parallel simulations run during each iteration [33]. Adaptive sampling can also reduce the total computer time necessary for a given model quality by a factor of two [33]. An exciting future direction will be to apply adaptive sampling to real systems of biological significance that are currently beyond the reach of computer simulations, like conformational changes in transcription elongation [67].

Even with adaptive sampling, however, MSMs are not without their limitations. For example, it is still unclear how to distinguish systematic errors from statistical uncertainty and correct for them too. Furthermore, while MSMs are excellent for describing free energy landscapes in terms of their basins, capturing transition states can also be quite important for understanding the mechanisms of conformational changes [46]. Recent work on the application of topological methods to understanding free energy landscapes shows how to capture such transition states [46, 68]. Indeed, progress on combining MSMs and topological methods to capture both free energy basins and transition states is already yielding insight into processes like RNA folding [16]. Moderate speedups may also be achieved by building MSMs from generalized ensemble simulations [6, 11, 15, 69], which perform a random walk in temperature to take advantage of broad sampling at high temperatures to improve mixing between states at lower temperatures [70, 71].

### Concluding Remarks

MSMs are a powerful means of mapping out the conformational space of both macromolecules and macromolecular complexes. While much of the literature on MSMs has focused on methods development and validation, they have already provided important insights into processes like protein folding, aggregation and vesicle fusion. Further progress in this direction will likely have important medical benefits, both by providing a deeper understanding of the molecular phenomena underlying higher-order biological processes (and especially human disease) and by allowing more effective drug and protein design. Methodological advances in the use of MSMs could also prove useful in systems biology and eventually find application in other fields relying on network representations, such as the study of gene networks and signaling pathways.

### Acknowledgments

We thank AN Bowman (Department of Developmental Biology, Stanford University) for help with Figure 1. This work was funded by NIH R01-GM062868, NIH U54 GM072970, and NSF EF-0623664. GRB was supported by the NSF GRFP.

## References

- Catalano SM, Dodson EC, Henze DA, *et al.* The role of amyloid-beta derived diffusible ligands (ADDLs) in Alzheimer's disease. *Curr Top Med Chem* 2006; **6**:597-608.
- Kelley NW, Vishal V, Krafft GA, Pande VS. Simulating oligomerization at experimental concentrations and long timescales: a Markov state model approach. *J Chem Phys* 2008; **129**:214707-214707-10.
- Schutte C. Conformational dynamics: modeling, theory, algorithm, and application to biomolecules. Department of Mathematics and Computer Science. Thesis, Freie Universitat Berlin, 1999.
- Bowman GR, Beauchamp KA, Boxer G, Pande VS. Progress and challenges in the automated construction of Markov state models for full protein systems. *J Chem Phys* 2009; **131**:124101.
- Noe F, Schutte C, Vanden-Eijnden E, Reich L, Weikl TR. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci USA* 2009; **106**:19011-19016.
- Huang X, Bowman GR, Bacallado S, Pande VS. Rapid equilibrium sampling initiated from nonequilibrium data. *Proc Natl Acad Sci USA* 2009; **106**:19765-19769.
- Schütte C, Fischer A, Huisinga W, Deuffhard P. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J Comput Phys* 1999; **151**:146-168.
- Yang S, Roux B. Src kinase conformational activation: thermodynamics, pathways, and mechanisms. *PLoS Comput Biol* 2008; **4**:e1000047.
- Yang S, Banavali NK, Roux B. Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proc Natl Acad Sci USA* 2009; **106**:3776-3781.
- Noe F, Fischer S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opin Struct Biol* 2008; **18**:154-162.
- Sriraman S, Kevrekidis LG, Hummer G. Coarse master equation from Bayesian analysis of replica molecular dynamics simulations. *J Phys Chem B* 2005; **109**:6479-6484.
- Gfeller D, De Los Rios P, Caflisch A, Rao F. Complex network analysis of free-energy landscapes. *Proc Natl Acad Sci USA* 2007; **104**:1817-1822.
- Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys* 2007; **126**:155101.
- Sriraman S, Kevrekidis IG, Hummer G. Coarse nonlinear dynamics and metastability of filling-emptying transitions: Water in carbon nanotubes. *Phys Rev Lett* 2005; **95**:130603.
- Bowman GR, Huang X, Pande VS. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* 2009; **49**:197-201.
- Huang X, Yao Y, Sun J, *et al.* Constructing multi-resolution Markov state models (MSMs) to elucidate RNA hairpin folding mechanisms. *Pac Symp Biocomput* 2010; **15**:228-239.
- Noe F, Horenko I, Schutte C, Smith JC. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J Chem Phys* 2007; **126**:155102.
- Sarich M, Noe F, Schutte C. On the approximation quality of Markov state models. *SIAM Multiscale Model Simul* 2010, in press.
- Rao F, Caflisch A. The protein folding network. *J Mol Biol* 2004; **342**:299-306.
- Schultheis V, Hirschberger T, Carstens H, Tavan P. Extracting Markov Models of peptide conformational dynamics from simulation data. *JCTC* 2005; **1**:515-526.
- Buchete NV, Hummer G. Coarse master equations for peptide folding dynamics. *J Phys Chem B* 2008; **112**:6057-6069.
- Elmer SP, Pande VS. Foldamer simulations: novel computational methods and applications to poly-phenylacetylene oligomers. *J Chem Phys* 2004; **121**:12760-12771.
- Andrec M, Felts AK, Gallicchio E, Levy RM. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc Natl Acad Sci USA* 2005; **102**:6801-6806.
- Pan AC, Roux B. Building Markov state models along pathways to determine free energies and rates of transitions. *J Chem Phys* 2008; **129**:064107.
- Kasson PM, Kelley NW, Singhal N, *et al.* Ensemble molecular dynamics yields submillisecond kinetics and intermediates of membrane fusion. *Proc Natl Acad Sci USA* 2006; **103**:11916-11921.
- Uversky VN. Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front Biosci* 2009; **14**:5188-5238.
- Bowman GR, Pande VS. The roles of entropy and kinetics in structure prediction. *PLoS One* 2009; **4**:e5840.
- Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. *Annu Rev Biophys* 2008; **37**:289-316.
- Klepeis JL, Lindorff-Larsen K, Dror RO, Shaw DE. Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol* 2009; **19**:120-127.
- Chodera JD, Swope WC, Pitera JW, Dill KA. Long-timescale protein folding dynamics from short-time molecular dynamics simulations. *Multi Mod Simul* 2006; **5**:1214-1226.
- Hinrichs NS, Pande VS. Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *J Chem Phys* 2007; **126**:244101.
- Roblitz S. Statistical error estimation and grid-free hierarchical refinement in conformation dynamics. Department of Mathematics and Computer Science. thesis, Freie Universitat Berlin 2008.
- Bowman GR, Ensign DL, Pande VS. Enhanced modeling via network theory: adaptive sampling of Markov state models. *JCTC* 2010; **6**:787-794.
- Swope WC, Pitera JW, Suits F, Pitman M, Eleftheriou M. Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and beta-hairpin peptide. *J Phys Chem B* 2004; **108**:6582-6594.
- Hummer G, Kevrekidis IG. Coarse molecular dynamics of a peptide fragment: free energy, kinetics, and long-time dynamics computations. *J Chem Phys* 2003; **118**:10762-10773.
- Singhal N, Snow CD, Pande VS. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J Chem Phys* 2004; **121**:415-425.
- Jayachandran G, Vishal V, Pande VS. Folding simulations of the villin headpiece in all-atom detail. *J Chem Phys* 2006; **124**:164902.



- 38 Chiu TK, Kubelka J, Herbst-Irmer R, *et al.* High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proc Natl Acad Sci USA* 2005; **102**:7517-7522.
- 39 Kubelka J, Chiu TK, Davies DR, Eaton WA, Hofrichter J. Sub-microsecond protein folding. *J Mol Biol* 2006; **359**:546-553.
- 40 Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 1997; **268**:209-225.
- 41 Bowman GR, Pande VS. Simulated tempering yields insight into the low-resolution Rosetta scoring functions. *Proteins* 2009; **74**:777-788.
- 42 Jager M, Nguyen H, Crane JC, Kelly JW, Gruebele M. The folding mechanism of a beta-sheet: the WW domain. *J Mol Biol* 2001; **311**:373-393.
- 43 Vanden Eijnden E. Toward a theory of transition paths. *J Stat Phys* 2006; **123**:503-523.
- 44 Berezhkovskii A, Hummer G, Szabo A. Reactive flux and folding pathways in network models of coarse-grained protein dynamics. *J Chem Phys* 2009; **130**:205102.
- 45 Chu VB, Herschlag D. Unwinding RNA's secrets: advances in the biology, physics, and modeling of complex RNAs. *Curr Opin Struct Biol* 2008; **18**:305-314.
- 46 Bowman GR, Huang X, Yao Y, *et al.* Structural insight into RNA hairpin folding intermediates. *J Am Chem Soc* 2008; **130**:9676-9678.
- 47 Koplin J, Mu Y, Richter C, Schwalbe H, Stock G. Structure and dynamics of an RNA tetraloop: a joint molecular dynamics and NMR study. *Structure* 2005; **13**:1255-1267.
- 48 Uhlenbeck OC. Tetraloops and RNA folding. *Nature* 1990; **346**:613-614.
- 49 Villa A, Widjajakusuma E, Stock G. Molecular dynamics simulation of the structure, dynamics, and thermostability of the RNA hairpins uCACGg and cUUCGg. *J Phys Chem B* 2008; **112**:134-142.
- 50 Garcia AE, Paschek D. Simulation of the pressure and temperature folding/unfolding equilibrium of a small RNA hairpin. *J Am Chem Soc* 2008; **130**:815-817.
- 51 Voelz VA, Luttmann E, Bowman GR, Pande VS. Probing the nanosecond dynamics of a designed three-stranded Beta-sheet with a massively parallel molecular dynamics simulation. *Int J Mol Sci* 2009; **10**:1013-1030.
- 52 Muff S, Caflisch A. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a beta-sheet miniprotein. *Proteins* 2008; **70**:1185-1195.
- 53 Kim YC, Wikstrom M, Hummer G. Kinetic gating of the proton pump in cytochrome c oxidase. *Proc Natl Acad Sci USA* 2009; **106**:13707-13712.
- 54 Voelz VA, Bowman GR, Beauchamp KA, Pande VS. Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9(1-39). *J Am Chem Soc* 2010; **132**:1526-1528.
- 55 Horng JC, Moroz V, Raleigh DP. Rapid cooperative two-state folding of a miniature alpha-beta protein and design of a thermostable variant. *J Mol Biol* 2003; **326**:1261-1270.
- 56 Kasson PM, Pande VS. Control of membrane fusion mechanism by lipid composition: predictions from ensemble molecular dynamics. *PLoS Comput Biol* 2007; **3**:e220.
- 57 Kelley NW, Huang X, Tam S, *et al.* The predicted structure of the headpiece of the Huntingtin protein and its implications on Huntingtin aggregation. *J Mol Biol* 2009; **388**:919-927.
- 58 Kasson PM, Pande VS. Predicting structure and dynamics of loosely-ordered protein complexes: influenza hemagglutinin fusion peptide. *Pac Symp Biocomput* 2007; **12**:40-50.
- 59 Bacallado S, Chodera JD, Pande V. Bayesian comparison of Markov models of molecular dynamics with detailed balance constraint. *J Chem Phys* 2009; **131**:045106.
- 60 Noe F. Probability distributions of molecular observables computed from Markov models. *J Chem Phys* 2008; **128**:244103.
- 61 Deuffhard P, Huisinga W, Fischer A, Schütte C. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin Alg Appl* 2000; **315**:39-59.
- 62 Deuffhard P, Weber M. Robust Perron cluster analysis in conformation dynamics. *Lin Alg Appl* 2005; **398**:161-184.
- 63 Weber M, Kube S. Robust Perron Cluster Analysis for various applications in computational life science. *Comput Life Sci Proc* 2005; **3695**:57-66.
- 64 Swope WC, Pitner JW, Suits F. Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J Phys Chem B* 2004; **108**:6571-6581.
- 65 Park S, Pande VS. Validation of Markov state models using Shannon's entropy. *J Chem Phys* 2006; **124**:054118.
- 66 Rao F, Caflisch A. Replica exchange molecular dynamics simulations of reversible folding. *J Chem Phys* 2003; **119**:4035-4042.
- 67 Wang D, Bushnell DA, Huang X, *et al.* Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution. *Science* 2009; **324**:1203-1206.
- 68 Yao Y, Sun J, Huang X, *et al.* Topological methods for exploring low-density states in biomolecular folding pathways. *J Chem Phys* 2009; **130**:144115.
- 69 Muff S, Caflisch A. ETNA: equilibrium transitions network and Arrhenius equation for extracting folding kinetics from REMD simulations. *J Phys Chem B* 2009; **113**:3218-3226.
- 70 Mitsutake A, Sugita Y, Okamoto Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* 2001; **60**:96-123.
- 71 Huang X, Bowman GR, Pande VS. Convergence of folding free energy landscapes via application of enhanced sampling methods in a distributed computing environment. *J Chem Phys* 2008; **128**:205106.