Methods 52 (2010) 99–105

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Review Article

Everything you wanted to know about Markov State Models but were afraid to ask

Vijay S. Pande [a,b,*], Kyle Beauchamp [a], Gregory R. Bowman [a]

[a] Program in Biophysics, Stanford University, USA
[b] Department of Chemistry, Stanford University, USA

ARTICLE INFO

Article history:
Available online 4 June 2010

Keywords:
Protein folding
Markov State Models
Computer simulation
Distributed computing
Molecular dynamics

ABSTRACT

Simulating protein folding has been a challenging problem for decades due to the long timescales involved (compared with what is possible to simulate) and the challenges of gaining insight from the complex nature of the resulting simulation data. Markov State Models (MSMs) present a means to tackle both of these challenges, yielding simulations on experimentally relevant timescales, statistical significance, and coarse grained representations that are readily humanly understandable. Here, we review this method with the intended audience of non-experts, in order to introduce the method to a broader audience. We review the motivations, methods, and caveats of MSMs, as well as some recent highlights of applications of the method. We conclude by discussing how this approach is part of a paradigm shift in how one uses simulations, away from anecdotal single-trajectory approaches to a more comprehensive statistical approach.

1. Introduction

Studying protein folding, either by experiment or simulation, is fraught with many challenges. In conjunction with its biological significance, these challenges make protein folding an important problem to study from a methodological perspective [1,2]. Moreover, approaches which have proven their utility in addressing folding related questions have often found broad applicability to wide range of other problems as well [3–5].

From the computational point of view, one of the primary challenges of protein folding simulation is the ability to reach experimentally relevant timescales, such as the millisecond to second timescale, with sufficiently detailed simulations in order to make quantitative predictions of experiment. However, often overlooked is the additional challenge that even if such simulations could be performed, one would need to have some means to analyze the resulting flood of data in a methodical and unbiased fashion. Finally, it is important to note that in many ways, these challenges are not unique to simulation, as single molecule experiments have similar challenges: one would like to ideally use as little data as possible to build models, long trajectories can at times be challenging (due to photobleaching or other technical challenges), and the analysis of the resulting data to gain insight is itself often a challenge.

Markov State Models (MSMs), kinetic models of the process understudy, typically constructed from detailed simulations such as Molecular Dynamics, have been proposed as a scheme to address these challenges. Moreover, this approach represents a paradigm shift in how one uses simulations, away from anecdotal single-trajectory approaches to a more comprehensive statistical approach. There have been many reviews of MSM methodology (e.g. see [6–8]), but these reviews have focused on theoretical and computational details, and are intended for theorists and practitioners of these methods. Here, our intention is to describe MSMs for primarily an experimentalist audience, with the primary goal of explaining in detail how MSMs work such that their strengths and weaknesses as applied to computer simulations of folding (and their predictions of experiment) can be understood. We stress that this is not meant to be a thorough review of the entire MSM field, but rather a basic "how to" guide to MSM construction for non-experts.

2. How does one build an MSM?

2.1. Goals

Before diving into the details of how one constructs an MSM, it is useful to remind the reader of the goals of MSM building. Here, we concentrate on three primary goals:

(1) The ability to quantitatively predict a broad array of experimental data.
(2) To use input data (either from simulations or experiment) as parsimoniously as possible.
(3) To build simplified models that are readily understood by human beings such that new insight can be gained; these models are not "cartoons" but rather coarse grained representations of the more detailed models employed for quantitative comparisons.

* Corresponding author at: Department of Chemistry, Stanford University, USA.
E-mail address: pande@stanford.edu (V.S. Pande).

1046-2023/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2010.06.002

2.2. Overall framework

The overall framework of an MSM (solving the Master equation) is, at its heart, similar to methods already familiar to biochemists and structural biologists for describing things like chemical reactions. Specifically, we wish to build a model with a series of N states and to parameterize the model with the rates between these states. However, unlike simple biochemical models, which typically have just a handful of states, MSMs often have many states, i.e. thousands to potentially millions. The rationale for having many states is that this allows one to construct a very high resolution model of the intrinsic dynamics as well as to more easily parameterize this model from relatively short molecular dynamics trajectories. That is, because the kinetic distance between adjacent states is small, short simulations are sufficient to observe transitions between them.

The specific challenges for building an MSM can be broken down into (1) how does one define states in a kinetically meaningful scheme and (2) how can one use this state decomposition in order to build a transition matrix in an efficient manner. Once this is performed, then the model is ready for both the goals of quantitative prediction of experiment as well as yielding qualitative insight into the mechanism at hand.

2.3. Initial data set

One must start with some initial data set. Below, we will concentrate on how MSMs are created from molecular dynamics simulation, but we stress that in principle, these methods are sufficiently general that they could be more broadly applied, for example also to other simulation methods interested in kinetics or thermodynamics. For molecular dynamics simulation, the initial data set could take several forms. In some cases, we are in an extremely data rich regime (e.g. many long trajectories that start unfolded and end folded [9]). In other cases, we are in a data poor regime, where there are a few trajectories that start unfolded (and perhaps some that start folded), but there are no (or few) single trajectories that traverse from the unfolded to folded states.

In the data rich regime, MSMs can help analyze the data set in a kinetically meaningful way. In the data poor regime, MSMs can also be used to direct future data collection (e.g. select the starting points of new MD simulations) in order to improve a model as efficiently as possible. Such methods have collectively been referred to as "adaptive sampling" methods. Since one must build an MSM before performing any additional adaptive sampling we will first describe more details of MSM building and then return to the subject of adaptive sampling.

Finally, we stress that many different types of simulations could be useful in creating the initial data set. One scheme is to "seed" MD simulations, i.e. start them in potentially relevant states *a priori*. For example, while thermodynamic sampling methods, such as Replica Exchange or Simulated Tempering [10–13], do not follow physical kinetics, they could be used to initially sample space for seeding (or in some cases to directly build an MSM [14]). Analogously, simplified force fields (e.g. coarse grained, implicit solvent, etc.) could be used to generate seeds, which would be followed by full-force field MD simulation.

One may be concerned that the errors in simpler methods (or the non-kinetically relevant aspects of thermodynamic sampling methods) would lead to seeds that are not useful. However, even a few useful seeds can be a significant speed up in the convergence of adaptive sampling (see below). Moreover, it is important to stress that the worst case scenario for "bad" seeds (i.e. seeds that do not lead to productive improvement of the MSM transition matrix) is that trajectories from these seeds do not make transitions between important states; while this uses computer time (and thus may affect the efficiency of the MSM creation process), this will not taint the calculation with any inaccurate data [15,16]. Finally, this seeding approach could also be used as a means to test the kinetic fidelity of simpler methods, so even results which lead to seeds which do not play a role could have scientifically important implications.

2.4. Building microstates

In order to build an MSM, i.e. to find a series of kinetically relevant states and the transition rates between them, we need to have a means to group structures in a kinetically meaningful manner. While structural clustering has been performed on simulation data for decades, previous methods have not explicitly considered kinetic properties [17,18].

MSM building techniques include kinetic information but begin with a traditional clustering method (e.g. k-means or k-centers) using a structural metric [8]. Considering the emphasis on kinetic clustering, this may sound strange; however, even though we want to define states by a kinetic criteria, it is important to keep in mind that one cannot define kinetics without some sense of geometric boundaries, i.e. one cannot define a rate between two objects without delineating where each starts and ends. Also, the kinetic relevance of this geometrical clustering can be tested (see Section 2.8). In fact, the structural resolution of this clustering can even make it appropriate for use in making quantitative connections with experiments [8,19,20].

This initial structural clustering is done to create many (e.g. 10,000–100,000 in protein examples so far) so-called "microstates" (where the initial criteria for clustering is the number of states). Due to the large number of microstates, conformations within the same microstate typically have RMSDs of no more than 2–3 Å [8,21]. This high degree of structural similarity implies a kinetic similarity, allowing for subsequent kinetic clustering of microstates into larger macrostates. Identifying kinetic relationships between microstates requires constructing a transition matrix.

One may be curious how the number of microstates would scale with the system size. This is currently known only for a range of system sizes from $\sim$10 to $\sim$100 residues (including unpublished results at the large scale). We have found so far that the number of states depends naturally directly on the complexity of the state space, but that the length of the protein is only a small part of this complexity. For example, comparing villin (36 residue alpha helical bundle protein), NTL9–39 (39 residue mini beta barrel), and lambda repressor (80 residue alpha helical protein), the number of microstates does not monotonically increase with protein length, with NTL9 having a much more complex space likely due to its beta sheet nature (requiring non-local many contacts). It remains to be seen how the number of microstates will scale with increasing chain length and is an important issue for future research.

2.5. Building a transition matrix

With the set of microstates, one can construct a microstate transition matrix. To do this, we take the MD data available (either from the initial data set or possibly also including data from adaptive sampling rounds) and assign each structure in the MD trajectory to a microstate. This "classification" step is comprised of comparing structures in MD trajectories to microstates, one by one, to find out which microstate is closest and then assigning the structure to that microstate. The result is a translation of the MD trajectory from a series of structures over time to a series of microstates over time.

Next, we then use these microstate trajectories to count how many transitions are seen between each pair of microstates i and j at some lag time τ, i.e. if a trajectory is at microstate i at time t,

then how many times did the simulation go to state j at time $t + \tau$? We call this set of quantities the count matrix $C_{ij}(\tau)$. Finally, one can estimate the probability of going from i to j in time τ (written as $p_{ij}(\tau)$) from the fraction of counts that started in i and went to j, compared to other possible states.

Given sufficient data, microstate transition matrices can be used to predict experimental observables or identify kinetically related microstates. However, by estimating the transition probability matrix from the counts, one can encounter problems when there are only a few counts. Discreteness effects and shot noise will lead to noise in the transition probabilities. At times, this does not matter, since some transitions are less important than others. However, one can improve on the estimate of the transition matrix by appealing to Bayesian techniques and including well-chosen prior probabilities [22–24]. For example, a prior (i.e. the imposition of a set of assumptions) that includes the effect of detailed balance can greatly enhance the effectiveness of data in the small count limit [23].

How much sampling does one need to build a reasonable converged MSM? One way to estimate this is to consider the number of transitions to calculate and the amount of simulation per transition. For a system with $N = 3 \times 10^4$ states (corresponding to a typical MSM for protein folding discussed herein), it is tempting to think that one would need N^2 or 10^9 states. However, it is important to note that not every microstate is connected to every other. Indeed, the matrix is very sparse and the connectivity does not appear to scale with N^2, but something on the order of $N \ln N$ or N. Thus, for 3×10^4 states, one may expect to have to calculate on the order of 10^5 transitions. Doing this with simple sampling would require on the order of 10^5 τ, where τ is the lag time of the MSM (on the order of 1–10 ns), yielding a total aggregate simulation time required on the 0.1–1 ms timescale; while this remains a challenging sampling problem, this degree of sampling has been demonstrated in recent cases for aggregate sampling from Folding@home.

However, we note that the estimate above is a bit simplistic. First, not all transitions need to be well-sampled, but rather only those which are uncertainty limiting. Indeed, adaptive methods take advantage of this fact and have estimated an increase of efficiency of 100× to 1000×, suggesting that 10s to 100s of microseconds of aggregate dynamics may be sufficient.

2.6. Coarse graining MSMs to gain human intuition

With a well-sampled microstate transition matrix, one has all the elements necessary for a functioning MSM. Indeed, with microstates (which are high resolution, well-defined states), the lag time (τ) is typically fairly small, e.g. on the ~10 ns timescale for MD simulations of protein folding. This results in a very high resolution model, which is especially useful for making quantitative comparisons to experiment (see below). However, in order to effectively use the resulting model (and especially to gain insight from it in a humanly understandable format), it is often useful to construct a coarse grained MSM.

This process consists of simplifying the microstate transition matrix into fewer states. This simplification can be done in a physically meaningful way by looking at timescales longer than the microstate lag time, i.e. 100 ns instead of 10 ns. At this slightly longer timescale, fewer states are kinetically relevant. Indeed, defining states in a "kinetically relevant" way requires that structures within a state can interconvert (i.e. kinetically reach each other) on timescales faster than the lag time. So, increasing the lag time means that states can get larger and more coarse grained. This makes it easier for the MSM to be humanly understandable, especially since the number of states can be arbitrarily small (and thus easier to comprehend), as long as one increases the lag time to match. We stress that such criteria above is necessary, but not sufficient, and thus we suggest that coarse grained models be tested (see below) or used primarily for visualization of the model.

But how can one perform this coarse graining (or "lumping") of states? While microstates are defined by a structural metric, this is where kinetic information must play a role. But where can we get kinetic information at this stage? The natural place is the microstate transition matrix, which directly encodes the kinetics between microstates. Typically, this is done via some sort of *spectral* clustering [25–27] of the microstate transition matrix, i.e. clustering methods which look at the eigenvalues and eigenvectors of the microstate transition matrix to identify kinetically similar states. These methods are well-developed and we will not go into further details, other than to stress that these methods allow one to define a coarse grained model at arbitrary resolution (high or low) depending on the goals for the model by lumping together kinetically related microstates.

2.7. Improving on the initial model: adaptive sampling

In the data rich regime, the previous steps can be sufficient for building an MSM. However, typically one is not in this regime and the previous steps result in an MSM which may qualitatively reflect the original dynamics, but many quantitative details could be improved. The natural way to improve the MSM is to perform more MD simulations to get better statistics, but how should one do this? One could choose to just continue the existing MD trajectories to make them longer, but this is not the most efficient scheme. Trajectories which have reached stable states (such as the native state or traps) will get stuck there for a very long time and thus further sampling these states will not greatly enhance the MSM.

Instead, it is natural to assign starting points for new simulations to optimize the ability of the MD data to improve the MSM. But how should one choose starting points? Adaptive sampling methods seek to answer this question by using the existing macrostate transition matrix (and more specifically the statistical uncertainty in its elements) to determine the ideal states to start additional simulations from. While we refer the reader to recent works for details [16,22,28,29], we will present the spirit of how this works below.

Recall that the count matrix ($C_{ij}(\tau)$) can tell us not just information about the transition probabilities ($p_{ij}(\tau)$), but also the uncertainty in these probabilities. Seeing 10^6 out of 2×10^6 counts reach a state is not the same as 1 out of 2. The low count regime will have a great deal of shot noise and will suffer from discreteness effects. Thus, the count matrix can also be used to predict the statistical error in the transition probability matrix, not just its values. By looking at which transitions contribute most to the statistical error of properties of interest (e.g. the folding rate), one can "on the fly" identify states which are limiting the accuracy of the model and start additional calculations from these states in order to improve the model. This approach is called adaptive sampling [15,22] since one adaptively modifies which simulations are run in order to optimize the data set used to build MSMs.

How well does this work? Previous tests [15,22] of adaptive sampling methods have shown a dramatic increase in efficiency, i.e. much less simulation data is needed when adaptive sampling is used. Indeed, recent work has quantitatively shown that adaptive sampling can reduce the time it takes to build a model by a factor of N, where N is the number of parallel simulations run during each round of sampling [30]. This makes physical sense, when one considers that traditional MD will get stuck in traps and over sample some areas and vastly under-sample others. Adaptive sampling opens the door for MSMs to be more than merely a way to

build a kinetic model from data, but rather to push the idea further, using the initial model ("knowing what you do not know") to speed convergence even further.

2.8. Validating the self-consistency of an MSM

Considering the challenges involved in constructing an MSM, it is important to first test that the MSM is self-consistent, i.e. agrees with the data used in its construction. These tests can be performed on both the fine-grained (microstate-based) and coarse grained (lumped) transition matrix. In particular, one challenge that often occurs is whether the lag time is sufficiently long to make the chosen state decomposition Markovian (i.e. do conformations within a state kinetically interconvert on timescales faster than the lag time and only make transitions to other states on slower timescales). There are several means to test this, including the Swope–Pitera eigenvalue test [31], information theoretic approaches [32], Chapman–Kolmogorov tests [19], and Bayesian Model selection approaches [23].

What can be done if the MSM fails the test? One natural approach is to run additional simulations in an adaptive manner in order to improve the model and pass the tests; this can be very efficient, but involves additional simulation which is not always practical. If one chooses not to use any addition simulation, the natural approaches are to either examine longer lag times (at the potential cost of some loss in temporal and/or spatial resolution) or construct a more fine-grained state decomposition (at the potential cost of some loss in statistical precision). We refer the reader to the works cited above for more details, but stress that such tests exist and are a critical step in MSM construction.

2.9. Connecting to experimental data

The quantitative prediction of experimental data is a primary goal of MSM creation. How is this done? Like any kinetic model involving states and rates, MSMs can take some initial conditions and report the state probabilities vs time. As in any simulation, in order to connect to experiment, one must be able to connect properties of the state to experimental observables. MSMs facilitate this by only requiring one to relate properties of the state to experiment (e.g. what is the IR spectra of the structures in this state?), then use the MSM to calculate the probability that each state is found at a given time t, and then perform a weighted average.

2.10. Visualization scheme

Finally, once one has done the above, it is natural to employ various visualization schemes to gain insight from the model. The most natural scheme is to examine the states and transitions in the coarse grained MSM for mechanistic properties. More recently, Transition Path Theory (TPT) [19,33,34] has been used to calculate the flux of relevant pathways, which aids in visualizing key mechanistic steps.

3. Examples of recent results

The past five years have seen a flurry of work on Markov State Models. Here we summarize some of the more recent advances. The discussion focuses on work from our own lab and collaborators, and so is by no means a comprehensive review. In particular, it is worth mentioning that works by Noe [7,19,24,35], Hummer [14,36], Roux [37–39], and Swope and Pitera [31,40] have also used MSMs or similar paradigms to study protein folding and dynamics.

Early attempts at MSM construction typically required one to build MSMs by hand, i.e. developing schemes to construct a state decomposition for a particular system [41–44]. The development of a fully-automated method [45] for MSM construction thus signaled a significant achievement. The automated algorithm described uses k-medoid clustering and iterative refinement to maximize a measure of state metastability. The success of their method is evident by several examples, including alanine dipeptide, Fs-peptide, and trpzip. In a particularly illustrative example, those authors performed a careful analysis of the alanine dipeptide results. Because the torsional preferences of this system are well understood, the authors could manually construct MSMs based on the Ramachandran plot for the peptide. A key result was that MSMs constructed by their automated algorithm performed equally well as the best hand-built models.

The past year has seen new MSM techniques come to fruition. The current generation of tools has proven capable of constructing accurate MSMs from some of the most extensive protein folding data sets available. In one case, using previously performed [9] simulations (545 trajectories of length up to 2 μs) of the double-norleucine mutant of the villin headpiece [46], the MSMBuilder package was used to construct a 10,000 state MSM. While working with such an extensive data sets can be challenging (due to the nature of clustering many conformations as well as the rare states which appear), the simple, automated algorithm in MSMBuilder produced a model that could quantitatively reproduce the original dynamics [8]. To demonstrate this, the authors compared several mock-observables (surrogate experimental quantities) as calculated two ways. First, they were calculated directly from the MD simulations, without using the MSM. Second, MSM dynamics were used to calculate the same observables, by propagating the initial state populations through time. State populations and ensemble average RMSD were among the observables computed. As an example, the folded microstate population as a function of time is shown (Fig. 1). In addition to reproducing the raw MD data, this model also showed success in making comparisons to experiment. First, the most populated MSM microstate was found to corre-

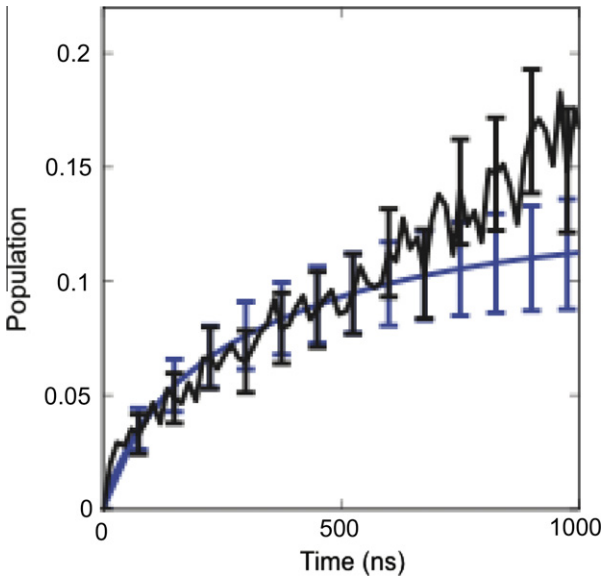

Fig. 1. The population of the folded state was monitored throughout the ensemble of 545 trajectories (black). The same population can be calculated by propagating the initial populations using MSM dynamics (blue). That both agree to within error suggests that the MSM accurately models the folding kinetics in this molecular dynamics data set [8].The model was constructed from simulations of a fast-folding villin headpiece mutant [9]. This figure was reproduced from Ref. [8].

spond to the experimentally determined crystal structure, with an RMSD of approximately 2 Å—thus, the combination of MD and MSMs can correctly predict the folded structure of villin (Fig. 2). Likewise, microsecond folding kinetics match the known experimentally observed timescales.

The past year has seen not just improvements in MSM construction, but also novel methods for interpreting MSMs. In a recent work [19], simulations of the pinWW protein were described using MSMs and Transition Path Theory (TPT). TPT provides a simple formalism for dissecting questions about reaction pathway and mechanism. TPT requires three things: an MSM, a starting state A, and an ending state B. A TPT analysis then decomposes the reaction A → B into individual pathways, along with the reactive flux for each pathway. The preferred mechanism of the reaction can be calculated by comparing reactive flux along pathways. An additional benefit of the TPT approach is that it allows representations that are both quantitatively accurate and visually compelling. The authors of that work used their TPT framework to quantitatively describe the folding pathways of the PinWW domain, a fast-folding beta protein that had previously been studied experimentally [47].

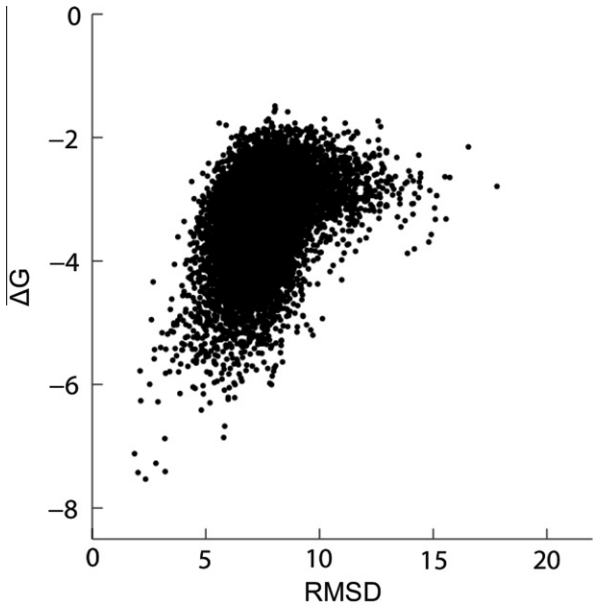

Fig. 2. The MSM-estimated equilibrium populations can be used to calculate the free energies of microstates. Here the lowest free energy states (most populated) are structurally similar to the experimentally determined crystal structure. The model was constructed from simulations of a fast-folding villin headpiece mutant [8]. This figure was reproduced from Ref. [8].

Combining the MSMBuilder approach to MSM construction with TPT methods, Voelz et al. sought to apply these methods to the NTL9 protein [21]. Compared to previously simulated proteins, NTL9 folds in a millisecond, which is as much as 1000 times slower. Its mixed alpha–beta topology makes it structurally more interesting than other protein folding model systems. Despite these challenges, Voelz et al. successfully built an MSM from implicit solvent simulations of NTL9. The model was able to recapitulate both the correct structure and the experimentally measured folding rate. TPT analysis (Fig. 3) revealed a heterogeneous folding mechanism, with a variety of misfolded and intermediate states. However, the native pairing of beta strands 1 and 2 occurred only for those states with pfold greater than 0.5, suggesting a rate-limiting role for the formation of this structural element. It is also interesting to note that the millisecond timescale of NTL9 is significantly longer than the individual simulation trajectories (each ~10 μs): this demonstrates that an MSM framework can help reach timescales longer than the individual MD trajectories.

4. Caveats of the MSM approach

As with any method, there are caveats to consider with taking an MSM approach. We detail the key caveats below, both to inform the reader interested in applying or evaluating these methods, as well as for those interested in advancing the existing methodology.

4.1. Sufficient sampling

While one of the primary goals of MSMs is to address the challenges involved with sampling, i.e. reaching sufficiently long

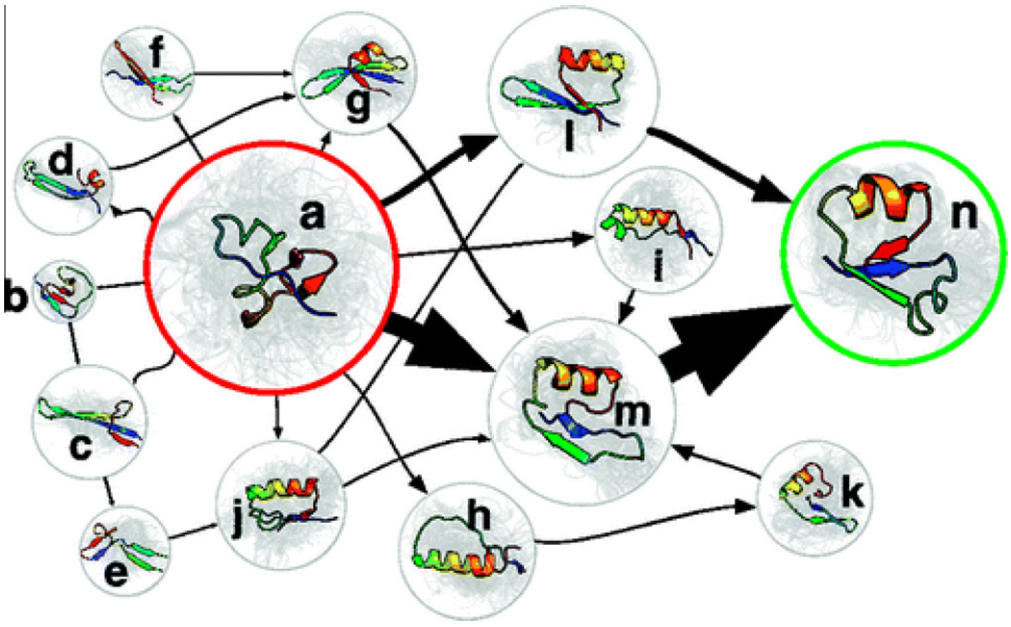

Fig. 3. The top 15 pathways of NTL9 folding were calculated using TPT. Each node represents a macrostate and is sized by the equilibrium free energy $-kt \log(P)$. The edges are sized by the folding flux through each segment, so the dominant pathways are depicted by larger arrows. This figure was reproduced from Ref. [21].

timescale phenomena with statistical significance, sampling is still a challenge. Indeed, while MSMs can greatly push the limits of what one can do with sampling, if an event occurs on very long timescales (beyond the longest timescales accessible by the MSM), then it will not be seen with simple approaches (although the combination of a state decomposition which can identify the relevant substates and adaptive sampling approaches could handle such problems efficiently). Moreover, we stress that the fact that MSMs seek to reflect true physical dynamics allows one to use physical reasoning to understand the limits of a given MSM.

With that said, it is important to stress that sampling with MSMs is of course considerably easier and more efficient than other methods. We consider it "easier" since MSMs are naturally amenable to parallel computation, so one could construct the necessary trajectories to parameterize an MSM on a simple cluster, rather than requiring an expensive supercomputer. Moreover, recent advances in using novel hardware, such as GPUs, opens the door to very long timescales, such as approximately 500 ns/day for a 36 residue protein (villin headpiece) with implicit solvent using OpenMM [48]. Thus, with even a medium sized cluster of 100 GPUs, one could simulate 50 μs/day of aggregate sampling or 5 ms of aggregate sampling over 100 days (a typical simulation run); thus, the combination of GPUs (e.g. running OpenMM) and MSM methods (e.g. using MSMBuilder) should allow many labs to study millisecond-scale protein folding phenomena. Finally, we stress that the discussion above has not included the effects of adaptive sampling, which would further increase the timescales one could reach, and thus greatly increase the efficiency and power of the calculation.

4.2. Accurate force fields

A general question for simulation is the accuracy of force fields. This is not an issue for MSMs in particular, although it is worth mentioning in any discussion of simulation caveats. Based on previous work cited above, it appears that the force fields have been sufficiently accurate for the systems examined so far, allowing for quantitative connection to experiment in many cases. Nevertheless, as one pushes these methods further and examines more complex systems, we may find new challenges and limitations of force fields.

However, in our opinion, the best way to tackle such problems is to ensure that we really can precisely understand the limits of the force fields; this comes from having sufficient sampling and statistical significance, which are the hallmarks of the MSM approach. Thus, we expect that even in situations where force fields fail, MSM approaches may prove valuable in understanding these limitations and potentially improving them.

4.3. Connections to experiment

Once one has constructed an MSM, it is natural to test the whole model (i.e. level of sampling combined with the force field accuracy) against experiment. However, a new challenge arises: how to make such quantitative connections? A detailed discussion of this area is beyond the scope of this work, but we mention this topic here nonetheless in order to highlight its significance. We stress that the ideal connection to experiments measure something as close to the experimental observable as possible, rather than trying to make a connection to the interpretation of those observables. While this direct connection is challenging, it avoids additional layers which could lead to erroneous connections. For example, simulations of villin agree quantitatively with analogs of experimental observables (e.g. analogs to fluorescence), but show differences with the experimental interpretation of the folding rate when more detailed observables (such as RMSD to the native state) are examined [9].

Moreover, an MSM approach also should prove valuable in making such direct connections, since this process is now simplified: one merely must make connections between the structures in a state and a given experimental observable and then those state properties can be propagated to yield thermodynamic, bulk-time resolved, or even time resolved single molecule like distributions analogous to experimental observables.

4.4. Well-constructed state decomposition

Finally, the greatest methodological challenge and area for future work with MSM construction methods is refining and improving our ability to construct a kinetically relevant state decomposition. Recent advances in automated methods (e.g. see [45] and [7]) have made great strides in this area. However, we suspect that as one applies MSMs to more complex systems, new challenges will arise. For example, for very long proteins, one may need to consider the role of knots and topological effects, which may be challenging (although not impossible) to handle with RMSD-metric based approaches. Also, applying these methods outside of the protein realm may require new metrics for structural similarity, especially when there are degrees of freedom which are fluid and their specific labels do not matter, such as structurally relevant water or lipid molecules. Luckily, methods established for checking the consistency of an MSM (described above) would reveal such problems and signal the need for further development.

5. Conclusions

We have walked the reader through the fundamentals of MSM construction, with an emphasis on discussing physical intuition over mathematical formalism. For additional details, we recommend recent reviews and research papers cited above. In a nutshell, MSMs represent a shift in how one thinks of computer simulation. Instead of creating a toy system, letting it go for a single or few long trajectories, and then reporting the (likely anecdotal) results, MSMs take a statistical approach. Indeed, the goal of simulations here is *not* creating trajectories which can be used to create mechanistic accounts, but rather first and foremost model building, where statistical methods are used to most efficiently build the best model for the kinetics possible, given the data at hand. We expect that in the end, it may be that the most significant contribution of the MSM approach may not be details of model building, but rather the shift in how one uses and conceptualizes simulations.

Acknowledgements

The authors thank NIH (R01-GM062868) and NSF (EF-0623664) for the funding of this work.

References

[1] K.A. Dill et al., Annu. Rev. Biophys. 37 (2008) 289–316.
[2] C.D. Snow et al., Annu. Rev. Biophys. Biomol. Struct. 34 (2005) 43–69.
[3] P.M. Kasson et al., Proc. Natl. Acad. Sci. USA 103 (32) (2006) 11916–11921.
[4] N.W. Kelley et al., J. Mol. Biol. 388 (5) (2009) 919–927.
[5] N.W. Kelley et al., J. Chem. Phys. 129 (21) (2008) 214707.
[6] G.R. Bowman, X. Huang, V.S. Pande, Methods 49 (2) (2009) 197–201.
[7] F. Noe, S. Fischer, Curr. Opin. Struct. Biol. 18 (2) (2008) 154–162.
[8] G.R. Bowman et al., J. Chem. Phys. 131 (12) (2009) 124101.
[9] D.L. Ensign, P.M. Kasson, V.S. Pande, J. Mol. Biol. 374 (3) (2007) 806–816.
[10] S. Gnanakaran et al., Curr. Opin. Struct. Biol. 13 (2) (2003) 168–174.
[11] A. Mitsutake, Y. Okamoto, J. Chem. Phys. 121 (6) (2004) 2491–2504.
[12] A. Mitsutake, Y. Sugita, Y. Okamoto, Biopolymers 60 (2) (2001) 96–123.
[13] Y. Okamoto, J. Mol. Graph. Model. 22 (5) (2004) 425–439.
[14] N.V. Buchete, G. Hummer, Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 77 (Pt. 1) (2008) 030902.

[15] X. Huang et al., Proc. Natl. Acad. Sci. USA 106 (47) (2009) 19765–19769.
[16] X. Huang, G.R. Bowman, S. Bacallado, V.S. Pande, Proc. Natl. Acad. Sci. USA 106 (2009) 19765–19769.
[17] M.E. Karpen, D.J. Tobias, C.L. Brooks 3rd, Biochemistry 32 (2) (1993) 412–420.
[18] J. Shao et al., J. Chem. Theory Comput. 3 (6) (2007) 2312–2334.
[19] F. Noe et al., Proc. Natl. Acad. Sci. USA 106 (45) (2009) 19011–19016.
[20] M. Sarich, F. Noe, C. Schutte, SIAM Multiscale Model Simul, in press. Available from: <http://publications.mi.fu-berlin.de/771/1/MSM_submitted.pdf>.
[21] V.A. Voelz et al., J. Am. Chem. Soc. 132 (5) (2010) 1526–1528.
[22] N. Singhal, V.S. Pande, J. Chem. Phys. 123 (20) (2005) 204909.
[23] S. Bacallado, J.D. Chodera, V. Pande, J. Chem. Phys. 131 (4) (2009) 045106.
[24] F. Noe, J. Chem. Phys. 128 (24) (2008) 244103.
[25] C. Schütte et al., J. Comput. Phys. 151 (1999) 146–168.
[26] P. Deuflhard et al., Lin. Alg. Appl. 315 (2000) 39–59.
[27] P. Deuflhard, M. Weber, Lin. Alg. Appl. 398 (2005) 161–184.
[28] N. Singhal, V.S. Pande, J. Chem. Phys. 127 (2007) 244101.
[29] S. Röblitz, Statistical Error Estimation and Grid-free Hierarchical Refinement in Conformation Dynamics, in: Fachbereich Mathematik und Informatik, Freien Universit, Berlin, 2008.
[30] G. Bowman, V.S. Pande, J. Am. Chem. Soc. 6 (2010) 787–794.
[31] W.C. Swope, J.W. Pitera, F. Suits, J. Phys. Chem. B 108 (21) (2004) 6571–6581.
[32] S. Park, V.S. Pande, J. Chem. Phys. 124 (5) (2006) 054118.
[33] P. Metzner, C. Schutte, E. Vanden-Eijnden, J. Chem. Phys. 125 (8) (2006) 084110.
[34] A. Berezhkovskii, G. Hummer, A. Szabo, J. Chem. Phys. 130 (20) (2009) 205102.
[35] F. Noe et al., J. Chem. Phys. 126 (15) (2007) 155102.
[36] S. Sriraman, I.G. Kevrekidis, G. Hummer, J. Phys. Chem. B 109 (14) (2005) 6479–6484.
[37] A.C. Pan, B. Roux, J. Chem. Phys. 129 (6) (2008) 064107.
[38] D. Sezer, J.H. Freed, B. Roux, J. Phys. Chem. B 112 (35) (2008) 11014–11027.
[39] S. Yang, B. Roux, PLoS Comput. Biol. 4 (3) (2008) e1000047.
[40] W.Y. Yang et al., J. Mol. Biol. 336 (1) (2004) 241–251.
[41] S.P. Elmer, S. Park, V.S. Pande, J. Chem. Phys. 123 (11) (2005) 114903.
[42] S.P. Elmer, S. Park, V.S. Pande, J. Chem. Phys. 123 (11) (2005) 114902.
[43] G. Jayachandran, V. Vishal, V.S. Pande, J. Chem. Phys. 124 (16) (2006) 164902.
[44] N. Singhal, C.D. Snow, V.S. Pande, J. Chem. Phys. 121 (1) (2004) 415–425.
[45] J.D. Chodera et al., J. Chem. Phys. 126 (15) (2007) 155101.
[46] J. Kubelka, W.A. Eaton, J. Hofrichter, J. Mol. Biol. 329 (4) (2003) 625–630.
[47] M. Jager et al., J. Mol. Biol. 311 (2) (2001) 373–393.
[48] M.S. Friedrichs et al., J. Comput. Chem. 30 (6) (2009) 864–872.