# Understanding Protein Folding Using Markov State Models

8

Vijay S. Pande

## 8.1 Introduction

### 8.1.1 What Is Protein Folding?

Proteins play a central role in biology, acting as catalysts, sources of molecular recognition, structural elements, among many other roles. But before they can carryout these functions, proteins must first assemble themselves, or "fold," into their biologically functional or "native" state. As proteins are long chain molecules constituting of tens to thousands of amino acids, the fact that proteins fold to essentially a unique fold is a triumph of natural selection, considering the enormous amount of conformational entropy that folding must overcome.

This leads to the natural question: how does this process occur? Answering this question would be a resolution to one of the greatest outstanding questions in molecular biophysics. Moreover, as self-assembly is at the heart of many biological processes as well as the inspiration for modern nanotechnology, understanding how proteins fold could have an impact on many other fields. Finally, how proteins fold has emerged as a central part of the molecular mechanism of many diseases, such as Alzheimer's Disease or Huntington's Disease, where it is believed that proteins fold incorrectly—or misfold—as a critical part of the disease pathology.

V.S. Pande (✉)
Stanford University, Stanford, CA 94305, USA
e-mail: pande@stanford.edu

### 8.1.2 Why Simulate Protein Folding?

The biophysical and biomedical aspects of protein folding has highlighted many challenges in understanding folding. First, we have found that even small changes, such as a mutation of a single amino acid, can lead to changes in how a protein folds or whether it even folds at all.

Studying protein folding experimentally is fraught with many challenges. In particular, we wish to understand folding at the atomic scale. This is particularly challenging for experimental methods, given the stochastic and heterogeneous nature of an ensemble of proteins folding in an experiment.

Therefore, this challenge suggests an opportunity—*simulating* protein folding is a means to gain new insight into this challenging problem. Ideally, simulations can shed new insight into how proteins fold, suggest new hypotheses, as well as suggest new interpretations of experiments. When tightly combined with experiments, simulations have the hope to address the ultimate question of how proteins fold. Below, we present recent advances deriving from MSM approaches.

### 8.1.3 Challenges in Simulating Protein Folding

There are three primary challenges in any simulation. First, is our model for interatomic interactions (i.e. the "force field") sufficiently accurate to predict the behavior of the system of

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_8, 

interest. This has been a challenge for decades, but recent work has suggested that current force fields are sufficiently accurate for the quantitative prediction of a wide-range of bimolecular properties, but within certain known limitations [1] (see Fig. 8.2). Second, can one simulate the timescales relevant for the phenomena of interest? This has been a central challenge, since until recently, experimentally relevant timescales (microseconds to milliseconds) could not be reached with modern computer power using sufficiently accurate, atomically detailed models. Finally, a third challenge arises now that one can simulate long timescales with sufficiently accurate models: how can one use the resulting sea of data to gain some new insight? With the first two challenges now within reach for small, fast folding proteins, the third challenge of gaining new insight has come into the forefront.

As we discuss below, MSMs can aid in both the push for longer timescales as well as for the development of means to gain new insight from the resulting simulation data, even for more conventional simulation methods. Moreover, we will see that there are some potentially unique challenges associated with the construction of MSMs for protein folding. In particular, the unfolded state of a protein is huge (many conformations) and thus sampling it can be a challenge for the construction of an MSM. Also, the potential exponential growth in the number of relevant MSM microstates is also a potential challenge for MSM construction of protein folding, as simple arguments suggest that the number of structures grows exponentially length of the chain.

### 8.1.4 Unanswered Questions to Which MSMs Can Yield Insight

The end goal of a simulation of protein folding is the elucidation of the mechanism by which a protein folds, i.e. what are the steps a protein takes in assembling itself. There are several questions associated with this, including

1. **Does a protein fold in a single pathway or in many parallel paths?** This question is both relevant for the basic biophysics of folding, but also potentially relevant for the biochemistry of chaperonins, which catalyze the folding of some protein substrates. If folding occurs via a single, well-defined path, then catalysis could naturally take the form of the recognition of some well-defined transition state in the folding process. If folding occurs via multiple paths, then the resolution of the mechanism of catalysis is considerably more complex.
2. **Are there intermediates along the way to folding?** A common paradigm in the protein folding field is that simple proteins fold in a "two-state" manner, i.e. with just the unfolded and folded states and no intermediates in-between. Another way to rephrase this question is to study the separation of timescales between the slowest timescale (corresponding to folding) and the next slowest timescale; is this gap large compared to the folding time itself (for "two-state" system) or not? Simulations can help probe this hypothesis in a way that experiments cannot, due to their limitations of signal to noise of accumulating intermediates.
3. **Is the protein folding mechanism robust?** The entire discussion of a protein folding "mechanism" is hinged on the concept that such details are robust to subtle changes in the experimental environment (pH, temperature, co-solvents, etc.) as well as to variations in force fields used to simulate folding. Mechanistic properties which are robust have the hope to be comparable to experiment and free of variations caused by either experimental or computational variations. Moreover, the identification of non-robust properties is itself an important contribution as well. For example folding rates are robust (Fig. 8.2).

## 8.2 MSMs Have Allowed the Direct Simulation of Protein Folding

Given that the sampling at millisecond timescales has been possible for only two years (see Fig. 8.1), and analysis methodology is still immature, unambiguous scientific results learned from atomic simulation have thus far been modest. It will be

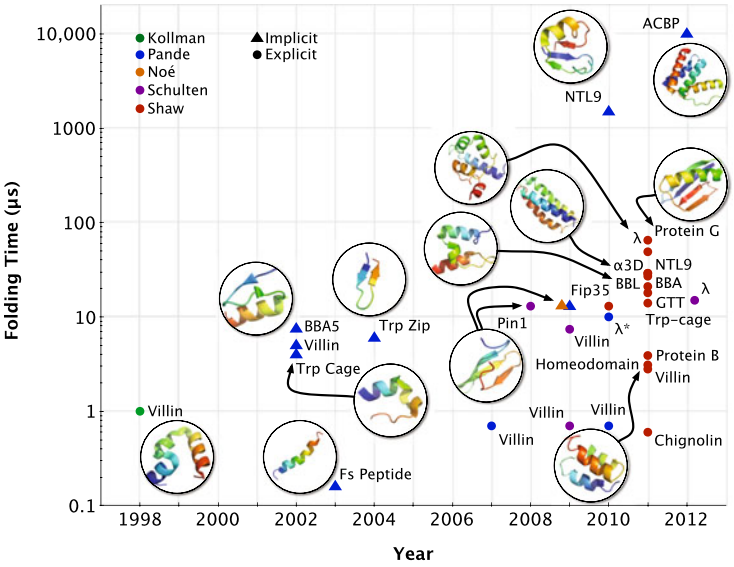


**Fig. 8.1** The folding times accessible by simulation have increased exponentially over the past decade. Shown are all protein folding simulations conducted using unbiased, all-atom MD in empirical force-fields reported in the literature. Some folding times for the same protein differ, due to various mutations. For lambda marked with a (*), the longest timescale seen in that simulation, which was not the folding time, occurred on the order of 10 ms

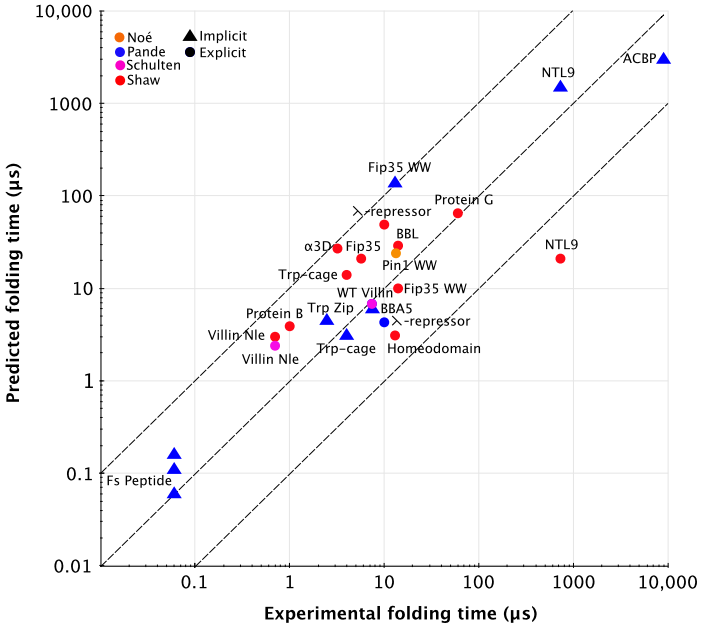


**Fig. 8.2** Comparison of predicted and experimentally measured folding times. *Central dashed line* is perfect agreement, *outside lines* are within one order of magnitude of perfect agreement. Given that experimental folding times can vary over more than an order of magnitude given different conditions (temperature, salt, pH, etc.), as well as uncertainties associated with measuring experimental and simulated folding times, an order of magnitude agreement is close to the upper limit of accuracy one might expect

a major challenge in the next five years to turn advances in sampling and accuracy into scientific insight about how proteins fold.

Despite this relative immaturity, MSMs built from atomistic simulation have already begun to influence our view of protein folding. Detailed comparisons to experiment have been performed for MSMs of many specific proteins, including villin [3], NTL9 [9], WW domains [6–8], lambda repressor [5], and ACBP [10]. Universally accepted generalities amongst these specific protein simulations have not yet emerged, though some have been suggested, for instance that folding kinetics might be hub-like [4].

## 8.3 What Have We Learned?

With the ability to simulate proteins which fold on long timescales (milliseconds) and for non-trivial sizes (approaching 100 amino acids), MSM

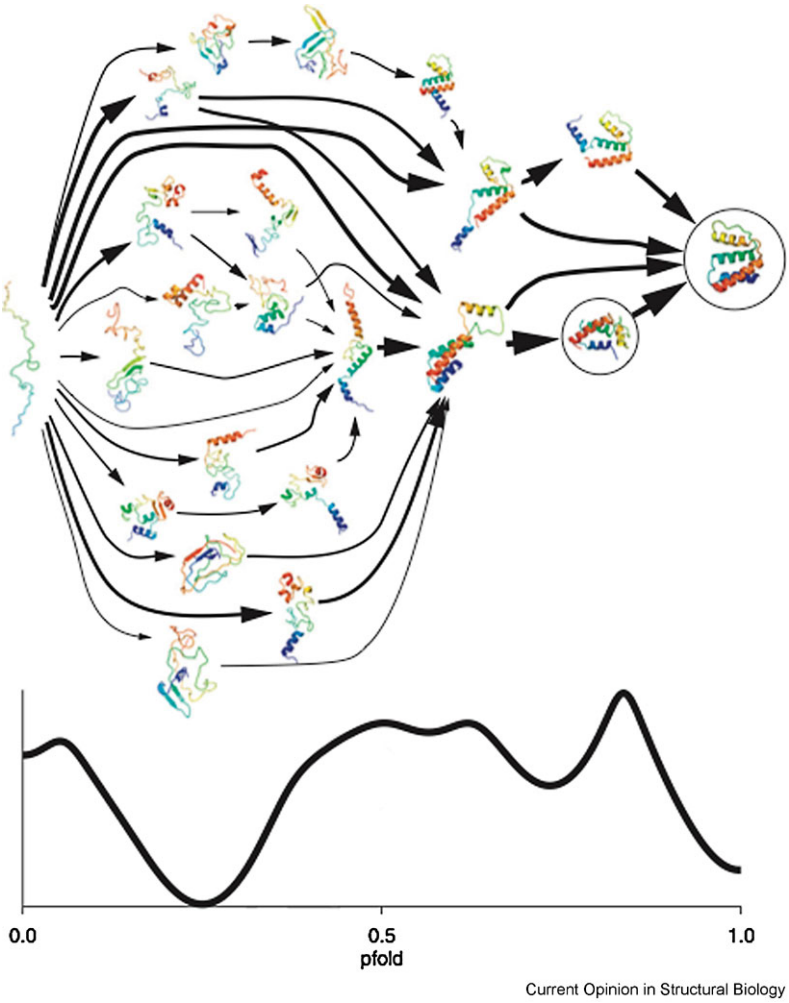


**Fig. 8.3** An MSM for the dynamics of ACBP, an 86 residue protein that folds on the 10 millisecond timescale. The size and long timescale (100 times longer than can be reached by traditional methods) make this calculation a landmark calculation in the simulation of protein folding. This diagram highlights the complexity of protein folding, showing the multiple paths that a protein can take to go from unfolded (*left*) to folded (*right*). The widths of the *arrows* denote how much flux each path carries

simulations have the hope to shed new insight into how do proteins fold. Below, we summarize three key results that have been seen so far.

### 8.3.1 Proteins Fold via Parallel Pathways Comprised of Metastable States

One of the principal results we have seen is that the mechanism of protein folding appears to be comprised of the interconversion of many metastable states. While an overall reaction may be dominated by a single slow timescale, leading to apparent "two-state" folding, more microscopically, folding looks much more detailed and complex. Where does this complexity go when examined experimentally? This complexity easily can be hidden when projected to a given reaction coordinate.

For example, consider Fig. 8.3, which shows an MSM for ACBP, which folds on the 10 millisecond timescale. While the MSM is complex, comprised of numerous states, ACBP appears to be only a three-state folder experimentally. However, when the MSM is projected to the fold reaction coordinate, we see that the MSM simplifies to look very much like a three-state folder [10]. This also opens the door to folding simulations helping predict new experiments which can more easily reveal this complexity.

### 8.3.2 These States Have Non-native Structural Elements: Register Shifts and Intramolecular Amyloids

With the illumination of these metastable states, one can interrogate the structural nature of these states to gain new insight into how proteins fold. One general property we find is that these states have an abundant degree of non-native structure. In particular, there are three forms of non-native structure which seems particularly common:

First, in beta sheet proteins, we often see states with register shifts. In these cases, the natural turn of a beta sheet is misplaced, leading to a different beta sheet structure. As turns can be formed in many places, sequences permit this reasonably easily in many cases [2].

Second, we often see elongated helices. In this case, a helix in a given intermediate state may be longer than in the native state. This is also natural given the commonality of helical propensity in amino acids, even in cases where the structure is not a helix natively.

Finally, and perhaps most strikingly, we have seen intramolecular amyloids,—cases where beta sheets form in alpha-helical proteins. This formation is not unlike the formation of intermolecular amyloids, where proteins spontaneously form beta sheet structures. Once a protein gets to be sufficiently long, we argue that it can act in the same fashion, *intramolecularly.*

### 8.3.3 The Connectivity of These States Suggest that the Native-State is a Kinetic Hub

Finally, how are these states "connected," i.e. that have non-zero conditional probabilities to go from one state to another? Addressing this question yields another aspect of the mechanism of protein folding. In MSM studies of protein folding, the native state has appeared to be a kinetic hub, i.e. there are many paths into the hub, compared with other states. This particular topology is common in other types of networks and suggest that the intrinsic kinetics of protein folding may have been evolutionarily optimized for kinetic properties including the kinetic network.

## 8.4 Next Challenges

MSM methods are sufficiently well developed to pursue many exciting applications. However, there is still a great deal of room for further methodological improvements. Here, we list a few of them.

1. **Longer timescales.** While MSMs have been able to simulate protein folding on the 10 millisecond timescale, proteins of interest for understanding how proteins fold can fold up to $1000\times$ longer. This could present new challenges for MSM sampling.
2. **Larger proteins.** Similarly, the largest proteins studied so far are just under 100 amino acids, while proteins of interest can be up to $2\times$ to $3\times$ longer in length. Larger proteins may present new challenges for MSM building due to the potential exponential growth in the configurational space involved.
3. **Better state decomposition.** One way to handle these challenges is to determine better methods for building states, allowing for fewer states to be used and thus enabling the ability to build more complex MSMs.

In the coming years, we expect that these challenges as well will be reached, yielding both new insights into how proteins fold but also new MSM methods which could be broadly applicable to many other applications as well.

## References

1. Beauchamp KA, Lin YS, Das R, Pande VS (2012) Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. J Chem Theory Comput 8(4):1409–1414
2. Beauchamp KA, McGibbon R, Lin YS, Pande VS (2012) Simple few-state models reveal hidden complexity in protein folding. Proc Natl Acad Sci USA 109(44):17,807–17,813
3. Bowman GR, Beauchamp KA, Boxer G, Pande VS (2009) Progress and challenges in the automated construction of Markov state models for full protein systems. J Chem Phys 131(12):124,101

4. Bowman GR, Pande VS (2010) Protein folded states are kinetic hubs. Proc Natl Acad Sci USA 107(24):10,890–10,895
5. Bowman GR, Voelz VA, Pande VS (2011) Atomistic folding simulations of the five-helix bundle protein (685). J Am Chem Soc 133(4):664–667
6. Lane TJ, Bowman GR, Beauchamp K, Voelz VA, Pande VS (2011) Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. J Am Chem Soc 133(45):18,413–18,419
7. Morcos F, Chatterjee S, McClendon CL, Brenner PR, Lopez-Rendon R, Zintsmaster J, Ercsey-Ravasz M, Sweet CR, Jacobson MP, Peng JW, Izaguirre JA (2010) Modeling conformational ensembles of slow functional motions in Pin1-WW. PLoS Comput Biol 6(12):e1001, 015
8. Noe F, Schutte C, Vanden-Eijnden E, Reich L, Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proc Natl Acad Sci USA 106(45):19,011–19,016
9. Voelz VA, Bowman GR, Beauchamp K, Pande VS (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). J Am Chem Soc 132(5):1526–1528
10. Voelz VA, Jager M, Yao S, Chen Y, Zhu L, Waldauer SA, Bowman GR, Friedrichs M, Bakajin O, Lapidus LJ, Weiss S, Pande VS (2012) Slow unfolded-state structuring in Acyl-CoA binding protein folding revealed by simulation and experiment. J Am Chem Soc 134(30):12,565–12,577