Emergence of Glass-like Behavior in Markov State Models of Protein Folding Dynamics

Jeffrey K. Weber†, Robert L. Jack‡, and Vijay S. Pande†*

†Department of Chemistry, Stanford University, Stanford, California 94305

‡Dept of Physics, University of Bath, Bath BA2 7AY, United Kingdom

*Corresponding Author. Email: pande@stanford.edu

ABSTRACT: The extent to which glass-like kinetics govern dynamics in protein folding has been heavily debated. Here, we address the subject with an application of space-time perturbation theory to the dynamics of protein folding Markov State Models (MSMs). Borrowing techniques from the *s*-ensemble method, we argue that distinct active and inactive phases exist for protein folding dynamics, and that kinetics for specific systems can fall into either dynamical regime. We do not, however, observe a true glass transition in any system studied. We go on to discuss how these inactive and active phases might relate to general protein folding properties.

As a grand challenge in statistical physics, the protein folding problem has been untangled to the degree that one can now claim some understanding of its components. The Levinthal "paradox" serves not as a true paradox, but rather as a reminder that proteins navigate a highly optimized free-energy landscape to find a unique native state. A rich body of literature exists concerning this conformational search over a rugged energy surface and its consequences for protein biology (1,2).

However, present challenges for the field of protein biophysics are not diminished in stature from those in years past. While the thermodynamics of the native state are well understood, knowledge about the kinetics involved in getting to this folded state is sparse. In recent years, systematically generated master equation-based models called Markov State Models (MSMs) have been successful in relating ensemble thermodynamics to a detailed description of kinetics (3-10). With each new insight that MSMs provide, however, new questions about the axioms of protein folding are raised. In particular, while some protein dynamics are well defined by a canonical two-state kinetics, MSMs demonstrate that many-state models are essential for describing dynamics in numerous systems (10). Additionally, few simple rules have emerged that relate properties like chain length and secondary structure to relaxation timescales (3-8, 10). Questions about connections between thermodynamics and folding kinetics thus remain open. What factors place a protein in one regime of kinetic behavior over another?

Central to this discussion are ideas about "glassiness" in protein dynamics. Early work in protein folding theory proposed a mapping between proteins and spin glasses, frustrated spin models that are largely applied to magnetic systems (11,12). The extent to which spin glass kinetics serve as a direct analogy to protein folding is debatable, but proteins do demonstrate some elements of glassy behavior in their folding dynamics. Notably, proteins and glasses share the characteristic of having rugged energy landscapes with deep valleys and potentially large barriers between states. Ideas of frustration in low energy states are prevalent in both classes of systems. Often, the division between random peptide heteropolymers and natural proteins is marked by a principle of least frustration. Evolved proteins exhibit single, highly optimized native states in which interactions are minimally frustrated; by contrast, random heteropolymers display more glass-like characteristics, folding into multiple, nearly-degenerate ground states that may lack structural correlation (11,12). Experimental attempts to observe glass transitions in protein folding systems have generated mixed results. Much evidence suggests that particular single- and multi-domain proteins exhibit kinetic traps typical of glass-like systems (13-15). In other single-domain systems, however, no evidence for a glass transition is found even at very low temperatures (16).

Given the success of MSMs in describing folding kinetics, a natural question arises from this discussion of glassiness in protein folding: how glassy are the dynamics of MSMs? Here, we study the kinetics of protein folding MSMs under the framework of non-equilibrium perturbation theory. Statistical mechanics in the space-time formalism has introduced the idea of different non-equilibrium phases, within which dynamical trajectories show distinct behavioral characteristics. In this Communication, we attempt to identify these behavioral regimes in protein folding trajectories gathered from MSMs. In particular, we borrow ideas from the "*s*-ensemble," a method for driving dynamics out of equilibrium using a biasing potential, *s* (17-20). In the text below, we discuss how one might apply this non-equilibrium perturbation theory to MSMs, and we go on to present the results of the *s*-ensemble approach carried out on 16 protein folding systems.

In using statistical mechanics in a path-based formalism, we define a trajectory, $x(t_{obs})$, as a time series of system configurations over some observation interval, t_{obs}. For the discrete-space, discrete-time Markov chains studied here, a trajectory is represented by a simple sequence of the system's Markov states, where transitions between states are determined by the model's transition probability matrix and occur at a fixed time interval, τ_{lag}.

To study dynamics in the *s*-ensemble, we introduce the real-valued biasing parameter, *s*, and the concept of a trajectory activity, *K*. The activity is an extensive measure of the "change" in a trajectory; in a spin system, for example, the

activity might be represented by a count of spin flips over a trajectory. Here, the activity will be measured by the number of conformational state-to-state transitions, $i \rightarrow j$, such that $i \neq j$, counted over a trajectory of length t_{obs}. The probability of a path $x(t_{obs})$ with activity K in a given s-ensemble is

$$P(x(t_{obs}), s) = P'(x(t_{obs}))\frac{\exp(-sK)}{Z(s, t_{obs})}$$

where P'(x(t_{obs})) is the unbiased trajectory probability and $Z(s, t_{obs})$ is called the dynamical partition function (17-20). An obvious analogy exists between s and the inverse temperature β; K thus assumes the role that energy plays in canonical equilibrium statistical mechanics. It is easily confirmed that unbiased dynamics are recovered when s = 0 (17-20). In practice, one can extract all information about biased trajectories from the s-ensemble transition matrix, written as

$$\mathbb{T}(\boldsymbol{s}) = \mathbb{U}\boldsymbol{e}^{-\boldsymbol{s}} + \mathbb{D}$$

where U and D are matrices containing the off-diagonal and diagonal elements of the unbiased MSM transition matrix, respectively. Using this matrix's partition function, we can calculate the mean activity per transition in the s-ensemble:

$$K(s) = -\frac{\partial}{\partial s}\log(Z(s, t_{obs}))$$

A detailed description of how one calculates this partition function and other quantities from T(s) is included in the SI (19,20).

To illustrate the effects of s-field perturbations on protein folding MSMs, we will first look at native state stabilities in our models as a function of the parameters s and t_{obs}. Figure 1 shows the native state probability versus s and t_{obs} for an MSM of the Fip-35 WW domain.

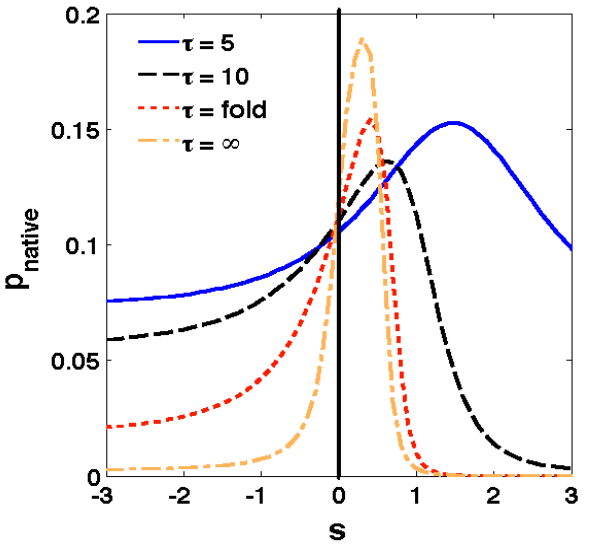


Figure 1. Native state probabilities as a function of s for various values of t_{obs} in the WW domain MSM (τ_{fold} = 28 τ_{lag}). For finite trajectories, native probabilities are calculated at $t = t_{obs}/2$.

Looking at Figure 1, one sees that the s-field has similar qualitative effects on the native state over all values of t_{obs}. At negative s (where activity is enhanced), native state stability is diminished, as transitions out of the folded state are favored. At positive s (where activity is restricted), the population of the folded state first increases, as unfolding transitions simply become less likely. As s becomes arbitrarily large, however, the stability of the native state disappears entirely. Here, we observe an inevitability of the s-ensemble: at large positive s, the system collapses onto its most "metastable" state (i.e., the state with the largest self-transition probability) (19, 20). In general, the curves in Figure 1 reflect the behavior of the native state in all the MSMs we studied. While native states are themselves quite metastable, select low-probability states had larger self-transition probabilities than the native state in all 16 models studied. This observation hints at the existence of highly metastable states in folding landscapes that are exceptionally difficult to access kinetically. Whether these states have physical relevance or are just artifacts of model construction, however, is a difficult question to answer. We leave such investigations, which will likely involve more extensive and targeted MD sampling, for future work. For the most part, we will limit ourselves here to looking at unbiased dynamics (s = 0) through the lens of the s-ensemble.

We should emphasize that, as is made clear in Figure 1, the quantitative nature of the s-ensemble can change drastically as a function of finite t_{obs}. The value of t_{obs} thus needs to be chosen carefully as a parameter. Since this study focuses on protein folding, we will from now on restrict ourselves to setting t_{obs} = τ_{fold} for all models, where τ_{fold} is defined by the longest relaxation timescale in a given MSM.

Plots of the mean activity K as a function of s are useful for studying the different regimes of dynamical behavior in the s-ensemble. Figure 2 shows three such curves for MSMs of the Fip-35 WW domain, Protein G, and the protein NTL9, all at their respective t_{obs} = τ_{fold} (6, 7, 10). By construction, $K(s)$ vs. s curves exhibit a crossover in the mean activity: the s-field drives the system into distinct active and inactive regimes, separated by a relatively fast decay in activity with s. The nature of these curves at t_{obs} = τ_{fold} varies from system to system. In many cases (as with Protein G and the WW domain), the transition is simple and smooth; in a few instances (like with NTL9), however, fine structure in $K(s)$ vs. s emerges that marks density transfer between specific Markov states.

In glass-forming systems, singularities in K(s) curves have been shown to indicate phase transitions between an active phase and an inactive "dynamical glass" (17). Our protein folding models, of course, were not designed a priori as glass-formers, and the transitions between the active and inactive dynamical regimes of MSMs at t_{obs} = τ_{fold} are much more gradual than those in glass systems (17,20). Indeed, since protein molecules are finite in size, they cannot support true dynamical phase transitions (17). However, one does expect the crossovers between active and inactive states to become increasingly sharp in protein systems with long-lived metastable states and more glassy dynamics.

We can easily discriminate between active (K≈1) and "glass-like" inactive (K≈0) regimes within a reasonable variation of the s-field for all MSMs studied. For the remainder of this Communication, we will call these regimes the "active" and "inactive" phases of the dynamics. To facilitate comparison between models, we label the midpoint of the $K(s)$ vs. s curve (K = 0.5) as s*, and we designate that point as the coexistence point between the two phases. We emphasize that the active and inactive regimes that we have found do not correspond directly with the folded and unfolded states of the protein: see Fig 1 and further discussion in the SI.

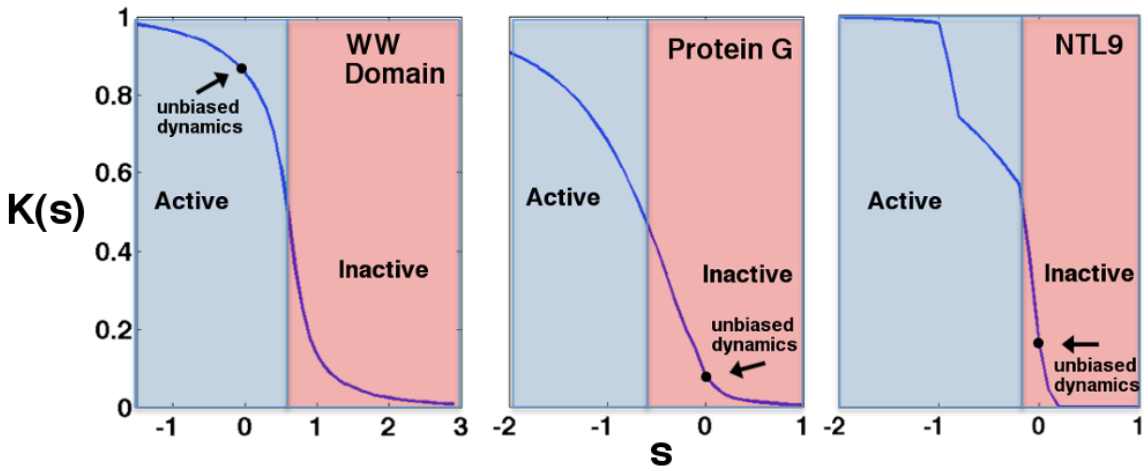


Figure 2: Mean activity per time step, $K(s)$, versus s curves for protein folding MSMs of the Fip35 WW-domain (left, $\tau_{fold} = 28\ \tau_{lag}$), Protein G (middle, $\tau_{fold} = 1110\ \tau_{lag}$), and NTL9 (right, $\tau_{fold} = 3332\ \tau_{lag}$). Similar curves were computed for 13 additional systems; the results of this analysis are shown in Figure 3.

Are the dynamics in protein folding MSMs inherently active or glass-like? As seen in Figure 2, the unbiased dynamics of Protein G and NTL9 are in the inactive phase, while those of Fip-35 are well into the active regime. The idea that this dynamical behavior would vary so drastically between proteins is intriguing, since such heterogeneity suggests that different processes are at work in driving each protein's folding kinetics.

Figure 3 shows the value of s at coexistence between the active and inactive phases as a function of chain length and folding time for all 16 proteins analyzed. Proteins with positive values of s^* are in the active regime under unbiased dynamics, while proteins with negative s^* are in the inactive regime. As noted in the SI, the molecular dynamics (MD) data for these 16 systems were collected under various force fields and at different temperatures; 15 models (villin being the only exception) were constructed from explicit solvent datasets (3 – 10).

We indeed see a wide variation in dynamical behavior among the proteins studied. In general, the smallest, fastest-folding proteins have the most active dynamics, while the largest, slowest folding systems are the most inactive. Proteins with intermediate folding timescales (≈10 μs) generally displayed dynamics near to their respective s^* values. We do not see a strong correlation between chain length and dynamical activity. The helix bundle α3D, for instance, is 73 residues in length, but resides firmly in the active phase. By contrast, all models with folding times greater than 10 μs were shown to exist in the inactive phase. Accordingly, a proposed boundary for dynamically active and inactive proteins in drawn as a horizontal line at 10 μs in Figure 3.

These observations would suggest that dynamical activity is largely independent of thermodynamic considerations, at least with respect to the extensive free energies of folding. We also note that the activity K measures relatively fast motion of the protein (on the time scale τ_{lag}), while the folding time τ_{fold} is a measure of much slower cooperative motion. Fig 3 indicates a strong correlation between molecular motion on these two widely separated time scales, across a range of systems whose molecular size and structure are very different. One might also posit that the proposed boundary in Fig. 3 arises because folding kinetics on time scales longer than 10 μs are somehow more complex and lead to a different dynamical regime. Just what factors contribute to the emergence of these two kinetic regimes, however, are up for debate. Interestingly, all proteins in the inactive phase (with the exception of villin, discussed below) either contain native β-sheet structures or have shown a propensity for forming β-rich misfolded states. It is possible that the emergence of β-sheet dynamics in protein folding is in part responsible for a restriction in dynamical activity. We should note that some proteins in the active regime (i.e. WW domain and BBA) also have native β-sheet structure, suggesting other factors are likely at play in determining the kinetic partitioning. From a molecular simulation point of view, the results concerning the dynamics of the villin headpiece domain are noteworthy. While the villin MSM constructed from explicit solvent MD data was in the active phase, the MSM constructed from implicit solvent data crosses the threshold to the inactive phase. This result suggests that dynamics in GBSA implicit solvent simulations are inherently more glass-like than dynamics in explicit solvent.

In summary, we have shown that s-field perturbations of protein folding MSMs bring to light two distinct regimes of kinetic behavior. We have characterized the unbiased dynamics of 16 protein systems through the lens of s-ensembles, and we have discussed how these active or inactive dynamics might relate to the properties of specific proteins.

Given that we observe a crossover from active to inactive behavior but no true phase transition to an inactive glass phase, our results agree with past consensus about the role of glassy dynamics in protein folding, i.e., that glass systems have marked similarities with, but are not directly applicable to, protein-folding systems. With respect to the active and inactive phases discussed in this paper, the proteins nearest to the proposed phase boundary in Figure 3 might warrant further study. In particular, one might see if temperature changes, specific mutations, or other perturbations would drive systems over the active-inactive threshold. As stated previously, the nature of the low-probability states onto which the MSM s-ensembles collapse is also potentially interesting. As s-ensemble methods are natural for studying perturbed MSM dynamics, a number of other intriguing extensions of this work could be imagined.

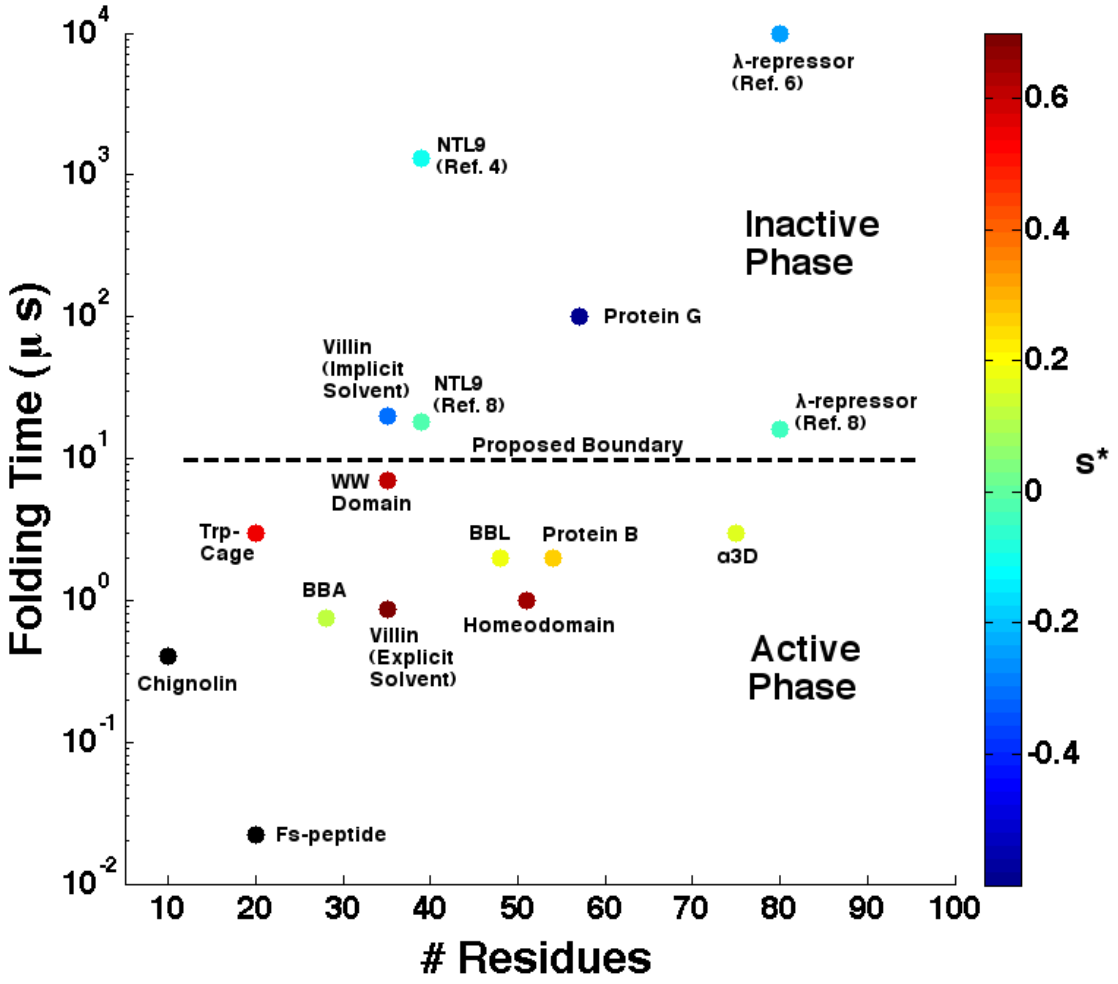


Figure 3: Plot of the *s*-ensemble parameter at coexistence, *s**, as a function of folding time (in μs) and chain length (in number of residues) for 16 protein folding MSMs. Values of *s** for Fs-peptide (1.20) and Chignolin (1.95) were omitted to preserve scale. The magnitude of an *s*-value suggests how far a model's unbiased dynamics deviate from coexistence between active and glassy phases.

Acknowledgement This work was supported in part by the NIH (R01-GM062868) and the NSF (MCB-0954714). JKW was funded by the Fannie and John Hertz Foundation on the endowed Yaser S. Abu-Mostafa Fellowship.

References

1. Dill KA, Ozkan SB, Shell MS, and Weikl TR. 2008. Annu. Rev. Biophys. 37 289-316.

2. Dobson CM. 2004. Sem. In Cell and Dev. Bio. 15(1) 3-16.

3. Sorin EJ and Pande VS. 2005. Biophys. J. 88(4) 2472-2493.

4. Ensign DL, Kasson PM, and Pande VS. 2007. J. Mol. Biol. 374(3) 806-816.

5. Bowman GR and Pande VS. 2010. Proc. Nat. Acad. Sci. 107(24) 10890 – 10895.

6. Voelz VA, Bowman GR, Beauchamp K, and Pande VS. 2010. J. Am. Chem. Soc. 132(5) 1526 – 1528.

7. Lane TJ, Bowman GR, Beauchamp K, Voelz VA, and Pande VS. 2011. J. Am. Chem. Soc. 133(45) 18413 – 18419.

8. Bowman GR, Voelz VA, and Pande VS. 2011. J. Am. Chem. Soc. 133(4) 664 – 667.

9. Lindorff-Larsen K, Piana S, Dror RO, and Shaw DE. 2011. Science 334(6055) 517 – 520.

10. Beauchamp KA, McGibbon R, Lin YS, and Pande VS. 2012. Proc. Nat. Acad. Sci. 109(44) 17807 – 17813.

11. Bryngelson JD and Wolynes PG. 1987. Proc. Nat. Acad. Sci. 84 7524 – 7528.

12. Shakhnovich EI and Gutin AM. 1989. Biophys. Chem. 34(3) 187- 189.

13. Morozova-Roche LA, Jones JA, Noppe W, and Dobson CM. 1999. J. Mol. Biol. 289(4) 1055 – 1073.

14. Sabelko J, Ervin J, and Gruebele M. 1999. Proc. Nat. Acad. Sci. 96(11) 6031 – 6036.

15. Prigozhin MB and Gruebele M. 2011. J. Am. Chem. Soc. 133(48) 19338 – 19341.

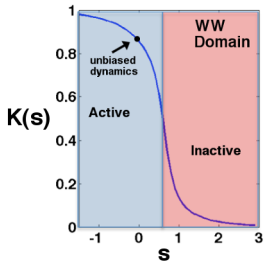
16. Gillespie B and Plaxco KW. 2000. Proc. Nat. Acad. Sci. 97(22) 12014 – 12019.

17. Garrahan JP et al. 2007. Phys. Rev. Lett . 98 195702.

18. Merolle M, Garrahan JP, and Chandler D. 2005. Proc. Nat. Acad. Sci. 102(31) 10837 – 10840.

19. Jack RL and Sollich P. 2010. Prog. Theor. Phys. Supp. 184 (Supplement 1): 304-317.

20. Hedges LO, Jack RL, Garrahan JP, and Chandler D. 2009. Science 323 (5919) 1309 – 1313.


K(s)
s
WW
Domain
unbiased
dynamics
Active
Inactive
1
0.8
0.6
0.4
0.2
0
-1
0
1
2
3