Simple few-state models reveal hidden complexity in protein folding

Kyle A. Beauchampa, Robert McGibbonb, Yu-Shan Linc, and Vijay S. Pandeh

aBiophysics Program, Stanford University, Stanford, CA 94305; and bChemistry Department, Stanford University, Stanford, CA 94305

Edited by Alan R. Fersht, Medical Research Council Laboratory of Molecular Biology, Cambridge, United Kingdom, and approved June 12, 2012 (received for review March 15, 2012)

Markov state models constructed from molecular dynamics simulations have recently shown success at modeling protein folding kinetics. Here we introduce two methods, flux PCCA+ (FPCCA+) and sliding constraint rate estimation (SCRE), that allow accurate rate models from protein folding simulations. We apply these techniques to fourteen massive simulation datasets generated by Anton and Folding@home. Our protocol quantitatively identifies the suitability of describing each system using two-state kinetics and predicts experimentally detectable deviations from two-state behavior. An analysis of the villin headpiece and FIP35 WW domain detects multiple native substates that are consistent with experimental data. Applying the same protocol to GTT, NTL9, and protein G suggests that some beta containing proteins can form long-lived native-like states with small register shifts. Even the simplest protein systems show folding and functional dynamics involving three or more states.

More than fifty years ago, Anfinsen (1) found that primary sequence encodes functional ribonuclease. Today, the process by which proteins self-assemble into their biologically-relevant states remains a key question in biophysics. A modern theory of protein folding must explain a variety of disparate experimental results. First, proteins can fold via a number of mechanisms, ranging from the simplest two-state model (2) to models with more complexity (3). Some proteins may not fold at all—recent work has shown that many eukaryotic proteins are either intrinsically disordered or fold only upon binding to their target (4). The misfolding and aggregation of other proteins have been associated with neurodegenerative disorders such as Alzheimer’s disease (5). A “solution” to the protein folding problem must capture not only the folded and unfolded states but intermediate, disordered, and misfolded states as well.

Kinetic models with two (2) or more (3, 6) states have been the dominant paradigm for understanding protein folding dynamics. This picture generalizes to an arbitrary number of states; the resulting dynamical model is known as the master equation (7). Recently, several labs (8–13) have used molecular dynamics simulations to parameterize discrete time master equations (also known as Markov state models) that describe the folding dynamics of proteins. Such procedures typically involve clustering the observed conformations to determine microstates, then counting transitions between microstates to estimate rates.

Markov state model approaches have typically been limited by two issues. First, the models so constructed have involved hundreds or thousands of states, leading to difficulty making direct connections to experimentally-derived models of folding kinetics. To address this problem, we have developed flux PCCA+ (FPCCA+). Like its precursors PCCA and PCCA+ (14–16), FPCCA+ allows one to construct macrostate models that optimally capture the slow dynamics observed in a microstate model; the advantage of FPCCA+ is in its ability to ignore slow but irrelevant (e.g. low population) dynamics by discarding eigenvectors that have low equilibrium flux. The second difficulty with MSM approaches is that imperfect state decompositions can lead to non-Markovian dynamics, or memory (8, 17, 18), which can bias rate predictions. To address this issue, we have developed a new rate estimation protocol, sliding constraint rate estimation (SCRE), that accurately estimates Markovian rates.

Combining these two approaches allows intuitive few-state models that accurately capture simulated protein folding kinetics in quantitative detail. While several previous studies have demonstrated accurate MSMs, such models have either involved thousands of states (11, 19) or have focused on simple peptide systems (12). The current protocol, however, constructs highly accurate few-state models, enabling quantitative and intuitive connections to experiment. Most importantly, our approach is equally applicable to two-state, three-state, or many-state behavior, with the optimal number of states selected algorithmically.

Applying these methods to 14 folding simulation datasets (20–22) reveals new insights into the nature of two and multi state kinetics. A flux analysis suggests that although model proteins show approximate two-state behavior, additional complexity is apparent. For example, three-state models of the villin headpiece and FIP35 WW domain reveal multiple native states. In contrast, three state models of the beta-containing proteins GTT, NTL9, and protein G (NuG2) show register-shifted native-like states with microsecond lifetimes. Simple but accurate multi-state models can reveal hidden complexity in protein folding.

Methods

Relaxation Spectra of Fast-Folding Proteins. How accurately can two-state models describe simulated protein folding? To answer this question, we analyzed a collection of 14 protein folding datasets (20–22) that were collected using either Anton (23) or Folding@home (24). For each system, a microstate MSM was used to calculate the relaxation timescales and eigenvector fluxes (see Methods). A system with two-state behavior is expected to show a large spectral gap—a single, high-flux timescale that is much slower than the remaining timescales. According to this analysis (Fig. 1), approximate two-state behavior is observed for chignolin, trp-cage (2JOF), HP35 (360 K), FIP35, GTT, NTL9, and A3D. For HP35 (300 K), homeodomain (UVF), BBL, BBA, protein G (NuG2), and lambda, however, two-state models cannot accurately describe the simulated dynamics. For this analysis, we classify a system as two-state if the slowest eigenvector shows a gap in both flux and timescale. More formally, we require that there is no other eigenvector with \( \frac{\Phi}{\Phi_{\text{slow}}} \geq \frac{1}{2} \) where \( \Phi \) is the relaxation timescale of the eigenvector, \( \tau_{\text{slow}} \) is the slowest relaxation in the system, \( \Phi \) is the eigenvector flux, and \( \Phi_{\text{slow}} \) is the eigenvector flux of the slowest relaxation in the system.

FPCCA+/SCRE Produces Accurate Two-State Models. To validate our analysis protocol, we estimated two-state models for the systems

Author contributions: K.A.B. and V.S.P. designed research; K.A.B. performed research; K.A.B. and R.M. contributed new reagents/analytic tools; K.A.B. and Y.-S.L. analyzed data; and K.A.B. and V.S.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

To whom correspondence should be addressed. E-mail: pande@stanford.edu.

This article contains supporting information online at www.pnas.orglookup/suppl/doi:10.1073/pnas.1201810109/-/DCSupplemental.
showing approximate two-state behavior. A two-state model is wholly described by two parameters, the fraction folded \( f = \frac{k_f}{k_f + k_u} \) and the relaxation timescale \( \tau_r = (k_f + k_u)^{-1} \). We compared our estimates of \( f \) and \( \tau_r \) to previous estimates (21, 22) made using the Hummer reaction coordinate formalism (25) and find quantitative agreement (Fig. 2). Further model validation is provided in Figs. S1–S7.

A key advantage of the FPCCA+/SCRE method, however, is its applicability to systems with multiple states. Indeed, a close examination of the microstate transition matrix reveals additional high-flux relaxations occurring on timescales somewhat faster than folding. We therefore set out to construct simple models capturing an intermediate level of detail, with the hope that such models might reveal multi-state behavior while still remaining intuitive.

**Multi-state Kinetics in BBA Folding Simulations at 325 K.** According to the relaxation spectrum analysis, BBA shows several slow but populated relaxations (Fig. 1); we therefore constructed a 5 state model using FPCCA+/SCRE. The resulting model (Fig. 3) consists of native, near-native, unfolded, and trapped states. The trapped states show non-native beta hairpins and each had microsecond lifetimes and 1% populations. The detected native state shows reasonable agreement with the NMR structure (26) and a population (17%) similar to a previous simulation based model (22%); the near-native state (population 3%) shows a native-like helix but a register shifted beta hairpin.

**The Villin Headpiece Populates Multiple Native States.** The relaxation spectrum analysis (Fig. 1, red) suggests the presence of two slow, high-flux timescales in a recent simulation of HP35-NLE-NLE at 360 K (22). We therefore used FPCCA+ to construct a three-state model that captures these relaxations. The resulting model consists of unfolded, near-native, and native states (Fig. 4A) with respective populations 77%, 6%, and 18%. The native state shows a strong resemblance to the crystallographic model (PDB: 1vii, 2ppz), with many conformations showing RMSDs under 2 Å. The near-native state is characterized by partial unraveling of the C terminal helix and shows some similarity to NMR structures (PDB: 1vii, 2ppz) of the wild-type (without norleucine) sequence, which also show partial unraveling of the C terminal helix. The rate matrix relaxation times (400 ns, 70 ns) differ by only a factor of 7, indicating that, at 360 K, HP35 does not show a strong separation of timescales. Given the triangular topology of the rate graph (Fig. 4A), a single reaction coordinate would give a poor description of the system.

Because the near-native N' is characterized by an unraveled C terminal helix, we asked whether chemical shifts might be sensitive to the difference between N and N'. To address this question, we used Sparta+ (27) and ShiftX2 (28) to estimate chemical shifts for each state. According to our analysis, amide chemical shifts of residues 32 and 33 show differences between the N and N' states (Fig. 5A and B). Although the differences are small, they are slightly larger than the systematic error inherent in chemical shift prediction. Furthermore, the difference is detected regardless of which chemical shift model is used. We also asked whether the near-native state would show experimentally observable differences in tertiary structure. To crudely simulate a proton nuclear overhauser experiment, we calculated the r^-6 weighted distance matrix, that is: \( d_{ij} = (r_{ij}^6)^{-1/6} \). Heat-maps of this quantity suggest that proton NOE experiments could detect subtle differences between the native and near-native states. In particular, the simulated native state shows residues 34 and 35 interacting with residues 29 and 30. In the near-native state, these interactions disappear as the C terminal helix unravels between residue 31 and 35 (Fig. 5D). Temperature melts of NMR observables in the C terminal helix may reveal multiple native substrates, although direct observation of N' may require experi-
mental conditions that perturb the equilibrium between N, N', and U. Direct kinetic measurements of an N to N' transition may be difficult due to the fast interconversion rate (precluding relaxation dispersion experiments) and the structural similarity of N and N'.

How robust is the predicted model to changes in force field and temperature? To address this, we analyzed an independent set (20) of 213 1.2 microsecond simulations. These simulations were held at 300 K and used the amber99sb-ildn (29) force field, which is less helical than the Charmm22* (30) used in the 360 K simulations (31, 32). According to the flux analysis, the 300 K dataset shows considerably less separation of timescales, with several high-flux relaxations occurring in the microsecond range. With many slow relaxations, we do not expect few-state FPCCA+ models to provide an accurate description of the dynamics. Despite the absence of strong three-state behavior, the 300 K simulations indeed sampled near-native conformations similar to those observed in the 360 K data (Fig. 6A). To probe the relative frequency of N versus N' in the 300 K data, we calculated a 2D histogram of the RMSD to native and near-native conformations (Fig. 6B). The population ratio of N to N' is approximately 9:1, roughly consistent with the 360 K data (4:1). We mention that similar heterogeneity has been observed in previous simulation studies (33, 34).

---

Fig. 3. The dominant kinetics of BBA at 325 K. Rate timescales are marked on each edge. The NMR structure (PDB: 1fme) is shown in black. Full rate matrices with error estimates are given in SI Text.

Fig. 4. The dominant kinetics of HP35 (A), FiP35 (B), and GTT (C) are summarized as networks. Simulated structures are shown in rainbow. For HP35 (A), experimental structures are shown in black (crystal structure 2f4k) and gray (NMR structure 1vii). For FiP35 (B), experimental structures are shown in black (holo PDB: 1i8g) and gray (apo PDB: 1i6c). For GTT (C), the experimental holo structure is shown in black. Full rate matrices with error estimates are given in SI Text.
Apo-Holo Dynamics of the FiP35 WW Domain. The spectrum of FiP35 (35) at 395 K (21) (Fig. 1, green) suggests the presence of high-flux relaxations on the microsecond and hundred nanosecond timescales. We therefore used FPCCA+ to construct a three-state model for the two dominant relaxations. The three-state model consists of holo (H), apo (A), and unfolded (U) states with respective populations 60%, 2%, and 38%; the observed states show good agreement with NMR structures (holo PDB:1i8g, apo PDB:1i6c) of the related Pin WW domain (36) (Fig. 4 B). The estimated rates in this model suggest that reaching the apo state most often occurs via the holo state, as no direct transitions between U and A are observed with the given level of sampling (Fig. 4 B). We point out that previous simulation analyses have identified additional near-native dynamics (19), traps (11), and short-lived intermediates (37) in simulations of WW domains.

Slow Register Shift Dynamics: a General Property of Beta Topologies? The relaxation spectra of the beta-containing protein G (NuG2) (38), NTL9 (39), and GTT (40) (a WW domain) each show a high-flux relaxation (Fig. 1, blue) that occurs on timescales 3–10 times faster than the slowest relaxation. We constructed three-state FPCCA+ models to capture these relaxations. In all three model systems, the observed relaxations correspond to native-like states with register-shifted beta strands; distance maps for GTT are shown in Fig. 7, while PDB structures are shown in Fig. 4C. The populations of the register-shifted states range from 0.1% (GTT) and 0.5% (NTL9) to 1.5% (NuG2); lifetimes of these states range from 0.3 to 3 μs. Finally, our analysis of FiP35 did not detect a similarly populated register shifted state. This may suggest that the populations of register-shifted states are either mutant-specific (GTT vs. FiP35) or dependent on the simulation force field (Charmm22* versus Amber99sb-ildn). A similar register-shifted state is observed in the five state model for BBA (Fig. 3). Register shifted states appear in 4/5 of the beta sheet proteins analyzed here as well as in two recent experimental studies (41, 42). We therefore suggest that slow register-shift dynamics may be a general phenomenon in beta topologies.

Experimental detection of register shifted states might be possible with isotope labeled 2DIR experiments (43). Alternatively, interconversion between native and near-native states may be detectable by relaxation dispersion NMR, which in principle can detect states with populations under 1.0% (44). Key questions are whether the native and near-native states have distinct chemical shifts and whether the interconversion timescale lies in the millisecond regime. Prediction of chemical shifts for the native and near-native states of NTL9 suggest that certain residues may show experimentally detectable chemical shift differences between the native and near-native states (Fig. S8).
high-temperature (≈350 K) simulations, the native and near-native states interconvert on the 1 to 10 μs timescale; at lower temperatures, interconversion times could approach one millisecond.

Discussion

Understanding the Relaxation Spectra of Model Proteins. Systems with approximate two-state behavior can hide remarkable complexity in their relaxation spectra. Although we have focused on simple model systems, our analysis has shown the presence of remarkably different three-state models. At 360 K, HP35 has only a weak separation of timescales, leading to a three-state model where N and N’ interconvert on the hundred nanosecond timescale, only marginally faster than the overall folding reaction. On the other hand, Fip35 shows a cleaner separation of timescales; it reaches its apo state only by first traversing the holo state. The other beta systems (GTT, NTL9, NuG2) show native-like, register-shifted states that might be described as either intermediates or kinetic traps; additional sampling will be required to conclusively decide which.

Comparing HP35 Simulation and Experiment. We first summarize the experimental consensus on HP35 folding. NMR (45) and x-ray crystallography (46) have revealed the structure of HP35 and its fast-folding (2) mutants. Circular dichroism (47) and other experiments have shown HP35 to be well-folded under ambient temperatures. The double norleucine mutant simulated herein was found (2) to have a melting temperature of 360 K. Finally, IR (48), NMR (49), and fluorescence (2) experiments have shown the folding rate of HP35 to be 3 μs or faster, with the double norleucine mutant showing an 800 ns folding timescale (2, 50).

Despite the success of two-state models in analyzing HP35 experimental results, many experiments demonstrated additional complexity. Laser temperature jump experiments have detected a 100 ns burst-phase (2, 51), believed to be a helix-coil transition within the folded state. A three-state fit (52) of available experi-

Fig. 6. (A) Near-native (gray: 361 K, rainbow: 300 K) conformations are compared for both HP35 datasets. The crystal structure (PDB:2f4k) is shown in black. (B) A 2D RMSD histogram suggests that the 300 K simulations populate both the native and near-native states, with a population ratio of approximately 9:1. The line x = y was used to separate the N and N’ states.

Fig. 7. Distance maps for GTT reveal a register-shifted near-native state. (A) Distance map (Cα) for state N. (B) Distance map for state N’ shows a register shift.
mental data leads to an intermediate state that is well-populated (10–50%) in the 300–350 K temperature range. Triplet-triplet energy transfer experiments detected multi-exponential kinetics in contact formation traces (53), leading to a model with native, near-native, intermediate, and unfolded states. In that model, the near-native state was native-like but lacking the same degree of compactness and terminal contact while the intermediate state was characterized by an unfolded C terminal helix and contact between residues 23 and 35.

Our three-state model recapitulates several previous experimental results. Structurally, the native state shows agreement with the crystallographic structure. Temperature jump experiments at 360 K detected equal amplitudes of slow (360 ns) and mental results. Structurally, the native state shows agreement between residues 23 and 35.

of compactness and terminal contact while the intermediate state was native-like but lacking the same degree

energy transfer experiments detected multi-exponential kinetics in

C

coordinates were available due to bandwidth limitations. We point out that

model.

The present analysis focuses on 13 folding datasets (21, 22) collected on the Anton folding machine (23) and one folding dataset (20) collected on Folding@home using Gromacs 4.5 (55). Simulations were performed using either the amber99sb-ildn (29) (HP35-300K, FIP35) or the charmm22* (30) (all other systems) force field. Each trajectory was first subsampled at 50 ns intervals (30 ns for the 300 K HP35 data; 10 ns for the 360 K HP35 data). For the Folding@home dataset, trajectories that were disconnected from all other trajectories (minimum RMSD >2.0 Å) were discarded prior to clustering. 41 of 254 trajectories were discarded. Antion trajectories were downloaded from DE Shaw Research. For trp-cage, HP35, and FIP35, all atom coordinates were available online. For the remaining systems, only Cα coordinates were available due to bandwidth limitations. We point out that DE Shaw Research has offered to provide all-atom coordinates for other systems via mail; we obtained all-atom coordinates for NTL9 in this manner. For Figs. 3 and 4C, we used the Pulchra software tool (56) to reconstruct all-atom PDB coordinates.

Model Construction. States are identified using a protocol similar to previous Markov state model approaches (8, 10–13, 57–59). First, the data are clustered into microstates using Ward's algorithm (60, 61) with the RMSD metric; Ward clustering provides better microstates than previous clustering approaches (Fig. 59). For HP35, FIP35, trp-cage, and NTL9, RMSD calculations were performed using Cα, Cβ, C, N, and O atoms. For the other systems, RMSD calculations were performed using Cα coordinates. The number of clusters is chosen such that \( n_{\text{clusters}} = n_{\text{mutations}} \). As our data shows agreement between residues 23 and 35, we find that this choice provides a good compromise between clustering accuracy (large \( n_{\text{clusters}} \)) and reliable statistics (small \( n_{\text{clusters}} \)). For BBA, somewhat more microstates were required. Clustering details (number of states, clusters, data storage frequency, and MSM lagtime) are provided in Table S1. A strongly ergodic, reversible transition matrix (57) is estimated for the microstate model. After clustering, FPCCA+ is used to construct a macrostate model that captures the slowest relevant relaxations in the data.

FPCCA+. The PCCA+ algorithm constructs a macrostate model that is most consistent with the underlying microstate model (15, 16). It does so by capturing the space spanned by the slowest eigenvectors of the microstate transition matrix. However, one possible problem with PCCA+ is that some slow eigenvectors may correspond to small population changes, which can be quantified by defining the eigenvector flux to be \( \Phi_i = ||\psi_i||^2 \), where \( \psi_i \) is the nth \( \alpha \)-normalized eigenvector (see SI Text for derivation). To construct macrostates that have significant populations, one can discard any slow eigenvectors that contribute insignificant amounts of equilibrium flux. We call this approach flux PCCA+ (FPCCA+). It is worth mentioning that FPCCA+ is essentially just using PCCA+ to selectively model a user-specified subset of eigenvectors; the original PCCA+ algorithm involved selecting just the slowest \( n \) eigenvectors. For the FPCCA+ models constructed here, we selected eigenvectors by choosing dual cutoffs for both implied timescale (\( \tau_c \)) and for flux (\( \Phi_i \)). We then used FPCCA+ to model all eigenmodes that satisfy both \( \tau > \tau_c \) and \( \Phi > \Phi_i \). A graphical depiction of these cutoffs is shown in Fig. S2. We point out that FPCCA+ is similar in spirit to "dynamical fingerprint" analysis (62).

Sliding Constraint Rate Estimation (SCRE). To estimate accurate rate constants that are unbiased by non-Markovian behavior, we have developed the SCRE approach to rate estimation. Given a set of simulations that have been assigned to states \( \{ i \}_{\pi} \), one can calculate a matrix of transition counts from state \( i \) to \( j \) (\( C_{ij} \)) that were observed with a lagtime (or sampling window) of \( \tau \). The log likelihood of a fixed lagtime transition matrix \( T(\tau) \) is given by \( f(T; \tau) = \sum_{\tau} C_{ij} \log(T_{ij}) \) (63). To estimate rate matrices, we express the transition matrix in terms of the generating rate matrix \( T(\tau) = \exp(-K\tau) \), leading to \( f(K; \tau) = \sum C_{ij} \log(\exp(-K\tau)) \); here \( \log(\exp) \) refers to matrix exponential, while log refers to the scalar logarithm. Maximizing the log likelihood with respect to \( K \) gives a maximum likelihood estimate of the rate matrix (12). In practice we use detailed balance to fix \( K_{ij} = K_{ji} \frac{p_i}{p_j} \) where the equilibrium populations \( p_i \) are estimated using a fixed-lagtime reversible transition matrix (57) with minimal lagtime. We also constrain the sparsity structure of \( K \) to that observed in the transition matrix; this reduces the number of parameters. Likelihood maximization is performed using the downhill simplex algorithm. It is important to estimate rates at a lagtime sufficiently long that the rates are Markovian, as otherwise rates will be biased by fast nonproductive fluctuations. On the other hand, too long a lagtime will lead to aliasing of the faster processes and large statistical uncertainties. Furthermore, the Markovian lagtime of one process may be longer than the other timescales in the system; this implies that a single lagtime might not allow accurate modeling all of the observed dynamics. As a solution, we propose the following algorithm to extract each rate constant from its shortest Markovian lagtime. For the first iteration, \( r \) is set to 1. Let \( \tau_r \) denote the time spacing between successive trajectory frames.

1. Calculate the counts \( C(\tau_r) \) at lagtime \( \tau_r \).
2. Use \( C(\tau_r) \) to estimate \( K(\tau_r) \) using likelihood maximization.
3. Check for (Markovian) convergence: if \( K_{ij}(\tau_r) = K_{ij}(\tau_r - 1) \), then fix \( K_{ij} \) to be the current values.
4. Increment \( r \) by 1 and repeat.

Presently, we estimate convergence graphically by manual inspection, but we are currently developing a fully automated approach. The lagtime for a given rate is chosen such to be the shortest lagtime where the given rate has plateaued. Good state decompositions will give rate elements that level immediately, while poor state decompositions may require long lagtimes to converge. Using this algorithm, each rate \( K_{ij} \) is estimated at its minimal Markovian lagtime. A key advantage of SCRE is that it can, if necessary,
estimate each $K_n$ with a different lagtime. In contrast, previous methods tend to estimate all $n^2$ rate elements at a single chosen lagtime. The use of a single lagtime can be problematic in systems where the rates span a large range of timescales; using too short a lagtime will lead to biased kinetics, while using too large a lagtime will also bias the dynamics observed in the system. We further validate the SCRE method in the SI text.

Software Availability. All analysis code (Ward clustering, FPCCA+, SCRE) is freely available in MSMBuilder 2.5 ([https://smitk.org/home/msmbuilder](https://smitk.org/home/msmbuilder)) (9, 57). The tutorial for MSMBuilder 2.5 applies Ward clustering, FPCCA+, and SCRE to analyze simulations of alanine dipeptide (see SI Text for details).

ACKNOWLEDGMENTS. We are very grateful to DE Shaw research for providing simulations. KAB was supported by a Stanford Graduate Fellowship. We thank NSF CNS-0619926 for computing resources and NIH R01-GM026868, NSF-0400700, and NSF-MCB-0954714 for funding. We thank Rhiju Das, TJ Lane, and Jesús Izaguirre for helpful discussions. We thank Folding@home donors for computing resources.
Supporting Information

Beauchamp et al. 10.1073/pnas.1201810109

SI Text

Supporting Methods. Rate matrices. Here, we summarize the rate matrices used in this work. Rate matrices are generated for macrostate models by applying the SCRE protocol described in Methods. Estimated error bars are given by

$$\frac{1}{\kappa_{ij}} [\mu] = \begin{pmatrix} -0.08 \pm 0.003 & 0.37 \pm 0.008 & 3.06 \pm 0.031 \\ 0.13 \pm 0.005 & -0.25 \pm 0.005 & 3.71 \pm 0.038 \\ 0.24 \pm 0.009 & 0.82 \pm 0.018 & -1.68 \pm 0.017 \end{pmatrix}$$

for States U, N, N

$$\frac{1}{\kappa_{ij}} [\mu] = \begin{pmatrix} -4.61 \pm 0.096 & 13.66 \pm 0.342 & 0.22 \pm 0.024 \\ 20.37 \pm 0.423 & -13.66 \pm 0.342 & \infty \\ 5.96 \pm 0.124 & \infty & -0.22 \pm 0.024 \end{pmatrix}$$

for States U, N, N

$$\frac{1}{\kappa_{ij}} [\mu] = \begin{pmatrix} -8.56 \pm 0.18 & 0.32 \pm 0.12 & 45.94 \pm 0.44 \\ 101.93 \pm 2.17 & -0.32 \pm 0.12 & \infty \\ 9.35 \pm 0.20 & \infty & -45.94 \pm 0.44 \end{pmatrix}$$

for States U, N, N

BBA. States N’, T1, T2, N, U

Derivation of flux. Suppose we have a rate matrix $K$. The population change is given by the master equation

$$\frac{d}{dt}x_i = \sum_j K_{ij}x_j$$

where $x_i$ denotes the population of state $i$ and $K_{ij}$ is a matrix of rates from state $j$ to state $i$. With this observation, we separate the incoming and outgoing flux: $\frac{d}{dt}x_i = \sum_j K_{ji}x_j + K_{ij}x_i$; we know that $K_{ij} < 0$ because the rate matrix conserves probability (e.g. $\sum_i K_{ij} = 0$). Define the flux of state $i$ to be the outgoing population, that is,

$$\Phi_i(x) = K_{i}x_i$$

Here, $x$ is some arbitrary set of starting populations, which we will eventually set to be equal to the equilibrium populations $\pi$. We now define the total flux as the sum of all the fluxes:

$$\Phi(x) = \sum_i K_{i}x_i$$

Now, suppose that $K$ has $\lambda_i$, $\phi_i$, and $\psi_i$ as eigenvalues, right eigenvectors, and left eigenvectors, respectively. If the eigenvectors are properly normalized such that $\phi_i^T\psi_j = \delta_{ij}$, $\phi_0 = \pi$, and $\psi_0 = 1$, one can express the rate matrix as $K_{ij} = \sum_n \lambda_n (\phi_n)(\psi_n)_j$. Inserting this into the total flux gives

$$\Phi(x) = \sum_i x_i \sum_n \lambda_n (\phi_n)(\psi_n)_i$$

Exchanging the order of summation, we have

$$\Phi(x) = \sum_n \lambda_n \sum_i (\phi_n)(\psi_n)_i$$

This expression gives the total outgoing flux. Suppose we prepare the system at equilibrium; that is, $x = \pi$. Recall that detailed balance implies a relationship between the left and right eigenvectors:

$$(\psi_n)_i = (\phi_n)_i / \pi_i$$

Inserting this expression in the flux calculation gives an expression for the total flux at equilibrium:

$$\Phi = \sum_n \sum_i (\phi_n)_i^2 = \sum_n \lambda_n ||\phi_n||^2$$

Thus, the total equilibrium flux $\Phi$ can be decomposed into contributions from each eigenvector. Eigenvectors with small values of $||\phi_n||^2$ will contribute little population flux at equilibrium.

Markov state model validation. The key question when applying MSM techniques is whether the model is able to accurately recapitulate the simulation data. Several consistency checks exist for MSM validation. Below, we describe the implied timescale, Chapman-Kolmogorov, and autocorrelation tests for MSM validation.

The eigenvalues of a fixed-lagtime transition matrix $T(\tau)$ are related to the eigenvalues of the generating rate matrix $K$ via the equation $\tau_j = \frac{1}{\lambda_j}$ Here, $-\frac{1}{\tau}$ is an eigenvalue of $K$, $\tau$ is the lagtime, and $\lambda_j$ is an eigenvalue of the transition matrix $T(\tau)$. The right and left sides of this expression should be equal and independent of the lagtime $\tau$, thus providing a necessary condition for an accurate model. In practice, the dynamics is Markovian only for large lagtimes, typically leading to increasing timescales that eventually level off. Due to the finite sampling window (aliasing), implied timescales also tend to grow linearly and noisily once $\tau \gg \tau_j$ (see below for proof). Ideally, a model would capture the converged value of the timescales.
Assuming Markovian dynamics, the transition matrices at various lagtimes should be related to the instantaneous rate matrix via the Chapman-Kolmogorov equation: $T(\tau) = T(1)^{\tau} = \exp(K\tau)$. This equivalence allows one to validate the Markovian nature of the transition matrix elements. Here, we focus on the self transition elements $T_{ii}(\tau)$; due to the metastability of the states, we expect self transition elements to have better statistics than the off-diagonal elements. At short lagtimes, we expect memory effects to dominate the observed self-transition probabilities. Indeed, we find that the sliding-constraint rate estimates agree with the observed self-transition probabilities for all lagtimes greater than some minimal lagtime. The minimal lagtime required for agreement depends on the state of interest and provides an estimate of the time required to reach equilibrium within each macrostate. In all systems examined, we find that fixed-lagtime transition matrices with short lagtimes underestimate the metastability, as compared to the rate estimates. This may suggest that non-productive fluctuations between the macrostates are the source of the inaccuracy of the fixed-lagtime models.

As a third test, we consider the ability of the model to recapitulate autocorrelation functions of the raw trajectory data. Let $x(t)$ be a trajectory of some observable and $z = x(t) - \langle x \rangle$. The time autocorrelation function of $x(t)$ is given by $C(t) = \frac{\langle z(t)z(0) \rangle}{\langle z^2 \rangle}$. For protein systems, we let $x(t)$ be the RMSD to the experimentally-determined structure. With this choice of $x$, the correlation function tracks the kinetics of folding, but in a way that is independent of our choice of state decomposition.

As a final test, we consider the autocorrelation functions of the eigenvectors. It has been shown that the autocorrelation function of a left eigenvector will decay exponentially with a time constant equal to the corresponding rate matrix eigenvalue. This provides a critical test of MSM accuracy.

In the supplementary figures, we validate FPCCA+/SCRE models of the following systems:

1. Alanine dipeptide (Fig. S3)
2. HP35 (361 K) (Fig. S4)
3. FIP35 (Fig. S5)
4. GTT (Fig. S6)
5. BBA (Fig. S7)

In part (A) of these figures, we compare the implied timescales as calculated two ways. First, the implied timescale is calculated by estimating a sequence of fixed-lagtime reversible transition matrices. Second, the implied timescales are calculated using the SCRE rate matrix.

In part (B) of these figures, we perform an RMSD autocorrelation analysis. The MSM autocorrelation function is estimated in a two-step approach. First, the MSM is used to generate a chain of states (e.g. a "pseudotrajectory" of state assignments). Second, for each state in chain, we randomly sample an RMSD value from the conformations assigned to that state. This results in a long trajectory of RMSD values, which can then be used to estimate the autocorrelation as described above.

In part (C) of these figures, we perform a Chapman Kolmogorov test of the self-transition probabilities. The test compares the self-transition probabilities as estimated two ways. First, the self-transition probabilities are estimated by estimating a fixed-lagtime model with the chosen lagtime ($T(\tau)$). Second, the self-transition probabilities are estimated by the matrix exponential of the SCRE rate matrix: $T_{ii}(\tau) = \exp(K\tau)_{ii}$. At long timescales, we expect agreement between these two estimates.

In part (D), we perform an autocorrelation analysis of the SCRE-estimated eigenvectors. The autocorrelation function of these eigenvectors are compared to the predicted exponential decays.

**Effect of aliasing.** We have claimed that aliasing effects are a major source of error in fixed-lagtime MSMs. In practice, implied timescales that are faster than the model lagtime will be aliased and tend to grow linearly with lagtime. For an analytical treatment of this problem, consider a two-state system. Suppose we estimate a transition matrix $\hat{T}(\tau)$, where $\tau$ is the lagtime. One can show that the nontrivial eigenvalue of $T$ is given by $\lambda(\tau) = T_{11}(\tau) + T_{22}(\tau) - 1$. The implied timescale is then given by

$$\hat{\tau} = -\frac{\tau}{\log(\lambda)} = \log(T_{11})^{-1} \log(T_{11} + T_{22} - 1)$$

In the limit of long lagtimes, we know that the self-transition probabilities will approach the equilibrium populations: $T_{ii} \rightarrow \pi_i$. However, there will remain some statistical noise, so that $\hat{T}_{11} + \hat{T}_{22} - 1 \approx \pi_1 + \pi_2 - 1 + e = e$. Thus, at large lagtimes, the argument of the logarithm is essentially a random variable centered near zero. Thus, the implied timescales will take the form

$$\hat{\tau} \approx \tau Z$$

where $Z$ is some (possibly complex) random variable. The prefactor of $\tau$ indicates that one expects linear growth in the implied timescales.

**Alanine dipeptide.** Fifty alanine dipeptide trajectories of length 500 ps were sampled at 1 ps intervals and clustered into 200 macrostates using Ward clustering with the RMSD metric. Macrostate relaxation timescales and their fluxes suggest that the relevant dynamics should be well-described by four relaxations (Fig. S2), leading us to build a five state model with PCCA+.

The resulting macrostates show good agreement with previously identified milestones (Fig. S1) With this five macrostate model, SCRE was used to estimate a rate matrix. The implied timescales, Chapman-Kolmogorov, and autocorrelation tests suggest that the resulting model accurately describes the dynamics at long timescales.

This alanine dipeptide data is included with MSMBuilder 2.5.0. The MSMBuilder tutorial walks users through the generation of the five-state macrostate model.

Fig. S1. PCCA+ was used to model the four slowest relaxations of alanine dipeptide, leading to five macrostates. A geometric state decomposition (4) is shown in black.
Fig. S2. Microstate implied timescales and eigenvector fluxes are shown for the FPCCA+ models analyzed in this work. In each system, we constructed models that capture the slowest and highest-flux timescales with a dual cutoff for timescale and flux; the modeled eigenvectors are shaded in gray. (A). Five-state model for alanine dipeptide. (B). Three-state model for HP35 (360 K). (C). Three state model for FIP35. (D). Three state model for GTT. (E). Five state model for BBA.
Fig. S3. Model validation for alanine dipeptide. (A) Macrostate implied timescales show agreement between SCRE and converged fixed-lagtime results. (B). Autocorrelation analysis shows agreement between MSM and raw trajectories. Error bars on autocorrelation functions are estimated as the results obtained from the first and second halves of the data (black circles). A different error estimate (black vertical lines) is given by assuming an exponential decay and approximating the autocorrelation function as a binomial experiment: $\sigma = \sqrt{p(1-p)}$, where $n$ is the total number of folding events and $p = e^{-\lambda t}$. (C) Self transition probabilities for fixed-lagtime and SCRE models show agreement at long lagtimes. Error bars on transition probabilities are estimated using a binomial approximation: $\sigma = \sqrt{p(1-p)}$. (D). Eigenvector autocorrelation functions for raw data and for MSM (e.g. exponential decays). Error bars on MSM predictions are estimated by exponential decays with timescales $\tau \pm \frac{\tau}{n}$, where $n$ is the total number of folding events observed in the dataset.
Fig. S4. Model validation for HP35. (A). Macrostate implied timescales show agreement between SCRE and converged fixed-lagtime results. (B). Autocorrelation analysis shows agreement between MSM and raw trajectories at long timescales. Error bars on autocorrelation functions are estimated as the results obtained from the first and second halves of the data (black circles). A different error estimate (black vertical lines) is given by assuming an exponential decay and approximating the autocorrelation function as a binomial experiment: $\sigma = \sqrt{\frac{n(1-p)}{n^2}}$, where $n$ is the total number of folding events and $p = e^{-\lambda t_{lag}}$. (C). Self transition probabilities for fixed-lagtime and SCRE models show agreement at long lagtimes. Error bars on transition probabilities are estimated using a binomial approximation: $\sigma = \sqrt{\frac{n(1-p)}{n^2}}$. (D). Eigenvector autocorrelation functions for raw data and for MSM (e.g. exponential decays). Error bars on MSM predictions are estimated by exponential decays with timescales $\tau \pm \frac{1}{\sqrt{2p}}$ where $n$ is the total number of folding events observed in the dataset.
Fig. S5. Model validation for FIP35. (A). Macrostate implied timescales show agreement between SCRE and converged fixed-lagtime results. (B). Autocorrelation analysis shows agreement between MSM and raw trajectories at long timescales. Error bars on autocorrelation functions are estimated as the results obtained from the first and second halves of the data (black circles). A different error estimate (black vertical lines) is given by assuming an exponential decay and approximating the autocorrelation function as a binomial experiment: $\sigma = \sqrt{n(1-p)}$, where $n$ is the total number of folding events and $p = e^{-\lambda t_{lag}}$. (C). Self transition probabilities for fixed-lagtime and SCRE models show agreement at long lagtimes. Error bars on transition probabilities are estimated using a binomial approximation: $\sigma = \sqrt{n(1-p)}$. (D). Eigenvector autocorrelation functions for raw data and for MSM (e.g., exponential decays). Error bars on MSM predictions are estimated by exponential decays with timescales $\tau = \frac{1}{\lambda t_{lag}}$ where $n$ is the total number of folding events observed in the dataset.
Fig. S6. Model validation for GTT. (A). Macrostate implied timescales show agreement between SCRE and converged fixed-lagtime results. (B). Autocorrelation analysis shows agreement between MSM and raw trajectories at long timescales. Error bars on autocorrelation functions are estimated as the results obtained from the first and second halves of the data (black circles). A different error estimate (black vertical lines) is given by assuming an exponential decay and approximating the autocorrelation function as a binomial experiment: $\sigma = \sqrt{\frac{p(1-p)}{n}}$, where $n$ is the total number of folding events and $p = e^{-\lambda t_{\text{lag}}}$. (C). Self transition probabilities for fixed-lagtime and SCRE models show agreement at long lagtimes. Error bars on transition probabilities are estimated using a binomial approximation: $\sigma = \sqrt{\frac{p(1-p)}{n}}$. (D). Eigenvector autocorrelation functions for raw data and for MSM (e.g. exponential decays). Error bars on MSM predictions are estimated by exponential decays with timescales $\tau \pm \frac{1}{\sqrt{n}}$, where $n$ is the total number of folding events observed in the dataset.
Fig. S7. Model validation for BBA. (A). Macrostate implied timescales show agreement between SCRE and converged fixed-lagtime results. (B). Autocorrelation analysis shows agreement between MSM and raw trajectories at long timescales. Error bars on autocorrelation functions are estimated as the results obtained from the first and second halves of the data (black circles). A different error estimate (black vertical lines) is given by assuming an exponential decay and approximating the autocorrelation function as a binomial experiment: $\sigma = \sqrt{\frac{1-p}{n}}$, where $n$ is the total number of folding events and $p = e^{-\lambda t_{\text{lag}}}$. (C). Self transition probabilities for fixed-lagtime and SCRE models show agreement at long lagtimes. Error bars on transition probabilities are estimated using a binomial approximation: $\sigma = \sqrt{\frac{1-p}{n}}$. (D). Eigenvector autocorrelation functions for raw data and for MSM (e.g. exponential decays). Error bars on MSM predictions are estimated by exponential decays with timescales $\tau \pm \frac{1}{\sqrt{n}}$ where $n$ is the total number of folding events observed in the dataset.
Fig. S8. Chemical shifts are calculated for the N, N', and U states of NTL9. For a handful of residues (those shown with error bars), the N and N' chemical shifts appear separated by a difference that is comparable to the systematic uncertainty in chemical shift estimation algorithms. These residues might be amenable to investigation by relaxation dispersion NMR experiment. (A). Estimates made using ShiftX2. (B). Estimates made using SPARTA+. Note that the estimates by ShiftX2 and SPARTA+ appear to be uncorrelated, so agreement between the two predictions leads to increased confidence. Plotted error bars depict the systematic uncertainty of SPARTA+ and ShiftX2 predictions.

Fig. S9. Ward, k-centers, and hybrid (1) microstate clusterings are compared for the HP35 (360 K) simulation. (A). The k-centers and hybrid clusterings show a large number of microstates with poor statistics, while the Ward clustering avoids this problem. (B). The Ward clustering produces implied timescales that are slower and more converged.

Table S1. The number of conformations, Ward microstates, frequency of data storage, and MSM lagtime are summarized for each system.

<table>
<thead>
<tr>
<th>System</th>
<th>$N_{\text{conf}}$</th>
<th>$N_{\text{microstates}}$</th>
<th>$\tau_{\text{output}}$ [μs]</th>
<th>$\tau_{\text{lag}}$ [μs]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chignolin</td>
<td>2139</td>
<td>400</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>2JO</td>
<td>4176</td>
<td>300</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>BBA</td>
<td>4443</td>
<td>550</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>HP35</td>
<td>12559</td>
<td>900</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>HP35 (300 K)</td>
<td>10267</td>
<td>730</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>FIP35</td>
<td>4000</td>
<td>290</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>GTT</td>
<td>13017</td>
<td>930</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>NTL9</td>
<td>43680</td>
<td>3120</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>BBL</td>
<td>8602</td>
<td>610</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>PRB</td>
<td>2081</td>
<td>150</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>UVP</td>
<td>6545</td>
<td>470</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>NuG2</td>
<td>23128</td>
<td>1650</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>A3D</td>
<td>14142</td>
<td>1010</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Lambda</td>
<td>12679</td>
<td>910</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>