Optimal Parameter Selection in Markov State Models for Biomolecular Conformational Dynamics

Christian R. Schwantes,[†,‖] Robert T. McGibbon,[†,‖] and Vijay S. Pande[*,†,‡,¶,§]

Department of Chemistry, Stanford University, Stanford CA 94305, USA, Biophysics Program, Stanford University, Stanford CA 94305, USA, Department of Computer Science, Stanford University, Stanford CA 94305, USA, and Department of Structural Biology, Stanford University, Stanford CA 94305, USA

E-mail: pande@stanford.edu

Abstract

Markov state models (MSMs) are a powerful tool for the analysis of molecular dynamics simulations, but have been hampered by the need for manual selection of the number of states. We report a new method for the optimal selection of the number of states in an MSM based on the Bayesian information criterion. We demonstrate the approach on three systems of increasing complexity...

Introduction

Protein are highly complex molecular machines, and their dynamics are an essential aspect of biomolecular function. These dynamics span a wide range of length scales, timescales and complexity, including folding and aggregation, conformational change between functional native substates, ligand binding, and allostery.[1–4] Whereas classical experimental probes have often been interpreted in two-state frameworks, ensemble measurements with increasingly high temporal resolution as well as sensitive single molecule probes have uncovered a vast array of complex multistate kinetics.[5,6] But the atomic-resolution characterization of these dynamics is often an enormous challenge – as molecular probes like Förster resonance energy transfer, small-angle x-ray scattering, and nuclear magnetic resonance techniques measure complex projections of the intrinsic structure, generally reporting simultaneously on many degrees of freedom.[7,8]

Computer simulations can complement experiments by providing atomic insight into conformational dynamics. With advances at the algorithmic, hardware, and software levels, modern molecular simulation paradigms, incorporating specialized or accelerated hardware, often in combination with highly parallel distributed computing frameworks, are capable of generating extensive simulation data sets.[9–13] In fact, the minimally-biased kinetic analysis of such simulations is often a central bottleneck and presents a major challenge to the field. The analysis paradigms often entail the construction of lower resolution models parametrized from the high resolution simulation data set which capture the essential features in an interpretable framework.[14,15] For example, by projecting the data down onto one or two degrees of freedom we create a simpler model for the system, such as one characterized by diffusion along a single reaction coordinate.[16]

Markov state models (MSMs) are one approach

[*]To whom correspondence should be addressed

[†]Department of Chemistry, Stanford University, Stanford CA 94305, USA

[‡]Biophysics Program, Stanford University, Stanford CA 94305, USA

[¶]Department of Computer Science, Stanford University, Stanford CA 94305, USA

[§]Department of Structural Biology, Stanford University, Stanford CA 94305, USA

[‖]Contributed equally to this work

for analyzing MD data sets and driving further MD simulations that are able to smoothly move between high and low-resolution models.[17–20] Such detailed models maintain quantitative agreement with the underlying simulation data, while low-resolution models capture the salient features of the potential energy landscape, sacrificing some degree of model complexity. In an MSM, the dynamics are modeled as a memory-less jump process between a discrete set of conformational states. The two key quantities which define the MSM are thus the state definitions, an indicator function basis over phase space, and the pairwise transition probabilities or transition rates, which parameterize the kinetics. The matrix of transition probabilities can be used to locate the systems transition paths,[21] and its dominant eigenvectors to identify the metastable states[22] and long-timescale dynamical modes.

A significant challenge in the automated construction of Markov state models is the choice of the number of states.[23] Although classical Hamiltonian dynamics form a continuous-time Markov chain in $\mathbb{R}^{6N}$, the Markov property does not hold after the projecting the dynamics onto a basis of discrete indicator functions. In particular, when states contain within them free energy barriers of substantial magnitude, the validity of the Markov assumption begins to suffer considerably. While this source of modeling error can be addressed by increasing the number of microstates, the reduction in one error comes at the expense of the increase in another. This second source of error is statistical in origin. As the number of states in the model grows, so does the number of parameters required to completely specify the kinetic model between all pairs of states. Because the amount of data is constant, each additional parameter leads to a decrease in the amount of data available per model parameter, which makes the approach susceptible to over-fitting.

Here, we seek to build models that are *suitably* complex, given the data, yielding complex descriptions of the system only to the extent that their additional parameters are implied by the observed dynamics. To that end, we introduce a procedure for scoring the likelihood of an MSM, which, together with standard statistical model selection techniques, enables the optimal selection of the state space, which we express both in terms of the number of states and the clustering algorithm employed to group sampled conformations into states. This approach complements validation procedures performed primarily based on human intuition, such as Chapman-Kolmogorov tests, and enables the treatment of model selection as an optimization problem amenable to automated methods.

Prior Work

In the context of MSMs, model selection has traditionally been performed by analyzing the self-consistency of the models with respect to changes in the Markov lag time. As suggested by Swope et al.[24], given a Markov model at a lag time of τ, $T(\tau)$, we expect that a model with a lag time of $n\tau$ should merely be the n^{th} power of $T(\tau)$:

$$T(n\tau) = T(\tau)^n$$

As such, the eigenvalues, and correspondingly the relaxation timescales, should remain constant as the lag time increases. The resulting test of a model's quality consists of calculating the eigenspectra of models built at many lag times and choosing models whose eigenvalues have converged.

In practice, this test has been useful for many,[25] however has some flaws. For instance, it is generally observed that most models' timescales do not converge, but steadily increase, so it is difficult to judge whether a particular model passes the test. Additionally, it is difficult to compare models built on two state decompositions, as neither may have convergent timescales. Often, the choice is made to use the model whose timescales better match an experimental observable.

Other evaluation strategies include a test based on comparison to a second-order Markov model.[26] Additionally, Bayesian methods that involve marginalizing over all possible MSMs have been developed.[27] However, these methods are often prohibitively expensive on larger datasets. In contrast, Kellogg et al.[28] proposed a tractable method for comparing MSMs built with different state decompositions, based on a likelihood function for

an MD trajectory given an MSM, discussed below.

Likelihood of a Markov State Model

Bayes' rule provides a foundation for model selection, by establishing the proportionality of the probability of a model given the data to the probability of the data given the model. Without a strong prior, model selection is reduced to the search for the model that maximizes the likelihood of the data.

Thus we seek to formulate this likelihood function for Markov state models. With the kinetic model expressed as a set of pairwise state to state transition probabilities at a given lag time, the likelihood of an ensemble of trajectories after projection into the MSM state space is given simply by the product of the transition matrix elements along the observed trajectories. However, as we vary the number of states, it is not permissible to simply compare these likelihoods as part of an optimization of the state definitions. In doing so, the optimal model would always be the trivial one state model, whose computed likelihood is unity regardless of the data.

The appropriate likelihood is instead a path action in phase space, on which the discrete states are merely an indicator function basis. With $s(\mathbf{x})$ as the function mapping conformations into the indicator function basis set, and $\{\mathbf{x}_t\}_{t=1}^N$ an observed trajectory, the likelihood can be written as:

$$P\left[\{\mathbf{x}_t\}_{t=1}^N\right] d\mathbf{x}^N = \prod_{t=1}^{N-1} T(s(\mathbf{x}_t) \to s(\mathbf{x}_{t+1})) \cdot \prod_{t=1}^{N} P(\mathbf{x}_t|s(\mathbf{x}_t)) \cdot d\mathbf{x}^N \tag{1}$$

With a discrete, non-overlapping state space, the likelihood of a sampled trajectory can be decomposed into a product of terms of two types: the state to state Markov transition probabilities, $T(s_i \to s_j) \equiv P(S_{t+1} = s_j | S_t = s_i)$, and so-called emission distributions of each state, $P(\mathbf{x}|s(\mathbf{x}))$, the conditional probability of observing a conformation at a given location in phase space given that the conformation, $\mathbf{x}$, is within a certain state, s_i. By convention, we define $P(\mathbf{x}|s(\mathbf{x}))$ to be zero for all $\mathbf{x}$ such that $s(\mathbf{x}) \neq s_i$.

This emission distribution can be potentially modeled in multiple ways, each of which, in combination with a transition matrix, parameterizes a different statistical model. Kellogg et al.[28] used a discriminative model with:

$$P(\mathbf{x}_t|s(\mathbf{x}_t)) = \left|\{\mathbf{x}_k : s(\mathbf{x}_k) = s(\mathbf{x}_t)\}\right|^{-1} \tag{2}$$

This model has the training dataset as its support rather than phase space, which makes it unable to generalize and assign probability to new data. In certain circumstances, this is an undesirable property. For example, protocols that involve fitting and validating models on separate datasets (e.g. cross-validation) are impossible when the statistical model lacks the capacity to describe new data.

Instead, we seek a generative model whose emission distributions are supported on phase space (i.e. standard distributions over $\mathbb{R}^{3N_{atoms}}$). We propose two possible emission distributions, $P(\mathbf{x}|s(\mathbf{x}))$: first, a uniform distribution over the (hyper)volume of each state, as shown in Eq. (3). This model has the advantage of appearing agnostic with respect to intra-state dynamics but is intractable in high dimensions when the states are defined by general polytopes, such as those produced by a data-driven Voronoi tessellation.

$$P(\mathbf{x}_t|s(\mathbf{x}_t)) = \frac{1}{V_{s(\mathbf{x}_t)}} \tag{3}$$

A more tractable alternative is a Gaussian emission model where $P(\mathbf{x}|s(\mathbf{x}))$ is modeled as multivariate normal Eq. (4).

$$P(\mathbf{x}_t|s(\mathbf{x}_t)) = \left[\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_{s(\mathbf{x}_t)}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{s(\mathbf{x}_t)})^T \Sigma^{-1}_{s(\mathbf{x}_t)} (\mathbf{x}_t - \boldsymbol{\mu}_{s(\mathbf{x}_t)})\right)\right] \tag{4}$$

where $\boldsymbol{\mu}_{s(\mathbf{x}_t)}$ is the cluster center assigned to $\mathbf{x}_t$ and $\Sigma_{s(\mathbf{x}_t)}$ is the corresponding covariance matrix. For simplicity, we follow Pelleg and Moore[29] and employ a spherical Gaussian model with a single shared variance parameter across the states esti-

mated by maximum likelihood, under which it reduces to:

$$P(\mathbf{x}_t|s(\mathbf{x}_t)) = \frac{1}{(2\pi\hat{\sigma}^2)^{\frac{d}{2}}} \exp\left(-\frac{||\mathbf{x}_t - \boldsymbol{\mu}_{s(\mathbf{x}_t)}||^2}{2\hat{\sigma}^2}\right) \tag{5}$$

where

$$\hat{\sigma}^2 = \frac{1}{N-k}\sum_{t=1}^{N} ||\mathbf{x}_t - \boldsymbol{\mu}_{s(\mathbf{x}_t)}||^2 \tag{6}$$

and k is the number of states.

Since the likelihood requires the emission distributions to have zero overlap – such that each conformation only has nonzero emission probability from a single state, the nonzero overlap between the Gaussian emission distributions introduces a truncation error. One approach to ameliorate this issue is to truncate the density functions at the state boundaries, and renormalize. We assume that the overlaps are small, and use the density defined in Eq. (5). An alternative approach for relaxing this approximation is possible with the forward-backward algorithm,[30] but is beyond the scope of this work.

Cross Validation and the Bayes Information Criterion

Likelihood maximization is insufficient for model selection when the number of parameters varies between proposed models, as more complex models generally exhibit higher empirical likelihoods, often at the cost of larger generalization errors due to overfitting.[31,32] Statistical learning theory provides a number of alternative approaches for this problem. Conceptually, the most straightforward is a full Bayesian treatment in which all unknown model parameters are represented by probability distributions. The evidence for a model is computed by formally integrating over the model parameters and the evidence ratio, or Bayes factor,[33] then provides a rigorous basis of model selection that appropriately punishes overly complex models as they become poorly constrained in parameter space. Unfortunately such approaches are intractable for problems of this size because of the need to integrate over all possible Markov models of a given size.

Instead, we explore three alternative procedures for choosing the number of states in an MSM: cross validation, Schwarz's Bayesian information criterion (BIC)[34] and the Akaike information criterion (AIC).[35] Cross-validation attempts to directly measure the Markov model's generalization error. First, the model is parameterized by building both the state space and transition matrix on a subset of the available molecular dynamics trajectories, then the likelihood is evaluated on the left-out portion. This scheme can be repeated many times with different partitions of the dataset. The AIC and BIC are augmented likelihood function approaches which do not require leaving out portions of the available data during fitting, and instead employ asymptotic approximations to the Bayesian evidence to directly penalize models with additional free parameters.

$$\text{BIC} \equiv -2 \cdot \ln L + \kappa \cdot \ln N \tag{7}$$

$$\text{AIC} \equiv -2 \cdot \ln L + 2 \cdot \kappa \tag{8}$$

where L is the likelihood, κ is the number of free parameters, and N is the number of data points, assumed to be independent and identically distributed. Model selection is performed by a minimization of the criterion.

Computational Methods

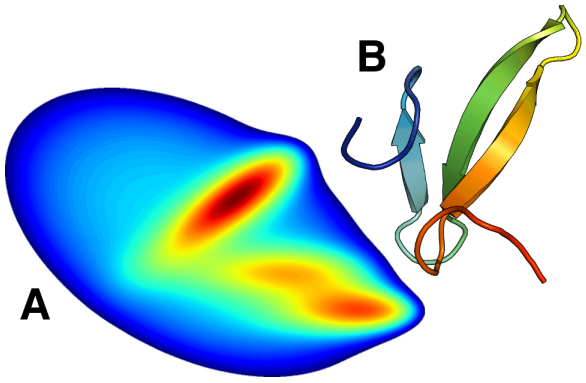

Figure 1: Systems studied in this work. (A) Langevin dynamics on the two dimensional Müller potential.[36] (b) 200 μs of dynamics of the Fip35 WW domain,[37] courtesy of D.E. Shaw research.[38]

The uniform distribution emission model

presents a computational challenge: its use requires the calculation of the (hyper)volume of the MSM's states, which, when defined by clustering, are high-dimensional Voronoi cells. While trivial in two or three dimensions, this computational geometry task becomes challenging in high-dimensional settings. The computation of such volumes has occupied significant attention in recent years in the computational geometry literature, especially via randomized algorithms.[39–41] We opt to approximate the volumes using naive Monte Carlo rejection sampling, which we find tractable for large systems only when the molecular dynamics dataset is first projected into a suitable small vector space of up to perhaps ten dimensions.

A further challenge is the description of the states that are at the edge of the MSM – whose Voronoi cells extend to infinity in at least one direction. In these cases, the Voronoi cells are of unbounded volume. Instead we wish to truncate these states by bounding them by the convex hull of the dataset. Because the convex hull of our simulation datasets are computationally inaccessible, we define the outer extent of our data sets to be the set of all trial points such that the nearest sampled conformation to the trial point is closer than a certain cutoff, R.

For cross validation, we find it essential to use a dense prior during transition matrix estimation. Because the maximum likelihood transition matrix fit on a subset of the data assigns probability zero to unobserved transitions, we observe cross validation log likelihoods of negative infinity in all but the most data-rich regimes. We used a Dirichlet prior with a pseudo-count of $\frac{1}{k}$ for each possible transition, where k is the number of states in the model.

Results and Discussion

Müller Potential

We simulated two trajectories with 10^6 steps of Langevin dynamics on the Müller potential.[36] Simulations were performed with a timestep $t = 0.1$, friction coefficient of $\gamma = 10^3$, and $kT = 15$. The first trajectory was clustered using the k-centers clustering algorithm, and state volumes were computed for the uniform emission model using 10^5 rounds of Monte Carlo rejection sampling with a cutoff of $R = 0.28$ from the 50 state model's centroids. The second trajectory was used as a test set. All MSMs were built using a lag time of $30 steps$.

The volumes did not change drastically upon doing ten times more Monte Carlo rejection sampling, and the likelihood remained essentially constant, which indicates that the use of approximation volumes is sufficient to obtain robust estimates of the likelihood.

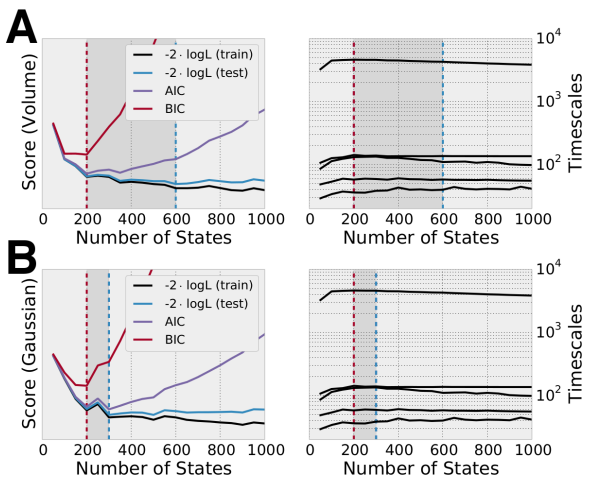

Figure 2: We calculated the training and test set log-likelihoods and model selection criteria for models built with 50-1000 states using both the uniform (**A**) and multivariate normal (**B**) emission distributions. Dashed lines represent the optimal model as defined by the AIC, BIC, or test set log-likelihood and form the boundaries of the optimal window (shaded dark gray). For both likelihoods, the BIC penalizes complexity more than the AIC and test set log-likelihood. The optimal window is consistent with the convergence of the implied timescales (right frames).

As shown in Fig. 2, models built with too few states achieve a drastically reduced likelihood, but above a threshold region the likelihood increases relatively slowly. The penalty on the number of parameters in Eq. (7) and Eq. (8) begins to dominate. The optimal models, according to the three methods are between 200 and 600 states for this system, which is consistent with the convergence of the relaxation timescales of the models.

The AIC and BIC penalize the larger state models much more than the test set log-likelihood. On the two-dimensional potential, in our data-rich regime it is difficult to produce an over-fit model. Cross-validation requires fitting multiple models on subsets of the data, and so is not feasible for larger systems in the data-poor regime.

We also tested the Gaussian likelihood on MSMs built using k-means clustering on the Müller potential. The training and test set log-likelihoods, BIC, and AIC are all optimized by a five state model (Fig. 3). The decrease in the training set log-likelihood can be explained by a trade-off inherent in Eq. (1), where adding more states leads to an improvement in the emission distribution term but causes the transition matrix term to suffer. This trade-off illustrates a deficiency in the step-wise approach in which neither the transition matrix nor the state space is parameterized to explicitly optimize the full Eq. (1).

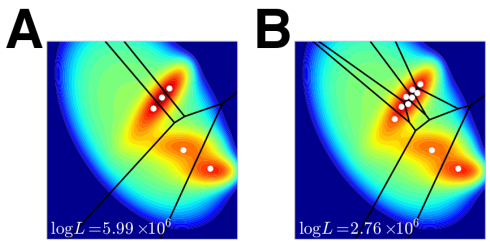

Figure 3: MSMs were built using k-means clustering on the Müller potential simulations. Since k-means optimizes an objective function that is related to the emission probability, our likelihood picks only a few states as the most likely model. The likelihood decreases as new states are introduced, due to a competition between the transition matrix term and emission distribution term in Eq. (1).

Fip35 WW Domain

We reanalyzed two ultra-long 100 μs molecular dynamics trajectories of the Fip35 WW domain,[37] provided courtesy of D.E. Shaw Research.[38] Because the likelihood calculations involving the uniform distribution emission model have exponential complexity in the dimensionality of the state space we first preprocess the trajectories with time-structure based independent components analysis (tICA),[42,43] extracting only the four slowest uncorrelated linear combinations of residue-residue distances. All MSMs were built using a lag time of 50ns.

The complexity of evaluating the Gaussian emission model likelihood does not increase with respect to dimension. Thus we used the Gaussian emission model with both the four-dimensional tICA models and those built using k-centers clustering with the minimal cartesian root-mean squared deviation (RMSD) over Fip35's 252 non-redundant heavy atoms, computed using the quaternion characteristic polynomial method.[44] For evaluating the Gaussian emission likelihood, we used a value of $3 \cdot 252 - 6 = 750$ for the dimension of the emission distribution.

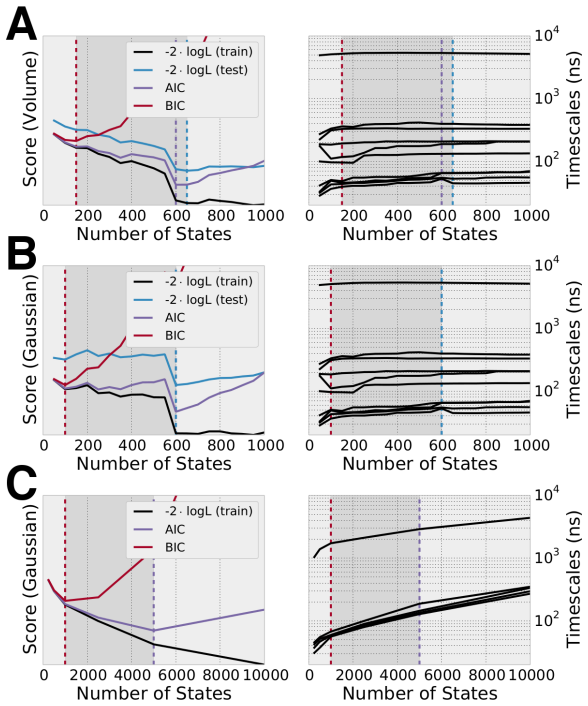

Figure 4: On models whose states were defined after projecting into the four-dimensional tICA subspace, we computed likelihoods based on the uniform (**A**) and Gaussian (**B**) emission models. Models built using the RMSD metric were evaluated with the Gaussian emission model (**C**). Dashed lines represent the optimal model as defined by the AIC, BIC, or test set log-likelihood and form the boundaries of the optimal window (shaded dark gray). Test set log-likelihoods are omitted in **C** as the dense matrix algebra is intractable for $k > 10^3$.

For the models built in the tICA subspace, the

AIC is comparable to the test set log-likelihood scores using either emission model, displaying a optimum at 600 states (Fig. 4). The BIC penalizes complexity more strongly with an optimum at 150 states. This window is consistent with the convergence of the implied timescales and the onset of Markovian behavior. Models built with RMSD require significantly more states than the tICA MSMs and show poor timescale convergence. These results indicate why model selection based on implied timescales is so difficult, as a less than optimal state decomposition can result in timescales that do not converge rapidly (if at all).

The two 100 μs molecular dynamics trajectories of the Fip35 WW domain simulated by Shaw et al.[38] have been used to parameterize multiple Markov state models. Lane et al.[45] proposed a 26,000 state MSM after clustering the data by RMSD, whereas Kellogg et al.,[28] clustering on contact maps or secondary structure, proposed 100 and 175 state models, respectively. Consistent with this spread, we find the optimal number of states to be highly dependent on the space in which the data is clustered. With RMSD, the Gaussian likelihood indicates that Markov models built with on the order of 5,000 states optimally balance model quality and generalization error. On the other hand, the tICA method's identification of slow order parameters in the simulations has the effect of reducing the number of states necessary when constructing Markov models. Even with 200-600 states, Markov models constructed using tICA are able to accurately identify the simulations folding process on the 1-10 μs timescale, as well as the faster near-native dynamics identified in this dataset by McGibbon and Pande.[23]

Limitations and Future Work

The likelihood functions described herein permit the comparison of Markov state models with varying number of states, however, they require that the compared models have the same support. As such, a direct comparison of the likelihoods between models built with tICA and models built with RMSD is not possible. Furthermore, the uniform emission model is intractable for all but the simplest systems. The Gaussian emission model does not have this limitation.

In the future, we plan to extend this work to the consideration of models without discrete states, where the requirement that states strictly partition phase space into a set of discrete indicator functions is relaxed and the models are parameterized by a direct optimization of Eq. (1). This strategy would complement approaches that generalize MSMs beyond discrete states.[?]

Conclusions

Markov State Models are powerful and popular frameworks for the analysis of molecular simulations, with a growing literature and two open source software tools.[46,47] There are, however a number of steps in the construction process that require hand-tuning, which limits the use of MSMs to experts and introduces significant biases into the model building process. Additionally, the ability to automatically construct MSMs on the fly while simulations are in progress, an important point for so-called adaptive sampling procedures,[48] is hampered when manual model selection is required.

These results take a step toward fully automating the MSM construction process and controlling the complexity and generalization error by providing a quantitative method for selecting the most suitably complex model, given the data.

References

(1) Dobson, C. M. Protein folding and misfolding. *Nature* **2003**, *426*, 884–890.

(2) Kim, S.; Born, B.; Havenith, M.; Gruebele, M. Real-Time Detection of Protein–Water Dynamics upon Protein Folding by Terahertz Absorption Spectroscopy. *Angew. Chem. Int. Ed.* **2008**, *47*, 6486–6489.

(3) Austin, R. H.; Beeson, K. W.; Eisenstein, L.; Frauenfelder, H.; Gunsalus, I. C. Dynamics of ligand binding to myoglobin. *Biochemistry* **1975**, *14*, 5355–5373.

(4) Bahar, I.; Chennubhotla, C.; Tobi, D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr. Opin. Struct. Biol.* **2007**, *17*, 633 – 640.

(5) Cosa, G.; Zeng, Y.; Liu, H.-W.; Landes, C. F.; Makarov, D. E.; Musier-Forsyth, K.; Barbara, P. F. Evidence for Non-Two-State Kinetics in the Nucleocapsid Protein Chaperoned Opening of DNA Hairpins. *J. Phys. Chem. B* **2006**, *110*, 2419–2426.

(6) Zhang, X.; Lam, V. Q.; Mou, Y.; Kimura, T.; Chung, J.; Chandrasekar, S.; Winkler, J. R.; Mayo, S. L.; Shan, S.-o. Direct visualization reveals dynamics of a transient intermediate during protein assembly. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 6450–6455.

(7) Mertens, H. D.; Svergun, D. I. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J. Struct. Biol.* **2010**, *172*, 128 – 141.

(8) Tzeng, S.-R.; Kalodimos, C. G. Protein dynamics and allostery: an {NMR} view. *Curr. Opin. Struct. Biol.* **2011**, *21*, 62 – 67.

(9) Eastman, P. et al. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9*, 461–469.

(10) Shirts, M.; Pande, V. S. Screen Savers of the World Unite! *Science* **2000**, *290*.

(11) Shaw, D. et al. *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on*; 2009; pp 1–11.

(12) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 116–122.

(13) Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; De Fabritiis, G. High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing. *J. Chem. Inf. Model.* **2010**, *50*, 397–403, PMID: 20199097.

(14) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in protein folding simulations: Timescale, representation, and analysis. *Nat. Phys.* **2010**, *6*.

(15) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58 – 65.

(16) Best, R. B.; Hummer, G. Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 1088–1093.

(17) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything you wanted to know about Markov state models but were afraid to ask. *Methods* **2010**, *52*, 99 – 105.

(18) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. Simple few-state models reveal hidden complexity in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17807–17813.

(19) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(20) Bowman, G. R.; Meng, L.; Huang, X. Quantitative comparison of alternative methods for coarse-graining biological networks. *J. Chem. Phys.* **2013**, *139*, 121905.

(21) Weinan, E.; Vanden-Eijnden, E. Towards a theory of transition paths. *J. Stat. Phys.* **2006**, *123*, 503–523.

(22) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra and its Applications* **2000**, *315*, 39–59.

(23) McGibbon, R. T.; Pande, V. S. Learning Kinetic Distance Metrics for Markov State Models of Protein Conformational Dynamics. *J. Chem. Theory Comput.* **2013**, *9*, 2900–2906.

(24) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T.

J. C.; Zhestkov, Y.; Zhou, R. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 2. Example Applications to Alanine Dipeptide and a β-Hairpin Peptide. *The Journal of Physical Chemistry B* **2004**, *108*, 6582–6594.

(25) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences* **2009**, *106*, 19011–19016.

(26) Park, S.; Pande, V. S. Validation of Markov state models using Shannon's entropy. *J. Chem. Phys.* **2006**, *124*.

(27) (a) Bacallado, S.; Chodera, J. D.; Pande, V. Bayesian comparison of Markov models of molecular dynamics with detailed balance constraint. *The Journal of Chemical Physics* **2009**, *131*, –; (b) Bacallado, S. Bayesian analysis of variable-order, reversible Markov chains. *Ann. Stat.* **2011**, *39*, 838–864.

(28) Kellogg, E. H.; Lange, O. F.; Baker, D. Evaluation and Optimization of Discrete State Models of Protein Folding. *J. Phys. Chem. B* **2012**, *116*, 11405–11413.

(29) Pelleg, D.; Moore, A. W. *ICML*; 2000; pp 727–734.

(30) Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **1989**, *77*, 257–286.

(31) Liddle, A. R. Information criteria for astrophysical model selection. *Mon. Not. R. Astron. Soc. Lett.* **2007**, *377*, L74–L78.

(32) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer New York Inc.: New York, NY, USA, 2001.

(33) Gelfand, A. E.; Dey, D. K. Bayesian Model Choice: Asymptotics and Exact Calculations. *J. R. Statistic. Soc. B* **1994**, *56*, pp. 501–514.

(34) Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464.

(35) Akaike, H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **1974**, *19*, 716–723.

(36) Müller, K. Reaction Paths on Multidimensional Energy Hypersurfaces. *Angew. Chem.* **1980**, *19*, 1–13.

(37) Liu, F.; Du, D.; Fuller, A. A.; Davoren, J. E.; Wipf, P.; Kelly, J. W.; Gruebele, M. An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 2369–2374.

(38) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341–346.

(39) Kannan, R.; Lovász, L.; Simonovits, M. Random walks and an O*(n5) volume algorithm for convex bodies. *Random Structures & Algorithms* **1997**, *11*, 1–50.

(40) Simonovits, M. How to compute the volume in high dimension? *Math. Program.* **2003**, *97*, 337–374.

(41) Lovász, L.; Vempala, S. Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm. *J. Comput. System Sci.* **2006**, *72*, 392 – 417.

(42) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009.

(43) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics* **2013**, *139*, –.

(44) Theobald, D. L. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallographica Section A* **2005**, *61*, 478–480.

(45) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *J. Am. Chem. Soc.* **2011**, *133*, 18413–18419.

(46) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale. *Journal of chemical theory and computation* **2011**, *7*, 3412–3419.

(47) Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S.; Schütte, C.; Noé, F. EMMA: A Software Package for Markov Model Building and Analysis. *Journal of Chemical Theory and Computation* **2012**, *8*, 2223–2238.

(48) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced modeling via network theory: Adaptive sampling of markov state models. *Journal of chemical theory and computation* **2010**, *6*, 787–794.