Statistical Model Selection for Markov Models of Biomolecular Dynamics

Robert T. McGibbon,[†,‖] Christian R. Schwantes,[†,‖] 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

[*] 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

Abstract

Markov state models provide a powerful framework for the analysis of biomolecular conformation dynamics in terms of their metastable states and transition rates. These models provide both a quantitative and comprehensible description of the long-timescale dynamics of large molecular dynamics with a Master equation, and have been successfully used to study protein folding, protein conformational change, and protein-ligand binding. However, to achieve satisfactory performance, existing methodologies often require expert intervention when defining the model's discrete state space. While standard model selection methodologies focus on the minimization of systematic bias and disregard statistical error, we show that by consideration of the states' conditional distribution over conformations, both sources of error can be balanced evenhandedly. Application of techniques that consider both systematic bias and statistical error on two $100\mu s$ molecular dynamics trajectories of the Fip35 WW domain shows agreement with existing techniques based on self-consistency of the model's relaxation timescales, with more suitable results in regimes in which those timescale-based techniques encourage over-fitting. By removing the need for expert tuning, these methods should reduce modeling bias and lower the barriers to entry in Markov state model construction.

KEYWORDS: MSM, molecular dynamics, AIC, BIC, Bayes factor, model selection

Introduction

Proteins 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.[7–9]

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 ensembles with relative ease.[10–15] A central challenge in the field of molecular simulations is now often the kinetic analysis, which typically involves constructing a lower resolution model that captures the system's essential features in an interpretable framework.[16,17] 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.[18]

Markov state models (MSMs) are one approach for analyzing MD data sets that are able to smoothly move between high and low-resolution models.[19–22] High-resolution 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 which partition the phase space. Two key quantities that define the MSM are thus the state definitions, an indicator function basis over phase space, and the pairwise transition probabilities, which parameterize the kinetics. The matrix of transition probabilities can be used to locate the systems transition paths, and its dominant eigenvectors identify the metastable states and long-timescale dynamical modes.[23,24]

Two competing sources of error govern the accuracy of an MSM. The first source of error is systematic and is common to all dimensionality reduction methods, in that simplified representations of complex systems discard information. The second source of error is statistical in nature, in that the parameters of an MSM must be estimated from a finite number of stochastic MD trajectories. As we discuss below, these two sources of error compete in a bias-variance tradeoff.[25] Given a finite size data set, as the number of states increases, the systematic error (the bias) decreases, while the statistical error (the variance) increases.

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 a balanced selection of the model's parameters.

Theory

Markov State Models

Let $\mathbf{x}_t$ be a ergodic time-homogeneous Markov process the phase space Ω with stationary density $\mu(\mathbf{x})$. Consider an ensemble of such processes at time t, described by a distribution $p_t(\mathbf{x})$. There exists an operator, termed the propagator, that evolves the a distribution forward in time at discrete intervals, τ.

$$p_{t+\tau}(\mathbf{x}) = \mathscr{Q}^{(\tau)} \circ p_t(\mathbf{x}) \tag{1}$$

The eigenfunctions of the propagator form a basis for $p(\mathbf{x})$, and the time evolution of the ensemble after n applications of the propagator can be decomposed into a sum of terms along each eigenfunction.

$$p_{t+n\tau}(\mathbf{x}) = \sum_{i=1}^{\infty} \exp\left(-\frac{n \cdot \tau}{t_i}\right) \langle \phi_i, p_t \rangle_{\mu^{-1}} \phi_i(\mathbf{x}), \tag{2}$$

where $t_i = -\dfrac{\tau}{\ln \lambda_i}$ are the propagator's relaxation timescales and ϕ_i and λ_i are the eigenfunctions and eigenvalues of the propagator, and $\langle \cdot, \cdot \rangle_{\mu^{-1}}$ denotes the inner product defined as

$$\langle \phi_i, p_t \rangle_{\mu^{-1}} = \int_{\Omega} d\mathbf{x}\; \phi_i(\mathbf{x}) p_t(\mathbf{x}) \frac{1}{\mu(\mathbf{x})}. \tag{3}$$

For a more detailed discussion of the propagator and its properties, we refer the reader to Prinz et al.[20].

For classical MD, $\mathscr{Q}^{(\tau)}$ is determined by the Hamiltonian and equations of motion.[26] However, it is not practical to explicitly construct the propagator from the Hamiltonian for complex systems. A Markov model provides an estimate for the propagator. A k-state Markov model is defined on a discrete set of states, $S = \{s_i\}_{i=1}^{k}$, where $s_i \subseteq \Omega$ and $\bigcup_{i=1}^{k} s_i = \Omega$. Furthermore, $s_i \cap s_j = \emptyset$ for all $i \neq j$. In words, every point in Ω is assigned to one (and only one) state in the MSM. Let $\sigma(\mathbf{x})$ be the function that maps a point $\mathbf{x} \in \Omega$ to the index i of the state in S such that $x \in s_i$.

The MSM models the dynamics of the propagator with a Markov jump process on $\{1,\ldots,k\}$. Let $\mathbf{p} \in (\mathbb{R} \cap [0,1])^k$ be a column vector whose entries sum to one. The elements in $\mathbf{p}$ are defined by a coarse-graining of $p(\mathbf{x})$ over S, as

$$p_i = \int_{\mathbf{x} \in s_i} d\mathbf{x}\, p(\mathbf{x}). \tag{4}$$

Consider a probability vector at time t denoted $\mathbf{p}(t)$. The time-evolution of $\mathbf{p}(t)$ is described by

$$\mathbf{p}(t+\tau)^T = \mathbf{p}(t)^T \mathbf{T}, \tag{5}$$

where T_{ij} is the probability of transiting to state j in time τ given that the system started in state i. With this construction, the eigenvalues and eigenvectors of $\mathbf{T}$ provide approximations for the eigenvalues and eigenfunctions of $\mathscr{Q}$. In a direct analogy with Eq. (2), the eigenvalues λ_i of T correspond to relaxation timescales of the Markov model

$$t_i = -\frac{\tau}{\ln \lambda_i}. \tag{6}$$

Let D be a set of molecular dynamics trajectories in Ω. To construct an MSM, both S and $\mathbf{T}$ must be determined from the data. The state-space, S, can be constructed by clustering the points in D or via a grid-based discretization of Ω. Given S, maximum-likelihood estimators for the $O(k^2)$ elements of a reversible transition probability matrix $\mathbf{T}$ exist.[20,27]

Using this procedure, it has been shown that MSM estimates of the eigenvalues λ_i are systematically underestimated in the limit of infinite data, and that the magnitude of this bias goes to zero as $k \to \infty$. This fact has inspired the development of variational methods for model selection, which optimize the MSM parameters in order to maximize the eigenvalues.[28]

Selecting the size of S is an important step in model construction, however there is not currently a widely-accepted method in use. Heuristic criteria, such as manually selecting $k = n_{\text{conformation}}/14$ or clustering D such that each conformation is within a pre-specified distance to its cluster center (e.g. 4.5Å or 1.2Å RMSD) have been previously employed.[21,29,30] Selection of τ based convergence of the relaxation timescales (Eq. (6)), as suggested by Swope et al.[31] and others is more well established. The rate of this convergence with respect to τ at fixed S has also been used as a metric for evaluating a model's state decomposition.[32]

The Bias-Variance Dilemma

For MSM methods to be valuable to practitioners, they must operate in limited data regimes. Thus, we seek a method for model selection that is sensitive to both the systematic error discussed above, as well as the statistical error associated with estimating the model's parameters. Whereas model selection based on the variational principle described above would tend to increase the number of states without bound, as the magnitude of the eigenvalue bias decreases with the number of states, practical application of this criteria would cause a catastrophic increase in statistical uncertainty in the model. This tradeoff between low-variance (small k) but biased estimators, and high-variance (large k) but unbiased estimators is a general feature of problems in statistical learning.[25] For example, it can be shown that the uncertainty in the posterior distribution of $\mathbf{T}$ given S and D approaches infinity as k approaches infinity.

Introduction to MSMs via the propagator theory motivates the view of the parametrization problem as one of *function estimation*. This perspective has a critical disadvantage: the natural error metric – the difference between the estimated MSM and true propagator is intrinsically uncomputable. The only access to the true propagator is via samples from the Hamiltonian with MD. Therefore, a compelling alternative is to view the problem as one of *prediction*. Since the MSM is a generative model for trajectories in Ω, it is possible to sample “pseudo-trajectories" directly from the MSM. If the MSM were perfect, these “pseudo-trajectories" would be indistinguishable from trajectories generated by the true propagator. Thus, the probability assigned by the MSM, to samples from the true propagator is a measure of the MSM’s accuracy. This argument is formalized by Bayes’ rule, which establishes a proportionality between the probability of a model given data to the probability of the data given the model

$$P\Big(\{S,\mathbf{T}\}\ \Big|\ D\Big) = P\Big(D\ \Big|\ \{S,\mathbf{T}\}\Big)\frac{P(\{S,\mathbf{T}\})}{P(D)}, \tag{7}$$

where $P\Big(\{S,\mathbf{T}\}\ \Big|\ D\Big)$ is the posterior probability of an MSM, $P\Big(D\ \Big|\ \{S,\mathbf{T}\}\Big)$ is the *likelihood* of the data given a model, $P(\{S,\mathbf{T}\})$ is the prior probability of an MSM before observing any data, and $P(D)$ is the probability of the data, a constant.

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. With $\{\mathbf{x}_t\}_{t=1}^N$ an observed trajectory of length N, the likelihood can be written as:

$$P\big(\{\mathbf{x}_t\}_{t=1}^N\ \big|\ \{S,T\}\big)d\mathbf{x}^N = \pi_{\sigma(\mathbf{x}_1)}\prod_{t=1}^{N-1} T_{\sigma(\mathbf{x}_t),\sigma(\mathbf{x}_{t+1})}\cdot\prod_{t=1}^{N} P(\mathbf{x}_t|\sigma(\mathbf{x}_t))\cdot d\mathbf{x}^N, \tag{8}$$

where π_i is the stationary probability of state i according to $\mathbf{T}$ and $P(\mathbf{x}_t|\sigma(\mathbf{x}_t))$ is the probability of sampling a conformation $\mathbf{x}_t$ given that the process is in the state $\sigma(\mathbf{x}_t)$, referred to as the emission distribution of state $\sigma(\mathbf{x}_t)$. There is some flexibility in the choice of emission distribution, though Eq. (8) is valid only when the emission distributions satisfy

$$P(\mathbf{x}|i) = 0 \quad \forall\,\mathbf{x} \text{ s.t. } \sigma(\mathbf{x}) \neq i. \tag{9}$$

An alternative approach involving fuzzy clustering, where conformations are not uniquely assigned to a single state, would not impose the constraint in Eq. (9), and would also involve a likelihood function that marginalizes over all possible paths.[33] In this work, we consider only models based on crisp partitioning. As such, given the coarse-graining of $p(\mathbf{x})$ in Eq. (4), the most natural choice emission distributions are normalized indicator functions

$$P(\mathbf{x} \mid i) = \frac{1}{V_i} \mathbf{1}_{s_i}(\mathbf{x}). \tag{10}$$

An alternative emission distribution was proposed by Kellogg et al.[34] Instead of having support on Ω, this distribution is supported on the training set $\{\mathbf{x}_t\}_{t=1}^{N}$. It is given by

$$P(\mathbf{x} \mid i) = \left|\{\mathbf{x}_k : \sigma(\mathbf{x}_k) = \sigma(\mathbf{x})\}\right|^{-1} \sum_{t=1}^{N} \delta(\mathbf{x} - \mathbf{x}_t), \tag{11}$$

where $|\cdot|$ denotes the cardinality of a set and $\delta(\cdot)$ is the Dirac delta function. As this is a discriminative model, it is 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 data sets (e.g. cross-validation) are impossible when the statistical model lacks the capacity to describe new data.

Statistical Model Selection

Likelihood maximization on the training data (i.e. the data used to fit the model) 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 over-fitting.[35,36] Statistical learning theory provides a number of methods to overcome 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 k-state model is computed by formally integrating Eq. (7) over the model parameters and the evidence ratio, or Bayes factor,[37] 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 impractical 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)[38] and the Akaike information criterion (AIC).[39] 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 data set.

$$\mathrm{BIC} \equiv -2 \cdot \ln L + (\ln N) \cdot \kappa \tag{12}$$

$$\mathrm{AIC} \equiv -2 \cdot \ln L + (2) \cdot \kappa \tag{13}$$

where L is the maximum likelihood, κ is the number of free parameters, and N is the number of data points. Model selection is performed by a minimization of the criterion. The BIC is derived based on the dominant terms in the Laplace approximation to the logarithm of the Bayes factor with a vague prior, while the functional form of the AIC comes from an asymptotic approximation to the Kullback-Leibler divergence between the model and the true distribution.[40,41] The appearance of number of data points N in Eq. (12) assumes that the data is independent and identically distributed given the model, which is often poorly justified for time series. Nonetheless, the value in the AIC and BIC comes from their simple form, that they do not require leaving out portions of the available data during fitting, and acceptable performance in practice.[42]

Methods

Volume Estimation

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.[43–45] We opt to approximate the volumes using Monte Carlo integration, which we find tractable for large systems only when the molecular dynamics data set 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 data set. Because the convex hull of our simulation data sets are computationally inaccessible, we defined the accessible volume of Ω denoted A, to be the set of all points within a distance R within a set of test points Y.

Algorithm 1 Monte Carlo Estimation of the State Volumes

```
procedure MC VOLUME ESTIMATION($\sigma$,$A$,$M$)
    $\mathbf{c} \leftarrow [0, 0, \ldots, 0]$                    ▷ c is a vector of length $k$
    while $\sum_i c_i \leq M$ do
        $\mathbf{x} \leftarrow$ sample from axis-aligned hyper cube containing $A$
        if $\mathbf{x} \in A$ then
            $c_{\sigma(\mathbf{x})} \leftarrow c_{\sigma(\mathbf{x})} + 1$
        end if
    end while
    $\mathbf{v} \leftarrow \mathbf{c} \cdot (\sum_{i=1}^{k} c_i)^{-1}$      ▷ v contains the relative volumes of each state
    return v
end procedure
```

The result of this scheme produces the volumes of each state relative to the volume of A. It's important, therefore, to use the same set A when comparing multiple models.

Simulation Protocol and MSM Construction

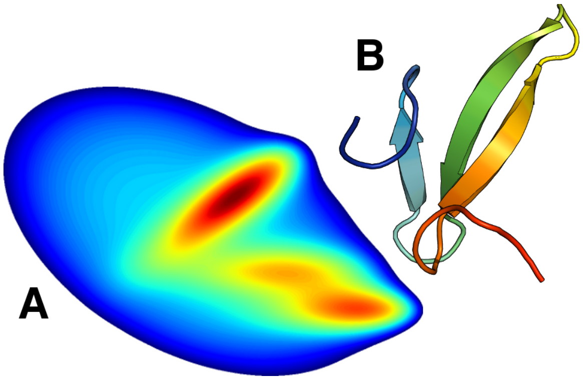

Figure 1: Systems studied in this work. (A) Brownian dynamics on the two-dimensional Müller potential.[46] (b) 200 μs of dynamics of the Fip35 WW domain,[47] courtesy of D.E. Shaw research.[48]

Two systems were investigated using this likelihood scheme (Fig. 1). The first was a simple two dimensional surface with three energy minima, called the Müller potential. The dynamics are governed by

$$\frac{d\mathbf{x}}{dt} = -\nabla V(\mathbf{x})\zeta + \sqrt{2kT\zeta}R(t),$$

where $\zeta = 10^{-3}$, $kT = 15$, $R(t)$ is a delta-correlated Gaussian process with zero mean, and $V(\mathbf{x})$ was defined as

$$V(\mathbf{x}) = \sum_{j=1}^{4} A_j \cdot \exp\left(a_j(x_1 - X_j)^2 + b_j(x_1 - X_j)(x_2 - Y_j) + c_j(x_2 - Y_j)^2\right),$$

where $a = (-1,-1,-6.5,0.7)$; $b = (0,0,11,0.6)$; $c = (-10,-10,-6.5,0.7)$; $A = (-200,-100,-170,15)$; $X = (1,0,-0.5,-1)$; $Y = (0,0.5,1.5,1)$ as suggested by Müller and Brown[46]. Using the Euler-Maruyama method and a time step of 0.1, we produced two trajectories of length 10^6 time steps. The initial positions were sampled via a uniform distribution over the box: $[-1.5,2.0] \times [-0.2,2.0]$.

The first trajectory was clustered using the k-centers clustering algorithm with the Euclidean distance. State volumes were computed for the uniform emission model using $M = 10^5$ rounds of Monte Carlo integration defining the set A using the cluster centers from the 50-state model as the test points Y with a cutoff of $R = 0.28$. The second trajectory was used as a test set. All MSMs were built using a lag time of 30 steps.

Next, we reanalyzed two ultra-long 100 μs molecular dynamics trajectories of the Fip35 WW domain,[47] provided courtesy of D.E. Shaw Research[48] (Amber ff99SB-ILDN force field,[49] TIP3P water model[50]). We projected the trajectories into a four dimensional space using time-structure based independent components analysis (tICA)[51,52] on the first trajectory. To calculate the tICs, each conformation was represented as a vector of distances between all pairs of residues separated by at least three amino acids. The distance $d(A,B)$ between residues A and B was determined as follows. Let $\{a_i\}_{i=1}^{n_a}$ and $\{b_i\}_{i=1}^{n_b}$ be the Cartesian coordinates of the heavy (non-hydrogen) atoms in A and B. Then we define

$$d(A,B) \equiv \min_{i,j} ||a_i - b_j||_2,$$

where $||\cdot||_2$ denotes the ℓ_2 norm. The tICs were computed using a correlation lag time of 200 ns. We also performed Principal Components Analysis (PCA) on the same residue-residue distance representation and built MSMs using the top three PCs. In each projection, k-centers was used with the Euclidean distance metric.

For the tICA MSMs, the cluster centers from the 100 state model were used as the test points Y to define A with a cutoff of $R = 0.81$. In the PCA MSMs, the cluster centers from the 500 state model were used as the test points with a cutoff of $R = 2.0$. In each case, $M = 10^6$ rounds of Monte Carlo integration were performed to compute the state volumes. All WW MSMs were built using a lag time of 50 ns, which is the same lag time used in previous analyses.[21]

For both the Müller potential and WW models, a maximum likelihood estimator for reversible transition matrices, with a pseudocount of $1/k$ was used to compute $\mathbf{T}$. The pseudocount ensures that all transitions are assigned nonzero probability, which is especially important for evaluating data that was not used to train the model, as often new transitions will be observed. All analysis and model construction was performed using MSMBuilder.[27]

The BIC and AIC were calculated according to Eq. (12) and Eq. (13), with N equal to the total number of transitions observed in the data. However, as transitions were counted using the “sliding window" approach, these transitions are not statistically independent.[20] Therefore, the BIC may be over-penalizing the number of parameters due to this over-estimate of N.

Results and Discussion

Müller Potential

How well can likelihood-based methods select parameters for an MSM in the data-rich regime? To answer this question, we simulated a simple two-dimensional potential and built a series of MSMs (see *Methods*). We compared four different model selection criteria: (1) test log-likelihood (2) BIC (3) AIC and (4) implied timescales convergence. The analysis (Fig. 2) indicates that the log-likelihood on the training data set continually increases with the addition of further states. The increase is most dramatic at low k, and levels off at high k, indicating that once k is large enough, the marginal increase in the likelihood with respect to further states is small. For this data set, the log-likelihoods computed on a separate test set exhibit a notably similar trend. Above $k = 600$, the test log-likelihood stops increasing. These two trends are to be expected: the additional parameters in the model at higher k never decreases the model's ability to fit the training data. On the other hand, this can lead to the model fitting the statistical noise in the training data set, rather than the system's true dynamics. As result, these "over fit" models at high k are less predictive, as measured by the log-likelihood they assign to new test data.

Leaving out a portion of a molecular dynamics data set from the training to serve as a test set is costly proposition for the applications of MSMs to the study of complex system. For large biomolecular simulations characterized by long intrinsic timescales, sampling the potential exhaustively is a significant challenge,[17] making it difficult to afford discarding half of the data set during the fitting of an MSM. For this reason, we also computed two penalized likelihood model selection criteria, the AIC and BIC, which augment the training set log-likelihood with an explicit penalty on model complexity to avoid over fitting. Applied to our simulations of the Müller potential, both the AIC and BIC penalize model complexity more strongly than the test set log-likelihood.

The likelihood-based methods are consistent with model selection based on maximization of the implied timescales. For this simple data set, the timescales are maximized by models with 200 states, which agrees with the results obtained with BIC. Model selection based on direct maximization of the implied timescales considers only the systematic error in the MSM, and neglects the statistical error. The consistency between the two approaches on this data set is an indication that the statistical uncertainty is low for these models, which is to be expected given the ease of sampling this two dimensional toy potential.

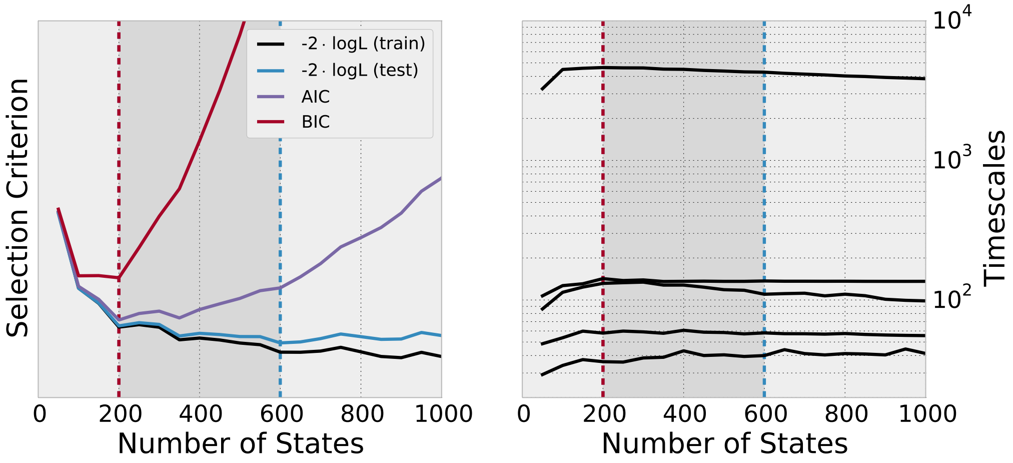

Figure 2: Four model selection criteria were used (left panel) to select the state discretization for MSMs built on simulations of the Müller potential. The log-likelihood was evaluated on the training set (black) and a separate test set (blue). These values are plotted as $-2 \cdot \log L$ for ease of comparison with the BIC (red) and AIC (purple). Lower scores indicate better models. Dashed lines are used to indicate criterion-minimizing models. These criteria are consistent with the convergence of the implied timescales (right panel).

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. (12) and Eq. (13) begins to dominate. The optimal models, according to the BIC, AIC, and test log-likelihood 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. It's not clear why the test log-likelihood is more lenient, however, it suggests that when possible, one should use a test log-likelihood approach as opposed to an approximate method like the AIC or BIC. This cross-validation requires fitting multiple models on subsets of the data, and so is not feasible for larger systems in the data-poor regime.

Fip35 WW Domain

How appropriate are likelihood-based model selection criteria for more realistic MD data sets? To answer this questions, we constructed Markov models based on simulations of the Fip35 WW domain[47] performed on the Anton supercomputer.[48] As a preprocessing step, we used time-structure based independent components analysis (tICA) to project the data into a four dimensional space. This procedure was necessary in order to compute the volumes of the MSM states, as described in Algorithm 1. MSMs were constructed in this lower dimensional space with differing number of states using the k-centers clustering algorithm (see *Methods*).

As was the case with the Müller MSMs, the log-likelihood of the training data set increases with the addition of more states, while the test log-likelihood, AIC and BIC lack a monotonic trend (**??**). The test log likelihood has a maximum at 650 states, and decreases with the addition of further states. This trend is matched by the AIC, which is computed only on the training data, while the BIC on the other hand seems to over-penalize model complexity. The implied timescales of these MSMs converge rapidly, and selection based on their convergence is consistent with the statistical methods.

We also built a series of models on the same data set by replacing the tICA preprocessing with Principal Components Analysis (PCA). We projected the data set onto the top three principal components, which are the highest-variance uncorrelated linear combinations of the input degrees of freedom (see *Methods*). Because PCA does not explicitly take into account the temporal structure of the data set, we expected that MSMs based on PCA would be less predictive than those based on tICA. Our results, shown in Fig. 4, confirm this hypothesis. We built MSMs with as many as 25,000 states, and saw a linear trend in the longest relaxation timescale with respect to the number of states.

The statistical model selection methods indicate that we are only justified in using models with

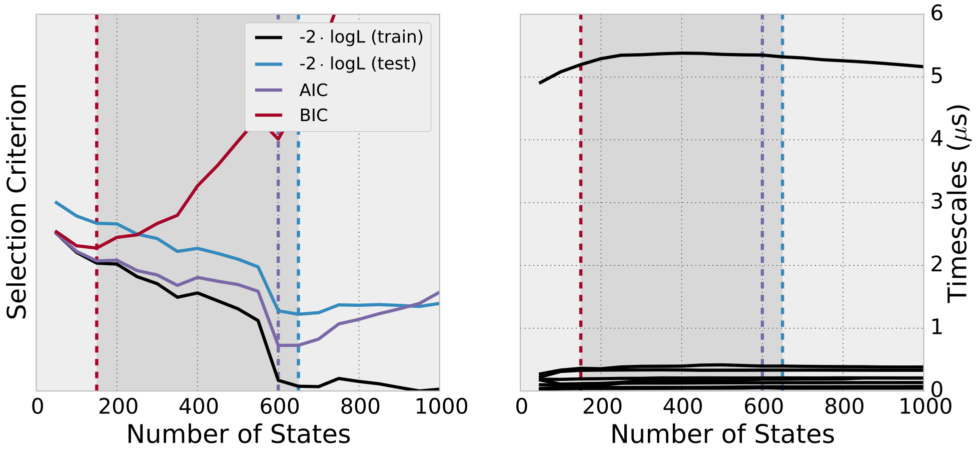

Figure 3: MSMs for the Fip35WW domain built with tICA were evaluated based on four model selection criteria. The log-likelihood was evaluated on the training set (black) and a separate test set (blue), and compared with the AIC (purple) and BIC (red) penalized log-likelihoods, computed on the training set. Dashed lines are used to indicate criterion-minimizing models. The models' ten slowest relaxation timescales (shown in the right panel) converge rapidly.

k on the order of 100. The AIC, BIC and test set log-likelihood criteria each recommend a model built with 200 states, which is also the point just after the most dramatic increase in the longest relaxation timescale occurs with respect to k. On the other hand, the relaxation timescales continue to increase further, even as the test set log likelihood begins to decrease after 1000 states.

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 PCA is not justified.

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.[27,53] 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 abil-

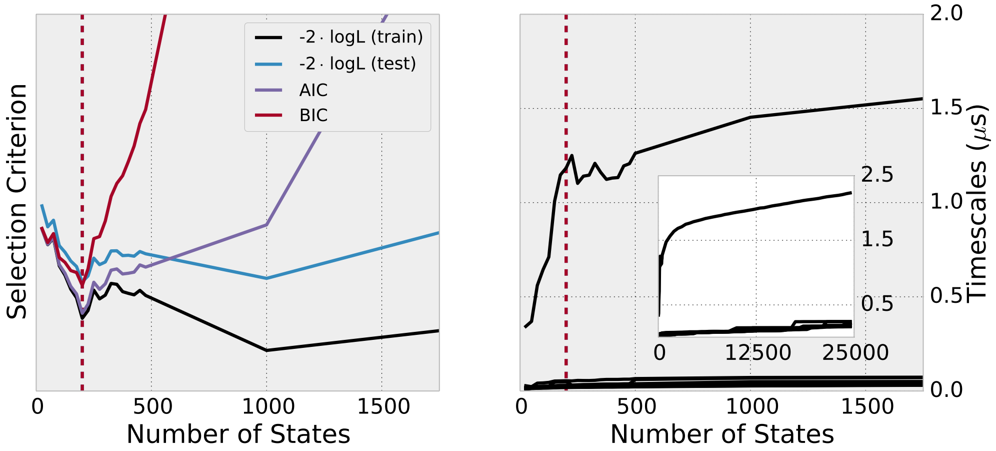

Figure 4: The four selection criteria were used to compare MSMs for the Fip35WW domain built with PCA. The longest relaxation timescale (right panel) exhibits a linear increase up to 25,000 states. The AIC (blue), BIC (red) and test set log-likelihood each recommend a model with 200 states.

ity to automatically construct MSMs on the fly while simulations are in progress, an important point for so-called adaptive sampling procedures,[54] is hampered when manual model selection is required.

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. (8). This strategy would complement approaches that generalize MSMs beyond discrete states.[28,55]

Acknowledgement

It is an honor and a pleasure to contribute this paper to the Swope Festschrift. I (V.S.P.) have been working with Bill since my early days at Stanford and have benefitted greatly from our numerous interactions with him. Bill's insight, clarity of thought, and careful approach has influenced my lab's work very broadly. I am very happy that our collaborations continue through today and look forward to many more in the future.

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. Direct Visualization Reveals Dynamics of a Transient Intermediate During Protein Assembly. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 6450–6455.

(7) Lipman, E. A.; Schuler, B.; Bakajin, O.; Eaton, W. A. Single-Molecule Measurement of Protein Folding Kinetics. *Science* **2003**, *301*, 1233–1235.

(8) 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.

(9) Tzeng, S.-R.; Kalodimos, C. G. Protein Dynamics and Allostery: an NMR View. *Curr. Opin. Struct. Biol.* **2011**, *21*, 62 – 67.

(10) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542–1555.

(11) 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.

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

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

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

(15) 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.

(16) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in Protein Folding Simulations: Timescale, Representation, and Analysis. *Nat. Phys.* **2010**, *6*.

(17) 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.

(18) Best, R. B.; Hummer, G. Coordinate-Dependent Diffusion in Protein Folding. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 1088–1093.

(19) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic Discovery of Metastable States for the Construction of Markov Models of Macromolecular Conformational Dynamics. *J. Chem. Phys.* **2007**, *126*, 155101–155101.

(20) 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.

(21) Beauchamp, K. A.; McGibbon, R. T.; 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.

(22) Bowman, G. R.; Meng, L.; Huang, X. Quantitative Comparison of Alternative Methods for Coarse-Graining Biological Networks. *J. Chem. Phys.* **2013**, *139*, 121905.

(23) Weinan, E.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123*, 503–523.

(24) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Identification of Almost Invariant Aggregates in Reversible Nearly Uncoupled Markov Chains. *Linear Algebra Appl.* **2000**, *315*, 39–59.

(25) Sammut, C.; Webb, G. I. *Encyclopedia of Machine Learning*; Springer, 2010.

(26) Schütte, C.; Huisinga, W.; Deuflhard, P. *Transfer Operator Approach to Conformational Dynamics in Biomolecular Systems*; Springer, 2001.

(27) 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. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419.

(28) Noé, F.; Nüske, F. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Model Sim.* **2013**, *11*, 635–655.

(29) 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.

(30) Bowman, G. R.; Geissler, P. L. Equilibrium Fluctuations of a Single Folded Protein Reveal a Multitude of Potential Cryptic Allosteric Sites. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 11681–11686.

(31) 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. *J. Phys. Chem. B* **2004**, *108*, 6582–6594.

(32) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131*, 124101.

(33) Gordon, H. L.; Somorjai, R. L. Fuzzy Cluster Analysis of Molecular Dynamics Trajectories. *Proteins Struct. Funct. Bioinf.* **1992**, *14*, 249–264.

(34) 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.

(35) Liddle, A. R. Information Criteria for Astrophysical Model Selection. *Mon. Not. R. Astron. Soc. Lett.* **2007**, *377*, L74–L78.

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

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

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

(39) Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723.

(40) Kass, R. E.; Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795.

(41) Liddle, A. R. How Many Cosmological Parameters? *Mon. Not. R. Astron. Soc. Lett.* **2004**, *351*, L49–L53.

(42) Kuha, J. AIC and BIC: Comparisons of Assumptions and Performance. *Sociol. Method Res.* **2004**, *33*, 188–229.

(43) Kannan, R.; Lovász, L.; Simonovits, M. Random Walks and an $O^*(n^5)$ Volume Algorithm for Convex Bodies. *Random Structures Algorithms* **1997**, *11*, 1–50.

(44) Simonovits, M. How to Compute the Volume in High Dimension? *Math. Program.* **2003**, *97*, 337–374.

(45) Lovász, L.; Vempala, S. Simulated Annealing in Convex Bodies and an $O^*(n^4)$ Volume Algorithm. *J. Comput. System Sci.* **2006**, *72*, 392 – 417.

(46) Müller, K.; Brown, L. D. Location of Saddle Points and Minimum Energy Paths by a Constrained Simplex Optimization Procedure. *Theor. Chim. Acta* **1979**, *53*, 75–93.

(47) 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.

(48) 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.

(49) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins Struct. Funct. Bioinf.* **2010**, *78*, 1950–1958.

(50) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(51) 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.

(52) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, –.

(53) Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S.; Schütte, C.; Noé, F. EMMA: A Software Package for Markov Model Building and Analysis. *J. Chem. Theory Comput.* **2012**, *8*, 2223–2238.

(54) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory Comput.* **2010**, *6*, 787–794.

(55) Chiang, T.-H.; Hsu, D.; Latombe, J.-C. Markov Dynamic Models for Long-Timescale Protein Motion. *Bioinformatics* **2010**, *26*, i269–i277.