Validation of Markov state models using Shannon's entropy

Sanghyun Park and Vijay S. Pande[a)]
Department of Chemistry, Stanford University, Stanford, California 94305
and Department of Structural Biology, Stanford University, Stanford, California 94305

(Received 1 August 2005; accepted 19 December 2005; published online 7 February 2006)

Markov state models are kinetic models built from the dynamics of molecular simulation trajectories by grouping similar configurations into states and examining the transition probabilities between states. Here we present a procedure for validating the underlying Markov assumption in Markov state models based on information theory using Shannon's entropy. This entropy method is applied to a simple system and is compared with the previous eigenvalue method. The entropy method also provides a way to identify states that are least Markovian, which can then be divided into finer states to improve the model. © *2006 American Institute of Physics*. [DOI: 10.1063/1.2166393]

I. INTRODUCTION

With the advance of computational resources, especially distributed computing,[1] massive parallel simulations have become possible. One of the challenges in parallel simulations of dynamical systems is the ability to reach long time scales through a multitude of independent short trajectories. This challenge is most evident in simulations of biomolecules where there is still a huge gap between the time scale of interest (milliseconds to seconds) and the time scale computationally accessible (nanoseconds to microseconds).

To address this challenge, a theoretical construct known as Markov state model (MSM) has been developed.[2–6] A MSM describes the kinetics of a system as Markovian transitions between a set of states. Transition probabilities, which are the building blocks of MSMs, are determined from simulation data collected at a certain time interval (referred to as *lag time* hereafter). Transition probabilities between each pair of states can be obtained from short trajectories, but the resulting MSM on the entire state space can describe long time scales. A crucial, yet often neglected, step in the use of MSMs is to verify that the transitions are indeed Markovian (i.e., history independent) at the given lag time. Swope *et al.*[2] suggested a criterion based on the eigenvalues of the transition matrix. In this paper we present a validation procedure based on Shannon's entropy,[7] and compare it with the eigenvalue method via a simple application.

II. A MEASURE OF MARKOVITY BASED ON SHANNON'S ENTROPY

Suppose that we build a MSM with M states, labeled $1,2,\ldots,M$, for a certain lag time τ. The states are typically defined by grouping similar molecular configurations according to structural properties such as the root-mean-square deviation, the number of native contacts, and so on.[3] We are interested in the stationary ensemble of trajectories and therefore assume that simulation data are collected after an equilibration period. In other words, we assume that initial configurations are drawn from an equilibrium distribution. This assumption, however, can be weakened as we will discuss later.

Markovity is defined as the independence of the transition probabilities on the previous history of visited states:

$$\Pr(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \Pr(X_n = x_n | X_{n-1} = x_{n-1}) \quad \text{for any } x_n, \ldots, x_0, \tag{1}$$

where X_n is the variable representing the state at nth time step, consecutive time steps being separated by τ. We follow the convention of denoting a variable itself by an uppercase letter and a possible value by a lowercase letter. Taking advantage of stationarity, we introduce the following shorthand notations:

$$\begin{aligned} &p(x) \equiv \Pr(X_n = x), \\ &p_\tau(x,y) \equiv \Pr(X_n = x, X_{n-1} = y), \\ &p_\tau(x|y) \equiv \Pr(X_n = x | X_{n-1} = y), \\ &p_\tau(x|y,z) \equiv \Pr(X_n = x | X_{n-1} = y, X_{n-2} = z), \end{aligned} \tag{2}$$

and so on.

Here we assume that the memory effect decreases monotonically in time so that it is sufficient to verify the independence on the immediate past only. This assumption is reasonable for most physical systems although one can construct exceptions.[8] Under this assumption, the kinetics on the state space is considered Markovian at the lag time τ if and only if

$$p_\tau(x|y,z) = p_\tau(x|y) \quad \text{for any } x,y,z. \tag{3}$$

In principle, one could use Eq. (3) to verify Markovity. This would be, however, rather inefficient since it involves as many as M^3 equalities to verify.

A necessary and sufficient condition for Eq. (3) can be written as a single equality in terms of Shannon's entropy:

$$H_\tau(X_n | X_{n-1}, X_{n-2}) = H_\tau(X_n | X_{n-1}), \tag{4}$$

where the first-order conditional entropy

a)Author to whom correspondence should be addressed. Electronic mail: pande@stanford.edu

$$H_\tau(X_n|X_{n-1}) = -\sum_{x,y} p_\tau(x,y)\ln p_\tau(x|y) \tag{5}$$

means the remaining information in X_n when X_{n-1} is known, and the second-order conditional entropy

$$H_\tau(X_n|X_{n-1},X_{n-2}) = -\sum_{x,y,z} p_\tau(x,y,z)\ln p_\tau(x|y,z) \tag{6}$$

means the remaining information in X_n when both X_{n-1} and X_{n-2} are known. For later uses, we also define the zeroth entropy:

$$H(X_n) = -\sum_{x} p(x)\ln p(x), \tag{7}$$

which is the information in X_n according to the stationary probability $p(x)$. We note again that, due to stationarity, references to absolute times are irrelevant, for instance, $H_\tau(X_n|X_{n-1}) = H_\tau(X_1|X_0)$.

The inequality $H_\tau(X_n|X_{n-1},X_{n-2}) \leqslant H_\tau(X_n|X_{n-1})$ holds between the two entropies because conditioning reduces information, where the equality holds only when X_n and X_{n-2} are conditionally independent given X_{n-1}.[9] Therefore, Eq. (4) is sufficient for the conditional independence expressed in Eq. (3). By substituting Eq. (3) into Eq. (6), it is straightforward to show that Eq. (4) is also necessary for Eq. (3). To recap, Markovity is equivalent to the property that when X_{n-1} is known the additional knowledge of X_{n-2} does not change the remaining information in X_n. In other words, Markovity is equivalent to the vanishing of the conditional mutual information[9]

$$I_\tau(X_n;X_{n-2}|X_{n-1}) = H_\tau(X_n|X_{n-1}) - H_\tau(X_n|X_{n-1},X_{n-2}) \tag{8}$$

between X_n and X_{n-2} given X_{n-1}.

When $I_\tau(X_n;X_{n-2}|X_{n-1})$ is zero, the Markovity of the given system is established. However, how do we interpret nonzero values of the mutual information? If $I_\tau(X_n;X_{n-2}|X_{n-1}) = 0.1$, for instance, what would be the conclusion about the Markovity of the system? For a quantitative measure of Markovity, we propose

$$R_\tau \equiv \frac{I_\tau(X_n;X_{n-2}|X_{n-1})}{H_\tau(X_n|X_{n-1})}, \tag{9}$$

which quantifies, given X_{n-1}, what fraction of the information in X_n is the mutual information between X_n and X_{n-2}. The range for the possible values of R_τ is between zero and one; $R_\tau = 0$ represents perfect Markovity and $R_\tau = 1$, which happens if X_n is determined with certainty by specifying X_{n-2} in addition to X_{n-1}, represents the least Markovian case possible. If $R_\tau = 0.1$, for instance, 10% of the information in X_n given X_{n-1} is the mutual information with X_{n-2}, which would indicate a 10% memory effect, i.e., non-Markovity.

The mutual information of Eq. (8) can be decomposed for each state:

$$\begin{aligned} I_\tau(X_n;X_{n-2}|X_{n-1}) &= \sum_y p(y) I_\tau(X_n;X_{n-2}|X_{n-1}=y), \\ I_\tau(X_n;X_{n-2}|X_{n-1}=y) &= H_\tau(X_n|X_{n-1}=y) \\ &\quad - H_\tau(X_n|X_{n-1}=y,X_{n-2}) \\ &= -\sum_x \frac{p_\tau(x,y)}{p(y)} \ln p_\tau(x|y) \\ &\quad + \sum_{x,z} \frac{p_\tau(x,y,z)}{p(y)} \ln p_\tau(x|y,z). \end{aligned} \tag{10}$$

Here we present two applications of this statewise decomposition. First, by inspecting

$$r_\tau(y) \equiv \frac{I_\tau(X_n;X_{n-2}|X_{n-1}=y)}{H_\tau(X_n|X_{n-1}=y)} \tag{11}$$

for each state y, one can identify states that are least Markovian, i.e., states within which equilibration is slowest. By dividing such states into finer states, a MSM can be improved in the sense that it will be Markovian at shorter lag times. Second, the statewise decomposition provides a more secure way to verify Markovity. It may happen that a certain state is significantly non-Markovian, but R_τ does not pick out its non-Markovity because it is a rare state $[p(y) \ll 1]$.[8] This will be particularly devastating if the rare state plays an important role in the kinetics of interest. However, this issue can be resolved by making sure that each and every $r_\tau(y)$ approaches zero.

III. BAYESIAN INFERENCE FROM TRANSITION DATA

In order to calculate the first- and second-order conditional entropies, we need four quantities: $p_\tau(x,y)$, $p_\tau(x,y,z)$, $p_\tau(x|y)$, and $p_\tau(x|y,z)$. Although all of these can be directly estimated from simulation data, the joint probabilities, $p_\tau(x,y)$ and $p_\tau(x,y,z)$, must be estimated from trajectories equilibrated *across* all the states (interstate equilibration), while the estimation of conditional probabilities requires only equilibration *within* each state (intrastate equilibration). Let us elaborate on this point. To estimate the first-order conditional probabilities, we can separate the first-order transition data into M disjoint sets in such a way that the first set contains $1 \to x$ transitions, the second set $2 \to x$ transitions, and so on, where x is any of the M states. Now the conditional probabilities $p_\tau(x|n)$ are estimated entirely from the nth set, without looking at the other sets. Therefore, equilibration is required only within each set; namely, initial configurations need to be equilibrated only within each state. The equilibration requirement is even weaker for the second-order conditional probabilities. The conditional probabilities $p_\tau(x|n,m)$ are estimated entirely from the set of $m \to n \to x$ transitions (x is any of the M states), which require equilibration only within the subspace of state m that contains the configurations (microstates) that can make a transition to state n after one step.

Therefore, we choose to estimate the second-order conditional probabilities from data and derive the other quantities from it, thereby weakening the equilibration require-

ment. The joint probability $p_\tau(x,y)$ is determined as the stationary distribution of the two-step propagator

$$A_{(x,w),(y,z)} \equiv \Pr(X_{n+1}=x, X_n=w | X_n=y, X_{n-1}=z) = \delta_{wy} p_\tau(x|y,z), \quad (12)$$

i.e., the eigenvector of the $M^2 \times M^2$ matrix A corresponding to the eigenvalue of 1. At this point, if desired, the condition of detailed balance, $p_\tau(x,y)=p_\tau(y,x)$, can be imposed by symmetrizing the $M \times M$ matrix $p_\tau(x,y)$.[5] The remaining quantities are determined by $p(x)=\Sigma_y p_\tau(x,y)$, $p_\tau(x,y,z)=p_\tau(x|y,z)p_\tau(y,z)$, and $p_\tau(x|y)=p_\tau(x,y)/p(y)$.

For the estimation of the second-order conditional probabilities from transition data, we use Bayesian inference.[10] In Bayesian theory, a probability means a state of knowledge. Bayesian inference calculates a posterior probability for parameters from prior probability and data. In our case, the second-order conditional probabilities are the parameters to estimate, and thus we are dealing with probabilities of probabilities. As explained above, the estimation of the second-order conditional probabilities, $p_\tau(x|y,z)$, is done separately for each (y,z) pair. For a given (y,z) pair, let us define $\theta_x \equiv p_\tau(x|y,z)$ and let n_x denote the observed number of the $z \to y \to x$ transition. Then, $\theta \equiv (\theta_1, \ldots, \theta_M)$ is the parameter to estimate and $n \equiv (n_1, \ldots, n_M)$ represents the data. θ_x can take values between zero and one, with the constraint $\Sigma_x \theta_x = 1$. As a prior probability that represents the state of knowledge on θ before seeing any data, we choose the uniform distribution $\Pr(\theta)=(\mathrm{M}-1)!$ which is properly normalized on the θ space. The application of Bayes theorem along with the sampling distribution

$$\Pr(n|\theta) = \frac{\left(\sum_x n_x\right)!}{n_1! \cdots n_M!} \theta_1^{n_1} \cdots \theta_M^{n_M} \quad (13)$$

leads to the posterior distribution

$$\Pr(\theta|n) = \frac{\Pr(n|\theta)\Pr(\theta)}{\int d\theta \Pr(n|\theta)\Pr(\theta)} = \frac{\left(\sum_x n_x + M - 1\right)!}{n_1! \cdots n_M!} \theta_1^{n_1} \cdots \theta_M^{n_M}, \quad (14)$$

which is a Dirichlet distribution.[6] This posterior distribution represents the state of knowledge on θ after seeing the transition data. Carrying out this procedure for each (y,z) pair, posterior distributions can be constructed for all the second-order transition probabilities, which we collectively denote by Θ. The posterior distribution for any quantity $f(\Theta)$ in question, such as the mutual information [Eq. (8)] and R_τ [Eq. (9)], can be obtained by sampling a number of values from the Θ posterior and calculating $f(\Theta)$ for each. Bayesian inference provides a distribution of parameters, not just a single estimate, from which the uncertainty of the estimation can be extracted, e.g., by taking 95% interval around the median as we do in the example below.

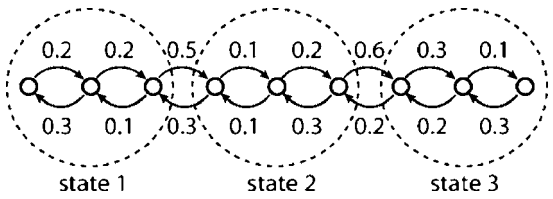

FIG. 1. A MSM of three states (macrostates) built from a stochastic model of nine configurations (microstates). The numbers next to arrows denote transition probabilities for unit time.

IV. A SIMPLE APPLICATION

For a demonstration, we constructed a MSM from a simple stochastic system consisting of nine configurations with transition probabilities as specified in Fig. 1, similar to the model studied by Swope *et al.*[2] We generated 100 000 trajectories, each 200 time steps in length, with initial configurations drawn from the equilibrium distribution among the nine configurations, although intrastate equilibration would be sufficient as explained above. The nine configurations (microstates) were lumped together into three states (macrostates) as shown in Fig. 1. From the trajectories generated, we computed the posterior distribution for entropies and other quantities necessary for the validation of Markovity following the Bayesian procedure outlined above, and obtained medians and 95% intervals from 1000 values sampled from the posterior distribution. Detailed balance was not imposed throughout this example.

Three entropies are plotted in Fig. 2. The zeroth-order entropy $H(X_n)$, estimated at each lag time τ, is independent of τ as it should be. As τ increases, the second-order conditional entropy $H_\tau(X_n|X_{n-1},X_{n-2})$ approaches the first-order conditional entropy $H_\tau(X_n|X_{n-1})$, and later both the first- and second-order conditional entropies approach $H(X_n)$. Thus, two distinct time scales exist: the intrastate equilibration time corresponding to the decay of $H_\tau(X_n|X_{n-1}) - H_\tau(X_n|X_{n-1},X_{n-2})$ and the interstate equilibration time corresponding to the decay of $H(X_n) - H_\tau(X_n|X_{n-1})$. Intrastate equilibration, which is relevant for Markovity, is further analyzed below.

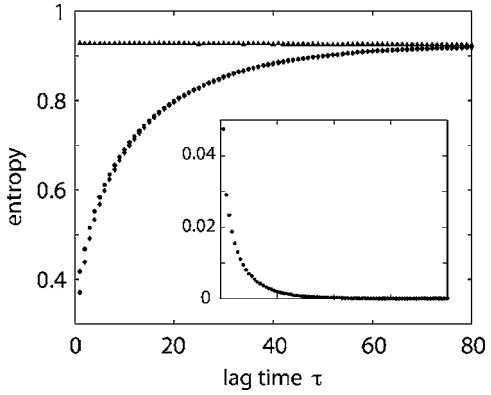

FIG. 2. Entropies calculated from 100 000 trajectories: Triangle: the zeroth-order entropy $H(X_n)$; circle: the first-order conditional entropy $H_\tau(X_n|X_{n-1})$; diamond: the second-order conditional entropy $H_\tau(X_n|X_{n-1},X_{n-2})$. The error bars calculated at 95% Bayesian intervals are plotted, but are indistinguishable from the symbols that denote the medians. The difference $H_\tau(X_n|X_{n-1}) - H_\tau(X_n|X_{n-1},X_{n-2})$ is shown in the inset.

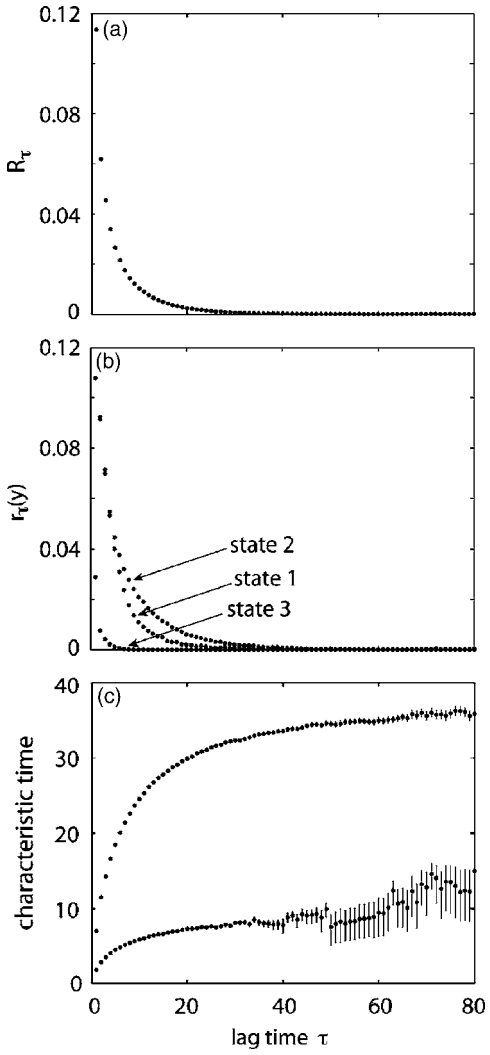

FIG. 3. Validation of Markovity using 100 000 trajectories: (a) The overall measure of Markovity, R_τ; (b) the state-by-state measure of Markovity, $r_\tau(y)$; and (c) characteristic times, $\chi_\tau^{(k)}$. In all three panels, error bars calculated at 95% Bayesian intervals are plotted, but some of them are indistinguishable from the symbols that denote the medians.

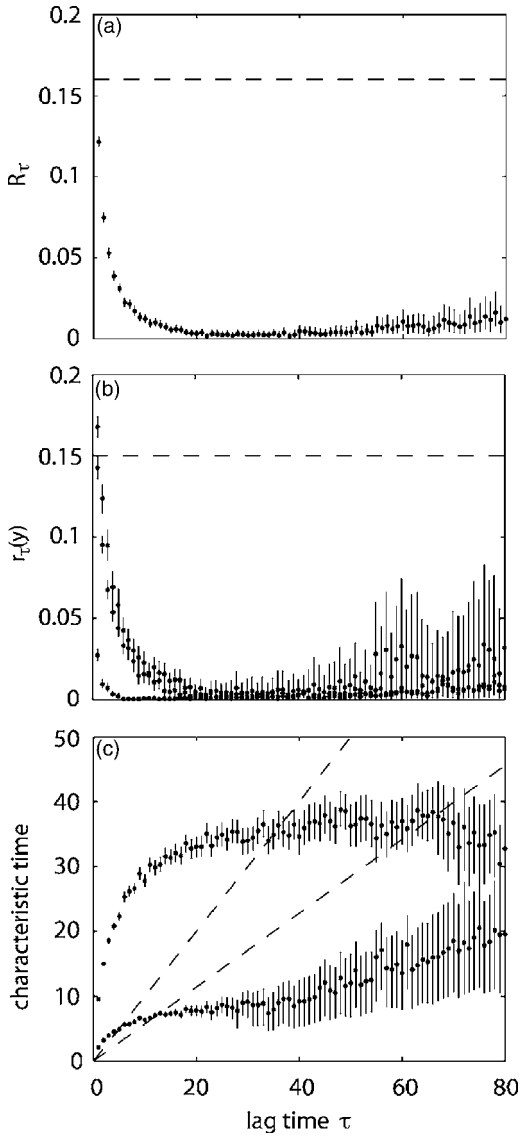

FIG. 4. Validation of Markovity using 1000 trajectories: (a) The overall measure of Markovity, R_τ; (b) the state-by-state measure of Markovity, $r_\tau(y)$; and (c) characteristic times, $\chi_\tau^{(k)}$. In all three panels, error bars calculated at 95% Bayesian intervals are plotted. The dashed lines denote the medians calculated purely from prior distribution.

The overall measure of Markovity, R_τ, is shown in Fig. 3(a). R_τ drops from 0.11 (non-Markovity of 11%) at $\tau=1$ to near zero after $\tau\approx 40$. The state-by-state measure r_τ shown in Fig. 3(b) indicates the Markovity of each state. State 3 is the most Markovian, and state 2 is the least Markovian and therefore should be the target for a refinement. This result makes intuitive sense if one takes a close look at the transition probabilities in Fig. 1; transitions are fastest in state 3, and slowest in state 2.

Figure 3(c) shows characteristic times, $\chi_\tau^{(k)} \equiv -\tau/\ln|\lambda_\tau^{(k)}|$, where $\lambda_\tau^{(k)}$ is the kth eigenvalue of the transition matrix $p_\tau(x|y)$, excluding the eigenvalue of 1. For Markovian systems, characteristic times should not depend on the lag time τ, which is the criterion suggested by Swope *et al.*[2] for the validation of MSMs. This criterion is a necessary condition for Markovity, but its sufficiency has not been proven. As can be seen in Fig. 3(c), the estimates of $\chi_\tau^{(k)}$ (especially the smaller one) seem noisier than those of R_τ and r_τ. This is due to the numerical instability of taking the logarithm of exponentially decaying functions. It is, therefore, both a statistical and a numerical issue; the noise cannot be reduced unless we improve the precision of the estimation of the transition matrix and the accuracy of the evaluation of the eigenvalues.

While distributed computing can naturally lead to many trajectories, the availability of 100 000 trajectories represents an uncommon case for typical computational capabilities today, particularly in simulations of macromolecules such as proteins. To assess the robustness of the entropy and eigenvalue methods with respect to the sampling size, we randomly selected 1000 trajectories from the total 100 000 and repeated the analysis with the reduced data set. The results are shown in Fig. 4. The same qualitative behavior is observed for the reduced data set, except that error bars are, of course, much bigger now. The posterior distribution determined by Bayesian inference depends on the data and the prior distribution, and as the amount of data is reduced, the influence of the prior distribution starts to dominate. In Fig.

4, the medians calculated purely from the prior distribution are plotted as dashed lines. Although the results from 1000 trajectories seem distant from the pure prior distribution effect, the effect of the prior distribution is noticeable in the rise of R_τ and r_τ at long lag times.

V. CONCLUDING REMARKS

Using Shannon's entropy, we have derived a method for validating the Markov assumption in MSMs. Our method provides an overall measure of Markovity and also a state-by-state measure which can be used to identify least Markovian states as targets for refinement. For the MSM of three states studied here, the entropy method seems less susceptible to statistical and numerical noise than the previous eigenvalue method. For a MSM of a large number of states, however, the entropy method might be less practical as it requires the estimation of second-order transition probabilities. Nevertheless, having known bounds (zero and one) for the entropy-based measures, R_τ and r_τ, is an advantage.

One might ask what value of R_τ is sufficient for verifying Markovity. R_τ can be used to compare different systems, or different models, but there cannot be any absolute verdict for a single R_τ value. Namely, we can say $R_\tau=0.1$ is better (more Markovian) than $R_\tau=0.2$, but "is $R_\tau=0.1$ good enough?" is not a well-posed question by itself. A well-posed question would be, for instance, "what value of R_τ is sufficient for accurately estimating a certain observable with a certain precision?" This is, however, a complicated problem which needs further investigation. Another way of interpreting R_τ is to extract the time scale for intrastate equilibration from the graph of R_τ as a function of τ, e.g., by fitting the graph with an exponential function. For lag times longer than the intrastate equilibration time, memory effects are likely to be negligible.

We note that Eq. (4) is not the only way to express Markovity in a single equality. For instance,

$$\sum_{x,y,z} [p_\tau(x|y,z) - p_\tau(x|y)]^2 = 0 \tag{15}$$

is also a necessary and sufficient condition of Eq. (3). Although we do not exclude the possibility that alternative *ad hoc* equalities such as Eq. (15) might be more suitable for certain systems, the advantage of the entropy-based criterion is that the quantities involved have clear interpretations in terms of information.

The development of MSM-based analysis methods for molecular simulations is still in its early stage, and much remain to be done. As for the validation of Markovity, it will be interesting to apply the present method to more realistic data such as molecular-dynamics trajectories of a protein. Such a study will give insight both on the practicality of the present method and on the time scale required for MSM-based analyses of realistic systems.

ACKNOWLEDGMENTS

We thank Nina Singhal, Sidney P. Elmer, Peter Kasson, and William C. Swope (IBM) for valuable suggestions and discussions, and acknowledge support from the Dreyfus Foundation and a grant from NIH (R01GM62868).

[1] D. Clery and D. Voss, Science **308**, 809 (2005).
[2] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[3] N. Singhal, C. D. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415 (2004).
[4] S. Sriraman, I. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).
[5] S. P. Elmer, S. Park, and V. S. Pande, J. Chem. Phys. **123**, 114902 (2005).
[6] N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[7] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[8] N. Singhal (private communication).
[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[10] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).