Coarse Master Equation from Bayesian Analysis of Replica Molecular Dynamics Simulations†

Saravanapriyan Sriraman,‡ Ioannis G. Kevrekidis,‡,§ and Gerhard Hummer*,#

Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08544, Program in Applied and Computational Mathematics and Department of Mathematics, Princeton University, Princeton, New Jersey 08544, and Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520

Received: August 6, 2004; In Final Form: November 23, 2004

We use Bayesian inference to derive the rate coefficients of a coarse master equation from molecular dynamics simulations. Results from multiple short simulation trajectories are used to estimate propagators. A likelihood function constructed as a product of the propagators provides a posterior distribution of the free coefficients in the rate matrix determining the Markovian master equation. Extensions to non-Markovian dynamics are discussed, using the trajectory "paths" as observations. The Markovian approach is illustrated for the filling and emptying transitions of short carbon nanotubes dissolved in water. We show that accurate thermodynamic and kinetic properties, such as free energy surfaces and kinetic rate coefficients, can be computed from coarse master equations obtained through Bayesian inference.

1. Introduction

Molecular dynamics (MD) simulations on classical or quantum mechanical energy surfaces can provide detailed insights into molecular processes. However, large system sizes, long-range interactions, and slow global dynamics combined with the necessity to integrate accurately even the fastest motions result in often severe time-scale limitations. As a consequence, rare transitions are often poorly sampled or remain inaccessible to conventional MD.[1] Formally, fast molecular motions can be integrated out using the projection operator formalism.[2] However, an analytic construction of generalized Langevin equations for specific dynamic variables in a complex molecular system remains a challenge. The coarse molecular dynamics (CMD) approach[3] provides a framework to extract the projected dynamics *on the fly* from unbiased simulations of multiple replicas. Running multiple MD simulations can result in a rapid exploration of configuration space[3−9] and is a key element of rate calculations using the reactive-flux method.[10−13] In the simplest CMD approach,[3] multiple unbiased and independent simulation trajectories are initialized with prescribed coarse variables (e.g., dihedral angles or the radius of gyration of a polymer) and then run for a time long enough for the distribution along the fast directions to saturate. From the observed collective behavior, i.e., the drift and spread along the coarse variables parametrizing the slow manifold, new initial conditions are created depending on the objective (e.g., finding free energy minima or saddle points). This approach immediately lends itself to a probabilistic interpretation in which ensembles of independent replica simulations sample local propagators.[3] With similar techniques, a master equation can be constructed when the simulator is not MD, but Monte Carlo.[14] Such "master equation" approaches can be used to estimate the long-time dynamics, and several formalisms related to the one developed here have been discussed.[15−20]

Here, we introduce a global Bayesian analysis of short simulation runs that will allow us to infer the slow dynamics in the projected coordinates. Likelihood-based and Bayesian approaches have also been used for the analysis of single-molecule experiments[21,22] and nonlinear dynamical systems.[23−25] We will construct a coarse master equation represented by a rate matrix, from which we estimate system properties and their statistical uncertainties. From the estimated dynamics projected onto coarse variables, we will obtain free energy surfaces and kinetic rate coefficients.[11] We will illustrate our general procedure by analyzing molecular dynamics simulations of the filling and emptying of a carbon nanotube in water.[26]

2. Theory

2.1. Master Equation. We start with Zwanzig's development of a generalized master equation for Newtonian dynamics in configuration space.[27] If the complete configuration space is divided into N nonoverlapping cells that together span the whole space, then the probability $p_i(t)$ of being in cell i at time t satisfies a generalized master equation,

$$\dot{p}_i(t) = -\int_0^t \mathrm{d}t' \sum_j K_{ij}(t-t')\, p_j(t') \qquad (1)$$

where $\dot{p}_i(t) \equiv \mathrm{d}p_i(t)/\mathrm{d}t$, and $K_{ij}(t)$ is the transition memory kernel given formally in terms of projections.[27] The absence of a "random force" (i.e., an inhomogeneous term) in eq 1 requires that nonequilibrium in the initial conditions ($t = 0$) is limited to the partitioning of replicas between different cells j. For replicas within a given cell j, the phase−space variables should accordingly be drawn from an equilibrium distribution.[27]

† Part of the special issue "David Chandler Festschrift".

* To whom correspondence should be addressed. E-mail: gerhard.hummer@nih.gov.

‡ Department of Chemical Engineering, Princeton University.

§ Program in Applied and Computational Mathematics and Department of Mathematics, Princeton University.

National Institutes of Health.

10.1021/jp046448u CCC: $30.25
Published on Web 01/25/2005

Zwanzig's generalized master equation, eq 1, forms the basis for a propagator based approach to configurational dynamics. Specifically, our goal is to estimate the elements K_{ij} of the transition memory kernel from multiple MD simulations. For simplicity, we will first assume that the memory is sufficiently short such that after a brief initial simulation period we can approximate eq 1 with a Markovian rate equation,

$$\dot{p}_i(t) = \sum_j R_{ij} p_j(t) \tag{2}$$

where $R_{ij} \geq 0$ are the constant elements of a rate matrix $\mathbf{R}$, $\sum_i R_{ij} \equiv 0$. Equation 2 can be solved formally in terms of a matrix exponential,

$$p_i(t) = \sum_j (\mathrm{e}^{t\mathbf{R}})_{ij} p_j(0) \tag{3}$$

From this relation we can identify the propagators,

$$p(i,t|j,0) = (\mathrm{e}^{t\mathbf{R}})_{ij} \tag{4}$$

where $p(i,t|j,0)$ is the conditional probability that a trajectory starting from cell j is in cell i at time t. The requirement of equilibrium initial conditions within cell j is satisfied if one uses all configurations in cell j during a long equilibrium run (or a random subset thereof). If instead initial conditions are prepared using short MD initialization runs,[3] the biasing potential driving the system into cell j should be flat within j, and the simulation should be long enough to obtain a representative sample. To correct for a nonuniform bias within cells, individual replica runs can be given a weight proportional to the inverse Boltzmann factor of the initial biasing potential.

2.2. Likelihood Function. In the following, we will analyze replica simulation runs using Bayesian inference to estimate the rate matrix $\mathbf{R}$. As will become clear, the same approach can be used to obtain also a maximum-likelihood estimate of the rate matrix.

According to the generalized master equation, eq 1, we set up initial conditions for independent replica runs within cells j. All other variables are drawn from equilibrium distributions. During the subsequent MD simulations, some of the replicas will cross into other cells i. Consider that we have made the following observations in replica runs $\alpha = 1, 2, \ldots$: At a time t_α after starting run α from cell j_α, the system is found in cell i_α. Under the assumption of the kinetic model eq 2, the likelihood of such an observation is $p(i_\alpha,t_\alpha|j_\alpha,0) = [\exp(t_\alpha\mathbf{R})]_{i_\alpha j_\alpha}$. Given the model eq 2, the likelihood of a series of such observations (assuming that they are independent!) is

$$L = \prod_\alpha p(i_\alpha,t_\alpha|j_\alpha,0) = \prod_\alpha (\mathrm{e}^{t_\alpha\mathbf{R}})_{i_\alpha j_\alpha} \tag{5}$$

In eq 5, we implicitly assume that the dynamics are Markovian on the time scale t_α of the observations. For non-Markovian dynamics, our observations are the actual paths, observed at discrete times along the simulation trajectories. The likelihood of observing a path[28] is given by a generalized path action functional. For the discrete time steps of a simulation, the probability to observe a particular path $i_0 \to i_1 \to \ldots \to i_n$ is given by a product

$$p(i_0 \to i_1 \to \ldots \to i_n) = p(i_0)\, p(i_1|i_0)\, p(i_2|i_0 \to i_1) \cdots p(i_n|i_0 \to i_1 \to \ldots \to i_{n-1}) \tag{6}$$

where $p(i_k|i_0 \to i_1 \to \ldots \to i_{k-1})$ is the conditional probability of reaching cell i_k for a given preceding path $i_0 \to i_1 \to \ldots \to i_{k-1}$. The likelihood of observing a series of paths in independent trajectories is then the product of the $p(i_0 \to i_1 \to \ldots \to i_n)$.

2.3. Bayesian Analysis of Replica Simulations. *Bayesian Inference*. Here, we restrict our analysis to the Markovian case. By monitoring the dynamics of the individual replicas *only after* a short relaxation time, we hope to avoid the initial non-Markovian dynamics. With this assumption, we use a Bayesian approach to obtain a posterior distribution of the model parameters (here: the coefficients of the rate matrix $\mathbf{R}$) from the trajectory data through[29]

$$p(\text{parameters}|\text{data}) \propto p(\text{data}|\text{parameters})\, p(\text{parameters}) \tag{7}$$

where $p(\text{data}|\text{parameters}) \equiv L$ is given by the likelihood function of eq 5. In the following, we assume a uniform prior distribution of the model parameters, $p(\text{parameters}) = 1$, such that

$$p(\text{parameters}|\text{data}) \propto L \tag{8}$$

Note that the assumption of a uniform prior reflects a particular choice of parameters, and correspondingly of an integration measure in parameter space.

Rate Matrix. Equation 8 forms the basis for inferring the unknown coefficients of the rate matrix $\mathbf{R}$ describing the coarse projected dynamics in the Markovian limit. As an important constraint, $\mathbf{R}$ must satisfy detailed balance, which reduces the number of free parameters. If we divide the configuration space into N cells i, the corresponding matrix $\mathbf{R}$ has dimensions $N \times N$. $\mathbf{R}$ can be expressed uniquely in terms of $N - 1$ equilibrium probabilities P_i (with $1 \geq P_i > 0$ and the Nth P_i given by normalization, $\sum_i P_i = 1$) and $N(N-1)/2$ rate coefficients:

$$R_{ij} = \begin{cases} R_{ij} > 0 & \text{if } i > j \\ -\sum_{l(\neq i)} R_{li} & \text{if } i = j \\ R_{ji} P_i / P_j & \text{if } i < j \end{cases} \tag{9}$$

The total number of free coefficients in a general $N \times N$ rate matrix $\mathbf{R}$ is thus $(N+2)(N-1)/2$. This number can be smaller if the structure of the rate matrix is known. For one-dimensional nonperiodic motion, for instance, $\mathbf{R}$ is tri-diagonal, resulting in only $2N - 2$ free coefficients.

Eigensystem Construction of Propagators. To obtain the propagators entering the likelihood function, we could numerically integrate the first-order ordinary differential equations of eq 2. Here, we instead use an eigenvalue technique. By constructing $\mathbf{R}$ from equilibrium probabilities P_i and the elements R_{ij} below the diagonal, we can use straightforward diagonalization of real symmetric matrixes to calculate the matrix exponential in eq 5. We first define a symmetric matrix $\tilde{R}_{ij} = P_i^{-1/2} R_{ij} P_j^{1/2}$. Its matrix $\mathbf{U}$ of real eigenvectors satisfies $\tilde{\mathbf{R}}\mathbf{U} = \mathbf{U}\Lambda$, where $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_N)$ is the diagonal matrix of real eigenvalues. With $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{1}$ (where T denotes the matrix transpose and $\mathbf{1}$ is the unity matrix), it follows that $\mathrm{e}^{t\tilde{\mathbf{R}}} = \mathbf{U}\mathrm{e}^{t\Lambda}\mathbf{U}^T$ with $\mathrm{e}^{t\Lambda} = \mathrm{diag}(\mathrm{e}^{t\lambda_1},\ldots,\mathrm{e}^{t\lambda_N})$. The matrix exponential describing the propagators then becomes $\mathrm{e}^{t\mathbf{R}} = \mathrm{diag}(P_1^{1/2},\ldots,P_N^{1/2})\mathrm{e}^{t\tilde{\mathbf{R}}}\,\mathrm{diag}(P_1^{-1/2},\ldots,P_N^{-1/2})$ and can be calculated by diagonalization of the real symmetric matrix $\tilde{\mathbf{R}}$.

Monte Carlo Sampling of Rate Coefficients. To infer the parameters of the coarse master equation, i.e., the rate coefficients R_{ij}, we will sample parameters according to eq 8 using the Metropolis Monte Carlo algorithm.[30] Random modifications of the parameters of $\mathbf{R}$ (i.e., R_{ij} for $i > j$ and P_i for $i > 1$) will be accepted and rejected according to the posterior distribution,

eq 8, given by the likelihood function, eq 5. This procedure allows us to sample parameter values consistent with the data. In this Monte Carlo approach, the log-likelihood, $-\ln L$, serves as an effective energy function in parameter space. From the Monte Carlo sampling in parameter space, we obtain distributions of the rate coefficients R_{ij}. Moreover, for the sampled parameters, we can calculate system properties determined by $\mathbf{R}$, such as the difference in free energy, $-\ln P_i/P_j$, between cells i and j, or the slowest rate of relaxation (i.e., the largest nonzero eigenvalue of $\mathbf{R}$). The distribution $p(\tau|\text{data})$ of a system property $\tau(\mathbf{R})$ can be calculated from the rate matrices $\mathbf{R}$ obtained during the Monte Carlo sampling in parameter space,

$$p(\tau|\text{data}) = \langle \delta[\tau - \tau(\mathbf{R})] \rangle_L \tag{10}$$

where the average is over parameters in $\mathbf{R}$ weighted with the likelihood function L according to eq 8. The distribution $p(\tau|\text{data})$ provides us with "error bars" for the inferred value of τ. In addition to the average, $\hat{\tau} = \int \tau p(\tau|\text{data})\, d\tau$, we will report credibility intervals that cover the equivalent of ± 1 standard deviations, as obtained from the points where the cumulative distribution, $P(\tau|\text{data}) = \int_{-\infty}^{\tau} p(\tau'|\text{data})\, d\tau'$, reaches 0.1587 and $1 - 0.1587$, respectively.

2.4. Maximum-Likelihood Estimate. Using essentially the same procedure as described above, we can also determine a maximum-likelihood estimate of the rate-matrix $\mathbf{R}$. In such an approach, the rate coefficients R_{ij} parametrizing the model are varied to maximize the likelihood function L given in eq 5. To find the maximum, we could adapt the Monte Carlo sampling of the parameters described above to perform simulated annealing.[31] The sampling in parameter space is again performed with a uniform prior, but now with an energy function, $-(\ln L)/T$, scaled by a "temperature" T that is slowly reduced to zero. Alternatively, one could also locate maxima of L with a gradient-based approach, adapting expressions from quantum-mechanical perturbation theory[32] for the derivatives of the propagators $p(i,t|j,0)$ with respect to the rate coefficients R_{ij}.

2.5. Direct Estimate of the Transition Matrix. Instead of first constructing a rate matrix $\mathbf{R}$ to obtain the transition matrix through $M_{ij}(t) = p(i,t|j,0) = (e^{t\mathbf{R}})_{ij}$, we could estimate $M_{ij}(t)$ directly as the fraction of MD replica runs starting from cell j at time 0 that end up in cell i at time t. At equilibrium, the total number of transitions from i to j, and from j to i is the same, which can be used to enforce detailed balance. From the transition matrix $\mathbf{M}(t)$, we can in turn estimate the limiting distribution(s) P_j as the eigenvector(s) corresponding to eigenvalue 1,

$$P_i = \sum_j M_{ij} P_j \tag{11}$$

If two or more eigenvalues are equal to 1, then the transition matrix is not connected. If detailed balance is not satisfied (i.e., $M_{ji}P_i \neq M_{ij}P_j$), the P_i correspond to a steady state, not an equilibrium distribution. To propagate the system by steps of time t, we could repeatedly apply the transition matrix $\mathbf{M}(t)$. This procedure would be the discrete analogue of the Chapman–Kolmogorov iterations used to propagate the structure of a peptide in dihedral-angle space.[3] We note that for short times t (compared to the fastest relaxation time), we can approximate $e^{t\mathbf{R}} \approx \mathbf{1} + t\mathbf{R}$, such that the off-diagonal rate-matrix elements are given by the number of transitions per unit time,

$$R_{ij} \approx M_{ij}(t)/t \qquad \text{for} \qquad i \neq j \text{ and } t \to 0 \tag{12}$$

This approximate relation can be used to estimate the rate matrix $\mathbf{R}$.

2.6. Testing the Assumption of Markovian Dynamics. Models with non-Markovian dynamics tend to be considerably more complicated than the rate model eq 2. By comparing the estimated rate matrices obtained at different observation intervals $\Delta t = t_\alpha$, we can investigate the contributions of non-Markovian effects. If the dynamics appear to be non-Markovian, one can either expand the state space (by increasing the number of cells and/or introducing new coarse variables) in the hope of obtaining an overall more Markovian dynamics, or attempt to estimate non-Markovian models such as eq 1.

2.7. On-the-Fly Construction of Coarse Master Equation. Above, we implicitly assumed that all necessary MD simulations were performed up front to construct the coarse master equation. In many practical situations, for instance, the folding of a protein, that may not be possible. Instead, one can perform the MD simulations *on demand*. In such an approach, one fills in elements of the transition matrix as new cells are visited. Assume that we have already obtained propagators $p(i,t|j,0)$ for cell $j \in J = \{j_1, j_2, ..., j_n\}$. If one of the replicas crosses into a cell i from which no runs have yet been started, $i \notin J$, we can next initiate simulations in cell i. The decision to initiate runs from i may also be based on the number of times i has been reached by simulations starting from $j \in J$, and the presumed relevance of i in the overall dynamics. As new cells are added, the dimension of the coarse master equation expands. If one can expect smoothness along the coarse coordinates, we can also attempt to interpolate the rate matrix elements,[3] or use an appropriate prior in the Bayesian formalism.[33]

3. Results

We have applied the Bayesian formalism to estimate the coarse master equation for the filling and emptying of short carbon nanotubes in a bath of water molecules. This system has been studied extensively by MD simulations[6,34] and has been shown to be representable by a lattice fluid model.[35] MD simulations showed that a narrow (~0.8 nm carbon–carbon diameter) tube at ambient conditions (300 K temperature and 1 bar pressure) filled with water molecules forming a hydrogen-bonded wire.[26,34] When the carbon–water attractive interactions were reduced, the system fluctuated between filled and empty states on a time scale of 0.1–1 ns for a tube of about 1.5 nm length.

Here, we use the number of water molecules $n(t)$ inside a nanotube surrounded by a water bath as a coarse observation variable for the filling process. A continuous occupancy number is defined as $n_{\text{cont}} = \sum_{i=1}^{N_{\text{wat}}} f(r_i,z_i)$, where the sum extends over the N_{wat} water molecules of the MD system with r_i and z_i being their radial and axial positions in the cylinder coordinate system (relative to the tube center) defined by instantaneous position and orientation of the freely moving nanotube. The weight function is given by $f(r,z) = \exp[-(2z/L)^6 - (r/R)^6]$ where $L = 1.35$ nm and $R = 0.405$ nm are the length and radius of the pore, respectively. To discretize, we round n_{cont} to the nearest integer n, which is our coarse observation variable.

Nanotube Filling with Water from Equilibrium MD. In a first illustrative example, we derive a coarse master equation from the $n(t)$ data of a long (47 ns) equilibrium run of a nanotube with modified carbon–water interactions.[26] Specifically, we count the number of transitions of n from one value to another within consecutive time intervals of a given length of $\Delta t = 1$, 2, 4, 8, and 16 ps. For the coarse master equation, we used both the complete rate matrix (20 free coefficients for $n = 0$, 1, ..., 5) and the reduced sequential model with transitions only between n and $n \pm 1$ (10 coefficients). The results for the full

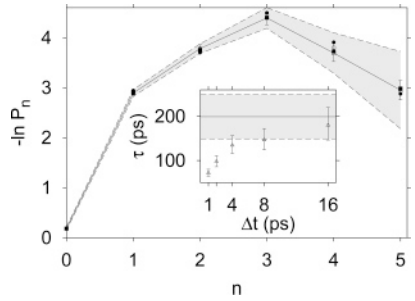

Figure 1. Free energy, $-\ln P_n$, and kinetics (inset) of water occupancy fluctuations in a carbon nanotube with "modified" carbon−water interactions corresponding to $\lambda \approx 0.752$.[26,34] The solid line shows the result of a 47 ns equilibrium simulation. The shaded gray area bounded by dashed lines corresponds to one standard deviation of the mean estimated from block averages. The symbols are the results of Bayesian inference using the kinetic model of eq 2 for sampling intervals of $\Delta t = 1, 2, 4, 8$, and 16 ps. The error bar (shown representatively for $\Delta t = 16$ ps) is the credibility range corresponding to ± 1 standard deviation (~68% credibility). The inset shows the relaxation time τ of the water occupancy number n. The solid line was obtained by integrating the normalized autocorrelation function of n from the 47 ns equilibrium run up to the point where it crosses the zero axis for the first time. The gray area bounded by dashed lines corresponds to ± 1 standard deviations estimated from block averages. The symbols are the relaxation times corresponding to the slowest decaying mode in the kinetic model eq 2, estimated for different sampling intervals Δt. Error bars correspond to 68% credibility intervals.

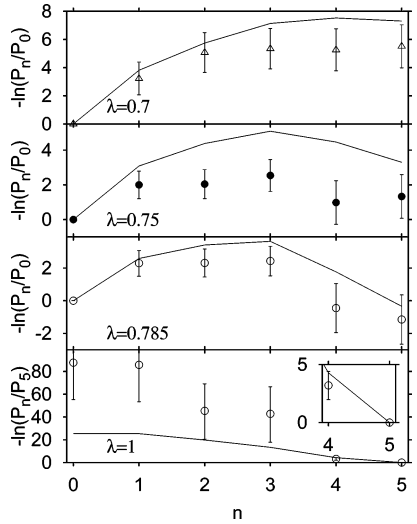

Figure 2. Free energy of water occupancy fluctuations in carbon nanotubes. Results for the free energy difference relative to the empty state, $-\ln(P_n/P_0)$, are shown for different strengths of the carbon−water Lennard-Jones attractive interactions ($\lambda = 0.7, 0.75, 0.785$ from top to bottom). For $\lambda = 1$ (bottom), the free energy difference relative to the filled state, $-\ln(P_n/P_5)$, is shown. Solid lines (as guides to the eye) indicate the results from a multiple-histogram analysis[36] of three equilibrium simulation runs (65.5 ns at $\lambda = 1$, 47 ns at $\lambda \approx 0.752$, and 19.5 ns at $\lambda = 0.785$). Symbols are the results of the Bayesian analysis (25×6 runs of 5 ps length per λ value). Error bars indicate 68% credibility ranges. The inset in the bottom panel shows the profile in the thermally accessible range ($n = 4, 5$) for $\lambda = 1$.

and reduced model are almost identical, and the rate coefficients between nonnearest neighbors are found to be vanishingly small. In Figure 1, we compare the free energy profiles from the equilibrium simulation to those calculated for the model given by the rate equations, eq 2, using Bayesian inference. We find that the free energy profile is accurately reproduced by the model even at the shortest sampling interval of 1 ps. For sampling intervals of $\Delta t \geq 4$ ps, the relaxation time τ of the occupancy number n is consistent with that estimated directly from the equilibrium run. We note that using the short-time approximation, eq 12, to estimate rate matrices from the long equilibrium MD run leads to very similar equilibrium populations and relaxation times. The increase of the inferred τ in Figure 1 with the length of the sampling interval Δt indicates that the estimated relaxation rate is more sensitive to effects of non-Markovian dynamics than the underlying free energy surface, which is practically independent of the sampling time. This sensitivity of time constants, but not equilibrium constants, is expected because a significant part of the non-Markovian dynamics can be attributed to "mis-assigning" states. For instance, the coarse variable n shows rapid fluctuations ± 1 that are caused by structural rearrangements of the water chain inside the nanotube and may not reflect an actual filling or emptying transition. Such correlated fluctuations result in an overestimate of the rate coefficients connecting neighboring states at small Δt.

Nanotube Filling with Water from CMD Replica Runs. The preceding example shows that we can use Bayesian inference to estimate a coarse master equation from a long equilibrium run. In the following, we will show that a Bayesian analysis of CMD simulations[3] will allow us to estimate the underlying *slow* dynamics, represented by a coarse master equation, from multiple short runs initialized at different values of the coarse variables. To illustrate this approach, we have used the same system, a nanotube in water. Initial conditions for different values of the water occupancy number n were created by adding a harmonic biasing potential $\kappa(n - n_0)^2/2$ to the MD Hamiltonian. This biasing potential drives the system near target occupancies n_0. We have studied four systems with different nanotube-water interaction potentials, measured by a factor λ that scales the r^{-6} carbon−water attractive Lennard-Jones interaction of the unmodified system.[34] We have found previously[34] that for $\lambda \approx 0.75$ and below, the nanotube surrounded by water remains mostly empty; for higher λ values, the water-filled state of the tube is preferred. For λ near 0.7−0.8, the tube fluctuates between a filled and empty state on the nanosecond time scale of the simulations.[26,34] Initial configurations for the replica simulations were created by running 15 ps long MD simulations biased toward to the target occupancy. We assigned five sets each of random Maxwell−Boltzmann velocities to five structures saved during the last 5 ps of the biased runs. From each of the 25 resulting initial configurations near the target occupancy, we initiated MD runs of 5 ps length, comparable to the characteristic times for filling and emptying transitions.[34] With six occupancy states ($n = 0, 1, ..., 5$), the combined length of the replica simulations at a given λ value is about 6×15 ps for equilibration, and $6 \times 25 \times 5$ ps for production, or 0.84 ns total, which is ~150 times less than the combined equilibrium runs.

In Figure 2, we compare the free energy profiles for water occupancy fluctuations at different values of λ obtained from multiple short replica runs to those estimated perturbatively from a combined analysis of three long equilibrium runs (132 ns total). For the perturbative analysis of the equilibrium MD, we used a weighted histogram analysis[36] to match the distributions of n obtained for different λ values. From three equilibrium simulations with different water−carbon Lennard-Jones interactions corresponding to states in which the tube is predominantly empty, filled, and half-filled/half-empty, respectively, we collected joint histograms of the water occupancy, the sum $\sum_{i,j} r_{ij}^{-6}$ determining attractive Lennard-Jones interactions between water oxygen atoms i and nanotube carbon atoms j at distance r_{ij}, and the corresponding sum $\sum_{i,j} r_{ij}^{-12}$ for repulsive Lennard-Jones interactions. The 3-dimensional histograms were matched using the weighted histogram method.[36] From the resulting "density

TABLE 1: Relaxation Time $\tau = \int_0^\infty C(t)\,dt$ of the Number n of Water Molecules Inside a Nanotube for Different Water−Carbon Interaction Strengths λ, Where $C(t) = \langle \delta n(t)\,\delta n(0)\rangle / \langle \delta n^2 \rangle$ Is the Normalized Autocorrelation Function of $\delta n(t) = n(t) - \langle n(t)\rangle$[a]

λ	τ_{eq} (ps)	$\hat{\tau}$ (ps)	range (ps)
0.75	199 ± 35	146	84−252
0.785	191 ± 40	173	93−322

[a] The second column lists reference values from long equilibrium runs with one standard deviation estimated from block averages. The third column lists the average value of $\hat{\tau}$ obtained from the Bayesian analysis of the 5 ps replica runs, calculated as minus the reciprocal of the smallest nonzero eigenvalue of **R**. The fourth column is the 68% credibility range of $\hat{\tau}$.

of states" (for the occupancy numbers and attractive/repulsive nanotube-water interactions) we estimated the occupancy distributions for different values of λ. Overall, the agreement of the free energies of occupancy fluctuations from the equilibrium runs (132 ns total) and the Bayesian analysis of 5 ps CMD runs is good. The Bayesian approach reproduces the character of the free-energy surfaces (bistable for $\lambda = 0.7$, 0.75, and 0.785; single well for $\lambda = 1$). Moreover, for the bistable systems ($\lambda = 0.7$, 0.75, and 0.785), the barrier height between the filled and empty states is reproduced with deviations of about $1-2\ k_BT$, where k_B is Boltzmann's constant and $T = 300$ K is the temperature. The largest deviations occur for the single-well system ($\lambda = 1$), where the Bayesian analysis predicts the empty state at a very high (unfavorable) free energy of $\sim 88\ k_BT$. However, the credibility ranges are also large, and the calculated equilibrium free energies are within the ∼95% credibility range, $29-121\ k_BT$, corresponding to ±2 standard deviations; not shown) of the estimated values. This means the relatively large errors in the estimated free energies for the highly unlikely empty state at $\lambda = 1$ are reflected in a large estimated uncertainty. The reason for these large uncertainties is the sharp decrease in free energy for increasing n, such that few replica simulations made transitions to n values smaller than at their starting points.

From the Bayesian analysis of the replica runs, we can also estimate the relaxation time as the time constant of the slowest decaying mode in **R**. Results are summarized in Table 1. For the bistable state $\lambda = 0.75$, we find a characteristic time of $\hat{\tau} = 146$ ps with a 68% credibility range of 84−252 ps. From the time integral of the autocorrelation function of a 47 ns equilibrium run at $\lambda \approx 0.752$,[26,34] we obtain $\tau_{eq} = 199 \pm 35$ ps, in good agreement with the Bayesian estimate from 5 ps replica runs. For $\lambda = 0.785$, the corresponding values are $\hat{\tau} = 173$ ps with a 93−322 ps credibility interval from the Bayesian analysis, in excellent agreement with $\tau_{eq} = 191 \pm 40$ ps from a 19.5 ns equilibrium run.

4. Conclusions

We have shown how observations from long equilibrium runs can be used to construct a coarse master equation. With the same formalism, we were also able to extract a master equation from multiple short MD runs, given a coarse observation variable for which the dynamics is sufficiently Markovian on the time scale of the replica runs. In our approach, a model of the underlying dynamics is assumed, motivated by the coarse dynamics derived from a projection operator approach,[27] and the parameters of that model are then estimated from explicit MD simulations using a Bayesian analysis. Here, we consistently used uniform prior distributions. However, a Bayesian approach easily accommodates additional information. In an extension of this work, position-dependent diffusion coefficients have been estimated,[33] with a nonuniform prior imposing smoothness.

The accurate construction of coarse master equations enables both equilibrium and kinetics analyses of molecular systems. By partitioning configuration space into cells, one can connect the relatively fast cell-to-cell transition dynamics to the global and slow configurational dynamics. In practical applications, one may not be able to include all relevant configurational states a priori. However, the approach is flexible enough that new configurational cells can be added *on the fly*, expanding the dimension of the coarse master equation as new cells are included.

In a first test of the Bayesian approach for a long equilibrium run of filling-emptying transitions of a nanotube in water, we obtained accurate free energies already at very short observation intervals of ∼1 ps. At observation intervals of $\Delta t \geq 4$ ps, we also obtained accurate relaxation times for the global filling-emptying transition (∼200 ps). We also showed that a Bayesian analysis of multiple short MD runs initiated at different tube occupancies n gave good free energies and rates of filling and emptying.

In all examples, we assumed that initial non-Markovian behavior is already integrated out at the observation time Δt, allowing us to use a Markovian model for the dynamics, eq 2. If non-Markovian effects are relevant, estimates of the rate matrix will depend on the time $\Delta t = t_\alpha$ at which we observe the replica simulations. Comparing rate matrixes **R** estimated at different time intervals Δt provides information about the neglected non-Markovian dynamics. Non-Markovian effects can be suppressed or eliminated by expanding the state space, and by using better coarse variables, obtained, for instance, via nonlinear principal component analysis[37,38] or path sampling approaches.[1,39] Alternatively, one can use non-Markovian models with corresponding path probabilities, eq 6.

The goal here was to estimate a coarse master equation described by a rate matrix. Implicitly, we assumed that the transitions between different states are sufficiently fast such that they can be sampled directly, either in long equilibrium simulations or in replica simulations initialized in the various states. In both cases, rate coefficients R_{ij} for transitions from one state j to many states i are probed simultaneously. However, if some of the transitions are inherently slow, then one may need either a partitioning into finer states, for instance, by introducing a continuous reaction coordinate,[3] or Bayesian estimates augmented by direct calculations of rate coefficients using, e.g., reactive flux calculations[10−13] or transition-path sampling methods.[40−42]

In conclusion, we want to address the interesting point of why it is at all possible to estimate the large number of coefficients in a multistate rate matrix. The success of the Bayesian approach can be rationalized by the fact that we have data for multiple different initial conditions, each probing a "column" of the rate matrix. As a matter of fact, the CMD approach[3,43] is ideally suited to reduce uncertainties in the estimated rate coefficients by selectively initiating new replica runs at states with the largest estimated errors. Studying systems with much larger numbers of states than the ones considered here should thus be possible.

Acknowledgment. G.H. thanks Dr. Attila Szabo and Dr. Alexander Berezhkovskii for many stimulating discussions. S.S. and I.G.K. acknowledge support through AFOSR and NSF/ITR grants.

References and Notes

(1) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291.

(2) Zwanzig, R. *Nonequilibrium Statistical Mechanics*; Oxford University Press: New York, 2001.
(3) Hummer, G.; Kevrekidis, I. G. *J. Chem. Phys.* **2003**, *118*, 10762.
(4) Grubmüller, H. *Phys. Rev. E* **1995**, *52*, 2893.
(5) Keasar, C.; Elber, R.; Skolnick, J. *Folding Design* **1997**, *2*, 247.
(6) Huber, T.; van Gunsteren, W. F. *J. Phys. Chem. A* **1998**, *102*, 5937.
(7) Voter, A. F. *Phys. Rev. B* **1998**, *57*, R13985.
(8) Yeh, I. C.; Hummer, G. *J. Am. Chem. Soc.* **2002**, *124*, 6563.
(9) Snow, C. D.; Nguyen, N.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102.
(10) Chandler, D. *Introduction to Modern Statistical Mechanics*; Oxford Unversity Press: New York, 1987, Chapter 8.3.
(11) Chandler, D. *J. Chem. Phys.* **1978**, *68*, 2959.
(12) Montgomery, J. A., Jr.; Chandler, D.; Berne, B. J. *J. Chem. Phys.* **1979**, *70*, 4056.
(13) Berne, B. J.; Borkovec, M.; Straub, J. E. *J. Phys. Chem.* **1988**, *92*, 3711.
(14) Kopelevich, D. I.; Panagiotopoulos, A. Z.; Kevrekidis, I. G. *J. Chem. Phys.* **2005**, in press.
(15) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J. Comput. Phys.* **1999**, *151*, 146.
(16) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571.
(17) 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. *J. Phys. Chem. B* **2004**, *108*, 6582.
(18) Chekmarev, D. S.; Ishida, T.; Levy, R. M. Submitted for publication.
(19) de Groot, B. L.; Daura, X.; Mark, A. E.; Grubmüller, H. *J. Mol. Biol.* **2001**, *309*, 299.
(20) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495.
(21) Andrec, M.; Levy, R. M.; Talaga, D. S. *J. Phys. Chem. A* **2003**, *107*, 7454.
(22) Kou, S. C.; Xie, X. S.; Liu, J. S. *Appl. Stat.* **2005**, *54*, 1.
(23) McSharry, P. E.; Smith, L. A. *Phys. Rev. Lett.* **1999**, *83*, 4285.
(24) Meyer, R.; Christensen, N. *Phys. Rev. E* **2000**, *62*, 3535.
(25) Smelyanskiy, V. N.; Timucin, D. A.; Bandrivskyy, A.; Luchinsky, D. G. Available at arxiv.org:physics/0310062.
(26) Hummer, G.; Rasaiah, J. C.; Noworyta, J. P. *Nature* **2001**, *414*, 188.
(27) Zwanzig, R. *J. Stat. Phys.* **1983**, *30*, 255.
(28) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964.
(29) O'Hagan, A. *Kendall's Advanced Theory of Statistics. Bayesian Inference*; John Wiley & Sons: New York, 1994; Vol. 2B.
(30) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.
(31) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. *Science* **1983**, *220*, 671.
(32) Messiah, A. *Quantum Mechanics*; Dover Publications: Mineola, NY, 1999.
(33) Hummer, G. *New J. Phys.* **2005**, in press.
(34) Waghe, A.; Rasaiah, J. C.; Hummer, G. *J. Chem. Phys.* **2002**, *117*, 10789.
(35) Maibaum, L.; Chandler, D. *J. Phys. Chem. B* **2003**, *107*, 1189.
(36) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195.
(37) Belkin, M.; Niyogi, P. *Neural Comput.* **2003**, *15*, 1373.
(38) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. *Proc. Natl. Acad. Sci. U.S.A.*, submitted for publication.
(39) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877.
(40) Dellago, C.; Bolhuis, P. G.; Chandler, D. *J. Chem. Phys.* **1999**, *110*, 6617.
(41) van Erp, T. S.; Moroni, D.; Bolhuis, P. G. *J. Chem. Phys.* **2003**, *118*, 7762.
(42) Hummer, G. *J. Chem. Phys.* **2004**, *120*, 516.
(43) Kevrekidis, I. G.; Gear, C. W.; Hummer, G. *AIChE J.* **2004**, *50*, 1346.