New Journal of Physics
The open–access journal for physics

Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations

Gerhard Hummer

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520, USA
E-mail: Gerhard.Hummer@nih.gov

New Journal of Physics **7** (2005) 34
Received 28 September 2004
Published 31 January 2005
Online at http://www.njp.org/
doi:10.1088/1367-2630/7/1/034

Abstract. Bayesian inference is used to obtain self-consistent estimates of free energies and position-dependent diffusion coefficients along complex reaction coordinates from molecular dynamics simulation trajectories. Effectively, exact solutions for the dynamics of a diffusive model are matched globally to the observed molecular dynamics data. The approach is first tested for a simple one-dimensional diffusion model, and then applied to the dihedral-angle dynamics of a peptide fragment dissolved in water. Both long equilibrium molecular dynamics simulations and short, appropriately initialized, replica simulations are used to sample the short-time dynamics of the peptide–water system. In both cases, accurate estimates of free energies and diffusion coefficients are obtained.

PII: S1367-2630(05)87003-X

Contents

1. Introduction

Over its 65-year history, the Kramers [1] model of diffusive barrier crossing has become one of the most powerful and widely used approaches to describe transitions in molecular systems [2]. In Kramers theory, the underlying many-body dynamics of a molecular system is condensed into a one-dimensional (1D) diffusive motion along a chosen reaction coordinate Q. The problem of obtaining rate coefficients for molecular transitions then reduces to (i) finding the free-energy surface $F(Q)$ along the reaction coordinate, in particular the height and shape of the barrier, and (ii) to estimating an effective, position-dependent diffusion coefficient $D(Q)$ that describes the local dynamics on the free-energy surface. Obtaining the free-energy surface, or potential of mean force, amounts to counting population densities $p(Q) \propto \mathrm{e}^{-\beta F(Q)}$ as a function of Q under equilibrium conditions, which can be performed rigorously and with a variety of straightforward techniques of classical molecular simulations [3, 4] and of non-equilibrium pulling experiments [5]. In contrast, estimating a local diffusion coefficient poses a more serious challenge because (i) diffusive dynamics is only an approximate assumption such that no rigorous expressions for $D(Q)$, unlike $F(Q)$, is expected; and (ii) the observed dynamics is determined by a combination of free energies and diffusion coefficients, requiring a 'deconvolution' step to remove contributions from a non-uniform free-energy surface $F(Q)$ when going from observed trajectories to $D(Q)$. In view of these difficulties in obtaining $D(Q)$, it is reassuring that the rates, while exponentially sensitive to the barrier height, are only linearly proportional to the diffusion coefficient in the Kramers theory.

How can one estimate accurate position-dependent diffusion coefficients? Following the earlier works of Berne *et al* [6], Woolf and Roux [7] proposed an elegant approach in which free energies and diffusion coefficients are calculated effectively from the same set of simulations. In their formalism, umbrella sampling [8] with a harmonic bias on Q is used to sample $p(Q)$ locally. From the Laplace transformation of the autocorrelation function of the velocity $\dot{Q}$ along

the reaction coordinate in the harmonically restrained simulations, one can then estimate $D(Q)$. As shown in appendix A, the expression of Woolf and Roux can be simplified considerably, and reduced exactly to

$$D(Q = \langle Q \rangle) = \frac{\mathrm{var}(Q)}{\tau_Q}, \tag{1}$$

where $\langle Q \rangle$ is the average of the reaction coordinate Q in the biased run, $\mathrm{var}(Q) = \langle Q^2 \rangle - \langle Q \rangle^2$ is its variance and τ_Q the characteristic time of its autocorrelation function, $\tau_Q = \int_0^\infty \langle \delta Q(t) \delta Q(0) \rangle \mathrm{d}t / \mathrm{var}(Q)$ with $\delta Q(t) = Q(t) - \langle Q \rangle$. Equation (1) is a relation between the diffusion coefficient and correlation time of a harmonic oscillator with overdamped Langevin dynamics [9] that has been used, for instance, to estimate the diffusion coefficients along protein [10] and peptide folding reaction coordinates [11].

Recently, Liu *et al* [12] proposed a different approach to calculate diffusion coefficients in anisotropic and inhomogeneous systems. For anisotropic systems, such as a liquid–vapour interface, the elements of a diagonal diffusion tensor are related to the conditional mean square displacement of only those particles that remain within a given position interval. If the system is inhomogeneous with a known free-energy surface $F(Q)$, the friction coefficient in independently run Langevin dynamics simulations on $F(Q)$ is varied to match the probability of remaining within a chosen interval, as obtained from the full molecular dynamics (MD) simulation.

Here, I follow a different approach and estimate diffusion coefficients and free energies self-consistently. The central idea is that long equilibrium simulation runs or, equivalently, an ensemble of independent and appropriately initialized short simulations can be used to probe the local 'propagators' along the coordinate Q. After coarse-graining in time, one can compare the observed motions along Q with those expected from diffusive dynamics. Such a global comparison will provide self-consistent information about the free-energy surface, $F(Q)$, and the diffusion coefficient, $D(Q)$. To perform this comparison, I use a global Bayesian analysis of the simulation data. Under the assumption of diffusive dynamics (after some short initial time accounting for fast molecular processes), and for a given free-energy surface and position-dependent diffusion coefficient, a likelihood function can be constructed that gives the probability of observing exactly the motions along Q seen in the simulation runs. Using Bayes' formula, the likelihood of observations given the parametrized diffusive model is turned into a posterior density of the unknown 'parameters' $F(Q)$ and $D(Q)$ of the diffusive model, given the simulation observations. Entering additionally into Bayes' formula is the 'prior', i.e., a distribution of the parameters that reflects what is known about them before the observations are made [13]. Implicit in the formalism is the assumption that Q is a 'good' reaction coordinate. If that is not the case, the dynamics along Q is not Markovian and the estimated diffusion coefficients will depend on the timescale at which the observations were made. Such time dependences can be used as a test of the underlying assumption of diffusive dynamics.

Within this Bayesian approach, one can infer the slow dynamics in the projected coordinate Q and construct a coarse master equation represented by a rate matrix [14]. At sufficiently fine discretizations along Q, this rate matrix contains the necessary information about free energies $F(Q)$ and diffusion coefficients $D(Q)$ in an inhomogeneous system. The general procedure will be illustrated for a simple 1D test system, and applied to the analysis of dihedral-angle transitions in a hydrated peptide fragment [15].

2. Theory

2.1. Master equation

As in [14], the starting point is Zwanzig's generalized master equation for Newtonian dynamics in configuration space [16]. If the complete configuration space is divided into N non-overlapping cells, the probability $p_i(t)$ of being in cell i at time t satisfies a generalized master equation,

$$\dot{p}_i(t) = -\int_0^t \mathrm{d}t' \sum_j K_{ij}(t-t')\, p_j(t'), \tag{2}$$

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 [16]. Here, the cells i correspond to intervals along the coordinate Q. Note in particular the absence of the 'random force' (i.e., an inhomogeneous term) in equation (2), which requires that initially (at time $t = 0$) all phase-space variables are at equilibrium within a given cell i, and non-equilibrium in the initial conditions is limited to the partitioning between different cells i.

Here, one is interested in diffusion, and I assume that after some time, the dynamics becomes Markovian. With this coarse graining in time, one can approximate equation (2) by a Markovian rate equation,

$$\dot{p}_i(t) = \sum_j R_{ij} p_j(t), \tag{3}$$

where the R_{ij} are the constant elements of a rate matrix $\mathbf{R}$, with $R_{ij} \geqslant 0$ for $i \neq j$, $R_{ii} \leqslant 0$, and $\sum_i R_{ij} \equiv 0$. One can solve equation (3) in terms of a matrix exponential,

$$p_i(t) = \sum_j (\mathrm{e}^{t\mathbf{R}})_{ij} p_j(0). \tag{4}$$

The 'propagators' of the Markovian model are accordingly given by

$$p(i, t|j, 0) = (\mathrm{e}^{t\mathbf{R}})_{ij}, \tag{5}$$

where $p(i, t|j, 0)$ is the conditional probability that a trajectory starting from cell i (with equilibrium initial conditions in phase space within i) is in cell j at a later time t.

2.2. Diffusive dynamics

In the following, I will connect this general rate formalism to diffusive dynamics by spatially discretizing the Smoluchowski diffusion equation and turning it into a system of rate equations. The resulting rate matrix $\mathbf{R}$ describes the dynamics between neighbouring intervals along Q, and contains information about free energies $F(Q)$ and diffusion coefficients $D(Q)$.

The Smoluchowski diffusion equation describes the time evolution of the probability density $p(Q, t)$ along the coordinate Q,

$$\frac{\partial p(Q,t)}{\partial t} = \frac{\partial}{\partial Q}\left\{ D(Q)\mathrm{e}^{-\beta F(Q)} \frac{\partial}{\partial Q}[\mathrm{e}^{\beta F(Q)} p(Q,t)] \right\}, \tag{6}$$

with $\beta^{-1} = k_B T$, where k_B is Boltzmann's constant and T the absolute temperature. Equation (6) can be discretized in space following Bicout and Szabo [9] to give a system of rate equations in the form of equation (3). In one dimension, the resulting rate equations take the following simple form

$$\dot{p}_i(t) = R_{i,i-1}\, p_{i-1}(t) - (R_{i-1,i} + R_{i+1,i})\, p_i(t) + R_{i,i+1}\, p_{i+1}(t) \quad \text{for } 1 \leqslant i \leqslant N, \tag{7}$$

where $p_i(t)$ is the probability of being in interval i around Q_i at time t and $R_{N,N+1} \equiv R_{N1}$ and $R_{N+1,N} \equiv R_{1N}$ for periodic boundary conditions, and $R_{N,N+1} = R_{N+1,N} = 0$ for reflecting boundary conditions, respectively.

The free-energy surface can be estimated from the equilibrium probabilities P_i of being in interval i around Q_i

$$F(Q_i) \approx -k_B T \ln \frac{P_i}{\Delta Q}, \tag{8}$$

where $\Delta Q = |Q_{i+1} - Q_i|$ is the bin width (assumed to be a constant for simplicity). The vector $\mathbf{P} = (P_1, \ldots, P_N)^T$ of equilibrium probabilities is an eigenvector of $\mathbf{R}$ with eigenvalue 0, $\mathbf{R} \cdot \mathbf{P} = 0$. The position-dependent diffusion coefficients, $D_{i+1/2} = D[(Q_i + Q_{i+1})/2]$, are related to the rate matrix and its equilibrium distribution through [9]

$$D_{i+1/2} \approx \Delta Q^2 R_{i+1,i} \left(\frac{P_i}{P_{i+1}}\right)^{1/2}, \tag{9}$$

where i and $i+1$ are two neighbouring cells at centre-to-centre distance $|Q_{i+1} - Q_i|$. Note that equation (9) is symmetric with respect to exchanging i and $i+1$ because the rate matrix satisfies detailed balance, $R_{i+1,i}/R_{i,i+1} = P_{i+1}/P_i = \mathrm{e}^{-\beta[F(Q_{i+1}) - F(Q_i)]}$. With equations (8) and (9), the positional coarse graining of the diffusion equation, equation (6), has turned the problem of finding $D(Q)$ and $F(Q)$ into that of estimating rate coefficients for transitions between adjacent intervals along Q. For systems that do not satisfy detailed balance (e.g., driven systems), but can be described by the kinetic scheme equation (3), the full rate matrix R_{ij} could be estimated.

2.3. Likelihood function

Following [14], Bayesian inference will be used to estimate the rate matrix $\mathbf{R}$ from either long equilibrium simulations or short replica simulations. Consider that one has made the following observations in either the short replica runs or the equilibrium trajectory: at a time t_α after the system was in state j_α, the system is found in state i_α. Under the assumption of the kinetic model equation (3), the likelihood of such an observation is $p(i_\alpha, t_\alpha | j_\alpha, 0) = (\mathrm{e}^{t_\alpha \mathbf{R}})_{i_\alpha j_\alpha}$. The likelihood L of a series of such observations is then simply the product of these probabilities,

$$L = \prod_\alpha p(i_\alpha, t_\alpha | j_\alpha, 0) = \prod_\alpha (\mathrm{e}^{t_\alpha \mathbf{R}})_{i_\alpha j_\alpha}, \tag{10}$$

assuming that the observations are independent and that the dynamics is given by the Markovian rate matrix $\mathbf{R}$. For non-Markovian dynamics, the observations are the actual paths, with a generalized path action functional determining their likelihood [14].

2.4. Bayesian analysis of simulation data

To estimate the rate coefficients R_{ij} in equation (7), I use the Bayesian-inference method of [14]. Alternatively, one could also use a maximum-likelihood approach. In the Bayesian formalism, a posterior distribution of the model parameters R_{ij} is constructed from the simulation data through

$$p(\text{parameters}|\text{data}) \propto p(\text{data}|\text{parameters})\, p(\text{parameters}), \tag{11}$$

where $p(\text{data}|\text{parameters}) \equiv L$ is given by the likelihood function L of equation (10). In the following, I will assume a uniform prior distribution of model parameters, $p(\text{parameters}) = \text{constant}$, such that

$$p(\text{parameters}|\text{data}) \propto L. \tag{12}$$

Implicit in the 'uniform' prior is the assumption of a particular integration measure because $p(\text{parameters}|\text{data})$ is a probability density.

The condition of detailed balance reduces the number of free parameters in $\mathbf{R}$, which can be written as

$$R_{ij} = \begin{cases} R_{ij} & \text{if } i > j, \\ -\sum_{l(\neq i)} R_{li} & \text{if } i = j, \\ R_{ji} P_i / P_j & \text{if } i < j. \end{cases} \tag{13}$$

For N states, there are $N-1$ free equilibrium probabilities P_i (with $1 \geqslant P_i > 0$ and the Nth P_i given by normalization, $\sum_i P_i = 1$) and $N(N-1)/2$ free rate coefficients in general. For 1D diffusion, the number of free off-diagonal rate coefficients is reduced to N and $N-1$ with periodic and reflecting boundary conditions, respectively.

To obtain the propagators $p(i, t|j, 0)$ for a given $\mathbf{R}$, I use an eigenvalue technique for the symmetric matrix $\tilde{R}_{ij} = {P_i}^{-1/2} R_{ij} {P_j}^{1/2}$. Its matrix $\mathbf{U}$ of eigenvectors satisfies $\tilde{\mathbf{R}}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_N)$ is the diagonal matrix of eigenvalues. The matrix exponential determining the propagators then becomes $\mathrm{e}^{t\mathbf{R}} = \text{diag}({P_1}^{1/2}, \ldots, {P_N}^{1/2}) \mathrm{e}^{t\tilde{\mathbf{R}}} \text{diag}({P_1}^{-1/2}, \ldots, {P_N}^{-1/2})$, where $\mathrm{e}^{t\tilde{\mathbf{R}}} = \mathbf{U}\mathrm{e}^{t\mathbf{\Lambda}}\mathbf{U}^T$ with $\mathrm{e}^{t\mathbf{\Lambda}} = \text{diag}(\mathrm{e}^{t\lambda_1}, \ldots, \mathrm{e}^{t\lambda_N})$. In some cases, it may also be possible to avoid the spatial discretization of the diffusion equation and instead use exact analytic expressions for the continuous propagators $p(Q, t|Q', 0)$, or at least accurate short-time expansions for locally smooth free-energy surfaces $F(Q)$ and diffusion coefficients $D(Q)$ (such as the short-time propagators used to generate diffusive trajectories [17]). Such expressions could be used to obtain the likelihood function without numerical matrix computations.

To construct posterior distributions of the parameters according to equation (11), I will sample parameters using Metropolis Monte Carlo simulations [14, 18] in which the negative log-likelihood, $-\ln L$, serves as an effective energy function in parameter space. In the Monte Carlo simulations, the equilibrium free energies, $g_i = -\ln P_i$, and the rate coefficients R_{ij} $(i > j)$ are randomly varied, the latter being restricted to the positive axis. For both R_{ij} and g_i, a uniform prior is assumed, unless specified otherwise. According to the Metropolis criterion [18], Monte Carlo moves are always accepted if the log-likelihood increases; if the log-likelihood decreases by Δ, the move is accepted with probability $\exp(-\Delta)$. From these Monte Carlo simulations in parameter space, one also obtains directly the posterior distributions of 'observables' derived from $\mathbf{R}$, in particular, $F(Q_i)$ and $D(Q_i)$, according to equations (8) and (9). I will report the mean of these distributions, and uncertainty intervals about the mean given by the points where the cumulative distributions reach 0.1587 and $1 - 0.1587$, corresponding to $\pm$SD for a Gaussian distribution.

2.5. Prior for smooth free energies and diffusion coefficients

In many practical situations, one may expect that the free-energy profile $F(Q)$ and the position-dependent diffusion coefficient $D(Q)$ are smooth, such that their values in neighbouring intervals should be similar. Such '*a priori*' expectations can be imposed by choosing an appropriate prior 'p(parameters)' in equation (11). Specifically for $D(Q)$, one can multiply the likelihood L by a weighting function that puts a harmonic penalty on deviations $|D(Q_i) - D(Q_{i+1})|$ of diffusion coefficients at adjacent grid points,

$$p(\text{parameters}|\text{data}) \propto L \prod_i \mathrm{e}^{-[D(Q_i)-D(Q_{i+1})]^2/2\gamma^2}, \tag{14}$$

where small values of the parameter γ impose smooth $D(Q)$. Alternatively, one could also impose smoothness by using low-order series expansions of the position-dependent $F(Q)$ and $D(Q)$, and then obtain Bayesian estimates of the series expansion coefficients.

3. Results

3.1. Test for 1D diffusion

To test the algorithm for reconstructing $F(Q)$ and $D(Q)$ from simulation data, I first analyse a simple model system. The purpose of this comparison is to explore whether the Bayesian procedure can accurately recover both the underlying free-energy surface and the position-dependent diffusion coefficients from noisy simulation data created for a known diffusive model. The model was designed to mimic the more complicated dihedral-angle dynamics of a peptide fragment studied below. Specifically, I simulate 1D diffusion along a coordinate $Q = \psi$ on a periodic free-energy surface, $\beta F(\psi) = -\cos(2\psi) + \text{const.}$, with free energy minima at $\psi = 0$ and $\psi = \pm\pi$ separated by barriers of height $2k_\mathrm{B}T$. The position-dependent diffusion coefficient $D(\psi) = (2 + \sin\psi)D_0$ oscillates between D_0 and $3D_0$, with $D_0 = 0.1\,\mathrm{rad}^2\,\mathrm{ps}^{-1}$. To create a diffusive equilibrium trajectory, I advance the angle ψ in finite time steps of $\Delta t = 0.001$ ps according to $\psi(t + \Delta t) = \psi(t) + \{D'[\psi(t)] - \beta D[\psi(t)]F'[\psi(t)]\}\Delta t + g_t\{2D[\psi(t)]\Delta t\}^{1/2}$, where the g_t are uncorrelated Gaussian random variables of mean zero and variance one, and primes denote first derivatives with respect to ψ [17].

Along a 100 ns equilibrium trajectory, I save the angle $\psi(t)$ every $\tau = 0.5$ ps. From the resulting time series of ψ values, I extract 0.5 ps transitions between cells defined as ψ-intervals of uniform width $\Delta\psi = 2\pi/n$ for $n = 12$, 16, 24, 36 and 48. To reduce the dependence on the number of bins n, I translate the initial point in a trajectory step to the centre of its interval, and shift the final point accordingly. In this way, I collect the number N_{ij} of transitions from cell j to cell i. For a given model $\mathbf{R}$, the logarithm of the likelihood then becomes a double sum over cell indices,

$$\ln L = \sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij} \ln(\mathrm{e}^{t\mathbf{R}})_{ij}. \tag{15}$$

This likelihood function is used in the Bayesian estimate together with a uniform prior distribution for the equilibrium free energies, $-k_\mathrm{B}T \ln P_i/\Delta\psi$, and rate coefficients, R_{ij} $(i > j)$. By sampling

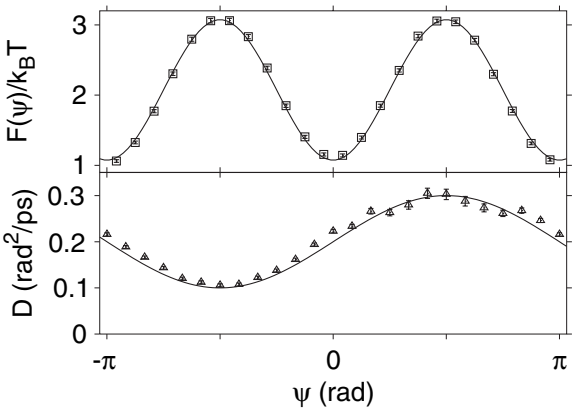

Figure 1. Free energy (top) and diffusion coefficient (bottom) for diffusion on a 1D periodic surface. The reference free energy $\beta F(\psi) = -\cos(2\psi) + \ln \int_{-\pi}^{\pi} \mathrm{e}^{\cos 2\psi'} \mathrm{d}\psi'$ and the position-dependent diffusion coefficient $D(\psi) = 0.1(2 + \sin\psi)\,\mathrm{rad}^2\,\mathrm{ps}^{-1}$ are shown as lines. Estimates obtained from a Bayesian analysis with $n = 24$ grid points are shown as symbols. Error bars indicate a credibility range of 68%.

the resulting posterior distribution of rate matrices, one obtains distributions of free energies, $-k_\mathrm{B}T \ln P_i/\Delta\psi$, and local diffusion coefficients $D(\psi_i)$ according to equation (9).

Figure 1 shows results for free energies and diffusion coefficients for $n = 24$ cells. One finds that the Bayesian estimate accurately recovers the underlying free-energy surface. Small deviations between the estimated and exact $F(\psi)$ in figure 1 are caused by the insufficient sampling of the 100 ns trajectory, with the minimum at $\psi = \pm\pi$ slightly more populated during the simulation run. The Bayesian estimate of the position-dependent diffusion coefficient also agrees well with the $D(\psi)$ used in the simulation. Note that the $t = 0.5$ ps observation time is comparable to the characteristic relaxation time in one of the free-energy wells, $[\beta F''(0)\, D(0)]^{-1} = 1.25$ ps. At such a relatively long timescale, estimates of $F(\psi)$ and $D(\psi)$ from the local average drift and spread of trajectories would require substantial curvature corrections. Estimating the local diffusion coefficient as $D(\psi_0) \approx \mathrm{var}(\psi; \psi_0, t)/2t$ from the variance of trajectories starting from a fixed initial point ψ_0 and evolving for time $t = 0.5$ ps results in over- and under-estimates of D by about 30% at the barrier tops and free energy minima, respectively.

As shown in figure 2, the estimated diffusion coefficients approach their reference value $D(\psi)$ with an error that depends quadratically on the bin width. The discretization of the Smoluchowski diffusion equation, equation (7), uses centred finite differences [9] with errors of the order of $\Delta\psi^2$. For $n = 24$ grid points, the estimated diffusion coefficients deviate by up to about 10% from the corresponding reference values. Overall, the results for the 1D test system show that the Bayesian analysis can produce accurate free energies and diffusion coefficients.

3.2. Alanine dipeptide in water

In the previous example, the underlying dynamics was diffusive by construction. In the following, I will use MD simulations of a small peptide fragment in water, and approximate its dihedral-angle dynamics by Smoluchowski diffusion. To estimate the free-energy surface and diffusion coefficient of the hydrated alanine dipeptide along the $Q = \psi$ dihedral angle, I again use the

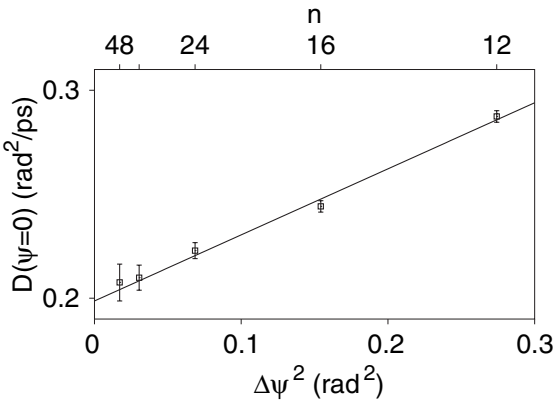

Figure 2. Dependence of the estimated local diffusion coefficient at $\psi = 0$ on the number of grid points (top scale) and the square of the bin width (bottom scale). The target value is $D(0) = 0.2\,\mathrm{rad}^2\,\mathrm{ps}^{-1}$. The solid line shows a linear fit of the estimated values with respect to the squared bin width, $\Delta\psi^2 = (2\pi/n)^2$.

Bayesian analysis with a uniform prior. Figure 3 compares the results for $F(\psi)$ from the Bayesian analysis of the replica runs of [15] to a long equilibrium MD simulation in explicit solvent. In a previous analysis [15], a related Chapman–Kolmogorov approach had been used in which local short-time propagators were calculated, but without imposing detailed balance on the coarse master equation. Without detailed balance, one obtains a steady state rather than an equilibrium solution. I find here that the Bayesian re-analysis, with detailed balance imposed, leads to improved estimates of the free-energy profile.

The Bayesian approach also produces a self-consistent estimate of the position-dependent diffusion coefficient $D(\psi)$, as shown in the centre panel of figure 3. For reference, a local diffusion coefficient calculated for harmonically biased equilibrium MD is included. As described in appendix A, the formalism of [7] was used to estimate $D(\psi)$ from the biased run. With the same data, I also used the analytical limit, equation (1) (see appendix A). As shown in the centre panel of figure 3, the results from the Bayesian analysis are in excellent agreement with the two estimates from the biased MD run. However, as discussed in appendix A, the two estimates from biased MD are subject to possible substantial systematic errors due to the particular choice of extrapolation schemes and integration cutoffs. The value of $D(\psi)$ near the global free energy minimum at $\psi \approx -0.3$ rad is also in good agreement with a previous estimate of $0.15\,\mathrm{rad}^2\,\mathrm{ps}^{-1}$ based on the equilibrium relaxation time [15].

In the bottom panel of figure 3, I show results for the diffusion coefficient estimated with a prior imposing 'smoothness'. In equation (14), γ was set to $0.05\,\mathrm{rad}^2\,\mathrm{ps}^{-1}$ for $n = 24$ cells. The results for the free-energy profile are essentially unchanged (not shown). The local diffusion coefficients also show no substantial changes, except for the removal of 'outliers' of $D(\psi)$ in the relatively poorly sampled barrier regions.

4. Conclusions

The objective of this paper was to extract diffusion coefficients and free energies along complex reaction coordinates self-consistently from many-particle molecular dynamics simulations.

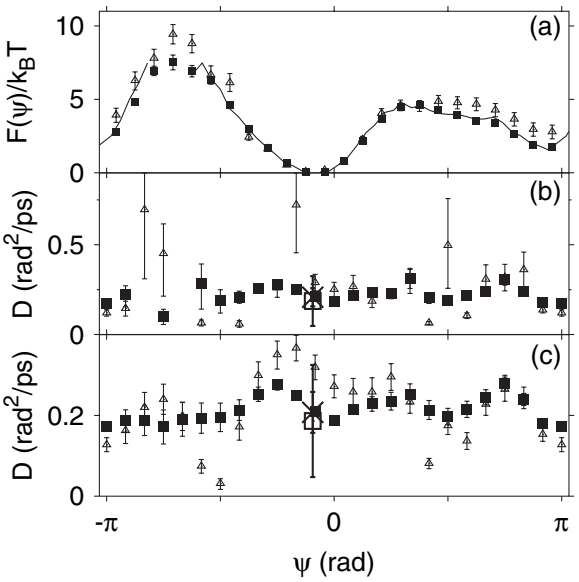

Figure 3. Free energy $F(\psi)$ and effective diffusion coefficient $D(\psi)$ of alanine dipeptide in a box of water as a function of the ψ dihedral angle. (a) Free energy: the solid line is the result of a ∼7 ns equilibrium MD run [15]. △, results from a Bayesian analysis with a uniform prior of 50×56 replica runs of 0.5 ps length; ■, Bayesian results for 0.5 ps propagators obtained from the ∼7 ns equilibrium MD run. (b) Diffusion coefficient: the position-dependent diffusion coefficients $D(\psi)$ obtained from the Bayesian analysis are shown with the same symbols as in panel (a). The × and □ show the diffusion coefficients estimated for $\psi \approx -0.3$ rad from a 240 ps simulation with a harmonic bias using equations (1) and (A.3), respectively. (c) Diffusion coefficients estimated with a smoothening prior, equation (14), for $\gamma = 0.05$ rad^2 ps^{-1}. Symbols as in panel (b). In all panels, error bars indicate a credibility range of ∼68% (corresponding to ±1 SD for a Gaussian).

The problem was approached by coarse-graining in space and time. Conformation space is divided into cells and MD trajectories are monitored at finite time intervals to identify transitions between cells. A model of diffusive dynamics is similarly coarse-grained in space, leading to coupled rate equations [9], and 'matched' to the observed dynamics with a Bayesian approach. The Bayesian formalism produces estimates of the underlying position-dependent free energies and diffusion coefficients, as well as their uncertainties. The approach is presented for diffusion along a 1D coordinate, but generalizations to higher dimensions are straightforward, following the spatial discretization procedure of Bicout and Szabo [9]. Estimates of $F(\mathbf{Q})$ and $D(\mathbf{Q})$ for a multidimensional coordinate $\mathbf{Q}$ will require trajectory data for a large number of initial conditions. However, the number of parameters can be effectively reduced if the free-energy surface or the diffusion coefficients are smooth functions of position.

To connect the diffusive model to the observed simulation data, a likelihood function is constructed. The likelihood L is defined as the probability of observing a set of transitions between conformation-space cells found in the MD simulations under the assumption of a diffusive model. Turning the likelihood around according to the Bayes theorem, one obtains a distribution of the

model parameters given the trajectory data. A 'prior' entering into the distribution of parameters is assumed to be uniform, but priors imposing smoothness on the estimated free energies and diffusion coefficients are also considered.

Approaches related to the one presented here have been used to estimate molecular rate processes from single-molecule photon-count statistics [19, 20]. Likelihood functions have also been developed for the analysis of dynamical systems [21], and have been used for Bayesian inference [22, 23]. However, the applicability of Bayesian approaches in the context of dynamical systems is still a matter of debate [24]. Clearly, with poorly chosen priors, or inappropriate models, a Bayesian approach is unlikely to lead to reliable results. Alternative approaches applied in the analysis of dynamical systems, for instance, directly relate the local average drift and the spread of trajectories at short times to the gradient of the free-energy surface and diffusion coefficient [25]–[27]. In the context of coarse molecular dynamics simulations, such expressions based on the short-time expansion of propagators have been used to estimate free energies and diffusion coefficients [15, 28].

Here, I obtained accurate estimates of the underlying free-energy surface and of local diffusion coefficients both for a simple 1D model, and for the dihedral-angle dynamics of a peptide fragment in explicit water. The diffusion coefficients agree well with those obtained by analysing a harmonically biased run using the method of Woolf and Roux [7] and the analytic limit of their formula (see appendix A). Compared to the previous estimate of the diffusion coefficient from the relaxation dynamics in a free-energy minimum [15], I found a slight speedup of the diffusion ($\sim$0.2 versus 0.15 rad^2 ps^{-1}). A relevant question is whether diffusive dynamics is useful to describe complex motions in a molecular system, such as peptide conformational changes. For alanine dipeptide in water, Bolhuis *et al* [29] used commitment probabilities to show that the ψ dihedral angle should be a relatively poor reaction coordinate for the α-to-extended transition. Nevertheless, with the diffusion coefficients estimated here, I expect a mean life time of $\sim$450 ps in the α-helical minimum. That value is bracketed by the two estimates from explicit MD simulations ($\sim$400 ps from a 7-ns run with 607 water molecules, and $\sim$800 ps from a 24-ns run with 265 water molecules). Even though this result suggests that diffusive dynamics, as estimated after coarse-graining in time, is useful here, it is essential to examine the system for significant deviations of the observed dynamics from the underlying model.

Acknowledgments

I want to thank Professor I G Kevrekidis and Dr A Szabo for many helpful discussions.

Appendix A. Diffusion coefficient from autocorrelation functions of harmonically restrained systems

From MD simulations with a harmonic biasing potential on the reaction coordinate Q, Woolf and Roux [7] estimated both free energies and diffusion coefficients. Their expression for the position-dependent diffusion coefficient uses the autocorrelation function of the 'velocity' $\dot{Q}$ along the reaction coordinate,

$$C_v(t;\, Q_i) = \langle \dot{Q}(t)\, \dot{Q}(0) \rangle_i, \tag{A.1}$$

where the harmonic bias restrains Q near Q_i. In terms of a Laplace transform, $\hat{C}_v(s;\, Q_i) = \int_0^\infty \mathrm{e}^{-st} C_v(t;\, Q_i)\, \mathrm{d}t$, Woolf and Roux arrive at the following expression [7]:

$$D(s;\, Q_i) = -\frac{\hat{C}_v(s;\, Q_i)\langle \delta Q^2\rangle_i \langle \dot{Q}^2\rangle_i}{\hat{C}_v(s;\, Q_i)[s\langle \delta Q^2\rangle_i + \langle \dot{Q}^2\rangle_i/s] - \langle \delta Q^2\rangle_i \langle \dot{Q}^2\rangle_i}, \tag{A.2}$$

where $\delta Q = Q - \langle Q\rangle_i$. This expression was obtained by integrating the memory function to estimate a friction coefficient in the presence of a harmonic potential. The local diffusion coefficient is then obtained by extrapolating $D(s)$ to $s \to 0$:

$$D(Q_i) = \lim_{s\to 0} D(s;\, Q_i). \tag{A.3}$$

One cannot immediately take this limit because, ideally, $C_v(s;\, Q_i) = 0$ for the harmonically restrained system and perfect sampling. Instead, a numerical extrapolation procedure is used in which $D(s;\, Q_i)$ is plotted as a function of s and extrapolated from some chosen range in s to $s = 0$.

However, it turns out that one can simplify the expressions given above, and bring them in a more familiar form, by using the autocorrelation function of the position instead of the velocity. With $\delta Q(t) - \delta Q(0) = \int_0^t \dot{Q}(t')\, \mathrm{d}t'$ one obtains

$$\frac{\partial}{\partial t}\langle [\delta Q(t) - \delta Q(0)]^2\rangle = 2\int_0^t \langle \dot{Q}(t')\dot{Q}(0)\rangle\, \mathrm{d}t' \tag{A.4}$$

and thus

$$\frac{\partial}{\partial t} C_Q(t;\, Q_i) = -\int_0^t C_v(t';\, Q_i)\, \mathrm{d}t', \tag{A.5}$$

where

$$C_Q(t;\, Q_i) \equiv \langle \delta Q(t)\delta Q(0)\rangle_i \tag{A.6}$$

is the autocorrelation function of the reaction coordinate (with the average removed) in biasing window i. Laplace transformation of equation (A.5) produces

$$\hat{C}_v(s;\, Q_i) = s\langle \delta Q^2\rangle_i - s^2 \hat{C}_Q(s;\, Q_i), \tag{A.7}$$

where I used that $C_Q(0;\, Q_i) = \langle \delta Q^2\rangle_i$. Substitution of equation (A.5) in equation (A.3) results in

$$D(s;\, Q_i) = \frac{\langle \delta Q^2\rangle_i \langle \dot{Q}^2\rangle_i [\langle \delta Q^2\rangle_i - s\hat{C}_Q(s;\, Q_i)]}{\hat{C}_Q(s;\, Q_i)[s^2\langle \delta Q^2\rangle_i + \langle \dot{Q}^2\rangle_i] - s\langle \delta Q^2\rangle_i^2}. \tag{A.8}$$

With $C_Q(s = 0;\, Q_i)$ finite and the divergence of equation (A.2) removed, one can now take the limit $s \to 0$:

$$D(Q_i) \equiv \lim_{s\to 0} D(s;\, Q_i) = \frac{\langle \delta Q^2\rangle_i^2}{\hat{C}_Q(0;\, Q_i)} = \frac{\langle \delta Q^2\rangle_i}{\tau_i}, \tag{A.9}$$

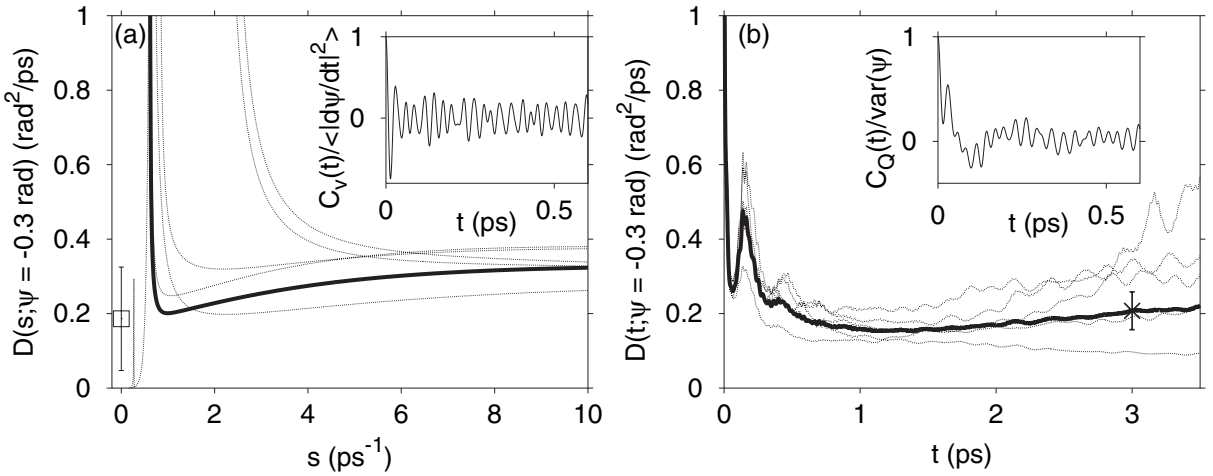

Figure A.1. Estimating diffusion coefficients from (a) the Laplace transform, equation (A.2) and (b) equation (1) as a function of the cutoff time t up to which the positional autocorrelation function was integrated to estimate the correlation time τ. Thick lines show results obtained by analysing a 240 ps biased MD run. Thin lines are the results obtained by using the same run, but subdivided into five blocks that were analysed separately. The open square in panel (a) indicates the value of D obtained by extrapolating to $s = 0$ with a quadratic fit. The cross in panel (b) is the value of D obtained by integration up to 3 ps. Those two values are included in figure 3. The insets in (a) and (b) show the normalized autocorrelation functions of the dihedral angle ψ and its velocity $\dot{\psi}$.

where τ_i is the correlation time of the reaction coordinate in biasing window i,

$$\tau_i = \frac{\int_0^\infty \langle \delta Q(t) \delta Q(0) \rangle_i \, \mathrm{d}t}{\langle \delta Q^2 \rangle_i}. \tag{A.10}$$

Equation (A.9) is exact for an overdamped harmonic oscillator, and has been used, for instance, to estimate diffusion coefficients in protein [10] and peptide folding [11].

Figure A.1 illustrates difficulties faced when estimating local diffusion coefficients from the autocorrelation functions calculated in harmonically biased simulations. A 240 ps MD simulation of alanine dipeptide in water was run as in [15], but with a harmonic restraint potential $k(\psi - \psi_0)^2/2$ added to keep ψ near $\psi_0 \approx -0.3$ rad. With a spring constant of $k = 100$ kcal mol^{-1} rad^{-2}, S.D. ~ 0.074 rad for ψ. As shown in figure A.1, the resulting autocorrelation functions of position (ψ) and velocity ($\dot{\psi}$) are highly oscillatory, making estimation of the correlation time in equations (1) and (A.9), and of $D(s)$ in equation (A.8) difficult. The resulting estimates of D from $\mathrm{var}(\psi)/\tau$ thus depend on the time range used to estimate τ. Equivalently, the estimated limit of $D(s)$ for $s \to 0$ depends on the range of s used in the extrapolation (with large s values giving high weight to the short-time correlations). Moreover, a singularity, caused by the numerical instability of equation (A.2), makes the result dependent on the chosen extrapolation method. Here, I used a quadratic fit for $3 < s < 10\,\mathrm{ps}^{-1}$. But despite these difficulties, both methods produced $D(\psi)$ values near those estimated with the Bayesian approach.

References

[1] Kramers H A 1940 *Physica* **7** 284
[2] Hänggi P, Talkner P and Borkovec M 1990 *Rev. Mod. Phys.* **62** 251
[3] Allen M P and Tildesley D J 1987 *Computer Simulation of Liquids* (Oxford: Clarendon)
[4] Frenkel D and Smit B 2002 *Understanding Molecular Simulation. From Algorithms to Applications* (San Diego: Academic)
[5] Hummer G and Szabo A 2001 *Proc. Natl Acad. Sci. USA* **98** 3658
[6] Berne B J, Borkovec M and Straub J E 1988 *J. Phys. Chem.* **92** 3711
[7] Woolf T B and Roux B 1994 *J. Am. Chem. Soc.* **116** 5916
[8] Torrie G M and Valleau J P 1974 *Chem. Phys. Lett.* **28** 578
[9] Bicout D J and Szabo A 1998 *J. Chem. Phys.* **109** 2325
[10] Socci N D, Onuchic J N and Wolynes P G 1996 *J. Chem. Phys.* **104** 5860
[11] Hummer G, García A E and Garde S 2000 *Phys. Rev. Lett.* **85** 2637
[12] Liu P, Harder E and Berne B J 2004 *J. Phys. Chem.* B **108** 6595
[13] O'Hagan A 1994 *Kendall's Advanced Theory of Statistics. Bayesian Inference*, vol 2B (New York: Wiley)
[14] Sriraman S, Kevrekidis I G and Hummer G 2005 *J. Phys. Chem.* B, at press
[15] Hummer G and Kevrekidis I G 2003 *J. Chem. Phys.* **118** 10762
[16] Zwanzig R 1983 *J. Stat. Phys.* **30** 255
[17] Ermak D L and McCammon J A 1978 *J. Chem. Phys.* **69** 1352
[18] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *J. Chem. Phys.* **21** 1087
[19] Andrec M, Levy R M and Talaga D S 2003 *J. Phys. Chem.* A **107** 7454
[20] Kou S C, Xie X S and Liu J S 2005 *Appl. Statist.* **54** 1
[21] McSharry P E and Smith L A 1999 *Phys. Rev. Lett.* **83** 4285
[22] Meyer R and Christensen N 2000 *Phys. Rev.* E **62** 3535
[23] Smelyanskiy V N, Timucin D A, Bandrivskyy A and Luchinsky D G 2003 *Preprint* physics/0310062
[24] Judd K 2003 *Phys. Rev.* E **67** 026212
[25] Siegert S, Friedrich R and Peinke J 1998 *Phys. Lett.* A **243** 275
[26] Gradišek J, Siegert S, Friedrich R and Grabec I 2000 *Phys. Rev.* E **62** 3146
[27] Friedrich R, Siegert S, Peinke J, Luck S, Siefert M, Lindemann M, Raethjen J, Deuschl G and Pfister G 2000 *Phys. Lett.* A **271** 217
[28] Kevrekidis I G, Gear C W and Hummer G 2004 *AIChE J.* **50** 1346
[29] Bolhuis P G, Dellago C and Chandler D 2000 *Proc. Natl Acad. Sci. USA* **97** 5877