Appl. Statist. (2005)
54, *Part* 3, *pp.* 469–506

Bayesian analysis of single-molecule experimental data

S. C. Kou, X. Sunney Xie and Jun S. Liu

Harvard University, Cambridge, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, October 13th, 2004*, Professor J. T. Kent *in the Chair*]

Summary. Recent advances in experimental technologies allow scientists to follow biochemical processes on a single-molecule basis, which provides much richer information about chemical dynamics than traditional ensemble-averaged experiments but also raises many new statistical challenges. The paper provides the first likelihood-based statistical analysis of the single-molecule fluorescence lifetime experiment designed to probe the conformational dynamics of a single deoxyribonucleic acid (DNA) hairpin molecule. The conformational change is initially treated as a continuous time two-state Markov chain, which is not observable and must be inferred from changes in photon emissions. This model is further complicated by unobserved molecular Brownian diffusions. Beyond the simple two-state model, a competing model that models the energy barrier between the two states of the DNA hairpin as an Ornstein–Uhlenbeck process has been suggested in the literature. We first derive the likelihood function of the simple two-state model and then generalize the method to handle complications such as unobserved molecular diffusions and the fluctuating energy barrier. The data augmentation technique and Markov chain Monte Carlo methods are developed to sample from the posterior distribution desired. The Bayes factor calculation and posterior estimates of relevant parameters indicate that the fluctuating barrier model fits the data better than the simple two-state model.

Keywords: Bayes factor; Brownian diffusion; Continuous time Markov chain; Cox process; Energy barrier; Likelihood; Ornstein–Uhlenbeck process; Scale transformation update

1. Introduction

Recent technological advances have allowed scientists to make observations on single-molecule dynamics, which was unthinkable just a few decades ago (Nie and Zare, 1997; Xie and Trautman, 1998; Weiss, 2000; Tamarat *et al.*, 2000; Moerner, 2002)—the famous physicist Richard Feynman once described that seeing the images of single atoms was a 'religious experience'. Complementary to the traditional experiments that are done on large ensembles of molecules, single-molecule experiments offer a great potential and many advantages for new scientific discoveries. First, one can directly measure the distributions of molecular properties, rather than only the ensemble average. Second, single-molecule experiments can capture transient intermediates of a biochemical process by following it in realtime, which previously could only be accomplished by synchronizing the actions of a large ensemble of molecules. Third, single-molecule trajectories provide detailed dynamic information, which is unavailable from the traditional ensemble experiments. The detailed dynamic information is particularly important for complex biomolecules that have intricate internal structures (Xie and Lu, 1999; Yang *et al.*, 2003) as the

Address for correspondence: Jun S. Liu, Department of Statistics, Harvard University, Science Center, One Oxford Street, Cambridge, MA 02138, USA.
E-mail: jliu@stat.harvard.edu

 0035–9254/05/54469

understanding of the dynamics of individual molecules is essential to reveal their biochemical properties and functions.

The new advances provide not only opportunities but also significant challenges, ranging from experimental and theoretical to statistical. The statistical challenges arise from the stochastic nature of single-molecule trajectories. Whereas the experimental and theoretical issues have attracted many researchers, statistical aspects of the single-molecule studies have received less attention. We consider here a particular type of single-molecule experiments for the study of deoxyribonucleic acid (DNA) hairpin kinetics to exemplify several statistical issues such as the modelling of single-molecule dynamics and experimental complications, efficient estimation of parameters of interest, model discrimination and statistical computation. To cope with the experimental complexity successfully, we use a Bayesian data augmentation approach, which provides much more precise estimates than the method-of-moment type of approach that is widely used in the field. We adopt the Bayesian approach here, not only because of its coherency and optimality in dealing with complex models and nuisance parameters, but also because we have quite detailed knowledge (prior) on the various parameters that are involved. The general strategies that are developed here can also be applied to other single-molecule experiments.

1.1. Deoxyribonucleic acid hairpin

A DNA hairpin is a single-stranded nucleic acid structure with bases at the two ends complementing each other so that the intramolecular pairing can form. Owing to the forming and breaking of the pairing, a DNA hairpin has two states, open and closed. In the closed state, the two ends pair together and the whole structure resembles a hairpin, whereas in the open states the intramolecular pairings are broken (Bonnet *et al.*, 1998; Krichevsky and Bonnet, 2002). Fig. 1 illustrates the closed and open states of the DNA hairpin $5'$–CCCAA–$(T)_{21}$–TTGGG–$3'$.

In a live cell, with the breaking of intermolecular pairing between the two DNA strands (double helix) the loose strand often forms a DNA hairpin structure. The DNA hairpin structure participates in many important biological functions including, for example, the regulation of gene expression (Zazopoulos *et al.*, 1997), DNA recombination (Froelich-Ammon *et al.*, 1994) and the facilitation of mutagenic events (Trinh and Sinden, 1993). The hairpin structure can also be a potential antisense drug (Tang *et al.*, 1993), as the injection into a live cell of a hairpin that has its nucleic acid bases complementing the messenger ribonucleic acid of a disease gene could potentially block the gene's expression. Because of its biological relevance, studying the conformational properties of a DNA hairpin, such as the conformational fluctuation and energy barrier between the open and closed states, serves as an important model system to understand more complicated biochemical processes.

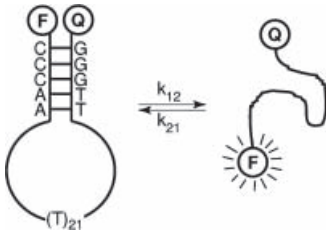

Fig. 1. The two states of a DNA hairpin: to infer the closed and open states, a fluorescence donor F and a quencher Q are attached to the two ends of the DNA hairpin

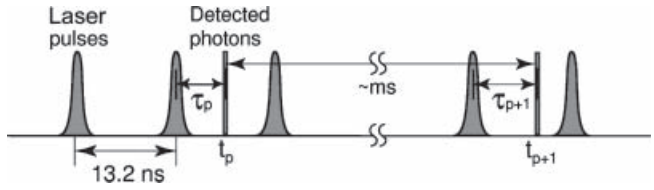

Fig. 2. The laser excites the donor dye, which releases photons: by comparing the detected photon arrival times with the regular laser pulse train, the photon delay times (with respect to their excitation pulse) are determined

1.2. Fluorescence lifetime experiments

The DNA hairpin spontaneously switches between the open and closed states. In the single-molecule fluorescence lifetime experiment under study, a fluorescence donor dye and a quencher are attached to the two ends of the DNA molecule (Fig. 1). The molecule is then placed in a focal volume illuminated by a laser pulse train. The donor dye emits photons when excited by a laser pulse, and the quencher annihilates the excitation. In the hairpin's closed state, the quenching is strong and very few photons from the donor are detected; in the open state, because the donor and quencher are far from each other, the quenching is weak and many photons from the donor are detected. For the photons detected their arrival times are recorded. By comparing the detected photon arrival time with the regular laser pulse train, we can further determine the time length that it takes the donor dye to release the photon from the moment that it is excited. This time length is termed the photon delay time. Fig. 2 illustrates the determination of the arrival times t and the delay times τ.

Let A denote closed, and let B denote open. The simplest model for the DNA hairpin dynamics is a continuous time two-state Markov chain termed the *two-state model* (Reilly and Skinner, 1993, 1994a, b): a molecule starting from state A will stay in A with an exponentially distributed waiting time and then jump to state B; and vice versa for state B. This can be depicted as

$$A \underset{k_{21}}{\overset{k_{12}}{\rightleftarrows}} B, \tag{1.1}$$

where k_{12} and k_{21} represent the transition rate constants between the two states. Let

$$Q = \begin{pmatrix} -k_{12} & k_{12} \\ k_{21} & -k_{21} \end{pmatrix}$$

be the intensity matrix (infinitesimal generator). The Kolmogorov forward equation (Fokker–Planck equation) gives the transition matrix

$$P(t) = \exp(Qt) = \begin{pmatrix} \pi_1 + \pi_2 \exp(-kt) & \pi_2\{1 - \exp(-kt)\} \\ \pi_1\{1 - \exp(-kt)\} & \pi_2 + \pi_1 \exp(-kt) \end{pmatrix}. \tag{1.2}$$

where $k = k_{12} + k_{21}$ and $(\pi_1, \pi_2) = (k_{21}/(k_{12} + k_{21}), k_{12}/(k_{12} + k_{21}))$ is the steady state distribution of the two-state Markov chain.

Both the photon arrival times t and the corresponding delay times τ depend on the underlying hidden two-state process. Let $\gamma(t)$ denote the fluorescence level of the DNA hairpin structure at time t, which takes values a and b respectively in states A and B with $a > b$. The photon arrival time t follows a doubly stochastic Poisson process with the arrival rate inversely proportional to $\gamma(t)$, where the term 'doubly stochastic' arises since both the arrival time and the arrival rate are stochastic. In the literature this process is also referred to as a Cox process (Cox, 1955). For a detected photon, its photon delay time τ follows an exponential distribution with rate

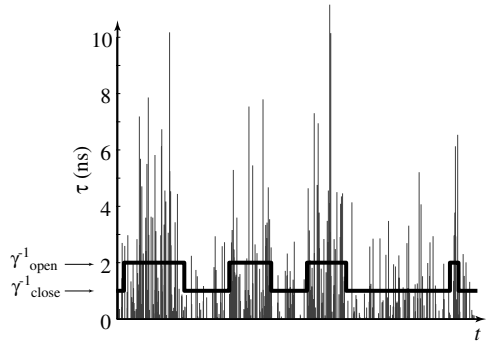

Fig. 3. Data structure in the single-molecule lifetime experiments: |, photon arrival times t in seconds (heights represent the delay time τ in nanoseconds); ⎍, unobserved two-state Markov chain, on which both t and τ depend

$\gamma(t)$. The dependence of the delay time on the underlying states is because, when the donor and quencher are close to each other (i.e. in the closed state), there is a transfer of energy from the donor to the quencher, which shortens the delay time. Fig. 3 illustrates the dependence of the data on the two-state process γ. The bold line depicts the unobserved two-state Markov chain corresponding to the open–closed states of the DNA hairpin. Each vertical bar in Fig. 3 represents the arrival (time) of a photon. We can see that in the open state (i.e. high value of γ^{-1}) there are more frequent photon arrivals. For each bar (i.e. each photon arrival), the height represents the corresponding delay time τ of that recorded photon. In the open state, the delay times τ tend to be larger. The open or closed state of the DNA hairpin can hence be inferred indirectly from the detected photon arrival times and the delay times (Lu *et al.*, 1998; Jia *et al.*, 1997; Eggeling *et al.*, 1998).

In addition to the hidden Markov structure, the *unobserved* Brownian diffusion trajectory of the molecule in the solvent adds another layer of complexity that significantly complicates the inference (see Section 3). To resolve these issues efficiently we use a Bayesian data augmentation approach, which 'imputes' the missing Brownian process (Tanner and Wong, 1987).

As increasingly more molecular data are accumulated, scientists begin to question the simple two-state model. For example, it has been argued that the energy barrier between the two states can also fluctuate dynamically or there may be substates within each of the two states and these substates may communicate at different rates (Cao, 2000; Yang and Cao, 2001, 2002). With the current data resolution and existing methods of inference, discerning the different models and assessing their fit to the experimental data have remained elusive beyond mere speculation (see Schenter *et al.* (1999) for further discussions). Using the Bayesian data augmentation approach, we are able to provide significant statistical evidence to conclude that the simple two-state model, which does *not* allow the energy barrier fluctuation between the two states, is not sufficient to explain the data. Shedding light on the nature of the energy barrier could be the first step towards a comprehensive understanding of a DNA hairpin's biophysical properties.

The paper is organized as follows. Section 2 details the two-state statistical model and the corresponding Bayesian analysis based on a closed form likelihood. Section 3 introduces the data augmentation approach for handling experimental complications, such as the molecular Brownian motion. Section 4 considers models beyond the two-state case. Section 5 analyses experimental as well as simulated data and discusses the issue of model assessment, where, by studying the continuous diffusive model, we demonstrate that the simple two-state model is insufficient to capture the fine details of the DNA hairpin's conformational dynamics. Section 6 concludes the paper with further discussion.

2. Bayesian analysis of the two-state model

Let $Y(t)$ be the total number of photon arrivals up to time t. Within an infinitesimal time interval $(t, t+\mathrm{d}t)$ the probability of observing a photon is proportional to $\gamma^{-1}(t)\,\mathrm{d}t$. Denoting $\triangle Y_t = Y(t+\mathrm{d}t) - Y(t)$, we have

$$P(\triangle Y_t = 1|\gamma_t) = A_0(t)\,\gamma^{-1}(t)\,\mathrm{d}t, \tag{2.1}$$

$$[\tau|\triangle Y_t = 1, \gamma_t] \sim \gamma(t)\exp\{-\gamma(t)\tau\} \tag{2.2}$$

where $A_0(t)$ is the photon arrival intensity at time t, which could vary from time to time. The stochastic nature of $A_0(t)$ will be addressed in Section 3. In the present section, to focus on the basic ideas, we treat $A_0(t)$ as a constant over time $A_0(t) \equiv A_0$. Following the convention in the stochastic processes literature, we use both parentheses and subscripts to index (the same) stochastic processes; for example $\gamma(t)$ and γ_t both denote the underlying two-state Markov chain, and $Y(t)$ and Y_t both refer to the photon counting process. It is important to point out that formulation (2.1) is equivalent to the photon arrivals' following the Cox process, but working with equation (2.1) allows an easy generalization as we shall demonstrate.

2.1. Likelihood calculation

Let $0 = t_0 < t_1 < \ldots < t_n$ be the observed photon arrival times and let τ_i be the corresponding photon delay times. The pairs $\{(t_i, \tau_i)\}_{i=0}^n$ are collected through the fluorescence lifetime experiments. Note that the likelihood consists of contributions from three parts:

(a) the photon arrival times t_i;
(b) the delay times τ_i;
(c) the fact that no photon arrives in time interval (t_i, t_{i+1}).

Note also that the Cox process can be viewed as a hidden Markov process. But it is distinct from the standard hidden Markov model (Elliott *et al.*, 1999) in that interval lengths between adjacent photon arrivals are also informative about the hidden process $\gamma(t)$. Given $\boldsymbol{\gamma} = (\gamma(t_0), \ldots, \gamma(t_n))$, the joint likelihood function can be written as

$$L(\mathbf{t}, \boldsymbol{\tau}, \boldsymbol{\gamma}|\boldsymbol{\theta}) = P\{\gamma(t_0)\}\left[\prod_{i=0}^{n}\left\{\frac{A_0}{\gamma(t_i)}\mathrm{d}t\right\}\right]\left[\prod_{i=0}^{n}\gamma(t_i)\exp\{-\gamma(t_i)\tau_i\}\right]$$
$$\times \prod_{i=0}^{n-1} P\{Y(t_{i+1}^-) - Y(t_i) = 0, \gamma(t_{i+1})|\gamma(t_i)\}.$$

Using an infinitesimal matrix approach, in Appendix A we show that

$$P\{Y(t_{i+1}^-) - Y(t_i) = 0,\, \gamma(t_{i+1})|\gamma(t_i)\} = [\exp\{(Q-H)(t_{i+1}-t_i)\}]_{(I(t_i), I(t_{i+1}))},$$

where Q is as in Section 1.2,

$$H = \begin{pmatrix} A_0/a & 0 \\ 0 & A_0/b \end{pmatrix}$$

and the index $I(t)$ is equal to 1 if $\gamma(t) = a$ and 2 otherwise. Hence, the unconditional likelihood can be computed by assuming that $\gamma(t_0)$ starts from the stationary distribution and summing out the $\gamma(t_i)$ by recursive matrix multiplications:

$$L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}) = (\pi_1, \pi_2) D_0 H \left[\prod_{i=0}^{n-1}\exp\{(Q-H)\Delta t_i\} D_{i+1} H\right]\begin{pmatrix}1\\1\end{pmatrix} \tag{2.3}$$

where (π_1, π_2) is the equilibrium distribution of $\gamma(t)$, $\Delta t_i = t_{i+1} - t_i$ and

$$D_i = \begin{pmatrix} a\exp(-a\tau_i) & 0 \\ 0 & b\exp(-b\tau_i) \end{pmatrix}.$$

Remark 1. We note that the likelihood function can also be obtained through the first-step analysis by starting from the infinitesimal generator and setting up a set of differential equations (Karlin and Taylor, 1998; Karr, 1986). Compared with the first-step analysis, the matrix approach that is detailed in Appendix A is more intuitive and easier to generalize to accommodate other complications such as the two-by-two model (Schenter *et al.*, 1999) and the molecular Brownian diffusions (see Section 3). For example, in the two-by-two model one assumes that there are two substates within each of the two states:

$$\begin{array}{ccc} A_1 & \underset{k_{21}}{\overset{k_{12}}{\rightleftarrows}} & B_1 \\ \alpha\downarrow\uparrow\beta & & \alpha\downarrow\uparrow\beta \\ A_2 & \underset{k'_{21}}{\overset{k'_{12}}{\rightleftarrows}} & B_2 \end{array} \tag{2.4}$$

where α and β are the transition rates between the substates. The $\gamma(t)$ process takes value a_1, a_2, b_1 and b_2 in the states A_1, A_2, B_1 and B_2 respectively. The same likelihood formula (2.3) holds with Q, H and D_i being replaced by the corresponding matrices in the new model:

$$Q = \begin{pmatrix} -(k_{12}+\alpha) & \alpha & k_{12} & 0 \\ \beta & -(k'_{12}+\beta) & 0 & k'_{12} \\ k_{21} & 0 & -(k_{21}+\alpha) & \alpha \\ 0 & k'_{21} & \beta & -(k'_{21}+\beta) \end{pmatrix} \begin{matrix} \leftarrow A_1 \\ \leftarrow A_2 \\ \leftarrow B_1 \\ \leftarrow B_2 \end{matrix}$$

$$H = \operatorname{diag}\left(\frac{A_0}{a_1}, \frac{A_0}{a_2}, \frac{A_0}{b_1}, \frac{A_0}{b_2}\right),$$

$$D_i = \operatorname{diag}\{a_1\exp(-a_1\tau_i), a_2\exp(-a_2\tau_i), b_1\exp(-b_1\tau_i), b_2\exp(-b_2\tau_i)\}.$$

We defer further discussion on the two-by-two model to Section 4, where models beyond the two-state model are considered. It is also worthwhile to note that Wolpert and Ickstadt (1998) have generalized the Cox process to doubly stochastic hierarchical models to account for uncertainty and spatial variations of the underlying rate in point process models.

2.2. A Markov chain Monte Carlo algorithm

The two-state model has five parameters: a, b, π_1, k and A_0. Let $\eta(\boldsymbol{\theta})$ denote the prior distribution on the parameters $\boldsymbol{\theta} = (a, b, \pi_1, k, A_0)$. Then the posterior distribution is

$$P(\boldsymbol{\theta}|\mathbf{t}, \boldsymbol{\tau}) \propto \eta(\boldsymbol{\theta})\, L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}),$$

where the following constraints hold:

(a) $a > b > 0$,
(b) $0 \leqslant \pi_1 \leqslant 1$,
(c) $k > 0$,
(d) $A_0 > 0$.

Since direct computation from $P(\boldsymbol{\theta}|\mathbf{t}, \boldsymbol{\tau})$ is infeasible, we design a Metropolis-type algorithm to sample from it. This algorithm iterates the following steps.

Step 1: given a and b,

(i) draw x from $\Gamma(1/c_1, ac_1)$ and y from $\Gamma(1/c_2, bc_2)$, where c_1 and c_2 are two tuning parameters, and

(ii) let $a' = \max(x, y)$ and $b' = \min(x, y)$.

The tuning parameters c_1 and c_2 control the step size of the proposal. Taking a' and b' to be the maximum and minimum respectively of x and y satisfies constraint (a) and consequently makes the proposal density a mixture of two product gamma densities.

Step 2: given π_1, draw π_1' from the beta distribution $B\{c_3\pi_1, c_3(1-\pi_1)\}$. Since the mean and variance of $B\{c_3\pi_1, c_3(1-\pi_1)\}$ are π_1 and $\pi_1(1-\pi_1)/(c_3+1)$ respectively, π_1' can be viewed as a local perturbation of π_1, and the tuning parameter c_3 controls the localness of the perturbation.

Step 3: given k, draw k' from $\Gamma(1/c_4, kc_4)$. The mean and variance of k'/k are 1 and c_4 respectively, letting c_4 finely tune the perturbation.

Step 4: given A_0, draw A_0' from $\Gamma(1/c_5, A_0c_5)$, whose mean is A_0, and whose variance is controlled by c_5

This algorithm works quite well in our simulation study in which 815 data pairs $\{(t_i, \tau_i)\}_{i=0}^{815}$ were generated from the hidden γ process (815 is a data size that is typical in real experiments). Fig. 4 summarizes the posterior sampling distributions (with a flat prior) for the parameters, where the vertical bars represent the true values. The algorithm, besides correctly identifying all the parameters, runs quite fast: it took less than 2 min on a Pentium 4 personal computer to draw

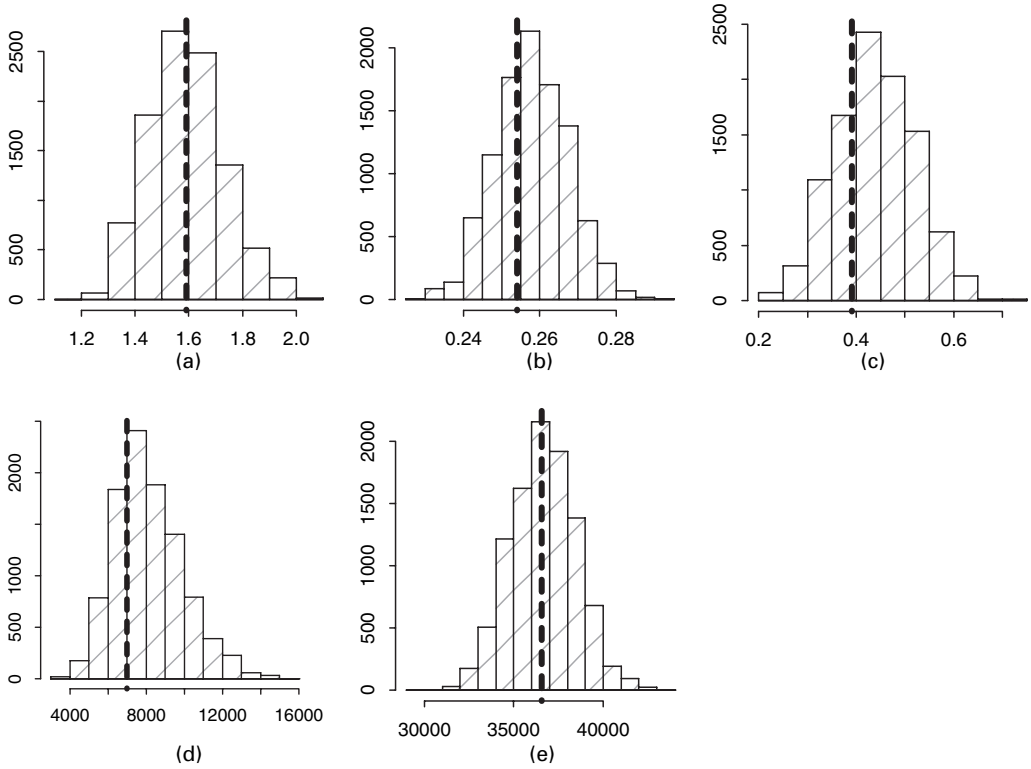

Fig. 4. Histograms of the posterior samples for the simulated two-state data (┊, true value): (a) a; (b) b; (c) π_1; (d) k; (e) A_0

the 10000 samples in Fig. 4. The fast computation is primarily because the direct evaluation of the likelihood function (2.3) requires only $O(n)$ operations.

3. Brownian diffusion in the two-state model

3.1. Statistical formulation

In the single-molecule lifetime experiment, as the laser-excited dye emits photons, the DNA hairpin molecule also diffuses in the focal volume. As a result, since the laser illuminating intensity that the DNA molecule receives is highest at the centre of the focal volume and decreases from the centre outwards, the actual photon arrival intensity $A_0(t)$ varies. Mathematically, we can write $A_0(t) = A_0\,\alpha(t)$ with

$$\alpha(t) = \exp\left\{-\frac{B_x^2(t) + B_y^2(t)}{2w_{xy}^2} - \frac{B_z^2(t)}{2w_z^2}\right\}, \tag{3.1}$$

where $(B_x(t), B_y(t)$ and $B_z(t))$ is the location of the DNA molecule, and the constants w_{xy} and w_z specify the x-, y- and z-axes of the ellipsoidal focal volume. Let σ denote the diffusion constant, i.e. $B_x(t) = \sigma\, W_1(t)$, $B_y(t) = \sigma\, W_2(t)$ and $B_z(t) = \sigma\, W_3(t)$, where $(W_1(t), W_2(t), W_3(t))$ are three independent standard Brownian motions. In the lifetime experiments, the constants w_{xy}, w_z and σ are calculated by fluorescence correlation spectroscopy theory (Magde *et al.*, 1974) and are assumed known. With the presence of diffusion, the probability formulations (2.1) and (2.2) are changed to

$$P(\triangle Y_t = 1 | \gamma_t, \alpha_t) = A_0\,\alpha(t)\,\gamma^{-1}(t)\,\mathrm{d}t, \tag{3.2}$$

$$[\tau | \triangle Y_t = 1, \gamma_t, \alpha_t] \sim \gamma(t) \exp\{-\gamma(t)\tau\}. \tag{3.3}$$

Owing to the extra conditionality on the intensity $\alpha(t)$, the likelihood changes as well. Because in the experiments $\alpha(t)$ is a slow-varying process relative to the photon arrivals, we made an approximation that $\alpha(t) \approx \alpha(t_i)$ for $t \in (t_i, t_{i+1})$. Then, using the same discretization technique as in Appendix A, we have

$$P\{Y(t_1^-) - Y(t_0) = 0, \gamma(t_1) | \gamma(t_0), \alpha(t_0)\} = (\exp[\{Q - \alpha(t_0)H\}(t_1 - t_0)])_{(I_{t_0}, I_{t_1})},$$

where the matrices Q and H are the same as before. Like in Section 2, combining the likelihood contributions from intervals $\{(t_i, t_{i+1}); i = 0, 1, \ldots, n-1\}$ with those from the pairs $\{t_i, \tau_i\}$ yields the conditional likelihood given α_t, a modification of equation (2.3):

$$L(\mathbf{t}, \boldsymbol{\tau} | \boldsymbol{\theta}, \alpha_t) = (\pi_1, \pi_2) D_0 H_0 \left[\prod_{i=0}^{n-1} \exp\{(Q - H_i)\Delta t_i\} D_{i+1} H_{i+1}\right] \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{3.4}$$

where H_i denotes $\alpha(t_i)H = \alpha(t_i)\,\mathrm{diag}(A_0/a, A_0/b)$. With a prior distribution $\eta(\boldsymbol{\theta})$ on the parameters, the posterior distribution of $\boldsymbol{\theta}$ given the observations $\{(t_i, \tau_i)\}_{i=0}^n$ is

$$P(\boldsymbol{\theta} | \mathbf{t}, \boldsymbol{\tau}) \propto \int \eta(\boldsymbol{\theta})\, L(\mathbf{t}, \boldsymbol{\tau} | \boldsymbol{\theta}, \alpha_t)\, P(\alpha_t)\, \mathrm{d}(\alpha_t), \tag{3.5}$$

where $P(a_t)$ denotes the probability law of $\alpha(t)$. Comparing expression (3.5) with the corresponding expression in Section 2, we note that the Brownian diffusion factor $\alpha(t)$ of the molecule is not observable and must be integrated out. This path integral cannot be calculated analytically and must be dealt with by the data augmentation approach (Tanner and Wong, 1987).

3.2. Data augmentation

Intuitively, if the Brownian diffusion (B_x, B_y, B_z) is given, the computation that was described in Section 2 can be easily carried out. Thus, by sampling from the joint posterior distribution of $\boldsymbol{\theta}$ and (B_x, B_y, B_z),

$$P(\boldsymbol{\theta}, B_x, B_y, B_z|\mathbf{t}, \boldsymbol{\tau}) \propto \eta(\boldsymbol{\theta})\, L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t)\, P(B_x)\, P(B_y)\, P(B_z), \tag{3.6}$$

where the dependence of α_t on (B_x, B_y, B_z) is given by equation (3.1), we effectively marginalized out the hidden diffusion process and obtain the correct inference. We only need to augment the three-dimensional Brownian motion at the photon arrival times and, correspondingly, each term of expression (3.6) has a simple form. For example, the transition density for B_x is

$$P(B_x) = (2\pi)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=0}^{n-1} \frac{\{B_x(t_{i+1}) - B_x(t_i)\}^2}{\Delta t_i} \right]. \tag{3.7}$$

Starting from an initial $\boldsymbol{\theta}$ and a configuration of (B_x, B_y, B_z), we iterate the following conditional sampling steps to simulate from expression (3.6).

Step 1: draw $\boldsymbol{\theta}$ conditioning on the current diffusion (B_x, B_y, B_z),

$$[\boldsymbol{\theta}|B_x, B_y, B_z, \mathbf{t}, \boldsymbol{\tau}] \sim P(\boldsymbol{\theta}|B_x, B_y, B_z, \mathbf{t}, \boldsymbol{\tau}) \propto \eta(\boldsymbol{\theta})\, L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t). \tag{3.8}$$

Step 2: draw the diffusion (B_x, B_y, B_z) conditioning on the current value of $\boldsymbol{\theta}$,

$$[B_x, B_y, B_z|\,\boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\tau}] \sim P(B_x, B_y, B_z|\boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\tau}) \propto L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t)\, P(B_x)\, P(B_y)\, P(B_z). \tag{3.9}$$

The first step is achieved by the Metropolis–Hastings (MH) algorithm as outlined in Section 2. Since the computation that is involved in the MH algorithm is much simpler than the sampling in step 2, especially when the chain is long, we conduct the MH iteration several times in step 1 before moving on to step 2.

Step 2 is achieved by updating the diffusion chain time point by time point. In other words, for each of $i = 0, 1, \ldots, n$, we propose a new location $\tilde{\mathbf{B}} = (\tilde{x}, \tilde{y}, \tilde{z})$ for the ith time point $\mathbf{B}(t_i) = (B_x(t_i), B_y(t_i), B_z(t_i))$ of the diffusion chain (with the others fixed) and accept the proposal according to the MH rule. Since only one time point is considered each time, the MH ratio is easy to compute. For example, if the ith time point $\mathbf{B}(t_i) = (B_x(t_i), B_y(t_i), B_z(t_i))$ is proposed to change to $\tilde{\mathbf{B}} = (\tilde{x}, \tilde{y}, \tilde{z})$, the MH probability is simply

$$\min\left[1, \frac{L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \tilde{\alpha}_t)}{L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t)} \frac{P(\tilde{B}_x)\, P(\tilde{B}_y)\, P(\tilde{B}_z)}{P(B_x)\, P(B_y)\, P(B_z)} \frac{T\{\tilde{\mathbf{B}} \to \mathbf{B}(t_i)\}}{T\{\mathbf{B}(t_i) \to \tilde{\mathbf{B}}\}}\right],$$

where $T\{\cdot \to \cdot\}$ is the transition density of the proposal and

$$\frac{P(\tilde{B}_x)\, P(\tilde{B}_y)\, P(\tilde{B}_z)}{P(B_x)\, P(B_y)\, P(B_z)}$$

follows from equation (3.7). To evaluate the likelihood function $L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t)$ efficiently, we implemented the following *forward–backward* recursive scheme: backward compute the partial sums (integrals) in the likelihood; forward update the diffusion chain one component (time point) at a time. This approach ensures that the computation cost is only $O(n)$ operations, whereas the naïve approach is of order $O(n^2)$.

To begin the forward–backward algorithm, we first compute *backwards* a sequence of matrices K_i by the recursion

$$\begin{cases} K_{n+1} = I, \quad K_n = D_n H_n, \\ K_i = D_i H_i \exp\{(Q - H_i)\Delta t_i\} K_{i+1}, & i < n, \end{cases}$$

where, as before, $D_i = \mathrm{diag}\{a\exp(-a\tau_i), b\exp(-b\tau_i)\}$, $H_i = \alpha(t_i)\,\mathrm{diag}(A_0/a, A_0/b)$ and $\Delta t_i = t_{i+1} - t_i$. Denote $v_0 = (\pi_1, \pi_2)$. Then the *forward* sampling algorithm can be described as follows, for $i = 0, 1, \ldots, n$.

Step 1: propose a change $\tilde{\mathbf{B}} = (\tilde{x}, \tilde{y}, \tilde{z})$ for the ith time point $\mathbf{B}(t_i)$.
Step 2: compute

$$R = \begin{cases} D_i H_i \exp\{(Q - H_i)\Delta t_i\} & \text{if } i < n, \\ D_n H_n & \text{if } i = n, \end{cases}$$

and

$$S = \begin{cases} D_i \tilde{H}_i \exp\{(Q - \tilde{H}_i)\Delta t_i\} & \text{if } i < n, \\ D_n \tilde{H}_n & \text{if } i = n, \end{cases}$$

where $\tilde{H}_i = \tilde{\alpha}(t_i)\,\mathrm{diag}(A_0/a, A_0/b)$ and

$$\tilde{\alpha}(t_i) = \exp\left(-\frac{\tilde{x}^2 + \tilde{y}^2}{2w_{xy}^2} - \frac{\tilde{z}^2}{2w_z^2}\right)$$

is the value of the illuminating intensity of the ith time point under the proposal $\tilde{\mathbf{B}}$. Therefore, $v_i R K_{i+1}\binom{1}{1}$ is now the likelihood $L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t)$ evaluated at the original diffusion chain, and $v_i S K_{i+1}\binom{1}{1}$ is the likelihood $L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \tilde{\alpha}_t)$ evaluated at the diffusion chain with newly proposed ith time point. The vector v_i is used to record the likelihood contribution from the first i time points; it will be updated at step 4.
Step 3: compute the MH ratio

$$r = \frac{L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \tilde{\alpha}_t)\, P(\tilde{B}_x)\, P(\tilde{B}_y)\, P(\tilde{B}_z)\, T\{\tilde{\mathbf{B}} \to \mathbf{B}(t_i)\}}{L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t)\, P(B_x)\, P(B_y)\, P(B_z)\, T\{\mathbf{B}(t_i) \to \tilde{\mathbf{B}}\}}.$$

Step 4: generate $u \sim U(0, 1)$. If $u < \min(1, r)$, we update $\mathbf{B}(t_i)$ to $\tilde{\mathbf{B}}$ and let $v_{i+1} = v_i S$; otherwise, we keep $\mathbf{B}(t_i)$ unchanged and let $v_{i+1} = v_i R$. The vector v_{i+1} now records the likelihood contribution from the first $i + 1$ time points.

Note that the MH approach is necessary since direct (conditional) sampling of $\mathbf{B}(t_i)$ is difficult. Although it is possible to conduct a 'blocking' operation on $\mathbf{B}(t)$, i.e. updating a segment of the process $(\mathbf{B}(t_i), \mathbf{B}(t_{i+1}), \ldots, \mathbf{B}(t_{i+b-1}))$ simultaneously, it is not clear whether this will substantially improve the convergence since a multidimensional MH algorithm is tricky to conduct. Besides, we found that the foregoing componentwise updating scheme takes advantage of an efficient likelihood evaluation scheme and is already quite efficient. In a later section, we shall show that the blockwise move is very useful for handling a more complicated model.

3.3. Other experimental complications

Owing to technological limitations, the actual experiments have more complications in addition to the molecular Brownian diffusion. The first additional complication comes from the background photons. So far in the analysis we have assumed that the photons that are recorded all come from the laser-excited DNA hairpin structure. But in the real experiment the laser also excites photons from the background, whose arrival is characterized as a time homogeneous Poisson process with rate ρ. The parameter ρ can be estimated by running a long controlled experiment with no molecule put in the focal volume (the photons detected thus all come from

the background) and is assumed known to us. Thus, we only need to change the overall photon arrival process (the Cox process) from equation (3.2) to

$$P(\triangle Y_t = 1|\gamma_t, \alpha_t) = \left\{\rho + \frac{A_0}{\gamma(t)}\alpha_t\right\} \mathrm{d}t. \tag{3.10}$$

The likelihood function correspondingly changes from equation (3.4) to

$$L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t) = (\pi_1, \pi_2) D_0 G_0 \left[\prod_{i=0}^{n-1} \exp\{(Q - G_i)\Delta t_i\} D_{i+1} G_{i+1}\right] \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{3.11}$$

where the matrix $D_i = \mathrm{diag}\{a\exp(-a\tau_i), b\exp(-b\tau_i)\}$ is as defined before and the new matrix

$$G_i = \mathrm{diag}\left\{\rho + \frac{A_0}{a}\alpha(t_i), \rho + \frac{A_0}{b}\alpha(t_i)\right\}. \tag{3.12}$$

The second experimental complication is the issue of time wrapping related to the determination of the photon delay time. As seen in Section 1.2, the delay times are determined by comparing the detected photon arrival times with the regular laser pulse train. However, a careful inspection of Fig. 2 suggests that, if the delay time exceeds the interval, 13.2 ns, between two adjacent laser pulses, we are no longer sure which laser pulse excites the donor dye. The machine reading assumes that the dye is excited by the pulse immediately preceding the arrival of the photon detected, but this creates the time wrapping effect. For example, if the actual delay time is 17 ns, the machine will determine it as $17 - 13.2 = 3.8$ ns. The third complication arises from negative readings: some photon delay times τ are recorded as negative. The negative τ indicates that a photon has arrived, but the delay time is missing. These two complications can be handled relatively easily by further modifying the likelihood function. Thanks to the memoryless property of the exponential distribution, the time wrapping changes the distribution of the delay time τ to a truncated exponential distribution, which alters expression (3.3) to

$$[\tau|\Delta Y_t = 1, \gamma_t, \alpha_t] \sim \frac{\gamma(t)\exp\{-\gamma(t)\tau\}}{1 - \exp\{-\gamma(t)M\}}, \tag{3.13}$$

where M is the point of wrapping (13.2 ns in the above example). The negative recording further alters distribution (3.13) to

$$[\tau|\Delta Y_t = 1, \gamma_t, \alpha_t, \tau > 0] \sim \frac{\gamma(t)\exp\{-\gamma(t)\tau\}}{1 - \exp\{-\gamma(t)M\}}.$$

Consequently replacing the D_i-matrix by

$$\tilde{D}_i = \begin{cases} \mathrm{diag}\left\{\dfrac{a\exp(-a\tau_i)}{1-\exp(-aM)}, \dfrac{b\exp(-b\tau_i)}{1-\exp(-bM)}\right\} & \text{if } \tau_i \geqslant 0, \\ I & \text{if } \tau_i < 0 \end{cases} \tag{3.14}$$

in formula (3.11) gives the likelihood for the real experimental data.

4. Beyond two-state models: the continuous diffusive model

4.1. The model

The two-state model (1.1) depicts the conformational dynamics as a two-state continuous time Markov chain. However, scientists have observed that for some proteins and enzymes the two-state model is not sufficiently accurate to describe the conformational details. For example, the two-state model implies a 'memoryless' property between consecutive waiting times at different

states, and this property has been seen to be violated in many single-molecule experiments involving proteins and enzymes (Schenter *et al.*, 1999; Lu *et al.*, 1998). This phenomenon, known as 'dynamic disorder' (Xie, 2002; Zwanzig, 1990), motivates models beyond the two-state model. The two-by-two model (2.4) that was briefly described in Section 2 is such an attempt. In this model the states A_1 and A_2 are experimentally indistinguishable and so are B_1 and B_2. What can be observed from the experiments are, hence, the aggregate effects of A_1 and A_2 (and respectively B_1 and B_2). In the statistics literature, this type of process is termed the aggregated Markov processes (Fredkin and Rice, 1986). The consequence of aggregation is that the times that a molecule spends in the collective states $\{A_1, A_2\}$ and $\{B_1, B_2\}$ are no longer independent, which could explain the dynamic disorder (Cao, 2000; Yang and Cao, 2001, 2002). As shown in Section 2, the likelihood function for the two-by-two model shares a simple form as the two-state model. Therefore the Bayesian data augmentation approach that is outlined in Sections 2 and 3 is readily available for the two-by-two model. The inference procedure is essentially identical, except that there are more parameters in the two-by-two model.

The two-by-two model can be further generalized to a continuous diffusive model, which cannot be directly handled by the data augmentation approach that was developed earlier and is the focus of the current section. In this model, instead of using the discrete aggregation to explain the dynamic disorder, a continuous stochastic control process $x(t)$ is introduced which 'controls' the transition rates as follows:

$$A \underset{k_{21}\exp\{-x(t)\}}{\overset{k_{12}\exp\{-x(t)\}}{\rightleftarrows}} B. \tag{4.1}$$

Physically, this can be seen as a result of a dynamically fluctuating energy barrier between the two states ($x(t)$ is related to the energy barrier $h(t)$ through $x(t) = h(t)/k_B T$, where k_B is the Boltzmann constant and T is the temperature). This fluctuating energy barrier (see Fig. 5 for an illustration) allows the transition rates to be time varying and stochastic. Correspondingly, the consecutive waiting times at states A and B are not independent.

As first suggested by Agmon and Hopfield (1983), an Ornstein–Uhlenbeck process is used to capture the fluctuation of the energy barrier between the two states:

$$\mathrm{d}x_t = -\lambda x_t \,\mathrm{d}t + \sqrt{(2\xi\lambda)}\,\mathrm{d}W_t. \tag{4.2}$$

A way to picture model (4.1) is to think of the DNA molecule as having infinitely many indistinguishable states $\{A_1, A_2, \dots\}$ and $\{B_1, B_2, \dots\}$, a continuous generalization of the two-by-two model. Although many debates have been initiated (Bonnet *et al.*, 1998; Ying *et al.*, 2001; Grunwell *et al.*, 2001; Ansari *et al.*, 2001, 2002), so far there is no clear evidence about whether the continuous diffusive model is definitively more appropriate than the simple two-state model

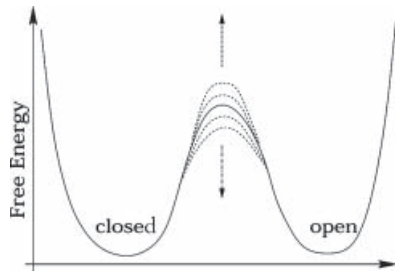

Fig. 5. Fluctuating barrier between the two states: the Ornstein–Uhlenbeck process models the dynamic oscillation of the energy barrier between the open and closed states of the DNA hairpin

for the DNA hairpin. It is thus of interest to synthesize all the available statistical evidence most efficiently for discerning between models.

4.2. Likelihood and data augmentation under the continuous diffusive model

Employing the observation that the control process $x(t)$ is stable between two successive photon arrivals, we impose the approximation that $x(t) \approx x(t_i)$ for $t \in (t_i, t_{i+1})$ and apply the technique in Section 2 to obtain the closed form conditional likelihood giving both $\alpha(t)$ and $x(t)$:

$$L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t, x_t) = (\pi_1, \pi_2)\tilde{D}_0 G_0 \left[\prod_{i=0}^{n-1} \exp\{(Q_i - G_i)\Delta t_i\}\tilde{D}_{i+1}G_{i+1}\right]\begin{pmatrix}1\\1\end{pmatrix}, \tag{4.3}$$

where G_i and $\tilde{D}_i$ are as defined before in equations (3.12) and (3.14), and

$$Q_i \triangleq \begin{pmatrix} -k_{12}\exp\{-x(t_i)\} & k_{12}\exp\{-x(t_i)\} \\ k_{21}\exp\{-x(t_i)\} & -k_{21}\exp\{-x(t_i)\}\end{pmatrix}.$$

The posterior distribution of the parameters $(\boldsymbol{\theta}, \lambda, \xi)$ given the observations $\{(t_i, \tau_i)\}$ is

$$P(\boldsymbol{\theta}, \lambda, \xi|\mathbf{t}, \boldsymbol{\tau}) \propto \int\int \eta(\boldsymbol{\theta}, \lambda, \xi) L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t, x_t) P(\alpha_t) P(x_t|\lambda, \xi)\, \mathrm{d}(\alpha_t)\, \mathrm{d}(x_t), \tag{4.4}$$

where $P(x_t|\lambda, \xi)$ is the transition density for the process $x(t)$ and η denotes the prior distribution on the parameters.

By augmenting both $x(t)$ and the Brownian diffusion $\mathbf{B}(t)$, we consider the joint distribution of $(\boldsymbol{\theta}, \lambda, \xi)$ and $(\mathbf{B}(t), x(t))$:

$$P(\boldsymbol{\theta}, \lambda, \xi, \mathbf{B}, x|\mathbf{t}, \boldsymbol{\tau}) \propto \eta(\boldsymbol{\theta}, \lambda, \xi) L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t, x_t) P\{\mathbf{B}(t)\} P(x_t|\lambda, \xi). \tag{4.5}$$

From the Kolmogorov backward equation

$$\frac{\partial P}{\partial t} = \xi\lambda\frac{\partial^2 P}{\partial x^2} - \lambda x\frac{\partial P}{\partial x},$$

we have $x(t_{i+1})|x(t_i) \sim N[x(t_i)\exp(-\lambda\Delta t_i), \xi\{1-\exp(-2\lambda\Delta t_i)\}]$ (see Karlin and Taylor (1981), page 218). Assuming that the Ornstein–Uhlenbeck process $x(t)$ starts from the stationary distribution, we have the joint density

$$\begin{aligned} P(x_t|\lambda, \xi) \propto{}& \xi^{-(n+1)/2}\left[\prod_{i=0}^{n-1}\{1-\exp(-2\lambda\Delta t_i)\}^{-1/2}\right] \\ &\times \exp\left[-\frac{x(t_0)^2}{2\xi} - \sum_{i=0}^{n-1}\frac{\{x(t_{i+1}) - x(t_i)\exp(-\lambda\Delta t_i)\}^2}{2\xi\{1-\exp(-2\lambda\Delta t_i)\}}\right]. \end{aligned} \tag{4.6}$$

To make the Markov chain Monte Carlo sampling more efficient, we reparameterize the control process $x(t)$ by replacing ξ with

$$\phi = \sqrt{(\xi\lambda)}.$$

The Ornstein–Uhlenbeck process is then changed to

$$\mathrm{d}x_t = -\lambda x_t\, \mathrm{d}t + \phi\sqrt{2}\, \mathrm{d}W_t. \tag{4.7}$$

The joint distribution (4.5) is then modified to

$$P(\boldsymbol{\theta}, \lambda, \phi, \mathbf{B}, x_t|\mathbf{t}, \boldsymbol{\tau}) \propto \eta'(\boldsymbol{\theta}, \lambda, \phi) L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t, x_t) P\{\mathbf{B}(t)\} P(x_t|\lambda, \phi^2/\lambda), \tag{4.8}$$

where η' is the prior distribution on $(\boldsymbol{\theta}, \lambda, \phi)$.

To obtain Monte Carlo samples from the joint posterior distribution, we iterate the following conditional sampling steps, starting from an initial configuration:

$$\begin{aligned}
\boldsymbol{\theta} &\sim [\boldsymbol{\theta}|\lambda, \phi, \mathbf{B}, x_t, \mathbf{t}, \boldsymbol{\tau}] \propto \eta'(\boldsymbol{\theta}, \lambda, \phi)\, L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t, x_t), \\
(\lambda, \phi) &\sim [\lambda, \phi|\boldsymbol{\theta}, \mathbf{B}, x_t, \mathbf{t}, \boldsymbol{\tau}] \propto \eta'(\boldsymbol{\theta}, \lambda, \phi)\, P(x_t|\lambda, \phi^2/\lambda), \\
\mathbf{B} &\sim [\mathbf{B}|\boldsymbol{\theta}, \lambda, \phi, x_t, \mathbf{t}, \boldsymbol{\tau}] \propto L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t, x_t)\, P(\mathbf{B}), \\
x(t) &\sim [x_t|\boldsymbol{\theta}, \lambda, \phi, \mathbf{B}, \mathbf{t}, \boldsymbol{\tau}] \propto L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t, x_t)\, P(x_t|\lambda, \phi^2/\lambda).
\end{aligned}$$

The sampling of $\boldsymbol{\theta}$ and (λ, ϕ) is achieved by the MH algorithm. The conditional sampling of $\mathbf{B}(t)$ and x_t cannot be achieved in closed form. We employ the same backward–forward algorithm as described in Section 3 to update them recursively componentwise.

4.3. Scale transformation update

To improve the computation efficiency further, we introduce an update that changes (λ, ϕ) and x_t simultaneously. This move is important because of the high correlation between (λ, ϕ) and x_t—it is seen from equation (4.7) that, conditional on large values of x_t, posterior draws of (λ, ϕ) tend to be large, whereas a large pair of (λ, ϕ) in turn tend to give rise to large values of x_t. Thus, if we can move (λ, ϕ) and x_t together, the sampler should be able to converge faster.

Given the current configuration of $(\lambda, \phi, \boldsymbol{\theta}, \mathbf{B}(t), x_t)$, the scale transformation proposes a move

$$(\phi, x_t) \to (s\phi, s x_t), \tag{4.9}$$

where s is a scalar. In other words, the proposal attempts to scale ϕ and the control process x_t up or down together. To preserve the joint distribution (4.8), the scalar s must be sampled from the following conditional distribution according to the 'generalized Gibbs sampling' rule in Liu and Sabatti (2000):

$$\begin{aligned}
p(s) &\propto s^{n+1} P(\boldsymbol{\theta}, \lambda, s\phi, B_x, B_y, B_z, s x_t|\mathbf{t}, \boldsymbol{\tau}) \\
&\propto s^{n+1} \eta'(\boldsymbol{\theta}, \lambda, s\phi)\, L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t, s x_t)\, P\{s x_t|\lambda, (s\phi)^2/\lambda\} \\
&\propto \eta'(\boldsymbol{\theta}, \lambda, s\phi)\, L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t, s x_t),
\end{aligned} \tag{4.10}$$

whose derivation is deferred to Appendix A. Since direct simulation from distribution (4.10) is not yet feasible, we apply the MH algorithm by noting that $s = 1$ means no scaling: first, propose s from the gamma density $\Gamma(s; 1/c, c)$, which has mean 1; then accept the proposed s with the MH-like probability

$$r = \min\left\{1, \frac{\Gamma(s^{-1}; 1/c, c)\, p(s) s}{\Gamma(s; 1/c, c)\, p(1)}\right\}. \tag{4.11}$$

If the proposal is accepted, we update (ϕ, x_t) by the scale move (4.9). Fig. 6 compares the autocorrelation of the samples with and without the scale move. The improvement over the standard approach is quite substantial. From a theoretical point of view, this MH move can be stated in a more general form that could find potential use in other applications. The following result gives a direct MH rule under the setting of general transformation groups. The proof is deferred to Appendix A.

Theorem 1. Let Λ be a locally compact group of transformations on the state space $\mathcal{X}$. Let $\mu(\mathrm{d}\nu)$ be the left Haar measure on Λ and let $\pi(\mathbf{x})$ be the target probability distribution that is of interest defined on $\mathcal{X}$. Suppose that we propose the following group action move: draw ν from $f(\nu)\,\mu(\mathrm{d}\nu)$ and update $\mathbf{x}' = \nu(\mathbf{x})$. Then, to maintain $\pi(\mathbf{x})$ as the stationary distribution

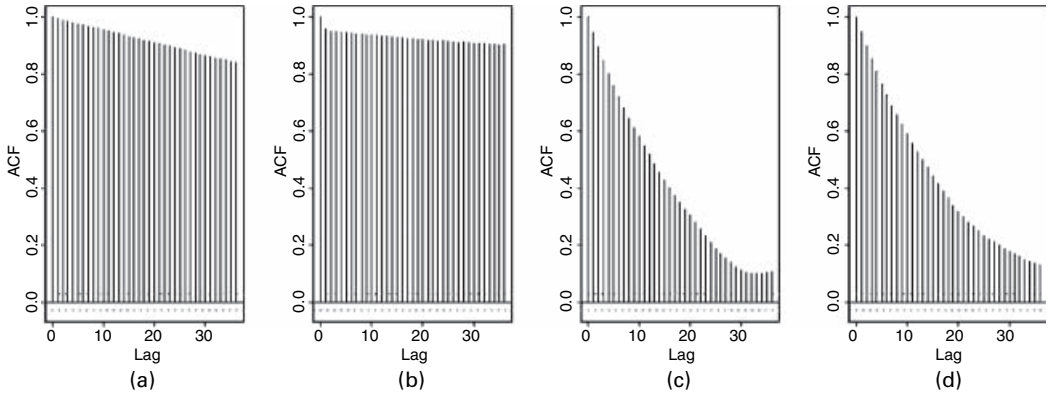

Fig. 6. Autocorrelations of the posterior samples with and without the scale transformation update: (a) λ-series without scale update; (b) ϕ-series without scale update; (c) λ-series with scale update; (d) ϕ-series with scale update

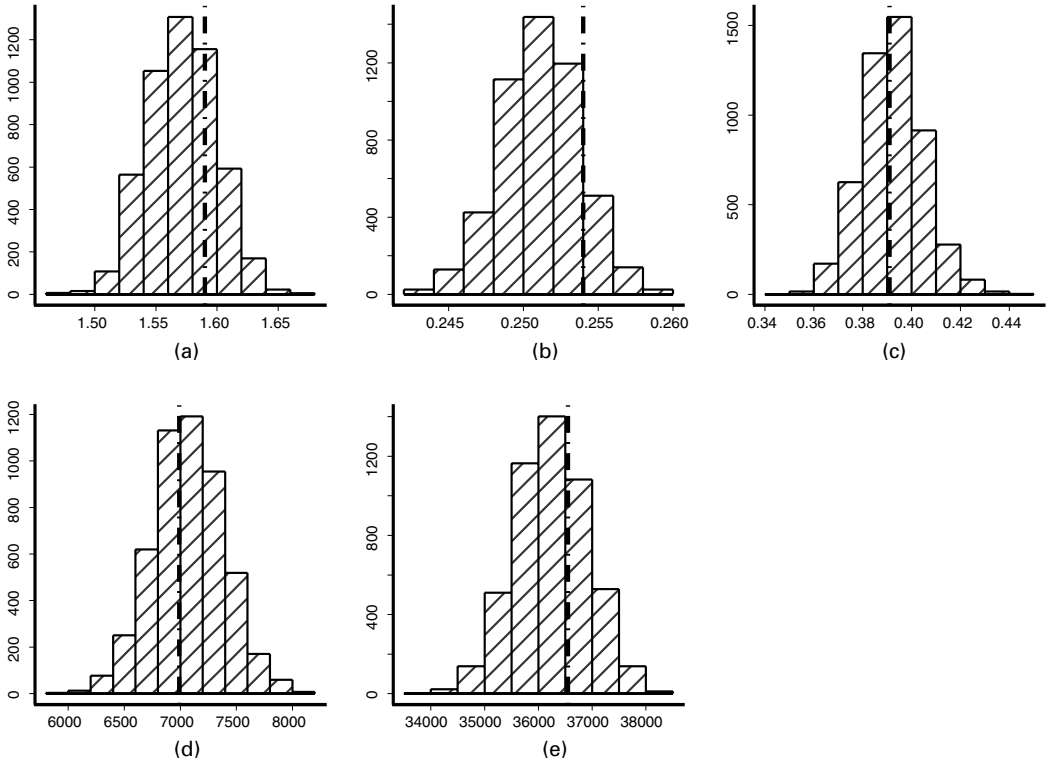

Fig. 7. Histograms of the posterior samples for the simulated data with Brownian diffusion (¦, true value): (a) a; (b) b; (c) π_1; (d) k; (e) A_0

after the move, the proposed move should be accepted with probability

$$r = \min\left\{1, \frac{f(\nu^{-1})\,\pi(\mathbf{x}')|J_\nu(\mathbf{x})|}{f(\nu)\,\pi(\mathbf{x})}\right\},$$

where $J_\nu(\mathbf{x})$ is the Jacobian of the transformation $\mathbf{x}' = \nu(\mathbf{x})$.

5. Empirical studies

5.1. Simulated data sets

As an illustration as well as to assess the efficacy of our approach, a two-state simulation experiment was carried out. First, we generate the Brownian diffusion process (B_x, B_y, B_z), which gives α_t according to equation (3.1); we then generate the two-state Markov process $\gamma(t)$; finally the pairs $\{(t_i, \tau_i)\}_{i=0}^{n}$ are generated according to Cox process (3.2) and exponential distribution (3.3). Using the same parameter $\boldsymbol{\theta}$, this data generation process is repeated 50 times, providing 50 independent data sets (as is typical in real experiments). The length of the data sets ranges from several hundred pairs to a few thousand. To make the simulation data as realistic as possible, the background photons and the time wrapping were both incorporated in the simulation. Our goal is to infer $\boldsymbol{\theta}$ from the generated (t_i, τ_i).

It is easy to apply the data augmentation approach that is described in Section 3.2 to analyse these 50 data sets jointly, simply by multiplying the likelihood functions from each individual experiment. For each data set, a Brownian diffusion chain is augmented. With a flat prior on $\boldsymbol{\theta}$, 5000 posterior samples are drawn from the joint distribution (3.6). Fig. 7 displays the sample posterior distribution of the parameters (the vertical bars are the true values that were used to generate the data). The algorithm is seen to identify all the parameters correctly. Fig. 8 plots the posterior samples pairwise. It is evident that a further reparameterization is not needed. Fig. 9 shows the autocorrelations between successive Monte Carlo samples for the parameters. The fast decay of the autocorrelations suggests speedy convergence of the algorithm. As a further

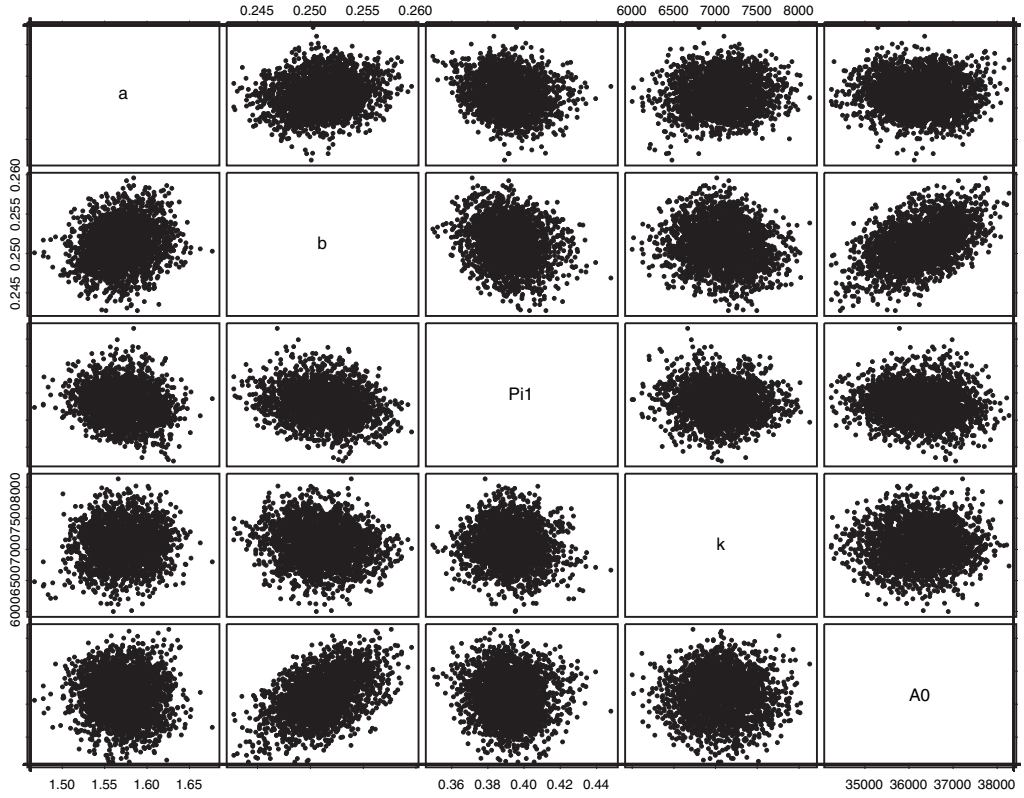

Fig. 8. Pairwise plot of the posterior samples for the simulated data

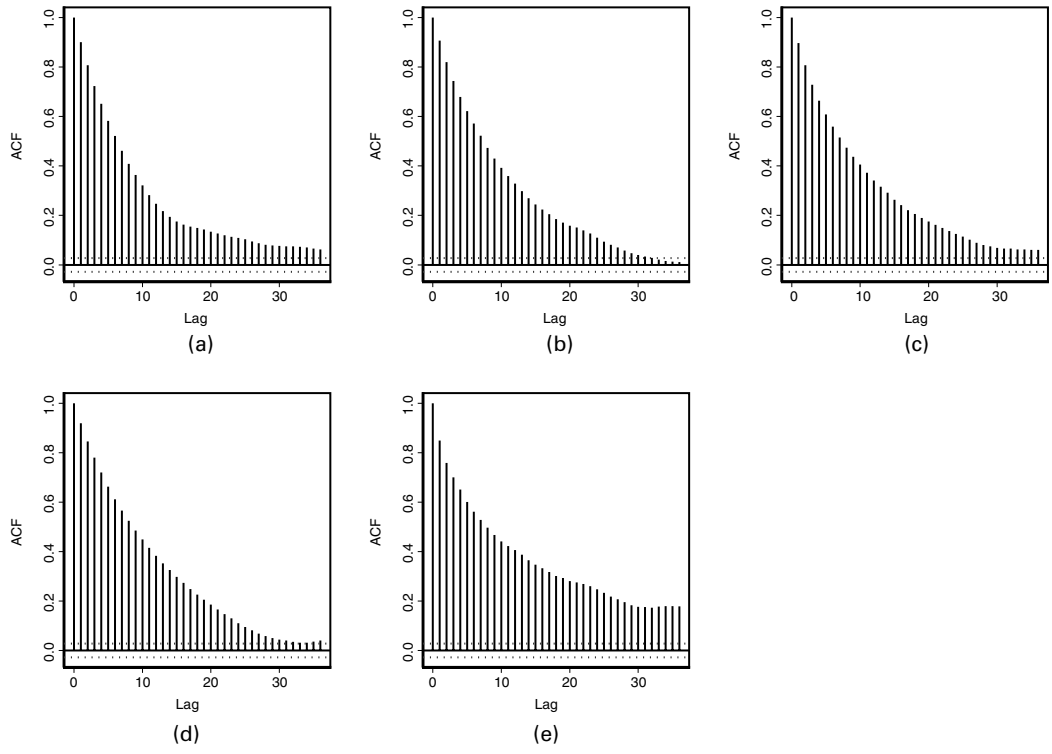

Fig. 9. Autocorrelation plot of the posterior samples for the simulated data: (a) a; (b) b; (c) π_1; (d) k; (e) A_0

test, we started the algorithm from different starting-points; all converge quickly to the posterior modal area.

To see how well the data augmentation approach unveils the hidden information, we compare the true Brownian motion paths with the augmented paths. Fig. 10 provides details for four typical data sets: the light curves (area) are the (multiple) augmented Brownian factor $\alpha(t)$; the true $\alpha(t)$ is shown as the thick curves. The data augmentation technique appears to recover the hidden factor quite well.

5.2. Analysing a real data set with the two-state model

Using the method that was developed in Section 3, we sought to analyse a data set that was obtained by the Xie laboratory at Harvard University. The data set has 1813 observation pairs. We obtained 5000 posterior samples for the five key parameters, whose histograms are shown in Fig. 11. In the analysis we used informative priors to integrate domain (biochemical) knowledge. For example, the knowledge from previous experiments that the signal-to-noise (i.e. donor photon to background photon) ratio is around from 3/1 to 8/1 leads us to put a wide gamma prior on A_0 with mean 35000 and standard deviation 25000; the knowledge that having a k that is much larger than 50000 is physically very unlikely leads to an exponential prior on k with mean 40000; the knowledge that the DNA hairpin switches roughly equally often between the open and closed states and that the two levels a and b are physically very unlikely to exceed 4 leads to a beta prior on π_1 with mean 0.5 and standard deviation 0.3, and wide gamma priors on a and b with mean 2 and 1 and standard deviation 2 and 0.8 respectively.

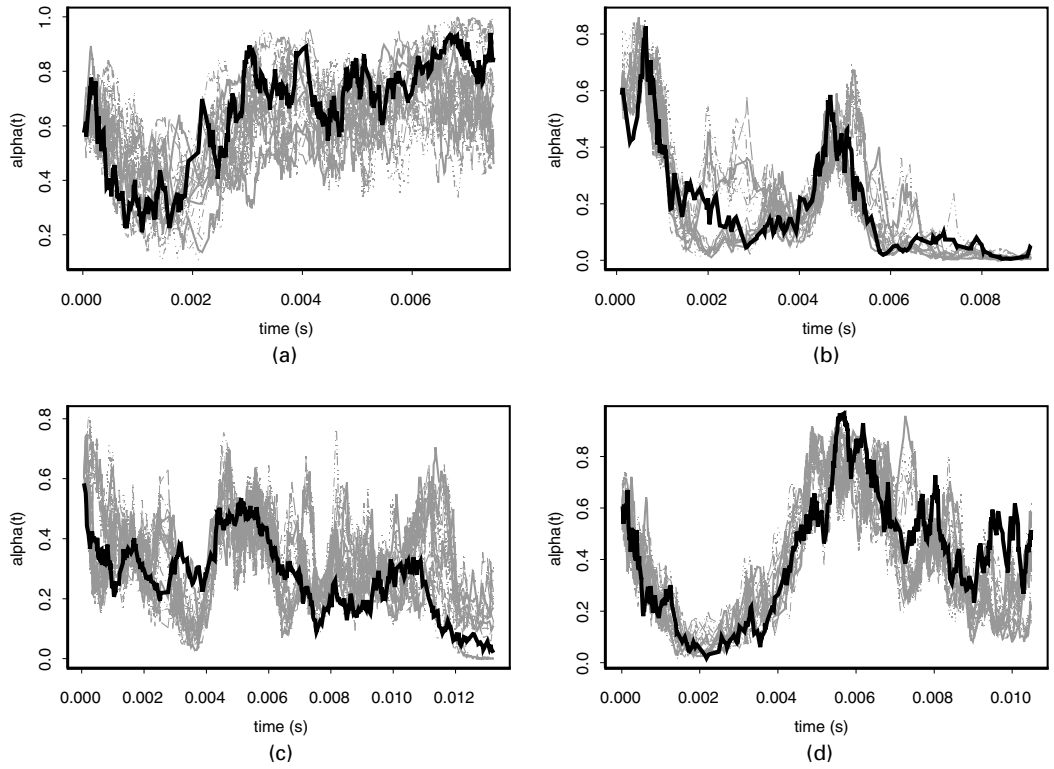

Fig. 10. Comparison of the augmented Brownian motion factors (———) with the actual factor (———) for the simulated data: (a) trajectory 5; (b) trajectory 25; (c) trajectory 35; (d) trajectory 45

Fig. 12 shows the posterior distribution of $1/k$, which, indicating the height of the energy barrier between the open and closed states, is termed the *decay time constant*. The 2.5-percentile and 97.5-percentile [39, 91] (microseconds) supply a 95% symmetric interval for the decay time constant $1/k$.

Since our method uses the likelihood, it is more efficient than the available method-of-moment type of estimation methods that are used in the chemistry literature. In the previous approaches arrival times must first be 'binned' together to smooth out the effect of the molecular Brownian diffusion, and then the binned arrival times are used to fit certain moment equations to estimate the parameters of interest (Pfluegl *et al.*, 1998; Brown and Silbey, 1998). Because what happens inside the binning time window is lost once the arrival times have been binned together, the binning approaches suffer a significant loss of time resolution. (In a sense, the binning approach is like measuring a distance by using a certain unit; if the real distance is shorter than the smallest unit, the measurement will not give a meaningful result.) For the same data set that we analysed, the methods based on binning have a maximum time resolution of 280 μs and indicate that the decay time constant $1/k$ is less than 280 μs. Though qualitatively consistent with the result from the binning approaches, our analysis provides a much sharper inference with the posterior median of $1/k$ being 59 μs and a 95% symmetric probability interval [39, 91] μs.

5.3. *Fitting the continuous diffusive model*

Since the stationary distribution of the control process (4.2) is $N(0, \xi)$, we note that the two-state model corresponds to a degenerate case of the continuous diffusive model with $\sqrt{\xi} = 0$. This

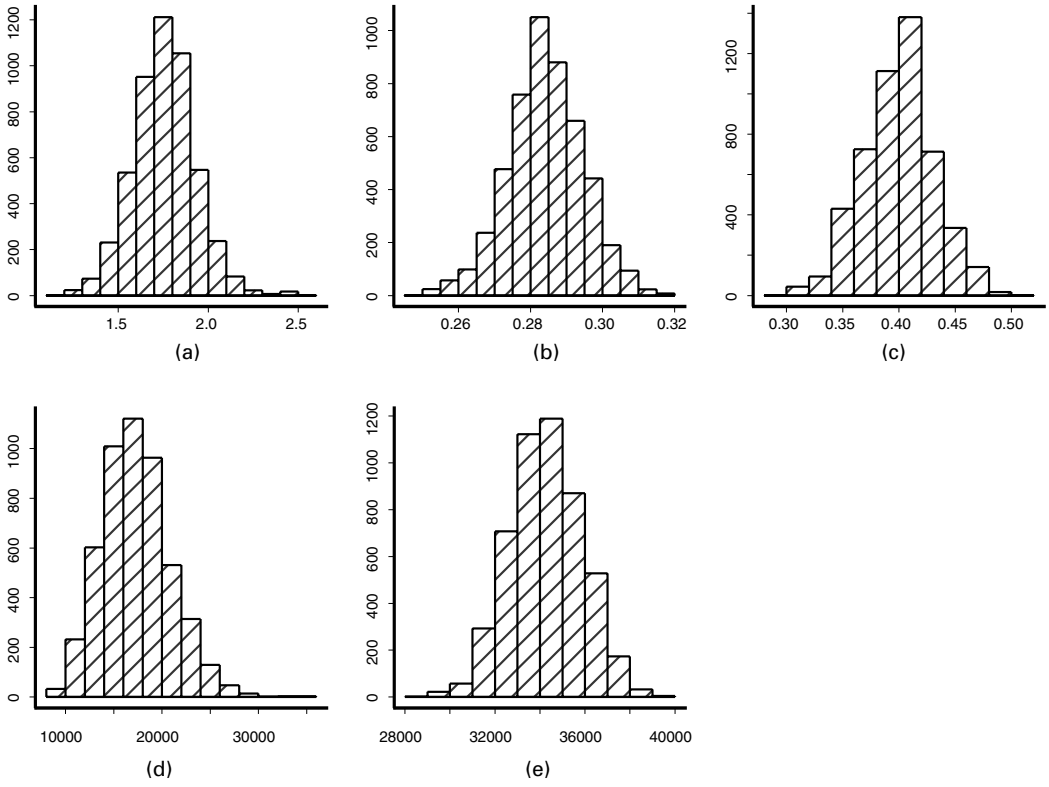

Fig. 11. Posterior histograms for the experimental data set with 1813 observation pairs: (a) a; (b) b; (c) π_1; (d) k; (e) A_0

suggests that, if the posterior probability mass of $\sqrt{\xi}$ is sufficiently far from 0, then the data support the diffusive model.

As an illustration, we apply the algorithm to analyse the same simulated data sets as were used in Section 5.1, which were generated according to the two-state model. 5000 samples were drawn from the posterior distribution (4.8) after a burn-in period of 1500 iterations. As shown in Fig. 13, the algorithm correctly identifies the parameters $\boldsymbol{\theta}$ as before. Furthermore, the posterior distribution of $\sqrt{\xi}$ suggests that there is no evidence against $\xi = 0$, indicating that the two-state model is indeed sufficient.

Next, we apply the method to the real experimental data set in Section 5.2 (with 1813 observation pairs) using the continuous diffusive model. Comparing Fig. 14 with Fig. 11, we observe that the posterior distributions of a, b and π_1 remain almost the same, whereas the distributions of k and A_0 shift slightly. Turning to the distribution of $\sqrt{\xi}$, we note that the mode is around 0.25. This comparison appears to offer some evidence in favour of the diffusive model, but not convincingly. We thus analyse 49 additional data sets that were acquired from the repeated single-molecule experiments on the same DNA hairpin to obtain a better model assessment.

Fig. 15, showing the results from eight typical data sets, represents the analysis of individual experiments. The results for $\boldsymbol{\theta}$ appear consistent across the experiments, whereas $\sqrt{\xi}$ appears to be distinct from 0 for some data sets and not so for others. Since all the 50 (independent) experiments should be governed by the same set of parameters, we next analyse these data sets jointly (the Brownian diffusion and the Ornstein–Uhlenbeck control process are augmented

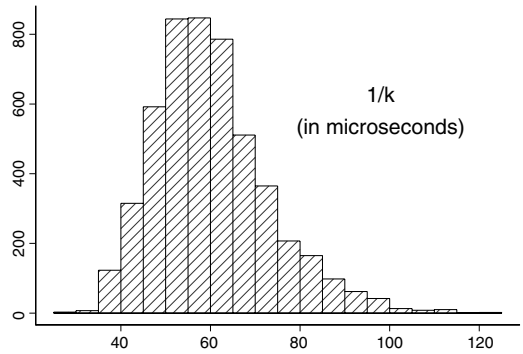

Fig. 12. Decay time constant $1/k$ of the experimental data set

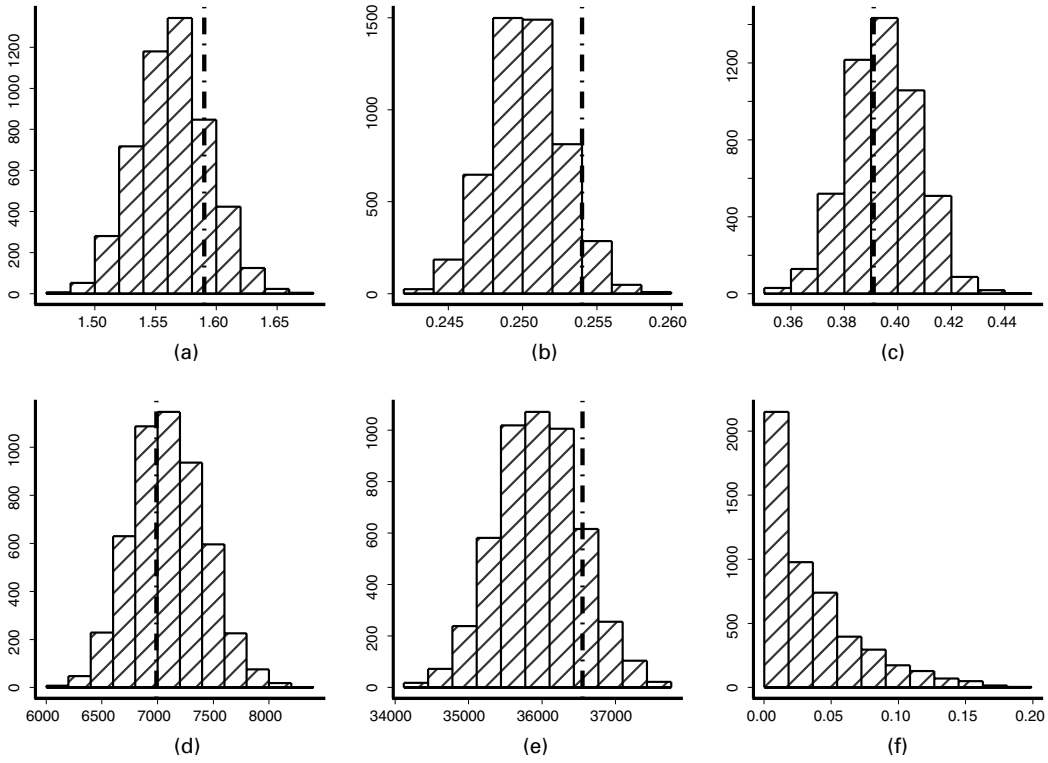

Fig. 13. Posterior histograms for the simulated two-state data sets ($\frac{1}{k}$, true value)—the distribution of $\sqrt{\xi}$, being highly concentrated around 0, strongly indicates the sufficiency of the two-state model: (a) a; (b) b; (c) π_1; (d) k; (e) A_0; (f) $\sqrt{\xi}$

for each data set). Fig. 16 shows the result from the 50 data sets combined. First we note that compared with the result from the individual data sets the posterior distributions of the parameters become narrower as more data provide more information. Second, and more importantly, Fig. 16 shows that the posterior samples of $\sqrt{\xi}$ are concentrated far from 0, which indicates strongly that the two-state model does not fit the data.

Though each individual experiment on its own offers only fragmental information, pooling the 50 experiments indicates that the two-state model is not sufficient to explain the experimental

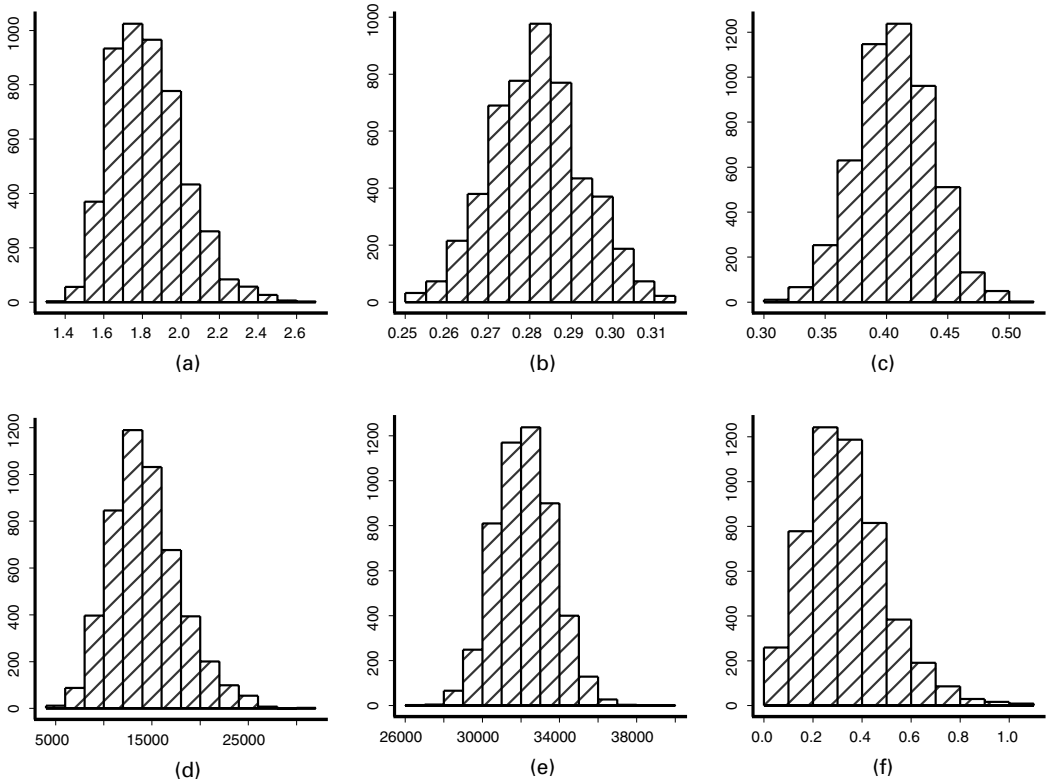

Fig. 14. Analysing the experimental data set (with 1813 data pairs) with the diffusive model: (a) a; (b) b; (c) π_1; (d) k; (e) A_0; (f) $\sqrt{\xi}$

data. To describe the detailed conformational dynamics of the DNA hairpin molecule more involved models appear necessary.

5.4. Monte Carlo estimation of the Bayes factor

To reinforce the intuitive result of Fig. 16, we further compute the Bayes factor between the two-state and the diffusive models, which is routinely used to aid Bayesian model selections (Kass and Raftery, 1995). In our case, the Bayes factor can be written as

$$\mathrm{BF} = \frac{P(\mathbf{t}, \boldsymbol{\tau}|M_1)}{P(\mathbf{t}, \boldsymbol{\tau}|M_2)} = \frac{\int \eta(\boldsymbol{\theta}) L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t) P\{\mathbf{B}(t)\} \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\mathbf{B}}{\int \eta'(\boldsymbol{\theta}, \lambda, \phi) L(\mathbf{t}, \boldsymbol{\tau}|\boldsymbol{\theta}, \alpha_t, x_t) P\{\mathbf{B}(t)\} P(x_t|\lambda, \phi^2/\lambda) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\lambda \, \mathrm{d}\phi \, \mathrm{d}\mathbf{B} \, \mathrm{d}x},$$

where M_1 denotes the two-state model and M_2 denotes the diffusive model. To simplify the notation, we denote the data as $\mathbf{y} = (\mathbf{t}, \boldsymbol{\tau})$ and write $\boldsymbol{\mu} = (\boldsymbol{\theta}, \mathbf{B})$ and $\boldsymbol{\zeta} = (\lambda, \phi, x_t)$, using which the above expression for the Bayes factor can be rewritten as

$$\mathrm{BF} = \frac{\int P(\mathbf{y}|M_1, \boldsymbol{\mu}) P(\boldsymbol{\mu}|M_1) \, \mathrm{d}\boldsymbol{\mu}}{\int P(\mathbf{y}|M_2, \boldsymbol{\mu}, \boldsymbol{\zeta}) P(\boldsymbol{\mu}, \boldsymbol{\zeta}|M_2) \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\boldsymbol{\zeta}},$$

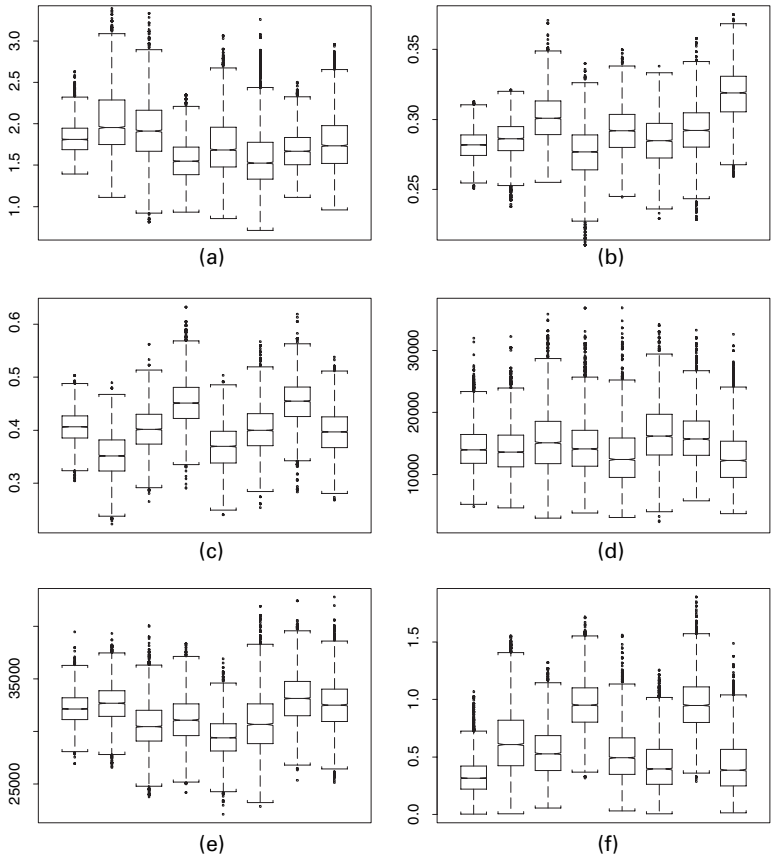

Fig. 15. Box plots of the results from analysing eight experimental data sets: (a) a; (b) b; (c) π_1; (d) k; (e) A_0; (f) $\sqrt{\xi}$

where we indicate that in model M_2 there is an additional set of parameters ($\boldsymbol{\zeta}$) in comparison with M_1.

Since a direct calculation of these integrals is infeasible, we seek to approximate the Bayes factor via Monte Carlo sampling. We note that, if the prior distributions of the two models are consistent, i.e.

$$P(\boldsymbol{\mu}|M_1) = \int P(\boldsymbol{\mu}, \boldsymbol{\zeta}|M_2)\,\mathrm{d}\boldsymbol{\zeta},$$

which is true in our case here, then the Bayes factor can be re-expressed as the posterior mean of the likelihood ratio:

$$\mathrm{BF} = \int \frac{P(\mathbf{y}|M_1, \boldsymbol{\mu})}{P(\mathbf{y}|M_2, \boldsymbol{\mu}, \boldsymbol{\zeta})} P(\boldsymbol{\mu}, \boldsymbol{\zeta}|\mathbf{y}, M_2)\,\mathrm{d}\boldsymbol{\mu}\,\mathrm{d}\boldsymbol{\zeta} = E\left[\frac{P(\mathbf{y}|M_1, \boldsymbol{\mu})}{P(\mathbf{y}|M_2, \boldsymbol{\mu}, \boldsymbol{\zeta})} \,\middle|\, \mathbf{y}, M_2 \right]. \tag{5.1}$$

This implies that, if we have posterior samples of the parameters $(\boldsymbol{\mu}^{(i)}, \boldsymbol{\zeta}^{(i)})$, $i = 1, \ldots, N$, drawn from the larger model M_2, we can estimate the Bayes factor by

$$\widehat{\mathrm{BF}} = \frac{1}{N} \sum_{i=1}^{N} \frac{P(\mathbf{y}|\boldsymbol{\mu}^{(i)})}{P(\mathbf{y}|\boldsymbol{\mu}^{(i)}, \boldsymbol{\zeta}^{(i)})}.$$

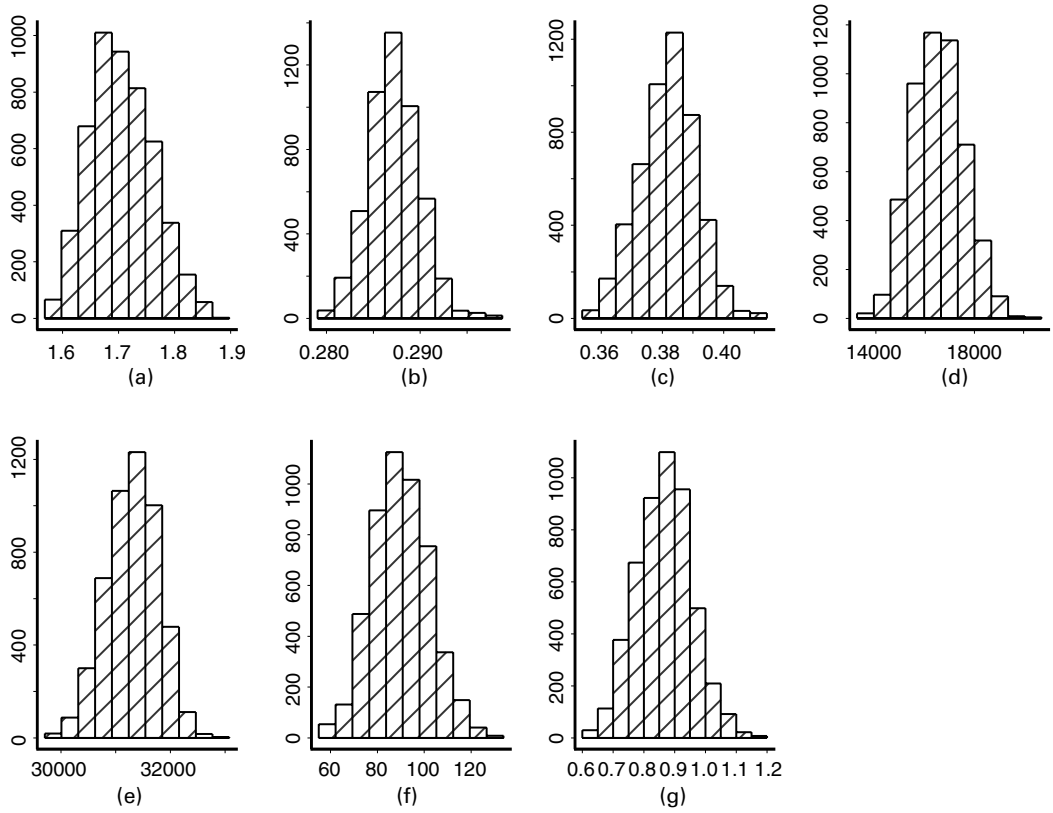

Fig. 16. Result from analysing 50 experimental data sets together (5000 samples were drawn from the posterior distribution): (a) a; (b) b; (c) π_1; (d) k; (e) A_0; (f) λ; (g) $\sqrt{\xi}$

Although this approach may not be the most efficient, it worked sufficiently well for all our data.

For the particular experimental data set (with 1813 data pairs) that was analysed in Section 5.2, calculation based on the 5000 posterior samples gives $\widehat{\mathrm{BF}} = 0.81 \pm 0.10$, which appears to favour the diffusive model slightly. To synthesize the available information, we calculate the Bayes factor on the basis of the combined analysis of the 50 experimental data sets as in Section 5.3, which gives $\widehat{\mathrm{BF}} = (3.43 \pm 0.29) \times 10^{-9}$ (the maximum likelihood ratio among the 5000 posterior samples is 5.79×10^{-7}). As a comparison, we also calculated the Bayes factor for the 50 data sets that were simulated from the two-state model in Section 5.1, and we obtained $\widehat{\mathrm{BF}} = 2.51 \pm 0.17$. This contrasts sharply with the Bayes factor from the real experimental data sets, which confirms our impression from Fig. 16.

The results from this subsection and the preceding subsection imply that the diffusive model describes the experimental data more accurately. From a scientific point of view the implication is that the energy barrier between the closed and open states of the DNA hairpin has more complex behaviour than the simple static picture that is depicted in the two-state model. The fluctuation of the energy barrier in this case may be due to conformational flexibility of the DNA molecule, which is subject to future investigation.

6. Discussion

This paper studies one type of single-molecule experiment: fluorescence lifetime fluctuation of

the DNA hairpin molecule. To analyse the experimental data efficiently, we have discussed three important statistical issues:

(a) synthesizing and comparing various stochastic models that are used in single-molecule conformational dynamics;
(b) deriving likelihood functions that are associated with the stochastic models involved;
(c) solving experimental complications, especially those involving unobserved stochastic processes.

The first issue is intrinsic to the nature of single-molecule experiments, whose stochasticity, unlike the traditional ensemble experiments, fundamentally requires stochastic modelling. Once the stochastic model has been constructed, efficient inference then relies on formulation of the likelihood; we demonstrate how to use the discretization and matrix technique to obtain closed form likelihoods. An important aspect for analysing real experimental data is the handling of experimental complications. Some complications (such as the time wrapping and negative reading of delay times) are relatively easy to handle, and the resulting modification to the likelihood is simple. Other complications such as those related to unobserved stochastic processes pose more difficult challenges, as computable analytical expression is impossible to obtain. Data augmentation techniques, aided with Markov chain Monte Carlo methods, prove to be a very powerful tool. They are not only conceptually simple but also provide a viable means to circumvent the analytical intractability.

Our Monte Carlo estimation of the Bayes factor relies on the simple and intuitive identity (5.1) that the posterior mean of the likelihood ratio (under the larger model) is equal to the Bayes factor of two nested models. It appears that this identity is well known (two of our colleagues, Xiao-Li Meng and Donald Rubin, have used variations of this identity privately before) and is very closely related to formulae that are used to derive the bridge sampling (Meng and Wong, 1996) and the ratio importance sampling methods (Chen and Shao, 1997; Chen *et al.*, 2000). However, we have neither been able to find a reference that explicitly states this result nor seen much of its use in Monte Carlo computations. In the special case when the smaller model is a fixed distribution (no unknown parameters), our method is equivalent to the 'harmonic mean' method of Newton and Raftery (1994). Although our method may not be as efficient as some other more delicate approaches in specific cases (e.g. the techniques that were described in Chen *et al.* (2000) and Meng and Wong (1996)), it works effectively for the study that is reported here—the variances of the likelihood ratios are well in control in all cases. In comparison, Chib's (1995) method does not work here because it requires an estimation of posterior density values at certain points, which is infeasible for our models. The reversible jump method (Green, 1995) is also unlikely to help because we know of no efficient jumping rules to add or remove a hidden Ornstein–Uhlenbeck process (i.e. jumping between the two-state and the continuous diffusive models).

Although the Markov chain Monte Carlo approaches that are illustrated in this paper have found wide acceptance in the statistics community in the past decade (Liu, 2001), their use in other scientific disciplines is relatively sparse. In the past, many researchers in physical sciences did not feel the necessity for delicate and efficient statistical inference methods in that their data on the ensemble were often overwhelmingly large and *ad hoc* methods such as moment matching would be more than sufficient to provide them with needed information. The single-molecule experiments that have been enabled by the advance of modern technology, as well as many large scale genomics experiments, seem to alter the landscape significantly. As shown in this paper, the Bayesian analysis provides much sharper estimates of the parameters that are associated with DNA hairpin dynamics compared with the existing moment matching and

binning methods. The data augmentation approach also enables us to study the goodness of fit of different models to the experimental data, which is especially difficult, if not impossible, for existing moment-based methods. Neither single-molecule experiments nor data augmentation (with Markov chain Monte Carlo) methods were thinkable even two decades ago. Therefore, it is our hope that this paper will generate further interest in applying modern statistical methodology to interesting and important scientific problems.

The experimental data sets that were used in this paper can be downloaded from the authors' Web pages `http://www.fas.harvard.edu/~skou` and `http://www.fas.Harvard.Edu/~junliu`.

Acknowledgements

The authors thank Haw Yang for providing both the experimental data and Figs 1 and 3, and Long Cai and Xiao-Li Meng for helpful discussions. The authors are grateful to the referees, whose comments and suggestions helped us greatly in clarifying many issues and improving the overall presentation. This work is supported in part by two National Science Foundation grants (DMS-0204674 and DBI-0138028), the Harvard University Clarke-Cooke Fund and a grant from the Office of Basic Energy Science, Office of Science, Department of Energy (DOE-FG02-00ER15072).

Appendix A: Mathematical derivations

A.1. Computation of $L(t_0, t_1) = P\{Y(t_1^-) - Y(t_0) = 0, \gamma(t_1) | \gamma(t_0)\}$ in Section 2.1

We first divide the interval (t_0, t_1) into N infinitesimal pieces: $(t_0, t_0+h), (t_0+h, t_0+2h), \ldots, (t_0+(N-1)h, t_1)$, where $h = (t_1 - t_0)/N$. The probability of observing 0 photons in the ith infinitesimal interval is

$$1 - \frac{A_0}{\gamma(t_0+ih)}h + o(h), \qquad i = 0, 1, \ldots, N-1.$$

This, together with the transition matrix (1.2), implies that with discretization resolution h the probability of not observing a photon in (t_0, t_1) is

$$\int \prod_{i=0}^{N-1} P(h)_{(I_{t_0+ih}, I_{t_0+(i+1)h})} \left\{ 1 - \frac{A_0}{\gamma(t_0+ih)}h + o(h) \right\} \mathrm{d}\gamma, \tag{A.1}$$

where I_t is equal to 1 if $\gamma_t = a$ and 2 if $\gamma_t = b$, and $P(h)_{(i_1,i_2)}$ denotes the (i_1,i_2)th entry of the transition matrix $P(h)$. The integral in expression (A.1) is taken with respect to the hidden γ process. Letting $h \to 0$ gives the probability

$$L(t_0, t_1) = \lim_{h \to 0} \left[\int \prod_{i=0}^{N-1} P(h)_{(I_{t_0+ih}, I_{t_0+(i+1)h})} \left\{ 1 - \frac{A_0}{\gamma(t_0+ih)}h + o(h) \right\} \mathrm{d}\gamma \right]. \tag{A.2}$$

By equation (1.2) the individual element in the above product can be written in a simple matrix form

$$P(h)_{(I_{t_0+ih}, I_{t_0+(i+1)h})} \left\{ 1 - \frac{A_0}{\gamma(t_0+ih)}h + o(h) \right\} = \left\{ \begin{pmatrix} 1 - \dfrac{A_0}{a}h & 0 \\ 0 & 1 - \dfrac{A_0}{b}h \end{pmatrix} \exp(Qh) + o(h) \right\}_{(I_{t_0+ih}, I_{t_0+(i+1)h})}.$$

Letting

$$H = \begin{pmatrix} A_0/a & 0 \\ 0 & A_0/b \end{pmatrix},$$

we obtain the matrix form of equation (A.2) as

$$[\{(I - Hh)\exp(Qh) + o(h)\}^{(t_1-t_0)/h}]_{(I_{t_0}, I_{t_1})}.$$

Noting that $(I - Hh)\exp(Qh) = I + (Q - H)h + o(h)$, we have

$$L(t_0, t_1) = \lim_{h \to 0} [\{I + (Q - H)h + o(h)\}^{(t_1 - t_0)/h}]_{(I_{t_0}, I_{t_1})} = [\exp\{(Q - H)(t_1 - t_0)\}]_{(I_{t_0}, I_{t_1})}.$$

A.2. Derivation of expression (4.10)

Liu and Sabatti (2000) considered a general framework: move a sample, say $\mathbf{x}$, to a transformed value $\nu(\mathbf{x})$, where the transformation ν is drawn from a group. They showed that, to maintain the invariance of the target distribution $\pi(\mathbf{x})$, the distribution on the transformation ν should be

$$p(\nu) \propto \pi\{\nu(\mathbf{x})\} |J_\nu(\mathbf{x})| \mu(\mathrm{d}\nu), \tag{A.3}$$

where $J_\nu(\mathbf{x})$ is the Jacobian of the transformation and μ is the left-Haar invariant measure on the group. For scale changes that we consider here, the left-Haar measure $\mu(\mathrm{d}s) = s^{-1}\,\mathrm{d}s$ and the Jacobian is simply s^d, where s is the scale factor and d is the dimension of the scale change. Plugging them into the general formula (A.3) and noting that

$$P\left\{ s x_t | \lambda, \frac{(s\phi)^2}{\lambda} \right\} = s^{-(n+1)} P\left(x_t | \lambda, \frac{\phi^2}{\lambda} \right)$$

yield expression (4.10).

A.3. Proof of theorem 1

The transition kernel of the group move is given by

$$\begin{aligned} K(\mathbf{x}, B) = & \int_{\{\nu: \nu(\mathbf{x}) \in B\}} f(\nu) \min\left[1, \frac{\pi\{\nu(\mathbf{x})\} f(\upsilon^{-1})}{\pi(\mathbf{x})\, f(\nu)} |J_\nu(\mathbf{x})|\right] \mu(\mathrm{d}\nu) \\ & + I_B(\mathbf{x}) \left(1 - \int_\Gamma f(\nu) \min\left[1, \frac{\pi\{\nu(\mathbf{x})\}\, f(\upsilon^{-1})}{\pi(\mathbf{x})\, f(\nu)} |J_\nu(\mathbf{x})|\right] \mu(\mathrm{d}\nu)\right). \end{aligned}$$

Next we verify that $\int_{\mathcal{X}} \pi(\mathbf{x}) K(\mathbf{x}, B)\,\mathrm{d}\mathbf{x} = \int_B \pi(\mathbf{x})\,\mathrm{d}\mathbf{x}$ for every B so that π is invariant under the move. Direct calculation gives

$$\begin{aligned} & \int_{\mathcal{X}} \pi(\mathbf{x}) K(\mathbf{x}, B)\,\mathrm{d}\mathbf{x} - \int_B \pi(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ & \quad = \int_{\mathcal{X}} \int_\Gamma I\{\nu(\mathbf{x}) \in B\}\, \pi(\mathbf{x})\, f(\nu) \min\left[1, \frac{\pi\{\nu(\mathbf{x})\}\, f(\upsilon^{-1})}{\pi(\mathbf{x})\, f(\nu)} |J_\nu(\mathbf{x})|\right] \mu(\mathrm{d}\nu)\,\mathrm{d}\mathbf{x} \\ & \quad\quad - \int_B \int_\Gamma \pi(\mathbf{x})\, f(\nu) \min\left[1, \frac{\pi\{\nu(\mathbf{x})\}\, f(\upsilon^{-1})}{\pi(\mathbf{x})\, f(\nu)} |J_\nu(\mathbf{x})|\right] \mu(\mathrm{d}\nu)\,\mathrm{d}\mathbf{x} \end{aligned} \tag{A.4}$$

Applying a one-to-one map

$$\begin{cases} g = \nu^{-1}, \\ \mathbf{y} = \nu(\mathbf{x}) \end{cases}$$

and using Fubini's theorem, the first term on the right-hand side of equation (A.4) becomes

$$\begin{aligned} & \int_{\mathcal{X}} \int_\Gamma I\{\nu(\mathbf{x}) \in B\}\, \pi(\mathbf{x})\, f(\nu) \min\left[1, \frac{\pi\{\nu(\mathbf{x})\}\, f(\upsilon^{-1})}{\pi(\mathbf{x})\, f(\nu)} |J_\nu(\mathbf{x})|\right] \mu(\mathrm{d}\nu)\,\mathrm{d}\mathbf{x} \\ & \quad = \int_{g \in \Gamma} \int_{\mathbf{y} \in \mathcal{X}} I(\mathbf{y} \in B)\, \pi\{g(\mathbf{y})\}\, f(g^{-1}) \min\left[1, \frac{\pi(\mathbf{y})\, f(g)}{\pi\{g(\mathbf{y})\} f(g^{-1})} \left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right|\right] \left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right| \mathrm{d}\mathbf{y}\, \mu(\mathrm{d}g^{-1}) \\ & \quad = \int_{\mathbf{y} \in B} \int_{g \in \Gamma} \min\left[\pi\{g(\mathbf{y})\} f(g^{-1}) \left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right|, \pi(\mathbf{y})\, f(g)\right] \mu(\mathrm{d}g^{-1})\,\mathrm{d}\mathbf{y} \\ & \quad = \int_B \int_\Gamma \min[\pi\{g(\mathbf{y})\}\, f(g^{-1}) |J_g(\mathbf{y})|, \pi(\mathbf{y})\, f(g)]\, \mu(\mathrm{d}g^{-1})\,\mathrm{d}\mathbf{y}, \end{aligned}$$

which is exactly the last term on the right-hand side of equation (A.4) on noting that for the Haar measure $\mu(\mathrm{d}v) = \mu(\mathrm{d}v^{-1})$. This cancellation in equation (A.4), therefore, implies that

$$\int_{\mathcal{X}} \pi(\mathbf{x})\, K(\mathbf{x}, B)\, \mathrm{d}\mathbf{x} = \int_{B} \pi(\mathbf{x})\, \mathrm{d}\mathbf{x}, \qquad \text{for any } B.$$

References

Agmon, N. and Hopfield, J. J. (1983) Transient kinetics of chemical reactions with bounded diffusion perpendicular to the reaction coordinate: intramolecular processes with slow conformational changes. *J. Chem. Phys.*, **78**, 6947–6959.

Ansari, A., Kuznetsov, S. V. and Shen, Y. (2001) Configurational diffusion down a folding funnel describes the dynamics of DNA hairpins. *Proc. Natn. Acad. Sci. USA*, **98**, 7771–7776.

Ansari, A., Shen, Y. and Kuznetsov, S. V. (2002) Misfolded loops decrease the effective rate of DNA hairpin formation. *Phys. Rev. Lett.*, **88**, 069801.

Bonnet, G., Krichevsky, O. and Libchaber, A. (1998) Kinetics of conformational fluctuations in DNA hairpin-loops. *Proc. Natn. Acad. Sci. USA*, **95**, 8602–8606.

Brown, F. L. and Silbey, R. J. (1998) An investigation of the effects of two level system coupling on single molecule lineshapes in low temperature glasses. *J. Chem. Phys.*, **108**, 7434–7450.

Cao, J. (2000) Event-averaged measurements of single molecule kinetics. *Chem. Phys. Lett.*, **327**, 38–44.

Chen, M.-H. and Shao, Q.-M. (1997) On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.*, **25**, 1563–1594.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000) *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *J. Am. Statist. Ass.*, **90**, 1313–1321.

Cox, D. R. (1955) The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proc. Camb. Phil. Soc.*, **51**, 433–441.

Eggeling, C., Fries, J., Brand, L., Günther, R. and Seidel, C. (1998) Monitoring conformational dynamics of a single molecule by selective fluorescence spectroscopy. *Proc. Natn. Acad. Sci. USA*, **95**, 1556–1561.

Elliott, R., Aggoun, L. and Moore, J. (1997) *Hidden Markov Models: Estimation and Control*. New York: Springer.

Fredkin, D. and Rice, J. (1986) On aggregated Markov processes. *J. Appl. Probab.*, **23**, 208–214.

Froelich-Ammon, S., Gale, K. and Osheroff, N. (1994) Site-specific cleavage of a DNA hairpin by topoisomerase II. DNA secondary structure as a determinant of enzyme recognition/cleavage. *J. Biol. Chem.*, **269**, 7719–7725.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Grunwell, J., Glass, J., Lacoste, T., Deniz, A., Chemla, D. and Schultz, P. (2001) Monitoring the conformational fluctuations of DNA hairpins using single-pair fluorescence energy transfer. *J. Am. Chem. Soc.*, **123**, 4295–4303.

Jia, Y., Sytnik, A., Li, L., Vladimirov, S., Cooperman, B. and Hochstrasser, R. (1997) Nonexponential kinetics of a single $\mathrm{tRNA}^{\mathrm{PHe}}$ molecule under physiological conditions. *Proc. Natn. Acad. Sci. USA*, **94**, 7932–7936.

Karlin, S. and Taylor, H. (1981) *A Second Course in Stochastic Processes*. New York: Academic Press.

Karlin, S. and Taylor, H. (1998) *An Introduction to Stochastic Modeling*, 3rd edn. New York: Academic Press.

Karr, A. (1986) *Point Processes and Their Statistical Inference*. New York: Dekker.

Kass, R. and Raftery, A. (1995) Bayes factors and model uncertainty. *J. Am. Statist. Ass.*, **90**, 773–795.

Krichevsky, O. and Bonnet, G. (2002) Fluorescence correlation spectroscopy: the technique and its applications. *Rep. Prog. Phys.*, **65**, 251–297.

Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Liu, J. S. and Sabatti, C. (2000) Generalized Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*, **87**, 353–369.

Lu, H. P., Xun, L. and Xie, X. S. (1998) Single-molecule enzymatic dynamics. *Science*, **282**, 1877–1882.

Magde, D., Elson, E. and Webb, W. (1974) Fluorescence correlation spectroscopy. *Biopolymers*, **13**, 1–61.

Meng, X.-L. and Wong, W. H. (1996) Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sin.*, **6**, 831–860.

Moerner, W. (2002) A dozen years of single-molecule spectroscopy in physics, chemistry, and biophysics. *J. Phys. Chem.* B, **106**, 910–927.

Newton, M. A. and Raftery, A. E. (1994) Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. R. Statist. Soc.* B, **56**, 3–48.

Nie, S. and Zare, R. (1997) Optical detection of single molecules. *Ann. Rev. Biophys. Biomol. Struct.*, **26**, 567–596.

Pfluegl, W., Brown, F. L. and Silbey, R. J. (1998) Variance and width of absorption lines of single molecules in low temperature glasses. *J. Chem. Phys.*, **108**, 6876–6883.

Reilly, P. D. and Skinner, J. L. (1993) Spectral diffusion of single molecule fluorescence: a probe of low-frequency localized excitations in disordered crystals. *Phys. Rev. Lett.*, **71**, 4257–4260.

Reilly, P. D. and Skinner, J. L. (1994a) Spectroscopy of a chromophore coupled to a lattice of dynamic two-level systems: I, Absorption line shape. *J. Chem. Phys.*, **101**, 959–964.

Reilly, P. D. and Skinner, J. L. (1994b) Spectroscopy of a chromophore coupled to a lattice of dynamic two-level systems: II, Spectral diffusion kernel. *J. Chem. Phys.*, **101**, 965–973.
Schenter, G. K., Lu, H. P. and Xie, X. S. (1999) Statistical analyses and theoretical models of single-molecule enzymatic dynamics. *J. Phys. Chem.* A, **103**, 10477–10488.
Tamarat, P., Maali, A., Lounis, B. and Orrit, M. (2000) Ten years of single-molecule spectroscopy. *J. Phys. Chem.* A, **104**, 1–16.
Tang, J., Temsamani, J. and Agrawal, S. (1993) Self-stabilized antisense oligodeoxynucleotide phosphorothioates: properties and anti-HIV activity. *Nucleic Acids Res.*, **21**, 2729–2735.
Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Ass.*, **82**, 528–540.
Trinh, T. and Sinden, R. (1993) The influence of primary and secondary DNA structure in deletion and duplication between direct repeats in Escherichia coli. *Genetics*, **134**, 409–422.
Weiss, S. (2000) Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nature Struct. Biol.*, **7**, 724–729.
Wolpert, R. L. and Ickstadt, K. (1998) Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**, 251–267.
Xie, X. S. (2002) Single-molecule approach to dispersed kinetics and dynamic disorder: probing conformational fluctuation and enzymatic dynamics. *J. Chem. Phys.*, **117**, 11024–11032.
Xie, X. S. and Lu, H. P. (1999) Single-molecule enzymology. *J. Biol. Chem.*, **274**, 15967–15970.
Xie, X. S. and Trautman, J. K. (1998) Optical studies of single molecules at room temperature. *A. Rev. Phys. Chem.*, **49**, 441–480.
Yang, H., Luo, G., Karnchanaphanurach, P., Louise, T.-M., Rech, I., Cova, S., Xun, L. and Xie, X. S. (2003) Protein conformational dynamics probed by single-molecule electron transfer. *Science*, **302**, 262–266.
Yang, S. and Cao, J. (2001) Two-event echos in single-molecule kinetics: a signature of conformational fluctuations. *J. Phys. Chem.* B, **105**, 6536–6549.
Yang, S. and Cao, J. (2002) Direct measurements of memory effects in single molecule kinetics. *J. Chem. Phys.*, **117**, 10996–11009.
Ying, L., Wallace, M. and Klenerman, D. (2001) Two-state model of conformational fluctuation in a DNA hairpin-loop. *Chem. Phys. Lett.*, **334**, 145–150.
Zazopoulos, E., Lalli, E., Stocco, D. and Sassone-Corsi, P. (1997) DNA binding and transcriptional repression by DAX-1 blocks steroidogenesis. *Nature*, **390**, 311–315.
Zwanzig, R. (1990) Rate processes with dynamical disorder. *Acc. Chem. Res.*, **23**, 148–152.

Discussion on the paper by Kou, Xie and Liu

Alan G. Hawkes (*University of Wales, Swansea*)

I welcome the, perhaps too infrequent, appearance of a stochastic modelling paper at a meeting of the Society, although the emphasis is perhaps more on the inferential aspects of the problem. For the simple model, without diffusion, my inclination would be to use good old-fashioned likelihood methods. However, once diffusion enters the model, I can well understand the attraction of Bayesian methods using Markov chain Monte Carlo methods: if you are going to augment the data with randomly generated values then you might as well generate parameter values also. I am not sufficiently expert in the practicalities of Markov chain Monte Carlo sampling to make serious comments about the details of that, although it seems to work out impressively well. From now on I shall restrict myself to discussing the stochastic modelling.

I would like to echo the authors' introductory remarks about the advantages of single-molecule observations in reaching an understanding of the structure and function of chemical processes. Together with David Colquhoun, from University College London, and others, I have been involved with the modelling of single ion channel behaviour for about 25 years. I would like to draw this briefly to your attention. An elementary introduction is given in Hawkes (2003).

Ion channels are large protein molecules, embedded in a cell membrane, that control the movement of electrically charged ions across the membrane. The flow of these charged ions constitutes a flow of electrical current. All electrical activity in the nervous system appears to be regulated by ion channels, thus playing a role in many diverse activities including thought processes, the transmission of nerve signals and their conversion into muscular contraction and controlling the release of insulin. Understanding their behaviour aids our understanding of normal physiology and the effect of drugs and toxins. It is therefore an important step towards developing treatments for a wide variety of medical conditions.

Remarkably, since the Nobel Prize winning work of Neher and Sakmann (1976), it has been possible to observe the current flowing through a single channel. In many cases, apart from a little noise, the current is either zero when the channel is shut or some constant value when the channel is open. It is usually not too difficult to extract an idealized signal of a sequence of alternating open–shut periods, as shown in the

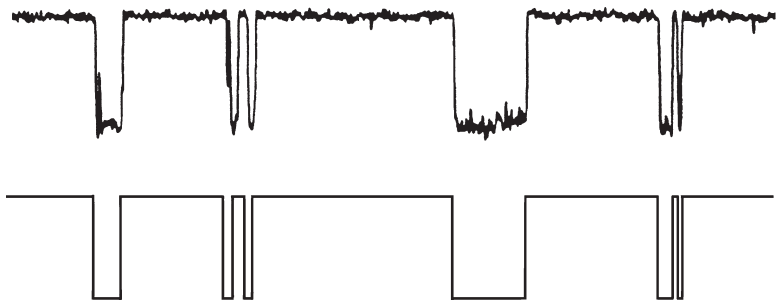

Fig. 17. Ideal open–closed sequence extracted from a noisy signal

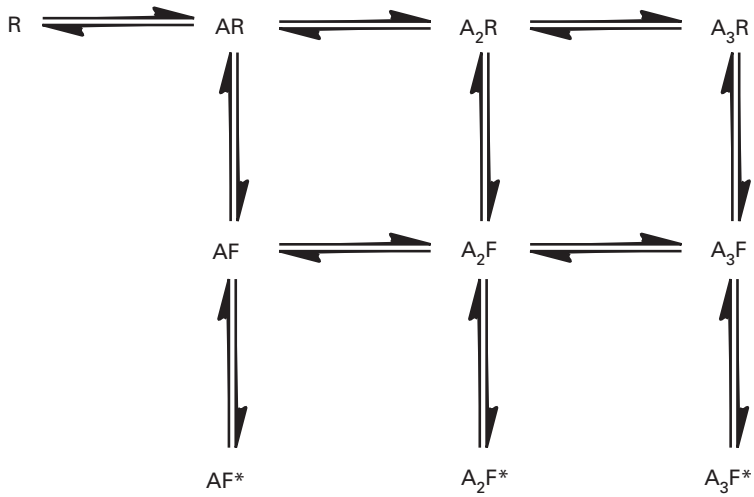

Fig. 18. Possible model for a glycine receptor: the three states on the bottom row are open states; the other seven states are closed

example in Fig. 17. In effect, we are lucky to be able to observe this molecule much more directly than is possible in fluorescence experiments.

Like the present paper, the underlying model is a finite state Markov process that describes various physical states that the molecule can adopt. Each state may be classified as open or shut. Fig. 18 shows a recent possible model of a glycine receptor (Burzomato *et al.*, 2004) with seven shut states and three open states (the three on the bottom row of Fig. 18). From any such model we can derive a description of the open–shut sequence in terms of a (hidden) semi-Markov process. What we observe is an aggregated Markov process, because we can observe only which set of states (open or shut) it is in, not the individual states. Actually, there are some complications concerning intervals that are too short to be resolved: the semi-Markov formulation can take these into account, but we need not detain you with details here. From this we can derive a likelihood for such a sequence and it has been shown to be quite effective in obtaining good maximum likelihood estimates of the parameters and in discriminating between potential underlying models.

Bayesian methods have also been used to derive parameter estimates directly from the original signal plus noise measurements, and also in extracting the signal from the noise (Fredkin and Rice, 1992). These can work quite well even for multilevel observations and quite high levels of noise (see the example in Fig. 19).

After that diversion, let me return to the model in this paper.

I am puzzled about the basic stochastic model, no doubt because of my weak grasp of physics.

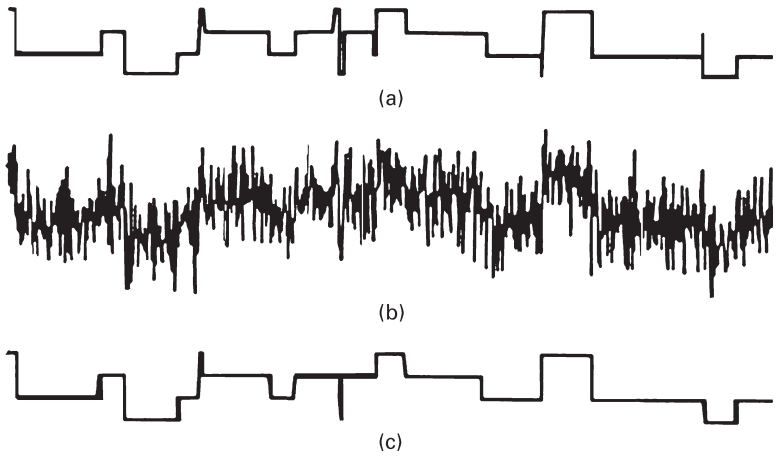

Fig. 19. Bayesian restoration: (a) simulated signal with four levels; (b) signal plus noise; (c) restored signal (four very short events missed)

If you look at the derivation of the likelihood in equation (2.3) you see that the process of photon observation depends on the underlying gamma process $\gamma(t)$, which itself is not dependent in any way on the sequence of pulses. And yet we are told that it is the pulses that stimulate photon emission: puzzling!

Then the delay times seem to be probabilistically independent of the photon arrival process, and of each other, and mathematically dependent on the value of the gamma process at the times of arrival, t_i, only. At first sight these assumptions seem most unlikely. I suppose the answer is that this is a reasonable approximation if the pulse rate is much faster than the rate at which $\gamma(t)$ changes and the rate at which photons arrive.

I would have welcomed more explanation of these points.

There are a couple of minor issues.

In remark 1 on page 474 we are told that the discretized approach to deriving the likelihood, as given in Appendix A, is much easier to generalize to more complicated underlying Markov models and the introduction of Brownian diffusions than it is if you derive the likelihood from a set of differential equations involving the infinitesimal generator. I disagree. The differential method is, to my mind, simpler and more elegant. Morever, using this method, the more complicated model is really no more difficult to solve than the simple model. A derivation along these lines follows.

To generalize slightly, let $X(t)$ denote the state that the molecule is in at time t and let the Poisson intensity at that time be $\alpha(t)\lambda_i$ if $X(t) = i$.

Let $\mathbf{H} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ and let the matrix $\mathbf{G}(t)$ have elements

$$g_{ij}(t) = P\{Y(t) - Y(0) = 0 \cap X(t) = j | X(0) = i\}.$$

Then it is easy to see that

$$\begin{aligned}\mathbf{G}(t+\delta t) &= \mathbf{G}(t)\{\mathbf{I} - \alpha(0)\mathbf{H}\,\delta t + o(\delta t)\}\{\mathbf{I} + \mathbf{Q}\,\delta t + o(\delta t)\} \\ &= \mathbf{G}(t)[\mathbf{I} + \{\mathbf{Q} - \alpha(0)\mathbf{H}\}\,\delta t + o(\delta t)]\end{aligned}$$

so that

$$\frac{\mathrm{d}\mathbf{G}}{\mathrm{d}t} = \mathbf{G}\{\mathbf{Q} - \alpha(0)\mathbf{H}\},$$

which is readily solved to give

$$\mathbf{G}(t) = \exp[\{\mathbf{Q} - \alpha(0)\mathbf{H}\}t]$$

as required.

I have another small quibble. In Section 4.1 we are told that, in the two-by-two model, the states A_1 and A_2 are experimentally indistinguishable as are the pair B_1 and B_2, thus forming a so-called aggregated

Markov process. This is only true if the constants $a_1 = a_2$ and $b_1 = b_2$ in the definitions following equation (2.4).

I am pleased to propose a vote of thanks to the authors for presenting their stimulating paper.

Mark Steel (*University of Warwick, Coventry*)

The authors are to be congratulated for a very interesting and stimulating paper which builds on an emerging literature in biochemistry on single-molecule experiments. I am afraid that I have no real expertise in bioinformatics, and, as a consequence, I will have to limit this discussion to rather more standard issues in Bayesian statistics.

Bayesian inference with latent processes

The authors propose models that depend on unobserved stochastic processes, and that unsurprisingly complicates the computational analysis. There is a growing literature (e.g. in finance) on Bayesian inference with latent stochastic processes, and it would be interesting, in my view, to compare the approach that is used here with some of the alternative proposals in the literature.

A first issue in this context is that the authors use componentwise updating of the diffusion chains. There is some mention of 'blocking' the sampler on the diffusions in Section 3.2 and a promise of coming back to these blockwise moves in a later section, but that presumably refers to the group move, rather than a 'real' blocking move, in the sense of for example Liu *et al.* (1994) or Shephard and Pitt (1997). Would the more usual types of blocking help in this setting?

A second, and possibly more critical, issue is how to deal with the overconditioning (i.e. the high correlation between the latent variables and the parameters). This occurs in the case of the $x(t)$ process that is introduced in the continuous diffusive model. The solution proposed is the use of a scale transformation which jointly updates the process and its diffusion parameter. One possible problem with such a scale transformation is that the whole process is scaled at once, and such moves could be difficult to accept if the likelihood is particularly informative about certain parts of the process. Here it seems to work, possibly as a consequence of the 'generalized Gibbs' step of Liu and Sabatti (2000), which brings the likelihood into play. This scaling is somewhat reminiscent of the approach of Roberts and Stramer (2001) who reparameterized the 'missing data' in a partially observed diffusion model (although they used a slightly different transformation to achieve continuity of the sample paths at the observed points). This can be given an interpretation as a so-called 'non-centred parameterization' as explained in Papaspiliopoulos *et al.* (2003). Basically, the non-centring resolves the situation where the augmented data contain much more information about the parameters than the observed data. Barndorff-Nielsen and Shephard (2001) introduced a class of stochastic volatility models where the (latent) volatility is modelled as an Ornstein–Uhlenbeck process, driven by a positive Lévy process without Gaussian component. In the context of these models non-centring was also proposed to solve the overconditioning problem in Roberts *et al.* (2004). For the same type of models, Griffin and Steel (2003) proposed a sampler where the volatility process is thinned in line with the parameter values proposed. Griffin and Steel (2003) implemented so-called 'dependent thinning', where the jumps that are added to or deleted from the process tend to be relatively small. This is achieved by the use of the Ferguson–Class representation of Lévy jump processes, which allows us to focus on small jumps when proposing changes to the process. Thus, the new process will be 'close' to the previous process, while also being compatible with the new value of the parameters. In fact, Roberts *et al.* (2004) mentioned that this approach can also be given an interpretation as a non-centred parameterization. Even though the models that were used in the papers mentioned are somewhat different, I wonder whether similar ideas could be used in the context of the present paper.

Prior issues

The authors mention in Section 1 that 'we have rather detailed knowledge (prior) on the various parameters that are involved'. Nevertheless, there is little discussion of prior elicitation or prior issues in general in the paper. In the simulated data examples a flat prior is used throughout. The authors do not make explicit what 'flat' means, but if it means uniform on (the logarithms of) the parameters defined on $\Re_+$ that makes the prior improper and raises the familiar spectre of posterior existence. Does the posterior exist with the chosen prior? In addition, why would uniformity on the particular parameterization that is used (rather than some other parameterization) be a reasonable choice to reflect a lack of prior information (which is presumably what the authors intend)? Is there perhaps an invariance argument that can motivate this choice? Also, it would be useful to have a feeling for the importance of the prior assumptions, through, for example, a prior sensitivity analysis. Whereas data sets of typical sizes could swamp the prior in certain directions, I am not convinced that this is necessarily so for all prior dimensions. It would be good to know

where to concentrate the effort in eliciting a prior for practical problems, and some ideas about how such elicitation might proceed could be quite valuable for applied users of these methods.

Estimation of the Bayes factor

The identity in equation (5.1) that is used for the estimation of the Bayes factor is interesting, and, as the authors state, has not been used much in the literature. I wonder whether the reason might be the same as for the 'harmonic mean' estimator, which constitutes the special case if the simpler model has no unknown parameters. Newton and Raftery (1994) commented that the harmonic mean does not, generally, satisfy a Gaussian central limit theorem and can be very unstable as a consequence. We would expect that such problems will be somewhat less severe for a likelihood ratio, but I wonder whether stability of the resulting estimates is still an issue, especially if the extra parameters in the larger model involve a latent process.

Finally, it gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

Chris Sherlock (*Lancaster University*)

We comment on the authors' use of flat priors in their simulation studies and offer a Gibbs sampler (see Fearnhead and Sherlock (2004)) as an alternative to the Metropolis–Hastings scheme that is described in Section 2.

Flat priors

We observe that the likelihood has a lower bound that is strictly greater than 0 over the entire range of certain parameters. For example

$$P(\text{data}) \geqslant P(\text{data} \cap \text{molecule always closed})$$

where the right-hand side is independent of the parameter b.

Thus an improper (on standard or log-scales) prior for b will lead to an improper posterior.

Gibbs sampler

In the present context *the state of the hidden (Markov) chain* is the DNA molecule being closed or open.

The Gibbs sampler has three stages.

Step 1: given parameter values at the end of the previous iteration, sample the state of the hidden Markov process at photon emission times from its exact conditionl distribution.
Step 2: for each time interval between emissions, given the state of the chain at the end points, sample a chain from the exact conditional distribution of such chains.
Step 3: given this instance of the full underlying Markov chain, sample from the exact parameter distribution given the data.

Step 1 uses the forward–backward algorithm and step 3 is simple if we use conjugate priors.

Step 2 is based on an idea from Fearnhead and Meligkotsidou (2004) and involves creating a dominating Poisson process, the rate of which is the maximum (over states of the chain) of the sum of the intensity of the state change process and the photon emission process. These dominating events correspond (via thinning) to possible state changes or photon emissions. Over each interval between two photon emissions we simulate

(a) the number of dominating events from its true conditional distribution,
(b) the times of each dominating event, which are independent and uniform over the interval, and
(c) the state change (if any) at each dominating event; here we again use the forward–backward algorithm.

We repeated the authors' simulation from Section 2 (with vague but proper priors). Fig. 20 shows results for two parameters. Parameter k_{21} exhibited the worst mixing of any of the five parameters. The simulation took less than twice as much central processor unit time as an additive random-walk Metropolis scheme with block updating.

Omiros Papaspiliopoulos (*Lancaster University*)

The authors are to be congratulated for this interesting paper. I would like to make some remarks on the computational methodology and to draw the attention of the authors to important work on inference for hidden stochastic processes by giving some key references which are missing from the paper.

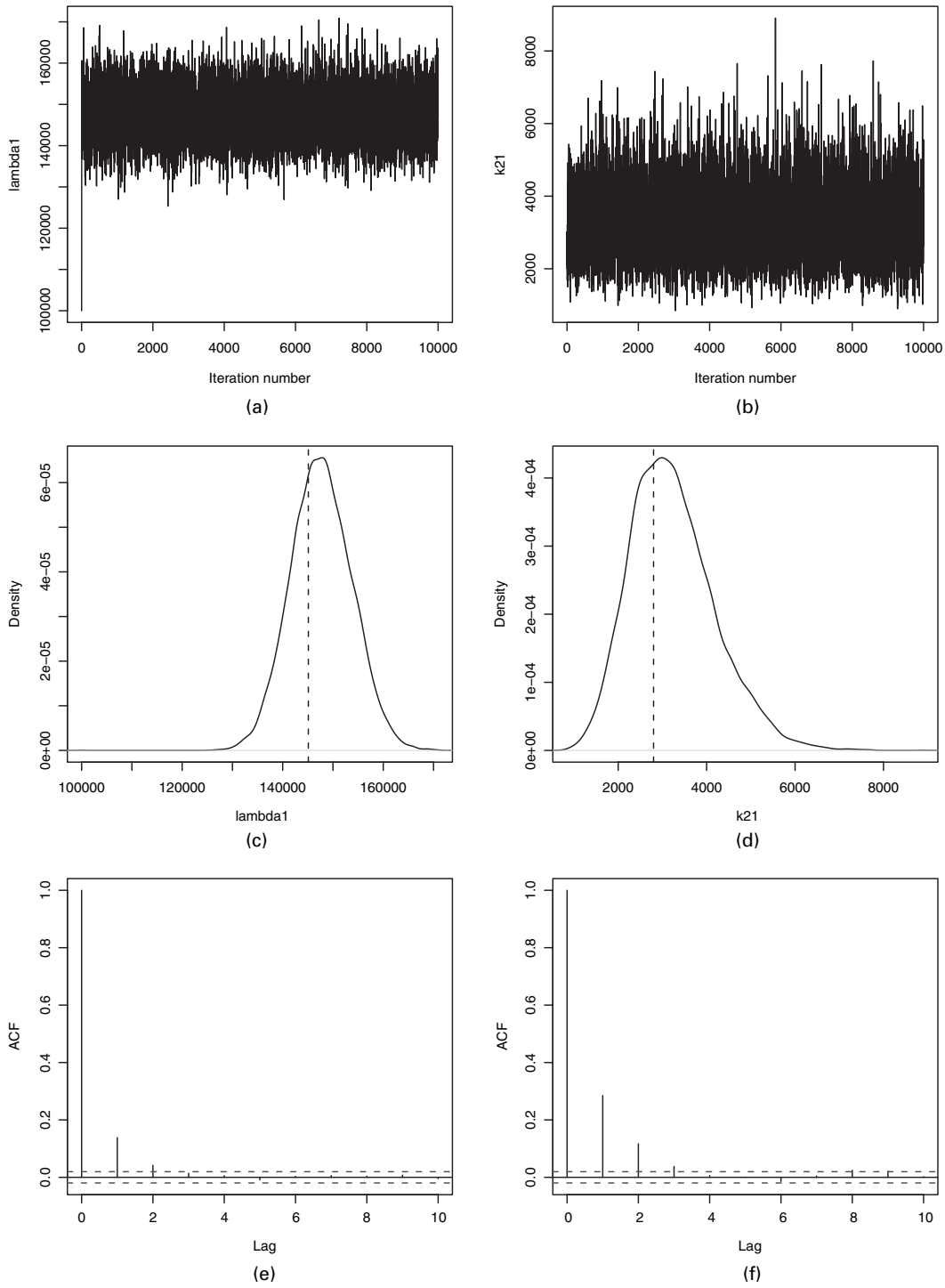

Fig. 20. (a), (b) Trace plots, (c), (d) density estimates (, true parameter values) and (e), (f) autocorrelation function for (a), (c), (e) λ_1 ($:= A_0/a$) and (b), (d), (f) k_{21}

The model that is described in Section 2 is a Markov modulated Poisson process, and there has been much work on fitting these models; see for example Asmussen (2000) and references therein. The likelihood in equation (2.3) is standard and can be derived (as well as equation (3.11)) using a much simpler (matrix analytic) argument than the approach that is adopted in Appendix A.

If $w_{xy} = w_z$ in equation (3.1) (or if we are willing to assume that the process lives in two dimensions) we could avoid augmenting the three Brownian motions and instead directly augment the $\alpha(t)$ process, which would be a transformation of the Bessel process which has known transition density. This might lead to much more efficient Markov chain Monte Carlo sampling, since there is a large amount of information about α, whereas the three processes are partially non-identifiable (all paths with the same distance from the origin have the same likelihood). Have the authors encountered mixing problems due to multiple modes in the path space? The time point by time point updating of hidden time series that is suggested in Section 3.2, although it results in fast updates, is known to lead to very slowly mixing Markov chain algorithms. Pitt and Shephard (l999), building on the results of Roberts and Sahu (1997), gave some analytic convergence results; see also the discussion in Jacquier *et al.* (1994) for empirical results from stochastic volatility models, and for example Elerian *et al.* (2001) for suggestions on efficient blocking schemes for updating diffusion paths. It is difficult to diagnose graphically on such high dimensions whether the mixing is good, but still I would be less convinced than the authors by Fig. 10, since in many cases the 'true' path lies well outside the range that is covered by the simulated paths.

Roberts and Stramer (2001) showed that the Gibbs sampler which updates the diffusion $\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t, \phi)\,\mathrm{d}W_t$ and the parameters θ has poor convergence; in fact if all the path is augmented the sampler is reducible. They suggested a non-centred reparameterization

$$X_t \to \tilde{X}_t = \eta(X_t; \phi),$$
$$\eta(x; \phi) = \int_0^x \frac{1}{\sigma(u; \phi)}\,\mathrm{d}u;$$

see also Roberts *et al.* (2004), Papaspiliopoulos *et al.* (2003) and Papaspiliopoulos (2003). There are close connections between reparameterizations and marginal augmentation methods, like those which were used in Section 4.3. In this context the former approach is more flexible since it can easily handle arbitrary diffusion functions and discretely observed diffusions, but it is interesting to know whether both approaches can be combined.

Martin Jacobsen and Michael Sørensen (*University of Copenhagen*)

We congratulate the authors on their interesting and stimulating paper and would like to indicate a rigorous derivation of their likelihood function in a more general setting (without any extra complication).

Suppose that a marked counting process has been observed in the time interval $[0,T]$. The data are the pairs of event times and marks $(t_1, \tau_1), \ldots, (t_n, \tau_n)$ $(0 < t_1 < \ldots < t_n < T)$. The intensity of the counting process is $\lambda_t = \lambda(M_{t-})$, where M is a continuous time Markov chain with state space $\{1, \ldots, d\}$, intensity matrix Q and stationary distribution $\pi_1, \ldots, \pi_d$. The probability density of a mark at time t conditional on the past of the marked counting process and on the Markov chain M is $f(\tau|t, M_{t-})$. The likelihood function of the marked counting process conditional on the Markov chain M is

$$L_c = \left\{ \prod_{i=1}^{n} \lambda_{t_i} f(\tau_i|t_i, M_{t_i-}) \right\} \exp\left(- \int_0^T \lambda_s\,\mathrm{d}s \right)$$

(with respect to a suitable dominating measure); see for example Bremaud (1981). Define iteratively

$$L_i(m) = E_M\{Z_i\, L_{i+1}(M_{t_i}) | M_{t_{i-1}} = m\}, \qquad i = 1, \ldots, n+1,$$

where E_M denotes expectation with respect to the distribution of M, $L_{n+2}(m) = 1$ and

$$Z_i = q_i(M_{t_i}) \exp\left\{ - \int_{t_{i-1}}^{t_i} \lambda(M_s)\,\mathrm{d}s \right\},$$

with $t_0 = 0, t_{n+1} = T$, $q_{n+1}(m) = 1$ and $q_i(m) = \lambda(m)\, f(\tau_i|t_i, m)$. Then, by iterating conditional expectations and using the Markov property, we obtain the unconditional likelihood

$$L = E_M(L_c) = E_M(L_1) = (\pi_1, \ldots, \pi_d)\mathbf{L}_1$$

where $\mathbf{L}_i = (L_i(1), \ldots, L_i(d))^{\mathrm{T}}$ (T denotes transposition). With $r_i(m) = q_i(m)\, L_{i+1}(m)$,

$$L_i(m) = E_M\left(\lim_{K\to\infty} \left[r_i(M_{t_i}) \exp\left\{ -\sum_{j=1}^{K} \lambda(M_{s_j})h \right\} \middle| M_{t_{i-1}} = m \right] \right)$$
$$= \lim_{K\to\infty} \left(E_M \left[r_i(M_{t_i}) \prod_{j=1}^{K} \exp\{-\lambda(M_{s_j})h\} \middle| M_{t_{i-1}} = m \right] \right)$$

where $s_j = t_{i-1} + jh$, $h = (t_i - t_{i-1})/K$, and where we have used the dominated convergence theorem. Using the same matrix calculations as the authors, we obtain that

$$L_i(m) = \lim_{K\to\infty} \left[\sum_{m_1,\ldots,m_K} r_i(m_K) \prod_{j=1}^{K} \exp\{-\lambda(m_j)h\}\, p(m_j|m_{j-1}) \right]$$
$$= \lim_{K\to\infty} \left[\sum_{m_1,\ldots,m_K} r_i(m_K) \prod_{j=1}^{K} \{\exp(Qh)\exp(-Hh)\}_{m_{j-1}m_j} \right]$$
$$= [\exp\{(Q-H)(t_i - t_{i-1})\} H D_i \mathbf{L}_{i+1}]_m,$$

for $i \leqslant n$, where $p(m_j|m_{j-1})$ is a transition probability for the Markov chain, $m_0 = m$, $H = \mathrm{diag}\{\lambda(1), \ldots, \lambda(d)\}$ and $D_i = \mathrm{diag}\{f(\tau_i|t_i, 1), \ldots, f(\tau_i|t_i, d)\}$. In conclusion,

$$\mathbf{L}_i = \exp\{(Q-H)(t_i - t_{i-1})\} H D_i \mathbf{L}_{i+1}$$

for $i \leqslant n$, and $\mathbf{L}_{n+1} = \exp\{(Q-H)(t_i - t_{i-1})\}\mathbf{e}$, where $\mathbf{e} = (1, \ldots 1)^{\mathrm{T}}$, so by iteration

$$L = (\pi_1, \ldots, \pi_d) \left[\prod_{i=1}^{n} \exp\{(Q-H)(t_i - t_{i-1})\} H D_i \right] \exp\{(Q-H)(T - t_n)\}\mathbf{e}.$$

It was assumed that the function $\lambda(m)$ is deterministic, but inspection of the proof shows that it holds also when the intensity is allowed to depend on the behaviour of the marked counting process before the previous event, whereas between events it depends additionally on the current value of M as above. The intensity can also depend on other sources of randomness, so the other models in the paper are covered also, but it is essential that between events the dependence on such extra randomness remains constant. Observation in the time interval $[t_1, t_n]$ is covered by taking $t_1 = 0$ and $T = t_n$.

Richard J. Boys (*University of Newcastle*)

The authors are to be congratulated on an impressive paper. Recent developments in biochemical experimentation now provide insight into the biological mechanisms within a single cell. Potentially, these techniques can give us a much fuller understanding of cellular biological processes than can be obtained by ensemble average experiments such as microarrays and polymerase chain reaction experiments. At Newcastle, Darren Wilkinson and I are developing Markov chain Monte Carlo methods for inferring stochastic kinetic rate constants in biochemical networks by using partially observed time course data on the numbers of molecules of chemical species within a single cell. This work (Boys *et al.*, 2004; Wilkinson *et al.*, 2004) also uses data augmentation, in our case to average over additional uncertainty due to the partial time course information.

Professor Steel's comments on model sensitivity, including the use of prior information, are interesting. It can be quite difficult to assess the influence of prior information, particularly when using diffuse priors in models with latent structures which are often only weakly identified by the data. In more straightforward scenarios, sensitivity to the prior specification can be assessed by looking at marginal likelihoods such as

$$\pi(y|\mu) \propto \hat{\pi}(\mu|y)/\pi(\mu),$$

where $\hat{\pi}$ is an empirical estimate of the marginal posterior density calculated from the Markov chain Monte Carlo output. In fact, this analysis uses diffuse priors and the model does not appear to suffer from identifiability problems and so the marginal posteriors are almost normalized marginal likelihoods. Turning to model sensitivity, one way of assessing the effect of the discretization that is used for their Ornstein–Uhlenbeck process would be to look at the sensitivity of inferences to models with a finer discretization, say by inserting time points $t_{i,1}, \ldots, t_{i,g}$ between each observed time point, and then using data augmentation techniques to integrate out the additional values $x(t_{i,j})$.

The authors comment that they have found no previous reference to their Bayes factor calculation (5.1). This is based on the posterior mean of the likelihood ratio for the hypothesis $H_0 : \zeta = 0$. In Aitkin *et al.* (2005), we study how this posterior distribution can be used to assess such hypotheses. Instead of using just the mean value, the strength of evidence against the hypothesis is gauged by using the full distribution. It would be interesting to hear whether, in this analysis, the posterior distribution of the likelihood ratio is particularly skewed and to have some measure of variability and some quantiles.

The **authors** replied later, in writing, as follows.

We thank the discussants for their thoughtful and constructive comments regarding our paper. We shall first discuss the issues that were most commonly raised and then make (necessarily) brief replies to questions which were asked by particular discussants.

Bayesian inference, prior specification and model sensitivity

To us, a main advantage of Bayesian inference for stochastic models is that all quantities, be they observed data, unobserved latent processes or parameters of interest, are given the same footing under a coherent probabilistic framework. The Bayesian approach is thus conceptually straightforward for inferring stochastic models with latent processes and has been applied in the finance literature to study partially observed diffusion processes, stochastic volatility models and term structure models, as pointed out by Steel and Papaspiliopoulos. Although all these models fall under the general umbrella of Markov or hidden Markov models, our model resembles the more classical hidden Markov models rather than the partially observed diffusion processes.

An integral part of the Bayesian analysis is to understand the effect of prior specifications. In analysing most natural science experiments, the prior distribution should be as diffuse as possible so that it is the scientist's data that lead to the main inferential conclusion. A 'good' experiment usually refers to the one that provides information easily overwhelming the prior distribution. In our analysis, we use 'flat' priors for all parameters, where the flatness simply implies a uniform distribution over a finite but relatively large region, which is determined by the domain scientific knowledge. Although our uniform prior is not transformation invariant, our sensitivity analysis showed that the prior influence is indeed very minimal. For the simulation study in Section 5.1, we generated 50 data sets, each containing hundreds to thousands of data pairs. For a wide range of prior distributions, there were essentially no differences in the posterior distributions.

The excellent point raised by Sherlock regarding the singularity of the likelihood function of our two-state model is analogous to that in a typical Gaussian mixture model, and can be avoided both practically and theoretically by using a proper prior.

Boys suggested an interesting way to test model sensitivity especially for the diffusive Ornstein–Uhlenbeck model. We only demonstrated in the paper that the simple two-state model is not sufficient to describe the DNA hairpin kinetics. The model sensitivity test that is suggested by Boys could be useful for further testing the validity of the diffusive model.

Likelihood calculation

We thank Jacobsen and Sørensen for providing a rigorous derivation of the likelihood. To us the discretization and matrix approach is quite straightforward and intuitive, as it provides a natural way to extend techniques for handling discrete time processes to continuous time ones. Another benefit of the discretization approach is that it can provide answers or insights to both analytical and computational issues: when the analytical form of the likelihood function exists, one can obtain it by letting the discretization length shrink to zero (as shown in the paper), whereas, if there is no analytical solution, this approach corresponds to the numerical 'Euler method' that is used in the econometrics literature and is the basis for the recent data augmentation approach for inferring partially observed diffusion processes (Elerian *et al.*, 2001).

Markov chain Monte Carlo issues

There are a few suggestions regarding the use of 'block' moves in augmenting the latent processes. We agree that this can be very helpful in the general state space model, especially when we can easily sample from the joint conditional distribution for the whole block of variables (Carter and Kohn, 1996). When such conditional block distributions cannot be handled at ease, Shephard and Pitt (1997) suggested some partial updating approaches, such as additive moves, which can improve computational efficiency. But in our experience these partial updates have a much smaller effect compared with the true block move. In Liu and Sabatti (2000), we introduced a multilevel partial block update approach

for partially observed diffusion processes, which further improves Shephard and Pitt's method, but the improvement is still not sufficiently large to merit additional programming effort for our current problem.

We do agree, however, with Papaspiliopoulos that our three-dimensional Brownian paths are unidentifiable—the posterior distribution is invariant under a rotation of the Brownian path around the z-axis. As a consequence, our single-site update sampler cannot explore the whole path space sufficiently. The reason that our method still seems to work well is that the likelihood is dependent on only $\alpha(t)$, and whether the sampler mixes well in the angular space is not important. On the basis of the invariance argument, we can also introduce a new move to speed up the convergence: $B'(t) = R \circ B(t)$, where R is a rotation-like transformation that leaves $\alpha(t)$ unchanged. Such a move should be easily accepted and should work much better than the block move.

We also agree with Steel and Papaspiliopoulos that reparameterization can drastically improve the convergence. We feel that the group move is perhaps more intuitive and easier to use than most reparameterizations because

(a) it can achieve almost all the effects of a reparameterization without actually changing the original distribution and
(b) we can do different types of group moves in one Markov chain Monte Carlo iteration easily, but it is much more involved to do several types of reparameterizations simultaneously.

However, some reparameterizations, such as that in Roberts and Stramer (2001), cannot be easily formulated as group moves, which suggests that there might be a useful and more general mathematical framework than the generalized Gibbs structure (Liu and Sabatti, 2000) to encompass non-group transformations.

Experimental background

We thank Hawkes for pointing out the connection between the current work and the literature on modelling and inferring ion channel behaviour, and for asking clarifying questions regarding the photon arrival time and delay time. In our experiments, since the laser pulse rate is much faster than the rate at which $\gamma(t)$ changes, as Hawkes rightly observed, the photon arrival rate is essentially determined by the underlying state of the molecule. The power of the laser affects the constant A_0. The photon delay time measures the length of time that it takes the molecule to emit a photon from the moment it is excited; this length of time is essentially independent of the laser pulse rate and is determined by the state of the molecule. In the experiment, the molecule is put in a laser focal volume. The focal volume is ellipsoidal, symmetric in the x- and y-axes but asymmetric in the z-axis, resulting in $w_{xy} = 310\,\text{nm}$ and $w_z = 1760\,\text{nm}$ in equation (3.1). Consequently, one Bessel process is not sufficient to augment the $\alpha(t)$ process, as Papaspiliopoulos suggested.

Finally, we thank all the discussants for their insightful contributions. We hope that the reader will enjoy reading these comments and thinking about various scientific, statistical and computational issues raised as much as we did.

References in the discussion

Aitkin, M., Boys, R. J. and Chadwick, T. J. (2005) Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statist. Comput.*, to be published.

Asmussen, S. (2000) Matrix-analytic models and their analysis. *Scand. J. Statist.*, **27**, 193–226.

Barndorff-Nielsen, O. E. and Shephard, N. (2001) Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics (with discussion). *J. R. Statist. Soc.* B, **63**, 167–241.

Boys, R. J., Wilkinson, D. J. and Kirkwood, T. B. L. (2004) Bayesian inference for a discretely observed stochastic kinetic model. *Research Report STA04.5*. Newcastle University, Newcastle upon Tyne.

Bremaud, P. (1981) *Point Processes and Queues: Martingale Dynamics*. New York: Springer.

Burzomato, V., Beato, M., Groot-Kormelink, P. J., Colquhoun, D. and Sivilotti, L. G. (2004) Single-channel behaviour of heteromeric $\alpha 1\beta$ glycine receptors: an attempt to detect a conformational change before the channel opens. *J. Neurosci.*, **24**, 10924–10940.

Carter, C. and Kohn, R. (1996) Markov chain Monte Carlo for conditionally Gaussian state space models. *Biometrika*, **83**, 589–601.

Elerian, O., Chib, S. and Shephard, N. (2001) Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, **69**, 959–993.

Fearnhead, P. and Meligkotsidou, L. (2004) Exact filtering for partially observed continuous time models. *J. R. Statist. Soc.* B, **66**, 771–789.

Fearnhead, P. and Sherlock, C. (2004) Bayesian inference for Markov modulated Poisson processes. To be published.
Fredkin, B. R. and Rice, J. A. (1992) Bayesian restoration of single-channel patch clamp recordings. *Biometrics*, **48**, 427–448.
Griffin, J. E. and Steel, M. F. J. (2003) Inference with non-Gaussian Ornstein-Uhlenbeck processes for stochastic volatility. *Mimeo*. University of Warwick, Coventry.
Hawkes, A. G. (2003) Stochastic modelling of single ion channels. In *Computational Neuroscience: a Comprehensive Approach* (ed. J. Feng), pp. 131–157. Boca Raton: Chapman and Hall–CRC.
Jacquier, E., Polson, N. G. and Rossi, P. E. (1994) Bayesian analysis of stochastic volatility models (with discussion). *J. Bus. Econ. Statist.*, **12**, 371–417.
Liu, J. S. and Sabatti, C. (2000) Generalized Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*, **87**, 353–369.
Liu, J. S., Wong, W. H. and Kong, A. (1994) Covariance structure of the Gibbs sampler with applications to the comparison of estimators and augmentation schemes. *Biometrika*, **81**, 27–40.
Neher, E., and Sakmann B. (1976) Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature*, **260**, 799–802.
Newton, M. A. and Raftery, A. E. (1994) Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. R. Statist. Soc.* B, **56**, 3–48.
Papaspiliopoulos, O. (2003) Non-centered parametrisations for hierarchical models and data augmentation. *PhD Thesis*. Department of Mathematics and Statistics, Lancaster University, Lancaster.
Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2003) Non-centered parameterizations for hierarchical models and data augmentation (with discussion). In *Bayesian Statistics 7* (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), pp. 307–326. New York: Oxford University Press.
Pitt, M. K. and Shephard, N. (1999) Analytic convergence rates and parameterization issues for the Gibbs sampler applied to state space models. *J. Time Ser. Anal.*, **20**, 63–85.
Roberts, G. O., Papaspiliopoulos, O. and Dellaportas, P. (2004) Bayesian inference for non-Gaussian Ornstein–Uhlenbeck stochastic volatility processes. *J. R. Statist. Soc.* B, **66**, 369–393.
Roberts, G. O. and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. R. Statist. Soc.* B, **59**, 291–317.
Roberts, G. O. and Stramer, O. (2001) On inference for partially observed non-linear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, **88**, 603–621.
Shephard, N. and Pitt, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653–667.
Wilkinson, D. J., Boys, R. J. and Kirkwood, T. B. L. (2004) Bayesian inference for general stochastic kinetic models using discretely observed data. *Research Report STA04.6*. Newcastle University, Newcastle upon Tyne.