Bayesian Inference of Accurate Population Sizes and FRET Efficiencies from Single Diffusing Biomolecules

Rebecca R. Murphy,[*,†] George Danezis,[*,‡] Mathew H. Horrocks,[*,†] Sophie E. Jackson,[*,†] and David Klenerman[*,†]

Department of Chemistry, University of Cambridge, U.K., and Dept. of Computer Science, University College London, U.K.

E-mail: rrm33@cam.ac.uk; g.danezis@ucl.ac.uk; mhh30@cam.ac.uk; sej13@cam.ac.uk; dk10012@cam.ac.uk

Abstract

It is of significant biophysical interest to obtain accurate intramolecular distance information and population sizes from single-molecule Förster resonance energy transfer (smFRET) data obtained from biomolecules in solution. Experimental methods of increasing cost and complexity are being developed to improve the accuracy and precision of data collection. However, the analysis of smFRET datasets currently relies on simplistic and often arbitrary methods, for the selection and denoising of fluorescent bursts. Although these methods are satisfactory for the analysis of simple, low-noise systems with intermediate FRET efficiencies, they display systematic inaccuracies when applied to more complex systems. We have developed an inference method

[*]To whom correspondence should be addressed
[†]Department of Chemistry, University of Cambridge, U.K.
[‡]Dept. of Computer Science, University College London, U.K.

for the analysis of smFRET data from solution studies, based on rigorous model-based Bayesian techniques. We implement a Monte-Carlo Markov Chain (MCMC) based algorithm that simultaneously estimates population sizes and intramolecular distance information directly from a raw smFRET dataset, with no intermediate event selection and denoising steps. Here, we present both our parametric model of the smFRET process and the algorithm developed for data analysis. We test the algorithm using a combination of simulated datasets and data from dual-labelled DNA molecules. We demonstrate that our model-based method systematically outperforms threshold-based techniques in accurately inferring both population sizes and intramolecular distances.

Introduction

Förster Resonance Energy Transfer (FRET) is a powerful technique for studying biological systems at the level of single molecules. Since the first demonstration that FRET could quantify the distance between two fluorescent dyes[1], single-molecule FRET (smFRET) became a popular tool to investigate the structure and dynamics of diffusing biomolecules[2–4]. FRET is a non-radiative energy transfer from a donor (D) to an acceptor fluorophore (A), where the efficiency of energy transfer (the FRET Efficiency, E) depends on their separation, r:

$$E = \frac{1}{1 + (\frac{r}{R_0})^6}, \tag{1}$$

where R_0 is the distance for which the transfer efficiency is 50%.

In a smFRET experiment, photons emitted from the donor and acceptor fluorophores are collected in a continuous stream and time-binned on a time-scale comparable with the average dwell-time of a molecule diffusing through the confocal volume. Time-bins containing photons from a fluorescent burst are identified by applying a threshold[5–7] and selected bursts

are denoised[8]. FRET efficiencies are calculated for the denoised bins using:

$$E = \frac{n_A}{n_A + \gamma \cdot n_D} \tag{2}$$

for n_A and n_D photons in the acceptor and donor channels respectively and γ an experimentally determined instrument-dependent correction factor. Histograms constructed from the calculated FRET efficiencies are fitted with Gaussian distributions to identify fluorescent populations[1].

Determining intramolecular distance information and population sizes from smFRET experiments however remains challenging[8]. Using smFRET data to constrain molecular dynamics simulations can provide structural information.[9,10] However, incomplete sample labelling, photophysical artifacts, unequal photon detection and the stochastic nature of diffusion through the confocal volume[8], as well as linker dynamics[11] hamper development of quantitative smFRET techniques. To overcome these challenges, linear flow has been used to reduce heterogeneity in confocal dwell-time and diffusion pathway[12,13], whilst methods to determine correction factors[14]; development of alternating-laser excitation[15–18] and multi-parameter fluorescence detection[19] as well as more sophisticated burst-selection algorithms[8,20] allow more accurate identification of fluorescent bursts.

However, despite their increasing cost and complexity, these techniques continue to use simplistic methods to identify and denoise fluorescent events. We show using simulated data that thresholding techniques can be biased[5,6]. They also assume that fluorescent bursts are clearly distinct from noise and can be separated using an arbitrary cut-off (Fig. 1 B). However, data from actual smFRET experiments (Fig. 1 C-D) are not linearly separable, exhibiting significant overlap between the number of noise photons and the number of photons emitted by a fluorescent molecule, meaning that no threshold can perfectly separate photons of interest from noise. Consequently, threshold choice is subjective and can significantly influence analytical outcomes.

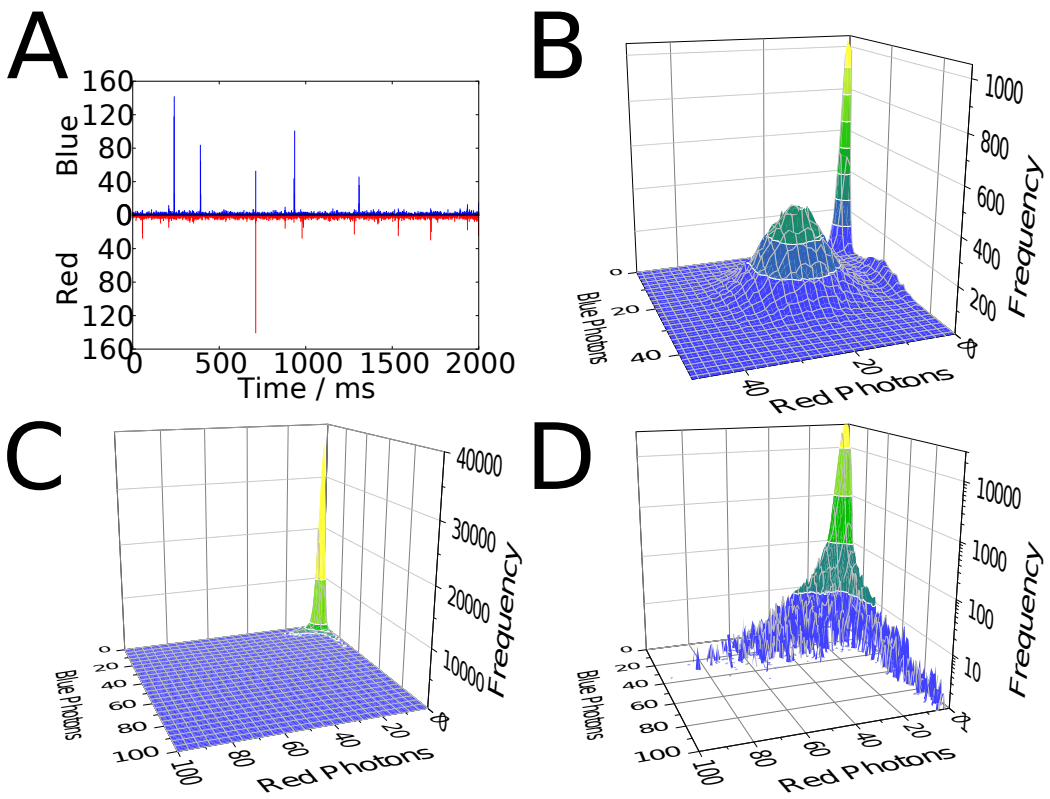

Figure 1: A typical smFRET dataset. (A) Snapshot of raw smFRET data from a high-FRET dual-labelled DNA. (B-D) show three-dimensional histograms of raw photon counts from smFRET datasets. (B) An idealised, simulated smFRET dataset, with signal and noise well separated, for which thresholding would be a suitable technique for event selection. (C) A real smFRET dataset. (D) The same dataset shown in (C) plotted using a logarithmic scale to show details of fluorescent bursts.

Besides burst selection, there is a need to denoise selected bursts. Typically, this involves subtracting an averaged, non-integral value from all bursts.[8] This frequently results in fractional or negative photon counts, negative FRET efficiencies and other analysis artifacts. Stochastic denoising methods[21–24], based on poisson statistics, are now used, but still assume that thresholding provides unbiased burst selection[21].

We address these issues by using model-based Bayesian inference to analyse smFRET data. Bayesian inference is a probabilistic method[25] that uses conditional probabilities based on Bayes' Theorem[26] to assess the likelihood that a series of observations were generated by a given model[27]. Analysis techniques based on Bayesian statistics are well established for analysis of smFRET data collected from immobilised molecules[28–33]. Bayesian methods have also been applied to single particle tracking [34]; analysis of diffusional trajectories[35,36]; fluorescence correlation spectroscopy[37–40] and fluorescence lifetime data[41,42]. Attempts have been made to apply Bayesian statistics to diffusion-based smFRET experiments[43,44]. However, these methods are computationally intractable[45] or apply only to removal of shot-noise from selected burst, so assume access to idealised simulated[44], or pre-selected[43] and denoised fluorescent traces[41].

Here we present a simple physical model of the FRET excitation/emission process within a diffusion experiment, incorporating both FRET based emission and background fluorescence events. We use the model as part of a custom-built inference algorithm based on MCMC Metropolis sampling[46] to infer values for all relevant physical parameters, including intramolecular distances and population sizes, conditioned on a smFRET dataset. We simultaneously infer all parameters directly from the raw time-binned data, with no intermediate burst selection or denoising steps. The model is summarized schematically in Fig. 2 and described in detail below. The mathematical model describes a Bayesian belief network and is shown as a directed acyclic graph in standard plate notation (Supplementary material Fig. S-1). We demonstrate this technique's effectiveness using realistic simulated datasets. We then analyse real smFRET data, generated from single populations and mixtures of

dual-labelled DNA molecules, showing that our technique can infer physically appropriate and experimentally informative parameters with high confidence across a wide range of conditions. In particular, we accurately infer absolute populations and FRET efficiencies of a mixture of two fluorescent species, where thresholding-based techniques fail.

Theory

A Physical Model of a smFRET Experiment

Thus far, analysis of smFRET data has not separated a defined model of the physical process from data analysis. As a consequence, implicit assumptions about the physical model may be reproduced during analysis[7]. Our key innovation is development of a model-based Bayesian analysis. This analysis uses a parametric model of the physical emission process. We then infer values for these parameters given a specific dataset, to learn information about intramolecular distances and population sizes for different fluorescent species. The model of photon emission in the presence of both dyes is inspired by the traditional model of FRET efficiency (Eq. 2). However, we model the energy transfer as altering the underlying rates of dye photon emission; whereas traditional techniques use the ratio of donor and acceptor photons observed.

In a basic smFRET experiment, fluorescently-labelled molecules in dilute solution diffuse freely through a laser beam focused with a high aperture objective onto a diffraction-limited focal point.[47] When a molecule diffuses into the confocal volume, the laser excites the donor fluorophore and photons are emitted. Emitted photons are collected through the objective and separated by a dichroic mirror into donor and acceptor photons for collection and analysis (Fig.S-1 A).

These experiments yield bursts of donor and acceptor fluorescence, caused by diffusion of a labelled molecule through the excitation volume, against a background of low to zero fluorescence detection. Although accurate arrival times can be recorded[48], raw data is often

collected as two synchronised streams of time-binned photons, corresponding to detected photons with wavelengths in the donor and acceptor emission regions (Fig. 1 A). The majority of bins ($> 95\%$) contain only background noise; the rest contain both background noise and photons from fluorescent bursts (Fig. 1 C-D).

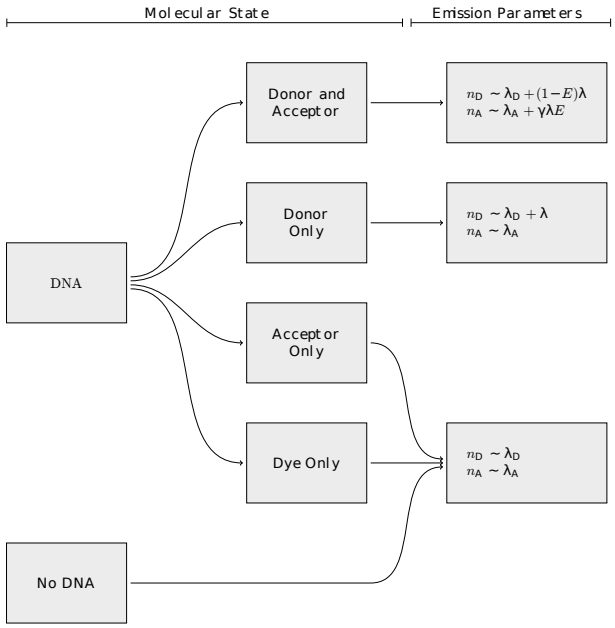

Figure 2: Flow diagram illustrating the generative model for a single FRET population. The molecular state is the underlying state of the current observation; the emission parameters are the Poisson parameters that result in observable photon emission - these are shown for both donor and acceptor channels.

We model a smFRET dataset as a sequence of pairs of measurements (f_D, f_A) of the number of photons observed in the donor and acceptor channels. Each pair of measurements is treated as an independent and identically distributed sample from a set of random variables describing the dataset. Each pair of data-points (f_D, f_A) in the data stream is the sum of noise photons and possibly some photons from a fluorescent event.

The number of noise photons is drawn from a Poisson distribution with rate parameter λ_D for the donor channel, and rate λ_A for the acceptor channel. The probability of observing

n_D noise photons in the donor channel and n_A in the acceptor channel is:

$$n_D \sim \text{Poisson}(n_D; \lambda_D) = \frac{\lambda_D^{n_D}}{n_D!} e^{-\lambda_D} \qquad n_A \sim \text{Poisson}(n_A; \lambda_A) = \frac{\lambda_A^{n_A}}{n_A!} e^{-\lambda_A} \tag{3}$$

In addition to noise, each observation may contain photons from one or more fluorescent molecules. For a dataset with a single fluorescent population, the number of molecules present in the excitation volume, n_{prot} follows a Poisson distribution with rate parameter λ_{prot}:

$$n_{\text{prot}} \sim \text{Poisson}(n_{\text{prot}}; \lambda_{\text{prot}}) = \frac{\lambda_{\text{prot}}^{n_{\text{prot}}}}{n_{\text{prot}}!} e^{-\lambda_{\text{prot}}} \tag{4}$$

The probability of seeing any molecule is typically low: λ_{prot} is small and $n_{\text{prot}} = 0$ for the majority of time-bins. However, multiple-occupancy events may occur.

We extend this model to describe two or more fluorescent species with different FRET efficiencies and population sizes. Here, the number of molecules of each species is determined independently, with n_{prot1} and n_{prot2}, the numbers of molecules observed of species 1 and 2 respectively, given by:

$$n_{\text{prot1}} \sim \text{Poisson}(n_{\text{prot1}}; \lambda_{\text{prot1}}) = \frac{\lambda_{\text{prot1}}^{n_{\text{prot1}}}}{n_{\text{prot1}}!} e^{-\lambda_{\text{prot1}}} \tag{5}$$

and

$$n_{\text{prot2}} \sim \text{Poisson}(n_{\text{prot2}}; \lambda_{\text{prot2}}) = \frac{\lambda_{\text{prot2}}^{n_{\text{prot2}}}}{n_{\text{prot2}}!} e^{-\lambda_{\text{prot2}}} \tag{6}$$

As before, most bins contain no fluorescent molecules ($n_{\text{prot2}} = n_{\text{prot2}} = 0$), but multiple occupancy can be modelled when $n_{\text{prot2}} + n_{\text{prot2}} > 1$.

Each molecule present may be in one of four labelling states: unlabelled, donor-only, acceptor-only or dual-labelled (Fig. S-1 B). We model the presence of donor and acceptor dyes as independent events with respective probabilities p_D and p_A. Thus, the molecule is unlabelled with probability $(1 - p_D)(1 - p_A)$; both dyes are present with probability $p_D p_A$; and only the acceptor or only the donor dye with probability $p_A(1 - p_D)$ and $(1 - p_A)p_D$

respectively. For multiple fluorescent populations, we assume that all species share the same labelling probabilities, p_D and p_A.

An unlabelled or acceptor-only labelled molecule is not excited, so only background noise is observed, thus $f_D = n_D$ and $f_A = n_A$.

When a donor dye is present, excitation potentially results in emission. This is modelled in two stages: first, a rate of donor emission, λ, is determined for the specific molecule as a random sample from a gamma distribution with shape parameter k_D and mean λ_B (Eq. 7). This captures the variation in the number of photons emitted by a molecule as a result of the diffusion path taken through the confocal volume and the effect of donor photobleaching partway through an observation.

$$\lambda \sim \text{Gamma}(\lambda; k_D, \theta) = \frac{1}{\Gamma(k_D)\theta^{k_D}} \lambda^{(k_D-1)} e^{-\frac{\lambda}{\theta}} \quad \text{for } \theta = \lambda_B / k_D \tag{7}$$

where Γ is the Gamma function.

As the confocal volume is fixed and emission from the dye is a fundamental property of the dye-laser interaction, the same k_D and λ_B are used for all fluorescent populations.

If only the donor dye is present, additional photons are observed in the donor channel only. These are drawn from a Poisson distribution with rate parameter λ, where λ is determined uniquely for each molecule using Eq. 7. The number of additional photons is then c_D:

$$c_D \sim \text{Poisson}(c_D; \lambda). \tag{8}$$

the total observed photons in the donor channel, f_D is then the sum $n_D + c_D$; in the acceptor channel only noise photons, n_A, are observed.

The interesting case is when both dyes are present. In this case, some of the excitation energy is transferred to the acceptor dye, resulting in emission of acceptor photons and attenuation of donor emission. Emission by both donor and acceptor dyes is modelled by drawing photons from Poisson distributions. In a single population dataset, the rate of donor

photon emission is now $\lambda \cdot (1 - E)$, whereas the acceptor rate of photon emission is $\lambda \cdot \gamma \cdot E$. Here, E is the efficiency of energy transfer (Eq. 1); λ is the unattenuated rate of donor photon emission associated with the observed molecule; and γ is an instrumental correction factor (Eq. 2). The additional photons in each channel, c_D and c_A are thus distributed as:

$$c_D \sim \text{Poisson}(c_D; \lambda(1 - E)) \qquad \text{and} \qquad c_A \sim \text{Poisson}(c_A; \gamma\lambda E) \tag{9}$$

The total number of photons in the donor and acceptor channels are thus $f_D = n_D + c_D$ and $f_A = n_A + c_A$ respectively.

For two populations, molecules from different populations exhibit different FRET efficiencies: respectively E_1 for the first population and E_2 for the second. This gives donor and acceptor emission rates, respectively c_{D1} and c_{A1}, from the first fluorescent population to be:

$$c_{D1} \sim \text{Poisson}(c_{D1}; \lambda(1 - E_1)) \qquad \text{and} \qquad c_{A1} \sim \text{Poisson}(c_{A1}; \gamma\lambda E_1) \tag{10}$$

Similarly, for the second population, c_{D2} and c_{A2}, are given by

$$c_{D2} \sim \text{Poisson}(c_{D2}; \lambda(1 - E_2)) \qquad \text{and} \qquad c_{A2} \sim \text{Poisson}(c_{A2}; \gamma\lambda E_2) \tag{11}$$

For simplicity, leakage and direct excitation are not currently considered, either for the single population case, or for multiple populations. However, these can be added to the model without introducing further complexity. Bleaching of the acceptor fluorophore partway through a bin is also not considered.

The total number of photons is then given by the sum of photons from all fluorescent species currently present, and any noise photons: $f_D = c_{D1} + c_{D2} + n_D$ for the donor and $f_A = c_{A1} + c_{A2} + n_A$ for the acceptor channel.

This process is then repeated for each time-bin in a dataset. This gives two data streams of integer photon counts – corresponding to the donor and acceptor channels in a FRET

experiment – representing background noise alone or a combination of noise and excitation events.

This model can be used, with appropriate parameters (see Table 1), to generate synthetic data. Comparison of these synthetic photon streams with experimental data reveals an excellent replication of all aspects of the experimental data (see Supplementary Material, Fig. S-4), suggesting that, despite its many simplifications, such as the neglect of direct excitation and leakage effects, this is an extremely good model of the FRET process.

Inference of Model Parameters

Our key innovation is the use of Bayesian model-based multi-variate statistical methods to infer the model parameters of the FRET experiment. Given the generative model of the physical process described in the previous section, we use the calculus of probabilities and Bayes' theorem to derive the joint distribution over all model parameters to be estimated.

Estimatuing the parameters of a complex model given some experimental observations is a typical inference problem. In a smFRET experiment, we want to determine the concentrations of the fluorescently labelled species and their respective inter-dye distances given some experimental data. We might also like to know other associated parameters, such as the rate of noise in each channel and the average brightness of fluorescent events. These values are described explicitly as parameters in our generative model. However, due to noise or a small amount of data, as well as the co-dependence of all observations on all parameters, it is difficult to determine the values of these parameters directly from observations. Consequently, a different strategy must be applied and probabilistic inference provides a solution.

Using probability theory this inference problem is expressed as determining the conditional probability distribution over the parameters given the observations, namely $\Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E|(f_I$ for a smFRET dataset with n time-bins. Given a generative model of the experiment that describes the probability of generating certain observations given known parameters, namely Pr[Obs.|Par.], we can apply Bayes' theorem to derive the required distribution over param-

eters:

$$\Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E|(f_D, f_A)_n] = \\ = \frac{\Pr[(f_D, f_A)_n|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E] \cdot \Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E]}{\sum_{\forall \lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E} \Pr[(f_D, f_A)_n|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E] \cdot \Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E]}. \quad (12)$$

The term $\Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E]$ encodes prior information about the parameters, whilst the denominator, $\sum_{\forall \lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E} \Pr[(f_D, f_A)_n|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E] \cdot \Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E]$ is a normalizing factor over all parameter space. Exact evaluation of this expression is often impossible because it is hard to derive an analytical expression for this denominator, or even compute it numerically. Consequently, exact evaluation of Eq. 12 and exact determination of the posterior distributionm, $\Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E|(f_D, f_A)_n]$ is not possible.

However, to estimate the distribution of values taken by the parameters of interest, it is not necessary to evaluate Eq. 12, exactly. It is sufficient to draw parameter samples distributed proportionally to the posterior distribution, $\Pr[(f_D, f_A)_n|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E] \cdot \Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E]$.[46] The mean, variance and quantiles of these samples can be used to estimate the required parameters. Consequently, we can determine the parameter distribution (dye-dye distance, concentration, etc) most likely to have generated a particular dataset by using a Monte Carlo method to sample many possible parameter values and calculating the probability that these parameters generated our data.

The Metropolis algorithm,[49 46] is an MCMC algorithm that can be used to sample parameter space for candidate parameter values. It defines the structure of a Markov chain that has as its stationary probability the posterior probability over the model parameters, here $\Pr[(f_D, f_A)_n|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E] \cdot \Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E]$. By performing long random walks over that chain we can generate independent samples of the parameters distributed according to Eq. 12. Metropolis[50] provides an introduction to the algorithm; our implementation is described in the Supplementary Material.

Methods

Generation of Simulated Data

Simulated datasets were generated using the model described above, using code written in Python. Code is available online (`https://bitbucket.org/rebecca_roisin/fret-inference`).

smFRET measurements

Single-molecule data were collected using a custom built system, as described previously[51]. Details of the instrumentation and DNA duplex preparation are found in the Supplementary Material. Data were collected for 30 minutes at room temperature using a 1 ms bin time, in frames of 10000 bins.

Analysis of single-molecule FRET data

Thresholding-based data-analysis was carried out in the standard manner[5] (see Supplementary Material). For the inference process, data were fitted in a single step. Raw data, prior to any denoising or event selection steps, were analysed using the Metropolis sampling process described above. Sampling occurred in two steps. First, two approximate samples were generated, with a burn-in of 3000 iterations and 1000 iterations between samples. Then, 100 further samples were made, with a burn-in of 1000 iterations and 100 iterations between samples. For all analyses, the initial parameters shown in Table 1 were used. The outcome of the inference was not sensitive to initial conditions (see Fig. S-3).

Table 1: Parameters used in the generation of synthetic data.

Parameter	λ_{prot}	λ_D	λ_A	p_D	p_A	k_D	λ_B	$R_0/\mathring{A}$	γ
Value	0.06	1.0	1.0	0.6	0.8	1.0	20.0	56.0	1.0

Results

Validation using Simulated Data

Single fluorescent species

To validate the inference method, we used the forward model to generate realistic simulated datasets with known parameters. We then analysed these datasets using the inference method to see how accurately the model parameters could be inferred. We varied several aspects of the simulated data, including mean dye-dye separation (altering E), dataset size, mean noise level and rate of observation of labelled molecules. Unless otherwise stated, parameters used in data generation are those shown in Table. 1. The results are summarised in Fig. 3. Fig. 3 (A) and (B) show the FRET efficiencies inferred for datasets with dye-dye distances across the spectrum of FRET efficiencies. From Fig. 3 (A) it can be seen that the inference method correctly reproduces the expected sigmoidal curve of FRET efficiency against dye separation, whilst Fig. 3 (B) shows a linear relationship between actual and inferred FRET efficiencies, with tight confidence intervals, demonstrating that the inference method exactly reproduces the values used to generate the simulated data. Similarly, Fig. 3 (D) shows that the inference method also correctly infers the rate at which fluorescent events are observed (analogous to concentration), demonstrating a linear relationship between the rate used for dataset generation and the rate inferred. The inferred value remains accurate even for very high and very low rates, showing that the method is robust over a wide range of conditions.

Fig. 3 (C) shows the variation in the size of the confidence interval with the number of time-bins in a dataset. Even for a small dataset of only 1000 time-bins, the inferred mean FRET efficiency was inferred exactly correctly (actual value 0.66, inferred mean 0.66), although the 98% confidence interval (CI98) is very wide (CI98: 056 - 0.75), as there are insufficient data to allow precise estimation. Making the dataset larger significantly reduces the size of the confidence interval, with very narrow intervals for datasets of 100000 bins or

larger (mean: 0.66, CI98: 0.65 - 0.67). Assuming a bin-time of 1 ms, a typical experimental dataset (10 - 20 minutes of data), would include six - 12 million bins. Consequently, it is a significant achievement of the inference method that it makes extremely accurate estimates of the FRET efficiency using only 100000 bins, corresponding to less than two minutes of data.

Fig. 3 (E) and (F) show the effect of noise and observation rate on the size of the confidence interval for the inferred FRET efficiency. As expected, both increased noise and a lower rate (lower concentration of fluorescent molecules) result in a wider confidence interval, reducing how accurately we can infer E. However even when a very low rate or very high noise is used, the size of the error remains small ($\pm$ 0.03 and $\pm$ 0.01 respectively), meaning that the inference method still gives accurate values.

These results are clear validation that the inference method works reliably across a wide range of datasets. However a more important question is whether inference can outperform thresholding. To determine this, we analysed a series of simulated datasets using AND and SUM thresholding and using inference. The results, summarized in Fig. 4, show that inference and thresholding are equally good at determining FRET efficiency, but that inference far outperforms thresholding in determining population sizes.

Fig. 4 (A) (top panel) shows the FRET efficiencies estimated by inference and by thresholding. All three techniques reproduce the characteristic sigmoidal relationship between dye-dye distance and E. However, AND thresholding (open black circles) overestimates E for the largest distances and was unable to be used for the two smallest separation intervals as too few events were selected to allow histogram construction. These discrepancies are however relatively minor and we see that thresholding performs similarly to inference in determining E.

A different story is told however, when population sizes are considered. Fig. 4 (A) (middle and bottom panels) and (B) compare the ability of thresholding and inference to accurately determine population sizes. The middle and bottom panels of Fig. 4 (A) show the relationship

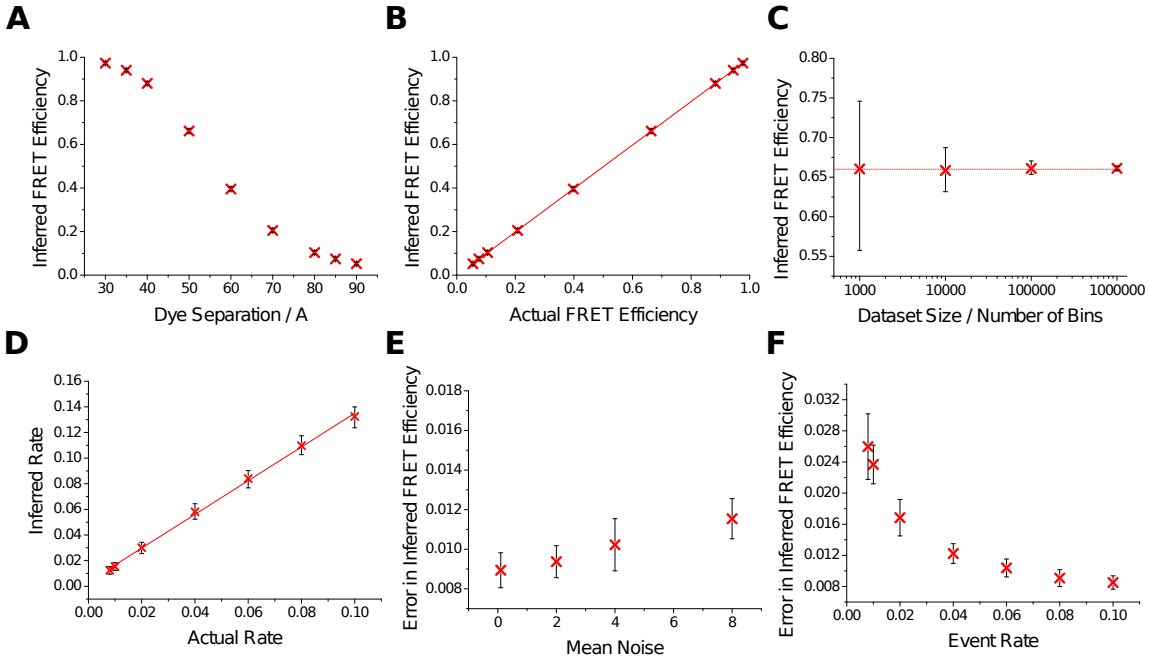

Figure 3: Validation of inference technique using realistic simulated datasets. (A) Inferred FRET efficiency plotted against dye-dye distance. (B) Inferred FRET efficiency plotted against calculated FRET efficiency. (C) Mean inferred FRET efficiency plotted against dataset size, for synthetic datasets with a dye-dye distance of 60 Å. (D) Inferred population size plotted against actual population size for synthetic datasets with a dye-dye distance of 60 Å. (E) Error in inferred FRET efficiency plotted against the mean value of background noise. The indicated mean noise was used in both the donor and the acceptor channels. The synthetic datasets used a dye-dye distance of 60 Å. (F) Error in inferred FRET efficiency plotted against the rate of observation of labelled molecules, for synthetic datasets with a dye-dye distance of 60 Å. All data points on all plots (A-F) were created using 10 synthetic datasets, generated independently from the same starting parameters. These datasets were analysed independently using the inference method, generating 98 accepted samples per dataset. Shown are the mean values of all accepted samples. The error bars are the values of the highest and the lowest accepted sample values, corresponding to a confidence interval within which the real value lies with probability $> 99\%$.

between actual and calculated population size for a range of different FRET efficiencies. Here, inference (Fig. 4 (A) middle) is clearly superior, showing no variation in the observed population size with FRET efficiency. Both thresholding techniques (Fig. 4 (A) bottom) show significant biases in their determined population sizes. The greatest problems arise from AND thresholding (open circles) where, although the peak areas of fluorescent species with intermediate FRET efficiencies are estimated correctly, there is significant underestimation of the peak sizes for both high- and low-FRET species. This bias is a direct result of the thresholding analysis: AND thresholding excludes fluorescent events that have a sub-threshold number of photons in one channel, but in a high- or low-FRET sample, this excludes most fluorescent events, causing huge underestimation of the population size. A smaller but still significant bias is observed in SUM thresholding (closed circles), which overestimates the peak area of low-FRET species. This is caused by inclusion of zero-peak events, which SUM thresholding cannot separate from real events.

A second illustration of this effect is shown in Fig. 4 (B), which shows, for a range of different FRET efficiencies, the relationship between actual and calculated population sizes. Both AND (top panel) and SUM (middle panel) thresholding show artifacts in the calculated population sizes. Thresholding results generally in an underestimate of the population size, due to exclusion of dim events. Furthermore, AND thresholding (top) considerably underestimates the population sizes for the higest ($E = 0.88$, open black circles) and lowest ($E = 0.11$, blue triangles) FRET efficiencies. Similarly, SUM thresholding overestimates the population sizes of the lowest-FRET species ($E = 0.11$, blue triangles and $E = 0.21$, green crosses), due to systematic inclusion of zero-peak events. In contrast, inference analysis (bottom) performs well across the full range of dye-dye separations, returning precisely the actual population size for all FRET efficiencies considered. This indicates that inference outperforms thresholding, correctly inferring both E and population size where thresholding cannot.

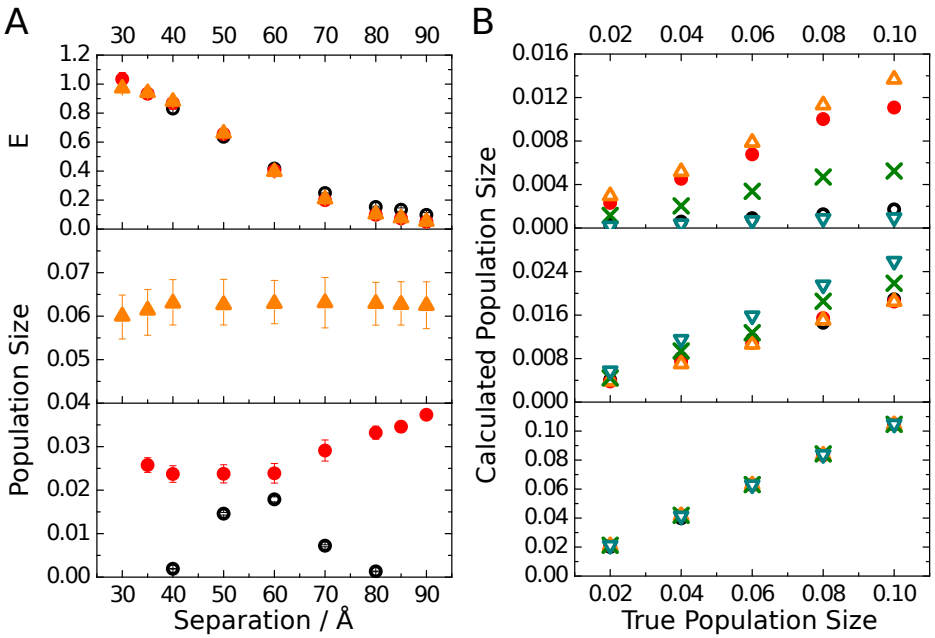

Figure 4: Comparison of the inference technique with thresholding-based methodologies. In all plots (A and B), values shown are from 10 datasets generated independently from the same parameters. For thresholding analsyes, error bars represent the standard deviation in the calculated mean from 10 independent datasets. For the inference technique, error bars are the values of the highest and the lowest accepted sample values, corresponding to a confidence interval within which the real value lies with probability $> 99\%$. (A - top) FRET efficiencies calculated for a series of simulated datasets using the inference methodologies and the AND and SUM thresholding techniques. Orange triangles are the inferred values, open black circles show AND thresholding, red circles show SUM thresholding. (A - middle, bottom) The effect of FRET efficiency on calculated population size. Orange triangles (middle) are the inferred values; open black circles and closed red circles (bottom) are values calculated using AND and SUM thresholding respectively. (B) The effect of FRET efficiency on calculated population size, as calculated using AND (top) and SUM (middle) thresholding and the inference method (bottom). The calculated population size is plotted against the value used in data generation. Open black circles, orange triangles (point up), red circles, green crosses and blue triangles (point down) correspond to FRET efficiencies of 0.88, 0.66, 0.40, 0.21 and 0.11 respectively.

Multiple fluorescent species

So far, we have considered simulated datasets containing a single fluorescent population. However, experimental datasets often contain a mixture of several fluorescent species. For a full analysis of these data, all populations must be correctly identified, both in terms of FRET efficiency and population size. To determine the utility of inference in these cases, we generated a total of 30 datasets simulating a mixture of two fluorescent populations, using three different population sizes and five different FRET efficiencies. Table 2 summarizes the parameters used. We then analysed these datasets using inference and using AND and SUM thresholding. The results, shown in Fig. 5, demonstrate that inference is significantly superior to both thresholding analyses. Fig. 5 (A) and (B) show the expected outcome of analysis of these data – there are five FRET efficiencies and three population sizes, resulting in a grid-like distribution of points. Both AND and SUM thresholding fail to reproduce this outcome. The bias of AND thresholding against high- and low-FRET species creates an inverted U-shaped distribution of calculated peak areas (Fig. 5 (C) and (D)) where species with intermediate FRET efficiencies (0.66 and 0.4) are calculated to have populations many times larger than those with high or low FRET efficiencies, even when these species were simulated with a rate three times higher. A different problem is observed in SUM analysis (Fig. 5 (E) and (F)). Here, although most populations are inferred correctly, peak areas of low-FRET species are enlarged by confounding with zero-peak events, significantly overestimating these populaton sizes. Furthermore, SUM thresholding entirely failed to separate mixtures of the two lowest-FRET species (E 0.21 and 0.11): only a single, broad peak could be fitted (not shown). In contrast, inference performs much better, although still imperfectly at this task. The results of the inference analysis, illustrated in Fig. 5 (G) and (H), show good separation of high, medium and low population sizes and very accurate inference of expected FRET efficiencies. For two datasets, inference does not infer correct values. These datasets both involve the lowest-FRET population ($E = 0.11$) at its lowest concentration, where it is very difficult to distinguish from noise. In one case, the magnitudes of the two populations ($E = 0.11$,

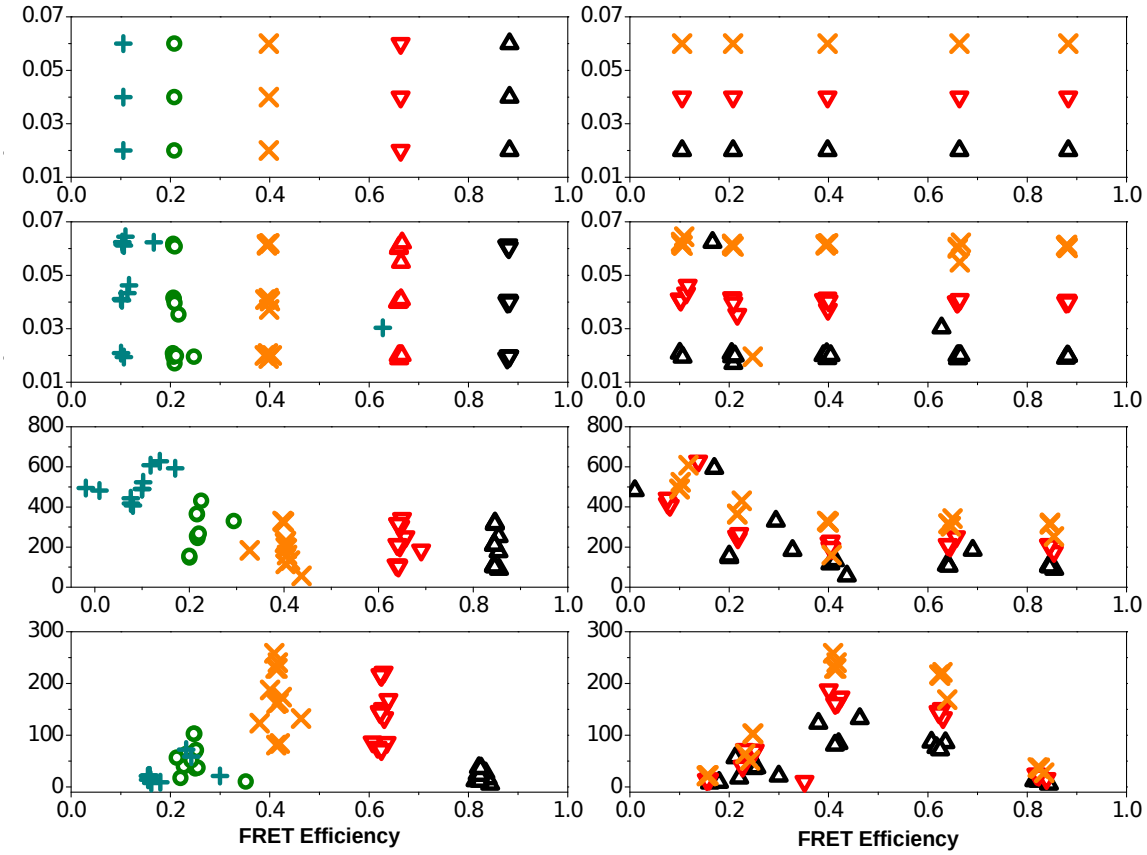

Figure 5: Comparison of inference and thresholding analysis on datasets generated to simulate mixtures of fluorescent species. Calculated population size is plotted against calculated FRET efficiency. (A) Idealised situation, in which all FRET efficiencies and population sizes are inferred correctly. (B) Results of analysis using the inference method. (C) Analysis using SUM thresholding. (D) Analysis using AND thresholding. In all panels A - D, graphs in the left hand column are coloured according to FRET efficiency: blue crosses (+) , green circles, orange crosses (x), red triangles (point down) and black triangles (point up) represent FRET efficiencies of 0.11, 0.21, 0.40, 0.66 and 0.88 respectively. Graphs in the right hand column are coloured according to population size: orange crosses, red triangles (point down) and black triangles (point up) represent respectively the large, medium and small population sizes.

$E = 0.21$, ratio 1:3) are switched. In the other case ($E = 0.11$, $E = 0.66$, ratio 1:3), the low-FRET population is ignored and the high-FRET population is split into two populations with similar values of E. Despite these two failures, it is important to note that whereas inference accurately infers the absolute size of both fluorescent populations in each dataset, not only do thresholding techniques fail to accurately estimate the absolute population sizes, they also frequently extimate incorrectly even the relative sizes of two populations, with inversion of estimated population sizes occurring.

Table 2: FRET efficiencies and population observation rates used in the generation of simulated datasets with two fluorescent populations.

E_1	0.88											
$\lambda_{\mathbf{prot1}}$	0.02	0.04	0.06	0.02	0.04	0.06	0.02	0.04	0.06	0.02	0.04	0.06
E_2	0.66			0.40			0.21			0.11		
$\lambda_{\mathbf{prot2}}$	0.06	0.04	0.02	0.06	0.04	0.02	0.06	0.04	0.02	0.06	0.04	0.02
E_1	0.66											
$\lambda_{\mathbf{prot1}}$	0.02	0.04	0.06	0.02	0.04	0.06	0.02	0.04	0.06			
E_2	0.40			0.21			0.11					
$\lambda_{\mathbf{prot2}}$	0.06	0.04	0.02	0.06	0.04	0.02	0.06	0.04	0.02			
E_1	0.40						0.21					
$\lambda_{\mathbf{prot1}}$	0.02	0.04	0.06	0.02	0.04	0.06	0.02	0.04	0.06			
E_2	0.21			0.11			0.11					
$\lambda_{\mathbf{prot2}}$	0.06	0.04	0.02	0.06	0.04	0.02	0.06	0.04	0.02			

Application to Experimental Data

DNA Duplexes

As a first test of the inference technique on experimental data, we determined the FRET efficiencies and population sizes of freely diffusing DNA duplexes labelled with the FRET pair of dyes Alexa Fluor® 488 and Alexa Fluor® 647. We also analysed these data using AND and SUM thresholding. We used a series of different DNA sequences, with dye attachment sites separated by between 4 and 12 base-pairs. As the separation between the dye attachment

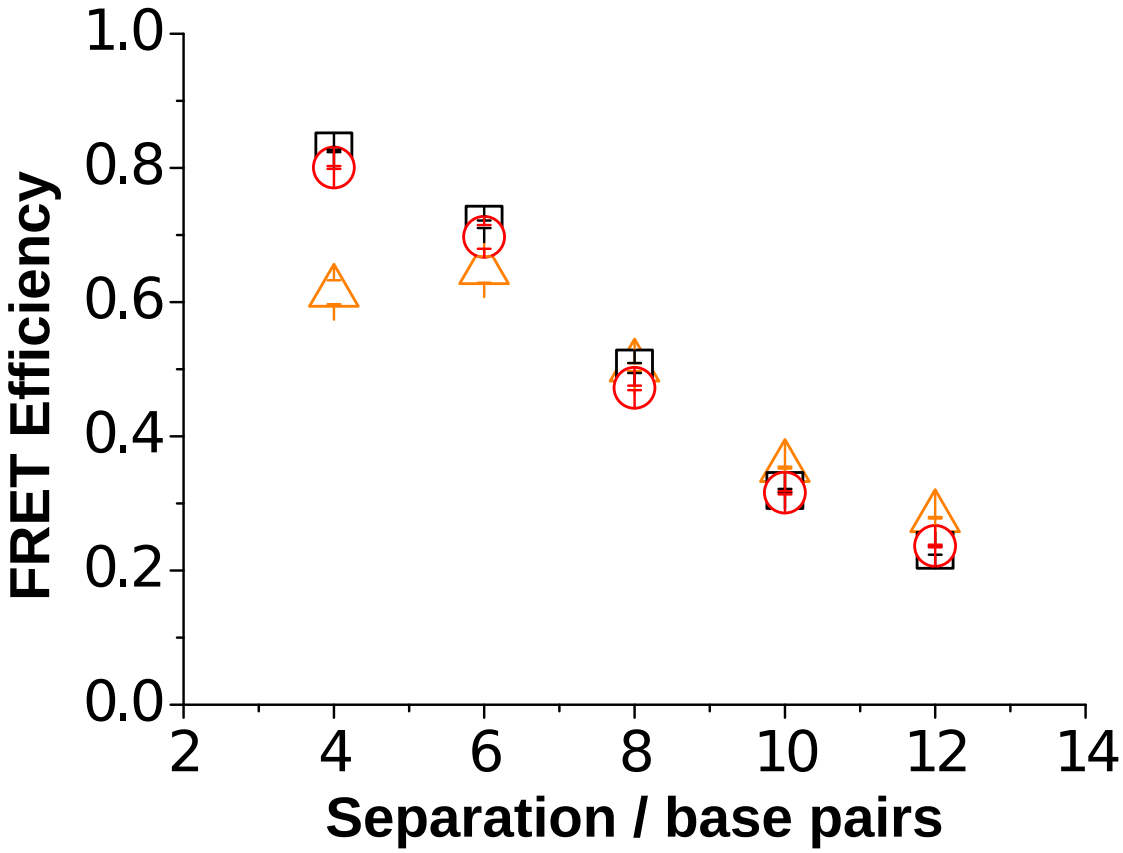

Figure 6: Results of AND, SUM and inference analysis of smFRET data from single populations of dual-labelled dsDNA. Orange triangles, black squares and red circles show respectively the results of the AND, SUM and inference based analyses. Error bars show the standard deviation of three independent repeats.

sites increases, the FRET efficiency is expected to decrease in a sigmoidal manner. As Fig. 6 shows, all three analysis methods reproduce this curve. The discrepancies between these curves are interesting. AND thresholding shows a somewhat squashed curve – with FRET efficiencies of the species with the highest FRET calculated to be lower than calculated by other methods and the species with the lowest FRET efficiencies calculated to have a slightly higher FRET efficiency than by other methods. This is explained by the bias towards intermediate-FRET species that results from the AND criterion. In contrast, both SUM thresholding and the inference process produce a smooth curve without demonstrating this bias.

Mixtures of DNA Duplexes

Finally, we applied two-population inference to mixtures of two DNA duplexes, combined, as in the synthetic examples, in an equimolar ratio (intermediate concentration), or with a three-fold excess of one duplex (high and low concentrations). We used a high- (4 bp separation), an intermediate- (10 bp separation) and a low-FRET duplex (12 bp separation). The datasets were also analysed using both AND and SUM thresholding. The results are displayed in Fig. 7).

Here, inference (Fig. 7 A - C) performs very well. In all three cases, the correct FRET efficiency was inferred, and a monotonic increase in event rate is seen between low, intermediate and high concentrations of duplex. In contrast, the thresholding analyses perform very poorly. FRET efficiencies calculated using AND thresholding (Fig. 7 D - F) are squashed towards intermediate FRET efficiencies. This also distorts the event distribution, meaning that the population sizes are inaccurately estimated. Similarly, although SUM thresholding (Fig. 7 G - J) accurately measures FRET efficiencies for two of the mixtures, it is unable to resolve the 10 bp - 12 bp mixture, so only a single fluorescent population can be resolved (Fig. 7 J) . Furthermore, SUM thresholding also distorts the population sizes, with populations of low FRET species (10 bp and 12 bp dupelexes) being significantly overesti-

mated, owing to zero-peak contributions. Consequently, inference analysis emerges as the most reliable method to analyse mixtures of fluorescent species. Note however, that even the inference method cannot fully resolve the the 10 bp - 12 bp mixture. Although population sizes are correctly inferred, the two FRET efficiencies are compressed towards each other, suggesting that the two species are difficult to distinguish. This indicates a resolution limit of approximately 5 Å for the inference method.

Summary and Conclusion

Model-based Bayesian inference is a powerful tool that is used in data analysis across many disciplines[27]. However, despite establishment of model-based inference methods to analyze FRET trajectories from immobilised molecules[28–32], a similar method had not been developed for smFRET data from molecules in solution. We have developed a model-based inference method, based on the Metropolis algorithm, suitable for the analysis of these datasets. This enables unbiased, single-step determination of FRET efficiencies and population sizes for one or more fluorescent species, as well as other parameters of the dataset. Raw data is analyzed in a single step directly, requiring neither biased thresholding, nor construction and subjective fitting of FRET histograms. It is extremely robust across a wide variety of conditions. Model-based inference is an exciting new avenue for analysis of smFRET datasets. With simple modifications, similar methods could be developed for analysis of data collected using alternating excitation methods, as well as for other types of smFRET experiment. The software for simulation of smFRET datasets and for the analysis of both real and simulated data is available publicly and can be downloaded from `https://bitbucket.org/rebecca_roisin/fret-inference`.

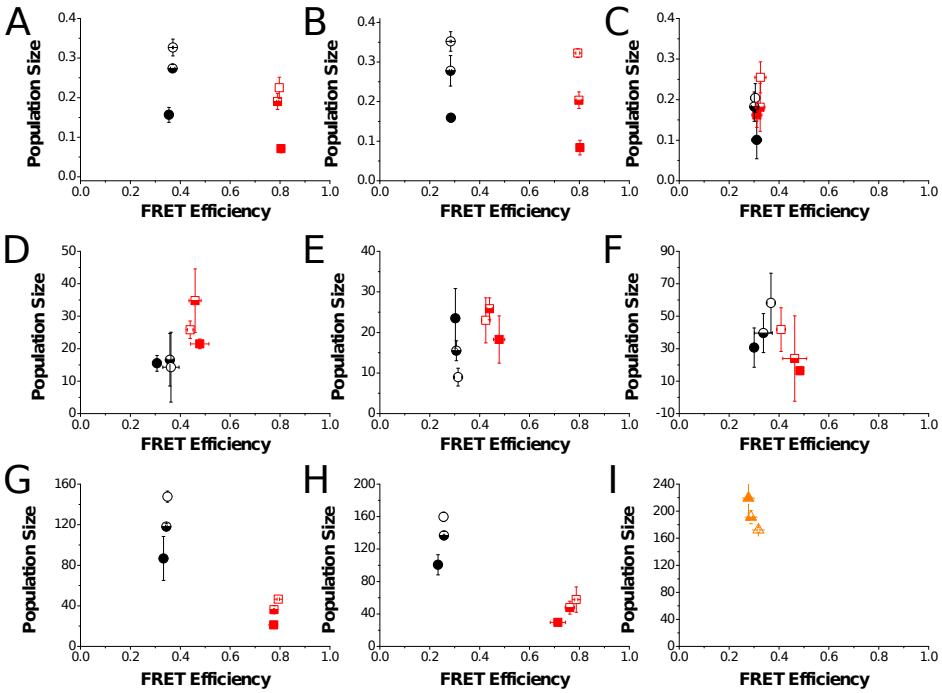

Figure 7: Analysis of a mixture of two populations of dual-labelled dsDNA, showing the calculated population sizes and FRET efficiencies. Three different DNA strands were used, with dye attachment sites separated by 4, 10 and 12 bp, corresponding to FRET Efficiencies of 0.79, 0.36 and 0.28 respectively as calculated using the inference method. Two DNA duplexes were combined to give a total DNA concentration of 80 pM, using either 20 pM (low concentration) of one duplex and 60 pM (high concentration) of the other duplex, or 40 pM (intermediate concentration) of both duplexes. Black triangles (point up), red triangles (point down) and orange crosses represent the low, intermediate and high concentrations of DNA respectively. A - C: Inference analysis of 4 and 10 bp, 4 and 12 bp, and 10 and 12 bp mixtures respectively. D - F: AND analysis of 4 and 10 bp, 4 and 12 bp, and 10 and 12 bp mixtures respectively. G - J: SUM analysis of 4 and 10 bp, 4 and 12 bp, and 10 and 12 bp mixtures respectively. Red squares represent the higher-FRET species in a mixture; black circles represent the lower-FRET duplex. Open shapes correspond to a concentration of 60 pM (high), whereas filled shapes correspond to a concentration of 20 pM (low). Half-filled shapes correspond to the intermediate duplex concentration of 40 pM. In J, SUM analysis was not able to resolve two peaks, so a single gaussian was fitted. The single peak area and FRET efficiency are shown with orange triangles. Error bars represent the standard deviation of three independent experiments, except in the case of the 3:1 4 bp : 10 bp mixture, where one repeat was excluded, due to an incorrect concentration of the 4 bp duplex being used.

Acknowledgement

RRM thanks BBSRC for research funding. GD was employed at Microsoft Research during part of this project.

Supporting Information Available

Supplementary methods and derivations, as well as a description of the Metropolis algorithm can be in the supporting information.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Ha, T.; Enderle, T.; Ogletree, D. F.; Chemla, D. S.; Selvin, P. R.; Weiss, S. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 6264–6268.

(2) Haran, G. *J. Phys.: Condens. Matter* **2003**, *15*, R1291–R1317.

(3) Schuler, B.; Lipman, E. A.; Eaton, W. A. *Nature* **2002**, *419*, 743–747.

(4) Weiss, S. *Nat. Struct. Mol. Biol* **2000**, *7*, 724–729.

(5) Deniz, A. A.; Lawrence, T. A.; Dahan, M.; Chemla, D. S.; Schultz, P. S.; Weiss, S. *Annu. Rev. Phys. Chem.* **2001**, *52*, 233–253.

(6) Gell, C.; Brockwell, D.; Smith, A. *Handbook of single molecule fluorescence*; Oxford University Press: Oxford, 2006.

(7) Ying, L.; Wallace, M. I.; Balsubramanian, S.; Klenerman, D. *J. Phys. Chem. B.* **2000**, *104*, 5171–5178.

(8) Nir, E.; Michalet, X.; Hamadani, K. M.; Laurence, T. A.; Neuhauser, D.; Kovchegov, Y.; Weiss, S. *J. Phys. Chem. B* **2006**, *110*, 22103–22124.

(9) Kalinin, S.; Peulen, T.; Sindbert, S.; Rothwell, P. J.; Berger, S.; Restle, T.; Goody, R. S.; Gohlke, C. A. M., H.Seidel *Nat Methods* **2012**, *9*, 1218–1225.

(10) Hoefling, M.; Lima, N.; Haenni, D.; Seidel, C. A. M.; Schuler, B.; Grubmller, H. *PLoS One* **2011**, *6*, e19791.

(11) Sindbert, S.; Kalinin, S.; Nguyen, H.; Kienzler, A.; Clima, L.; Bannwarth, W.; Appel, B.; Muller, S.; Seidel, C. A. M. *J. Am. Chem. Soc.* **2011**, *133*, 2463–2480.

(12) Vogelsang, J.; Doose, S.; Sauer, M.; Tinnefeld, P. *Anal. Chem* **2007**, *79*, 7367–7375.

(13) Horrocks, M. H.; Li, H.; Shim, J.; Ranasinghe, R. T.; Clarke, R. W.; Huck, W. T. S.; Abell, C.; Klenerman, D. *Anal. Chem.* **2012**, *84*, 179–185.

(14) Deniz, A.; Laurence, T.; Grunwell, J.; Ha, A., T.J. an Faulhaber; Chemla, D.; Weiss, S.; Schultz, P. *Proc. Natl. Acad. Sci. USA.* **1999**, *96*, 3670–3675.

(15) Kapanidis, A.; Laurence, T.; Lee, N.; Margeat, E.; Kong, X.; Weiss, S. *Acc. Chem. Res.* **2005**, *38*, 532–533.

(16) Muller, B. K.; Zaychikov, E.; Brauchle, C.; Lamb, D. C. *Biophys. J.* **2005**, *89*, 3508–3522.

(17) Doose, S.; Heilemann, M.; Michalet, X.; Weiss, S.; Kapanidis, A. N. *Eur. Biophys. J.* **2006**, *36*, 669–674.

(18) Kudryavtsev, V.; Sikor, M.; Kalinin, S.; Mokranjac, D.; Seidel, C. A. M.; Lamb, D. C. *ChemPhysChem* **2012**, *13*, 1060–1078.

(19) Sisamakis, E.; Valeri, A.; Kalinin, S.; Rothwell, P. J.; Seidel, C. A. M. *Methods in Enzymology* **2010**, *475*, 455–514.

(20) Eggeling, C.; Berger, S.; Brand, L.; Fries, J.; Schaffer, J.; Volkmer, A.; C.A.M., S. *J. Biotechnol.* **2001**, *86*, 163–180.

(21) Kalinin, S.; Felekyan, S.; Antonik, M.; Seidel, C. A. M. *J. Phys. Chem. B.* **2007**, *111*, 10253–10262.

(22) Antonik, M.; Felekyan, S.; Gaiduk, A.; Seidel, C. A. M. *J. Phys. Chem. B.* **2006**, *110*, 6970–6978.

(23) Santoso, Y.; Torella, J. P.; Kapanidis, A. N. *ChemPhysChem* **2010**, *11*, 2209–2219.

(24) Torella, J. P.; Holden, S. J.; Santoso, Y.; Hohlbein, J.; Kapanidis, A. N. *Biophysical Journal* **2011**, *107*, 5058–5063.

(25) Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.

(26) Bayes, T.; Price, R. *Phil. Trans. R. Soc. London* **1763**, *53*, 370–418.

(27) MacKay, D. J. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.

(28) McKinney, S. A.; Joo, C.; Ha, T. *Biophysical Journal* **2006**, *91*, 1941–1951.

(29) Bronson, J. E.; Fei, J.; Hofman, J. M.; Gonzalez, R. N.; Wiggins, C. H. *Biophys. J.* **2009**, *97*, 3196–3205.

(30) Bronson, J. E.; Hofman, J. M.; Fei, J.; Gonzales, R. L.; Wiggins, C. H. *BMC Bioinformatics* **2010**, *11*, 2–10.

(31) Taylor, J. N.; Makarov, D. E.; Landes, C. F. *Biophys. J.* **2010**, *98*, 164–173.

(32) Taylor, J. N.; Landes, C. F. *J. Phys. Chem. B.* **2011**, *115*, 1105–1114.

(33) Uphoff, S.; Gryte, K.; Evans, G.; Kapanidis, A. N. *ChemPhysChem* **2011**, *12*, 571579.

(34) Yoon, J. W.; Bruckbauer, A.; Fitzgerald, W. J.; Klenerman, D. *Biophys. J.* **2008**, *94*, 4932–4947.

(35) Turkcan, S.; Alexandrou, A.; Masson, J.-B. *Biophys. J.* **2012**, *102*, 2288–2298.

(36) Stigler, J.; Rief, M. *ChemPhysChem* **2012**, *13*, 1079–1086.

(37) Kugel, W.; Muschielok, A.; Michaelis, J. *Chemphyschem* **2011**, *13*, 1013–1022.

(38) Guo, S.-M.; He, J.; Monnier, N.; Sun, G.; Wohland, T.; Bathe, M. *Anal. Chem.* **2011**, *84*, 3880–3888.

(39) He, J.; Guo, S.-M.; Bathe, M. *Anal. Chem.* **2011**, *84*, 3871–3879.

(40) Guo, S.-M.; Bag, N.; Mishra, A.; Wohland, T.; Bathe, M. *Biophys. J.* **2014**, *106*, 190–200.

(41) Kou, S. C.; Xie, X. S.; Liu, J. S. *J. R. Stat. Soc.* **2005**, *54*, 469–496.

(42) Kalinin, S.; Felekyan, S.; Valeri, A.; Seidel, C. A. M. *J. Phys. Chem. B.* **2007**, *112*, 8361–8374.

(43) DeVore, M. S.; Gull, S. F.; Johnson, C. K. *J. Phys. Chem. B.* **2012**, *116*, 4006–4015.

(44) Gopich, I. V.; Szabo, A. *J. Phys. Chem. B.* **2007**, *111*, 12925–12932.

(45) Gopich, I. V.; Szabo, A. *J. Phys. Chem. B.* **2003**, *107*, 5058–5063.

(46) Hastings, W. *Biometrika* **1970**, *57*, 97–109.

(47) Schuler, B. *Chemphyschem* **2005**, *6*, 1206–1220.

(48) Chung, H. S.; Louis, J. M.; Eaton, W. M. *Proc. Natl. Acad. Sci. USA.* **2009**, *106*, 11837–11844.

(49) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *Journal of Chemical Physics* **1953**, *21*, 1087–1092.

(50) Chib, S.; Greenberg, E. *Am. Stat.* **1995**, *49*, 327–335.

(51) Li, H.; Ying, L.; Green, J. J.; Balasubramanian, S.; Klenerman, D. *Analytical Chemistry* **2003**, *75*, 1664–1670.

Graphical TOC Entry