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

Contents

[*]To whom correspondence should be addressed

[†]Department of Chemistry, University of Cambridge, U.K.

[‡]Dept. of Computer Science, University College London, U.K.

Supplementary Theory: Inference of Model Parameters

Inference of Model Parameters

For each data point $(f_d, f_a)_i$ in a dataset, the probability that it was generated by a given set of parameters can be calculated as the sum of the probabilities that it was generated from each of the distinct states, described in the generative model:

$$
\begin{aligned}
&\Pr[(f_d, f_a)_i | \lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E] \\
&= \Pr[(f_d, f_a)_i | \lambda_D, \lambda_A, \text{noise only}] \cdot \Pr[\text{noise only}] \\
&+ \Pr[(f_d, f_a)_i | \lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, p_D, p_A, \text{donor event}] \cdot \Pr[\text{donor event}] \\
&+ \Pr[(f_d, f_a)_i | \lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A, \text{FRET event}] \cdot \Pr[\text{FRET event}] \\
&+ \Pr[(f_d, f_a)_i | \lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A, \text{multiple occupancy}] \cdot \Pr[\text{multiple occupancy}],
\end{aligned}
\tag{1}
$$

where $(f_d, f_a)_i$ is the ith pair of observations in the dataset and Pr[noise only], Pr[donor event], Pr[FRET event] and Pr[multiple events] are the probabilities of observing noise photons only; of observing a protein carrying just the donor dye; of observing a protein carrying both donor and acceptor dyes and of observing multiple proteins present in the excitation volume. These probabilities, for the single fluorescent population case, are then:

$$\begin{aligned}&\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \text{noise only}]\\&= ((1-p_{\text{prot}}) + p_{\text{prot}}(1-p_D)(1-p_A) + p_{\text{prot}}(1-p_D)p_A)\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \text{noise only}]\end{aligned} \tag{2}$$

$$\begin{aligned}&\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, p_D, p_A, \text{donor event}]\\&= p_{\text{prot}}p_D(1-p_A)\Pr[(f_d, f_a)_i|\lambda_{n_D} + \lambda, \lambda_{n_A}, \text{Donor only}]\end{aligned} \tag{3}$$

$$\begin{aligned}&\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A, \text{FRET event}]\\&= p_{\text{prot}}p_D p_A \Pr[(f_d, f_a)_i|\lambda_{n_D} + \lambda(1-E), \lambda_{n_A} + \lambda E\gamma, \text{FRET}]\end{aligned} \tag{4}$$

where p_D and p_A are the labelling probabilities with the donor and acceptor dyes respectively and p_{prot} is the probability mass function of a Poisson distribution with mean λ at $k = 1$: $p_{\text{prot}} = \lambda e^{-\lambda}$, giving the probability that the confocal volume is occupied by exactly one protein molecule.

For the two population case, Eq. 2 and Eq. 3 can still be used to describe the probability that an event is generated by noise only (Eq. 2) or by a donor-only fluorescent event (Eq. 3), provided that the p_{prot} terms are replaced by the sum $p_{\text{prot}} = p_{\text{prot1}} + p_{\text{prot2}}$. However, equation 4 needs modification to accommodate the multiple FRET efficiencies E_1 and E_2 as well as their respective population sizes:

$$\begin{aligned}&\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A, \text{FRET event}]\\&= p_{\text{prot}}p_D p_A \Pr[(f_d, f_a)_i|\lambda_{n_D} + \lambda(1-E_1), \lambda_{n_A} + \lambda E_1\gamma, \text{FRET}_1]\\&+ p_{\text{prot}}p_D p_A \Pr[(f_d, f_a)_i|\lambda_{n_D} + \lambda(1-E_2), \lambda_{n_A} + \lambda E_2\gamma, \text{FRET}_2].\end{aligned} \tag{5}$$

For a dataset with a single fluorescent population, if a single labelled molecule is present in the confocal volume, the emission probabilities for observed photons $(f_d, f_a)_i$ are then given by the integrals:

$$\Pr[(f_d, f_a|\lambda, \lambda_D, \lambda_A] = \text{Poisson}(f_A; \lambda_A) \cdot \int_0^\infty \text{Poisson}(f_D; \lambda + \lambda_D) \cdot \text{Gamma}(\lambda; k_D, \theta)\, d\lambda \tag{6}$$

$$= \int_0^\infty \frac{(\lambda + \lambda_D)^{-f_D} e^{-(\lambda+\lambda_D)}}{f_D!} \frac{\lambda_A^{-f_A} e^{-\lambda_A}}{f_A!} \lambda^{k_D-1} \frac{e^{-\frac{\lambda}{\theta}}}{\theta^{k_D}\Gamma(k_D)}\, d\lambda, \tag{7}$$

for a molecule labelled with only the donor dye, and:

$$\Pr[f_d, f_a|\lambda, E, \gamma, \lambda_D, \lambda_A] = \tag{8}$$

$$= \int_0^\infty \text{Poisson}(f_D; \lambda(1-E) + \lambda_D) \cdot \text{Poisson}(f_A; \lambda E\gamma + \lambda_A) \cdot \text{Gamma}(\lambda; k_D, \theta)\, d\lambda \tag{9}$$

$$= \int_0^\infty \frac{(\lambda(1-E) + \lambda_D)^{-f_D} e^{-(\lambda(1-E)+\lambda_D)}}{f_D!} \frac{(\lambda E\gamma + \lambda_A)^{-f_A} e^{-(\lambda E\gamma+\lambda_A)}}{f_A!} \lambda^{k_D-1} \frac{e^{-\frac{\lambda}{\theta}}}{\theta^{k_D}\Gamma(k_D)}\, d\lambda, \tag{10}$$

for a molecule labelled with both donor and acceptor dyes. The donor-only probabilities are unchanged in the case of two fluorescent populations. However, the FRET case, in which both the donor and acceptor dyes are present, becomes:

$$\Pr[(f_d, f_a)_i|\lambda, E_1, E_2, \gamma, \lambda_D, \lambda_A] = P_1 \cdot \Pr[(f_D, f_A)_i|\lambda, E_1, \gamma, \lambda_D, \lambda_A] + P_2 \cdot \Pr[(f_D, f_A)_i|\lambda, E_2, \gamma, \lambda_D, \lambda_A] \tag{11}$$

where the two terms $\Pr[(f_D, f_A)_i|\lambda, E_1, \gamma, \lambda_D, \lambda_A]$ and $\Pr[(f_D, f_A)_i|\lambda, E_2, \gamma, \lambda_D, \lambda_A]$ can be determined as for the single population case using eqn 10 and P_1 and P_2, given by Eqn. 12 below, describe the relative sizes of the two fluorescent populations.

$$P_1 = \frac{\lambda_{\text{prot1}}}{\lambda_{\text{total}}} \cdot \lambda_{\text{total}} \cdot e^{-\lambda_{\text{total}}} \qquad P_2 = \frac{\lambda_{\text{prot2}}}{\lambda_{\text{total}}} \cdot \lambda_{\text{total}} \cdot e^{-\lambda_{\text{total}}} \tag{12}$$

for

$$\lambda_{\text{total}} = \lambda_{\text{prot1}} + \lambda_{\text{prot2}} \tag{13}$$

These integrals are computed numerically.

Finally, the Pr[multiple events] term represents a simplification of the inference process compared with the forward model. Whereas the generative process modelled explicitly multiple occupancy of the excitation volume; in the inference process, parameters are inferred assuming only a single protein is present in the confocal volume. In the single-population case, the potential to observe multiple proteins in the excitation volume at the same time is collapsed into a single negative binomial term, with a single averaged parameter:

$$\begin{aligned}&\Pr[(f_d, f_a)_i | \lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A, \text{multiple events}] \\ &= (1 - \lambda_{\text{prot}} e^{-2\lambda_{\text{prot}}}) \frac{\Gamma(f_D + r)!}{f_D! \Gamma(r)} \frac{\Gamma(f_A + r)!}{f_A! \Gamma(r)} (1 - \frac{\mu_D}{r + \mu_D})^r (\frac{\mu_D}{r + \mu_D})^{f_D} (1 - \frac{\mu_A}{r + \mu_A})^r (\frac{\mu_A}{r + \mu_A})^{f_A}\end{aligned} \tag{14}$$

where r is a fixed over-dispersion parameter, $r = 4$, and μ_D and μ_A are the mean number of photons expected in the donor and acceptor channels, respectively, when two or three proteins are observed:

$$\mu_D = \frac{p_2(2\lambda(1-E)) + p_3(3\lambda(1-E))}{p_2 + p_3} \qquad \mu_A = \frac{p_2(2\lambda\gamma E) + p_3(3\lambda\gamma E)}{p_2 + p_3}, \tag{15}$$

where:

$$p_2 = \frac{\lambda_{\text{prot}}^2}{2!} e^{-\lambda_{\text{prot}}} \qquad p_3 = \frac{\lambda_{\text{prot}}^3}{3!} e^{-\lambda_{\text{prot}}}, \tag{16}$$

In the multiple population case, accounting for multiple occupancy is made more complex by the potential for molecules in different states to co-occupy the confocal volume. For this reason, co-occupancy by up to four molecules is treated explicitly. The mean number of photons expected from two, three or four fluorescent molecules, in all possible

labelling and configurational states is calculated. These values are then used to calculate a total mean for multiple occupancy events, which is used as above in eqn (14).

The total probability that the pair of datapoints f_D, f_A was generated by a certain set of parameters is than computed using Equation 1. The probability that the whole dataset was generated by a those parameters is then the product:

$$\Pr[\text{Obs.}|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A] = \prod_i \Pr[(f_D, f_A)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A] \quad (17)$$

Comparing the total probability values for different sets of parameters allows identification of parameter sets that have a high probability of having generated the observed dataset. Repeated sampling of parameters using the Metropolis-Hastings algorithm allows determination of mean parameter values and associated confidence intervals.

The Metropolis-Hastings Algorithm

The section above describes a method to sample values of the FRET efficiencies, population sizes an so on by sampling from a Markov Chain that has as its stationary probability the posterior probability over the parameters, $\Pr[\text{Obs.}|\text{Par.}] \cdot \Pr[\text{Par.}]$. We use a custom implementation of the Metropolis algorithm to achieve this. The Metropolis algorithm works as follows.

- Each variable that we wish to infer (namely $\lambda_A, \lambda_D, \lambda_{\text{prot}}, \lambda$ and E for the single population case, replacing λ_{prot} with λ_{prot1} and λ_{pro2} and E with E_1 and E_2 for the two popultion inference) is sampled from an arbitrary probability distribution, typically a Gaussian distribution, centred around the current value of that variable: $x' \sim Q(x'|x)$, where x and x' are the current and newly sampled values of variable x, and $Q(x'|x)$ is a symmetric proposal density, with the property that $Q(x'|x) = Q(x|x')$.

- In each sampling event, the probability $\Pr(\text{Obs.}|\text{Par.})$ is evaluated for the current set of parameters, using equation 1 and calculating the components of the sum using

the equations described above.

- A new value is then drawn for one of the variable parameters, chosen at random from the sampled variables, and the probability $\Pr(\text{Obs.}|\text{Par.})$ is recalculated for new set of parameters and the results compared by computing the acceptance ratio, a, which defines how probable the new sample value is, compared with the current value of the parameter:

$$a = \frac{\Pr(\text{Obs.}|x')}{\Pr(\text{Obs.}|x)} \tag{18}$$

  where $\Pr(\text{Obs.}|x')$ and $\Pr(\text{Obs.}|x)$ are the total probabilities that the dataset was generated by the new parameters and the old parameters respectively.

- If $a \geq 1$, the new value, x' is accepted and the value of the parameter updated. Otherwise, if $a < 1$, x' is accepted with probability a; with probability $1 - a$, the parameter's value is unchanged.

- This process is initialised using arbitrary values for all the parameters, and the sampling process is then repeated multiple times, selecting a fixed distribution to vary at each sampling event. After many iterations (the burn-in period, typically 4000 iterations), the initial values are forgotten and drawing further samples allows sampling from the distribution $\Pr(\text{Obs.}|\text{Par.})$. This allows us to sample repeatedly from regions of the parameter sample space that have a much higher probability density - giving parameters that have a high probability of having generated the data observed. We found that all areas of the sample space were accessible given an arbitrary starting value, meaning that the parameter values inferred were independent of their initial values (Fig. S3).

Supplementary Experimental Methods

DNA Sample Preparation

Single-stranded DNA labelled with either Alexa Fluor® 488 or Alexa Fluor® 647 were purchased from Sigma. Two 488-labelled donor sequences were used, whose sequences are shown in Table 1. These were annealed to one of the five 647-labelled acceptor sequences shown in Table 2. Annealing was performed by mixing an aliquot of donor sequence with a 1.1 molar excess of acceptor sequence and heating to 90° for 30 minutes, then cooling gradually to room temperature over a period of three hours. The final concentration of dsDNA was 2 μM. For smFRET measurements, a total dsDNA concentration of 60 pM was used.

Table 1: DNA sequences of the donor-labelled strands, where 5 is a deoxy-T nucleotide, labelled with Alexa Fluor® 488 at the C6 amino position.

Donor Construct	Sequence
Donor 1	5AAATCTAAAGTAACATAAGGTAACATAACGGTAAGTCCA
Donor 2	5AAATCGCTAAAGTAACATAAGGTAACATAACGGTAAGTCCA

Table 2: Preparing the dual-labelled dsDNA. An acceptor-labelled ssDNA, with the sequence shown was annealed to the indicated donor construct, to yield a dual-labelled construct with the labels separated by the given number of base pairs. In the displayed acceptor-strand sequences, 6 is a deoxy-T nucleotide, labelled with Alexa Fluor® 647 at the C6 amino position.

Dye Separation / bp	Acceptor Construct Sequence	Annealed Donor
7	TGGACTTACCGTTATGTTACCTTATGTTACTT6AGATTTA	Donor 1
12	TGGACTTACCGTTATGTTACCTTATGT6ACTTTAGATTTA	Donor 1
14	TGGACTTACCGTTATGTTACCTTATGT6ACTTTAGCGATTTA	Donor 2
17	TGGACTTACCGTTATGTTACCT6ATGTTACTTTAGATTTA	Donor 1

Instrumentation

A Gaussian laser beam of wavelength 488 nm (Qioptiq) and 75 μW power was directed *via* a fibre-optic cable (iFLEX Viper) into the back port of an inverted microscope (Nikon Eclipse TE2000-U). The beam was focused 5 μm into 350 μL of the sample in a 0.6 mL Laboratory Tek chambered cover slide (Scientific Laboratory Suppliers Ltd., Surrey, UK) through a high numerical aperture oil immersion objective (Appochromat 60 x, NA 1.40 Nikon). Sample fluorescence was collected by the same objective and imaged onto a 50 μm pinhole (Melles Griot) to exclude out of focus fluorescence. Donor and acceptor photons were then separated using a dichroic mirror (58DRLP, Omega Optical Filters).

Donor fluorescence was filtered by long-pass and band-pass filters (510ALP and 535AF45, Omega Optical Filters), then focused onto an avalanche photodiode (APD, SPCM AQ-161, EG&G, Canada). Acceptor fluorescence was similarly filtered using both long pass and band-pass filters (565ALP and 695AF55, Omega Optical Filters) before being focused on a second APD device (SPCM AQR-141, EG&G, Canada). Outputs from the two APDs were coupled to a PC-implemented Fluorescence Correlation Card (FPGA Celoxica RC10). The cross-talk from the donor to the acceptor channel has been found to be 3%, the acceptor-to-donor cross-talk is negligible.

Thresholding Analysis

For AND thresholding, time-bins were denoised by subtraction of an averaged autofluorescence value for each channel and for cross-talk by subtraction of 3 % of the donor channel value from the acceptor channel. Time-bins containing fluorescent events were identified using the criterion $n_D > 10$ and $n_A > 10$ for n_D and n_A photons in the donor and acceptor channels respectively. The FRET efficiency for each selected event was then calculated using an instrumental γ-factor of 1.0. Frequency histograms were then constructed of the calculated FRET efficiencies and fitted with a single Gaussian (for single fluorescent

species) or two Gaussains (for two fluorescent species). The mean of the fit was taken to be the mean FRET efficiency of the species and the area under the curve was taken to be proportional to the population size. For SUM thresholding, denoised time-bins were selected if $n_D + n_A > 20$. FRET Efficiency histograms were constructed and then fitted. If the data peaks were well separated from the zero peak, the zero peak was not fitted and one or two Gaussians were used as above to fit the histogram. If the data peaks were not distinct from the zero peak, an additional Gaussian was used to fit the zero-peak. Fitting was carried out using graphical fitting software (Origin 8.1 from OriginLab).

Determining Labelling Efficiency

To determine the fraction of labelled DNA molecules, an alternating laser excitation (ALEX) method was used over a data collection period of 10 minutes. The fraction of donor-labelled molecules and acceptor-labelled molecules, fr_D and fr_A, equivalent to p_D and p_A were found by calculating the ratios:

$$fr_D = \frac{n_{\text{donor}}}{n_{\text{total}}} \qquad fr_A = \frac{n_{\text{acceptor}}}{n_{\text{total}}} \tag{19}$$

where n_{donor} and n_{acceptor} are respectively the total number of donor and acceptor events in the dataset, and n_{total} is the total number of molecules seen in the dataset and is given by:

$$n_{\text{total}} = n_{\text{donor}} + n_{\text{acceptor}} - n_{\text{FRET}} \tag{20}$$

where n_{FRET} is the number of events for which an event was observed in both the donor and acceptor channels.

For these ALEX measurements, a bin time of 1 ms was used, with 10 laser modulations per bin. Analysis was carried out using an initial threshold of 10 donor and 10 acceptor photons, followed by an application of the ALEX thresholding criterion.[?] Software for

implementation of the ALEX analysis was written in Python.

Supplementary Figures

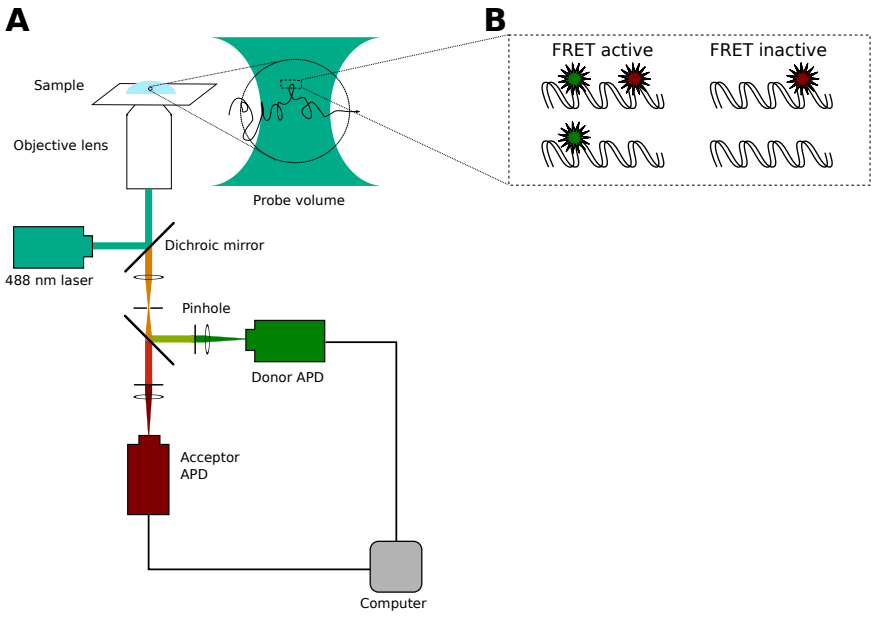

Fig. S - 1: A typical smFRET experiment. (A) Microscope set-up for smFRET. APD: Avalanche Photodiode. (B) The four possible labelling states for a single molecule in the confocal volume.

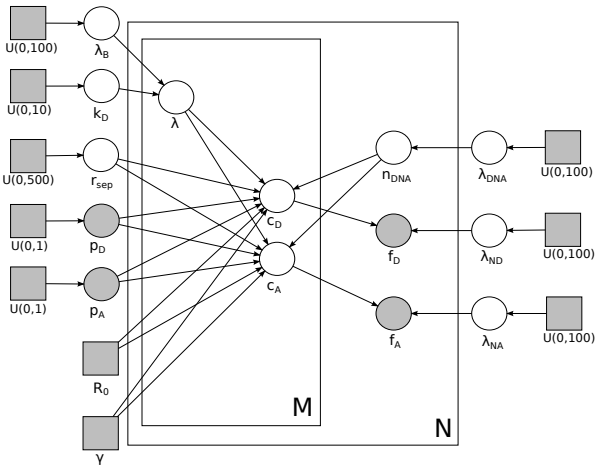

Fig. S - 2: Directed Acyclic Graph illustrating the interrelation of parameters in the inference model. In this notation, circles represent random variables, while squares represent constants. Known or observed values are shaded, while hidden variables are not. For each time bin in a dataset of size N, f_D and f_A are respectively the number of donor and acceptor photons observed; n_{prot} is the number of molecules present in the confocal volume. For each of M molecules present per bin, c_D and c_A are respectively the number of donor and acceptor photons emitted, r_{sep} is the dye separation interval and λ is the emission rate of the donor dye. The global variables λ_D and λ_A are the background emission rates of donor and acceptor photons; λ_{prot} is the rate of observation of fluorescent molecules; p_D and p_A are the probability that a molecule carries respectively a donor and an acceptor dye; λ_B and k_D are the parameters of the Gamma-distribution, from which the local donor emission rate, λ is selected. Each random variable is initialized using a prior selected from a uniform distribution across the indicated ranges. The two known constants R_0 and γ are the dye-separation for which FRET efficiency is 50 % and the instrumental γ-factor discussed above.

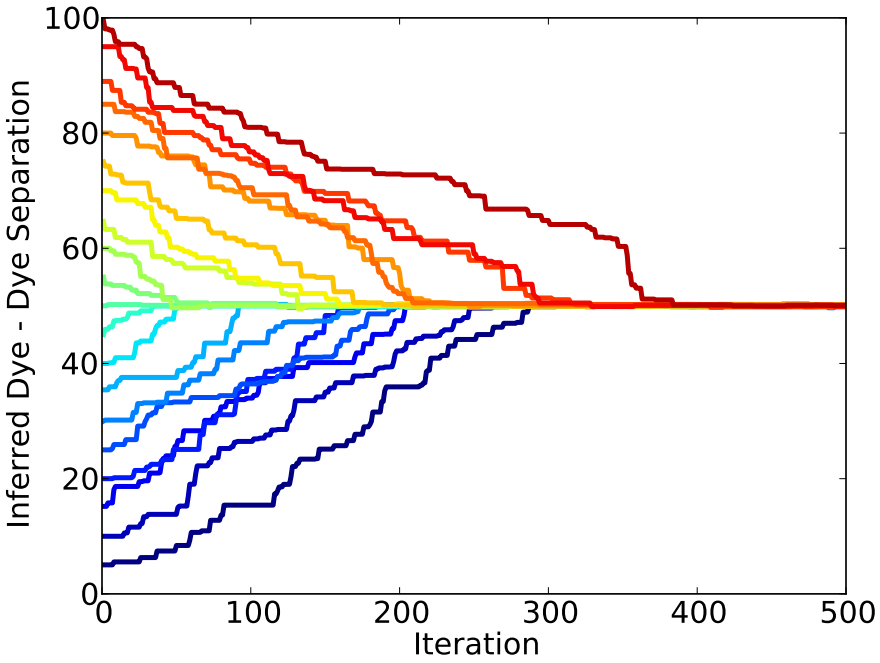

Fig. S - 3: Graph illustrating the convergence of the Metropolis sampler during the initial burn-in period of the sampling. Analysis of the same dataset was initiated using different values of the dye-dye separation from 5 Å(in blue) to 100 Å(in red), in steps of 5 Å. After 500 iterations of the sampler (within the 1000 iterations used in the burn-in period, during which no samples are stored) all initial values have converged on the correct value for the dye-dye separation.

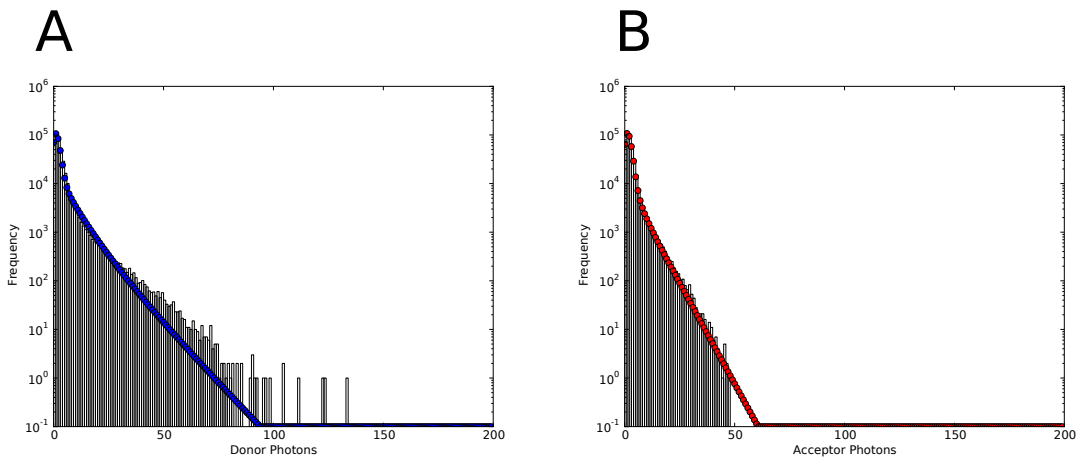

Fig. S - 4: Histograms showing the marginal distributions of donor and acceptor photons in a smFRET dataset (1:1 mixture of 4bp and 12 bp DNA duplexes). (A) Marginal distributuon photons in the donor channel. Histogram shows the number of time bins observed to contain this many donor photons. Blue circles show the number of donor photons predicted by our model, using parameters inferred from the dataset using the inference method. (B) Marginal distributuon photons in the acceptor channel. Histogram shows the number of time bins observed to contain this many acceptor photons. Red circles show the number of acceptor photons predicted by our model, using parameters inferred from the dataset.

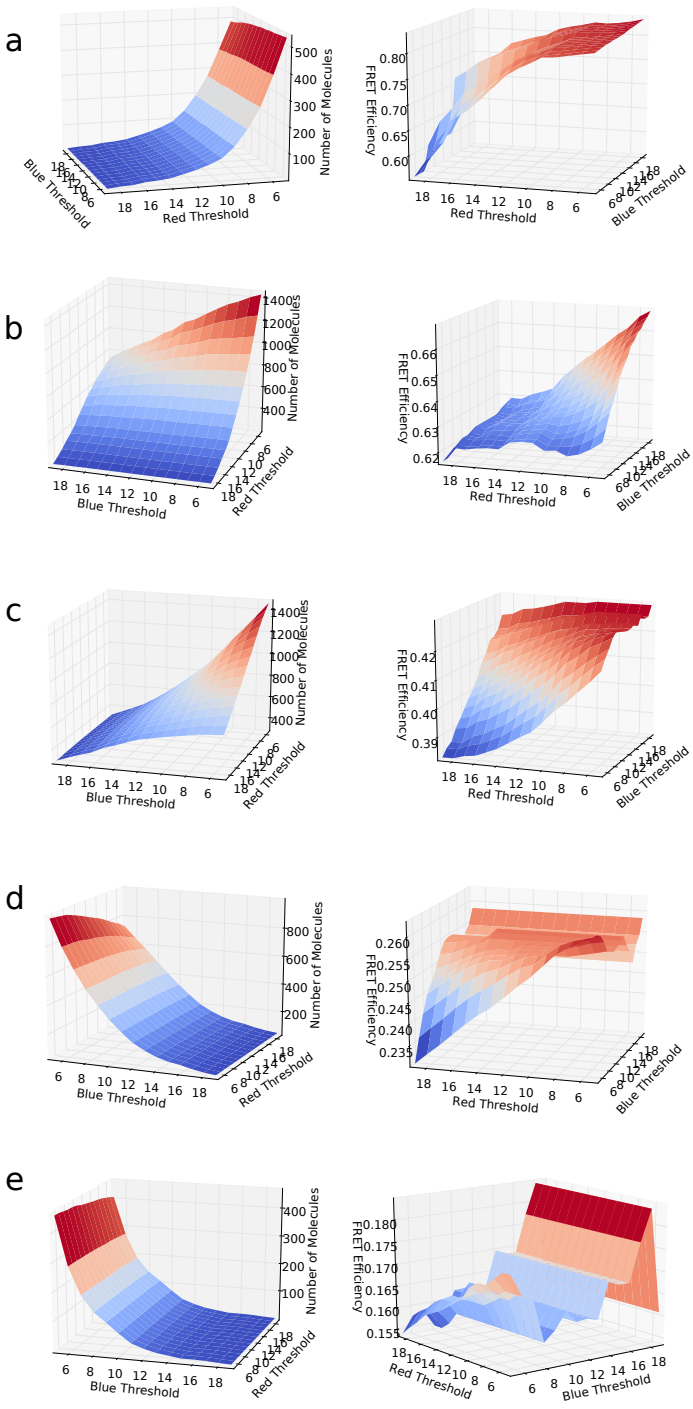

Fig. S - 5: Graph illustrating the effect of threshold choice on the number of molecules detected (left) and the calculated FRET efficiency (right) of synthetic datasets with FRET efficiencies of 0.88 (A), 0.66 (B), 0.40 (C), 0.21 (D) and 0.11 (E) respectively. For all FRET efficiencies, the chosen threshold has a large effect on the number of molecules detected. The threshold also influences the calculated FRET efficiency, with the effect being particularly large for the highest FRET efficiencies.