Thesis of DOOM

Rebecca Roisin Murphy
Department of Chemistry
University of Cambridge
Lensfield Road, Cambridge, CB2 1EW

May 14, 2015

Abstract

My PhD thesis.

Contents

Acknowledgements

Had I the heavens' embroidered cloths,
Enwrought with golden and silver light,
The blue and the dim and the dark cloths
Of night and light and the half-light,
I would spread the cloths under your feet:
But I being poor have only my dreams;
I have spread my dreams under your feet;
Tread softly because you tread on my dreams.
W. B. Yeates

This thesis presents more than three years of work at the University of Cambridge, during which time I have had the privilege to meet and work with many fantastic people. Please allow me to begin by thanking those of you who have supported and encouraged me in my research work – without you, nothing that I achieved would have been possible.

First and foremost, I would like to thank my supervisors, David Klenerman and Sophie Jackson. Thank you both for your continued support and encouragement, despite the somewhat unusual structure and focus of my PhD. Thank you for giving me the freedom to pursue my research interests and for allowing me to find a path that is my own, if somewhat unconventional.

To my colleagues in the Klenerman and Jackson groups, thank you all for making my time in the Chemistry Department so enjoyable. Particular thanks are due to:

> Mathew, for many many invaluable conversations about data, sizing and thresholding. Your enthusiasm for research, your attention to detail and your ability to get things done are all phenomenal. I wish you great success and great happiness in your future

research career. Enjoy your life as an expatriate.

Vladas, you were a fantastic collaborator and Masters student throughout your research project, even though it was somewhat "wetter" than you might have hoped. The Holliday Junctions that you prepared have been invaluable to our research on oligomer sizing and have given the group many useful insights. Thank you also for sharing Marmalade!

Alex, for your amazing skills with hardware description languages and all of your custom FPGA correlators. Without your expertise, none of this research could have been done. Thank you also for your patience in the face of my lack of physics knowledge and your assisstance with signal generators.

Kristina, for your help with TIRF experiments, as well as for many interesting discussions about data analysis and statistics. As a scientist, you are an inspiration.

Magnus, for all your support with pyFRET, in particular your enthusiasm to test new functionality and all your suggestions for improvements. Your encouragement took this from just an idea to something real and made me a much better sofware engineer in the process.

Yu, thank you for many interesting discussions about data, your honesty, and your eventual agreement that actually I am quite clumsy in the lab!

Steve, thank you for many helpful discussions. Your enthusiasm and insight have dramatically improved my understanding of optics, light and physics generally. Your guidance and collaboration helped to bring many ideas from theoretical concept to concrete reality.

Thank you also to everyone who proof-read parts of this thesis. Your constructive comments and helpful advice have been exceedingly useful.

Gratitude is also owing to the Algorithms Group at Illumina, who made my internship there such an interesting and memorable experience. In particular, I would like to thank Ole and Jared for all their patience and understanding, as well as for your support in bringing such an interesting project to reality.

To George, my partner. Thank you for your unconditional support, your infinite love and the best conversation in the world. Thank you for introducing me to the magic that is programming and for your many hours debugging my early spaghetti attempts at writing

code. You have brought more to my life than you can imagine.

To James, my colleague and friend, thanks are owing for many years of great discussions, tea and cynicism. Thanks for your support and for telling me to get my act together.

Finally to my family: Granny, Mum and Peter, thank you all so much for your help and support throughout my life. Thank you for teaching me about the truly important things: kindess, integrity and honesty. And Grandad, thank you for your early enouragement on my scientific journey. Thank you for teaching me about evidence and proof. I hope you would be proud of me today.

Summary

Single molecule fluorescence microscopy describes a number of experimental techniques for the study of individual molecules using fluorescence detection. Single molecule microscopy has found a wide range of research applications. These applications include the study of protein folding, protein aggregation and intermolecular associations. Since the first demonstration of the detection of a single, fluorescently-labelled molecule, considerable effort has been made to improve the quality of the data obtained, through the development of novel experimental methodologies.

However, the development of novel experimental methods requires the concommitant development and rigorous evaluation of analysis tools that can maximise the information available from the data obtained. This thesis describes the development and thorough evaluation of data analsysis methodologies for confocal single molecule fluorescence microscopy. We describe the implementation of an open source software library for the analysis of confocal single molecule fluorescence data, and the systematic evaluation of existing data analysis methodologies. We then describe the development of a novel method of data analsysis, based on Monte Carlo sampling. We apply this analysis methodology to the calculation of intramolecular distances and to the determination of oligomer sizes using confocal fluorescence microscopy. We then describe a thorough evaluation of the performance of these analysis methods. A final chapter describes the development of an error correction tool for genome assemblies generated using fluorescence-based sequencing technology from Illumina.

Overall, the work described in this thesis describes the development and systematic evaluation of computational techniques for the analysis of fluorescence data. We describe the theoretical basis of novel techniques and the deployment and analysis of both novel and existing methodologies. In completing this work, we hope to provide a stable foundation on which further research can be built, both through application of the techniques described here and through their extension to further experimental techniques.

Chapter 1

Introduction

1.1 Introduction

This chapter provides a general introduction to the contextual background of the work presented in this thesis. First, we provide a general overview of the physical phenomena of fluorescence and Förster resonance energy transfer (FRET) and describe how these pheonomena can be applied to the study of biological molecules. We also review the common experimental and analytical techniques used in these research areas. We then present an overview of techniques of statistical analysis, with a particular focus on model-based Bayesain inference. We present common sampling-based techniques for Bayesian statistical analysis and describe why Bayesian inference is a useful tool in the analysis of single-molecule fluorescence data.

1.2 Fluorescence and FRET

Since FRET was first used to measure the distance between two fluorescent dyes on individual molecules bound to a surface [1], single-molecule FRET (smFRET) has become a popular tool to investigate the structure and dynamics of biomolecules, both on a surface and diffusing freely in solution [2, 3, 4]. We open this chapter with a discussion of the physical basis of molecular fluorescence and FRET. We then describe some of the applications of fluorescence in biological research, and describe the constraints and experimental techniques commonly used in single molecule fluorescence microscopy.

1.2.1 The Physical Basis of Fluorescence

Fluorescence Fluorescence is the physical phenomenon by which a molecule that has absorbed a photon of electromagnetic radiation and which is in an excited state relaxes via non-radiative processes, such as vibrational interactions, rotation and translation in its medium and then emits a photon, returning to its ground state [5]. Typically, a fluorescence excitation-emission cycle involves transitions between the singlet ground state (S_0) and first excited state (S_1) of the fluorophore (Fig. 1.1). Owing to the intramolecular transitions occurring prior to emission, the photon emitted is of a lower energy than the photon absorbed. The difference in energy between the absorbed and emitted photons is termed the Stokes shift (Fig. 1.2, [6]) and is a characteristic of the energy levels available to the excited molecule.

The typical timescale for a single fluorescence excitation-emission cycle is approximately $10^{-10} - 10^{-7}$ s. This timescale is dominated by the lifetime of the excited state, (typically $10^{-10} - 10^{-7}$ s); photon excitation and emission processes, as well as vibrational transitions, occur on a timescale that is several orders of magniude shorter (absorption 10^{-15} s, relaxation $10^{-12} - 10^{-10}$ s).

Competing Processes In addition to the S_1 to S_0 transition that results in fluorescence emission, other competing energetic transitions can also occur, resulting in non-radiative relaxation to the ground state S_0, or permanent changes to the electronic structure of the fluorophore (photobleaching). Internal conversion – a non-radiative transition to a lower energy state of the same spin multiplicity – is possible, although the large energy gap between the S_1 and S_0 states makes it much less frequent than the more accessible fluorescent pathway. A more significant competitor is inter-system crossing, involving a non-radiative transition from the first singlet excited state, S_1 to the first triplet excited state, T_1. Although forbidden by spin selection rules, inter-system crossing is facilitated by spin-orbit coupling, allowing it to compete with fluorescence (rate $10^{-10} - 10^{-8}$ s) as a relaxation pathway.

From the T_1 state there are several accessible pathways to the ground state (S_0). Phosphorescence is the radiative decay from T_1 to T_0; alternatively, a second inter-system crossing back to S_1, followed by delayed fluorescence emission to reach the S_0 state is more common. As the lifetime of the T_1 state is long (on the order of $10^{-6} - 1$ s), this can be observed as photoblinking. These two processes are illustrated in Fig. 1.1.

A third relaxation pathway from the T_1 state is photobleaching. If the fluorophore is in a

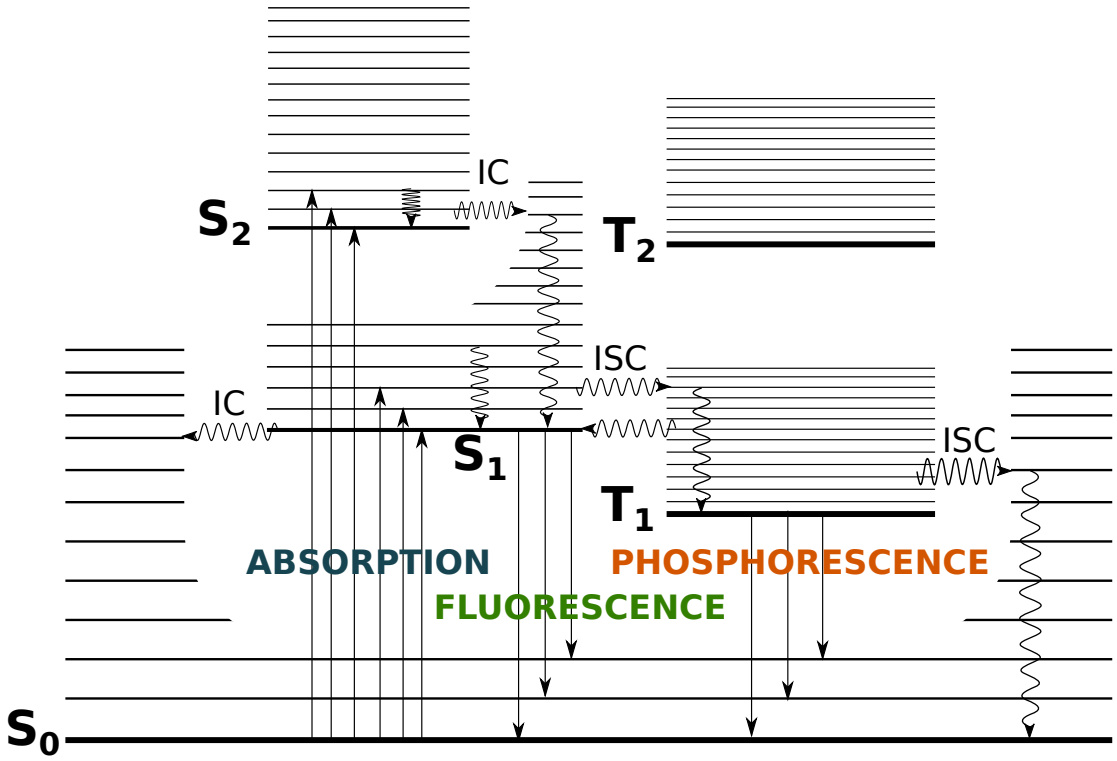

Figure 1.1: Schematic Jablonski diagram depicting the possible electronic processes undergone by an excited fluorophore. Solid lines indicate electronic transitions, wavy lines indicate vibrational transitions. Absorption of a high-energy photon promotes an electron from the ground state S_0 to an excited state, typically S_1. Fluorescence emission occurs during the electronic relaxation from S_1 to S_0. Non-radiative relaxation via internal conversion (IC) or other competing processes, such as inter-system crossing (ISC) to the triplet state T_1, reduce the quantum yield. Figure adapted from [7], with permission.

solution containing dissolved oxygen, the T_1 state can interact with the triplet ground state of molecular oxygen, leading to de-excitation via triplet-triplet annihilation:

$$^3T_1 + ^3O_2 \longrightarrow ^1S_0 + ^3O_2 \tag{1.1}$$

The resultant singlet oxygen can then irreversibly oxidise the fluorophore, leaving it unable to undergo further fluorescent excitation-emission cycles.

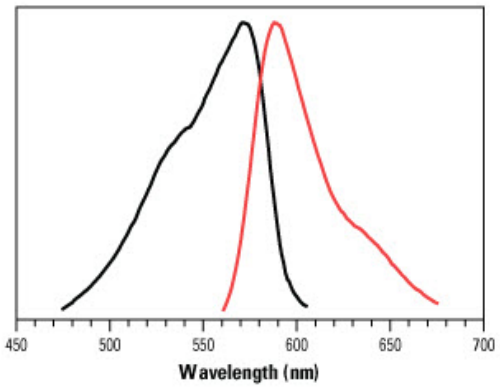

Figure 1.2: The absorption (black line) and emission (red line) spectra for the Quanta Red dye. Due to fast vibrational relaxation, the fluorescence emission is shifter to longer wavelengths than the absorption. This is termed the Stokes shift. Figure from [8]

The Extinction Coefficient and Quantum Yield In choosing a fluorophore for a single molecule fluorescence experiment, it is desirable to maximise the number of successful absorption-emisssion cycles undergone per unit time. Hence, the best fluorophores both absorb many exciting photons, increasing the number of possible cycles, and return to the ground state via fluorescence emission following the majority of excitations, maximising the number of photons actually emitted. Two key parameters that enable quantitative evaluation of fluorescent properties are the molar extinction coefficient, ϵ, and quantum yield, ϕ.

The extinction coefficient, ϵ, measures how strongly a fluorophore absorbs light at a given wavelength λ. It can be calculated for a fluorophore f, from the Beer-Lambert law [9] as:

$$I = I_0 e^{-\epsilon_f l} \tag{1.2}$$

where for an initial excitation intensity I_0, I is the excitation intensity at a distance l into a solution of fluorophore f and ϵ_f is the extinction coefficient of the fluorophore.

Similarly, the fluorescence quantum yield for fluorophore f, ϕ_f, describes the fraction of photon absorption events that result in emission of a photon. It can be defined as:

$$\phi_f = \frac{k_f}{k_f + k_{nr}} \tag{1.3}$$

where k_f and k_{nr} are the rates of fluorescence emission and non-radiative emission respectively. When all molecules in the excited state were promoted by photon absorption, ϕ_f can be equivalently defined as the fraction of absorption events that result in fluorescence emission:

$$\phi_f = \frac{n_{\text{emitted}}}{n_{\text{absorbed}}} \tag{1.4}$$

where n_{emitted} and n_{absorbed} are respectively the number of emitted and absorbed photons.

1.2.2 The Physical Basis of FRET

Förster Resonance Energy Transfer Förster resonance energy transfer (FRET) is a non-radiative energy transfer process that can occur between chromophoric molecules [10]. The degree of energy transfer, E, is dependent on the fluorophore distance, varying inversely with the sixth power of the dye-dye distance, r:

$$E = \frac{1}{1 + (\frac{r}{R_0})^6} \tag{1.5}$$

where, R_0 is the Förster distance, a dye-dependent constant that defines the dye-dye distance for which the energetic transfer efficiency is 50 %. R_0 is defined, for a donor (D) and acceptor (A) fluorophore pair, by their degree of spectral overlap and their relative orientation:

$$R_0 = \sqrt[6]{\frac{9000\phi_D \ln(10)\kappa^2 J(\lambda)}{128\pi^2 n^4 N_A}} \tag{1.6}$$

Here, ϕ_D is the quantum yield of the donor fluorophore, κ is the dipole orientation factor, N_A is Avogadro's number, n is the refractive index of the medium, and $J(\lambda)$ is the spectral overlap integral:

$$J(\lambda) = \int f_D(\lambda)\epsilon_A(\lambda)\lambda^4 \, d\lambda \tag{1.7}$$

where $f_D(\lambda)$ and $\epsilon_A(\lambda)$ are respectively the normalised emission spectrum of the donor and the molar extinction coefficient of the acceptor at wavelength λ. The dipole orientation factor, κ^2 is typically assumed to be $\frac{2}{3}$, the value observed if both dyes are freely rotating

and hence can be assumed to be isotropically oriented during the lifetime of the S_1 excited state [11].

Deriving the FRET Equation The distance dependence of the energy transfer allows FRET to be used as a "molecular ruler" [12], to determine intramolecular distances. For a given FRET enrgy transfer event, the FRET Efficiency, E, defined with respect to the intramolecular distance r in Eq. 1.5, is given ratiometrically by:

$$E = \frac{n_A}{\gamma \cdot n_D + n_A} \tag{1.8}$$

where n_A and n_D are the number of observed photons emitted by the acceptor and donor fluorophores respectively, and γ is an instrument-dependent constant that corrects for unequal detection efficiencies. The equivalence of Eq. 1.5 and Eq. 1.8 is derived below.

As defined above, the quantum yield of a (donor) fluorophore is given by the ratio of the rates of decay from the excited state S_1 via radiative and non-radiative processes:

$$\phi_D = \frac{k_f}{k_f + k_{nr}} \tag{1.9}$$

Similar ratios can be used to describe the quantum yield of a donor fluorophore that can undergo FRET energy transfer (Eq. 1.10) and the FRET efficiency (Eq. 1.11):

$$\phi_{DA} = \frac{k_f}{k_f + k_{nr} + k_{ET}} \tag{1.10}$$

$$E = \frac{k_{ET}}{k_f + k_{nr} + k_{ET}} \tag{1.11}$$

where k_{ET}, the rate constant for the FRET energy transfer process, is given by:

$$k_{ET} = \frac{1}{t_D}\left(\frac{R_0}{r}\right)^6 \tag{1.12}$$

and t_D is the lifetime of the excited state in the absence of an acceptor fluorophore:

$$t_D = \frac{1}{k_f + k_{nr}} \tag{1.13}$$

Combining Equations 1.9, 1.10 and 1.11 yields:

$$\begin{aligned} E &= 1 - \frac{k_f + k_{nr}}{k_f + k_{nr} + k_{ET}} \\ &= 1 - \frac{\phi_{DA}}{\phi_D} \\ &= 1 - \frac{n_{DA}}{n_D} \end{aligned} \tag{1.14}$$

where n_{DA} and n_D are respectively the number of observed photons emitted by the donor fluorophore in the presence and absence of a FRET acceptor at distance r.

In the presence of a FRET acceptor, it is not possible to directly observe n_D. However, this can be derived from the number of photons observed to be emitted by the acceptor fluorophore, n_A. The difference in the number of photons emitted by the donor in the presence and absence of the acceptor is $n_D - n_{DA}$. This is related to the number of photons emitted by the acceptor, n_A, as follows:

$$\frac{n_D - n_{DA}}{n_A} = \frac{\phi_D \eta_D}{\phi_A \eta_A} \tag{1.15}$$

where ϕ_D and ϕ_A are the quantum yields of the donor and acceptor fluorophores respectively, and η_D and η_A are the respective instrumental detection efficiencies of donor and acceptor photons.

If we call the ratio $\frac{\phi_D \eta_D}{\phi_A \eta_A}$ γ, Eq. 1.15 can be re-arranged to give:

$$n_D = n_{DA} - \frac{n_A}{\gamma} \tag{1.16}$$

Substituting this definition of n_D back into Eq. 1.14 yields the expected definition of E, 1.8, that allows the inter-fluorophore distance to be related to an experimentally observed ratio of donor and acceptor photons.

1.3 Single Molecule Fluorescence Microscopy

Both fluorescence and FRET have found many applications in the study of biological systems. Fluorescence microscopy is a well-established field of research, which uses fluorescence

emission to generate an image. Synthetic or naturally occurring fluorophores are conjugated to biomolecules of interest and then excited using monochromcatic light matched to the fluorophore's absorption spectrum [13]. Photons emitted from the fluorophores are collected and used to gain information about the structures and associations of the labelled molecules.

Advances in the fields of optical microscopy, fluorophore synthesis and data analysis have facilitated development of optical techniques that are sensitive enough to detect emission from individual fluorophores. This has created the field of single molecule fluorescence microscopy and the capacity to characterise the behaviour of individual biomolecules. The following section describes the benefits of single molecule fluorescence microscopy, introduces the main optical systems used for single molecule imgaing, and discusses the experimental and analytical challenges involved in a single molecule experiment. As the main experimental technique used in this thesis, we focus our discussion on single molecule confocal spectroscopy; however, for context, we include an overview of other single molecule fluorescent techniques.

1.3.1 Techniques of Single Molecule Imaging

The greatest challenge in single molecule microscopy is to be able to detect dim fluorescence emission from individual molecules against a background of photon emission from other sources, such as Raman scattering by solution molecules and residual fluorescence from solution impurities [14]. Consequently, a large amount of work has been done to maximise the signal-to-noise ratio (SNR) of fluorescent emission data. This ratio can be optimised from both ends; namely maximising the signal and minimising the background noise. Maximising the signal involves the development and selection of fluorophores that achieve many fluorescent emission cycles during an excitation event. This is discussed in Section 1.3.2.

An effective way to minimize the background noise is to reduce the emission detection volume; as the number of photons detected from irrelevant species will be reduced with a smaller detection volume, whereas detection of photons from a single fluorophore of interest within the detection volume will be unaffected [15]. Single molecule fluorescence techniques can hence be characterised based on their method of detection volume reduction: Wide-field techniques, including Epifluorescence Microscopy and Total Internal Fluorescence Microscopy (TIRFM) illuminate and collect data from a wide image plane with a large area but minimal depth; by contrast confocal microscopy techniques attempt to collect data from the smallest detection volume possible – a voxel of volume approximately 1 fL. These techniques and the optical

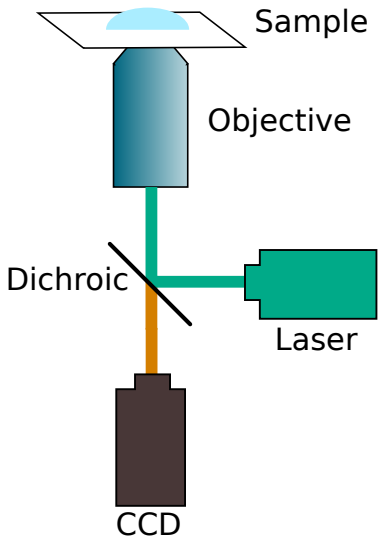

Figure 1.3: Schematic of an epifluorescence microscope. Laser light is focused into a sample using an objective lens with a high N.A. Emitted fluorescence is captured using the same objective, separated from reflected exciting photons using a dichroic mirror and directed onto a CCD.

geometry required for successful data collection, are summarized below.

Epifluorescence Microscopy The simplest geometry for single molecule fluorescence detection is epifluorescence microscopy (Fig. 1.3). In an epifluorescence microscope monochromatic, collimated light is focused into the back apperture of an objective lens with a high Numerical Aperture (N.A.). A large volume of the fluorescently labelled sample, which is placed above the objective, is illuminated; emitted photons are collected through the same objective. Fluorescence emission is separated from excitation light using various filters before being focused onto a charge coupled device (CCD) camera which generates a current at each pixel proportional to the intensity of incident light.

The principal advantage of epifluorescent illumination, in addition to its ease of set-up, is the large illumination area. The detection volume is typically seveal microns in diameter, allowing multiple emitters to be imaged at the same time and to be tracked across multiple image frames [16]. However, because the depth of the illumination volume in epifluorescence microscopy is relatively large, the SNR is quite low: many molecules at different depths into the sample volume are illuminated, resulting in a large amount of out-of-focus fluorescence reaching the CCD camera. Consequently, although epifluorescence microscopes are the most

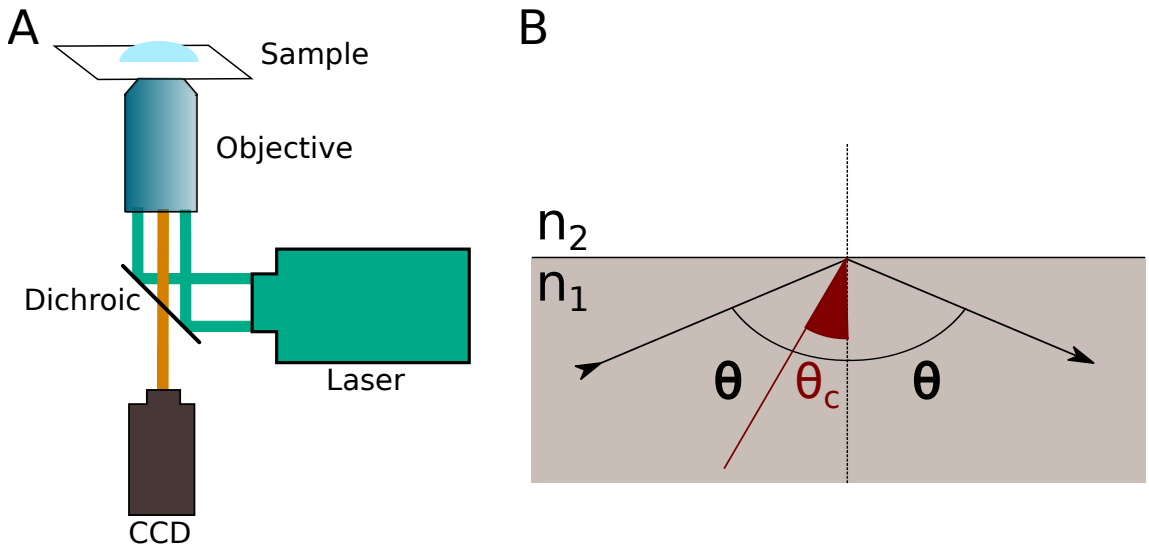

Figure 1.4: A) Schematic of a TIRF microscope. As in epifluorescence, laser light is focused into a sample using an objective lens with a high N.A. Emitted fluorescence is captured using the same objective, separated from reflected exciting photons using a dichroic mirror and directed onto a CCD. However, because the excitation wavelength meets the glass coverslip at an angle exceeding the critical angle, only molecules very close to the coverslip are excited. B) For two materials with refractive indices $n_2 < n_1$, TIR occurs when the incident angle, θ, exceeds the critical angle θ_c.

commonly used variant of the fluorescence microscope, they are less useful for single molecule detection, as the SNR is too low to efficiently resolve fluorescence emission from individual molecules.

Total Internal Fluorescence Microscopy Total Internal Fluorescence Microscopy (TIRFM) (Fig. 1.4 A) uses a similar optical geometry to epifluorescence microscopy: as in epifluorescence imaging, the sample is illuminated via the back port of an objective and fluorescence emission is collected via the same objective and focused onto a CCD camera. However, TIRFM exploits the phenomenon of Total Internal Reflection (TIR) to reduce the depth of the illumination volume and hence increase the SNR.

TIR occurs when a light wave passing through a medium with refractive index n_1 meets a dielectric medium with refractive index $n_2 < n_1$ with an incident angle θ, where $\theta > \theta_c$, the critical angle (Fig. 1.4 B). Under these conditions, rather than some of the light being refracted at the boundary between the two media, all of the light is reflected internally.

As a result of the TIR, an evanescent wave is generated at the boundary. This wave is propagated along the boundary, but decays exponentially with distance from the boundary. The distance that the wave penetrates beyond the boundary is dependent on the incident angle θ, decreasing as θ increases.

The critical angle, θ_c is determined using Snell's Law [17] using the ratio of the refractive indices n_1 and n_2:

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \tag{1.17}$$

For a typical borosilicate coverslide ($n = 1.518$) in contact with water ($n = 1.33$), a TIRF angle of 61.2° is required, enabling fluorescence excitation to be limited to a depth of ~ 100 nm. The narrow excitation depth of TIRFM allows all of the advantages of epifluorescence (specifically large excitation area and particle tracking), but increases the SNR sufficiently that fluorescence emission from individual molecules can be effectively resolved.

Confocal Microscopy Confocal microscopy takes the reduction in excitation volume a further step, minimizing the detection volume in all dimensions and collecting fluorescence emission from a solution volume of ~ 1 fL. In confocal microscopy (Fig. 1.5), a collimated laser beam is focused through an infinity corrected objective with a high N.A. onto a diffraction limited spot (diameter $> 1\mu$m) a few microns into the sample volume. Fluorescence emission from the sample is collected and recollimated by the same objective, passed through a dichroic mirror to remove reflected incident light and then directed through a small aperture pinhole (diameter $50 - 100\mu$m) placed in the conjugate focal plane to remove out-of-focus fluorescence. The remaining photons are directed onto an Avalanche Photodiode (APD) detector, which can detect individual photons with nanosecond resolution. To enable detection of fluorescence from individual fluorescent molecules, analyte solutions are used at extremely low concentrations (typically $10 - 100$ pM), such that the probability of more than one molecule occupying the confocal volume simultaneously is extremely small. Single molecule confocal microscopy is typically used in this manner to image labelled molecules diffusing freely in solution [4] or flowed rapidly through the exciting laser beam [18]; in conjunction with a moveable sample stage or movable laser mount, it can also be used to image surfaces and surface-tethered molecules [19].

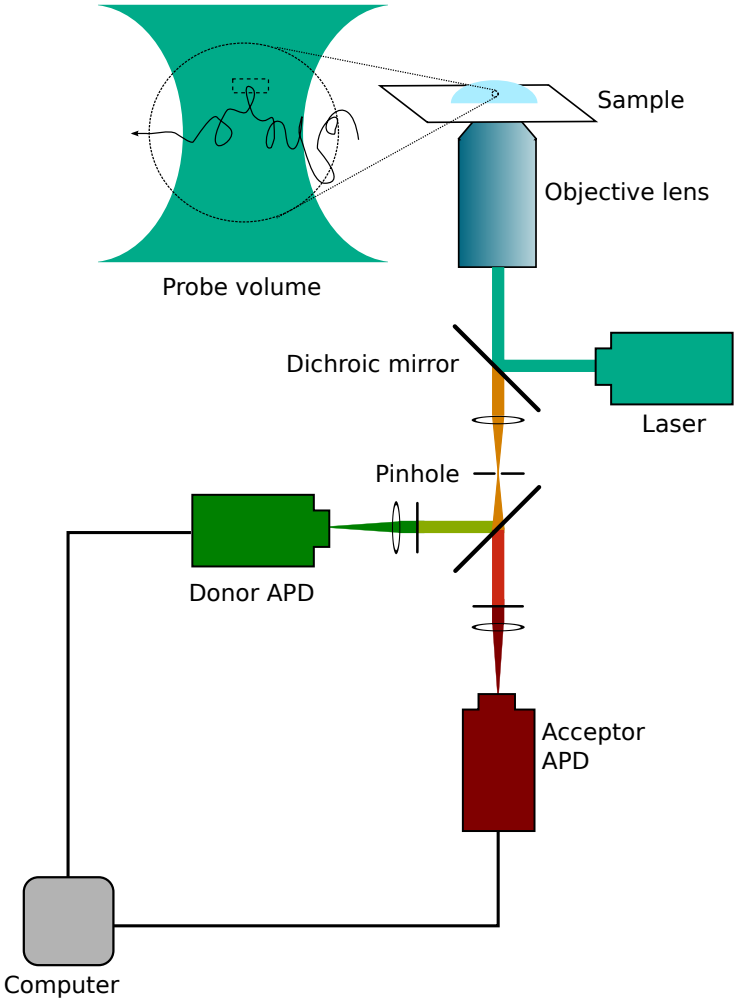

Figure 1.5: Schematic of a confocal microscope. As in epifluorescence, laser light is focused into a sample using an objective lens with a high N.A. Emitted fluorescence is captured using the same objective and separated from reflected exciting photons using a dichroic mirror. The addition of a pinhole enables removal of out-of-focus light, improving the SNR as emitted fluorescence is now collected from only the fL-scale confocal volume. A second dichroic separates the remaining photons into donor and acceptor wavelengths. These photons are then directed onto APDs for counting and anlysis.

Fluorescence Correlation Spectroscopy Fluorescence Correlation Spectroscopy (FCS) [20] is an extension of the Confocal method described above. In FCS, a higher concentration ($1-100$ nM) of analyte solution is used, such that multiple fluorophores occupy the confocal volume an any time. Temporal fluctuations in the detected fluorescence emission can be used to determine physical paramters of the analyte solution, such as molecular concentrations and diffusion constants. These are determined by fitting a fluorescence autocorrelation function (Eq. 1.18),

$$G(\tau) = \frac{\langle \delta F(t) \cdot \delta F(t+\tau) \rangle}{\langle F(t)^2 \rangle} \tag{1.18}$$

where $F(t)$ and $\delta F(t)$ are the absolute fluorescence intensity and the fluctuation about the mean intensity at a time t, respectively, using a three-dimensional diffusion model corrected for photoblinking effects (Eq. 1.19):

$$G(\tau) = \left(1 - T + T\exp(\frac{-\tau}{-\tau_T})\right) \frac{1}{N} \left(\frac{1}{1+\frac{\tau}{\tau_T}}\right) \left(\frac{1}{1+(\frac{\omega}{z_0})^2} \frac{t}{\tau_D}\right)^{0.5} \tag{1.19}$$

Here, N is the average number of molecules residing in the confocal volume, τ_D is the characteristic residence time of the molecules, ω is the beam-waist and z_0 is the length of the measurement volume. The term $\left(1 - T + T\exp(\frac{-\tau}{-\tau_T})\right)$ corrects for photoblinking caused by population of the triplet excited state; T is the fraction of molecules occupying the triplet state and τ_T is the triplet state lifetime.

FCS is used in Fluorescent Cross Correlation Spectroscopy [21] to determine the stoichiometry of molecular complexes; it is also used to study chemical reaction processes such as protein aggregation [22, 23] and to study the behaviour of fluorescently labelled molecules inside living cells [24].

1.3.2 Fluorophores for Single Molecule Imaging

All of the techniques of single molecule fluorescent microscopy described thus far are reliant on the availability of fluorophores that are bright enough that emission from individual fluorophores can be detected against background noise. This requires the fluorophores to undergo sufficient cycles of excitation and fluorescent emission that a large number of photons can be detected. Several factors affect the suitability of fluorophores for this purpose.

As described above (Section 1.2.1), to maximise the number of excitation-emission cycles, a fluorophore should have both a high extinction coefficient and a high quantum yield, to maximise the number of photons emitted per unit time. The lifetime of the excited state should also be short, to increase the number of excitation-emission cycles per unit time. Furthermore, the fluorophore should have a high extinction coefficient at the excitation wavelength, to maximise the number of photon absorption events. Furthermore, the fluorophore should be resistant to both permanent photobleaching and to transient photoblinking, to prevent premature attenuation of fluorescence emission. These considerations are particularly pertinent for fluorophores used in confocal techniques, where the number of photons observed from each fluorophore is additionally limited by the dwell-time of the molecule in the confocal volume (typically less than 1 ms).

Considerable effort has been made to synthesize fluorophores that fulfil these requirements for the excitation wavelengths accessible to single molecule microscopy. Many of the most popular dyes are derived from fluorone (Fig. 4.20): fluorescein, the fluorophore used in the first demonstration of single molecule detection [25] and the green-excited Rhodamine 6G are both built around fluorone's chromophoric heterocyclic structure; functionalised derivatives of these molecules form the basis of a range of blue- and green-excited fluorophores in the commercial Alexa Fluor and Atto ranges (Fig. 1.6). Red-excited fluorophores are typically derived from cyanine structures, although the polymethine backbone makes them susceptible to photobleaching via cis/trans isomerisation. To address this issue, conformationally locked derivatives, such as the commercially available Atto647N, have been developed; however both locked and unlocked cyanine derivatives are prone to photoblinking via population of long-lived triplet states [26].

1.4 Confocal Microscopy

Thus far, we have given a general overview of the different applications of fluorescence in single molecule imaging. Howerver, this thesis primarily focuses on the analysis of data single molecule confocal microscopy. Hence, we shall now give a more detailed description of the challenges and experiemntal methodologies specific to this single molecule imaging technique.

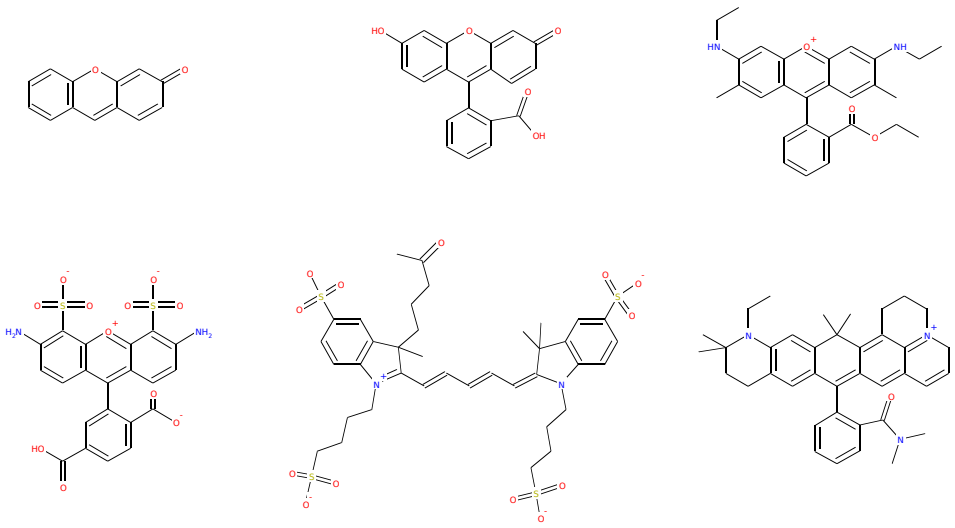

Figure 1.6: Chemical structures of some common fluorescent dyes. Top row, left to right: fluorone forms the basis of many flurorescnent dyes; fluorescein was used in the first demonstration of single molecule detection; Rhodamine 6G was XXX? Bottom row, left to right: the commercially-available dyes Alexa Fluor 488; Alexa Fluor 647; and the conformationally locked variant Atto 647N.

1.4.1 Data Acquisition and Analysis

Single Molecule FRET The most common form of confocal single microscopy exploits the phenomenon of FRET (Section 1.2.2) to determine intramolecular distances in a single molecule FRET (smFRET) experiment. In such an experiment, each molecule is site-specifically labelled with two fluorescent dyes, a donor (D) and an acceptor (A). The dyes are chosen such that the emission spectrum of the donor overlaps with the excitation spectrum of the acceptor, enabling FRET emission to occur when the two dyes are sufficiently close in space.

A collimated laser beam, matched to the excitation spectrum of the donor, is used to illuminate an extremely dilute solution of these labelled molecules. When a labelled molecule diffuses through the laser beam, the fluorophore is excited and photons are emitted. Emitted photons are collected, separated from excitation photons, recollimated, then further separated into donor and acceptor wavelengths using a dichroic mirror and directed onto two APD detectors for collection and analysis (Fig. 1.5).

In the simplest smFRET experiment, all photons reaching the detectors during data acquisition binned as they are received into time-bins of length similar to the expected dwell-time

of a molecule in the confocal volume. Fluorescent events are identified as the time-bins that contain sufficient photons to meet a specified criterion [4]. Alternatively, to avoid missing or double counting fluorescent events that are split over more than one time-bin, photons can be binned a time-scale much shorter than the typical dwell time. A burst search algorithm [27] is then used to scan the resultant photon stream for bursts of a specified duration and brightness. Details of these two methods of event selection are provided in Chapter 2.

Identified fluorescent events are then used to generate FRET histograms: the FRET efficiency of each event is calculated as described above:

$$E = \frac{n_A}{n_A + \gamma \cdot n_D} \tag{1.20}$$

Histograms of the observed FRET efficiencies (Fig. 1.7 A) can be fitted with gaussian distributions to calculate approximate intramolecular distances and to identify conformational sub-populations that might have been invisible to an ensemble-based technique [28, 29].

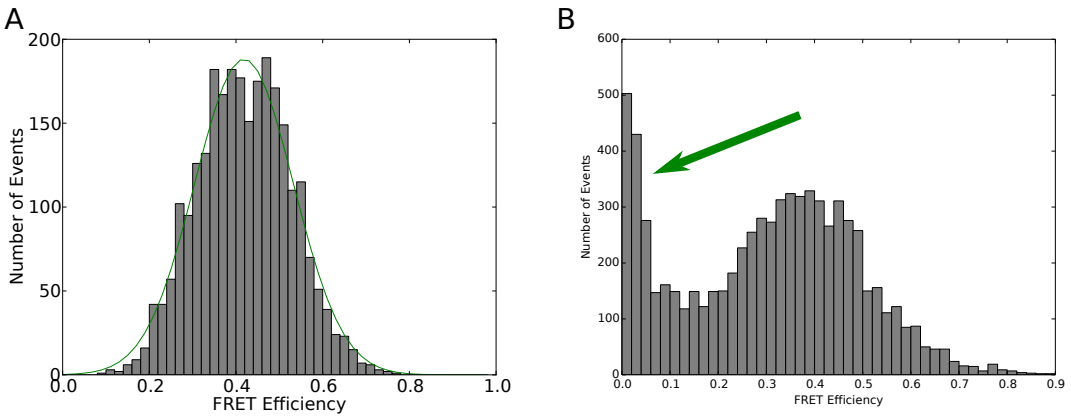

Figure 1.7: Example FRET histograms from dual-labelled DNA duplexes. A) Following event selection and FRET Efficiency calculation, a histogram of the FRET efficiencies is prepared and fitted with a gaussian distribution. Parameters of the fit give information about intramolecular distances and population heterogeneity. B) a FRET histogram that contains a zero peak artifact, indicated by the green arrow, caused by improperly labelled molecules or acceptor photobleaching.

Direct Acceptor Excitation One drawback of the experimental technique summarized above is that the acceptor fluorophore can only be observed indirectly via FRET energy

transfer from an excited donor fluorophore. Acceptor photobleaching and imperfect labelling of the analyte molecules can therefore lead to experimental artifacts such as the zero peak (Fig. 1.7 B), which complicate or distort downstream analysis. As a consequence, several experimental modifications of the simple FRET experiment have been developed to enable direct observation of the acceptor fluorophore.

One simple modification is Two Colour Coincidence Detection (TCCD) [30, 31]. In this technique, two collimated, overlapped lasers are used to simultaneously excite both the donor and the acceptor fluorophores, enabling observation of the acceptor fluorophore even when the donor fluorophore is not present or when the inter-fluorophore distance is too great for FRET to occur. TCCD has been used in conjuction with simple FRET to study protein aggregation [32] and molecular conformation [33]. It is useful for identifying multimeric species, but extracting molecular distance can be challenging, as direct acceptor excitation cannot be simply separated from FRET emission [34].

A more complex modification uses Alternating Laser Exciation (ALEX). In an ALEX experiment the diffusing molecules are excitated by two lasers in rapid alternation [35]. One laser excites the donor fluorophore and the other can directly excite the acceptor fluorophore. The modulation is fast on the timescale of molecular dwell-time, allowing a single fluorescent molecule to receive multiple cycles of both donor excitation and direct acceptor excitation. The additional information obtained from direct excitation of the acceptor fluorophore allows fluorescent events originating from improperly labelled molecules to be excluded from further analysis.

A further modification of this ALEX technique is Periodic Acceptor Excitation (PAX) [36]. PAX is a simplified version of ALEX, in which donor excitation is continuous, but there is rapid modulation of the acceptor excitation power. As in an ALEX experiment, this enables information about the molecular labelling state to be determined, however the experimental setup is simpler as rapid modulation is required for only one of the two lasers used.

Application to Oligomer Sizing As described above, a smFRET experiment is used to obtain information about intramolecular distances. However, single molecule confocal spectroscopy can also be applied to the problem of determining the size and stoichiometry of oligomeric species [31, 32]. In these experiments, individual monomer molecules are labelled with a single fluorophore. Following some period of time during which the aggregation reaction proceeds, the reaction mixture is diluted to picomolar concentrations and subjected

to TCCD illumination. Oligomer size and stoichiometry is then calculated by comparing the brightness of observed fluorescent events with the brightness of a labelled monomer:

$$\text{Size} = 2 \cdot \left(\frac{n_D + \gamma^{-1} n_A}{\langle n_{\text{monomer}} \rangle} \right) \tag{1.21}$$

where n_D and n_A are the number of observed donor and acceptor photons, γ is the instrumental gamma-factor described above and $\langle n_{\text{monomer}} \rangle$ the mean number of photons observed in fluoreoscence events from labelled monomeric species.

A further extension of this sizing technique uses microfluidic flow [37] combined with increased excitation intensities to increase the data collection rate. Automated dilution can also be used to rapidly dilute analyte solutions by up to six orders of magnitude [38]. This enables unstable mulitmeric complexes with weaker binding affinities to be successfully analysed using smFRET before dissociation occurs.

1.4.2 Computational Challenges

A confocal smFRET experiment, as described above, presents several computational challenges during data analysis. Based on an experimentally observed stream of observed photon counts, analysis should be able to determine many properties of the analyte and experiment, such as the number and concentration of fluorescent populations, the intramolecular distances between the dye attachment sites and, in an aggregation experiment, the stoichiometry of a multimeric complex. The following chapter (Chapter 2) describes implementation of the most common methods for this form of data analsysis, involving deterministic event selection algorithms, simple subtraction based denoising, and the construction and fitting of FRET histograms to determine the number of fluorescent populations. Similar methods for oligomer sizing are briefly discussed in Chapter 4.

However, as photon emission by fluorophores is inherently a probabilistic phenomenon, these problems are also well-posed as problems of probabilistic inference: given an observed sequence of fluorescent bursts, what are the most likely properties of the experimental system that generated that dataset. A considerable part of this thesis deals with the application of the tools of probabilistic analysis to the analysis of smFRET data, both to infer intramolecular distances (Chapter 3) and to oligomer sizing (Chapter 4). Consequently, we now give a brief overview of the research discipline of Bayesian inference and outline the sampling

techniques used in this work.

1.5 Probabilistic Inference and Bayesian Statistics

This section introdeces the core concepts of probability theory and relates them to the analysis of smFRET datasets. We first introduce the concepts of model-based inference and Bayesian statistics. We then demonstrate how these concepts can be used to describe the physical processes underlying a smFRET experiment. We show that this describes a physical model of the smFRET experiment. Next, we introdue the sampling-based techniques that can be used to infer the parameters of these models in order to determine information from the experiments performed.

1.5.1 Probabily Theory

Probabilty Probability theory is the study of stochastic processes and the mathematical functions, termed probability distributions, that describe them. In probability theory, a variable whose value is subject to chance variations is termed a random variable. A discrete random variable, such as the outcome of a die roll or a coin flip, can take one of a fixed number of values and is described by a probability mass function (PMF); a continuous random variable can take any of a continous range of values and is described by a probability density function (PDF) [39, 40].

In the field of single molecule fluorescence, the number of photons emitted by a fluorophore undergoing continuous excitation can be described as a discrete random variable. It is discrete since the number of photons emitted must be integral: it is unphysical to consider fractions of a photon. Photon emission can therefore be modeled using a discrete probability distribution. The most common choice is the Poisson distribution [41]. The Poisson distribution, given by the below equation [39]:

$$\Pr(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} \tag{1.22}$$

for random varaible X and integer values of x, describes the probability of a given number of events occurring in a fixed time interval, so is a good model for photon emission.

Conditional Probability and Bayes Theorem Conditional probability is the probability that some event occurs, conditioned on the value of some other observation. From set theory, the conditional probability of event A, given event B, written $\Pr(A \mid B)$, is given by:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{1.23}$$

This can be simply rearranged to give an expression for the probability of both events A and B occurring:

$$\Pr(A \cap B) = \Pr(A \mid B) \cdot \Pr(B) \tag{1.24}$$

An analagous expression can be derived from the probability of event B given event A:

$$\Pr(A \cap B) = \Pr(B \mid A) \cdot \Pr(A) \tag{1.25}$$

A simple substitution and rearrangement allows the derivation of Bayes' theorem, which relates the conditional probability $\Pr(A \mid B)$ to the conditional probability $\Pr(B \mid A)$ [42]:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)} \tag{1.26}$$

Bayesian Statistics Bayes Theorem (Eq. 1.26) provides the basis for Bayesian inference – a method of statistical hypothesis testing. In Bayesian inference, the conditional probability $\Pr(A \mid B)$ is interpreted as a hypothesis, or model, the likelihood of which is evaluated based on observational evidence. Rewriting Eq. 1.26 as the probability of a specific, parametric, model given some experimental observations,

$$\Pr(\text{model}|\text{data}) = \frac{\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})}{\Pr(\text{data})} = \frac{\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})}{\int_{\forall \text{parameter values}} \Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})} \tag{1.27}$$

we see that it is possible to use a forward, generative model of an observed dataset, $\Pr(\text{data}|\text{model})$, coupled with a prior, $\Pr(\text{model})$, that describes appropriate values for the model parameters,

to infer the posterior probability, Pr(model|data), of the model paramters [42]. This relationship, which forms the basis of Bayesian statistics, allows parameter values of a parametric model to be inferred conditioned on observed experimental evidence [43].

1.5.2 Monte Carlo Sampling

For an extremely simple model, it may be possible to compute the solution to Eq. 1.27 exactly, by enumerating all possible states of the model. However, computing the normalisation factor in the denominator requires summing over all possible values of parameter of the model. Consequently, it quickly becomes computationally intractable either to derive an analytical expression for the denominator, or to compute it numerically.

To overcome this issue, sampling methods have been developed, which allow estimation of the distribution of values taken by the model parameters. To estimate the distribution of parameter values, it is sufficient to draw parameter samples distributed proportionally to the posterior distribution, $\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})$ [44]. The mean, variance and quantiles of these samples can be used to estimate the parameter values.

Monte Carlo Markov chain (MCMC) algorithms are a class of algorithms that facilitate sampling from the posterior probability over the model parameters, $\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})$. A Markov chain is a memoryless random walk [45], that transitions between different, enumerable, states in a manner that depends only on the state currently occupied. This is termed the Markov property [46]:

$$\Pr(X_{n+1} = x_{n+1}|X_0 = x_0, \ldots, X_n = x_n) = \Pr(X_{n+1} = x_{n+1}|X_n = x_n) \tag{1.28}$$

A Markov Chain Monte Carlo algorithm constructs a Markov chain that has as its equilibrium distribution the desired posterior distribution $\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})$, but from which samples can easily be made [47]. Running such a Markov chain for many thousands of iterations allows it to reach equilibrium; once this state has been reached, sampling further values with long inter-sample intervals is equivalent to drawing independent samples of the model parameter values, conditioned on the observed data, allowing inference of the distribution of parameter values.

The following sections provide a brief overview of some of the most widely used MCMC algorithms.

Metropolis Hastings Sampling The Metropolis-Hastings algorithm [44], an extension of the earlier Metropolis Algorithm [48], is a sampling algorithm for approximating a probability distribution, $\Pr(x)$, by sampling from a distribution, $f(x)$ that is propoprtional to the probability density of $\Pr(x)$. As with all MCMC sampling methods, Metropolis-Hastings sampling iteratively samples parameter values such that, as more samples are made, the distribution $f(x)$ comes to approximate $\Pr(x)$ more and more closely [44].

At each step of the Metropolis-Hastings algorithm, one of the model parameters, x^i is selected at random and its value modified by a small amount. The probability density $f(x, x^i)$, which is proportional to $\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})$, is evaluated using both the old and the new value of the parameter x^i and the new parameter value is accepted with probability proportional to the acceptance ratio α, defined as [49, 42]:

$$\alpha = \frac{f(x, x^i_{\text{new}})}{f(x, x^i_{\text{old}})} \sim \frac{\Pr(\text{data}|\text{model}, x^i_{\text{new}}) \cdot \Pr(\text{model}, x^i_{\text{new}})}{\Pr(\text{data}|\text{model}, x^i_{\text{old}}) \cdot \Pr(\text{model}, x^i_{\text{old}})} \tag{1.29}$$

Using this definition of the acceptance rate, α, x^i_{new} is always accepted if it improves the probability that the observed dataset was generated by the model parameters including x^i_{new} ($alpha \geq 1$); accepting x^i_{new} with probability proportional to α ($0 < \alpha < 1$) even when it makes the models worse prevents the sampler from getting stuck in a local optimum. Sampling in this manner allows the Markov chain to converge towards a better estimates of the model parameter values, by sampling from high-density regions of $\Pr x$.

Gibbs Sampling Gibbs sampling is a MCMC algorithm that is, in its simplest implementation, closely related to the Metropolis-Hastings algorithm. In Gibbs sampling, starting from a parametric model with a random initialisation of all parameter values, the sampling algorithm cycles through each parameter and draws a new sample value from a probability distribution, $\pi(x)$, conditioned on the current values of all other parameters [47]:

$$x_i \sim \pi(x_i \mid x_0, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) \tag{1.30}$$

In contrast to the Metropolis Hastings algorithm, where a new sample value is accepted dependent on the value of the acceptance ration, α; in Gibbs sampling the new sample value is always accepted [50]. Multiple iterations of this cycle produces a Markov chain whose stationary distribution, as for other MCMC sampling techinques, is the the desired posterior distribution $\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})$ [51].

Slice Sampling A third method of MCMC sampling is slice sampling. Slice sampling samples approximates the value of a random variable by sampling points uniformly from the are under the curve of its probability distribution function and then storing the x-coordinates of the samples obtained [52]. To approximate a univariate random varaible distributed according to some distribution, $f(x)$, the slice sampling algorithm is initialised with a random value of x, x_0, for which $f(x_0) > 0$. An auxiliary variable, y is then chosen by drawing from the uniform distribution $U(0, f(x_0))$. Drawing a horizontal line across the distribution $f(x)$ at the value y defines a horizontal slice, above which $f(x) > y$. An interval $I = (x_L, x_R)$ is defined such that the horizontal slice lies within I and a new value of x, x_1 is chosen by sampling uniformly from I. This defines a Markov chain that converges to a uniform distribution over the area under the curve of the desired probability distribution [52]. Extensions of the slice sampling procedure to infer the values of multiple parameters can be achieved either by applying the univariate sampling algorithm to each variable in turn; or by sampling directly from the multivariate distribution using hyperrectangles to define the sample space for uniform sampling [52].

Considerations for MCMC Sampling In order to obtain valid estimates of model parameters, several important aspects of the MCMC process must be considered. Firstly, to avoid perturbation of the stationary distribution away from the desired equilibrium distribution (namely the probability distribution $\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})$), it is necessary that the property of Detailed Balance is maintained [53]. Detailed Balance refers to the microscopic reversibility of the Markov chain, such that the probability of being in some state x and transitioning to some other state x' must be equal to the probability of being in state x' and transitioning to state x:

$$\Pr(x) \cdot \Pr(x \to x') = \Pr(x') \cdot \Pr(x' \to x) \tag{1.31}$$

This consideration is of particular importance for both Metropolis-Hastings and Gibbs sampling, where failure to select proposal distributions such that detailed balance is maintained may lead to the sampler not converging to the correct stationary distribution [54].

Two further important considerations are the related concepts of burn-in and sample thinning. In MCMC sampling, the Markov chain is typically intialised with random values that result in a model that approximates very poorly the posterior distribution $\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})$. Hence, in sampling parameters using Markov chain sampling, it is necessary to

run the sampler through many sampling iterations before beginning to store parameter estimates, in order to ensure that the initial values have been forgotten and the Markov chain has reached equilibrium. Similarly, once equlibrium has been reached, parameter samples that are drawn in quick succession will be correlated, as the Markov chain explores only slowly the parameter space. Therefore, in order to obtain many independent samples from the stationary distribution, it is important to "thin" the samples made by storing samples on only a small fraction of iterations, in order to ensure samples are taken from the whole parameter space and are not correlated with each other [50].

1.6 Thesis Objectives

This thesis describes the development of software for the analysis of fluorescence data. We implement and rigorously evaluate standard algorithms for the analysis of smFRET data. We then describe the implementation and evaluation of novel techniques for the analysis of smFRET data using sampling-based probabilistic inference. Finally, we apply methods of probabilistic analysis to the problem of error correction in genome assemblies. Although this research was undertaken primarily to explore analysis methodologies appropriate to smFRET data, the methods and implementations described here have found considerable application, both within the Klenerman research group and beyond.

The primary objectives of this thesis are as follows:

1. To develop open-source software for the analysis of confocal smFRET data and to evaluate the performance of different algorithms for data analysis, in order to facilitate reproducible research practises within the smFRET research community and ensure best-practice data analysis.

2. To develop data analysis tools for confocal smFRET data that use model-based probabilistic inference and to evaluate the performance of these tools.

3. To understand the relationships between oligomer size and photon emission for fluorescently labelled oligomers, in order to improve the algorithms used for determination of oligomer size.

4. To develop and release tools for error correction in *de novo* genome assemblies constructed using reads from fluourescence-based sequencing technologies.

Each chapter in this thesis is presented in a stand-alone manner, to accurately reflect their relative independence. In each chapter, we first provide a more detailed overview of the specific context within which the research was performed. We then introduce any required theoretical concepts and experimental techniques, before presenting the relevant results and appropriate conclusions. The rest of this thesis is structured as follows.

Chapter 2 introduces the pyFRET library, which we developed for analysis of confocal smFRET data. We describe the theory and implementation of different analysis algorithms for smFRET datasets. Data from both continuous and alternating excitation experiments are considered. In the second part of Chapter 3, we provide a comprehensive evaluation of different smFRET analysis algorithms, using a combination of simulated and experimental datasets. We benchmark popular analysis techniques, demonstrating their relative utility under different data collection regimes.

Chapter 3 considers a Bayesian method for the analysis of data from continuous excitation smFRET datasets. First, we introduce a model-based theory of the smFRET experiment. Then, we describe how this parametric model can be used to infer intramolecular distances and population sizes from time-binned smFRET data. We benchmark our Bayesian analysis technique against thresholding techniques used for time-binned data, showing the superior performance of the inference technique.

Chapter 4 extends this Bayesian analysis to sizing of labelled protein aggregates. We describe how a simplified model can be used to describe photon emission from multiple fluorophores. Using a combination of real and simulated datasets, we then show that there is a complex, non-linear relationship between aggregate size and the number of photons emitted in a fluorescent event, making inference of aggregate sizes extremely challenging. We explore the sources of this emission heterogeneity and suggest a novel excitation regime that could mitigate some of these sources of error.

Chapter 5, the final results chapter, is somewhat different. This chapter describes a Bayesian analysis tool for error correction in genome assemblies generated using Illumina Nextera mate pairs. Illumina sequencing technology uses fluorescence emission from multiple fluorophores in its base calling algorithm. However, base calling errors and complex repeat structures can make reassembly of short reads challenging. In this chapter, we introduce NxRepair, an error correction tool that can identify mistakes in *de novo* assemblies of bacterial genomes. We benchmark NxRepair against existing tools, demonstrating its superior performance. Although not specifically related to smFRET data, this tool falls into the remit of this

thesis, as it applies similar techniques of probabilistic analysis to a problem encountered in downstream analysis of fluorescence data.

Finally, Chapter 6 provides the conclusion to the thesis. Here, we summarise the overall contribution of this thesis and relate the work described to its wider research context. We also discuss possible extensions of the research described, indicating future applications of the research. Overall, the work presented in this thesis describes a rigorous evaluation of the analysis methodologies used in an experimental discipline, identifies several shortcomings and implements novel tools to mitigate overcome these issues. We hope that this work proves useful and informative to other researchers working in this field.

Chapter 2

Analysis Tools for Single Molecule Confocal Microscopy

2.1 Overview

This chapter describes the development of pyFRET, an open source library of analysis tools for confocal single molecule spectroscopy. The chaper is structured as follows. Firstly, we describe the analysis algorithms supported by pyFRET, including their theoretical basis and programmatic implementation. Next, we describe the specifics of their implemenation and deployment using pyFRET. Finally, we compare the performance of different analysis techniques, as implemented using pyFRET, to understand their strengths.

The contributions of this chapter are twofold. Firstly, pyFRET is the first open source software released for smFRET data analysis. This is important as open software facilitates effective benchmarking and comparison of different analysis methods. Secondly, we present the first detailed comparison of different analysis techniques for confocal smFRET. The data collection and analysis methods used by different research groups are currently highly heterogeneous, so this comparison makes an important contribution to our understanding of the best techniques to use for accuarate evaluation of smFRET data.

2.2 Introduction

2.2.1 The Single Molecule Fluorescence Experiment

As described in Chapter 1, in a confocal smFRET experiment, molecules are labelled with two fluorescent dyes. The emission spectrum of the donor dye (D) is chosen to overlap with the excitation spectrum of the acceptor (A). When the donor and acceptor are sufficiently close in space, exciting the donor dye results in FRET and fluorescent emission from the acceptor dye. The FRET efficiency, E, which describes the proportion of excitation energy transferred from the donor to the acceptor, depends on the distance, r between the two dyes (Eq. 2.1) and R_0, the Förster distance, a dye dependent constant that describes the dye separation at which 50% energy transfer is achieved (Fig. 2.1 C)

$$E = \frac{1}{1 + (\frac{r}{R_0})^6} \tag{2.1}$$

Consequently, the distance between the two fluorophores can be determined from the ratio of donor and acceptor photons emitted during an excitation event (Eq. 3.2).

Experimentally, a collimated laser beam, is used to illuminate an extremely dilute solution of labelled molecules. When a labelled molecule diffuses through the laser beam, the donor dye is excited and photons are emitted from both donor and acceptor dyes. Emitted photons are collected through the objective and separated into donor and acceptor streams for collection and analysis (Fig. 2.1 A, B).

For a single fluorescent burst, the FRET efficiency, E, can be calculated as (Eq 3.2):

$$E = \frac{n_A}{n_A + \gamma \cdot n_D} \tag{2.2}$$

for n_A and n_D detected acceptor and donor photons respectively and γ an experimentally determined instrument-dependent correction factor. During the course of a smFRET experiment, several thousand fluorescent bursts are collected and used to construct FRET efficiency histograms, which cab then be used to identify populations of fluorescent species [1].

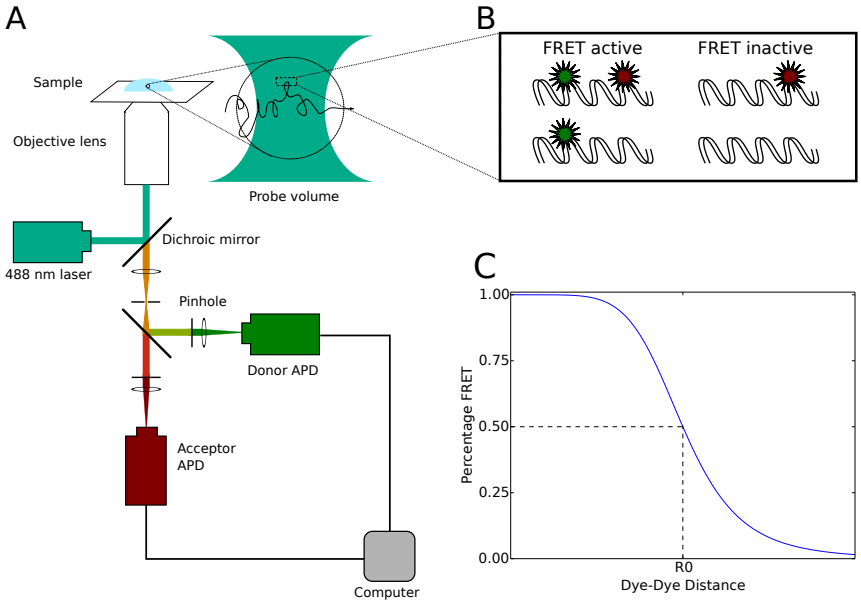

Figure 2.1: **Instrumentation for a smFRET experiment.** A) The confocal microscope, excitation and detection apparatus. B) Labelled molecules diffuse through the excitation volume. C) The characteristic sigmoidal dependence of FRET efficiency on dye-dye distance.

2.3 Data Analysis in Confocal smFRET Experiments

Analysis of smFRET data involves several computational challenges. Firstly, photons emitted by a fluorescent molecule diffusing through the excitation volume must be identified against a noisy background. Secondly, identified bursts must be denoised, including removal of background auto-fluorescence and donor-acceptor crosstalk. Fluorescent bursts that are distorted by photobleaching or other photophysical artifacts should be identified and excluded. Multiple methods of burst selection and analysis have been developed and applied to the analysis of smFRET data [4, 55, 56, 27, 35, 57, 36, 58, 59]. However, software for analysis of smFRET data has thus far been developed on an ad hoc basis, with individual groups preparing and maintaining their own analysis scripts, leading to problems typical of research programming projects [60, 61].

2.3.1 Development of Scientific Software

A lot has been written concerning the development and maintenance of software used by academic researchers [60, 61]. Compared with code written by professional software engineers, scientists are often described as writing poor quality code that is inefficient, badly documented and difficult to use. Software for smFRET analysis has historically experienced several of these issues, which has slowed innovation in the field.

Firstly, smFRET software used by different research groups often requires "reinventing the wheel" [62]. Within smFRET research groups, programming ability is not a standard skill, despite the need for sophisticated data analysis and use of custom data collection hardware. It is common for researchers with programming skills to maintain their own series of data-analysis scripts which may be wholly dependent on particular hardware tools or analysis packages. Other researchers, who may lack the skills to maintain and develop even simple scripts, are dependent on these black-box techniques provided by their colleagues. Consequently, data analysis is dependent on scripts written and maintained by just a few researchers. Loss of programming expertise when these team members leave can result in significant difficulties for the remaining group members, who are then dependent on poorly documented code that they do not fully understand how to use. Furthermore, the lack of available open source software often requires new researchers in the field of smFRET to completely reimplement standard analysis techniques in order to become independently productive.

Secondly, the need for many researchers to develop and maintain their own analyis tools has significant impact on research productivity. The requirement to reimplement standard analysis techniques consumes valuable time that could better be used in experimental research or in developing and benchmarking improved analysis tools. Furthermore, most researchers have no formal training in software engineering, with the result that analysis software can vary hugely in quality and is frequently poorly documented and maintained, making it difficult for other researchers to understand and use. New analysis scipts are often added in an ad hoc manner, transforming simple modifications into complex undertakings requring significant time investmenent. Poorly maintained code and undocumented code adds an additional barrier to open sharing of resources.

Finally, there is the issue of research reproducibility. Different research groups use widely differing tools to complete relatively similar tasks. New methods of data collection and

analysis are frequently developed [35, 27, 63]. However, when software is not released to the community, it is difficult for researchers, who must often implement poorly described methodologies entirely from scratch, to verify results or to adopt new techniques in their own research. As a consequence, new techniques are poorly benchmarked, making it difficult to understand whether a new analysis adds quality or merely complexity, whilst adoption of useful new methods is relatively slow. These three issues of productivity, reliability and reproducibility, all linked to the problem of poorly maintained softwared and lack of software development skills, are now becoming a key bottleneck in smFRET research.

2.3.2 Continuous Excitation

Time-binned Data In the most simple smFRET experiment, fluorescently labelled molecules are excited by a laser that will excite the donor dye; all photons reaching the detectors during data acquisition binned as they are received into time-bins of length similar to the expected dwell-time of a molecule in the confocal volume (for small, freely diffusing molecules, a bin-time of 1 ms is typically used). Event selection then simply involves identifying time-bins that contain sufficient photons to meet a specified criterion. Two thresholding criteria are in common use. AND thresholding selects time bins for which $n_D > T_D$ AND $n_A > T_A$ for n_D and n_A photons in the donor and acceptor channels respectively, and T_D and T_A the donor and acceptor thresholds. In a similar manner, SUM thresholding considers the sum of photons observed in the donor and acceptor channels, selecting time bins for which $n_D + n_A > T$.

Fluorescent Burst Data Although using time-bins that are matched to the dwell-time of molecules in the confocal volume is simple, it is not ideal, as some bursts will be split over several bins, so may be counted as separate events, or not considered for analysis. More sophisticated event selection algorithms, typically called burst search algorithms [27], bin photons on a time-scale much shorter than the typical dwell time in the confocal volume and then scan the resultant photon stream for bursts of a specified duration and brightness. pyFRET implements both All Photons Burst Search (APBS) and a Dual Channel Burst Search (DCBS) algorithms both for ALEX data and for simple FRET data, as originally described [27].

In APBS burst search for FRET data, photons from both donor and acceptor channels are considered together. A burst is defined according to three constants: T, the averaging

window; M, the minimum number of photons within window T; and L, the minimum total number of photons required for an identified burst to be retained. These three values are used in a two-step process for burst identification.

Firstly, "the start (respectively, the end) of a potential burst is detected when the number of photons in the averaging window of duration T is larger (respectively, smaller) than the minimum number of photons M." [27].

The DCBS burst search is similar, but considers the donor and acceptor channels separately. For a burst to be accepted in DCBS, both channels must simultaneoulsy meet the running sum criterion, allowing exclusion of single colour bursts and bursts where one fluorophore bleaches.

2.3.3 Alternating Laser Excitation

Time-binned Data A more sophisticated smFRET experiment uses Alternating Laser Excitation (ALEX) during data acquisition. In this method, the diffusing fluorescent molecules are subjected to excitation from both two lasers in rapid alternation [35]. One laser excites the donor fluorophore and the other can directly excite the acceptor fluorophore. The alternation of the laser excitation is fast on the timescale of molecular dwell-time in the confocal volume, allowing a single fluorescent molecule to receive multiple cycles of donor-acceptor direct excitation.

Instead of the two photon streams – donor and acceptor photons – observed in a simple smFRET experiment, in a confocal ALEX experiment, there are four streams: $F_{D_{ex}}^{D_{em}}$, $F_{D_{ex}}^{A_{em}}$, $F_{A_{ex}}^{D_{em}}$ and $F_{A_{ex}}^{A_{em}}$. $F_{D_{ex}}^{D_{em}}$ and $F_{D_{ex}}^{A_{em}}$, respectively donor and acceptor emission during donor excitation, are analogous to the two original donor and acceptor photon streams in a smFRET experiment. $F_{A_{ex}}^{A_{em}}$ records acceptor emission during direct acceptor excitation, whilst $F_{A_{ex}}^{D_{em}}$ records donor emission during acceptor excitation. Initial event selection is performed based on the number of photons observed during both donor excitation and acceptor excitation: $F_{D_{ex}}^{D_{em}} + F_{D_{ex}}^{A_{em}} > T_{D_{ex}}$ AND $F_{A_{ex}}^{A_{em}} > T_{A_{ex}}$ for threshold $T_{D_{ex}}$ during donor excitation and $T_{A_{ex}}$ during acceptor excitation provides an equivalent to AND thresholding; $F_{D_{ex}}^{D_{em}} + F_{D_{ex}}^{A_{em}} + F_{A_{ex}}^{A_{em}} > T$ for overall threshold T is analogous to SUM thresholding. Performing event selection in this manner, based on direct excitation of both fluorophores, should remove the biases caused by simple AND or SUM thresholding. The presence of these extra channels also provides additional information about the labelling state of molecules giving rise to

fluorescent bursts, allowing calculation of an emission stoichiometry, S:

$$S = \frac{F_{D_{ex}}^{D_{em}} + F_{D_{ex}}^{A_{em}}}{F_{D_{ex}}^{D_{em}} + F_{D_{ex}}^{A_{em}} + F_{A_{ex}}^{A_{em}}} \tag{2.3}$$

Values of S that are very close to one or very close to zero, indicate presence of only the donor or acceptor fluorophore respectively, so can be excluded from further analysis using a second event selection criterion: $S_{min} < S < S_{max}$.

Fluorescent Burst Data The burst search algorithms implemented for ALEX data work in a similar manner to those implemented for simple FRET data. In the ALEX APBS method, bursts are identified by considering the total number of fluorescent photons $F_{total} = F_{D_{ex}}^{D_{em}} + F_{D_{ex}}^{A_{em}} + F_{A_{ex}}^{A_{em}}$ in each time bin. Bursts are identified where the running sum (calculated using the F_{total} photon stream) in the averaging window T exceeds the threshold M. The DCBS method considers donor excitation photons $F_{D_{ex}}^{D_{em}} + F_{D_{ex}}^{A_{em}}$ separately from photons emitted during direct acceptor excitation $F_{donor} = F_{A_{ex}}^{A_{em}}$, requiring that the running sum exceeds M for both F_{donor} and $F_{A_{ex}}^{A_{em}}$.

2.4 pyFRET: Design and Implementation

This chpter presents pyFRET, an open-source library, written in the python programming language, for the analysis of smFRET data. To our knowledge, this is the first open source software ever released by the smFRET research community. pyFRET is a small library that provides a toolkit facilitating all key steps in analysis of smFRET data: burst selection; cross-talk subtraction and burst denoising; data visualisation; and construction and simple fitting of FRET efficiency histograms. In providing this toolkit to the smFRET research community, we hope to facilitate the wider adoption of smFRET techniques in biological research as well as providing a framework for open communication about and sharing of data analysis tools.

2.4.1 Code Layout and Design

pyFRET provides four key data structures (classes) for manipulation of smFRET data. The FRET data object describes two fluorescence channels, corresponding to time-bins containing

photons collected from donor (the donor channel, D) and acceptor (the acceptor channel, A) fluorophores. The ALEX data object describes four fluorescence channels, corresponding to the four temporal states in a smFRET experiment using Alternating Laser Excitation (ALEX), namely the donor channel when the donor laser is switched on (D_D); the donor channel when the acceptor laser is switched on (D_A); the acceptor channel when the donor laser is on (A_D); and the acceptor channel when the acceptor laser is on (A_A). These data channels are implemented as numpy arrays [64], allowing efficient computations and selection operations. The data structure can readily be expanded to include data from more detectors, which is needed in e.g. three-colour or anisotropy measurements.

Two similar classes are used for fluorescence bursts identified using the burst search algorithms. In addition to the donor and acceptor channels, the FRET bursts class type holds three additional arrays, giving the first and last bin of each burst, and the duration of each burst. These three new attributes are similarly present in the ALEX bursts object, in addition to the four fluorescent channels in the ALEX data object. In addition to the methods present for the simple FRET and ALEX objects, the burst data objects also implement methods to plot burst duration and to analyse recurrent bursts (RASP) [65].

The data analysis workflow is illustrated in Fig. 2.2 Following initialization of data objects, a single line of code performs background subtraction, event selection, cross-talk correction and calculation of the FRET efficiency. Likewise, simple but high-quality figures (see Fig. 2.3 for examples) can be generated in a single step.

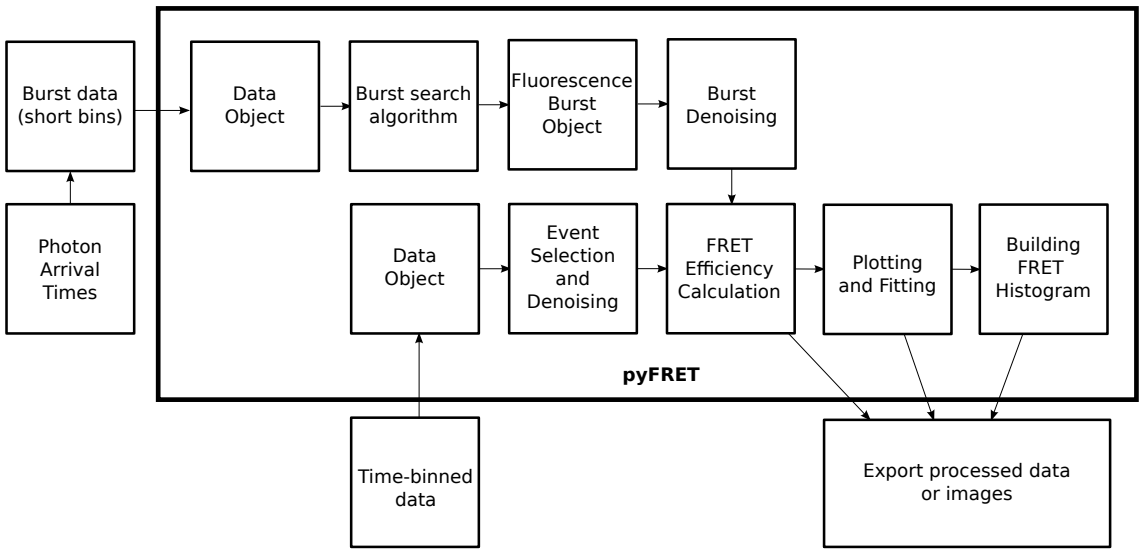

Figure 2.2: **Typical workflow for data analysis using pyFRET.**

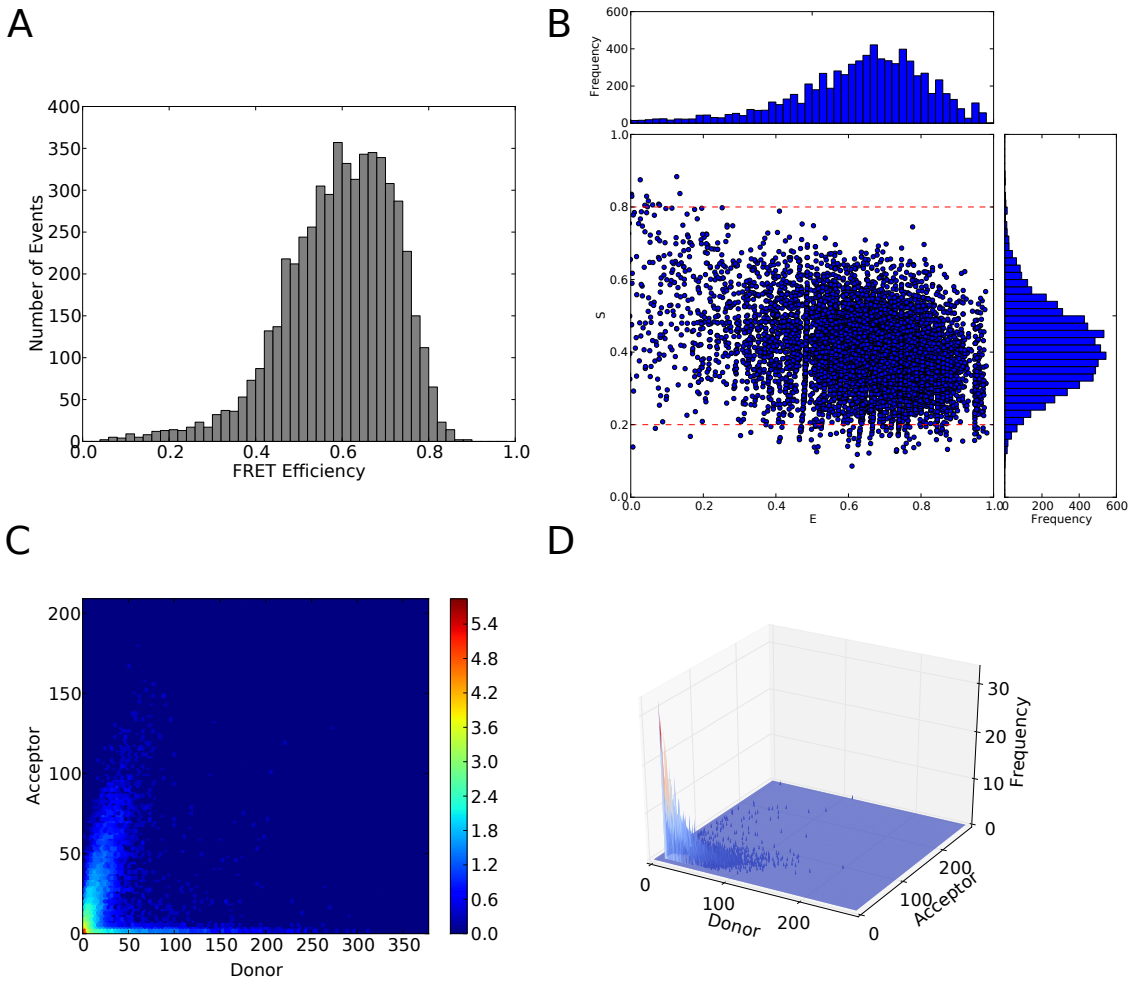

Figure 2.3: **Figures made using pyFRET.** A) A Proximity Ratio histogram. B) A scatterplot of FRET efficiency and fluorophore stoichiometry from ALEX data. C) A heatmap of event frequencies. D) A 3D plot of event frequencies.

2.4.2 Dependencies

In addition to the numpy arrays used for efficient data processing, pyFRET uses several other python modules to enable efficient data analysis. Specifically, we use matplotlib [66] for plotting and data visualisation; whilst gaussian fitting of the FRET histograms is implemented using scikit-learn [67]. These dependencies are available for free anonymous download for all platforms (see Section 2.8). We used the free Anaconda package bundle (`https://store.continuum.io/cshop/anaconda/`) to install these dependencies. Detailed installation instructions can be found in the pyFRET documentation (`http://pyfret.readthedocs.org/en/latest/tutorial.html#installing-pyfret`).

2.4.3 Simple Event Selection and Denoising

FRET Data As described above, the most straighforward method of smFRET data analysis of data from a continuous excitation experiement uses time-binned data and simply selects time-bins where the photon count exceeds some stated threshold. This threshold can be over both the donor and acceptor channels (AND thresholding) or over the sum of photons in both donor and acceptor channels. These thresholding techniques can be implemented using a single call to a pyFRET function:

```
# Simple thresholding

# define thresholds
Td = 20 # donor threshold
Ta = 20 # acceptor threshold
T = 50  # combined threshold

# AND thresholding
data.threshold_AND(Td, Ta)

# SUM thresholding
data.threshold_SUM(T)
```

Following event selection, a simple method of denoising is to subtract from each selected event the average background autofluorescence observed in each channel and the average cross-talk between the two channels:

```
# Simple denoising

# removing autofluorescence
auto_donor = 0.5   # donor autofluorescence
auto_acceptor = 0.3 # acceptor autofluorescence
my_data.subtract_bckd(auto_donor, auto_acceptor)

# removing cross-talk
cross_DtoA = 0.05 # fractional cross-talk from donor to acceptor
cross_AtoD = 0.01 # fractional cross-talk from acceptor to donor
my_data.subtract_crosstalk(cross_DtoA, cross_AtoD)
```

This is the simplest method for event selection and denoising. However, it has several limitations. In particular, simple subtraction of constant values can lead to unphysical artifacts such as negative and fractional photon counts. Furthermore, the simple thresholding criteria for event selection are known to be biased [27], so can distort downstream data analysis.

ALEX Data pyFRET implements ALEX event selection as described in the original publication [68]. In brief, bursts are initially selected using a selection criterion based on the total number of photons emitted during donor and acceptor excitation: $F_{D_{ex}}^{D_{em}} + F_{D_{ex}}^{A_{em}} > T_D$ AND $F_{A_{ex}}^{A_{em}} > T_A$.

Following this initial event selection, a second selection step is performed, based on the ratio of photons emitted during donor and acceptor excitation periods. The photon stoichiometry, S is calculated for each burst as:

$$S = \frac{F_{D_{ex}}^{D_{em}} + F_{D_{ex}}^{A_{em}}}{F_{D_{ex}}^{D_{em}} + F_{D_{ex}}^{A_{em}} + F_{A_{ex}}^{A_{em}}} \tag{2.4}$$

Events for which the stoichiometry is either very close to one or very close to zero, indicating presence of only the donor or acceptor fluorophore respectively can be excluded using a second event selection criterion: $S_{min} < S < S_{max}$

Following event selection, remaining bursts can be corrected for photon leakage and direct excitation contributions. A two-dimensional scatter plot of FRET efficiency E and stoichiometery S is then produced, including one-dimensional histograms of both E and S.

pyFRET allows each of these steps to be performed separately, however they can also be combined into a single step combining event selection, denoising, FRET efficiency calculation and plotting:

```
# Simple ALEX analysis

g_factor = 0.95 # instrumental gamma factor
S_min = 0.2   # min accepted value of S
S_max = 0.8   # max accepted value of S
filepath = "path\to\my\file"
filename = "scatter_plot"
```

```
ALEX_data.scatter_hist(S_min, S_max, gamma=g_factor, save=True,
    filepath=filepath, imgname=filename, imgtype="png")
```

Example ALEX analysis scripts can be found in the iPython notebooks available online (???).

2.4.4 Burst Search Algorithms

pyFRET implements busrt search algorithms for both continuous excitation and ALEX experiments. As described above, to use a burst search algorithm effectively, photons must be binned into time-bins of duration much shorter than the average dwell time of a molecule in the confocal volume. Bursts are first identified through an initial scan of the burst stream, to identify windows in which the number of photons observed exceeds the threshold M. In pyFRET, this initial search is performed using the convolve method from numpy [69] to provide a running sum across windows of T time-bins. Following initial burst identification, a burst is retained if it contains more than L photons [27].

Example code for running a burst search algorithm is shown below:

```
# Burst search using FRET burst data

# required parameters
T = 50          # time window (bins)
M = 50          # first threshold
L = 60          # second threshold

# calling APBS algorithm
bursts_APBS = FRET_data.APBS(T, M, L)
```

The procedure for running a burst search algorithm on ALEX data is precisely analogous:

```
# Burst search using ALEX burst data

# required parameters
T = 50          # time window (bins)
M = 50          # first threshold
L = 60          # second threshold
```

```
# calling APBS algorithm
bursts_APBS = ALEX_data.APBS(T, M, L)
```

2.5 RASP: Recurrence Analysis of Single Particles

A recent innovation in confocal smFRET is Recurrence Analysis of Single Particles (RASP) [65]. RASP exploits the fact that fluorescent bursts occurring close in time are more likely to be from the same molecule diffusing back through the confocal volume than from a newly observed molecule. This can be used to determine interconversion kinetics and to test whether broad peaks in the FRET histogram derive from overlapping static populations.

RASP is a two-step process. First, initial bursts (b_1) with a FRET efficiency E_{b1} within some defined range $\Delta(E_{b1})$ are identified. Secondly, bursts (b_2) occurring within a time interval (called the recurrence interval) $T = (t_1, t_2)$ of b_1 are identified. Analysis of the distribution of FRET efficiencies in b_2, the population of recurrent bursts, provides information about the interconversion rate between subpopulations. The rate constants of interconversion can be extracted by fitting the relative subpopulations as a function of the recurrence interval T.

pyFRET implements RASP using array masking, to allow efficient selection of relevant bursts. RASP can be called in a single step from a FRET bursts or ALEX bursts object, and a loop can readily be made to repeat the process at different time intervals:

```
# RASP

# initial E range: 0.4 < E < 0.6
Emin = 0.4
Emax = 0.6

# Time interval for re-occurrence
# given in number of bins
Tmin = 1000
Tmax = 10000
```

```
# selecting re-occurring bursts
recurrent_bursts = bursts_APBS.RASP(Emin, Emax, Tmin, Tmax)

# histogram of re-occurring bursts
recurrent_bursts.build_histogram(filepath, csvname, gamma=g_factor)
```

2.5.1 Compatibilities

pyFRET is written in Python. Both python 2 (v2.7) and python 3 (v3.3) are supported. pyFRET requires four further python libraries, namely numpy and scipy for data manipulation, matplotlib for data visualisation and scikit-learn for peak fitting. pyFRET was written and tested in a Linux environment. However, it was written to be platform independent and has also been used successfully on both iOS and Windows computers.

The lack of Open Source software in the smFRET community has led to a proliferation of esoteric file-types used for data collection and storage. To make pyFRET as usable as possible for a wide range of smFRET researchers, we provide file parsers for simple .csv and .txt file formats, as well as the custom binary format used in the Klenerman group. The pyFRET data structures can be initialised using simple python arrays of time-binned photons, for users whose file format is not currently supported. The tutorial provides example scripts for parsing common filetypes into pyFRET objects.

2.6 Experimental Methods

In order to effectively evaluate the performance of the pyFRET library, we generated some simulated data with known parameters. We also collected some smFRET data under various different excitation and data collection regimes. We then analysed the datasets using pyFRET and compared the results. The following section describes the experimental protocols used in the generation of simulated datasets and the methods used in the collection and analysis of smFRET data.

2.6.1 Benchmarking the Gaussian Fitting Using Simulated Datasets

To test the ability of the fitting algorithms to distinguish fluorescent populations with similar FRET efficiencies, we generated simulated datasets consisting of mixtures of FRET bursts with a known FRET efficiency and population size. These bursts were then fitted using pyFRET's two component mixture model and the results compared to the known input FRET efficiencies and population sizes. The python script used to generate and fit these datasets can be found in the online repository.

2.6.2 Data to Evaluate the Simple Event Selection Algorithms

We tested the pyFRET library using DNA duplexes dual-labelled with Alexa Fluor 488 and Alexa Fluor 647. The duplex sequences and labelling sites are shown in Tables 3.2 and 2.2. Labelled duplexes were diluted to a concentration of 50 pM in TEN buffer (10 mM Tris, 1mM EDTA, 100 mM Nacl), pH 8.0, containing 0.0001 % Tween-20. FRET data were collected for 15 minutes using continuous excitation at 488 nm at a power of 80 mW. Collected photons were binned online in intervals of 1 ms and stored in files of 10000 bins. To process the data, AND thresholding was performed using thresholds $N_D = 10$ and $N_A = 10$. SUM thresholding used a threshold $N = 20$. For FRET efficiency calculation the γ-factor was 0.95.

ALEX data were collected for 15 minutes using alternating excitation at 488 and 640 nm, with respective laser powers of 80 and 70 mW, and a modulation rate of 0.1 ms, a dead-time of 0.1 μs and a delay compensation of 3 μs. ALEX data were then binned in intervals of 1 ms. The scripts and configuration files used to analyse these data using pyFRET can be found in the online repository.

Table 2.1: **Donor DNA Sequence** DNA sequence of the donor-labelled strand, where **5** is a deoxy-T nucleotide, labelled with Alexa Fluor 488 at the C6 amino position.

Construct	Sequence
Donor	TACTGCCTTTCTGTATCGC**5**TATCGCGTAGTTACCTGCCTTGCATAGCCACTCATAGCCT

Table 2.2: **Acceptor DNA Sequences**. 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.

Separation / bp	Acceptor Sequence
4	AGGCTATGAGTGGCTATGCAAGGCAGGTAACTACGCGATAAGCGA**6**
6	AGGCTATGAGTGGCTATGCAAGGCAGGTAACTACGCGATAAGCGATA**6**
8	AGGCTATGAGTGGCTATGCAAGGCAGGTAACTACGCGATAAGCGATACA**6**
10	AGGCTATGAGTGGCTATGCAAGGCAGGTAACTACGCGATAAGCGATACAGA**6**
12	AGGCTATGAGTGGCTATGCAAGGCAGGTAACTACGCGATAAGCGATACAGAAA**6**

2.6.3 Data to Evaluate Event Selection Using the Burst Search Algorithms

To test the burst search algorithms, we collected data under both ALEX and FRET conditions. Data were collected for 10 minutes, using laser powers of 130 μW in both donor and acceptor excitation. For FRET data collection, laser illumination by the donor laser was continuous and the data were binned online into time bins of 50 μs, roughly 5 % of the duration of the average burst from a freely diffusing molecule. The acceptor laser was not used. For ALEX data collection, photons were similarly binned online into time bins of 50 μs length, but the laser modulation rate was increased to 50 μs, with a dead-time of 0.1 μs and a delay compensation of 3 μs, corresponding to 2 modulations per short time bin.

To evaluate the effect of bin-time on performance of the burst search algorithms, a further dataset was collected using FRET excitation on the 6 bp duplex. For this dataset, data were collected for 10 minutes using the 488 nm laser at 130 μW. Data were binned into time bins of 10 μs, allowing for re-binning into longer time-bins as required.

Unless otherwise specified, DCBS was carried out using parameters $T = 20$ bins, $L = 10$ and $M = 10$; APBS was carried out using parameters $T = 20$ bins, $L = 20$ and $M = 20$.

2.6.4 Performance Analysis Using Mixtures of DNA Duplexes

To test the accuracy of the burst search algorithms, we collected ALEX data from mixtures of two DNA duplexes at known concentration ratios. The 4, 8 and 12 bp duplexes were used (see Table 2.2). Samples were prepared containing a mixture of two duplexes in TEN buffer with a total DNA concentration of 80 pM, divided in either a 3 : 1 ratio (high concentration 60 pM, low concentration 20 pM) or a 1 : 1 ratio (both components 40 pM). Data were collected under ALEX excitation for 10 minutes, using 50 μs time-bins as described above. The data were later re-binned into 1 ms time-bins, to compare performance of burst search and simple thresholding in separating the two populations.

2.6.5 Testing the RASP Algorithm

To test the RASP Algorithm, a 1:1 mixture of 6 bp and 12 bp duplex was prepared to a total DNA concentration of 50 pM. A 400 μL aliquot of the dilute solution was placed in one chamber of a lidded, chambered coverslide (LabTex) to reduce evaporation during the measurement. FRET data were collected for 600 minutes using continuous excitation at 488 nm at a power of 140 μW. Collected photons were binned online in intervals of 50 μs and stored in files of 1000000 bins. Events were identified using APBS burst selection was used, with the lower thresholds of $L = 10$ and $M = 10$ to maximise the number of detected events. RASP was performed on the selected events, using the parameters given in Table 2.3.

Table 2.3: **RASP Parameters**. Parameters used in RASP analysis of the 6 bp duplex - 12 bp duplex mixture.

Parameter	Value
T	20
M	10
L	10
E_{min}	0.65
E_{max}	0.85
ΔT	1 ms

2.7 Performance Analysis of smFRET Analysis Algorithms

To evaluate the performance of the smFRET analysis algorithms, we used experimental datasets collected as described above. The following sections describe the data analysis pathway using pyFRET and present the results of pyFRET analysis. We also provide a comparison of the performance of different algorithms on smFRET data collected from the same DNA constructs but under different excitation and photon collection regimes.

2.7.1 Evaluating Performance with DNA Duplexes

Simple Algorithms We tested the pyFRET library using DNA duplexes dual-labelled with Alexa Fluor 488 and Alexa Fluor 647. The duplex sequences, dye attachment sites and dye-dye separations are shown in Tables 2.1 and 2.2. Event selection and denoising, calculation of FRET efficiency and the plotting and fitting of FRET efficiency histograms were performed using pyFRET. The results, shown in Fig. 2.4 for FRET data analysed using AND thresholding, and Fig. 2.5 for ALEX data, demonstrate that even using the simplest event selection and denoising techniques, pyFRET is able to effectively fit histograms from single fluorescent populations (Fig. 2.4 and Fig. 2.5 A - E), to reproduce the characteristic sigmoidal FRET efficiency curve (Fig. 2.4 and Fig. 2.5 F). However, as can be seen from Fig. 2.7 C), where the FRET efficiency curves from four different analysis algorithms are overlaid, the simple FRET analysis results in a flattened curve, caused by the bias towards events with intermediate FRET efficiencies displayed by the simple AND thresholding algorithm.

Burst Search Algorithms We tested the pyFRET burst search algorithms using data collected from the same DNA duplexes but using a shorter bin-time. Sample results, from DCBS analysis of both FRET and ALEX data, and APBS analsysis of ALEX data are shown in Fig. 2.6. APBS analysis of the FRET burst data is not shown, as the presence of a zero peak hampered fitting of the low-FRET species. The characteristic sigmoidal FRET efficiency curves generated from DCBS analysis of all five duplexes are shown in Fig. 2.7 A (FRET data) and B (ALEX data).

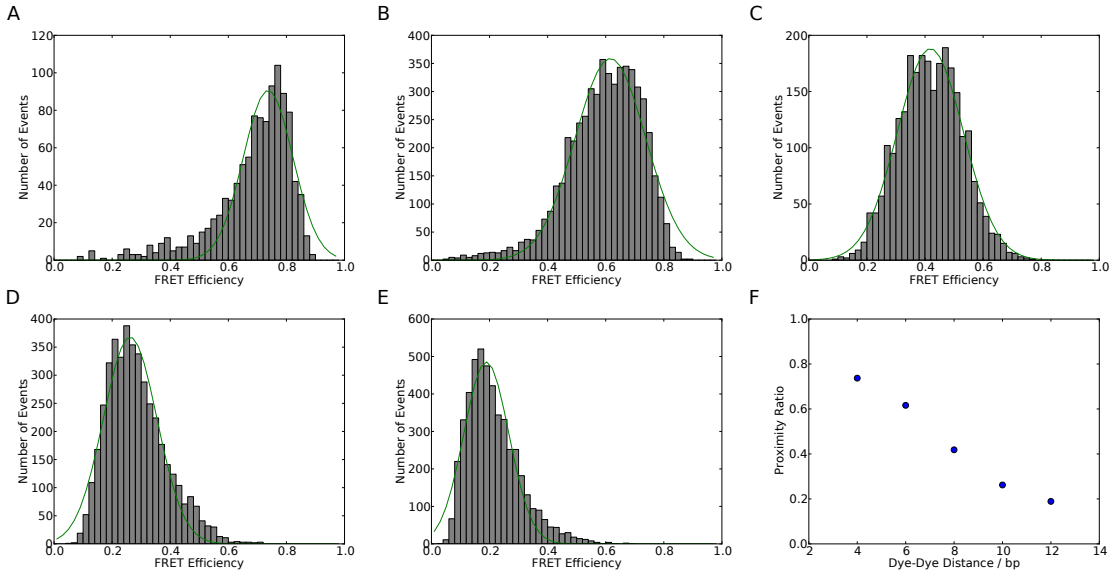

Figure 2.4: **Analysis of FRET data from DNA duplexes using pyFRET.** A - E: Fitted FRET histograms from DNA duplexes labelled with a dye-dye separation of 4, 6, 8, 10 and 12 base pairs respectively. F) Characteristic sigmoidal curve of FRET efficiency against dye-dye distance.

2.7.2 Evaluating the Burst Search Algorithms

In addition to demonstrating the functionality of the burst search algorithms, we have performed a comprehensive analysis of the impact of bin-time, threshold and detection window on the performance of the burst search algorithms. To our knowledge, this is the first time that such an analysis has been performed, and hence provides useful insight to other researchers depending on the performance of these algorithms.

Burst search algorithms were developed as an improvement to the simple thresholding technique, designed to reduce inaccuracies caused by photobleaching and by long bursts being split over multiple time-bins. To assess the improvement in data quality as a result of using the burst search algorithm, we collected data from the 6 bp duplex using short time-bins of 10 μs, which could then be re-binned into longer time-bins for comparative analysis. Analysis was performed using both APBS and DCBS burst search algorithms on FRET data. For APBS, thresholds of $M = 20$ and $L = 20$ were used; the thresholds used for DCBS were $M = 10$ and $L = 10$. As the bin-time was varied, the minimum burst duration T, given in number of bins, was also varied, to keep the minimum burst duration to a constant time of 1 ms. The bin lengths and their corresponding value of T are shown in Table 2.4. The results,

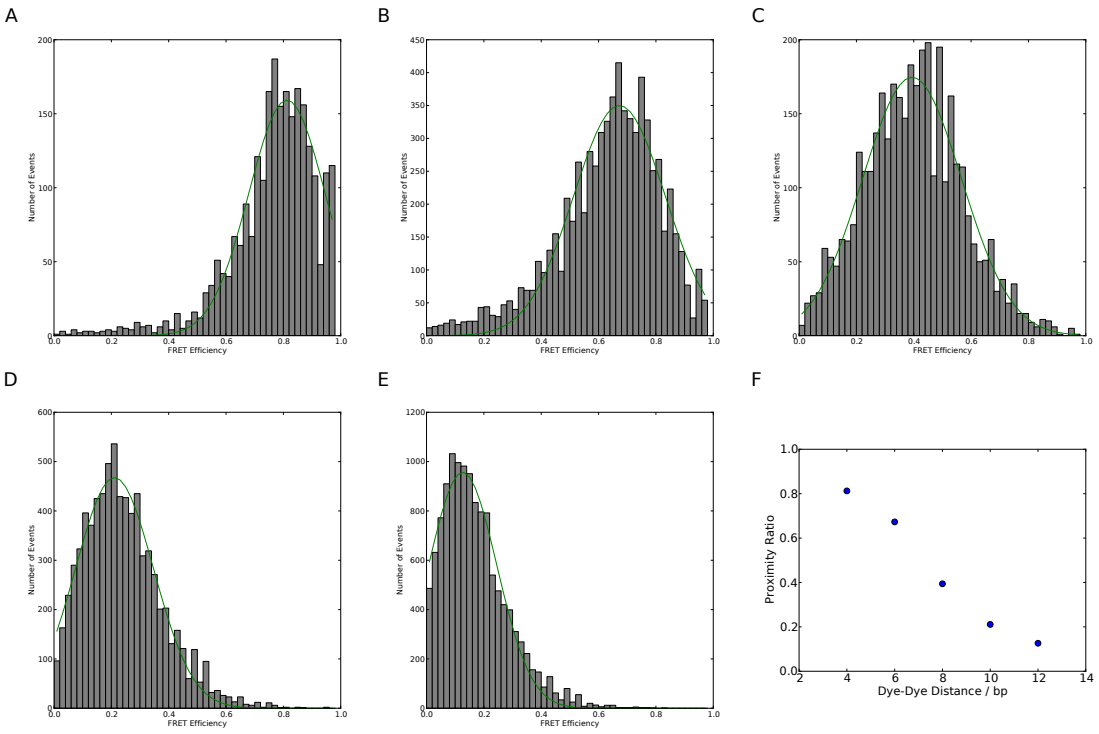

Figure 2.5: **Analysis of ALEX data from DNA duplexes using pyALEX.** A - E: Fitted FRET histograms from DNA duplexes labelled with a dye-dye separation of 4, 6, 8, 10 and 12 base pairs respectively. F) Characteristic sigmoidal curve of FRET efficiency against dye-dye distance.

shown in Fig. 2.8 A (DCBS) and B (APBS) are surprising. When the minimum time-interval for burst detection is not varied, the performance of the algorithm is essentially unaffected by the bin-times used: for both APBS and DCBS algorithms, the resultant FRET efficiency histograms are extremely similar, whether many short bins or a few long bins are searched.

Secondly, we evaluated the effect of the search window T on burst search performance. For a fixed bin-time of 50 μs, we varied the required burst duration T between 100 μs (2 bins) and 1000 μs (10 bins). The results, shown in Fig. 2.8 C for the DCBS algorithm are again surprising. Across all values tested, the shape of the FRET efficiency histogram is unaffectd by the size of the burst search window. The peak areas are also relatively unaffected by the search window size: the very shortest window of 100 μs retains only the brightest bursts, resulting in a slightly reduced peak area; other than this the number of detected events is not significantly altered. An analysis of DCBS and APBS on ALEX data showed similar results

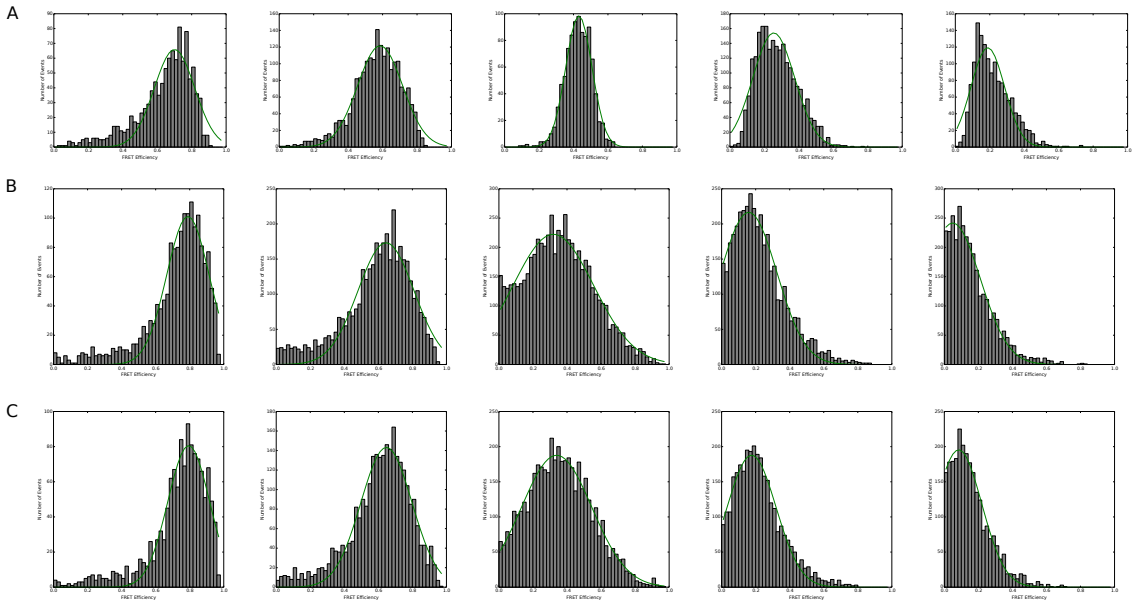

Figure 2.6: **Testing the burst search algorithms.** A) Fitted FRET histograms from DCBS on FRET data. B) Fitted FRET histograms from APBS on ALEX data. C) Fitted FRET histograms from DCBS on ALEX data. In A-C histograms from left to right are collected from 12 bp, 10 bp, 8 bp, 6 bp and 4 bp duplexes respectively. Histograms from APBS on FRET data are not shown, as the large zero peak hampers gaussian fitting when the FRET efficiency is low.

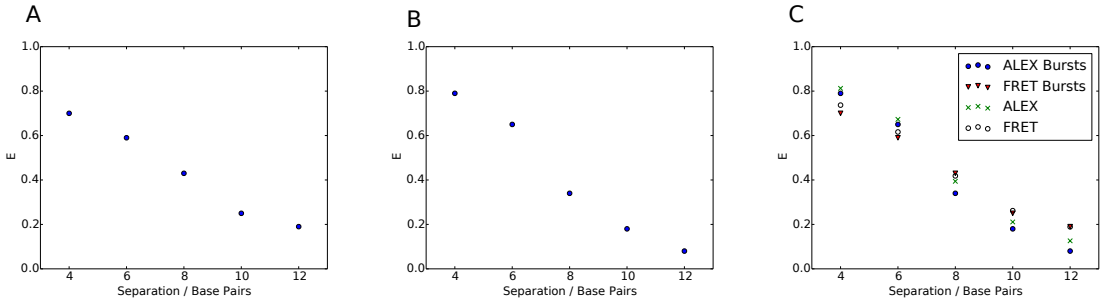

Figure 2.7: **Plot of FRET Efficiency vs dye-dye separation.** A) FRET data, analysed using the DCBS burst search algorithm. B) ALEX data, analysed using the DCBS burst search algorithm. C) Comparison of different methods. Blue circles and red triangles show ALEX and FRET burst data respectively; green crosses and open circles show simple thresholded data from ALEX and FRET experiments respectively.

(Fig. 2.8 D - F), although the reduction in peak area when the time window is shortest (100 μs) is more pronounced (Fig. 2.8 F). This lack of dependence on the burst search parameters

Table 2.4: **Bin-times.** Bin lengths and the corresponding minimum burst duration used to evaluate the effect of bin-time on burst search performance.

Bin-time / μs	T / bins
10	100
20	50
40	25
50	20
100	10
500	2
1000	1

ensures that conclusions drawn are independent of the thresholding parameters used.

Finally, we evaluated the effect of the thresholds M and L on the performance of the DCBS algorithm for both FRET and ALEX data using the high-FRET 6 bp duplex. We systematically varied the thresholds used in burst search analysis, then evaluated their effect on the fitted peak area and FRET efficiency. The results, shown in Fig. 2.9, display several interesting features. Firstly, the decline in peak area with increased threshold is striking for both FRET (Fig. 2.9 C) and ALEX (Fig. 2.9 D) data, suggesting that the lowest possible values of L and M should be used, to maximise the data retained. Secondly, increasing the thresholds systematically reducuces the calculated FRET efficiency for the FRET dataset (Fig. 2.9 A). This is caused by the DCBS algorithm selecting against bursts that have a low donor count, as they do not meet the initial threshold M. This effect is not seen in DCBS on ALEX data (Fig. 2.9 B), as events are selected based on photons emitted during direct donor and acceptor excitation, so there is no bias towards intermediate FRET efficiencies.

We note that, when used on FRET data, the APBS and DCBS burst search algorithms perform in an analogous manner to simple AND and SUM thresholding. Consequently, they retain the well-known disadvantages of these thresholding methods [70]: specifically, APBS retains a zero-peak caused by donor-only labelled molecules, whereas DCBS is biased against extreme FRET efficiencies, distorting the FRET efficiency histogram. ALEX data does not display these biases, as the direct excitation of both fluorophores allows events to be selected independently of their FRET efficiencies. The effect of these biases in analysis of simple FRET data can be seen in Fig. 2.7 C, which overlays the calculated FRET efficiency curves for DCBS analysis of FRET and ALEX data. The curves for both AND-thresholded FRET

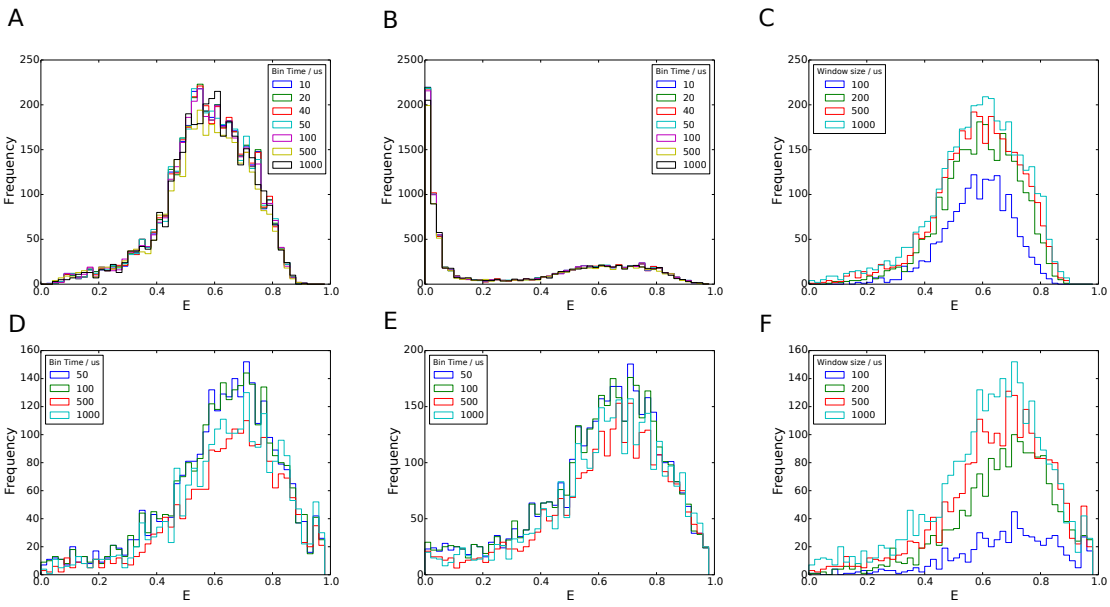

Figure 2.8: **Evaluating the effect of detection window and bin time on the DCBS burst search algorithm for FRET and ALEX data.** A - C show analysis of FRET data. D - F show analysis of ALEX data. A) Varying the length of the time-bin has little effect on the performance of the DCBS algorithm. B) Varying the length of the time-bin has little effect on the performance of the APBS algorithm. C) Reducing the length of the minimum detection window used in DCBS selects for very bright bursts. For very short windows, this reduces the number of detected bursts but does not affect the calculated FRET efficiency. D) Varying the length of the time-bin has little effect on the performance of the DCBS algorithm. E) Varying the length of the time-bin has little effect on the performance of the APBS algorithm. D) Reducing the length of the minimum detection window used in DCBS selects for very bright bursts.

data and DCBS burst selection of FRET data are distorted towards intermediate FRET efficiencies; ALEX data does not show this distortion. Consequently, ALEX is a superior technique and should be used when available.

2.7.3 Evaluating the Gaussian Fitting

As part of pyFRET, we used the gaussian mixture model implemented in scikit-learn [67] to fit the FRET histograms with one or more gaussian distributions. The fits shown in the preceeding figures use this mixture model. However, we also wanted to understand

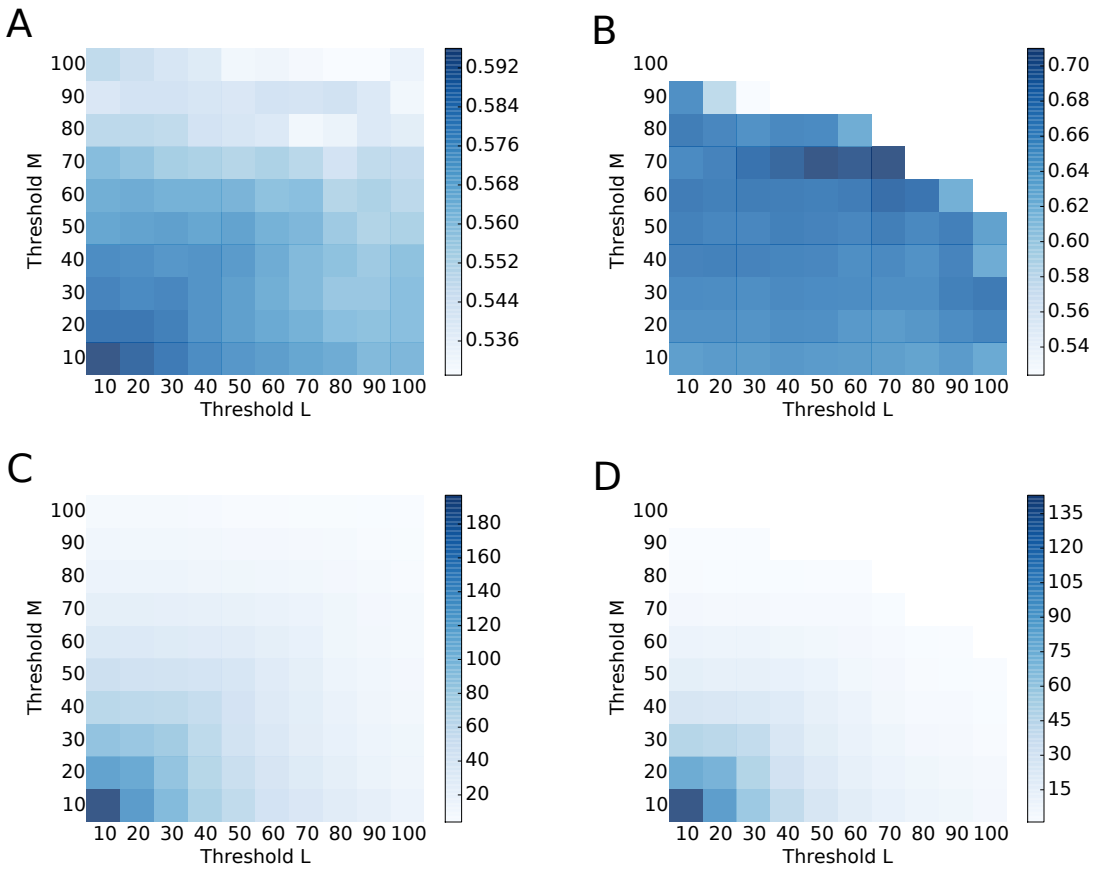

Figure 2.9: **Heatmaps showing the effect on calculated FRET efficiency and peak area of varying the burst search thresholds L and M.** A) Calculated FRET efficiency from DCBS analysis of FRET data. B) Calculated FRET efficiency from DCBS analysis of ALEX data. Missing values in B) are the result of insufficient events being retained for a gaussian fit to be performed. C) Calculated peak area from DCBS analysis of FRET data. D) Calculated FRET efficiency from DCBS analysis of ALEX data.

the limitations of the fitting method and to demarcate the conditions under which multiple fluorescent populations could not be effectively distinguished using pyFRET.

We used a combination of simulated and experimental data to benchmark the performance of the Gaussian fitting method when presented with multiple FRET populations. Firstly, we simulated FRET bursts drawn from a mixture of two FRET populations, with mean FRET efficiencies ranging from 0.1 to 0.9. These bursts were fitted with a two component Gaussian mixture model and the fits and FRET histograms overlaid. Example fits, shown in Fig. 2.10 A and B, demonstrate that where peaks are well separated (Fig. 2.10 A), they are correctly distinguished by the fitting protocol. However, when there is significant overlap

between the FRET peaks (Fig. 2.10 B), there is insufficient information to distinguish the two populations, so the two gaussians converge to very similar means, whose values lie in between the true mean values. This is an inherent problem with fitting multiple gaussians to peaks that are poorly resolved, and is thus not a problem specific to the fitting procedure.

To further quantify the limits of the fitting performance on two population mixtures, we quantified the discrepancy between the true and calculated FRET efficiencies for all peaks fitted from our simulated datasets, then plotted this discrepancy as a function of the FRET efficiency of the partner peak. The results are shown in Fig. 2.10 C and D. When the two peaks have extremely similar mean FRET efficiencies, the fitting algorithm cannot effectively distinguish them, so both calculated values are distorted towards an intermediate value. However, when the two peaks are well separated, both peaks are fitted correctly, independent of the mean of the partner peak.

Finally, we evaluated the performance of the gaussian fitting on experimental datasets, consisting of a two-component mixture of DNA duplexes. Three duplexes were used for this experiment, namely the duplexes with a 4, 6 and 12 bp dye-dye separation, corresponding to a high-, intermediate- and low-FRET species respectively. The duplexes were mixed in either a 1:3 or 1:1 concentration ratio for data collection using ALEX We analysed these data using the DCBS algorithm and then re-binned the raw data into 1 ms time-bins and re-analysed using simple ALEX event selection on these re-binned data. Additionally, we separated out the two photon channels (D_D and A_D) that record donor and acceptor photons that arrive during donor excitation and hence correspond to the simple FRET part of an ALEX experiment. We used these photon counts, from both the re-binned and the raw data to compare the performance of the DCBS algorithm for FRET with simple FRET event selection using AND thresholding. Identified bursts were fitted with a two-component mixture of gaussians.

The resulting FRET efficiencies and fractional peak areas are shown in Fig. 2.11 for the ALEX data and in Fig. 2.12 for the simple FRET data. A comparison of these results shows several interesting points. Only DCBS on ALEX data can reliably identify three separate FRET efficiencies (Fig. 2.11 A) and concentrations (Fig. 2.11 B), independent of the concentration or FRET efficiency of the partner duplex. In contrast, the FRET efficiencies (Fig. 2.11 C) and concentrations (Fig. 2.11 D) identified by simple ALEX analysis of long time bins are distorted by the partner peak values.

The results are even more distorted when the simple FRET data is considered (Fig. 2.12).

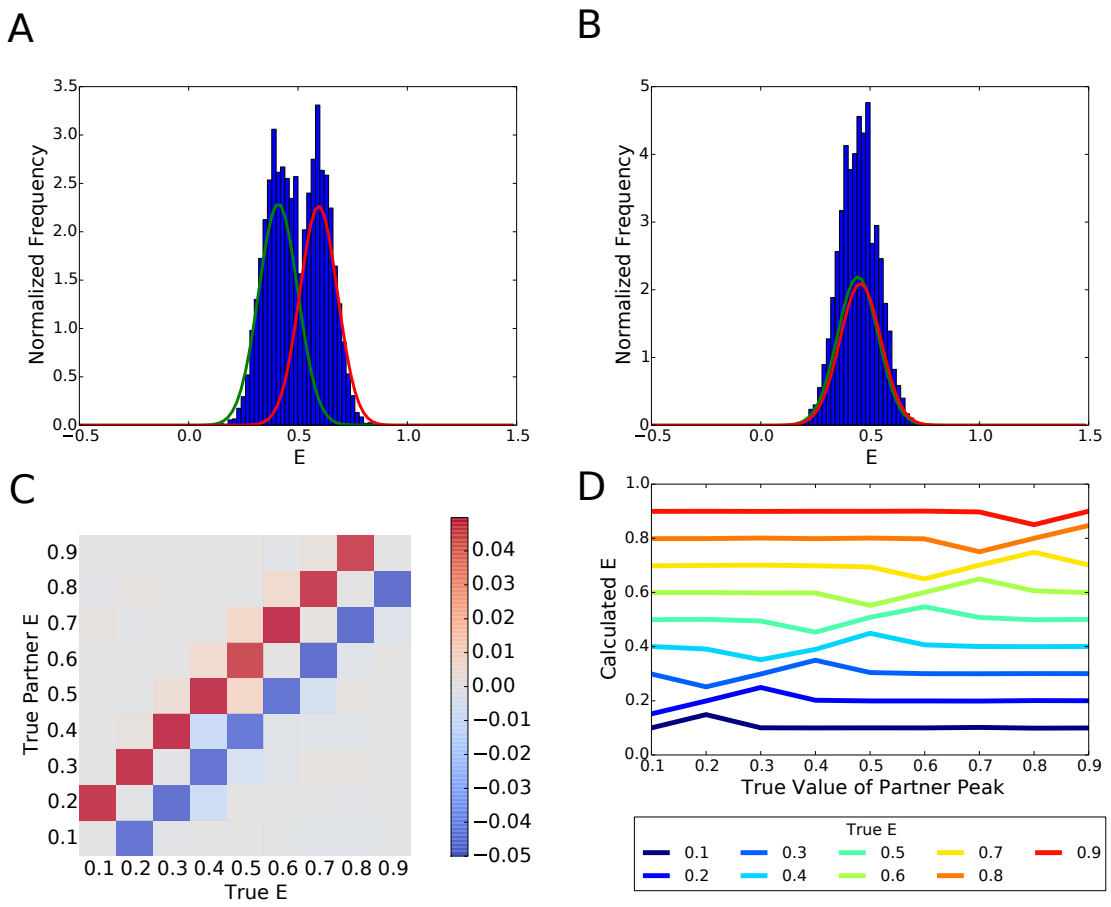

Figure 2.10: **Benchmarking the gaussian fitting process using simulated data.** A) Fitting a two-component Gaussian mixture to a simulated dataset simulating a 1:1 ratio of FRET populations, with FRET efficiencies of 0.4 and 0.6. The two peaks are easily distinguished and fitted correctly. B) Fitting a two-component Gaussian mixture to a simulated dataset simulating a 1:1 ratio of FRET populations, with FRET efficiencies of 0.4 and 0.5. There is significant overlap between the two peaks, so they cannot be distinguished by the fitting algorithm. C) Heatmap showing the relationship between partner FRET efficiency and the error in calculated FRET efficiency. Gaussian fits of datasets where the FRET efficiencies of the two peaks are extremely close to each other are distorted towards an intermediate value. D) One-dimensional representation of the heatmap shown in C), showing the distortion in calculated FRET efficiency when the two components of the mixture have very similar FRET efficiencies.

Compared with the ALEX data, the calculated FRET efficiencies are clearly distorted towards intermediate FRET efficiency values for both DCBS burst search (Fig. 2.12 A) and simple thresholding (Fig. 2.12 C) analsyes. Furthermore, the population sizes calculated from

the simple FRET data (Fig. 2.12 B and D) are entirely scrambled, showing an extremely poor reflection of the underlying concentrations used. Concentrations in this experiment is a proxy for the relative population of states in typical smFRET experiment. Overall, from these four analyses of the same dataset, it is clear that the additional information available from an ALEX excitation regime is crucial to unbiased event selection and hence to accurate calculation of FRET efficiencies and population sizes. Furthermore, this additional information is best exploited using a burst search analysis rather than long time bins. Consequently, we conclude that the performance of the gaussian fits are only as good as the prior event selection, and that, when there is explicit access to acceptor excitation information, burst search across short bins outperforms simple thresholding on long bins.

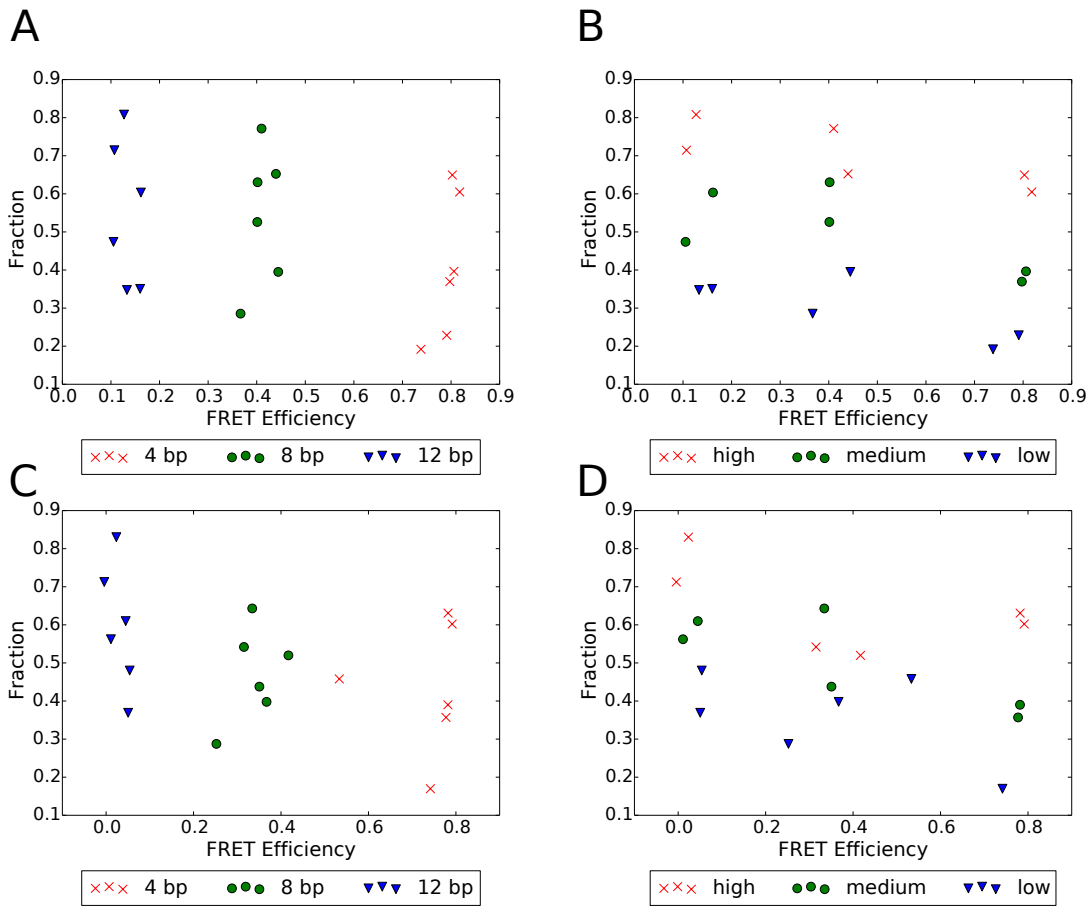

Figure 2.11: **Using pyFRET to fit multiple fluorescent populations from ALEX data.** ALEX data from mixtures of two DNA duplexes were analysed using pyFRET DCBS (A) and B)) or simple ALEX thresholding on re-binned data (C) and D)), then fitted using a two-component Gaussian mixture model. Three duplexes were used, to provide high-, intermediate- and low-FRET peaks. A) and C) Plot of FRET efficiency vs fractional population, coloured according to the dye-dye separation in the duplex. B) and D) Plot of FRET efficiency vs fractional population, coloured according to the population size.

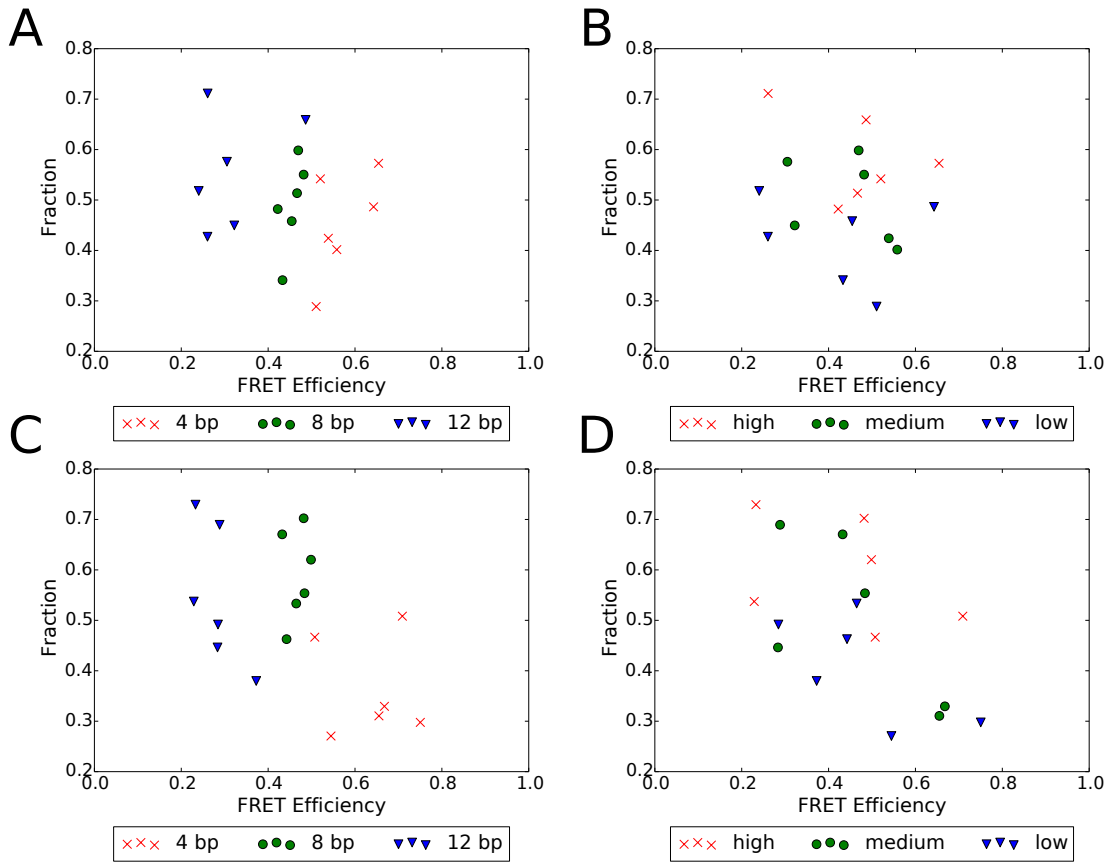

Figure 2.12: **Using pyFRET to fit multiple fluorescent populations from FRET data.** FRET data from mixtures of two DNA duplexes were analysed using pyFRET DCBS (A) and B)) or simple AND thresholding on re-binned data (C) and D)), then fitted using a two-component Gaussian mixture model. Three duplexes were used, to provide high-, intermediate- and low-FRET peaks. A) and C) Plot of FRET efficiency vs fractional population, coloured according to the dye-dye separation in the duplex. B) and D) Plot of FRET efficiency vs fractional population, coloured according to the population size.

2.7.4 Benchmarking the RASP Algorithm

We benchmarked the pyFRET implementaton of the RASP algorithm using data collected from a 1:1 mixture of 6 bp (high FRET) and 12 bp (low FRET) duplexes over a period of 10 hours (600 minutes). The parameters used in RASP analysis are shown in Table 2.3. RASP allows selection and analysis of bursts that occur within a time interval ΔT after bursts with FRET efficiency in the range $E_{min} - E_{max}$. We demonstrate the correct performance of the RASP algorithm by analysing bursts that occur in 1 ms intervals, centred at the stated times, following a high FRET burst. The resultant FRET histograms are shown in Fig. 2.13. At short recurrence intervals ($T_{max} < 8$ ms), not only do the great majority of events show a high FRET efficiency, but the number of events is greatly increased compared with longer recurrence intervals, demonstrating the enrichment of these events by molecules that have diffused back into the confocal detection volume.

2.8 Availability and Future Directions

pyFRET is available to download from PyPI under an open source BSD licence from the Python Package Index (`https://pypi.python.org/pypi/pyfret0.1.0`). Documentation can also be found here, whilst a more extensive tutorial, including example scripts, can be found on our website (`http://pyfret.readthedocs.org/en/latest/tutorial.html`).

The data from DNA duplexes used in evaluating pyFRET can be found in the DataDryad online repository: XXX DOI XXX. iPython Notebooks that reproduce the analyses performed here are also included.

smFRET is a fast-developing and active research field and although pyFRET provides the core tools for analysis of smFRET data, we do not currently implement all the algorithms used in analysis of smFRET data. In particular, we are currently not able to parse files generated using the picoQuant instrumentation and we do not implement stochastic denoising or photon distribution analysis [71, 72, 73, 74]. We are keen to extend the functionality of pyFRET and are happy to work with others to enable their use of and contribution to the pyFRET library.

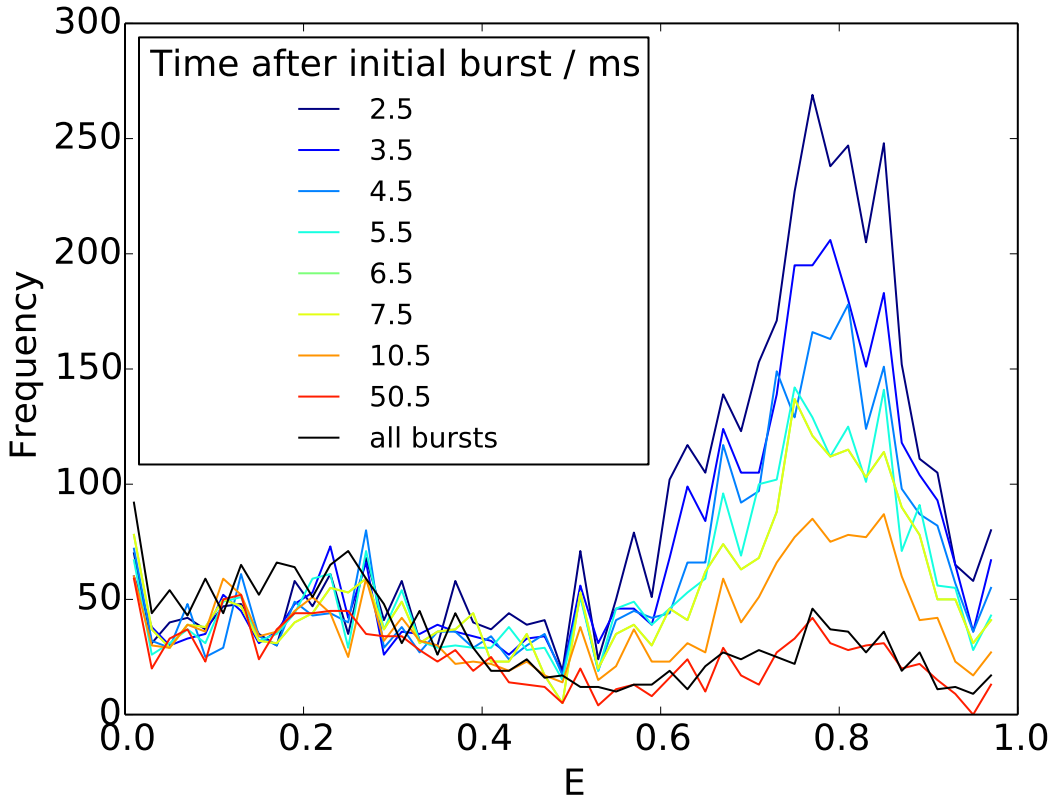

Figure 2.13: **RASP analysis of FRET data from a mixture of a high-FRET and low-FRET duplex.** The black line shows the baseline peak areas for the two duplexes. Other lines show the FRET histograms generated from bursts that occurred at the indicated time following a high-FRET burst. The Recurrence Interval ΔT was 1 ms, centred at the times shown in the legend. Note the greatly increased peak area for high-FRET bursts, demonstrating the recurrence of high-FRET events as molecules diffuse back into the confocal volume.

2.9 Conclusions

In this chapter, we presented pyFRET, a versatile open source library for analysis of smFRET data. In addition to demonstrating the usage of our event selection and fitting algorithms, we also thoroughly evaluated their performance. We note that even the sophisticated burst search algorithms are not able to overcome the biases introduced through event selection in simple FRET datasets. ALEX data collection, on the other hand, does not suffer from these biases, so is able to analyse a wide variety of smFRET datasets with considerable accuracy. Furthermore, the burst search algorithms, especially when combined with ALEX

data collection, are robust to changes in the values of parameters T, L and M, ensuring that conclusions drawn from smFRET data are not influenced by the thresholding parameters used. We hope that the evaluations and the pyFRET library will be of use to many researchers in the smFRET research field.

Chapter 3

Bayesian Inference of Intramolecular Distances Using Single Molecule FRET

3.1 Overview

This chapter introduces a novel method of analysing data from single moleclue FRET (smFRET) experiments using model-based Bayesian inference. The introduction opens with an overview of the smFRET experiment and describes historical approaches to data analysis. Following this introduction to the experimental environment, we provide an overview of model-based inference and explain why this is an appropriate technique for understanding data from smFRET experiments. The rest of the chapter comprises a detailed description of our parametric model of the smFRET experiment and our implementation of a Monte-Carlo Markov Chain (MCMC) algorithm that can simultaneously estimate population sizes and intramolecular distance information directly from a raw smFRET dataset, with no intermediate event selection and denoising steps.

The main results described in this chapter are as follows. Firstly, in the introductory section, we illustrate the limitations of a thresholding approach to data analysis through detailed examination of a smFRET dataset. We then present a parametric model of a smFRET experiment. Despite its apparent simplicity, this model is able to capture all key features of a smFRET dataset. Further, we introduce a model-based inference technique that infers the

parameters of this model, conditioned on an observed dataset. Following introduction of our parametric model and sampling algorithm, we validate our sampling-based analysis using simulated datasets. We show that our Monte-Carlo analysis can accurately infer parameter values over a wide range of dataset sizes, FRET efficiencies and molecular concentrations. Subsequently, we use further simulated datasets to compare the performance of the inference technique with simple thresholding techniques, showing that inference is not subject to the biases and inaccuracies seen in thresholding analyses. Following this, we use data from fluorescently labelled DNA duplexes to demonstrate the superior efficacy of our inference tool in analysing experimental smFRET datasets. We show that molecular concentrations and FRET efficiencies can be more accurately calculated using inference than using thresholding. Finally, we justify our model choice by comparing its performance with that of a simplified model, showing that this simplified model is insufficient to explain the heterogeneity observed in a smFRET experiment. The chapter concludes by considering possible extensions to the model-based analysis that would enable a wider range of datasets to be successfully analysed using these tools.

3.2 Introduction

3.2.1 A smFRET Experiment

The Experiment As described in the previous chapter, smFRET experiments rely on the phenomenon of Förster Resonance Energy Transfer (FRET), a distance-dependent non-radiative transfer of energy between fluorescent molecules [10, 1]. In a confocal smFRET experiment, molecules labelled with a FRET pair of dyes (a donor and acceptor) are diluted to pico-molar (pM) concentrations and allowed to diffuse freely in a dilute solution illuminated by a laser beam of the correct wavelenth th excite the donor dye. When a labelled molecule diffuses into the confocal volume, the laser excites the donor fluorophore and photons are emitted. Emitted photons are collected and separated by a dichroic mirror into donor and acceptor photons for analysis (Fig. 3.1 A).

In this chapter, we consider only experiments in which diffusing molecules are subject to continuous illumination from a single laser that excites the donor dye. More sophisticated experimental techniques have been developed. For example, ALEX excitation [35, 75] and Pulsed Interleaved Excitation (PIE) [57, 58] use rapid alternation of the exciting wavelength

to directly excite both the donor and the acceptor fluorophores, allowing specific identification of correctly labelled fluorescent molecules. Similary, Multi-Paramater Fluorescence Detection (MFD) [63, 76] separates photons based on both their wavelength and their polarisation, allowing concommitant separation of molecules based on their fluorescence anisotropy. We chose to focus on only the simple case of continuous excitation of the donor dye for two reasons. Firstly, in developing and evaluating a model-based analysis, it is advantageous to focus on the simplest incarnation of a problem. This enables us to quickly identify the key parts of the model necessary to accurately describe the data observed, without requiring huge computational complexity in order to model all features. Secondly, this is the type of experiment that is most commonly performed in our research environment. Consequently, we wished to make a tool that would be directly applicable to the types of experimental data that we encounter in our research. For similar reasons, we choose to consider only time-binned datasets, for which photons have been grouped into time-bins of length comparable to the dwell-time in the confocal volume, rather than considering the more sophisticated burst-search approaches that are popular elsewhere [27].

The Data As described above, 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 [77]. When a molecule diffuses into the confocal volume, the laser excites the donor fluorophore and photons are emitted. 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. The majority of bins ($> 95\%$) contain only background noise; the rest contain both background noise and photons from fluorescent bursts (Fig. 3.5 C-D). Consequently, the raw data from a confocal smFRET experiment consists of two lists of integers, corresponding to donor and acceptor photon counts. A typical experiment, lasting a minimum of 10 minutes, will consist of at least six million time-bins; many datasets contain upwards of 30 million time-bins, most of which contain only noise photons.

Aside from the delicate experimental apparatus, one of the key challenges in smFRET experiments is therefore the selection of photon bursts originating from flourescent excitation events, and the accurate denoising of these identified bursts by removal of background contributions. The following section describes this analytical challenge in detail and explains why simple thresholding approaches are not appropriate to accurately select fluorescent events.

3.2.2 Approaches to Analysis of smFRET Data

Event Selection: Thresholding Analysing a smFRET dataset has two main challenges: identifying fluorescent events and separating the fluorescence emission contribution from noise photons. We call these two steps event selection and event denoising. Event selection is typically achieved by applying a photon count threshold to the time-bins and retaining as fluorescent events only the time-bins for which that threshold is exceeded [55, 56, 78]. AND thresholding applies a threshold across both donor and acceptor time-bins: $n_D > T_D$ AND $n_A > T_A$; SUM thresholding applies a threshold to the sum of the photons observed in the donor and acceptor time-bins: $n_D + n_A > T$, for n_D and n_A donor and acceptor photon counts; and T_D, T_A, and T the photon count thresholds applied to the donor channel, acceptor channel and total photon counts respectively. We note that, as described in the previous chapter, the APBS and DCBS burst search algorithms [27], as applied to simple FRET data, are analagous to SUM and AND thresholding respectively.

These thresholding-based approaches to event selection rely on the assumption that fluorescence events and background events are linearly separable, allowing application of a simple threshold to effectively isolate the fluorescent events (Fig. 3.2). However, when data from a smFRET experiment are examined in close detail, it is clear that this is not the case. Fig. 3.3 (A) shows the raw data from a typical smFRET experiment, plotted as a two-dimensional scatter-plot. Unlike the idealised dataset illustrated in Fig. 3.2, in the real dataset there is no clear separation between signal and noise. Consequently, no threshold can accurately separate true events from noise. This problem is exacerbated at very high and very low FRET efficiencies, where thresholds not only remove a large fraction of fluorescent bursts, but do so in a manner that will distort the resulting FRET histogram (Fig. 3.4).

Event Denoising The second challenge in analysing a smFRET dataset is event denoising. Time-bins selected as containing fluorescent bursts still contain a noise contribution to their total photon count. The most straigtforward method to remove this contribution is to calculate the average photon count in each channel across time-bins from a dataset containing no fluorophores, and then subtract these averaged (usually non-integral) values from each fluorescent event [27]. If the average noise is different between the donor and the acceptor channels, this has the effect of linearly shifting the FRET efficiency histogram by a small amount and can result in analysis artifacts, such as negative, fractional photon counts and negative FRET efficiencies.

Removal of cross-talk photons and direct excitation contributions to the acceptor channel are achieved in a similar manner. The average ratio of photons in the donor and acceptor channels is calculated for a) a concentrated sample of donor dye and b) a concentrated sample of acceptor dye. These ratios, termed the "leakage" and "direct excitation" constants, l and d respectively, are then used to correct the ratio of observed donor and acceptor photons in each retained fluorescent burst (Eq. 3.1):

$$n_A = n_A - l \cdot n_D \qquad n_A = n_A - d \cdot n_A \tag{3.1}$$

for n_A and n_D photons in the acceptor and donor channels respectively

FRET Efficiency Calculation Following these event selection and denoising steps, FRET efficiencies are calculated for the denoised bins using:

$$E = \frac{n_A}{n_A + \gamma \cdot n_D} \tag{3.2}$$

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].

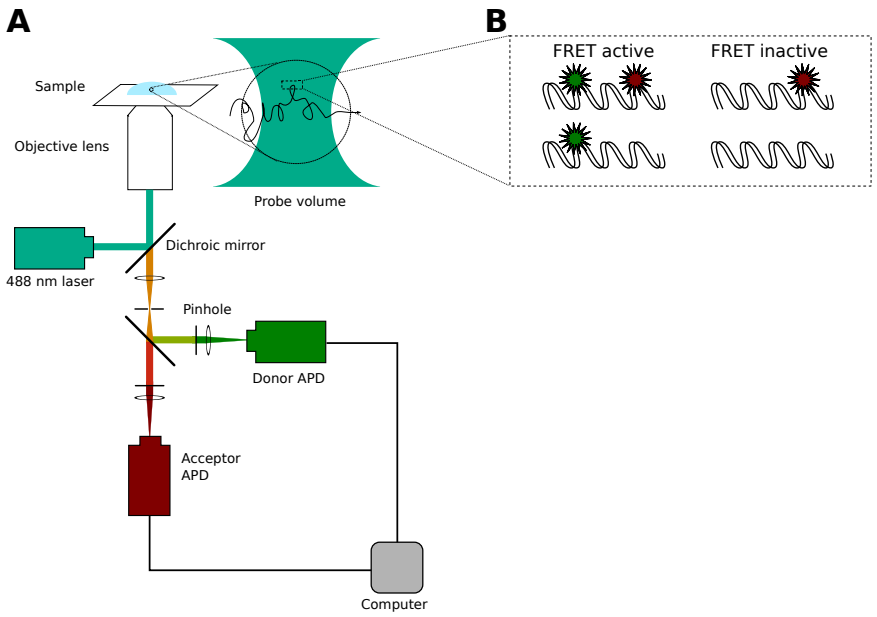

Figure 3.1: A typical smFRET experiment. (A) Microscope set-up for smFRET. APD: Avalanche Photodiode. (B) The four possible labelling states for a single molecule in the confocal volume.

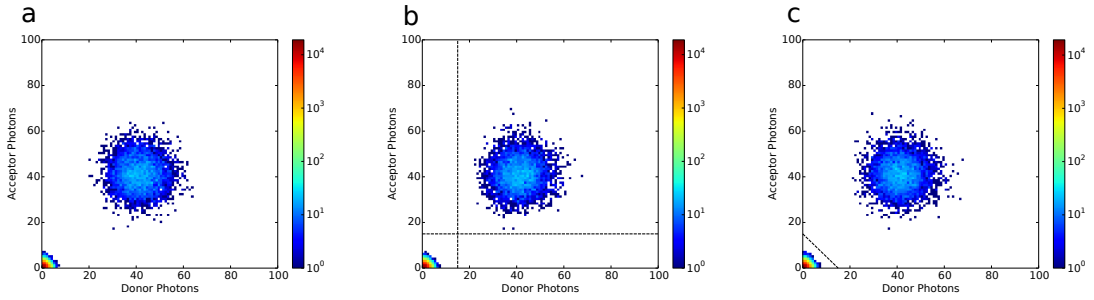

Figure 3.2: A simulated smFRET dataset, for which thresholding is appropriate to separate fluorescent bursts from background noise. a) There is a clear separation between noise emission and fluorescent burst, allowing correct isolation of burst data using both b) AND thresholding (thresholds 15, 15) and c) SUM thresholding (threshold 15)

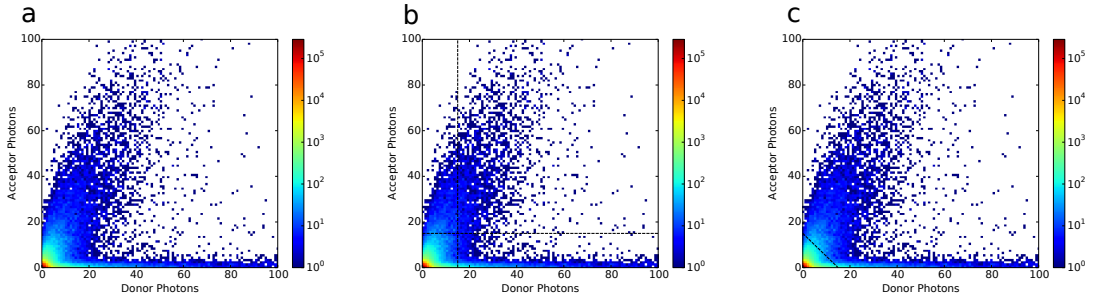

Figure 3.3: A real smFRET dataset, from a DNA duplex with a 6 bp separation between donor and acceptor dye attachment sites. a) In this dataset, there is no clear separation between noise photons and fluorescent events. Consequently neither b) AND thresholding (thresholds 15, 15) nor c) SUM thresholding (threshold 15) can accurately isolate burst data.

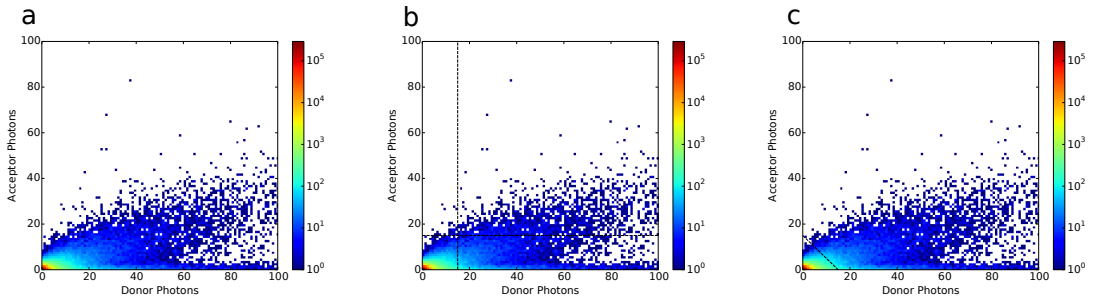

Figure 3.4: A real smFRET dataset, from a DNA duplex with a 10 bp separation between donor and acceptor dye attachment sites. a) This duplex has a low FRET efficiency, meaning that not only is there is no clear separation between noise photons and fluorescent events, but also that both b) AND thresholding (thresholds 15, 15) and c) SUM thresholding (threshold 15) distort the observed FRET histograms by removing a selection of fluorescent events that is not randomly distributed across all observed FRET effiencies.

Stochastic Denoising As described above, these methods of event selection and denoising are simple but prone to artifacts that distort the outcome of downstream analyses. To mitigate this problem, several stochastic denoising methods [71, 72, 73, 74], based on poisson statistics, have been developed. The most popular of these is a Probability Distribution Analysis (PDA) [72], which attempts to derive the shot-noise limited FRET histogram by explicitly considering the origin of photons observed on the donor and acceptor channels.

PDA considers observed ratios of donor and acceptor photons, termed the signal ratio, $\frac{S_D}{S_A}$, and total number of photons due to fluorescence emission, F. The probability of observing a certain signal ratio, is given by summing, over all combination of noise and fluorescent photons that yield such a ratio, the probability of observing that combination of photons:

$$Pr(\frac{S_G}{S_R})_i = \sum_{\text{all } F, F_{RT} B_G, B_R \text{ yielding } (\frac{S_G}{S_R})_i} \Pr(F) \Pr(F_{RT}|F) \Pr(B_G) \Pr(B_R) \tag{3.3}$$

where B_G and B_R are respectively the number of background photons in the green (donor) and red (acceptor) channels, F is the total number of fluorescence photons observed, and F_{RT} is the number of fluorescent photons observed in the red channel (due to FRET).

The noise emission distributions are assumed to be poisson, such that the probability of observing B background photons is given by Eq. 3.4 for a mean number of background photons $\langle B \rangle$. Similarly, the ratio of donor and acceptor fluorescence photons can be described by a binomial distribution with parameter ϵ, an apparent FRET efficiency (Eq. 3.5.

$$\Pr_{\langle B \rangle}(B) = \frac{\langle B \rangle^B \cdot e^{-\langle B \rangle}}{B!} \tag{3.4}$$

$$\Pr_{\epsilon}(F_{RT}|F) = \frac{F!}{F_{RT}!(F - F_{RT})!} \cdot \epsilon^{F_{RT}} \cdot (1 - \epsilon)^{F - F_{RT}} \tag{3.5}$$

Consequently, Eq. 3.3 can be rewritten as:

$$Pr(\frac{S_G}{S_R})_i = \sum_{\text{all } N, F_{RT} B_G, B_R \text{ yielding } (\frac{S_G}{S_R})_i} \Pr(N) \Pr_{\epsilon}(N - B_G - B_R|F) \Pr_{\langle B_G \rangle}(B_G) \Pr_{\langle B_R \rangle}(B_R) \tag{3.6}$$

$\Pr(N)$ can be estimated from the dataset whilst $\langle B_G \rangle$ and $\langle B_R \rangle$ are estimated from the count

rates. Consequently, the only unknown in this equation is the mean value of ϵ, which is then estimated using a Levenberg-Marquardt χ^2 optimisation.

Despite its sophisticated treatment of noise photons in determining the underlying FRET parameter ϵ, the PDA analysis still uses a simple threshold to select events and assumes that this provides an unbiased sample of fluorescent bursts [71, 72]. As we have illustrated above, in the absence of information from direct acceptor excitation, this is not a valid assumption. Consequently, we would like to find a method of performing event selection and denoising that does not suffer from these biases.

3.2.3 Model Based Inference

As an alternative to the thresholding analyses described above, we would like to use a model-based inference method for analysis of smFRET data. This would enable simultaneous inference of all useful parameters – such as population sizes, FRET efficiencies and mean noise levels – directly from the raw smFRET dataset, without requiring multiple event selection and denoising steps. The following sections provide an overview of Bayesian statistics and describe analysis of smFRET data in terms of a model-based inference problem.

Bayesian Inference Model-based Bayesian inference is a probabilistic method [79] that uses conditional probabilities based on Bayes' Theorem [80] to assess the likelihood that a series of observations were generated by a given model [81]. Bayes' Rule (Eq. 3.8) derives the posterior probability of some event A given some observed event B, by linking their conditional probabilities, $\Pr(A|B)$ and $\Pr(B|A)$:

$$\begin{aligned} \Pr(A|B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \qquad \Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \\ \Pr(A \cap B) &= \Pr(A|B) \cdot \Pr(B) = \Pr(B|A) \cdot \Pr(A) \end{aligned} \tag{3.7}$$

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)} \tag{3.8}$$

In model-based inference, this is reframed to describe the probability that a given model generated an observed dataset, conditioned on the data observed:

$$\Pr(\text{model}|\text{data}) = \frac{\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})}{\Pr(\text{data})} \tag{3.9}$$

Application to Confocal smFRET Data Analysis techniques based on Bayesian statistics are well established for analysis of smFRET data collected from immobilised molecules [82, 83, 84, 85, 86, 87]. Bayesian methods have also been applied to single particle tracking [88]; analysis of diffusional trajectories [89, 90]; fluorescence correlation spectroscopy [91, 92, 93, 94] and fluorescence lifetime data [95, 96]. An excellent theoretical understanding of the physical FRET process has been developed by Gopich and Szabo [97, 98, 99], which has facilitated a Maximum Likelihood approach to analysis of fluorescent bursts from diffusing molecules [99, 100]. However, these methods do not consider burst selection [99] or apply only to removal of shot-noise from selected bursts [97], so assume access to idealised simulated [97], or pre-selected [100, 99] and denoised fluorescent traces [95].

Nevertheless, confocal smFRET experiments are an excellent candidate for successful model-based data analysis. Inference analysis allows us to simultaneously infer all parameters of a model, conditioned on the observed dataset. Consequently, we can accurately infer parameters of interest, such as molecular concentrations and inter-dye distances, in a single step, removing the need for event selection heuristics. The underlying physical processes of excitation and emission are well studied, allowing a simple but accurate model to capture all key features of a smFRET dataset using only a small number of parameters. Furthermore, a large amount of data can be collected in a short period of time, allowing parameters to be inferred with high confidence.

In the following section, we describe our physical model of the smFRET experiment. We explain the structure of this parametric model and relate its parameters to the physical experiment. We then describe the sampling technique used to infer parameters of the model, conditioned on a smFRET datset. Following this theoretical discussion, we apply model-based analysis to infer parameters from smFRET datasets. Firstly, we use simulated datasets, for which the underlying parameters are known. This enables us to evaluate the performance of the model under different conditions. We also present a comparison with standard thresholding based tools for event selection and denoising, showing that model-based inference is superior over a wide range of datasets. Secondly, we compare the performance of model-based inference and thresholding when applied to experimental data, again showing that inference analysis more accurately estimates the sizes and intramolecular distances of

multiple fluorescent populations. Finally, we present a brief discussion of our design choices in the parametric model, justifying in particular our use of a gamma-poisson mixture model to describe fluorescence emission. The chapter concludes with a discussion of potential extensions to our model, which would allow model-based inference to be used in the analysis of a wider range of confocal smFRET datasets.

3.3 Theory

3.3.1 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 [78]. The key innovation we present here is the development of a model-based Bayesian analysis of smFRET data. 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. 3.2). However, whereas traditional techniques use the ratio of donor and acceptor photons observed, we model the energy transfer as altering the underlying rates of dye photon emission, using directly the distance-dependence of FRET energy transfer:

$$E = \frac{1}{1 + (\frac{r}{R_0})^6} \tag{3.10}$$

for r, the inter-fluorophore distance and R_0 the Förster distance characteristic of the fluorophores used.

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 [77]. 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. 3.1 A).

As described in Chapter 2, 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 [101], 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. 3.5 A). The majority of bins ($> 95\%$) contain only background noise; the rest contain both background noise and photons from fluorescent bursts (Fig. 3.5 C-D).

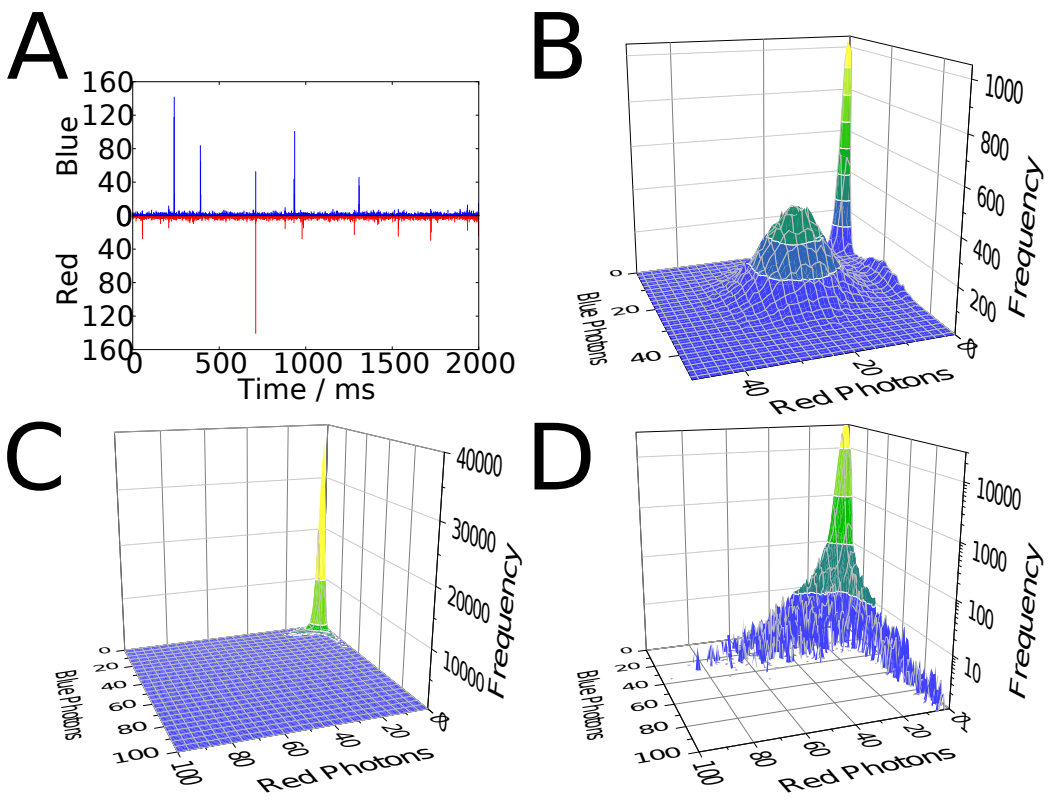

Figure 3.5: A typical smFRET dataset. (A) Snapshot of raw smFRET data from a high-FRET dual-labelled DNA duplex. (B-D) show three-dimensional histograms of raw photon counts from smFRET datasets. (B) An idealised, simulated smFRET dataset, with signal and noise well separated, for which thresholding 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.

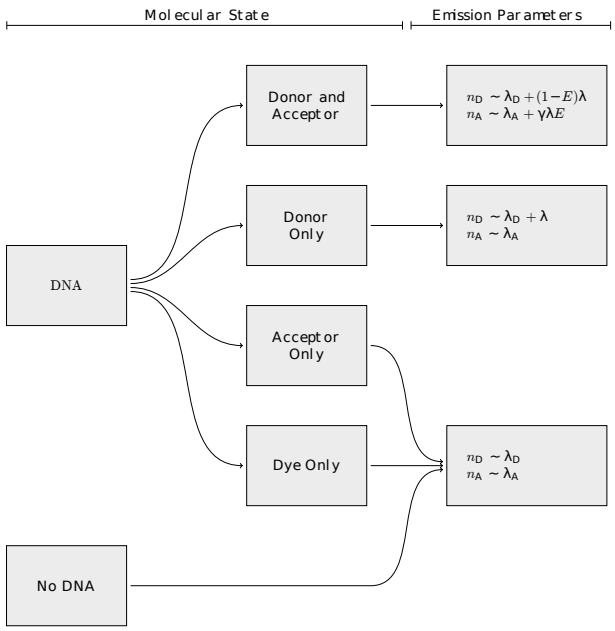

Figure 3.6: Flow diagram illustrating the generative model for a single FRET population. The molecular state is the underlying state of the current observation; the emission parameters are the Poisson parameters that result in observable photon emission - these are shown for both donor and acceptor channels.

We model such 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:

$$n_D \sim \text{Poisson}(n_D; \lambda_D) = \frac{\lambda_D^{n_D}}{n_D!} e^{-\lambda_D} \qquad n_A \sim \text{Poisson}(n_A; \lambda_A) = \frac{\lambda_A^{n_A}}{n_A!} e^{-\lambda_A} \tag{3.11}$$

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}:

$$n_{\text{prot}} \sim \text{Poisson}(n_{\text{prot}}; \lambda_{\text{prot}}) = \frac{\lambda_{\text{prot}}^{n_{\text{prot}}}}{n_{\text{prot}}!} e^{-\lambda_{\text{prot}}} \tag{3.12}$$

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:

$$n_{\text{prot1}} \sim \text{Poisson}(n_{\text{prot1}}; \lambda_{\text{prot1}}) = \frac{\lambda_{\text{prot1}}^{n_{\text{prot1}}}}{n_{\text{prot1}}!} e^{-\lambda_{\text{prot1}}} \tag{3.13}$$

and

$$n_{\text{prot2}} \sim \text{Poisson}(n_{\text{prot2}}; \lambda_{\text{prot2}}) = \frac{\lambda_{\text{prot2}}^{n_{\text{prot2}}}}{n_{\text{prot2}}!} e^{-\lambda_{\text{prot2}}} \tag{3.14}$$

As before, most bins contain no fluorescent molecules ($n_{\text{prot2}} = n_{\text{prot2}} = 0$), but multiple occupancy can be modelled when $n_{\text{prot2}} + n_{\text{prot2}} > 1$. Note that we are using n_{prot} as the confocal occupation parameter to reflect the common usage of smFRET to study protein dynamics and for continuity with Chapter 4 which discusses protein aggregation; this of course does not preclude use of the model for systems involving DNA or other (biological) molecules.

Each molecule present may be in one of four labelling states: unlabelled, donor-only, acceptor-only or dual-labelled (Fig. 3.1 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. 3.15). 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.

$$\lambda \sim \text{Gamma}(\lambda; k_D, \theta) = \frac{1}{\Gamma(k_D)\theta^{k_D}} \lambda^{(k_D-1)} e^{-\frac{\lambda}{\theta}} \quad \text{for } \theta = \lambda_B / k_D \tag{3.15}$$

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.

The choice of a gamma-poisson mixture model to describe the fluorescence emission behaviour was not arbitrary. It is well understood that photon emission from molecules passing through the confocal volume displays super-poissonian behaviour [102], where the population variance exceeds the mean, owing to the inhomogeneous excitation profile. The negative binomial distribution is a very well characterised and popular probability distribution for modeling overdispersed data for which all observations are positive [103, 104]. It can be parameterised as a gamma-poisson mixture [103], where emission behaviour follows a Poisson distribution with gamma-distributed intensity. Under this parameterisation, λ_B is the global mean emission rate of photons, whilst the shape parameter k_D describes the degree of overdispersion caused by confocal volume inhomogeneity. At the end of this chapter (Section 3.5.1) we briefly compare performance of inference analysis using the gamma-poisson mixture model with a simplified analysis that assumes pure poisson emission, demonstrating the need for overdispersion in the model to accurately fit the tails of the photon emission distributions.

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. 3.15. The number of additional photons is then c_D:

$$c_D \sim \text{Poisson}(c_D; \lambda). \tag{3.16}$$

and the total observed photons in the donor channel, f_D is the sum of noise and fluorescence photons, namely $n_D + c_D$; 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. 2.1); λ is the unattenuated rate of donor photon emission associated with the observed molecule; and γ is an instrumental correction factor (Eq. 3.2). The additional photons in each channel, c_D and c_A are thus distributed as:

$$c_D \sim \text{Poisson}(c_D; \lambda(1-E)) \qquad \text{and} \qquad c_A \sim \text{Poisson}(c_A; \gamma\lambda E) \tag{3.17}$$

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:

$$c_{D1} \sim \text{Poisson}(c_{D1}; \lambda(1-E_1)) \qquad \text{and} \qquad c_{A1} \sim \text{Poisson}(c_{A1}; \gamma\lambda E_1) \tag{3.18}$$

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

$$c_{D2} \sim \text{Poisson}(c_{D2}; \lambda(1-E_2)) \qquad \text{and} \qquad c_{A2} \sim \text{Poisson}(c_{A2}; \gamma\lambda E_2) \tag{3.19}$$

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 mathematical model describes a Bayesian belief network and is shown as a directed acyclic graph in standard plate notation in Fig. 3.7. It can be used, with appropriate parameters (see Table 3.1), 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 Fig. 3.8), suggesting that, despite its many simplifications, such as the

neglect of direct excitation and leakage effects, this model captures the most important features sufficient to analyse the FRET process. In the following section, we describe our custom-built inference algorithm, based on MCMC Metropolis sampling [44], which can be used to infer values for all relevant physical parameters of the model.

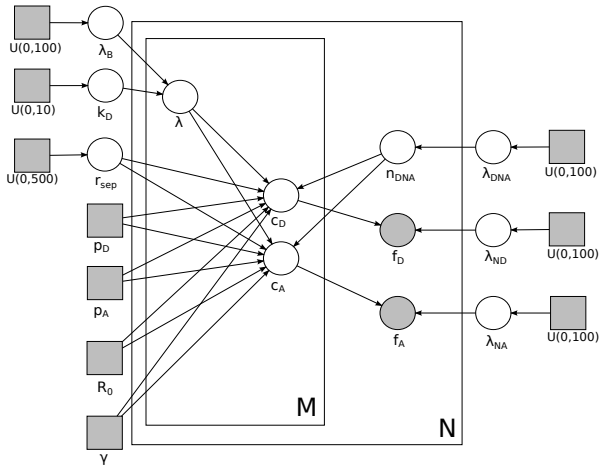

Figure 3.7: Directed Acyclic Graph illustrating the interrelation of parameters in the inference model. In this notation, circles represent random variables, while squares represent constants. Known or observed values are shaded, while hidden variables are not. For each time bin in a dataset of size 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.

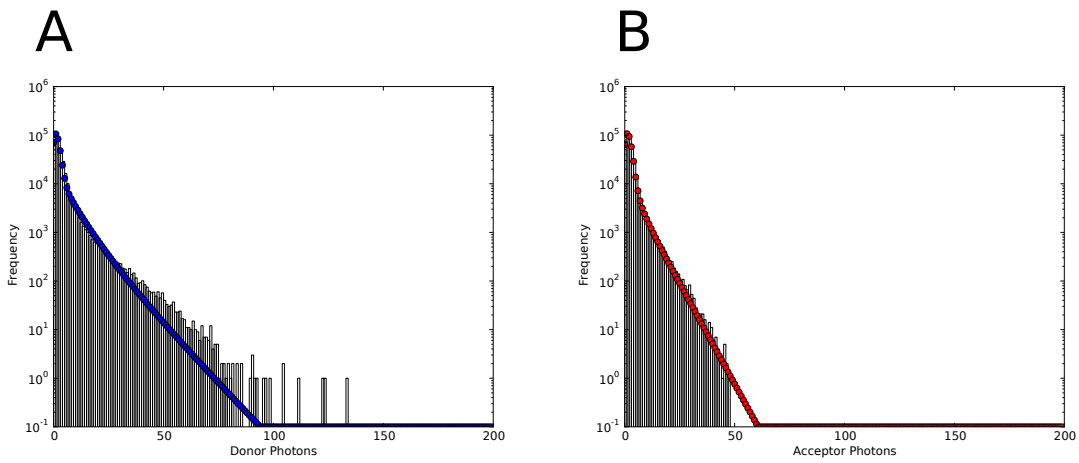

Figure 3.8: Histograms showing the marginal distributions of donor and acceptor photons in a smFRET dataset (1:1 mixture of 4bp and 12 bp DNA duplexes). (A) Marginal distributuon photons in the donor channel. Histogram shows the number of time bins observed to contain this many donor photons. Blue circles 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.

3.3.2 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.

Estimating 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 would also to know other associated parameters, such as the rate of noise emission in each channel and the average brightness of fluorescent events. These values are described explicitly as parameters in the 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

$$\Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E | (f_D, f_A)_n]$$

for a smFRET dataset with n time-bins, as described in Section 3.3.1. As described above (Section 3.2.3), 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 (Eq. 3.9) to derive the required distribution over the parameters of our model:

$$\begin{aligned}\Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E | (f_D, f_A)_n] = \\ = \frac{\Pr[(f_D, f_A)_n | \lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E] \cdot \Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E]}{\sum_{\forall \lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E} \Pr[(f_D, f_A)_n | \lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E] \cdot \Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E]}.\end{aligned} \quad (3.20)$$

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. 3.20 and exact determination of the posterior distribution, $\Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E|(f_D, f_A)_n]$ is not possible.

However, to estimate the distribution of values taken by the parameters of interest, it is not necessary to evaluate Eq. 3.20, exactly. It is sufficient to draw parameter samples distributed proportionally to the posterior distribution, $\Pr[(f_D, f_A)_n|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E] \cdot \Pr[\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E]$ [44]. 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 [48], [44] is a Monte Carlo Markov Chain (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. 3.20. Metropolis [49] provides an introduction to the algorithm; our implementation is described below.

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:

$$
\begin{aligned}
&\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E] \\
&= \Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \text{noise only}] \cdot \Pr[\text{noise only}] \\
&+ \Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, p_D, p_A, \text{donor event}] \cdot \Pr[\text{donor event}] \\
&+ \Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A, \text{FRET event}] \cdot \Pr[\text{FRET event}] \\
&+ \Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A, \text{multiple occupancy}] \cdot \Pr[\text{multiple occupancy}],
\end{aligned}
\tag{3.21}
$$

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:

$$\begin{aligned}&\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \text{noise only}]\\&= ((1-p_{\text{prot}}) + p_{\text{prot}}(1-p_D)(1-p_A) + p_{\text{prot}}(1-p_D)p_A)\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \text{noise only}]\end{aligned} \tag{3.22}$$

$$\begin{aligned}&\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, p_D, p_A, \text{donor event}]\\&= p_{\text{prot}}p_D(1-p_A)\Pr[(f_d, f_a)_i|\lambda_{n_D} + \lambda, \lambda_{n_A}, \text{Donor only}]\end{aligned} \tag{3.23}$$

$$\begin{aligned}&\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A, \text{FRET event}]\\&= p_{\text{prot}}p_D p_A \Pr[(f_d, f_a)_i|\lambda_{n_D} + \lambda(1-E), \lambda_{n_A} + \lambda E\gamma, \text{FRET}]\end{aligned} \tag{3.24}$$

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. 3.22 and Eq. 3.23 can still be used to describe the probability that an event is generated by noise only (Eq. 3.22) or by a donor-only fluorescent event (Eq. 3.23), provided that the p_{prot} terms are replaced by the sum $p_{\text{prot}} = p_{\text{prot1}} + p_{\text{prot2}}$. However, equation 3.24 needs modification to accommodate the multiple FRET efficiencies E_1 and E_2 as well as their respective population sizes:

$$\begin{aligned}&\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A, \text{FRET event}]\\&= p_{\text{prot}}p_D p_A \Pr[(f_d, f_a)_i|\lambda_{n_D} + \lambda(1-E_1), \lambda_{n_A} + \lambda E_1\gamma, \text{FRET}_1]\\&+ p_{\text{prot}}p_D p_A \Pr[(f_d, f_a)_i|\lambda_{n_D} + \lambda(1-E_2), \lambda_{n_A} + \lambda E_2\gamma, \text{FRET}_2].\end{aligned} \tag{3.25}$$

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:

$$\Pr[(f_d, f_a|\lambda, \lambda_D, \lambda_A] = \text{Poisson}(f_A; \lambda_A) \cdot \int_0^\infty \text{Poisson}(f_D; \lambda + \lambda_D) \cdot \text{Gamma}(\lambda; k_D, \theta)\, d\lambda \tag{3.26}$$

$$= \int_0^\infty \frac{(\lambda + \lambda_D)^{-f_D} e^{-(\lambda+\lambda_D)}}{f_D!} \frac{\lambda_A^{-f_A} e^{-\lambda_A}}{f_A!} \lambda^{k_D-1} \frac{e^{-\frac{\lambda}{\theta}}}{\theta^{k_D}\Gamma(k_D)}\, d\lambda, \tag{3.27}$$

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

$$\Pr[f_d, f_a|\lambda, E, \gamma, \lambda_D, \lambda_A] = \tag{3.28}$$

$$= \int_0^\infty \text{Poisson}(f_D; \lambda(1-E) + \lambda_D) \cdot \text{Poisson}(f_A; \lambda E\gamma + \lambda_A) \cdot \text{Gamma}(\lambda; k_D, \theta)\, d\lambda \tag{3.29}$$

$$= \int_0^\infty \frac{(\lambda(1-E) + \lambda_D)^{-f_D} e^{-(\lambda(1-E)+\lambda_D)}}{f_D!} \frac{(\lambda E\gamma + \lambda_A)^{-f_A} e^{-(\lambda E\gamma+\lambda_A)}}{f_A!} \lambda^{k_D-1} \frac{e^{-\frac{\lambda}{\theta}}}{\theta^{k_D}\Gamma(k_D)}\, d\lambda, \tag{3.30}$$

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:

$$\Pr[(f_d, f_a)_i|\lambda, E_1, E_2, \gamma, \lambda_D, \lambda_A] = P_1 \cdot \Pr[(f_D, f_A)_i|\lambda, E_1, \gamma, \lambda_D, \lambda_A] + P_2 \cdot \Pr[(f_D, f_A)_i|\lambda, E_2, \gamma, \lambda_D, \lambda_A] \tag{3.31}$$

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 3.30 and P_1 and P_2, given by Eqn. 3.32 below, describe the relative sizes of the two fluorescent populations.

$$P_1 = \frac{\lambda E_1}{\lambda E_1 + \lambda E_2} \qquad P_2 = \frac{\lambda E_2}{\lambda E_1 + \lambda E_2} \tag{3.32}$$

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 explicitly modelled 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:

$$\begin{aligned}&\Pr[(f_d, f_a)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A, \text{multiple events}] \\ &= (1 - \lambda_{\text{prot}} e^{-2\lambda_{\text{prot}}}) \frac{\Gamma(f_D + r)!}{f_D! \Gamma(r)} \frac{\Gamma(f_A + r)!}{f_A! \Gamma(r)} (1 - \frac{\mu_D}{r + \mu_D})^r (\frac{\mu_D}{r + \mu_D})^{f_D} (1 - \frac{\mu_A}{r + \mu_A})^r (\frac{\mu_A}{r + \mu_A})^{f_A}\end{aligned} \tag{3.33}$$

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:

$$\mu_D = \frac{p_2(2\lambda(1 - E)) + p_3(3\lambda(1 - E))}{p_2 + p_3} \qquad \mu_A = \frac{p_2(2\lambda\gamma E) + p_3(3\lambda\gamma E)}{p_2 + p_3}, \tag{3.34}$$

where:

$$p_2 = \frac{\lambda_{\text{prot}}^2}{2!} e^{-\lambda_{\text{prot}}} \qquad p_3 = \frac{\lambda_{\text{prot}}^3}{3!} e^{-\lambda_{\text{prot}}}, \tag{3.35}$$

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 (3.33).

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 3.21. The probability that the whole dataset was generated by a those parameters is then the product:

$$\Pr[\text{Obs.}|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A] = \prod_i \Pr[(f_D, f_A)_i|\lambda_D, \lambda_A, \lambda_{\text{prot}}, \lambda_B, E, p_D, p_A] \tag{3.36}$$

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.

3.3.3 The Metropolis-Hastings Algorithm

Section 3.3.1 described how an observed dataset (Obs.) can be related to the values of a parametric model ($\lambda_D, \lambda_A, \lambda_{\text{prot}}$ etc.), showing that it is possible to describe probabilistically the likely parameter values, conditioned on the observed data. As described in Section 1.5, parameter inference can be achieved by sampling from a Markov Chain that has as its stationary probability the posterior probability over the parameters, Pr[Obs.|Par.]·Pr[Par.]. We now describe the development of a custom implementation of the Metropolis algorithm that can infer the parameters of our model conditioned on a smFRET dataset. 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(Obs.|Par.) is evaluated for the current set of parameters, using equation 3.21 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(Obs.|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:

$$a = \frac{\Pr(\text{Obs.}|x')}{\Pr(\text{Obs.}|x)} \tag{3.37}$$

  where Pr(Obs.|x') and Pr(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(Obs.|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. 3.9).

3.4 Experimental 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 (see XXX introduction). Details of the instrumentation and DNA duplex preparation are described below. 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 [55]. 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 3.1 were used. The outcome of the inference was not sensitive to initial conditions (see Fig. 3.9).

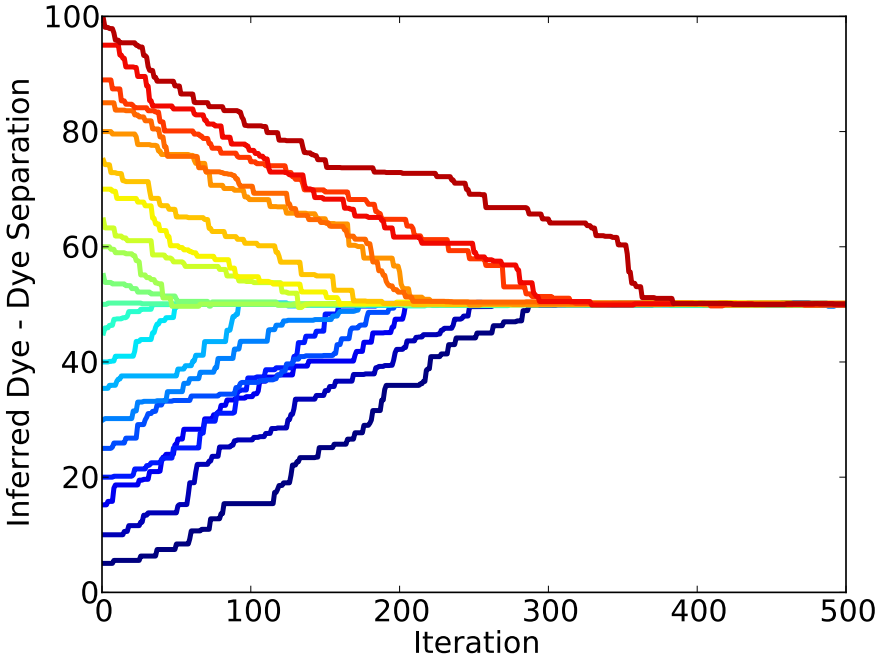

Figure 3.9: Graph illustrating the convergence of the Metropolis sampler during the initial burn-in period of the sampling. Analysis of the same dataset was initiated using different values of the dye-dye separation from 5 Å(in blue) to 100 Å(in red), in steps of 5 Å. After 500 iterations of the sampler (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.

Table 3.1: **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

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 3.2. These were annealed to one of the five 647-labelled acceptor sequences shown in Table 3.3. 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 3.2: **DNA sequence** 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	TACTGCCTTTCTGTATCGC**5**TATCGCGTAGTTACCTGCCTTGCATAGCCACTCATAGCCT

Table 3.3: **Preparing the dual-labelled dsDNA.** An acceptor-labelled ssDNA, with the sequence shown was annealed to the donor construct described in Table S1, 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**
4	AGGCTATGAGTGGCTATGCAAGGCAGGTAACTACGCGATAAGCGA**6**
6	AGGCTATGAGTGGCTATGCAAGGCAGGTAACTACGCGATAAGCGATA**6**
8	AGGCTATGAGTGGCTATGCAAGGCAGGTAACTACGCGATAAGCGATACA**6**
10	AGGCTATGAGTGGCTATGCAAGGCAGGTAACTACGCGATAAGCGATACAGA**6**
12	AGGCTATGAGTGGCTATGCAAGGCAGGTAACTACGCGATAAGCGATACAGAAA**6**

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:

$$fr_D = \frac{n_{\text{donor}}}{n_{\text{total}}} \qquad fr_A = \frac{n_{\text{acceptor}}}{n_{\text{total}}} \tag{3.38}$$

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:

$$n_{\text{total}} = n_{\text{donor}} + n_{\text{acceptor}} - n_{\text{FRET}} \tag{3.39}$$

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 [35]. Software for implementation of the ALEX analysis was written in Python.

3.5 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. 3.1. The results are summarised in Fig. 3.10. Fig. 3.10 (A) and (B) show the FRET efficiencies inferred for datasets with dye-dye distances across the spectrum of FRET efficiencies. From Fig. 3.10 (A) it can be seen that the inference method correctly reproduces the expected sigmoidal curve of FRET efficiency against dye separation, whilst Fig. 3.10 (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. 3.10 (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. 3.10 (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. 3.10 (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. 3.12, show that inference and thresholding are equally good at determining FRET efficiency, but that inference far outperforms thresholding in determining population sizes.

Fig. 3.12 (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. 3.12 (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. 3.12 (A) show the relationship between actual and calculated population size for a range of different FRET efficiencies. Here, inference (Fig. 3.12 (A) middle) is clearly superior, showing no variation in the observed population size with FRET efficiency. Both thresholding techniques (Fig. 3.12 (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. 3.12 (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 (Fig. 3.11). 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

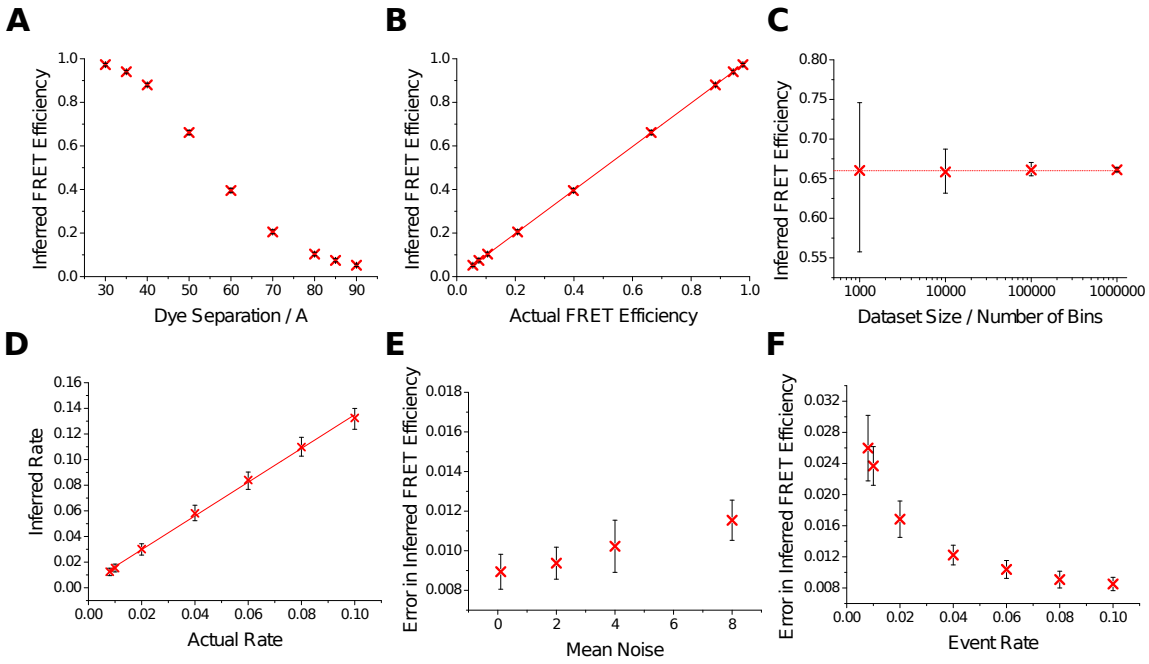

Figure 3.10: Validation of inference technique using realistic simulated datasets. (A) Inferred FRET efficiency plotted against dye-dye distance. (B) Inferred FRET efficiency plotted against calculated FRET efficiency. (C) Mean inferred FRET efficiency plotted against dataset size, for synthetic datasets 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 $> 99\%$.

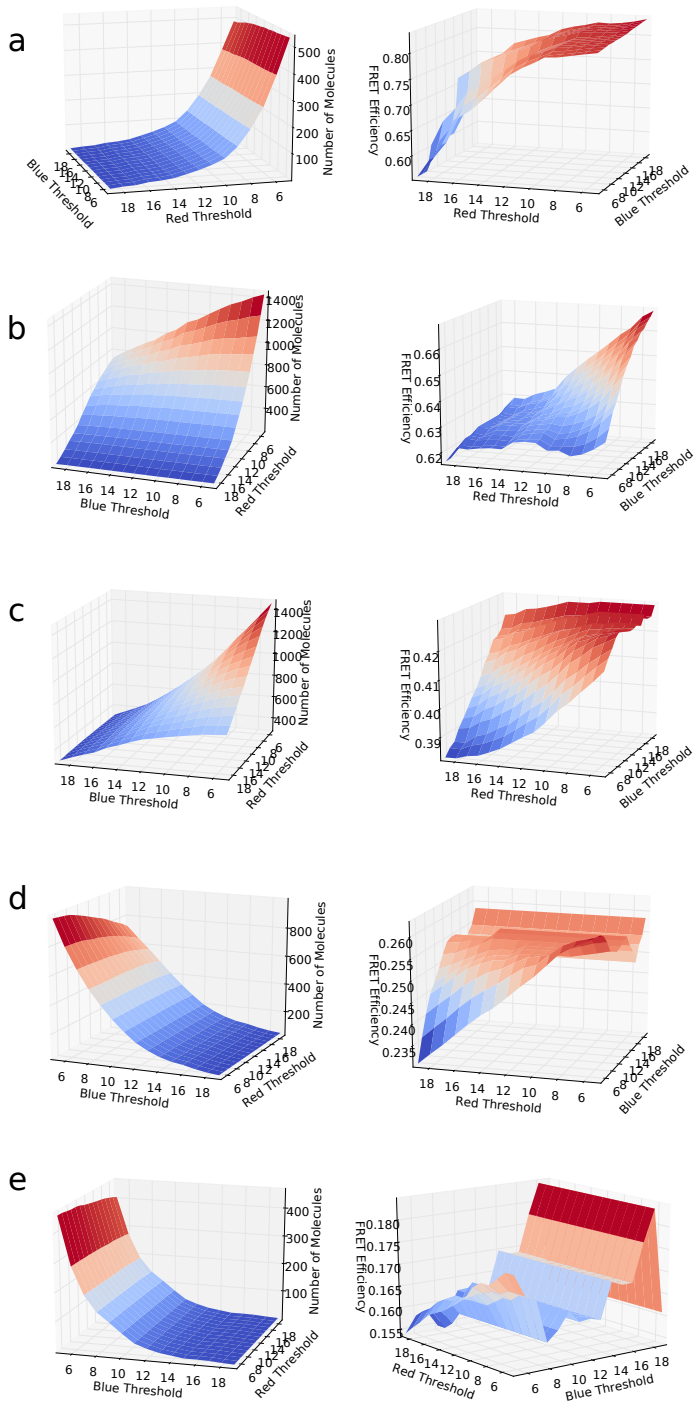

Figure 3.11: Graph illustrating the effect of threshold choice on the number of molecules detected (left) and the calculated FRET efficiency (right) of synthetic datasets with FRET efficiencies of 0.88 (A), 0.66 (B), 0.40 (C), 0.21 (D) and 0.11 (E) respectively. For all FRET efficiencies, the chosen 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.

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.

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 3.4 summarizes the parameters used. We then analysed these datasets using inference and using AND and SUM thresholding. The results, shown in Fig. 3.13, demonstrate that inference is significantly superior to both thresholding analyses. Fig. 3.13 (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. 3.13 (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. 3.13 (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. 3.13 (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

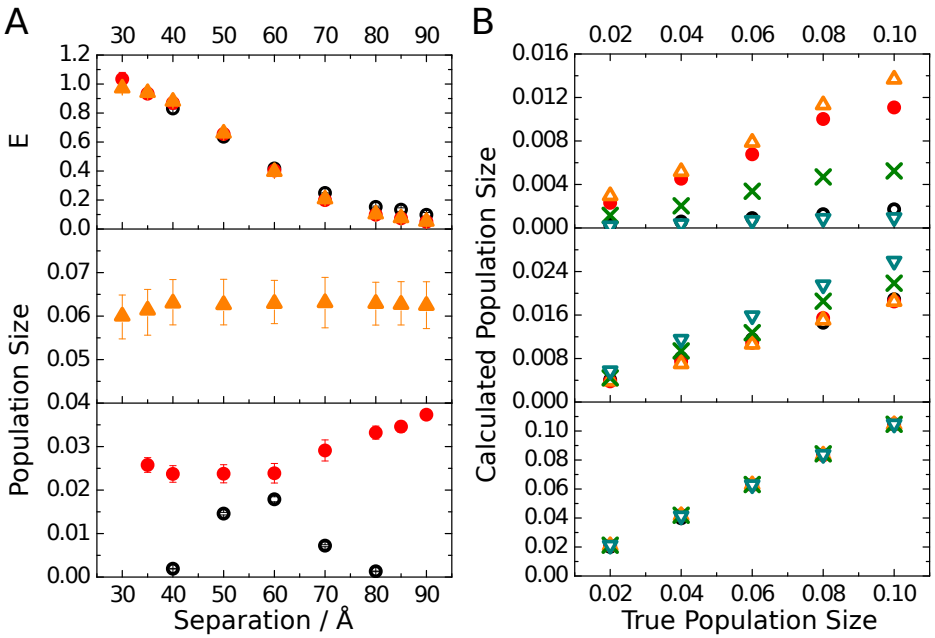

Figure 3.12: Comparison of the inference technique with thresholding-based methodologies. In all plots (A and B), values shown are from 10 datasets generated independently from the same parameters. For thresholding analsyes, error bars represent the standard deviation in the calculated mean from 10 independent 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.

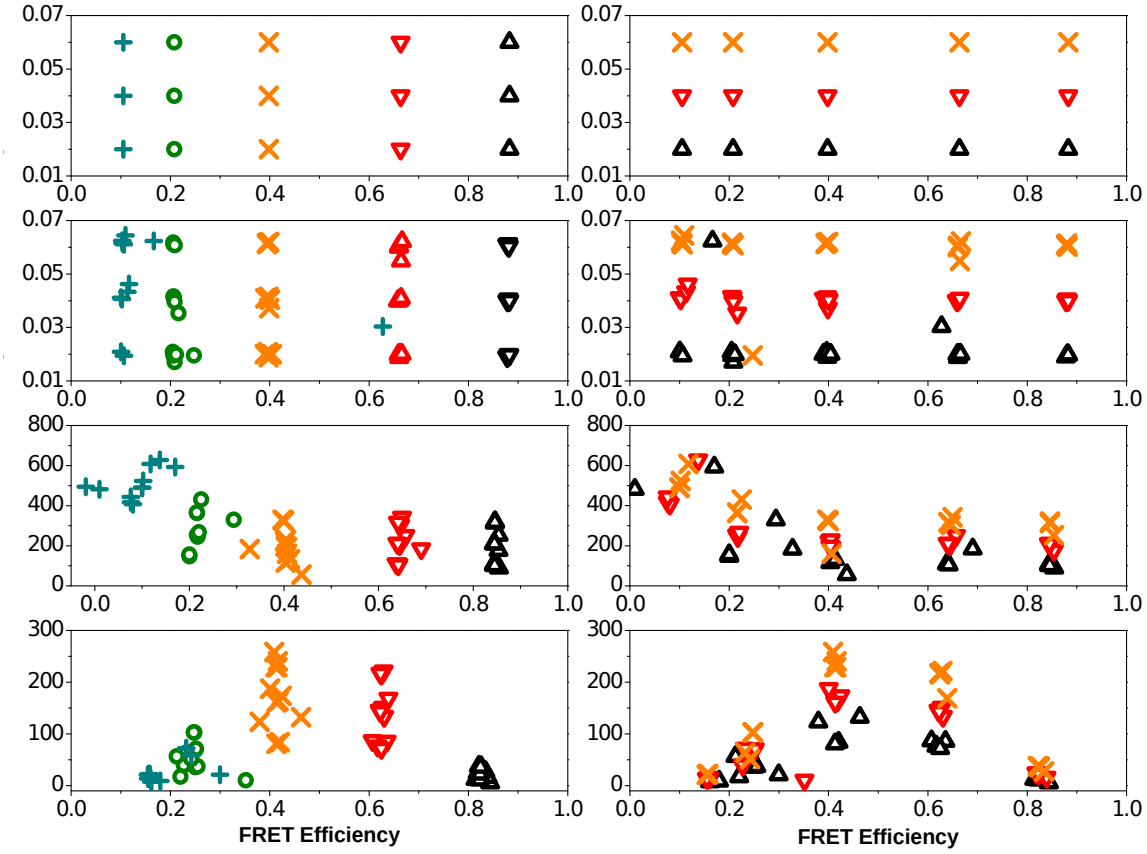

Figure 3.13: Comparison of inference and thresholding analysis on datasets generated to simulate mixtures of fluorescent species. Calculated population size is plotted against calculated FRET efficiency. (A) Idealised situation, in which all FRET efficiencies and population sizes are inferred correctly. (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.

lowest concentration, where it is very difficult to distinguish from noise. In one case, the magnitudes of the two populations ($E = 0.11$, $E = 0.21$, 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.

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			

Table 3.4: FRET efficiencies and population observation rates used in the generation of simulated datasets with two fluorescent populations.

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

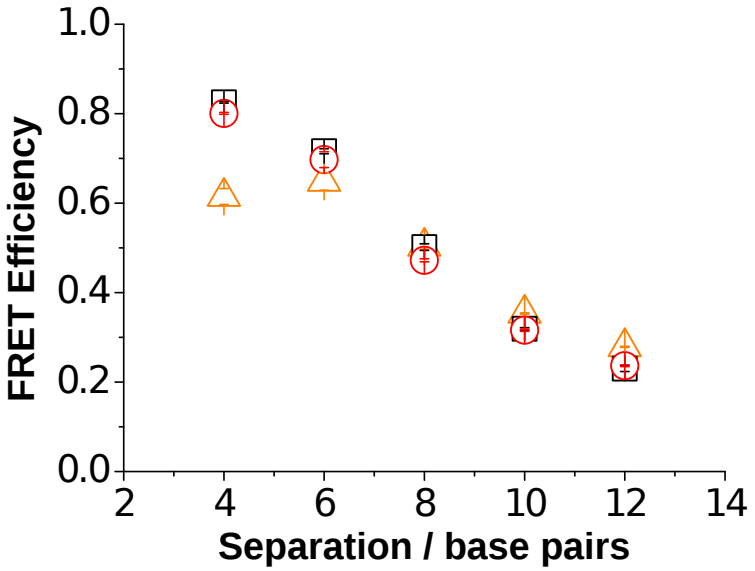

Figure 3.14: Results of AND, SUM and inference analysis of smFRET data from single populations of dual-labelled dsDNA. Orange triangles, black squares and red circles show respectively the results of the AND, SUM and inference based analyses. Error bars show the standard deviation of three independent repeats.

sites increases, the FRET efficiency is expected to decrease in a sigmoidal manner. As Fig. 3.14 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. 3.15).

Here, inference (Fig. 3.15 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. 3.15 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. 3.15 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. 3.15 J) . Furthermore, SUM thresholding also distorts the population sizes, with populations of low FRET species (10 bp and 12 bp dupelexes) being significantly overestimated, 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.

3.5.1 Justification of the Gamma-Poisson Mixture Model

Thus far, we have shown that MCMC sampling to infer the parameters of the model described in Fig. 3.7 is a very effective method to determine key experimental values, conditioned on a smFRET dataset. However, we have not yet given a detailed justification for using a Gamma-Poisson mixture distribution to describe fluorescence emission. As introduced above, it is well-known that photon emission from molecules passing through the confocal volume displays super-poissonian behaviour [102]. Consequently, although the emission from individual fluorophores can be described using a poisson distribution, a single poisson distribution is insufficient to describe the emission behaviour of a population of fluorescent molecules. This is illustrated in Fig. 3.16 A and B: the distribution of observed photon counts when data is simuated using simple poisson emission behaviour has a much smaller variance than an experimental dataset. Consequently, attempting to fit an experimental dataset using a simple poisson emission model results in poor fits and unphysical parameter values (Fig. 3.16 C and D.

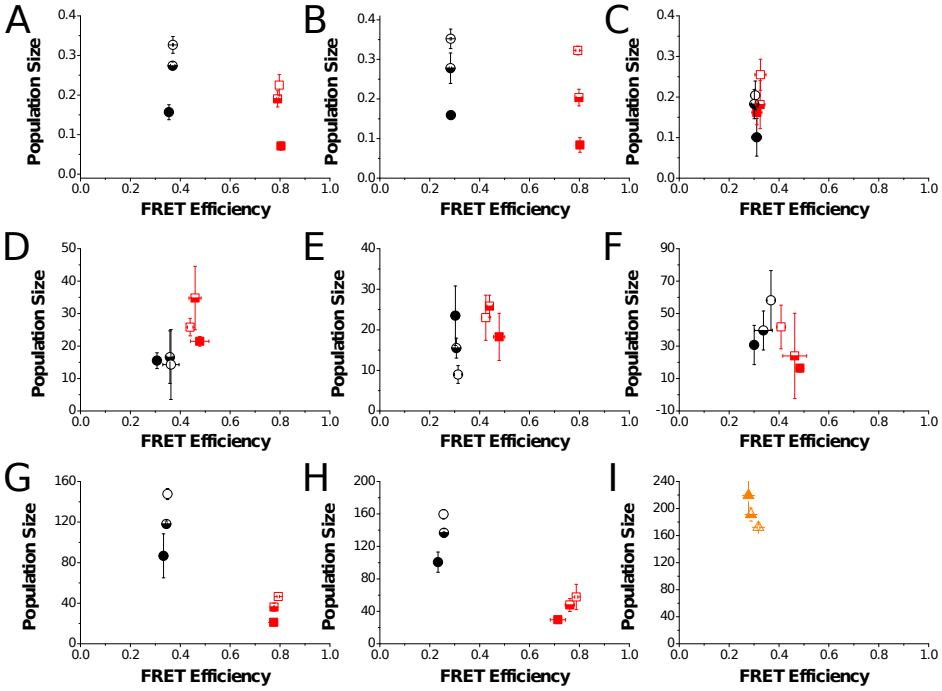

Figure 3.15: Analysis of a mixture of two populations of dual-labelled dsDNA, showing the calculated population sizes and FRET efficiencies. Three different DNA strands were used, with dye attachment sites separated by 4, 10 and 12 bp, corresponding to FRET Efficiencies of 0.79, 0.36 and 0.28 respectively 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.

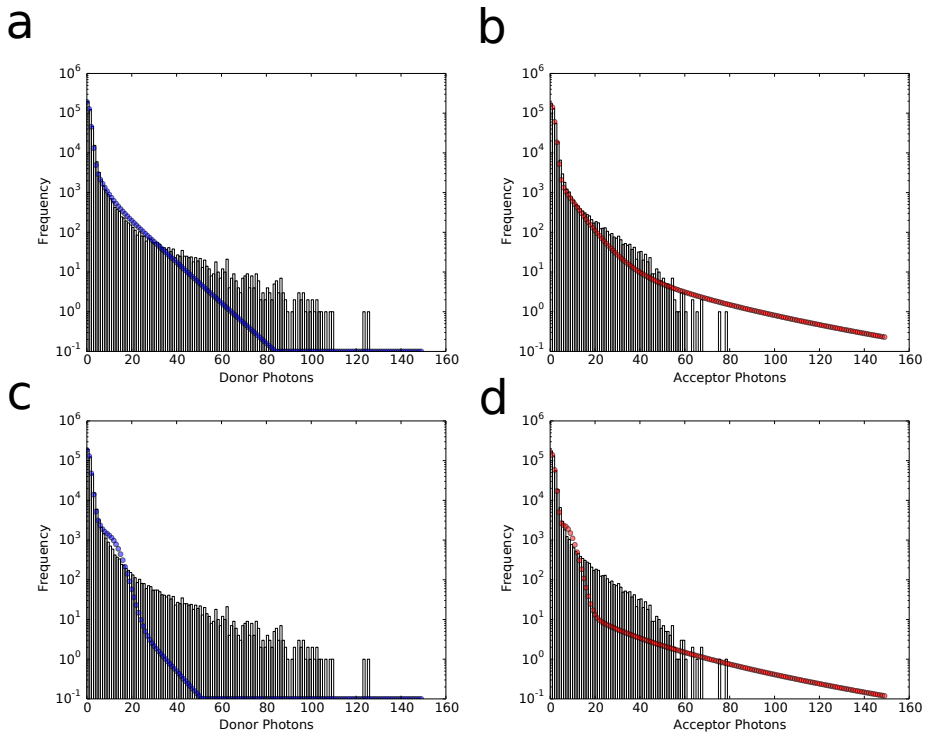

Figure 3.16: Analysis of an experimental dataset with a 6bp dye-separation distance, using a gamma-poisson mixture model (A, B) and a simple poisson model (C, D). When the simple poisson model is used, the shape of the inferred photon emission distribution (blue and red circles) for both donor (C) and acceptor (D) photons is heavily influenced by the shape of the poisson distribution and does not reflect the true photon distribtuion (grey histogram). In contrast, when a gamma-poisson mixture model is used, the shape of the inferred photon emission distribution (blue and red circles) for both donor (C) and acceptor (D) photons is much closer to the true emission distributions (grey histogram).

As described above however, the gamma-poisson mixture model is able to fit experimental datasets much better. The inclusion of overdispersal allows datasets with a much greater variance than can be described using the single poisson emission parameter to be accurately fitted. Moreover, the features (specifically positive, integral photon counts) of a poisson distribution required to accurately describe a smFRET dataset are retained.

3.6 Conclusions and Future Work

Model-based Bayesian inference is a powerful tool that is used in data analysis across many disciplines [81]. However, despite establishment of model-based inference methods to analyze FRET trajectories from immobilised molecules [82, 83, 84, 85, 86], a similar method had not been developed for smFRET data from molecules in solution. This chapter described the development of a model-based inference method, based on the Metropolis algorithm, suitable for the analysis of smFRET datasets. We described how using this inference analysis 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. Using both simulated and experimental data, we show how raw data is analyzed in a single step, requiring neither biased thresholding, nor construction and subjective fitting of FRET histograms.

Extension to ALEX Data However, despite these gains, it is necessary to acknowledge the considerable limitations of our inference tool in its current state. Firstly, as described above, we currently consider only time-binned data from FRET experiments. The additional information that would be available from an ALEX experiment is not currently exploited. One of the main weaknesses of the inference analysis is that we must estimate the labelling efficiencies of the two dyes in a separate experiment. Consequently, one feasible and highly effective extension of our inference analysis would be to incorporate the information that would be gained about labelling state from directly exciting the acceptor dye in an ALEX experiment, direclty into our model of fluorescence emission.

However, despite the apparent simplicity of this extension, incorporating ALEX information into the inference process will lead to a quite significant increase in the time taken for data analysis. This is because our inference analysis takes advantage of the fact that in a smFRET dataset, many pairs of donor and acceptor photons, (f_D, f_A) will be seen many times. Consequently, once the probability $\Pr(\text{params.}|\text{data}_i)$ has been determined for a specific pair of

photon counts $(f_D, f_A)_i$, its contribution to the total probability can be scaled by the number of times that pair is observed, without recomputing its probability, leading to a significant saving in compute-time. When the number of photon streams is doubled, as in an ALEX experiment, the number of different tuples of photon counts, $(f_{DD}, f_{DA}, f_{AD}, f_{AA})$, observed in a dataset of comparable duration, is significantly increased. This means that the computational saving achieved from exploiting duplicated photon counts would be significantly reduced. As a consequence, we have not yet implemented such a model, as the increased compute-time required to analyse even a small dataset limits its attraction for researchers using ALEX data collection.

Extension to Photon Arrival Times A second piece of information that is currently not exploited in the inference analysis described here is the additional information about burst shapes and durations that could be obtained from either recording the precise arrival times of individual photons, or from binning photons over a much shorter timescale. As described in the previous chapter, burst search algorithms [27] make use of this information to exclude fluorescent bursts that are distorted by acceptor photobleaching, removing one source of FRET histogram broadening. It would be possible to modify the forward model to capture rates of photon arrival, instead of per-bin photon counts, enabling this more fine-grained information to be used in the inference analysis.

Extension to Inference of Fluorescent Populations In its current implementation, we can use the inference analysis to infer the parameters of either a single fluorescent population, or of multiple fluorescent populations. What we are not able to do, however, is to infer the number of fluorescent populations. Instead, the number of populations to be inferred must be given as a parameter before analysis begins. In principle, it would be possible to also include the number of fluorescent populations as a parameter to be inferred. Reversible Jump Monte Carlo sampling is a well-established tool for this form of model selection in an inference analysis [105]. Hence, a feasible extension of the inference method would be to include a model selection step that allows inference of the number of fluorescent populations. However, this is currently not implemented, as there has been insufficient interest to prioritize this extension.

Extension to Oligomer Sizing A further application of this method of inference analysis is to the determination of the oligomerisation state of labelled proteins. Such an analysis

requires both an understanding of the emission behaviour of a single fluorophore and how emission scales with the number of fluorophores. We have attempted such an analysis and evaluated its performance using labelled DNA constructs with a known number of attached fluorophores. Implementation and analysis of this model-based approach to oligomer sizing forms the subject of the following chapter.

Conclusion This chapter described a novel method for the analysis of smFRET data. We described a generative model for photon emission in a smFRET experiment, involving one or more fluorescent populations. We then implemented a custom-buit Metropolis sampler to infer the parameters of this model, conditioned on experimental observations. We showed, through analysis of both simulated and experimental data, that inference can correctly determine intramolecular distances and population sizes for one or more fluorescent populations. We have also shown that model-based inference does not suffer from the biases observed in thresholding-based event selection techniques.

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`.

Chapter 4

Bayesian Inference of Oligomer Sizes Using Single Molecule FRET

4.1 Overview

This chapter describes the extension of the Bayesian model based inference described in the previous chapter to to problem of oligomer sizing in the study of protein aggregation. The introductory section opens with a brief overview of the diseases of protein aggregation as motivation for the study of small aggregates. We then describe prior research into protein aggregation using single molecule microscopy, paying particular attention to the assumed relationship between aggregate size and emitted photons. Next, we describe sources of heterogeneity that complicate this relationship and describe acousto-optic modulation as a method to reduce heterogeneity. Finally, we introduce the Holliday Junction [106] as a model oligomer, which can be used to test sizing tools.

This chapter presents results in two main areas. Firstly, we describe the adaptation of our FRET emission model to a simplified model of oligomer emission. We present a number of simple simulations modeling smFRET experiments on oligomers of known size, or in mixures of known size distribution. We show that when the emission rate is assumed to be poisson, the number of emitted photons is a good metric for oligomer size. However, when the emission rate is over-dispersed, the number of photons emitted is no longer dominated by the number of fluorophores present, so is a bad measure of oligomer size. We then use fluorescently labelled DNA Holliday Junctions as model oligomers of known size. Using

these oligomer models, we next show, through comparison of simulations with experimental data from single colour single molecule fluorescence experiments, that the photon emission distribution is indeed over-dispersed.

Secondly, we present an attempt to fit the parameters of our model using both simulated and experimental data. We show that accurate inference of even the mean monomer brightness is difficult, because the relationship between fluorophore emission and confocal occupancy is challenging to decouple. As a consequence, we show that our model of the emission process unfortunately cannot be used to reliably infer the oligomer size distribution, with the result that we cannot accurately determine oligomer size from the number of observed photons.

Finally, we attempt to understand the sources of the emission broadening. We undertake a photobleaching analysis of our synthetic oligomers. From these data, we show that the dye Alexa Fluor 488 undergoes reversible photoblinking on a time-scale comparable with the dwell time in the confocal volume, identifying an important additional source of emission heterogeneity. We also discuss the effect of unequal excitation power on photon emission and suggest experimental modifications that could mitigate this problem.

4.2 Introduction

4.2.1 Diseases of Protein Aggregation

Several human diseases, such as Alzheimer's Disease [107] and Parkinson's Disease [108] are diseases of protein aggregation. The neurodegenerative Alzheimer's Disease is characterised by loss of neuronal cells and the presence of large protein aggregates both in the extra-cellular space and within cells. The extra-cellular deposits are primarily composed of polymers of the peptide Amyloid Beta (hereafter Aβ) [109], whilst the intracellular aggregates, named neurofibrillary tangles, are dominated by the protein Tau [110]. Similarly, at the cellular level, Parkinson's Disease is charaterised by the presence of Lewy Bodies [111], solid depositions of protein, predominantly the intrinsically disordered protein α-synuclein [108], inside nerve cells.

In both Alzheimer's Disease and Parkinson's Disease, the protein aggregates, when examined at the microscopic level, are found to have a similar structure: long, polymeric chains of proteins form fibrils with a regular structure comprised of inter-molecular (cross) beta

sheets [112]. These amyloid fibres are also found in patients suffering from other forms of neurodegenerative disease, such as Transmissible Spongiform Encephalopathy [113] and Fatal Familial Insomnia [114], as well as in other human diseases that do not include neurodegerative symptoms [115] and some animal diseases [116].

The role that amyloidosis plays in neurodegenerative disease is currently not well understood. Heritable mutations, such as the A30P and A53T mutations found in α-synuclein, which predispose the peptide to aggregate formation in vitro [117] are associated with early onset disease. However, it is curently unknown whether there is a common mechanism of cytotoxicity associated with the structurally similar oligomeric species found in various diseases of protein aggregation, [118]; whether the accumulation of amyloid aggregate overwhelms the cellular degradation and molecular chaperone pathways [119], or whether the mechanisms of toxicity are complex and disease specific [120].

4.2.2 Studying Protein Aggregation

As a consequence, understanding the chemical pathway of protein aggregation, particularly of disease-associated peptides such as Aβ and α-synuclein, is of considerable biological interest. Current understanding is that aggregation follows a "nucleated growth" mechanism [121]. Under the nucleated growth mechanism, there is an initial lag phase, during which a few protein molecules convert from their native structure to an aggregation nucleus. Following their formation, fibril growth occurs rapidly via polymerisation from these molecular seeds (see Fig. 4.1). Analytical solution of the kinetic equations of aggregation suggests that, under certain conditions, secondary nucleation via templating [122], as well as fibril fragmentation [123], dominates the rate of aggregation. This theoretical result has been used to explain the differing experimentally observed aggregation rates of the two amyloidogenic peptides Aβ-40 and Aβ-42 [124, 125].

To fully understand the protein aggregation pathway and the nature of the biochemical species involved, it is necessary to study both the kinetics of the aggregation reaction and the biophysical properties of the different species on the aggregation pathway. Owing to its relative abundance, the native structure of the monomeric species can be characterised relatively effectively using, for example, NMR solution studies [127]. Similarly, excellent insight into the morphology of mature fibrils can be obtained through experimental techniques such as TEM microscopy [128] and SAXS [129]. Furthermore, the chemical Thioflavin T (ThT

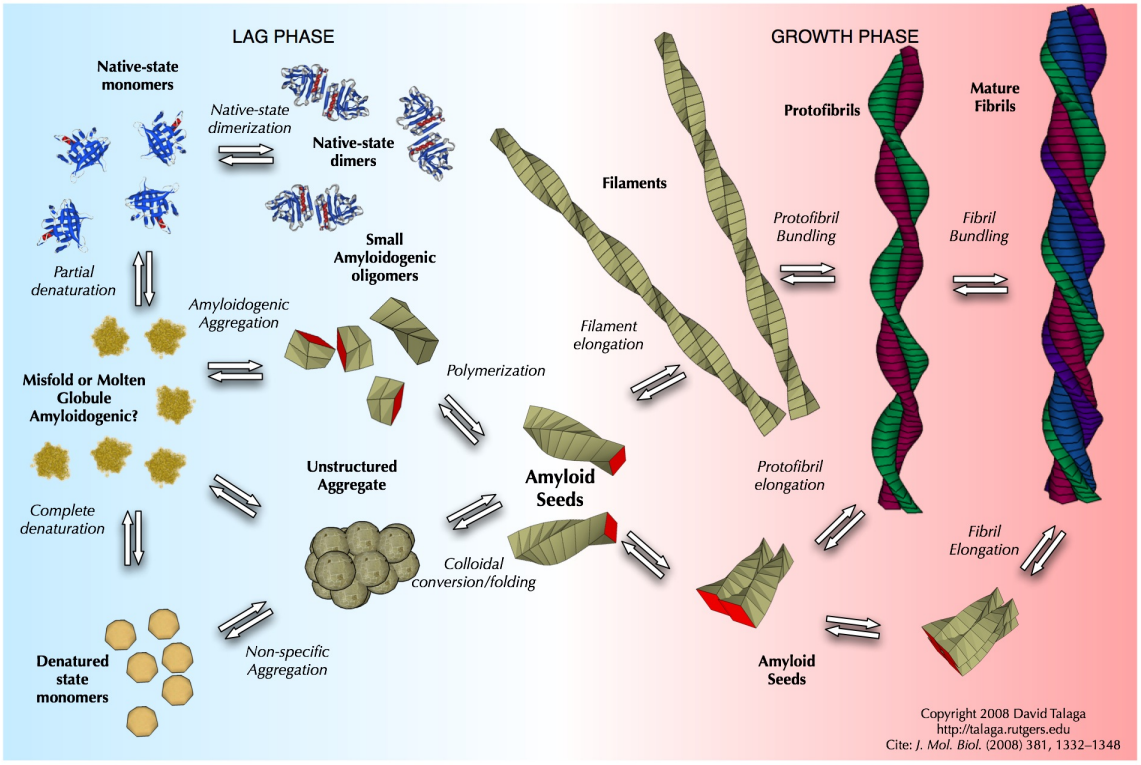

Figure 4.1: Schematic showing pathways of amyloid aggregation. Figure from [126] and used with permission.

hereafter)(Fig. 4.2 A), which is weakly fluorescent when free in solution, becomes several orders of magnitude more fluorescent on binding to the beta sheet structure of an amyloid fibril, allowing it to be used to study the kinetics of fibril formation ([130], Fig. 4.2 B).

Using smFRET to Identify Oligomers Unfortunately, the all important oligomeric species involved in nucleation remain relatively elusive to scientific study. Although the presence of oligomers can be detected through ELISA assay [131] and synthetic oligomers can be prepared in vitro [132, 133], direct charaterisation of oligomeric species is hindered by the difficulty of accurately detecting oligomers against a background of monomeric and fibrillar structures. One promising method of oligomer observation is through the use of single molecule fluorescence studies.

Previous work using two colour single-molecule microscopy has been able to both track the kinetics of aggregation of α-synuclein and to identify different oligomeric states [134, 32]. This research used α-synuclein monomers labelled with one of the fluorescent dyes Alexa Fluor 488 and Alexa Fluor 647. Mixtures of monomers labelled with the dyes were incubated under fibrillizing conditions and then subjected to single molecule analysis at time points distributed throughout the duration of the aggregation reaction. AND-based thresholding was used to identify oligomeric species against an overwhelming background of monomers, as only oligomers carrying at least one of each dye type would be able to display photon emission in both donor and acceptor channels. These experiments identified two fluorescent species, displaying different FRET efficiencies. Moreover, the high-FRET species emerged later in the aggregation reaction, was associated with oligomers of larger size and was further demonstrated to show increased resistance to Proteinase-K degradation, suggesting an interconversion between two oligomeric species as the aggregation progressed.

4.2.3 The Relationship Between Size and Photon Emission is Complex

This research provides a fascinating insight into the biophysical changes that occur during amyloid aggregation. However, one limitation is that the method used for determining the size of an observed oligomer is essentially a simple heuristic. Oligomer sizes were determined by comparison with the brightness of a monomer event, correcting for the increased dwell-time of larger molecules in the confocal volume, but otherwise assuming a linear relationship

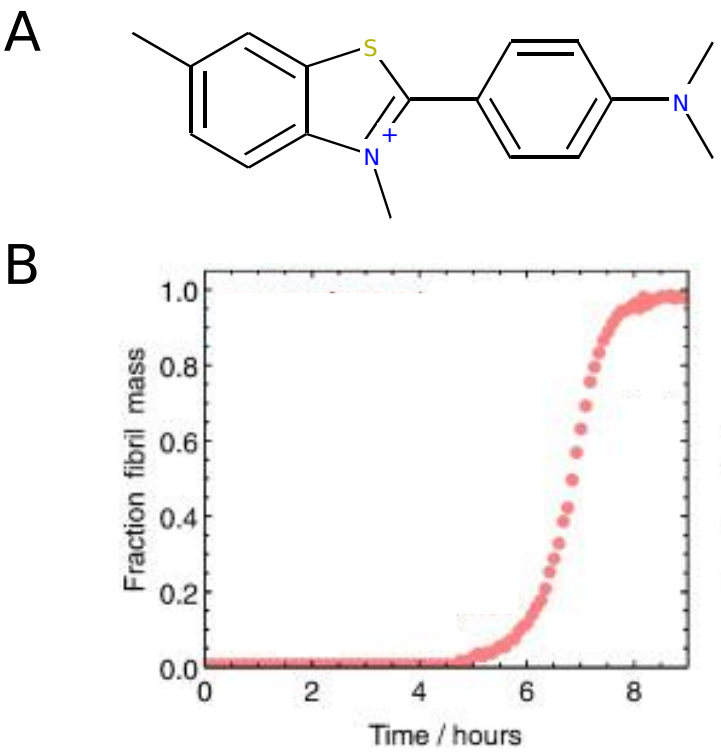

Figure 4.2: ThT fluorescence is used to follow the kinetics of amyloid formation. A) The molecular structure of ThT. When free in solution, rotation around the central bond between the benzylamine and benzathiol rings within the molecule enables twisted intramolecular charge transfer in the excited state, and a low quantum yield. On binding to extended beta-sheet structures, the rotation is prevented and the fluorescence quantum yield increases significantly. B) Using ThT fluorescence to track the kinetics of amyloid formation reveals a characteristic sigmoidal curve, displaying an initial lag phase, followed by a rapid growth phase that plateaus once the aggregation reaction reaches equilibrium. B) is adapted from [130].

between oligomer size and photon emission rate [134]:

$$\text{size} = \frac{2 \cdot I_D + \gamma^{-1} I_A}{I_{\text{monomer}}} \tag{4.1}$$

for I_D and I_A the intensities (number of photons) in the donor and acceptor channels respectively, γ the instrumental gamma factor and I_{monomer} monomer event brightness, determined from fluorescent bursts that consisted of photons of a single colour only by selecting events for which $I_D \geq T_D$ but $I_A < T_A$.

However, as described in the previous chapter (Section 3.5.1), not all fluorescently labelled molecules passing through the confocal volume result in emission of the same number of photons. Multiple factors, including the confocal dwell time, the pathway taken through the gaussian laser beam and the effect of photobleaching during excitation result in broadening of the photon emission distribution becoming super-poissonian. The emision behaviour is well-approximated by a gamma-Poisson mixture model [41], as we use in our inference-based analysis tool [70]. Single molecule fluorescence studies of oligomeric species are subject to similar behaviours, meaning that the relationship between oligomer size and burst brightness is complex and non-linear.

4.2.4 The DNA Holliday Junction as a Model Oligomer

To understand the relationship between the number of fluorescent labels and the number of emitted photons and hence to evaluate the accuracy of oligomer sizing using single molecule fluorescence, it was necessary to have access to synthetic oligomers carrying a known number of fluorescent labels. For this model oligomer, we used the DNA Holliday Junction [106]. The Holliday Junction is a four-way structure that is observed biologically during DNA recombination [135]. Synthetic Holliday junctions can be prepared using four partially complementary strands of DNA (Fig. 4.3) and have frequently been used to study junction dynamics using single molecule fluorescence [136, 137, 138]. These synthetic structures provide an ideal model for small oligomers in fluorescence studies, as each strand can be independently labelled with a fluorescent dye prior to construction of the junction, enabling precise control of the number of fluorescent dyes present (Fig. 4.4). We prepared synthetic Holliday junctions labelled with between one and four fluorescent dyes as described below. The sequences of the four arms used are shown in Table 4.1 These synthetic species were then used in single

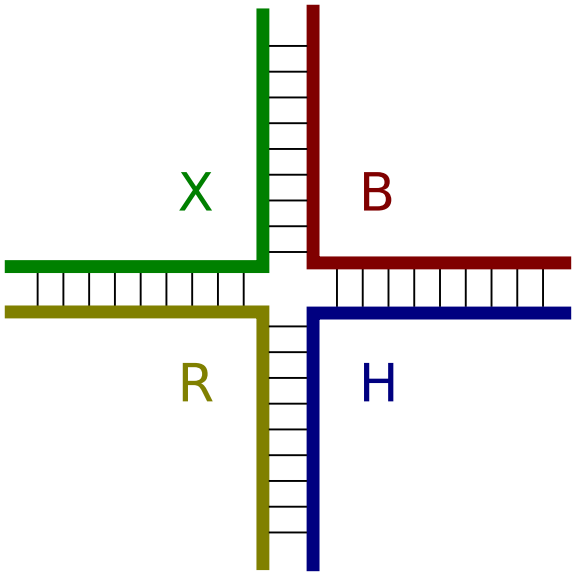

Figure 4.3: Schematic of a synthetic Holliday junction. The four arms, named B, H, R and X are shown in red, blue, yellow and green respectively.

molecule fluorescence studies to probe the relationship between oligomer size and photon emisson.

4.3 Theory

This section describes the extension of our generative model of the smFRET experiment to the case of fluorescently labelled oligomers, carrying multiple fluorescent dyes. Here, we describe two different models of photon emission from a labelled oligomer. First, we describe a simple poisson model in which sources of distribution broadening are not considered. We then extend this model to a gamma-poisson mixture model, which includes overdispersion. We show, through comparison of simulations and experimental data that the gamma-poisson mixture is a more appropriate model for oligomer fluorescence.

For simplicity, we consider only molecules labelled with a single dye type (experimentally, either Alexa Fluor 488 or Alexa Fluor 647), allowing FRET effects to be ignored. Several other assumptions are also made, namely that the relative concentration of oligomeric species is modelled by a geometric distribution and that all monomers carry a fluorescent dye (a labellling efficiency of 100%.) Further, we consider only events binned on the order of the dwell time in the confocal volume and do not explicitly consider either the effect of

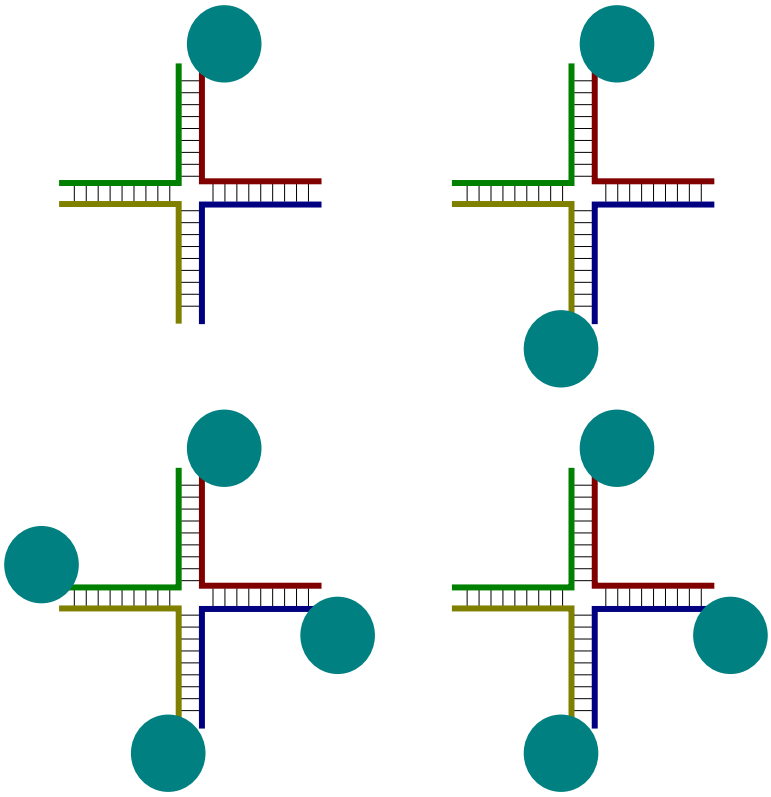

Figure 4.4: Schematic showing the four labelling states of our synthetic Holliday Junction construct. Shown clockwise from top left: The monomer model has only Arm B labelled (top left); the dimer has both arms B and X labelled (top right); the trimer model has labels on arms B, H and R (bottom right); in the tetramer model all four arms are labelled (bottom left).

photobleaching partway through a bin or changes in the dwell time caused by slower diffusion of larger molecules. Details of the model, and the roles of these assumptions are given below.

4.3.1 A Simple Poisson Model of Oligomer Photon Emission

In the simplest fluorescence experiment from an oligomeric species, the fluorescently-labelled oligomers diffuse freely through the confocla detection volume. When a molecule diffuses into the confocal volume, the laser excites the attached fluorophore(s) and photons are emitted (Fig. 4.5. Typically, both donor and acceptor fluorophores would be used to allowing identification of oligomeric species based on the presence of coincident donor and acceptor bursts. However, we consider a simplified experiment in which monomers are labelled with one dye type only. This enables us to ignore the effect of FRET on the number of photons observed and to consider only the total number of photons emitted in a burst. We also assume full labelling of monomers (100 % labelling efficiency), so that the number of fluorophores present is equivalent to the oligomer size (in monomers). Under these conditions, we expect intuitively that the number of emitted photons increases with the number of fluorophores present. As for dual-labelled monomers, we consider photons binned into time-bins of length on the order of the dwell time in the confocal volume; we expect the great majority of time-bins to contain noise photons only.

We model this simple experiment as a sequence of measurements (f_D) of the number of photons observed in the fluorescence emission channel. Each time-bin is treated as an independent and identically distributed sample from a set of random variables describing the dataset. Each data point (f_D) in the data stream is the sum of noise photons and, where an oligomer is present, some photons from a fluorescent event.

The number of noise photons is drawn from a Poisson distribution with rate parameter λ_D. The probability of observing n_D noise photons in the donor channel is then:

$$n_D \sim \text{Poisson}(n_D; \lambda_D) = \frac{\lambda_D^{n_D}}{n_D!} e^{-\lambda_D} \tag{4.2}$$

In addition to noise, each observation may contain photons from one or more fluorescently labelled oligomers. As before, the number of molecules present in the excitation volume, n_{prot} can be described using a Poisson distribution with rate parameter λ_{prot}:

$$n_{\text{prot}} \sim \text{Poisson}(n_{\text{prot}}; \lambda_{\text{prot}}) = \frac{\lambda_{\text{prot}}^{n_{\text{prot}}}}{n_{\text{prot}}!} e^{-\lambda_{\text{prot}}} \tag{4.3}$$

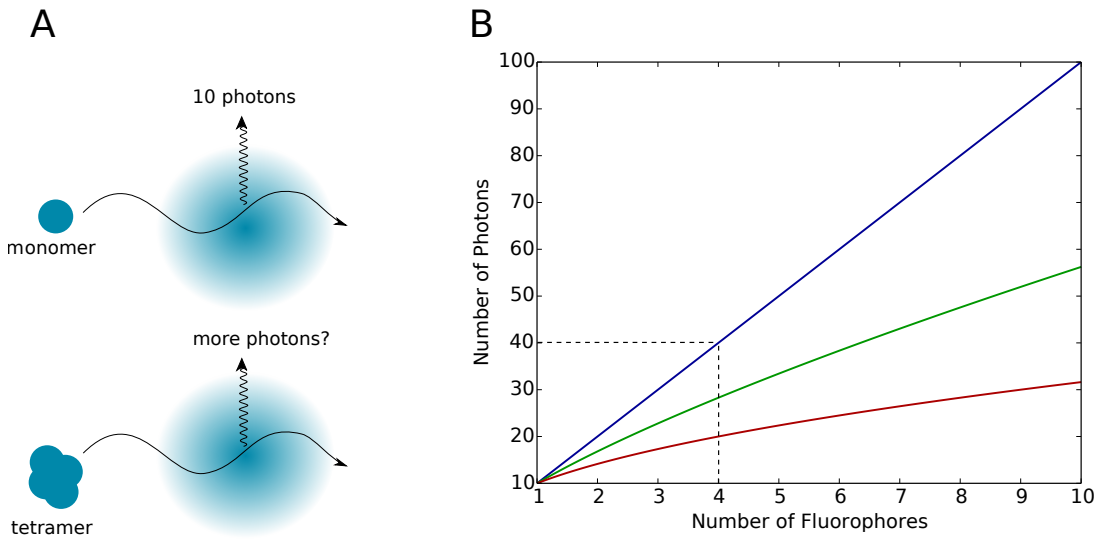

Figure 4.5: Schematic showing the problem of determining oligomer size from the number of observed photons. A) If we observe an average of 10 photons when a monomer diffuses through the confocal volume, how many more should we expect if a tetramer takes the same pathway through the confocal volume? B) Possible relationships between the number of fluorophores and the average number of photons observed. If there is a linear relationship (blue line) then we expect a tetramer to emit an average of 40 photons; if the relationship is not linear, we may expect fewer photons (red and green lines).

However, now we must also consider the oligomer size. In the absence of evidence to the contrary, we assume that fluorophores on an oligomer are non-interacting, such that each fluorophore is excited independently and does not experience quenching effects from the proximity of other fluorophores. We begin with a simple, purely poisson model of oligomer fluorescence. We model each fluorophores as displaying poisson emission with rate parameter λ. As the sum of poisson random variables is itself a poisson random variable [139], an oligomer of size n can be described as having an emission rate of $n \cdot \lambda$. Therefore, the dditional photons present in a time-bin contributued by an oligomer of size n is given by:

$$c_D \sim \text{Poisson}(c_D; n \cdot \lambda) \tag{4.4}$$

The total number observed photons in that time bin, f_D, is then the sum $n_D + c_D$.

Given this simple model, we can easily generate multiple datasets with different properties. Below, we describe generation of simulated datasets from oligomers of known size, or from mixtures of oligomer sizes of known size distributuion. However, first we introduce the more complex gamma-poisson mixture model that is a direct extension of the model described in the previous chapter.

4.3.2 A Gamma-Poisson Mixture Model of Oligomer Photon Emission

The previous section described a simple poisson model of photon emission from labelled oligomers. However, as discussed in the previous chapter, we have found a gamma-poisson (negative binomial) distribution to be a better model for the overdispersion seen in emission from dual-labelled monomers [70]. This section describes the extension of this model of fluorescence emission to the case of labelled protein oligomers.

The system modelled is equivalent to that described in the previous section. The noise distribution and probability of confocal volume occupancy are unchanged. However, now we choose to include over-dipsersion in our model of fluorophore excitation, by drawing event specific emission parameters from a gamma distribution, moving from a single-step model of emission to a two stage emission process.

Firstly, given that a fluorescently labelled molecule is present in the confocal volume, we determine a rate of donor emission, λ for monomers in that specific molecule as a random

sample from a gamma distribution with shape parameter k_D and mean λ_B (Eq. 4.5). 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 individual fluorophores photobleaching partway through an observation.

$$\lambda \sim \text{Gamma}(\lambda; k_D, \theta) = \frac{1}{\Gamma(k_D)\theta^{k_D}} \lambda^{(k_D-1)} e^{-\frac{\lambda}{\theta}} \quad \text{for } \theta = \lambda_B/k_D \tag{4.5}$$

Here, Γ is the Gamma function.

In the second step, we consider the oligomer size. As above, we consider labelleing efficiency to be 100 %, such that an oligomer containing n monomers is assumed to have n attached fluorescent dyes. We assume these dyes to be non-interacting. Furthermore, although we are now considering the impact of the pathway through the confocal volume on the number of photons emitted by a fluorophore, the size of an oligomer is still small relative to the size of the confocal volume. Hence, we assume that all fluorophores on a specific oligomer have the same rate of donor emission, λ, determined according to equation 4.5. Consequently, the number of photons emitted by an oligomer of size n with emission rate λ is given, as before by

$$c_D \sim \text{Poisson}(c_D; n \cdot \lambda) \tag{4.6}$$

and the total number of observed photons, f_D is then the sum $n_D + c_D$. Note, however that whereas in the simple poisson model λ was a constant, now λ itself is a random variable, which is drawn afresh from equation 4.5 for each oligomer emission event.

The rest of this chapter describes the comparison of simulated datasets generated using these two models with real data from model oligomers of known size and from true protein aggregation experiments with unknown oligomer size distributions. In the next section, we summarize the experimental techniques used to collect single molecule fluorescence data from aggregation timecourses; the preparation of labelled Holliday Junctions as model oligomers; and the modification of the confocal excitation set-up to modify the confocal excitation field.

4.4 Experimental Methods

4.4.1 Preparation of DNA Holliday Junctions

Holliday Junction Preparation *Holliday Junction purification was carried out by Vladimiras Oleinikovas under the supervision of the author.* Four oligonucleotides labeled at the 5-prime end by the fluorescent dye Alexa Fluor 488 and with the correct sequence to anneal to form a Holliday Junction were purchased (atdbio), along wih equivalent unlabelled strands (Sigma Life Science). Their sequences are given in Table 4.1. The four strands required to generate a Holliday junction were combined in Tris buffer (10 mM, pH 8.0, 50 mM NaCl), to give a final total DNA concentration of 5 μM in a total volume of 20 μL, using an equimolar concentration of the four DNA strands. The single-stranded DNAs were annealed into the Holliday Junction by heating to 95∘C for 10 minutes, followed by slow cooling overnight to 25∘C.

Table 4.1: DNA sequences of the four arms of the Holliday Junction. Each sequence is shown in the 3' - 5' direction/ **5** is a maleimide linkage to Alexa Fluor 488, present in the fluorescently labelled DNA strands, but absent in their unlabelled partners.

Arm	**Sequence**
B	CCCTAGCAAGCCGCTGCTACGG**5**
H	CCGTAGCAGCGAGAGCGGTGGG**5**
R	CCCACCGCTCTTCTCAACTGGG**5**
X	CCCAGTTGAGAGCTTGCTAGGG**5**

Holliday Junction Purification *Holliday Junction purification was carried out by Vladimiras Oleinikovas under the supervision of the author.* After annealing, the Holliday Junctionss were purified from the reaction mixture using gel electrophoresis on an 8% TBE acrylamide gel in TBE buffer (Novex, Life Technologies Ltd.). The complete Holliday Junction was identified by visualising the bands using low intensity (PMT 300V) excitation at 526 nm and comparison with a 10 bp DNA ladder. The band corresponding to the complete Holliday Junction was cut from the gel and eluted into 200 μL of Tris buffer by passive difusion overnight. Holliday Junctions were then stored in the dark at 4∘C until required.

4.4.2 Simple FRET Measurements of DNA Holliday Junctions

Labelled Holliday Junctions were diluted to a concentration of 50 pM in TEN buffer (10 mM Tris, 1mM EDTA, 100 mM Nacl), pH 8.0, containing 0.01 % Tween-20. FRET data were collected for 15 minutes using continuous excitation at 488 nm at a power of 80 mW. Collected photons were binned online in intervals of 1 ms and stored in files of 10000 bins.

The microscope used for experiments into the effet of unequal excitation was similar to that described in previous chapters. However, a number of modifications were used to allow alteration of the detection volume. Firslty, a voltage controlled oscillator (DRFA10Y-B-0, PhotonLines) was coupled via an RF amplifier (AMPA-B-30, PhotonLines) to an acousto-optic modulator (MT80-A1-VIS, AA Optolectronincs). The modulator was placed in the beam path of the 488 nm laser () and a triangular wave, provided using a signal generator (), was used to deflect the beam in the transverse direction at a speed of 100 kHz. Modifying the amplitude of the wave allowed the degree of deflection to be precisely controlled. This enabled us to quantify the effect of unequal excitation power on the photon emission distribution. The AO modulator deflects only the first harmonic of the 488 nm laser, which is spatially separated from the zeroth and other order beams. Consequently, we direct only this beam into the back port of an inverted microscope (Nikon Eclipse Ti-U). This harmonic corresponds to approximately 10 % of the total power output of the laser.

4.4.3 Counting Photobleaching Steps Using TIRF Imaging

These experiments were performed with the assistancne of Kristina A. Ganzinger. Data analysis was completed by Kristina A. Ganziner. Borosilicate glass coverslips (VWR international, 20 by 20 mm, 63 1-0122) were cleaned using an argon plasma cleaner (PDC-002, Harrick Plasma) for 1 hour to remove any fluorescent residues. The slides were demarcated using slide stickers (XXX?) and a 200 μL of a 50 pM solution of the labelled Holliday Junction was introduced to the demarcated well. Photobleaching analysis was carried out using a 48 nm laser at XXX μW. Data were collected using a XXX camera (XXX brand) using a frame length of XXX ms. Data collection proceeded until the majority of dyes had been bleached (XXX s for monomers and dimers; XXX s for trimers and tetramers). Three separate slides were prepared for each Holliday Junction and XXXnineXXX? locations imaged on each slide. Following data collection, photobleaching trajectories were extracted and analysed using custom software cite???.

4.5 Results

The major results of this chapter are now presented. Firstly, we show that the relationship between oligomer size and rate of photon emission shows considerable over-dispersion. We demonstrate this through comparison of simulated and real datasets from oligomers of known size. We then present our attempts to reduce excitation heterogeneity by acousto-optic modulation of the laser beam position. Finally, we present the results of a photobleaching step analsyis of the labelled Holliday junctions, which reveals that photoblinking is a major source of emision heterogeneity.

4.5.1 The need for a Generative Model

To open the results section of this chapter, we present a justification for taking a model-based approach to studying aggregation. It is clear that there are many unknowns and many simplifications in our models. Here, we present some simple simulations to demonstrate the importance of understanding the processes underlying fluorescent emission. Although it is true that our models contain simplifications, it is also true that any method of analsing experimental data contains simplifications and assumptions about the structure of those data. In the case of oligomer sizing experiments, these assumptions are often not made explicit, but are based on fluorophores demontsrating poissonian emission behaviour. Here, we compare the performance of this data analysis methodology, using datasets simulated using both the poisson model of emission behaviour, and the gamma-poisson mixture model. We show that the assumptions made in the data analysis workflow do not hold when the photon emission distribution is overdispersed, leading to calculated oligomer size distributions that are both quantitatively and qualitatively incorrect.

The largest assumption made in prior work to determine oligomer sizes using single molecule fluorescence is that there is a linear relationship between oligomer size and photon emission. In prior work [32], oligomers were identified by using the AND criterion on time-binnned TCCD data (see Section 4.2.2 for details) to select time-bins that exceeded some thresholds T_D and T_A in the donor and acceptor channels respectively. Following event selection, the approximate size of each identified oligomer was calculated using the relationship:

$$\text{size} = \frac{2 \cdot I_D + \gamma^{-1} I_A}{I_{\text{monomer}}} \tag{4.7}$$

where I_D and I_A are the intensities (number of photons) in the donor and acceptor channels respectively, γ is the instrumental gamma factor and I_{monomer} is the mean brighness of monomer events (selected by identifying events for which $I_D \geq T_D$ but $I_A < T_A$). This means that the calculated oligomer size is simply the oserved event brightness divided by the mean event brightness for monomer events.

Fig. 4.6 shows the expected emission distributions from small oligomers of various sizes, when emission behaviour is purely poisson. This simple simulation ignores contributions from noise and assumes a mean monomer emission of 10 photons. The code used to generate these data can be found in Appendix XXX. As seen from this figure, when the emission behaviour is purely poisson, there are clear, distinct peaks in photon emission frequency for each oligomer size. As a consequence, using Eq. 4.7 to estimate the oligomer size gives approximately accurate results. We simulated oligomer emission events, sampling a uniform distribution of sizes with a monomer monomer brightness of $I_{\text{monomer}} = 10$. As we consider only single-colour excitation and emission, event selection was performed using $T_D = 10$ only and the identified fluorescent events were sized according to a simplified version of Eq. 4.7:

$$\text{size} = \frac{I_D}{I_{\text{monomer}}} \tag{4.8}$$

The results, shown in Fig. 4.7 show that, apart from at the very ends of the size distribution, the calculated size distribution (blue curve) is an excellent approximation of the true size distribution (grey histogram).

However, when oligomers are simulated using the gamma-poisson mixture model, the assumption of linearity is no longer valid. As Fig. 4.8 shows, overdipsersal results in much broader distributions for which clear peaks corresponding to different oligomer sizes can no longer be discerned. As a result, estimating the oligomer size using Eq. 4.7 gives wildly inaccurate results (Fig. 4.9).

This simple analysis of simulated data illustrates the importance of understanding the models that underlie a data analysis technique. If oligomers are pure poisson emitters, then Eq. 4.7 is an appropriate method to approximate the oligomer size distribution. However, if the effects of excitation heterogeneity and photobleaching cause significant over-dispersal of the emission distribution, Eq. 4.7 produces an estimation of the oligomer size distribution that is extremely inaccurate. In the following section, we compare photon emission histograms from model oligomers of known size with simulated datasets, showing the poisson distribution is

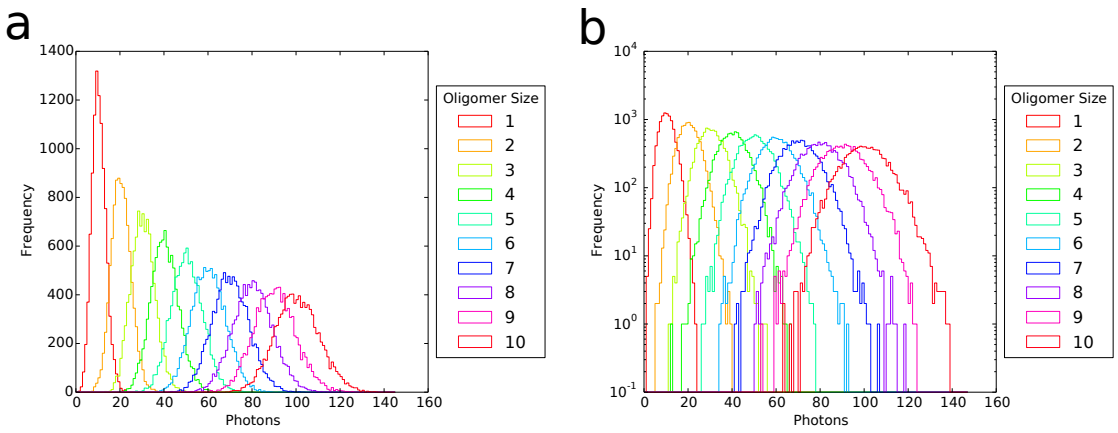

Figure 4.6: Simulating oligomer photon emission using the simple poisson model. a) Simulated emission distributions from oligomers of size 1 (monomer) to 10, for simple poisson emission. b) The same distributions as in a), plotted using a logarithmic scale.

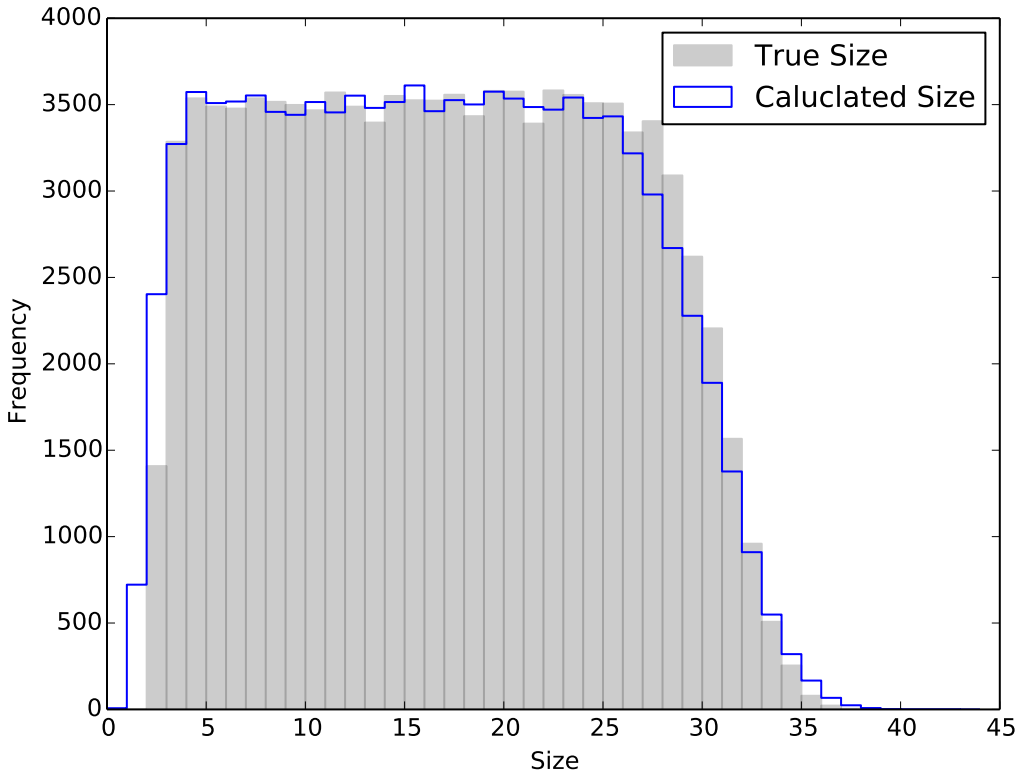

Figure 4.7: True and calculated oligomer sizes for data simulated using poisson emission behaviour. Under these conditions, the calculated oligomer size distribution (blue line) closely approximates the true size distribution (grey histogram).

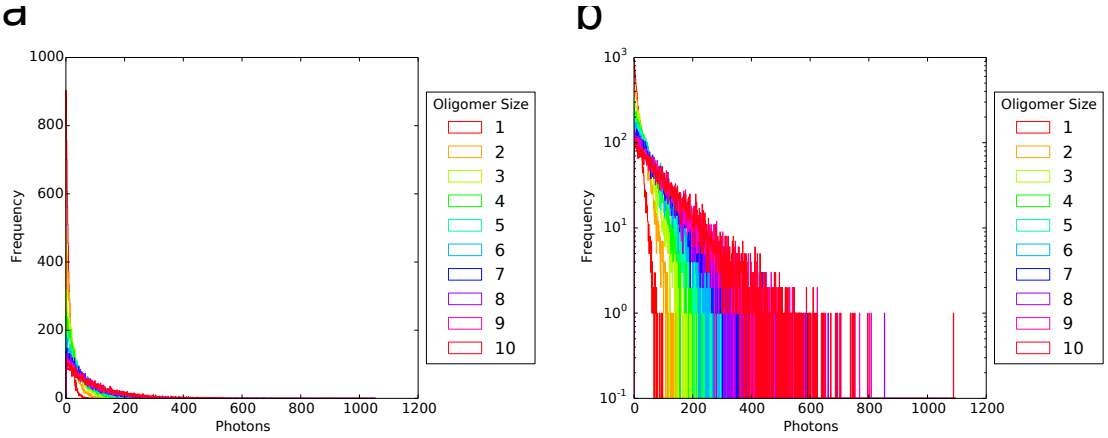

Figure 4.8: Simulating oligomer photon emission using the gamma-poisson mixure model. a) Simulated emission distributions from oligomers of size 1 (monomer) to 10, for gamma-poisson emission. b) The same distributions as in a), plotted using a logarithmic scale to show emission detail.

a poor model of the observed photon emission distributions.

4.5.2 Understanding the Relationship Between Size and Photon Emission

To understand the relationship between oligomer size and photon emission in a single molecule fluorescence experiment, it is necessary to have experimental datasets for which the size, or size distribution, of oligomers present is known. In a protein aggregation experiment, the aggregation reaction generates an heterogeneous solution of oligomers, but their size distribution is unknown, making them a poor model for understanding photon emission behaviour.

To overcome this problem, we use a labelled DNA Holliday Junction as a model oligomer. We prepared Holliday Junctions labelled with between one and four fluorophores (Alexa Fluor 488), enabling us to collect data from solutions of pure monomer, dimer, trimer and tetramer models. We compare experimental data collected using these oligomer models from datasets simulated using either the pure poisson or the gamma-poisson emission models. The results are very informative.

Fig. 4.10 displays histograms of the full photon frequency distribution for Holliday Junctions labelled with between one and four fluorophores. These data are displayed on a logarithmic

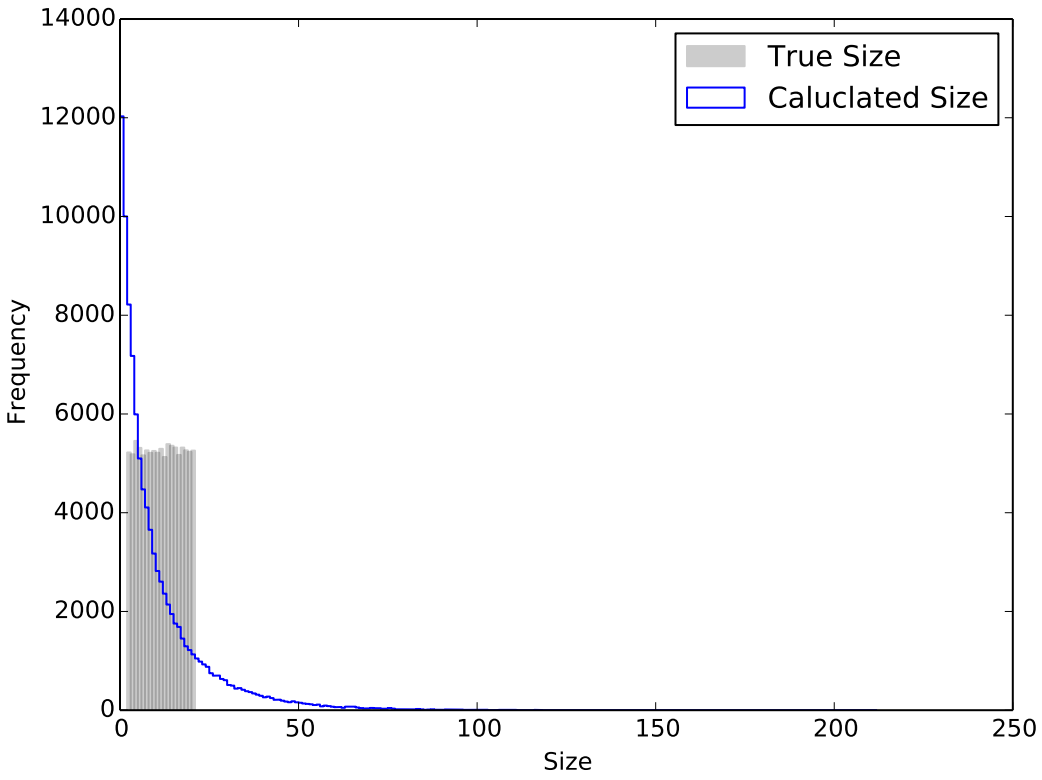

Figure 4.9: True and calculated oligomer sizes for data simulated using gamma-poisson emission behaviour. Under these conditions, the calculated oligomer size distribution (blue line) is an extremely poor approximation of the true size distribution (grey histogram).

scale, so that key features of the emission distribution can be identified. Several points of interest can be noted. Firstly, similarly to the dual-labelled monomeric species described in the previous chapter, no clear distinction is observed between the poissonian noise distribution and the distribution of photons from fluorescent events. This is true even when multiple fluorophores are present. Consequently, we present the entire dataset, including time-bins containing only noise photons, to avoid distorting the observed brightness distributions. Secondly, no clear peaks – corresponding to oligomers of different sizes – are observed. Instead, there is extensive overlap between the photon emission distributions for oligomers of different sizes, with the main differentiator being an increase in the length of the tail for species carrying more flurophores. Finally, it can also be observed that at the opposite end of the distribution, the frequency of events also increases for multimeric species: the monomer model has the highest number of bins containing zero photons. For all other photon counts, multimeric species show an increased frequency of observation.

This is in contrast to datasets simulated using the pure poisson emission model (Fig. 4.11). In these datasets, separate peaks corresponding to oligomers of different sizes can be clearly distinguished. Furthermore, rather than a monotonic decrease in frequency with number of observed photons, the poisson model shows an initial decrease in event frequency at low numbers of photons, corresponding to the noise distribtuion, followed by a second peak, corresponding to oligomer events. This produces a clear separation between the noise and event emission distributions, which would allow a simple threshold to select oligomer events.

On the other hand, the Holliday Junction photon distributions show better resemblance to datasets simulated using the gamma-poisson mixture model (Fig. 4.13). Like the Holliday Junction data, these datasets show a monotonic decrease in event frequency with increasing photons and display longer tails of rare, bright events when more fluorophores are present Furthermore, these datasets do not distinguish peaks corresponding to oligomers of different sizes.

This comparison raises two important points. Firstly, the emission behaviour of our oligomeric models in a single molecule fluorescence experiment is extremely poorly approximated by the poisson model for which an assumption of linear increase in photon emission with oligomer size would be valid. It is therefore clearly inappropriate to attempt using a simple thresholding method to calculate oligomer size distributions in aggregation reactions. These are likely to lead to the wildly inaccurate estimations of oligomer size distributions shown in Fig. 4.9.

Secondly, as the gamma-poisson mixture model appears to better approximate the emission

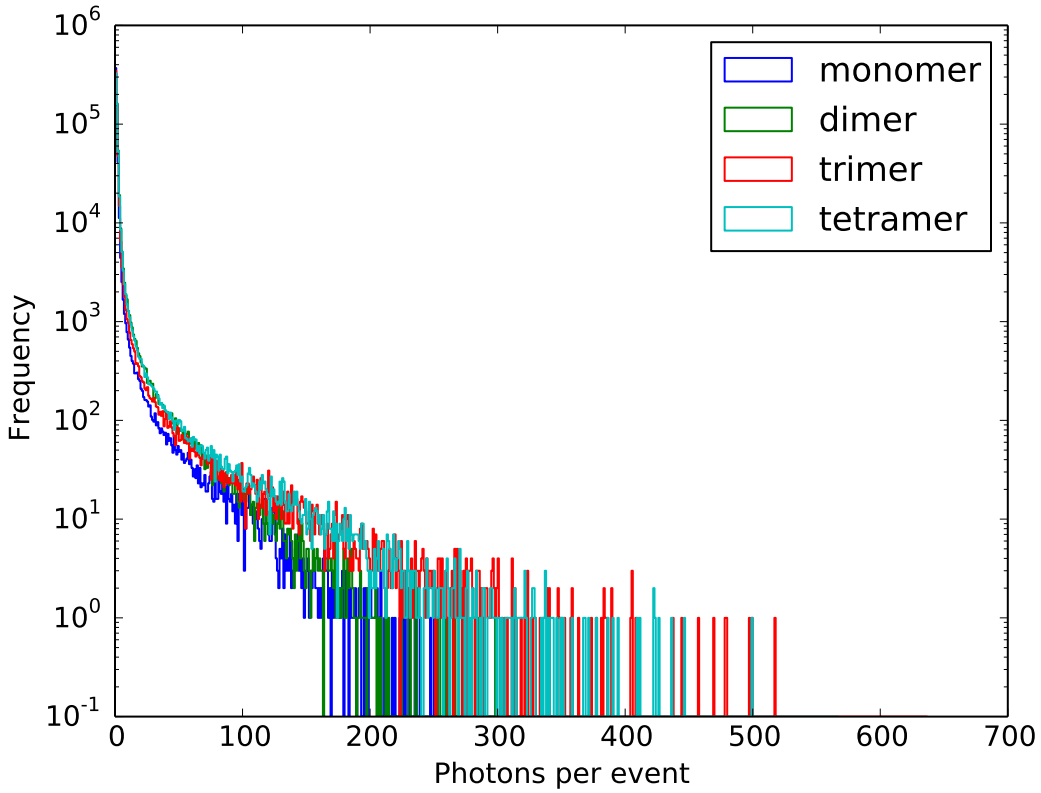

Figure 4.10: Histograms of experimental photon emission distributions for labelled Holliday Junctions. All time-bins are shown. Data from monomer, dimer, trimer and tetramer models are show in dark blue, green, red and cyan respectively.

distribution actually observered, it is possible that inferring the parameters of this model conditioned on an oligomeric dataset would enable accurate determination of the emission paramaters of at least oligomers of a single size. The next section describes our attempt to infer the parameters of the single-colour model, conditioned on datasets from the synthetic Holliday Junctions.

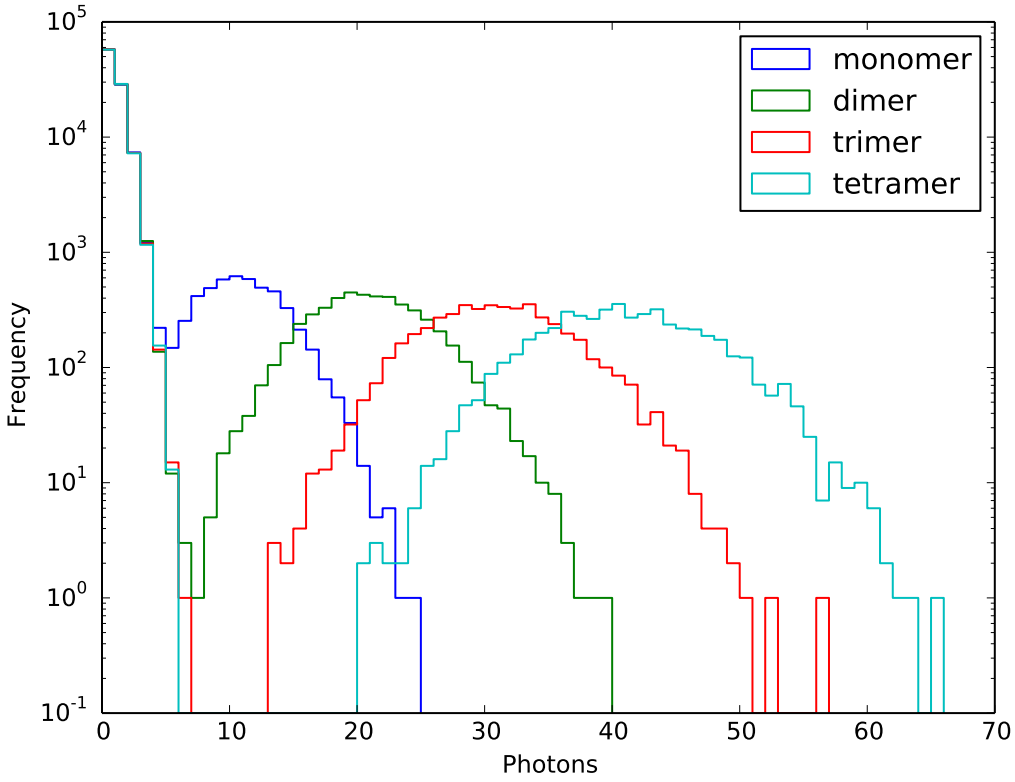

Figure 4.11: Histograms of photon emission distributions for model oligomers simulated using a pure poisson model. Data from monomer, dimer, trimer and tetramer models are show in dark blue, green, red and cyan respectively.

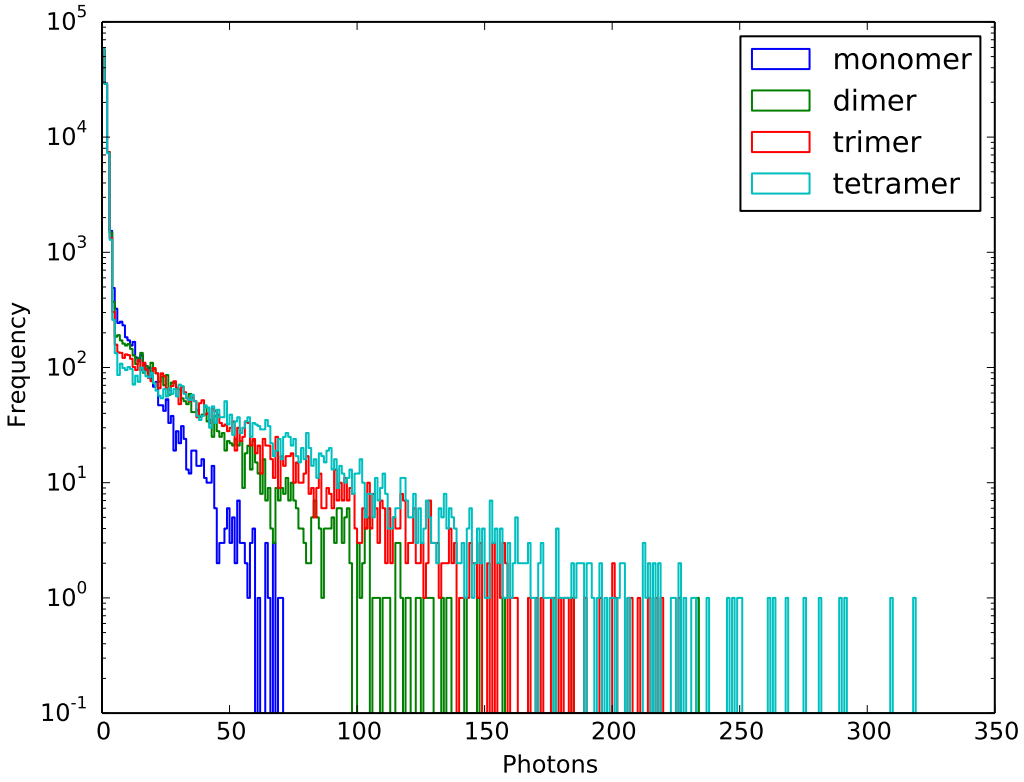

Figure 4.12: Histograms of photon emission distributions for model oligomers simulated using a gamma-poisson mixture model. Data from monomer, dimer, trimer and tetramer models are show in dark blue, green, red and cyan respectively.

4.5.3 Inferring Event Brightness Using the Gamma-Poisson Model

The previous section described comparing experimental data from model oligomers of known sizes with simulated datasets. We showed that poisson emission is an extremely poor model of the behaviour of fluorescently labelled oligomers in a single molecule fluorescence experiment. In this section, we use our Metropolis-Hasting sampler, described in Chapter chapter 3 to infer the parameters of our gamma-poisson mixture model, conditioned on single molecule datasets. We first present parameter inference on datasets simulated using this model, demonstrating effective performance. We then present the results of parameter inference conditioned on experimental datasets, showing that our lack of knowledge of key model parameters means that the model is underspecified, making accurate parameter inference difficult.

We simulated four datasets, consisting of oligomers of known size (from monomers to tetramers) using the parameters shown in Table 4.2. Then, we used the Metropolis Hastings sampler to infer the mean oligomer and noise emission rates, as well as the oligomer concentration. The initial values used are shown in Table 4.3; the inferred values are shown in Fig. 4.13. Two example fits are shown in Fig. 4.14

Table 4.2: Parameters used in simulating emission datasets for oligomers of known size.

Parameter	**Value**
λ_{prot}	0.05
λ_{NB}	1.0
λ_{DB} (monomer)	10.0
λ_{DB} (dimer)	20.0
λ_{DB} (trimer)	30.0
λ_{DB} (tetramer)	40.0
R_{blue}	1.0
γ_{ins}	1.0
n_{samples}	10000

The results are mixed. Although inference performs well for small oligomers (monomers and dimers), when the mean event brightness is very high (pure trimer and pure tetramer

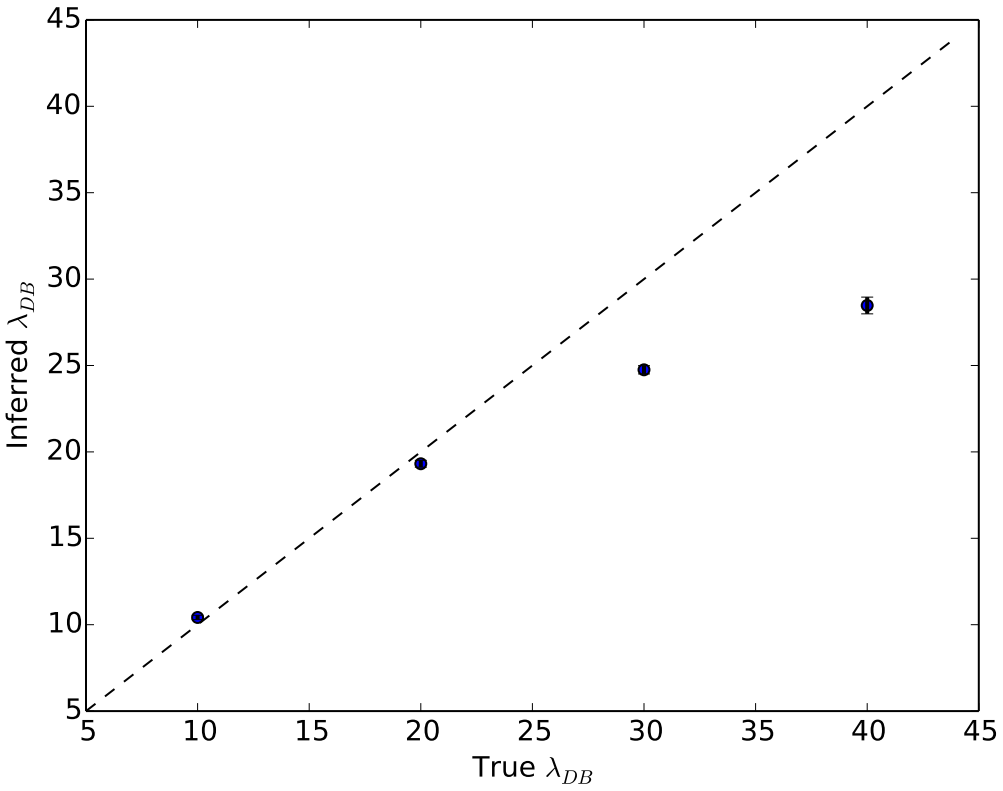

Figure 4.13: Plot of inferred vs true values of λ_{DB}. The dashed line shows the expected trend, but the inferred values are underestimated for high values of λ_{DB}. Error bars show the standard deviation of 100 samples.

Table 4.3: Initial values used in inferring oligomers of known size.

Parameter	Value
λ_{prot}	0.1
λ_{NB}	1.0
λ_{DB}	20.0

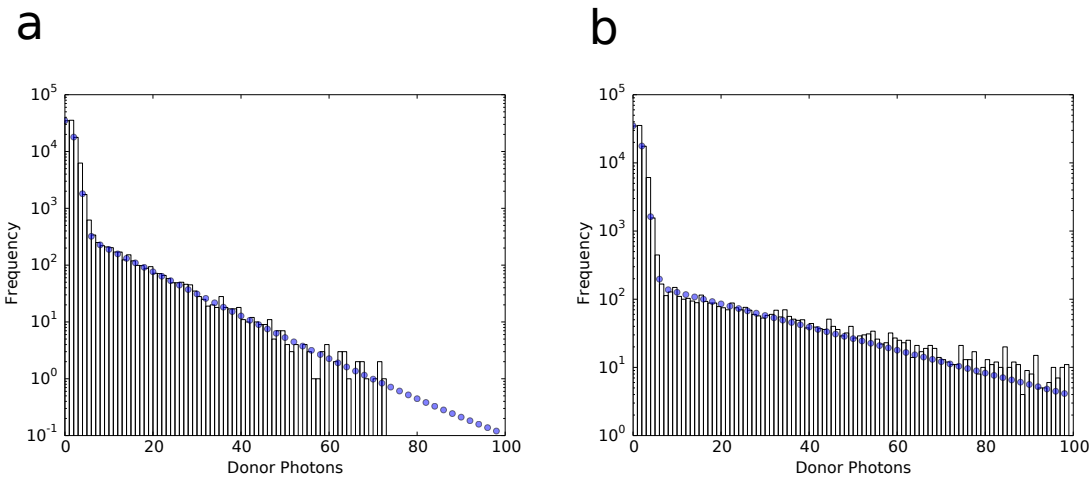

Figure 4.14: Histograms of photon emission distributions for simulated a) monomers and b) trimers. The overlaid blue circles show the number of photons predicted by the gamma-poisson mixture model, using parameters inferred from the dataset using the Metropolis sampler.

simulations), the true value of λ_{DB} is underestimated. Despite the considerable error, this still produces expected emission frequencies that are a good approximation of the true distribution (compare the histogram and the overlaid blue circles in Fig. 4.14 B). The likely cause of this estimation error is the overdispersal. In these simulations, the value of R_{blue}, the overdispersal parameter was 1.0. This produces an extremely broad distribution: in particular, for high mean values λ_{DB}, the distribution has a very long tail and the modal observed value is far from the mean of the distribution. As a consequence of this, the possible range of parameters values that provide a good fit to the observed data is very wide and the stationary distribution is broad and flat, allowing many values of λ_{DB} to provide a good approximation of the dataset.

It should be noted that this poor performance is not caused by the sampler not converging. As can be seen from the tight error bars in Fig. 4.13, the accepted samples are drawn from a small range of values. Furthermore, the initial conditions do not affect the final value

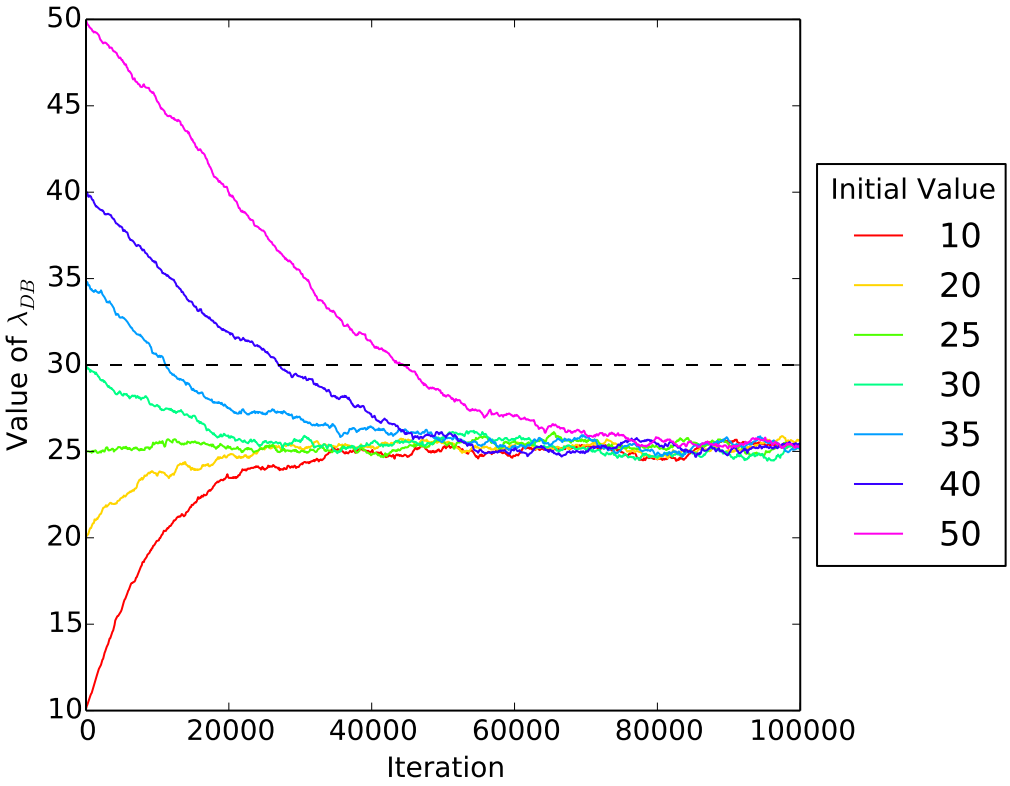

Figure 4.15: Convergence of the inferred value of λ_{DB} during burn-in for simulations of the trimeric oligomer. The black dashed line represents the true value of λ_{DB}.

reached by the sampler: although a longer burn-in phase is required when the sampler is initialised using values far from the true parameter values, sampling converges to the same (incorrect) values (Fig. 4.15).

This performance is disappointing. The datasets we have simulated are a greatly simplified approximation of a true single molecule fluorescence dataset from a protein aggregation experiment: they contain "oligomers" of only a single size; many sources of error, such as cross-talk, are entirely neglected and all model parameters are known. Despite knowing that the model for which we are inferring parameters is precisely the model that generated our datasets, we are unable to recover the correct parameter values. This suggests that our model-based approach will struggle similarly to extract meaningful parameters from experimental datasets. Our attempt to do so is summarised in the next section.

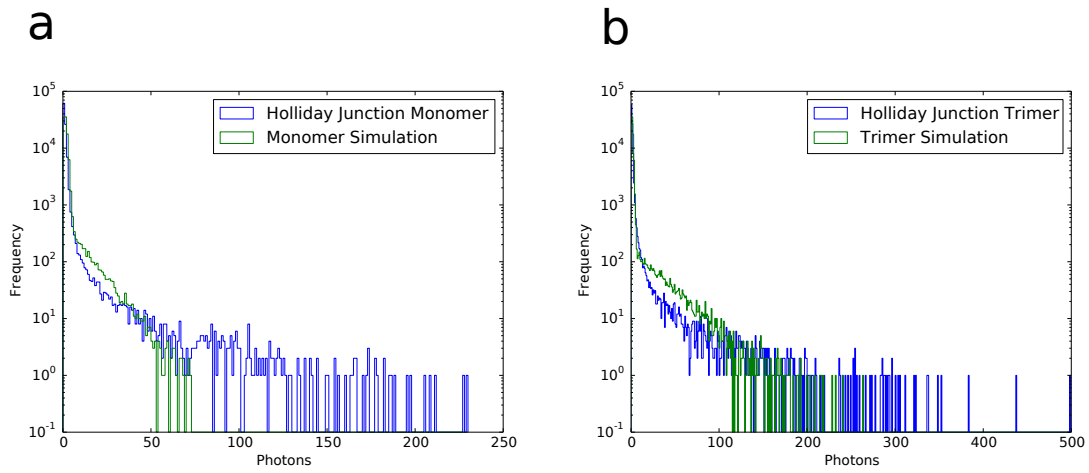

Figure 4.16: Overlays of simulated and experimental photon emission distributions for a) monomer and c) trimer oligomers. The blue histograms display experimental data from the Holliday Junction model oligomers. The green histograms display data from simulations using the gamma-poisson mixture model.

4.5.4 How Bright Are Holliday Junction Events

This section discusses our attempt to infer parameters of the gamma-poisson mixture model conditioned on data from our Holliday Junction model oligomers. Firstly, we provide a detailed comparison of the simulated and experimental datasets, identifying probable errors in the model. Secondly, we compare different methods of estimating the event brightness of fluorescence emission from the Holliday Junction oligomers. Finally, we consider experimental techniques to understand sources of discrepancy between our model and the observed data.

Fig. 4.16 A and B show emission distribution histograms for the monomer and trimer model oligomers, overlaid by simulated monomer and trimer emission distributions, simulated using the gamma-poisson mixture model. Although there is rough agreement between the two distributions, several discrepancies can be noted. Firstly, as is particularly evident in the comparison of monomer data (Fig. 4.16 A), the experimental distribution shows a much longer tail of bright events. Secondly, the experimental datasets display fewer events of intermediate fluorescence. These discrepancies suggest that the gamma-poisson mixture model, as parameterised in our simulations, is not a perfect model for fluorescent emission.

The most likely source of this discrepancy is incorrect parameterisation of the gamma-poisson mixture distribution. This occurs because the true values of several parameters of this

Table 4.4: Inferred parameter values conditioned on a dataset from the monomer model Holliday Junction, using different fixed values of R_{blue}. At low values of R_{blue}, overfiting occurs and the inferred parameters are unphysical. When we try to infer the value of R_{blue} (bottom row), the data is also overfitted.

$R_{\mathbf{blue}}$	$\lambda_{\mathbf{noise}}$	$\lambda_{\mathbf{prot}}$	$\lambda_{\mathbf{DB}}$
0.1	0.44	0.215	2.769
0.5	0.46	0.059	9.784
1.0	0.47	0.042	13.156
2.0	0.48	0.035	15.193
inferred = 0.029	0.44	1.006	0.591

distribution are unknown. In particular, in our simulations, we assume that the overdispersal parameter, R_{blue}, is 1.0. However, for expermimental data, this parameter is unknown. By comparison of datasets simulated using different values of R_{blue} with experimental data, it is possible to estimate the experimental value of R_{blue} (Fig. 4.17). Unfortunately, however, the parameter inference is extremely sensitive to the value of R_{blue} (see Fig. 4.18 and Table 4.4), meaning that without a good estimate of this value, it is very difficult to infer physically meaningful values.

To overcome this issue, we attempted to co-infer R_{blue} along with other model parameters, by allowing it to become a variable rather than a constant in our model. Unfortunately, this was also unsuccessful, as the model converges to extremely low values of R_{blue}, for which the inferred oligomer concentration is concommitantly unphysically high and the mean event brightness very low (Table 4.4).

This poor behaviour is caused by overfitting. There is insufficient information in the datasets to distinguish between a solution with manny molecules each of which emit only a few photons and a model in which few molecules emit many photons in each rare excitation event. Moreover, a model in which there are multiple dim emitters in the confocal volume at any one time gives greater flexibility to exactly fit small, variations in the observed emission distributions. As a result, dim emitters are preferred as they allow the model to more closely approximate this random noise, resulting in clearly incorrect and unphysical fits.

This has considerable implications for our ability to fit oligomer data using the Metropolis sampler, as the results of sampling are extremely dependent on the value of the overdisper-

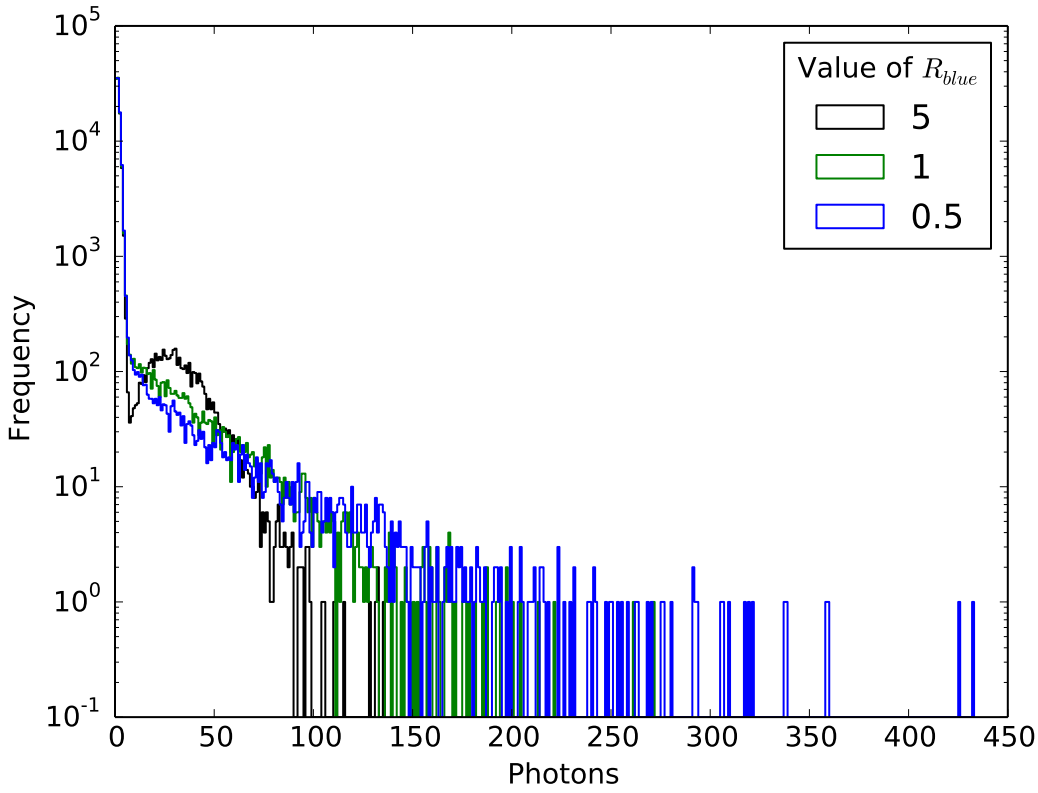

Figure 4.17: Photon emission distributions for datasets simulated with different values of the overdispersion parameter R_{blue}. The values used are shown in the legend. For high values of R_{blue}, the degree of overdispersal is small, and the fluorescence emission distribution resembles a pure poisson distribuion (black histogram). For small values of R_{blue}, the amount of overdispersal is much greater and the tail of the distribution is lengthened (green histogram).

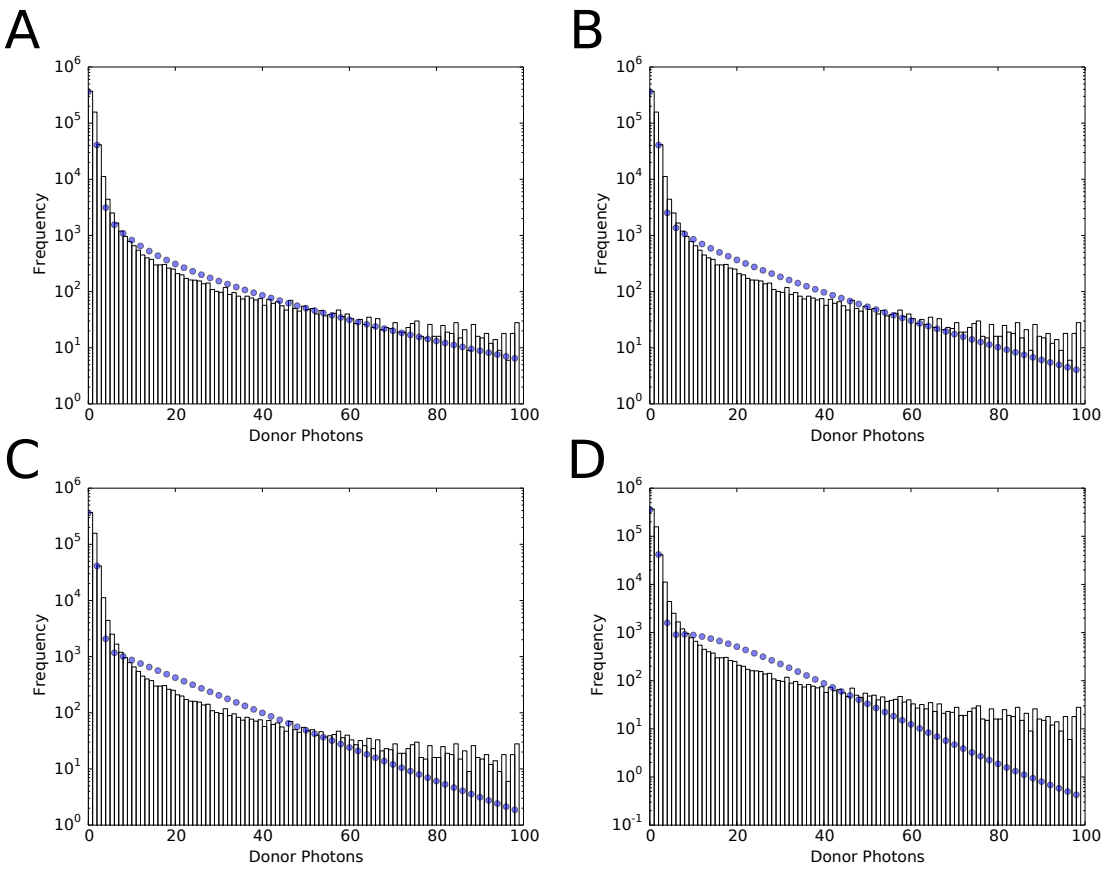

Figure 4.18: Fitting a dataset from the monomer model Holliday Junction, using different fixed values of R_{blue}: a) $R_{\text{blue}} = 0.1$, b) $R_{\text{blue}} = 0.5$. c) $R_{\text{blue}} = 1.0$, d) $R_{\text{blue}} = 2.0$. Although the data superficially seems better fitted when R_{blue} is smaller, low values of R_{blue} overfit random variations and produce unphysical parameter values (see Table 4.4).

sion parameter R_{blue}, which we are not able to infer accurately. Consequently, extension of the inference tool to oligomer sizing is unsuccessful. There is insufficient information obtained from an experiment to accurately calculate the mean event emission brightness or to separate the noise and fluorescence emission distributions, even for molecules carrying a known, uniform number of fluorophores. However, as we have shown (Fig. 4.9), the degree of overdispersal means that simple thresholding analyses are also entirely inappropriate to accurately calculate oligomer size distributions. To obtain accurate size information from a single molecule fluorescence experiment, it is necessary to reduce the sources of heterogeneity, so that photon emission can be well approximated using a pure poisson distribution (Fig. 4.7). The following section describes our attempts to understand sources of emission overdispersal. The chapter concludes with a discussion of possible methods to reduce this heterogeneity.

4.5.5 Photobleaching Steps Analysis Reveals Additional Source of Overdispersal

This section describes the results of a photobleaching step analysis of the Holliday Junction model oligomers. We inially performed these experiments to verify the presence of multiple fluorophores on the dimer, trimer and tetramer constructs; however it also revealed that photoblinking is likely to be a major source of emission heterogeneity in single molecule fluorescence experiments.

The final results of these experiments, summarized in Fig. 4.19, show that, although we do observe an increase in the mean number of photobleaching steps for Holliday Junctions labelled with more fluorophores, there is considerable variation in the number of photobleaching steps seen. This could imply that the Holliday Junctions are badly formed and therefore not a homogeneous sample. However, this result should be considered in conjunction with the fact that, of all the photobleaching traces collected 0.81 ± 0.01 (i.e. over 80%) had to be discarded during quality control as being unsuitable for photobleaching step counting, because they contained unclear transitions. A typical example of a trace that could not be used is shown in Fig. **??**. Both increases and decreases in the observed photon emission are clearly observed. This behaviour is caused by photoblinking [26], reversible transitions of the fluorophores between bright and dim or dark emission states.

The extensive photoblinking behaviour that we observe here in Alexa Fluor 488 is cause for

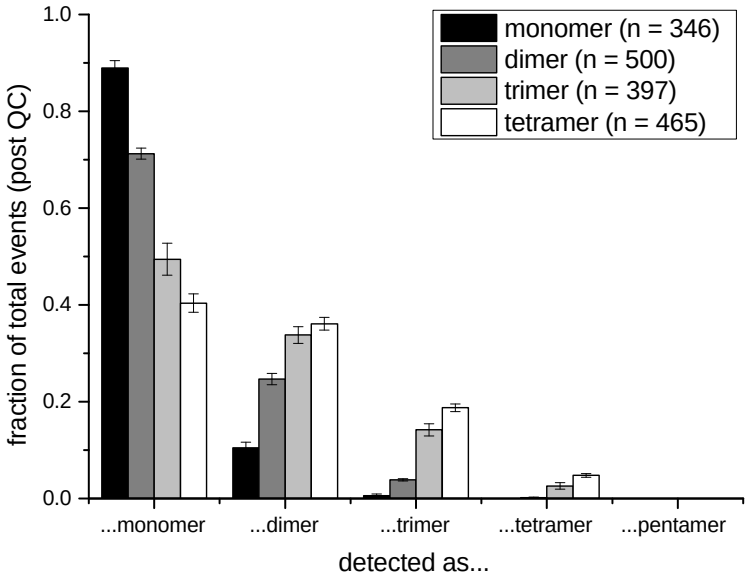

Figure 4.19: Photobleaching analysis of the Holliday Junction oligomer models. Although the mean number of photobleaching steps observed does increase with the number of attached dyes, there is considerable heterogeneity. *This figure was prepared by Kristina A. Ganzinger and used with permission.*

concern. This blinking occurs on a timescale comparable with dwell-time in the confocal volume during a diffusion-based single molecule experiment. Experimental techniques that aim to reduce the excitation field heterogeneity will not affect heterogeneity caused by this photo-physical effect. Consequently, any attempt to accurately size individual oligomeric molecules, or to characterize a size distribution based on the observed fluorescence emission brightness distribution must take this source of heterogeneity into account.

4.6 Conclusions

4.6.1 Complex Relationship between Size and Photon Emission

This chapter described a model-based approach to understanding the size distribution of oligomers in a single molecule fluorescence study of protein aggregation. Through a comparison of experimental datasets with simple simulations, we were able to show that the photon emission distribution observed, even from a simple, homogenous solution of labelled molecules, is much better modeled by an overdispersed poisson distribution (gamma-poisson mixture) that a simple poisson distribution.

We showed, using simulated datasets, that this overdispersion makes the number of observed photons and extremely poor estimator for molecular size (Fig. 4.9), resulting in a huge divergence between true and calculated size distributions. Furthermore, because we were not able to accurately estimate or measure the value of the overdispersion parameter R_{blue}, we were unable to use the Metropolis sampler described in the previous chapter to infer reliable estimates of our model parameters. Consequently, we conclude that unless stragegies can be found for reducing photon emission heterogeneity, it is extremely difficult to accurately infer oligomer sizes based on their emission behaviour in single molecule fluorescence experiments.

4.6.2 Implications for Future Work on Molecular Sizing

Reducing emission hetereogenity in order to accurately infer oligomer sizes requires a dual approach. Firstly, it is necessary to suppress photoblinking behaviour, so that reversible transitions between light and dark states, which would broaden the emission distribution, do not occur. Certain fluorophores, such as Rhodamine 6G, are known to show extensive photoblinking behaviour, and a heterogeneous distribution of dark state lifetimes [26]. Alexa

Fluor 488, the fluorophore used in our study of Holliday Junctions, has extensive structural similarity to Rhodamine G6, with both dyes having a heterocyclic aromatic structure, derived from fluorone (Fig. 4.20). Consequently, it is likely that the photoblinking behaviour observed in Alexa 488 is caused by a single electron transfer allowing population of a radical state, similar to that observed in Rhodamine [26].

Several systems have been derived that reduce the occurrence of photoblinking in organic fluorophores. These systems, including the glucose oxidase - catalase (GODCAT) system [140], Trolox-based buffers [141] and protocatechuic acid (PCA)/protocatechuate-3,4-dioxygenase (PCD) [142], typically scavenge triplet oxygen from the solution, removing the main source of electrons for the dark state entry reaction. Although our buffers were freshly degassed, to reduce the concentration of dissolved oxygen, we did not use an oxygen scavenging system. We suggest that use of such a system could help to reduce the extensive photoblinking that prevented oligomer quantification.

Secondly, emission heterogeneity can also be reduced by creating a more uniform exciation within the confocal detection volume. The gaussian shape of the exciting laser beam means that the pathway taken by molecules diffusing through the confocal volume affects both the intensity and duration of excitation, broadening the observed emission distribution. Several recent studies identify methods for reducing the confocal excitation heterogeneity. Liu and Wang [143] use a cylindrical lens to create a uniform excitation field in one dimension; they then use a square pinhole combined with nanofluidic channels to detect only molecules that pass precisely through this uniform excitation field. A similar approach is taken by Tyagi and coworkers [144], who use reversible flattening of microfluidic channels to observe emission from fluorescent molecules confined in the TIRF field. In this work, the channels were also perfused with nitrogen gas, enabling removal of oxygen molecules capable of inducing blinking. We suggest that adopting such techniques, which homogenize both excitation duration and intensity, could be used to reduce emission heterogeneity and hence enable accurate quantification of oligomer sizes based on fluorescence emission behaviour.

Finally therefore, we would like to conclude this chapter on an optimistic note. Although we were unsuccessful in quantifying oligomer sizes, our model based approach has enabled us to identify limitations in current experimental practices that are currently preventing single molecule fluorescence approaches from accurately determining oligomer sizes. Based on these limitations, we are able to suggest several possible modifications to the experimental protocol that could improve the accuracy of this technique. We hope that these suggestions

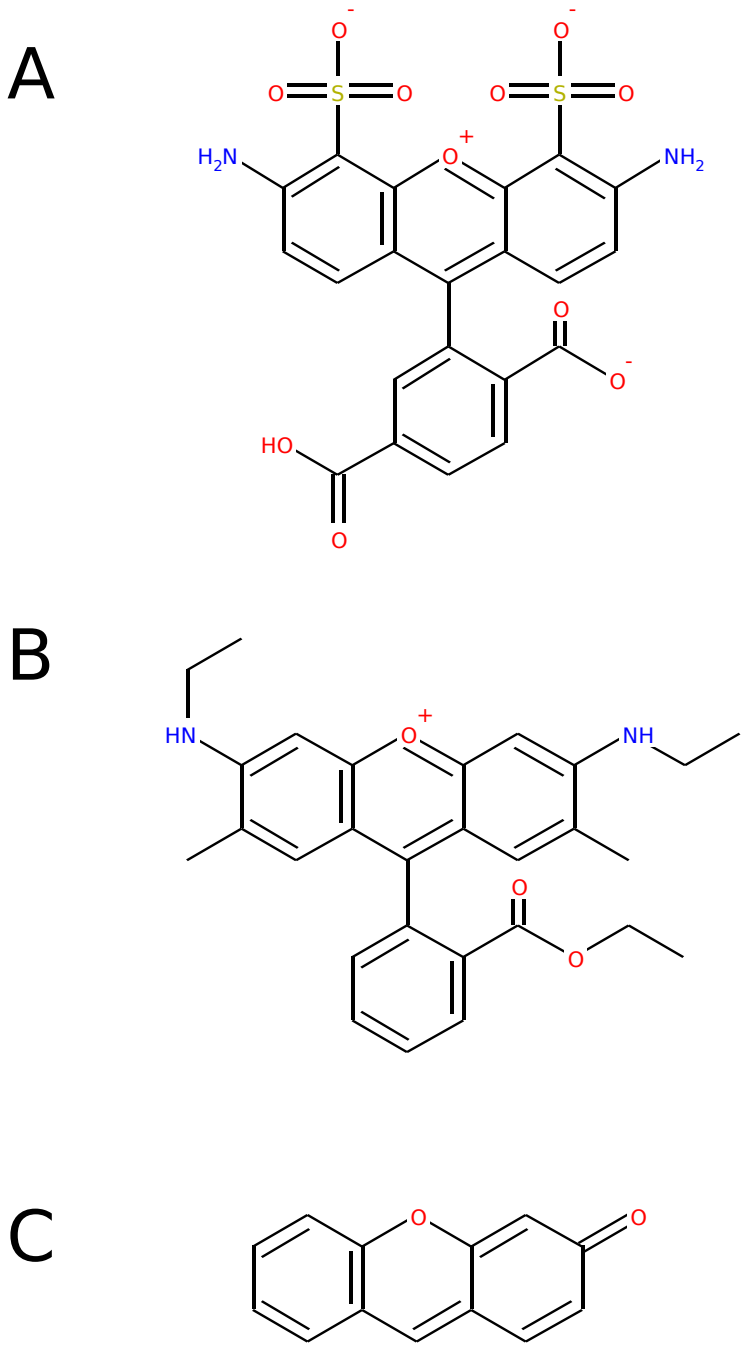

Figure 4.20: The chemical structures of A) the para isomer of Alexa Fluor 488 and B) Rhodamine 6G are very similar. Both are derived from the flurone ring system shown in C).

prove helpful to future researchers working in this field.

Chapter 5

Probabilistic Inference for Error Detection in De Novo Genome Assemblies

5.1 Overview

This chapter presents results in a different research domain than the previous three chapters. Instead of confocal single molecule fluorescence, here we consider gene sequencing using Illumina's fluorescence-based sequencing technology [145]. We describe the development and evaluation of an error detection tool, NxRepair, designed to locate and correct misassemblies in *de novo* assemblies of bacterial genomes constructed using Illumina Nextera mate pair [146] sequencing. In a *de novo* sequence assembly, scaffolding errors and incorrect repeat disambiguation during *de novo* assembly can result in large scale misassemblies in draft genomes. Nextera mate pair sequencing data provide additional information to resolve assembly ambiguities and prevent these errors. However, despite this additional information, mistakes can still occur. We show, through comparison with matched refernence genomes, that the NxRepair tool can quantitatively improve the quality of *de novo* assemblies and that it outperforms other error detection tools in identifying large-scale misassemblies.

This chapter opens with a brief overview of sequencing technologies and their relationship to fluorescence microscopy. We then introduce the problem of *de novo* genome assembly and discuss prior methods for error detection in genome assemblies. Following this introduction,

we describe the development of our NxRepair tool, which uses a probabilistic analysis to identify regions of a *de novo* assembly that have a high probability of containing an error. Finally, we evaluate the performance of NxRepair on a series of nine bacterial genomes. We show that NxRepair can identify and correct large scaffolding errors, without use of a reference sequence, resulting in quantitative improvements in the assembly quality. We close this chapter by discussing the potential applicatons of this tool, but also the limitations of this approach to error detection. NxRepair can be downloaded from GitHub; a tutorial and user documentation are also available.

5.2 Introduction

5.2.1 Next-Generation Sequencing Technologies

Sequencing is the determination of the sequence of base pairs in a strand of DNA. Since the completion of the first complete genome sequence [147], there have been huge developments in the technology used both to read the seqeunce of DNA bases and to assemble those reads into a complete genome. The first genome sequencing experiments used polymerisation reactions with a reaction mixture containing both radiolabelled deoxynucleotidetriphosphates (dNTPs) and radiolabelled di-deoxynucleotidetriphosphates (ddNTPs), which terminate chain elongation at specific bases. The elongation reaction was carried out four times, each time using a different ddNTP, so that fragments in one reaction mixture would all terminate with the same base. Following elongation, the newly constructed framgments from the four reaction mixtures were run out on a high-resolution gel and the sequence read out manually based on the fragment position and the ddNTP used in the elongation reaction (Fig. 5.1, [147]). However, this method of sequencing was laborious, slow and expensive, significantly limiting the widespread application of DNA sequencing technologies [148]. The development of fluorescent nucleotide labels [149], initially enabling sequencing from a single elongation reaction (Fig. 5.1 [150]) combined with methods for automating and paralellizing the sequencing reactions created the field of high-throughput, or next-generation, sequencing (NGS).

Illumina Sequencing Technology The most popular, and best known, NGS technology is Illumina's sequencing by synthesis [151]. In the Illumina sequencing pipeline, DNA

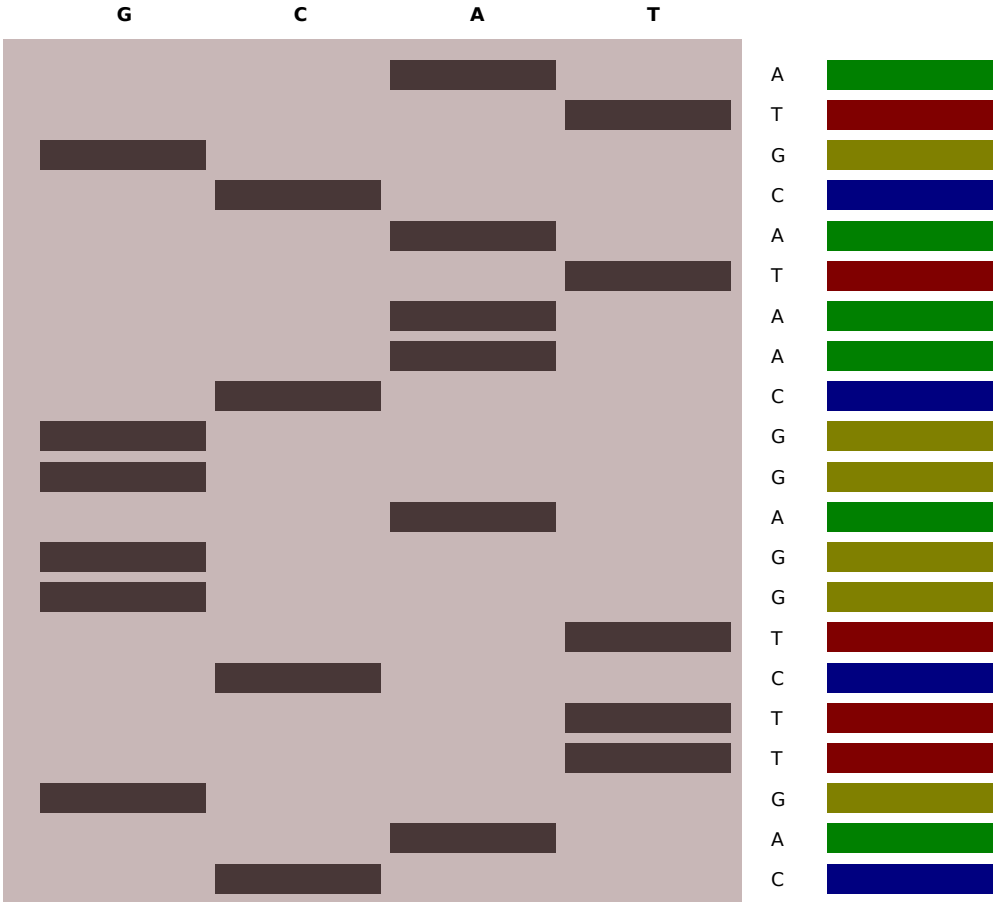

Figure 5.1: Schematic of Sanger Sequencing using terminator chemistry. Left: sequencing on a four-lane gel. Four lanes represent sequencing mixtures containing different ddNTPs. Right: Sanger sequencing in a single lane using fluorescently labelled ddNTPs. Middle: the corresponding DNA sequence.

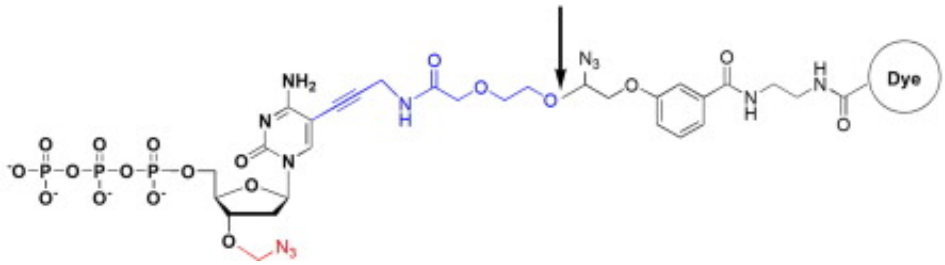

Figure 5.2: A modified nucleotide showing Illumina terminator chemistry. The 3'-hydroxyl group is reversibly blocked by the 3-O-azidomethy terminator (red atoms), whilst the fluorescent indicator dye can be reversibly cleaved from the modified DNA base at the bond indicated by the arrow. Figure adapted from [153]

strands are fragmented into short pieces and attached to specific adapter sequences. The modified sequences are introduced to the Illumina flowcell, locally amplified in a "clustering" reaction [152] and then cycled through repeated elongation and imaging steps, to read out the DNA sequence one nucleotide at a time (Fig. 5.3). Illumina sequencing is a fluorescence-based technology. It uses reversible terminator chemistry to elongate a templated DNA molecule by a single base, using a dye-labelled nucleotide bearing a terminator group to prevent further elongation. The DNA base is identified based on the fluorescence emission of the attached dye. Following imaging, the terminator group and dye are removed and the next terminated nucleotide introduced at the now exposed hydroxyl group (Fig. 5.2).

This is a fluorescence based imaging technique. However, there are several major differences between this sequencing process and the single molecule confocal experiments described in the previous chapters. Firstly, rather than freely diffusing molecules imaged by a laser beam, the fluorescently labelled sequences are tethered to the flow-cell base and imaged using multiple cameras and epifluorescent illumination (see Section 1.3.1). Secondly, unlike in a smFRET experiment, the imaging step does not attempt to distinguish individual fluorophores against a noisy background signal. Instead, the clustering step prior to imaging produces multiple, spatially localised copies of the same DNA sequence, hugely amplifying the available fluorescent signal.

Single Molecule Sequencing Technologies Although Illumina does not use single molecule imaging, several other sequencing methods do sequence individual DNA molecules. Single Molecule Real Time Sequencing (SMRT) [154] is another fluorescence based sequencing tech-

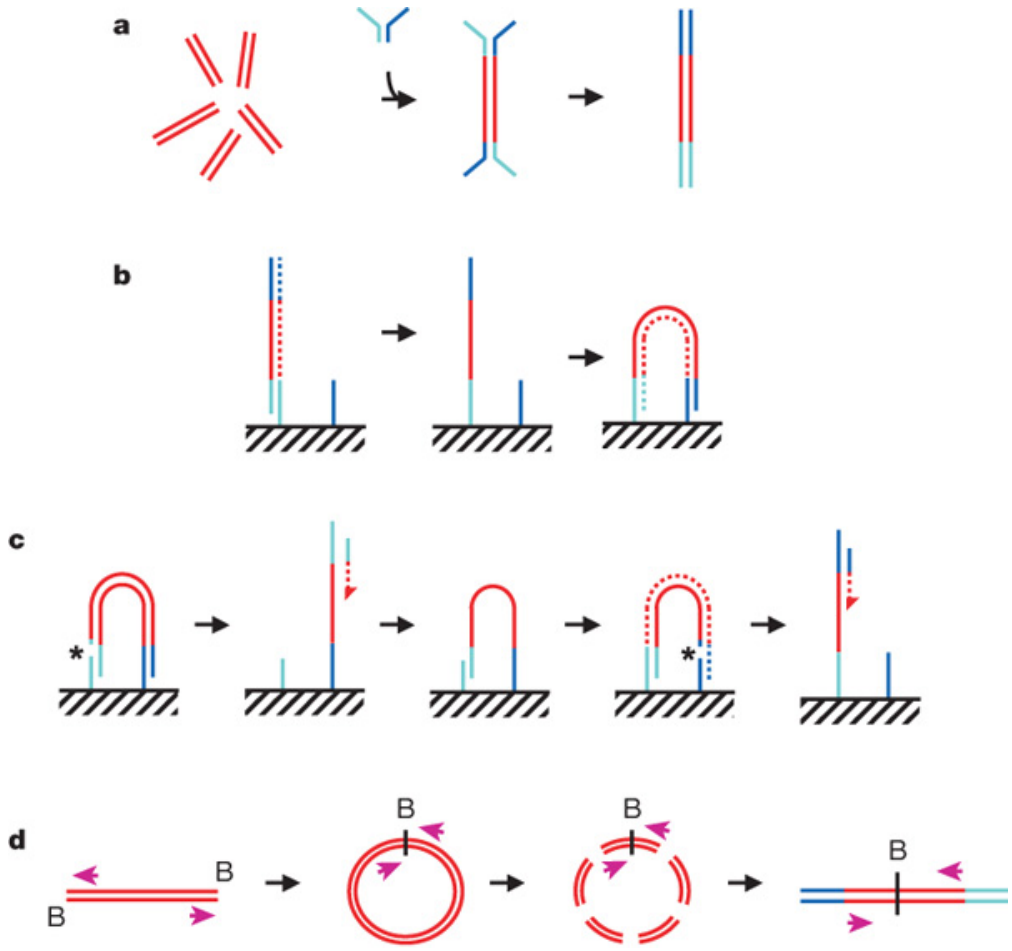

Figure 5.3: Sequencing by synthesis. A) DNA fragments are generated and joined to a pair of oligonucleotides. The ligated products are amplified using two oligonucleotide primers, resulting in double-stranded blunt-ended material with a different adaptor sequence on either end. B) DNA fragments are denatured and single strands are annealed to complementary oligonucleotides on the flow-cell surface. A new strand is copied from the original strand in an extension reaction; the original strand is then removed by denaturation. Multiple cycles of annealing, extension and denaturation in isothermal conditions result in cluster growth. C) The DNA in each cluster is linearized by cleavage within one adaptor sequence and denatured, generating single-stranded templates. D) Long-range paired-end sample preparation. The ends of each long fragment are tagged by incorporation of biotinylated (B) nucleotide and then circularized, forming a junction between the two ends. Circularized DNA is randomly fragmented and the biotinylated junction fragments are recovered and used as starting material in the standard sample preparation procedure. Figure and (adapted) caption from [145].

nology. In SMRT, DNA polymerase molecules are immobilized within the detection region of a zero-mode waveguide (ZMW). Fluorescently labelled dNTPs are incorporated by the polymerase and imaged in real time. The ZMW, an aluminium-coated glass slide, containing multiple nano-scale pinhole apertures, creates a zeptolitre detection volume, enabling detection of the nucleotide currently under incorporation, despite a relatively high (μM) concentration of labelled oligonucleotides [155]. SMRT, used in Pacific Biosciences sequencing technologies, is currently extremely expensive, but produce sequences with an average read length of 3000 nucleotides [156], compared with the 151 or 251 bases typical of an Illumina read.

A second, currently experimental, single molecule sequencing technique is under development by Oxford Nanopore [157]. This sequencing method uses electrophoresis to drive long DNA molecules through a nano-scale pore. Application of a voltage across the nanopore causes a flow of ions through the pore, creating a measurable current. DNA is negatively charged, so can also be drawn through the pore, causing variations in the observed current that are dependent on the DNA bases currently occupying the pore [158]. In principle, this allows a direct read-out of the base sequence; in practice several challenges, including control of the speed of strand progression through the pore and achieving individual base resolution [159], mean that although nanopore sequencing can produce very long reads (average 5000 bases), these have a high error rate including systematic errors [160].

5.2.2 De Novo Sequence Assembly

As described above, Illumina NGS produces reads of length 151 or 251 bases. Even a small bacterial genome has a total genomic size of several million bases. Consequently, one of the most challenging aspects of genome sequencing is to piece together the short fragments of sequence obtained from the sequencer into one or more contiguous sequences, corresponding to correctly reconstructed sections of the genome. This process of attempting to generate a complete genome from short reads is termed *de novo* sequence assembly; it is distinct from, for example, sequence alignent [161], in which short reads are aligned against a previously constructed reference genome for the same organism, allowing identification of polymorphisms and small differences from the reference genome.

De novo assembly is the construction of a long, contiguous genomic sequence from short DNA reads, without using a reference genome. The most popular method of *de novo* assembly

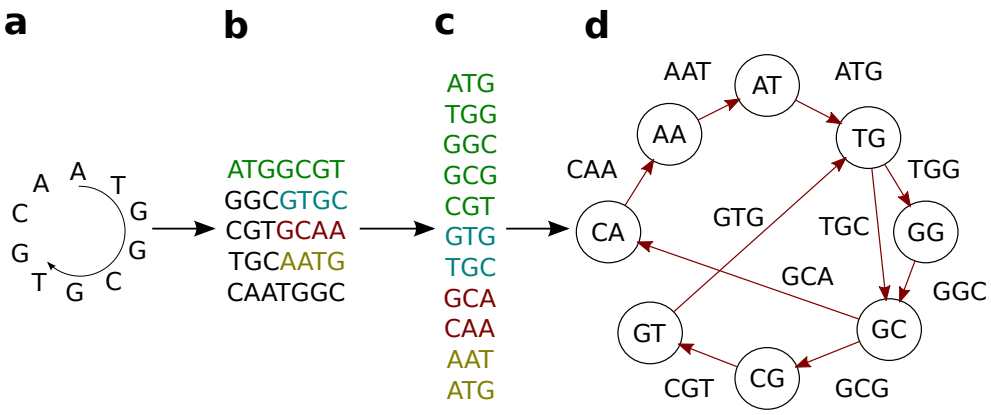

Figure 5.4: De Bruijn graph assembly from k-mers. (A) A small, circular genome (B) Some short reads sequences. (C) Breaking the reads into k-mers ($k = 3$). k-mers are coloured according to the sequence from which they are derived. (D) Constructing a De Bruijn Graph from the k-mers. Figure adapted from [162].

contstructs a de Bruijn overlap graph [162] from even shorter fragments, termed k-mers, generated by fragmenting the short reads, then traverses this graph to reconstruct the genome sequence. De Bruijn graph traversal as a method of assembly was proposed as early as 2001 [163]; however the method only gained popularity once the first de Bruijn assemblers were released [164]. In the simplest case, the graph is constructed from single end reads. However, with only single end reads, disambiguating repeat regions, which tangle the de Bruijn graph, remains challenging.

In de Bruijn graph assembly, the sequencing reads are subdivided into all possilbe shorter reads of length k, termed k-mers. A k-mer is described as having a prefix (the first $k-1$ nucleotides) and a suffix (the last $k-1$ nucleotides). A graph is then constructed, in which observed prefixes and suffixes are represented as nodes; a directed edge is drawn between two nodes if the k-mer representing that prefix-suffix pair is present in the sequencing data (Fig. 5.5). The challenge of constructing the *de novo* assembly is then addressed by constructing a continuous path through the resulting de Bruijn graph, in which every edge is visited. Such a path is termed an Eulerian cycle; efficient algorithms exist to find such cycles and many have been implemented for the *de novo* assembly application [164, 165, 166].

Even using the most sophisticated assemblers, it is rare for the entire genome to be recovered as a single, contiguous unit. Consequently, assembly is typically a two-stage process. First, long contiguous sections, named contigs, are constructed. Second, once the contigs cannot be extended any further, "scaffolding" algorithms [167] attempt to join multiple contigs, using

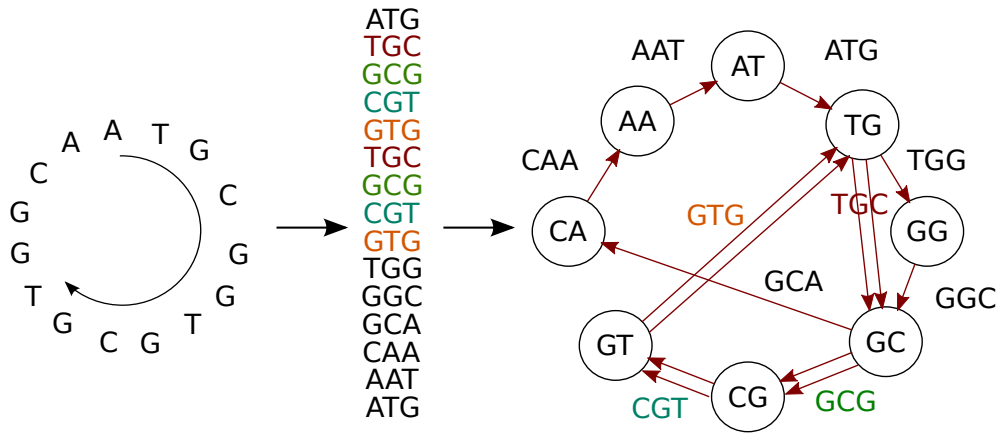

Figure 5.5: A se Bruijn graph with repeats. The larger circular genome ATGCGTGCGTGGCA contains the repeated section TGCGTG, resulting in several 3-mers (TGC, GCG, CGT and GTG, coloured red, green blue and orange respectively) appearing more than once. This results in a tangled de Bruijn graph. Figure adapted from [162].

insert size information to determine contig order and approximate gap size.

One important challenge that is not resolved by de Bruijn graph assembly is the disambiguation of repeats – sequences found in multiple places in the genome – which tangle the de Bruijn graph (Fig. **??**): certain k-mer sequences, corresponding to the start and end of repeat regions, have multiple entry and exit paths, the ordering of which cannot be determined. Repeat disambiguation is a major challenge for correct *de novo* assembly. Consequently, several alterations to the sequencing methodology have been developed to provide more data about the positioning of repeat sequences within a genome.

5.2.3 Paired End Reads and Mate Pair Sequencing

Paired End and Single End Reads The description of *de novo* assembly given above has assumed that reads are "single-end" – that is, the fragments of genomic DNA attached to the flowcell surface are sequenced using one strand only. However, a simple modification to the protocol additionally allows sequencing of the complementary strand. This provides two sequences, corresponding to the two ends of the genome fragment, separated by some unsequenced central region. These two sequences can be matched [168], to give a pair of reads (a read pair), separated by some small distance within the genome. The genomic distance from the start of one read to the end of the other is termed the insert size. The genomic

libraries used for this paired end sequencing typically contain fragments with an insert size of around 500 bases in length, although this can be controlled during library preparation.

When the insert size (fragment length) is longer than the length of a repeat unit, paired end reads provide extra information that can help to disentangle a de Bruijn graph. By creating a bridge between the start and end sections of a repeat unit, the paired end reads can enable the order of repeat sections to be determined. Some assemblers [169, 165] directly incorporate pairing information into the de Bruijn graph; in other assembly pipelines, pairing information is used in the subsequent scaffolding step to order determine the correct ordering of the contigs.

Mate Pair Sequening Many repeat regions are larger than the average insert size of a paired end library, with the result that fully constructing a genome from short reads is still not possible. To enable further disambiguation of the repeat structure, protocols to create read pairs with a much larger insert size have been developed [146], allowing construction of libraries containing fragments that are several kilobases (kb) in length. These "mate pair" libraries can be used either alone, or in combination with a paired end library, to provide additional information for a *de novo* assembly.

5.2.4 Evaluating Assembly Quality

Many assemblers incorporate mate pair insert size information into either both the contig assembly and scaffolding processes [169, 165], or just into the scaffolding step [164] but errors can still occur. These errors can significantly affect the quality of the assembled genome, with repercussions for downstream research. Consequently, error correction and quality evaluation of *de novo* assemblies are problems receiving considerable research interest.

The most serious mistakes found in a *de novo* assembly are large scale scaffolding or extension errors (Fig. 5.6 (A)), in which two two disparate regions of a genome are incorrectly joined together. Similarly, large insertion or deletion errors (indels) create structural irregularities in the *de novo* assembly; whereas mistakes in base calling lead to errors at a single position only.

The most common method of evaluating assembly quality is to measure its $N50$ score, defined as "the length of the shortest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the total assembly" [170]. However, this

evaluation method optimises only for contig length, without considering the correctness of the assembly. When a reference genome is available, the modified $NA50$ statistic [171] can be used, which considers only contig sections that correctly align to the reference. Recent work, such as the Assemblathon [172] and GAGE [173] collaborations used this metric to compare the quality of the *de novo* assemblies prepared by various assemblers by comparing them with their equivalent reference genomes. However, in the absence of a high-quality reference genome, other methods of quality evaluation must be used.

5.2.5 Error Detection Methods

Ghodsi et al. [174] have developed a Bayesian method of assembly quality evaluation, which can calculate an assembly quality score, without requiring a reference genome. However, this provides only an overall quality score and cannot be used to identify errors or low-quality regions. Several recent papers have developed error identification and correction methods, which perform a fine-grained analysis of assembly quality. The most well-known of these is the A5 Assembly Pipeline [175, 176], which includes an error detection and rescaffolding step that makes use of mate pair alignment information. Two new tools, REAPR [177] and ALE [178] have also been developed to use read pair data to identify misassemblies. A similar tool is currently under development at the Broad Institute [179]. However, with the exception of ALE, which is no longer actively maintained, these newer tools are not optimised to use mate pair information.

This chapter describes NxRepair, an assembly error detection tool that we designed to identify the most serious misassemblies by examining the distribution of Nextera mate pair insert sizes. NxRepair does not require a reference genome and can be used with assemblies prepared with a single mate pair library alone. It specifically targets the most serious scaffolding errors and large-scale indels by identifying regions with a high number of anomalous insert sizes, or very few supporting reads, breaking the scaffold and optionally trimming out the misassembled region. NxRepair also provides a fine-grained quality score, allowing researchers to visualise poor-quality regions.

In the following sections, we describe the theoretical basis of NxRepair and its efficient implementation using an interval tree data structure. We then demonstrate the use of NxRepair on bacterial genomes assembled from a single Nextera mate pair library. For the *de novo* assembly process, we use the state of the art SPAdes assembler [165], which

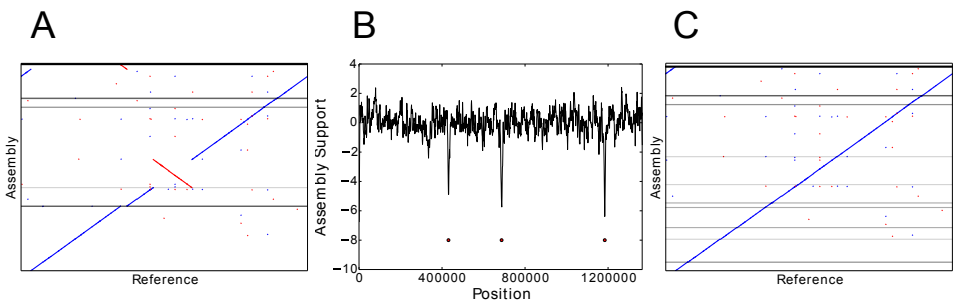

Figure 5.6: Using NxRepair to remove large misassemblies. (A) Alignment of the *de novo* assembly of the M. tuberculosis genome to its reference genome. The assembly contains several large misassemblies. (B) A plot of NxRepair's support metric against scaffold position for the assembly. Low support for the assembly is identified in three regions of a contig. (C) Breaking the contigs at the identified positions resolves the most significant misassemblies. In (A) and (C), horizontal lines demarcate the scaffold boundaries.

explicitly uses insert size information during contig construction, as well as for scaffolding. Using these genomes, we benchmark NxRepair against the error correction module of the A5 assembler, A5qc [176], which is currently the most widely used error correction tool, demonstrating NxRepair's superior performance. We conclude the chapter with a discussion of the performance and limitations of NxRepair as a quality control tool and discuss possible extensions that could improve performance.

5.3 Theory

5.3.1 Statistical Analysis of Mate Pair Insert Sizes

Nextera mate pair libraries are prepared to have a certain insert size, typically between 1 and 10 kb. Following construction of a *de novo* assembly, it is possible to align the mate pair reads used to construct the assembly back to the assembly itself. The actual distance along the *de novo* assembly between the two reads that form a mate pair can then be calculated. Large-scale errors in the assembly will cause mate pair reads aligning near to the error to have unusual insert sizes and read orientations. This means that we can use the insert size distribution displayed by mate pair reads aligning at different sites along the assembly to evaluate the local assembly quality. We identify assembly errors at sites where the local insert size distribution deviates too far from that observed in the rest of the genome.

We model the observed insert sizes of mate pairs aligned to the *de novo* assembly using a two-component mixture distribution. The first component of this mixture is the insert size distribution of correctly aligned mate pairs. We model the distribution of insert sizes, Y, as a normal distribution with mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$: $Y \sim N(\hat{\mu}, \hat{\sigma}^2)$. We estimate $\hat{\mu}$ and $\hat{\sigma}$ for the entire genome by aligning reads back to the assembly and using robust estimators (see below). The second component, defined as a uniform distribution across the contig size $U(0, L)$ for a contig of length L, captures anomalous insert sizes.

To calculate the degree of support for the assembly at each site across a contig, NxRepair retrieves all mate pairs spanning the region $[i - W, i + W]$, of size $2W - 1$ at position i on the contig, where spanning is defined to mean that one read ends entirely before the region $[i - W, i + W]$ and the other read begins entirely after (see Fig. 5.7.) The default value of W is 200 bases (see Table 5.1).

A uniform distribution was selected to model anomalous insert sizes, as it makes no assumption about the cause of an anomaly. It is uniform over the contig length, L, as opposed to over all possible sites in the assembly, as only pairs where both members align fully to that contig are considered. Similarly, even though the insert size distribution for correctly aligned mate pairs will typically display a longer tail than the normal distribution (unless a gel-extraction protocol is used), we found that using a normal distribution to model correct insert sizes did not adversely affect NxRepair's error detection. This is because each site is spanned by many mate pairs and the insert size of correctly aligned mate pairs is not

correlated with location. Consequently, despite these assumptions, the small fraction of correctly aligned mate pairs with a very large insert size do not lead to false positives in error identification.

We define a latent indicator variable $X_l \in \{0, 1\}$ for each pair of reads, l, which takes the value 1 if the insert size came from the null distribution, and 0 otherwise. Within each window queried, the probability that each retrieved read, r_l is drawn from the null distribution is given by:

$$P(X_l = x|Y_l) \quad = \quad \frac{\pi_x(Y_l|X_l = x)}{\sum_{k=0}^{1} \pi_k(Y_l|X_l = k)} \tag{5.1}$$

where Y_l is the insert size of read pair l, π_k is the user defined prior probability of class k and $\pi_1 + \pi_0 = 1$. The default value of π_0 is 0.01 (see Table 5.1), meaning that in the absence of any insert size information, 99 % of read pairs are expected to arise from the null distribution.

Within each window, the total support for a correct assembly at position i can be calculated as:

$$D_i = \sum_{l=1}^{N} P(X_l = 1|Y_l) \cdot C_l \tag{5.2}$$

where the summation is over all read pairs aligning across position i and C_l is an indicator variable, reporting pairing orientation:

$$C_l = \begin{cases} 1, & \text{if mate pairs have correct orientation and strand alignment} \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

Within each contig, the contig assembly support mean μ_D and variance s_D are calculated from all reads aligning to the contig,

$$\hat{\mu}_D = \frac{\sum_{l=1}^{N} D_l}{N} \qquad s_D = \frac{\sum_{l=1}^{N} \sqrt{(D_l - \hat{\mu})^2}}{N} \tag{5.4}$$

We use these contig specific mean and variance, rather than the global values, to prevent local variations in coverage from either causing false positives or masking changes in the

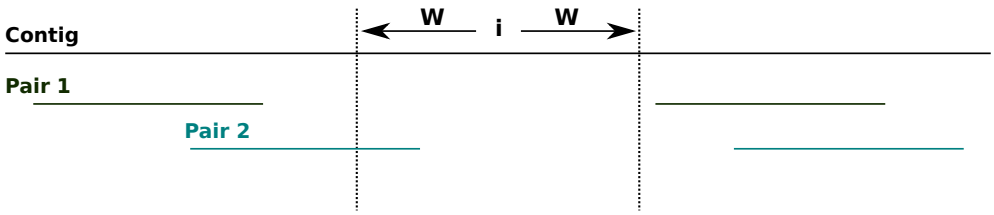

Figure 5.7: Schematic illustrating mate pairs spanning a selected region. The upper read pair (green) spans the indicated region, as both reads align entirely outside of the region $[i-W, i+W]$. The lower read pair (cyan) do not span the indicated region, as the left-hand read overlaps the target area.

insert size distribution. Using these values, we calculate the Z-score z_l within each queried interval as:

$$z_l = \frac{D_l - \hat{\mu}_D}{s_D} \tag{5.5}$$

The Z-score is sensitive both to local changes in the insert size distribution, and to large variations in the number of correctly aligned mate pairs, for example caused by a large number of reads with a mate aligning to a different contig. This ensures that NxRepair can identify misassemblies occurring both within and between contigs.

A misassembly is identified if $z_l < T$ for a user-defined threshold T (default value -4). This threshold describes the number of standard deviations below the mean assembly support that is required to identify an anomaly. The default value of -4 will flag only positions whose assembly support is less than four standard deviations below the mean level of support.

5.3.2 Global Assembly Parameters

NxRepair identifies misassemblies by identifying regions where the mate pair insert size distribution differs significantly from the insert size distribution across the entirety of the *de novo* assembly. Consequently, it is necessary to have a robust estimate of the global mate pair insert size distribution. For calculation of population statistics, mate pairs that align to different contigs are excluded, as are mate pairs with an incorrect strand or pairing orientation and pairs whose insert size exceed 30 kb (approximately 10 times the mean insert size). Pairs whose mapping quality falls below a user specified threshold (minmapq, default value 40) are also excluded, removing reads that are not uniquely mapped from the

calculation of global parameters. The global mean $\hat{\mu}$ and median absolute deviation (MAD) are calculated across all contigs in the assembly as:

$$\hat{\mu} = \frac{\sum_{l=1}^{N} Y_l}{N} \qquad \text{MAD} = \text{median}_m(|Y_m - \text{median}_l(Y_l)|) \tag{5.6}$$

where $\text{median}_l(Y_l)$ is the median insert size of reads with correct pairing behaviour and $|Y_m - \text{median}_l(Y_l)|$ is the absolute value of the residual from the median of the mth on N reads. The standard deviation was then calculated by scaling the MAD using:

$$\hat{\sigma} = K \cdot \text{MAD} \tag{5.7}$$

where

$$K = \frac{1}{\frac{3}{4} \cdot \phi^{-1}} \sim 1.4826 \tag{5.8}$$

and ϕ^{-1} is the quantile function [180]

The MAD is a robust estimator for the standard deviation, as it is not sensitive to outliers, such as the long tail of the mate pair insert size distribution. Using the MAD as an estimator prevents over-estimating the variance of the insert size, allowing anomalously large insert sizes to be correctly identified.

These were then used as the parameters of the null distribution, as described above (Section 5.3.1.

5.3.3 Interval Tree Construction

To facilitate rapid lookup of mate pair properties, we construct an interval tree [181] for each contig in the *de novo* assembly. An interval tree is a data structure that facilitates $O(\log n + m)$ lookup of intervals that span a given point or interval, for n total entries and m spanning entries. The interval tree contains the start and end positions of each mate pair aligned to that contig, as well as a flag variable indicating whether that mate pair had correct strand and pairing orientation. Mate pairs where the two reads align to different contigs were excluded. Mapping quality is currently not considered at this stage - reads are retained regardless of mapq score. For each position i for which the Z-score is to be

calculated, the tree is queried with a start position $i - W$ and end position $i + W$, to retrieve read pairs spanning the interval between positions $i - W$ and $i + W$ (exclusive). The insert sizes of retrieved read pairs are then used to calculate the Z-score for position i. This allows NxRepair to rapidly query positions across a contig to discover the insert size distribution at the queried position. Use of the interval tree significantly increases the efficiency of Z-score calculation, as each pair of reads is fetched only once in order to build the tree. All relevant parameters are then stored in the tree for rapid look-up when a position is queried. This has several advantages. Firstly, it is significantly faster than fetching reads only when a position is queried. Secondly, it is more space efficient than a frequency array of all positions on all contigs but does not lose any information about the exact alignment positions. Finally, once construction of the tree is complete, multiple passes across a contig (for example with different spatial resolutions, or using different window sizes) can rapidly be made using the same tree.

5.3.4 Misassembly Location and Contig Breaking

To improve the quality of the *de novo* assembly, once a misassembly has been located, a contig is broken into two separate pieces at the site of a misassembly. The broken ends of the two new contigs can optionally be trimmed by a user defined length (default 4 kb) to remove the misassembled region. Trimming allows removal of the incorrectly assembled regions around a break-point, but can be switched off if a user does not want any sequence to be removed from the assembly. To prevent excessive clipping, misassemblies separated by less than the trimming distance are grouped together, the contig is broken at the start and end of the misassembled region and the misassembled section is discarded. Low-scoring regions within the trimming distance of the ends of contigs are not considered misassemblies, as the high proportion of mate pairs aligning here whose mate maps to a different contig reduces the number of pairs under consideration and hence lowers the observed Z-score. This also ensures that circular molecules, such as small plasmids, which are assembled into a single contig, are not truncated because of mate pairs at either end of the assembly that appear to span the entire contig, but which are spatially close when circularisation is considered.

5.3.5 Availability and Dependencies

NxRepair is available for free anonymous download from the Python Package Index (PyPI) here: `https://pypi.python.org/pypi/nxrepair`. The source code, written in python is hosted on GitHub: `https://github.com/rebeccaroisin/nxrepair`. A full tutorial and API can be found on ReadTheDocs: `http://nxrepair.readthedocs.org/en/latest/`.

NxRepair makes use of several further open source libraries, specifically:

Numpy [64] (`http://www.numpy.org/`)

Scipy [182] (`http://www.scipy.org/`)

Matplotlib [66] (`http://matplotlib.org/`)

Pysam (`https://pypi.python.org/pypi/pysam`), the python wrapper for Samtools

Samtools [183] (`http://samtools.sourceforge.net/`)

We installed the numpy, scipy and matplotlib libraries via Anaconda (`https://store.continuum.io/cshop/anaconda/`).

We have used the Interval Tree implementation from the bx-python library (`https://bitbucket.org/james_taylor/bx-python/wiki/Home`).

5.4 Experimental Methods

5.4.1 Data

Preparation of the genome libraries was carried out by Emma Carlson and Niall Gormley at Illumina Cambridge.

Nine bacterial genomes were prepared according to the Nextera mate pair protocol and sequenced in duplicate in a single MiSeq run using 2×151 bp reads. The organisms sequenced are shown in Table 5.4. Reads were trimmed using the MiSeq inbuilt trimmer. Table 5.5 gives an overview of sequencing yield, mean quality and read length after trimming. The untrimmed reads are available from BaseSpace via `https://basespace.illumina.com/s/TXv32Ve6wTl9`. In addition, the trimmed reads are available at the European Nucleotide Archive (ENA) at `http://www.ebi.ac.uk/ena/data/view/PRJEB8559`. Note that only

these Nextera mate pair libraries were used. No additional single end or paired end libraries were required. For performance optimisation, the first replicate from each genome sequenced was used as a training set. The test set, for performance evaluation, was formed from the second replicate of each genome.

5.4.2 Performance Optimisation

ROC Plots To optimise the threshold in Z below which to identify a misassembled region, we prepared ROC plots using Replicate 1 of each genome, varying the threshold value, T, in steps of 1 between -10 and 0.

The positions of true misassemblies were identified by aligning each *de novo* assembly to its reference genome using QUAST [171]. To correctly compare the sites of true misassemblies with those identified by NxRepair, we divided each contig of the assembly into short stretches of 1 kb length. We then prepared an array, A_{Nx} of size $\frac{L}{1000}$ for contig length L, corresponding to misassemblies identified by NxRepair. A_{Nx} was filled as follows:

$$A_{Nx} = \begin{cases} 1, & \text{if NxRepair identified a misassembly in stretch } i \\ 0, & \text{otherwise} \end{cases} \tag{5.9}$$

To prepare the ROCs, each position i in A_{Nx} was labeled as true positive (TP) if $A_{Nx}[i] = 1$ and a true misassembly fell within it, true negative (TN) if $A_{Nx}[i] = 0$ and no true misassembly occurred within the interval, false positive (FP) if $A_{Nx}[i] = 1$ but no true misassembly had occurred, or false negative (FN) if $A_{Nx}[i] = 0$ but the interval contained a true misassembly. The 1 kb interval used was the same interval used in error identification, ensuring that the resolution of the evaluation matched the error detection resolution. The true positive rate (TPR) and false positive rate (FPR) were then calculated as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{5.10}$$

Based on the resultant ROC plots, shown in Fig. 5.8, a threshold in Z of -4 was found to detect true misassemblies with minimal false positives, so was used for all subsequent analyses.

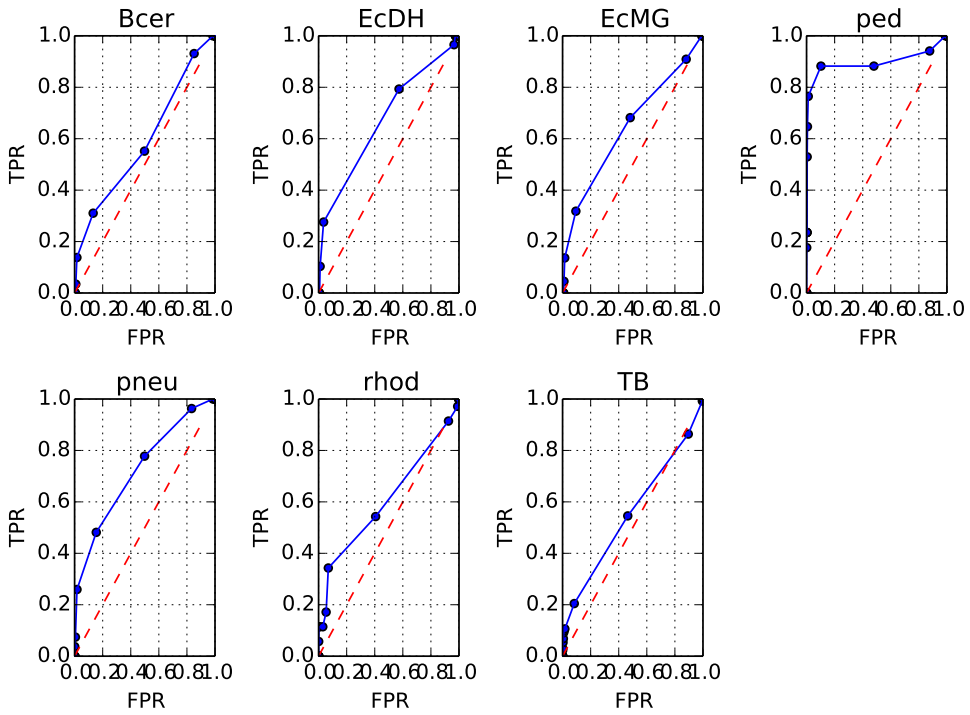

Figure 5.8: ROC plots for the seven genomes containing misassemblies.

Profiling Performance analysis was performed on a single core with 8 GB RAM available. Runtime analysis was performed using the python cProfile module. The memoryprofiler python module was used to analyse memory usage.

5.4.3 Workflow Pipeline

De novo assemblies were prepared using the SPAdes Assembler, version 3.1.1 [165]:

```
spades.py -k 21,33,55,77 -t 4
--hqmp1-12 bacteria.fastq.gz --hqmp1-fr -o assembly
```

The initial assembly quality was evaluated using QUAST [171] (version 2.3) to align the *de novo* assembly to a reference genome:

```
python quast.py -o results_sample -t 16
-R ref/reference.fna sample_new.fasta
```

Following assembly, the same reads used to generate the assembly were aligned back to the *de novo* assembly using BWA-MEM [184] (BWA version 0.7.10). A sorted BAM file of the

resulting alignment was then prepared using SAMtools (version 1.1) [183]:

```
bwa index sample/scaffolds.fasta
```

```
bwa mem sample/scaffolds.fasta -p bacteria.fastq.gz | samtools view -bS - | samtools sort - sample
```

```
samtools index sample.bam
```

We identified misassemblies using NxRepair (version 0.13) as follows:

```
python nxrepair.py sample.bam sample/scaffolds.fasta sample_scores.csv sample_new.fasta -img_name sample_new
```

The default parameters used and their meanings are shown in Table 5.1. These have been optimised for Illumina Nextera mate pair libraries with a mean insert size of approximately 4 kb. For mate pair libraries with a much larger (smaller) insert size, the maxinsert and trim parameters may need to be increased (decreased).

Finally we used QUAST [171] to evaluate the assembly quality following NxRepair by aligning the *de novo* assembly to a reference genome as described above.

5.5 Results

We used NxRepair to correct *de novo* assemblies from Replicate 2 of each of the nine bacterial genomes described above. Mate pair reads were trimmed, assembled using the SPAdes assembler (version 3.1.1) [165] and then aligned back to the assembled scaffold using BWA-MEM [184]. We used QUAST [171] to evaluate the assembly quality before and after NxRepair correction by aligning to an appropriate reference genome. For all NxRepair analyses, the default parameters, shown in Table 5.1 were used. Fig. 5.6 (A) shows a misassembled genome that contained several scaffolding errors identified by NxRepair (Fig. 5.6 (B)). Following NxRepair correction, the most significant structural misassemblies were resolved (Fig. 5.6 (C)). The improvement following NxRepair correction is shown for all nine genomes in Table 5.2 (middle column). For two assemblies, errors were removed without reducing NGA50; for one genome, errors were removed but NGA50 was slightly reduced; for six genomes, three of which contained no large errors, no errors were found and the assembly was unchanged. We are not able to correct all misassemblies, as not all misassemblies exhibit

Table 5.1: NxRepair Parameters

Parameter	Default Value	Meaning
imgname	None	Prefix under which to save plots.
maxinsert	30000	Maximum insert size, below which a read pair is included in calculating population statistics.
minmapq	40	Minimum MapQ value, above which a read pair is included in calculating population statistics.
minsize	10000	Minimum contig size to analyse.
prior	0.01	Prior probability that the insert size is anomalous.
stepsize	1000	Step-size in bases to traverse contigs.
trim	4000	Number of bases to trim from each side of an identified misassembly.
T	-4.0	Threshold in Z score (number of standard deviations from the mean) below which a misassembly is called.
window	200	Window size across which bridging mate pairs are evaluated.

a change in Z-score large enough to identify an error against the background score fluctuation caused by the wide insert size distribution of the Nextera mate pairs.

To benchmark NxRepair's performance, we also attempted to identify assembly errors using the A5qc error correction module of the A5 Assembly pipeline [176]. The results are shown in Table 5.2 (right hand column). For eight of the nine genomes evaluated, A5qc was unable to detect any errors. For the final genome (K. pneumoniae), A5qc did detect errors, but the contig-breaking process left Quast unable to align the resultant assembly to the reference genome. Re-scaffolding using the A5 scaffolder did allow reference alignment, but the assembly contained more misassemblies (15) than the original assembly. Consequently, we are confident that, for these high-quality alignments, NxRepair is a superior error-detection tool.

Despite this good performance, it is clear that NxRepair is not able to find all misassemblies present. There are several reasons for this. Firstly, NxRepair's resolution is limited to relatively large-scale errors, as a very large disruption in mate pair insert sizes over a region of approximately 1 kb to significantly reduce the Z-score. Consequently, indel errors with a displacement smaller than 1 kb will not be detected. Secondly, NxRepair is limited by the intrinsic error rate of the mate pair library used. If the insert size distribution has a very wide variance, large fluctuations caused by assembly errors will be masked, making error correction more challenging.

A number of improvements to NxRepair might mitigate these issues. Firstly, NxRepair currently identifies errors using a simple threshold applied to the total assembly support from spanning mate pairs, D. A more rigorous approach would implement a fully probabilistic method of error detection, using the distribution of spanning mate pair insert sizes to evaluate the relative probability of an error.

Furthermore, NxRepair currently uses a user-defined prior probability of incorrect pairing and uses some simplistic thresholding to determine the parameters of the global insert size distribution. It would be possible to implement simultaneous co-estimatation of the mate pair error rate and the insert size distribution. In addition to relieving the user of estimating the error rate of their mate pair library, this would improve the accuracy of parameter estimation for the correct mate pairs, particularly for libraries with a large mate pair error rate. However, this would not necessarily translate into improved accuracy of error detection, as noise from the mate pair library would still mask true errors.

Table 5.2: Number of large misassemblies and NGA50 as reported by QUAST before (left) and after correction by NxRepair (middle) and A5qc (right).

		Before NxRepair		**After NxRepair**		After A5qc	
Genome	Genome size	No.	NGA50	No.	NGA50	No.	NGA50
B. cereus ATCC 10987	5,432,652	0	1,157,846	0	1,157,846	0	1,157,846
E. coli K-12 substr. DH10B	4,686,137	7	573,003	6	573,003	7	573,003
E. coli K-12 substr. MG1655	4,641,652	3	693,692	3	693,692	3	693,692
L. monocytogenes EGDe	2,944,528	0	1,496,613	0	1,496,613	0	1,496,613
M. ruber DSM 2366	4,839,203	0	2,702,549	0	2,702,549	0	2,702,549
P. heparinus DSM 2366	5,167,383	1	1,269,147	0	952,558	1	1,269,147
K. pneumoniae MGH 78578	5,694,894	8	578,813	8	578,813	-	-
R. sphaeroides 2.4.1	4,602,977	8	2,715,434	8	2,715,434	8	2,715,434
M. tuberculosis H37Ra	4,411,532	63	186,136	57	186,136	63	186,136

5.5.1 Performance

We evaluated the runtime and peak memory usage of NxRepair on each of the nine genomes analysed. The results are shown in Table 5.3. The most memory and computationally intensive part of the NxRepair analysis is construction of the interval trees. The size of each interval tree is dependent on the contig size. Consequently, we expect both runtime and memory usage to scale with the size of the largest contig.

5.6 Conclusions

NxRepair is a simple error correction module that can be used to rapidly identify and remove large scale errors from *de novo* assemblies using Nextera mate pair reads. We evaluated NxRepair using *de novo* assemblies of nine bacterial genomes prepared using the SPAdes assembler, showing that of the six genomes containing misassemblies, three could be improved by NxRepair correction; compared with no improvements made by the A5qc module. SPAdes is the current state of the art in bacterial genome assembly and explicitly uses mate pair information during both contig construction and scaffolding. Even in these excellent assemblies, NxRepair could identify misassemblies and improve the assembly quality. We predict that NxRepair will be even more useful for identifying errors in *de novo* assemblies where mate pair information was used only at the scaffolding stage.

Table 5.3: NxRepair performance analysis.

Bacterium	Total Time (s)	Memory Usage (MiB)
B. cereus ATCC 10987	78	271
E. coli K-12 substr. DH10B	123	444
E. coli K-12 substr. MG1655	70	260
L. monocytogenes EGDe	97	383
M. ruber DSM 2366	259	565
P. heparinus DSM 2366	123	417
K. pneumoniae MGH 78578	59	227
R. sphaeroides 2.4.1	190	463
M. tuberculosis H37RaTB	155	411

NxRepair is freely available online. It can be downloaded from the Python Package Index (`https://pypi.python.org/pypi/nxrepair`) and run with a single call from the command line, making it an attractive option for fast evaluations of and improvements to assembly quality. The source code is available on GitHub (`https://github.com/rebeccaroisin/nxrepair`), facilitating easy incorporation into user assembly pipelines.

Table 5.4: Summary of bacteria analysed and the relevant NCBI information on their reference genomes. There were two repeats of each strain. All 18 samples were prepared with the Nextera mate pair protocol and sequenced in a single MiSeq run using 2×151 bp reads. The untrimmed reads we used as input to NxTrim (3.9Gbp in all) are available from BaseSpace via `https://basespace.illumina.com/s/TXv32Ve6wTl9`.

Abbreviation:	Bcer
Bacteria:	*Bacillus cereus ATCC 10987*
Accession ID:	NC_003909, NC_005707
NCBI FTP:	`ftp.ncbi.nih.gov/genomes/Bacteria/Bacillus_cereus_ATCC_10987_uid57673/`
Abbreviation:	EcDH
Bacteria:	*Escherichia coli str. K-12 substr. DH10B*
Accession ID:	NC_010473
NCBI FTP:	`ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__DH10B_uid58979/`
Abbreviation:	EcMG
Bacteria:	*Escherichia coli str. K-12 substr. MG1655*
Accession ID:	NC_000913
NCBI FTP:	`ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/`
Abbreviation:	list
Bacteria:	*Listeria monocytogenes*
Accession ID:	NC_003210
NCBI FTP:	`ftp.ncbi.nih.gov/genomes/Bacteria/Listeria_monocytogenes_EGD_e_uid61583/`
Abbreviation:	meio
Bacteria:	*Meiothermus ruber DSM 1279*
Accession ID:	NC_013946
NCBI FTP:	`ftp.ncbi.nih.gov/genomes/Bacteria/Meiothermus_ruber_DSM_1279_uid46661/`
Abbreviation:	ped
Bacteria:	*Pedobacter heparinus DSM 2366*
Accession ID:	NC_013061
NCBI FTP:	`ftp.ncbi.nih.gov/genomes/Bacteria/Pedobacter_heparinus_DSM_2366_uid59111/`
Abbreviation:	pneu
Bacteria:	*Klebsiella pneumoniae subsp. pneumoniae MGH 78578*
Accession ID:	NC_009648, NC_009649, NC_009650, NC_009651, NC_009652, NC_009653
NCBI FTP:	`ftp.ncbi.nih.gov/genomes/Bacteria/Klebsiella_pneumoniae_MGH_78578_uid57619/`
Abbreviation:	rhod
Bacteria:	*Rhodobacter sphaeroides 2.4.1*
Accession ID:	NC_007488, NC_007489, NC_007490, NC_007493, NC_007494, NC_009007, NC_009008
NCBI FTP:	`ftp.ncbi.nih.gov/genomes/Bacteria/Rhodobacter_sphaeroides_2_4_1_uid57653/`
Abbreviation:	TB
Bacteria:	*Mycobacterium tuberculosis H37Ra*
Accession ID:	NC_009525
NCBI FTP:	`ftp.ncbi.nih.gov/genomes/Bacteria/Mycobacterium_tuberculosis_H37Ra_uid58853/`

Table 5.5: Yield in bp, mean base quality and average read length after adapter removal, as well as the raw NGA50 score prior to NxRepair analysis for all genomes analysed.

Genome	Yield [bp]	Mean base quality	Mean read length	Raw NGA50
B. cereus lib1	140,034,231	30.83	120.50	1,157,404
B. cereus lib2	150,883,336	31.69	124.08	1,157,846
E. coli DH10B lib1	229,164,175	31.50	127.55	576,143
E. coli DH10B lib2	167,955,255	31.19	126.49	573,003
E. coli MG1655 lib1	138,893,204	NaN	104.56	640,732
E. coli MG1655 lib2	164,490,239	31.93	129.67	693,692
L. monocytogenes lib1	197,796,210	32.66	129.79	1,496,615
L. monocytogenes lib2	161,114,700	31.81	125.79	1,496,613
M. ruber lib1	180,542,545	29.96	123.69	3,095,733
M. ruber lib2	150,298,958	31.09	129.40	2,702,549
P. heparinus lib1	186,070,764	32.21	127.30	1,269,259
P. heparinus lib2	146,448,694	31.32	124.00	1,269,147
K. pneumoniae lib1	182,614,602	31.86	131.70	577,220
K. pneumoniae lib2	166,306,322	31.82	130.28	578,813
R. sphaeroides lib1	184,138,610	30.08	127.99	3,181,390
R. sphaeroides lib2	210,961,284	30.12	129.79	2,715,434
M. tuberculosis lib1	211,892,634	30.43	127.37	184,170
M. tuberculosis lib2	177,615,358	30.06	126.82	186,136

Chapter 6

Conclusions and Future Work

6.1 General Conclusions

This thesis presents three years' work on understanding and improving the analysis tools used in confocal smFRET experiments. This work has involved both implementing and benchmarking existing techniques, as well as implementing and thoroughly evaluating novel methodologies. We hope that this work provides a strong foundation for rigorous work by other researchers in this field.

6.1.1 pyFRET

Chapter 2 described the development and release of an open source software library for the analysis of confocal smFRET data. For researchers working in different research environments to be able to cross-compare their results effectively, it is necessary that they have an effective method to describe and share their analysis tools and to select the best analysis tools for their specific research problem. Furthermore, smFRET data is quite sensitive to the analysis methods used. As we have shown, both the specific analysis tools used and the method of their usage can quite significantly influence the appearance of the processed data and hence the outcome of donwstream analyses. Unfortunately, to date there has been no culture of publishing the code or smFRET data analysis tools, which has made it difficult for researchers to compare the effectiveness and efficiency of different analysis methods.

To overcome these challenges, we have developed and released pyFRET, a software library

that implements many of the main methodologies for analysis of smFRET data. pyFRET is open source and fully documented. This provides both a toolkit of analysis methods that can be freely accessed by other researchers and a platform for researchers to contribute and benchmark novel analysis methods. We also describe a rigorous comparison of different data collection and analysis methods for confocal smFRET data. The evaluation presented here identifies the clear superiority of alternating excitation schemes that provide direct access to acceptor excitation information, enabling thresholding-based event selection to be carried out without bias. We show that without direct access to information about acceptor emission, the more complex APBS and DCBS burst search algorithms display biases comparable to those seen from the simple AND and SUM thresholding algorithms, despite significantly increased space requirements and computational complexity. We hope that this work will facilitate research by other scientists, who can now quickly and effectively select the correct analysis tool for their data, as well as providing a platform for the publication and benchmarking of novel analysis methods.

6.1.2 Inference Analysis of smFRET Data

Thresholding-based event selection for confocal smFRET data necessarily discards a significant fraction of the smFRET dataset, including not only noise but also a large fraction of fluorescence emission events. Furthermore, as we show in Chapter 2, in the absence of direct information about acceptor emisson behaviour, thresholding criteria are biased and do not select a random subset of fluorescent events, resulting in distortion of downstream analyses. To overcome these challenges we developed a novel analysis method, based on model-based Bayesian inference, that can simultaneously co-infer the values of all parameters of interest in a confocal smFRET dataset.

Chapter 3 presents the development and evaluation of this novel tool. Our evaluation shows that model-based inference is an effective tool for analysis of smFRET data and that we can accurately infer the concentrations and intramolecular distances of single fluorescent populations or of a mixture of two fluorescent populations. The full source code for our inference analysis is freely available online. This work provides a strong foundation for the application of model-based inference to a wider range of smFRET datasets and demonstrates the proof of concept that model-based inference is both feasible and fruitful as a method of data analysis for this research field.

6.1.3 Inference Analysis of Oligomer Sizing

Chapter 4 extends the concept of model-based inference and applies this form of analysis to the problem of accurately sizing heterogeneous mixtures of fluorescently labelled oligomers. A significant research effort has been made in the Klenerman group to develop experimental methodologies to allow accurate characterisation of the oligomerisation reactions observed in amyloidosis. The data analysis side of this effort has been somewhat neglected, allowing unfounded assumptions about the data collected to significantly reduce the accuracy of calculated oligomer size distributions.

The work presented in this chapter demonstrates, through a comparison of simulated datasets and data from simple, well-controlled experiments, that the heterogeneity observed in photon emission is not dominated, as would be hoped, by the number of fluorophores attached to a molecule. Instead, underlying photophysical processes, including, but not limited to, photobleaching, photoblinking, and the diffusion pathway of molecules through the confocal volume, have a more significant effect on the number of photons observed. Although this research was not able to provide a greatly improved analysis technique, we hope that our thorough evaluation of the challenges encountered in attempting to accuarately determine oligomer sizes, even in well-characterised, homogeneous samples, demonstrates the importance of thoroughly examining all assumptions implicit in any method of data analysis.

6.1.4 NxRepair

The final results chapter, Chapter 5, presented work performed in a different research field; namely the development of NxRepair, an error correction tool for *de novo* assemblies of bacterial genomes assembled from Illumina's Nextera nate pair reads. Although long-read technologies are being developed, these are either still experimental or prohibitively expensive for normal use. Consequently, the challenge of accurately reconstructing the sequence of an entire genome from short read sequences remains current.

The work presented in this chapter demonstrates a novel method for error identification and error correction based on probabilistic analysis of the insert size distribution when mate pair reads are aligned back to the *de novo* assembly. NxRepair is available online for free download and the source code is hosted on the popular code repository GitHub. This chapter presents a comprehensive evaluation of the performcance of NxRepair using multiple bacterial genomes,

demonstrating its superior performance when compared to the most popular existing error correction tool.

6.2 Applications and Future Work

The work presented in this thesis, and summarized above, provides a stable foundation for further work to consolidate and improve the analysis of smFRET data. Some of the results presented have encouraged experimental method development to allow more rigorous analysis of smFRET data. To conclude this thesis, we now summarize some of the applications and extensions that are being considered as a result of the work described here.

6.2.1 Open Source Software for smFRET

pyFRET, our open source library of data analysis tools for smFRET currently provides sufficient functionality for a researcher to perform a full analysis of smFRET data collected from a range of different experiment types using a number of different methods. As the source code that implements the analysis tools, as well as the tools themselves, is freely available online it is straightforward for researchers to understand how the analysis works and to modify the tools as they require. This prevents new users of smFRET technology from needing to develop their analysis tools from scratch in order to be able to analsye their data.

Furthermore, as experimental technologies develop and improve, it is to be expected that new analysis tools will also be required. pyFRET provides a central repository to the community to which new techniques can be added, to allow easy access and evaluations; and against which new techniques can be benchmarked; enabling researchers to make fully informed descisions about which analysis tools and experiemental methodologies to use in their research.

pyFRET already has users outside of the Klenerman research group and we have responded to requests for additional features. There is clearly considerable work to do to allow pyFRET to become useful to a wider swathe of the smFRET research community. However, we hope that its existence encourages a more open, collaborative approach to smFRET data analysis and method development.

6.2.2 Inference Analysis of smFRET Data

The work on inference analysis on smFRET data that we have presented here clearly shows the feasibility of inference analysis as a method of analysing smFRET data. It also highlights considerable extensions and improvements to the inference method as presented here, the implementation of which would considerably increase the utility of inference as an analysis tool. The two most important of these are:

1. The implementation of revesible jump Monte Carlo inference [105].
2. The release of the inference analysis as a software package.

In its current implementation, the inference analysis cannot infer the number of fluorescent populations in a mixture; the number of populations must be given. This is a considerable limitation, as the number of fluorescent populations is frequently unknown. At present, determining the number of fluorescent populations can only be carried out by performing inference multiple times, modifying the number of expected fluorescent populations, and then manually inspecting the results obtained. Implementation of reversible jump Monte Carlo sampling would make this model selection step part of the inference analysis and would greatly increase the attraction of inference as a method of smFRET data analysis.

Similarly, although the source code for the inference analysis can be viewed and downloaded from an open source code repository, the code is not packaged or documented in a manner that makes it easy to use. Packaging the analysis code as a software package, for example by incorporating it into the pyFRET framework, would make this analysis tool more accessible to other researchers. Making both of these improvements is likely to significantly improve the reception of inference analysis by smFRET researchers.

6.2.3 Accurate Sizing of Fluorescent Oligomers

Although the work on accurate oligomer sizing using inference methods was the least successful piece of research presented in this thesis in terms of successful method development, it is the work that has presented the greatest number of opportunities for further work. The results that we present here conclusively demonstrate the role of both photophysics and the underlying experimental methodology in preventing accurate assessment of oligomer size distributions based on fluorescence data. These conclusions have led to significant re-evaluation

of the experimental methods used and to the proposal of several modifications to the research methodology that may enable a more quantitative evaluation of oligomer sizes.

Specifically, the Klenerman group is currently constructing a novel confocal microscope that uses acousto-optic modulation to allow uniform excitation of the confocal volume in the x-dimension. By flowing labelled molecules rapidly across this region of uniform excitation using nano-fluidic channels that confine the molecules to a narrow region in the y-dimension, we hope to be able to reduce the experimental sources of emission heterogeneity. Similarly, we are currently evaluating fluorophore performance, to find commercially available structures that are more robust to photobleaching and photoblinking behaviour. A thorough evaluation of the new experimental set-up is currently in progress and we are hopeful that it may lead to a more quantitative experimental methodology for determining sizes and size distributions.

6.2.4 Error Detection in *de novo* Assemblies

The work presented on error correction in *de novo* assemblies is predominantly self contained. We developed and released an open-source error correction tool, which is currently used by bioinformaticians at Illumina; an evaluation of the tool's performance has been submitted for peer reviewed publication. Extensions of this tool include modifications to enable larger genome assemblies to be evaluated and to allow evaluation of *de novo* assemblies prepared using alternative sequencing methodologies, such as the long-reads from Nanopore sequencing that can be "threaded" through a de Bruijn graph to resolve ambiguities [185]. Modification of the tool to enable chaining with other assembly and quality control tools is also possible. Implementation of these additions depends on demand and uptake from other bioinformaticians. We hope that NxRepair finds an interested user-base within the bioinformatics research community.

Bibliography

[1] T. Ha, T. Enderle, D. F. Ogletree, D. S. Chemla, P. R. Selvin, and S. Weiss. Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci. USA*, 93(13):6264–6268, JUN 25 1996.

[2] G Haran. Single-molecule fluorescence spectroscopy of biomolecular folding. *J. Phys.: Condens. Matter*, 15(32):R1291–R1317, AUG 20 2003.

[3] B. Schuler, E. A. Lipman, and W. A. Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419(6908):743–747, OCT 17 2002.

[4] S Weiss. Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nat. Struct. Mol. Biol*, 7(9):724–729, SEP 2000.

[5] B. Valeur, editor. *Molecular Fluorescence: Principles and Applications*. Wiley-VCH, 2001.

[6] J. R. Albani. *Structure and Dynamics of Macromolecules: Absorption and Fluorescence Studies: Absorption and Fluorescence Studies*. Elselvier, Amsterdam, 2011.

[7] M. H. Horrocks. *Development of single-molecule techniques to study the aggregation of α-synuclein*. PhD thesis, University of Cambridge, 2014.

[8] Quanta red spectra. `http://tools.lifetechnologies.com/content/sfs/gallery/high/QuantaRed-Spectra.jpg`. Accessed: 2015-04-30.

[9] *IUPAC Compendium of Chemical Terminology*. IUPAC, 2006. Accessed: 2015-04-30.

[10] T. Forster. Zweischenmolekulare energiewanderung undfluoreszenz. *Annalen der Physik*, 2:55–75, 1948.

[11] A. P. Demchenko. *Development of single-molecule techniques to study the aggregation of α-synuclein*. Springer, 2008.

[12] L. Stryer and R. P. Haughland. Energy transfer: a spectroscopic ruler. *Proceedings of the National Academy of Sciences of the United States of America*, 58:719–726, 1967.

[13] X. Michalet, A. N. Kapanidis, T. Laurence, F. Pinaud, S. Doose, M. Pflughoefft, and S. Weiss. The power and prospects of fluorescence microscopies and spectroscopies. *Annu. Rev. Biophys. Biomol. Struct.*, 31:161–182, 2003.

[14] W. E. Moerner and D. P. Fromm. Methods of single-molecule fluorescence spectroscopy and microscopy. *Rev. Sci. Instrum.*, 74:3597–3619, 2003.

[15] M. S. Dillingham and M. I. Wallace. Protein modification for single molecule fluorescence microscopy. *Org. Biomol. Chem.*, 6:3031–3037, 2008.

[16] D. J. Webb and C. M. Brown. Epi-fluorescence microscopy. *Methods Mol Biol.*, 931:29–59, 2014.

[17] K. B. Wolf and G. Krotzsch. Geometry and dynamics in refracting systems. *Eur. J. Phys.*, 16:14–20, 1995.

[18] M. H. Horrocks, H. Li, J. Shim, R. T. Ranasinghe, R. W. Clarke, W. T. S. Huck, C. Abell, and D. Klenerman. Single molecule fluorescence under conditions of fast flow. *Anal. Chem.*, 84:179–185, 2012.

[19] C. M. Brown, R. B. Dalal, B. Hebert, M. A. Digman, A. R. Horwitz, and E. Gratton. Raster image correlation spectroscopy (rics) for measuring fast protein dynamics and concentrations with a commercial laser scanning confocal microscope. *J Microsc.*, 229:78–91, 2008.

[20] D. Magde, E. Elson, and W. Webb. Thermodynamic fluctuations in a reacting system: Measurement by fluorescence correlation spectroscopy. *Phys. Rev. Lett.*, 29:705–708, 1972.

[21] K. Bacia and P. Schwille. Practical guidelines for dual-color fluorescence cross-correlation spectroscopy. *Nat. Protoc.*, 2:2842–2856, 2007.

[22] K. M. Berland, P. T. C. So, Y. Chen, W. W. Mantulin, and E. Gratton. canning two-photon fluctuation correlation spectroscopy: particle counting measurements for detection of molecular aggregation. *Biophys. J.*, 71:410–420, 1996.

[23] L. O. Tjernberg, A. Pramanik, S. Bjorling, P. Thyberg, J. Thyberg, C. Nordstedt, K. D. Berndt, L. Terenius, and R. Rigler. Amyloid beta-peptide polymerization studied using fluorescence correlation spectroscopy. *Chem. Biol.*, 6:53–62, 1999.

[24] K. Bacia, S. A. Kim, and P. Schwille. Fluorescence cross-correlation spectroscopy in living cells. *Nat. Methods*, 3:83–89, 2006.

[25] W.-C. Sun, K. R. Gee, D. H. Klaubert, and R. P. Haugland. Synthesis of fluorinated fluoresceins. *J. Org. Chem.*, 62:6469–6475, 1997.

[26] R. Zondervan, F. Kulzer, S. B. Orlinskii, and M. Orrit. Photoblinking of rhodamine 6g in poly(vinyl alcohol): Radical dark state formed through the triplet. *J. Phys. Chem. A.*, 107:6770 – 6776, 2003.

[27] E. Nir, X. Michalet, K. M. Hamadani, T. A. Laurence, D. Neuhauser, Y. Kovchegov, and S. Weiss. Shot-noise limited single-molecule FRET histograms: Comparison between theory and experiments. *J. Phys. Chem. B*, 110(44):22103–22124, NOV 9 2006.

[28] C. Joo, Balci H., Y. Ishitsuka, C. Buranachai, and T. Ha. Advances in single-molecule fluorescence methods for molecular biology. *Annual Reviews of Biochemistry*, 77:51–76, 2008.

[29] N. G. Walter, C. Y. Huang, A. J. Manzo, and M. A. Sobhy. Do-it-yourself guide: how to use the modern single-molecule toolkit. *Nat Methods.*, 5:475–489, 2008.

[30] A. Orte, R. Clarke, S. Balasubramanian, and D. Klenerman. Determination of the fraction and stoichiometry of femtomolar levels of biomolecular complexes in an excess of monomer using single-molecule, two-colour coincidence detection. *Analytical Chemistry*, 78:7706–7715, 2006.

[31] A. Orte, R. Clarke, and D. Klenerman. Single-molecule two-colour coincidence detection to probe bimolecular associations. *Biochemical Society Transactions*, 38:914–918, 2010.

[32] N. Cremades, S. I. A Cohen, E. Deas, A. Y. Abramov, A. Y. Chen, A. Orte, M. Sandal, R. W. Clarke, P. Dunne, F. A. Aprile, C. W. Bertoncini, N. W. Wood, T. P. J. Knowles, C. M. Dobson, and D. Klenerman. Direct observation of the interconversion of normal and toxic forms of a-synuclein. *Cell*, 149:1048 – 1059, 2012.

[33] Ye Y., G. Blaser, M. H. Horrocks, M. J. Ruedas-Rama, S. Ibrahim, A. A. Zhukov, A. Orte, D. Klenerman, S. E. Jackson, and Komander D. Ubiquitin chain conformation regulates recognition and activity of interacting proteins. *Nature*, 492:266–270, 2012.

[34] A. Orte, R. W. Clarke, and D. Klenerman. Fluorescence coincidence spectroscopy for single-molecule fluorescence energy-transfer measurements. *Analytical Chemistry*, 80:8389–8397, 2008.

[35] A.N. Kapanidis, T.A. Laurence, N.K. Lee, E. Margeat, X. Kong, and S. Weiss. Alternating-laser excitation of single molecules. *Acc. Chem. Res.*, 38:532–533, 2005.

[36] S. Doose, M. Heilemann, X. Michalet, S. Weiss, and A. N. Kapanidis. Periodic acceptor excitation spectroscopy of single molecules. *Eur. Biophys. J.*, 36:669–674, 2006.

[37] M. H. Horrocks, H. Li, J. Shim, R. T. Ranasinghe, R. W. Clarke, W. T. S. Huck, C. Abell, and D. Klenerman. Single molecule fluorescence under conditions of fast flow. *Anal. Chem.*, 84:179–185, 2012.

[38] M. H. Horrocks, L. Rajah, P. Jönsson, M. Kjaergaard, M. Vendruscolo, T. P. J. Knowles, and D. Klenerman. Single-molecule measurements of transient biomolecular complexes through microfluidic dilution. *Anal. Chem.*, 85:6855–6859, 2013.

[39] R. Meester. *A Natural Introduction to Probability Theory*. Springer, 2003.

[40] W. Feller. *An introduction to probability theory and its applications*. Wiley, third edition, 1968.

[41] Y. Chen, J. D. Muller, P. T. C. So, and E. Gratton. The photon counting histogram in fluorescence fluctuation spectroscopy. *Biophys. J.*, 77:553 – 567, 1999.

[42] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[43] C. N. Morris. Parametric empirical bayes inference: Theory and applications. *J. Am. Statist. Assoc.*, 78:57 – 55, 1982.

[44] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[45] J. R. Norris. *Markov chains*. Cambridge University Press, 1998.

[46] P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to Stochastic Processes*. Houghton Mifflin Company, 1972.

[47] A. F. M. Smith and G. 0. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo method. *J. R. Statist. Soc. B*, 55:3–23, 1993.

[48] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[49] S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *Am. Stat.*, 49:327–335, 1995.

[50] G. Casella and E. I. George. Explaining the gibbs sampler. *Journal of Molecular Biology*, 46:167–174, 1992.

[51] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2003.

[52] R. M. Neal. Slice sampling. *Ann. Stat.*, 31:705–767, 2003.

[53] A. O'Hagan and J. Forster. *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference.* OUP, 2004.

[54] F. P. Kelly. *Reversibility and Stochastic Networks*. OUP, 2011.

[55] A. A. Deniz, T. A. Lawrence, M. Dahan, D. S. Chemla, P. S. Schultz, and S. Weiss. Ratiometric single-molecule studies of freely diffusing molecules. *Annu. Rev. Phys. Chem.*, 52:233–253, 2001.

[56] C. Gell, D. Brockwell, and A. Smith. *Handbook of single molecule fluorescence*. Oxford University Press, Oxford, 2006.

[57] B. K. Muller, E. Zaychikov, C. Brauchle, and D. C. Lamb. Pulsed interleaved excitation. *Biophys. J.*, 89(5):3508–3522, NOV 2005.

[58] V. Kudryavtsev, M. Sikor, S. Kalinin, D. Mokranjac, C. A. M. Seidel, and D. C. Lamb. Combining mfd and pie for accurate single-pair frster resonance energy transfer measurements. *ChemPhysChem*, 13:1060–1078, 2012.

[59] C. Eggeling, S. Berger, L. Brand, J.R. Fries, J. Schaffer, A. Volkmer, and Seidel C.A.M. Data registration and selective single-molecule analysis using multi-parameter fluorescence detection. *J. Biotechnol.*, 86:163–180, 2001.

[60] G. Wilson. Where's the real bottleneck in scientific computing? *American Scientist*, 94:5–6, 2006.

[61] Z. Merali. Computational science: Error, why scientific programming does not compute. *Nature*, 467:775–777, 2010.

[62] G. R. Mirams, C. J. Arthurs, M. O. Bernabeu, R. Bordas, J. Cooper, A. Corrias, Y. Davit, S.-J. Dunn, A. G. Fletcher, D. G. Harvey, M. E. Marsh, J. M. Osborne, P. Pathmanathan, J. Pitt-Francis, J. Southern, N. Zemzemi, and D. J. Gavaghan. Chaste: An open source c++ library for computational physiology and biology. *PLOS Comp. Biol.*, 9:e1002970– e1002970, 2013.

[63] E. Sisamakis, A. Valeri, S. Kalinin, P. J. Rothwell, and C. A. M. Seidel. Accurate single-molecule fret studies using multiparameter fluorescence detection. *Methods in Enzymology*, 475:455–514, 2010.

[64] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13:22–30, 2011.

[65] A. Hoffmann, D. Nettels, J. Clark, A. Borgia, S. E. Radford, J. Clarke, and Schuler B. Quantifying heterogeneity and conformational dynamics from single molecule fret of diffusing molecules: recurrence analysis of single particles (rasp). *Phys Chem Chem Phys*, 13:1857–1871, 2011.

[66] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing In Science and Engineering*, 9(3):90–95, 2007.

[67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[68] N. K Lee, A. N. Kapanidis, Y. Wang, X. Michalet, J. Mukhopadhyay, R. H. Ebright, and S. Weiss. Accurate fret measurements within single diffusing biomolecules using alternating-laser excitation. *Biophys. J.*, 88:2939–2953, 2006.

[69] S. van der Walt, S. C. Colbert, and Varoquaux G. The numpy array: A structure for efficient numerical computation. *Comput. Sci. Eng.*, 13:9–12, 2011.

[70] R. R. Murphy, G. Danezis, M. H. Horrocks, S. E. Jackson, and D. Klenerman. Bayesian inference of accurate population sizes and fret efficiencies from single diffusing biomolecules. *Anal. Chem. (submitted April 2014)*, 86:8603–8612, 2014.

[71] S. Kalinin, S. Felekyan, M. Antonik, and C. A. M. Seidel. Probability distribution analysis of single-molecule fluorescence anisotropy and resonance energy transfer. *J. Phys. Chem. B.*, 111:10253–10262, 2007.

[72] M. Antonik, S. Felekyan, A. Gaiduk, and C. A. M. Seidel. Separating structural heterogeneities from stochastic variations in fluorescence resonance energy transfer distributions via photon distribution analysis. *J. Phys. Chem. B.*, 110:6970–6978, 2006.

[73] Y. Santoso, J. P. Torella, and A. N. Kapanidis. Characterizing single-molecule fret dynamics with probability distribution analysis. *ChemPhysChem*, 11:2209–2219, 2010.

[74] J. P. Torella, S. J. Holden, Y. Santoso, J. Hohlbein, and A. N. Kapanidis. Identifying molecular dynamics in single-molecule fret experiments with burst variance analysis. *Biophysical Journal*, 107:5058–5063, 2011.

[75] N.K. Lee, A.N. Kapanidis, H.R. Koh, Y. Korlann, S.O. Ho, N. Kim Y. andGassman, S.K. Kim, and S. Weiss. Three-color alternating-laser excitation of single molecules: simultaneous monitoring of multiple interactions and distances. *Biophys. J.*, 92:303–12, 2007.

[76] P. J. Rothwell, S. Berger, O. Kensch, S. Felekyan, M. Antonik, B. M. Whrl, T. Restle, R. S. Goody, and C. A. M. Seidel. Multiparameter single-molecule fluorescence spectroscopy reveals heterogeneity of hiv-1 reverse transcriptase:primer/template complexes. *Proc. Natl. Acad. Sci. USA*, 100:1655–1660, 2002.

[77] B. Schuler. Single-molecule fluorescence spectroscopy of protein folding. *Chemphyschem*, 6:1206–1220, 2005.

[78] L. Ying, M. I. Wallace, S. Balsubramanian, and D. Klenerman. Ratiometric analysis of single-molecule fluorescence resonance energy transfer using logical combinations of threshold criteria: a study of 12-mer dna. *J. Phys. Chem. B.*, 104:5171–5178, 2000.

[79] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge, UK, 2012.

[80] T. Bayes and R. Price. An essay towards solving a problem in the doctrine of chance. by the late rev. mr. bayes, communicated by mr. price, in a letter to john canton, m. a. and f. r. s. *Phil. Trans. R. Soc. London*, 53:370–418, 1763.

[81] D. J. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003.

[82] S. A. McKinney, C. Joo, and T. Ha. Analysis of single molecule fret trajectories using hidden markov modeling. *Biophysical Journal*, 91:1941–1951, 2006.

[83] J. E. Bronson, J. Fei, J. M. Hofman, R. N. Gonzalez, and C. H. Wiggins. Learning rates and states from biophysical time series: A bayesian approach to model selection and single-molecule fret data. *Biophys. J.*, 97:3196–3205, 2009.

[84] J. E. Bronson, J. M. Hofman, J. Fei, R. L. Gonzales, and C. H. Wiggins. Graphical models for inferring single molecule dynamics. *BMC Bioinformatics*, 11:2–10, 2010.

[85] J. N. Taylor, D. E. Makarov, and C. F. Landes. Denoising single-molecule fret trajectories with wavelets and bayesian inference. *Biophys. J.*, 98:164–173, 2010.

[86] J. N. Taylor and C. F. Landes. Improved resolution of complex single-molecule fret systems via wavelet shrinkage. *J. Phys. Chem. B.*, 115:1105–1114, 2011.

[87] S. Uphoff, K. Gryte, G. Evans, and A. N. Kapanidis. Improved temporal resolution and linked hidden markov modeling for switchable single-molecule fret. *ChemPhysChem*, 12:571579, 2011.

[88] J. W. Yoon, A. Bruckbauer, W. J. Fitzgerald, and D. Klenerman. Bayesian inference for improved single molecule fluorescence tracking. *Biophys. J.*, 94:4932–4947, 2008.

[89] S. Turkcan, A. Alexandrou, and J.-B. Masson. A bayesian inference scheme to extract diffusivity and potenetial fields from confined single-molecule trajectories. *Biophys. J.*, 102:2288–2298, 2012.

[90] J. Stigler and M. Rief. Hidden markov analysis of trajectories in single-molecule experiments and the effects of missed events. *ChemPhysChem*, 13:1079–1086, 2012.

[91] W. Kugel, A. Muschielok, and J. Michaelis. Bayesian-inference-based fluorescence correlation spectroscopy and single-molecule burst analysis reveal the influence of dye selection on dna hairpin dynamics. *Chemphyschem*, 13:1013–1022, 2011.

[92] S.-M. Guo, J. He, N. Monnier, G. Sun, T. Wohland, and M. Bathe. Bayesian approach to the analysis of fluorescence correlation spectroscopy data ii: Application to simulated and in vitro data. *Anal. Chem.*, 84:3880–3888, 2011.

[93] J. He, S.-M. Guo, and M. Bathe. Bayesian approach to the analysis of fluorescence correlation spectroscopy data ii: Theory. *Anal. Chem.*, 84:3871–3879, 2011.

[94] S.-M. Guo, N. Bag, A. Mishra, T. Wohland, and M. Bathe. Bayesian total internal reflection fluorescence correlation spectroscopy reveals hiapp-induced plasma membrane domain organization in live cells. *Biophys. J.*, 106:190–200, 2014.

[95] S. C. Kou, X. S. Xie, and J. S. Liu. Bayesian analysis of single-molecule experimental data. *J. R. Stat. Soc.*, 54(3):469–496, 2005.

[96] S. Kalinin, S. Felekyan, A. Valeri, and C. A. M. Seidel. Characterizing multiple molecular states in single-molecule multiparameter fluorescence detection by probability distribution analysis. *J. Phys. Chem. B.*, 112:8361–8374, 2007.

[97] I. V. Gopich and A. Szabo. Single-molecule fret with diffusion and conformational dynamics. *J. Phys. Chem. B.*, 111:12925–12932, 2007.

[98] I. V. Gopich and A. Szabo. Decoding the pattern of photon colors in single-molecule fret. *J. Phys. Chem. B*, 113, 2009.

[99] I. V. Gopich and A. Szabo. Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule fret. *Proc. Natl. Acad. Sci. USA*, 109, 2012.

[100] M. S. DeVore, S. F. Gull, and C. K. Johnson. Classic maximum entropy recovery of the average joint distribution of apparent fret efficiency and fluorescence photons for single-molecule burst measurements. *J. Phys. Chem. B.*, 116:4006–4015, 2012.

[101] H. S Chung, J. M. Louis, and W. M. Eaton. Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proc. Natl. Acad. Sci. USA.*, 106:11837–11844, 2009.

[102] Y. Chen, J. D. Muller, P. T. So, and E. Gratton. The photon counting histogram in fluorescence fluctuation spectroscopy. *Biophys. J.*, 77:553–567, 1999.

[103] J. O. Lloyd-Smith. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLOS ONE*, 2:e180, 2007.

[104] C. I. Bliss and R. A. Fisher. Fitting the negative binomial distribution to biological data. *Biometrics*, 9:176–200, 1953.

[105] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–32, 1995.

[106] R. A. Holliday. A mechanism for gene conversion in fungi. *Genet. Res. Camb.*, 5:282 – 304, 1964.

[107] K. Uéda, H. Fukushima, E. Masliah, Y. Xia, A. Iwai, M. Yoshimoto, D. A. Otero, J. Kondo, Y. Ihara, and T. Saitoh. Molecular cloning of cdna encoding an unrecognized component of amyloid in alzheimer disease. *Proc. Natl. Acad. Sci. USA*, 90:11282–11286, 1993.

[108] M. G. Spillantini, M. L. Schmidt, V. M.-Y. Lee, T. Q. Trojanowski, R. Jakes, and M. Goedert. α-synuclein in lewy bodies. *Nature*, 388:839–840, 1997.

[109] E. Noe, K. Marder, K. L. Bell, D. M. Jacobs, J. J. Manly, and Y. Stern. Comparison of dementia with lewy bodies to alzheimers disease and parkinsons disease with dementia. *Mov. Disord.*, 19(1):60–67, 2004.

[110] M. Goedert, M. G. Spillantini, R. Jakes, D. Rutherford, and Crowther R. A. Multiple isoforms of human microtubule-associated protein tau: sequences and localization in neurofibrillary tangles of alzheimer's disease. *Neuron*, 3(4):519–526, 1989.

[111] W. R. Gibb and A. J. Lees. The relevance of the lewy body to the pathogenesis of idiopathic parkinson's disease. *J. Neurol. Neurosurg. Psychiatry*, 51:745–752, 1988.

[112] L. C. Serpell. Alzheimers amyloid fibrils: structure and assembly. *BBA-MOL Basis Dis.*, 1502:16–30, 2000.

[113] G. A. Wells, A. C. Scott, C. T. Johnson, R. F. Gunning, R. D. Hancock, M. Jeffrey, M. Dawson, and R. Bradley. A novel progressive spongiform encephalopathy in cattle. *Vet. Rec.*, 121:419–420, 1987.

[114] L. G. Goldfarb, R. B. Petersen, M. Tabaton, P. Brown, A. C. LeBlanc, P. Montagna, P. Cortelli, J. Julien, C. Vital, and W. W. Pendelbury. Fatal familial insomnia and familial creutzfeldt-jakob disease: disease phenotype determined by a dna polymorphism. *Science*, 258:806–808, 1992.

[115] A. Clark, C. A. Wells, I. D. Buley, J. K. Cruickshank, R. I. Vanhegan, D. R. Matthews, G. J. Cooper, R. R. Holman, and R. C. Turner. Islet amyloid, increased a-cells, reduced b-cells and exocrine fibrosis: quantitative changes in the pancreas in type 2 diabetes. *Diabetes Res.*, 9:151–159, 1988.

[116] J. Safar, P. P. Roller, D. C. Gajdusek, and C. J. (Jr) Gibbs. Conformational transitions, dissociation, and unfolding of scrapie amyloid (prion) protein. *J. Biol. Chem.*, 268:20276–20284, 1993.

[117] K. A. Conway, S.-J. Lee, J.-C. Rochet, T. T. Ding, R. E. Williamson, and P. T. Lansbury. Acceleration of oligomerization, not fibrillization, is a shared property of both -synuclein mutations linked to early-onset parkinson's disease: Implications for pathogenesis and therapy. *Proc. Natl. Acac. Sci. USA*, 97:571–576, 1999.

[118] C. Haass and D. J. Selkoe. Soluble protein oligomers in neurodegeneration: lessons from the alzheimer's amyloid bold beta-peptide. *Nat. Rev. Mol. Cell Biol.*, 8:101–112, 2007.

[119] P.J. Muchowski and J. L. Wacker. Modulation of neurodegeneration by molecular chaperones. *Nat. Rev. Neurosci.*, 6:11–22, 2005.

[120] I. Benilova, E. Karran, and B. De Strooper. The toxic aβ oligomer and alzheimer's disease: an emperor in need of clothes. *Methods Mol. Biol.*, 15:349–357, 2012.

[121] F. Chiti and C. M. Dobson. Protein misfolding, functional amyloid and human disease. *Annu. Rev. Biochem.*, 75:333– 366, 2006.

[122] T. P. J. TKnowles, C. A. Waudby, G. L. Devlin, S. I. A. Cohen, A. Aguzzi, M. Vendruscolo, E. M. Terentjev, M. E. Welland, and C. M. Dobson3. An analytical solution to the kinetics of breakable filament assembly. *Annu. Rev. Biochem.*, 326:1533 – 1537, 2009.

[123] T. P. J. Knowles and M. J. Buehler. Nanomechanics of functional and pathological amyloid materials. *Proceedings of the National Academy of Sciences of the United States of America*, 6:469–1479, 2011.

[124] S. A. I. Cohen, S. Linse, L. M. Luheshi, E. Hellstrand, D. A. White, D. E. Rajaha, L.and Otzen, M. Vendruscolo, C. M. Dobson, and T. P. J. Knowles. Proliferation of amyloid-42 aggregates occurs through a secondary nucleation mechanism. *Proc. Natl. Acac. Sci. USA*, 110:9758 – 9763, 2013.

[125] G. Meisl, X. Yang, E. Hellstrand, B. Frohm, J. B. Kirkegaard, S. I. A. Cohen, C. M. Dobson, S. Linse, and T. P. J. Knowles. Differences in nucleation behavior underlie the contrasting aggregation kinetics of the a40 and a42 peptides. *Proc. Natl. Acac. Sci. USA*, 111:9384 – 9389, 2014.

[126] J. T. Giurleo, X. He, and D. S. Talaga. β-lactoglobulin assembles into amyloid through sequential aggregated intermediates. *J. Mol. Biol.*, 381:1332–1348, 2008.

[127] M. D. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, and C. M. Dobson. Mapping long-range interactions in -synuclein using spin-label nmr and ensemble molecular dynamics simulations. *J. Am. Chem. Soc.*, 127:476–477, 2005.

[128] S. L. Gras, L. J. Waddington, and K. N. Goldie. Transmission electron microscopy of amyloid fibrils. *Methods Mol Biol.*, 752:197–214, 2011.

[129] B. Langkilde, A. E.and Vestergaard. Methods for structural characterization of pre-fibrillar intermediates and amyloid fibrils. *FEBS Lett.*, 583:2600–2609, 2009.

[130] S. I. Cohen, M. Vendruscolo, C. M. Dobson, and T. P. Knowles. From macroscopic measurements to microscopic mechanisms of protein aggregation. *J. Biol. Chem.*, 421:160–171, 2012.

[131] S. D. Schmidt, R. A. Nixon, and P. M. Mathews. Elisa method for measurement of amyloid-beta levels. *Methods Mol. Biol.*, 299:279–297, 2005.

[132] H. A. Lashuel, B. M. Petre, J. Wall, M. Simon, R. J. Nowak, T. Walz, and Lansbury P. T. (Jr). Alpha-synuclein, especially the parkinson's disease-associated mutants, forms pore-like annular and tubular protofibrils. *J Mol Biol*, 322:1089–1102, 2002.

[133] I. V. J Murray, B. I. Giasson, S. M. Quinn, V. Koppaka, P. H. Axelsen, H. Ischiropoulos, J. H. Trojanowski, and V. M.-Y. Lee. Role of α-synuclein carboxy-terminus on fibril formation in vitro. *Methods Mol. Biol.*, 42:8530–8540, 2003.

[134] A. Orte, N. R. Birkett, R. W. Clarke, G. L. Devlin, C. M. Dobson, and D. Klenerman. Direct characterization of amyloidogenic oligomers by single-molecule fluorescence. *Proceedings of the National Academy of Sciences of the United States of America*, 105:14424–14429, 2008.

[135] H. Potter and D. Dressler. On the mechanism of genetic recombination: Electron microscopic observation of recombination intermediates. *Proc. Natl. Acac. Sci. USA*, 73:3000 – 3004, 1976.

[136] S. A. McKinney, A.-C. Declais, D. M. J. Lilley, and T. Ha. Structural dynamics of individual holliday junctions. *Nature Structural Biology*, 10:93–97, 2008.

[137] S. Uphoff, S. J. Holden, J. Le Reste, L. andd Periz, S. van de Linde, M. Heilmann, and A. N. Kapanidis. Monitoring multiple distances within a single molecule using switchable fret. *Nature Methods*, 7:831 – 836, 2010.

[138] C. Hyeon, J. Lee, J. Yoon, S. Hohng, and D. Thirumalai. Hidden complexity in the isomerization dynamics of holliday junctions. *Nature Chemistry*, 4:904 –914, 2012.

[139] E. L. Lehmann. *Testing Statistical Hypotheses*. Springer Verlag, New York, 1986.

[140] Joo C., McKinney S. A., Nakamura M., Rasnik I., Myong S., and Ha T. Real-time observation of reca filament dynamics with single monomer resolution. *Cell*, 126:515–527, 2006.

[141] Rasnik I., McKinney S. A., and Ha T. Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat. Methods*, 3:891–893, 2006.

[142] C. E. Aitken, R. A. Marshall, and J. D. Puglisi. An oxygen scavenging system for improvement of dye stability in single-molecule fluorescence experiments. *Biophys. J.*, 94:1826 – 1835, 2008.

[143] K. J. Liu and T.-H. Wang. Cylindrical illumination confocal spectroscopy: Rectifying the limitations of confocal single molecule spectroscopy through one-dimensional beam shaping. *Biophys J.*, 95:2964–2975, 2008.

[144] S. Tyagi, V. VanDelinder, N. Banterle, G. Fuertes, S. Milles, M. Agez, and E. A. Lemke. Continuous throughput and long-term observation of single-molecule fret without immobilization. *Nat. Methods*, 11:297–300, 2014.

[145] D. R. Bentley, S. Balasubramanian, H. P Swerdlow, G. P. Smith, J. Milton, G. Brown, C, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, K. Cheetham, R, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M.T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, Scally A., G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. Earnshaw, C. Egbujor, U, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, K. V. Fraser, L. J. andFuentes Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. G. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. H. Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, Z. Kindwall, A. P. an Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. M. Mullens, T. Newington, Z. Ning, B. L. Ng, S. M. Novo, M. J. ONeill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. C. Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. E. S. Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. vandeVondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, G. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, M. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59, 2008.

[146] Illumina. Data procesing of nextera mate pair reads on illumina sequencing platforms. http://www.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing.pdf, 2012. Accessed: 2015-02-05.

[147] F. Sanger, S. Nicklen, and A. R. Coulson. Dna sequencing with chain-terminating inhibitors. *Proc. Nati. Acad. Sci. USA*, 74:5463–5467, 1977.

[148] M. L. Metzker. Emerging technologies in dna sequencing. *Genome Res.*, 15:1767–1776, 2005.

[149] L. M. Smith, S. Fung, M. W. Hunkapiller, T. J. Hunkapiller, and L. E. Hood. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent dna primers for use in dna sequence analysis. *Nucleic Acids Res.*, 13:2399–2412, 1985.

[150] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. Fluorescence detection in automated dna sequence analysis. *Nature*, 321:674–679, 1986.

[151] Illumina. Illumina sequencing technology. http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf, 2010. Accessed: 2015-03-25.

[152] C. Adessi, G. Matton, G. Ayala, G. Turcatti, J.-J. Mermod, P. Mayer, and E. Kawashima. Solid phase dna amplification: characterisation of primer attachment and amplification mechanisms. *Nucl. Acids Res.*, 28:e87, 2000.

[153] F. Chen, M. Dong, M. Ge, L. Zhu, L. Ren, G. Liu, and R. Mu. The history and advances of reversible terminators used in new generations of sequencing technology. *Genomics Proteomics Bioinformatics*, 11:34–40, 2013.

[154] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, Z. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner.

Real-time dna sequencing from single polymerase molecules. *Science*, 323:133–138, 2008.

[155] M. J. Levene, J. Korlach, W. Turner, M. Foquet, H. G. Craighead, and W. W. Webb. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299:682–686, 2003.

[156] R. J. Roberts, M. O. Carneiro, and M. C. Schatz. The advantages of smrt sequencing. *Genome Biology*, 14:405–409, 2013.

[157] H. Bayley. Nanopore sequencing: From imagination to reality. *Clin. Chem.*, 61:25–31, 2014.

[158] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, K. Krstic, P, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss. The potential and challenges of nanopore sequencing. *Nature Biotechnol.*, 26:1146–1153, 2008.

[159] A. H. Laszlo, I. M. Derrington, B. C. Ross, H. Brinkerhoff, A. Adey, I. C. Nova, J. M. Craig, K. W. Langford, J. M. Samson, R. Daza, K. Doering, Shendure J., and J. H. Gundlach. Decoding long nanopore sequencing reads of natural dna. *Nature Biotechnol.*, 32:829–833, 2014.

[160] A. S. Mikheyev and M. M. Y. Tin. A first look at the oxford nanopore minion sequencer. *Mol. Ecol. Resour.*, 14:1097–1102, 2014.

[161] H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010.

[162] P. E. C. Compeau, P. A. Pevzner, and G. Tesler. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnol.*, 29:987–991, 2011.

[163] P. A. Pevzner, H. Tang, and M. S. Waterman. An eulerian path approach to dna fragment assembly. *Proc. Natl. Acac. Sci. USA*, 98(17):9748–9753, 2001.

[164] D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.*, 18:821–829, 2008.

[165] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comp. Biol.*, 19(5):455–477, 2012.

[166] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, 20(2):265–272, 2010.

[167] M. Hunt, C. Newbold, M. Berriman, and T. D. Otto. A comprehensive evaluation of assembly scaffolding tools. *Genome Biol.*, 15:R42, 2014.

[168] A. Edwards, H. Voss, P. Rice, A. Civitello, J. Stegemann, C. Schwager, J. Zimmermann, H. Erfle, C. T. Caskey, and W. Ansorge. Automated dna sequencing of the human hprt locus. *Genomics*, 6(4):593–608, 1990.

[169] P. Medvedev, S. Pham, M. Chaisson, G. Tesler, and P. Pevzner. Paired de Bruijn graphs: A novel approach for incorporating mate pair information into genome assemblers. *J Comput Biol.*, 18(11):1625–1634, 2011.

[170] J. R. Miller, S. Koren, and G. Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[171] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.

[172] K.R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W. C. Chou, J. Corbeil, C. Del Fabbro, T. R. Docking, R. Durbin, Earl D., S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, E. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D.. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J.O. Kitzman, J.R. Knight, S. Koren, T. W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. Maccallum, M.D. Macmanes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O.S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure,

Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S. M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*, 2(1):10, 2013.

[173] S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marais, M. Pop, and J. A. Yorke. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.*, 22(3):557–567, 2012.

[174] M. Ghodsi, C. M. Hill, I. Astrovskaya, H. Lin, and D. D. Sommer. De novo likelihood-based measures for comparing genome assemblies. *BMC Research Notes*, 6:334, 2013.

[175] D. Coil, G. Jospin, and A. E. Darling. A5-miseq: an updated pipeline to assemble microbial genomes from illumina MiSeq data. 2014.

[176] A. Tritt, J. A. Eisen, M. Facciotti, and A. E. Darline. An integrated pipeline for de novo assembly of microbial genomes. *PLoS ONE*, 7(9):e42304, 2012.

[177] M. Hunt, T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T. D. Otto. REAPR: a universal tool for genome assembly evaluation. *Genome Biol.*, 14:R47, 2013.

[178] S. C. Clark, R. Egan, P. I. Frazier, and Z. Wang. ALE: a generic assembly likelihood evaluation framework for asessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, 29(4):435–443, 2013.

[179] B. Walker. Pilon. `https://github.com/broadinstitute/pilon/releases`, 2014.

[180] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.*, 49(4):764–766, 2013.

[181] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press and McGraw-Hill, 2009.

[182] K. J. Millman and M. Aivazis. Python for scientists and engineers. *Computing in Science and Engineering*, 13:9–12, 2011.

[183] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[184] H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.

[185] S. Koren and A. M. Phillippy. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.*, 23:110–120, 2015.