

# Contents

<b>1</b>	<b>Thesis Overview</b>	<b>6</b>
1.1	General Introduction . . . . .	6
1.2	Contributions . . . . .	6
1.3	Thesis Overview . . . . .	6
<b>2</b>	<b>Introduction</b>	<b>8</b>
2.1	Introduction . . . . .	8
2.2	Fluorescence Microscopy . . . . .	8
2.3	Forster Resonance Energy Transfer . . . . .	8
2.4	Single Molecule Fluorescence Microscopy . . . . .	9
2.4.1	Overview . . . . .	9
2.4.2	Confocal Microscopy . . . . .	9
2.4.3	Total Internal Fluorescence Microscopy . . . . .	10
2.4.4	Epifluorescent Microscopy . . . . .	10
2.4.5	Super-Resolution Microscopy . . . . .	10
2.5	Probabilistic Inference and Bayesian Statistics . . . . .	10
2.5.1	Bayesian Statistics . . . . .	10
2.5.2	Sampling Techniques . . . . .	10
<b>3</b>	<b>Analysis Tools for Single Molecule Confocal Microscopy</b>	<b>11</b>
3.1	Overview . . . . .	11
3.2	Introduction . . . . .	12
3.2.1	The Single Molecule Fluorescence Experiment . . . . .	12
3.3	Data Analysis in Confocal smFRET Experiments . . . . .	13
3.3.1	Continuous Excitation . . . . .	14
3.3.2	Alternating Laser Excitation . . . . .	15

3.4	Development of Scientific Software . . . . .	16
3.5	pyFRET: Design and Implementation . . . . .	17
3.5.1	Code Layout and Design . . . . .	17
3.5.2	Simple Event Selection and Denoising . . . . .	19
3.5.3	Burst Search Algorithms . . . . .	22
3.6	RASP: Recurrence Analysis of Single Particles . . . . .	22
3.6.1	Compatibilities . . . . .	23
3.7	Experimental Methods . . . . .	24
3.7.1	Benchmarking the Gaussian Fitting Using Simulated Datasets . . . . .	24
3.7.2	Data to Evaluate the Simple Event Selection Algorithms . . . . .	24
3.7.3	Data to Evaluate Event Selection Using the Burst Search Algorithms . . . . .	25
3.7.4	Performance Analysis Using Mixtures of DNA Duplexes . . . . .	26
3.7.5	Testing the RASP Algorithm . . . . .	26
3.8	Performance Analysis of smFRET Analysis Algorithms . . . . .	27
3.8.1	Evaluating Performance with DNA Duplexes . . . . .	27
3.8.2	Evaluating the Burst Search Algorithms . . . . .	28
3.8.3	Evaluating Performance of the Gaussian Mixture Model . . . . .	32
3.8.4	Benchmarking the RASP Algorithm . . . . .	35
3.9	Conclusions . . . . .	36
3.10	Availability and Future Directions . . . . .	38
<b>4</b>	<b>Bayesian Inference of Intramolecular Distances Using Single Molecule FRET</b>	<b>39</b>
4.1	Overview . . . . .	39
4.2	Introduction . . . . .	40
4.2.1	A smFRET Experiment . . . . .	40
4.2.2	Approaches to Analysis of smFRET Data . . . . .	42
4.2.3	Model Based Inference . . . . .	44
4.3	Theory . . . . .	49
4.3.1	A Physical Model of a smFRET Experiment . . . . .	49
4.3.2	Inference of Model Parameters . . . . .	57
4.3.3	The Metropolis-Hastings Algorithm . . . . .	61
4.4	Experimental Methods . . . . .	63
4.5	Results . . . . .	67
4.5.1	Justification of the Gamma-Poisson Mixture Model . . . . .	77
4.6	Conclusions and Future Work . . . . .	79

<b>5</b>	<b>Bayesian Inference of Oligomer Sizes Using Single Molecule FRET</b>	<b>81</b>
5.1	Overview . . . . .	81
5.2	Introduction . . . . .	82
5.2.1	Diseases of Protein Aggregation . . . . .	82
5.2.2	Studying Protein Aggregation . . . . .	83
5.2.3	The Relationship Between Size and Photon Emission is Complex . . .	84
5.2.4	The Effect of Confocal Excitation Heterogeneity on Photon Emission	85
5.2.5	Controlling Confocal Excitation Heterogeneity . . . . .	86
5.2.6	The DNA Holliday Junction as a Model Oligomer . . . . .	87
5.3	Theory . . . . .	87
5.3.1	A Simple Poisson Model of Oligomer Photon Emission . . . . .	90
5.3.2	A Gamma-Poisson Mixture Model of Oligomer Photon Emission . . .	91
5.4	Experimental Methods . . . . .	93
5.4.1	Labelling of Protein Monomers . . . . .	93
5.4.2	Protein Aggregation Experiments . . . . .	93
5.4.3	Preparation of DNA Holliday Junctions . . . . .	93
5.4.4	Simple FRET Measurements of DNA Holliday Junctions . . . . .	94
5.4.5	Flattening the Confocal Volume Using Acousto-Optic Deflection: A Modified Single Molecule Fluorescence Microscope . . . . .	94
5.4.6	Preparation of Microfluidic Channels . . . . .	95
5.4.7	smFRET Measurements to Determine the Effect of Unequal Excitation on Photon Emission . . . . .	95
5.4.8	Counting Photobleaching Steps Using TIRF Imaging . . . . .	96
5.5	Results . . . . .	96
5.5.1	The need for a Generative Model . . . . .	96
5.5.2	Understanding the Relationship Between Size and Photon Emission .	99
5.5.3	Inferring Event Brightness Using the Gamma-Poisson Model . . . . .	102
5.5.4	How Bright Are Holliday Junction Events . . . . .	108
5.5.5	Photobleaching Steps Analysis Reveals Additional Source of Overdis- persal . . . . .	112
5.6	Conclusions . . . . .	113
5.6.1	Complex Relationship between Size and Photon Emission . . . . .	113
5.6.2	Implications for Future Work on Molecular Sizing . . . . .	114

## 6 Probabilistic Inference for Error Detection in De Novo Genome Assem-

<b>blies</b>	<b>116</b>
6.1 Overview . . . . .	116
6.2 Introduction . . . . .	116
6.3 Theory . . . . .	119
6.4 Experimental Methods . . . . .	124
6.5 Results . . . . .	128
6.6 Conclusions . . . . .	130
<b>7 Conclusions and Future Work</b>	<b>134</b>
7.1 General Conclusions . . . . .	134
7.2 Applications . . . . .	134
7.3 Future Work . . . . .	134

# Chapter 1

## Thesis Overview

### 1.1 General Introduction

### 1.2 Contributions

### 1.3 Thesis Overview

The rest of this thesis is structured as follows. Chapter 2 provides a introduction to other research that has been undertaken in the field of fluorescence microscopy. General experimental and analysis techniques for fluorescence microscopy of single molecules are introduced, and the work presented in this thesis is contextualised. Chapter 2 also provides an overview of current research in Bayesian statistics and probabilistic analysis, introducing the statistical methods that are used in later chapters. Following this introductory section, we present our results over four separate chapters.

Chapter 3 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 4 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 5 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 this model is degenerate, making inference of aggregate sizes undecidable. We further show that a single emission event has multiple sources of heterogeneity, creating a complex, non-linear relationship between aggregate size and the number of photons emitted in a fluorescent event.

Chapter 6, 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.

Finally, Chapter 7 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.