# MARKOVIAN DYNAMICS IN CHROMATIN LOOP EXTRUSION FACTORS

KATHLEEN ENG

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTER OF SCIENCE

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

SEPTEMBER 2019

# Abstract

Markov properties can be used to model different dynamic processes at various stages of the loop extrusion process. The current methods are proposed to gain insight on how Markov models may illustrate chromatin behaviour once the appropriate observed data becomes available. We find that single molecule FRET experiments are able to identify the conformational states of chromatin using Gaussian mixture models. The unbinding and binding rates of loop extrusion factors (LEFs) were applied in an immigration-death model, and found to play a role in influencing the frequency of loop extrusion. By including the additional parameter of the presence of nucleosomes with LEF binding on a strand of DNA, we find that the theoretical timescale of DNA exposure decreased upon LEF binding. The binding behaviour of LEFs is also dependent on the location of nucleosomes on a strand of DNA. This is modelled with the Gillespie algorithm to simulate LEF binding activity with single cell dynamics.

# Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Jorg Grigull, for his patient guidance, support, and immense knowledge. I am particularly grateful for his willingness to give his time so generously, and encouragement to keeping my progress on schedule. His contributions have been very much appreciated.

I would also like to thank my committee members, Dr. Xin Gao, Dr. John McDermott, and Dr. Iain Moyles for their help in offering me their time, comments, critiques and suggestions regarding my research.

Finally, I wish to thank my family and friends for their support and encouragement throughout my study. They have continuously made sacrifices in order for me to pursue my own goals. Their support and encouragement is worth more than I can express on paper.

# Table of Contents

**Abstract** **ii**

**Acknowledgements** **iii**

**Table of Contents** **iv**

**List of Tables** **viii**

**List of Figures** **ix**

**1 General Introduction** **1**

1.1 Objectives and Aims . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.1.1 Research Question . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.1.2 Research Design . . . . . . . . . . . . . . . . . . . . . . . . . . 2

**2 Mixture Models in Chromatin Compaction** **4**

2.1 Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.2 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.3 Materials and methods . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.3.1 Single-molecule FRET . . . . . . . . . . . . . . . . . . . . . . 5

2.3.2 EM algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

2.3.3 EM application: Gaussian Mixture Models . . . . . . . . . . . . . . . 10

2.3.4 Previous applications of the EM mixture model in enzyme kinetics . . 15

2.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

2.4.1 Gaussian Mixture Model Application: Chromatin dynamics . . . . . 17

2.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

2.6 Future directions . . . . . . . . . . . . . . . . . . . . . . . . . . . 28

2.7 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29

**3 CTMC Application: Loop Extrusion Factor Stacking 30**

3.1 Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

3.2 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

3.3 Materials and methods . . . . . . . . . . . . . . . . . . . . . . . . . 31

3.3.1 Continuous Time Markov Chains . . . . . . . . . . . . . . . . . . . 31

3.3.2 Poisson Process . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32

3.3.3 Immigration-Death Process . . . . . . . . . . . . . . . . . . . . . . 34

3.3.4 Gillespie Algorithm for Chromatin Loop Dynamics . . . . . . . . . . 36

3.3.5 Rates of Loop Extrusion . . . . . . . . . . . . . . . . . . . . . . . 39

3.3.6 Binding Behaviour of LEFs . . . . . . . . . . . . . . . . . . . . . . 41

3.3.7 Combining LEF Binding with Loop Extrusion Activity . . . . . . . . 43

3.3.8 Burst Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 46

3.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 49

3.4.1 Steady state binding . . . . . . . . . . . . . . . . . . . . . . . . . 49

3.4.2 Varying the rates of binding and unbinding . . . . . . . . . . . . . . 52

3.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

3.6 Future directions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 58

3.7 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59

**4 Theoretical Exposure Time for Loop Extrusion 60**

4.1 Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 60

4.2 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 60

4.3 Materials and Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61

4.3.1 Random Walks in Chromatin Organization . . . . . . . . . . . . . . . 61

4.3.2 Transition Rates . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62

4.3.3 Random Walk Application: LEF Binding . . . . . . . . . . . . . . . . 64

4.3.4 Derivation of Exposure Time on Chromatin . . . . . . . . . . . . . . 69

4.3.5 Mean First-Passage Time of DNA Exposure . . . . . . . . . . . . . . 71

4.3.6 LEF Exposure Cases on DNA . . . . . . . . . . . . . . . . . . . . . . 73

4.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 85

4.4.1 DNA Exposure Time Theoretical Equations . . . . . . . . . . . . . . 85

4.4.2 DNA Exposure Time Simulations . . . . . . . . . . . . . . . . . . . . 88

4.4.3 Case A: No LEF Binding . . . . . . . . . . . . . . . . . . . . . . . . 90

4.4.4 Case B: Binding of one LEF . . . . . . . . . . . . . . . . . . . . . . 92

4.4.5 Case C: Binding of one LEF and one nucleosome . . . . . . . . . . . 96

4.4.6 Case D: Binding of one LEF and two nucleosomes . . . . . . . . . . . 98

4.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 100

4.6 Future directions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102

4.7 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103

**5 General Conclusion 104**

5.1 Limitations of research . . . . . . . . . . . . . . . . . . . . . . . . . . . 104

5.2 Future directions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 107

**Bibliography 108**

**A R Code for Gaussian Mixture Models 115**

**B Python Code for Markovian Models 117**

**C Python Code for Loop Visualization 121**

# List of Tables

2.1 Parameters from smTIRF- FRET data for 2-component model . . . . . . . . 19

2.2 Results from smTIRF- FRET data for 2-component model . . . . . . . . . . 19

2.3 Parameters from smTIRF- FRET data for adjusted 2-component model . . . 20

2.4 Results from smTIRF- FRET data for adjusted 2-component model . . . . . 20

2.5 Parameters from smTIRF- FRET data with smaller tolerance . . . . . . . . 21

2.6 Results from smTIRF- FRET data with smaller tolerance . . . . . . . . . . . 21

2.7 Parameters from smTIRF- FRET data for 4-component model . . . . . . . . 23

2.8 Results from smTIRF- FRET data for 4-component model . . . . . . . . . . 23

2.9 Parameters from smTIRF- FRET data for 5-component model . . . . . . . . 24

2.10 Results from smTIRF- FRET data for 5-component model . . . . . . . . . . 24

3.1 Parameters from observed data . . . . . . . . . . . . . . . . . . . . . . . . 41

4.1 Nucleosome Kinetic Parameters ($s^{-1}$) . . . . . . . . . . . . . . . . . . . . . 90

# List of Figures

1.1 ‘Beads on a string” illustration of chromatin. The orange circles depict the nucleosomes, and the blue ‘string’ depicts the strand of DNA. . . . . . . . . 1

2.1 Single iteration of the EM algorithm. Figure obtained from literature [9] . . 9

2.2 FRET populations with 2-component Gaussian Distribution Fit . . . . . . . 20

2.3 FRET populations with 4-component Gaussian Distribution Fit . . . . . . . 22

2.4 FRET populations with 5-component Gaussian Distribution Fit . . . . . . . 23

3.1 Diagram of M/M/$\infty$ queue . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

3.2 Simulated loop extrusion with overlapping loops in a single time frame. The height of the loop demonstrates the amount of chromatin that has been extruded from the loop. The colour of the loop is also significant such that darker the loop, the more stacking has occurred. . . . . . . . . . . . . . . . . . 38

3.3 Simulated loop extrusion with overlapping loops in a single time frame with more LEFs in the system . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39

3.4 Steady states of DNA loops. **A.** Sparse state ; **B.** Dense state . . . . . . . . 40

3.5 Rates of loop extrusion. The left plot denotes the rate of loop division, and the right plot denotes the rate of loop death. . . . . . . . . . . . . . . . . . . . 42

3.6 Binding behaviour of LEFs on DNA. This plot illustrates the relationship between the copy number of binding sites to the amount of LEFs stably bound to DNA. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44

3.7 Loop Extrusion Dynamics with LEF binding. Components illustrate the relationship between the rate of binding ($k_{\text{on}}$) and unbinding ($k_{\text{off}}$) from the unbound (UB) to the dynamically bound (DB) LEFs. The dynamically bound LEFs then become stably bound (SB) and allow for loop extrusion (LE) to occur. This activity is regulated by their rate of loop division ($R_{\text{div}}$) and loop death ($R_{\text{death}}$). . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44

3.8 Simplified loop extrusion model . . . . . . . . . . . . . . . . . . . . . . . 46

3.9 Low steady state binding rate, where $k_{\text{on}}/k_{\text{off}} = 0.04$. The upper plot denotes the activity of the LEFs bound and unbound from the strand of DNA by their on/off states. The bottom plot denotes the loop extrusion activity that corresponds to the state of the LEF binding. . . . . . . . . . . . . . . . . . . 51

3.10 High steady state binding rate, where $k_{\text{on}}/k_{\text{off}} = 0.4$. The upper plot denotes the activity of the LEFs bound and unbound from the strand of DNA by their on/off states. The bottom plot denotes the loop extrusion activity that corresponds to the state of the LEF binding. . . . . . . . . . . . . . . . . . . 51

3.11 The rate of $k_{\text{on}}$ is greater than that of $k_{\text{off}}$ in an unsteady state. The upper plot denotes the activity of the LEFs bound and unbound from the strand of DNA by their on/off states. The bottom plot denotes the loop extrusion activity that corresponds to the state of the LEF binding. . . . . . . . . . . 53

3.12 The rate of $k_{\text{off}}$ is greater than that of $k_{\text{on}}$ in an unsteady state. The upper plot denotes the activity of the LEFs bound and unbound from the strand of DNA by their on/off states. The bottom plot denotes the loop extrusion activity that corresponds to the state of the LEF binding. . . . . . . . . . . 53

4.1 Diagram of nucleosome structure. The left of the arrow shows the “beads on a string” view of the nucleosome, whereas the right of the arrow depicts the side view of a nucleosome . . . . . . . . . . . . . . . . . . . . . . . . . 61

4.2 Random walk of LEF binding . . . . . . . . . . . . . . . . . . . . . . . . 62

4.3 Random walk model of binding potential . . . . . . . . . . . . . . . . . . 66

4.4 Loop extrusion in the presence of nucleosomes . . . . . . . . . . . . . . . . 70

4.5 Theoretical cases of LEF binding between nucleosomes. **A.** Exposed strand of DNA, with no LEF binding. The target patch , $m$, is bounded by two nucleosomes, $N_1$ and $N_2$. **B.** Binding of one LEF ($C$) between two nucleosomes. **C.** Binding of one LEF and one nucleosome between two nucleosomes. **D.** Binding of one LEF and two nucleosomes between two nucleosomes. . . . . . 74

4.6 No LEF or nucleosome binding . . . . . . . . . . . . . . . . . . . . . . 76

4.7 Binding of one LEF . . . . . . . . . . . . . . . . . . . . . . . . . . . 78

4.8 Binding of one LEF and one nucleosome . . . . . . . . . . . . . . . . . 80

4.9 Binding of one LEF and two nucleosomes . . . . . . . . . . . . . . . . . 82

4.10 Theoretical exposure timescale plot. **Case A** No LEF or nucleosome binding. **Case B** Binding of one LEF. **Case C** Binding of one LEF and one nucleosome. **Case D** Binding of one LEF and two nucleosomes. . . . . . . . . . . . . . 85

4.11 Theoretical exposure timescale plot on a log axis. **Case A** No LEF or nucleosome binding. **Case B** Binding of one LEF. **Case C** Binding of one LEF and one nucleosome. **Case D** Binding of one LEF and two nucleosomes. . . 86

4.12 Theoretical exposure timescale plot with a greater on rate. Plot A depicts the theoretical exposure timescale plot, and plot B depicts the same plot on a log axis. **Case A** No LEF or nucleosome binding. **Case B** Binding of one LEF. **Case C** Binding of one LEF and one nucleosome. **Case D** Binding of one LEF and two nucleosomes. . . . . . . . . . . . . . . . . . . . . . . . . . 87

4.13 Theoretical exposure timescale plot with a greater off rate. Plot A depicts the theoretical exposure timescale plot, and plot B depicts the same plot on a log axis. **Case A** No LEF or nucleosome binding. **Case B** Binding of one LEF. **Case C** Binding of one LEF and one nucleosome. **Case D** Binding of one LEF and two nucleosomes. . . . . . . . . . . . . . . . . . . . . . 88

4.14 No LEF or additional nucleosome binding. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.** Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot. . . . . . . . . . . . . . . . . . . . . . . . . 91

4.15 No LEF, additional nucleosome binding, nor sliding of nucleomes. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.** Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot. . . . . . . . . . . 93

4.16 Binding of one LEF. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.**Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 94

4.17 Binding of one LEF. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.**Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95

4.18 Binding of one LEF and one nucleosome. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.**Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot. . . . . . . . . . . . . . . . . . . . . . . . . . 96

4.19 Binding of one LEF and two nucleosomes. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.**Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot. . . . . . . . . . . . . . . . . . . . . . . . . . 98

# Abbreviations

| | |
|---|---|
| BDP | Birth-Death process |
| ChIA-PET | Chromatin Interaction analysis with Paired-End Tag |
| ChIP | Chromatin immunoprecipitation |
| CTCF | CCCTC-binding factor |
| CTMC | Continuous-time Markov chain |
| DNA | Deoxyribosenucleic acid |
| EM | Expectation Maximization |
| FRET | Förster (or Fluorescence) Resonance Energy Transfer |
| GMM | Gaussian Mixture Model |
| IDP | Immigration-Death process |
| LEF | Loop extrusion factor |
| L-BFGS | Limited-memory Broyden-Fletcher-Goldfarb-Shanno |
| MCMC | Monte Carlo Markov chain |
| MLE | Maximum Likelihood Estimation |
| mRNA | Messenger RNA |
| PCR | Polymerase chain reaction |
| pdf | Probability density function |
| pgf | Probability generating function |
| PTM | Post-translational modifications |
| RNA | Ribonucleic acid |
| $R_{\text{division}}$ | Rate of division |
| $R_{\text{death}}$ | Rate of death |
| smFRET | Single molecule FRET |
| smTIRF | Single-molecule total internal reflection fluorescence |
| SMC | Structural maintenance chromosome |
| TAD | Topologically associating domain |
| TF | Transcription factor |

# 1 General Introduction

A cell's genome contains the genetic material of an organism in a collection of chromosomes. Chromatin is a mass of genetic material composed of DNA that condense to form these chromosomes that make up the genome. Its arrangement may seem chaotic, but it is far from random. Research has revealed that the genome may be organized through chromosomal folding in the form of DNA loops. This is performed with a loop-forming process, called extrusion, where the human genome consists of 10,000 loops [3]. This process is significant in understanding what determines which genes get expressed, or activated, in different cells. It ultimately influences the functions that the cell performs.

During interphase, chromatin is classified as either euchromatin or heterochromatin depending on its level of compaction [68]. Euchromatin has a less compact structure, and may be referred as chromatin in an open or sparse state. The chromatin fiber visually has an appearance of 'beads on a string' (Figure 1.1) where the beads represent nucleosomes and the string represents DNA.

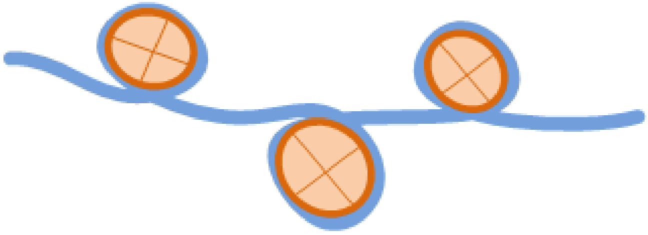

Figure 1.1: 'Beads on a string" illustration of chromatin. The orange circles depict the nucleosomes, and the blue 'string' depicts the strand of DNA.

Heterochromatin has a more compact structure, and may be considered chromatin in a closed, or dense state. As chromatin is compacted, DNA becomes less accessible for transcription factors. By loosening the chromatin, transcription machinery is better able to access the genomic DNA, and thus promote transcription [68]. Therefore, nucleosome organization and dynamics are regularly modified by the combined influence of covalent post-translational modifications, histone chaperones, ATP-dependent nucleosome remodelers, etc. The differential attraction between euchromatin and heterochromatin leads to phase separation and reproduces compartmentalization. Heterochromatin is phase separated from euchromatin, but not collapsed.

In this thesis, the possibility of modelling the chromosome compaction process with Markovian dynamics in different stages of the DNA looping mechanism is explored. The overall process may be observed through examining the conformational states of chromatin macroscopically. By examining the process more closely, other Markovian properties can be examined through the binding kinetics of loop extrusion factors. The specific contributions of these loop extrusion factors may be investigated through the development of a proper mathematical model to describe this process. Ultimately, the organizational process of chromatin will influence transcription activity, and gene expression. This work will contribute to understanding how information is stored in our DNA by the manner in which chromatin is structured.

## 1.1 Objectives and Aims

### 1.1.1 Research Question

The purpose of this research is to explore the mathematical processes involved in the chromatin compaction procedure. The specific loop extrusion factor properties are analyzed to determine the contributions that affect its looping behaviour.

### 1.1.2 Research Design

In this research, the methods of data analyses are emphasized for the chromatin loop compaction process. The perspective in which chromatin data is analyzed will allow for further mathematical applications to predict chromatin folding behaviour with several loop extrusion factors. This will allow for the promotion of future motivation into researching other mathematical methods to illustrate the behaviour of DNA.

To conduct this investigation, the chromatin looping process is modelled mathematically with the use of Markov properties. Single-molecule FRET experiments will be used to examine chromatin conformations. With the use of the Expectation-Maximization algorithm, these conformations may be more easily identified by employing Gaussian mixture models. The immigration-death process is then used to model the behaviour of reinforcing chromatin loops through stacking loop extrusion factors. The specificity of the binding behaviour of loop extrusion factors is also included to determine the information extruded in a loop. The

exposure timescale of DNA is also explored to determine the behaviour of LEF binding in the presence of nucleosomes on a strand of DNA. These results give just a few interpretations of chromosome organization and highlight how chromatin is a complex, active matter shaped by loop extrusion.

# 2 Mixture Models in Chromatin Compaction

## 2.1 Abstract

Gaussian mixture models can be used to predict the specific areas of single-molecule FRET data that represent different stages of chromatin folding states. We found that some understanding of the chromatin folding process is required to establish the necessary number of components for the model. The Gaussian mixture model is effective in conjunction with the use of the Expectation-Maximization algorithm to iteratively cluster the components of FRET data. However, with a better understanding of the steps required to fold chromatin into stable loops, the Gaussian mixture model can successfully identify the single-molecule FRET properties associated with each folding state.

## 2.2 Introduction

Chromatin is a complex system due to its molecular organization, heterogeneous structure, and multiscale dynamics induced by post-translational modifications on chromatin itself and other considerations, such as transcription factors. As previously determined [15], single molecule FRET experiments examine the interconversion kinetics of discrete tetranucleosome units and the impact on post-translational modifications by ubiquitylation on chromatin structure. These experiments are able to determine the conformational dynamics through single-molecule total internal reflection fluorescence (smTIRF) microscopy data sets.

The Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. This allows the data itself to dictate how many mixture components are required to model it, and provides a measure of the probability that some data may share common characteristics. It is characterized as a clustering algorithm, since it can be used to find distinct clusters in the data. The Förster (or Fluorescence) Resonance Energy Transfer (FRET) distribution of all, or a subset, of molecules from multiple movie acquisitions can be analyzed using Gaussian mixture distributions. The aim of this FRET distribution analysis is to determine the number of state components in the global data, their relative weights, and the FRET efficiency of these components. In this application, Gaussian mixture models are fitted to chromatin datasets to analyze how the Expectation-Maximization algorithm can identify some of its dynamic conformations.

## 2.3 Materials and methods

### 2.3.1 Single-molecule FRET

Förster (or Fluorescence) Resonance Energy Transfer (FRET) at the single molecule level focuses on the study of immobilized molecules that allow measurements of single molecule reaction trajectories from about 1 millisecond to many minutes. This is becoming an essential tool for characterizing proteins, signaling pathways or any biological phenomenon. Single-molecule FRET (smFRET) has rapidly developed to answer fundamental questions about replication, recombination, transcription, translation, RNA folding and catalysis, non-canonical DNA dynamics, protein folding and conformational changes, to name a few [60].

In the case of chromatin looping, smFRET is used to examine the evolution of nanometre-length scale conformational changes of protein-DNA and protein-protein complexes at the single-molecule level. In smFRET experiments, two fluorophores, or dyes rather, are placed at known positions on double-stranded DNA molecules with complementary overhangs (sticky ends) are immobilised onto the glass coverslip. Fluorescence signals from smFRET are observed when molecules are trapped in the looped state due to base pairing between the sticky ends. Looping and unlooping of DNA lead to fluorescence intensity fluctuations, where low FRET signals correspond to the unlooped state and high FRET signals correspond to the looped state [35]. Some interesting looping kinetic properties can also arise in these experiments, such as the looping probability density, looping rate, annealing rate, etc.

Briefly, FRET measures the extent of non-radiative energy transfer between two fluorescent dye molecules, termed donor and acceptor, and reports the intervening distance which can be estimated from the ratio of acceptor to total emission intensity. This efficiency of energy transfer, $E$, is given as $E = [1 + (R/R_0)^6]^{-1}$, where $R$ represents the inter-dye distance and $R_0$ is the Förster radius at which $E = 0.5$ [62]. Conformational dynamics of single molecules can be observed in real-time by tracking FRET changes. FRET is advantageous compared to other imaging techniques as it is a ratiometric method that enables the measurement of the internal distance in the molecular frame rather than in the laboratory frame. This prevents interference from instrumental noise and drift. The smFRET time trajectories are acquired by imaging surface immobilized molecules with the aid of total internal reflection miscroscopy that allows high throughput data sampling.

In this chapter, we utilize a two colour FRET scheme as performed by Kilic et al. [39], but it is noted that higher order FRET schemes can also be applied to probe multi-component interactions between conformational changes in large molecular complexes. In order to process the observed data to apply the GMM, iSMS software is used [55]. This single-molecule FRET microscopy software is used to study the structure and conformational dynamics of biomolecules where the data was collected in the form of a movie file. These movie files essentially show movies of the single molecule events of interest. iSMS integrates and automates common procedures in smFRET data analysis including: molecule localization, intensity-trace integration, quantitative FRET determination, FRET distribution analysis, molecule subpopulation analysis and transition state dynamics analysis [55]. In the present context of studying chromatin looping behaviour, the goal is to determine the number of looping conformations present in the observed dataset with the EM algorithm.

#### 2.3.1.1 SmTIRF data set

The observed chromatin data was obtained from smFRET experiments conducted previously by Kilic et.al. [39] to reveal structural states and their interconversion kinetics in chromatin fibers. Plasmids for chromatin DNA production were generated in DH5$\alpha$ cells grown in a medium and isolated by alkaline lysis followed by gel filtration. These plasmids were then prepared for flurorescent labelling in smFRET experiments at three distinct sets of internal positions yielding structural information from several vantage points. To begin investigating the conformational and dynamic properties of the assembled chromatin fibers with single molecule imaging, smTIRF is applied. With this type of imaging, it is possible to view complex biological interactions in real time. TIRF allows for visualization of single molecules by eliminating out-of-focus fluorescence and enhancing signal-to-noise ratio [42]. The excitation is restricted to a very thin section near the coverslip to achieve this precision for single molecule detection. This also reduces the photobleaching of fluorophores in solution and prevents harmful light damage when imaging live cells. For the purpose of examining the effectiveness of the GMM on chromatin data, only one out of the three dye configurations on Donor-Acceptor positions were utilized to demonstrate the performance of the GMM in identifying chromatin states. It is presumed that the GMM would identify fluorescent populations in the other two dye configurations in a similar manner. The DA2 labeled chromatin fibers in flow channels were immobilized and the donor and acceptor fluorescence emissions were measured. The time traces of FRET efficiency are then generated.

Chromatin fiber was examined in vivo with Alexa Fluor 568 as FRET donor and Alexa Fluor 647 as FRET acceptor. The goal of Kilic's experimentation [39] was to determine the extent of the impact of the heterochromatin protein 1$\alpha$ in establishing a compact chromatin state. Ultimately, this would be significant, as a compact chromatin state would contribute to gene silencing. Each dye pair was positioned in the center of the 12-mer nucleosome array to probe distinct contacts and motions. The DA2 dataset was used from Kilic et al's [39] work, where the bivalent cation of magnesium posseses a concentration of 0.5mM $Mg^{2+}$ is present in the system. This pair measured inter-nucleosome interactions closer to the dyad. The results for this particular dataset in the study showed that these conditions promoted structural dynamics in chromatin upon examining the fluctuations in the time traces of donor and acceptor fluorescence emission. This specific dataset was used due to its presentation, as the information can be interpreted more clearly for the purposes of our investigation. The GMM can still be applied to the other datasets in this study, but only this data was utilized here to better illustrate the effectiveness of the GMM.

### 2.3.2 EM algorithm

The Expectation-Maximization (EM) algorithm is presented prior to applying this to the smFRET dataset. It provides an iterative method in determining maximum-likelihood estimates for model parameters when data is incomplete, has missing data points, or has unobserved (hidden) latent variables. It is strongly dependent on the selection of initial values of model parameters and increases the values for the likelihood function at each iteration. As well, it produces a local rather than global maximum of the likelihood function [34].

Each iteration of the EM algorithm consists of two processes: the E-step, and the M-step. The expectation step (E-step), estimates the complete-data sufficient statistics by finding the missing data given the observed data, and current estimate of the model parameters. The maximization step (M-step), maximizes the likelihood function under the assumption that the missing data are known. This is where the missing data from the E-step is used to complete the observed dataset. Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

To begin the derivation of the EM algorithm, let $\mathbf{X}$ be a random vector which results from a parameterized family. The goal is to find $\theta$ such that $P(\mathbf{X}|\theta)$ is a maximum. This is referred to as the maximum likelihood estimate for $\theta$. The log likelihood function is ultimately required to estimate the parameter $\theta$,

$$L(\theta) = \ln P(\mathbf{X}|\theta). \tag{2.1}$$

The likelihood function is considered to be a function of the parameter $\theta$ given $\mathbf{X}$. Since $\ln(x)$ is a strictly increasing function, the value of $\theta$ which maximizes $P(\mathbf{X}|\theta)$ also maximizes $L(\theta)$.

The goal of the EM algorithm is to maximize $L(\theta)$, so we want to obtain an updated estimate for $\theta$ at each iteration such that,

$$L(\theta) > L(\theta_n), \tag{2.2}$$

where after the $n^{th}$ iteration the current estimate for $\theta$ is denoted by $\theta_n$. Equivalently, we want to maximize the difference,

$$L(\theta) - L(\theta_n) = \ln P(\mathbf{X}|\theta) - \ln P(\mathbf{X}|\theta_n). \tag{2.3}$$

At this point, we have not considered any unobserved or missing variables. We denote the hidden random vector by $\mathbf{Z}$. The total probability $P(\mathbf{X}|\theta)$ may now be written in terms of the hidden variables $\mathbf{z}$ as,

$$P(\mathbf{X}|\theta) = \sum_{\mathbf{z}} P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta). \tag{2.4}$$

This allows for Equation 2.3 to be re-written as,

$$L(\theta) - L(\theta_n) = \ln \sum_{\mathbf{z}} P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta) - \ln P(\mathbf{X}|\theta_n). \tag{2.5}$$

As observed, this involves the logarithm of a sum. Now by employing Jensen's inequality, we can obtain

$$\ln \sum_{i=1}^{n} \lambda_i x_i \geq \sum_{i=1}^{n} \lambda_i \ln(x_i), \tag{2.6}$$

for constants $\lambda_i \geq 0$ with $\sum_{i=1}^{n} \lambda_i = 1$. Jensen's inequality is a general result in convexity. It states that for a convex function $f$, if $\lambda \in [0, 1]$, then

$$\lambda f(x) + (1-\lambda)f(y) \geq f(\lambda x + (1-\lambda)). \tag{2.7}$$

Ultimately, we can generalize the result to expectation: $\mathbb{E}[f(\mathbf{X})] \geq f([\mathbf{X}])$

Further from Equation 2.5, the constants $P(\mathbf{z}|\mathbf{X},\theta_n)$ are denoted as,

$$\begin{aligned}
L(\theta) - L(\theta_n) &= \ln \sum_{\mathbf{z}} P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta) - \ln P(\mathbf{X}|\theta_n) \\
&= \ln \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X},\theta_n)\Big(\frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_n)}\Big) - \ln P(\mathbf{X}|\theta_n) \\
&\geq \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X},\theta_n)\ln\Big(\frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_n)}\Big) - \ln P(\mathbf{X}|\theta_n) \\
&= \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X},\theta_n)\ln\Big(\frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_n)P(\mathbf{X}|\theta_n)}\Big) \\
&\triangleq \Delta(\theta|\theta_n)
\end{aligned} \tag{2.8}$$

This can be simplified and rearranged to write $L(\theta) \geq L(\theta_n) + \Delta(\theta|\theta_n)$. However, to further simplify this, let $l(\theta|\theta_n) \triangleq L(\theta_n) + \Delta(\theta|\theta_n)$. By combining these equations, this states

$$L(\theta) \geq l(\theta|\theta_n) \tag{2.9}$$

The function $l(\theta|\theta_n)$ is bounded above by the likelihood function $L(\theta)$. Additionally, observe that,

$$\begin{aligned} l(\theta_n|\theta_n) &= L(\theta_n) + \Delta(\theta_n|\theta_n) \\ &= L(\theta_n) + \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta)\ln\frac{P(\mathbf{X}|\mathbf{z}, \theta_n)P(\mathbf{z}|\theta_n)}{P(\mathbf{z}|\mathbf{X}, \theta_n)P(\mathbf{X}|\theta_n)} \\ &= L(\theta_n) + \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_n)\ln\frac{P(\mathbf{X}|\mathbf{z}, \theta_n)}{P(\mathbf{X}|\mathbf{z}, \theta_n)} \qquad (2.10) \\ &= L(\theta_n) + \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_n)\ln 1 \\ &= L(\theta_n). \end{aligned}$$

This implies that for $\theta = \theta_n$ the functions $l(\theta|\theta_n)$ and $L(\theta)$ are equal.

With the EM algorithm, we aim to choose values of $\theta$ such that the likelihood function, $L(\theta)$, is maximized. The function $l(\theta|\theta_n)$ is bounded above by the likelihood function and the values of the functions $l(\theta|\theta_n)$ and $L(\theta)$ are equal at the current estimate for $\theta = \theta_n$. Therefore, any $\theta$ which increases $l(\theta|\theta_n)$ in turn increases the likelihood function. In order to achieve the greatest possible increase in the value of $L(\theta)$, the EM algorithm calls for selecting $\theta$ such that $l(\theta|\theta_n)$ is maximized. We denote this updated value as $\theta_{n+1}$. An illustration of a single iteration of the EM algorithm is provided in Figure 2.1, as given previously [9]. The

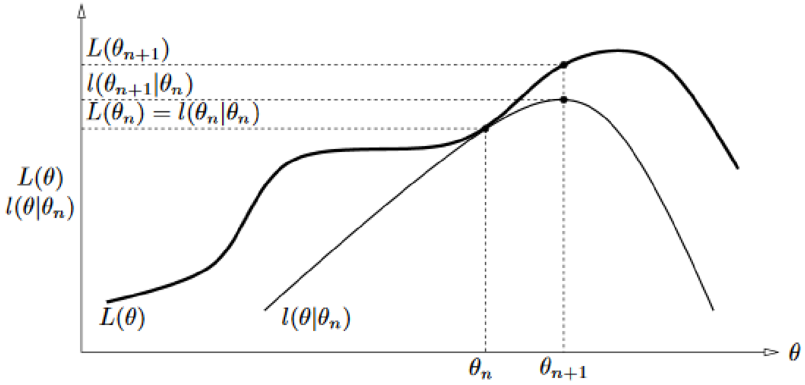

Figure 2.1: Single iteration of the EM algorithm. Figure obtained from literature [9]

value of $\theta_{n+1}$ may be derived such that,

$$\begin{aligned}\theta_{n+1} &= \underset{\theta}{\text{argmax}}\{l(\theta|\theta_n)\} \\ &= \underset{\theta}{\text{argmax}}\left\{L(\theta_n) + \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_n)\text{ln}\frac{P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X}, \theta)P(\mathbf{X}|\theta)}\right\} \\ &= \underset{\theta}{\text{argmax}}\left\{\sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_n)\text{ln}P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)\right\} \\ &= \underset{\theta}{\text{argmax}}\left\{\sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_n)\text{ln}\frac{P(\mathbf{X}, \mathbf{z}, \theta)}{P(\mathbf{z}, \theta)}\frac{P(\mathbf{z}, \theta)}{P(\theta)}\right\} \\ &= \underset{\theta}{\text{argmax}}\left\{\sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_n)\text{ln}P(\mathbf{X}, \mathbf{z}|\theta)\right\} \\ &= \underset{\theta}{\text{argmax}}\left\{\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_n}\{\text{ln}P(\mathbf{X}, \mathbf{z}|\theta)\}\right\}\end{aligned} \tag{2.11}$$

It is observed from Equation 2.11 that the E-step and the M-step of the EM algorithm have been fulfilled. In the E-step, the conditional expectation $Q(\theta, \theta_n) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_n}\{\text{ln}P(\mathbf{X}, \mathbf{z}|\theta)\}$ is determined. In the M-step, this expression was maximized with respect to $\theta$, where $\hat{\theta} = \text{argmax}_\theta\, Q(\theta, \theta_n)$.

Note that for the EM algorithm, the log likelihood function gives a formal measure of how well a particular parameter, $\theta$ fits the observed sample. This represents a function of both the observed data and the desired parameters for the model. The higher the log likelihood, the higher the probability is assigned under the model to the observation data. If we could efficiently search for $\hat{\theta} = \text{argmax}_\theta\, l(\theta)$, then this would result in the "correct" parameters to represent the data.

When running the EM algorithm, the procedure assures that learning occurs at each iterate. In order for it to stop learning, the stopping criterion is established,

$$\frac{|l(\theta^{n+1}) - l(\theta^n)|}{|l(\theta^{n+1})|} < \epsilon, \tag{2.12}$$

where $\epsilon$ is a preset threshold.

This completes the derivation of the EM algorithm. Its application in Gaussian mixture models will now be explained to further its use for parameter estimation in genetic datasets.

### 2.3.3 EM application: Gaussian Mixture Models

The Gaussian mixture model (GMM) is a finite mixture probability distribution model, where the generic probability density function is a weighted sum of independent processes that adds to a total density function with a total area of 1. The GMM is presented in this case to find clusters in a dataset from which we know, or assume to know, the number of clusters in the dataset, but we do not know whether these clusters are as well as how they are shaped.

To begin with the derivation of the GMM, consider a univariate Gaussian distribution, with mean $\mu$ and variance $\sigma$,

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{exp}\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{2.13}$$

For the multivariate distribution, rather than an equation with variance, this considers covariance, $\Sigma$,

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi|\Sigma|)}}\mathrm{exp}\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \tag{2.14}$$

From Equation 2.14, the parameters can be estimated by using Maximum Likelihood Estimation (MLE). The log of the multivariate Gaussian distribution is considered,

$$\mathrm{log}p\big(x|\mu,\sum\big) = -\frac{1}{2}\mathrm{log}(2\pi) - \frac{1}{2}\mathrm{log}\big|\sum\big| - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \tag{2.15}$$

Note that by applying the log function to the likelihood, this aids in decomposing the product and removing the exponential function so that the MLE can be more easily solved. In order to perform the MLE, the gradient of Equation 2.15 is taken with respect to each the mean and the covariance. The gradient with respect to the mean, $\mu$, is determined,

$$\begin{aligned}\frac{\partial \mathrm{log}p(x|\mu,\Sigma)}{\partial\mu} &= 0\\ \mu_{ML} &= \frac{1}{N}\sum_{n=1}^{N} x_n.\end{aligned} \tag{2.16}$$

The MLE of the mean was denoted by $\mu_{ML}$ and $N$ is the number of samples or data points. This will be needed for the maximum likelihood of the covariance, $\Sigma$. The gradient of Equation 2.15 is now determined with respect to $\Sigma$,

$$\begin{aligned}\frac{\partial \mathrm{log}p(x|\mu,\Sigma)}{\partial\Sigma} &= 0\\ \Sigma_{ML} &= 1N\sum_{n=1}^{N}(x_n-\mu_{ML})(x_n-\mu_{ML})^T\end{aligned} \tag{2.17}$$

These MLE derivations can also be applied to Gaussian mixture models. The GMM can be visualized as Gaussian distributions centered at different means. The EM algorithm is applied to the GMM to better search for the optimal parameters of the mixture distribution. This search is initialized with a clustering algorithm to identify areas of the dataset to start identifying the central means for the GMMs. Once these clusters are identified, the GMM can be optimized. The Gaussian mixture distribution can be written as a linear superposition of Gaussians in the form,

$$p(x_{1:N}|c_{1:N}) = \prod_n \mathcal{N}(x_n|\mu_n, \sigma^2 I), \tag{2.18}$$

where $c$ refers to the cluster for data point $n$, as identified from the initialization step. Let the covariance be denoted by $\sigma^2 I$, where $I$ is the unit matrix. However, instead of using $c_n$, let $z_n \in \{0,1\}^K$ be an indicator vector for data point $n$. The model for the GMM is generated by choosing among one of the $k$ clusters for each data point, each with probability $\pi_k$. Then the data point is generated by a Gaussian centered at $\mu_k$. This is portrayed by the following,

$$\begin{aligned} p(x_{1:N}, z_{1:N,1:K}|\mu_{1:K}, \sigma^2, \pi) &= \prod_n \prod_k \{\pi_k \mathcal{N}(x_n|\mu_k, \sigma^2 I)\}^{z_{n,k}} \\ &= \prod_n \prod_k \left\{\pi_k (2\pi(\sigma^2)^d)^{-1/2} \text{exp}\left[-\frac{1}{2\sigma^2}||x_n - \mu_k||^2\right]\right\}^{z_{n,k}}, \end{aligned} \tag{2.19}$$

where $\pi$ denotes the weight of each distribution. Equation 2.19 can now be used to get the likelihood of the data by summing over the unknown $z$ vector,

$$\begin{aligned} p(x|\mu, \sigma^2, \pi) &= \sum_{z_{1:N,1:K}} \prod_n \prod_k \left\{\pi_k (2\pi\sigma^{2d})^{-1/2} \text{exp}\left[-\frac{1}{2\sigma^2}||x_n - \mu_k||^2\right]\right\}^{z_{n,k}} \\ &= \prod_n \sum_{z_{n,1:K}} \prod_k \left\{\pi_k (2\pi\sigma^{2d})^{-1/2} \text{exp}\left[-\frac{1}{2\sigma^2}||x_n - \mu_k||^2\right]\right\}^{z_{n,k}} \end{aligned} \tag{2.20}$$

The log likelihood of this equation then becomes,

$$\text{log}p(x|\mu, \sigma^2, \pi) = \sum_n \text{log} \sum_{z_n} \prod_k \left\{\pi_k (2\pi\sigma^{2d})^{-1/2} \text{exp}\left[-\frac{1}{2\sigma^2}||x_n - \mu_k||^2\right]\right\}^{z_{n,k}} \tag{2.21}$$

The complete log likelihood can then be determined, which implies that $z$ is known,

$$\text{log}p(x, z|\mu, \sigma^2, \pi) = \sum_n \sum_k \{\text{log}\pi_k + \text{log}\mathcal{N}(x_n|\mu_k, \sigma^2 I)\} \tag{2.22}$$

However, when $z$ is not known, this is represented by the incomplete log likelihood. The EM algorithm can now be employed to iteratively refine the initial guesses for $z$. The $z$ vector will satisfy $z_{n,k} \geq 0$ for all $n, k$ and $\sum_k z_{n,k} = 1$ for all $n$. The values for $z$ can be predicted by taking the expectation of the following by using Bayes' theorem,

$$\begin{aligned}
\mathbb{E}_{p(z|x,\mu,\sigma^2,\pi)} z_{n,k} &= 1 \times p(z_{n,k} = 1|x_n, \mu, \sigma^2, \pi) + 0 \times p(z_{n,k} = 0|x_n, \mu, \sigma^2, \pi) \\
&= p(z_n, k = 1|x_n, \mu, \sigma^2, \pi) \\
&= \frac{p(x_n|z_{n,k} = 1, \mu, \sigma^2)p(z_{n,k} = 1|\pi)}{\sum_{k'} p(x_n|z_{n,k'} = 1, \mu, \sigma^2)p(z_{n,k'} = 1|\pi)} \\
&= \frac{\mathcal{N}(x_n|\mu_k, \sigma^2)\pi_k}{\sum_{k'} \mathcal{N}(x_n|\mu_{k'}, \sigma^2)\pi_{k'}}
\end{aligned} \tag{2.23}$$

This represented the "Expectation step", or E-step of the algorithm, where the expected values for the latent variable are estimated. This solves the inference problem to essentially determine which Gaussian generated each datapoint. This is a distribution over all probabilities because we cannot be sure. Now, Equation 2.23 can be used to maximize the complete data log likelihood with respect to $\mu$ and $\sigma^2$. In order to do this, the gradient of the complete likelihood with respect to $\pi$, $\mu$, and $\sigma^2$ is required.

First, the maximization of $\pi$ is done, where the constraint which the weights sum to one must be obeyed. In order to do this, a Lagrange multiplier is needed, which gives an augmented likelihood function,

$$\sum_n \sum_k z_{n,k} \text{log}\pi_k - \lambda \left( \sum_k \pi_k - 1 \right) \tag{2.24}$$

Equation 2.24 is then differentiated with respect to $\pi_k$ to obtain,

$$\sum_n \frac{z_{n,k}}{\pi_k} - \lambda = \sum_n z_{n,k} - \lambda \pi_k = 0 \tag{2.25}$$

Finally, by summing over all $K$, we see that $\lambda = \sum_n \sum_k z_{z,k} = N$, so

$$\pi_k = \frac{1}{N} \sum_n z_{n,k}, \tag{2.26}$$

which completes the "Maximization step", or M-step, partially. The maximized values for both $\mu_k$ and $\sigma^2$ must also be determined.

Now, the maximization of $\mu_k$ is performed. This does not have any constraints, therefore the Lagrange multiplier is not required. The gradient of the complete log likelihood is done with respect to $\mu_k$,

$$\begin{aligned}
\nabla_{\mu_k} \text{log} p(x, z|\mu, \sigma^2, \pi) &= \sum_n z_{n,k} \nabla_{\mu_k} \text{log} \mathcal{N}(x_n|\mu_k, \sigma^2 I) \\
&= \sum_n z_{n,k} \nabla_{\mu_k} \frac{-1}{2\sigma^2} ||x_n - \mu_k||^2 \\
&= -\sum_n z_{n,k} \frac{1}{\sigma^2}(x_n - \mu_k)
\end{aligned} \tag{2.27}$$

In order to solve this, Equation 2.27 is equated to zero to give,

$$\begin{aligned} \sum_n z_{n,k} \frac{1}{\sigma^2}(\mu_k - x_n) &= 0 \\ \sum_n z_{n,k} \frac{1}{\sigma^2} \mu_k &= \sum_n z_{n,k} \frac{1}{\sigma^2} x_n \\ \mu_k \sum_n z_{n,k} &= \sum_n z_{n,k} x_n \\ \mu_k &= \sum_n \frac{z_{n,k}}{\sum_{n'} z_{n',k}} x_n \end{aligned} \tag{2.28}$$

Which completes the maximization step for $\mu_k$. Finally, $\sigma^2$ can be maximized, as well.

$$\begin{aligned} \nabla_{\sigma^2} \mathrm{log} p(x, z|\mu, \sigma^2, \pi) &= \sum_n \sum_k z_{n,k} \nabla_{\sigma^2} \mathrm{log} \mathcal{N}(x_n|\mu_k, \sigma^2) \\ &= \sum_n \sum_k z_{n,k} \nabla_{\sigma^2} \left[ -\frac{d}{2} \mathrm{log}(\sigma^2) - \frac{1}{2\sigma^2} ||x_n - \mu_k||^2 \right] \\ &= \sum_n \sum_k z_{n,k} \left[ -\frac{d}{2\sigma^2} + \frac{1}{2\sigma^4} ||x_n - \mu_k||^2 \right] \end{aligned} \tag{2.29}$$

Again, by setting Equation 2.29 to zero, we can obtain

$$\begin{aligned} \sum_n \sum_k z_{n,k} \frac{d}{2\sigma^2} &= \sum_n \sum_k z_{n,k} \frac{1}{2\sigma^4} ||x_n - \mu_k||^2 \\ \frac{d}{\sigma^2} \sum_n \sum_k z_{n,k} &= \frac{1}{\sigma^4} \sum_n \sum_k z_{n,k} ||x_n - \mu_k||^2 \\ dN\sigma^2 &= \sum_n \sum_k z_{n,k} ||x_n - \mu_k||^2 \\ \sigma^2 &= \frac{1}{dN} \sum_n \sum_k z_{n,k} ||x_n - \mu_k||^2 \end{aligned} \tag{2.30}$$

Which completes all components of the M-step.

Formally, the EM-GMM can be summarized in the following algorithm,

1. Initialize cluster centers $\mu_{1:K}$, $\pi$ and $\sigma^2$
2. E step:
   $\mathbb{E}_{p(z|x,\mu,\sigma^2,\pi)} z_{n,k} = \frac{\mathcal{N}(x_n|\mu_k,\sigma^2)\pi_k}{\sum_{k'} \mathcal{N}(x_n|\mu_{k'},\sigma^2)\pi_{k'}}$

3. M step:
   $\pi_k = \frac{1}{N} \sum_n z_{n,k}$
   $\mu_k = \sum_n \frac{z_{n,k}}{\sum_{n'} z_{n',k}} x_n$
   $\sigma^2 = \frac{1}{dN} \sum_n \sum_k z_{n,k} ||x_n - \mu_k||^2$
4. Evaluate log likelihood until convergence is achieved, otherwise return to step 2
   $\log p(x, z|\mu, \sigma^2, \pi) = \sum_n \sum_k \{\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \sigma^2 I)\}$

The fourth step is repeated until the convergence criteria is met. As given in Equation 2.12, learning is stopped when the log likelihood is below a certain preset threshold.

### 2.3.4 Previous applications of the EM mixture model in enzyme kinetics

The concept of applying the EM algorithm with mixture models for enzyme data is not new, as this has been done in several different applications. This is not restricted to only Gaussian mixture models, as it can be used as a tool to cluster data and identify significant trends.

A Bayesian mixture model was introduced previously as a method of clustering when the number of components is unknown[58]. In this model, the number of components and the mixture component parameters are first modelled jointly. Then, a base inference about these quantities on their posterior probabilities is made. The posterior distributions of the target's objects of inference is then presented, and not just the 'best estimates'. The concept in which a sample-based approach can be used to compute mixture models allowed for a more subtle extraction of information from posterior distributions. It was determined that the experiments and discussion conducted utilized a hierarchical model for mixtures that aim to provide a simple and generalizable way of being weakly informative about parameters of mixture models. Ultimately, Bayesian mixture models have previously been used successfully to determine the number of unknown components present in a sample. This may also be done as an alternative to applying the GMM. While the Bayesian approach typically offers a more apealing strategy of determining the number of components in a model, it may not be representative of a "true" number of clusters. The Bayesian approach allows an unbounded number of components. As data gets larger, the number of components to the model will naturally increase. In this case, since we are investigating the behaviour of the Gaussian mixture model on the smFRET data, the Bayesian mixture model will not be explored. We are interested in examining a small number of clusters in the model, so the GMM should suffice in identifying the overall conformations of chromatin present in the model. However, one may be able to explore whether the Bayesian mixture model will identify the correct components of the model in the future.

Mixture models have also been explored in Birnbaum-Saunders distributions with $G$ components, to improve upon previous work where there were only two components considered [5].

The identifiability of the model with $G$ components was proven and an EM algorithm for the maximum likelihood estimation of the mixture parameters was developed. This is where the $k$-bumps algorithm was used as an initialization strategy in the EM algorithm. In the $k$-bumps algorithm, the $k$ bumps are detected in the observed data. The maximum point for each bump is determined, where the maximum point is the mode for each bump. Each observation is assigned to the cluster with the closest maximum point. The maximum point and the $k$ bump can then be used to calculate the initial value of the shape parameter, which completes the clustering algorithm [5]. This implies there has been some exploration in the impact of the initialization values for the EM algorithm. It was determined that mixture models can provide a multimodal log likelihood function, which suggests that the method of maximum likelihood estimation through the EM algorithm may not give maximum global solutions if the starting values are far from the real parameters. Based on this work, good initial values are necessary to hasten or enable the convergence in the EM algorithm. Therefore, the importance of the initialization parameters were determined with $k$-means as a method of estimating the initial starting conditions, to minimize the uncertainty in the initial search values.

The speed of convergence for the EM algorithm has also been previously examined [50], and concluded that convergence is dependent on the amount of overlap among the mixture components. When the mixture components exhibit some overlap, the convergence of EM in GMMs become slower as the dynamic range among the mixing coefficients increases. This can be examined by running the EM algorithm and observing its impact when changing the stopping criterion. The number of components in the model may also be shown to be significant, as this adds more complexity to the model. Therefore, choosing a proper stopping criteria will impact the confidence in the output. The number of components actually present in the model must also be addressed, for fear of obtaining inaccurate results.

A mixture model with detection limits was previously proposed to obtain regression analyses of antibody response to vaccines [48]. Standard analyses typically assume the data arise from a single lognormal response distribution, but this may not always be considered appropriate, as more observations are censored than would be expected under such a model. A mixture model with censored lognormal distribution and a point distribution located below the detection limit was proposed for these situations. It was concluded in this study that this could have been modelled by two lognormal distributions. However, more informative regression analyses were required to explain the mixing of the behaviour of the data to appropriately model the distributions. In our model, lognormal distributions can also be considered as possible distributions in the mixture. However, we focus on the GMM and its applicability to this dataset to test its applicability to the different components of data. It is acknowledged that other distributions may be more appropriate as mixtures in this dataset.

The concept of applying the EM algorithm in enzyme data was also not limited to vaccines, in that Gaussian mixture models have previously been applied to chromatin interaction mapping studies [71]. In other words, Gaussian mixture models have been applied to model the behaviour of chromatin. To be more specific, they were used to map topologically associating domains and subdomains in the genome. This provided a further understanding of the three dimensional structure-function relationship of the genome. The use of Gaussian mixture models was advantageous for this publication [71], since it allows for the flexibility of modeling a wide range of probability distributions. The input to the algorithm was a normalized Hi-C contact matrix. By fitting a 2-component Gaussian mixture model to the count matrix, the contacts are distinguished within a chromatin domain from contacts that are outside of a chromatin domain. This serves to reduce the amount of noise in the normalized Hi-C data. Mixture models have also been previously applied for chromatin interaction data [52], where the dependency among ChIA-PET data (count of DNA fragment pairs) and the available information on protein binding sites and gene promoters was accounted for to reduce false positives. The EM algorithm is employed in this case to identify the optimal parameters for both determining the number of state components in the global data, and to maximize the difference in the proportion of intra-domain contacts in putative domains and outside of putative domains. The EM algorithm has been shown to be successful in optimizing parameters for the application of GMMs in chromatin interaction data, so it should successfully be transferable in this investigation. In our research, this raised the question of how many components should be used to model the dataset. This was investigated in terms of the known steps in chromatin organization, and will be further discussed.

The use of Gaussian mixture models in chromatin interaction data has widely been used in biological analyses of chromatin data. In this case, applying GMMs to chromatin data will allow for a better understanding of the macroscopic properties associated with identifying chromatin states in this type of data. We use the GMM in a unique manner such that is tested against some known steps in the chromatin folding process. The significance of the central means identified in the GMM in this application will give insight toward how the data supports these configurations.

## 2.4 Results

### 2.4.1 Gaussian Mixture Model Application: Chromatin dynamics

The analysis of individual, noisy signal trajectories has been greatly facilitated by the use of hidden Markov models (HMMs). This has been introduced within the context of patch-clamp experiments on ion channels, and have since been applied within a variety of single-molecule experimental platforms, including smFRET experiments. In HMM approaches, a statistical model defines an expected distribution of measurement values in terms of a set of parameters, such as the centers and widths of Gaussian peaks representing the signal values associated with each conformation state, and the transition probabilities between states. Given this model, the EM algorithm is applied to determine the maximum likelihood of the parameters and conformational trajectory for each measured signal trajectory.

After conducting an extensive literature search, the number of chromatin states that can summarize the loop formation procedure can be summarized in either two [71], four [61], or five states [59] depending on the imaging detection methods that were made available in these studies to measure chromatin looping activity. Since all of these methods are accepted to measure chromatin activity, the GMM will optimize the data based on a 2-component, 4-component, and a 5-component model. While there may be other conformation states that may occur, these have already been identified in the available literature.

#### 2.4.1.1 Initialization

In order to initialize the EM algorithm for its initial search parameters, the $k$-means algorithm was applied to the observed dataset. The importance of a good estimate for the initial parameters were previously emphasized by Benites, et. al [5]. It was stated that this affected the search with the EM algorithm such that it may determine whether the achieved result may be either a local or global maximum. $K$-means is one of the most popular clustering algorithms. This method stores $k$ centroids that is used to define clusters. A point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid.

$K$-means finds the best centroids by alternating between assigning data points to clusters based on the current centroids, and choosing centroids (points which are the center of a cluster) based on the current assignment of data points to clusters. The $k$-means algorithm partitions $n$ objects into $k$ clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly $k$ different clusters. The optimal number of $k$ clusters leading to the greatest separation is not known and must be computed from the given data. First, the number of cluster centroids are initialized randomly. Then, the following equation is used to determine the total Euclidean distances between each data point and the corresponding centroid.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2, \tag{2.31}$$

where $J$ is the objective function with the total distances. In Equation 2.31, $k$ represents the number of clusters, $n$ represents the number of points belonging to cluster $j$, $x_i$ represents the specific point $i$, and $c_j$ is the centroid of cluster $j$. The new centroid of each cluster is now defined by calculating the mean of all points assigned to that cluster,

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{2.32}$$

This procedure is repeated from the initialization step before employing Equation 2.31 until the positions of the centroids no longer move and the assignments stay the same. Using the $k$-means algorithm to initialize the EM algorithm will allow for the search of optimized parameters and ultimately achieve convergence. The EM algorithm may now be implemented for a mixture model with the observed dataset provided by Kilic et. al. [39].

#### 2.4.1.2 Implementation

As mentioned, the observed dataset used was obtained from a previous publication [39], where they analyzed chromatin structure dynamics. The smTIRF data was deposited, and made available publicly at www.zenodo.org under the accession code 1069675.

To analyze the smTIRF dataset, the FRET donor ($F_D$) and acceptor fluorescence emission intensity ($F_A$) traces FRET efficiency ($E_{\text{FRET}}$) traces, were calculated using the following equations,

$$E_{\text{FRET}} = \frac{F_A - \beta F_D}{F_A - \beta F_D + \gamma F_D} \text{ and } \gamma = \frac{\Delta F_{A,bleach}}{\Delta F_{D,bleach}} \tag{2.33}$$

The values of $\beta = 0.141$ and $\gamma = 0.468$ were experimentally determined for the dye pair Alexa568/647 in this dataset. The bin size for the histogram was set to 0.02, as previously discussed in the original publication [39]. The final histogram was then fitted using a 2-component Gaussian function. This was performed using iSMS [55], and then extracted to modify the raw data to fit the EM GMM with R software. The code to perform this fit is available in the Appendix.

The results of the implementation of the EM algorithm with a 2-component Gaussian mixture model is given in Figure 2.2, where the histogram depicts the raw data count density of FRET populations against FRET emissions, and the red line depicts the EM GMM fit to the data. Numerically, the mean ($\mu$), standard deviation ($\sigma$), and weight ($\pi$) of each Gaussian distribution fit from Figure 2.2 is given in Table 2.1. A summary of the EM results is given in Table 2.2.

Table 2.1: Parameters from smTIRF- FRET data for 2-component model

| k | $\mu$ | $\sigma$ | $\pi$ |
|---|---|---|---|
| 1 | -0.00833 | 0.0485 | 0.707 |
| 2 | 0.605 | 0.200 | 0.293 |

Table 2.2: Results from smTIRF- FRET data for 2-component model

| Iterations | Log-likelihood | Tolerance |
|---|---|---|
| 23 | 607.57 | 1e-08 |

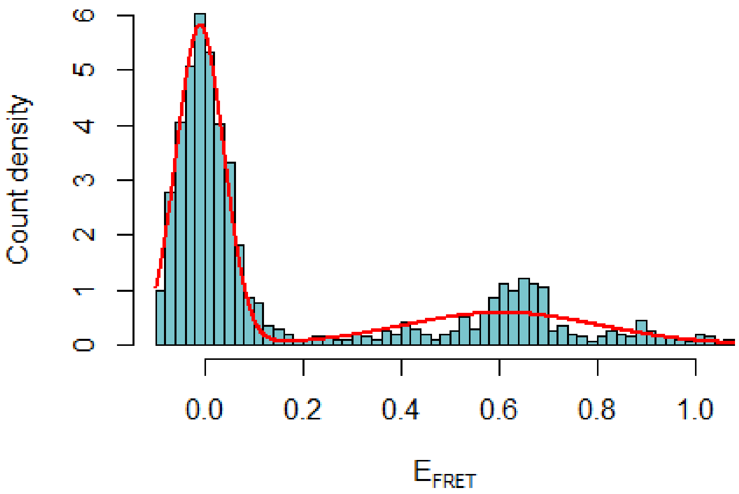

Figure 2.2: FRET populations with 2-component Gaussian Distribution Fit

It is observed that this EM algorithm run took 23 iterations at a stopping criterion of 1e-08. The log-likelihood value was returned for comparison of other tested models. However, upon examining Figure 2.2 visually, one may wonder if this could be fit better by perhaps adjusting the stopping criterion to examine the effectiveness of the EM algorithm on the 2-component GMM. The stopping criterion may also affect the speed of convergence [50]. A larger stopping criterion, or tolerance, is now implemented with 1e-01. A summary of the Gaussian fit parameters is given in Table 2.3, and the corresponding results of the EM implementation is given in Table 2.4.

Table 2.3: Parameters from smTIRF- FRET data for adjusted 2-component model

| k | $\mu$ | $\sigma$ | $\pi$ |
|---|---|---|---|
| 1 | -0.00847 | 0.0484 | 0.706 |
| 2 | 0.604 | 0.201 | 0.294 |

Table 2.4: Results from smTIRF- FRET data for adjusted 2-component model

| Iterations | Log-likelihood | Tolerance |
|---|---|---|
| 10 | 607.56 | 1e-01 |

It is observed that both Tables 2.1 and 2.3 have yielded similar results in terms of the Gaussian fit parameter values, therefore Figure 2.2 will remain the same shape. However, what is interesting to note is that the log-likelihood value that returned was also similar to that of Table 2.2, but it has converged in fewer iterations. Since this has affected the number of iterations, it seemed that by having a larger tolerance, the problem required fewer iterations to converge. Therefore, one may wonder if the converse is true; whether a smaller tolerance will cause the problem to use even more iterations to achieve convergence. To examine this outcome, a smaller tolerance of 1e-10 was used to run the EM algorithm again. The results of the Gaussian fit is given in Table 2.5, and the results of the EM algorithm run is given in Table 2.6.

Table 2.5: Parameters from smTIRF- FRET data with smaller tolerance

| k | $\mu$ | $\sigma$ | $\pi$ |
|---|---|---|---|
| 1 | -0.00833 | 0.0485 | 0.707 |
| 2 | 0.605 | 0.200 | 0.293 |

Table 2.6: Results from smTIRF- FRET data with smaller tolerance

| Iterations | Log-likelihood | Tolerance |
|---|---|---|
| 23 | 607.57 | 1e-10 |

It was observed that the results of the GMM parameters with the smaller tolerance still yielded similar results to the original run in Table 2.1, and the fit to Figure 2.2 remains unchanged. However, there were more iterations performed with a smaller established stopping criterion, as predicted. In addition, similar to the previous run, the log-likelihood that returned remains unchanged. By comparing the three models with their respective log-likelihood values, it seems that it does not matter how long it takes for the EM algorithm to achieve convergence, as they have all yielded the same result. In order to minimize the computations made by the computer, the larger tolerance could have been chosen for this problem, and would still achieve accurate results.

In addition, from both 2-component model results, it is noted that there was one peak where $\mu$ returned a negative value. Negative efficiency values indicate that the fluorescence of the donor in the prescence of the acceptor is enhanced instead of quenched. This demonstrates the weakness of indirect measurements of FRET. The method is only stable with absolutely repeatable detection of the donor signals before and after adding acceptor and thus, between two different samples. Small changes in the fluorescence, whether noise- or sample-induced, can dramatically alter the efficiency and yield nonsensical negative values [6]. The negative FRET emission signals will be discussed further in Section 2.5. This implied that more of the data was skewed toward lower values of $E_{\text{FRET}}$. This should be further investigated by introducing more components to the model, to verify its validity in this dataset. This suggests that the model may still be improved by examining the log-likelihood values, and the values of $\mu$ that returned. Therefore, either more components should be added to the model, or the initial starting parameters should be modified. In this case, the model was already initialized with the $k$-means algorithm with the original intention of minimizing this problem beforehand. To modify the initial search parameters, a different algorithm could then be used to mitigate this uncertainty. However, we continue this investigation by adding more components, as this may ultimately affect the initial search parameters by splitting up the dataset.

The number of components were now increased to four to model the dataset. This result is illustrated in Figure 2.3.

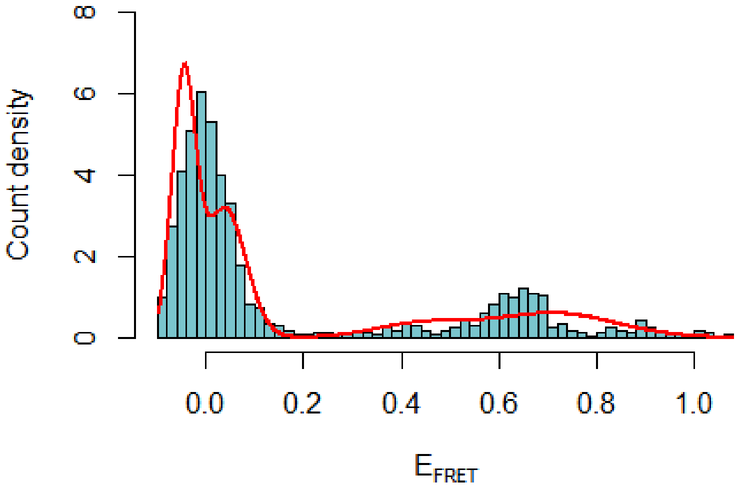

Figure 2.3: FRET populations with 4-component Gaussian Distribution Fit

It is observed that at lower $E_{\text{FRET}}$ emission values, there is more distinction in the peak than what was identified from the 2-component model. By running the EM algorithm with the 4-component model, there are some similarities observed between the two peaks. A summary of the results of this model is presented in Tables 2.7 and 2.8.

It is noted that there are now more values tending toward the lower $E_{\text{FRET}}$ values. In order to add more confidence to the interpretation of this model, one more component is added to test which component model is better suited to represent the dataset. A five component

Table 2.7: Parameters from smTIRF- FRET data for 4-component model

| k | $\mu$ | $\sigma$ | $\pi$ |
|---|---|---|---|
| 1 | -0.007 | 0.037 | 0.508 |
| 2 | -0.065 | 0.019 | 0.111 |
| 3 | 0.065 | 0.060 | 0.100 |
| 4 | 0.623 | 0.181 | 0.282 |

Table 2.8: Results from smTIRF- FRET data for 4-component model

| Iterations | Log-likelihood | Tolerance |
|---|---|---|
| 673 | 631.28 | 1e-08 |

model is now fitted to identify whether this will determine another significant Gaussian peak in the lower $E_{\text{FRET}}$ emissions values. The values of $\mu$ are also compared to determine whether these results will output similar trends in their results, in that they may identify values in similar regions.

After increasing the number of components and visually examining the fit compared with the observed data, the GMM was modelled with five components. This is illustrated in Figure 2.4.

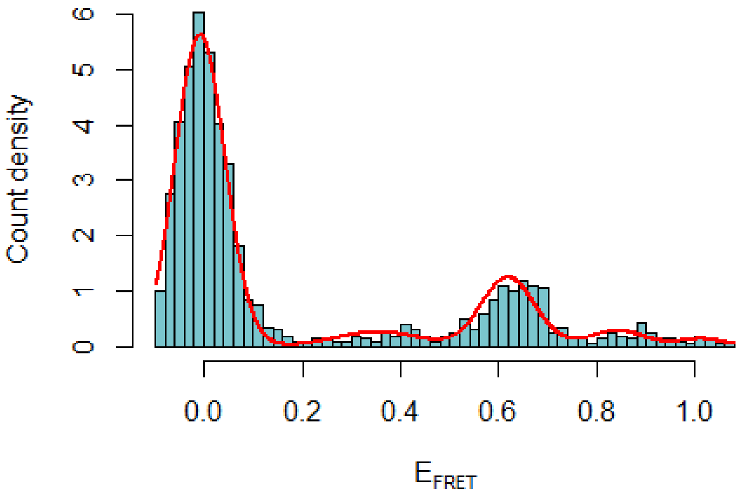

Figure 2.4: FRET populations with 5-component Gaussian Distribution Fit

The GMM parameters for the 5-component model is summarized in Table 2.9. By comparing the results from the previous models to the 5-component model, it is observed that the original locations of the mean $\mu$ has shifted. The weight $\pi$ parameter has also significantly decreased. The EM algorithm results for this GMM has also been provided in Table 2.10. As expected, the log-likelihood value has increased, and is bigger than that of both the 2-component and 4-component models. This implies that the 5-component GMM is generally a better model for this dataset.

Table 2.9: Parameters from smTIRF- FRET data for 5-component model

| k | $\mu$ | $\sigma$ | $\pi$ |
|---|---|---|---|
| 1 | -0.011 | 0.035 | 0.473 |
| 2 | -0.068 | 0.018 | 0.098 |
| 3 | 0.049 | 0.051 | 0.139 |
| 4 | 0.598 | 0.241 | 0.192 |
| 5 | 0.629 | 0.041 | 0.099 |

Table 2.10: Results from smTIRF- FRET data for 5-component model

| Iterations | Log-likelihood | Tolerance |
|---|---|---|
| 931 | 661.04 | 1e-08 |

It was observed that by increasing the number of components to the dataset, this was able to capture a better shape, and return a larger log-likelihood value. This provides strong evidence that the larger component model was better representative of the data. It was also noted that by changing the stopping criterion, this affected the number of iterations required for the model to achieve convergence. The log-likelihood value returned from the alternate stopping criteron did not change, nor did the GMM parameters.

As mentioned, the likelihood can always be improved by adding more states to the kinetic model which makes it difficult to distinguish real conformational states from states that arise from overfitting the inherently noisy individual signal trajectories. However, this still implies there are different distinct states present in this dataset. Imaging methods would need to be conducted to validate these proposed conformations. Perhaps adding more components may also imply that in the future, the chromatin organization/folding process can be broken down more granularly in more detail for each component. If this is always true, it means we can examine each individual stage in a more detailed manner.

## 2.5 Discussion

The purpose of smFRET experiments is to observe fluorescence signals on DNA to detect chromatin looping activity. Based on the fluorescence intensities, this indicates the presence of whether chromatin looping has occurred on DNA. Each stage of the chromatin looping procedure will also possess different biophysical properties [21]. These properties rationalize how soluble protein factors dissolved in the liquid nuclear phase, the nucleoplasm, bind and organize transcriptionally active or silenced chromatin domains. There are also different mechanisms for the formation of phase-separated chromatin subcompartments [21]. The distinguishment of these biophysical properties during the chromatin compaction process are important to make predictions on the assembly of chromatin. The GMM in this chapter has been proposed as a method to distinguish the steps required to compact chromatin into organized loops. What is novel is that the GMM was applied to existing proposed chromatin folding procedures, and tested to analyze the folding properties of chromatin in each step. Its efficacy is now discussed to gain further insight on how this can be a powerful tool in chromatin looping analyses.

In an smFRET dataset, the output is interpreted such that low FRET signals correspond to unlooped DNA, and high FRET signals imply the DNA is in its looped state. In this section, the plot of the count density to the FRET emissions was analyzed. The EM algorithm was implemented to identify significant components of the data. The Gaussian mixture model developed from the EM algorithm contained components were employed to identify the number of states in which the strand of DNA was structured. If the means of the Gaussian components identified more trends with lower FRET emissions, this would imply that the DNA was largely in the unlooped state. If higher Gaussian means were identified, then the data was organized into loops. This determined the frequency of interactions within the strand of DNA. In the environment where this chromatin was analyzed, it allowed for a specific number of chromatin configurations to occur at that moment. This could give information pertaining to whether chromatin looping would have been promoted or discouraged in this type of environment. Overall, the GMM was able to cluster the data into the pre-defined number of components. The smFRET experiments can be organized into a detailed ensemble of disordered states within biological networks as categorized into specific components via the GMM and EM method.

It is important to emphasize that smFRET experiments only identifies the conformational states of chromosomes. Therefore, by identifying the significant peaks in the data, one is able to determine specific values of $E_{\text{FRET}}$ associated with a specific state. Donor fluorophore and acceptor fluorophore are placed at the specific positions on neighbouring nucleosomes on DNA to allow for detection of FRET fluctuations to measure chromatin folding activity [7]. This allows for the generation of specific ranges of FRET emissions to be established to determine what kind of activity has occurred. This does not necessarily specify the details that contribute to this state, such as the external factors, but it does allow for experiments to validate whether the strand of DNA will achieve this topological structure.

There are two states in chromatin data analysis with Hi-C data, as conducted previously by Yu et al [71]: intra-domain contact frequency and inter-domain contact frequency. The difference between these conformations is that intra-domain contacts describe contacts that are within a chromatin domain, whereas inter-domain contacts describe contacts that are outside of a chromatin domain. Some examples of intrachromosomal interactions include interactions between promoters and enhancers over several kilobases up to megabases away[18]. Another example includes insulator-mediated contacts that appear to contribute to the organization of the genome into functionally distinct regions by separating differently regulated regions from each other[18]. Interchromosomal interactions are not as well researched, and are mostly involved in promoting the formation of chromatin domains [18]. Some examples may be involved in gene regulation. Interchromosomal domains also affect chromosomal translocations because the rejoining of broken chromosomes requires their physical interaction [18].

In the two component model, we aim to identify areas of the data that identify where $E_{\mathrm{FRET}}$ shows both inter and intra domain contact frequencies. There were two distinct peaks identified around $E_{\mathrm{FRET}}$ at 0.05 and 0.6. From this results, it suggests that there is are generally chromatin in a looped and unlooped state based on smFRET interpretation. This may be difficult to interpret the intra domain and inter domain contact frequencies of chromatin, as both states contain looped and unlooped chromatin. In general, it was discussed by Brackley et al that inter-domain interactions are weaker than intra-domain ones in Hi-C data [11]. One advantage to smFRET technology is it is able to measure inter-domain motions in proteins [65]. Therefore, inter- and intra-domain contacts these can both be factors in contributing to chromatin looping. Since the $E_{\mathrm{FRET}}$ data is not able to distinguish between an unfolded and folded state of chromatin, this does not provide insight toward whether inter or intradomain interactions have contributed to the chromatin organization. Therefore, other proposed components to the model are further scrutinized to see if the GMM can distinguish betweeen other folding activities. In addition, the two component model itself was not successful in presenting accurate clusters around these $E_{\mathrm{FRET}}$ values. With the help of the higher component models, these states were able to be identified as important, and should be more closely evaluated. There were some $\mu$ values in which the EM algorithm ouputted negative values. Mathematically, this information implies that the data is more prominant toward the left side of the plot, where there are lower values of $E_{\mathrm{FRET}}$, signifying an unlooped chromatin state. Therefore, the given dataset for this time frame showed that this strand of DNA was largely unlooped. While there were some significant peaks around the $\mu = 0.6$ state, this implied that some chromatin looping has also occurred.

Another study described the chromatin folding process into four steps. Loop extrusion occurs when an extrusion complex loads onto a fiber at a random locus, forming an extremely short range loop. As the two subunits move in opposite directions along the fiber, the loop grows and the extruded fiber forms a domain. When a subunit detects a motif on the appropriate strand, it can then stop sliding [61]. It is observed that some of the outputted central means of the GMM in the 4-component model shares some similarities to those determined in the 2-component model. Visually, it seems that some of the components have merged closer together within the smFRET dataset. When clustering the data into more components, it did not seem as though there was much distance between the clusters. By this interpretation biologically, there may not be much conformational difference between the properties of DNA between the exposed strand, and the strand upon binding of an LEF. There was some indication that there was some looped chromatin in this dataset, but there was a stronger presence of LEF in an unlooped DNA strand since more components identified more central data in the lower $E_{\text{FRET}}$ values. This could imply that $E_{\text{FRET}}$ is detecting the small loops in the same manner as the large loops, as well. From the four steps of the chromatin process described, the third step of the procedure should result in looped chromatin. The results from the four-component model in our GMM run were not reflective of this result, therefore the four component model described would not be effective with the GMM-EM method. Since this model still provides vague information toward the conformational state of chromatin in this model, one more component model is analyzed.

The general steps in chromatin assembly can also be summarized in five stages [59]. In the first step, assembly begins with the incorporation of the H3/H4 tetramer. Second, two H2A-H2B dimers are added to form the core particle. The newly synthesized histones utilized are then modified, where histone H4 is typically acetylated at Lys5 andLys 12. Thus, the third step required ATP to establish a regular spacing, and histones are de-acetylated. The incorporation of linker histones is accompanied by folding of the nucleofilament. This establishes the fourth step, where the structure is configured into a solenoid shape, in which there are six nucleosomes per gyre. Finally, folding events ultimately lead to a defined domain organization within the nucleus. By using a 5-component GMM, this attempts to identify the areas in which these five steps may be occuring in the data. As described, it isn't until the fourth step of this process until the chromatin obtains a more established shape. In the first three steps of this interpretation of the chromatin folding process, the histones are binding to the chromatin to prepare the DNA for looping. This is where the enzymes have prepared the DNA to allow for the biophysical properties required for DNA interaction. As observed from the five component GMM results, it isn't until the fourth component where the $E_{\text{FRET}}$ value becomes greater than 0.5. This implies that there is some organization of chromatin loops. Therefore, this model is more appropriate to model the chromatin folding process, as given in conjunction with the procedure described by Ridgeway and Almouzni [59].

The GMM combined with the EM algorithm maximizes the likelihood such that it will not bias the means towards zero, or bias the cluster sizes to have specific structures that may or may not apply. Due to the nature of the size of smFRET data, there is little bias, as it often collects large samples of data. Knowing the number of chromatin folding stages will also make the GMM a powerful tool for chromatin data, as it will be able to successfully identify the FRET emissions associated with each specific stage of the looping process. One limitation of this method is that the likelihood can always be improved by adding more states to the kinetic model, making it difficult to distinguish real conformational states from states that arise from overfitting the inherently noisy individual signal trajectories [69]. This is important to note when considering the number of stages of chromatin organisation to include in the GMM components. This stresses the importance of having observed data for each stage of the chromatin compaction procedure to understand the true number of components required to create the GMM model. This becomes a further limitation, since the model requires prior knowledge of the biological processes involved.

## 2.6 Future directions

As discussed, one of the disadvantages of the GMM involves the identification of the number of components. The number of components for the GMM were set based on the known chromatin states in this chapter that are available based on current observational experiments. This becomes a limitation to this model, as this is a supervised learning procedure. This allows for the understanding of the specific chromatin states that are already known. Should an unsupervised method be used to cluster the data, there could be many more states identified that are not already known. Therefore, some knowledge of the chromatin looping stages would need to be known in order to use this model. This may also be advantageous, as this will reinforce the understanding of the already established chromatin looping states. The EM algorithm will also always use all the components it has access to, needing held-out data or information theoretical criteria to decide how many components to use in the absence of external cues. Selecting the number of components in the GMM is then important to identify the clusters. One way to determine this would be to add more components until overfitting occurs. However, without observed experimental evidence to support these components, there would not be added meaning to the identified components. Therefore, observed experimental data and more work conducted into identifying the steps of chromatin folding and the unique biochemical properties at each step would allow for more meaningful results.

The BIC criterion can be used to select the number of components in a GMM in an efficient way. In theory, it recovers the true number of components only in the asymptotic regime. However, using a variation Bayesian Gaussian mixture model avoids the specification of the number of components for a GMM. It was previously shown by Cordunneanu and Bishop [16] that by setting the mixing coefficients to maximize the marginal log-likelihood, unwanted components can be suppressed, and the appropriate number of components for the mixture can be determined in a single training run without recourse to cross-validation. Variational treatment based on factorized approximation to the posterior distribution can be used to accomplish this task. However, this was not performed here as this would not be supportive of the unknown number of components that may be required to support the chromatin compaction procedure. More experimental evidence would be needed to support this method of validating the smFRET components.

Biologically, there is much research to be conducted in visualizing specific steps in the chromatin loop process. Specified stages should be identified in order to successfully implement the GMM with the EM algorithm. The steps that have been researched in the literature are vague due to the imaging limitations in current technology. For example, upon any chromatin modifications, this will modify the genetic makeup of chromatin in a similar manner to nucleosome modification behaviour. Any binding of an enzyme will ultimately influence whether changes to the makeup of chromatin will be influenced. The environment in which the chromatin is modified will also influence the conformation. Once there is more information available in the specific steps in chromatin conformation, the GMM applied in conjunction with the EM algorithm can be a powerful tool in identifying significant $E_{\mathrm{FRET}}$ data that may identify unique properties of chromatin in each state.

Some $E_{\mathrm{FRET}}$ components may have a wider range corresponding to a specific step in the chromatin organisation process. Therefore, other mixture distributions should also be fit to the data to gain more information on which $E_{\mathrm{FRET}}$ values correspond to a specific area of data. Previously, it was discussed that experimental smFRET histograms have been well-fitted by both Gaussian functions and log-normal functions [51]. It was also stated by Deniz et al [19] that "One must exercise great caution in interpreting the width of FRET distribution", due to the interplay of many factors. Therefore, when interpreting the central means of each distribution for each component of the GMM, it would be helpful to know some additional biological data to understand what needs to be included in the model.

## 2.7 Conclusion

The GMM assigns a probability to each point to see whether it should belong to the identified potential cluster, rather than assigning a flag to a point that belongs to certain cluster, as performed in the classical $k$-means method. Then, GMM is producing non-convex clusters, which can be controlled with the variance of the distribution. The EM algorithm has aided in optimizing the loss function of the GMM. The clustering work conducted with both the GMM and the EM algorithm is a powerful tool in identifying significant clusters of data. This can be a powerful tool in identifying the areas of $E_{\mathrm{FRET}}$ data where the specific chromatin conformations are located. This will help in understanding the specific functionalities and properties of chromatin in each defined state. While more work needs to be conducted in identifying the number of stages required to model the chromatin compaction procedure, this tool can be powerful in identifying the components in smFRET data.

# 3 CTMC Application: Loop Extrusion Factor Stacking

## 3.1 Abstract

The binding behaviour of loop extrusion factor stacking can be modelled using several continuous-time Markov chain (CTMC) models. The kinetic rates including the rates of binding, unbinding, loop division and loop death are used to drive these CTMC applications for loop extrusion. The immigration-death model is used to model the stacking behaviour of loop extrusion factors to reinforce chromatin structure. The Gillespie algorithm is used to simulate the number of fluctuating LEFs in the IDP. The burst model is also used to demonstrate the abilities of the chromatin compaction process beginning from the binding of an LEF to the creation of chromatin loops. We found that the burst model applied with non-equilibrium kinetic rates of the LEF exhibited more consistent results with current available experimental data. Future work to investigate the burst model with other LEF candidates would be helpful in gaining a better more general understanding of the behaviour of LEFs in chromatin organization.

## 3.2 Introduction

Elements of chromatin looping have been modelled with stochastic processes due to the dynamic nature of loop extrusion factors. This suggests that Markov chains may be suitable for describing some of their structural folding behaviour. Markov chains describe a sequence of possible events in which the probability of each event depends on the state attained in the previous event. In this chapter, Markov properties are explored to model various stages of the chromatin looping process with loop extrusion factors through chromatin binding, unbinding, loop division and loop degradation activities.

The Immigration-Death process (IDP) is an example of a continuous-time Markov chain (CTMC), which is originally derived from the Birth-Death process (BDP). This model is used to describe the fluctuations of chromatin loops in DNA compaction. The BDP is first introduced in order to derive the IDP. This is then evolved and further applied to model the behaviour of loop formation and degradation in chromatin compaction. This is the main model in this section that describes the loop organization behaviour.

## 3.3 Materials and methods

### 3.3.1 Continuous Time Markov Chains

Since the Birth-Death process (BDP) is an example of a continuous-time Markov chain (CTMC), the properties of a CTMC are presented. To model a process in continuous time, the limit of a discrete-time process is taken. A Markov chain in discrete time remains in any state for exactly one unit of time before making a transition, or a change of state. When the process enters a state, the time it spends before leaving the state is referred as the holding time. Some conditions are placed on the holding time to ensure that the CTMC satisfies the Markov property. The Markov property refers to the memoryless property of the CTMC. This is described as $\{X(t), t \geq 0\}$, which is a stochastic process that takes values on a finite or countable set (0, 1, 2, ...,). The memoryless property is described as the following,

$$P[X(t+s) = j | X(s) = i, X(u) = x(u) \text{ for } 0 \leq u \leq s] = P[X(t+s) = j | X(s) = i] \tag{3.1}$$

Time homogeneous chains are also considered, where the memoryless property is employed. As mentioned, a condition is placed on the holding time to ensure this property holds, wherein they will have to be exponentially distributed, such that

$$P[X(t+s) = j | X(s) = i] = P[X(t) = j | X(0) = i] \tag{3.2}$$

By time homogeneity, this refers to whenever the process enters state $i$; the way it evolves probabilitically from that point is the same as if the process started in state $i$ at time 0. When the process enters state $i$, the time it spends in that state before it leaves is referred to as the holding time in state $i$. By time homogeneity, the holding time distribution is the same every time the process enters state $i$, which is also exponentially distributed. This explores the Markov property (Equation 3.1), where the behaviour of the future of the process is only dependent upon the current state and not of the past. The models are generalized to allow for time to be continuous, and the state space is either finite or countably finite. The sequence of states is written as $\{X_n, n \geq 0\}$ that $\{X(t)\}$ arrives in, and let $S_n$ be the corresponding arrival times. The Markov property for $\{X(t)\}$ implies the (discrete-time) Markov property for $\{X_n\}$, thus $\{X_n\}$ is an embedded Markov chain, with transition matrix $P = [P_{ij}]$. The key quantities that specify a discrete-time Markov chain are the transition probabilities. In continuous time, the corresponding key quantities are the transition rates.

CTMCs can each be modeled by representing the passage from one state to another as a

Poisson process. This process is now described in more detail to gain a better understanding of its role in the CTMC.

### 3.3.2 Poisson Process

Briefly, the Poisson distribution by itself describes the probability of an event occurring after a given amount of time. It uses a single parameter, which expresses the average rate at which the event occurs. A recurring event, or process, whose probability follows this distribution is called a Poisson process. To fully classify something as a Poisson process, specifically for Markov chains, the shift invariance and independent inter-arrival times must be derived. Shift invariance describes the probability of an event occurring within a given amount of time and does not change depending on the point at which you start counting time. Independent inter-arrival times indicate that once an event in a Poisson process has occurred once, the probability of it occuring a second time can be modelled with the same Poisson distribution, as if from the initial occurrence. The specific properties of the Poisson process are now explained in more detail to understand the criteria for this model.

The Poisson distribution occurs by way of the *Law of rare events.* Consider a large number $N$ of independent Bernoulli trials where the probability $p$ of success on each trial is small and constant from trial to trial. Let $X_{N,p}$ be the total number of successes in $N$ trials, where $X_{N,p}$ follows the binomial distribution for $k = 0, 1, ..., N$.

$$P(X_{N,p} = k) = \binom{N}{k} p^k (1-p)^{N-k} \tag{3.3}$$

If we assume that $N \to \infty$ and $p \to 0$, so that $Np = \mu$, then the distribution for $X_{N,p}$ becomes the Poisson distribution:

$$P(X_\mu = k) = \frac{e^{-\mu}\mu^k}{k!} \text{ for } k = 0, 1, ... \tag{3.4}$$

In stochastic modelling, this law is used to suggest circumstances under which the Poisson distribution might be expected to dominate approximately.

The Poisson distribution given in Equation 3.4 describes the probability of having $k$ events over a time period in $\mu$. The random variable $X$ having a Poisson distribution has the mean $\mathbb{E}[X] = \mu$ and the variance $Var[X] = \mu$.

The Poisson process gives the notion of the Poisson distribution together with independence. A Poisson process of intensity $\lambda$>0, which describes the expected number of events per unit of time, is an integer-values Stochastic process $\{X(t); t \geq 0\}$ for which the following criteria holds:

1. For any arbitrary time points $t_0<t_1<t_2<...<t_n$ and $t_0 = 0$, the number of events happening in disjoint intervals (process increments)

$$X(t_1) - X(t_0), X(t_2) - X(t_1), X(t_3) - X(t_2), ..., X(t_n) - X(t_{n-1})$$

   are independent random variables. This implies that the number of events in one time interval is independent from the number of events in an interval that is disjoint from the first interval. This is the *independent increments* property of the Poisson process.

2. For $s \geq 0$ and $t \geq 0$, the random variable $X(s+t) - X(s)$, which describes the number of events occurring between time $s$ and $s+t$ (independent increment), follows the Poisson distribution,

$$P(X(s+t) - X(s) = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}.$$

3. We assume that at time zero the number of events that have happened already is zero

The parameters of the Poisson distribution is $\lambda t$, $\mathbb{E}[X(t)] = \lambda t$, $Var[X(t)] = \lambda t$. We fix a short interval of time $h$ to begin the derivation of exactly one event over the time period $h$:

$$\begin{aligned}
P(X(t+h) - X(t) = 1) &= \frac{(\lambda h) e^{-\lambda h}}{1!} \\
&= (\lambda h) \sum_{n=0}^{\infty} \frac{(-\lambda h)^n}{n!} \\
&= (\lambda h) \left( \frac{(-\lambda h)^0}{0!} + \frac{(-\lambda h)^1}{1!} + \frac{(-\lambda h)^2}{2!} + ... \right) \\
&= (\lambda h)(1 - \lambda h + \frac{1}{2}\lambda^2 h^2 - ...) \\
&= \lambda h + o(h)
\end{aligned} \tag{3.5}$$

where $o(h)$ denotes a general and unspecified remainder term of smaller order than $h$. We can view the rate $\lambda$ in Poisson process $X(t)$ as the proportionality constant in the probability of an event occurring during an arbitrary small interval $h$.

In a Poisson process, the waiting time between consecutive events is referred to as a Sojourn time, $S_i - W_{i+1} - W_i$, where $W_i$ is the time of occurrence of the $i$th event. This means that $S_i$ measured the duration that the Poisson process sojourns in state $i$. The Sojourn times $S_0, S_1, ..., S_{n-1}$ are independent random variables, each having the exponential probability density function as given in the following equation,

$$f_{S_k}(s) - \lambda e^{-\lambda s} \tag{3.6}$$

The Poisson process may also be referred to as the pure birth process. In other words, only forward transitions apply. The Birth-Death process (BDP) then introduces the term where backward transitions are allowed. The applicability of the Poisson process is now extended to the BDP. The BDP is a special case of a CTMC where the state transitions are of only two types: "births", which increases the state variable by one and "deaths", which decreases the state variable by one. There have been several extensions of the BDP to introduce other variables such as immigration, shifts, migration, etc., to name a few. The BDP may also be described with queueing theory, as well. For the purpose of this investigation, the Immigration-Death process is now derived and analyzed for its applicability in chromatin loop dynamics.

### 3.3.3 Immigration-Death Process

In order to derive the Immigration-Death process (IDP), the Birth-Death process (BDP) must be explained to compare its modifications in the IDP. The BDP describes a CTMC with discrete state spaces in which state transitions may only occur between neighbouring states. Originally, this was used to model the number of "particles" in a system, where each particle can "give birth" to another particle or "die" ([22],[38]). The rates of birth and death at any given time are dependent on the number of particles present in the system. This is a type of branching process, where the history of the trajectory of the particle is insignificant, but the total number of particles in a system is maintained at a given time.

When there are $i$ particles in the system, a birth occurs with an instantaneous rate, $\lambda_i$ and a death occurs with an instantaneous rate, $\mu_i$. In the "simple linear" BDP, $\lambda_i = i\lambda$ and $\mu_i = i\mu$ to allow for constant birth and death rates. In the "general" BDP, the values for $\lambda_i$ and $\mu_i$ can be any function of $i$ but are time-homogeneous. This allows for modifications to be made to generate the IDP. To adjust the properties of the BDP to create the IDP, the birth rates are instead given by $\lambda_i = \alpha$. Note the death rates are still given by $\mu_i = i\mu$ for all $i$. The IDP describes $\{N(t)\}_{t\geq 0}$ as a time-homogeneous irreducible CTMC where the possible states for which transitions $i \to j$ may occur are supplied by the state space $E = \{0, 1, ...\}$. The birth and death rates propel the IDP with $\theta = (\alpha, \mu)$. This is assumed to take values in some parameter space $\Theta$ which is a compact subset of $\mathbb{R}_+^2$. In order to view $\{N(t)\}_{t\geq 0}$ as a BDP, the *Law of the total probability* is applied, where the total probability of an outcome can be realized via several distinct events. For $i > 0$, the infinitesimal transition probabilities are given by the following set of equations:

$$p_{ij}(t;\theta) = P(N(h+t) = j|N(h) = i) = \begin{cases} \lambda_i t + o(t), & \text{if } j = i+1 \\ 1 - (\lambda_i + \mu_i)t + o(t), & \text{if } j = i \\ \mu_i t + o(t), & \text{if } j = i-1 \\ o(t) & \text{if } |j-i| > 1 \end{cases} \tag{3.7}$$

By extension, the IDP can also be extended to queueing theory, where the concepts of the IDP parallel those of an M/M/$\infty$ system. The M/M/$\infty$ queue is a multi server queueing model where every arrival experiences immediate service and does not wait. Each customer arrives according to a Poisson process with intensity $\alpha$. They are individually handled by its own server so that its Sojourn time in the system is exponential with intensity $\mu$ and independent of all other customers. This means that the waiting customers are either served, or removed from the queue. In this case, there are infinitely many servers, so customers do not need to wait for a server.

The transition matrix is then described as the following equation $\Delta p(t) = p(t)Q$, where the matrix is given in Equation 3.8, and the process is illustrated in Figure 3.1, where $\alpha$ represents the immigration rate and $\mu$ represents the death rate.

$$Q = \begin{pmatrix} -\alpha & \alpha & & & \\ \mu & -(\mu+\alpha) & \alpha & & \\ & 2\mu & -(2\mu+\alpha) & \alpha & \\ & & 3\mu & -(3\mu+\alpha) & \alpha \\ & & & & \ddots \end{pmatrix} \tag{3.8}$$

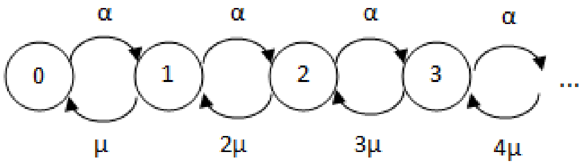

Figure 3.1: Diagram of M/M/$\infty$ queue

The IDP may now be applied to model the chromatin looping process. The immigration and death rates in the process will be denoted by the rate of division and degradation of loops on a strand of DNA . By modelling the chromatin compaction activity in this manner, it is predicted that the rates of loop formation and degradation may be established for different loop extrusion factors used to perform extrusion. While chromatin loop formation is still unknown due to the limitations of experimentation, this model may hopefully contribute to the understanding of DNA storage information once appropriate data becomes available in the future. For the purpose of this investigation, simulations conducted with the Gillespie algorithm are used in the place of observed data to test the proposed systems. This is now introduced to understand how it may provide a proposed trajectory of single cell processes.

### 3.3.4 Gillespie Algorithm for Chromatin Loop Dynamics

The Gillespie algorithm was created by Dan Gillespie, with the intention of using it to simulate chemical or biochemical systems of reactions efficiently and accurately with limited computational power [27]. This has been useful for simulating reactions within cells where the number of reagents is low and it is computationally feasible to keep track of the position and behaviour of individual molecules.

While there are generally two ways of implementing the Gillespie algorithm (the deterministic approach and the stochastic approach), the algorithm may be altogether summarized in the following manner,

1. **Initialization.** Initialize the number of molecules in the system, reaction constants, and random number generators.
2. **Monte Carlo step.** Generate random numbers to determine the next reaction to occur as well as the time interval. The probability of a given reaction to be chosen is proportional to the number of substrate molecules, and the time interval is exponentially distributed with mean $1/R_{TOT}$ where $R_{TOT}$ is the total number of substrate molecules.
3. **Update.** Increase the time step by the randomly generated time in Step 2. Update the molecule count based on the reaction that occured.
4. **Iterate.** Go back to Step 2 unless the number of reactants is zero or the simulation time has been exceeded.

For the purpose of this investigation, the stochastic approach is favoured due to the dynamic nature of the loop extrusion factor (LEF) binding to chromatin. This algorithm provides a systematic method for obtaining a sample of trajectories that are consistent with the master equation that describes the stochastic system. When the chemical rates are known functions of time, then the simple and intuitive way to augment the Gillespie algorithm is to assume that the propensities are known functions of time. However, this leads to an approximate, rather than an exact stochastic algorithm. The deterministic approach regards the time evolution as a continuous, wholly predictable process governed by a set of coupled, ordinary differential equations. These are generally represented with reaction-rate equations. The stochastic approach regards the time evolution as a kind of random-walk process, which is governed by a single differential-difference equation, and represented by the master equation.

Biochemical experiments have widely accepted and used the Gillespie algorithm to simulate biochemical networks due to its inclusion of noise, which is especially important in gene regulation [45]. This algorithm is used when fluctuations arise from the small number of reactant molecules. The advantage is that it generates an ensemble of trajectories with correct statistics for a set of biochemical reactions. This allows for favourable conditions to generate loop extrusion trajectories.

Prior to applying the Gillespie algorithm in a chromatin loop process, the following assumptions were made, as described previously in literature [30].

1. The two heads of each LEF stochastically step away from each other with the average rate, $\alpha$, the immigration rate of the IDP.
2. The heads of different LEFs cannot step over each other and thus stop extrusion upon reaching another LEF. However, the two heads of the same LEF may extrude loops independently and if one head of a LEF is blocked, another head continues extrusion.
3. LEFs stochastically unbind from the fiber with the rate $\frac{1}{\tau}$, where $\tau$ represents the average residence time.
4. Free LEFs immediately rebind to the chromatin fiber at a random uniformly chosen pair of adjacent sites.

It is assumed in this LEF model that upon unbinding from chromatin, an LEF immediately rebinds to another randomly chosen site along the chromosome to maintain the steady state assumption. With this assumption, the number of LEFs bound to a chromosome remains constant throughout the simulation. Another strong assumption is that an LEF extrudes two chromatin strands independently, such that blocking extrusion of one strand does not stop extrusion of another. Coordinated blocking of extrusion prevents the growth of the average loop size with $\lambda/d$ and the average loop size saturates at $\sim d$, where $d$ is the LEF separation, and $\lambda$ is the LEF processivity. This shows that in order to efficiently compact a chromosome, LEFs must be able to move one of its contact points when the other is blocked.

The implementation of the Gillespie algorithm in this application with LEFs for loop extrusion is summarized. Suppose the system is known at time $t$, which means the number of DNA reacting with LEFs are known, and consequently, the set of reactions $a_\mu(t)$ with a given number of LEFs are known for each reaction. Call $a_0(t)$ the sum of all $a_\mu(t)$. Let reaction 1 be the reaction with probability per unit time of $a_1$, reaction 2 be the reaction with probability per unit time of $a_2$, and so on.

Then the following steps are followed:

1. Find the time $\tau$ after $t$ at which the next reaction will take place, by drawing a random number from an exponential probability density function (pdf) of rate $a_0 p(\tau) = a_0 \exp(-a_0 \tau)$.

2. Choose at random the reaction which will occur at time $t+\tau$. Draw a random number from a uniform distribution between 0 and 1. If that number falls between 0 and $a_1/a_0$ reaction 1 is chosen. If that number falls between $a_1/a_0$ and $(a_1 + a_2)/a_0$ reaction 2. This pattern then continues in a similar manner.

3. The occurrence of the chosen reaction at time $t + \tau$ changes the number of DNA molecules involved in the reaction. Thus the values of the $a_\mu$ which depend on any of these numbers change. One then goes back to point 1 of the algorithmic implementation with a new distribution of molecules at time $t+\tau$. The process is reiterated for as long as one wishes to follow the evolution of the system.

The observed data parameters were simulated using the Gillespie algorithm from the Python code made available from a previous publication [30]. This software was used to create illustrative diagrams of the loop extrusion process, where it was used to demonstrate the loop formation over time. Figure 3.2 gives a representation of chromosome compaction by LEFs in a single frame of time.

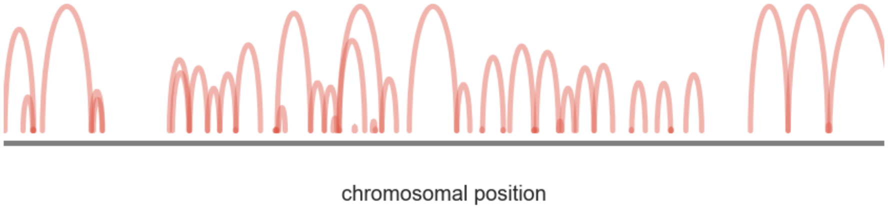

Figure 3.2: Simulated loop extrusion with overlapping loops in a single time frame. The height of the loop demonstrates the amount of chromatin that has been extruded from the loop. The colour of the loop is also significant such that darker the loop, the more stacking has occurred.

The height and colour of the loops represent the LEF stacking process that occurs in chromatin compaction over time. Since loop extrusion is a stochastic process, there is much variability in the chromatin loop process at any time frame. The darker the loop, the more stacking has occurred. The height of the loop denotes the amount of chromatin that has been extruded from the loop. In initial simulations, LEFs generated tightly stacked loops with a high degree of chromatin compaction, despite their constant dissociation. Simulations converged to states with degree of compaction and distribution of loop size that depended on the control parameters, but were independent of initial states. This supports the existence of a well-defined, loop stacked steady state.

By running this same simulation again with more LEFs present in the system, Figure 3.3 illustrates a different arc diagram than that from Figure 3.2. It is observed that by increasing the number of LEFs present in the system, there was more overlap in the loops formed. This indicates that a greater degree of loop extrusion activity, and ultimately DNA compaction, has occurred. What also differred was that more stacking has occurred due to the increased number of darker loops formed. This could suggest that the model favours LEF stacking when more LEFs are available.

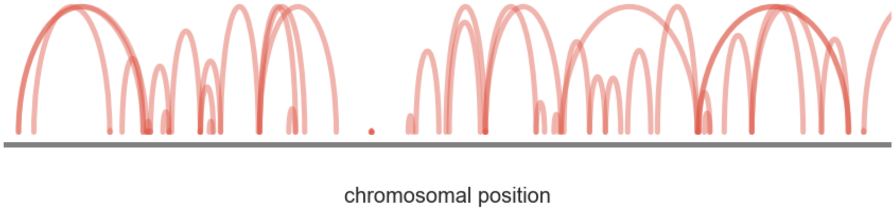

Figure 3.3: Simulated loop extrusion with overlapping loops in a single time frame with more LEFs in the system

In this case, the chromatin loop activity describes a 1D lattice model for loop division and degradation. The rates for the IDP are now derived and supported with Gillespie simulated activity to test the model.

### 3.3.5 Rates of Loop Extrusion

The underlying mechanisms that drive chromatin loop extrusion is still unknown. There are several models proposed to describe this process, however experimentation is still limited to validate this data. The rates of loop division and degradation presented in this section are derived from a previous publication by Goloborodko, et al [30]. It was discussed that loop extrusion factors (LEFs) dynamically exchange between the nucleoplasm and chromatin fiber. LEFs self-organize into a dynamic array of consecutive loops with two distinct steady states: a sparse state, where loops are separated by gaps and provide moderate compaction; and a dense state, where jammed LEFs drastically compact a long chromatin fiber [30]. The analytical model is based on microscopic properties of LEFs and their abundance. An illustration of the sparse and dense state is presented in Figure 3.4

There are several theories available regarding chromatin looping and DNA compaction. One example of chromatin looping has been studied by Alipour and Marko [1], where they introduced a quantitative model of loop extrusion and considered the dynamics of solvent-exchanging LEFs on a short chromosomal segment. They found that extruded chromatin loops can be stabilized by multiple stacked LEFs, thus making loops robust enough for the exchange of individual LEFs. A small system size however, prevented them from obtaining a complete picture of self-organization. The remaining key question is whether LEFs alone are sufficient to form arrays of non-overlapping loops on a long chromosome or if other factors are required to define the loop bases.

In this investigation, the model proposed by Goloborodko, et. al. [30] is favoured to describe

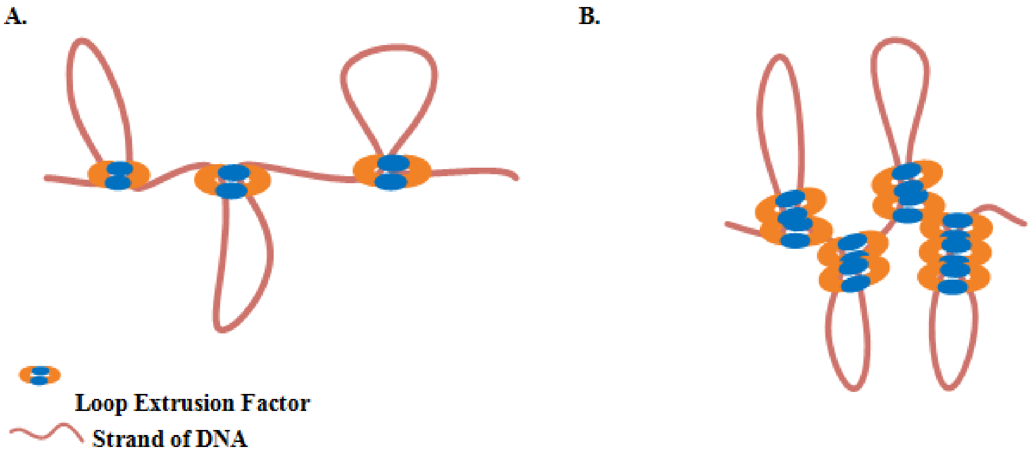

Figure 3.4: Steady states of DNA loops. **A.** Sparse state ; **B.** Dense state

the loop extrusion process because it demonstrated that efficient chromosome compaction can be achieved solely by an active loop-extrusion process. This method is also further investigated because it has been acknowledged as one of the favoured methods of chromatin organization in literature [56].

In the model of Goloborodko et al, the rate of loop division is given by

$$R_{divsion} \approx \left(\frac{1}{\tau}\right)\left(\frac{l}{d}\right)^{3}\left(\frac{d}{\lambda}\right), \tag{3.9}$$

while the rate of loop death is given by

$$R_{death} \approx \left(\frac{1}{\tau}\right)\left(\frac{l}{d}\right)e^{-l/d} \tag{3.10}$$

As observed, the rates are made up of several other conditions contributing to chromatin loop activity. These factors are defined by the following variables,

$L$ = length of the chromosome<br>
$N$ = number of LEFs in the system<br>
$d = L/N$, LEF separation, the average spacing between LEFs if they were randomly dispersed along the chromosome<br>
$\lambda = 2\nu\tau$, LEF processivity, the average length of a chromatin loop that a single unobstructed LEF can extrude over its residence time<br>
$\nu$ = the average speed with which a LEF motor translocates chromatin fiber<br>
$\tau$ = the average time that a LEF stays continuously bound to the chromosome, also referred to as the residence time<br>
$l$ = length of a loop

It is noted that these rates are largely determined by the LEF separation, and the LEF processivity. To visually observe the behaviour of the rate parameters, the values summarized in Table 3.1 were used to plot an example of the rate behaviour determined by these factors.

Table 3.1: Parameters from observed data

| d | $l$ | $\lambda$ | $\tau$ |
|---|---|---|---|
| 30 kb | 100 kb | 830 | 83 |

The values chosen in Table 3.1 were used to represent chains of 5000 monomers and corresponds to 30 Mb, close to the size of the smallest human chromosomal arm. We focus on the effect of two parameters, the linear separation of LEFs and their processivity, which have been previously considered in the context of a 1D model [30]. These parameters control the fraction of a chromosome extruded into loops and the average loop length. This publication used condensin behaviour to generalize all enzymes that perform LEF activity [30]. The separation, denoted by $d$ is 30 kb, which is given by 1000 condensins per 30 Mb as measured in observed data [24]. The processivity, denoted by $\lambda$ is 830 kb, where condensins form a dense array of gapless loops with the average loop length ($l$) of 100 kb. In addition, the compaction density of $\sim$ 5 kb/nm was used in order to calculate the residence time, $\tau$ [72]. For further details on the experimental methods used to obtain these values, refer to the Goloborodko, et. al. [30] publication.

These values were then plotted to observe the extent of loop length and separation given in Figure 3.5. It is observed that the polynomial function of the rate of loop division will always be increasing, whereas the rate of loop death will reach a maximum rate then rapidly decrease as the ratio of loop length to loop separation increases. Therefore, at larger loop lengths, there will be more loop division activity, which will allow for compaction. This is consistent with the current understanding for DNA compaction, as there is more available DNA that will allow for loop extrusion. When there is less DNA available, LEFs are less likely to remain stably bound and promote loop extrusion.

By drawing from these concepts, these are further explored in the exposure timescale for binding to DNA in Chapter 4. For now, LEF binding will be explored by analyzing theoretical kinetics of unbound LEFs up until it has conducted loop extrusion activity and unbinds from the strand of DNA.

### 3.3.6 Binding Behaviour of LEFs

As mentioned, the binding behaviour of loop extrusion factors (LEFs) is unknown. While there are several theories and models proposed for the loop extrusion process, there is little to no observed data available to validate these models. The chromatin loop extrusion model

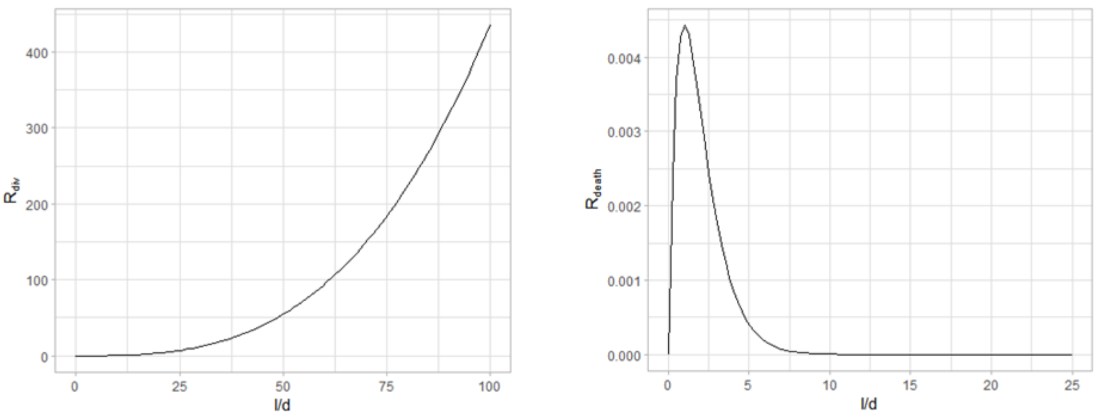

Figure 3.5: Rates of loop extrusion. The left plot denotes the rate of loop division, and the right plot denotes the rate of loop death.

is favoured because it allows for the formation of CTCF loops and topological domains. This explains the arrangement of CTCF binding motifs that stabilize loops, and extrusion is the only model available so far that explains this. The model requires a motor to generate the loops, and although cohesin is a strong candidate for an extruding factor, a suitable motor protein has yet to be found. However, for the purpose of this investigation, CTCF is used as a loop extrusion boundary element to stop loop extrusion, and cohesin is used as an example of an LEF.

To begin the derivation of LEF binding kinetics to DNA, one must consider the binding behaviour of an enzyme to DNA. There are two states in this case: bound (B) and unbound (UB). Once bound, the enzyme can either remain bound or immediately unbind from the strand of DNA. This gives rise to two options in the bound case: dynamically bound (DB) and stably bound (SB). Therefore, the rate that an LEF can bind to DNA may be given by $k_{\text{on}}$ for the forward reaction, and the rate that an LEF unbinds can be given by $k_{\text{off}}$ for the reverse reaction. This was previously described to model the behaviour of cohesin binding to chromatin [32]. It is interesting to note that this model was applied to the growth stages of the cell cycle (G1 and G2). In the steady-state, similar to the concepts applied from Michaelis-Menten properties, the following equation can be used to describe the kinetics.

$$k_{\text{on}} UB = k_{\text{off}} DB \tag{3.11}$$

As described in the previous paragraph, the total copy number of LEFs is the sum of both bound, and unbound LEFs. This can be depicted mathematically as $C_T = B_T + UB_T$. By rearranging this equation, the number of unbound (UB) LEFs can be denoted as $UB = C_T - B_T$, which can be substituted into the left hand side of Equation 3.11. The total number of bound LEFs can also be denoted as the fraction of dynamically bound LEFs over the stably bound LEFs. This is denoted mathematically as $B_T = DB/SB$. By rearranging this equation for the dynamically bound (DB) LEFs, where $DB = B_T SB$, this may be substituted into the right hand side of Equation 3.11. These rearrangements were made in Equation 3.12 below.

$$k_{\text{on}}(C_T - B_T) = k_{\text{off}}(B_T SB). \tag{3.12}$$

Since we are interested in the binding behaviour of LEFs, Equation 3.12 is now rearranged to determine the manner in which the on-rates and off-rates affect LEF binding. This will allow for the examination of how these factors make up the total number of LEFs bound to DNA, and perhaps predict loop extrusion activity. In Equation 3.13, $B_T$ is now isolated.

$$B_T = \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} C_T + \frac{k_{\text{off}}}{k_{\text{on}} + k_{\text{off}}} SB \tag{3.13}$$

As observed, the total number of LEFs bound is made up of the sum of the portion of LEFs that bind to DNA and the portion of stably bound LEFs that unbind from DNA. This equation may be further normalized with respect to the copy number of LEFs bound to DNA. This is done by dividing both sides of the equation by the copy number by substituting $b_T = B_T/C_T$ and $sb = SB/C_T$ such that,

$$b_T = \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} + \frac{k_{\text{off}}}{k_{\text{on}} + k_{\text{off}}} sb. \tag{3.14}$$

In order to gain a better understanding of this equation, assuming a steady state binding of $k_{\text{on}}$ and $k_{\text{off}}$, the relationship between the number of LEFs bound to DNA to the number of stably bound LEFs is illustrated in Figure 3.6. This concept has now been extended to show that the number of stably bound LEFs contribute largely to the number of bound LEFs, which will skew the number of bound LEFs. As this model was previously tested with the binding behaviour of cohesin [32], it is important to note that they found that the only major distinction between cohesin dynamics in G1 and G2 phase cells is that a fraction of cohesin becomes stably bound in G2 [32]. Therefore, the number of stably bound LEFs contribute to the understanding of which stage of the cell cycle this will impact.

### 3.3.7 Combining LEF Binding with Loop Extrusion Activity

The rates of loop extrusion activity and LEF binding activity have now been established separately. Since these two processes contribute to DNA compaction, it is only fitting that these models be combined in some manner. LEF binding can be described with a reversible chemical reaction, whereas loop extrusion from this binding is described with an IDP. These mathematical procedures are now combined sequentially to create a larger illustration of the chromatin compaction process.

To describe this process simply, an LEF must stably bind to DNA and then extrude a loop. This may parallel the burst model [46] to describe stochastic single-cell transcription. The burst model is described in more detail in Section 3.4.1. Transcription may occur in both

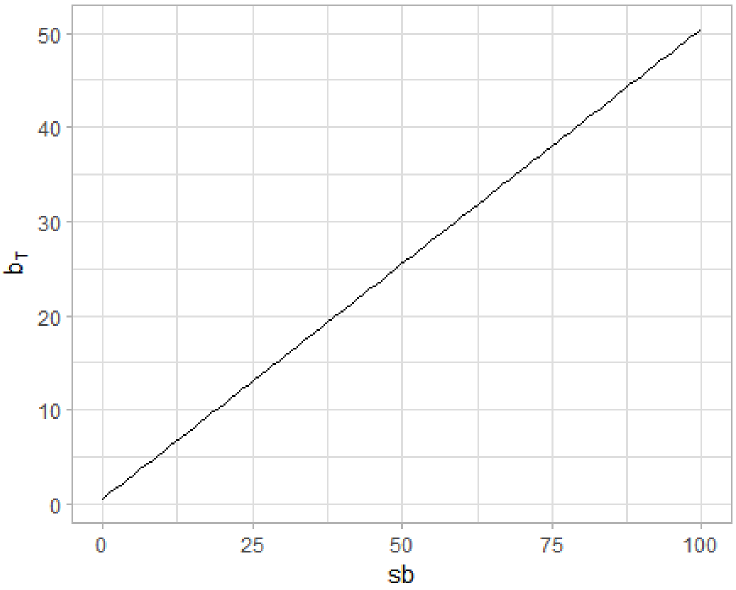

Figure 3.6: Binding behaviour of LEFs on DNA. This plot illustrates the relationship between the copy number of binding sites to the amount of LEFs stably bound to DNA.

a bursty and non-bursty manner, which depends on the parameter values describing the enzyme rates performing transcription. However, this differs from the model in that there are two conditions in the bound state: dynamically bound and stably bound. For the purpose of this investigation, the bound state will be simplified and modelled as one state to simplify the interpretation of the model. This process is illustrated in Figure 3.7.

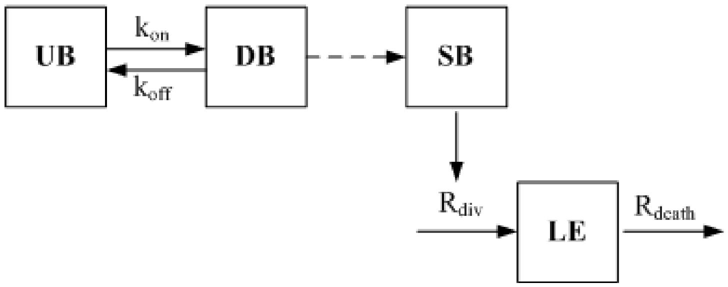

Figure 3.7: Loop Extrusion Dynamics with LEF binding. Components illustrate the relationship between the rate of binding ($k_{\text{on}}$) and unbinding ($k_{\text{off}}$) from the unbound (UB) to the dynamically bound (DB) LEFs. The dynamically bound LEFs then become stably bound (SB) and allow for loop extrusion (LE) to occur. This activity is regulated by their rate of loop division ($R_{\text{div}}$) and loop death ($R_{\text{death}}$).

As illustrated, loop extrusion is a highly regulated sequence of intrinsically stochastic processes. In LEF binding or unbinding, an LEF may bind dynamically to DNA, with a residence time unique to each factor. The biological significance of the effect of the residence time is still unclear for the loop extrusion process, however this will be explored later in terms of how it affects the exposure timescale of the target DNA in which an LEF binds. Therefore, the on-time and off-time of an LEF is given by $\tau_{\text{on}} = 1/k_{\text{off}}$ and $\tau_{\text{off}} = 1/k_{\text{on}}$, respectively.

In order to derive the mean number of occurrences for loop extrusion, the contributing factors from enzyme activity as described in previous sections is considered. The number of loops is affected by both the number of LEFs bound to DNA, and the rate of loop extrusion activity. Thus, both the rates of loop formation, and the rate of binding of the LEF must be utilized. The fraction of loop division and the fraction of LEF binding is combined as a product to yield the mean number of occurences for loop extrusion. Mathematically, this is given by,

$$\begin{aligned} L_{\text{T}} &= \frac{R_{\text{div}}}{R_{\text{death}}} \cdot \frac{\tau_{\text{on}}}{\tau_{\text{on}} + \tau_{\text{off}}} \\ &= \frac{R_{\text{div}}}{R_{\text{death}}} \cdot \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} \end{aligned} \tag{3.15}$$

While the on-time and off-time is originally used to model the situation, the equations for $\tau_{\text{on}}$ and $\tau_{\text{off}}$ may be substituted to depict the behaviour of the LEF in terms of its on-rate and off-rate of binding to DNA. In other words, this gives the mean copy number of LEFs per cell as a function of the fraction of time spent in the on- and off-state from chromatin.

Further, from the binding properties of an LEF, loop extrusion activity is increased by the amount of LEFs bound and extrude a loop. The loop extrusion activity from an LEF can then be given as a product of the rate of loop division and the amount of time it spends on a strand of DNA,

$$b = R_{\text{div}}\tau_{\text{on}} = \frac{R_{\text{div}}}{k_{\text{off}}} \tag{3.16}$$

The frequency of the loop extrusion activity is then determined by the rates that an LEF remains bound to DNA. By substituting the residence time equations of an LEF bound to DNA with its kinetic on- and off-rates, we determine the frequency $f$ as,

$$f = \frac{1}{\tau_{\text{on}} + \tau_{\text{off}}} = \frac{k_{\text{on}}k_{\text{off}}}{k_{\text{on}} + k_{\text{off}}} \tag{3.17}$$

Finally, Equation 3.15 may be rewritten in terms of the factors given in Equation 3.16 and 3.17, to yield

$$L_T = \frac{bf}{R_{\text{death}}} \tag{3.18}$$

As observed from these theoretical calculations, the importance of the four parameters given in Figure 3.7 largely determine loop extrusion activity: $k_{\text{on}}$, $k_{\text{off}}$, $R_{\text{div}}$, and $R_{\text{death}}$. These parameters will now be tested with simulations to model the trajectories of loop extrusion and residence time of LEFs to promote extrusion. However, prior to testing the model, the burst model is introduced to describe how significant peaks are determined in the data. To analyze the dataset, significant bursts, or time periods of data with significant activity, are identified to narrow the area of data for examination.

### 3.3.8 Burst Model

As previously mentioned, the binding process for an LEF may parallel that of the burst model. Burst detection was previously described by Kleinberg [40], where a burst detection algorithm was described with the purpose of identifying time periods in which a target event is uncharacteristically frequent, or "bursty". Burst detection may be used to detect bursts in a continuous stream of events, or in discrete batches of events.

Recall in Figure 3.7 the loop extrusion dynamics with LEF binding. This process can also be simplified into three stages, as illustrated below. To represent Figure 3.8 mathematically, the process can be summarized by two differential equations,

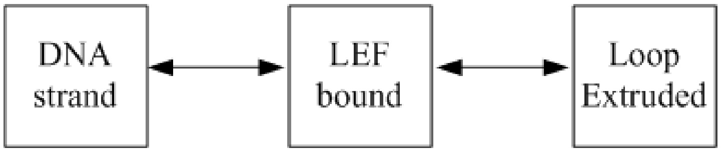

Figure 3.8: Simplified loop extrusion model

$$\begin{aligned} \frac{dm}{dt} &= \alpha_m - \tau_m^{-1} m \\ \frac{dn}{dt} &= \alpha_n m - \tau_n^{-1} n, \end{aligned} \tag{3.19}$$

where $m$ represents the rate of LEF binding to the DNA strand, and $n$ represents the rate of loop extrusion. Note that these equations ignore the fluctuations of the burst model.

In the steady-state, these are represented by

$$\begin{aligned} \bar{m} &= \alpha_m \tau_m \\ \bar{n} = \alpha_n \tau_n \bar{m} &= \alpha_m \tau_n \alpha_n \tau_m \end{aligned} \tag{3.20}$$

Two dimensionless parameters are now defined to ease the future calculations,

$$\begin{aligned} \bar{b} &= \alpha_n \tau_m \\ a &= \alpha_m \tau_n \end{aligned} \tag{3.21}$$

It is observed that the value of $\bar{b}$ is the mean number of loops produced from the LEFs bound to DNA. When there is more than one loop present in the system, the loops are produced in "bursts". This quantity will be denoted as the mean-burst size. This value was been approximated by a geometric distribution, which is a discrete version of the exponential distribution. The quantity $a$ measures the mean-number of bursts per cycle.

The burst sizes are distributed in a geometric manner, as mentioned, since it is a discrete version of the exponential distribution. This is now derived. We start by considering a single LEF molecule that can be bound to DNA at rate $\alpha_m$, and can be destroyed at rate $\tau_m^{-1}$. Most of the time, nothing will typically happen. In fact, using the same intuition as the Gillespie algorithm, we can calculate the waiting time distribution between events. However, for calculating the number of loops produced from each LEF molecule, this distribution plays no role. Instead, we need to know the probability that when an event happens, it is the production of a loop rather than an LEF unbinding. This probability can be written by the following,

$$q = \frac{\alpha_n}{\alpha_n + \tau_m^{-1}} = \frac{\bar{b}}{\bar{b}+1}, \tag{3.22}$$

which is the second step of the Gillespie algorithm.

The probability of producing exactly $b$ bursts is then given by

$$P_{\text{burst}}(b) = q^b(1-q) = \frac{1}{1+\bar{b}}\left(\frac{\bar{b}}{\bar{b}+1}\right)^b, \tag{3.23}$$

which represents the Geometric distribution. Since the geometric distribution is the discrete analogue of the exponential distribution, we can write,

$$\begin{aligned} P_{\text{burst}}(b) &= \frac{e^{-\ln(1+1/\bar{b})b}}{1+\bar{b}} \\ &\approx \frac{e^{-b/\bar{b}}}{\bar{b}} \end{aligned} \tag{3.24}$$

Note that in the second line, a Taylor expansion was performed in $1/\bar{b}$ which is valid when $\bar{b}$>>1.

This is now used to derive the Gamma distribution for loop abundances. To do so, we will write a loop-only master equation. We will call the number of loops $x$.

$$\frac{dn(x,t)}{dt} = \underbrace{\alpha_m \int db P_{\text{burst}}(b) p(x-b,t)}_{\text{term 1}} - \underbrace{\alpha_m p(x,t)}_{\text{term 2}} + \underbrace{\tau_n(x+1)p(x,t)}_{\text{term 3}} - \underbrace{\tau_m x m(x,t)}_{\text{term 4}} \tag{3.25}$$

The first term represents the probability that you have $x-b$ loops and produce burst of size $b$ which is integrated and summed over all burst sizes $b$. The second term is the probability that you have $x$ loops and produce a burst of any size. The third term is the probability that you have $x+1$ loops and degrade a loop. The fourth term is the probability that you have $x$ loops and degrade a loop. When $x>>1$, we can approximate this Master equation by a Fokker-Planck equation of the form,

$$\frac{dm(x,t)}{dt} = \alpha_m \int db P_{\text{burst}}(b) m(x-b,t) - \alpha_m m(x,t) + \partial_x(\tau_m^{-1} x m(x,t)) \tag{3.26}$$

At the steady-state, this becomes

$$a \int bd[P_{\text{burst}}(b) - \delta(x)]p(x-b) = -\partial_x(xm(x,t)), \tag{3.27}$$

with $a = a_m \tau_n$ the mean number of bursts per cycle. Note that the left hand side of Equation 3.27 is just a convolution of two distributions: an exponential distribution and the distribution we want to solve for. Equation 3.27 can then be solved, and isolated for $\hat{P}_{\text{burst}}(s)$ such that,

$$\hat{P}_{\text{burst}}(s) = \frac{1\bar{b}}{s + 1/\bar{b}} \tag{3.28}$$

Thus, we can take the Laplace transform of both side to obtain

$$-\frac{as}{s + 1/\bar{b}} \hat{p}(s) = s \partial_s \hat{p}(s) \tag{3.29}$$

which yields

$$\hat{p}(s) = \frac{1}{(s + 1/\bar{b})^a}. \tag{3.30}$$

Note that this is the Laplace transform of the Gamma distribution. The distribution of loop abundances on a strand of DNA is then described by the following gamma distribution,

$$m(x) = \frac{x^{a-1} e^{-x/\bar{b}}}{\Gamma(a) \bar{b}^a} \tag{3.31}$$

It was previously stated that Gamma distributions are able to model a wide range of molecular mechanisms relevant for gene switching and transcription initiation [64]. It was also stated that an LEF, cohesin, has exponentially distributed binding times to DNA and are highly stochastic [31]. This implies that the above burst model to model the binding and looping behaviour of chromatin is appropriate based on the available experiments conducted. However, due to the limited experiments available for consistent results, the burst model is not without limitations. Imaging was only able to confirm that the binding behaviour of cohesin to cognate sites are exponentially distributed because few events exhibit the mean value[31]. This does not necessarily imply that all LEFs will exhibit this behaviour because this result was not widely repeated with many experiments. This will be further discussed in the Discussion section.

The burst detection algorithm was implemented with the Stochpy package in Python [46] to detect bursts in one of Stochpy's built-in modules. This is implemented in the analysis of LEFs on loop kinetics both in steady state and non steady state binding behaviour. In this application, the burst model is used to model the binding of LEFs and extrude the loop. It is illustrated in a manner that shows when an LEF is either bound or unbound on a strand of DNA.

## 3.4 Results

### 3.4.1 Steady state binding

The steady state assumption in enzyme kinetics was proposed by George Briggs and John Haldane in 1942. This assumes that the concentrations of the intermediates of a reaction remain the same, even when the concentrations of starting materials and products are changing. In other words, the rate of formation and breakdown of the intermediate are equal. Both the formation of the intermediate from reactions and the formation of products from the intermediate have rates much higher than their correponding reverse reactions. An application of this example is found in Michaelis-Menten enzyme kinetics, where there is a closed-form solution for the concentration of reactants and products in an enzymatic reaction. The steady-state assumption assumes a negligible rate of change in the concentration of the enzyme-substrate complex during the course of the reaction.

While the binding of an LEF to DNA may not wholly exhibit Michaelis-Menten properties since there are many activation sites on a single strand of DNA, the steady state assumption is applied in this model. This assumption was previously discussed by Goloborodko et al [30], where loop dynamics are controlled by competition between loop death and division. This implies that the rate of binding and unbinding of an LEF onto a strand of DNA is equal. Since the loop stacking process may be given by an immigration-death process, as previously described, the steady state is given by the following equation,

$$R_{\text{div}} - R_{\text{death}}(n_{\text{loops}}) = 0 \tag{3.32}$$

The average number of loops can then be given by the ratio between the rate of loop division and the rate of loop death, as denoted in Equation 3.33, without the consideration of the residence time of the LEFs.

$$n_{\text{loops}} = \frac{R_{\text{div}}}{R_{\text{death}}} \tag{3.33}$$

Since the LEF stacking activity is modelled with an IDP, it follows a Poisson process. The arrival of an event is independent of the prior event, which satisfies the Markov property. Therefore, the probability mass function of loop extrusion is given by,

$$f(n, \lambda t) = P(LE = n) = \frac{\lambda t^n e^{-\lambda t}}{n!} \tag{3.34}$$

where $n = L_T$, and $\lambda = n_{\text{loops}} = \frac{R_{\text{div}}}{R_{\text{death}}}$. The waiting time until the next event may also be established. This can then be denoted by a decaying exponential function, as the probability of waiting a given amount of time between successive events decreases exponentially as the time increases. The following equation then denotes the probability of waiting more than a specified time,

$$P(T{>}t) = e^{-\lambda t} \tag{3.35}$$

by using the same parameters as defined in Equation 3.34.

Now that these functions have been established, the relationship between the LEF binding and the loop extrusion activity may be simulated with the Gillespie algorithm. This activity is observed using Stochpy [46] software, as they have pre-established modules that utilize this algorithm in different biological processes.

To begin investigating the kinetic rates of the burst model, a literature search was conducted to determine appropriate parameters. The kinetic rates of loop activity were previously studied by Brackley et al [13], which will be applied in our models. Brackley et al [13] studied the formation of CTCF-mediated chromatin loops. A model was proposed for the formation of chromatin loops based on the diffusive sliding of molecular slip links. For the equilibrium state, $k_{\text{on}} = k_{\text{off}} = 0.04$. Throughout these investigations, the rate of loop extrusion and degradation will be fixed at $R_{\text{div}} = 5$, and $R_{\text{death}} = 10$, respectively, as defined by the parameters in Table 3.1. This result is illustrated in Figure 3.9. The upper plot of the two plots shown illustrates the amount of time an LEF remains bound to the strand of DNA. There are only two states in this plot: on and off. This plot is interpreted in a manner in which the size of the each rectangle shows the amount of time the LEF extrudes the loop. The bottom plot is a simulation of the loop extrusion activity via the immigration death process.

As desired, in Figure 3.9, loop extrusion activity occurs when the LEF is in the “on” state. Loop extrusion activity remains at 0 when the LEF binding is in the “off” state. However, the loop extrusion activity does not always behave in the same manner each time the loop is bound to DNA. In other words, there are some times in which the loop extrusion activity may peak higher, or it may not be as active. When a loop has been bound to DNA for a longer period of time, the loop extrusion activity is still constantly fluctuating. It is neither consistently increasing nor decreasing. With these observations in mind, a higher kinetic rate for the steady state condition of the binding of LEF to DNA is now used, which was also explained in the study conducted by Brackley et al [13]. In this case, $k_{\text{on}} = k_{\text{off}} = 0.4$.

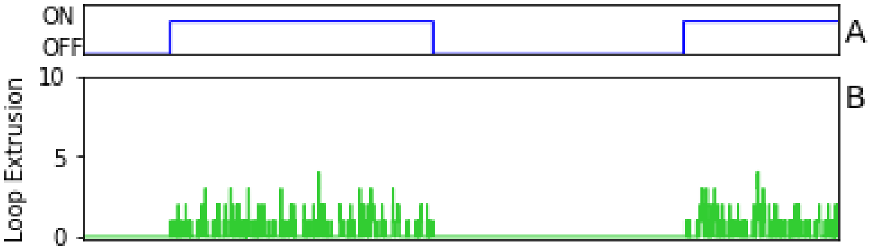

Figure 3.9: Low steady state binding rate, where $k_{\text{on}}/k_{\text{off}} = 0.04$. The upper plot denotes the activity of the LEFs bound and unbound from the strand of DNA by their on/off states. The bottom plot denotes the loop extrusion activity that corresponds to the state of the LEF binding.

It is perhaps expected that similar activity will occur, since these scenarios are in a steady state, and the kinetic parameters do not differ as a whole. The results are illustrated in Figure 3.10, with similar plots to Figure 3.9 to show loop extrusion activity.

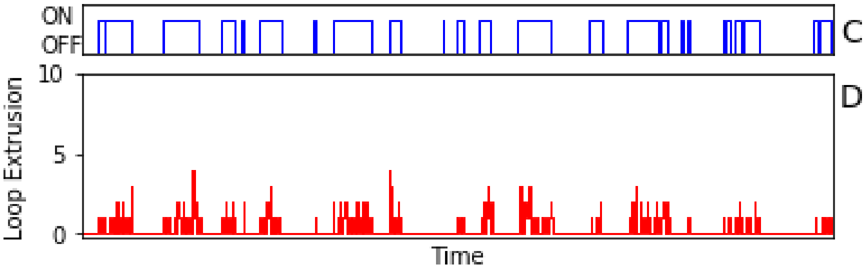

Figure 3.10: High steady state binding rate, where $k_{\text{on}}/k_{\text{off}} = 0.4$. The upper plot denotes the activity of the LEFs bound and unbound from the strand of DNA by their on/off states. The bottom plot denotes the loop extrusion activity that corresponds to the state of the LEF binding.

The steady state kinetics for the rate of LEF binding and unbinding to DNA is higher for this case, so it is expected that the frequency of binding would have increased. As observed from Figure 3.10, there is barely any time in which an LEF is not bound to DNA. As soon as one LEF is found in the off-state, another LEF would bind immediately and overlap with the previous LEF's exposure time on the DNA strand. This is suggestive of a stacked LEF state, in which there may be several LEFs that extrude a loop. This makes this model appropriate for the immigration death model proposed by Goloborodko et. al. [30]. However, this was not observed in Figure 3.9, in which the kinetic rates were much lower. While it is possible that the residence time of the LEF is much shorter in the present case, and the rate of binding of new LEFs may not overlap at all, it is not possible to distinguish between the states in this plot for the proposed model. However, previous publications have discussed the possibility of reinforced loops, in which this is supportive of this idea, theoretically. Note that the plot depicting the IDP (in red) in Figure 3.10 shows similar behaviour in terms of the loop extrusion activity in Figure 3.9. It is also not evident of the activity due to the corresponding LEF bound to DNA, therefore reinforced looping must be occurring.

Since the higher rates depict a model suggestive of chromatin looping, it implies that the rate of kinetics for the on and off rates of LEF binding should be greater than $k_{on} = k_{off} = 0.4$. Should there be experimental proof of loop reinforced extrusion, then the LEF rates should also show higher kinetic rates. However, LEF binding in a steady state is only suggestive of one possible method of LEF binding. LEF binding activity is now explored with non steady state kinetics to explore how this model performs in these conditions.

### 3.4.2 Varying the rates of binding and unbinding

Suppose the previous assumption for steady state binding kinetics do not apply, and the binding rates of LEFs do not occur in the same manner as described above. This may be possible due to the uncertainty of LEF stacking activity, and that there may be several contributors to loop extrusion. Since the properties and identities of an LEF are still unknown, it is possible that non steady state kinetics may occur for loop extrusion activity. This implies that upon unbinding from chromatin, the LEF may not immediately rebind to another randomly chosen site along the chromosome. In Brackley et al's research [13] with the proposed kinetic rates for sliding, they even proposed a nonequilibrium model where the binding and unbinding kinetics of cohesin violate detailed balance. This models the fact that both its loading and unloading onto chromatin requires ATP. It was emphasized that passive LEF sliding would be the driver of loop formation and degradation. The merit of this assumption should not be ignored, since there is evidence that suggests that this enzyme behaves in a non-steady state manner.

To begin this investigation, let the parameters of $R_{\text{div}}$ and $R_{\text{death}}$ remain identical to the values used in the previous assessment. These values remained unchanged because we wish to isolate the steady state component of LEF binding to DNA. This will allow for a better understanding of how the presence of LEFs may impact loop extrusion activity. The values of $k_{\text{on}}$ and $k_{\text{off}}$ used in this case is then 0.4 and 0.04, respectively, where $k_{\text{on}} > k_{\text{off}}$. The results of this simulation is illustrated in Figure 3.11.

Similar to what was observed in Figure 3.10, there is a higher rate of LEFs bound to DNA, and there is very little area which indicates there was no LEF bound at any specific time. The maximum peaks of the loop extrusion activity (green) in this plot does not indicate a correlation with the binding of LEF activity. It is noted that the loop extrusion activity behaves in a similar manner to that of the steady state representation. This is to be expected, as the IDP is still used to model the same procedure. The only difference is the frequency of the on and off rate of binding for loop extrusion activity on the loop itself. The rates of loop

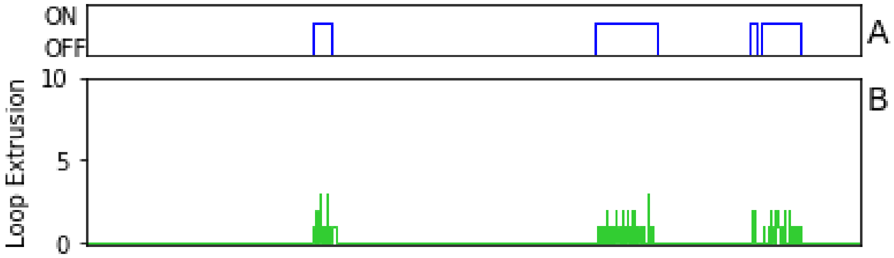

Figure 3.11: The rate of $k_{\mathrm{on}}$ is greater than that of $k_{\mathrm{off}}$ in an unsteady state. The upper plot denotes the activity of the LEFs bound and unbound from the strand of DNA by their on/off states. The bottom plot denotes the loop extrusion activity that corresponds to the state of the LEF binding.

divison and death have remain unchanged. While there are some sections in which there are narrower rectangles for binding activity, it may be indicative that loop reinforcement activity will occur less frequently.

The behaviour of the unbinding LEF activity in the non steady-state is now examined. While the rates of LEF extrusion activity remains the same, such that the parameters of $R_{\mathrm{div}}$ and $R_{\mathrm{death}}$ remain consistent to the previous plots, the rates $k_{\mathrm{on}}$ and $k_{\mathrm{off}}$ are now reversed in their roles. The values of $k_{\mathrm{on}}$ and $k_{\mathrm{off}}$ are now 0.04 and 0.4, respectively, and $k_{\mathrm{on}} < k_{\mathrm{off}}$. This simulation is illustrated in Figure 3.12.

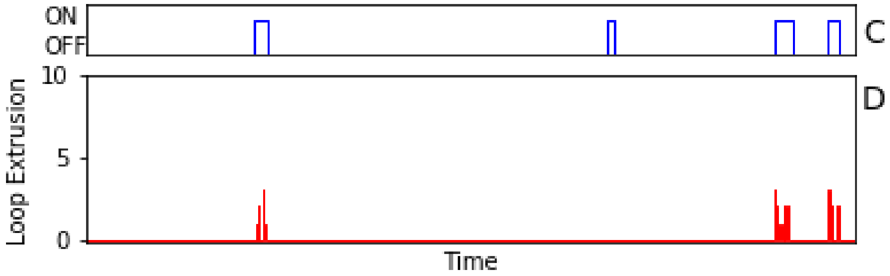

Figure 3.12: The rate of $k_{\mathrm{off}}$ is greater than that of $k_{\mathrm{on}}$ in an unsteady state. The upper plot denotes the activity of the LEFs bound and unbound from the strand of DNA by their on/off states. The bottom plot denotes the loop extrusion activity that corresponds to the state of the LEF binding.

It is observed that Figure 3.12 shows results that are much different that what was explored in the previous plots. It was thought that as soon as an LEF is bound to DNA, it will immediately extrude a loop. The scenarios in which loop extrusion may not occur once an LEF is bound to DNA is not yet known, however this plot suggests this could be a possible event. As observed from the binding plot of LEF, there is very short and somewhat frequent activity of binding to DNA. Although, the width of the rectangles are not very wide as expected, since the rate of unbinding is greater than that of binding. It is also observed that there is much more space between the rectangles, suggesting a longer waiting time for another LEF to bind to DNA successfully to perform loop extrusion. The IDP activity also shows very little activity, but does not seem to wholly conform to the binding behaviour of LEFs. While it still only extrudes the loop while the LEF is bound to DNA, it is not always guaranteed.

To continue investigating the binding of LEFs, the residence time is now examined theoretically. By observing the results in Figure 3.12, it is interesting to think about the impact of having an LEF bind to DNA and not actually perform loop extrusion activity. While available literature is only speculating about the true properties of an LEF, we only consider cohesin as an example. It is thought that cohesin binds specifically to DNA in a specific target area that may be rich in A-T sequences. Therefore, loop extrusion may only occur when an LEF binds specifically to DNA. Should an LEF bind unspecifically in DNA, this may result in false results that suggest loop extrusion activity may occur. Therefore, the probability of this occurring may be considered mathematically in the following manner,

$$P(t) = Ae^{-k_{\mathrm{ns}}t} + Be^{-k_{\mathrm{sp}}t} \tag{3.36}$$

where $A$ and $B$ are some constants, and the rates are defined as,

- $k_{\mathrm{ns}}$ = unbinding rate for non-specific binding
- $k_{\mathrm{sp}}$ = unbinding for specific binding

Intuitively, the residence time of the model should be defined as,

$$\tau_s \approx 1/k_{\mathrm{off}} \tag{3.37}$$

where $\tau_s$ is the residence time of specific binding, which is of interest because only specific binding to the DNA strand will promote loop extrusion. Equation 3.37 is only an approximation, as the off rate may be influenced by non-specific binding as well. The impact is not shown here, as the amount of bias this will ensue for the residence time. As observed by the simulations, it is unclear how one may be able to distinguish between specific and non-specific binding. This affect may not be negligible, and therefore should be acknowledged when interpreting these results.

Therefore, the IDP for loop extrusion may be influenced by both specific and non-specific binding. Nonetheless, the rates of LEF binding to DNA, as well as the rates of loop extrusion will influence the loop extrusion activity that will perform chromatin compaction. After running these simulations, it would be interesting to explore the specific positions of LEF activity, to further explore whether loops in chromatin are reinforced by several LEFs. Since these results only suggest that LEF binding may occur, the actual rates of binding of LEFs are still uncertain, since it is unsure of which enzymes are confirmed to be LEFs. It would be interesting to explore the range of rates that these LEFs would bind and unbind to DNA, as this is what drives the chromatin compaction activity. While this work is simply theoretical and speculatory, there have been other publications that suggest that these findings would support their conclusions regarding reinforced loop extrusion activity {[1], [23], [30]}. The exposure timescales of LEFs will be further explored in the next chapter.

## 3.5 Discussion

There are several kinetic rates that drive loop extrusion activity. These include the rates of: loop division, death, LEF processivity, and LEF separation. Through the use of continuous-time Markov chains, these rates can be utilized in stochastic equations to model chromatin activity depicting the shape, size, and stability of chromatin loops. The behaviour of LEFs can be modelled with the help of the Immigration-Death process, the Gillespie algorithm, and the burst model. What is novel is that existing binding kinetics have been tested against the burst model to determine how this can be used to show the fluctuations in LEF stacking to reinforce chromatin loops. This can allow for further simulations to be conducted when experimentation becomes limited.

The Immigration-Death model is appropriate to model chromatin loop division and loop death due to the manner in which it allows an LEF to transition from one state to another. The traditional Birth-Death model is not appropriate, because LEF binding activity is suspected not to be dependent on the current number of LEFs on the extruded loop based on the potential stacking ability to stabilize a loop [30]. Loop death, however, is dependent on the number of LEFs bound to the loop since it is less likely to die when it is reinforced by several LEFs. Goloborodko et al [30] have discussed how the rates of binding and unbinding of LEFs are also dependent on environmental factors. In the Immigration-Death model, one thing that was not consistent between the rates of division and death, and the rates of binding and unbinding is the specificity of the location of DNA in which this occurs. While the rate of loop division is dependent on the number of LEFs bound to DNA, it is not specific to the location of DNA in which it is bound. The only influence of spatial location of these LEFs is addressed in the LEF separation. However the affinity of the DNA is not considered. This could be that once the LEF has stably bound to DNA, the aspect of the specificity was already taken care of. Thus, the specificity of the rates of loop division and loop death is not required as a factor in these rates. The number of loops extruded has already included both of these rates in the same equation. Therefore, this should be treated as a whole system, rather than as separate components when considering the complete process of chromatin looping.

One idea that was explored was the activity of multiple LEFs bound to the chromatin loop extrusion. When there were more LEFs bound to the system, it was difficult to observe whether this would have increased loop extrusion activity due to the amount of noise generated from the IDP. The number of LEFs may only affect the strength of the loop rather than its activity. It may be that the residence time of the loop has a greater influence on the loop extrusion activity. Some factors that can be individually explored within the rates would be whether the specific components such as the sequence of DNA on the loop would promote an increased velocity of loop extrusion. If only specific DNA sequences should be extruded from the loop, this would raise questions on the kind of information stored during the chromatin loop process, and what information is lost should it not be extruded.

A dynamical model was proposed by Brackley et al [13] where a molecular slip link might organize chromosomal loops. It was shown that diffusive sliding of cohesin explains the experimentally observed bias favouring convergent over divergent CTCF loops. Second, the probability of formation of cohesin/CTCF-mediated loops does not obey a power law, in stark contrast with the case of polymer loops in thermodynamic equilibrium. Finally, when multiple links slip links bind to chromatin at a “loading site” rather than randomly, a ratchet effect arises, which favours the formation of much larger loops than are possible with single slip links. These observations were dependent on their assumption that that the cohesin binding kinetics violate detailed balance, which is motivated by the fact that its loading and unloading requires ATP. While these observations were not incorporated into our current model, it is still important to consider in future experimental observations. Since this is only one perspective on external factors that influence chromatin organization, it may not be valid for all LEF candidates for active loop extrusion. Nonetheless, it is still important to consider in the future, when there is more extensive experimental observations that can validate its inclusion for the kinetic rates that motivate chromatin looping.

Based on the parameters that contribute to the rates of loop division and loop death established by Fudenberg et al. [23], they have deduced that LEF stacking activity are based on LEF processivity and LEF separation. It was discussed that the protein link between each binding domain leads to extrusion of a DNA loop. This association of machines and DNA, known as infinite processivity, forms a disordered distribution of small loops. However, if dissociation of the machine and DNA occurs, known as finite processivity, highly stable and large DNA loops are formed with few fluctuations. The importance of LEF processivity is similar to that discussed by Brackley et al [13] where LEF sliding propels the formation of chromatin loops, rather than the actual binding of LEFs. Bonato et al have used Brownian dynamics to study the diffusive sliding of molecular slip links, thus mimicking the behaviour of cohesin molecules [8]. They have proposed that diffusive sliding is sufficient to explain the chromatin loop extrusion of hundreds of kilo-base pairs, which may then be stabilised by interactions between cohesin and CTCF proteins. The loop extrusion rate, or processivity here has shown that the rates of the immigration-death process are crucial in the formation of stable chromatin loops. From the burst model, the actual rate of binding and unbinding had a small impact on the behaviour of chromatin organisation.

The specificity of binding to the strand of DNA to allow for loop extrusion to occur should be incorporated into the model, based on the work that has already been established for potential LEF candidates. While it seems that the actual LEF processivity will propel the model, it may be that the LEF specifically will promote processivity based on specific sequences that need to be conserved in the strand of DNA. This will influence the fluctuation caused by the immigration-death model of the stability of the loops. In a similar manner, factors that promote the rate of loop formation, or loop extrusion, should be further examined and included in the model. There are some limitations in the rates of division and death as derived by Golobordko et al [30]. Some research has explored the possibilities of ATP propelling loop formation, which may need to be included into the rates. Given the presence of ATP, this may increase the rate of loop extrusion.

In our model, the burst model would appropriately support Brackley et al's [13] experimental observations. In the nonequilibrium model, it would appropriately show that upon binding of an LEF onto a strand of DNA, it does not necessarily mean loop extrusion will occur. The sliding rate is more dependent on loop formation and degradation. Although the individual rates of loop binding, unbinding, division, and death were fixed, the Gillespie algorithm allowed for varied results for each run. The shape of the data was not necessarily duplicated in the next simulation run. This allowed for a better interpretation of the random output that can be yielded. The nonequilibrium model will also allow for a better interpretation of real world behaviour, as the loop behaviour may vary according to what kind of information in stored in a chromatin loop and for what purpose that loop serves.

An external factor that may influence chromatin looping activity with cohesin as a potential LEF is the influence of chaperone enzymes. It was determined by Garcia-Luis et al [26] that the role of cohesin in chromosome organization requires the histone chaperone FACT ('facilitates chromatin transcription') in yeast. It was determined that FACT interacts directly with cohesin, and is dynamically required for its localization on chromatin. Depletion of FACT in metaphase cells prevents cohesin accumulation at pericentric regions and causes reduced binding on chromosome arms. This implies that without the presence of FACT, cohesin would not be stabilized, and would not be able to extrude the cohesin loop, even if it is bound on the strand of DNA. This is was represented in some of the burst model experimentation, where the LEF may be bound, or "ON" , in the model, but does not perform any immigration-death activity to extrude the loop. This implies that an external factor may influence the extrusion activity in terms of the sliding rate, as the cohesin must be stabilized by the histone chaperone in order to perform its activity.

The binding behaviour of cohesin in human cells was analyzed by Holzmann et al [32]. They have measured absolute copy numbers and dynamics of cohesin, CTCF, NIPBL, WAPL and sororin by mass spectrometry, fluorescence-correlation spectroscopy and fluorescence recovery after photobleaching in HeLa cells. It was determined that if cohesin extrudes loops of chromatin, it is possible that it does so without topological entrapment, raising the possibility that two populations of cohesin exists in cells, one that is competent for loop extrusion and the other for cohesion. The number and position of TADs and loops does not difer significantly between G1 and G2 cells, that is chromatin architecture does not detectably change even though many cohesin complexes are stably bound to chromatin in G2. They deduced that considering around half of chromatin-bound cohesin is stably-bound in G2 and may not function in loop extrusion, they needed further analysis with LC-MS, FCS and FRAD data to compare the number of dynamically chromatin-bound cohesin complexes in cells synchronised in G1 and G2. They deduced that either stably-bound cohesin participates in genome organisation in some way - without any of the changes in chromatin structure observed following WAPL depletion - or the two-fold increase in DNA content in G2 must be organised by relatively fewer cohesin complexes. In our burst model, this was demonstrated in the non-equilibrium model where the rate of LEF binding was greater than the rate of LEF unbinding. This showed behaviour where LEF bound did not exhibit loop extrusion activity. This also enforces that the steady-state assumption does not apply when modelling the behaviour of LEF loop extrusion activity. The assumption that every LEF bound to DNA will perform loop extrusion activity will also not be applicable when creating a model. Therefore, the burst model with non equilibrium binding characteristics is well supported by current assumptions in chromatin organization.

## 3.6 Future directions

The kinetic rates of LEF binding are still speculatory in today's research and more work is needed to confirm proposed models. In this research however, it was determined that the given evidence was appropriate in supporting the present literature that explores the behaviour of chromatin formation. Although, it is acknowledged that this model contains assumptions that may not hold true in the future. There could be other key factors or steps that are crucial in modelling the chromatin compaction process that are not in this current model.

The specificity of binding of an LEF to extrude the loop should also be incorporated into this model. Since it is difficult to identify a potential enzyme that acts as the known characteristics of an LEF, it is challenging to assign the chemical properties that should be included in this model. In addition, it has been speculated that loop extrusion is driven by post-translational modifications, ATP concentration, etc. The purpose of the specific loop extruded should also be considered, such that perhaps the type of DNA sequences it contains may be more conserved than other genetic information. This could serve some purposes such as transcription, gene modification, duplication, etc. More experiments are required to support the motivation that drives loop extrusion.

There are now robust methods for evaluating likelihoods for realizations of Birth-Death processes: finite-time transition, first passage, equlibrium probabilities, and distributions of summary statistics that arise commonly in applications. Recent work has also exploited the connection between continuously- and discretely-observed BDPs to derive EM algorithms for maximum likelihood estimation. Likelihood-based inference for previously intractable BDPs is much easier than previously thought and regression approaches analogous to Poisson regression are straightforward to derive [17]. As these tools are starting to develop, this can be applied to the future for the Immigration-Death model to optimize the loop divison and death rates. Given the limited observed data able to predict these parameters, it is not possible to obtain an accurate representation of the model currently. However, when the tools become available to observe the chromatin looping procedure to gain a comprehensive understanding of its organization, this will provide more insight in the rates of LEF behaviour for chromatin loop organization.

As discussed, there has been recent extensive research toward examining the specific drivers involved in the chromatin looping process. Racko et al [56] have shown that growing plectonemes resulting from transcription-induced supercoiling have the ability to actively push cohesin rings along chromatin fibers. This evidence supported earlier explanations proposing why TADs flanked by convergent CTCF binding sites form more stable chromatin loops than TADs flanked by divergent CTCF binding sites. It was proposed that transcription of eRNA (enhancer RNA) sends the first wave of supercoiling that can activate mRNA transcription in a given TAD. If this is true, then the effect of supercoiling should be introduced into the burst model to add additional components to account for potential transcription activity that may propel the chromatin looping process. This may add more steps to the model, so that looping activity can be controlled depending on specific factors. This implies if this model were true, that chromatin fibers passing through cohesin rings experience significant hydrodynamic drag limiting their free rotation [56]. This decreases the diffusion rate of cohesin. Should all LEFs behave in this manner in which some external factors will influence their kinetic rates, some variable should be included to account for these influences.

## 3.7 Conclusion

Several CTMC models were successfully applied to predict the binding behaviour of LEF to DNA for chromatin loop formation. The immigration-death model was able to predict the stability of chromatin loops, with the aid of the Gillespie algorithm to simulated its binding activities. The burst model was able to show the trajectory of an LEF upon binding to DNA to the actual loop extrusion procedure. It was determined that the non equilibrium model was able to show the real application of chromatin behaviour since the potential LEF candidates do not follow the equilibrium assumption. For future development of this model, external factors that contribute to the loop extrusion kinetic rates should be incorporated into the model to account for different LEF candidates.

# 4 Theoretical Exposure Time for Loop Extrusion

## 4.1 Abstract

The presence of nucleosomes on a strand of DNA can impact loop extrusion factor activity by hindering its binding behaviour. The exposure timescale of LEF binding to a strand of DNA was derived, and verified through simulations. The specificity of binding is considered in the kinetic rates to include sequence specifications to promote loop extrusion activity. The impact of nucleosome activity impedes LEF binding and extrusion behaviour. Further research is required to fully understand the barriers in loop extrusion activity and the extent of what controls their activity, however this provides a preliminary understanding of how LEF activity is impacted by the presence of other enzymes on a strand of DNA.

## 4.2 Introduction

The topology of chromosomal DNA molecules is controlled by cell processes that have yet to be fully understood. One theory is that there is some external loop extrusion factor that organizes chromatin structure. This loop extrusion factor may bind to DNA on a specific sequence site in order for extrusion to occur. There are several other factors bound to DNA, or may bind to DNA, that may impede loop extrusion activity. In this chapter, the amount of time this specific sequence site remains exposed and ready for binding, is explored. Since nucleosomes share the same environment as LEFs, the impact of their presence on the strand of DNA is also explored for loop extrusion activity. This can play a role in transcription regulation, as loop extrusion extrudes some genetic information for storage.

Nucleosomes perform DNA packaging functions in eukaryotes, consisting of a segment of DNA wound in sequence around eight histone protein cores [3]. This is often visualized as thread wrapped around a spool (Figure 4.1). LEFs, on the other hand, also interact with DNA, but they pull DNA fiber into a loop through specific contacts with its protein scaffold. Since both nucleosomes and LEFs may be present on a single strand of DNA at the same time, one may wonder how they may perform similar functions in parallel.

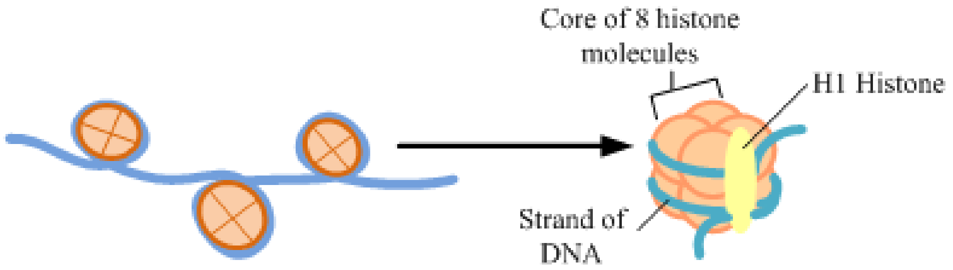

Figure 4.1: Diagram of nucleosome structure. The left of the arrow shows the "beads on a string" view of the nucleosome, whereas the right of the arrow depicts the side view of a nucleosome

An example of an LEF that has been studied in conjunction with the presence of nucleosomes is cohesin. It was discussed previously that one of the conditions for cohesin loading to achieve chromatin remodeling is nucleosome-free DNA [49]. There is a correlation between chromatin remodeling and cohesin loading onto chromosomes, which describe the entry point by which cohesin accesses DNA in the context of chromatin [49]. Based on this, it is important to consider the presence of nucleosomes when deriving the theoretical exposure times for LEF binding to DNA. It was also established under single molecule microscopy that cohesin can undergo rapid one dimensional diffusion along DNA, but individual nucleosomes, nucleosome arrays, and other protein obstacles significantly restrict its mobility [67]. It was also determined that cohesin prefers binding to A-T rich sites, therefore when considering its binding behaviour its sequence dependence should also be considered.

The impact of nucleosome presence for LEF extrusion factor mechanisms becomes important to understand both biological and mathematical perspectives to determine the extent of their impact for modelling purposes. Whether a nucleosome free section of DNA becomes a requirement for proper LEF binding could be significant for defining the probability of loop extrusion.

## 4.3 Materials and Methods

### 4.3.1 Random Walks in Chromatin Organization

The Gillespie algorithm is often used to model chemical reactions, which includes the behaviour of Markovian random walks in particle number space. In the previous chapter, the Gillespie algorithm was used to model the behaviour of LEF binding on DNA to promote loop extrusion activity. The burst model showed the significant trends in which the LEF binding may promote activity. In the previous chapter also, the burst model was used to show the significant trends of the IDP in a bursty manner. This showed LEF binding on a continuous time scale. In this case, the random walk will be presented to show on a discrete time scale how LEF binding may be modelled on a specific sequence of DNA. This will be used later when analyzing the behaviour of LEF binding and how its binding activity may be affected in the presence of nucleosomes.

### 4.3.2 Transition Rates

In the previous chapter, the burst model was explored and applied to chromatin loop organization. This behaviour is also comparable to a random walk model. Recall in Figure 3.1, the M/M/$\infty$ queue was illustrated as a Markov chain to model the IDP. This describes a stochastic process, where a sequence of possible events in which the probability of each event depends on the state attained from the previous event. A random walk is also a stochastic process that describes a path that consists of a succession of random steps on some mathematical space such as the integers. A random walk occurs in discrete time intervals, whereas the IDP occurs in continuous time. By examining the behaviour of LEF binding in the IDP in the previous section, there are some interesting properties that connect the random walk model and the IDP for the LEF behaviour in chromatin compaction. This was modelled on several occasions to represent ion-channel gating behaviour [47], where there is one open state as an absorption state, and either one or several closed states. The purpose of presenting the random walk in this case is to demonstrate how upon binding of an LEF for loop extrusion activity, the LEF may not remain in this state such that the LEF may become stacked, and the state of the loop will "walk" or wander away from its original exposed state. This may allow for the loop formation for chromatin compaction.

The transition behaviour of the IDP is analogous to a random walk with the difference that the transitions occur at random times, as opposed to fixed time periods in random walks. To begin implementing the random walk model, consider Figure 4.2 depicting the LEF reinforcement model,

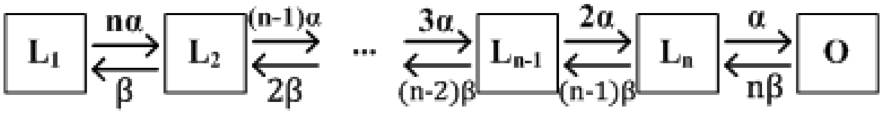

Figure 4.2: Random walk of LEF binding

The illustration may be represented mathematically by a master set of equations:

$$\begin{aligned}\frac{dp_{L_n}}{dt} &= (2\alpha + n\beta)p_{L_{n-1}} - ((n-1)\beta + \alpha)p_{L_n} \\ \frac{dp_{L_k}}{dt} &= (k-1)\beta p_{L_{k-1}} + (n-k)\alpha p_{L_{k+1}} - ((n-k)\alpha + (k-1)\beta)p_{L_k} \\ \frac{dp_{L_1}}{dt} &= \beta p_{L_2} - n\alpha p_{L_1},\end{aligned} \tag{4.1}$$

where $p_n(t)$ denotes the probability of occupying any state at time $t$, and $k \in (1<k<n)$.

Since random walks are represented in discrete time, the total probability is conserved and the rate of probability redistribution is determined by the set of transition rate constants. A compact formulation of the master equation is then given by,

$$\frac{d}{dt}P(t) = P(t)A \tag{4.2}$$

The transition probability matrix is given by the following Matrix A:

$$A = \begin{bmatrix} -n\beta & \alpha & & & & \\ n\beta & -((n-1)\beta - \alpha) & 2\alpha & & & \\ & (n-1)\beta & -(2\alpha + (n-2)\beta) & & & \\ & & (n-2)\beta & & & \\ & & & \ddots & & \\ & & & & (n-1)\alpha & \\ & & & & -((n-1)\alpha + \beta) & n\alpha \\ & & & & \beta & -n\alpha \end{bmatrix} \tag{4.3}$$

where the solution to this matrix would be $P(t) = \exp(At)$.

For this rate scheme, standard Markov analysis yields the average times for when the strand of DNA is occupied by an LEF, or when it is exposed, as given by the following equations,

$$\begin{aligned} E[T_O] &= 1/(n\beta), \\ E[T_L] &= \frac{(\alpha+\beta)^n - \alpha^n}{n\alpha^n\beta} \end{aligned} \tag{4.4}$$

For n > 1, when there is more than one LEF present on the loop and the state moves by chance to the left toward the stable loop state, rate constants become progressively slower. Thus, the average “dwell time” in states $L_0$ or $L_1$ is much longer than the dwell time in state $L_n$ or $L_{n-1}$. In the limit as $n \to \infty$, this model converges to power-law distributions of dwell times. Because this scheme follows the Markov assumption of memoryless performance, we are able to simulate the behaviour of such LEFs using the Gillespie algorithm.

The Gillespie algorithm is conducted by choosing an exponentially distributed random number for the dwell time that accounts for all of the possible transitions out of the current state, which uses the sum of the transition probabilities. After the length of the dwell time in the current state is determined, the destination state is selected by choosing a uniformly distributed random variable on an approximated partitioned interval. This will be observed in the later section when conducting simulations regarding the behaviour of LEF binding.

### 4.3.3 Random Walk Application: LEF Binding

In Section 3.3, the binding behaviour of LEFs was described with the burst model. This behaviour may also be described with a random walk model, as introduced above in a different context, such that the impact of the search time of the LEF upon binding DNA is explored. This will later become impactful when considering the exposure time of DNA for LEF binding in the presence of nucleosomes. The random walk model is derived by explaining the search time of LEF on DNA, how it impacts protein-DNA binding energy, then finally the mean first-passage time is presented. The theoretical random walk is presented here to gain an understanding of the movement of the LEF on a strand of DNA. Simulations for this behaviour will be presented in the later section in presence of nucleosomes.

Previous work conducted by Slutsky [66] discussed the problem of how a protein finds its target site on DNA. When a protein binds to a DNA site, it binds via three-dimensional diffusion, and hits the right site of the target DNA. If a protein performs both three-dimensional (3D) and one-dimensional (1D) diffusion, then the total search process can be considered as a 3D search followed by binding DNA and a round of 1D diffusion. Upon dissociation from the DNA, the protein continues 3D diffusion until it binds DNA in a different place, and the cycle continues.

The search time is explored first to describe how upon LEF binding, it may slide on the strand of DNA to find a specific sequence of DNA prior to extruding a loop. This process would then include $N$ rounds of 1D searches which each take $\tau_{\alpha,i}$, where $i = 1, ...N$, separated by rounds of 3D diffusion ($\tau_{\beta,i}$). The total search time is the sum of the times of individual search rounds:

$$t_s = \sum_{i=1}^{N} (\tau_{\alpha,i} + \tau_{\beta,i}) \tag{4.5}$$

The total number $N$ of such rounds occuring before the target site is eventually found is very large, so probability distributions are required. One simplification that can be made is for $\tau_{\beta,i}$, where it is replaced by its average $\tau_\beta$. This is valid when the distribution of 3D diffusion times inside the DNA nucleoid is sufficiently narrow. Each round of 1D diffusion scans a region of $n$ sites, where $n$ stems from some distribution $p(n)$. The time $\tau_\alpha(n)$ it takes to scan $n$ sites can be obtained from the exact form of the 1D diffusion law. If, on average,

$n$ sites are scanned in each round, then the average number of such rounds to find the site on a strand of DNA of length $M$ is $N = M/\bar{n}$. Equation 4.5 can now be substituted with these average values to obtain the following,

$$t_s(n, M) = \frac{M}{\bar{n}}[t_\alpha(n) + \bar{\tau}_\beta] \tag{4.6}$$

As observed from this equation, when $\bar{n}$ is small, then $t_s(\bar{n}, M)$ is large. When $\bar{n}$ is small, this implies that very few sites are scanned in each round of 1D search and a large number of rounds are required to find the site. Conversely, if $\bar{n}$ is large, more time is spent scanning a single stretch of DNA, which would render the search inefficient.

The binding potential is now explored for its implications in the random walk. When diffusing along DNA, the LEF may experience different binding potential $U(\vec{s})$ at every site $\vec{s}$ it encounters. The energy of the protein-DNA interaction requires consideration from both specific and non-specific binding, as discussed in the previous chapter. This is portrayed in the following equation, where $\vec{s}$ describes a DNA sequence of length $l$.

$$U_i = U(\vec{s} = s_i, ...s_{i+l-1}) + E_{\text{ns}} \tag{4.7}$$

The non-specific binding energy $E_{ns}$ arises from interactions that do not depend on the DNA sequence that the LEF is bound. The specific part of the interaction energy exhibits a strong dependence on the actual sequence. For example, cohesin loading possesses an affinity for A-T rich sequences [10]. Energy is referred to the change in the free energy related to binding, $\Delta G_b$.

The energy of specific protein-DNA interactions can be approximated by a weight matrix, where each nucleotide contributes independently to the binding energy,

$$U(\vec{s} = s_i, ...s_{i+l-1}) = \sum_{j=1}^{l} \epsilon(j, s_j) \tag{4.8}$$

where $s_j$ is a base-pair in position $j$ of the site and $\epsilon(j, x)$ is the contribution of base-pair $x$ in position $j$. For a sufficiently long site, the distribution of the binding energy of random sites can be closely approximated by a Gaussian distribution with a certain mean $\langle U \rangle$ and variance $\sigma^2$ [66]:

$$f(U_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big[-\frac{(U_i - \langle U \rangle)^2}{2\sigma^2}\Big] \tag{4.9}$$

By combining all of these concepts, it is observed that the whole DNA molecule can be mapped onto 1D array of sites $\{\vec{s}\}$, each corresponding to a certain binding sequence comprising bases from the $i$th to the $(i + l - 1)$th, where $l$ is the length of the motif. The random walk is thus observed becauses there is a probability $p_i$ of moving forward one site $i + 1$ and a probability $q_i$ to moving backward one site $i - 1$. These probabilities depend on the specific binding energies $U_i$, $U_{i+1}$, and $U_{i-1}$ at the $i$th site and at the adjacent sites, respectively. These are proportional to the corresponding transition rates, $w_{i,i+1}$, and $w_{i,i-1}$. The transition rates can be represented by the following equation,

$$w_{i,i\pm 1} = \begin{cases} \nu e^{-\beta(U_{i\pm 1} - U_i)} & \text{if } U_{i\pm 1} > U_i \\ \nu & \text{otherwise} \end{cases} \tag{4.10}$$

where $\nu$ is the effective attempt frequency, $\beta \equiv (k_B T)^{-1}$, $k_B$ is the Boltzmann constant and $T$ is the ambient temperature. This is a one-dimensional random walk with position dependent hopping probabilities,

$$\begin{aligned} p_i &= \frac{w_{i,i+1}}{w_{i,i+1} + w_{i,i-1}} \\ q_i &= 1 - p_i \end{aligned} \tag{4.11}$$

Recall Figure 4.2, where the random walk model is depicted in terms of compartments. Figure 4.3 is now shown to illustrate the random walk model graphically given the hopping probabilities from Equation 4.11. The behaviour of the plot shows the change of the binding potential, where each $U_i$ is represented differently for each base pair on a strand of DNA. This plot was illustrated to demonstrate how the random walk model can effectively show the fluctuations of the binding potential on a DNA strand. Parameters were chosen based on the previous publication by Eeftens et al [20] that examined the behaviour of condensin.

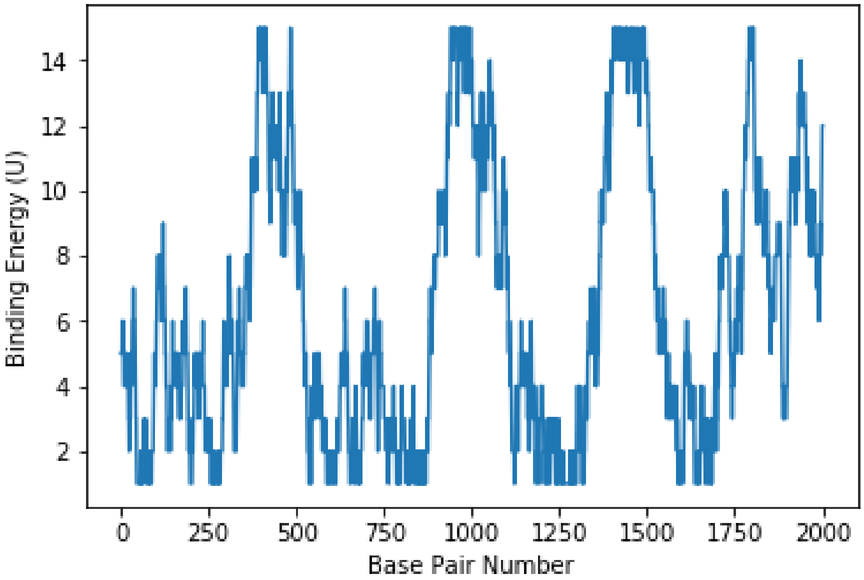

Figure 4.3: Random walk model of binding potential

As proposed by Eeftens et al [20], condensin compacts DNA in a stepwise manner. As one of the proposed LEFs includes condensin, the random walk is suitable to propose the compaction behaviour of chromatin. This figure is also similar to what was observed by Kurkcuoglu and Bates [43] for cohesin. This random walk plot parallels the behaviour of the flutuations depicting the functional sites of cohesin loading with respect to the hinge and coil structure of cohesin. The higher binding potential correspond to that of the coil structure of cohesin, where they will exhibit higher mobility than that of the head and hinge regions. However, it was noted that in this study one limitation is that existing modeling techniques are incapable of accurately constructing the long coiled-coil arms based on amino-acid sequence. Nonetheless, it is observed here that the random walk model is able to duplicate the results as predicted for the sequence binding kinetics of a proposed LEF, cohesin. Additionally, this will further emphasize the importance of binding specificity of the LEF on the strand of DNA. Therefore, the structural properties of an LEF should be considered when constructing a mathematical model for chromatin loop behaviour.

#### 4.3.3.1 Mean First Passage Time of the Random Walk

One property of the random walk is that it is dependent on the probabilities $\{p_i\}$ of either stepping left or right. The mean first-passage time (MFPT) is now presented. It is noted that it can be further used to derive the diffusion law, however we will focus on the MFPT to observe the random walk behaviour. The diffusion law may be useful to illustrate protein sliding along the DNA given the sequence-dependent binding energy.

The MFPT is calculated from site 0 to site $L$, defined as the mean number of steps the LEF needs to make in order to reach the site $L$ for the first time. This was described in more detail in a previous publication [66]. A summary of this concept, and an extension for LEF binding is presented.

Let $P_{i,j}(n)$ denote the probability to start at site $i$ and reach site $j$ in exactly $n$ steps. Then,

$$P_{i,i+1}(n) = p_i T_i(n-1) \tag{4.12}$$

where $T_i(n)$ is defined as the probability of returning to the $i$th site after $n$ steps without stepping to the right of it. All of the paths contributing to $T_i(n-1)$ should now start with the step to the left and then reach the site $i$ in $n-2$ steps, not necessarily for the first time. Thus, $T_i(n-1)$ can be written as,

$$T_i(n-1) = q_i \sum_{m,l} P_{i-1,i}(m) T_i(l) \delta_{m+l,n-2}, \tag{4.13}$$

where $\delta_{m+l,n-2}$ denotes the left jump site probability for the site $i$ at $n-2$ steps. The following generating functions, $\tilde{P}_{i,j}(z) = \sum_{n=0}^{\infty} z^n P_{i,j}(n)$ and $\tilde{T}_i(z) = \sum_{n=0}^{\infty} z^n T_i(n)$, can be used to show that

$$\tilde{P}_{0,L}(z) = \prod_{i=0}^{L-1} \tilde{P}_{i,i+1}(z) \tag{4.14}$$

This now eases the MFPT calculation with the following,

$$\bar{t}_{0,L} = \frac{\sum_n nP_{0,L}(n)}{\sum_n} P_{0,L}(n) = \left[\frac{d}{dz}\ln\tilde{P}_{0,L}(z)\right]_{z=1} \\ = \sum_{i=0}^{L-1}\left[\frac{d}{dz}\ln\tilde{P}_{i,i+1}(z)\right]_{z=1} \tag{4.15}$$

Using Equations 4.12 and 4.13, the following recursion may be obtained,

$$\tilde{P}_{i,i+1}(z) = \frac{zp_i}{1 - zq_i\tilde{P}_{i-1,i}(z)} \tag{4.16}$$

To solve for $\bar{t}_{0,L}$, we must introduce boundary conditions. Let $p_0 = 1$, $q_0 = 0$, which is equivalent to having symmetry occur at $i = 0$. This gives,

$$\bar{t}_{0,L} = \sum_{i=0}^{L-1}\tilde{P}'_{i,i+1}(1) \tag{4.17}$$

The recursion relation for $P'_{i,i+1}(1)$ may be obtained from Equation 4.16 to yield,

$$P'_{i,i+1}(1) = \frac{1}{p_i} + \frac{q_i}{p_i}\tilde{P}'_{i-1,i}(1) = 1 + \alpha_i\left[1 + \tilde{P}'_{i-1,i}(1)\right] \tag{4.18}$$

where $\alpha_i \equiv q_i/p_i$. Thus, the expression for $\bar{t}_{0,L}$ is obtained in closed form

$$\bar{t}_{0,L} = L + \sum_{k=0}^{L-1}\alpha_k + \sum_{k=0}^{L-2}\sum_{i=k+1}^{L-1}(1+\alpha_k)\prod_{j=k+1}^{i}\alpha_j \tag{4.19}$$

This gives the MFPT in terms of a given realization of disorder producing a certain set of probabilities $\{p_i\}$, whereas we are interested in the behaviour average over all realizations of disorder. The cumulative products in Equation 4.19 reduce to the two form $e^{\beta(U_i - U_j)}$, which after being averaged over uncorrelated Gaussian disorder produce a factor of $e^{\beta^2\sigma^2}$. After the summations are carried out, the expression for MFPT becomes for $L >> 1$,

$$\langle\bar{t}_{0,L}\rangle \simeq L^2 e^{\beta^2\sigma^2} \tag{4.20}$$

The aim of this section was to describe the random walk model for when an LEF will find the optimal section of DNA to extrude a loop. In the next section, MFPT will be described in terms of the exposure timescale of the DNA strand. The binding of an LEF is now explored in terms of the theoretical timescale in which the DNA remains exposed and available for LEF binding in the presence of nucleosomes.

### 4.3.4 Derivation of Exposure Time on Chromatin

Previous work by Parmar, et al.[54] investigated the theoretical estimates of exposure timescales of protein binding sites on DNA regulated by nucleosome kinetics. A theoretical method to estimate the time of continuous exposure of binding sites of non-histone proteins along any genome was investigated. It was concluded that exposure timescales are determined by cooperative dynamics of multiple nucleosomes, and their behaviour is often different from expectations based on static nucleosome occupancy. Drawing inspiration from the concepts proposed in this paper, these will be further modified and applied to loop extrusion activity.

While it is unknown which enzymes are actually LEFs, it was proposed in other work that LEFs are structural maintenance of chromosome (SMC) protein complexes, in particular cohesin during interphase. Cohesin topologically entraps DNA then slides along the strand, and over small DNA-bound proteins and nucleosomes. It is enriched at TAD boundaries and corner peaks [53]. In addition, it was shown in vitro that a closely related SMC, yeast condensin, has ATP-dependent motor activity, and growing loops were directly visualized [25]. By comparing the looping behaviour of LEFs and nucleosomes conceptually, they may mathematically be modelled in a similar manner. There are several similarities and differences between the behaviour of LEFs and nucleosomes, therefore the previous work done for nucleosomes [54] may be adapted to loop extrusion activity with some similarities.

Nucleosomes organize the eukaryotic genome into chromatin, where nucleosome assembly in cells rely on the activity of histone chaperones. Nucleosomes are modular and dynamic structures composed of an octameric core of histone proteins, wrapped by 147 bp of DNA. These histones act as spools around which DNA winds, and plays a role in gene regulation. These histones wind the DNA in chromosomes. Histone modifications are what drive the gene regulation, in that this is regulated by histone acetylation and deacetylation activity. Acetylation transforms the condensed chromatin into a more relaxed structure that is associated with greater levels of gene transcription. A conserved ATPase motor is also able to increase nucleosome sliding activity along DNA. Therefore, nucleosome activity is stochastic, and includes binding, diassembling, and sliding along DNA.

The specific location of these nucleosome kinetics are important to consider mathematically. For example, disassembly of nucleosomes is known to be important for exposure of TATA sites in promoters, while dynamics of binding nucleosomes is likely to influence the assembly near the transcription start site. Sliding of nucleosomes as well as partial wrapping/unwrapping of DNA at nucleosome edges may also contribute toward creating exposed regions along DNA. These factors affect gene expression of DNA in different ways.

In the case of loop extrusion, the LEF may behave in a similar manner to histones. LEFs are able to bind to an exposed section of DNA and compact it through loop extrusion. Histones perform this function by winding the DNA. Since LEFs bind to DNA in the presence of nucleosomes, as well, the nucleosome activity is also included in this model. The presence of boundary elements are also present on the strand of DNA. Once a boundary element has attached onto a strand of DNA at its specific site, this will halt the loop extrusion process. These factors are also present on the same strand of DNA in which nucleosomes are present. This is illustrated in Figure 4.4.

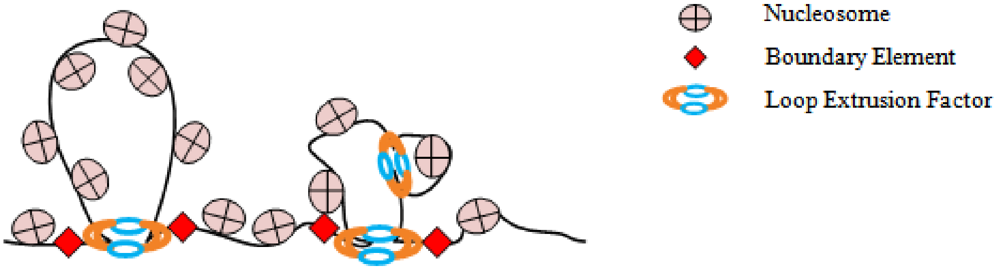

Figure 4.4: Loop extrusion in the presence of nucleosomes

The rate of binding and unbinding enzymes from chromatin can now be explored. It was previously discussed by Parmar et al [54] that the effect of spatial heterogeneity in histone-DNA interaction is captured through spatially varying off-rates. There are many different enzymes that are capable of binding to DNA, and influencing the rate of unbinding from DNA. This includes LEFs and nucleosomes activity. For the purpose of this investigation, only these two enzymes will be the main focus. The rate of unbinding is explored, as it is expected that they will unbind in a similar fashion. To clarify, they may not necessarily unbind at the same rate, but they are sequence-dependent in their rate of unbinding. Therefore, the unbinding rates ($k_{\text{off}}$) for nuclesomes and LEFs can be generalized with Equation 4.21.

$$k_{\text{off}}^{(i)} = k_{\text{on}}\text{exp}\left\{\left(\frac{V_i + U_i}{k_B T}\right)\right\} \tag{4.21}$$

where

- $V_i = -k_B T ln(P_i)$ represents the sequence-dependent effective potential
- $U_i$ represents ATPases, histone modifications, histone exchanges or DNA methylation that modifies enzyme stability

While this represents a generalization of the unbinding rates for sequence specific unbinding of nucleosomes and LEFs, it does not imply they will unbind at the same rate. The factors $V_i$ and $U_i$ influence the specific location and state of the enzyme, in that they may not be unbinding at the same location, or as easily. The value of $V_i$ is obtained from the previous work conducted by Kaplan et al. [37] representing the nucleosome-DNA binding affinity between sites $i$ and $i + k - 1$ in the presence of basal remodelling activity. The value $k_B$ represents the Boltzmann constant in $V_i$, and the value $T$ represents the absolute temperature, in K. Their work allows for one to be able to acquire the relative affinity, which represents the probability $P_i$ that an enzyme starts at the $i^{th}$ bp. The value of $U_i$ denotes the remodelling activity that may occur from locally recruited ATPases, histone modifications, histone exchange, or DNA methylation, to name a few. The magnitude of $U_i$ being positive or negative would result in lesser or greater local stability of the enzyme attached to the DNA.

### 4.3.5 Mean First-Passage Time of DNA Exposure

The mean first-passage time (MFPT) was explored in Section 4.2 to describe the specificity of LEF binding on a strand of DNA. The MFPT may also be calculated to represent the amount of time a specific section of DNA may stay exposed and available for LEF binding. This is derived here to show the impact of external factors on LEF activity.

The MFPT was previously derived by Parmar et al [54] for nucleosome activity in the presence of transcription factors. In this case, this will be modified for the purpose of incorporating the presence of LEF binding, which may be represented mathematically in a similar manner. Rather than the binding of transcription factors, this will explore the binding of LEFs. From the open state, two events are possible: binding of an LEF or binding of a nucleosome. Upon LEF binding, the first passage happens, while if a nucleosome binds, the first passage is delayed by additional mean time $T_f^{close}$. Therefore, we obtain

$$T_f^{open} = \frac{1}{k_{\text{on}}^{(n)} + k_{\text{on}}^{(c)}} + \frac{k_{\text{on}}^{(n)}}{k_{\text{on}}^{(n)} + k_{\text{on}}^{(c)}} T_f^{close}, \tag{4.22}$$

where $k_{on}^{(n)}$ and $k_{on}^{(c)}$ represent the binding rate of nucleosomes and LEFs, respectively. Similarly, from the closed state, the only possible event is nucleosome unbinding, and there is a subsequent delay of mean time $T_f^{open}$. This gives,

$$T_f^{close} = \frac{1}{k_{\text{off}}^{(n)}} + T_f^{open}, \tag{4.23}$$

and $k_{\text{off}}^{(n)}$ represents the unbinding rate of nucleosomes. Rearranging the above equation to isolate for $T_f^{open}$ gives,

$$\begin{aligned} T_f^{open} &= \frac{1}{k_{\text{on}}^{(c)}}\left(1+\frac{k_{\text{on}}^{(n)}}{k_{\text{off}}^{(n)}}\right) \\ T_f^{close} &= \frac{1}{k_{\text{off}}^{(n)}}+\frac{1}{k_{\text{on}}^{(c)}}\left(1+\frac{k_{\text{on}}^{(c)}}{k_{\text{off}}^{(c)}}\right) \end{aligned} \tag{4.24}$$

In case at $t = 0$, the site may be either open or closed by the nucleosome, the MFPT $T_f$ would be a weighted average of $T_f^{open}$ and $T_f^{close}$, such that

$$T_f = \frac{k_{\text{off}}^{(n)}T_f^{open} + k_{\text{on}}^{(n)}T_f^{close}}{k_{\text{on}}^{(n)} + k_{\text{off}}^{(n)}} \tag{4.25}$$

Although it was observed from Equation 4.24 that $T_f^{open}$ depends only on the ratio $k_{\text{on}}^{(n)}/k_{\text{off}}^{(n)}$, which can be predicted through the nucleosome occupancy, the same is not the case for $T_f^{close}$ or $T_f$. Moreover, when Parmar et al [54] analyzed the full probability distributions of the open and closed states, they determined that those depend on individual values of $k_{\text{on}}^{(n)}$ and $k_{\text{off}}^{(n)}$, and not just the ratio $k_{\text{on}}^{(n)}/k_{\text{off}}^{(n)}$. Thus, the protein binding histories will be distinct for different nucleosome kinetics.

By deriving the MFPT for the exposure time of DNA, these concepts may be introduced into the theoretical exposure timescales of LEF binding in the presence of nucleosomes in different scenarios for binding. In the next section, the specific area of LEF binding is considered when determining the factors contributing to LEF binding.

#### 4.3.5.1 Distribution length of DNA strand

If the length of DNA, $l$, is not held fixed but drawn from a distribution $P_{\text{in}}(l)$, the mean exposure time $T_{\text{av}}$ of the $m$-patch would be a weighted average over the mean times $T_l$.

$$T_{\text{av}} = \sum_{l \geq l_{\min}} P_{in}(l)T_l \tag{4.26}$$

If no experimental bias is introduced, $P_{in}(l)$ is expected to be the steady state gap distribution, which is an exponential function. If the distribution of the initial nucleosome location ($l$) from the barrier is chosen in steady state, then the initial gap distribution is equivalent to the gap distribution in steady state, which is denoted by $P_{ss} = (1-C)C^l$ for $l \geq 0$ and $P_{ss} = 0$ for $l$<0. The constant $C$ is determined by the ratio between the dissociation rate to binding rate, as given by

$$\frac{k_{\text{off}}}{k_{\text{on}}} = \frac{C^k}{1-C} \tag{4.27}$$

which was described by Parmar et al [54]. This distribution is introduced in the theoretical derivations to demonstrate the impact of how the length of the DNA is not held fixed.

The length of the DNA strand is not the only factor that may not be held fixed. Depending on environmental factors, the kinetic rates of the enzymes binding and unbinding to the strand of DNA can also be derived from a distribution. These equations are further derived in the following section.

### 4.3.6 LEF Exposure Cases on DNA

To begin investigating the binding kinetics, one must consider whether the enzyme is binding in a specific, or non-specific manner. As observed from the previous chapter, an LEF may bind to the strand of DNA, but may not extrude a loop. However, since the exposure time of the specific patch of DNA is investigated here, these kinetics must be considered. In the case of non-specific binding, then the on-rate ($k_{\text{on}}$) may simply be given by Equation 4.28, where $k_0$ is merely some constant. This will apply for both LEF and nucleosome non-specific binding kinetics.

$$k_{\text{on}} = k_0 \tag{4.28}$$

However, we suspect that LEFs may bind to DNA on specific nucleotide sequences. This case will not apply to nucleosomes, as nucleosomes may bind to DNA non-specifically then slide to the promoter site whereupon it will commence winding the strand of DNA. An example of a possible LEF that binds to DNA specifically is cohesin, in which its loading complex has been found at A-T rich DNA both in vitro and in vivo [10]. Therefore, it is assumed that specific binding may also occur for some LEFs. The rate of specific binding is then given by Equation 4.29.

$$k_{\text{on}} = k_0 e^{\left(\frac{V_i+U_i}{k_B T}\right)} \tag{4.29}$$

To begin exploring the effect of these binding kinetics on the exposure timescales of the target DNA, several cases are considered. These cases are modified from the previous work performed by Parmar, et al. [54]. However, there are other factors and considerations included in our model. Since nucleosomes are able to bind, dissociate, and slide from DNA that may affect the exposure time, this information must be included as factors that may contribute to LEF binding. It is assumed that the rate of nucleosome sliding is constant and non-specific. The four cases of binding activity that will be explored is illustrated in Figure 4.5. Note that the target patch will be on the right side of N1. Should the equations be visualized as the patch on the left side of N2, the distances to the left will be negative. For consistency and simplicity, the distances to the right of N1 will be shown.

In Figure 4.5, there are several cases that depict factors that may influence and affect LEF

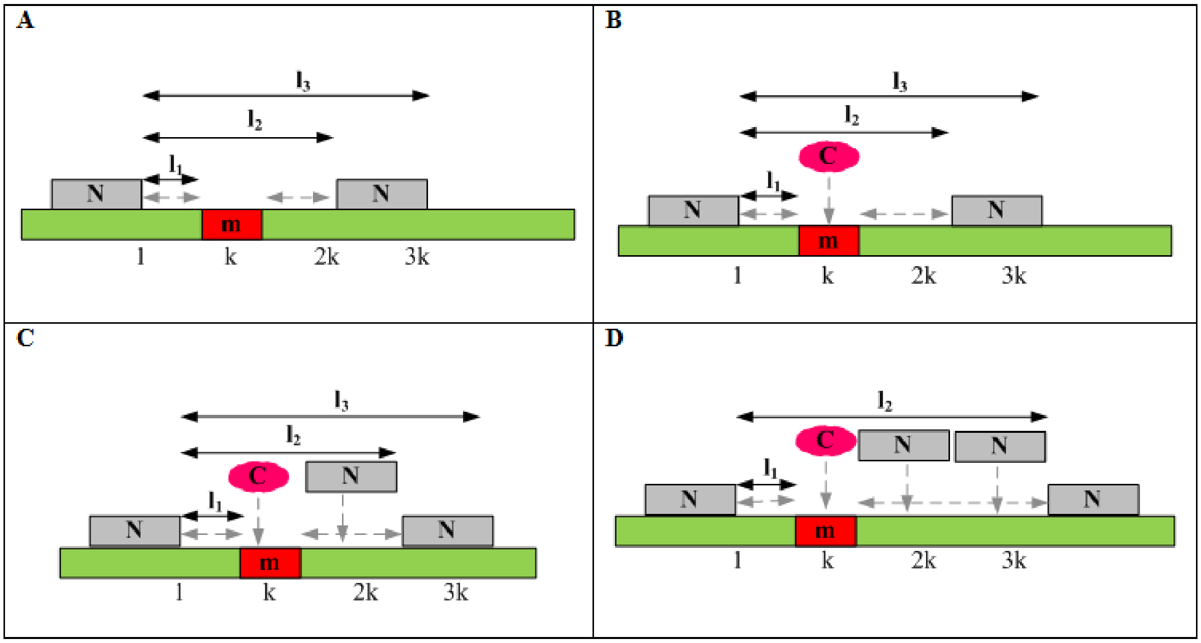

Figure 4.5: Theoretical cases of LEF binding between nucleosomes. **A.** Exposed strand of DNA, with no LEF binding. The target patch , $m$, is bounded by two nucleosomes, $N_1$ and $N_2$. **B.** Binding of one LEF ($C$) between two nucleosomes. **C.** Binding of one LEF and one nucleosome between two nucleosomes. **D.** Binding of one LEF and two nucleosomes between two nucleosomes.

binding to the DNA that could impede loop extrusion activity. The green rectangle illustrates the strand of DNA in which enzymes will bind. The gray rectangle with the "$N$" at the center depicts a nucleosome. As mentioned, this will be around 147 bp long on the strand of DNA. The length of DNA is based on the size of nucleosomes that can bind to the DNA. Therefore, $k$ is based on the width of the LEF that may bind to the DNA. The length of the strand strand of DNA ($l$) studied is only $3k$. The reason is that large gaps are rare [54], so this behaviour is expected to be repeated on the strand of DNA for any case. The LEF used for these cases will be cohesin, as an example. This is denoted by the oval shape with the "$C$" above the target ("m") patch of DNA. The relative size of cohesin is 0.8 kb, which is significantly larger than a nucleosome [29]. While the figure does not adjust the size of cohesin and nucleosomes relative to size, it is still implied that binding will be impeded should a nucleosome not allow for enough exposed DNA for anything to bind.

In order to further examine the exposure time of the target patch of DNA, the nucleosome and LEF activity is simulated again with the Gillespie algorithm. The Nucleosome Tool plug-in [2] from Stochpy will be used to create plots of this activity. This plug-in also models four different states of histones: acetylated (A), unmodified (U), methylated (M) or occupied by a methyltransferase (Mt). Other rates used in this tool that affects nucleosome activity include: $k_{\text{on}}$, $k_{\text{off}}$, $k_{\text{transferase}}$, $k_{\text{slide}}$, $k_{\text{recruitment}}$, $k_{\text{demodification}}$, and $k_{\text{interaction}}$. Only some of the rates are used in these simulations to create a more realistic situation of nucleosome activity that affect the binding of LEFs to DNA. There are several kinetic rates included in the tool's capabilities described by Anink-Groenen et al [2], including the following:

- $k_{\text{on}}$: Influx of transferase at initiation site
- $k_{\text{off}}$: Release rate of transferase from nucleosome
- $k_{\text{transferase}}$: Modification rate of nucleosome
- $k_{\text{neighbour}}$: Modification rate of neighbouring nucleosome
- $k_{\text{slide}}$: 1D difusion rate over the chromatin
- $k_{\text{recruitment}}$: Influx of transferases at modified nucleosome
- $k_{\text{demodification}}$: Rate constant of demodification
- $k_{\text{interaction}}$: Interaction frequency

In the simulations run to examine these individual cases, the rate of specific binding of LEFs was also added to observe the exposure times. However, only the $k_{\text{on}}$, $k_{\text{off}}$, and $k_{\text{slide}}$ rates were utilized, as the other factors were not required because they may not affect LEF activity in any manner. These were not included in the theoretical equations, as the focus of this investigation was to determine how the LEFs would behave in the presence of other nucleosomes in terms of the exposure time on DNA, and whether it would be able to bind. These four cases are now examined in more detail by deriving the theoretical exposure rate, then testing the model by running simulations. The theoretical exposure time of each case is first derived, however to simplify the theoretical cases, nucleosome behaviour will be restricted to three options: binding, unbinding, and sliding.

#### 4.3.6.1 Case A: No LEF binding

The theoretical case in which there is no LEF binding is first derived. A detailed illustration of this case is presented in Figure 4.6. The gray arrows represent the possibility of nucleosomes sliding within the area of the target DNA. If the nucleosomes slide too closely to the DNA, there will not be enough space for the LEF to bind to the target, and thus will not bind successfully. Since an LEF binds between two nucleosomes, the distance between these nucleosomes must be examined more closely. The black arrows in the Figure depict relevant lengths between different attributes that are found on the strand of DNA. Since

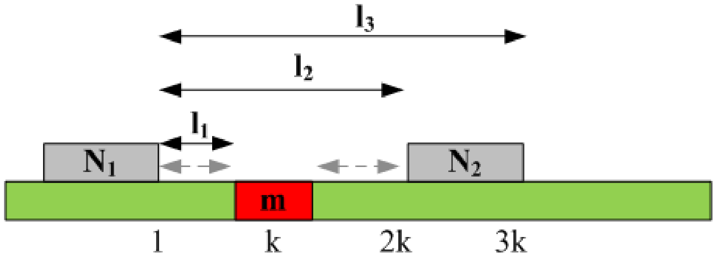

Figure 4.6: No LEF or nucleosome binding

there is no LEF binding in this case, only the behaviour of the nucleosomes will be examined to determine its movement along the strand of DNA.

In this Figure, let $N_1$ be the left nucleosome that surrounds the target patch of DNA. This will be the initial starting point of reference for counting the distance of the attributes found on the strand of DNA. Let $l_1$ depict the length between $N_1$ and the target patch $m$. The target patch of DNA is the set of sequences that the LEF will bind specifically. Let $l_2$ depict the length of DNA between $N_1$ and $N_2$. This represents the length of exposed DNA that encompasses the LEF binding activity. Note the target patch of DNA maintains exposed within this length, which will be measured. Since the binding activity may be specific, the length $l_3$ depicts the length which measures from the right side of $N_1$ to the right side of $N_2$. Note this length of DNA is restricted to $3k$, which represents the size of three LEFs.

To analyze the enzyme behaviour of this figure, all aspects of nucleosome activity is considered. Since no new enzymes will be entering the system, the only other possible activity occurring will be the rate of unbinding of the current nucleosomes. Ultimately, that is expected to increase the exposure time of the DNA. The exposure time of the DNA is given by $T_{l_j}$, where $j$ represents the specified segment of on the DNA strand in the figure. The theoretical equation expected to represent this activity is given in Equation 4.30, which is simply a rearrangement of Equation 4.23.

$$\begin{aligned} T_f^{close} &= \frac{1}{k_{\text{off}}} + T_f^{open} \\ T_f^{close} - T_f^{open} &= \frac{1}{k_{\text{off}}} \end{aligned} \tag{4.30}$$

Further from Equation 4.26, if the length of DNA is not held fixed, but drawn from a distribution, the exposure time would be represented by a weighted average over average exposure times. This exposure time is then represented by the open state of exposure, and is substitued into $T_f^{open}$. The open state also depicts the probability weight to make a transition to a new gapped state ($l_3$) with new exposure time $T_{l_3}$. This gives the following,

$$T_f^{close} - \sum_{l_3=l_2+k}^{3k-1} P(l_3)T_{l_3} = \frac{1}{k_{\text{off}}} \tag{4.31}$$

It is also known that nucleosomes are able to slide along the length of DNA, which should then be considered into the model. This will be depicted at a constant rate, $k_{\text{s}}$. It was previously stated by Lequieu et al [44] that the mechanism of nucleosome repositioning is shown to be strongly linked to DNA sequence and directly related to the binding energy of a given DNA sequence to the histone core. This will be interpreted such that the rate of sliding is directly proportional to the binding of the DNA strand in the open state. This is only added to the existing distribution without the effect of sliding, such that if there was no sliding activity, the exposure time is still affected in its open state. As well, the effect of sliding may be anywhere between $[1, 3k]$, where the probability distribution of the exposed DNA along $l_3$ is affected, thus the sum of the effect of this region is included. It is assumed that the desired DNA binding sequence lies in this area of the DNA strand. Therefore, this gives Equation 4.32,

$$T_f^{close} - (k_s + 1) \sum_{l_3=l_2+k}^{3k-1} P(l_3)T_{l_3} = \frac{1}{k_{\text{off}}} \tag{4.32}$$

The effect of binding specificity is also affected by the off rate of LEFs. The unbinding rate, $k_{\text{off}}^{(i)}$, denotes the off-rate where $i$ is the left most base pair of the enzyme positioned between $[i, i+k-1]$. This rate was previously defined in Equation 4.21, where the sequence-dependent effective binding potential influences the rate.

$$T_{l_2} - (1 + k_s) \sum_{l_3=l_{2+k}}^{3k-1} P(l_3)T_{l_3} = \frac{1}{k_{\text{off}}^{(l_2+1)}} \tag{4.33}$$

The mean persistence time of the initial gap state is given by this unbinding rate for either nucleosome, as denoted on the right side of the equation. This equation describes the exposure time of the target DNA. This is given by the exposed time of the length of DNA, $l_2$, since this is the distance between the two nucleosomes that may impede binding. This is affected by the possible sliding activity of nucleosomes as well, so they need to be taken into consideration. On the right hand side of Equation 4.33, this denotes the rate of unbinding of any of the nucleosomes. Therefore, the rate of exposure is dependent on the rate of unbinding. Since either nucleosome can unbind from the strand of DNA and increase the exposed timescale of DNA, the off- rate is dependent on the specific nucleosome activity. It is assumed that the off-rate for each nucleosome will have the same potential, otherwise two off-rates would need to be specified in the equation for each nucleosome.

#### 4.3.6.2 Case B: Binding of one LEF

Further building on Case A, the binding of an LEF is now considered. In the previous case, there was no LEF binding activity. Here, the binding of an LEF, namely cohesin, will be investigated. This case is illustrated in Figure 4.7.

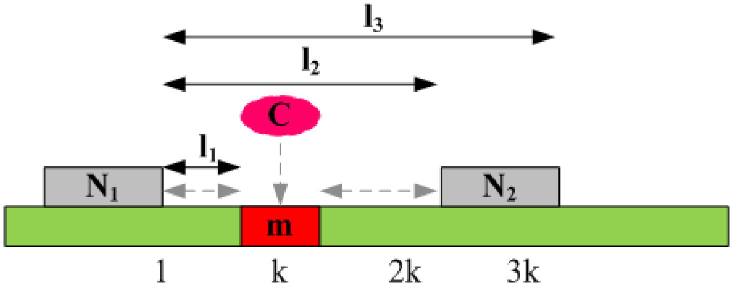

Figure 4.7: Binding of one LEF

By comparing Case A and Case B, the only new aspect introduced in this case is the binding of the LEF. Starting from the closed state, the enzymatic activities only include the binding of the LEF and the unbinding of the nucleosome to create an exposed strand of DNA. Therefore, to expand on Equation 4.30, the effect of LEF binding must be included, which will be given by $k_{\text{on}}^{(c)}$. Note that this rate is represented by Euqation 4.29, which includes factors affecting the specificity of binding for the LEF. This is significant, as cohesin binds at A-T rich DNA sites[10]. The probability of unbinding activity of the nucleosome and the binding activity of the LEF is then given by $\frac{1}{k_{\text{on}}^{(c)}+k_{\text{off}}}$. This is given by Equation 4.34,

$$\begin{aligned}
T_f^{close} &= \frac{1}{k_{\text{on}}^{(c)}+k_{\text{off}}} + \frac{k_{\text{off}}}{k_{\text{on}}^{(c)}+k_{\text{off}}} T_f^{open} \\
(k_{\text{on}}^{(c)}+k_{\text{off}}) T_f^{close} &= k_{\text{off}} T_f^{open} + 1 \\
(k_{\text{on}}^{(c)}+k_{\text{off}}) T_f^{close} - k_{\text{off}} T_f^{open} &= 1 \\
T_f^{close} - \frac{k_{\text{off}}}{k_{\text{on}}^{(c)}+k_{\text{off}}} T_f^{open} &= \frac{1}{k_{\text{on}}^{(c)}+k_{\text{off}}}
\end{aligned} \tag{4.34}$$

The probability weights of the transitions are now modified to include the binding rate of the LEF as $(\frac{k_{\text{off}}}{k_{\text{on}}^{(c)}+k_{\text{off}}})P(l_3)T_{l_3}$. Note that in a similar manner to Case A, the open state of the DNA is now modified to include the binding distribution, as described by Equation 4.26.

$$T_f^{close} - \frac{k_{\text{off}}}{k_{\text{on}}^{(c)}+k_{\text{off}}} \left( \sum_{l_3=l_{2+k}}^{3k-1} P(l_3) T_{l_3} \right) = \frac{1}{k_{\text{on}}^{(c)}+k_{\text{off}}} \tag{4.35}$$

Since we are interested in the first time the LEF binds to DNA, the sliding rate of this enzyme is not required. However, the sliding rates of the nucleosomes already bound to the DNA is still necessary, and its inclusion follows the same rationale as that of Case A. The equation depicting this case is given by Equation 4.36.

$$T_f^{close} - \frac{k_{\text{off}}}{k_{\text{on}}^{(c)} + k_{\text{off}}}(1+k_s)(\sum_{l_3=l_{2+k}}^{3k-1} P(l_3)T_{l_3}) = \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}} \tag{4.36}$$

Finally, the binding specificity of the LEF is now included, as it affects the binding rate of the enzyme. This is now included in Equation 4.37.

$$T_{l_2} - (1+k_s)\left(\frac{k_{\text{off}}^{(l_2+1)}}{k_{\text{off}}^{(l_2+1)} + k_{\text{on}}^{(c)}(l_2-k+1)}\right)\sum_{l_3=l_{2+k}}^{3k-1} P(l_3)T_{l_3} = \frac{1}{k_{\text{off}}^{(l_2+1)} + k_{\text{on}}^{(c)}(l_2-k+1)} \tag{4.37}$$

As previously discussed, in order for binding to occur, there needs to be enough space between the two nucleosomes to allow for LEF binding. The mean persistence time is now given by $\frac{1}{k_{\text{on}}^{(c)}+k_{\text{off}}}$, which corresponds to the dissociation events of the nucleosomes, and the binding event of the LEF. The binding events immediately cover the target patch, while the dissociation events lead to exposure in new gapped states, as previously discussed for Case A. It was also specified in this equation the distinction between the binding activity of nucleosomes and the LEFs. It is assumed that all rates correspond to that of the nucleosome activity, except for $k_{\text{on}}^{(c)}$, which denotes the binding rate of the LEF.

#### 4.3.6.3 Case C: Binding of one LEF and one nucleosome

The case has now been further developed to include the binding of a nucleosome, in addition to the binding of an LEF from the initial case. Since nucleosomes do not necessarily bind specifically to DNA, competition is not considered for the binding in the target patch. The new nucleosome will bind to a region at some distance away from the target patch. This situation is illustrated in Figure 4.8 below.

Similar to Case B, the sliding rate of the LEF that will bind to the DNA is not considered. The sliding rate of the nucleosome that will bind to the DNA is also not considered because it has not yet had the chance to bind to DNA to affect the exposure time. Since we are interested in the first passage time, this would have already delayed the exposure time upon initial binding. The Equation depicting this case is given in Equation 4.38.

$$T_f^{close} = \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}} + \frac{k_{\text{off}}}{k_{\text{on}}^{(c)} + k_{\text{off}}}T_f^{open} + \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} + k_{\text{off}}}T_{fn}^{close} \tag{4.38}$$

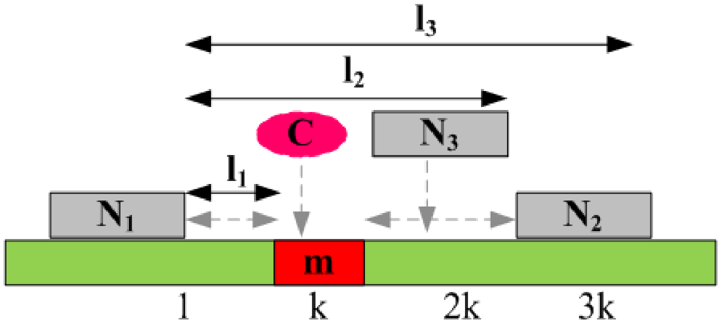

Figure 4.8: Binding of one LEF and one nucleosome

In addition to the same terms as denoted in Equation 4.37, the binding of the additional nucleosome is now considered. This binding may lead to a new gap of length $l_4 \in [l_1{+}m, l_2{-}k]$ without covering the $m$ patch. The weight of transitions in this case is $\frac{k_{on}}{k_{\text{on}}^{(c)}+k_{\text{off}}}$. The average delay after direct binding events missing the $m$-patch is now considered. The delay is denoted by the probability that each nucleosome unbinds from the DNA strand, which is represented by the following,

$$\frac{1 + k_{\text{off}}^{(l_4+1)} T_{l_2} + k_{\text{off}}^{(l_2+1)} T_{l_4}}{k_{\text{off}}^{(l_2+1)} + k_{\text{off}}^{(l_3+1)}}. \tag{4.39}$$

This represents the value of $T_{fn}^{close}$ because only the binding of a nucleosome in the exposed section can delay LEF binding.

After the binding of a nucleosome between N1 and N2, a gap of length $l_4$ is created. Now, either N3 dissociates, and contributes to a further delay of average time $T_{l_2}$, or if nucleosome N2 dissociates, the delay is $T_{l_1}$. The average waiting time for neither of these two events to happen is $1/(k_{\text{off}}^{(l_2+1)} + k_{\text{off}}^{(l_4+1)})$, which completes the rationale for $T_{fn}^{close}$. The other closed state is that of the state of the portion of exposed DNA between N1 and N2, as represented by $T_f^{close}$. This state of DNA can simply be represented by $T_{l_2}$ since we have started this scenario in the closed state with the binding events of either a nucleosome or an LEF to the exposed section of DNA, $l_2$. Therefore, the values for $T_f^{close}$, and $T_{fn}^{close}$ from Equation 4.39, can be substituted into Equation 4.38 to obtain,

$$\begin{aligned}
T_f^{close} &= \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}} + \frac{k_{\text{off}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} T_f^{open} + \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} \sum_{l_4=m+l_1}^{l_2-k} \left[ \frac{1 + k_{\text{off}}^{(l_4+1)} T_{l_2} + k_{\text{off}}^{(l_2+1)} T_{l_4}}{k_{\text{off}}^{(l_2+1)} + k_{\text{off}}^{(l_4+1)}} \right] \\
T_f^{close} - \frac{k_{\text{off}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} T_f^{open} &+ \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} \sum_{l_4=m+l_1}^{l_2-k} \left[ \frac{1 + k_{\text{off}}^{(l_4+1)} T_{l_2} + k_{\text{off}}^{(l_2+1)} T_{l_4}}{k_{\text{off}}^{(l_2+1)} + k_{\text{off}}^{(l_4+1)}} \right] = \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}} \\
T_{l_2} - \frac{k_{\text{off}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} T_f^{open} &+ \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} \sum_{l_4=m+l_1}^{l_2-k} \left[ \frac{1 + k_{\text{off}}^{(l_4+1)} T_{l_2} + k_{\text{off}}^{(l_2+1)} T_{l_4}}{k_{\text{off}}^{(l_2+1)} + k_{\text{off}}^{(l_4+1)}} \right] = \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}}
\end{aligned} \tag{4.40}$$

Note that Equation 4.39 was also manipulated to show the factors that contribute to the probability of unbinding activity.

The final component is the exposure time in the open state, where the area affected by the binding distribution of the DNA strand, as previously explained with Equation 4.26, is now substituted into $T_f^{open}$ to obtain,

$$T_{l_2} - \frac{k_{\text{off}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} \sum_{l_4=l_2+k}^{3k-1} P(l_3) T_{l_4} + \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} \sum_{l_4=m+l_1}^{l_2-k} \left[ \frac{1 + k_{\text{off}}^{(l_4+1)} T_{l_2} + k_{\text{off}}^{(l_2+1)} T_{l_4}}{k_{\text{off}}^{(l_2+1)} + k_{\text{off}}^{(l_4+1)}} \right] = \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}} \tag{4.41}$$

Finally, the effect of sliding of the existing nucleosomes that were already on the strand of DNA is incorporated in the equation in a similar manner to what was explained for Case A.

$$\begin{aligned}
&T_{l_2} - (1 + k_s) \frac{k_{\text{off}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} \sum_{l_4=l_2+k}^{3k-1} P(l_4) T_{l_4} \\
&+ \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} \sum_{l_4=m+l_1}^{l_2-k} \left[ \frac{1 + k_{\text{off}}^{(l_4+1)} T_{l_4} + k_{\text{off}}^{(l_2+1)} T_{l_4}}{k_{\text{off}}^{(l_2+1)} + k_{\text{off}}^{(l_4+1)}} \right] = \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}}
\end{aligned} \tag{4.42}$$

There are other ways in which this delay could have been approximated, but due to the presence of N3 at $l_2$, the gap distribution is so strongly conditioned away from steady state that the actual delays are far from $T_{l_1}$. The value of $\frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}}$ remains identical to that of Case B, as the same dissociation events of the nucleosomes and the binding event of the LEF are the same conditions that contribute to the mean persistence time.

#### 4.3.6.4 Case D: Binding of one LEF and two nucleosomes

Finally, one last case is considered. A larger gap for $l_2$ that has now been created, where more nucleosomes are able to bind to the strand of DNA. This is the case in which N2 is located outside of $l_2$, so the distance between N1 and N2 is much wider than the previous cases. Competition between nucleosomes and the LEF is again not considered for the binding in the target patch. This case is illustrated in Figure 4.9 below.

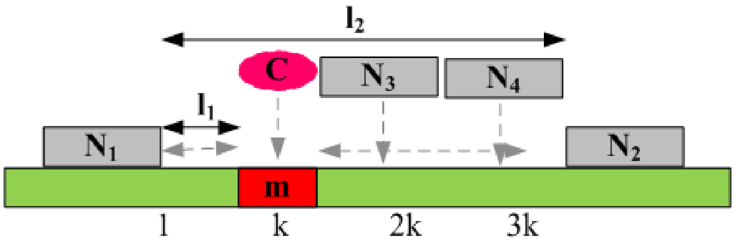

Figure 4.9: Binding of one LEF and two nucleosomes

The binding contribution of the second nucleosome within $l_2$ must now be incorporated, as Equation 4.42 is further developed. This yields Equation 4.43. Dissociation events are ignored altogether as they produce gaps of size $l_4 \geq 3k$, which is not considered. This was established prior to investigating these cases, as the nucleosome activity is studied in these smaller sections.

$$\begin{gathered} T_f^{close} = \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}} + T_f^{open} + \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} T_{fn1}^{close} + \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} + k_{\text{off}}} T_{fn2}^{close} \\ T_f^{close} - T_f^{open} - \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} - k_{\text{off}}} T_{fn1}^{close} - \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} - k_{\text{off}}} T_{fn2}^{close} = \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}} \end{gathered} \tag{4.43}$$

As previously mentioned, the effect of sliding is now incorporated into the model. Also, in the open state, the length of DNA may not be held fixed, but drawn from a distribution as previously explained in Case A. Therefore, Equation 4.26 is substituted into $T_f^{open}$, which yields

$$T_f^{close} - (1 + k_s) \sum_{l_2}^{3k-1} P(l_2) T_{l2} - \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} - k_{\text{off}}} T_{fn1}^{close} - \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} - k_{\text{off}}} T_{fn2}^{close} = \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}} \tag{4.44}$$

The values for $T_{fn1}^{close}$ and $T_{fn1}^{close}$ are now derived. These values are a little more complicated due to the limited amount of space available on the portion of DNA available for binding. While this scenario should be similar to the previous case, there is the additional nucleosome that can be available for binding in this space of exposed DNA. This affects the closed state of the exposure time for the DNA strand. In the closed states, the average delay after direct binding events missing the $m$-patch has now expanded to include the influence of both nucleosomes. These nucleosomes can bind in the gap between $m + l_1$ and $l$, which are both of size $k$. This can then be denoted by the following,

$$T_f^{close} - (1 + k_s) \sum_{l_2}^{3k-1} P(l_2) T_{l2} - \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} - k_{\text{off}}} \sum_{l_4=k}^{3k-1} T_{l_4} - \frac{k_{\text{on}}}{k_{\text{on}}^{(c)} - k_{\text{off}}} \sum_{l_4=m+l_1}^{3k-1} T_{l_4} = \frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}} \tag{4.45}$$

Recall Equation 4.39 for Case C, which represented the average delay after direct binding events missing the $m$-patch. However now that there is an additional nucleosome, the effects of the second nucleosome must be incorporated. Binding events happen with weight factor $\frac{k_{\text{on}}}{k_{\text{on}}^{(c)} + k_{\text{off}}}$, similar to the previous cases. If the patch gets directly covered, the contribution to $T_{l_2}$ is only $\frac{1}{k_{\text{on}}^{(c)} + k_{\text{off}}}$. But quite often the newly bound nucleosome misses the target patch, which results in larger delays in coverage like the cases described above. In this case, there are two nucleosomes binding in the gap between $m + l_1$ and $l$. The possibility of a second binding arises in three different ways. First, if the new gap $l_1 \geq k$, the remaining gap $l_2 - (l_1 + k)$ is not big enough to accomodate a second nucleosome to its right, as $l_2$ has a maximum value of $3k - 1$ within the approximate. But a second nucleosome can bind in the space $l_1$ with an average timescale of $T_{l_1}$. Second, if the new gap $l_1 < k$, a second nucleosome can only bind the gap $l_2 - (l_1 + k) \geq k$. Assuming that is the case, the first binding is followed by either its immediate dissociation, dissociation of N2, or binding of a second nucleosome in $l_2 - l_1, -2k + 1$ positions, where $\delta$ is the space between N3 and N4. Lastly, for $l_1 < k$ and $l_2 - (l_1 + k) < k$, there is not enough space for second binding.

To mathematically show the binding effects of nucleosomes N3 and N4, let $\delta$ denote the distance between N3 and N4. Equation 4.39 for Case C can now be modified according to the binding distance between N3 and N4 for this case where $T_{fn2}^{close}$ can be,

$$\frac{1 + k_{\text{off}}^{(l_4+1)} T_{l_4+k+\delta} + k_{\text{off}}^{(l_4+k+\delta+1)} T_{l_4}}{k_{\text{off}}^{(l_4+1)} + k_{\text{off}}^{(l_4+k+\delta+1)}} \tag{4.46}$$

However, as mentioned in the previous paragraph, binding events have weight factor, $\frac{k_{\text{on}}}{k_{\text{on}}^{(c)} + k_{\text{off}}}$, so Equation 4.46 becomes,

$$\frac{\widetilde{k}_{\text{on}}}{k_{\text{off}}^{(l+1)} + k_{\text{off}}^{(\tilde{l}+1)} + \widetilde{k_{\text{on}}}(l - l_4 - 2k + 1)} \sum_{\delta=0}^{l_2-l_4-2k} \left\{ \frac{1 + k_{\text{off}}^{(l_4+1)} T_{l_4+k+\delta} + k_{\text{off}}^{(l_4+k+\delta+1)} T_{l_4}}{k_{\text{off}}^{(l_4+1)} + k_{\text{off}}^{(l_4+k+\delta+1)}} \right\} \tag{4.47}$$

Equation 4.46 is now added and modified to accomodate the binding of the second nucleosome.

$$\begin{aligned}&\frac{1+k_{\text{off}}^{(l_4+1)}T_{l_2}+k_{\text{off}}^{(l_2+1)}T_{l_4}}{k_{\text{off}}^{(l+1)}+k_{\text{off}}^{(\tilde{l}+1)}+\widetilde{k_{\text{on}}}(l-l_4-2k+1)}\\&+\frac{\widetilde{k}_{\text{on}}}{k_{\text{off}}^{(l+1)}+k_{\text{off}}^{(\tilde{l}+1)}+\widetilde{k_{\text{on}}}(l-l_4-2k+1)}\sum_{\delta=0}^{l_2-l_4-2k}\left\{\frac{1+k_{\text{off}}^{(l_4+1)}T_{l_4+k+\delta}+k_{\text{off}}^{(l_4+k+\delta+1)}T_{l_4}}{k_{\text{off}}^{(l_4+1)}+k_{\text{off}}^{(l_4+k+\delta+1)}}\right\}\end{aligned}\tag{4.48}$$

By including the weight factor again and apply it to the binding events of both N3 and N4, then $T_{fn2}^{close}$ is fully defined and can be substituted into the following equation to model Case D,

$$\begin{aligned}&T_{l_2}-(1+k_s)\sum_{l_2}^{3k-1}P(l_2)T_{l_2}-\frac{k_{\text{on}}}{k_{\text{on}}^{(c)}+k_{\text{off}}}\sum_{l_4=k}^{l_2=k}T_{l_4}\\&-\frac{k_{\text{on}}}{k_{\text{on}}^{(c)}+k_{\text{off}}}\sum_{l_4=m+l_1}\Big[\frac{1+k_{\text{off}}^{(l_4+1)}T_{l_2}+k_{\text{off}}^{(l_2+1)}T_{l_4}}{k_{\text{off}}^{(l+1)}+k_{\text{off}}^{(\tilde{l}+1)}+\widetilde{k_{\text{on}}}(l-l_4-2k+1)}\\&+\frac{\widetilde{k}_{\text{on}}}{k_{\text{off}}^{(l+1)}+k_{\text{off}}^{(\tilde{l}+1)}+\widetilde{k_{\text{on}}}(l-l_4-2k+1)}\sum_{\delta=0}^{l_2-l_4-2k}\left\{\frac{1+k_{\text{off}}^{(l_4+1)}T_{l_4+k+\delta}+k_{\text{off}}^{(l_4+k+\delta+1)}T_{l_4}}{k_{\text{off}}^{(l_4+1)}+k_{\text{off}}^{(l_4+k+\delta+1)}}\right\}\Big]\\&=\frac{1}{k_{\text{on}}^{(c)}+k_{\text{off}}}\end{aligned}\tag{4.49}$$

and

$$\widetilde{k_{\text{on}}}=\begin{cases}0 & \text{if } l_4{<}0\\ k_{\text{on}} & \text{if } l_4\geq 0\end{cases}$$

which concludes the derivation of Case D.

#### 4.3.6.5 Remarks

While the activity of only nucleosomes and LEFs were examined in the above-mentioned cases, there are several other factors that may also affect the exposure timescale of DNA including the binding of barrier elements, ATPase modifications, transcription factors, etc. In addition, there has yet to be specific enzymes that have been confirmed to be LEFs. This is not to say that cohesin is representative of all LEF behaviour, nor that its size is consistent for all LEFs. There is much research that has yet to be conducted to fully confirm the precise mechanism for loop extrusion behaviour. For now, these theoretical cases for the exposure times should only be interpreted as one method of modelling the binding behaviour of enzymes, and how it will affect the time a specific patch of DNA remains exposed and available for binding.

## 4.4 Results

### 4.4.1 DNA Exposure Time Theoretical Equations

In Section 4.3, the exposure time of a strand of DNA in four different cases were derived. These cases include: A) No LEF binding, B) Binding of one LEF, C) Binding of one LEF and one nucleosome, and finally, D) Binding of one LEF and two nucleosomes. The final theoretical exposure equations were finalized in equations 4.33, 4.37, 4.42, and 4.49, respectively. The theoretical exposure timescales are now plotted to gain a better understanding of the behaviour of binding activity (Figure 4.10). The kinetic rates used to run these equations were the same as those used in Chapter 3 for the burst model, as derived from Brackley et al [13]. Specifically, for this run, the lower steady state kinetic rates were used, where $k_{\text{on}} = k_{\text{off}} = 0.04$.

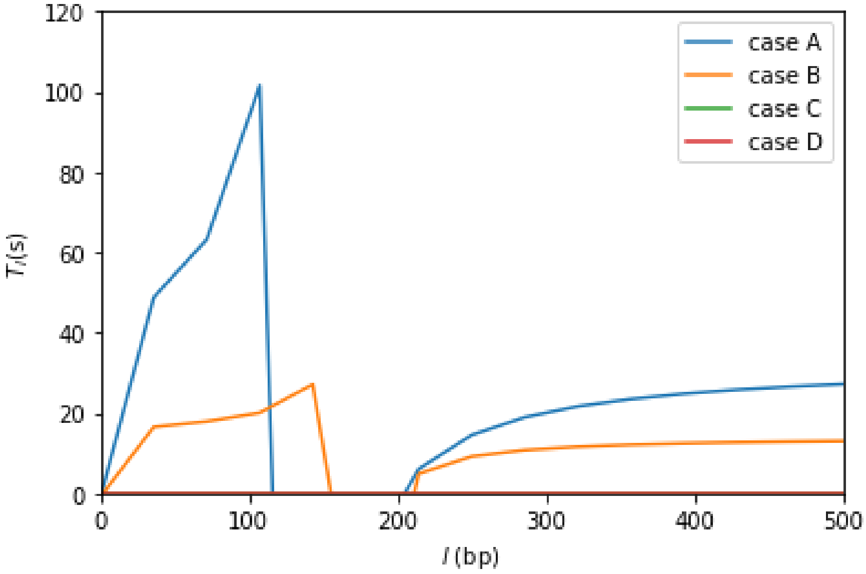

Figure 4.10: Theoretical exposure timescale plot. **Case A** No LEF or nucleosome binding. **Case B** Binding of one LEF. **Case C** Binding of one LEF and one nucleosome. **Case D** Binding of one LEF and two nucleosomes.

The shape of the curves show the desired “butterfly-like” shape with two ‘wings’, similar to what was plotted by Parmar et al [54]. This represents the situation where $l<k$ and

$l \geq (k+l+1+m)$. As observed from Figure 4.10, case A, where there is no binding activity within the two boundary nucleosomes exhibits the highest amount of exposure time on the strand of DNA. Case B, where there is binding of only one LEF, illustrates a scenario where there is a much smaller exposure time on the strand of DNA. The gap in the exposure time has also decreased and shifted toward the right side of the plot. This may be indicative of the sliding activity of the enzyme along the strand of DNA. Note that four lines should have been plotted in this figure. It seems that the activity for cases C and D are hidden along the y=0 axis, which requires further investigation. If this exposure timescale was actually 0, this implies that there can be no more binding activity allowed on this strand of DNA once one LEF binds to the strand. Or, it will be extremely difficult for another enzyme to bind on this section of DNA.

Since Figure 4.10 does not necessarily show the activity for cases C and D, the activity was plotted again on a log axis for the exposure time, as illustrated in Figure 4.11. Since cases C and D are closer to y=0, this is investigated more closely.

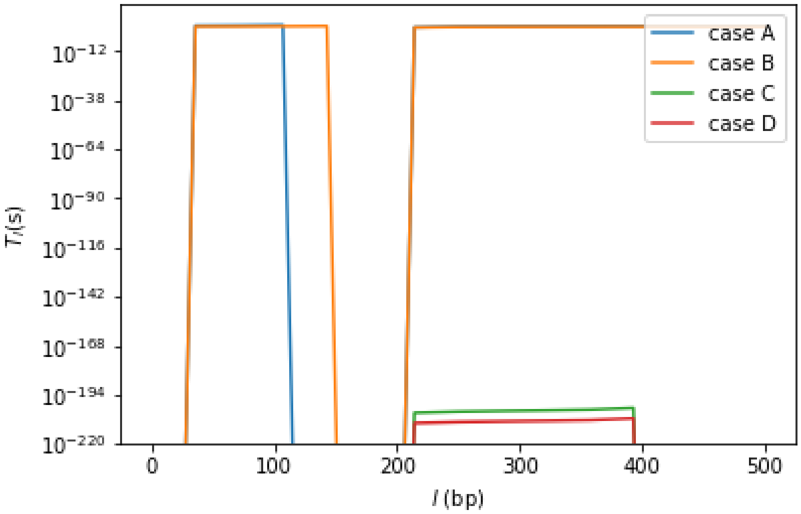

Figure 4.11: Theoretical exposure timescale plot on a log axis. **Case A** No LEF or nucleosome binding. **Case B** Binding of one LEF. **Case C** Binding of one LEF and one nucleosome. **Case D** Binding of one LEF and two nucleosomes.

The theoretical exposure time for cases C and D are now more visible in Figure 4.11. Case C has a slightly higher exposure time than that of Case D, but they are both significantly lower than those of cases A and B. The theoretical exposure time in these cases are found between $T_l(s) = 10^{-220}$ and $T_l(s) = 10^{-194}$, which is extremely small. While it does not completely eliminate the possibility of having two enzymes bind to this section of DNA, it decreases the chance of this occurring. Note that both of these cases contain the presence of nucleosomes within the proximity of binding of an LEF. The implications of their presence will be further discussed in Section 4.5.

In Chapter 3, the idea was explored where the impact of the specificity of binding an LEF to the strand of DNA was explored. Due to the considerations of possible LEF candidates and the specificity of binding for the kinetic rates, this was run again but with a larger $k_{on} = 0.4$ rate (Figure 4.12), as previously used in Chapter 3 for the non steady state kinetics, based on the experiments conducted by Brackley et al [13]. The reason for imploring a larger on rate was the possibility of the LEF having a larger affinity for the target rate, which increases the rate of binding, as described in Equation 4.29.

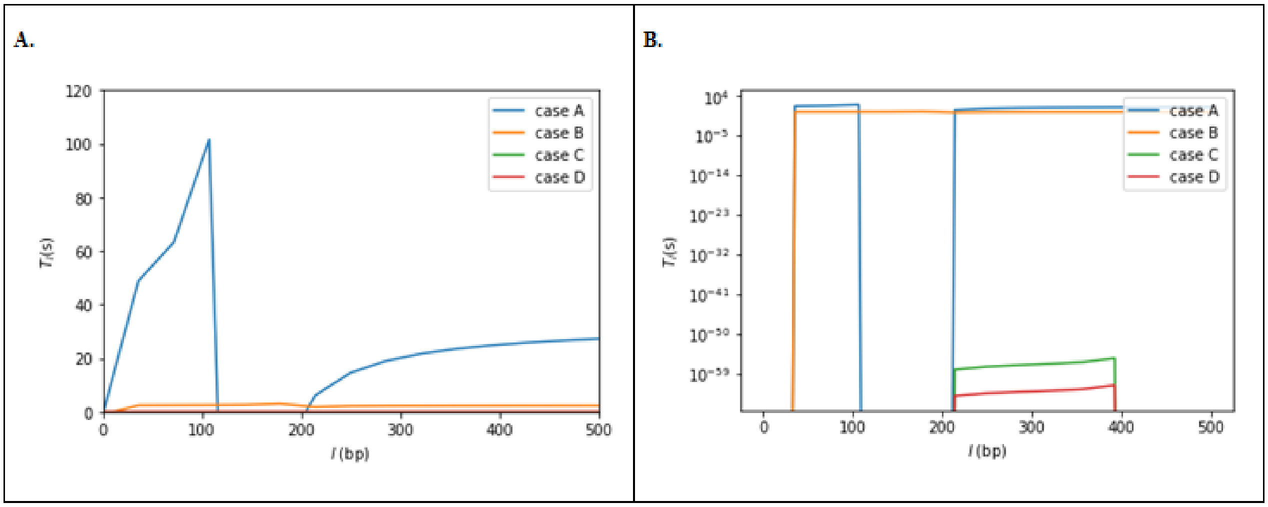

Figure 4.12: Theoretical exposure timescale plot with a greater on rate. Plot A depicts the theoretical exposure timescale plot, and plot B depicts the same plot on a log axis. **Case A** No LEF or nucleosome binding. **Case B** Binding of one LEF. **Case C** Binding of one LEF and one nucleosome. **Case D** Binding of one LEF and two nucleosomes.

It is observed in Figure 4.12 that only case A has maintained its shape. This is due to the fact that no enzyme binds to the DNA in this scenario, so it is not influenced by binding kinetics. However, case B has now significantly decreased and lost its "butterfly-like" shape. The log plot also reflects this observation, where it is more of a horizontal line. It is also noticeable how cases C and D now have a larger separation in exposure, but interestingly, has actually increased in exposure. In the previous plot (Figure 4.11), the theoretical exposure time in these cases are found between $T_l(s) = 10^{-220}$ and $T_l(s) = 10^{-194}$, but in this case, the exposure time was found between $T_l(s) = 10^{-60}$ and $T_l(s) = 10^{-50}$. This can be influenced by the possibility of overcrowding on the strand, which may allow for the binding of both LEFs and nucleosomes. However it will be difficult to initiate activity when there are too many enzyme on the same strand of DNA.

Alternatively, since $k_{off}$ may also exhibit binding specificity, this was run again but with a large $k_{off} = 0.4$ value (Figure 4.13) again based on Brackley et al's experiments [13]. In a similar manner to the increased on rate, a larger off rate was run due to the possibility of a stronger affinity for sequence specificity of unbinding, as proposed in Equation 4.21.

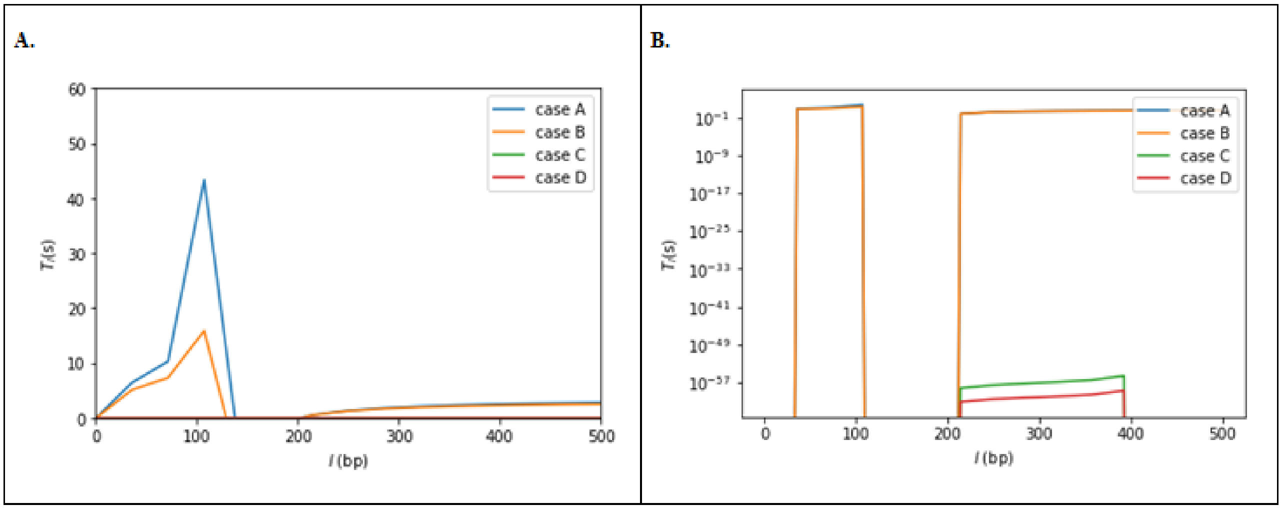

Figure 4.13: Theoretical exposure timescale plot with a greater off rate. Plot A depicts the theoretical exposure timescale plot, and plot B depicts the same plot on a log axis. **Case A** No LEF or nucleosome binding. **Case B** Binding of one LEF. **Case C** Binding of one LEF and one nucleosome. **Case D** Binding of one LEF and two nucleosomes.

In Figure 4.13, since all equations are dependent on the off rate kinetics, it was predicted that all cases would have been impacted. It was observed that the exposure time has decreased for cases A and B. The exposure time then decreases to reach the mean persistence time, $1/k_{\text{off}}$. It was stated by Parmar et al [54] that the exposure time decreases as the adjacent nucleosomes become more unstable. This concept is consistent with Figure 4.13 since the rate of nucleosome unbinding has increased, they do not remain stably bound to the strand of DNA. What was observed was the exposure has decreased in this scenario overall, as well.

The theoretical equations show the possible activity of LEF binding as demonstrated in the proposed cases. The kinetic rates of binding and unbinding ultimately affected the overall shape and exposure time of the DNA strand, as predicted. The possibility of other enzymes binding to this DNA strand will also affect the binding activity of LEFs, as demonstrated by the varying rate of binding and unbinding. This concept will be further discussed in the future directions for this model.

### 4.4.2 DNA Exposure Time Simulations

As observed from the theoretical exposure timescale derivations, the activity of DNA is influenced by the composition of the chromatin in which it is embedded. These factors include nucleosome turnover, conformational dynamics, and covalent histone modifications, which each induce changes in the structure of chromatin and its affinity for regulatory proteins.

The dynamics of histone modifications and the persistence of modification patterns for long periods are still largely unknown. Simulations are now conducted to validate the theoretical equations derived in the previous section with the Nucleosome Tool plug-in within the Stochpy software [2] to visualize the impact of these modifications to DNA with both nucleosome and LEF activity. The code was modified from the plug-in to include the effects of LEF binding.

Further from the previous section, the same four cases will be explored. The trajectory of the DNA is modelled again using the Gillespie algorithm to observed the effects of LEF and nucleosome binding. From this tool, the behaviour of the nucleosome has not been restricted to just binding, unbinding, and sliding. This will also include the recruitment of transferase, interaction, recruitment, interaction, and neighbour nucleosome influences, for a more realistic scenario for LEF binding, as described in Section 4.1.

In each case, four different plots will be shown to summarize the results of each simulation: Pattern Time Series, Pattern Distributions, Species Time Series, and Species Distributions. While the intent of this plot was originally used for nucleosome behaviour, the interpretations of these plots will be different than as described in the Nucleosome Tool documentation [2]. The pattern time series plot will be used to illustrate the modification state of each position on the strand of DNA. The pattern distributions plot will illustrate the behaviour of the probability mass function (pmf) of each nucleosome state at each position on the strand of DNA. The species time series plot illustrates the time it takes for the specific state of the nucleosome to reach a steady state. The species distributions plot illustrates the distribution of the pmf of the copy number of these nucleosomes from the species time series plot. By showing the information from these plot side by side, it will make it easier to show a complete illustration of the nucleosome and LEF dynamics.

In order to interpret the plots, the model was simplified to treat each of the nucleosomes as a single unit that can exist in only one out of three modification states: acetylated (A), unmodified (U), or methylated (M). These modifications are used for active, neutral, and silenced chromatin. The histone modifications are modified with the methylation and acetylation enzyme reactions. This considers the binding, 1D diffusion, and recruitment of the transferases. The methyl- and acetyletransferases can only modify the nucleosome in its unmodified state (from U to M or U to A), and that each nucleosome can be bound by only one transferase at a time. The conversions from a modified to unmodified state are assumed to exhibit basal activity of the demodification enzymes, so the tracking of binding, diffusion, and recruitment of demodification enzymes are not tracked. While this may not seem applicable to this model, the code was modified to interpret this behaviour in a mathematical manner to consider LEF activity.

In this example, let A denote the available DNA, M denote the nucleosome activity, U denote the available binding site for LEF activity, and Mt denotes the movement of the nucleosome. A summary of each case has been simulated, and illustrated in the following sections.

Further the following table summarizes the rates used to model nucleosome kinetics. These parameters were previously described and is not explained here. Refer to the publication for the development is thie iSMS tool for the derivation of these parameters [2].

Table 4.1: Nucleosome Kinetic Parameters ($s^{-1}$)

| $k_{\text{on}}$ | $k_{\text{off}}$ | $k_{\text{slide}}$ |
|---|---|---|
| 2.4 (one enzyme) | 0.1 | 0.6 |
| 0.01 (two enzymes) | | |

The rates of binding and unbinding of the LEF is also fixed at 0.001 $s^{-1}$ based on previous literature and the steady state assumption made previously [57]. This is very speculatory, as true properties and the identity of LEFs have yet to be confirmed. This value is based on the behaviour of the residence time of cohesin, yet it is not known whether cohesin is truly an LEF. Therefore, we are mindful that this rate is not indicative of all LEF behaviour.

### 4.4.3 Case A: No LEF Binding

In this case, there is only the behaviour of two nucleosomes examined, as observed in the theoretical case. Recall Figure 4.6 in the previous section where two nucleosomes surround the target area of DNA where an LEF may bind to the sequence of interest. A summary of the results of this simulation is illustrated in Figure 4.14.

By first examining plot i.), which denotes the pattern time series, it is observed that most of the plot is green with the exception of two areas which show activity in red. This shows all of the free DNA available for binding. The red area depicts the nucleosome activity of that was initially bound to the DNA. Once bound, the nucleosome is free to slide within any free area of DNA. As observed the trajectory of the red area deviates over time, which indicates sliding activity has occurred. However, note that the red area never overlaps with one another, as they have been positioned far enough away from each other that there is enough space between them that it will not be sliding close enough to interact with one another. While the space between them does become narrow at some points in time, namely around 40, it does not overlap. Since the size of the LEF is not completely known, this may not be enough space to allow for LEF binding to occur. In the future, the size of the specific LEF should be considered for these simulations. Since these are not definitely supported by observed experments at this time, it remains theoretical.

The corresponding pattern distributions plot to plot i.) is illustrated in plot ii.). As observed from the previous plot, as the nucleosome activity increases, then the area of exposed DNA is decreased. It is observed from the pmf that the area of DNA exposed peaks is surrounded by the nucleosome activity. The area of exposed DNA is still influenced by the nucleosome

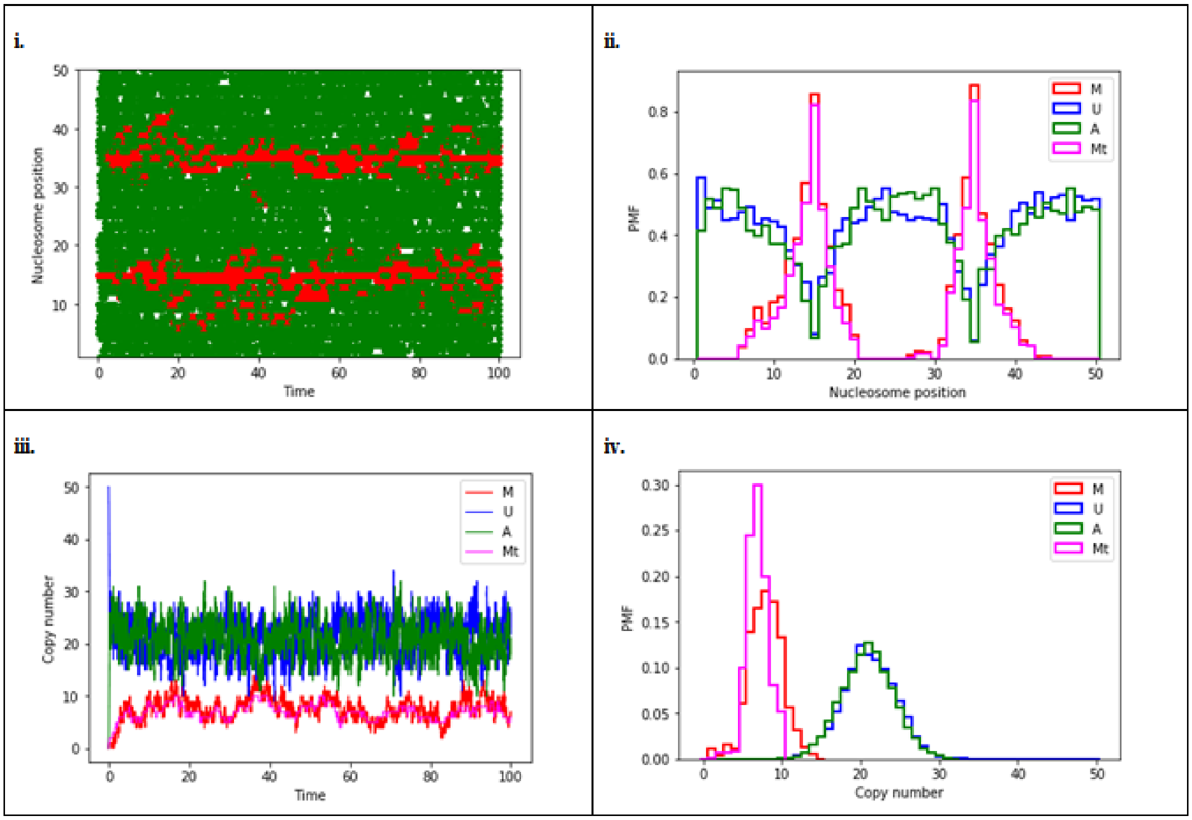

Figure 4.14: No LEF or additional nucleosome binding. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.** Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot.

activity. Had the nucleosome been positioned much closer together, the target area of DNA between these nucleosomes would not be as wide. Therefore, it is important to keep in mind how much space is needed for the proposed LEF to bind onto the strand of DNA. Thus, the importance of the location of the nucleosomes, the LEF, and the target patch must be emphasized.

The species time series plot is depicted in plot iii.). The copy number of the nucleosomes and free DNA binding sites are observed in this plot. It is observed that over time, there will always be more exposed DNA binding sites than that for the nucleosome, as there is no additional binding of other factors.

The corresponding species distribution plot to that of the species time series plot is denoted in plot iv.). As observed from the pattern species plot, the largest pmf stems from the nucleosome activity, as it is the only movement present in the system. There is a higher copy number of available DNA for binding activity to occur, with a lower pmf value. In this system, since the only activity stems from the nucleosomes that are already bound to DNA, the simulated results were as expected.

For comparison, $k_{\text{slide}}$ is now set to zero to understand the extent of the effect of sliding in this system. This is illustrated in Figure 4.15. It is observed from this results that there is far less noise, and nucleosome activity is restricted to two different types of activity: binding or unbinding. As observed from plot i.) in this figure, the movement of the red line shows the nucleosome activity. Once a nucleosome binds, it can only remain bound and does not slide to other areas of the DNA. However, this may unbind, as shown by the gap in the red lines. Since this activity was limited to the nucleosome at positions 15 and 35, it is not known whether this specific nucleosome may then decide to bind to other exposed sections of the DNA. By analyzing the red line at nucleosome position 15 of plot i.), it is observed that the exposed DNA does not remain unbound for a long period of time. A nucleosome then finds itself bound to the section of DNA no later than 5 time points. By examining the other plots ii.), iii.), and iv.), the nucleosome activity is consistent with the assumptions with no sliding, such that the activity is more restricted, and does not have as much noise as that in Figure 4.14.

While removing the sliding parameter was shown for the purpose of comparing the nucleosome activity with that of sliding activity, it is not representative of a realistic biochemical system with nucleosome activity. Nonetheless, this demonstrates that mathematically, it impacts the model for the exposure time for LEFs. The sliding parameter is important to introduce a certain level of noise in the system, as an obstable that will affect the probability of LEF binding to occur.

### 4.4.4 Case B: Binding of one LEF

The binding of an LEF is now included in this case, as a further development of Case A. Recall Figure 4.7 in the previous section where two nucleosomes surround the target area, and now an LEF may bind between them onto the sequence of interest. A summary of this simulation is illustrated in Figure 4.16.

As observed in the plot i.) from this Figure, there is now an inclusion of a white space between the (red) nucleosome activity. This denotes the LEF binding opportunity and exposure time. The nucleosome activity is consistent with that from the previous case, in that the sliding activity allows for the nucleosome to slide onto the exposed DNA area. Over time, the nucleosome sliding behaviour is consistent, although it does not take the same trajectory as the first case. This is one advantage of the Gillespie algorithm, as it introduces a certain level of noise whcih allows for slight variation in each result, but will present a general understanding of what may occur in the biochemical system. In the initial time,

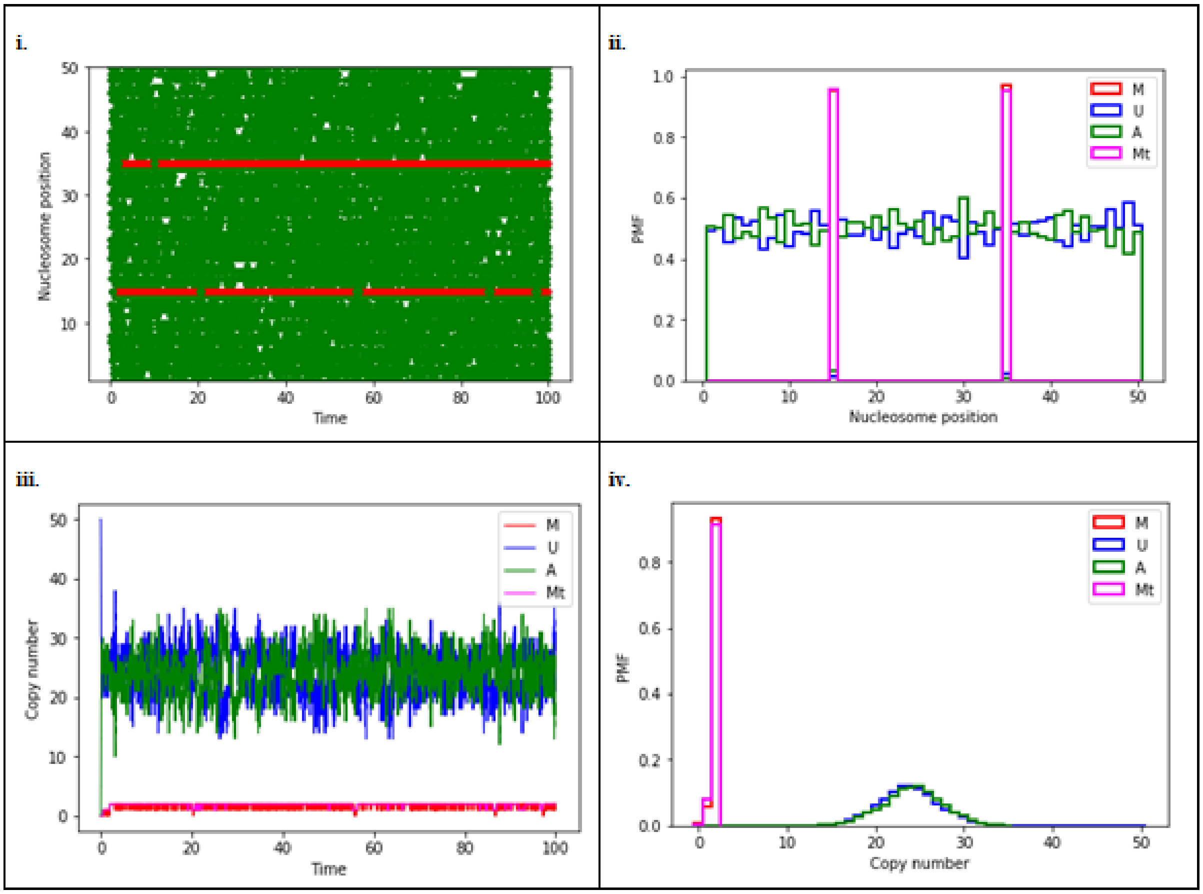

Figure 4.15: No LEF, additional nucleosome binding, nor sliding of nucleomes. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.** Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot.

there is an opportunity for LEF binding, but does not occur until approximately time 35, where there is more white activity in the middle of the plot. This implies LEF an LEF has bound to the DNA. The nucleosome activity is still consistent however, and does not alter its behaviour necessarily, as it still slides within free regions of DNA.

Further from this plot, plot ii.) depicts a similar illustration to that of the first case albeit the distance between the nucleosome activity is much narrower in distance. The reason is that the initial case was set to a more narrow area to see whether the distance would affect LEF activity. As observed, if the nucleosome activity overlaps with the exposed area of DNA at some point in time, the LEF may become more unstable and will be less likely to bind or remain bound to the strand of DNA. It would be interesting to further develop this model in this case to allow for multiple LEFs on the same loop, as this would affect whether the

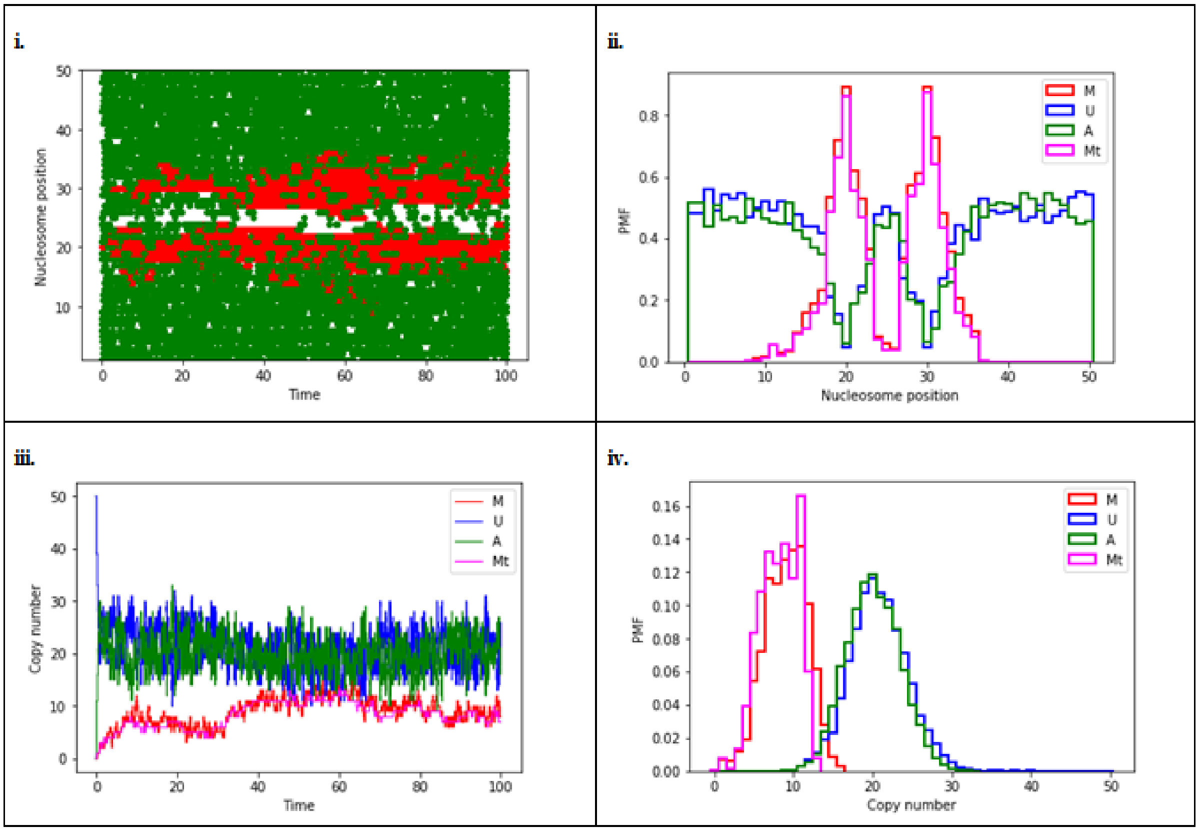

Figure 4.16: Binding of one LEF. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.**Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot.

loop will remain there over time. However, this is beyond the scope of this investigation for the first passage time activity of LEFs.

In the species time series plot iii.), it is observed that this differs from the first case such that the nucleosome activity at one point approaches the same number of free, exposed DNA. This implies that the nucleosome activity may have been affected by the LEF binding at this point, such that there is less exposed DNA. Since it seems at this point, the nucleosome activity has maintained a steady copy number, and the amount of freely exposed DNA has slightly decreased at this point, it implies that the LEF has bound to the DNA, but the nucleosome activity may interact with the LEF at this point. Later, the LEF would have begun to extrude the loop, and the nucleosome activity will decrease, and the amount of free exposed DNA will increase, which is observed at the later time period, around 90.

Finally, the species distributions plot iv.) has also differed from the previous case as the pmf has leveled, and the two peaks occur at almost the same value. However, the copy number of freely exposed DNA still remains higher than that of the nucleosomes. At some point, the copy number of the nucleosome and the exposed DNA overlaps, which is consistent if the nucleosome activity impedes the binding of exposed DNA.

Since it was observed from Figure 4.16, the effect of distance of the nucleosomes may impede LEF binding, this example was re-simulated with a greater distance between neighbouring nucleosome to allow for better LEF binding. This result is illustrated in Figure 4.17 with the same four plots.

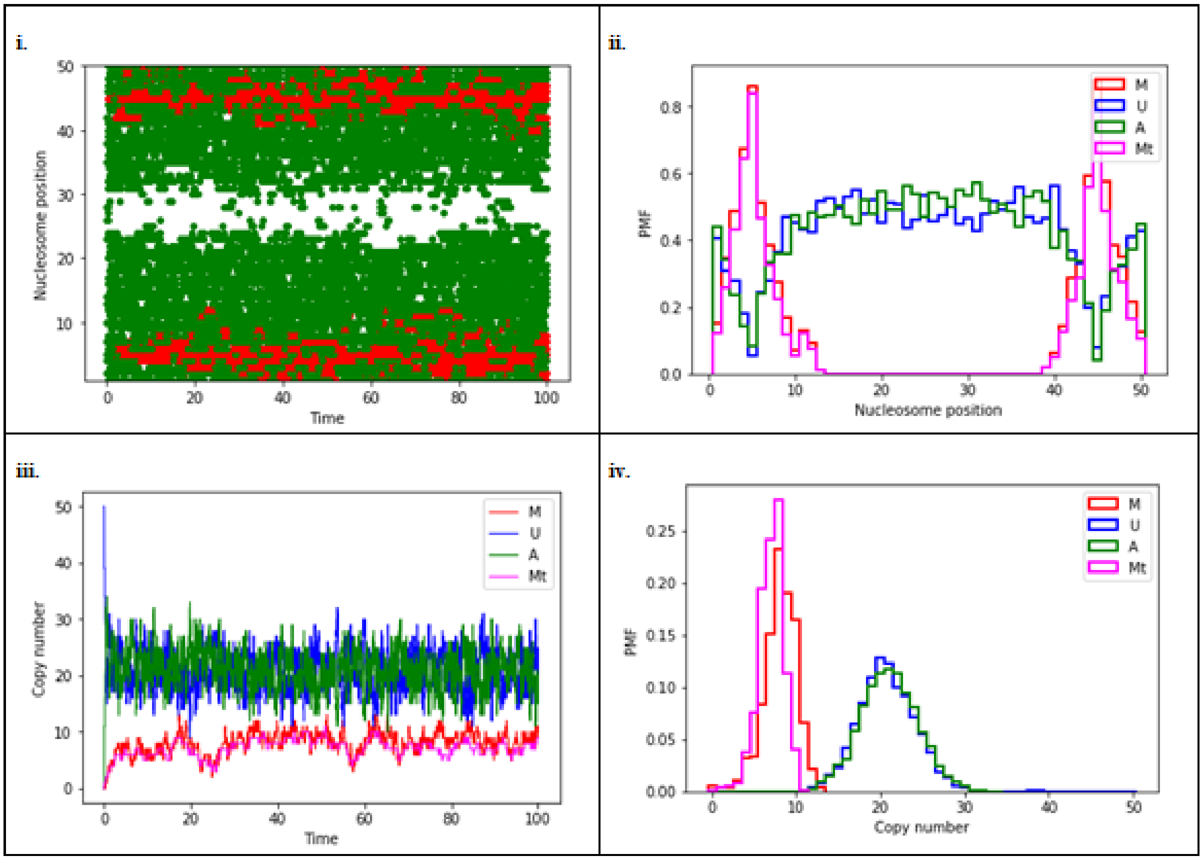

Figure 4.17: Binding of one LEF. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.**Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot.

It is observed from this result that the LEF is not affected by the impact of nucleosomes, as there is enough distance between them to allow for better ease of binding. The risk of unbinding is lower because there are no nucleosome obstacles in the path of the LEF. The nucleosomes and the LEF are able to slide more freely in their respective positions on the strand of DNA. Therefore, the exposure time of the strand of DNA for LEF binding would be smaller should the distance between the LEF and obstructing nucleosomes be large, as the presence of nucleosomes will unlikely contribute to LEF unbinding.

### 4.4.5 Case C: Binding of one LEF and one nucleosome

Further from the previous case, the binding of additional nucleosome is now simulated into the system. Recall Figure 4.8 in the previous section for the theoretical case. Now an additional nucleosome will also bind within the area between the two nucleosomes that are already bound to the DNA. The binding of an LEF will also be included, where it will bind to the target patch of DNA. A summary of this simulation is illustrated in Figure 4.18.

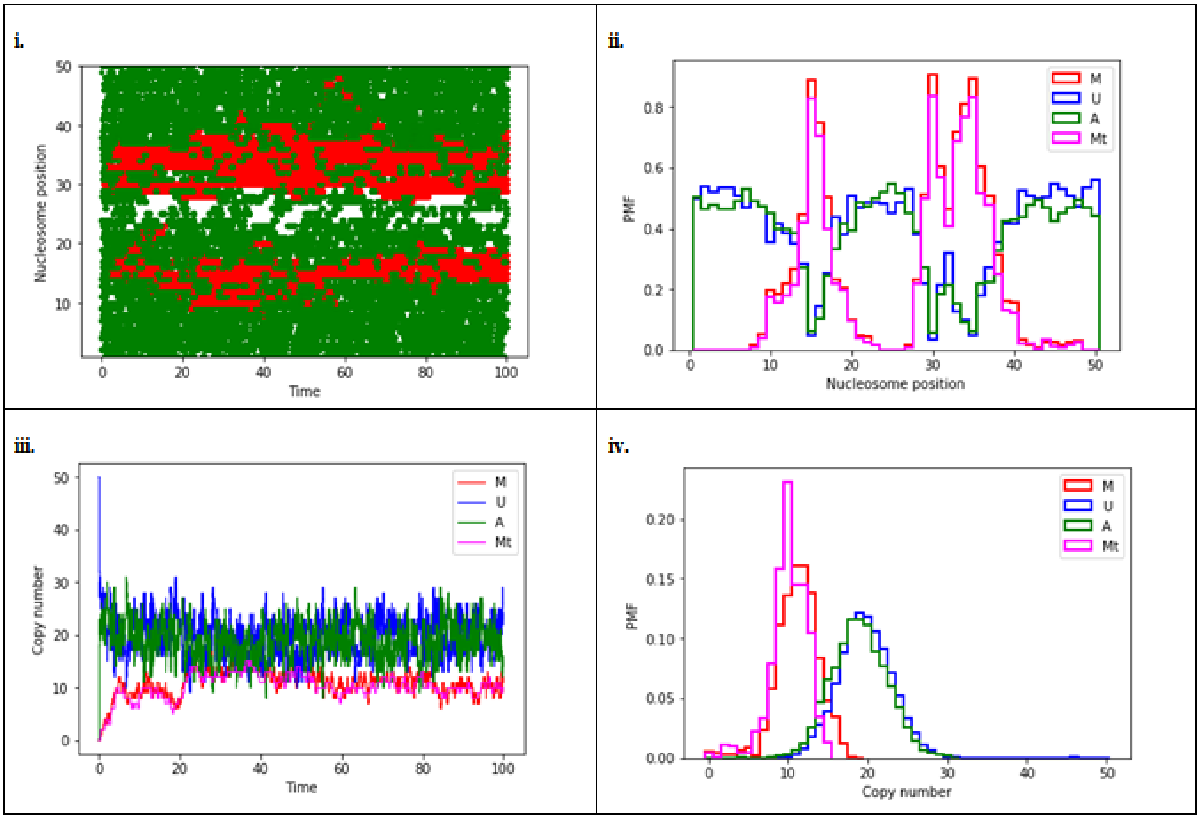

Figure 4.18: Binding of one LEF and one nucleosome. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.**Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot.

It is observed in the pattern time series plot that the nucleosome activity is much heavier than from previously, as expected. It would also seem that when the nucleosomes are side by side, as in the case between the nucleosomes between position 30 and 40, the sliding activity occurs more dense. When there is only one nucleosome bound, as in the case between nucleosome position 10 and 20, the nucleosome seems to deviate away from the free patch of DNA in the middle of the plot. The LEF bound, is not as stable as was previously observed in the first plot, when the activity of the nucleosome closest to it becomes closer. However, it would seem that the sole nucleosome, as it is sliding away from the area of activity, compensates for the second nucleosome bound and creates a wider area of exposed DNA. One may question how this may affect loop extrusion activity. Since one side has a more dense area of nucleosome activity, would the LEF simply extrude on the area of DNA that has more area between nucleosomes, or extrude the loop symmetrically? Although, due to the unstability of the LEF due to the dense nucleosome activity, it may eventually unbind and loop extrusion activity would not occur altogether. The exposed section of DNA may possibly be consumed by the nucleosome activity.

In the pattern distributions plot ii.), the three nucleosomes are observed from the three different peaks of the nucleosome activity. The free exposed DNA is observed, and shares the same pmf as the previous plots. This corresponds to the activity that may occur for LEF binding. Of note, as the two nucleosomes are bound side by side, there may be more sliding that occurs as they interact with one another. Thus, the peaks of these nucleosomes have blended together, but both peaks are still separate and distinct from one another. In the case of the one nucleosome surrounded by free DNA, this peaks is much more narrow and does not slide as often to take over the other exposed DNA regions. Thus it would be more favourable for LEFs to extrude in a one-sided manner to allow for more movement in terms of loop extrusion. With more nucleosome activity, this minimizes the free space that the LEF may occupy once it slides onto the DNA into its new position after loop extrusion.

In the species time series plot iii.), this differs from the previous plots as the nucleosome activity has reached the same copy number of the exposed DNA. This implies that this has overlapped with the region in which the LEF may bind. However, as this deviates again and almost interacts again, this is interpreted as observing the unstability of the LEF. As Goloborodko, et al. [30], previously explained, the loops are more stabilized when there are multiple LEFs on the same loop to reinforce the structure. Due to the specificity of the LEF binding, the nucleosome may have interacted with the LEF at some point and may obstruct the structure. Another interpretation is that loop extrusion would have commenced at time 40, as the nucleosome activity has decreased afterward. This implied it would have stabilized and the DNA would have been more free and exposed once extruded in the loop.

In the species distribution plot iv.) it is interesting to note the overlap between the peaks denoting the amount of exposed DNA and the amount of nucleosome activity. There is more activity for the nucleosomes present in the system, thus the higher pmf is shown. The peak is significantly higher in terms of nucleosome activity, and it almost overlaps with the amount of free DNA. However, there is still more exposed DNA, as it has a higher copy number. Since there are only 3 nucleosomes present in the system, it is expected that the copy number remains lower.

### 4.4.6 Case D: Binding of one LEF and two nucleosomes

Finally, the binding of one more additional nucleosome in now introduced in the system. Recall from Figure 4.9 in the previous section for the theoretical case that two additional nucleosomes will bind within the area between the two nucleosomes already bound to the strand of DNA. The binding of the LEF will also be included, where it will bind to the target patch of DNA. A summary of this simulation is illustrated in Figure 4.19.

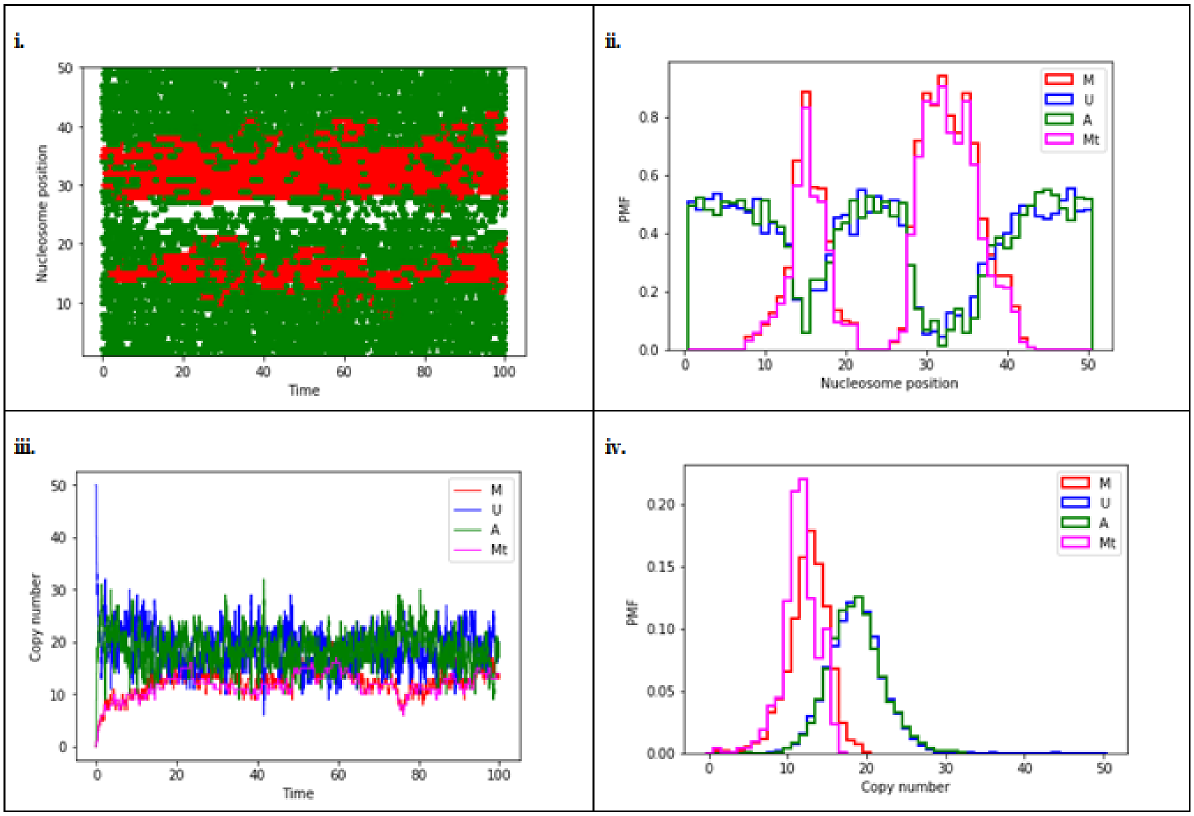

Figure 4.19: Binding of one LEF and two nucleosomes. The values of M, U, A, Mt are as follows: nucleosome activity, available binding regions for LEFs, free exposed DNA, and the movement of nucleosomes. **i.**Pattern time series plot. The area for free, exposed DNA is denoted in green. The area of nucleosome activity is denoted in red. **ii.** Pattern distributions plot. **iii.** Species time series plot. **iv.** Species distribution plot.

By examining the pattern time series plot i.), it is observed that the binding of all three nucleosomes in close proximity would affect the binding of LEFs. If those nucleosomes bind to the DNA before the LEF, this will delay the exposure time of DNA, as explained in the theoretical case. It is also observed that the nucleosome activity seems to close the gap from the exposed DNA between them. This could be due to the nucleosome sliding activity. Since one side of the target patch has the possibility of having 3 nucleosomes, this could be too crowded for that portion of DNA. These nucleosomes could then slide into the exposed patch of DNA and prevent LEF binding. The target patch of DNA, if it is still exposed, may not have enough space for the LEF to bind. Instead, the 4 nucleosomes present in the system may even out the distance between them. As observed from the LEF activity in white, it disappears at one point, which may be due to the crowding from the nucleosomes. Therefore, there is not enough space in this system to allow for the LEF binding to occur, as the nucleosome activity have take presedence in the system. Although the target patch of DNA may still be exposed, the nucleosomes may be too close together.

In the pattern distributions plot ii.), it is observed that the peak on the right depicting the activity of the 3 nucleosomes is now much wider. There are now 2 distinct peaks for the nucleosome activity, as it seems the three nucleosomes on the right position of the target patch has overlapped. The single nucleosome on the left of the target patch has, however remained narrow. There is very little space between the two nucleosomes, which may allow for LEF activity.

In the species time series plot iii.), there is now very little distinction between the free DNA and the nucleosome activity. It is observed that there is more interaction between the two activities. This implies that there is now an even number of free DNA in the system as there is for nucleosome activity. In terms of LEF activity, it seems that at one point, around time 60, in which the LEF would have bound to the DNA and started extruding the loop, as the nucleosome activity decreased. However, the nucleosome activity has then risen once again, which implies there could have been competition between the LEF and the nucleosome and the LEF had finally unbound from the DNA as there were too many elements bound to the DNA. It is challenging to have definite interpretations of the plot, as the observed data is not available to validate these speculations.

In the species distribution plot iv.), there is now a larger overlap between the nucleosome activity and the free DNA. The pmf of the nucleosome activity is much larger than that of the free DNA, which implies more activity with the nucleosomes. Since the nucleosome activity overlaps with that of the free DNA, it is observed that there will be less free DNA for elements to bind. This is consistent with the idea that by binding more nucleosomes, this will delay the exposure time for the DNA and delay the binding of an LEF. Due to the sliding nature of nucleosomes, this will also add to delaying the exposure time of DNA.

## 4.5 Discussion

These findings were consistent with those as discussed previously by Parmar et al. [54] for the theoretical estimates of exposure timescales of protein binding sites on DNA by nucleosome kinetics, despite the inclusion of the effects of LEF binding in this case. It is still important to stress the locations of both nucleosome and LEF binding kinetics. The specificity of binding is also important, as it determines the type of activity that may occur in that region of DNA. This work was novel as research on chromatin compaction is limited in the influence of nucleosomes, and how the presence of other enzymes on a strand of DNA will impede loop extrusion activity.

As observed from the simulations conducted above, nucleosome activity will delay the exposure time of DNA, and it may prevent the binding of LEFs. Some other considerations is how this nucleosome activity will affect loop extrusion. In the Nucleosome Tool [2] plug-in used, there was already a loop rate however, this model focuses on chromatin interactions, and how two non-adjacent nucleosomes are connected to create a loop. In the loop extrusion model, an LEF will bind to DNA to extrude the loop with motor activity. In this tool, it was described that if transferases are present on the nucleosomes, they are able to hop to the connected nucleosomes, where a methylated nucleosome will interact with an acetylated chromosome. Therefore, this is more of a fold on the DNA rather than extruded. This rate was ultimately not required, since we focus mainly on LEF behaviour, not of nucleosomes. However, the concept of nucleosome hopping would be interesting to explore in future models to understand how nucleosomes position themselves onto DNA loops.

The concept of the chromatin as performing self-organization without a motor is similar to the concept of loop extrusion without a motor[12]. While loop extrusion is generally favoured, there has yet to be a suitable motor protein, or a motor activity in cohesin itself that will allow this to occur. In the model proposed for loop extrusion without a motor, it was speculated that a thermal motion within the nucleus drives extrusion. Perhaps this concept may be explored further in the future with more information regarding the specific manner in which loops may be formed. So far, recent HiC data has revealed a strong bias in favour of a particular arrangement of the CTCF binding motif that stabilize loops, and extrusion is the only model to date that can explain this. This is the reason why the loop extrusion model was heavily studied in this thesis.

One question that was raised during the simulations was the effect of nucleosome presence on LEF binding. As observed from the results, when there were more nucleosomes present on the strand of DNA, loop extrusion was more challenging, and became unstable. It is not clear whether the LEF would be able to remain bound to DNA in this case. Since it was proposed that cohesin may be an LEF, the behaviour of this enzyme was explored. It was previoulsy determined by Liu et al [36] that in budding yeast, cohesin is loaded onto the chromosome during the late G1 phase, established sister chromatid cohesin concomitant with DNA replication, and dissociates in telophase. In their experiments, it was shown that at anaphase, nearly all of the cohesin binding sites contain nucleosome-free regions. The majority of these sites remain nucleosome-free throughout the cell cycle, which is consistent with the suggestion of a DNA binding anchoring protein present at these sites, although such a region could also serve as part of a marker for the binding of cohesin in the next cell cycle. Therefore, the regions must be wider to allow for LEF binding from the simulations in our experiments conducted.

The behaviour of cohesin with nucleosome activity was previously explored by Stigler, et al.[67]. Single molecule microscopy was used to observed the dynamic and functional characteristics of cohesin bound to DNA. Cohesin is able to undergo rapid one-dimensional diffusion along DNA, but individual nucleosomes, nucleosome arrays, and other protein obstacles significantly restrict its mobility. DNA motor proteins can readily push cohesin along DNA, but they cannot pass through the interior of the cohesin ring. It was revealed that DNA-bound cohesin has a central pore that is substantially smaller than anticipated. Therefore, one should consider whether the DNA loop extrusion model should include the impact of nucleosome presence on LEF binding to DNA. While the binding may not be possible on a nucleosome, and DNA must be exposed without other proteins already bound, the loop extrusion procedure itself still remains unknown in literature.

The use of the fixed rate of binding and unbinding of LEFs is also a limitation in this study. Since the true rate of binding activity of either cohesin and condensin varies depending on several factors such as temperature, solvent concentration, ATP activity, etc., the rate may vary from cell to cell. Therefore, the binding activity may appear more frequenty than that of nucleosome binding depending on the conditions of the cell. In this model, only one perception of LEF binding behaviour is explored. This is not indicative of all cell conditions, but it should present an understanding of one possible method of LEF binding behaviour.

In most models when analyzing loop extrusion, the illustrations do not include the presence of nucleosomes in the loop. It is unclear whether the LEF will simply jump over the nucleosome present on the strand, if they will bind to the chromatin loop after it is extruded, or if they are simply not present at all. Further research exploring the effect of nucleosomes in loop extrusion is required to validate the loop extrusion process. Recent single-molecule imaging experiments conducted by Kong et al [41] have shown that both condesin I and II exhibit ATP-dependent motor activity and promote extensive and reversible compaction of double-stranded DNA. Nucleosomes are incorporated into DNA loops during compaction without being displaced from the DNA, indicating that condensin complexes can readily act upon nucleosome fibers. Based on this finding, it may be a novel characteristic of LEF binding behaviour such that the binding activity of both an LEF and nucleosomes may be occurring at the same time during loop extrusion acitivity. In this model, this characteristic was incorporated by considering the mean first passage time for LEF binding to allow for loop extrusion. However, since the timing of nucleosome and LEF binding is not fully understood, this will impact the model such that nucleosome binding and competition may have a smaller role in the loop extrusion process. If competition of LEF binding does not occur, it may not have to be included in the model at all. More research is needed to fully map out the timing of LEF binding and nucleosome activity, and whether these activity can occur simultaneously without competition, but the model developed is advantageous for examining cases where competing enzymes can promote loop extrusion activity. Our model was based on cohesin binding activity. While cohesin and condensin may exhibit similar loop extrusion properties, their individual influence on nucleosome binding may also differ. Therefore, specific LEF characteristics need to be established prior to applying this model and determining whether they are in agreeance.

As observed from our results, this is consistent with what is currently understood and stated regarding the relationship between nucleosome activity and cohesin. Therefore, the Gillespie algorithm was successful in simulating the exposure timescale of DNA with respect to LEF and nucleosome activity for the motor model proposed for chromatin loop extrusion. It is observed that nucleosome presence impedes exposure and LEF binding activity. However, this is only speculatory based on the theoretical equations explored in this chapter. More experimental evidence would need to be performed to fully support this assumption.

## 4.6 Future directions

There are several different factors that influence enzyme binding on a strand of DNA. Loop extrusion activity is dependent on the environment in which the LEF binding activity is promoted.

Munoz et al [49] have determined that chromatin remodeling is required to generate a nucleosome-free region that is the substrate for cohesin loading. An engineered cohesin loading module can be created by fusing the Scc2 C terminus (a cohesin loader) to the RSC chromatin remodeling complex or to other chromatin remodelers, but not to unrelated DNA binding proteins. This demonstrates the importance of nucleosome-free DNA for cohesin loading and provide insight into how cohesin accesses DNA during its varied chromosomal activities. To incorporate this into our model, for the future, the presence of the cohesin loading enzymes should be incorporated into the kinetic rates that determine the LEF binding parameters. This will influence the likelihood of whether a region of DNA will be free of nucleosomes and increases LEF binding opportunity.

It was also previously determined by Horlbeck et al [33] that nucleosomes also impede Cas9 access to DNA in vivo and in vitro. Cas9 is a prokaryotic CRISPR (clustered regularly interspaced palindromic repeats)-associated protein. It has been widely adopted as a tool for editing, imaging, and regulating eukaryotic genomes. This study aimed to understand the selection of single-guide RNAs (sgRNAs) that mediate efficient Cas9 activity, since there was an information gap into how chromatin impacts Cas9 targeting. It was also determined that highly active sgRNAs for Cas9 and dCas9 were found almost exclusively in regions of low nucleosome occupancy. If this is the case, since our current understanding is that cohesins also bind to DNA in nucleosome free regions, perhaps competition should also be considered in this model. The in vitro experiments conducted by Horlbeck et al [33] have observed that nucleosomes in fact directly impede Cas9 binding and cleavage, while chromatin remodeling can restore Cas9 access.

Future direction for this model would include testing potential LEF candidates and the extent of its behaviour in this model. Some LEF candidates include condensin and cohesin, however they may not bind to DNA and extrude loops in the same manner such that they will require different cell conditions, or environments, in order to extrude a loop. It was previously examined by Baxter et al [4] that in silico simulations predicted that if SMC complexes are acting as part of LEF machines, then they should display distinct biochemical and cell biological characteristics. First, ablation of SMC activity should lead to a cessation of sequencial *cis*-looping along chromosomes, but not inhibit chromosome loops acting across other looped domains or in *trans*. Second, generation of loops by LEFs should be associated with the translocation of SMCs along DNA. Third, the coverage and size of looping should be a function of the number and processivity of SMCs on chromatin. Therefore, when experimental parameters for potential LEF candidates become available, this can help gain clarity on the behaviour of LEFs in the presence of nucleosomes in the future.

## 4.7 Conclusion

Loop extrusion factor activity is hindered by the presence of nucleosomes on DNA. The kinetic rates to describe the activity of both LEFs and nucleosomes are influenced by the specificity of bindng on a strand of exposed DNA which will aid in determining whether loop extrusion activity can occur. Due to the limited experimentation conducted toward the understanding of chromatin compaction, this model is to be used as a potential guide toward the understand of loop extrusion activity in the presence of nucleosomes.

# 5 General Conclusion

Aspects of the chromatin loop extrusion process were able to be modelled with Markovian dynamics. The GMM was successfully able to highlight specific states of chromatin folding in smFRET data, macroscopically. The EM algorithm was able to optimize the GMM parameters for this data by identifying the significant clusters. Microscopic folding properties of chromatin configurations were also explored through LEF binding and unbinding kinetics. The specificity of its binding properties is an important parameter that determines whether the loops may be extruded in that section of DNA. The number of LEFs bound to the loop may deduce whether the loop is stably bound, and thus promote extrusion. The amount of time in which the region of DNA explored is also dictated by the number of nucleosomes present in the system. Therefore, future works may explore the impact of several LEF candidates on loop extrusion for chromatin organization.

## 5.1 Limitations of research

Limited research has been conducted regarding the chromatin organization procedure. This research was based on the current understanding of loop extrusion factor binding on a strand of DNA to extrude chromatin loops. This area of research is rapidly developing and gaining more interest. Even as this work was conducted, there has been continuous development in the study of chromatin compaction.

There are many assumptions in these models that have yet to be verified with observed experimentation. Simulations runs in these models are limited in the assumptions of what is currently known of the behaviour of LEFs in chromatin compaction. When this information becomes available for the future, this can help validate and adjust the proposed models accordingly. However, this work is still important to consider in this field, as it allows us to raise questions about some areas of research that can influence the mathematical modeling. This also allows us to propose possible conclusions and assumptions that may be expected regarding the behaviour of chromatin.

As discussed, there is no concrete list of LEFs that has been developed. Their properties are developed based on what is speculated for loop extrusion activity and what should be occuring based on our current knowledge of the DNA loop formation process. There are several factors that may not be considered in their properties, such as the manner in which external factors affect the loop extrusion activity.The following are research models for loop extrusion by: the structual maintenance of chromosomes (SMC) family of ATPases, DNA-loop-extruding enzymes, transcription-induced supercoiling, osmosis, and thermal motion (diffusive extrusion) [63]. In this research, most of the models relied solely on loop extrusion by the structural maintenance of chromsomes (SMC) family of ATPases, and loop extrusion by DNA-loop-extruding enzymes. Mathematically, this influences the manner in which the behaviour of loop extrusion can be modelled. This was preferred as there was more information available to support these assumptions over the other models.

The postulation that loop extrusion by the SMC family of ATPases is consistent with what has been discussed to formulate these models. The SMC family of ATPases such as condensin and cohesin, extrude DNA into large loops. Energy from hydrolysis helps the SMC proteins to translocate along the chromatin fibers. A single complex of these LEFs could extrude DNA and the growth of the loop can stop when the LEF encounters a barrier element in the convergent direction, which are typically present at the border of topologically associating domains (TADs). This was ultimately the motivation for this research, as there was real-time evidence observed in previous experiments that validate this finding. However, other models have also been researched that can change this point of view. These models are discussed as a form of limitation of this research, as they alter the assumptions in our models.

The model that proposed that loop extrusion occurs via DNA-loop-extruding enzymes postulate that they are bound to a DNA lattice of finite length. Each enzyme is assumed to have two binding domains that can bind and bridge two DNA sites. It is thought that ATP can hydrolyze each binding domain to move along the DNA. The protein link between each binding domain leads to extrusion of a DNA loop. This association of machines and DNA, known as infinite processivity, forms a disordered distribution of small loops. However, if dissociation of the machine and DNA occurs, known as finite processivity, highly stable and large DNA loops are formed with very less fluctuations. Our models become impacted in this case such that the size of the loops and the fluctuations of the loops are affected. The immigration-death model may no longer be appropriate to model the behaviour of chromatin loops, as the stability may not be influenced by the number of LEFs present in the system. External environmental factors may be the driving factors that affect the size of the loops. The immigration-death process would have to be adjusted, or it may simply be a matter of adjusiting the kinetic rates of loop formation and division.

In the case that loop extrusion by transcription-induced supercoiling, loop extrusion occurs when TAD cohesin rings are actively pushed along chromatin fibers, forming cohesin 'handcuffs'. These handcuffs are then released by the DNA topoisomerase IIB (TOPIIB) enzyme present on the TADs borders to form DNA extrusion loops. The motivation for this scenario is due to the lack of evidence that cohesin rings act as active DNA translocases to give rise to loop extrusions. However, transcription-induced supercoiling is a well-documented process in living organisms. This involves the presence of an enzyme chaperone, TOPIIB, required to form DNA loop extrusion loops. The quantity of this chaperone enzyme, and the understanding of its own mechanism of binding to the LEF is required for the models.

Yamamoto and Schiessel [70] examined the behaviour of cohesin ring dynamics on a loop with cohesin loaders in the middle of the loop and unloaders at the loop ends. This mechanism assumes that cohesin monomers bind to the loader more frequently than cohesin dimers. The cohesin dimers facilitated DNA loop extrusion due to osmotic pressure exerted by the cohesin monomers. However, this theory is only true if: 1) there is rare loading of cohesin dimers onto the chromatin fiber, and 2) cohesin monomers do not form dimers on the chromatin fiber. This affects the proposed enzyme kinetic rates. The assumption that loop extrusion activity occurs once an LEF binds would be incorrect, and two enzyme would have to be bound to DNA in order for loop extrusion activity to initiate. This would also affect the theoretical exposure timescale equations, as described in Chapter 4, as this assumes that only one LEF would be neccesary for loop extrusion. Not only would the number of nucleosomes affect the theoretical exposure time, but DNA activity would have to wait until two cohesin monomers were bound to the DNA strand.

Finally, the last proposed theory is that loop extrusion is driven by thermal motion, which is also referred to as diffusive extrusion. This is motivated by the fact that there is no suitable motor protein that generated loops have been identified in this research, despite how HiC experiments have suggested that CTCF binding motifs and cohesin rings lead to loop extrusion. Brackley et al [12] used HiC to carry out simulation experiments in order to shed more light on whether diffusive extrusion could generate DNA loops. Their simple 1D simulation demonstrated a ratchet effect. In their simulation, different cohesin handcuffs were continually loaded and uploaded from a chromatin fiber. It was observed that if the handcuffs were loaded at random locations, a number of loops were formed side-by-side competing for space. However, if the handcuffs were loaded at a single location, it led to a ratchet effect. This ratchet effect prevented the first handcuff from diffusing back towards the loading site, thus creating an osmotic pressure leading to diffusive loop extrusion. The handcuffs are ultimately not included in this research. This would affect the theoretical exposure equations such that the position of the target patch would not be applicable. There would not be a target location for loop extrusion activity to occur if the cohesin binds to DNA in a completely random manner. The specificity of binding would have to be re-defined, as there will not be an affinity to a specific location.

As discussed, there are several loop extrusion theories that researchers have been continuously working to validate. Depending on the definitive evidence that becomes available in the future, there will have to be modifications to the proposed models as discussed in this research. However, these models can aid toward understanding what is currently speculated regarding LEF behaviour. While it is not necessarily correct, it also may not be completely invalid depending on the model that holds true. This work is still relevant to consider ways to incorporate specific elements into the mathematical modeling of chromatin behaviour.

## 5.2 Future directions

In general there have been several future directions proposed for each model as described at the end of each chapter. In terms of the general future directions required with this research, this involves more work required in the broader perspective for loop extrusion. The motivation of mathematically modelling the behaviour of chromatin loop formation was to contribute to the current understanding of the DNA organization process. However, this still raises a few questions. From the second chapter, the specific number of components that would be optimal to apply the GMM to smFRET experiments remains unknown, as current research has limited observed data to support any proposed value. This would be useful to answer questions regarding the impact on transcription activity and the type of information that would be conserved in these DNA loops. In the third chapter, the proper parameters required to derive the LEF kinetic rates for the burst model are still limited due to the unknown environmental requirements for LEFs to perform loop extrusion activity. In the fourth chapter, the kinetics are also affected to determine the full impact of LEF behaviour in conjunction with nucleosome activity. While this research is lacking in the resources required to answer these questions, it can contribute in understanding the behaviour when the necessary resources become available.

Loop extrusion impacts transcription activity such that genomic DNA must be compacted to fit in the cell, but must simultaneously remain accessible for transcription [14]. As Racko et al have previously discussed, transcription may be needed for the formation of TADs [56]. Recent studies of chromatin structure in inactive X chromosomes have shown that a few genes that were transcriptionally active in these chromosomes, were all located in chromosome portions forming TADs, whereas the rest of the chromosome was free [28]. However, once TADs are formed and chromatin loops are extruded and stabilized by interaction between cohesin and CTCF, the ongoing transcription would not be necessary anymore for TADs maintenance. It is important to know the impact of transcription on this process, as this may be needed to initiate the proposed models. In the future, this factor can be implemented in models necessary for loop extrusion. The integration of transcription activity should be included in these models with a more significant impact in the future. Transcription activity should be incorporated as an additional component to the GMM model, to identify at which stage during the loop extrusion process that the cell begins to use the information stored in these chromatin loops. In the burst model, the necessary environmental that would promote transcription activity would have an impact on the kinetic rates of LEF activity. The presence of nucleosomes would also be affected by transcription activity such that it may also interfere with the procedure. By understanding the impact of transcription on loop extrusion, this will contribute to understanding the broader field of genome organization and how it influences gene expression. The spatial organization between the LEF and DNA will allow for an understanding of how the LEF overcomes some of the DNA-bound obstacles to initiate transcription activity.

In the future, these methods would have to be adapted according to the information conserved within each chromatin loop. By understanding how loop extrusion occurs and the motivation for storing specific types of information, the proper parameters can be included when deriving the kinetic rates for loop extrusion activity. This would introduce specific limitation parameters in the configuration procedure. For example, Zhang et al [73] found with dCas9 that additional, undefined, chromatin-based mechanisms may enhance the synapsis of functional cis-elements via loop extrusion more generally. There was also scanning that was impeded by targeted binding of nuclease-dead Cas9 based on the position of the loop anchor. This implies that the position of transcription promoting elements on the loop can impact what information should be extruded in a loop to prevent or enhance accessibility to this information. Mathematically, this could impact the probability of loop extrusion occurring. Perhaps this can be defined by a specific distribution based on the presence of specific elements that are present on the loop. This can impact the rate of binding and loop extrusion activity that will affect every model proposed in this research.

In general, our work was unique in applying Markov models to illustrate its application in different stages of the chromatin loop extrusion process. There is more work required to gain a full understanding of the mechanism and motivation behind loop extrusion in itself, but ultimately, this procedure can be successfully modelled with the specific mathematical methods presented. As the loop extrusion process becomes more defined in the biological community, the mathematical models should be able to adapt accordignly.

# Bibliography

[1] Alipour, E. and Marko, J. F. (2012). Self-Organization of Domain Structured by DNA-Loop-Extruding Enzymes. *Nucleic Acids Research*, 40:11202–11212.

[2] Anink-Groenen, L. C. M., Maarleveld, T. R., Verschure, P. J., and Bruggeman, F. J. (2014). Mechanistic Stochastic Model of Histone Modification Pattern Formation. *Epigenetics & Chromatin*, 7:1–16.

[3] Annunziato, A. T. (2008). DNA Packaging: Nucleosomes and Chromatin. *Nature Education*, 1:26.

[4] Baxter, J., Oliver, A. W., and Schalbetter, S. A. (2019). Are SMC Complexes Loop Extruding Factos? Linking Theory With Fact. *BioEssays*, 41:1–10.

[5] Benites, L., Maehara, R., Vilca, F., and Marmolejo-Ramos, F. (2017). Finite Mixture of Birnbaum-Saunders Distributions Using the $k$-bumps Algorithm. pages 1–23.

[6] Berney, C. and Danuser, G. (2003). FRET or No FRET: A Quantitative Comparison. *Biophysical Journal*, 84:3992–4010.

[7] Boichenko, J. and Fierz, B. (2019). Chemical and Biophysical Methods to Explore Dynamic Mechanisms of Chromatin Silencing. *Current Opinion in Chemical Biology*, 51:1–10.

[8] Bonato, A., Brackley, C. A., Johnson, J., Michieletto, D., and Marenduzzo, D. (2019). Chromosome Compaction and Chromatin Stiffness Enhance Diffuse Loop Extrusion by Slip-Link Proteins. *arXiv*, page 1909.07108.

[9] Borman, S. (2006). The Expectation Maximization Algorithm. *University of Utah*, pages 1–9.

[10] Borrie, M. S., Campor, J. S., Joshi, H., and Gartenberg, M. R. (2017). Binding, Sliding, and Function of Cohesin During Trancriptional Activation. *PNAS*, 114:E1062 – E1071.

[11] Brackley, C., Johnson, J., Kelly, S., Cook, P. R., and Marenduzzo, D. (2016). Simulated Binding of Transcription Factors to Active and Inactive Regions Folds Human Chromosomes Into Loops, Rosettes and Topological Domains. *Nucleic Acids Research*, 44:3503–3512.

[12] Brackley, C., Johnson, J., Michieletto, D., Morozov, A., Nicodemi, M., Cook, P., and Marenduzzo, D. (2018). Extrusion Without a Motor: a New Take on the Loop Extrusion Model of Genome Organization. *Nucleus*, 9:95–103.

[13] Brackley, C. A., Johnson, J., Morozov, A. N., Nicodemi, M., Cook, P. R., and Michieletto, D. (2017). Nonequilibrium Chromosome Looping via Molecular Slip Links. *American Physical Society*, 19:138101–1 – 138101–5.

[14] Brandao, H. B., Paul, P., van den Berg, A. A., Rudner, D. Z., Wang, X., and Mirny, L. (2019). RNA Polymerases as Moving Barriers to Condensin Loop Extrusion. *Proceedings of the National Academy of Sciences of the United Stated of America*, 116:20489–20499.

[15] Cao, A.-M., Mivelaz, M., Boichenko, I., Bryan, L., Kubik, S., and Shore, D. (2019). Chromatin Remodeling Induced by the Invasion of Yeast Pioneer Transcription Factor *Rap1* revealed by single-molecule fret. *Biophysical journal*, 116:39A–40A.

[16] Corduneanu, A. and Bishop, C. (2001). Variational Bayesian Model Selection for Mixture Distributions. *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, 1:27–34.

[17] Crawford, F., Ho, L. S. T., and Suchard, M. A. (2018). Computational Methods for Birth-Death Processes. *Wiley Interdisciplinary Reviews. Computational Statistics*, 10:e1423.

[18] Dekker, J. and Misteli, T. (2015). Long-Range Chromatin Interactions. *Cold Spring Harbor Perspectives in Biology*, 7:1–24.

[19] Deniz, A. A., Laurence, T. A., Beligere, G. S., Dahan, M., Martin, A. B., Chemla, D. S., Dawson, P. E., Schultz, P. G., and Weiss, S. (2000). Single-Molecule Protein Folding: Diffusion Fluorescence Resonance Energy Transfer Studies of the Denaturation of Chymotrypsin Inhibitor 2. *Proceedings of the National Academy of Sciences*, 97:5179–5184.

[20] Eeftens, J., Bisht, S., Kerssemakers, J., Kshonsak, M., Haering, C., and Dekker, C. (2017). Real-Time Detection of Condensin-Driven DNA Compaction Reveals a Multistep Binding Mechanism. *EMBO Journal*, 36:3448–3457.

[21] Erdel, F. and Rippe, K. (2018). Formation of Chromatin Subcompartments by Phase Separation. *Biophysical Journal*, 114:2262–2270.

[22] Feller, W. (1971). An Introduction to Probability Theory and its Applications. *Wiley New York*.

[23] Fudenberg, G., Imakev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, 15:2038–2049.

[24] Fukui, K. and Uchiyama, S. (2007). Chromosome Protein Framework From Proteome Analysis of Isolated Human Metaphase Chromosomes. *The Chemical Record*, 7:230–237.

[25] Ganji, M., Shaltiel, I., Bisht, S., Kim, E., Kalichava, A., Haering, C., and Dekker, C. (2018). Real-Time Imaging of DNA Loop Extrusion by Condensin. *Science*, 360:102–105.

[26] Garcia-Luis, J., Lazar-Stefanita, L., Guitierrez-Escribano, P., Thierry, A., Cournac, A., Garcia, A., Gonzalez, S., Sanchez, M., Jarmuz, A., Montoya, A., Dore, M., Kramer, H., Karimi, M. M., Antequera, F., Koszul, R., and Aragon, L. (2019). FACT Mediates Cohesin Function on Chromatin. *Nature Structural & Molecular Biology*, 26:970–979.

[27] Gillespie, D. (1977). *The Journal of Physical Chemistry*, 81:2340–2361.

[28] Giorgetti, L., Lajoie, B. R., Carter, A. C., Attia, M., Zhan, Y., Chen, C. J., Kaplan, N., Chang, H. Y., Heard, E., and Dekker, J. (2017). Structural Organization of the Inactive X Chromosome in the Mouse. *Nature*, 535:575–579.

[29] Glynn, E. F., Megee, P. C., Yu, H.-G., Mistrot, C., Unal, E., Koshland, D. E., DeRisi, J. L., and Gerton, J. L. (2004). Genome-Wide Mapping of the Cohesin Complex in the Yeast *Saccharomyces cerevisiae*. *PLoS Biology*, 9:E259.

[30] Goloborodko, A., Marko, J. F., and Mirny, L. A. (2016). Chromosome Compaction by Active Loop Extrusion. *Biophysical Journal*, 110:2162–2168.

[31] Hansen, A. S., Cattoglio, C., Darzacq, X., and Tijan, R. (2018). Recent Evidence that TADs and Chromatin Loops are Dynamic Structures. *Nucleus*, 9:20–32.

[32] Holzmann, J., Politi, A. Z., Nagasaka, K., Hantsche-Grininger, M., Walther, N., Koch, B., Fuchs, J., Durnberger, G., Tang, W., Ladurner, R., Stocsits, R. R., Busslinger, G. A., Novak, B., Mechtler, K., Davidson, I. F., Ellenberg, J., and Peters, J.-M. (2019). Absolute Quantification of Cohesin, CTCF, and Their Regulators in Human Cells. *eLife*, 8:e46269.

[33] Horlbeck, M. A., Witkowsky, L. B., Guglielmi, B., Replogle, J. M., Gilbert, L. A., Villalta, J. E., Toringoe, S. E., Tjian, R., and Weissman, J. S. (2016). Nucleosomes Impede Cas9 Access to DNA in vivo and in vitro. *eLife*, 5:e12677.

[34] Huda, S., Yearwood, J., and Togneri, R. (2009). A Stochastic Version of Expectation Maximization Algorithm for Better Estimation of Hidden Markov Model. *Pattern Recognition Letters*, 30:1301–1309.

[35] Jeong, J., Le, T. T., and Kim, H. D. (2017). Single-molecule Fluorescence Studies on DNA Looping. *Methods*, 105:34–43.

[36] Jie Liu, D. M. C., Liang, S., and Shao, Z. (2008). Cell Cycle Dependent Nucleosome Occupancy at Cohesin Binding Sites in Yeast Chromosomes. *Genomics*, 91:274–280.

[37] Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J., and Segal, E. (2009). The DNA-Encoded Nucleosome Organization of a Eukaryotic Genome. *Nature*, 458:362–366.

[38] Karlin, S. and Taylor, H. (1975). A First Course in Stochastic Processes. *Academic Press.*

[39] Kilic, S., Felekyan, S., Doroshenko, O., Boichenko, I., Dimura, M., Vardanyan, H., Bryan, L. C., Seidel, C. A., and Fierz, B. (2018). Single-Molecule FRET Reveals Multicale Chromatin Dynamics Modulated by HP1$\alpha$. *Nature Communications*, 9:1–14.

[40] Kleinberg, J. (2002). Bursty and Hierarchical Structures in Streams. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101.

[41] Kong, M., Cutts, E., Pan, D., Beuron, F., Kaliyappan, T., Xue, C., Morris, E., Musacchio, A., Vannini, A., and Greene, E. C. (2019). Human Condensin I and II Drive Extensive ATP-Dependent Compaction of Nucleosome-Bound DNA. *bioRXiv*, 1:1–52.

[42] Kudalkar, E. M., Davis, T. N., and Ashbury, C. L. (2016). Single-Molecule Total Internal Reflection Fluorescence Microscopy. *Cold Spring Harbor Protocols*, pages 1–6.

[43] Kurkcupglu, O. and Bates, P. A. (2010). Mechanism of Cohesin Loading onto Chromosomes: A Conformational Dynamics Study. *Biophysical Journal*, 99:1212–1220.

[44] Leqieu, J., Scwartz, D. C., and de Pablo, J. J. (2017). In Silico Evidence for Sequence-Dependent Nucleosome Sliding. *PNAS*, 114:E9197–E9205.

[45] Lu, T., Volfson, D., Tsimring, L., and Hasty, J. (2004). Cellular Growth and Division in the Gillespie Algorithm. *Systems biology*, 1:121–128.

[46] Maarleveld, T., Olivier, B., and Bruggeman, F. (2018). Stochpy: A Comprehensive, User-Friendly Tool for Simulating Stochastic Biological Processes. *PLOS*, pages e79345–e79345.

[47] Millhauser, G. L., Salpeter, E. E., and Oswald, R. E. (1988). Diffusion Models of Ion-Channel Gating and the Origin of Power-Law Distributions From Single-Channel Recording. *Proceedings of the National Academy of Science of the United States of America*, 85:1503–1507.

[48] Moulton, L. and Halsey, N. (1995). A Mixture Model with Detection Limits for Regression Analyses of Antibody Response to Vaccine. *Biometrics*, 51:1570–1578.

[49] Munoz, S., Minamino, M., Casas-Delucchi, C. S., Patel, H., and Uhlmann, F. (2019). A Role for Chromatin Remodeling in Cohesin Loading Onto Chromosomes. *Molecular Cell*, 74:664–673.

[50] Naim, I. and Gildea, D. (2012). Convergence of the EM Algorithm for Gaussian Mixtures with Unbalanced Mixing Coefficients. *Proceedings of the 29th International Conference on Machine Learning*, pages 1655–1662.

[51] Nir, E., Michalet, X., Hamadani, K. M., Laurence, T. A., Neuhauser, D., Kovchegov, Y., and Weiss, S. (2006). Shot-Noise Limited Single-Molecule FRET Histograms: Comparison Between Theory and Experiments. *The Journal of Physical Chemistry B*, 110:22103–22124.

[52] Niu, L. and Lin, S. (2015). A Bayesian Mixture Model for Chromatin Interaction Data. *Statistical Applications in Genetics and Molecular Biology*, 14:53–64.

[53] Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., and Mirny, L. (2018). Chromatin Organization by an Interplay of Loop Extrusion and Compartmental Segregation. *Proceedings of the National Academy of Sciences of the United States of America*, 115:E6697–E6706.

[54] Parmar, J. J., Das, D., and Padinhateeri, R. (2016). Theoretical Estimates of Exposure Timescales of Protein Binding Sites on DNA Regulated by Nucleosome Kinetics. *Nucleic Acids Research*, 44:1630–1641.

[55] Preus, S., Noer, S. L., Hildebrandt, L. L., Gudnason, D., and Birkedal, V. (2015). iSMS: Single-Molecule FRET Miscroscopy Software. *Nature Methods*, 12:593–594.

[56] Racko, D., Benedetti, F., Dorier, J., and Stasiak, A. (2018). Transcription-Induced Supercoiling as the Driving Force of Chromatin Loop Extrusion During Formation of TADs in Interphase Chromosomes. *Nucleic Acids Research*, 46:1648–1660.

[57] Rhodes, J. D., Haarhuis, J. H., Grimm, J. B., Rowland, B. D., Lavis, L. D., and Nasmyth, K. A. (2017). Cohesin Can Remain Associated With Chromosomes During DNA Replication. *Cell Reports*, 20:2749–2755.

[58] Richardson, S. and Green, P. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Royal Statistical Society*, 59:731–792.

[59] Ridgway, P. and Almouzni, G. (2001). Chromatin Assembly and Organization. *Journal of Cell Science*, 114:2711–2712.

[60] Roy, R., Hohng, S., and Ha, T. (2013). A Practical Guide to Single Molecule FRET. *Nature Methods*, 5:507–516.

[61] Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnapan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., and Aiden, E. L. (2015). Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 112:E6456–E6465.

[62] Sasmal, D. K., Pulido, L., Kasal, S., and Huang, J. (2016). Single-Molecule Fluorescence Resonance Energy Transfer in Molecular Biology. *Nanoscale*, 8:9928–19944.

[63] Sathyajith, D. (2019). DNA Loop Extrusion Mechanisms. *News-medical*, https://www.news-medical.net/life-sciences/DNA-Loop-Extrusion-Mechanisms.aspx.

[64] Schwabe, A., Rybakova, K. N., and Bruggeman, F. J. (2012). Transcription Stochasticity of Complex Gene Regulation Models. *Biophysical Journal*, 103:1152–1161.

[65] Sielaff, H., Singh, D., Gruber, G., and Borsch, M. (2018). Analyzing Conformational Changes in Single FRET-Labeled $A_1$ Parts of Archeal $A_1A_0$-ATP Synthase. *Proceedings of SPIE*, 1:1–16.

[66] Slutsky, M. (2005). Protein-DNA Interaction, Random Walks and Polymer Statistics. *Massachusetts Institute of Technology*, pages 1–124.

[67] Stigler, J., Camdere, G. O., Koshland, D. E., and Greene, E. C. (2016). Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin. *Cell Reports*, 15:988–998.

[68] Tamaru, H. (2010). Confining Euchromatin/Heterochromatin Territory: *Jumonji* Crosses The Line. *Genes Development*, 24:1465–1478.

[69] van de Meent, J.-W., Bronson, J. E., Wiggins, C. H., and Jr, R. L. G. (2014). Empirical Bayes Methods Enable Advanced Population-Level Analyses of Single-Molecule FRET Experiments. *Biophysical Journal*, 106:1327–1337.

[70] Yamamoto, T. and Schiessel, H. (2017). Osmotic Mechanism of the Loop Extrusion Process. *Physical Review*, 96:030402.

[71] Yu, W., He, B., and Tan, K. (2017). Identifying Topologically Associating Domains and Subdomains by Gaussian Mixture Model and Proportion Test. *Nature*, 535:1–9.

[72] Yunis, J. (1981). Mid-Prophase Human Chromosomes. The Attainment of 2000 Bands. *Human Genetics*, 56:293–298.

[73] Zhang, Y., Ba, X. Z. Z., Liang, Z., Dring, E. W., Hu, H., Lou, J., Kyritsis, N., Zurita, J., Shamim, M. S., Aiden, A. P., Aiden, E. L., and Alt, F. W. (2019). The Fundamental Role of Chromatin Loop Extrusion in Physiological V(D)J Recombination. *Nature*, 573:600–604.

# A R Code for Gaussian Mixture Models

Listing A.1: R Code to determine two component Gaussian mixture models

```
df= as.data.frame(cbind(Overall.Cond= data$V1, Freq= data$V2))
df
df.freq= as.vector(rep(df$Overall.Cond, df$Freq))
hist(df.freq)
x=df.freq

mem = kmeans(x,2)$cluster
mu1 = mean(x[mem==1])
mu2 = mean(x[mem==2])
sigma1 = sd(x[mem==1])
sigma2 = sd(x[mem==2])
pi1 = sum(mem==1)/length(mem)
pi2 = sum(mem==2)/length(mem)

sum.finite = function(x) {
  sum(x[is.finite(x)])
}

Q = 0

Q[2]= sum.finite(log(pi1)+log(dnorm(x, mu1, sigma1)))
+ sum.finite(log(pi2)+log(dnorm(x, mu2, sigma2)))

k = 2

while (abs(Q[k]-Q[k-1])>=1e-8) {
  comp1 = pi1 * dnorm(x, mu1, sigma1)
  comp2 = pi2 * dnorm(x, mu2, sigma2)

  comp.sum = comp1 + comp2

  p1 = comp1/comp.sum
  p2 = comp2/comp.sum
  pi1 = sum.finite(p1) / length(x)
  pi2 = sum.finite(p2) / length(x)
  mu1 = sum.finite(p1 * x) / sum.finite(p1)
  mu2 = sum.finite(p2 * x) / sum.finite(p2)
  sigma1 = sqrt(sum.finite(p1 * (x−mu1)^2) / sum.finite(p1))
  sigma2 = sqrt(sum.finite(p2 * (x−mu2)^2) / sum.finite(p2))

  p1 = pi1
  p2 = pi2

  k <− k + 1
  Q[k] <− sum(log(comp.sum))
}

library(mixtools)
gm=normalmixEM(x,k=2, epsilon=1e−08)
gm$mu
gm$sigma
gm$lambda
gm$loglik
hist(x, prob=T, breaks=59, xlim=c(range(x)[1], range(x)[2]), main=’’,
ylim=c(0, 6), xlab=expression(’E’[’FRET’]), ylab=”Count␣density”, col=”cadetblue3”)
lines(density(x), col=”dark␣blue”, lwd=2)
x1 = seq(from=range(x)[1], to=range(x)[2], length.out=1000)
y = pi1 * dnorm(x1, mean=mu1, sd=sigma1) + pi2 * dnorm(x1, mean=mu2, sd=sigma2)
```

# B Python Code for Markovian Models

Listing B.1: Python code for random walk implementation

```
import random
import numpy as np
import matplotlib.pyplot as plt

prob = [0.25, 0.75]
start = 5
positions = [start]

r = np.random.random(2000)
downp = r < prob[0]
upp = r > prob[1]

for idownp, iupp in zip(downp, upp):
    down = idownp and positions[-1] > 1
    up = iupp and positions[-1] < 15
    positions.append(positions[-1] - down + up)

%matplotlib inline
plt.xlabel('Base␣Pair␣Number')
plt.ylabel('Binding␣Energy␣(U)')
plt.plot(positions)
plt.show()
```

In order to use the Stochpy modules, the Stochpy github (https://github.com/SystemsBioinformatics/stochpy) was cloned and saved onto the desktop. The Stochastic Simulation Algorithm (SSA) module was used to employ the Burst Model, and modified to comply with the parameters of this research. Specifically, the Immigration Death module was modified for the purpose of this simulation. Listing B.2 specifies the final commands used to initiate the simulations.

Listing B.2: Python code for Burst Model (example modified from Stochpy package)

```
import stochpy
import matplotlib.gridspec as gridspec
sim_end = 100
smod = stochpy.SSA()
smod.Model('Burstmodel.psc')
smod.ChangeParameter("kon",0.4)
smod.ChangeParameter("koff",0.4)
smod.ChangeParameter("ksyn", 4.8)
smod.ChangeParameter("kdeg", 2.4)
smod.DoStochSim(end=sim_end,mode='time',trajectories = 1)
gs = gridspec.GridSpec(4,1,width_ratios=[1],height_ratios=[0.3,1,0.3,1])
ax1 = stochpy.plt.subplot(gs[0])

smod.PlotSpeciesTimeSeries(species2plot ='ONstate',xlabel ='',ylabel ='')
stochpy.plt.legend('',frameon=False)
stochpy.plt.xlim([0,sim_end])
stochpy.plt.xticks([])
stochpy.plt.ylim([0,1.5])
stochpy.plt.yticks([])
stochpy.plt.text(-5.5,0.9,'ON')
stochpy.plt.text(-5.5,0,'OFF')
stochpy.plt.text(101,0.35,'A',fontsize = 14)

ax2 = stochpy.plt.subplot(gs[1])
smod.plot.ResetPlotnum()
smod.PlotSpeciesTimeSeries(species2plot ='mRNA',colors = ['#32CD32'])
stochpy.plt.xlim([0,sim_end])
stochpy.plt.legend('',frameon=False)
stochpy.plt.xticks([])
stochpy.plt.title('')
stochpy.plt.xlabel('')
stochpy.plt.ylabel('Loop_Extrusion')
stochpy.plt.yticks([0,5,10])
stochpy.plt.text(101,8,'B',fontsize = 14)
```

Listing B.3: Python code for LEF Exposure Timescale Equations

```
import numpy as np
import matplotlib.pyplot as plt
import math
from scipy import misc

L=200
f=0.04
s=5
o=0.04

def C_series(x, k):
    n = np.arange(k)
    X, N = np.meshgrid(x, n)
    val =1/(L-(X**N))*8
    return np.sum(val, axis=0)

val2=1/(f*(1-((1+s)*C_series(x, k))))
val3=1/((f+o)*(1-((1+s)*((f/(f+o))*C_series(x, k)))))

g=0.04
def E_series(x, k):
    n = np.arange(k)
    X, N = np.meshgrid(x, n)
    val =(1+((o**N)*X)+((g**N)*X))/((g**N)+(o**N))
    return np.sum(val, axis=0)
val4=1/((f+o)*(1-((1+s)*((f/(f+o))*C_series(x, k)*(s*o/(f+o))*E_series(x, k)))))

def D_series(x, k):
    n = np.arange(k)
    X, N = np.meshgrid(x, n)
    val =(1+f**N*X+g**N*X)/(f+g)
    return np.sum(val, axis=0)

def F_series(x, k):
    n = np.arange(k)
    X, N = np.meshgrid(x, n)
    val =((1+g**N*X+f**N*X)/(g+f+o))+((o/(g+f+o)*D_series(x, k)))
    return np.sum(val, axis=0)
val5=1/((f+o)*(1-((1+s)*((f/(f+o))*C_series(x, k)*(s*o/(f+o))*E_series(x, k)*(o/(o+f))*F_s

x0 = 0
xf= 500

x = np.linspace(x0, xf, 15)

for k in [147]:
    plt.plot(x, val2, label="case␣A")
    plt.plot(x, val3, label="case␣B")
    plt.plot(x, val4, label="case␣C")
    plt.plot(x, val5, label="case␣D")

plt.ylim([0, 120])
plt.xlim([0, 500])
#plt.yscale('log')
plt.xlabel('$l$␣(bp)')
plt.ylabel('$T_l$(s)')
plt.legend(loc="upper␣right")
plt.show()
```

The NucleosomeTool, a Stochpy plug in, was downloaded and cloned it into the Stochpy package, as obtained from the NucleosomeTool download (https://sourceforge.net/projects/stochpy/files/StochPyPlugins/NucleosomeTool/). The NucleosomeTool file within the plug in was then modified to include the impact of LEFs, as specified in Chapter 4. This ultimately affected the NucleosomeBuilder command, which was used to simulate the trajectory of the DNA folding procedure. The specific parameters were then included to comply with the specifications of this research. Listing B.4 describes the final commands used to execute the simulation.

Listing B.4: Python code for LEF Exposure Timescale in the presence of nucleosome example (no sliding)

```
import stochpy
bmod=stochpy.NucleosomeModelBuilder(Kslide=0)
bmod.SetInitiationSites({'M':[15,35]})
bmod.SetSliding(False)
bmod.SetRecruitment(False)
bmod.SetInitialState('U')
bmod.BuildModel('model_noslide')

smod=stochpy.NucleosomeSimulator(File='model_noslide.psc')
smod.DoMesoscopicStochSim(mode='time',end=100)
smod.PlotPatternTimeSeries()
smod.PlotPatternDistributions()
smod.PlotSpeciesTimeSeries()
smod.PlotSpeciesDistribution()
```

# C Python Code for Loop Visualization

Prior to loading the github, the grin package, and the mirnylib package were required. The looplib github (https://github.com/golobor/looplib/tree/master/looplib) was then also cloned and saved it onto the desktop. The looplib package was then installed and initalized by downloading the individual files in the folder. The simlef_onesided document was modified to include the parameters required to conduct this research. The sample_simulation_onesided document with Jupyter notebook was run, but simply by using the commands to create the simulations and the plots. The background documents were modified to comply with these research parameters, and to adapt the machine used to run the simulations. For example, the looptools document was modified as the numbers were not properly defined in the original, cloned github package. In order to yield the figure in the document, the points required to simulate the loop activity was performed with the simlef document, and then the loopviz file was required to actually visualize the simulation. Listing C.1 depicts the commands used to execute the simulation.

Listing C.1: Adapted from Goloborodko’s Github, looplib [23]

```
!easy_install grin
!pip install https://bitbucket.org/mirnylab/mirnylib/get/tip.tar.gz

import sys
import numpy as np
from mirnylib import h5dict
import pyximport; pyximport.install(
    setup_args={”include_dirs”:np.get_include()},
    reload_support=True)

from looplib import loopviz, looptools, simlef_onesided, simlef
import os, sys, glob, shelve, time

p = {}
p[’L’] = 50000
p[’N’] = 1000
p[’R_OFF’] = 1.0/692.0
```

```
p['R_EXTEND'] = float(5.0)
p['R_SHRINK'] = 0

p['R_SWITCH'] = p['R_OFF'] * 10

p['T_MAX_LIFETIMES'] = 100.0
p['T_MAX'] = p['T_MAX_LIFETIMES'] / p['R_OFF']
p['N_SNAPSHOTS'] = 200
p['PROCESS_NAME'] = b'proc'

l_sites, r_sites, leading_legs, ts = simlef_onesided.simulate(p)
l_sites, r_sites, ts = simlef.simulate(p)

ts4plot = [0, 100, 1, -10]
for t in ts4plot:
    loopviz.plot_lefs(
        l_sites=l_sites[t],
        r_sites=r_sites[t],
        L=p['L'],
        colors=[(223.0/255.0,90/255.0,73/255.0)] *500,
        site_width_bp = 10,
        max_height=200,
        plot_text=False,
        height_factor=2.0)
    plt.xlim(0,p['L']//10)
    plt.xticks([])
    plt.xlabel('chromosomal position')
```