RECONSTRUCTING DNA REPLICATION KINETICS FROM SMALL DNA FRAGMENTS

by

Haiyang Zhang
B. Sc. Physics, Lanzhou University, China, 1999

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN THE DEPARTMENT OF
PHYSICS

© Haiyang Zhang 2005 SIMON FRASER UNIVERSITY Summer, 2005

All rights reserved. This work may not be reproduced in whole or in part, by photocopy or other means, without permission of the author.
APPROVAL

Name: Haiyang Zhang

Degree: Master of Science

Title of Thesis: Reconstructing DNA replication kinetics from small DNA fragments

Examining Committee:
Dr. David Broun
Assistant Professor, Department of Physics (Chair)

Dr. John Bechhoefer, Senior Supervisor
Professor, Department of Physics

Dr. Martin Zuckermann, Supervisor
Adjunct Professor, Department of Physics

Dr. Eldon Emberly, Supervisor
Assistant Professor, Department of Physics

Dr. Michael Wortis, Internal Examiner
Professor Emeritus, Department of Physics

Date Approved: July 29, 2005
Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for circulation via the Library’s website.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author’s written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.

Bennett Library
Simon Fraser University
Burnaby, BC, Canada
Abstract

In higher organisms, DNA replicates simultaneously from many origins. Recent in vitro experiments have yielded large amounts of data on the state of replication of DNA fragments. From measurements of the time dependence of the average size of replicated and non-replicated domains, one can estimate the rate of initiation of DNA replication origins. One problem with such estimates is that, in the experiments, the DNA is broken up into small fragments, whose finite size can bias the measured averages. Here, I present a systematic way of accounting for this bias. In particular, I derive theoretical relationships between the original domain-length distributions and fragment-domain length distributions. I also derive unbiased average-domain-length estimators, which can yield accurate results even in cases where the replicated (or non-replicated) domains are larger than the average DNA fragment. Then I apply these estimators to previously obtained experimental data to extract replication kinetics parameters.
To My Girlfriend Xi Wang, My Parents Xingzhen Yang and Mingfu Zhang
Contents

Approval ii

Abstract iii

Dedication iv

Contents v

List of Tables viii

List of Figures ix

1 Introduction 1

1.1 DNA replication ........................................... 1

1.2 Experiments on DNA replication .......................... 2

1.3 Double-labeling Experiment .............................. 5

1.4 KJMA model .................................................. 6

1.5 About this thesis ........................................... 7

2 Background On the KJMA Model 9

2.1 Generalized KJMA Model ................................ 9

2.2 Replicated Fraction f(t) .................................. 10

2.3 Extracting DNA Replication Kinetics ................... 12

2.3.1 Ideal Case .............................................. 13

2.3.2 Application to the Double-labeling Experiment ..... 13

2.4 Finite-size Effect ........................................... 14
CONTENTS

3 Numerical Simulation Algorithm 16
  3.1 The Lattice-model Algorithm ....................................... 16
  3.2 The Double-list Algorithm ........................................... 17
  3.3 The Phantom-Nuclei Algorithm ....................................... 17

4 Finite-size Effect on Length Distribution 21
  4.1 Definition of Domains ............................................... 21
  4.2 Simplified Cut Model ............................................... 23
    4.2.1 Distribution of Interior Domains .............................. 24
    4.2.2 Distribution of Edge Domains ................................ 28
    4.2.3 Distribution of Oversize Domains .............................. 34
  4.3 General Cut Model .................................................. 36
    4.3.1 Distribution of Interior Domains .............................. 37
    4.3.2 Distribution of Edge Domains ................................ 40
    4.3.3 Distribution of Oversize Domains .............................. 42
  4.4 Reconstructing the Original Domain Length Distribution ....... 45
    4.4.1 Reconstructing the Original Eye and Hole Domain Length Dis-
        tributions ....................................................... 46
    4.4.2 Reconstruction Method for Eye-to-eye Domains ............... 48
  4.5 Summary .................................................................. 48

5 Average Domain Length Estimators 50
  5.1 Interior Average Domain Length Estimator ......................... 50
  5.2 Reconstructed Distribution Estimator ............................... 54
  5.3 Unbiased Estimator I ................................................. 57
  5.4 Unbiased Estimator II ............................................... 61
  5.5 Interior-Edge Estimator .............................................. 64
  5.6 Summary ................................................................ 67

6 Application to Experimental data 70
  6.1 Introduction ................................................................ 70
  6.2 Test of Uniform Distribution Assumption ......................... 70
    6.2.1 Test on experimental data ...................................... 72
CONTENTS

6.3 Application to experimental data ........................................ 77
6.4 Summary ........................................................................... 80

7 Conclusion ........................................................................... 83

Bibliography ........................................................................... 85
List of Tables

4.1 Definition of symbols for the simplified cut model ............... 24
4.2 Monte Carlo results for oversize domains in the simplified cut model .. 36
4.3 Definition of symbols for the general cut model. ................. 37
4.4 Major formulas for simplified cut and general cut models ........ 49

5.1 Minimum length ($L_{min}$) of fragments ......................... 67

6.1 Extracted nucleation rate $I(t)$ and fork velocity $v$ .............. 79
6.2 Extracted average start time. ..................................... 80
List of Figures

1.1 Electron micrograph of replication bubbles .................. 3
1.2 Schematic description of the double-labeling experiment of DNA .... 5
1.3 Image of DNA fragments at different stages of replication ........ 6

2.1 Mapping DNA replication onto the one-dimensional KJMA model ... 10
2.2 Kolmogorov's method to solve for $f(t)$ ..................... 11

3.1 Schematic description of the phantom-nuclei algorithm .......... 19
3.2 Comparison of simulation times for the three algorithms .......... 20

4.1 Definition of domains in a DNA molecule ..................... 22
4.2 Definition of subclasses of fragment domains .................. 22
4.3 Strategy of reconstructing domain statistics ................... 23
4.4 Simplified cut model with equal-sized fragments ............... 24
4.5 Length over which an interior domain can be contained in a fragment without being chopped .................................. 25
4.6 Method for making fake DNA molecule .......................... 26
4.7 Monte Carlo simulation for interior domain length distribution in the simplified cut model (Example I) ......................... 27
4.8 Monte Carlo simulation for interior domain length distribution in the simplified cut model (Example II) ....................... 28
4.9 Length of a domain smaller than the fragment size available to be cut by one cut position ................................ 29
4.10 Distribution of edge domains when a domain of size $X$ is cut into two pieces .................................................. 30
LIST OF FIGURES

4.11 Length for a domain bigger than fragment size to be cut ............. 31
4.12 Monte Carlo simulation for edge domain length distribution in the simplified cut model (Example I) ............................................. 33
4.13 Monte Carlo simulation for edge domain length distribution in the simplified cut model (Example II) ........................................... 34
4.14 Expected number of oversize domains from a big domain ........... 35
4.15 Monte Carlo simulation for interior domain length distribution in the general cut model (Example I) ............................................. 39
4.16 Monte Carlo simulation for interior domain length distribution in the general cut model (Example II) ............................................ 40
4.17 Monte Carlo simulation for edge domain length distribution in the general cut model (Example I) ............................................. 41
4.18 Monte Carlo simulation for edge domain length distribution in the general cut model (Example II) ............................................ 43
4.19 Symmetrical relationship between interior and oversize domains. .. 44
4.20 Monte Carlo simulation for oversize domain length distribution in the general cut model (Example I) ............................................. 45
4.21 Monte Carlo simulation for oversize domain length distribution in the general cut model (Example II) ............................................ 46
4.22 Comparison between the reconstructed and original domain length distributions .............................................................................. 47
4.23 Length for a interior eye-to-eye domain to be contained within a fragment 48

5.1 Performance of the interior estimator ........................................... 53
5.2 Performance of the reconstructed estimator ................................... 55
5.3 Illustration of the use of Unbiased Estimator I .............................. 59
5.4 Test of Unbiased Estimator I ...................................................... 60
5.5 Test of Unbiased Estimator II ..................................................... 63
5.6 Performance of the Interior-Edge estimator ................................... 66
5.7 Comparison of interior, reconstructed, and I-E estimators ............... 67
5.8 Comparison between the relative standard deviation of Unbiased Estimators I and II ................................................................. 68
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Distribution of lengths and fraction of combed DNA fragments</td>
<td>73</td>
</tr>
<tr>
<td>6.2</td>
<td>Fit to edge and interior domain-length distributions</td>
<td>74</td>
</tr>
<tr>
<td>6.3</td>
<td>$\langle X_e \rangle / \langle X_i \rangle$ vs. $f$</td>
<td>75</td>
</tr>
<tr>
<td>6.4</td>
<td>Comparison between $f_a$ and $f$</td>
<td>76</td>
</tr>
<tr>
<td>6.5</td>
<td>Average domain length vs. fraction replicated</td>
<td>78</td>
</tr>
<tr>
<td>6.6</td>
<td>The extracted $I(t)$ vs. $t$</td>
<td>79</td>
</tr>
<tr>
<td>6.7</td>
<td>Mean quantities vs. replication fraction</td>
<td>80</td>
</tr>
<tr>
<td>6.8</td>
<td>Mean quantities vs. replication fraction (interior estimator)</td>
<td>81</td>
</tr>
<tr>
<td>6.9</td>
<td>Extracted starting-time distribution</td>
<td>82</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 DNA replication

After Watson and Crick discovered the double-helix structure of DNA [1], recognition of the complementary base pairing of the DNA double-helix immediately gave rise to the notion that a template was involved in the transfer of information between generations. Since DNA serves as the genetic link between generations, the base sequence must be copied correctly during replication. DNA replication is highly regulated both temporally and spatially. Many questions about the structure and biochemistry of DNA replication follow.

Researchers have found several general features of DNA replication [2]:

1. DNA replication is semiconservative; that is, a newly formed duplex contains one old and one new DNA chain;

2. Most DNA replication is bidirectional, proceeding in both directions from a given starting site, with both strands being copied at each fork.\(^1\)

\(^1\)The cell has complex machinery in DNA replication. Since the DNA polymerase can read in only one direction (5' to 3'), and the DNA strands are antiparallel, the situations for the two strands are different. For one strand, the fork moves in the same direction as replication, and DNA can be replicated smoothly. For the other strand, DNA polymerase produces short "Okazaki fragments," which can be pieced together using a backstitching technique [2]. Note that the two replication forks are symmetric: each has a "leading strand," where polymerase proceeds in the forward direction
CHAPTER 1. INTRODUCTION

But given that a cell’s DNA is replicated once per cell cycle, when and where does initiation actually occur? How many replication origins are there along the genome? In prokaryotes, these issues are well-understood. For example, *E. coli* has one specific initiation site called *oriC* where replication starts. The replication bubble grows bidirectionally with a fork velocity of \( \approx 1000 \) b/s and terminates at another site called *terC*. Since the genome of *E. coli* contains 4 Mb, the replication of its DNA—the “S phase” of its cell cycle—takes about 25 min \( (\approx \frac{4 \text{Mb}}{2 \times 1000 \text{b/s}}) \).

But the situation for eukaryotes is much more complex. Eukaryotic genomes are usually much longer than prokaryotic genomes. In *Xenopus* embryos, for example, there is about 1000 times more DNA than in *E. coli*. Also, the speed at which replication occurs is about 100 times slower. (The speed is much slower because of proofreading and error correction, which reduce the replication rate.) Yet, in embryonic cells, the time for replication is about the same. There is thus a factor of roughly 100,000 to account for. Nature solves this problem by having a few hundred thousand distinct origins of replication Fig. 1.1.

1.2 Experiments on DNA replication

In the 1960s, autoradiography was first applied to visualize DNA fibers and measure the length of DNA molecules. In 1966, Huberman and Riggs measured DNA fibers from Chinese Hamster cells [3]. Their method was to incorporate \(^3\text{H}\)-thymidine (\(\beta\) decay, half-life 12.33 yr) into DNA molecules and then autoradiograph them. A track of the DNA molecule was generated on photographic emulsion by \(\beta\) rays given off by \(^3\text{H}\)-thymidine. After development, the length was measured under microscope. Soon, they found that this technique can be used to study DNA replication, to provide information about the directions and the velocity of replication fork movement and the locations of replication origins [4]. The grain density of the autoradiogram is proportional to the activity of \(^3\text{H}\)-thymidine incorporated in the experiment. By incorporating \(^3\text{H}\)-thymidine of different activities at different times in experiment, the newly replicated DNA strand marked by different activity was replicated at the and a “lagging strand.” Thus, it is natural to expect equal fork velocities at both forks around a replication origin.
corresponding time. One can find when the DNA strand was replicated by measuring the grain density of the autoradiogram. Thus, they could measure the direction and velocity of replication forks and could find the replication origins.

In the meantime, electron micrographs were also used to study DNA replication. Kriegstein and Hogness found evidence for bidirectionality of DNA replication by analyzing the fine structure of replication forks in 1974 [5]. Blumenthal analyzed spatio-temporal distribution of replication bubbles of *Drosophila* early embryos [6].

Figure 1.1: Electron micrograph showing multiple replication bubbles of *Drosophila melanogaster* DNA. Arrows indicate the position of replication bubbles, areas on the DNA that have been already duplicated. From [6], courtesy of D. S. Hogness.

Another method that has been used to study replication kinetics is 2D-gel electrophoresis. Electrophoresis is a method for the separation and identification of small charged particles. Two-dimensional gel electrophoresis separates proteins/DNA fragments first according to their isoelectric point and in the second dimension according to their molecular mass. Thus, it can separate DNA molecules containing branches (bubble or Y shapes—fragments contain a replication fork) from linear molecules.
CHAPTER 1. INTRODUCTION

The ratio of bubble-shaped molecules to Y-shaped molecules provides a qualitative estimate of origin efficiency [7, 8].

Fluorescence in situ hybridization (FISH) for visualization of nucleic acids developed as an alternative to costly and low-resolution older methods that used radio-labeled probes in 1969 [9]. In the last five years, the introduction of easily accessed methods for probe preparation and detection has led to widespread application of FISH [10]. Much more data have been gathered than previous experiments could possibly have obtained. In this technique, the DNA probe is either labeled directly by incorporation of a fluorescent-labeled nucleotide precursor, or indirectly by incorporation of a nucleotide containing a reporter molecule (such as biotin or digoxigenin) which after incorporation into the DNA is then bound to a fluorescently labeled affinity molecule [10, 11].

FISH can also be combined with other powerful techniques to achieve high-resolution mapping. New high-resolution FISH techniques include hybridization of probes to free chromatin, DNA fibers, or mechanically stretched chromosomes [12, 13]. These targets have improved the resolution of FISH to detect distances to a resolution of a few kilobases. They also have significantly sped up high-resolution physical mapping [12].

Another recently developed technique that has had impact on replication experiments is molecular combing. In the molecular-combing technique, DNA molecules attached at one end to a solid surface are extended and aligned by a receding air-water interface and left to dry on the surface [14]. Molecular combing has the advantage not only of producing large quantities of data but also of reproducibly stretching the DNA to a controlled extension [15]. Thus, there is an accurate mapping between distances measured on a digital image and lengths along the genome. In this thesis, we will discuss one particular molecular-combing experiment by Herrick et al. [16].

DNA microarrays are another new tool to study replication kinetics. An array consists of a large matrix of samples. It provides a medium for matching known and unknown DNA samples based on base-pairing rules and automates the process of identifying the unknowns. Thus, an ordered set of unique-sequence DNA probes can be used to monitor DNA replication [17–19]. Arrays have been used to generate large quantities of data on yeast replication [18]. Their unique advantage is that one can identify where on the genome a particular DNA fragment comes from. By contrast,
the combing techniques discussed above do not give such information.

Figure 1.2: Schematic description of the double-labeling experiment of DNA [16]. The steps of the experiment are as follows: (a) Before replication starts, one adds “red dye” (biotin-dUTP) into the solution of Xenopus egg extracts and sperm chromatin. (b) “Eyes” then grow while more replication origins fire. (c) At chosen time points, one adds “green dye” (dig-dUTP) and waits until the DNA is completely replicated. (d) One then stretches the replicated DNA molecules in solution onto a glass surface (“molecular combing”). In this last process, the long DNA molecules are broken up into relatively small pieces. From [20], courtesy of Suckjoon Jun.

1.3 Double-labeling Experiment

Here, we see how the double-labeling experiment of DNA works (Fig. 1.2). The experimental results obtained on the kinetics of DNA replication in vitro in the well-characterized Xenopus laevis cell-free system are based on this experimental method [16].

The combed molecules are visualized using fluorescent antibodies. Figure 1.3 is a picture of combed molecules as viewed in two-color microscopy. The red and green regions correspond to replicated and un-replicated DNA – the state of the DNA when the second (green) dye was added. Thus, these pictures are like snapshots of the replication states at the time that the second dye was added. The typical
Figure 1.3: DNA fragments at different stages of replication as they appear in an epifluorescence microscope [16]. The non-replicated domains (holes) have been fluorescently labeled. Images B-F show the progression (through S phase) from the early to the late stages of replication. Image A shows two-color labeling with both replicated and non-replicated domains indicated. Reprinted from Journal of Molecular Biology, Vol 300, pp. 1133-1142, Copyright (2000), with permission from Elsevier.

The length of combed DNA fragments is 100 kb, which is very short compared with the chromosome length of about 100 Mb. Since we have only rather small fragments of each chromosome, we can have problems in measuring the average lengths of replicated and non-replicated domains. These problems become severe once the domain sizes are comparable to the cut size. Qualitatively, we expect that such a situation will arise for the non-replicated domains near the start of S phase and for replicated domains near the end of S phase.

1.4 KJMA model

The goal of the *Xenopus* experiment was to deduce the number of DNA origins, their rate of firing throughout S phase, and the fork velocities. Jun et al. [21–23], have showed that one can infer these quantities from replication experiments using a
widely known stochastic model (the “Kolmogorov-Johnson-Mehl-Avrami,” or KJMA model [24–28]) that was originally introduced to study the kinetics of phase transitions. I will discuss the KJMA model in more detail in Chapter 2. The analysis of Jun et al. uses as an input the average replicated length (and non-replicated length) as a function of either time or replication fraction to extract these key quantities. But the measurement of these quantities is complicated when the underlying molecule is broken up into small fragments.

1.5 About this thesis

These “finite-size effects” discussed above are a limiting factor in the analysis of the DNA experiments. In this thesis, we will consider mathematical problems that are due to such finite-size effects: What is the relationship between what is observed on fragments and the original distribution? How can one estimate the original average domain length from distributions measured on the fragments? Our goal here is to devise a way of getting around these problems. Then, I will use the theory to reanalyze the experimental data of Herrick et al. [16, 23].

In Chapter 2, we review the generalized KJMA model theory. We focus on how to use experimental data to extract key parameters of DNA replication, such as the initiation rate $I(t)$ and the fork velocity $v$. Then, in Chapter 3, I introduce three numerical simulation algorithms for the KJMA model and its variants. They are the lattice model, double-list, and phantom-nuclei algorithms. In Chapter 4, we define three subclasses of fragment domains—interior, edge, and oversize—and give the theoretical relationships between the original and fragment domain-length distributions and test each relationship by Monte Carlo simulation. In Chapter 5, we compute the bias of the average domain length from interior average domain length estimator, which is the data processing method that has been used up to now. Next, we derive the reconstructed distribution average domain length estimators. This estimator is biased but has less bias than the interior estimator. We also derive two different unbiased estimators. They are unbiased but cannot be used in the analysis of present experimental data. Then, we derive the interior-edge estimator, which has less bias than either the interior or reconstructed estimator. Each estimator has been tested
against Monte Carlo simulation. In Chapter 6, we first test the assumption of uniform cuts along the DNA genome on experimental data from Herrick et al. [16], which was a crucial assumption underlying the work in Chapters 4 and 5. The test results show that this assumption is reasonable. Next, we use the interior-edge estimator to reanalyze the experimental data. Our new estimator does give more accurate values for the average domain lengths.
Chapter 2

Background On the KJMA Model

As we discussed in Chapter 1, a generalized version of the KJMA model was used to extract key quantities of DNA replication from the experiment of Herrick et al. [21–23]. In particular, they measured the initiation rate of replication origins $I(t)$ and fork velocity $v$. The KJMA model uses as an input the average replicated length (and non-replicated length) as a function of time or of replication fraction. The purpose of this thesis is to find a better way to compile experimental data to serve as input to the KJMA model. But, first, we review the KJMA model theory to give a feel for the uses to which the input quantities are put. This chapter will follow closely the work of Jun et al. [20–22]

2.1 Generalized KJMA Model

As we discussed in Chapter 1, several general features of DNA replication are shared by most species [23]:

1. Each replication origin is activated not more than once during the cell-division cycle;

2. DNA synthesis propagates at replication forks bidirectionally from each origin;

3. DNA synthesis stops when two newly replicated regions of DNA meet.
CHAPTER 2. BACKGROUND ON THE KJMA MODEL

These three processes—initiation, growth, and coalescence—are shared by a wide range of other physical processes, notably crystal growth. In crystal growth, there are three processes: (a) nucleation of solid domains (islands); (b) growth of existing islands; (c) coalescence, which occurs when two expanding islands merge [29].

Inspired by this analogy between DNA replication and crystal growth, the process of DNA replication can be understood as a kind of 1-D nucleation-growth problem. See Fig. 2.1. The “Kolmogorov-Johnson-Mehl-Avrami” (KJMA) model has been widely used in physics to analyze nucleation-growth problems. Jun et al. generalized the traditional KJMA model to apply it to the DNA replication problem [20–22].

The following two sections will discuss the KJMA theory for DNA replication and how the model may be used to extract the initiation rate \( I(t) \) and the replication fork velocity \( v \) from experimental data.

### 2.2 Replicated Fraction \( f(t) \)

The replicated fraction \( f(t) \) denotes the portion of DNA that has been replicated. To calculate \( f(t) \), we first calculate the fraction of DNA that has not replicated, \( S(t) \). Then \( f(t) = 1 - S(t) \). In another words, \( S(t) \) is the probability for an arbitrary point \( X \) at time \( t \) to remain uncovered, that is, for no initiation to have occurred at \( X \) and no other growing eyes to have covered \( X \) at time \( t \).

In Fig. 2.2(a), we see that the probability for an arbitrary point \( X \) at time \( t \) to
Figure 2.2: Kolmogorov’s method. (a) Spacetime diagram. In the small square box, the probability of initiation is \( I(x,t) \Delta x \Delta t \), where \( I(x,t) \) is the initiation rate. In order for the point \( X \) to remain uncovered by eyes, there should be no initiation in the shaded triangle in spacetime. (b) Kinetic curve for constant initiation rate \( I_0 \) Eq. 2.2b. From [21], Copyright © The American Physical Society (2005).

remain uncovered is

\[
S(t) = \lim_{\Delta x, \Delta t \to 0} \prod_{x,t \in \Delta} \left( 1 - I(x,t) \Delta x \Delta t \right) = \exp \left( - \int \int_{x,t \in \Delta} I(x,t) dx dt \right).
\]  

(2.1)

where \( I(x,t) \) is the initiation rate function. For a constant initiation rate \( I_0 \) and fork velocity \( v \), we have

\[
S(t) = e^{-I_0 vt^2},
\]

(2.2a)

\[
f(t) = 1 - e^{-I_0 vt^2},
\]

(2.2b)

which has a sigmoidal shape, see Fig. 2.2(b). For a variable initiation rate \( I(t) \), we have

\[
S(t) = e^{-2v \int_0^t g(t') dt'},
\]

(2.3a)

\[
f(t) = 1 - e^{-2v \int_0^t g(t') dt'},
\]

(2.3b)

where \( g(t') = \int_0^{t'} I(t'') dt'' \). For the domain density \( n(t) \), the number of domains per unit length at time \( t \), we have [21]
As we shall see in more detail in Chapter 4, we are interested in three different quantities: replicated domains, which are referred to as eyes; unreplicated domains, which are referred to as holes; and the distance between the centers of two adjacent eyes, which is defined as the eye-to-eye length.

Since all three types of domains exist in equal numbers, we have the following general relationships

\begin{equation}
C_{2i}(t) = \bar{\ell}_{i}(t) + jh(t),
\end{equation}

where \(\bar{\ell}_{i}, jh\) and \(li_{2i}\) are the average lengths of eyes, holes, and eye-to-eye domains [22].

One can also calculate their values explicitly:

\begin{equation}
\bar{\ell}_{i}(t) = \frac{1}{g(t)} \left[ e^{2v\int_{0}^{t} g(t')dt'} - 1 \right],
\end{equation}

\begin{equation}
\bar{\ell}_{h}(t) = \frac{1}{g(t)},
\end{equation}

\begin{equation}
\bar{\ell}_{i_{2i}}(t) = \frac{1}{g(t)} e^{2v\int_{0}^{t} g(t')dt'}.
\end{equation}

Note that \(\bar{\ell}_{i}(t) [\bar{\ell}_{h}(t)]\) are monotonically increasing (decreasing) functions of time.

### 2.3 Extracting DNA Replication Kinetics

In the previous section, we showed how, given an initiation rate \(I(t)\) and fork velocity \(v\), we can calculate quantities such as the replication fraction \(f(t)\). In experiments, the challenge is to do the reverse: from measurements on quantities such as the average hole size \(\bar{\ell}_{h}(t)\), how can one estimate \(I(t)\) and \(v\)?
2.3.1 Ideal Case

In the ideal case, origin initiation occurs at random sites along the genome, and the fork velocity \( v \) is constant. And, we know exactly the average hole length as a function of time, that is, we know \( \bar{h}(t) \).

From Eqs. 2.3 and 2.7, we can then directly invert the mean quantities to obtain the initiation rate \( I(t) \) and the fork velocity \( v \):

\[
I(t) = \frac{d}{dt} \frac{1}{\bar{h}(t)}
\]

\[
v = -\frac{1}{2} \ln S(t) \frac{1}{\bar{h}(t')} dt'.
\] (2.8)

This method is straightforward and has been tested by computer simulation [22].

2.3.2 Application to the Double-labeling Experiment

In the DNA replication experiment, the poor synchronization of the cycles of differing cells means that it is not possible to know the exact replication time for every fragment. We use the average domain length as a function of replicated fraction \( f \) as the input to extract \( I(t) \) and \( v \).

Once the data have been sorted by \( f \), we have \( \bar{l}_i \) and \( \bar{h}_i \) as functions of \( f \). Then we can extract the initiation frequency \( I \) as a function of \( f \). Using Eqs. 2.3-2.7, we can directly obtain

\[
\frac{I(f)}{2v} = \frac{1}{\bar{l}_i + \bar{h}_i} \frac{d}{df} \frac{1}{\bar{h}_i}.
\]

\[
2vt(f) = \int_0^f (\bar{l}_i + \bar{h}_i) df'.
\] (2.9)

In other words, we have a direct inversion \( I/2v \) vs. \( 2vt \) from the data. Note that both \( I \) and \( t \) are accompanied by the factor \( 2v \), which has to be determined as below.

From the double-labeling-experiment data, we can compile histograms of the distribution \( \rho(f, t_i) \) of replicated fractions \( f \) at time \( t_i \), where \( t_i \) is the time when the
second dye was added (Fig. 2.1). The distribution \( \rho(f, t_i) \) depends on the starting-time distribution \( \phi(\tau) \) and \( f(t) \), where \( \tau \) is the laboratory time that each DNA starts replicating, and \( t \) is the time since the onset of replication. Since

\[
\phi(\tau) d\tau = \rho(f(t'), t_i) \cdot df(t'),
\]

(2.10)

where \( t' = t_i - \tau \), we obtain

\[
\rho(f, t_i) = \phi(\tau) \times \left( \frac{df}{d\tau} \bigg|_{t=t_i-\tau} \right)^{-1}.
\]

(2.11)

Since the distribution of starting times is unknown, a Gaussian starting time distribution \( \phi(\tau) \) can be a good guess. We can fit all \( \rho(f, t_i) \)'s using three fitting parameters, the fork velocity \( v \), the average starting time \( \tau_0 \), and the starting time width \( \sigma_\tau \). Then, we follow next steps to obtain a coarse-grained \( v \) [22]: First, guess \( v \), trace \( \rho(f, t_i) \) back to find the starting time \( \phi(\tau) \) (Eq. 2.10), and fit the starting time by a Gaussian. Then, using the fitted Gaussian starting time function to generate the theoretical \( \rho(f, t_i) \) (Eq. 2.11), calculate \( \chi^2 \) between the experimental \( \rho(f, t_i) \) and the theoretical \( \rho(f, t_i) \). This is also a global fit, as the \( \chi^2 \) statistic is summed over data from all time points \( t_i \). We change \( v \) to \( v + \Delta v \) in the range of \( v_{\text{min}} \) and \( v_{\text{max}} \) and repeat the above steps, to find a well-defined minimum of the \( \chi^2(v) \). The corresponding \( v \) is our estimation of the fork velocity. Knowing \( v \), we can find \( I \) vs. \( t \) (Eq. 2.9).

### 2.4 Finite-size Effect

The above analysis is mainly due to Jun and Bechhoefer. There are still some unsolved problems in their model, which were also mentioned in their papers. In the double-labeling experiment, long-stranded DNA is broken up into small fragments. In this process, a lot of information is lost, which affects the estimation of all parameters. Such effects are called "finite-size effects," and they include mainly the following three aspects:

First, the distributions of eyes, holes, and eye-to-eyes in a fragment are different from the original distributions. An obvious example is that a domain size larger than...
 CHAPTER 2. BACKGROUND ON THE KJMA MODEL

$L$ cannot be observed on a fragment of size $L$. In this thesis, I try to understand the relationships between the fragment domain distributions and the original domain distributions in a theoretical way, and I will propose a data processing method to reconstruct the original domain length distribution.

Second, if one uses only information from fragment domains to compute averages, the average eye and hole sizes will be underestimated. In this thesis, I use the theoretical relation between the distribution of fragment domains and the original domain distribution to invent several average domain length estimators.

In the following several chapters, I will address these problems and present a systematic way of accounting and numerical approach for those problems. And, I will use an improved data-processing method to reanalyze experimental data from the work of Herrick et al. [16].
Chapter 3

Numerical Simulation Algorithm

As we have seen, analytic solutions to the KJMA model in realistic situations sometimes are very difficult to obtain. In many cases, computer simulation is the most direct and practical approach. Here, I will present three different simulation algorithms. These simulations begin with a length of unreplicated DNA and simulate into state of replication at time $t$ later. They were all implemented using the programming language of Igor Pro [30]. The work presented in this chapter has been published [21] and the discussion here closely follows that work. Bechhoefer contributed the method discussed in Sec. 3.1, Jun that of Sec. 3.2, and I came up with the method in Sec. 3.3.

3.1 The Lattice-model Algorithm

For one-dimensional KJMA processes, the most straightforward simulation method is to use an Ising-model-like lattice, where each lattice site is assigned either 1 or 0 (or $-1$, for the Ising model) representing eye and hole, respectively. The natural lattice size is $\Delta x = v\Delta t$, with $v$ the growth velocity. At each timestep $\Delta t$ of the simulation, every lattice site is examined. If 0, the site can be initiated by the standard Monte Carlo procedure; i.e., a random number is generated and compared with the initiation probability $I(t)\Delta x\Delta t$. If the random number is larger than the initiation probability, the lattice site switches from 0 to 1. Once initiation is done, the eyes grow by $\Delta x$, namely, by one lattice size at each end. For more details see [21].
CHAPTER 3. NUMERICAL SIMULATION ALGORITHM

3.2 The Double-list Algorithm

The lattice-model algorithm can work well, but it is slow and takes up too much memory. In particular, at the beginning and end of replication, the DNA replication state is modeled by long strings of zeros (or ones), which seems wasteful. In the double-list algorithm, one in effect keeps track of only the replication forks.

The basic idea here is to maintain two separate lists of lengths: \( \{i\} \) for eyes, \( \{h\} \) for holes. The second list \( \{h\} \) records the cumulative lengths of holes, while \( \{i\} \) lists the individual eye sizes. Using cumulative hole lengths simplifies the initiation routine dramatically. For instance, for times \( t \) ranging between \( \tau \) and \( \tau + \Delta \tau \), the average number of new initiations is \( \bar{N} = I(\tau) \Delta x \Delta t \). Since the initiation process is Poissonian, we obtain the actual number of new initiations \( N = p(\bar{N}) \) from the Poisson distribution \( p \). We then generate \( N \) random numbers between 0 and the total hole size, namely, the largest cumulative length of holes \( h_{max} \) (the last element of \( \{h\}\)). The list \( \{h\} \) is then updated by inserting the \( N \) generated numbers in their rank order. Accordingly, \( \{i\} \) is automatically updated by inserting zeros at the corresponding places. If \( \{h\} \) were to record the actual domain sizes as \( \{i\} \) does, the initiation routine would become much more complicated because the individual hole sizes would have to be taken into account as weighting factors in distributing the initiation positions along the template. For more details, see [21].

3.3 The Phantom-Nuclei Algorithm

The double-list algorithm is fast, but sometimes one needs the information from only several time points. In the double-list algorithm, one has to compute the state of the molecule explicitly at each simulation time step. I came up with the phantom-nuclei algorithm, which can directly obtain the statistical information at the desired time point without having to calculate the molecule’s state at intermediate times.

Figure 3.1 describes schematically the phantom-nuclei algorithm. The basic idea is to capitalize on the ability to specify when and where in the two-dimensional spacetime plane lie potential initiation sites, in advance of the actual simulation. Thus, in Fig. 3.1, the circles, which represent potential initiation sites, are laid down in the
first part of the simulation. One then uses an algorithm, described below, to determine which of these potential sites actually initiates (these are denoted by open circles) and which cannot fire because the system has already been transformed. (These “phantom nuclei” are denoted by closed circles.)

The principal advantage of the phantom-nuclei algorithm is that one can find the state of the system at a particular time \( t \) without having to calculate the system’s state at intermediate time steps. In the previous methods the replication state of the molecule had to be determined at every intermediate time step, which is very costly. If one is interested in the system state at a small number of time points, then the method can be significantly faster than the double-list algorithm. The filled triangles in Fig. 3.2 illustrate a hundred-fold improvement compared to the double-list algorithm (and a \( 10^5 \)-fold improvement relative to the lattice algorithm). On the other hand, if information is needed at every time step (or if the number of phantom nuclei is very large), the algorithm slows. The open triangles in Fig. 3.2 show a simulation where information is collected at each time step. The run time is comparable to the double-list algorithm. Because there is no sorting operation, a linear time scaling is maintained. The phantom-nuclei and double-list algorithms thus cross over at about two million sites.

The phantom-nuclei algorithm is performed as follows:

1. One generates the potential initiation sites in the two-dimensional spacetime plane. At each time, this is done in the same way as in the double-list algorithm. The only difference is that here, the number of sites at any time is calculated over the whole length of the system (regardless of its state of transformation). One uses two vectors to store the position and initiation time of every potential site.

2. One removes all initiation sites that are in positions that have already transformed before their initiation time. (They lie in the “shadows” in Fig. 3.1.) Because the growth velocity is known at each time, it is straightforward to implement this. Briefly, one first sorts the potential initiation sites by space. Then for each potential site (indexed by \( i \), with position \( x_i \) and nominal initiation time \( t_i \)), one calculates the position of the right-hand boundary \( r_i \) at the reference
Figure 3.1: Schematic description of the phantom-nuclei algorithm. The figure shows the distribution of potential initiation sites in the spacetime plane. The open circles denote sites that do initiate, while the “phantom” filled circles, lying in the “shadow” of the open circles, do not initiate.

3. One calculates the desired statistics concerning domain sizes at the reference time point. This time point is arbitrary. For the filled triangles of Fig. 3.2, it is the last time step of the transformation process, while in the open triangles, it occurs at the next time step of the double-list simulation. (For the latter case, the statistics were then repeatedly calculated at each time interval.)

In conclusion, we note that both the double-list and the phantom-nuclei algorithms are significant improvements on the more straightforward lattice algorithm. For simple initiation schemes, where it is possible to give a function $I(t)$ for the initiation sites, the phantom-nuclei algorithm will generally be preferable. For more complicated initiation schemes, where the initiation of sites is correlated with the activation of earlier sites, the double-list algorithm may be preferable. In this thesis, we present simulation results based on the “phantom-nuclei” algorithm.
CHAPTER 3. NUMERICAL SIMULATION ALGORITHM

Figure 3.2: Comparison of simulation times for the three algorithms discussed in the text. Circles are used for the lattice-model algorithm, squares for the double-list algorithm, and triangles for the phantom-nuclei algorithm. For each system size, the number of Monte Carlo realizations ranges from 5–20, and the lines connect the average simulation times. The double-list algorithm is two to three orders of magnitude faster than the lattice algorithm, while the phantom-nuclei algorithm ranges from three to five orders of magnitude faster, depending on the number of time points at which one records data. The filled triangles show the fastest case, with the molecule's replication state at only one time point requested, while the open triangles show the slowest case, where data are recorded at each intermediate time step.
Chapter 4

The Effect of Finite-size Fragments on Measured Length Distributions

Inspired by the problems of accounting for finite-size effects on the measurement of length distributions in the double-labeling DNA replication experiment, we will consider in this chapter the following mathematical problems: what is the relationship between the distributions measured on fragments and the original distributions on the unbroken DNA? Can we estimate without bias the original average domain length from the fragment domain distribution?

4.1 Definition of Domains

Generally, we will refer to a replicated domain as an eye. The terminology comes from the appearance of replicated domains in electron micrographs such as Fig. 1.1 (a). An unreplicated domain is defined as a hole, and the distance between the centers of two adjacent eyes is defined as the eye-to-eye distance. To be consistent with the two-domain replication pattern in DNA, the mathematical model for the theory and the computer simulation will have two different types of domains that are connected to each other alternately to form a long line. From the mathematical point of view, the two types of domains could be any kind of quantity. For convenience, I will use “eye” and “hole” to represent “domain #1” and “domain #2” below.

In the experiment, after replication is complete, a very long DNA molecule is
CHAPTER 4. FINITE-SIZE EFFECT ON LENGTH DISTRIBUTION

Figure 4.1: Definition of domains in a DNA molecule. The hollow flat black ellipses represent eyes, the black solid line represents holes.

broken up into many relatively small fragments. Some domains are cut off at an edge, while others fall entirely within a fragment. To describe the different types of a domain in a chopped fragment, we define three subclasses of domains (Fig. 4.2). A domain with both ends inside a fragment is defined to be an interior domain; a domain with one end inside the fragment and one chopped end on the edge of the fragment is an edge domain; and an oversize domain has two chopped ends covering the whole fragment. We call the three subclasses fragment domains.

Figure 4.2: Definition of subclasses of fragment domains. The vertical wedges denote places where the DNA molecule is cut.

The three subclasses of domains each contain part of the information from the original domain distribution. In this chapter, I will explore the relationships between the original domain length distribution and the fragment domain length distributions (see Fig. 4.3) and then give a method to estimate the original domain length distribution.
Figure 4.3: The distributions of interior domains, edge domains, and oversize domains all derive from the original distribution. Each contains a different part of information from the original distribution. In this chapter, we will derive relationships (solid vectors) between the original distribution and the three subclasses of fragment domains. In the next chapter, we shall try to find possible ways (dashed vectors) to use the information from fragment domains to estimate the original average domain length, which is the quantity used in the KJMA analysis.

4.2 Simplified Cut Model

In this chapter, we assume that the cut positions along the DNA are evenly distributed over the DNA molecule; that is, the probability that the DNA has been cut at one position is equal to that at anywhere else. I will explain why we make and test on experimental data this assumption of a uniform cut distribution in Chapter 6. With this basic assumption, we can derive the distributions of interior domains, edge domains, and oversize domains.

Let us begin with the simplest case, where the size of each fragment is equal (Fig. 4.4), and the cuts form a periodic one-dimensional lattice with spacing $L$. Let $X$ be the length of a domain (either eye or hole). There are a total of $N_e$ eye or hole
domains on a DNA molecule, whose total length of $S$. The original domain length distribution is $\rho(X)$. There are a total of $M$ cut positions in series covering the whole molecule, with $L$ the fragment size. We then have the relation $M \cdot L = S$. See Table 4.1.

![Figure 4.4: Simplified cut model with equal-sized fragments.](image)

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>Length of a domain</td>
</tr>
<tr>
<td>$N_t$</td>
<td>Total number of domains</td>
</tr>
<tr>
<td>$S$</td>
<td>Total length of all domains</td>
</tr>
<tr>
<td>$\rho(X)$</td>
<td>Original domain length distribution</td>
</tr>
<tr>
<td>$M$</td>
<td>Total number of fragments</td>
</tr>
<tr>
<td>$L$</td>
<td>Length of a fragment</td>
</tr>
</tbody>
</table>

Table 4.1: Definition of symbols for the simplified cut model. Note that a domain can be either a hole or an eye.

### 4.2.1 Distribution of Interior Domains

We denote the interior domain length distribution as $\rho_i(X)$ and the expected number of interior domains with size between $X$ and $X + \Delta X$ as $N_i(X)\Delta X$. Again, a domain can be either a hole or an eye.

First, the interior domain number $N_i(X)$ is obviously proportional to the original domain number $N_t \cdot \rho(X)$:

$$N_i(X) \propto N_t \cdot \rho(X). \quad (4.1)$$
Second, consider a domain of length $X$ and cut fragment of length $L$. If $X > L$, the cut fragment cannot contain this domain, and the interior domain number for $X > L$ is zero:

$$N_i(X > L) = 0. \quad (4.2)$$

If $X \leq L$, the probability of having an interior domain is proportional to the "free" length within this fragment, i.e. to $L - X$ (Fig. 4.5). This means that a big domain has a relatively small likelihood of being contained within a fragment; a little domain has relatively large likelihood; and a domain that is longer than the fragment cannot be contained at all. Since there are $M$ fragments, the interior domain number is proportional to the total probability for one domain to be within a fragment:

$$N_i(X \leq L) \propto \frac{M \cdot (L - X)}{S} = \frac{L - X}{S/M} = \frac{L - X}{L}. \quad (4.3)$$

Thus, combining Eqs. 4.1, 4.2, and 4.3, we find for the interior domain number:

$$N_i(X) = \begin{cases} N_i \cdot \frac{L - X}{L} \cdot \rho(X), & X \leq L, \\ 0, & X > L. \end{cases} \quad (4.4)$$

Since $N_i$ is a constant, from the interior domain number, we can directly get the interior domain length distribution:

$$\rho_i(X) = \begin{cases} K \cdot \frac{L - X}{L} \cdot \rho(X), & X \leq L, \\ 0, & X > L, \end{cases} \quad (4.5)$$
where \( K = 1 / \left( \int_{0}^{L} \frac{L - X}{L} \cdot \rho(X) dX \right) \) is the normalization factor.

Equation 4.4 should be valid for any distribution. Next, I will use Monte Carlo simulation to test this equation on a variety of particular distributions. The simulation is done as follows:

1. Sample lengths of eyes and holes from given distributions [31, 32];
2. Connect fake eye and hole length data alternatively into a long line (Fig. 4.6);
3. Make histograms of overall eye and hole lengths, to get the original eye and hole length distribution;
4. Cut the fake long line at a series of positions with identical spacing; then compile statistics for the distribution of interior, edge, and oversize domains within the fragments.

Figure 4.6: Method for generating fake data: connect eyes and holes alternatively to form a long line. Black and white blocks represent eye and hole domains.

I have tested many different kinds of distributions with this Monte Carlo simulation method. It turns out that the theoretical curves always fit the computer experiment data. Here I show two typical examples, one for a uniform domain distribution and another for an exponential domain distribution.

I. The original distribution is

\[
\rho(X) = \begin{cases} 
\frac{1}{100}, & 0 \leq X \leq 100, \\
0, & X > 100.
\end{cases}
\]

In this example, the fragment length \( L = 50 \). Using Eq. 4.4, we find the expression for the expected number of interior domains with lengths between \( X \) and \( X + \Delta X \),

\[
N_i(X) \cdot \Delta X = \begin{cases} 
N_t \cdot \frac{50 - X}{5000} \cdot \Delta X, & 0 \leq X \leq 50, \\
0, & X > 50,
\end{cases}
\]
where \( N_t = 10^4 \) is the total number of domains in the simulation and \( \Delta X = 2 \) is the histogram interval. The theoretical prediction and the Monte Carlo simulation data are plotted in Fig. 4.7.

![Figure 4.7: Comparison between the theoretical expected interior domain number and Monte Carlo simulation of the interior domain histogram. The original distribution is uniform. Here, and elsewhere below in this chapter, domain sizes are in arbitrary units.](image)

We can see from Fig. 4.7 that because of the finite-size effect, the interior domain length distribution becomes a first-order polynomial function, which is totally different from the original uniform distribution. We also can see that the theoretical curve fits the simulated data very well. Note that there are no adjustable parameters.

II. Here, the original distribution is an exponential distribution,

\[
\rho(X) = \frac{1}{100} \cdot e^{-\frac{X}{100}}.
\]

In this example, the fragment length is equal to 200. Using Eq. 4.4, we find

\[
N_i(X) \cdot \Delta X = \begin{cases} 
N_t \cdot \frac{200-X}{20000} \cdot e^{-\frac{X}{100}} \cdot \Delta X, & 0 \leq X \leq 200, \\
0, & X > 200,
\end{cases}
\]
where \( N_t = 6668 \) is the number of total domains and \( \Delta X = 2 \) is the histogram bin size. The theoretical prediction and the Monte Carlo simulation data are plotted in Fig. 4.8.

![Figure 4.8](image)

**Figure 4.8:** Comparison between the theoretical expected interior domain number and Monte Carlo simulation of the interior domain histogram. The original distribution is exponential.

We can see from Fig. 4.8 that the theoretical curve fits the simulated interior domain histogram data very well. Again, there are no free parameters.

In summary, from these Monte Carlo simulations and others not shown, Eq. 4.4 is consistent with the observed relationship between the original distribution and the resulting interior domain length distribution.

### 4.2.2 Distribution of Edge Domains

Let us denote the edge domain length distribution as \( \rho_e(y) \) and the expected number of edge domains of size between \( y \) and \( y + \Delta y \) as \( N_e(y)\Delta y \).

Consider a domain of length \( X \) and a fragment of length \( L \). The probability for one domain to be cut and to create an edge domain of length \( y \) results from considering two different cases.
CHAPTER 4. FINITE-SIZE EFFECT ON LENGTH DISTRIBUTION

Figure 4.9: Length of a domain smaller than the fragment size available to be cut by one cut position. Within a length $X$, the domain is cut. Thus the amount of the domain $X$ available to be cut is equal to its length $X$. This means a large domain has a relatively high probability of being cut, while a small domain has a relatively low probability of being cut.

First, if $X < L$, see Fig. 4.9. Since there are $M$ cut positions, the probability for this domain’s being cut, denoted as $P_{X \leq L}^{\text{cut}}$, is equal to

$$P_{X \leq L}^{\text{cut}}(X) = \frac{M \cdot X}{S} = \frac{X}{L}. \quad (4.6)$$

Equation 4.6 is part of the probability of having an edge domain of length $y$. It is easy to see that an edge domain of length $y$ can only come from a domain of length $X \geq y$. From Fig. 4.10, since we have two ways to create an edge domain of size $y$, (or since one cut makes two edge domains), the probability density for a domain of size $X < L$ to be cut to create an edge domain of size $y$, denoted as $P_{X \leq L}^{\text{edge}}$, is

$$P_{X \leq L}^{\text{edge}}(y) = \frac{2}{X}. \quad (4.7)$$

Multiplying the original domain number $N_t \cdot \rho(X)$ (the number of domains of size $X$), Eq. 4.6 (the probability that domain of size $X$ is cut) and Eq. 4.7 (the probability of creating an edge domain of size $y$ from domain $X \leq L$), we find the number of edge domains of length $y$ created from a domain of length $X \leq L$,

$$N_e(y, X \leq L) = N_t \cdot \rho(X) \cdot \frac{X}{L} \cdot \frac{2}{X} = N_t \cdot \rho(X) \cdot \frac{2}{L}. \quad (4.8)$$
CHAPTER 4. FINITE-SIZE EFFECT ON LENGTH DISTRIBUTION

Figure 4.10: Distribution of edge domains when a domain of size $X$ is cut into two pieces according to the uniform-cut assumption. When one cut is on one domain (eye or hole) of length $X$, we get two edge domains. The length of each edge domain depends on the cut position $y$ (with respect to the origin at one of the two ends of the domain). The lengths of the two edge domains are equal to $(X - y)$ and $y$, respectively. Based on the random-cut assumption, the probability of the cut position is evenly distributed over the whole DNA molecule. It is also evenly distributed along the domain. That means that $y$ is a uniform random number between 0 and $X$, and so is $(X - y)$. Then the distributions of the two edge domains are uniform between 0 and $X$.

If $X > L$, no matter where the domain is put, it will always be cut (Fig. 4.11). That is, the probability of this domain's being cut is equal to one

$$P_{X>L}^{\text{cut}}(X) = 1.$$  

(4.9)

From Fig. 4.11, a domain of length $X > L$ can create edge domains whose length is between 0 and $L$. Following the same idea as for small domains, the edge domain length is distributed uniformly between 0 and $L$. The probability density for a domain of size $X > L$ being cut to create an edge domain of size $y$, denoted as $P_{X>L}^{\text{edge}}$, is
CHAPTER 4. FINITE-SIZE EFFECT ON LENGTH DISTRIBUTION

Figure 4.11: Length for a domain bigger than fragment size to be cut. We can see that no matter where one puts the big domain, there will be at least one cut on this domain and two edge domains created that are uniformly distributed between 0 and L.

\[ P_{X>\ell}^{edge}(y) = \frac{2}{L}. \]  

(4.10)

Combining the original domain number \( N_t \cdot \rho(X) \) (the number of domains of size \( X \)), Eq. 4.9 (the probability that a domain of size \( X \) is cut) and Eq. 4.10 (the probability of having an edge domain of size \( y \) from domain \( X > \ell \)), we find the number of edge domains of length \( y \) created from a domain of length \( X > \ell \),

\[ N_e(y, X > \ell) = N_t \cdot \rho(X) \cdot \frac{2}{L}. \]  

(4.11)

Since Eqs. 4.8 and 4.11 are the same, we write them in one equation

\[ N_e(y, X) = N_t \cdot \rho(X) \cdot \frac{2}{L}. \]  

(4.12)

The overall probability of an edge domain \( y \) can be obtained by integrating Eq. 4.12 over \( X \). Obviously, only a domain that is bigger than the edge domain length can create this edge domain. Thus,

\[ N_e(y) = N_t \cdot \frac{2}{L} \cdot \int_{y}^{+\infty} \rho(X) \cdot dX. \]  

(4.13)

Since \( N_t \) is a constant, from the edge domain number \( N_e(y) \), we can directly find the edge domain length distribution
CHAPTER 4. FINITE-SIZE EFFECT ON LENGTH DISTRIBUTION

$$\rho_e(y) = K \cdot \frac{2}{L} \cdot \int_y^{+\infty} \rho(X) \cdot dX. \quad (4.14)$$

where $K$ is the normalization factor. To evaluate $K$, we first integrate by parts, noting that edge domain length can only be between 0 and $L$:

$$\int_0^L \left( \int_y^{+\infty} \rho(X) \cdot dX \right) dy = y \cdot \int_0^{+\infty} \rho(X) \cdot dX \bigg|_0^L + \int_0^L X \cdot \rho(X) dX$$

$$= \int_0^{+\infty} \rho(X) dX + \int_0^L X \cdot \rho(X) dX \quad (4.15)$$

We can write $K$ as

$$K = \frac{1}{\left( \int_L^{+\infty} 2 \cdot \rho(X) dX + \int_0^L \frac{2X}{L} \cdot \rho(X) dX \right)} \quad (4.16)$$

Next, these theoretical relations will be tested against Monte Carlo simulations. I have tested many random distributions, and it turns out that the theoretical curves always fit well the computer simulation data. Here, I show two typical examples, one for a sine domain length distribution and another for an exponential domain length distribution.

I. The original distribution is

$$\rho(X) = \begin{cases} \frac{\pi}{200} \cdot \sin\left(\frac{\pi}{200} \cdot X\right), & 0 \leq X \leq 200, \\ 0, & X > 200. \end{cases}$$

In the Monte Carlo simulation, the fragment length is equal to 200. Using Eq. 4.13, we find the expected number of edge domains with length between $y$ and $y + \Delta y$ is

$$N_e(y) \cdot \Delta y = \begin{cases} N_t \cdot \frac{1}{100} \cdot (1 + \cos\left(\frac{\pi}{200} \cdot y\right)) \cdot \Delta y, & 0 \leq X \leq 200, \\ 0, & X > 200. \end{cases}$$

where $N_t = 6668$ is the number of total domains and $\Delta y = 2$ is the histogram bin size. The theoretical prediction and the Monte Carlo simulation results are plotted in Fig. 4.12.
We can see from Fig. 4.12 that the edge domain length distribution becomes a cosine function, which is totally different from the original sine distribution. We also can see that the theoretical curve fits the simulated data very well.

II. The original distribution is an exponential function,

\[
\rho(X) = \frac{1}{100} \cdot e^{-\frac{x}{100}}.
\]

In this example, the fragment length \( L \) is equal to 300. Using Eq. 4.13, we find that the expected number of edge domains with length between \( y \) and \( y + \Delta y \) is

\[
N_e(y) \cdot \Delta y = N_t \cdot \frac{1}{150} \cdot e^{-\frac{x}{100}} \cdot \Delta y,
\]

where \( N_t = 6668 \) is the total number of domains obtained from experimental data, \( \Delta y = 2 \) is the histogram bin size. The theoretical prediction and the Monte Carlo simulation results are plotted in Fig. 4.13.

We can see from Fig. 4.13, that the theoretical curve fits the experiment edge domain histogram data very well.
CHAPTER 4.  FINITE-SIZE EFFECT ON LENGTH DISTRIBUTION

Figure 4.13: Comparison between the theoretical edge domain length distribution and the Monte Carlo simulated edge domain histogram. The original distribution is an exponential distribution.

In summary, these Monte Carlo simulations are all consistent with the theoretical equation (Eq. 4.13) of the relationship between the original distribution and the edge domain length distribution.

4.2.3  Distribution of Oversize Domains

In the simplified cut model discussed so far, there is only one identical fragment length; that is, there is only one identical oversize domain length. The issue for oversize domains is thus to calculate their expected number, given the original domain length distribution and cut size.

Let us denote the number of oversize domains as $N_o$. Since the length of oversize domains in the simplified cut model is equal to $L$, the distribution of oversize domains is

$$\rho_o(y) = \delta(y - L),$$

(4.17)

where $\delta(X)$ is the Dirac delta function.
The series of cut positions constitutes a periodic one-dimension lattice. If we can compute the expected number of oversize domains in one period, we can also compute their overall expected number. See an example shown in Fig. 4.14.

Figure 4.14: Expected number of oversize domains from a big domain of size $X$ between $2L$ and $3L$. This domain scans over a period length $L$, from its original position (solid line ellipse) to the end of the period (dashed line ellipse) which is also the start position for the next period. This domain can be cut to create one or two oversize domains, depending on the position of the domain. The length of finding one oversize domain in one period is $3L - X$, and the length for obtaining two oversize domains is $X - 2L$. Since the period length is $L$, the overall probability for getting one and two oversize domains is $\frac{3L - X}{L}$ and $\frac{X - 2L}{L}$ respectively.

For convenience, we start from an example to show how to find the expected number of oversize domains created by a domain whose size $X$ is between $2L$ and $3L$ (Fig. 4.14). Then we find the expected number of oversize domains

$$N_X = 1 \cdot \frac{3L - X}{L} + 2 \cdot \frac{X - 2L}{L} = \frac{X - L}{L}. \quad (4.18)$$

In fact, this equation is general. If the length of the domain is between $nL$ and $(n + 1)L$, $n = 1, 2, 3...$, the probability for having $(n - 1)$ and $n$ oversize domains is $\frac{(n+1)L - X}{L}$ and $\frac{X - nL}{L}$, respectively, and the expression of expected number of oversize domains is

$$N_X = (n - 1) \cdot \frac{(n+1)L - X}{L} + n \cdot \frac{X - nL}{L} = \frac{X - L}{L}. \quad (4.19)$$
Combining the original domain number $N_t \cdot \rho(X)$ and Eq. 4.19, we find the expected number of oversize domains for domains of length $X > L$,

$$N_o(X) = N_t \cdot \rho(X) \cdot \frac{X - L}{L}.$$  (4.20)

Integrating Eq. 4.20 over $X$, we get the overall expected oversize domain number. Since only domains with size bigger than the cut size can create oversize domains, the integration range is from $L$ to $+\infty$. We then have

$$N_o = N_t \cdot \int_{L}^{+\infty} \frac{X - L}{L} \cdot \rho(X) \cdot dX.$$  (4.21)

Equation 4.21 has been tested in Monte Carlo simulations. Here, I chose three typical examples which all show that the theory fits the Monte Carlo simulation data very well.

<table>
<thead>
<tr>
<th>Original Distribution</th>
<th>Cut Size</th>
<th>Computer Exp Data</th>
<th>Theoretical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform distribution (0 → 200)</td>
<td>100</td>
<td>$\overline{N}_o = 1661$</td>
<td>$E(N_o) = 1667$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sigma_o = 5$</td>
<td></td>
</tr>
<tr>
<td>Linear distribution (0 → 200)</td>
<td>100</td>
<td>$\overline{N}_o = 713$</td>
<td>$E(N_o) = 714$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sigma_o = 6$</td>
<td></td>
</tr>
<tr>
<td>Exponential (x ≥ 0)</td>
<td>200</td>
<td>$\overline{N}_o = 951$</td>
<td>$E(N_o) = 967$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sigma_o = 6$</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Comparison between the theoretical expected oversize domain number and computer simulation data. Here, $\overline{N}_o$ is the average number of oversize domains from 30 experiments, $\sigma_o$ is the standard deviation of the mean, and $E(N_o)$ is the theoretical expected number of oversize domains, based on Eq. 4.21.

### 4.3 General Cut Model

In DNA replication experiments, the DNA fragment size is typically not constant but follows some kind of distribution. In the experiment of Herrick et al. [23], the measured length distribution of fragments is approximately log-normal see Fig. 6.1 (a).
Building on the simplified cut model, I will discuss the general cut model in this section. The basic idea to find the expression for fragment domain length distribution is, first, find the joint distribution of domain $X$ and fragment $L$ and then, second, integrate over $L$ to obtain the overall fragment domain length distribution.

Here, I define symbols that will be used in the general cut model. Let $X$ be the length of a domain; let there be a total of $N_t$ eye or hole domains on a loop with total length $S$. The original distribution is $\rho(X)$. There are a total of $M$ cut positions in series covering the whole line with $L$ the fragment size, fragment length distribution is $\rho_f(L)$. The average fragment length is equal to $\bar{L}$, that is, $M \cdot \bar{L} = S$. See Table 4.3.

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>Length of a domain</td>
</tr>
<tr>
<td>$N_t$</td>
<td>Total number of domains</td>
</tr>
<tr>
<td>$S$</td>
<td>Total length of all domains</td>
</tr>
<tr>
<td>$\rho(X)$</td>
<td>Original domain length distribution</td>
</tr>
<tr>
<td>$\rho_f(X)$</td>
<td>Original fragment length distribution</td>
</tr>
<tr>
<td>$M$</td>
<td>Total number of fragments</td>
</tr>
<tr>
<td>$L$</td>
<td>Length of a fragment</td>
</tr>
<tr>
<td>$\bar{L}$</td>
<td>Average length of all fragments</td>
</tr>
</tbody>
</table>

Table 4.3: Definition of symbols for the general cut model.

### 4.3.1 Distribution of Interior Domains

As before, the interior domain length distribution is $\rho_i(X)$, and the expected number of interior domains of length between $X$ and $X + \Delta X$ in fragments of length between $L$ and $L + \Delta L$ is $N_i(X, L)\Delta X\Delta L$. Referring to Eq. 4.4, we can write

$$N_i(X, L) = N_t \cdot \frac{L - X}{L} \cdot \rho(X),$$

where $N_i$ is the number of domains that can be covered by fragments of length $L$. If the total number of domains is $N_t$, then $N_i$ is proportional to the total length of all fragments of length $L$. Since the total number of fragments is $M$, the total length of fragments with size $L$ is equal to $M \cdot \rho_f(L) \cdot L$. Then we can write $N_i$ as
Combining Eqs. 4.22 and 4.23, we find the number of interior domains,

\[ N_i = N_t \cdot \frac{M \cdot \rho_f(L) \cdot L}{S} = N_t \cdot \frac{\rho_f(L) \cdot L}{S/M} = N_t \cdot \frac{\rho_f(L) \cdot L}{L}. \]  

(4.23)

Integrating Eq. 4.24 over \( L \), we can find the overall distribution of interior domains. Since only a fragment whose length is longer than the domain \( X \) can contain this domain, the integration range will be \( (X \to +\infty) \), and thus,

\[ N_i(X, L) = \frac{N_t}{L} \cdot \rho(X) \cdot (L - X) \cdot \rho_f(L). \]  

(4.24)

Integrating Eq. 4.24 over \( L \), we can find the overall distribution of interior domains. Since \( N_t \) and \( \overline{L} \) are constant, we can write the distribution of interior domains as

\[ \rho_i(X) = K \cdot \rho(X) \cdot \int_X^\infty (L - X) \cdot \rho_f(L) dL. \]  

(4.25)

where \( K \) is the normalization factor.

Equation 4.25 has been tested in many Monte Carlo simulations. I show here two typical examples.

I. The domain length and fragment distributions are both uniform distributions. Explicitly, we have, for the original domain length distribution,

\[ \rho(X) = \begin{cases} \frac{1}{120}, & 0 \leq X \leq 120, \\ 0, & X > 120, \end{cases} \]  

(4.27)

and for the fragment distribution,

\[ \rho_f(X) = \begin{cases} \frac{1}{120}, & 0 \leq L \leq 120, \\ 0, & L > 120. \end{cases} \]  

(4.28)

Referring to Eq. 4.25, we find for the interior domain number,

\[ N_i(X) = \begin{cases} \frac{N_t}{60} \cdot \left( \frac{1}{2} \cdot \frac{1}{120} \cdot X + \frac{1}{2 \cdot 120^2} \cdot X^2 \right), & 0 \leq X \leq 120, \\ 0, & X > 120, \end{cases} \]  

(4.29)
Figure 4.15: Theoretical number of interior domains for the general cut model and Monte Carlo simulation results. The original domain and the fragment length distributions are both uniform distributions.

where \( N_t = 8334 \) is the total number of domains. The test result is shown in Fig. 4.15.

II. The original domain length distribution is an exponential function,

\[
\rho(X) = \frac{1}{100} \cdot e^{-\frac{X}{100}}
\]  \hspace{1cm} (4.30)

In the DNA replication experiment, the fragment length has a lower limit. To simulate this, we set the fragment distribution to be a uniform distribution starting at 80,

\[
\rho_f(X) = \begin{cases} 
0, & L < 80, \\
\frac{1}{120}, & 80 \leq L \leq 200, \\
0, & L > 200.
\end{cases}
\]  \hspace{1cm} (4.31)

Referring to Eq. 4.25, we find for the distribution of interior domains,

\[
N_i(X) = \begin{cases} 
\frac{N_t}{140} \cdot \frac{1}{100} \cdot e^{-x/100} \cdot (140 - X), & X < 80, \\
\frac{N_t}{140} \cdot \frac{1}{100} \cdot e^{-x/100} \cdot \left( \frac{200^2}{240} - \frac{200X}{120} + \frac{X^2}{240} \right), & 80 \leq X \leq 200, \\
0, & X > 200.
\end{cases}
\]
4.3.2 Distribution of Edge Domains

We now turn our attention to the edge domains. Let the edge domain length distribution be denoted as \( \rho_e(y) \). The expected number of edge domains of length between \( y \) and \( y + \Delta y \) in fragments of size between \( L \) and \( L + \Delta L \) is \( N_e(y, L) \Delta y \Delta L \). Referring to Eq. 4.14, we can write the number of edge domains of size \( y \) in a specified fragment \( L \) as

\[
N_e(y, L) = N_t \cdot \frac{2}{L} \cdot \int_y^{+\infty} \rho(X) dX.
\]  

(4.32)
Referring to Eq. 4.23, we then have

$$N_e(y, L) = \frac{N_t}{L} \cdot \rho_f(L) \cdot L \cdot \frac{2}{L} \cdot \int_y^{+\infty} \rho(X) dX. \quad (4.33)$$

Integrating this equation over fragment length $L$, we can get the overall number of edge domains. Since only a fragment longer than the edge domain length $y$ can contain this edge domain, the integration range is $(y \to +\infty)$, and

$$N_e(y) = \frac{2 \cdot N_t}{L} \cdot \int_y^{+\infty} \rho_f(L) dL \cdot \int_y^{+\infty} \rho(X) dX. \quad (4.34)$$

For $N_t$ and $\bar{L}$ constant, we then find

$$\rho_e(y) = K \cdot \int_y^{+\infty} \rho_f(L) dL \cdot \int_y^{+\infty} \rho(X) dX, \quad (4.35)$$

where $K$ is the normalization factor.

Equation 4.34 has been tested in many Monte Carlo simulations. I show two typical examples here.

![Figure 4.17: Theoretical number of edge domains for the general-cut model and results from Monte Carlo simulation. The original domain and the fragment length distributions are both uniform.](image)
I. We use the same domain and fragment distribution as in Eqs. 4.27 and 4.27. Referring to Eq. 4.34, we predict

\[ N_e(y) = \begin{cases} \frac{N_t}{15} \cdot \left( \frac{1}{2} - \frac{1}{120} \cdot y + \frac{1}{2 \cdot 120^2} \cdot y^2 \right), & 0 \leq y \leq 120, \\ 0, & y > 120. \end{cases} \]

We can see that the edge domain number is similar to the equation for the interior domain number, except that the total number of edge domains is four times greater than the number of interior domains. Intuitively, some of the large domains have been cut into smaller edge domains.

II. This example is the same as the second example of the last subsection; the domain and fragment distribution follow Eqs. 4.30 and 4.31. Referring to Eq. 4.34, we predict

\[ N_e(y) = \begin{cases} \frac{2N_t}{140} \cdot e^{-y/100}, & y < 80, \\ \frac{2N_t}{140} \cdot e^{-y/100} \cdot \frac{200-y}{120}, & 0 \leq y \leq 120, \\ 0, & y > 120. \end{cases} \]

There is a slope discontinuity at \( y = 80 \), which is shown clearly in the edge domain histogram data in Fig. 4.18.

In summary, these Monte Carlo simulations are consistent with the theoretical equations for the general cut model for the relationship between the original domain length distribution and the edge domain length distribution.

### 4.3.3 Distribution of Oversize Domains

In the simplified cut model, fragments all have the same length; thus the length of oversize domains is also fixed number. In the general cut model, fragments follow a distribution and thus oversize domains also follow a (different) distribution. We use the same notation as before and denote the oversize domain length distribution as \( \rho_o(y) \); the expected number of oversize domains of length between \( y \) and \( y + \Delta y \) in fragments of size between \( L \) and \( L + \Delta L \) is \( N_o(y, L) \Delta y \Delta L \).

We refer to Eq. 4.21 for the number of oversize domains for fragments of size \( L \) and note that the length of the oversize domain is equal to the length of the fragment. We use \( y \) to replace \( L \) in Eq. 4.21 and find
Figure 4.18: Theoretical number of edge domains for the general cut model and results from Monte Carlo simulation. The original domain length distribution is exponential and the fragment length distribution is uniform. The arrow marks the slope discontinuity in the edge domain distribution.

\[ N_o(y) = N_t \cdot \int_y^{+\infty} \frac{X - y}{y} \cdot \rho(X) \cdot dX. \]  \hspace{1cm} (4.36)

Referring to Eq. 4.23 and noting that we also use \( y \) to replace \( L \), we can write Eq. 4.36 as

\[ N_o(y) = N_t \cdot \frac{\rho_f(y)}{\bar{L}} \cdot \int_y^{+\infty} \frac{X - y}{y} \cdot \rho(X) \cdot dX \]

\[ = \frac{N_t}{\bar{L}} \cdot \rho_f(y) \cdot \int_y^{+\infty} (X - y) \cdot \rho(X) \cdot dX \]  \hspace{1cm} (4.37)

Since \( N_t \) and \( \bar{L} \) are constant, we find the oversize domain length distribution

\[ \rho_o(y) = K \cdot \rho_f(y) \cdot \int_y^{+\infty} (X - y) \cdot \rho(X) \cdot dX, \]  \hspace{1cm} (4.38)

where \( K \) is the normalization factor.
CHAPTER 4. FINITE-SIZE EFFECT ON LENGTH DISTRIBUTION

Comparing Eqs. 4.37 and 4.25, we find that if we switch the symbols of domain and fragment size in Eqs. 4.37 and 4.25, they are the same. That is there is a symmetric relationship between interior and oversize domains. If we view domains as cut fragments, then the oversize domain is the “interior domain,” while the interior domain is the “oversize domain”. See Fig. 4.19.

Next, we test the theory of oversize domain length distributions by Monte Carlo simulation.

1. This example is the same as before, with domain and fragment length distributions that follow Eqs. 4.27 and 4.28. Using Eq. 4.37, we find the theoretical number of oversize domains in this simulation

\[
N_o(y) = \begin{cases} 
\frac{N_t}{60} \cdot \left( \frac{1}{2} - \frac{1}{120} \cdot y + \frac{1}{2 \cdot 120^2} \cdot y^2 \right), & 0 \leq y \leq 120, \\
0, & y > 120.
\end{cases}
\]

Since the domain length distribution is the same as the fragment distribution in this simulation, we find that \(N_o(y)\) is the same as the theoretical interior domain number \(N_t(X)\) (Eq. 4.29). This reflects the symmetric relationship between interior domains and oversize domains discussed above. The result is shown in Fig. 4.20.

From the above example, we can see that the theory of oversize domain length distribution of the general cut model fits the Monte Carlo simulation very well.

II. This example is the same as the second example of the last subsection. The
domain and fragment distributions are the same as in Eqs. 4.30 and 4.31. Using Eq. 4.37, we find the expected number of oversize domains in this simulation

\[
N_o(y) = \begin{cases} 
\frac{N_e}{120} \cdot \frac{1}{120} \cdot 100 \cdot e^{-y/100}, & 80 \leq y \leq 200, \\
0, & y > 200.
\end{cases}
\]

The test results are shown in Fig. 4.21.

There are many other examples that cannot be all shown, but all the Monte Carlo simulations are consistent with the theoretical equation of the general cut model for the relationship between original distribution and oversize domain length distribution.

4.4 Reconstructing the Original Domain Length Distribution

In this section, we will discus how to use the interior domain and fragment length distributions to reconstruct the original domain length distribution. Note that be-
cause the length of interior domains is shorter than the fragment length, we can only reconstruct the original distribution for $X < L_{max}$, where $L_{max}$ is the maximal length of all fragments. Since there will be few fragments with $L \gg \bar{L}$, the variance of any estimation of $\rho(X)$ will be increasingly unreliable for $L > \bar{L}$.

### 4.4.1 Reconstructing the Original Eye and Hole Domain Length Distributions

We can directly use the expression for the interior domain length distribution to find a method to reconstruct the original domain length distribution from the interior domain length distribution. Referring to Eq. 4.25, we write the original domain length distribution as

$$\rho(X) = \frac{\bar{L}}{N_t \cdot \int_X^\infty (L - X) \cdot \rho_f(L) dL} \cdot N_i(X),$$  \hspace{1cm} (4.39)$$

where $N_i(X)$ is the number of domains of length $X$, $\rho_f(L)$ is the length distribution.
of fragments. In the actual experimental data analysis, both \( N_i(X) \) and \( \rho_f(L) \) can be measured directly by histograming experimental data. If the fragment length data is histogramed into \( n \) bins, we can write Eq. 4.39 as

\[
\rho(X) = N_i(X) \cdot \frac{\overline{L}}{N_t \cdot \sum_{L_j > X} (L_j - X) \cdot P(L_j)}.
\]  

(4.40)

Since \( N_t \) and \( \overline{L} \) are constant, we can rewrite Eq. 4.40 as

\[
\rho(X) = K \cdot \frac{N_i(X)}{\sum_{L_j > X} (L_j - X) \cdot P(L_j)},
\]  

(4.41)

where \( K \) is the normalization factor and \( P(L_j) \) is the probability of having fragments of length \( L_j \). I tested Eq. 4.41 in Monte Carlo simulation. In the following example, domain and fragment length distributions follow Eqs. 4.27 and 4.28.

![Figure 4.22: Comparison between the reconstructed and original domain length distributions. The original domain and cut fragment length distributions are both uniform over lengths between 0 and 120.](image)

We can see from Fig. 4.22 that the reconstructed distribution for small domain lengths is stable and close to the original distribution, while for large domain lengths,
it fluctuates wildly. This happens because the scale factor is very large for big domains, and, thus, magnifies the fluctuations.

4.4.2 Reconstruction Method for Eye-to-eye Domains

An eye-to-eye domain is made up from two eyes and one hole domain. Its length is the distance between the centers of two adjacent eyes.

![Diagram of eye-to-eye domain](image)

Figure 4.23: Length for interior eye-to-eye domain to be contained within a fragment. $X$ is the eye-to-eye domain length, $L$ the cut fragment length, $T$ the total length of the three domains that make up the eye-to-eye domain, and $L - T$ the length for which this eye-to-eye domain can stay in the fragment.

From Fig. 4.23, the actual length for one eye-to-eye domain to be in a fragment is $L - T$. Thus we should rewrite Eq. 4.41 for reconstructing the original eye-to-eye length distribution as

$$
\rho(X) = K \cdot \frac{N_{23}(X)}{\sum_{L_j > T} (L_j - T_X) \cdot P(L_j)},
$$

(4.42)

where $T_X$ is the total length of the three domains that make up the interior eye-to-eye domain of length $X$.

4.5 Summary

In this chapter, I have derived expressions for the fragment domain distributions, given original domain length distributions and given cut fragment length distribution. I have then confirmed these relationships by extensive Monte Carlo simulation. I collect the
major formulas in Table 4.4. In the next chapter, we will show how to use this model to compute average hole or eye lengths given only data from the three fragment domain distributions.

<table>
<thead>
<tr>
<th>Fragment Distributions</th>
<th>Simplified cut model</th>
<th>General cut model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interior</td>
<td>( \rho_i(X) = K \cdot \frac{L-X}{L} \cdot \rho(X) )</td>
<td>( \rho_i(X) = K \cdot \rho(X) \cdot \int_{X}^{\infty} (L-X) \cdot \rho_f(L)dL )</td>
</tr>
<tr>
<td>Edge</td>
<td>( \rho_e(y) = K \cdot \frac{2}{L} \cdot \int_{y}^{\infty} \rho(X) \cdot dX )</td>
<td>( \rho_e(y) = K \cdot \int_{y}^{\infty} \rho_f(L)dL \cdot \int_{y}^{\infty} \rho(X)dX )</td>
</tr>
<tr>
<td>Oversize</td>
<td>( N_o = N_i \cdot \int_{\frac{L}{L}}^{\infty} \frac{X-L}{L} \cdot \rho(X) \cdot dX )</td>
<td>( \rho_o(y) = K \cdot \rho_f(y) \cdot \int_{y}^{\infty} (X-y) \cdot \rho(X) \cdot dX )</td>
</tr>
</tbody>
</table>

Table 4.4: Major formulas for the relationships between the fragment domain length distributions and the original domain length distributions. Definitions of all symbols can be found in the former sections.
Chapter 5

Average Domain Length Estimators

In Chapter 2, we saw that the experimental analysis of DNA replication data is based on the evaluation of averages of replicated and unreplicated lengths of DNA. But, as we have seen in Chapter 2, using only statistics compiled from interior domain from DNA fragments will bias downward the naive estimate of average domain size. In this chapter, I will estimate how large this bias is, and I will show how to find better, unbiased estimators for the average domain length.

To be consistent with the discussion of the last chapter, I define $X$ to be the length of a domain whose distribution is $p(X)$. There are a total $N_t$ domains and $M$ cut positions in series covering the DNA molecule, with $L$ the fragment length and $M \cdot L = S$.

5.1 Interior Average Domain Length Estimator

If we sum the lengths of the interior domains to compute the average domain length, we have the interior average domain length estimator\(^1\). This is what has been commonly used in most experiments up to now. Obviously, the average computed using the interior estimator lacks contribution from domains bigger than the cut size and thus will be biased down—but by how much? First, we look at the simplest case,

\(^1\)A statistic used to estimate some property of a p.d.f (e.g. its mean, variance, or other parameters) is called an estimator. Applied to a particular set of data, an estimator gives an estimate.
with constant fragment size $L$. Referring to Eq. 4.5, we recall that the interior domain length distribution is

$$
\rho_i(X) = \begin{cases} 
K \cdot \frac{L-X}{L} \cdot \rho(X), & X \leq L, \\
0, & X > L,
\end{cases}
$$

where $K = 1/ \left( \int_0^L \frac{L-X}{L} \cdot \rho(X) dX \right)$ is the normalization factor. We denote the average interior domain size as $\langle X_i \rangle$, which can be expressed as

$$
\langle X_i \rangle = K \cdot \int_0^L X \cdot \frac{L-X}{L} \cdot \rho(X) dX
$$

$$
= K \cdot \left( \int_0^L X \cdot \rho(X) dX - \int_0^L \frac{X^2}{L} \cdot \rho(X) dX \right).
$$

In Eq. 5.2, when $L \to 0$, we see that $\langle X_i \rangle \to 0$; while when $L \to +\infty$, we have $\langle X_i \rangle \to \langle X \rangle$. These two limits are consistent with common sense. For experimental data, we should use the discrete form

$$
\langle X_i \rangle = \frac{1}{N} \sum_{j=0}^{N} X_j.
$$

where $X_j$ is the length of one domain and $N$ is the total number of interior domains.

The original overall average domain length can be written as

$$
\langle X \rangle = \int_0^{+\infty} X \cdot \rho(X) dX.
$$

Comparing Eq. 5.2 with Eq. 5.4, we can see the bias between interior estimator and original average length comes from two effects: first, the range of the interior domain distribution is limited to $(0, L)$, which may not cover the whole range of domain lengths; second, the shape of the interior domain length distribution (Eq. 5.1) is different from the original domain length distribution.

I have tested Eq. 5.2 by Monte Carlo simulation using many different distributions. The Monte Carlo simulations are done as follows:

1. Sample sizes of eyes and holes from chosen distribution. Connect eyes and holes alternately into a long line. See Sec. 4.2.1.
2. Since the bias of the average domain length estimator depends on the cut fragment length, test cut fragments of lengths ranging from \( L = 0.05 \cdot \langle X \rangle \) to \( L = 20 \cdot \langle X \rangle \).

3. Randomly choose cut positions and record lengths of fragment domains for each fragment. Here, we used the importance-sampling method. (Also called “biased sampling,” this is one of several variance-reducing techniques in Monte Carlo methods [33–35].) In our Monte Carlo simulation, when testing the interior estimator, only fragments that contain at least one interior domain are included in data compiling. Thus, the variance can be reduced even if we use a small sampling number.

4. Compute the theoretical expected average length for different estimators, the average domain length and the standard deviation for every estimator from the simulated data, and plot the theoretical curves and simulated data on one figure.

As an example, let the original domain distribution be an exponential function, which is what one expects for the hole length distribution in DNA replication with random initiation [21]. In the Monte Carlo simulation, the original distribution function is

\[
\rho(X) = \frac{1}{\mu} \cdot e^{-\frac{X}{\mu}}, \tag{5.5}
\]

where \( \mu = \langle X \rangle = 100 \) is the overall average domain length of this distribution. Using Eq. 5.2, we obtain the theoretical expected average interior domain length as a function of fragment length,

\[
\frac{\langle X_i \rangle}{\mu} = \frac{(L' - 2) + (L' + 2) \cdot e^{-L'}}{L' - (1 - e^{-L'})}, \tag{5.6}
\]

where \( L' = \frac{L}{\mu} \) and \( L \) is the cut fragment length. From Eq. 5.6, we can see that as \( L' \to 0 \), the interior average domain size \( \langle X_i \rangle = \frac{L}{\mu} \to 0 \). This means that when the cut fragment \( L \) is very small, the average interior domain length will be very small, too. By contrast, as \( L' \to +\infty \), \( \langle X_i \rangle \to \mu = \langle X \rangle \), i.e., when the cut fragment \( L \) is
Figure 5.1: Performance of the interior estimator. The original average domain length is shown by the dashed line; the Monte Carlo simulated average interior domain length is shown by circles with error bars; and the theoretical average interior domain length is shown by the solid line. The original domain length distribution is exponential (Eq. 5.5), with $\mu = 100$, and the theoretical curve is from Eq. 5.6. The normalized fragment length $L' = L/\mu$.

very large, the average interior domain length will equal the overall average domain length. These features are shown on Fig. 5.1.

In Fig. 5.1, the horizontal axis is rescaled to be the ratio of the cut fragment length to the original average domain length. The vertical axis is also rescaled to be the ratio of the measured fragment average domain length to the original average domain length. As predicted, when $L \to 0$, the ratio of average interior domain length and cut fragment length $\langle X_i \rangle / \langle X \rangle \to L/(3 \langle X \rangle) \to 0$; and when $L \to +\infty$, the average interior domain length is approached to overall average domain length, that is $\langle X_i \rangle / \langle X \rangle \to 1$. We can see that the theoretical line fits simulated data very well.
5.2 Reconstructed Distribution Estimator

From the last section, the bias of the interior domain estimator comes from the finite cut-fragment length and from the difference between the interior domain distribution and that of the original domain. From the theory developed in the last chapter, the original distribution can be reconstructed from the interior domain distribution. In this section, I will discuss how to use the reconstructed distribution to compute the average domain length. We call this the reconstructed distribution estimator. One direct advantage of this estimator is its bias from the original average domain length is smaller.

I denote the reconstructed domain average length as \( \langle X_R \rangle \) and the interior domain distribution as \( \rho_i(X) \). Referring to Eq. 4.5, we can write the original distribution between 0 and \( L \) as

\[
\rho(X) \propto \begin{cases} \frac{L}{L-X} \cdot \rho_i(X), & X \leq L, \\ \text{Undefined}, & X > L, \end{cases}
\]  

(5.7)

Since the length of each interior domain must be smaller than the length of the cut fragment, we cannot find the original distribution of the entire range of domain length. We can only reconstruct the original distribution for \( 0 < X \leq L \) and cannot find the original distribution for \( X > L \) from information from the interior domains. Thus, this equation only partly reconstructs the original domain length distribution. Then the reconstructed domain average length can be written as

\[
\langle X_R \rangle = \frac{\int_0^L X \cdot \frac{L}{L-X} \cdot \rho_i(X) dX}{\int_0^L \frac{L}{L-X} \cdot \rho_i(X) dX} = \frac{\int_0^L X \cdot \rho(X) dX}{\int_0^L \rho(X) dX}.
\]  

(5.8)

For experimental data, we should use the equation in its discrete form,

\[
\langle X_R \rangle = \frac{\sum_{j=0}^{N} X_j \cdot \frac{L}{L-X_j}}{\sum_{j=0}^{N} \frac{L}{L-X_j}}.
\]  

(5.9)

where \( X_j \) is the length of one domain, \( L \) is the length of cut fragment and \( N \) is the total number of interior domains. In this data-processing approach, we actually use
the original domain length distribution to compute the average domain length; thus
the bias comes only from the finite size of cut fragments, which restricts the integra-
tion range to $(0, L)$. Compared to the interior domain estimator, the reconstructed
distribution estimator reduces the bias and improves the accuracy of the estimation of
the average domain length. We will now test the reconstructed distribution estimator
by Monte Carlo simulation.

Figure 5.2: Performance of the reconstructed estimator. The original average domain
length is shown by the dashed line; the Monte Carlo simulated average interior do-
main length is shown by circles with error bars; and the theoretical average interior
domain length is shown by the solid line. The original domain length distribution is
exponential (Eq. 5.5), with $\mu = 100$, and the theoretical curve is from Eq. 5.10.

We use the same exponential distribution that we did in Eq. 5.5, with $\langle X \rangle = \mu =
100$. Using Eq. 5.8, we find the reconstructed distribution estimator

$$\langle X_R \rangle = \frac{1 - (1 + L') \cdot e^{-L'}}{1 - e^{-L'}}.$$

From Eq. 5.10, we can see that as $L' \to 0$, the interior average domain size
$\langle X_R \rangle \to \frac{L}{2} \to 0$. Thus, when the cut fragment $L$ is very small, the average interior
domain length will be very small too; while when $L' \to +\infty$, $\langle X_R \rangle \to \mu = \langle X \rangle$. That
CHAPTER 5. AVERAGE DOMAIN LENGTH ESTIMATORS

is, when the cut fragment $L$ is very big, the average interior domain length will equal to the overall average domain length. These features are shown on Fig. 5.2, which gives our Monte Carlo simulation results.

In Fig. 5.2, the horizontal axis and vertical axis are rescaled as in the previous section. The range of cut fragment lengths is also the same as in the last section. As predicted by Eq. 5.10, when $L' \to 0$, $\langle X_R \rangle / \langle X \rangle = L/(2 \langle X \rangle) \to 0$; and when $L' \to +\infty$, the average interior domain length approaches the overall average domain length; that is, $\langle X_R \rangle / \langle X \rangle \to 1$. We can see that the theoretical line fits the simulated data very well.

If the cut lengths are not all equal, we refer to Eq. 4.25. The interior domain length distribution can be then written as

$$N_i(X) = \frac{N_t}{\bar{L}} \cdot \rho(X) \cdot \left( \int_X^\infty L \cdot \rho_f(L)dL - \int_X^\infty X \cdot \rho_f(L)dL \right)$$

$$= \frac{N_t}{\bar{L}} \cdot \rho(X) \cdot (\bar{L}_X - P_X \cdot X),$$

(5.11)

where $\bar{L} = \int_0^\infty L \cdot \rho_f(L)dL$ is the overall average cut fragment length, $\bar{L}_X = \int_X^\infty L \cdot \rho_f(L)dL$, and $P_X = \int_X^\infty \rho_f(L)dL$ is the cumulative probability for $L \geq X$. Since the cut fragment length distribution can be measured directly from experimental data, these two factors are known.

Thus we obtain for the original distribution,

$$\rho(X) = \frac{\bar{L}}{N_t(\bar{L}_X - P_X \cdot X)} \cdot N_i(X),$$

(5.12)

where $N_i$ is a constant and may be not known from the experimental data. We should then compute the reconstructed average domain length,

$$\langle X_R \rangle = \frac{\int_0^{L_{\text{max}}} X \cdot \frac{\bar{L}}{\bar{L}_X - P_X \cdot X} \cdot \rho_i(X)dX}{\int_0^{L_{\text{max}}} \frac{\bar{L}}{\bar{L}_X - P_X \cdot X} \cdot \rho_i(X)dX},$$

(5.13)

where $L_{\text{max}}$ is the maximal cut fragment length. For experimental data, we should use the equation in discrete form,
where $X_j$ is the length of one domain and $N$ is the total number of interior domains. The form of Eq. 5.14 is very close to Eq. 5.9. We tested this equation against Monte Carlo simulations; The theory fits the simulation data very well and gives a cut fragment length-dependent bias similar to that found in the simple cut model.

### 5.3 Unbiased Estimator I

Both the interior and reconstructed average length estimators are biased away from the original average domain length. According to our theory of fragment domain distributions, the interior domain distribution contains information only about domains of length smaller than the cut fragment length; the edge domain distribution contains information about the whole range of domain lengths; and the oversize domain distribution contains information only about domains of length larger than the cut fragment length. If we organize the information from the three fragment domains in a proper way, it is possible to make an unbiased estimator of the average domain length.

Referring to the interior domain number Eq. 4.4, we can write the expected total number of interior domains, denoted as $N_i^{tot}$, as

$$
N_i^{tot} = \int_0^L N_i(X) \cdot dX
$$

$$
= N_i \cdot \int_0^L \rho(X)dX - \frac{N_i}{L} \cdot \int_0^L X \cdot \rho(X)dX,
$$

(5.15)

where $\rho(X)$ is the original domain length distribution and $N_i$ is the total amount of domains. Referring to the edge domain number Eq. 4.13, we can write the expected total number of edge domains, denoted as $N_e^{tot}$, as
CHAPTER 5. AVERAGE DOMAIN LENGTH ESTIMATORS

Referring to the oversize domain number Eq. 4.21, we can write the expected total number of oversize domains, denoted as $N_{o}^{\text{tot}}$, as

\[
N_{o}^{\text{tot}} = \frac{N_{t}}{L} \cdot \int_{0}^{L} X \cdot \rho(X) dX - N_{t} \cdot \int_{0}^{L} \rho(X) dX.
\]  \hspace{1cm} (5.17)

Combining Eqs. 5.15 and 5.16, we have

\[
N_{i}^{\text{tot}} + N_{e}^{\text{tot}} / 2 = N_{t} \cdot \int_{0}^{L} \rho(X) dX + N_{t} \cdot \int_{0}^{+\infty} \rho(X) dX
\]

\[
= N_{t} \cdot \int_{0}^{+\infty} \rho(X) dX.
\]  \hspace{1cm} (5.18)

Combining Eqs. 5.16 and 5.17, the sum of the expected total oversize domain number and half of expected total edge domain number is

\[
N_{o}^{\text{tot}} + N_{e}^{\text{tot}} / 2 = \frac{N_{t}}{L} \cdot \int_{0}^{+\infty} X \cdot \rho(X) dX + \frac{N_{t}}{L} \cdot \int_{0}^{L} X \cdot \rho(X) dX
\]

\[
= \frac{N_{t}}{L} \cdot \int_{0}^{+\infty} X \cdot \rho(X) dX.
\]  \hspace{1cm} (5.19)

Using $\langle X \rangle = \int_{0}^{+\infty} X \cdot \rho(X) dX$, we can combine Eqs. 5.18 and 5.19 to find the unbiased domain average length estimator,

\[
\langle X \rangle = \left( \frac{N_{o}^{\text{tot}} + N_{e}^{\text{tot}} / 2}{N_{i}^{\text{tot}} + N_{e}^{\text{tot}} / 2} \right) \cdot L.
\]  \hspace{1cm} (5.20)

Figure 5.3 illustrates how to use this estimator. Equation 5.20 makes sense: if the cut fragment length is smaller than the average domain length, then we expect
to have many oversize domains and relatively few interior domains. This means that the denominator will be smaller than the numerator, and thus the estimator will give a length that is larger than $L$. If the cut fragment length is larger than the average domain length, we expect to find many interior domains and few oversize domains. This means the denominator will be bigger than the numerator, and the estimator will give a length smaller than $L$.

We still need Monte Carlo simulations to test this estimator. We use the same original distribution as in Eq. 5.5 with $\langle X \rangle = \mu = 100$. The test results are shown in Fig. 5.4.

We can see that the estimated average domain length agrees with the true average domain length. The relative standard deviation increases as the cut fragment length decreases. This happens because when the cut fragment length is small, the number of interior domains is small and the number of oversize domains is big, a small change in the number of interior domains will cause a huge change in the average domain length; when the cut fragment length is big, the change in the number of interior domains causes a small change in the estimated average domain length.

If the cut fragment length is not a fixed number, we should use the general cut model. For a specified cut fragment length $L$, we refer to Eq. 4.23 and replace the $N_t$ in Eq. 5.18 by $N_L = N_t \cdot \frac{\rho_f(L)}{L}$. Here, $N_i^L$, $N_e^L$ and $N_o^L$ denote the expected number of interior, edge, and oversize domains for cut fragments of a given size $L$.

$$N_i^L + N_e^L/2 = N_t \cdot \frac{\rho_f(L)}{L} \cdot \int_0^{+\infty} \rho(X) dX. \quad (5.21)$$

Integrating Eq. 5.21 over $L$ and defining $\bar{L} = \int_0^{+\infty} L \cdot \rho_f(L) dL$, we have
Figure 5.4: Performance of the Unbiased Estimator I. The original average domain length is shown by the dashed line and the Monte Carlo simulated average interior domain length is shown by circles with error bar. The original domain length distribution is exponential (Eq. 5.5), with $\mu = 100$.

\[
N^t_{o} + N^e_{o}/2 = N_t \cdot \int_0^{+\infty} \frac{\rho_f(L) \cdot LdL}{L} \cdot \int_0^{+\infty} \rho(X)dX
\]

\[
= N_t \cdot \int_0^{+\infty} \rho(X)dX. \tag{5.22}
\]

We rewrite Eq. 5.19 as

\[
N^t_{o} + N^e_{o}/2 = \frac{N_t \cdot \rho_f(L) \cdot L}{L} \cdot \int_0^{+\infty} X \cdot \rho(X)dX
\]

\[
= \frac{N_t \cdot \rho_f(L)}{L} \cdot \int_0^{+\infty} X \cdot \rho(X)dX. \tag{5.23}
\]

Integrating Eq. 5.23 over $L$ and using $1 = \int_0^{+\infty} \rho_f(L)dL$, we have
\[ N_{o}^{tot} + N_{e}^{tot}/2 = \frac{N_{i}}{L} \cdot \int_{0}^{+\infty} \rho_{f}(L)dL \cdot \int_{0}^{+\infty} X \cdot \rho(X)dX \]
\[ = \frac{N_{i}}{L} \cdot \int_{0}^{+\infty} X \cdot \rho(X)dX. \]  

Combining Eqs. 5.22 and 5.24, we find the unbiased domain average length estimator,
\[ \langle X \rangle = \left( \frac{N_{o}^{tot} + N_{e}^{tot}/2}{N_{i}^{tot} + N_{e}^{tot}/2} \right) \cdot \bar{L}. \]  

We can see that the form of estimator for variable cut fragment length is the same as Eq. 5.20 for the fixed-cut-fragment-length estimator, except that \( L \) is replaced by \( \bar{L} \). I have also tested the estimator Eq. 5.25 in Monte Carlo simulation. The test results shows that the theory fits the simulated data very well with no big difference between estimators for fixed cut fragment length. The standard deviations are similar to those observed for a single cut size.

\section*{5.4 Unbiased Estimator II}

The Unbiased Estimator I uses information about the number of fragment domains, but it ignores the information coming from the length of fragment domains. In this section, we will also make use of the length information to derive a better unbiased estimator.

Referring to the interior domain number Eq. 4.4, we can write the expected total length of interior domains, denoted as \( X_{i}^{tot} \), as
\[ X_{i}^{tot} = \int_{0}^{L} X \cdot N_{i}(X) \cdot dX \]
\[ = N_{i} \cdot \int_{0}^{L} X \cdot \rho(X)dX - \frac{N_{i}}{L} \cdot \int_{0}^{L} X^{2} \cdot \rho(X)dX, \]  
where \( \rho(X) \) is the original domain length distribution and \( N_{i} \) is the total number of domains. Referring to the edge domain number Eq. 4.13, we can write the expected total length of edge domains, denoted as \( X_{e}^{tot} \), as
CHAPTER 5. AVERAGE DOMAIN LENGTH ESTIMATORS

Referring to the oversize domain number Eq. 5.17, we denote the expected total length of oversize domains as
\[ X_{o}^{\text{tot}} = N_{o}^{\text{tot}} \cdot L, \]

Combining Eqs. 5.26, 5.27 and 5.28, we write the expected total length of all fragment domains as
\[ X_{i}^{\text{tot}} + X_{e}^{\text{tot}} + X_{o}^{\text{tot}} = N_{i} \cdot \int_{0}^{+\infty} X \cdot \rho(X)dX. \]  

Combining Eqs. 5.18 and 5.29, we see that the unbiased domain average length estimator is
\[ \langle X \rangle = \frac{X_{i}^{\text{tot}} + X_{e}^{\text{tot}} + X_{o}^{\text{tot}}}{N_{i}^{\text{tot}} + N_{e}^{\text{tot}}/2}. \]

The form of the Unbiased Estimator II is very simple. Taking the experimental data, we can sum up the total length of interior domains, edge domains, and oversize domains, and divide the total length by total number of interior domains and half of edge domains. In fact, what we do here is to view an interior domain as one entire domain and count 1; we look at an edge domain as one half domain and count 0.5; and we view an oversize domain as part of a big domain and only add its length without counting its number. Although this is all consistent with common sense, that it is an unbiased estimator may be not obvious.

Next, we test this estimator by Monte Carlo simulation. The original distribution is the same as in Eq. 5.5 with \( \langle X \rangle = \mu = 100 \). The test results are shown in Fig. 5.5.
CHAPTER 5. AVERAGE DOMAIN LENGTH ESTIMATORS

Figure 5.5: Performance of the Unbiased Estimator II. The original average domain length is shown by the dashed line and the Monte Carlo simulated average interior domain length is shown by circles with error bar. The original domain length distribution is exponential (Eq. 5.5), with $\mu = 100$.

From Fig. 5.5, we can see that the estimated average domain length agrees with the true average domain length. The standard deviation increases as the cut fragment length decreases.

If the cut fragment length is not a fixed number, we should use the general cut model. As before, the $N_t$ in Eq. 5.18 should be replaced by $N_L = N_t \cdot \frac{\rho_f(L) \cdot L}{L}$. We denote by $X_t^L$, $X_e^L$ and $X_o^L$ the expected total length of interior domains, edge domains, and oversize domains for all cut fragments of length $L$. Then, from Eq. 5.29, we have

$$X_t^L + X_e^L + X_o^L = N_t \cdot \frac{\rho_f(L) \cdot L}{L} \cdot \int_0^{+\infty} X \cdot \rho(X) dX.$$  \hspace{1cm} (5.31)

Integrating Eq. 5.31 over $L$ and using $\bar{L} = \int_0^{+\infty} L \cdot \rho_f(L) dL$, we have
Combining Eqs. 5.22 and 5.32, we can write the unbiased domain average length estimator as

\[ X_i^{tot} + X_e^{tot} + X_o^{tot} = N_t \cdot \int_0^{\infty} \frac{\rho_f(L) \cdot L dL}{L} \cdot \int_0^{\infty} X \cdot \rho(X) dX \]

\[ = N_t \cdot \int_0^{\infty} X \cdot \rho(X) dX. \quad (5.32) \]

Equation 5.33 is the same as Eq. 5.30 for fixed cut fragment lengths. I have also tested the estimator Eq. 5.33 by Monte Carlo simulation. The test results shows that the theory agrees with the simulated data and there is no difference between estimators for fixed cut fragment length and for variable cut fragment length.

### 5.5 Interior-Edge Estimator

If the only experimental “problem” with the DNA combing experiments were that the DNA is broken up into finite-size fragments, we would have all the tools we need to analyze the data properly. Unfortunately, the problem is that experiments are conducted on populations of cells that are not well-synchronized. Thus, when one observes a fragment of DNA, one is not sure when the cell it came from started replication. One solution to this difficulty is to use the local replication fraction of the fragment itself as a kind of clock. Thus, one can sort data by binning with respect to local replication fragment. While this works well when there are at least a few domains on the fragment, it clearly cannot work for an oversize domain. Then information of oversize domains is not available for computing the average domain length.

Thus, we have to consider how to make an estimator using information from only interior domains and edge domains. Our strategy is to use the Unbiased Estimator II from Eq. 5.34 and simply delete the contribution from oversize domains. Thus, we have the “Interior-Edge” (IE) estimator,
CHAPTER 5. AVERAGE DOMAIN LENGTH ESTIMATORS

\[ \langle X_{ie} \rangle = \frac{X_{i}^{tot} + X_{e}^{tot}}{N_{i}^{tot} + N_{e}^{tot}/2}. \]  

(5.34)

Using Eqs. 5.26 and 5.27, we have

\[ X_{i}^{tot} + X_{e}^{tot} = N_{i} \cdot \int_{0}^{L} X \cdot \rho(X) dX + N_{i} \cdot L \cdot \int_{L}^{+\infty} \rho(X) dX, \]  

(5.35)

Combining Eqs. 5.35 and 5.22, we find

\[ \langle X_{ie} \rangle = \int_{0}^{L} X \cdot \rho(X) dX + L \cdot \int_{L}^{+\infty} \rho(X) dX. \]  

(5.36)

We can see the IE estimator is unbiased if \( L \geq X_{\text{max}} \), while it is biased if \( L < X_{\text{max}} \). Though it is not totally unbiased, it is better than the reconstructed distribution estimator, for the second term of Eq. 5.36 gives a minimal estimation of the effect of domains of length longer than the cut fragment length. We test this equation by Monte Carlo simulation. The original distribution is the same as in Eq. 5.5, and we set the overall average length \( \langle X \rangle = \mu = 100 \). Using Eq. 5.36, we find

\[ \frac{\langle X_{ie} \rangle}{\mu} = 1 - e^{-L'}, \]  

(5.37)

with again \( L' = \frac{L}{\mu} \). From Eq. 5.37, we can see that as \( L' \to 0 \), the interior average domain size \( \langle X_{ie} \rangle \to L \to 0 \). On the other hand, for large \( L' \), \( \langle X_{ie} \rangle \to \mu \): the estimator is unbiased in this limit. There features are shown in Fig. 5.6, which gives the Monte Carlo test results.

If the cut fragment length is not a fixed number, the form of interior-edge estimator is the same as Eq. 5.34. We should use the general cut model to find out the theoretical expected average domain length. Again, the \( N_{i} \) in Eq. 5.35 should be replaced by \( N_{L} = N_{i} \cdot \frac{\rho(L) \cdot L}{L} \), and \( X_{i}^{L} \) and \( X_{e}^{L} \) denote the expected total length of interior domains and edge domains for a cut fragment of size \( L \).

\[ X_{i}^{L} + X_{e}^{L} = N_{i} \cdot \frac{\rho(L) \cdot L}{L} \cdot \int_{0}^{L} X \cdot \rho(X) dX + N_{i} \cdot \frac{\rho(L) \cdot L}{L} \cdot L \cdot \int_{L}^{+\infty} \rho(X) dX. \]  

(5.38)

We integrate Eq. 5.38 over \( L \) from 0 to \( L_{\text{max}} \) to get the total length of domains. As the integration result is complex, we give here the approximate result.
Figure 5.6: Performance of the Interior-Edge estimator. The original average domain length is shown by the dashed line; the Monte Carlo simulated average interior domain length is shown by circles with error bar and the theoretical average interior domain length is shown by the solid line. The original domain length distribution is exponential (Eq. 5.5), with $\mu = 100$, and the theoretical curve is from Eq. 5.36.

\[
X_i^L + X_e^L \approx N_t \cdot \int_0^{L_{\max}} X \cdot \rho(X) dX + N_t \cdot \frac{\overline{L^2}}{L} \cdot \int_{L_{\max}}^{+\infty} \rho(X) dX, \tag{5.39}
\]

where $L_{\max}$ is the maximal cut fragment length. Combining with Eq. 5.22, we find

\[
\langle X_{ie} \rangle \approx \int_0^{L_{\max}} X \cdot \rho(X) dX + \frac{\overline{L^2}}{L} \cdot \int_{L_{\max}}^{+\infty} \rho(X) dX. \tag{5.40}
\]

where $\overline{L^2} = \int_0^{L_{\max}} L^2 \cdot \rho_f(L) dL$. We can see that Eq. 5.40 is similar to Eq. 5.36. Its accuracy of estimation depends on the maximal cut fragment length and the standard deviation of the cut fragment length distribution. The longer the maximal cut fragment length is, the smaller the bias; We have tested this equation by Monte Carlo simulation, and the theory fits the simulation data very well.
5.6 Summary

We have discussed five different domain average length estimators in this Chapter. To compare the different estimators, we plot the interior domain, reconstructed distribution and interior-edge estimators and Unbiased Estimator II in Fig. 5.7; to compare the unbiased estimators, we plot the standard deviation of Unbiased Estimator I and Unbiased Estimator II in Fig. 5.8.

Figure 5.7: Comparison of Interior estimator, reconstructed estimator and interior-edge estimator. We can see that the Interior-Edge estimator is the best of the biased estimators.

<table>
<thead>
<tr>
<th>$L_{min}$</th>
<th>Interior Estimator</th>
<th>Reconstructed Estimator</th>
<th>Interior-Edge Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>$12 \langle X \rangle$</td>
<td>$4 \langle X \rangle$</td>
<td>$2.5 \langle X \rangle$</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: Minimal length ($L_{min}$) of fragment needed to achieve a bias of 10% or less.

Figure 5.7 clearly shows that the interior-edge estimator has a lower bias than
either the interior or the reconstructed distribution estimators. See Table 5.1 for the minimal length of fragment required to reduce the bias to 10%. For DNA replication experiments that are not well-synchronized, one should use the interior-edge estimator. There are two benefits to using the interior-edge estimator: first, the interior-edge estimator has lower bias; second, more of the data can be used for the interior-edge estimator than for the interior estimator, since some fragments with only edge domains cannot be used by the interior estimator. We will see how much this improves the data estimation in the next chapter.

![Figure 5.8: Comparison between the relative standard deviations (σ/μ) of Unbiased Estimators I and II.](image)

We can see that Unbiased Estimator II has a smaller deviation than Unbiased Estimator I for big fragment and the same deviation as Unbiased Estimator I for small fragments. The deviation of Unbiased Estimator II was fit to $\sigma/\mu = 0.169(L')^{-0.5}$, which is consistent with the usual law, $\sigma_{\text{average}} = \sigma_{\text{distribution}}/\sqrt{N_{\text{samples}}}$, since here, $L' \propto N_{\text{domains}}$.

If the DNA replication experiment is well-synchronized, all the information from the three fragment domains will be available. We can use either of the unbiased estimators. They each have their own benefits. Unbiased Estimator I is simpler to compute—one only needs to count domains and measure the length of fragments. Unbiased Estimator II has a lower variance than Unbiased Estimator I, but one needs
to measure the length of all domains.
Chapter 6

Application to Experimental data

6.1 Introduction

In this chapter, we will apply our methods for treating finite-size effects to the experimental data collected by Herrick et al. [16] on replication in Xenopus cell embryos. Since our theory relies on the assumption of a uniform distribution of cuts, we will first test this assumption against the experimental data. Having shown that the assumption of uniform cut sites is a reasonable one, we use the new estimators we have derived to recompute the average domain lengths of both replicated and non-replicated regions. Then, we apply the recomputed average domain length to the KJMA model [22], to extract important quantities concerning DNA replication. In particular, we focus on $I(t)$, which is the rate, per length of unreplicated DNA, with which new origins are initiated.

6.2 Test of Uniform Distribution Assumption

In the analysis of Chapters 4 and 5, we assumed that the cut positions were uniformly distributed along the genome. In the DNA replication experiment of Herrick et al. [16, 23], the combed molecules were broken up after having fully replicated, this means that at the time the chromosome is cut into fragments, the eyes and holes have both completely replicated, with the only difference between them being the type (color) of
fluorophore attached to the altered bases of the complementary strand. In particular, it is NOT the case that the replicated regions have two double-strand fragments while the unreplicated regions have one. (Such is the case at the time the second labeled base is added but the DNA is fragmented after completion.) Thus, there is little physical difference between eyes and holes, and the random-cut hypothesis is at least plausible.

However, we still need to test this assumption by checking the statistical properties of the experimental data. Although there are an infinite number of ways for the cut positions to be non-uniformly distributed, we focus on two reasonable scenarios. The first is that cuts always take place at the boundaries between domains ("replication forks"). The second is that breaks are more likely to take place within one or the other types of domains.

The first hypothesis is that the DNA molecules break up at the boundaries of attached different fluorophores, where the replication forks are at the time when the second dye was added. If this assumption were true, an edge domain would not actually be an "edge" domain, it would be an entire domain. Then the average edge domain length should be equal to the average interior domain length, that is,

$$\frac{\langle X_i \rangle}{\langle X_e \rangle} = 1,$$

where $\langle X_i \rangle$ is the average interior domain length and $\langle X_e \rangle$ is the average edge domain length.

On the other hand, assuming a uniform distribution of cuts of equal size (Sec. 4.2.1), we recall that the interior domain length distribution is

$$\rho_i(X) = K_i \cdot \frac{L - X}{L} \cdot \rho(X),$$

where $\rho(X)$ is the original domain length distribution, $K_i = 1/\left( \int_0^L \frac{L - X}{L} \cdot \rho(X) dX \right)$ is the normalization factor, and $L$ is the cut fragment length. Thus, the average interior domain length is

$$\langle X_i \rangle = K_i \cdot \left( \int_0^L X \cdot \rho(X) dX - \int_0^L \frac{X^2}{L} \cdot \rho(X) dX \right),$$
Referring to Sec. 4.2.2, the edge domain length distribution is

\[ \rho_e(y) = K_e \cdot \frac{2}{L} \cdot \int_y^{+\infty} \rho(X) \cdot dX, \]  

where \( K_e = 1/\left(\int_{L}^{+\infty} 2 \cdot \rho(X) dX + \int_{0}^{L} \frac{2X}{L} \cdot \rho(X) dX\right) \) is the normalization factor.

Knowing the edge domain distribution, and referring to Eq. 5.27, we find the average edge domain length,

\[ \langle X_e \rangle = K_e \cdot \left( \int_{L}^{+\infty} L \cdot \rho(X) dX + \int_{0}^{L} \frac{X^2}{L} \cdot \rho(X) dX \right). \]  

(6.5)

These expression for \( \langle X_i \rangle \) and \( \langle X_e \rangle \) depend both on the cut size \( L \) and on the form of the original parent distribution \( \rho(X) \). To understand their implications better, we look at the limits of very small and very large cut sizes. First, when \( L \rightarrow 0 \), we can easily see that, for any original distributions \( \rho(X) \)

\[ \langle X_e \rangle \rightarrow \frac{L}{2}, \quad \langle X_i \rangle \rightarrow \frac{L}{3}. \]  

(6.6)

More precisely, for \( L' = L/\langle X \rangle \rightarrow 0 \), we expect to see \( \langle X_e \rangle / \langle X_i \rangle \rightarrow 1.5 \).

As \( L \rightarrow \infty \),

\[ \langle X_e \rangle \rightarrow \frac{\langle X^2 \rangle}{2 \langle X \rangle}, \quad \langle X_i \rangle \rightarrow \langle X \rangle. \]  

(6.7)

where \( \langle X \rangle \) is the original average domain length. Since the variance \( \sigma^2 = \langle X^2 \rangle - \langle X \rangle^2 \), we can write Eq. 6.7 as

\[ \frac{\langle X_e \rangle}{\langle X_i \rangle} = \frac{1}{2} \left[ 1 + \left( \frac{\sigma}{\langle X \rangle} \right)^2 \right]. \]  

(6.8)

Note that, depending on \( \sigma/\langle X \rangle \), this ratio can be larger or smaller than 1. Thus, for example, if \( \rho(X) \) is an exponential distribution, then \( \sigma = \langle X \rangle \), and we expect \( \langle X_e \rangle / \langle X_i \rangle \) to vary from 1.5 at small \( L' \) to 1 at large \( L' \).

6.2.1 Test on experimental data

I use the same experimental data as in Herrick et al. in [23]. If replication origins are uniformly distributed, then the hole-length distribution should be an exponential. In
fact, we see clearly this distribution in the experimental data.\textsuperscript{1} Since the distribution of eyes is more complex (because of coalescence of growing domains), our test will focus on hole length.

For an exponential distribution, \( p(X) = \frac{1}{\mu} \cdot e^{-\frac{X}{\mu}} \) and \( \sigma = \langle X \rangle = \mu \). Then, as we have discussed above, for \( L/\langle X \rangle \to 0 \), we expect to have \( \langle X_e \rangle / \langle X_i \rangle = 1.5 \); when \( L/\langle X \rangle \to +\infty \), we expect to have \( \langle X_e \rangle / \langle X_i \rangle = 1 \).

The ratio of average edge domain length to average interior domain length depends on both the cut fragment length \( L \) and the original average domain length \( \langle X \rangle \). This last quantity in turn depends on the fraction \( f \) replicated [in the early stage of replication (small \( f \)), holes are large; while at later stages (large \( f \)), holes are small]. Since the cut fragment length and fraction replicated vary in experimental data, we cannot use all the data for the test. See the distribution of cut fragment lengths and replication fractions in Fig. 6.1.

\begin{figure}[h]
\centering
\begin{subfigure}[b]{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{a.png}
\caption{(a) Distribution of lengths of combed DNA fragments. The average length is 102 kb and the standard deviation is 75 kb.}
\end{subfigure}
\begin{subfigure}[b]{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{b.png}
\caption{(b) Distribution of fraction replicated of individual DNA fragment.}
\end{subfigure}
\caption{Figure 6.1: (a) Distribution of lengths of combed DNA fragments. The average length is 102 kb and the standard deviation is 75 kb. (b) Distribution of fraction replicated of individual DNA fragment.}
\end{figure}

There are a total of 1142 DNA fragments. The data we used for the test are taken for a relatively small range of fragment lengths, \( 40 \text{ kb} < L < 120 \text{ kb} \), in which range

\textsuperscript{1}The presence of weak correlations in the size of neighboring eyes may alter slightly the form of the hole-length distribution [36]. We do not see any clear indication of this in the experimental hole-length histograms.
a majority of the data lies (612 fragments). Compared with the wide range of original average hole length, we can ignore the change of cut fragment length in this small range. Data are also sorted by replication fraction $f$ into 5 uniform bins of width 0.2.

Figure 6.2: Fit to edge and interior domain length distributions. The data come from $f$ between 0 and 0.2. Triangles represent the data for edge domain lengths and hollow-circles represent interior domain lengths. The solid line is the fitted exponential curve for edge domain data, and the dashed line is for interior data. We can see that the first points of both edge and interior are much smaller than the second and third points, because of the finite-resolution effect. Fits to the edge and interior domain length data exclude the first points.

An example of the data thus plotted is shown in Fig. 6.2, which shows the distribution of both interior and edge domains. Except for the first point, both distributions are reasonably well-fit by an exponential function. We can understand the deviation from an exponential distribution at small length scales for both domain sizes by recalling that the measurement of domains were done by optical microscopy. The spatial resolution of fluorescence microscopy is $\sim 1$ kb, which means that small domains will not be detectable. Thus, the number of small domains will be biased downward. We
use an exponential function to fit the histogram of interior or edge data without data from the first bin. As we know, the decay length of an exponential function is also the average length of the distribution. Then we can avoid the problem of missing small-scale lengths by using the decay constants from edge and interior data to evaluate the ratio of the average edge domain length to average interior domain length. See Fig. 6.2.

Using this method, I find the ratio of average edge domains to average interior domains as a function of replication fraction $f$ (Fig. 6.3).

![Figure 6.3](image)

Figure 6.3: Ratio of average edge domain length to average interior domain length vs. fraction. The dashed line show the upper and lower limit expected for an exponential distribution ($1.5 \rightarrow 1$). Each point corresponds to one of the five bins of replication fraction. The error bars are evaluated from the errors of the function fit parameters.

We can see the ratio is not always equal to 1, which is what we would expect given the hypothesis that the cut positions always occur at the boundaries between eyes and holes. Since we have limited the test data to a small range of lengths, we expect that when $f$ is close to 0, that is the average hole length is very large, $(X_e)/(X_i) \approx 1.5$; and when $f$ is close to 1, that is the average hole length is very small, $(X_e)/(X_i) \approx 1$. We can thus reject the hypothesis that all cuts are at boundaries of attached different fluorophores. Furthermore, the data are qualitatively consistent with expectations given uniform cuts. At this point, we cannot pursue this further without detailed,
model-dependent simulations.

Based on the experimental data we have, we can reject the hypothesis that cut positions always lie at the boundaries between eyes and holes. But we still need to consider the second hypothesis, that cuts lie preferentially in either eye or hole domains.

\[ f = \frac{L_i^{tot}}{L_i^{tot} + L_h^{tot}} = \frac{n_i}{n_i + n_h}, \]  

(6.9)

where \( L_i^{tot} \) and \( L_h^{tot} \) are the total lengths of eyes and holes, \( n_i \) and \( n_h \) are the number of cuts located on eye and hole domains, respectively. Since we can reject that large
number of the cuts are located at the boundaries of eyes and holes, we can assume
that cut positions are located either on eyes or holes. Then \( n_i \) and \( n_h \) are equal to
the number of eye edge domains and the number of hole edge domains. The test will
be to use all DNA fragments that come from the same range of fractions \( f \), compute
\[ f_n = \frac{n_i}{n_i + n_h}, \]
and then compare \( f_n \) with the value of \( f \) that is directly estimated from
each molecular fragment (i.e. from \( L_i^{\text{tot}} \) and \( L_h^{\text{tot}} \)). The test result is shown on Fig. 6.4.

We can see that \( f_n \) increases when \( f \) increases; they follow the same trend. But
when \( f < 0.5 \), \( f_n \) tends to be bigger than \( f \); while when \( f > 0.5 \), \( f_n \) tends to be
smaller than \( f \). This may happen because when \( f < 0.5 \), the hole domain length
is big and there are some oversize hole domains that are not counted. Thus \( n_h \) is
smaller than the actual number, making \( f_n \) bigger than \( f \). For the same reason, when
\( f > 0.5 \), eye domains are big, and one counts fewer \( n_i \), which makes \( f_n \) smaller than
\( f \).

After testing the experimental data, we can say that the uniform distribution
assumption is reasonable and acceptable. But since we do not have enough data
to do a very accurate test, further tests on more experimental data would still be
desirable.

6.3 Application to experimental data

Since the assumption of uniform cuts seems reasonable, we can proceed to analyze the
experimental data of Herrick et al. in [23] using the corrected estimators for domain
lengths formed in Chapter 5. Specifically, we will compare the analysis based on the
interior estimator with that based on our new Interior-Edge (I-E) estimator.

First, I compute the average domain length and the fraction replicated from the
experimental data using the interior domain length estimator and I-E average domain
length estimator. Figure 6.5 shows the results. We can see that the difference between
interior average domain lengths and I-E average domain lengths is not significant when
the average domain length is relatively small. Only when the average domain length
is comparable to or bigger than the fragment length is the difference relatively big.
That is, we see relatively large differences only when \( f \) is small for hole lengths and
when \( f \) is big for eye lengths. In other words, from Eq. 5.2 and Eq. 5.36, we know
CHAPTER 6. APPLICATION TO EXPERIMENTAL DATA

Figure 6.5: Average domain length vs. fraction replicated. Triangles represent average domain lengths using the interior average domain length estimator, while circles represent average domain length calculated using the Interior-Edge estimator. (a) Average hole length; (b) Average eye length. We can see that the difference between the two estimators is not significant for most points, except for the first few points of holes at $f$ close to 0 and the last few points of eyes at $f$ close to 1. This suggests that the finite-size effect is not severe in this experiment.

that as $L/\langle X \rangle \to 0$, the finite-size effect is significant, $\langle X_{\text{it}} \rangle \approx 2 \langle X_i \rangle$. Thus, we can examine the difference between average domain lengths of two kinds of estimators, to determine how big the finite-size effect is. In the experimental data, the finite-size effect is not significant except at the beginning and the end of replication data. This agrees with earlier work by Jun, who estimated the importance of finite-size effects by direct Monte Carlo simulation [23].

Next, I will use both average domain length curves to extract the nucleation rate $I(t)$ and replication fork velocity $v$. The methods we use here were introduced in Chapter 2, and I use the computer code written by Jun [23].

We can see in Fig. 6.6 that the $I(t)$ curves for the I-E data and for the interior data are similar.

From Fig. 6.6, the data are flat and with little scatter at the beginning. After some time, it jumps up and shows more scatter. As in Herrick et al., we can fit the
data to a bilinear curve (two connected straight lines of slopes $I_1$ and $I_2$). To fit the extracted $I(t)$, we need three fit parameters: $I_1$, $I_2$, and $T_c$. The extracted parameters are given below in Table 6.1.

Figure 6.7 shows the theoretical average domain length curve computed from the fitted bilinear $I(t)$. Comparing with the average domain length vs. fraction replicated curve Fig. 6.8 using old method in [23]. We can see that the curve using I-E estimator fits the theoretical curve better than does the interior average domain length curve. This is also what we expected.

<table>
<thead>
<tr>
<th></th>
<th>$v$ (kb/min)</th>
<th>$I_1$</th>
<th>$I_2$</th>
<th>$T_c$ min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interior Estimator</td>
<td>0.58</td>
<td>$5.06 \times 10^{-4}$</td>
<td>$8.75 \times 10^{-3}$</td>
<td>18.1</td>
</tr>
<tr>
<td>I-E Estimator</td>
<td>0.62</td>
<td>$4.07 \times 10^{-4}$</td>
<td>$4.96 \times 10^{-3}$</td>
<td>17.4</td>
</tr>
</tbody>
</table>

Table 6.1: Extracted nucleation rate $I(t)$ and fork velocity $v$ from data analyzed using two estimators of average domain sizes.

The starting-time distributions for the two estimators were extracted using Eq. 2.10 are shown in Fig. 6.9, and Table 6.2 gives the corresponding average start times and

\[2\text{The last few points are an artefact: Monte Carlo simulations show large fluctuations as } f \to 1.\]
In this chapter, we have tested the assumption of a uniform distribution of cut positions by comparing the average interior and edge domain lengths. Based on the data we have, the uniform distribution of cut position is reasonable and thus gives us the confidence to use our previously developed theory. We then applied the theory to the experimental data of Herrick et al. and extracted various replication parameters. For this experiment, the finite-size effects are not very significant, and the extracted
Figure 6.8: Mean quantities vs. replication fraction. Triangles represent experimental average domain length calculated using the interior estimator, and the solid line is the theoretical curve derived from Eq. 2.7 using a bilinear $I(t)$, Fig. 6.6 (a). (a) Average hole size $l_h(f)$; (b) Average eye size $l_i(f)$.

parameters from the two data analysis schemes not very different. But we can see that our analysis does improve the computation of average domain lengths in two respects: first, it reduces statistical fluctuations in the average-domain-length data; second, it improves the value of the average lengths of large domains. Furthermore, we gain very detailed theoretical insight into finite-size effects. The theory given here can be used in future experiments.
Figure 6.9: Extracted starting-time distribution. Solid line is a least-squares fit to a Gaussian distribution (a) using interior data; (b) using both interior and edge data.
Chapter 7

Conclusion

In this thesis, I have introduced several mathematical problems that are inspired by finite-size effects on the measurement of length distributions in the DNA replication experiments of Herrick et al. [16]. I first showed how, given an original length distribution on an infinite length of DNA, one can compute the three fragment domain distributions (interior, edge, and oversize). Then I showed how, from measurements made on the three fragment domain distributions, one can construct an unbiased estimator for the average of the original length distributions. Unfortunately, other technical limitations of the Xenopus experiment make it impossible to compute the unbiased estimator. (One cannot use the information from oversize domains.) As a result, we introduced a biased estimator (the interior-edge estimator) that nonetheless shows lower bias than the “naive” interior estimator that had been previously used in data analysis from this experiment. As anticipated, the major improvement in the experimental results occurred at the start and end of S phase, where holes and eyes respectively are large. In addition, we have provided a method to check the experimental finite-size effects from just the experimental data. The theory can also be used to determine how big the finite-size is, by checking the difference between the interior average domain length and I-E average domain length.

The Xenopus experiments of Herrick et al. were conducted on embryonic cells, where the average origin spacing is about 6.3 kb [23]. In the cells of fully developed organisms (somatic cells), the average origin spacing is much larger. For example, the typical origin spacing for somatic human cells is 100 kb [37]. Thus, even though
recent advances in combing technologies have increased the average cut size from about 100 kb in the *Xenopus* experiment to about 500 kb [38], we still expect that the method of analysis developed here will be relevant to future experiments. In addition, the generality of the mathematical problem—the two domains can have arbitrary physical meaning—implies that there may be other, unexpected settings where the ideas developed in this thesis may be applied.
Bibliography


BIBLIOGRAPHY


