2010

Mathematical modeling of eukaryotic gene expression

Tang, Shouchun (Terry)

Lethbridge, Alta. : University of Lethbridge, Dept. of Chemistry and Biochemistry, 2010

http://hdl.handle.net/10133/2567

MATHEMATICAL MODELING OF EUKARYOTIC GENE EXPRESSION

Terry Tang
Bachelor of Science, University of Alberta, 2003
Master of Science, University of Lethbridge, 2006

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfillment of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

Department of Chemistry and Biochemistry
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

To Ren Xiaomeng

Abstract

Using the Gillespie algorithm, the export of the mRNA molecules from their transcription site to the nuclear pore complex is simulated. The effect of various structures in the nucleus on the efficiency of export is discussed. The results show that having some of the space filled by chromatin near the mRNA synthesis site shortens the transport time. Next, the complete eukaryotic gene expression including transcription, splicing, mRNA export, translation, and mRNA degradation is modeled using delay stochastic simulation. This allows for the study of stochastic effects during the process and on the protein production rate patterns. Various protein production patterns can be produced by adjusting the poly-A tail length of the mRNA and the promoter efficiency of the gene. After that, the opposing effects of the chromatin density on the seeking time of the transcription factors for the promoter and the exit time of the mRNA product are discussed.

Acknowledgments

My supervisor, Marc Roussel, through his vast scientific knowledge, has always been very helpful in guiding the path of my research, answering my technical questions, and being supportive of my research ideas. Rui Zhu, Marc's former post-doctoral fellow, provided me with part of the Matlab code for my simulations. My committee members, Steve Mosimann and Brent Selinger, provided me with feedback on my research results every semester. Paul Agutter and Denys Wheatley from Britain advised me on the first draft of my paper. Tony Russell, Doug Bray, and Roy Golsteyn were helpful in answering my biology questions, especially before my comprehensive examination. I also thank all the past and present group members for our discussions together.

Contents

Approval/Signature Page ii

Dedication iii

Abstract iv

Acknowledgments v

Table of Contents vi

List of Tables viii

List of Figures ix

1 Introduction 1
- 1.1 Eukaryotes versus Prokaryotes 1
- 1.2 Traffic through the Nuclear Envelope 4
- 1.3 Gene Expression 13

2 Messenger RNA Nuclear Export 16
- 2.1 Introduction 16
- 2.2 Methods 16
- 2.3 Results 23
  - 2.3.1 Varied chromatin depth with density gradient 23
  - 2.3.2 Varied chromatin depth and mRNA synthesis site with density gradient 30
  - 2.3.3 Varied chromatin depth with constant density 32
  - 2.3.4 Varied chromatin depth with fixed synthesis site 33
  - 2.3.5 Varied chromatin density 33
  - 2.3.6 Varied mRNA synthesis site with fixed chromatin layer 34
  - 2.3.7 Varied mRNA synthesis site 35
  - 2.3.8 PLF proficiency 37
  - 2.3.9 Discussion 38

3 Model Fitting and Tail Study 39
- 3.1 Method 39
- 3.2 Results 41
  - 3.2.1 Weibull fitting 41
  - 3.2.2 ccdf test 42
  - 3.2.3 Mean residual lifetime test 43
  - 3.2.4 Kozubowski algorithm 45
  - 3.2.5 Jackson method 45

3.3 Discussion 45

4 Delay SSA **49**
4.1 Method 49
4.1.1 Reaction equations 49
4.1.2 Parameter adjustment 51
4.2 Implementation 53
4.3 Results 55
4.3.1 Protein production distribution 55
4.3.2 Stochastic nature of gene expression 59
4.3.3 Change in algorithm 64
4.3.4 Noise in transcription 69
4.3.5 Noise in transcription with longer activation 72
4.3.6 Protein expression pattern 73
4.3.7 Discussion 79

5 Activation and Exit **81**
5.1 Method 81
5.2 Results 83
5.3 Discussion 90

6 Conclusion **93**

Bibliography **97**

List of Tables

2.1 Summary of parameter setups for chapter 2 24

3.1 Jackson method results for constant chromatin density 46
3.2 Jackson method results for gradient chromatin density 46
3.3 Results of Jackson's method for section 2.3.2. 46
3.4 Results of Jackson's method for section 2.3.8. 46

5.1 Rate constants for different species of molecules 83
5.2 The standard errors of the seeking times at various chromatin-layer DPCs . 86

List of Figures

2.1 10×10 model space . . . 18

2.2 Fitting with a Weibull distribution . . . 25

2.3 Visitation time map for the molecule that takes the longest to exit . . . 26

2.4 Visitation frequency for the molecule that made the most movements . . . 26

2.5 Overall visitation time map with chromatin density gradient . . . 27

2.6 Cumulative exit time distribution at various dense chromatin depths with a chromatin density gradient. . . . 29

2.7 Comparison between simulations in a 50×50 and a 100×100 space . . . 29

2.8 Comparison between 2D and 3D space . . . 30

2.9 Same as figure 2.6 except that the mRNA synthesis site moves to stay above the dense chromatin layer . . . 31

2.10 Same as figure 2.9 except with a constant diffusion probability in the chromatin layer . . . 32

2.11 Overall visitation time map with constant-density chromatin . . . 33

2.12 Same as figure 2.10 except that the mRNA synthesis site is fixed . . . 34

2.13 The cumulative exit time distributions at different DPCs . . . 35

2.14 Cumulative exit time distributions as the mRNA synthesis site varies . . . 36

2.15 Cumulative exit time distributions as the mRNA synthesis site varies . . . 36

2.16 Cumulative distribution of exit times with varying left-right to up-down diffusion probability ratio . . . 37

3.1 Weibull fitting of exit times from section 2.3.3 . . . 42

3.2 Same as figure 3.1 except that the y-axis is replaced by the negative of the natural logarithm of the ccdf . . . 43

3.3 Mean residual lifetime test . . . 44

3.4 Same data as figure 3.1 fitted with the lognormal distribution. . . . 48

4.1 Protein production distribution with 70 ribosomes . . . 55

4.2 Protein production distribution with 700 ribosomes . . . 57

4.3 Cytoplasmic mRNA level with 70 ribosomes . . . 58

4.4 Cytoplasmic mRNA level with 700 ribosomes . . . 59

4.5 Two simulations run with the same parameters and initial conditions as in figure 4.1. . . . 62

4.6 Two simulations run with the same parameters and initial conditions as in figure 4.2 . . . 63

4.7 Protein quantity comparison between fixed and random transcriptional delays 65

4.8 Protein and mRNA quantity distribution with identical parameters and initial conditions . . . 70

4.9 Protein distribution with τ_1 and τ_2 drawn from distributions . . . 71

4.10 The protein quantity distribution of the set of simulations with 3 pre-mRNA molecules and no active transcription. . . . 72

4.11 Same as the lower panel of figure 4.8 except that the activation time is 10 instead of 1 minute. 73
4.12 Same as figure 4.10 except the starting pre-mRNA number is 13 instead of 3 74
4.13 Protein synthesis pattern at default parameters 76
4.14 Protein synthesis pattern with weaker promoter 77
4.15 Protein synthesis pattern with very weak promoter and longer activation time 77
4.16 Protein synthesis pattern with weaker promoter and shorter poly-A tail . . . 78

5.1 Exit times with DPC ratio of 1.3 between TF and mRNP 84
5.2 The lower part of figure 5.1. 85
5.3 Normalized histogram of the seeking time for various DPCs 87
5.4 The same as figure 5.2 except with more runs 88
5.5 Same as figure 5.4 except with 10-fold slower mRNP addition rate. 89
5.6 Same as figure 5.2 except that the DPC ratio between the mRNP and the TF is 2 90
5.7 Same as figure 5.2 except that the DPC ratio between the mRNP and the TF is 0.5 91

List of Abbreviations

ccdf	complementary cumulative distribution function
cdf	cumulative distribution function
CV	coefficient of variation
DPC	diffusion probability constant
Dscam	down syndrome cell adhesion module
MRL	mean residual lifetime
mRNP	messenger ribonucleic protein
NES	nuclear export signals
NLS	nuclear localization signal
NPC	nuclear pore complex
NTP	ribonucleoside triphosphate
PD	probability density
pdf	probability distribution function
PLF	pore-linked filament
RBS	ribosome binding site
SE	standard error
SSA	stochastic simulation algorithm
TBP	TATA-binding protein
TF	transcription factor

Chapter 1

Introduction

1.1 Eukaryotes versus Prokaryotes

Ever since the discovery of deoxyribonucleic acid (DNA) as the genetic material [4] and Watson and Crick's demonstration of the double helical structure of DNA [71], a tremendous amount of work has been devoted to further the study of its structure and function. As a result, we have gained a wealth of knowledge on DNA sequences and their meanings.

However, billions of years of natural history with constant pressure on the struggle for survival in a constantly changing environment has created solutions that are not always simple for humans to understand. In the eukaryotic nucleus, for example, despite the progress made on sequencing genomes, the three-dimensional structures of the chromosomes and of the nucleus encasing them are still not well understood, especially as these structures relate to their functions. These structures are important in shaping or facilitating the biological processes that take place in the nucleus. In this thesis, I am interested in studying how the various components and their positions affect the export process of messenger ribonucleic acid (mRNA) molecules. For example, are there specific reasons for a gene to be at a certain place in the nucleus or is the physical location of any piece of DNA totally random?

A eukaryotic cell, unlike a prokaryotic cell, has most of its genetic material enclosed in one compartment, the nucleus. The reason for this arrangement might be historical in that the nucleus is probably the descendant of one prokaryote that was internalized into another; over time a symbiotic relationship developed and the genetic material of the host cell was eventually lost [44]. Similar stories can be told for mitochondria and chloroplasts, both of which have DNA of their own. Other theories on the origin of the nucleus include the "viral eukaryogenesis" [63] and the "exomembrane hypothesis" [19]. Regardless of the origin of the nucleus, the reason that it exists is that it serves a function that is favorable to the survival unit that contains it.

One of the advantages of having a nucleus is that it allows the separation of transcription and translation. The transcription and translation of a prokaryotic mRNA often take place concurrently by having translation started at the 5' end while the 3' end of the mRNA is still being transcribed [47]. This allows for the quick production of proteins with a drawback: for a cell to have more genes, it will generally be necessary to increase the size of its DNA, and the increase in DNA size will roughly be linearly proportional to the increase in gene number. This inhibits the increase in gene number for prokaryotes. Eukaryotes, on the other hand, have their transcription and translation separated physically by confining the DNA in the nucleus and having translation occur in the cytoplasm. This separation allows the mRNA to be processed before being translated, and through alternative splicing, one gene can produce more than one polypeptide. An extreme case of this feature involves a gene called Dscam (Down Syndrome Cell Adhesion Module, whose proteins, DSCAM, are linked to Down Syndrome in humans) studied mainly in *Drosophila* [73]. This gene has 16 exons and potentially can produce 38016 different protein isoforms. During neuron formation, to ensure that the axons from a neuron branch out properly, the Dscam mRNA is spliced randomly so each neuron contains its own set of Dscam protein isoforms, all of which are mainly extracellular due to a transmembrane domain. The extracellular domains of identical isoforms bind to one another to induce a contact-dependent repulsive reaction, and because the chance of two neighboring neurons carrying the same isoform is small, the repulsive interaction ensures that the sister branches from the same neuron do not cluster. For a prokaryotic system to produce the same number of different proteins, it would typically have to possess the exact number of genes, whereas in this case the eukaryotic system does it with only one gene. This is perhaps one of the reasons why there are no truly multicellular prokaryotes: the length of DNA alone can be a prohibiting factor.

Surely, one could argue that to have a larger genome is not at all impossible if only the cell increases its size. This bring up the second advantage of the nucleus: compartmentalization. To increase the genome size in prokaryotes in order to obtain more functions is plausible but then the size of the cell would have to grow accordingly to accommodate the DNA material. Besides the fact that the surface to volume ratio decreases with increasing cell size, which puts a higher burden on the plasma membrane to traffic substances and on the membrane-bound enzymes to carry out reactions, a bigger cell also makes it harder for potentially interacting molecules to find each other. The expected value of the search time to find a target that has a radius of A in a spherical space of radius R is [39]

$$\frac{R^3}{3AD}\left(1-\frac{9A}{5R}\right),$$

with D being the diffusion coefficient, which means that the search time increases roughly proportionally to the cube of the cell radius. Even if this equation does not apply quantitatively to the intracellular space, it still shows qualitatively the correlation between volume and search time. For prokaryotes, every freely diffusing molecule potentially has the entire cell to cover in order to find its targets, even though its targets make up only a tiny portion of all the molecules it could potentially meet during the progress of the search. Despite some ingenious strategies to reduce the need for target finding such as in the case of the *lac* operon [68] where the three genes responsible for lactose metabolism are together and controlled by one promoter and one operator to reduce the need for RNA polymerase to find them separately, an increase in cell size still drastically increases the delay for biochemical reactions. The solution adopted by a eukaryotic cell is to physically compartmentalize its functions by confining molecules required for a common purpose at one location that is bounded by membranes. In the case of the nucleus, the function is to supply the cell with RNA and RNA-related products (such as ribosome subunits) and to do that, it contains

DNA and all the components required for transcription, post-transcriptional processing, and export so that an RNA polymerase, for example, only has to search the volume of the nucleus, which is comparable to that of a prokaryotic cell, to find its target despite the fact that the size of the entire cell is much larger.

The advantages of a eukaryotic cell over a prokaryotic cell might seem overwhelming but the reality is that prokaryotic life forms as a whole are nothing short of thriving. There are 10^{14} bacterial cells in a typical human body whereas the same body only contains about 10^{13} human cells [61] so it might not be entirely unreasonable to say that a human (and many other animals, for that matter), is a symbiotic life form containing mainly bacterial cells with 10% eukaryotic cells. One advantage of a prokaryote is that simultaneous transcription and translation allow for faster response to the environment, and, while the environment is suitable, a smaller genome and a simpler architecture allow for faster division. On the other hand, a eukaryotic cell in most cases has to accept the combined delay in transcription, processing, mRNA export, and translation. This thesis will use various mathematical tools to study gene expression in a typical eukaryotic cell with emphasis on the mRNA transport process, which has not been focused on as much as the other steps.

1.2 Traffic through the Nuclear Envelope

The nucleus is enclosed by the nuclear envelope which is made of a double lipid bilayer. Perhaps due to its prokaryotic origin and to the process of endocytosis, the nuclear envelope consists of two layers of membrane. The lipid bilayer is mainly made of phospholipid with hydrophilic heads on the surface and hydrophobic tails at the center of the membrane. The physical properties of the bilayer mean that it is impermeable to macromolecules such as proteins and mRNAs. The plasma membrane, made of the same material as the nuclear membrane, has ways to exchange these macromolecules through endocytosis and exocytosis. Exocytosis is performed by confining the macromolecules to be exported in a cytoplasmic vesicle with a phospholipid membrane. By inducing a fusion between the membrane of the vesicle and the plasma membrane, the content inside the vesicle is released to the extracellular space. Endocytosis works in reverse. This is a great solution to transport macromolecules through the plasma membrane and the same strategy could be adopted to transport macromolecules from the nucleus to the cytoplasm: this would involve first budding a vesicle from the inner nuclear membrane into the perinuclear space and then inducing a fusion between the vesicle and the outer membrane. However, since there is no other reason for the vesicle to exist in the perinuclear space, it would be more convenient for one vesicle to be budding and fusing at the same time and that introduces a pore. The nuclear pores are supported by nuclear pore complexes (NPCs).

The NPC connects both of the nuclear membranes and makes them continuous locally. The number of NPCs varies significantly between organisms and even between different cell phases depending on the demand for molecule trafficking between the cytoplasm and the nucleoplasm. The number of NPCs in human cervix tissue nuclei is about 4000 [46] and that in yeast during early mitosis is about 140 [72] which means that the NPC densities are on the same order of magnitude because the volume ratio between the two cells is about 100. The shape of an NPC is symmetrically octagonal. The length of an NPC (the dimension orthogonal to the membrane) is 15 nm and the diameter is 120 nm [72]. The size of the channel, however, depends on perspective in the following ways. Molecules that are smaller than 5 kDa are allowed to freely diffuse through the pore. In this sense, the diameter of the channel is about 9 nm [2]. Molecules that are between 5 and 60 kDa can still rely on passive diffusion to pass through but with increasing resistance. For example, a protein of 17 kDa takes 2 minutes to equilibrate between cytoplasm and nucleoplasm [2, 18]; a protein of 44 kDa takes 30 minutes to equilibrate [18]. In this sense, the diameter of the NPC is between 10 and 20 nm. The nuclear membrane is considered impermeable to molecules that are larger than 60 kDa [2, 18, 43]. To put it in perspective, a mammalian ribosomal subunit has a molecular mass in the range of megadaltons [68]. These subunits are assembled in the nucleus and exported to the cytoplasm to participate in translation so there has to be a way of getting these large molecules through the nuclear envelope.

If one were to design the NPC to keep the nucleoplasm and the cytoplasm segregated in terms of their solute concentrations while at the same time selectively letting macromolecules pass through, there would be a dilemma: if the diameter of the NPC were large enough, the entire contents of both sides would be mixed by way of Brownian diffusion; if on the other hand the diameter of the NPC were small, the macromolecules could not get through. The same situation is faced elsewhere in biology. An example is in the human digestive system: food enters the stomach through the esophagus while the gastric juice in the stomach in most cases has to be prevented from entering the esophagus where it would cause inflammation, but the gastric juice contains particles that are much smaller than a food bolus. The solution to this is to have a valve at the inlet of the stomach called the esophageal sphincter that permits materials through selectively. A similar mechanism is believed to operate in the NPC. As demonstrated by Panté and Aebi with gold-labeled proteins that have sizes larger than allowed to diffuse through the NPC, it seems that the center of an NPC is gated and that the gate opens in response to signals because proteins of similar size without the signal cannot get through the NPC. Moreover, several binding sites along the channel can be identified under the electron microscope [51].

The signal for protein import is called the Nuclear Localization Signal (NLS) , first discovered in 1984 [36]. Proteins contain the signal in one or two clusters [20]. The sequence of the signal is not highly conserved and can contain many combinations of arginine and lysine although not all combinations are equally effective. After all, the NLS is simply a signal to be recognized by another protein called the nuclear import receptors, which facilitate import. The NLS and the nuclear import receptors are subject to coevolution. The sequence of the NLS can mutate as long as its receptor can recognize it. The location of the NLS in the protein to be imported does not seem to have much significance. The signal sequence can be inserted almost anywhere in the protein sequence (before folding) and will function [2]. Some signal sequences can function properly even when covalently linked to a side chain of an otherwise cytosolic protein [35]. Given that the signal recognition process requires protein-protein interaction, it is natural to assume that the NLS must be located at the surface of a protein. The positively charged lysines and arginines strongly favor a surface location and this is perhaps the reason that charged amino acids evolved to be part of the sequence. Protein export from the nucleus to the cytoplasm takes place in similar fashion as import with the nuclear export signal and the nuclear export receptor.

The mRNA molecules pass through the NPC from the nucleus to the cytoplasm in different fashions than do the common proteins (even the ribosomes) for several reasons. First, the size of an mRNP (messenger ribonucleic protein), which is mRNA bundled with proteins, can easily have diameters of over 30 nm or even over 40 nm [14]. To have all the mRNPs pass the same way as the proteins is impossible because the channel only allows through proteins up to about 25-30 nm [2] in diameter. Second, the mRNA does not need to go through the channel the same way as the proteins do. The reason that the NPC channel has to accommodate the diameters of the proteins is that proteins have to function in a folded state, which means that their structures, from primary to quaternary, are all critical in their functions. The folding process of proteins is an elaborate process and preserving the tertiary and quaternary structure of a folded protein is important in keeping its function. On the other hand, an mRNA is useful mainly through its primary sequence so it can be folded and unfolded many times and in many ways, and so long as its sequence is intact, it is fully functional. This nullifies the reason to try and keep an mRNP's tertiary and quaternary structure consistent across the nuclear envelope. The same reasoning can also determine the way by which a ribosomal subunit exits the nucleus. Although a ribosomal subunit resembles an mRNP in that they are both a combination of RNA and proteins, it functions in very different ways from an mRNA. A ribosome, similar to a common protein, has to rely on its tertiary and quaternary structure to function whereas an mRNA is only useful in its primary sequence. This difference in function determines that a ribosomal subunit has to pass through the NPC channel while folded and an mRNA does not. It would be interesting to speculate as to whether it is a coincidence that the maximum size of a ribosome subunit is about 30 nm [2] and the maximum diameter of the NPC channel is about the same.

The export of the mRNP is not completely understood but at least for some of them, the molecule is partially unfolded and the 5' end of the mRNA goes through the channel first and then the rest of the strand follows. As the strand is still going through, the emerging 5' end is already bound by ribosomes on the cytoplasmic end [14, 17]. This scene of the 5' end of an mRNA being bound to the ribosomes while the rest of the strand is still emerging might remind people of the transcription of prokaryotic mRNA in that the 5' end emerges first and signals for translation to start while the rest of the molecule is still being transcribed. This commonality might not be a complete coincidence. First of all, since all translations take place from the 5' to the 3' end, having the 5' end available first to the translational machinery will make the translation process start earlier than having any other part of the mRNA emerge first from the nucleus. If the symbiotic theory for the origin of the nucleus is true, one simple way for the guest organism (the ancestor of the nucleus) to have its genes expressed when it first entered the host would have been to have its mRNA behave to the host in a way that mimics the host mRNA by having the 5' end emerge first. This way, it requires no special functions on the side of the host to express the guest gene. If the mRNA is instead exported as a folded structure like the proteins are, possibly inside an mRNP, there has to a mechanism in the host cytoplasm to unfold it to a degree so that the ribosome can bind either to an initiation sequence or to the 5' end.

Before going through the nuclear pore complex, an mRNA molecule has to find its way through the nuclear environment. This process is still not well understood. The classical model of the structure of the nucleus is that the chromatin is loosely organized and is distributed randomly throughout the subnuclear space and that the substances within the space rely on Brownian diffusion to reach their targets to carry out their functions. This model has been repeatedly shown to be overly simplified [42]. A network of filamentous structures emanating from the nuclear pore complexes (NPCs) and extending to the nucleoplasmic side has long been observed under the electron microscope [22]. It features a fibrogranular network whose structure contains ribonucleoprotein [62] and is very dynamic [66], which is different from the rigid cytoskeleton. The presence of actin and its ability to polymerize within the nucleus have been shown [37, 41] which suggests that it could be a component of the filaments. These filaments are sometimes considered to be an extension of the NPC [25], but to avoid confusion with the NPC in the classical sense, they are given a separate name: pore-linked filaments (PLFs). Very clear images of PLFs were recently taken inside *Xenopus* oocyte nuclei [37]. It is also known that particles coated with a protein containing nuclear export signals (NES) diffuse to the nuclear pore complex (NPC) more quickly than otherwise [23]. One of the proposed mechanisms for this quick exit is by having the NES-containing particles attached to the PLFs which provide tracks along which particles can move [23, 24, 37] although it is not well understood whether the NES itself is directly attached. Another possible mechanism is for the PLFs to compartmentalize the sub-nuclear envelope region to locally trap the particles as some microfilaments do in cytoplasm [6] because compartmentalization enables a molecule in the nucleus to find its target more efficiently by means of non-directed random movements [39].

Outside of the PLF layer, the most dominant structure in the nucleus is the chromatin. There are, in general, two types of chromatin: heterochromatin and euchromatin. While most of a cell's genes are located in the euchromatin region, heterochromatin exists to provide structural support. Heterochromatin was distinguished from euchromatin initially because it appears darker under a microscope. The reason is that its DNA is densely packed and can be considered impermeable to macromolecules, which is one of the reasons why its genes are not expressed. Euchromatin, on the other hand, is organized into loops and contains active genes. It is likely that mRNP molecules, on the way to finding the NPC, can move into a region where the chromatin filaments are dense and become immobile. The stalled molecules require ATP expenditure to free them even though the mRNA transport process itself is not active [67]. The exact mechanism in freeing stalled mRNP molecules is not clear but what is known is that ATP depletion in the nucleus decreases the fraction of mobile mRNP particles [48]. There are suggestions of possible mechanisms for this observation: one is that the ATP is directly involved in breaking the association between mRNP and the chromatin; another is that the chromatin structure is dynamic and reorganizing it, requiring ATP, is able to free the stalled mRNP [67].

The chromatin tends to organize itself into discrete territories called chromosome territories because the chromatin in each territory is from the same chromosome. Outside the chromosome territories is the continuous interchromatin compartment [16]. An active gene, in addition to being already located at the euchromatin portion of the chromosome, is often spatially relocated out of its chromosome territory and associates with other distant active genes from the same or different chromosomes to form a transcription factory [74]. A mammalian cell typically has 15000 active genes at any given time whereas the number of transcription sites is of the order of thousands [2]. Because mRNA splicing is co-transcriptional, a transcription factory is also where splicing takes place. The advantage of having a transcription factory over the classical view of transcription, which entails the RNA polymerase randomly searching for active sites, is probably efficiency. In the classical model, an RNA polymerase spends some time scanning DNA in search of an active site, transcribes it, and searches again. The entire workload is on the polymerase. By allowing the active gene to relocate, a polymerase could be transcribing a gene with another gene queuing while a third gene with some of the activation factors bound is searching for the factory. This way, the burden of searching does not interfere with the actual process of transcription. It also allows several transcription factors to co-localize to reduce the time needed for them to find genes. With the active genes associating and dissociating frequently, a transcription factory is highly dynamic.

There could be another reason why an active gene is relocated out of its chromosome territory. A chromosome territory contains chromatin fiber that presumably exerts a higher level of hindrance to the movement of macromolecules such as RNA polymerase and mature mRNP. This makes it difficult for RNA polymerase to enter and to find the active gene; at the same time, it is also harder for mRNP transcribed in this environment to exit in order to find an NPC. In order for the mRNP to be freed from entanglement, ATP is involved which is costly to the cell. One possible remedy to this problem is to make the euchromatin less dense in order for molecules to get in or out more easily. One problem with a less dense euchromatin is that it allows more mRNP to enter to be entangled, though the entanglement is not as severe. The effect of the adjustment of the density of the chromatin is tested in chapter 2 of this thesis. Another problem is that a less dense chromatin means a bigger nucleus to house the same amount of DNA content, which means more difficulties in target-finding for the molecules in the nucleus. On the other hand, another possible remedy is to make the euchromatin more dense to exclude mature mRNP from entering. This, however, would make it more difficult for the activators to move in it to find the active genes. A high-density chromatin does exist and it is called the heterochromatin, which contains fewer genes, many of which are silent. The activity of a gene in relation to the physical density of the chromatin allows the cell to add a higher level of gene control by organizing some genes in the heterochromatin region to silence them. While some parts of the DNA in a cell are permanently organized as heterochromatin, others can dynamically switch between being euchromatin and heterochromatin in time [2]. The current euchromatin system is a compromise to ensure a reasonably high level of transcription and an acceptable speed of mRNA exit. This is achieved by keeping the euchromatin at a certain density, relocating the active genes away from the chromatin, and freeing the stalled mRNP at the expense of energy. The level of improvement in mRNA exit time by having the transcription factory located outside of the chromosome territory is shown in section 2.3.

While it is clear that a higher chromatin density exerts a higher hindrance on the diffusion coefficient of the molecules buried in it, arguments can be made as to whether the density of the chromatin territory is homogeneous or whether there is a density gradient. The euchromatin is made up of DNA loops extended outward from the inner regions of a chromosome territory, so if the outer edge has the same number of loops as does the inner region, because these DNA loops cover more area near the edge, the density in that region should be less. However, the material of the euchromatin is not so rigid that it extends in a straight line; instead, it is flexible and can fold onto itself. Almost like free diffusion, the chromatin material can diffuse in space to make its distribution tend towards being homogeneous in its chromosome territory. Microscopic images show that chromatin in reality is a combination of both effects: the diffusive effect to achieve homogeneity and the extension to make the center more dense. For the most part, it has a homogeneous density, and near the edge, its density decreases [67]. In this thesis, both of these situations are simulated.

With chromatin that has a density gradient, it is not clear whether a thick layer will help or hinder mRNA exit because while the mRNA molecules can diffuse into the chromatin region and experience hindrance, the gradient also favors the molecules' existing in low density regions because between the possibilities of diffusing into either a low or a high density region, the former is easier. This effectively decreases the space that an mRNA has to explore to find the NPC. The overall effect of having a thicker layer of gradient chromatin region on the efficiency of export is studied in section 2.3.1.

With a chromatin layer that has homogenous density, it could also assist the molecules' exit by excluding some of them from the space filled by the chromatin. Unlike the gradient chromatin which favors the existence of molecules in low density regions, the homogenous chromatin provides equal probability in every direction for the molecules that are in it. If the chromatin layer is thin enough, a molecule that is in it still has a reasonable probability to get out within a reasonable amount of time; and once out, it is not so easy to diffuse back in due to the density differential at the interface. However, diffusion of the molecules that are in a thicker layer of chromatin could be impeded for a long time. Finding a function to fit the exit probability distribution could be difficult because it may not bear significant resemblance to any common models for exit time distributions. What makes this model unique is the fact that the molecules behave differently in different regions which is uncommon for most waiting situations. The study of the effect of a homogenous chromatin layer is shown in section 2.3.3, and the effort in finding a probability distribution function is recorded in chapter 3.

1.3 Gene Expression

The mRNA exit time is one of the factors that could have an effect on the protein production rate. Another factor is the transcription time which can be controlled by the length of the transcription unit. Though it may seem that gene expression speed is important to an organism (after all, it is one of the reasons why the prokaryotes are successful), it is more important for the eukaryotes to set the timing and rhythm correctly. Prolonging transcription as a means of delaying gene expression is important in a system that demands accurate timing such as in the developmental stage of a multicellular organism. For example, one strategy to ensure that one gene is always expressed earlier than another is to have the two physically located near each other (or even share part of the transcription unit) and also to ensure that the one that is supposed to appear later has the longer transcription unit. In

Drosophila, the gene *E74A* appears after *E74B* by sharing the same 3' end but with different promoters and transcription unit lengths. The lengths of the two genes are conserved between two species of *Drosophila* whereas the intron sequences are not [34, 64].

Other major factors affecting the synthesis rate of the protein product of a gene are translation delay and the mRNA decay in the cytoplasm. Each step during the gene expression process is constituted of simpler steps. For example, the transcriptional process involves individual movements of the RNA polymerase along the DNA template while adding nucleotides to the elongating RNA; each addition of a nucleotide is subsequently dependent on the acquisition of the correct ribonucleoside triphosphate (NTP) which in turn depends on the local concentration of the NTPs. The stochastic nature of the positions of the molecules involved makes certain that transcription is a stochastic process. Overall, the entire gene expression is also a stochastic process.

If one were to model transcription with the finest detail, one should at least consider the diffusion of RNA polymerase and the NTPs and the probabilities of them binding each time they collide for each nucleotide addition. However, this level of detail is computationally costly, to implement. One way to conceal the less important details in exchange for a significant gain in speed is by a stepwise model [58] that uses the Gillespie stochastic simulation algorithm (SSA) [28, 29]. This model characterizes each nucleotide of the coding strand as being unoccupied by the polymerase, occupied, or active, and a stochastic rate constant is associated with the conversion between two states. The SSA is then employed to stochastically determine the next step to occur and the associated time. This model is able to simulate transcription of one or a few copies of gene(s) within a reasonable amount of time.

The fact that the biological processes at cellular or subcellular level are intricately interconnected dictates researchers' goals in studying them. Once a process is understood well enough at the micro-level with enough details, the next step is almost always to grasp how this process inter-plays with others and to appreciate the significance it plays in the biological system at large. While the detailed stepwise transcription model using the SSA models transcription well, to place it in the larger gene expression process with many genes would require quite some time to complete. This is especially so because each gene being expressed gives rise to a number of copies of mRNA, each of which has the potential to be translated. Translation bears enough resemblance, in terms of delays, to transcription mechanistically to also require the use of this model. Moreover, to characterize the time distribution of a gene from transcription to the appearance of its protein products, the simulation needs to be repeated many times to obtain a statistically significant set of results. These factors warrant a faster model.

One way to simplify the model is by realizing that stochasticity is involved in the onset of a process, such as the binding of the polymerase in the case of transcription. With no regard to what happens during the elongation process, the mRNA is released after a waiting period. In addition to elongation, this period includes transcription initiation (failed or successful), polymerase pausing (if any), and transcription termination. The binding process has the same attributes as two molecules finding each other in any chemical reaction; and the waiting period can be accounted for by having a time delay associated with the release of the mRNA. In this sense, the components in the gene expression process are equivalent to a delayed reaction, which differs from regular chemical reactions in that the reaction rate constant only determines the consumption of the reactants, and the products appear some time after the reactants are consumed. The method and results in implementing the delay SSA on the eukaryotic gene expression process is recorded in chapter 4. The impacts of various other factors on protein production such as noise, promoter efficiency, and length of poly-A tail are also shown in chapter 3.

Chapter 2

Messenger RNA Nuclear Export

2.1 Introduction

This chapter shows a mathematical model for mRNA migration in the nucleus from the site of synthesis to a nuclear pore where it is exported. Various factors can affect the mRNA export efficiency. These factors include the thickness and density of the chromatin layer, the thickness and efficiency of the proposed PLF layer [23, 24, 37], and the location of the mRNA synthesis site relative to the nuclear envelope. The method used to simulate the molecular movements is the Gillespie algorithm [28, 29]. Because mRNA molecules have different sizes and size is one of the main factors that influences the diffusion coefficient, the stochastic rate constant that determines the frequency of movement in the Gillespie algorithm is dimensionless and chosen arbitrarily. This means that the simulated time is also dimensionless. The details of the model are described in the next section.

2.2 Methods

The mRNA exit time is modeled with a two-dimensional square space that represents a portion of the nucleus whose top and bottom are bound by the normally impenetrable nuclear envelope and the heterochromatin layer, respectively. The sides are periodic boundaries to account for the input from and output to the nearby space. The simulation method used here is stochastic simulation of individual mRNAs. This method is used over solving the diffusion equations because the full probability distribution for exit times is not easily recoverable from solutions of the diffusion equation. The space has a lattice-like structure: Square cells are the basic units and each cell is in contact with four others as shown in figure 2.1. Each cell has two states: free and occupied. A particle has the possibility to diffuse to a nearby cell only if that cell is vacant. The left and right boundaries are periodic and the upper and lower ones are impermeable except at the mRNA exit sites as indicated in the figure. The only way for a molecule to exit is by occupying one of the exit sites and moving upward as the next step. The Gillespie algorithm [28, 29] is used to generate the diffusive motion on this lattice. It takes the reaction probabilities used in Gardiner's spatio-temporal master equation [26] to determine the next event (reaction or diffusion) and its time of occurrence. Therefore, the Gillespie algorithm is a stochastic realization of the probability evolution given by the master equation. The stochastic rate constant in the Gillespie algorithm is 0.4 in the space between the dense chromatin and the PLF layer (call it the middle space). The rate of synthesis at the addition site is also governed by the Gillespie algorithm and its rate constant is 8×10^{-5} unless otherwise stated. This small number ensures that there is a small number of molecules in the system at any given time and thus that the effect of collisions is negligible. The nuclear environment is crowded with macromolecules and the diffusion of macromolecules in such a space is complex. It depends on many factors that are difficult (but possible) to measure such as the aggregation level of the background molecules [30]. It is assumed here that the hindrance to a particular molecule caused by collisions with other macromolecules can be factored into the density of the nuclear environment. Interactions between mRNA molecules from the same gene are negligible.

The Gillespie algorithm works as follows: first, determine all the possible combinations of molecules in a space that can react; second, determine the likelihood (propensity) for each reaction; third, based on the total propensity and a randomly generated number, generate a random time for the next reaction to occur; last, based on the magnitudes of the propensities and another randomly generated number, determine which combination is the

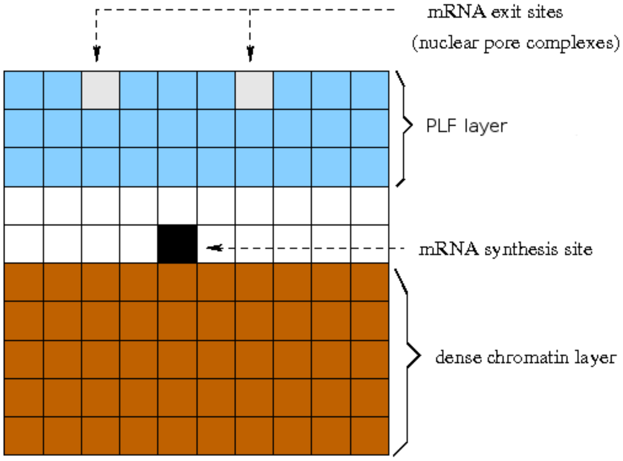

Figure 2.1: A 10×10 lattice-like space which contains two mRNA exit points and one mRNA addition point with actin filaments on the upper moiety and chromatin on the lower one.

next reaction. For example,

$$A + B \xrightarrow{k_1} C \xrightarrow{k_2} D$$

has two reactions. If a, b, c, and d represent the molecule numbers of their respective species, the propensity for the first reaction is k_1ab, and that of the second reaction is k_2c, where k_1 and k_2 are stochastic rate constants. The time it takes for the next reaction to occur is

$$\tau = (1/a_0)\ln(1/r_1)$$

where $a_0 = k_1ab + k_2b$ and r_1 is the random number drawn from the uniform distribution in the unit interval. Using another uniformly distributed random number r_2, the choice of the next reaction is determined: if $r_2 < \frac{k_1ab}{a_0}$, it is the first reaction; if $r_2 > \frac{k_1ab}{a_0}$, it is the second reaction.

For a diffusing molecule the diffusion probability constant (DPC) of each possible movement is determined by the environment the molecule is in. For a molecule immersed in the middle space, the probability constants for it to move in all directions are identical; for a molecule immersed in a chromatin layer with uniform density, its probability constants are uniformly lower; and for one that is at the interface between the middle space and the chromatin layer, there are different probability constants with respect to ascending and descending movements because it takes more time to travel the same distance into a denser region (the chromatin layer) than into a less dense one (the middle space). Table 2.1 gives a summary of the DPC settings. More details can be found in later sections.

To calculate the time τ:

$$\tau = (1/a)\ln(1/r_1)$$

where a is the sum of the propensities and r_1 is the random number from the uniform distribution in the unit interval.

The goal of the simulation is to obtain the individual times for a molecule in the system to exit through one of the openings while different parameters are varied in order to understand the influences of various nuclear structures on the exit times.

The possible role of the PLFs is tested by having a bias in the diffusion probability constant at the top filamentous layer in the directions perpendicular to the nuclear envelope. The dense chromatin is located at the bottom of the simulation space: the diffusion probability constant is smaller there than in the middle space (between the PLF and dense chromatin). This is implemented in one of two ways:

- There is a probability constant gradient across the depth of the chromatin layer ranging from that of the middle space at the top to a value of 0 at the bottom of the dense chromatin layer.

- There is a fixed, smaller probability constant in the entire chromatin layer.

In some of the scenarios, the mRNA synthesis site is consistently located outside of the chromatin layer whereas in others its position can be either inside or outside of the dense chromatin layer.

The coordinate system is thus: The top-left cell in the square space in figure 2.1 has the coordinate (1,1); the *x* axis extends towards the right and the *y* axis extends downwards so that the bottom right cell has the coordinate (10,10) in the lattice illustrated in figure 2.1. The coordinates of a cell are listed as the *y* axis followed by the *x* axis, to be consistent with the way Matlab organizes matrices: an element in a matrix is called upon by specifying its row number before its column number. For example, (1,10) refers to the cell at the top right corner in figure 2.1. The PLF layer is at the ceiling so its depth is an indication of the *y* coordinate it extends to; the dense chromatin layer is at the bottom so its depth indicates number of rows counting from the bottom.

To make the square simulation space better resemble the nuclear space where each molecular movement covers a small distance compared to the size of the nucleus, it is desirable to let the simulation space contain more cells. On the other hand, each simulated molecule could potentially cover every cell in the space before finding the exit site so the more cells there are, the more steps there are to simulate; and with each step taking the same computational time, the longer the simulation process takes. With a modern computer (four-core Xeon Mac OS X system), the simulation time for each molecule might take tens of seconds but in order to obtain a set of data that has statistical significance, tens of thousands of molecules need to be added, which can be very time-consuming. To balance the need to have more cells with the necessity to keep the simulation time realistic, I decided to make the square space contain 50×50 cells. With this setup, the exit sites are at (1,17) and (1,33) in the aforementioned coordinate system.

In the first case for which results are shown in subsection 2.3.1, the chromatin depth is varied in this 50×50 space and 10 000 molecules are added for each depth. There is a density gradient in the chromatin layer which ranges from being impenetrable at its bottom to that of the middle space at the top. For each cell in the chromatin layer, its diffusion probability constants (DPCs) governing the left and right movements are the same, and they decrease linearly for the cells that represent regions deeper into the chromatin region; the downward DPC of the cell is the same as the left and right DPCs for the cell directly below it; and the upward DPC is the same as the left and right DPCs for the cell directly above it. One could argue that the step for a molecule to move to a less dense environment is as difficult as to move to an equally dense environment because in both cases, the resistances surrounding the molecule are the same. In this sense, the upward DPC should be equal to that in the left and right direction. However, for a molecule undergoing obstructed Brownian diffusion, its observed movement from one point to another almost always involves its moving between these points many times, so the propensity to move in a direction is reflective of the preference of this molecule after it has explored both its starting point and its destination. Effectively, there is a bias for a molecule in the chromatin layer to move upward rather than downward.

To test the validity and scalability of the program, the square simulation space is scaled up to 100×100 so there are four times as many cells. In order to theoretically obtain the same exit time, the diffusion probability is also quadrupled. The exit sites reside at the same places except that each is two cells wide. Since in the 50×50 space an NPC-residing molecule's next movement must either lead to its exit or to its leaving the NPC before it has another chance to exit, the 2-cell-wide exit sites pose a difficulty because there is a possibility that a molecule could move from one cell to the other *within* one exit site leading to a higher exit probability. This problem is dealt with by requiring a molecule that moves within the exit site to leave and re-enter it as a prerequisite to exit. Another adjustment made to the 100×100 space is by reducing the diffusion probability of the process of entering one of the exit sites to $3/4$ because by making each exit site 2 cells wide, there are 4 instead of 3 adjacent cells in the vicinity of each exit site to absorb molecules.

The program is also implemented in three-dimensional space with 50 cells in each dimension. This is to prove that the two dimensional square space can produce qualitatively the same result as the three dimensional model. The mRNA addition site is at

$(x,y,z) = (25,28,25)$ in which the y coordinate is the depth and there are 25 exit sites evenly distributed at the surface defined by $y = 1$.

In consideration of the computational speed, the rest of the results were computed in the 50×50 space unless otherwise noted.

The case of homogenous chromatin density is studied by having a layer with reduced DPC (40% of that for the middle space in most cases) at the bottom of the square space. The thickness of the chromatin layer is varied. At the interface between the middle space and the chromatin layer, the row of cells directly above the chromatin layer assumes the DPC of the middle space except that its downward value is the same as for the chromatin layer; the row of cells directly below the middle space assumes the DPC of the chromatin layer except that its upward value is the same as for the middle space. This setup is consistent with that for the gradient chromatin and it is done this way for the same reason as discussed on page 21. In one of the tests, the mRNA synthesis site relocates with the chromatin depth so that it is always immediately outside of the chromatin layer. In the other test, the synthesis site is located at (15,25) so that it shows only the effect of the chromatin depth. These results are shown in sections 2.3.3 and 2.3.4.

The effect of having the mRNA molecules being synthesized outside as opposed to inside of the constant-density chromatin layer is studied by setting the chromatin layer at a constant depth of 30 with the DPC in it being 10% that in the middle space, and then varying the y coordinate of the synthesis site. The results are shown in section 2.3.6. In this case, the exit time is affected by both the hindrance of the chromatin layer and by its proximity to the nuclear envelope. The effect of the proximity alone is tested by varying the mRNA synthesis site in a square space with no chromatin or PLF layer. Section 2.3.7 shows the results.

The impact of the PLF on the export process of the mRNA molecules is modeled by having smaller horizontal probability constants than the vertical ones, favoring the up-down movements in the PLF layer. To test whether there is an ideal affinity between the mRNA molecules and the PLF, the ratio of the left-right to the up-down diffusion probability constants is varied. Because a molecule that is bound to the PLF has the same amount of thermal energy as an unbound one, the sum of the diffusion probability constants in all four directions is the same in both cases. The difference is that for the bound molecule, the DPC in the up-down directions is greater than in the left-right directions. The results are shown in section 2.3.8.

Table 2.1 summarizes the parameter setups.

2.3 Results

2.3.1 Varied chromatin depth with density gradient

With the addition site at (28,25), no PLF layer, and with chromatin density gradient, as shown in table 2.1, the normalized distribution of exit times for a chromatin depth of 35 is obtained. Among the distributions that have the positive real semi-axis for their support, the Weibull distribution has the versatility of being either heavy-tailed, exponential, or light-tailed. The probability density function of a Weibull random variable $t \geq 0$ (in this case, exit time) is

$$f(x) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-(t/\lambda)^k}, \tag{2.1}$$

where λ is the scale parameter and k is the shape parameter. Figure 2.2 is the fitted plot for one set of simulations. The fitted equation is able to capture the rise and the fall of the distribution although the peak of the data is not adequately represented.

The prediction that says a thicker layer of gradient chromatin can help mRNA molecules in the system to exit is based on the assertion that a thicker layer of chromatin makes it more difficult for the molecules to reach further into the chromatin layer, hence it would, in ef-

Table 2.1: Summary of the parameter setups for various subsections of the Results. The chromatin DPC refers to the DPC in the chromatin layer compared to that in the middle space in all directions.

Subsection	Synthesis Site	Chromatin Depth	Chromatin DPC	PLF depth	Varied Parameter
2.3.1	(28,25)	varied	gradient	0	chromatin depth
2.3.2	varied	varied	gradient	0	chromatin depth + synthesis site
2.3.3	varied	varied	0.4×	0	chromatin depth + synthesis site
2.3.4	(15,25)	varied	0.4×	0	chromatin depth
2.3.5	(20,25)	30	varied	0	DPC
2.3.6	varied	30	0.1×	0	synthesis site
2.3.7	varied	0	N/A	0	synthesis site
2.3.8	(28,25)	0	N/A	20	PLF proficiency

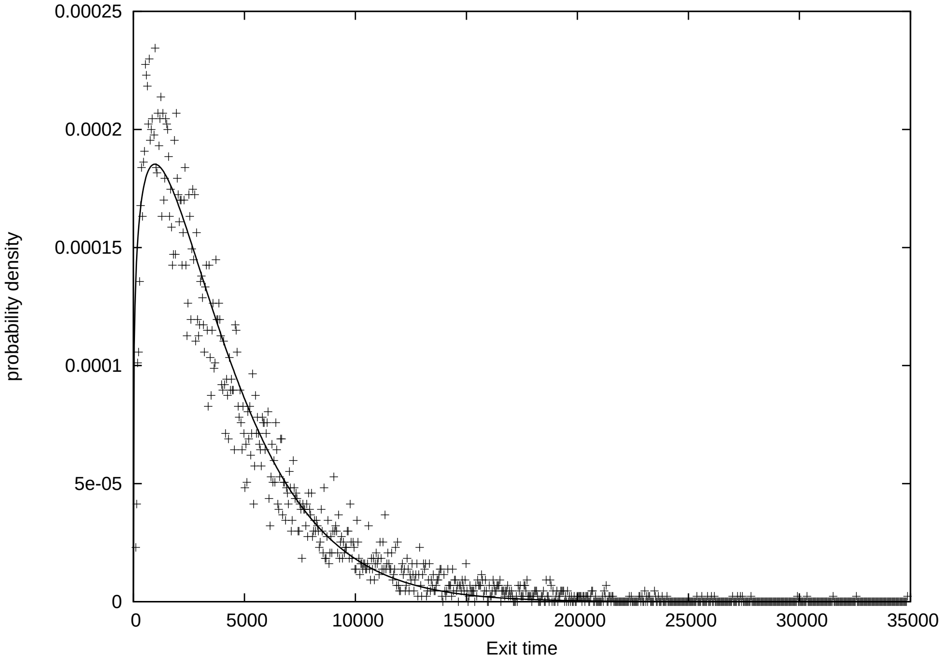

Figure 2.2: Exit time probability distribution with chromatin layer depth=35 along with its Weibull distribution fit. In the Weibull distribution, equation 2.1, $\lambda = 4045$ and $k = 1.214$.

fect, reduce the volume that a molecule has to search to find the NPC; the other side of the argument says that a thicker layer of chromatin can hinder the movement of the molecules in it and therefore increase the time of searching. By recording the molecule that spends the most time in the space from the simulated ensemble, figure 2.3 shows the time it spends in each cell. It shows that for this particular molecule, the latter argument seems to have merit because this molecule does spend quite some time in the chromatin layer. To show how much the hindrance by the chromatin layer contributes to the longer dwelling time, figure 2.4 is the visitation frequency map by the same molecule. Note that the lower right quarter of the map in figure 2.3 that shows a higher level of visitation time than the upper right quarter is much less frequently visited in figure 2.4, implying that each visit takes longer.

To see if this one molecule is representative of the overall effect of the chromatin layer, figure 2.5 shows in relative terms the collective duration that each cell hosts a total of

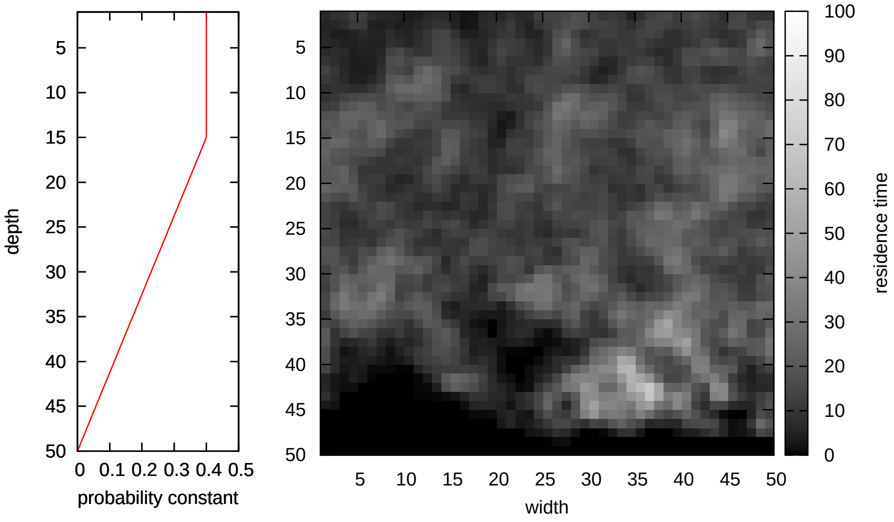

Figure 2.3: The time spent in each cell of the square space by the molecule that takes the longest to exit (right). The panel on the left shows the chromatin density gradient.

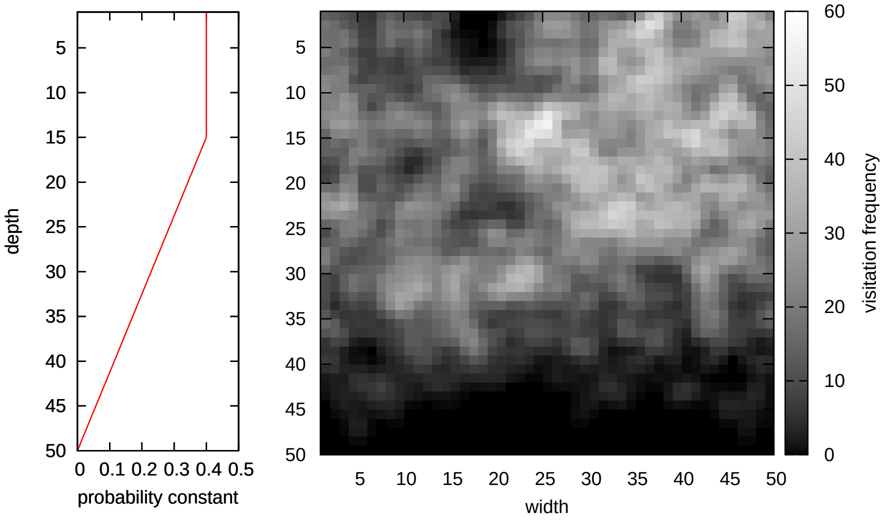

Figure 2.4: Same as figure 2.3 except that the visitation time is replaced by visitation frequency, i.e. the number of simulation steps spent at a site.

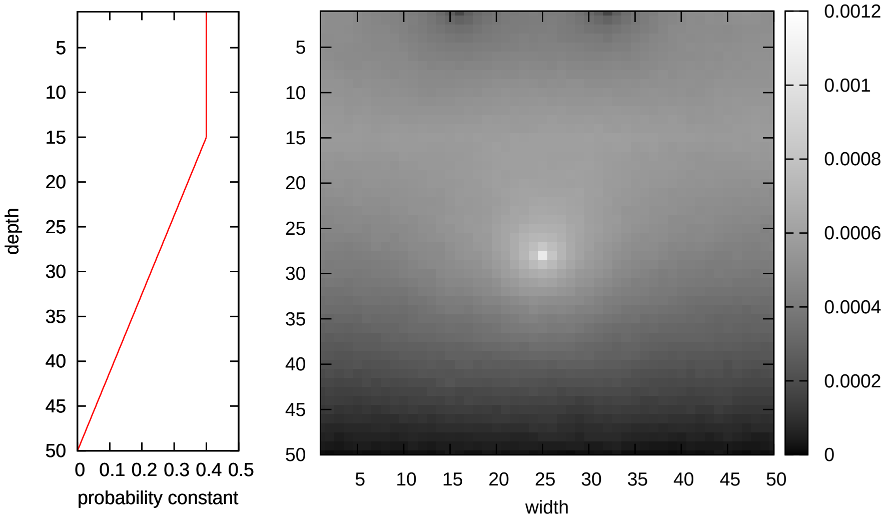

Figure 2.5: The fraction of time that an ensemble of 10 000 molecules spends in each cell. The bright spot at (28,25) is the synthesis site and it is more visited because every molecule added to the system has to visit the spot at least once. The two dark spots at the top represent the exit sites.

10 000 molecules. Contrary to the point made with the longest-dwelling molecule, figure 2.5 shows that the argument that says a gradient chromatin layer at the bottom shortens exit times is more valid in characterizing the overall trend of a statistically significant collection of molecules because the deeper into the chromatin layer, the less time is spent there by the molecules. In other words, a chromatin layer with gradient helps to exclude the diffusing molecules.

The reason that the vicinity of the chromatin exit sites are less visited than their surroundings is that there is a selection effect taking place as follows. Each molecule covers the square space unevenly in that is covers some areas more than others even if there is no chromatin layer. The one that happens to spend much of its time within the lower right corner, for example, is allowed to do so; but the one that happens to diffuse near the NPC is less likely to spend much time there because diffusing near the NPC favors exiting, and once it has exited, the subsequent movements near the NPC that could have happened should there be no NPC would not. This selection effect is analogous to the explanations of some of the phenomena in evolution such as why flying birds know to avoid crashing into trees. By chance, the instinct that leads a bird to crash into trees has equal, if not more, probability to be first conceived than the one guiding the bird to avoid trees, but the former instinct is quickly extinguished by the death of the body that hosts it whereas the latter is passed on through the generations along with other instincts that positively serve their hosting bodies.

Another more quantitative way of observing how much the gradient chromatin layer assists in improving the exit time is to plot the cumulative probability distributions at different chromatin depths (figure 2.6). Each point on the graph represents the exit probability for a molecule (the ordinate) and the time (the abscissa) to reach that probability. The result shows that by increasing the chromatin depth from 5 to 35, there is a shortening of exit times by about 30% for the most part of the cumulative probability because the time to reach each probability for chromatin depth at 35 is 70% that for chromatin depth at 5. Though it doesn't seem to be a huge improvement, a speed increase of this magnitude can be significant in evolution especially since for many genes the export process is the most time consuming part of gene expression from activation to the appearance of the protein product [3].

The comparative cumulative exit time distribution between 50×50 and 100×100 is shown in figure 2.7. The small discrepancy between the two different sets of simulations could be due to the faster speed, caused by the higher diffusion probability, in moving back into the exit site after leaving it as a result of an unsuccessful exit attempt. The close agreement between the 50×50 and 100×100 systems means that the 50×50 system is representative of a system with more grids.

The cumulative exit time distribution for the molecules in the three dimensional space is shown on figure 2.8. The curves representing the exit probabilities at different chro-

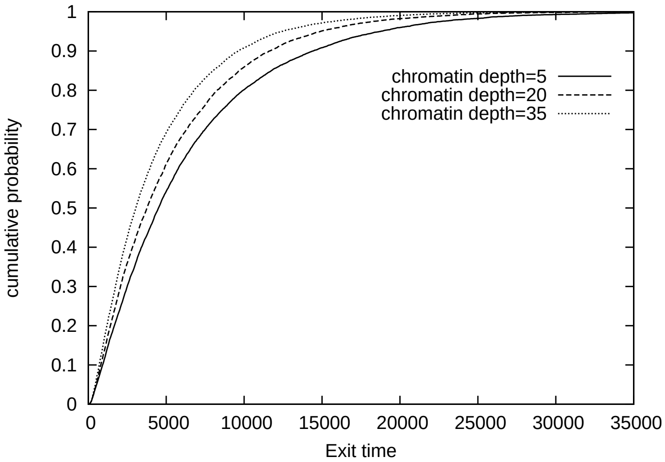

Figure 2.6: Cumulative exit time distribution at various dense chromatin depths with a chromatin density gradient.

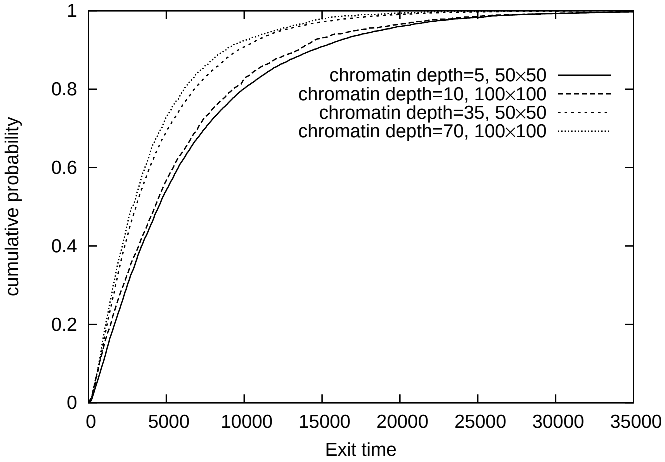

Figure 2.7: Cumulative distribution of exit times in both 50×50 and 100×100 square spaces. PLF depth=0; chromatin layer depths are shown in the graph itself. The parameters for 50×50 space are the same as in figure 2.6. As for 100×100 space, the parameters are as follows: Counting from top down and left to right, the molecule addition site is (56, 50). The DPC outside the dense chromatin layer is 1.6 (4× that for 50×50 system). The exit positions are (1, 33), (1, 34), (1, 66), and (1, 67). For each 100×100 curve plotted, 2000 molecules go through the system as opposed to 10 000 for the 50×50 systems.

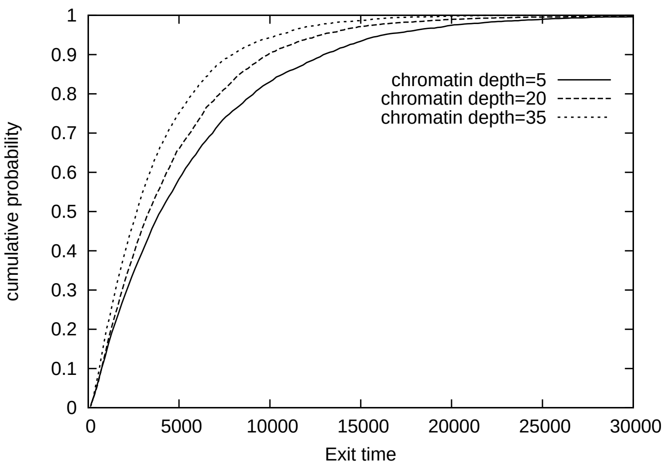

Figure 2.8: Cumulative distribution of exit times in the $50 \times 50 \times 50$ three-dimensional space.

matin depths display qualitatively the same chromatin depth effect as does figure 2.6. This means that the results from two-dimensional simulations also apply qualitatively to three-dimensional space.

2.3.2 *Varied chromatin depth and mRNA synthesis site with density gradient*

The nucleolus is well known to be the center for rRNA production. Similarly, it is proposed that mRNA molecules are also synthesized in transcription factories which are located outside of the chromatin region. This means that despite the dynamic nature of the chromosomes, the transcription factories always stay outside but near the surface of the chromatin.

In the case of a varied gradient chromatin depth with an mRNA synthesis site that always stays directly outside of the chromatin layer, the cumulative exit time distributions are shown in figure 2.9.

The difference between Figs. 2.9 and 2.6 is mainly at the beginning of the curves where

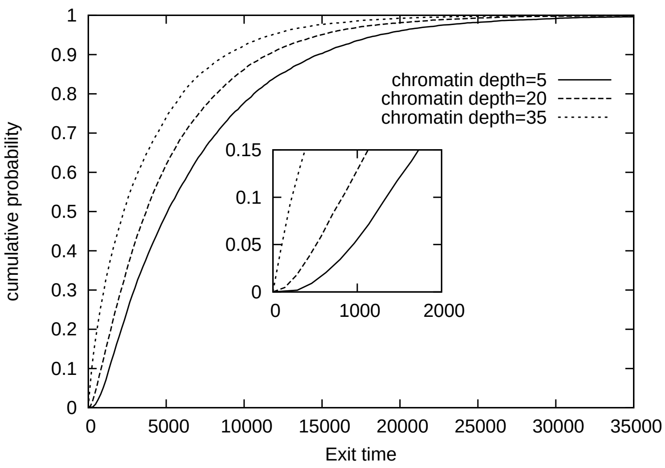

Figure 2.9: Same as figure 2.6 except that the mRNA synthesis site moves to stay above the dense chromatin layer. The inset is a magnified view of the beginning of the curves.

in figure 2.9 the curves are farther apart than those in figure 2.6. The reason is that varying the position of the addition site, in the case shown in figure 2.9, affects the early-exiting molecules simply due to the different distance that has to be travelled to exit from the addition point at each chromatin depth. This is supported by the inset graph which shows the early exits for all chromatin depths. To reach the 5th percentile, which is the beginning of the maximum exit rate, it takes approximately 9 times as long for the system with a chromatin depth of 5 (synthesis site at $y = 45$) as it does when the chromatin depth is 35 (synthesis site at $y = 15$). The ratio of 5th percentile exit times is therefore the square of the ratio of synthesis site depths. This is consistent with the well-known equation $\chi^2 = 2Dt$ where χ is the distance travelled, D is the diffusion coefficient and t is the time. Here $\chi \approx y$. On the other hand, late-exiting molecules survey much of the square space often more than once before exiting so the varied distance between addition site and NPC plays a relatively small role in the exit time for them.

In figure 2.9, the reason for the curve at chromatin depth=5 to have an ≈ 0 slope at the beginning is that, due to the long distance between the synthesis site and either of the exit sites, the rate at which the molecules arrive at an exit site increases more slowly at

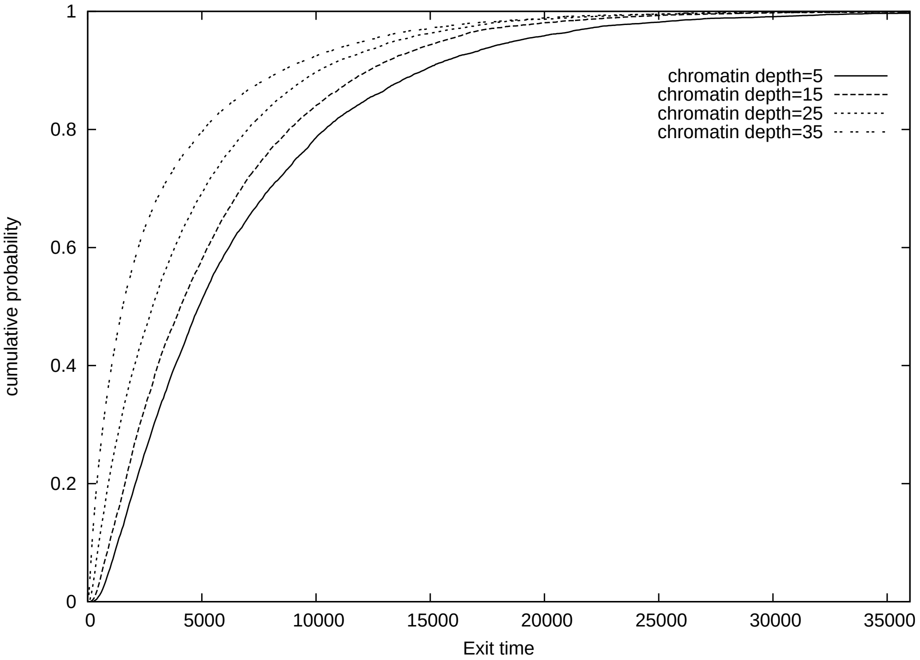

Figure 2.10: Same as figure 2.9 except that the diffusion probability constant gradient is replaced by a constant value that is 40% of that in the middle space.

the beginning. On the other hand, the early-to-exit molecules in the case of chromatin depth=35 are closer to one another before the rate of arrival at the exit sites reaches the maximum. Therefore, there is no qualitative difference between these two curves.

2.3.3 Varied chromatin depth with constant density

The cumulative distribution of the case where the chromatin layer has a homogeneous density with the mRNA synthesis site staying outside of the chromatin layer is shown in figure 2.10. The effect on exit time in this case is caused by both the chromatin depth and the proximity of the synthesis site to the exit sites. The visitation time map at chromatin depth = 35 is shown in figure 2.11. It shows that the diffusing molecules are by and large excluded from the chromatin layer and those molecules that do diffuse in it are not able to compensate for the low visitation frequency by staying longer with each visit.

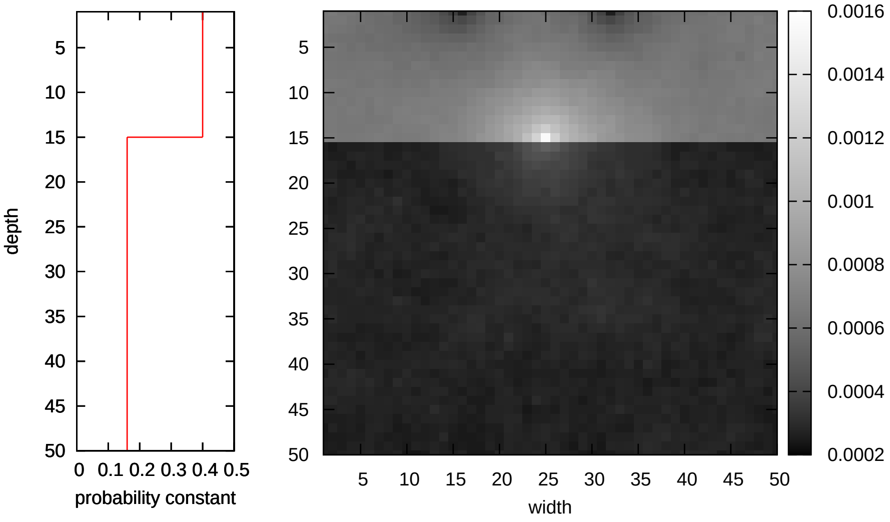

Figure 2.11: The fraction of time the molecules spends in each cell. The chromatin depth is 35, the synthesis site is at (15,25), and the chromatin layer has a constant density.

2.3.4 Varied chromatin depth with fixed synthesis site

The effect of varying chromatin depth alone on the exit time is shown in figure 2.12. This shows that a thicker chromatin layer with homogeneous density assists the exit of the mRNA molecules. The hindrance to the movement of the molecules due to the chromatin density in this case is only set at 40%. Section 2.3.5 shows the effect of the chromatin density. Comparing with figure 2.10, the major difference in this figure is at the beginning of the curves which are closer to each other in the case of figure 2.12. The reason is due to the locations of the mRNA synthesis site and were previously explained in section 2.3.2.

2.3.5 Varied chromatin density

The diffusion probability constant for the molecules in the chromatin layer is varied and the result is shown in figure 2.13. This shows that mRNA exit time is shortened with higher

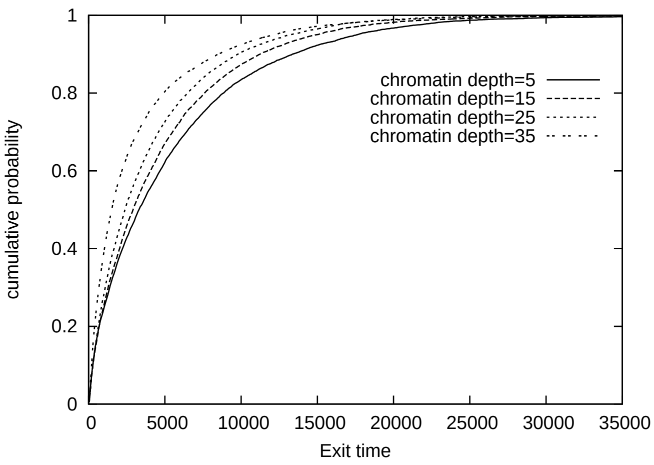

Figure 2.12: Same as figure 2.10 except that the mRNA synthesis site is fixed at (15,25) which is always outside of the chromatin layer.

density of the chromatin but only up to a point beyond which an even higher density would not make any difference. This is due to the exclusion of the overwhelming majority of the molecules from the chromatin layer because exclusion is the only conceivable means by which the chromatin layer shortens exit time.

2.3.6 Varied mRNA synthesis site with fixed chromatin layer

The cumulative exit time distribution with fixed chromatin depth and varied mRNA synthesis site is shown in figure 2.14. For the cases of the mRNA synthesis site being in the middle space, changing its proximity to the nuclear envelope does not have a significant impact on the exit time even though a factor of ten seems to be quite a difference in terms of distance. On the other hand, varying the distance from 16 to 28 caused the molecules to exit much more slowly. The reason is that at $y = 28$, the synthesis site is buried in the chromatin layer, and the chromatin hinders the movement of the molecules. At $y = 40$, the exit time is even longer. This shows the incentive for the eukaryotic system to have the mRNA molecules manufactured at a place that is different from the place where their genes

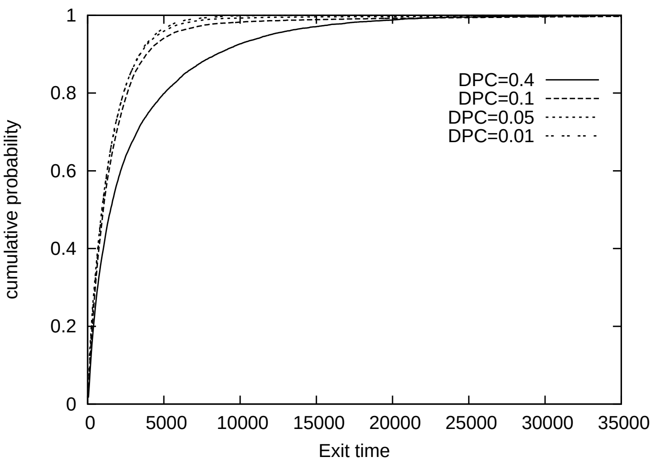

Figure 2.13: The cumulative exit time distributions at different DPCs. The chromatin layer thickness is 30 and the mRNA synthesis site is directly above it.

reside normally.

2.3.7 Varied mRNA synthesis site

The cumulative exit time distribution with varied mRNA synthesis site in a square space with no chromatin or PLF layer is shown in figure 2.15. While the closer the synthesis site is to the nuclear envelope, the faster the molecules exit, the difference in exit times between having the synthesis site at 4 and 16 (8 and 32% of the total depth of the simulation space) is greater than that between at 28 and 40 (56 and 80%). The further away from the exit sites, the less of an impact synthesis site position is expected to have on the exit time of the molecules. On the assumption that the nucleus is much bigger than the space modeled here, this result is consistent with the study that finds that the diffusing mRNA molecules in the nucleus tend to cover much of the inter-chromatin space before exit with little regard to where the physical release site is for these molecules [67].

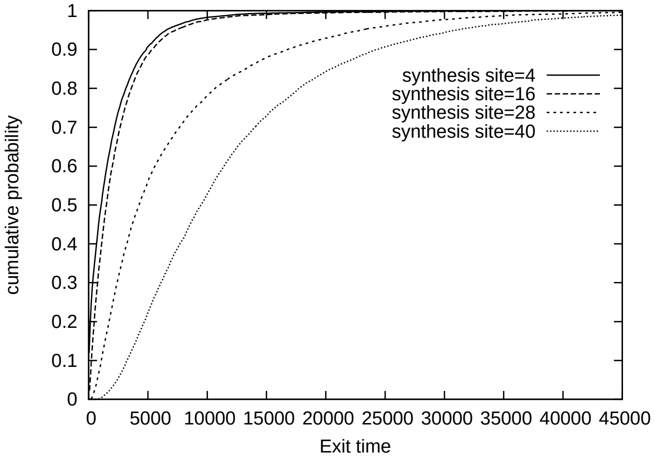

Figure 2.14: Cumulative exit time distributions as the mRNA synthesis site varies. The chromatin thickness is at 30 so the synthesis sites of 28 and 40 are inside the chromatin layer. The diffusion probability constant in the chromatin layer is 10% of that in the middle space. There is no PLF layer.

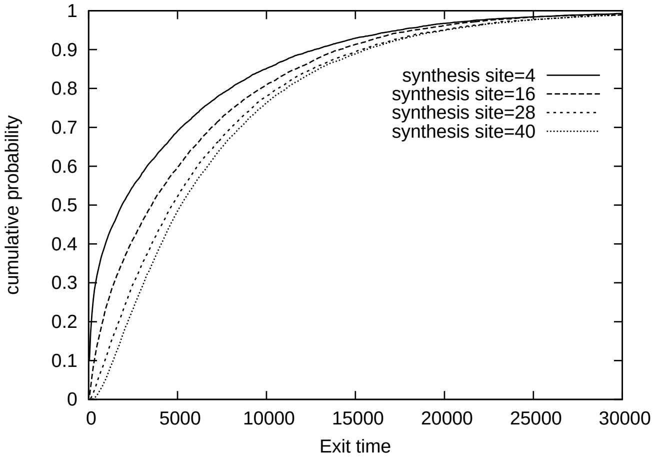

Figure 2.15: Cumulative exit time distributions as the mRNA synthesis site varies. There is no PLF or chromatin layer.

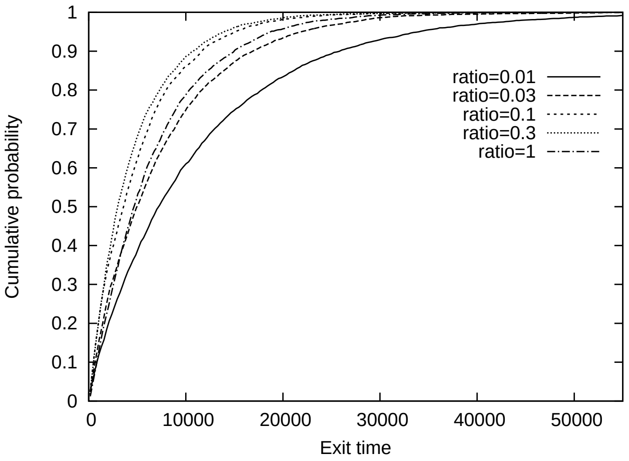

Figure 2.16: Cumulative time distribution of exit times with varying left-right to up-down diffusion probability ratio. PLF depth=20; dense chromatin depth=0.

2.3.8 PLF proficiency

With the PLF depth at 20 and no chromatin layer, the cumulative exit time distributions at various left-right to up-down DPC ratios are shown in figure 2.16. As the ratio increases from 0.01 to 0.1, the exit time increases because although the very small ratio (strong up-down movements) may allow many molecules to diffuse upward towards the nuclear envelope, the deficiency of left-right movements cannot move them to one of the NPCs. On the other hand, as the ratio approaches 1, the bias for a molecule to move upward is lost so its path tends to cover the two dimensional area where the PLF resides in search of an NPC. The preferred ratio is ≈ 0.1.

As the result of the model shows, there is a preferred ratio to promote the shortest exit time. This ratio can be biologically achieved either by having the PLFs branch laterally which is already known to occur [37] or by adjusting the affinity of the transported molecules for the PLFs. A combination of both means is also possible.

2.3.9 Discussion

The goal of this chapter was to gain a sense as to how an mRNA molecule is transported from the synthesis site to the NPC, considering only passive diffusion. The chromatin, which is the predominant structure of the nucleus, can affect the exit efficiency with its density. A denser chromatin layer favors mRNA exit if it is synthesized outside of the chromatin layer so that not too many molecules are trapped in the chromatin layer. This idea of manufacturing mRNA outside of the chromatin layer is supported by the experimental evidence of transcription factories. The benefit of doing so with regard to the exit time is that the mRNA molecules are less likely to be trapped in the chromatin layer. The consequence of an mRNA being trapped is discussed in section 2.3.6. One of the factors that prevents the chromatin from assuming an extremely dense structure to facilitate the exit of mRNA molecule is that after the density reaches a point, a higher density would not make a significant difference; the other factor is that the genes need to be accessed by molecules such as the activators in order to be useful.

Solid state transport of the mRNA molecules along the PLF is a controversial concept [1]. The result in section 2.3.8 shows that although favoring the movements of the molecules in one direction improves exit time, the improvement is moderate. The structure that is required to allow the attachment of molecules seems to be demanding. It should allow the molecules to adhere to it without too much hindrance of their mobility because the simulation does not assume any hindrance from the PLF layer. With hindrance, the advantage of having the PLF would be decreased. An even more controversial idea about mRNA transport is the one that suggests active transport. However, it has been shown that the mobilities of mobile mRNA molecules do not depend on ATP which is a prerequisite for active transport [67].

Efforts have been made to fit the exit time distributions to an analytic probability distribution function (pdf) and the results are discussed in the next chapter.

Chapter 3

Model Fitting and Tail Study

In the pursuit of further understanding of the factors that affect the exit process, the next step was to fit the exit probabilities to an explicit probability density function (pdf). Because each pdf represents a set of underlying statistical assumptions, a good fit between a pdf and several exit time distributions would suggest that the assumptions underlying a pdf might apply to the exit time. Being able to find such a pdf would also allow the exit times to be drawn from a distribution that is fully defined with, say, 2 parameters for the purpose of simulation. This becomes useful when integrating the export process in a larger model such as the entire gene expression as studied in chapter 4.

3.1 Method

The probability of a molecule exiting the space representing part of the nucleus with respect to time looks like a potentially heavy-tailed distribution so the Weibull distribution was chosen to fit the data. There are two parameters in the Weibull distribution: scale (λ) and shape (k). The probability density (PD) function of Weibull is heavy-tailed when $k < 1$ and *vice versa*. The results indicate $k < 1$ in some scenarios and $k > 1$ in others (see subsection 3.2.1). The question raised by k is whether the distribution is heavy-tailed because they are not perfect fits. In the case of figure 3.1, the rising part of the distribution is not taken into account by the pdf. For the export of mRNAs, it is important to know when the mRNA molecules first arrive at the cytoplasm.

One way to test for a heavy tail is by using the fact that the cumulative distribution function (cdf) of a Weibull distribution is: $\text{cdf} = 1 - e^{-(x/\lambda)^k}$, which means that $-\ln(\text{ccdf}) = (t/\lambda)^k$ where ccdf (complementary cumulative distribution function) is defined as $\text{ccdf} =$

$1 - \text{cdf}$. If the distribution is exponential, k should be 1 and the graph of $-\ln(\text{ccdf})$ vs t should be linear; the case of $k < 1$ (heavy-tailed) is represented by the graph being concave down and *vice versa*. This is if the data indeed fit the Weibull distribution. Otherwise, it's only the tail for which these properties apply with this analysis. The results of this method are discussed in subsection 3.2.2.

Another heavy-tail testing method is described by Bryson [10] that entails calculating the mean residual lifetime (MRL) of the tail as a function of time (t) where the mean residual lifetime is the mean of the distribution after t. A heavy-tailed distribution is represented by an increasing function of MRL vs t and *vice versa*. The results are discussed in subsection 3.2.3.

A third method used here to test for a heavy tail is provided by Kozubowski et al. [38]. In it, the survival function of the classical Pareto distribution is re-parameterized as:

$$S(x) = \left(\frac{1}{1 + \frac{\omega x}{s}} \right)^{1/\omega}.$$

Maximization of the log-likelihood function requires finding the maximum of

$$Q(\sigma) = -n\{1 + \log\sigma + \log[\frac{1}{n}\sum_{i=1}^{n}\log(1 + X_i/\sigma)] + \frac{1}{n}\sum_{i=1}^{n}\log(1 + X_i/\sigma)\},$$

with

$$s(\sigma) = \frac{\sigma}{n}\sum_{i=1}^{n}\log(1 + X_i/\sigma) \in (0, \infty)$$

and $\omega = s(\sigma)/\sigma$. Then the values of $\hat{\omega}_n$ and $\hat{s}_n$ that maximize $Q(\sigma)$ can be obtained through σ. The likelihood ratio is

$$\lambda_n = \frac{(e\overline{X}_n)^{-n}}{\prod_{j=1}^{n} \frac{1}{s} \left(\frac{1}{1 + \frac{\omega x_j}{s}} \right)^{1+1/\omega}}.$$

The deviance statistic $-2\log\lambda_n$ is compared to the critical value $C_{\alpha,n}$ where α is the significance level. The null hypothesis that a distribution is exponential is rejected at a confidence level α if $-2\log(\lambda_n) > C_{\alpha,n}$. For example, $C_{0.10,\infty}$ is 1.64, meaning that if a data set with a very large number of data produces $-2\log(\lambda_n) > 1.64$, then the null hypothesis can be rejected with 90% confidence.

Because the Kozubowski method tests the Pareto against exponential distribution, it works best when the data is either Pareto or exponential. Another method by Jackson [33, 38] tests both light tail and heavy tail against exponential distribution. It has the advantage of not assuming any particular distribution for data that do not fit the null hypothesis. It involves first calculating

$$T_n = \frac{\sum_{r=1}^{n} t_{rn} X_{(r)}}{\sum_{r=1}^{n} X_r},$$

where $t_{rn} = \sum_{i=1}^{r} (n-i+1)^{-1}$ with X being the data set with n points. For X sampled from an exponential distribution, T_n deviates from a mean value of 2 following a normal distribution with the standard deviation of $n^{-1/2}$. For light-tailed or heavy-tailed distributions, T_n is expected to be significantly less or more than 2, respectively. In the case of the latter, the normal cumulative distribution function with mean$=2$ and variance$=n^{-1/2}$ evaluated at T_n gives the probability that the data is heavy-tailed.

3.2 Results

3.2.1 Weibull fitting

The scale and shape parameters of the Weibull distribution as a result of the fitting shown in figure 2.2 are $\lambda = 4045.19$ and $k = 1.2136$, respectively. The Weibull distribution is heavy-tailed when $k < 1$ and *vice versa*. The result shows that the exit time distribution is

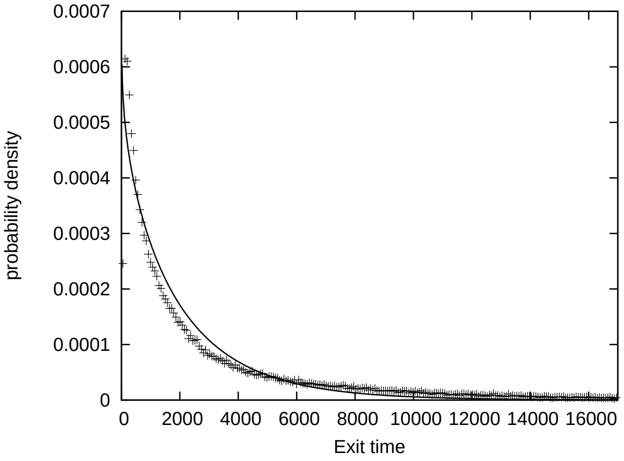

Figure 3.1: The probability density distribution of the exit time with the set up given in section 2.3.3 with the chromatin depth at 35. The solid curve is the fitted function.

probably light-tailed.

In order to confirm that it is the case for all the parameter settings in the simulations, a few more data sets were fit to the Weibull distribution and one of the examples is shown in figure 3.1 with $\lambda = 2117$ and $k = 0.9319$. This figure shows that it is not a good fit especially because the rising part of data distribution is not represented by the function. The shape parameter of the fitted function indicates that it has a heavy tail ($k < 1$) which is inconsistent with the fitting result shown in figure 2.2. Note that the data lie above the fitted curve suggesting that it could be heavy-tailed. This warrants an investigation into whether or not the tail is really heavy and how many other cases give a heavy-tailed distribution.

3.2.2 ccdf test

Figure 3.2 is the plot of $-\ln(\text{ccdf})$ as a function of exit time. The straight line fits the later part of the curve quite well; however, the beginning part is obviously curved down ($k < 1$). This means that if the probability distribution of the exit time is to be fit by Weibull, the beginning part would require a different k than the latter part which means

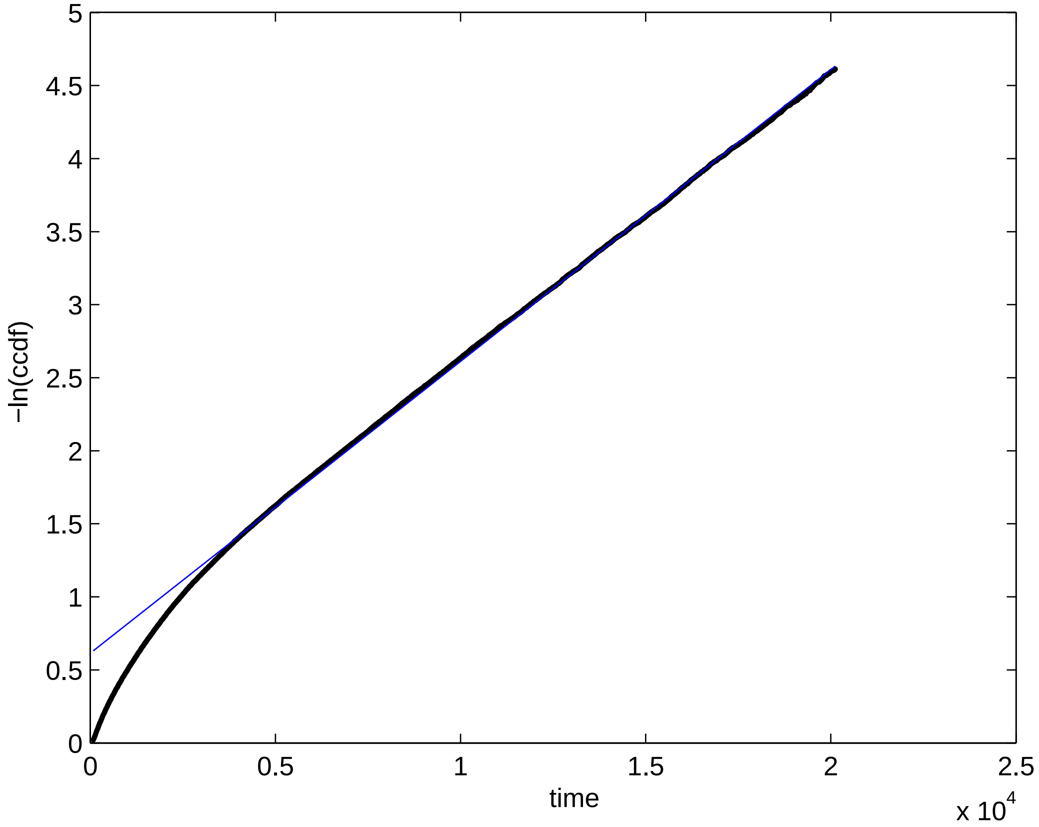

Figure 3.2: Same as figure 3.1 except that the *y*-axis is replaced by the negative of the natural logarithm of the ccdf. The curved line is from the data and the straight line though its later part is the linear fit to the probability density of the exit times that are greater than the median value.

that Weibull might not be the ideal distribution to fit the data. It also means that the tail of the distribution is apparently exponential.

3.2.3 Mean residual lifetime test

Figure 3.3 is one example of the MRL with respect to the exit time using the same data set as figure 3.1. The tail portion is noisy but flat overall indicating that it is close to being exponential.

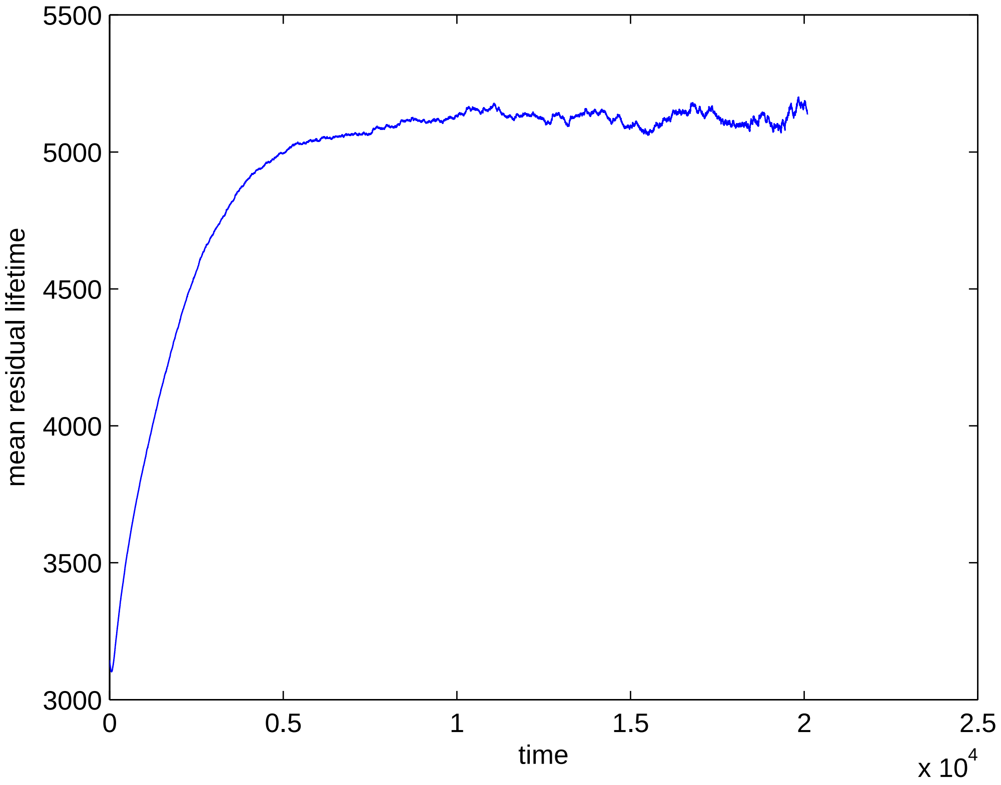

Figure 3.3: Mean residual lifetime after time t plotted against t. The data set is the same as for figure 3.1.

3.2.4 Kozubowski algorithm

The data from the simulation, according to Kozubowski's Pareto test [38] gives different results in different parts of the tail. Analyzing the data above the 60th percentile from a data set with a total of 100 000 exits, $-2\log(\lambda_n)$ is 474.89 which is by far bigger than $C_{\alpha,n}$ at the 99.5% confidence level (∼6.63), which allows one to say with more than 99.5% confidence that the distribution is not exponential. For data above the 80th percentile, $-2\log(\lambda_n)$ drops to 6.37 which is approximately at the 99.5% confidence level; At 90th percentile, $-2\log(\lambda_n)$ becomes 0.17, which is essentially exponential. This result is consistent with the results from using the previous methods in that the latter part of the tail appears to be exponential which cannot be said about the entire tail.

3.2.5 Jackson method

The result of the Jackson method tested with data from different chromatin depths with all other parameters the same as for figure 3.1 is shown in table 3.1. P is the probability of generating a normally distributed number with mean 2 and variance $n^{-1/2}$ that is bigger (for $T_n > 2$) or smaller (for $T_n < 2$) than T_n. The data set contains the exit times of 100 000 molecules except for the depths of 15 and 20 which contains 200 000 data points each. For the rest of the results using Jackson's method, 100 000 data points are used for each parameter setting unless otherwise notified.

The same results with different parameter settings are presented in tables 3.2 to 3.4.

3.3 Discussion

The Jackson method shows that the case that convincingly gives a heavy tail is the one in section 2.3.3 with a thick layer of chromatin of constant density. In this case, some

Table 3.1: The assessment of the shape of the tail (T_n) and the probability (P) that the data is generated from a normal distribution. All parameters except for the chromatin depth, which is varied, are the same as figure 3.1.

Depths	T_n	P
0	1.9977	0.3581
5	1.9996	0.4776
10	2.0021	0.3694
15	1.9951	0.1358
20	2.0008	0.4300
25	2.0099	0.0585
30	2.0214	3.580×10^{-4}
35	2.0328	1.0712×10^{-7}

Table 3.2: Results of Jackson's method for section 2.3.1.

Depths	T_n	P
5	2.0049	0.2202
15	2.0041	0.2562
25	1.9928	0.1259
35	2.0048	0.2247

Table 3.3: Results of Jackson's method for section 2.3.2.

Depths	T_n	P
0	1.9977	0.3582
5	1.9888	0.0388
20	2.0073	0.1230
35	2.0080	0.1016

Table 3.4: Results of Jackson's method for section 2.3.8.

PLF Ratios	T_n	P
0.3	1.9878	0.0269
0.1	2.0073	0.1366
0.03	2.0022	0.3735
0.01	1.9918	0.1105

molecules can get into the chromatin layer and could stay there for some time which prolongs the exit time. This is the cause of the heavy tail. With a thin chromatin layer with the same nature, though a molecule has similar probability to enter the layer, it has more chances to leave it within a reasonable period of time, hence no heavy tail is observed.

Trying to find a distribution to fit the heavy-tailed data has not turned out to be fruitful. Several distribution functions supported on the range 0 to ∞ have been tried. The best fits are Weibull and lognormal (figure 3.4).

There is a possibility that none of the two-parameter distribution functions would be able to fit the data perfectly simply because a two-parameter distribution has limited flexibility. The data gathered in mRNA exit model might require more flexibility because their underlying statistical assumptions are more complex than what a two-parameter distribution typically attempts to capture. The complexity derives from that of the system which is heterogeneous in its structure.

In a complete eukaryotic gene expression model, especially one that embraces the stochastic side of biochemistry, the mRNA export time has to be randomly drawn. Because no one distribution function can fit all the data, the exit times can be drawn from the data themselves. With each data set consisting of about 100 000 points, the bias in the drawing process can be kept low. One of the reasons that real time is not used in simulations in this chapter and chapter 2 is that, in the nucleus, each species of mRNA has its average export time, depending on its size. A distribution in real time can be easily obtained by aligning the average time in the simulated data with the average export time from experiments. This point is demonstrated in chapter 4.

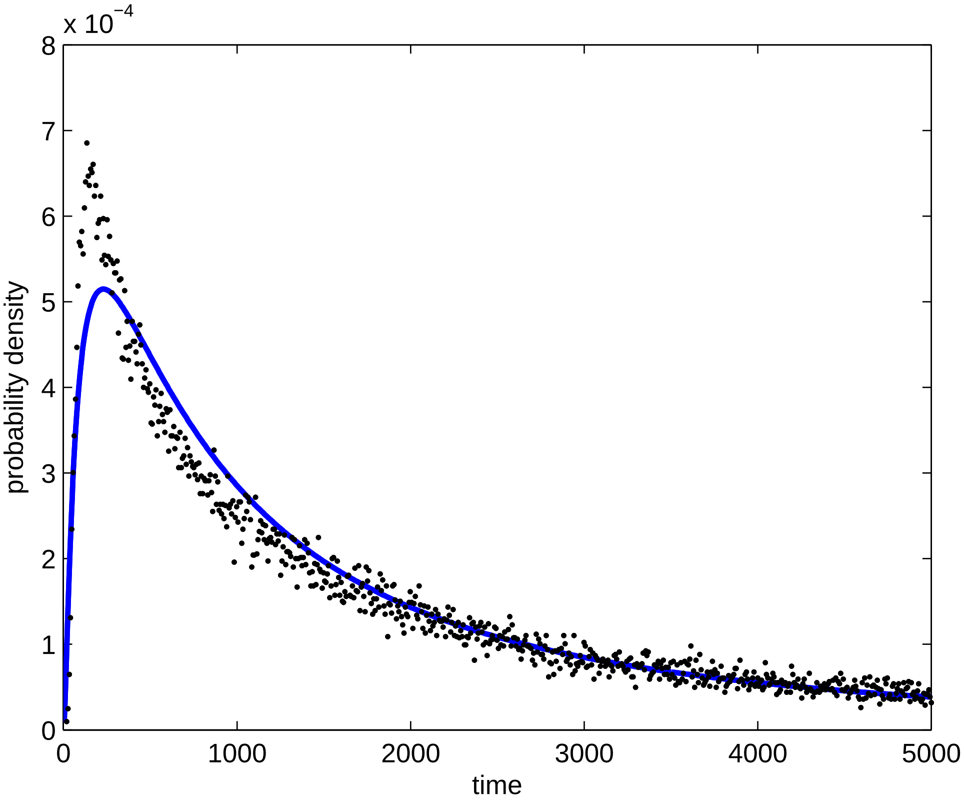

Figure 3.4: Same data as figure 3.1 fitted with the lognormal distribution.

Chapter 4

Delay Stochastic Simulation of a Complete Eukaryotic Gene Expression

With the individual steps in the eukaryotic gene expression pathway well studied and well understood by researchers, the next task is to put the steps in context with one another and to study their behaviors and influences on one another as a system. This chapter presents the development of a model to simulate the eukaryotic gene expression process from the activation of a gene to the production of the proteins. It emphasizes the noise of the system and the protein production pattern at various parameter combinations.

4.1 Method

4.1.1 Reaction equations

The gene expression process is divided into 3 steps: DNA-to-mRNA transcription, mRNA splicing and mRNA transport, and mRNA translation. The system starts with transcriptional promoters, RNA polymerases, and ribosomes. The transcription step is modeled as a delayed process. In explicit form, it is

$$\mathrm{Pro}(t)+\mathrm{RNAP}(t)\xrightarrow{k_1}\mathrm{Pro}(t+\tau_1)+\mathrm{RNAP}(t+\tau_2)+\text{pre-mRNA}(t+\tau_2) \qquad (4.1)$$

in which Pro is a transcriptional promoter and RNAP is RNA polymerase. The delayed mass-action notation [57] means that for the transcriptional promoter and the polymerase that bind at time t, the promoter is cleared after time τ_1 and the RNA polymerase along with the pre-mRNA are released when transcription is complete after time τ_2.

Due to the relatively constant and high concentration of spliceosomes in the transcription factory, the mRNA splicing process is expressed by a pseudo-first-order chemical reaction:

$$\text{pre-mRNA} \xrightarrow{k_2} \text{mRNAn} \tag{4.2}$$

where mRNAn means nuclear mRNA.

There is another delay in mRNAn export:

$$\text{mRNAn}(t) \xrightarrow{k_3} \text{RBS}(t+\tau_3) \tag{4.3}$$

where RBS is a ribosome binding site. Though the RBS exists in an mRNA when it is in the nucleus, it is after its export to the cytoplasm that it can actually bind to ribosomes.

The translation step is:

$$\text{RBS}(t)+\text{Ribosome}(t) \xrightarrow{k_4} \text{RBS}(t+\tau_4)+\text{Ribosome}(t+\tau_5)+\text{Protein}(t+\tau_5). \tag{4.4}$$

The last step is the decay of RBS:

$$\text{RBS} \xrightarrow{k_5} . \tag{4.5}$$

All the delays are randomly generated according to distributions. All except for τ_3 are drawn from gamma distributions with variance to mean ratio arbitrarily chosen as 30%. τ_3 is drawn from the distribution obtained in chapter 2. The details are described in section 4.1.2.

An mRNA can be exported as soon as it has been spliced and packaged into an mRNP. Thus, there is no distinct "commitment to export" step implied by the rate constant k_3, so equations 4.2 and 4.3 can be combined into one reaction:

$$\text{pre-mRNA} \xrightarrow{k_2} \text{RBS}(t+\tau_3). \tag{4.6}$$

In the end, equations 4.1, 4.4, 4.5, and 4.6 constitute the delay SSA model.

4.1.2 Parameter adjustment

The parameters such as the number of polymerase molecules, the reaction coefficients, and the delays need to be adjusted to accord with biological facts so that the results are comparable to experimental data. The organism chosen to obtain the data is *Saccharomyces cerevisiae* (yeast) because it is the simplest eukaryotic model organism in biology. The notation k_i is usually used for deterministic rate constants while c_i is used for stochastic rate constants

The maximal transcription initiation rate is one initiation every 6-8 seconds, achieved by using an efficient promoter [32]. Assume that half of this time is due to the polymerase finding the promoter and half is due to the regeneration of the promoter, then $\tau_1 = 4$s. The next task is to obtain c_1 from the other half of the 8 seconds. According to the theory of stochastic chemical reactions, the reaction probability density function is given by [29]:

$$P(t) = ae^{-at}$$

with $a = c \cdot N_1 \cdot N_2$, where the Ns are the numbers of molecules of each species. The first moment (the average) of t is given by

$$\int_0^{+\infty} tP(t) = \int_0^{+\infty} ate^{-at} = \frac{1}{a}.$$

Therefore, $1/a_1 = 4s$. The number of copies of the promoter here is 1. Since the number of RNA polymerase II molecules is about 5 times the number of genes [8, 11] and assuming that half of the genes are concurrently active and are dependent on RNA polymerase II for transcription, the number of RNA polymerase II molecules devoted to each active gene is about 10. Hence, $c_1 = 1/40$s. With other promoters that are less efficient, we could have

$$\frac{1}{a_1} = 50 \Longrightarrow c_1 = \frac{1}{500} s^{-1}.$$

The average gene length in yeast is 1.6kb and the transcription rate is between 1.1kb/min and 2kb/min [32, 45] so the average length of time for transcription is about 70 s. The termination process takes about 50 s [76]. Therefore, $\tau_2 \approx 120$ s.

The reaction coefficient for splicing of introns is between 5 and 30 per hour [3]. The relationship between the mean reaction time $\langle t \rangle$ and the propensity a is still $\langle t \rangle = 1/a$. With a pseudo-first order reaction, c is the same as a, hence c_2 is between 0.0014 and 0.0083 s^{-1}. The reason for modeling the splicing step as a pseudo-first order reaction is that while the mRNA is being transcribed, there is a fair chance that the splicing machinery can assemble on it. The halflife of a mature mRNA in the nucleus due to the export process is 2.5 to 4.4 minutes [3]. With the correlation between half life ($t_{1/2}$) and mean life time (τ) being $t_{1/2} = \tau \ln 2$, the mean life time is 3.6 to 6.3 minutes. The exit time distributions, as discussed in chapter 3, cannot be fitted to one particular distribution which means that the random variable that is the exit time cannot be drawn from an explicit probability distribution function. However, each set of data collected at a parameter setting in chapter 2 is large enough to constitute a statistically significant collection, and the mRNA exit times can be drawn from one of these data sets. Note that the times produced in chapter 2 are dimensionless – partly because the model is a two-dimensional representation of a three-dimensional space – so the time drawn must be scaled to real time. The mean life time is between 3.6 to 6.3 minutes, and assume that the scenario shown in section 2.3.1 with the chromatin depth at 35 has a mean life time of 4 minutes. The quotient of the mean life time (4 minutes) over the mean value of the data set is the scale factor, which multiplied to the drawn value from the data set produces the exit time in real time.

The number of ribosomes in a yeast cell is about 310 000 [55] and the number of mRNA is about 60 000 [76] so the number of ribosomes is 5 times that of mRNA. This is roughly consistent with the result that says that each ribosome, on average, is associated with 154 nucleotides [76] and that an average gene is 1.6 kb in length, meaning that an mRNA can hold 10 ribosomes at maximum capacity. Since the ratio between ribosomes and mRNA is only 5, the ribosome is the limiting species.

Under fast growth conditions, there are 13 000 translation initiations per cell per second [69] and under slower growth conditions, this number can be expected to be lower. With the number of mRNA molecules in a cell being 60 000 [76], it can be calculated that each translation initiation on an mRNA takes about 10 seconds. This duration is caused by the process of ribosomes finding the mRNA as well as the process of initiating translation. Since the ribosomes are limiting, I assume that most of the 10 seconds are spent on the process of finding the mRNA. Set $1/a_4 = 7$ s so $c_4 = 1/70\ \mathrm{s}^{-1}$ and $\tau_4 = 3$ s.

The translation speed is between 5 and 10 amino acids per second [5] and the average gene length is 1.6 kb [40] so the translation process takes about 1 minute which is τ_5.

The median mRNA decay constant is $5.6 \times 10^{-4}\ \mathrm{s}^{-1}$ [70] which is also c_5.

For lack of numerically precise experimental data, the variances of the delays that follow the gamma distribution are set arbitrarily to 30% of their respective means.

4.2 Implementation

The standard Gillespie SSA typically deals with elementary reactions for which the emergence of the products occurs at the same time as the consumption of the reaction. This way, the Gillespie algorithm simply decides, based on the reaction propensities, which reaction will be next and the time it takes; and then the reaction finishes by updating the number of molecules of the involved species and the time. The delay complicates the situation in that a reaction usually does not complete in one step because the consumption of the reactants occurs significantly earlier than the output of the products. In some cases, different products appear at different times. In equation 4.1, for example, the promoter (Pro) appears at τ_1 after the reaction starts, and the RNA polymerase (RNAP) and the pre-mRNA appear at τ_2.

To accommodate the delay, the reactions are divided into *reacting* and *generating* events [9, 27, 54, 59]. The Gillespie algorithm is used for the reacting events by deciding which reaction is next and its associated reacting time. The time delay associated with the generating step is stored in an array so before the next reacting event can occur, the Δt from the Gillespie algorithm must be checked against the delay array. If the smallest time remaining in the delay array is smaller than Δt, the corresponding generating event will take place, and the smallest delay time is subtracted from every remaining element in the delay array. Otherwise, if Δt is smaller than the smallest element in the delay array, the corresponding reacting event will take place, and in addition to updating the overall time, Δt is subtracted from every element in the delay array.

As stated in section 4.1.1, all the delays except for τ_3 are drawn from the gamma distribution. In reality, the RNA polymerases cannot pass one another during the transcription process. However, in the simulation, there is a chance that the next τ_2 is smaller than one of the existing ones that represents transcription on the same chain of mRNA. This is dealt with by drawing the delay repeatedly until it is longer than any delay of the same type on the same chain. This rejection process also applies to τ_5.

Often in nature, a gene is activated for a limited period of time in response to external stimuli. This feature is simulated by allowing the promoter (Pro) to exist for a length of time. When the time expires the variable that keeps record of the promoter quantity is set to 0. The transcription processes that have already started are allowed to finish, and the promoter that is bound to the polymerase (RNAP) is not disabled immediately but will be

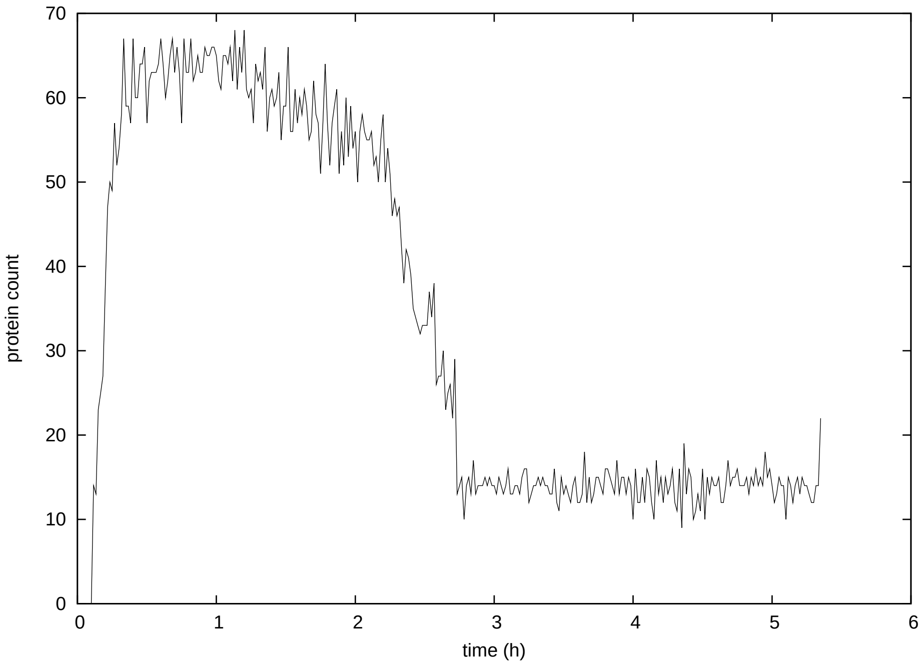

Figure 4.1: Protein production with respect to time. The bin width is fixed at 60 seconds, and there are 70 ribosomes.

when τ_1 expires. Similarly, a translation in progress is also allowed to finish in the event that the RBS that started it is degraded.

4.3 Results

4.3.1 Protein production distribution

The distribution of the protein production quantity with respect to time is shown in figure 4.1. The gene is activated for one hour. The protein quantity is the number of proteins produced in 60-second intervals and the time is since the beginning of the gene activation. The number of ribosomes devoted to the translation of this gene is 70.

At the beginning after the gene activation, there is a short period of time during which there is no protein production. This is simply due to the delays. Then the mRNAs start to arrive in the cytoplasm and translation starts. The arrival times for the mRNAs are different and before the amount of mRNA molecules is able to exhaust the free ribosome stock, protein production is at a lower level. Since τ_4 is much smaller than τ_5, it does not take many mRNAs to keep almost all the ribosomes busy at translation which explains the fast rise of the protein production rate.

The maximum protein production rate after that is when there are enough mRNAs to bind to almost all of the ribosomes, and the maximum is defined by the number of ribosomes available. To show that the ribosomes are indeed limiting, their number was increased to 700 with the results shown in figure 4.2. With all the other parameters and initial conditions being the same between figures 4.2 and 4.1, the fact that 9 times more ribosomes are associated with a nearly 9 times higher maximum protein production rate indicates that the ribosome is limiting.

Some time after the activation time has expired, the availability of the mRNA to the ribosomes starts to decrease. It then reaches a point where the ribosome quantity is no longer limiting and mRNA becomes limiting. This transition is shown by a sharp decrease in protein production. After the transition, the number of cytoplasmic mRNAs determines the protein production rate which stays at a fairly constant level for a fixed number of mRNAs. The last horizontal cluster of points from $\approx$2.75 to $\approx$5.25 hours in figure 4.1 is when there is only one cytoplasmic mRNA left. The cytoplasmic mRNA levels for figures 4.1 and 4.2 are shown in figure 4.3 and 4.4 respectively. Note that the mRNA counts represent the existing number of mRNAs in the cytoplasm whereas the protein counts represent the number of proteins being produced in an interval of 60 seconds.

One of the features of figures 4.1 and 4.2 is that even though their activation times are the same and the difference is only in ribosome number, the overall gene expression process lasted much longer in the case of the latter. This is at first glance unexpected because

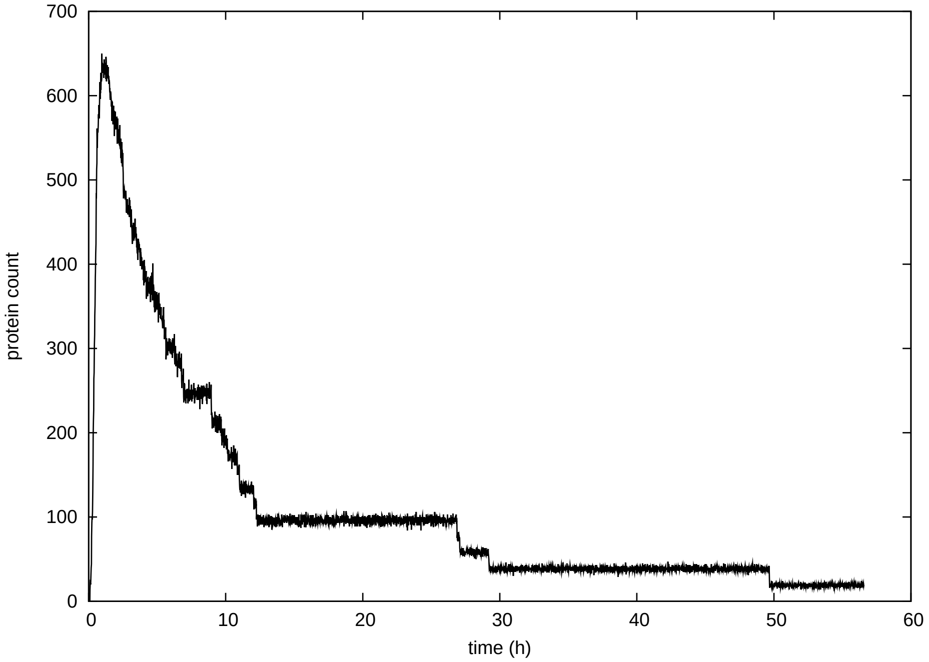

Figure 4.2: Protein production with respect to time. The parameter settings are the same as for figure 4.1 except that the number of ribosome is 700 instead of 70. Note that the time axis is on a greater scale.

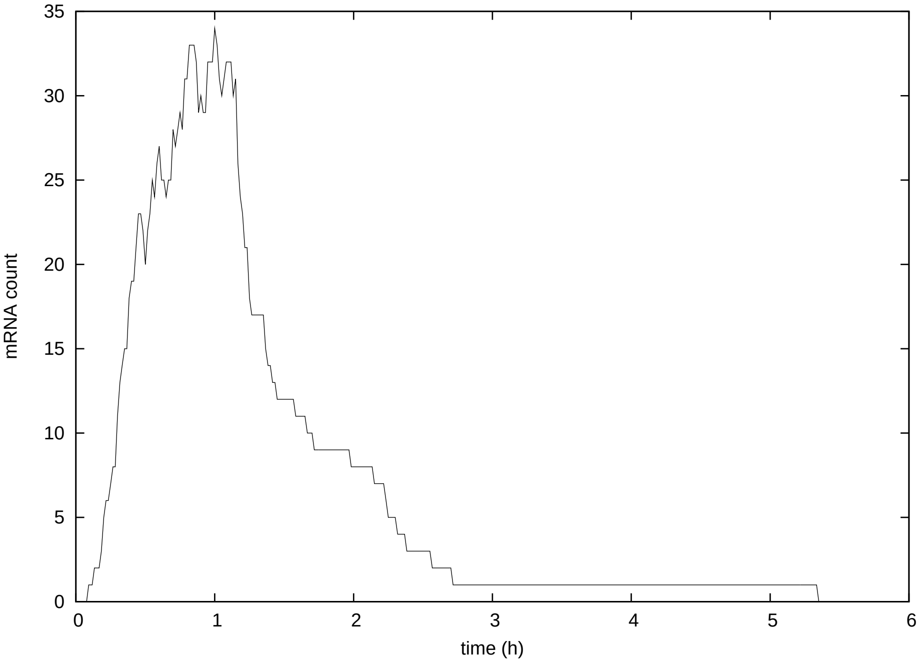

Figure 4.3: Cytoplasmic mRNA level with respect to time. It is obtained from the same simulation that generated figure 4.1.

every process in the nucleus is the same between the two cases except for minor stochastic effects so there should be similar quantities of mRNA exported to the cytoplasm. With the decay rate being the same, the mRNA molecules should disappear at about the same time counting from the moment that the gene is activated. The cause for this effect is that there is a competitive relationship between the ribosome (including translation initiation factor eIF4F) and the decapping enzymes [65]. When the ribosome is bound to the CAP structure at the 5' end of the mRNA, the decapping enzymes do not have access until the ribosome moves away. In the simulation, τ_4 needs to end before its RBS can go through the decay process.

This accounts for an ingenious way to control the level of mRNA in the cytoplasm: If the level is much higher than the translation capacity of the ribosome, its degradation level is high so as to decrease its number and recycle the nucleotides. It then reaches a

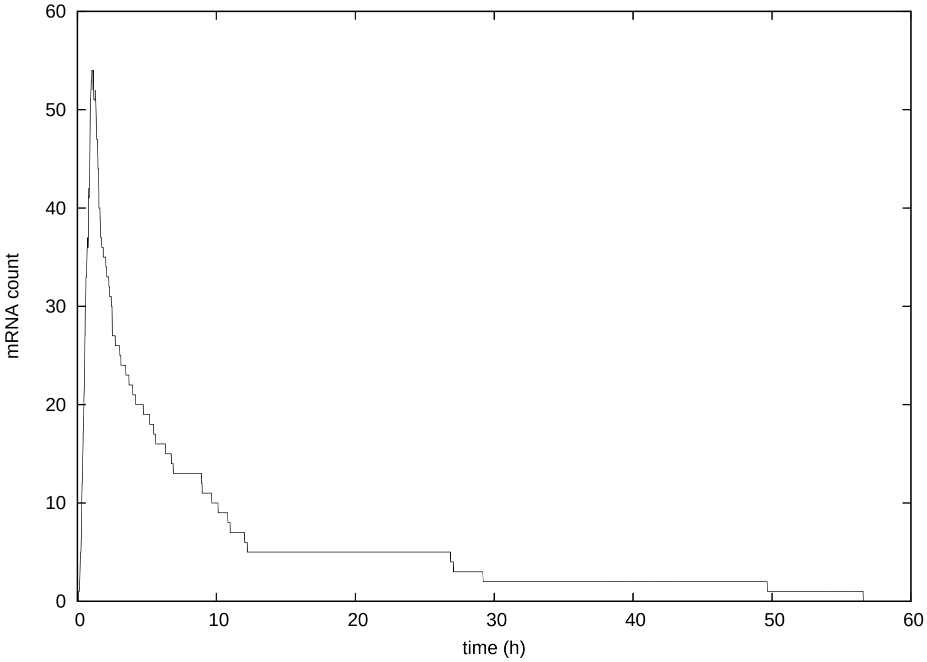

Figure 4.4: Cytoplasmic mRNA level with respect to time. It is obtained from the same simulation that generated figure 4.2.

level where each mRNA molecule is being translated at nearly full capacity and it will remain at that level for some time before the molecules are degraded slowly. Therefore, the ribosome, along with the translation initiation factors, can not only translate the mRNAs but also regulate their levels. This is an example of the intricacy of biological systems.

4.3.2 Stochastic nature of gene expression

A cell, whether prokaryotic or eukaryotic, is a system that is able to respond to external stimuli. For example, when a bacterium that normally uses glucose as its energy source is put in a lactose-only environment, it switches to using lactose by activating the *lac* operon. The ability of a bacterium to make this switch is important to its survival. In this sense, the molecular system in a cell needs to have a deterministic feature. However, at low concentration of lactose, genetically identical bacteria make β-galactosidase, a key enzyme in lactose metabolism, at different rates: some do not make this enzyme at all [53]. The differences between the cells can only be caused by the randomness in the underlying processes at the level of the individual molecules involved. More specifically, molecular movements follow Brownian motion and a probability governs whether each collision between two reactants will lead to a reaction. In this sense, a biological process or system can never achieve the exact optimal response to an event no matter how long it is subjected to the evolutionary pressure because randomness is the nature of every molecule of any process. One obvious way to minimize randomness is by having a large number of the same species of molecules involved in a reaction so that the molecule-level probability is multiplied to a population-level proportion. Nonetheless, a cellular process usually involves a small number of at least some species so stochastic variations can significantly impact the dynamics of the entire process. Therefore, the presence of stochasticity in biological systems is unavoidable. Because every organism is subjected to it, it is not necessarily a disadvantage in the competition for survival. In the case of a bacterial population, though some might starve to death for not being able to use lactose as an energy source, enough do survive to recover the loss; in a multicellular organism, though some cells might not respond to a stimulus in a way that is expected of them, others do respond and ensure the survival of the organism.

Though evolution is not able to eliminate the consequences of the stochastic effects in cellular processes, what it can do and is good at doing is adaptation. Often, adaptation can turn disadvantages to advantages. In the case of the *lac* operon, one paradox is that lactose needs to enter the bacterial cell in order to turn on the *lac* operon; the only way for the lactose to enter the cell is through β-galactoside permease which is encoded by the *lac* operon. If the *lac* operon is initially off, there should be no β-galactoside permease, hence the lactose cannot enter the cell and the *lac* operon will stay off. This constitutes a case of the classic chicken and egg problem. Logically, if neither exists initially, neither will exist in the future. The solution to it as employed by the cellular system is by resorting to the stochasticity of the individual molecules. The primary reason why the operon is off is that there is a repressor protein that binds to the operator sequence upstream of the genes but downstream from the promoter sequence so that the RNA polymerase is not able to pass the operator to start transcription. However, even though the affinity between the repressor and the operator is very high, the thermal energy in the repressor allows it to break free from the operator once in a relatively long period of time. During this window of opportunity, if there happens to be an RNA polymerase bound to the promoter and ready to start transcription, there is a chance that an mRNA would be made. This mRNA would then proceed to make several copies of each protein encoded by the *lac* operon including the β-galactoside permease.

Random fluctuations can be especially conspicuous during gene expression. The variation caused by the randomness at the mRNA level is magnified when it reaches the protein level because each mRNA molecule gives rise to several protein molecules. To see the difference in protein production due to the randomness of the process using the delay SSA, the same simulation was run twice with exactly the same parameters and initial conditions. The result is shown in figure 4.5. The first run produced 10 661 protein and 69 mRNA molecules, and the second run produced 9120 protein and 65 mRNA molecules. The time at which the number of mRNA molecules in the first run is reduced to 1 is earlier than for the second run. Despite this, the last mRNA in the first run decayed much later ($\sim$ 5.5 hours) than its counterpart in the second run ($\sim$ 3 hours), displaying significant stochastic effect.

The same simulation except with 700 ribosomes instead of 70 was run and the result is shown in figure 4.6. The first run produced 382 333 protein and 67 mRNA molecules, and the second run produced 391 536 protein and 70 mRNA molecules. The first run ended

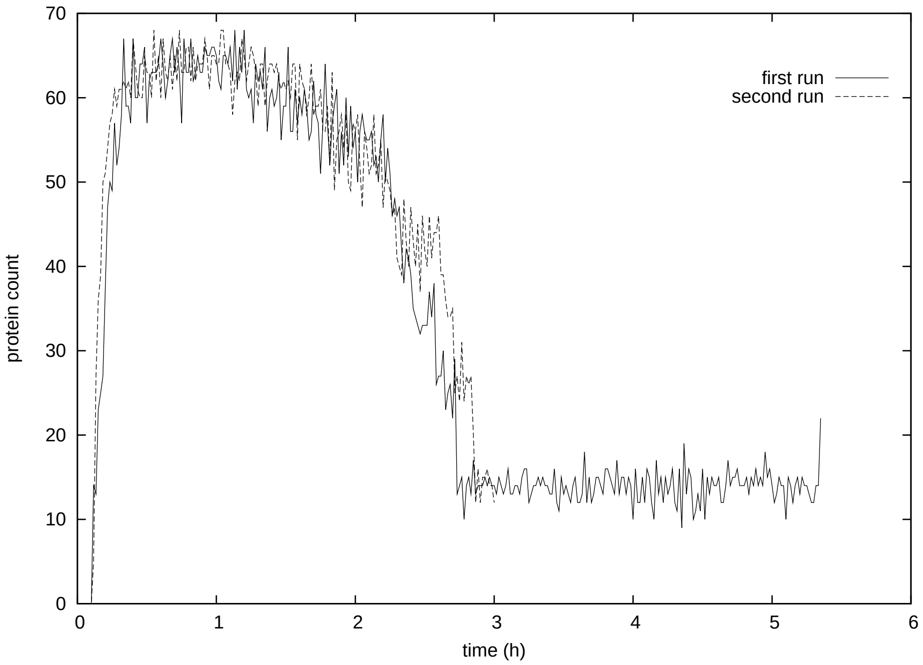

Figure 4.5: Two simulations run with the same parameters and initial conditions as in figure 4.1.

after $\sim$ 57 hours and the second run ended after $\sim$ 59 hours.

Though in most runs more mRNA molecules gives more protein molecules, there are instances where the reverse is true, showing the randomness in the translation process.

To see whether the randomness in the transcription process has an impact on the eventual protein quantity, the setup is as follows: all the reacting events are the same as before; τ_3, τ_4, and τ_5 are kept at constant values which are their respective means as discussed in section 4.1.2. In the first case, τ_1 and τ_2, the delays in transcription, are kept constant as well. The process is run 200 times with the same parameter setup. The activation time of the gene is 10 minutes; there are 700 ribosomes; and the number of the protein product is collected from each run. If there is no randomness at all, the protein quantities should be identical. In this case, the randomness is introduced only by the reacting events.

The second simulation is performed the same way as the first one except that τ_1 and τ_2 are drawn from a gamma distribution with a coefficient of variation being 0.3. Between

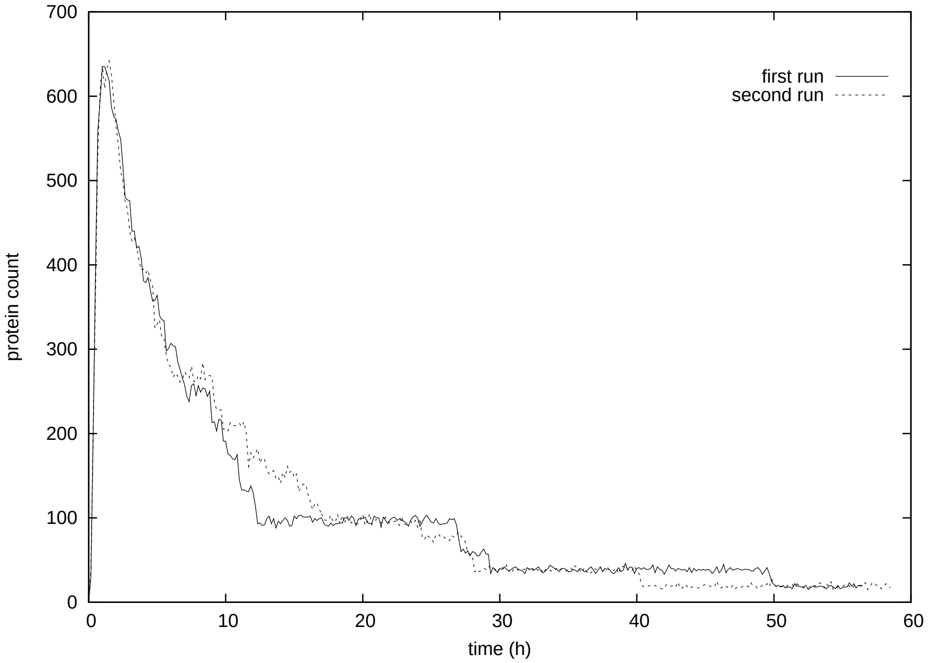

Figure 4.6: Two simulations run with the same parameters and initial conditions as in figure 4.2. In fact, the first run uses the same data set as figure 4.2 itself. To make the two lines more distinguishable, they are produced using every 10th data points from the respective data sets.

the two simulations, the extra randomness in the second one is introduced only at the transcriptional level.

The coefficient of variation (CV) , which is defined as the standard deviation over the mean, is used to show variations in the protein quantity. Each set is run 200 times. In the first set of simulations, the mean protein number as a result of 10 minutes of gene activation is 185 310, and the standard deviation is 54 130, giving a CV of 0.3042. The mean for the second simulation is 178 050, and the standard deviation is 62 212, giving a CV of 0.3495. The two-sample t-test gives a p-value of 21%. The null hypothesis that the two distributions have the same mean cannot be rejected. The two-sample Kolmogorov-Smirnov test, which compares entire distributions (not just the means), gives a p-value of 13%. This means the hypothesis that the two samples are drawn from the same distribution cannot be rejected with great confidence. Therefore, the variability in transcription delays only has a small (if any) impact on the overall randomness.

The same pair of simulations is then performed with 70 instead of 700 ribosomes. The simulated protein distributions are shown in figure 4.7. The one with constant τ_1 and τ_2 gives a mean protein quantity of 7689, a standard deviation of 2581, hence a CV of 0.3357; the one with variable τ_1 and τ_2 gives a mean protein quantity of 7209, a standard deviation of 2465, hence a CV of 0.3420. The t-test gives a p-value of 4.99% which is the probability that the two distributions have the same mean. The two-sample Kolmogorov-Smirnov test gives a p-value of 26%. This means the hypothesis that the two samples are drawn from the same distribution cannot be rejected with great confidence.

4.3.3 Change in algorithm

The result in section 4.3.2 shows that the local number of ribosomes has an enormous impact on protein production because a higher number of ribosomes allows for a higher

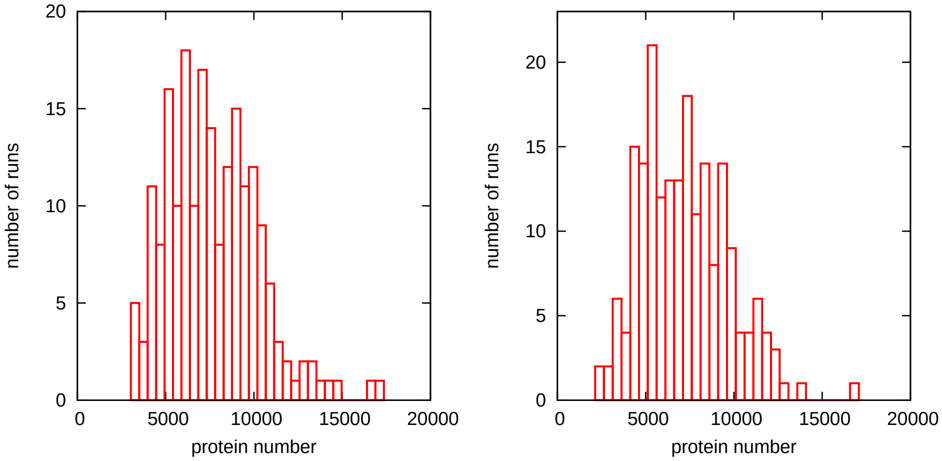

Figure 4.7: Protein quantity comparison between fixed and random transcriptional delays with 70 ribosomes. All other parameters are the same for both cases. Each histogram presents data from 200 runs with identical parameters and initial conditions. In the left histogram, τ_1 and τ_2 are fixed at 4 and 120; in the right histogram, they are drawn from gamma distributions.

frequency of mRNA binding; this not only increases protein production rate directly but also shields the cap of an mRNA for a longer time, thereby allowing the mRNA to survive longer. This implies that by being translated, an mRNA molecule decreases the number of ribosomes available to protect its and others' caps, thereby decreasing their life times. This is a valid and important argument while considering the entire translational system of the cell. However, from the perspective of only one gene, it may only have $\sim$10 mRNA molecules at any given time; the number of ribosomes bound to these molecules is insignificant compared to the number that are available in the cytoplasm. Moreover, the diffusion of the mRNA molecules in the nucleus favors their exits in different NPCs so that they are physically separated in the cytoplasm. In all, the binding of the ribosome to one mRNA molecule should have very little effect on the ribosomal availability to the other mRNA molecules that are from the same gene. This calls for a change in the reaction equations in section 4.1. Transcription is better represented by the current model than translation because transcription is at a transcription factory and the local RNA polymerase level is, to some extent, fixed.

Another part of the model that needs to be changed to make it more biologically sound is the degradation of mRNA in the cytoplasm (RBS). In the current model, the mRNA is subjected to degradation as soon as it enters the cytoplasm. This is inaccurate because the decapping is usually inhibited by a protein called Pab1p that also binds to the poly-A tail [12, 50], and its protection on the mRNA is weakened after the poly-A tail is shortened [65]. Therefore, the RBS is protected from degradation for a period of time. This calls for two species of RBS: one that is resistant to degradation; and the other that is prone to it. Both of them are equally efficient in participation in translation.

After taking into account these changes, the model still starts with equation 4.1,

$$\mathrm{Pro}(t)+\mathrm{RNAP}(t)\xrightarrow{k_1}\mathrm{Pro}(t+\tau_1)+\mathrm{RNAP}(t+\tau_2)+\text{pre-mRNA}(t+\tau_2). \tag{4.1}$$

The splicing and export step now incorporates the delay in trimming of the poly-A tail,

$$\text{pre-mRNA}(t)\xrightarrow{k_2}\mathrm{RBS}(t+\tau_3)+\mathrm{polyA_s}(t+\tau_3+\tau_6), \tag{4.7}$$

where $\mathrm{polyA_s}$ represent shortened poly-A tail. The step in shortening the poly-A tail is included in the splicing and export step because it is considered to start immediately after the mRNA is in the cytoplasm. The original implementation is to have the RBS go through a delay before it becomes susceptible to decay: $RBS(t)\longrightarrow RBS'(t+\tau)$ where only RBS' can decay. This implementation is incorrect because it implies that the RBS must be consumed for the delay to initiate. In reality, however, the RBS can bind to the ribosome *while* the tail of the mRNA is being shortened. To be consistent with the previous model in terms of symbols, τ_6 is introduced to represent the time for poly-A degradation.

The translation step is

$$\mathrm{RBS}(t)\xrightarrow{k_4}\mathrm{RBS}(t+\tau_4)+\mathrm{Protein}(t+\tau_5), \tag{4.8}$$

which is a pseudo-first order reaction due to the constant free ribosome concentration.

The mRNA (RBS) degradation step is:

$$\mathrm{RBS}\xrightarrow[\mathrm{polyA_s}]{k_5}, \tag{4.9}$$

which means that the RBS has a chance to be degraded in the presence of a shortened poly-A tail on the same mRNA molecule. Though the decay requires two species of reactants, it is of first-order because RBS and $\mathrm{polyA_s}$ do not need to find each other. The coding of this conditional first order reaction requires some extra steps as follows. Each RBS-polyA$_s$ pair is given a unique identity for the export step (equation 4.7). Whenever an RBS or a polyA$_s$ is generated, check the inventory to find its corresponding polyA$_s$ or RBS, respectively. If its pair cannot be found (because the other molecule is in a delay), it enters the inventory to wait to be checked when its pair is generated. If its pair can be found, they together produce a new species (call it virtual-RBS) by consuming polyA$_s$ but not RBS: $\mathrm{RBS} + \mathrm{polyA_s} \longrightarrow$virtual-RBS+RBS. The new species, virtual-RBS, assumes the same identity as the RBS and polyA$_s$ pair. Because the production of virtual-RBS does not consume RBS, it can only decay as an RBS in equation 4.9 but cannot be translated in equation 4.8; the RBS, on the other hand, can be translated but cannot decay. In short, RBS participates in the translation aspect of the mRNA and virtual-RBS participate in the decay aspect of the mRNA. Whenever an RBS is consumed, its corresponding virtual-RBS also disappears and whenever a virtual-RBS is consumed, its corresponding RBS disappears. The generating step produces both RBS and virtual-RBS. This ensures that an RBS cannot decay while being used in translation.

The decay of the poly-A tail depends on several factors, chief among which is its length and the mRNA sequence. For typical mRNAs in yeast, the length of the poly-A tail is about 200 nucleotides; it needs to be shortened to about 30 A's for decapping to occur [2]; and the deadenylation rate is between 4 and 13 residues per minute [12]. Here I assume the time for deadenylation to be 20 minutes, making $\tau_6 = 1200$s.

After the digestion of the poly-A tail, the removal of the cap simply involves the assembly of the decapping machinery at the 5' end of the mRNA molecule. Though the decapping process and the turnover of the mRNA molecule may take some time, because the ribosome is in competition with the decapping enzyme, once the latter is bound to the cap, translation stops. The exact time it takes for the decapping enzyme to bind to an mRNA molecule is not known but it is a simple searching process so it is assumed here to be 5 seconds ($c_5 = 0.2$/s). This is on the condition that the RBS's poly-A tail has been shortened enough, i.e. once polyA$_s$ has been formed.

As mentioned in section 4.1, it takes about 7 seconds for a ribosome to bind to the cap. With the translation process being modeled as a first-order reaction, c_4 becomes 0.14 s^{-1}.

4.3.4 Noise in transcription

According to the delay SSA model, there are two causes of variability in transcription: the reaction step that involves the promoter and the RNA polymerase; and the generation step that involves the two delays in generating the pre-mRNA and in the regeneration of the polymerase and the promoter.

The randomness in the generating step affects the final protein quantity through the delays in re-generating the promoter and the RNA polymerase but not through the synthesis of the pre-mRNA because the expected number of protein molecules synthesized from each mRNA does not fluctuate in response to the emergence time of the pre-mRNA or mRNA. The noise produced by the generating step is simulated by the same method as described in section 4.3.2 which is: first keep τ_1 and τ_2 constant throughout the first set of simulations; then the second set of simulations uses τ_1 and τ_2 drawn from a gamma distribution. Each set repeats the same process of gene expression many times in order to obtain a statistically significant number of protein levels. The noise level of a set is represented by the CV of the protein number distribution; and the noise caused by the delays in transcription is shown by the comparison between the CVs of the two sets with fixed and randomly drawn τ_1's and τ_2's.

The set with constant τ_1 and τ_2 gives a mean protein quantity of 276.3 after 60 seconds of gene activation; the standard deviation is 133.6; so the CV is 0.4836. The distribution is shown in figure 4.8. Note that it is only the gene activation that is 60 seconds and all

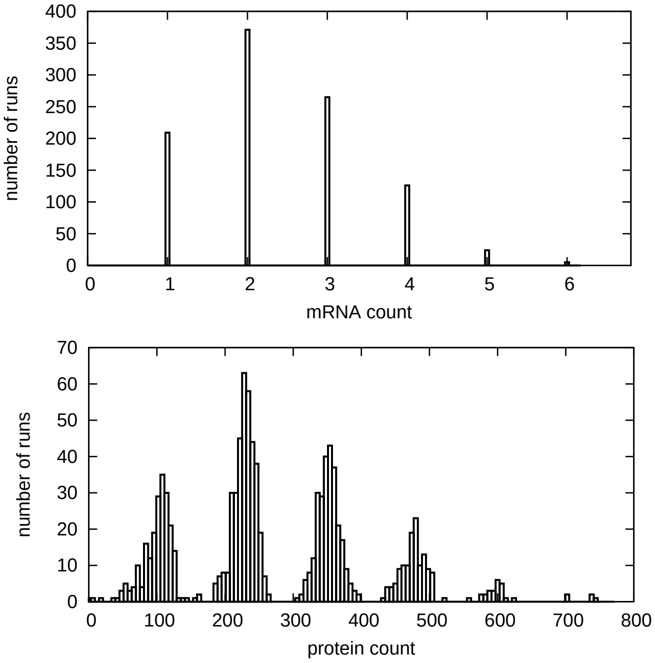

Figure 4.8: Protein and mRNA quantity distribution of 1000 runs with identical parameters and initial conditions: τ_1 and τ_2 fixed at 4 and 120 seconds; all the other τ's are drawn from their corresponding distributions; gene activation time is 1 minute.

the other quantities are accounted for the entire process (from the beginning to the time when there is active promoter, RBS, or mRNA-bound ribosomes). The most predominant feature of this figure is that the protein quantities are distributed in clusters. This is because each run has its number of mRNA molecules and the number of protein molecules produced from each mRNA molecule is also distributed. Figure 4.8 shows mRNA quantity distribution from the same set of simulations.

The set with randomly drawn τ_1 and τ_2, after 60 seconds of gene activation, has a mean protein quantity of 277.6 with a standard deviation of 135.9 so the CV is 0.4894. There is virtually no difference between the results of this set and the set with constant τ_1 and τ_2. The protein quantity distribution is shown in figure 4.9. This shows that the randomness seen at the final protein level is not significantly attributable to the dispersion in transcription times.

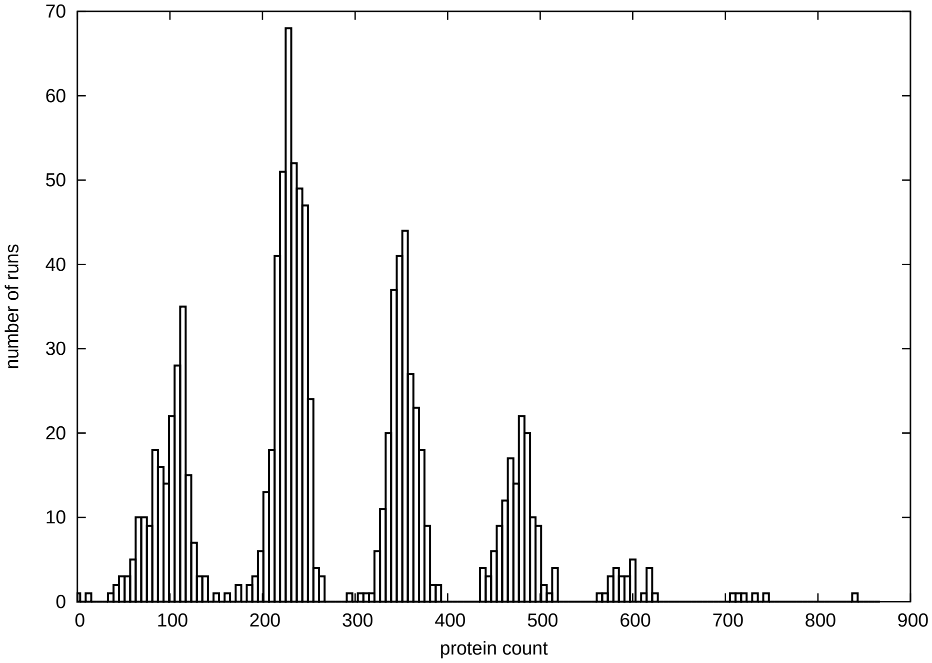

Figure 4.9: Same as the protein distribution in figure 4.8 except that τ_1 and τ_2 are drawn from gamma distributions with a variance to mean ratio of 0.3.

The impact of the entire transcription process on the randomness of the final protein quantity is studied by setting the initial conditions of the simulation so it has no promoter or RNA polymerase, hence no transcription. To still be able to produce the protein molecules, a fixed number of pre-mRNA is present in the nucleus at the beginning. In the case of starting with 3 pre-mRNAs, which is close to the mean number of mRNA produced after 1 minute of gene activation, the mean protein quantity is 352.8 with a standard deviation of only 16.1 so the CV is 0.0457. The protein quantity distribution is shown in figure 4.10.

This confirms that the dominant noise in the gene expression process after a short window of activation comes from transcription initiation. Because each transcript gives rise to a number of protein molecules before its poly-A tail is shortened, a variation in transcript number can significantly impact the final protein quantity, especially due to the fact that the short activation time only produces a few transcripts. The delays in transcription in this

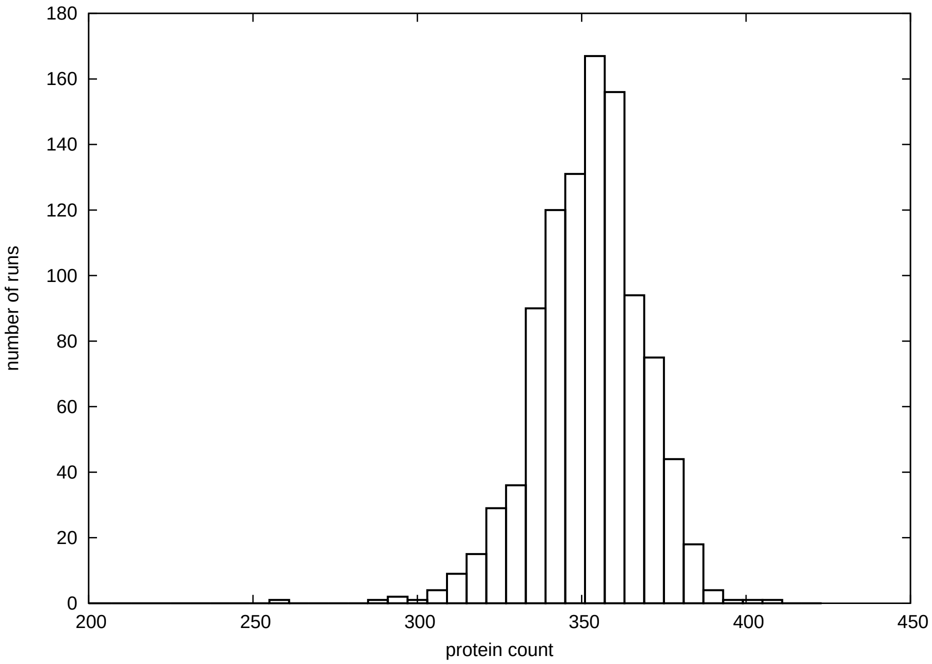

Figure 4.10: The protein quantity distribution of the set of simulations with 3 pre-mRNA molecules and no active transcription.

parameter regime have virtually no impact on the noise level of protein quantities but with the gene being activated for only 60 seconds which is shorter than $< \tau_2 >$, τ_2 does not have a chance to exert its impact on the overall noise. This calls for a repeat of the simulations above with a longer activation time.

4.3.5 Noise in transcription with longer activation

With a longer activation of 10 minutes, figure 4.8 in section 4.3.4 where τ_1 and τ_2 are fixed becomes figure 4.11. The modes in the distribution of protein counts are less clearly separated than their counterparts in section 4.3.4 even though the influence of the mRNA number is still visible. To see how much of the randomness is still caused by transcription, the entire transcription process is once again replaced by a fixed number of pre-mRNA

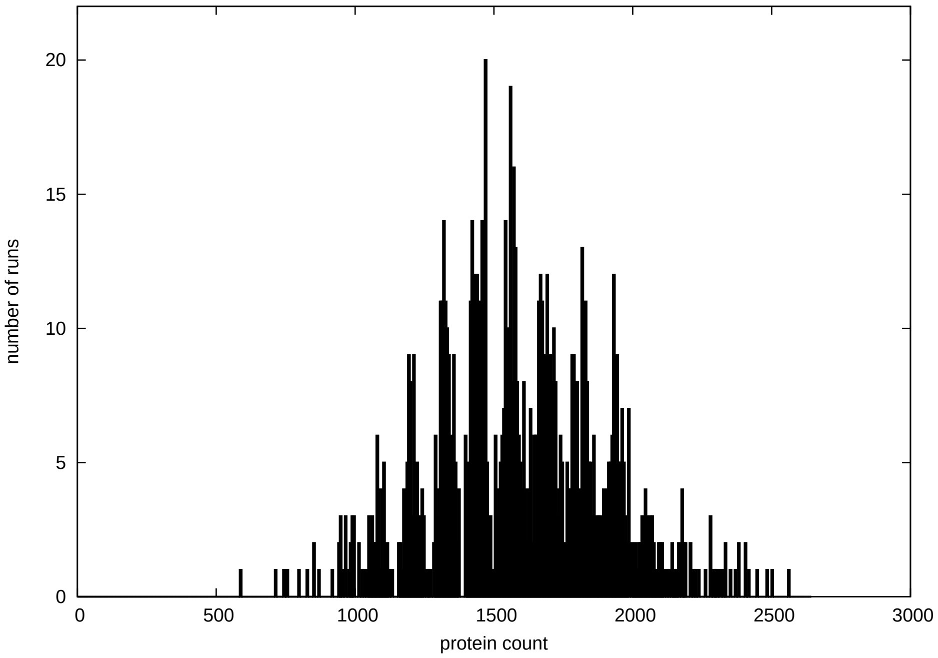

Figure 4.11: Same as the lower panel of figure 4.8 except that the activation time is 10 instead of 1 minute.

molecules.

The 10-minutes-activation-time counterpart of figure 4.10 is figure 4.12. Judging from the range of the distribution in comparison to figure 4.11, the major factor in the randomness of the protein quantity after a reasonable length of gene activation time is still transcription initiation.

4.3.6 Protein expression pattern

The protein production pattern with respect to time from the moment the gene is activated can take different shapes. Since, as demonstrated in section 4.3.4, the mRNA quantity has a multiplying effect on the protein production, the factors that control the synthesis and decay of the mRNA can affect the overall pattern. The factor that controls mRNA synthesis is the effectiveness of the promoter which in the model is represented by k_1. This in a sense

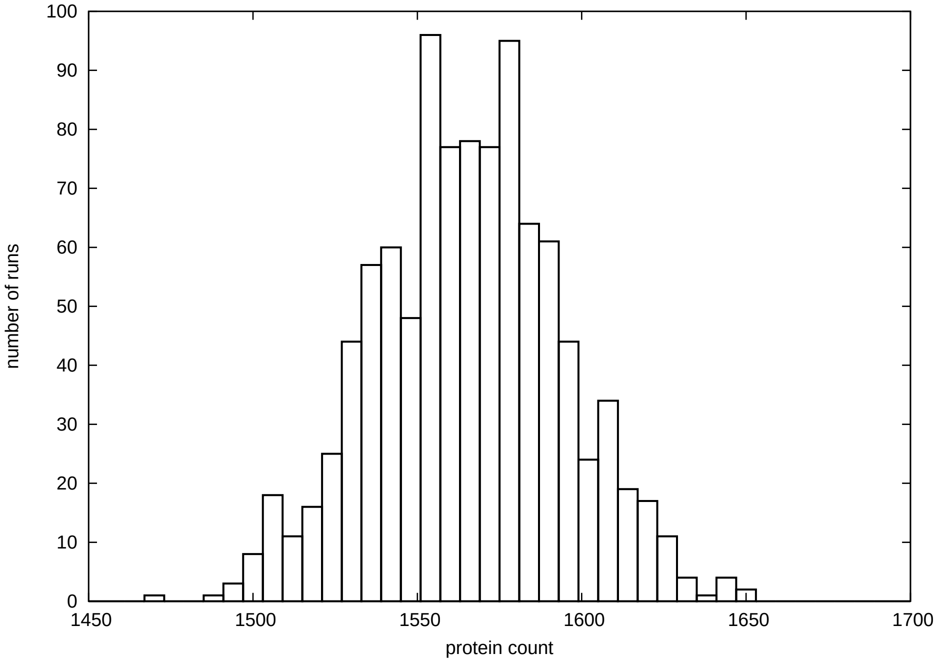

Figure 4.12: Same as figure 4.10 except that the starting pre-mRNA number is 13 instead of 3. This number is close to the mean number of pre-mRNA synthesized after 10 minutes of gene activation.

is similar to the decay process which is also determined by the rate for the decapping proteins to bind to the cap. However, the cap is protected by Pab1p which is weakened by the shortening of the poly-A tail [65] which means that the decay process is eventually controlled by the length of (or the time it takes to digest) the poly-A tail. In the model, this is represented by τ_6. Promoter efficiency and poly-A tail length are two important and common ways selected by evolution to fine-tune the expression of a gene.

Currently, in the model described in section 4.1.2, $k_1 = 0.002$ s^{-1} and this number reflects an efficient promoter. As discussed in section 4.3.3, τ_6 is drawn from a gamma distribution with a mean of 1200 s which is the digestion time for a common length of poly-A tail. By varying these two parameters, various patterns of protein synthesis can be produced. With the parameters as they are, figure 4.13 is the histogram showing the protein production rate. It shows a rise and a fall. The rise is due to the accumulation of the mRNA in the cytoplasm. Because the mRNA synthesis rate is high and its lifetime in the cytoplasm is quite long, the molecules exported earlier are still in existence when the ones exported later arrive. After that, due to the expiration of the active period of the gene, there is no longer a supply of mRNA from the nucleus; after the poly-A tails are shortened, the number of mRNA molecules in the cytoplasm falls.

If the efficiency of the promoter (k_1) is decreased, and the expected lifetime of the mRNA in the cytoplasm remains the same, the protein synthesis rate is more constant as shown in figure 4.14. This is because the accumulation of the mRNA in the cytoplasm does not have the same magnitude due to its slower production rate. At even smaller k_1, the mRNA production rate would be so low that an mRNA molecule in the cytoplasm would have already decayed before the next one appears and the protein molecules would be synthesized in bursts as shown in figure 4.15 . In the simulation, k_1 is 2×10^{-5} s^{-1} and the gene is activated for 1000 minutes because at such small k_1, the interval between mRNA synthesis is long.

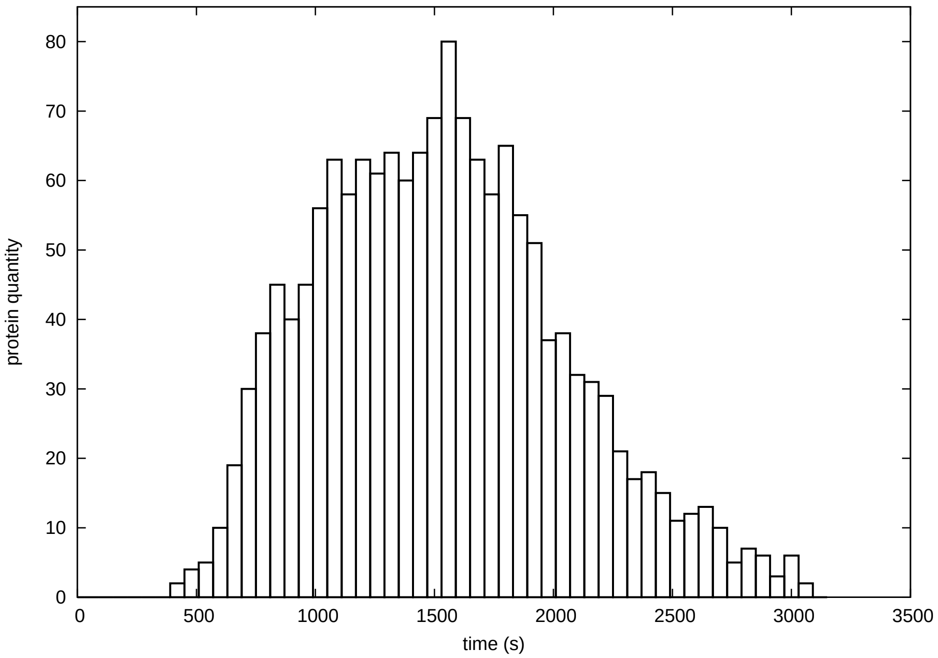

Figure 4.13: The number of protein synthesized at intervals of 60 seconds since gene activation. The parameters are set at default values that are discussed in sections 4.1.2 and 4.3.3. The gene is turned on for 10 minutes.

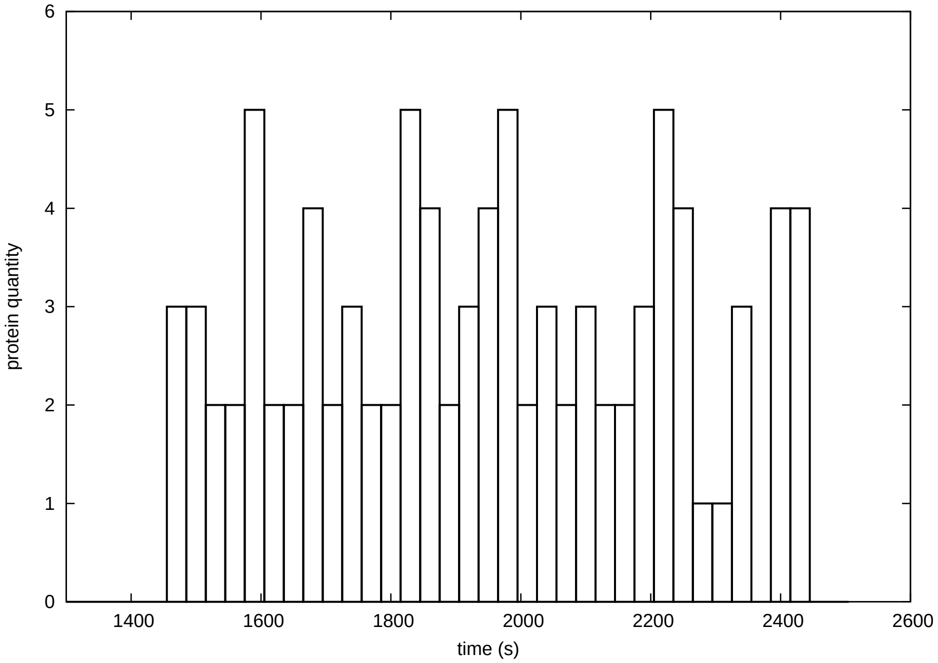

Figure 4.14: Same as figure 4.13 except that k_1 is decreased to 2×10^{-4} s^{-1} and that the bin width for the histogram is 30 seconds.

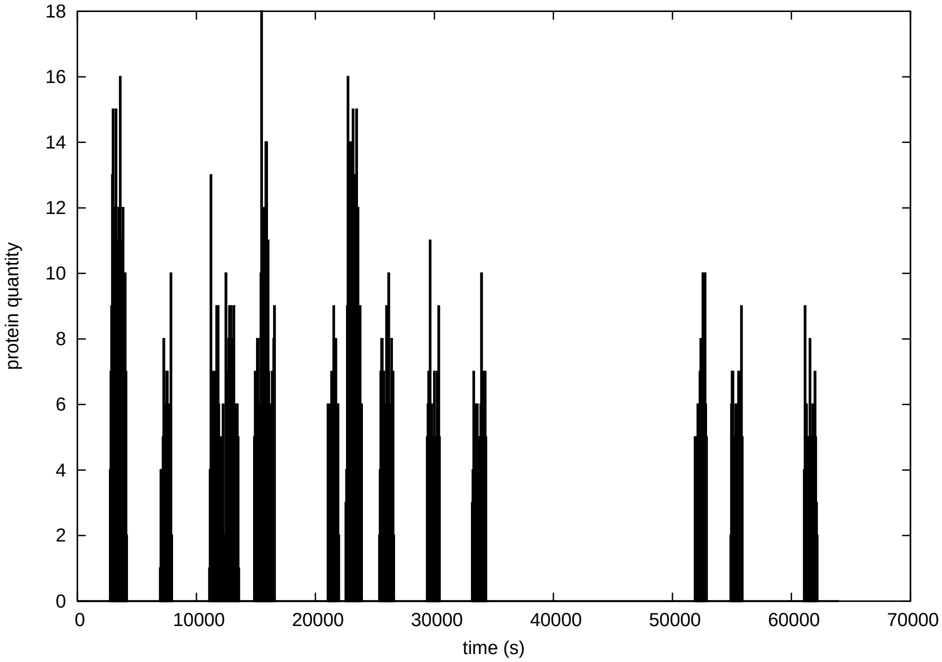

Figure 4.15: Same as figure 4.13 except that k_1 is 2×10^{-5} s^{-1} and that the gene is activated for 1000 minutes instead of 1 minute.

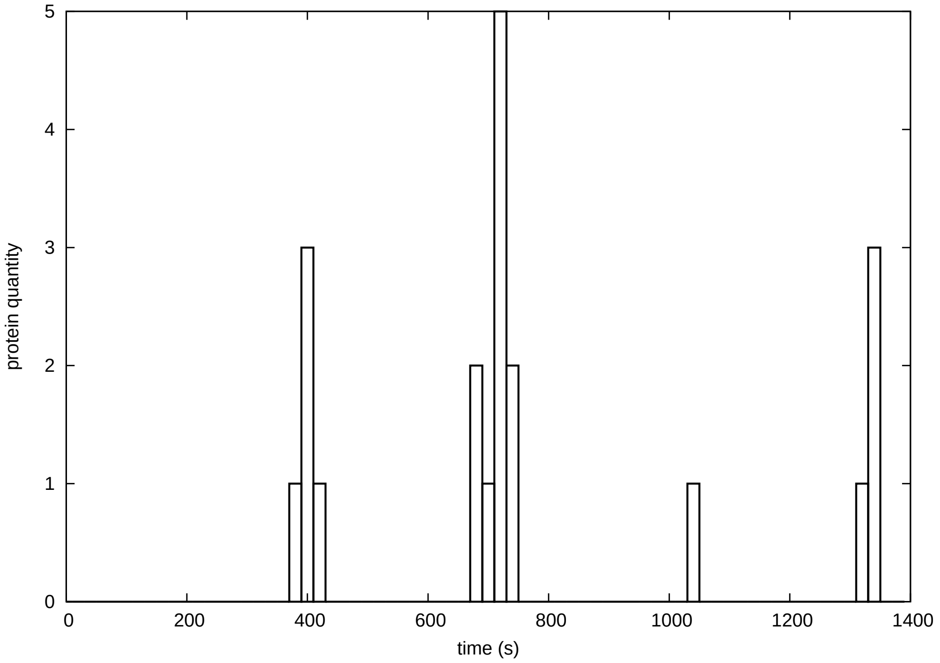

Figure 4.16: The number of protein synthesized within intervals of 20 seconds. $k_1 = 2 \times 10^{-4}$ s^{-1}, $\tau_6 = 120$s, and the activation time is 10 minutes.

A similar pattern with bursts also can be observed when k_1 is moderate (2×10^{-4} s^{-1}) and τ_6 is small as shown in figure 4.16. In this figure, the number of protein molecules in each burst is lower because the expected survival time for each mRNA molecule is only 2 minutes. This simulation produces 5 mRNA molecules.

The bursts in protein production are a result of discrete mRNA export time with most of the bursts each corresponding to a single mRNA molecule. The quantity of mRNA, in turn, is decided by the transcriptional process. All the steps before the translation have some control over the timing of the arrival of the mRNA molecules, hence the time separation of the bursts. All the steps following translation, especially the survival time of an mRNA molecule, decide the properties of the individual bursts.

4.3.7 Discussion

This delay stochastic simulation algorithm for the complete gene expression process can successfully produce significant variation in protein quantity from conditions that are identical. It is also demonstrated that the variation is primarily caused by the randomness of the transcriptional process which agrees with Hasty and Collins [31]. To reach beyond the experimental results, this model shows that the binding of the RNA polymerase to the promoter, rather than the rest of the transcription process, is the main cause of the noise. The bursts [75] in protein production are observed with the delay SSA by choosing certain sets of combinations of parameters, especially the promoter efficiency and the poly-A tail length.

If a model is a metaphor, the metaphor in this chapter has its usefulness not only in agreeing with experimental results but also in its predictive power. However, there are details where the analogy is not perfectly accurate and therefore leaves room for improvements in the future. Because the quantitative experimental data on the various steps of the gene expression process is limited, even for the well-studied *Saccharomyces cerevisiae*, certain assumptions have to be made such as the distribution function that several delays follow. Moreover, most of the reaction times are taken from different genes assuming they are typical. As there is no "typical" cell type in a multicellular organism, there might not be a typical gene in a cell so the most important improvement is to have all numerical data of the model come from a specific gene of interest.

This chapter also shows the evolution of a delay SSA model of gene expression along with the reasons behind the changes. The modification of the algorithm to suit reality should be an on-going process. For example, many genes have multiple introns so their splicings are multi-step processes. The model assumes that the splicing machinery assembly on the mRNA is co-transcriptional. In fact, the entire splicing process is co-transcriptional [15] which means there is not a delay specifically devoted to splicing and it may finish before or after the completion of transcription depending on factors such as where the splice site is. Another part of the model that needs to be modified is the delay distributions of transcription and translation. In transcription, for example, with events like abortive initiation and pausing, the distribution must be more complex than what's assumed in this chapter. Detailed transcription model can be used to produce the distribution.

Because it is the mRNA that carries the genetic information from the nucleus to the cytoplasm, its name is used in the text. However, it must be noted that mRNA is packaged into mRNP before the export process and it is the mRNP that diffuses in the nucleoplasm.

A similar model is reviewed by Ribeiro that applies delay stochastic simulation to prokaryotic gene expression [56]. It considers the translation step to be a second-order reaction that involves RBS and ribosome which is reasonable in a prokaryotic system because translation occurs near where transcription is and all the mRNAs share the same ribosome pool.

Chapter 5

Combining activation time with exit time

As observed in chapter 2, given the assumption that the mRNA molecules are synthesized outside of the uniformly dense chromatin layer, a denser chromatin layer shortens the exit time. Without a factor that counter-balances the incentive to pack the chromatin as tightly as possible to accommodate the exit speed, there would not have been the less dense euchromatin. The counter-balancing factor is the need to access the DNA in the chromatin. This requires various molecules such as the DNA glycosylase and the AP endonuclease (both involved in DNA repair) to enter the chromatin and to move with reasonable speed in order to find their targets. Their movements are favored with a less dense chromatin layer. In the gene expression process alone, the necessity for the transcription factors (TF) to find the promoter balances the extra time cost in having a less dense chromatin layer during the exit of the mRNA molecules.

In this chapter, the balancing of the opposite factors in deciding the density of the chromatin is tested. The purpose is to see if there is a chromatin density that can minimize the overall time of gene expression in the nucleus (i.e. from the time a TF enters the nucleus to the time the mRNA molecules exit).

5.1 Method

The method to simulate the finding of the promoter by the TF and the exit of the subsequent mRNA is similar to the method used in chapter 2 (more specifically section 2.3.3). For this setup in general, the density at the chromatin layer is constant and there is no PLF layer. In this case, the chromatin depth is 25, the promoter is located at (35,25), and the transcription factory is at (19,25). It starts by having a TF at one of the nuclear pores. This represents the external stimulus that is imported to the nucleus for the purpose of activating a group of genes. This TF presumably carries the nuclear localization signal so that even when close to the NPC, it cannot exit the nucleus [36].

Because the TF is different in size from an mRNA, their diffusion probability constants are different. According to the Stokes-Einstein equation, the diffusion coefficient (D) is inversely proportional to the radius of the solute [21]. The root mean square displacement equals to $\sqrt{2Dt}$, which means that the time required to move the same distance is inversely proportional to the diffusion coefficient. According to the theory of stochastic reactions [29], $\tau = {a_0}^{-1} \ln {r_1}^{-1}$ where a_0 is the overall propensity, which is proportional to the diffusion probability constant; r_1 is a random number drawn from the uniform distribution in the unit interval. The TF, therefore, has a set of DPCs that is different from that for the mRNP. The ratios of DPCs between the middle space and the chromatin layer for both molecules are the same.

The transcription factor II D (TFIID) that is involved in promoter binding and initiation of transcription has a size of 1.2 MDa [7, 60]. TFIID is the first of a series of transcription factors to recognize a gene because it has a subunit called the TATA-binding protein (TBP). The size of an mRNP molecule is highly variable. The average size of mRNA in *Saccharomyces cerevisiae* is 1.6 kb [40]. Each base in the mRNA weighs about 340 daltons; each mRNA is bound to 4 times as much protein (in weight) to become an mRNP [52]. The average mass of the mRNPs is therefore 2.7 Mda and is 2.3 times as massive as TFIID. Assuming that both the mRNP and TFIID are spherical in shape, the ratio of their radii is $\sqrt[3]{2.3} = 1.3$. This is also the ratio in their DPCs. Because the mRNPs have many sizes, three DPC ratios based on three mRNP sizes (2.7, 9.6, and 0.15 MDa) are tested. Their DPCs are summarized in table 5.1.

The transcription factor TFIID diffuses in the square space to find the promoter at (35,25). Once this happens, the TF is consumed and 50 mRNPs are synthesized at the

Table 5.1: The rate constants for various steps in the system. The values for TFIID and mRNPs of various sizes are their DPCs in the middle space; the one for mRNP addition is the stochastic rate constant for adding an mRNP to the system. The rate constants in the chromatin layer are compared to these. For example, by saying that the DPC in the chromatin layer is 0.5, it is in fact 50% of that in the middle space.

	TFIID	mRNP diffusion			mRNP addition
		2.7MDa	9.6MDa	0.15MDa	
rate constants	0.56	0.4	0.28	1.12	0.112

transcription factory. This number is to demonstrate the long-term trend of the exit time; it is further discussed in section 5.3. The mRNP addition rate is chosen so that it takes a relatively short period of time to add all the mRNPs. Even though it takes some time in reality for the promoter to associate with the transcription factory after it is found, and it also takes time for the RNA polymerase to move along the gene before the first mRNA appears, neither process is, presumably, influenced by the density of the chromatin. This simulation addresses the influence of the chromatin on the time of gene expression. Therefore, these two processes are not taken into account in the simulation and the mRNA release starts as soon as the promoter is found. The simulation completes when every mRNP is out of the system through one of the pores.

5.2 Results

The exit times of the mRNPs at various chromatin-layer DPCs (in comparison to the DPC in the middle space, i.e. expressed as a ratio to the chromatin DPC) were recorded. Each exit time is the average over 60 runs with identical parameters and initial conditions. Plotted against the exit order (i.e. the first to the fiftieth to exit at each activation), the exit times are shown in figure 5.1. It shows that the lines representing DPCs from 1 to 0.4 are closer together than the three that represent DPCs of 0.1 to 0.3. Figure 5.2 is a plot of DPCs only from 1 to 0.4. To draw any conclusions from these figures, one has to be sure

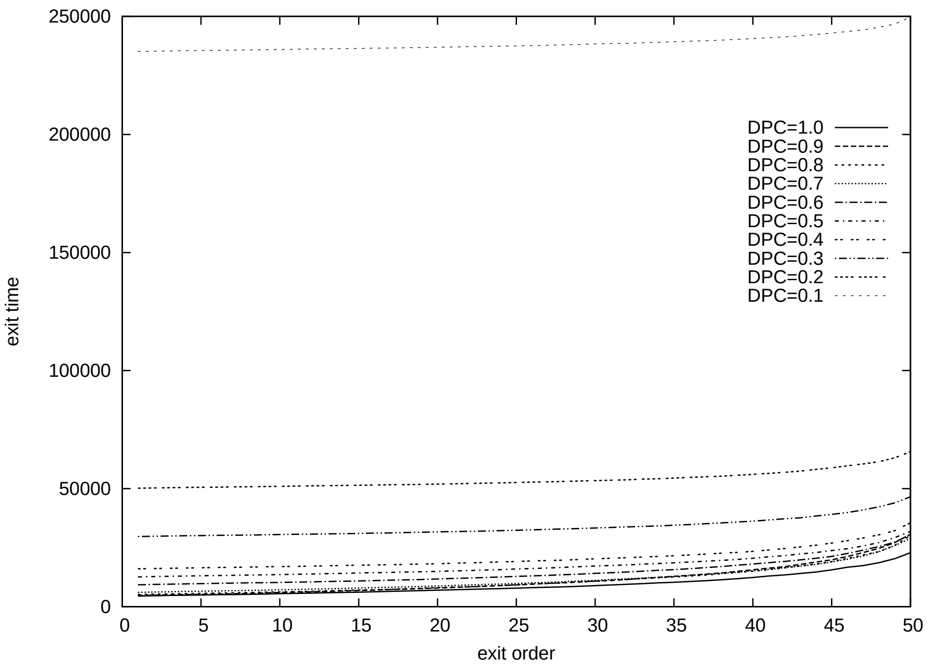

Figure 5.1: The exit times at various chromatin-layer DPCs. The DPC ratio between a TF and an mRNP both in the middle space and in the chromatin layer is 1.3. The abscissa is the exit order, i.e. the first, the second, and so on to exit the system.

that the result is not contingent on the particular realizations generated by the underlying stochastic process. From the looks of the curves, they seem reasonably smooth, which is an indication that the statistical fluctuations are not particularly large. The standard errors (SEs), however, do not concur with this notion. The SEs for all the points on the figure calculated over 60 runs are very high (many of them are over 10% of the mean). The reason for such high SEs for such smooth curves is that each curve consists of two steps: the seeking for the promoter by TFIID and the exit of the mRNPs released after the promoter is found. As there are 50 mRNP molecules for each curve and they are released within a relatively short period of time, the time it takes to exit for each exit order is reasonably consistent. For example, the first exit can be expected to happen shortly after most of the mRNPs have been released even though in one instance it was the third mRNP released that exited first, and in another instance, the eighth. This produces a smooth curve. On

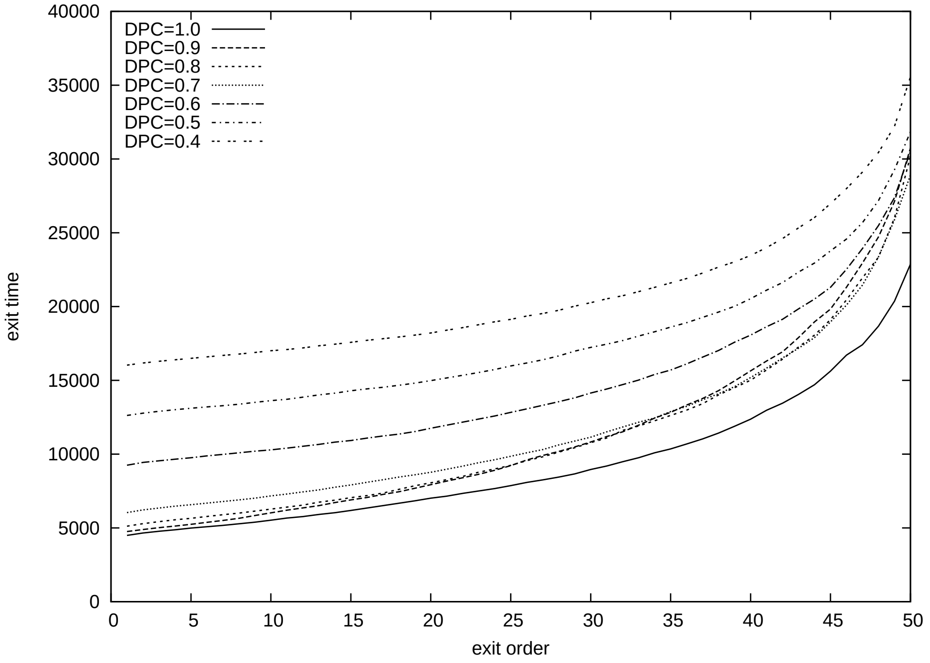

Figure 5.2: The lower part of figure 5.1.

the other hand, however, each curve consists of only one finding of the promoter by one TFIID molecule (although averaged over 60 runs) so it makes a huge difference in searching time if the TFIID molecule spends more time surveying the space than going to the promoter following a relatively straight path. Even though the same process is simulated 60 times and averaged, it is not enough to significantly diminish the stochastic effect. In summary, each curve in figure 5.2 has a wide range to start at but once started, the shape of it is reproducible. The obvious solution is to produce each curve with much more than 60 runs but because the simulation as it stands already takes more than a day to run on a four-core Xeon Mac OS X system, there is a time constraint on how many runs there can be. However, since the high SE only occurs in the seeking step, it alone can be repeated many more times than the exit step (i.e. from the moment TFIID found the promoter to when the last mRNP exited the system). This separation is possible because even though the two steps take place consecutively, the time it takes for one to complete is independent of the other. The seeking step is expected to take much less computation power than the

Table 5.2: The standard errors of the seeking times at various chromatin-layer DPCs. Each value is produced from a sample of 3000 identical runs

DPC	SE	Mean
1.0	66.4	4105
0.9	80.4	4744
0.8	91.5	5608
0.7	111.1	6727
0.6	144.0	8662
0.5	194.8	11356
0.4	277.6	16236
0.3	461.3	26699
0.2	975.7	55207
0.1	3782.4	212560

exit step because it is suspected that most of the time in the exit step is spent simulating the exit of the last few mRNP molecules which might be mired in the chromatin layer. This notion is proven to be consistent with reality because repeating the seeking step 3000 times takes about a day of computational time. From now on, unless otherwise indicated, the number of repetitions for the seeking step is 3000 and 150 for the exit step.

The seeking time after 3000 runs still has a CV of over 0.5 because the variation from one run to the next is wide. In a situation like this, one cannot expect the time it takes for a TFIID molecule to find a promoter to be reproducible because it is a stochastic system. However, the result is consistent when there is a large number of molecules being considered because the mean of the first 1500 runs is very close to the second 1500 runs (their difference is less than 1%). A more convincing measure of the reproducibility of the mean of a sample is its standard error which is the standard deviation of the sample divided by the square root of the sample size. The SEs of the exit step range from 11.0 to 472.6 with all the ones greater than 300 in the last three in exit order (48th, 49th, and 50th). Table 5.2 summarizes the standard errors for the seeking step. Figure 5.3 shows the seeking time distribution at DPC= 1, 0.6, and 0.1. It demonstrates the difference that the chromatin-layer DPC makes in finding the promoter by the TF at the beginning.

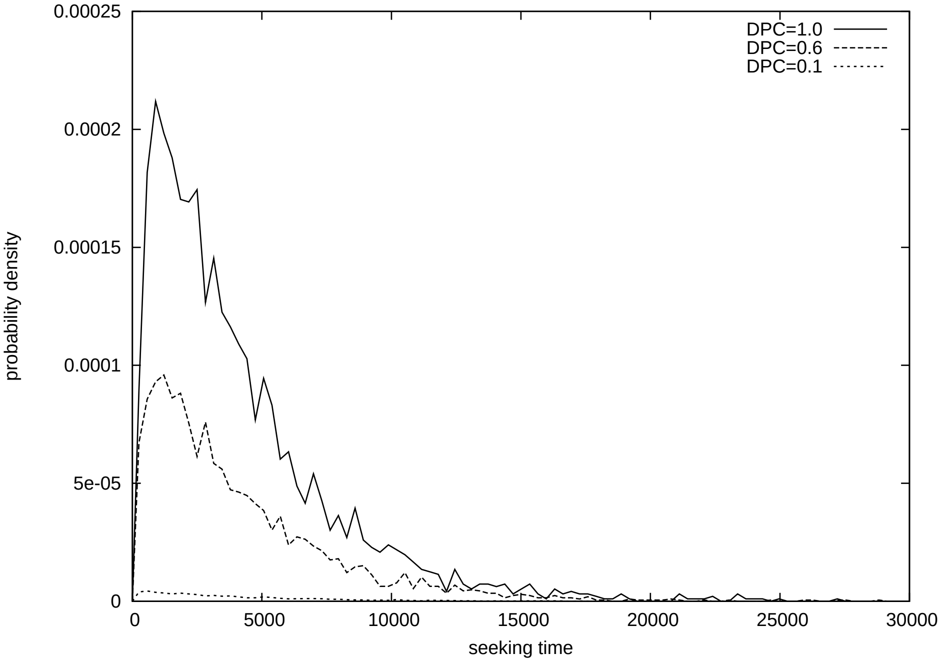

Figure 5.3: The normalized histogram of the seeking time for the transcription factor to find the promoter generated using 3000 data points at DPC= 1, 0.6, and 0.1.

Figure 5.4 is the same plot as figure 5.2 with more runs. It shows that although the general trend is that exit time increases with lower chromatin-layer DPC, there are exceptions. DPCs 1 to 0.7 are close together for the first few mRNPs. Although DPC= 0.8 is above DPC= 0.9 for the beginning and middle parts, it goes beneath towards the end. This is indicative of the opposing effects of having a lower DPC: it hinders the finding of the promoter by the TF but promotes the exit of the mRNP molecules. Overall, the ideal DPC value to help the exit of the mRNP is 1. However, to have the density in the chromatin layer be the same as that in the middle space is almost impossible because it requires the chromatin to thin out over a large volume. To provide such a volume, the nucleus itself must be big enough. A larger volume would mean more time required to find a target by diffusion. With the same amount of chromatin material, a more densely packed structure allows for faster exit simply by decreasing the volume needed to contain it. Although in a constant volume, an increase in density tends to increase the exit time, there are exceptions

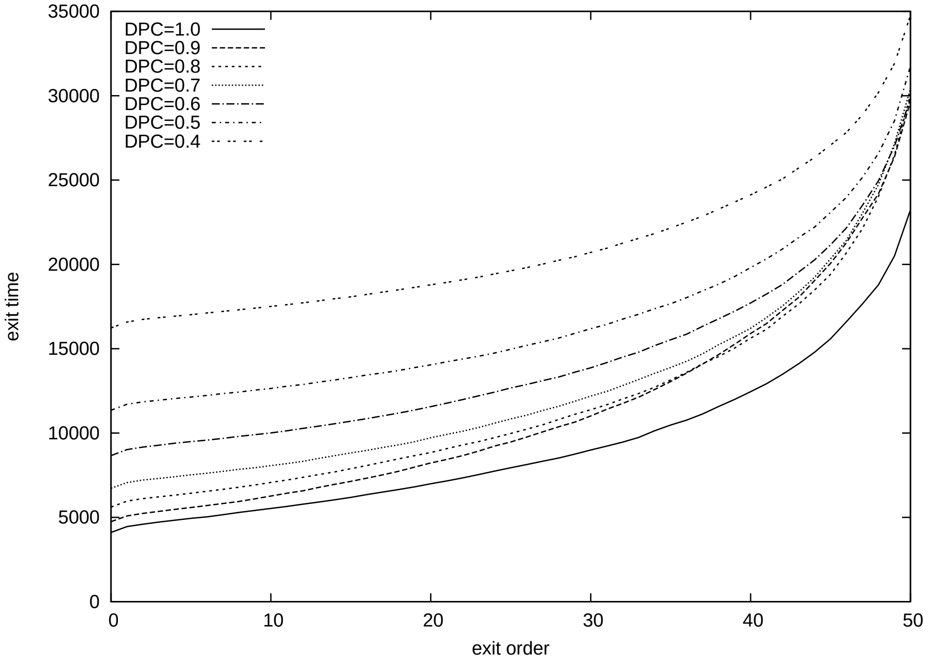

Figure 5.4: The same as figure 5.2 except that the seeking times are each the average of 3000 runs and the exit times are each the average of 150 runs. The number 0 on the abscissa indicates the time at which the promoter is found by TFIID.

according to the simulation. With a large number of mRNP molecules produced each time a gene becomes active, having the DPC at 0.8 can shorten the time it takes to export most of the molecules. This does not take into account the necessity to expand the volume of the nucleus with a higher DPC (lower density). With it, DPC= 0.8 should be even better favored over the higher DPCs. DPCs at 0.7 and 0.6 both have the almost the same time in getting all 50 of the mRNP out as DPC= 0.9. With the volume effect, these two DPCs can be even more desirable.

Figure 5.5 shows the same plot as figure 5.4 except that the mRNP addition rate is 10-fold slower. These two figures look very similar and the little difference is at the beginning of the curves: onward from the first in exit order, figure 5.5 shows a phase where the curves slightly concave down; whereas figure 5.4 shows no such phase. The reason is that it takes longer for the first few mRNPs to exit with a slower addition rate. This difference is

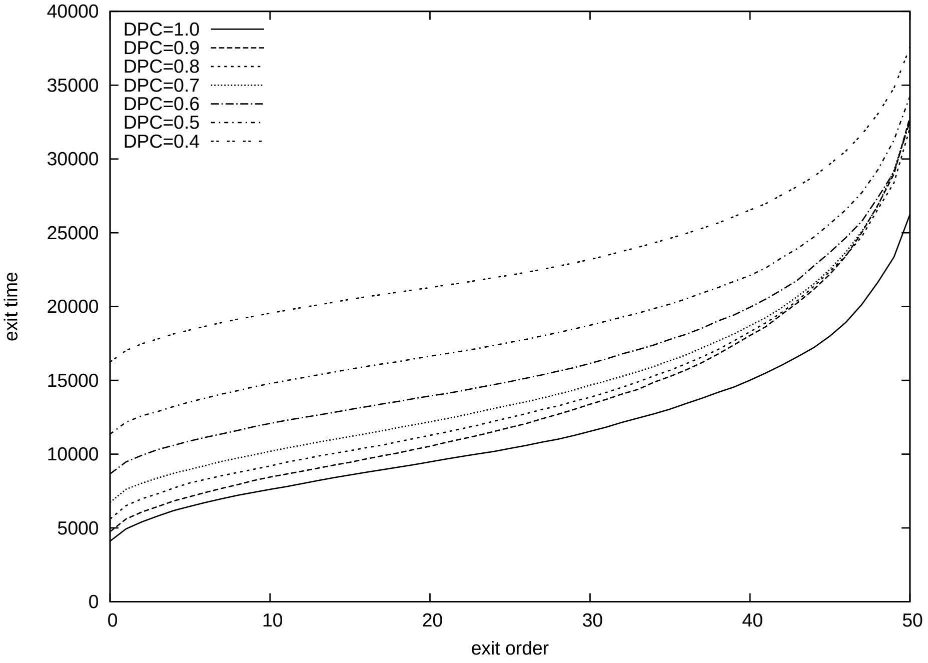

Figure 5.5: Same as figure 5.4 except with 10-fold slower mRNP addition rate.

observed in all the curves so it has little impact on the optimal density of the chromatin.

Figure 5.6 shows the exit times of a bigger mRNP. The DPC ratio between the TF and the mRNP is 2 instead of 1.3. The size of the mRNP, therefore, is $2^3 \times 1.2 = 9.6$ MDa. The major difference between this figure and figure 5.4 is that the times for the later-to-exit molecules are bigger in this case, reflecting a smaller DPC due to a larger molecular size. Because the change in the DPC for the mRNP does not affect the seeking time of the TF for the promoter, all the curves in figure 5.6 start at the same points as their counterparts in figure 5.2. Because a smaller DPC for the mRNPs prolongs the exit time, the exit time makes up a larger proportion in the overall time (including seeking time). Therefore, a means to shorten the exit time by the same percentage is more helpful in improving the overall time. This is why the lines towards the end are closer to each other in figure 5.6 than in figure 5.2. The diverging of DPC= 1 and DPC= 0.9 is indicative of the long-tailed behavior discussed in chapters 2 and 3.

The exit times of a smaller mRNP is shown in figure 5.7. The DPC of the mRNP is

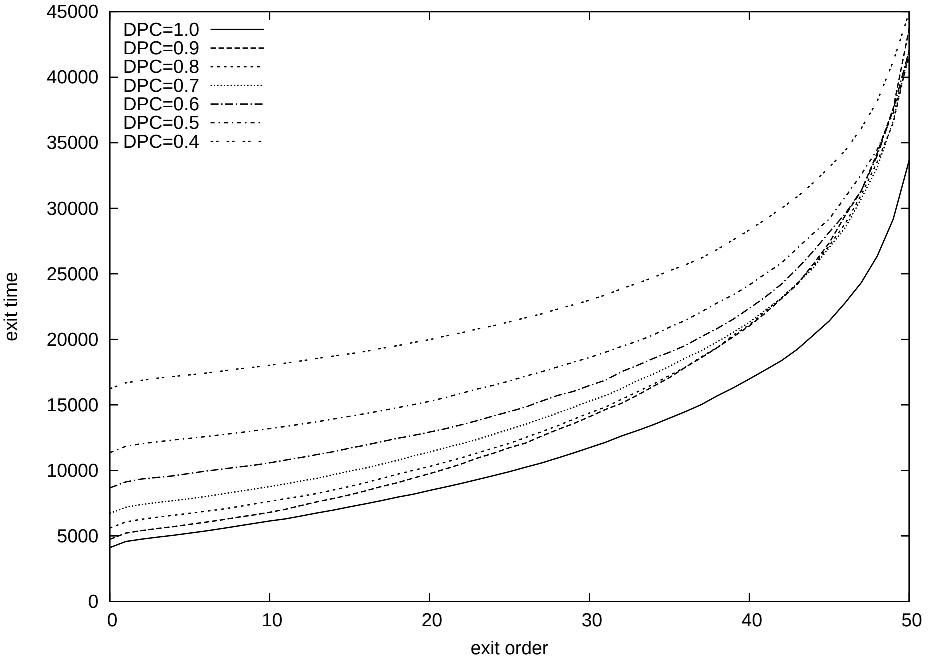

Figure 5.6: Similar to figure 5.2. The only difference is that the size of the mRNP in this case is 9.6 MDa so the DPC ratio between the mRNP and the TF is 2.

twice that of the TF so the size is one-eighth (1200/8 = 150 kDa). Because, in this case, the DPC of the mRNP is relatively large, the separation of the curves is predominantly due to the seeking of the TF for the promoter. Therefore, in the case of a smaller mRNP size, the system favors a lower chromatin density.

5.3 Discussion

The trend shown in the simulations serves as a proof of concept in that it shows what is possible under certain conditions. The actual nucleus differs from the simulation in many ways. For example, although the export of the mRNA is considered one of the primary functions of the nucleus, there are other important factors that may influence the organization of the chromatin such as the need for molecules involved in DNA repair to enter and move within the chromatin, therefore favoring a less dense chromatin. It might

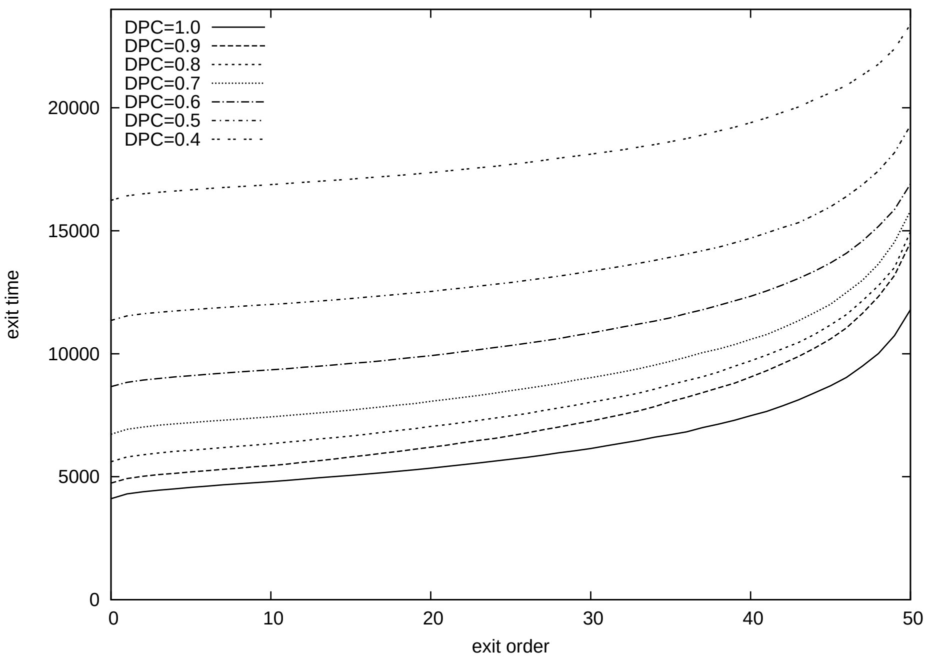

Figure 5.7: Similar to figure 5.4. The only difference is that the size of the mRNP in this case is 150 kDa so the DPC ratio between the mRNP and the TF is 0.5.

seem desirable to organize the chromatin in such a way that it shortens the time of gene expression, but sometimes the opposite is true. One level of control for the gene activity is to have the active genes being accessible to the TF while keeping the less active genes physically hidden [13].

In the simulation, the TBP is assumed to be bound to TFIID. This might not always be the case. While TFIID is important in the transcription of some promoters, its absence has little effect on others [49]. On this regard, TBP perhaps is able to search for the genes by itself. Because the size of TBP is much smaller, it would allow a nucleus with denser chromatin to be functional. It would also allow genes in the denser chromatin region to be active.

What can be learned from this chapter is that there are two opposing factors in shaping the density of the chromatin: one is to allow the molecules to enter the chromatin; the other is to prevent molecules to be exported out of the nucleus from getting hindered.

A natural follow up to the work in this chapter is to change other parameters such as the chromatin depth and the locations of the transcription factory and the promoter. As far as the exit time is concerned the difference between this chapter and chapter 2 is that in this chapter mRNP complexes are only synthesized after the promoter has been found. In other words, the time is greater than 0 when the first mRNP molecule enters the system. In chapter 2, on the other hand, time is counted strictly between a molecule's release into the system and its exit. To reconcile these two chapters, provided that the mRNP addition rate is large enough (i.e. the time gaps between successive mRNP releases are small enough), the results in chapter 2 can be used instead of the exit time in this chapter because there are more runs in more scenarios in chapter 2.

This chapter does not explicitly take into account the volume effect mentioned earlier i.e. at higher chromatin density, the volume occupied by chromatin is smaller. In future work, it could be considered. One difficulty in doing so is the find the correlation between the DPC and the volume. For example, if the DPC is halved, how much volume can it save? To formulate this concept, suppose V_0 is the volume for certain amount of chromatin if it is in heterochromatin state (i.e. it is impermeable to diffusion), the volume occupied by the same amount of chromatin at a less dense concentration can be expressed as: $V_c = f \cdot V_0$, where f is the factor by which heterochromatin expands and it is a function of DPC and the molecular size of the molecule of interest.

A different way to model diffusion in the nucleus is to treat the chromatin as a porous region so that molecules above a certain size are excluded from the chromatin territory. This might explain why the mRNA molecules are packaged with so much protein to form the mRNP. This way, the mRNP molecules tend to be excluded from the chromatin region. The large quantity of protein associated with each mRNA is curious especially because the tertiary structure of the mRNA in the nucleus is irrelevant to its function.

Chapter 6

Conclusion

Gene expression in the eukaryotic cell system is a complex process that involves many major steps with each step influenced by several factors. The overall ramifications of changing some of the factors can be obvious: for example, moving a transcription location closer to the NPC certainly will shorten the average time for the mRNA molecule to exit; the effect of other changes might not to so simple: an example would be to replace the patch of empty nucleoplasm near the transcription location with a layer of chromatin with a gradient of density that decreases towards the direction of the nuclear envelope. This thesis addresses some of the factors and shows their influence on either the entire gene expression or its major steps.

Using the Gillespie algorithm, chapter 2 simulates the exit of the mRNA in several nuclear conditions. In a space where the exit sites are located at the top and the bottom is layered with chromatin whose density increases towards the bottom, the effect of limiting the space that a molecule has to explore by filling some of the space with chromatin overcomes the effect of hindering the frequency of molecular movements. As the result, the molecule in the chromatin environment that has a density gradient takes less time to exit on average. The gradient not only fills the space but also exerts a tendency to move the molecules towards its surface. Moving of the mRNA release site closer to the exit sites mainly decreases the exit times of the molecules that exit early because their paths are more or less straight. The rest of the molecules, to different degrees, explore the space more thoroughly. A constant-density chromatin layer at the bottom also shortens the average exit time despite the fact that the movement of the molecules in the chromatin layer is slowed. The increase of chromatin density at the bottom decreases the exit time up to a point beyond which a further increase of density would have no effect. This is due to the exclusion of the majority of the molecules from the chromatin layer. Without the chromatin layer, having the release site closer to the exit sites on average does not cause a significant improvement in the exit time; with a constant-density chromatin layer at the bottom, however, the location of the mRNA synthesis site relative to the chromatin matters significantly. As long as the release site is above the chromatin layer, its location has little impact; having the release site in the chromatin layer, on the other hand, increases the exit time many-fold. The depth of the release site into the chromatin layer also has a great impact. Restricting the molecular movements more in the up-down direction and less in the left-right direction shortens the exit time but there is an optimum level in the restriction beyond which its influence on the exit time is the opposite.

There is no probability distribution function that can fit the exit time distribution of all the scenarios described above. The one that does the best is Weibull which can fit, to a satisfactory degree, almost all the cases except the one that has a thick layer of constant-density chromatin at the bottom. The chromatin traps some of the molecules and delays their exit. As a result, the distribution in this case is heavy-tailed.

Chapter 4 models the complete eukaryotic gene expression pathway using delay stochastic simulation. With a fixed period of gene activation, having more ribosomes not only produces more protein but also keeps the life span of the mRNAs longer. This is because there is a competitive relationship between the ribosome and the decapping enzymes that are involved in the degradation of mRNA. With its 5’ CAP structure spending more time bound to a ribosome, an mRNA has less chance to bind to the decapping enzyme, hence prolonging its life span. There is then a change in the equations to account for, first, the notion that the number of ribosomes used by the mRNAs of one gene has very little impact on the overall population of ribosome and, second, the fact that the mRNA molecules newly arrived at the cytoplasm are not immediately subjected to degradation. The stochasticity of the simulation captures the stochastic nature of the biological system. The randomness in generating the delays has little impact on the overall randomness of the system. The main source of stochasticity comes from the reacting event (i.e. the uncertainty of two species binding to each other). By varying the promoter efficiency and the length of the poly-A tail which are two common ways a cellular system adopts to fine-tune its gene expression, the model can produce various protein production patterns that include peaked distribution, flat and consistent distribution, and a pattern with many bursts.

Given that a higher chromatin density at the bottom shortens the exit time, the counter-balancing factor that prevents the chromatin from adopting an extremely high density is the need for molecules to access the DNA. As far as gene expression is concerned, the main reason to access DNA is for transcription factors to find the promoters. Chapter 5 studies the balance between having loosely packed chromatin to allow access and having densely packed chromatin to encourage the molecules' exit. The result shows that having a denser chromatin layer delays the finding of the promoter. On the other hand, a denser chromatin does not help improve the exit time of the first few mRNPs that find the exit sites because their paths are reasonably straight from the release site and the exit site. Therefore, the time intervals from the entering of the transcription factor (or transcription factor complex) to the exit of mRNP for the first few molecules increase with increasing chromatin density. However, for the mRNP molecules that survey the space before finding an exit site, a denser chromatin layer helps in limiting the space they have to explore hence shortening their exit times. As the result, the time from the entering of the TF to the exit of all the molecules that are released due to one association of the gene to a transcription factory is similar between several chromatin densities. The details depends on the size of the mRNP molecules. For larger-sized mRNP molecules, it takes longer for them to exit overall. The help by the chromatin layer to decrease their exit times would be more prominent comparing to the time it takes for the TF to find the gene.

In the cases of chapter 2 and 5, a two-dimensional square space is used to represent a three-dimensional space in the nucleus. Although some of the results are checked against the 3D model and they are consistent with each other, it would be an improvement to obtain all the results using the 3D model. The current drawback is the computational speed of modern computers. Ultimately, instead of using a square space (or cubic space for 3D), one should be able to model the entire nuclear space with a sphere. In addition to the nuclear structures modeled in chapter 2, another structure that is worth investigating is the nuclear-envelope-associated heterochromatin. Due to its high density, heterochromatin excludes the massive molecules from entering. One could study the effect of its presence near the nuclear envelope between the NPCs on the exit time of mRNA molecules. Since the chromatin layer (heterochromatin or euchromatin) discourages the entering of massive molecules, it could be interesting to view the macromolecules in the nucleus as members of two groups: the ones that need to enter the chromatin and the ones that need to be excluded from the chromatin. Then one would characterize members of each group by measuring their sizes. One hypothesis is that those that need access to the chromatin tend to have smaller sizes than those that do not. If this hypothesis holds, it would favor the notion that TBP searches for the promoter alone without binding to TFIID because TFIID is more massive and may be difficult to move in the chromatin.

Bibliography

[1] Paul S. Agutter. Models for solid-state transport: messenger RNA movement from nucleus to cytoplasm. *Cell Biol. Int.*, 18:849–858, 1994.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell.* Garland, fourth edition, 2002.

[3] A. Audibert, D. Weil, and F. Dautry. In vivo kinetics of mRNA splicing and transport in mammalian cells. *Mol. Cell. Biol.*, 22:6706–6718, 2002.

[4] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med*, 79:137–158, 1944.

[5] Bonven B. and Gulløv K. Peptide chain elongation rate and ribosomal activity in *Saccharomyces cerevisiae* as a function of the growth rate. *Mol. Gen. Genet.*, 170:225–230, 1979.

[6] T. Berleth, M. Burri, G. Thoma, D. Bopp, S. Richstein, G. Frigerio, M. Noll, and C. Nusslein-Volhard. The role of localization of bicoid RNA in organizing the anterior pattern of the *Drosophila* embryo. *EMBO J.*, 7:1749–1756, 1988.

[7] Suparna Bhattacharya, Shinako Takada, and Raymond H. Jacobson. Structural analysis and dimerization potential of the human TAF5 subunit of TFIID. *PNAS*, 104:1189–1194, 2007.

[8] T. Borggrefe, R. Davis, A. Bareket-Samish, and R. D. Kornberg. Quantitation of the RNA polymerase II transcription machinery in yeast. *J. Biol. Chem.*, 267:47150–47153, 2001.

[9] Dmitri Bratsun, Dmitri Volfson, Lev S. Tsimring, and Jeff Hasty. Delay-induced stochastic oscillations in gene regulation. *Proc. Natl. Acad. Sci. U.S.A.*, 102:14593–14598, 2005.

[10] Maurice C. Bryson. Heavy-tailed distributions: Properties and tests. *Technometrics*, 14:61–68, 1974.

[11] Kevin P. Byrne and Kenneth H. Wolfe. The yeast gene order browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, 15:1456–1461, 2005.

[12] Giordano Caponigro and Roy Parker. Multiple functions for the poly(A)-binding protein in mRNA decapping and deadenylation in yeast. *Gene. Dev.*, 9:2421–2432, 1995.

[13] Lyubomira Chakalova, Emmanuel Debrand, Jennifer A. Mitchell, Cameron S. Osborne, and Peter Fraser. Replication and transcription: Shaping the landscape of the genome. *Nature Reviews Genetics*, 6:669 – 678, 2005.

[14] Charles N. Cole and John J. Scarcelli. Transport of messenger RNA from the nucleus to the cytoplasm. *Curr. Opin. Cell Biol.*, 18:299–306, 2006.

[15] P. Cramer, A. Srebrow, S. Kadener, S. Werbajh, M. de la Mata, G. Melen, G. Nogus, and A. R. Kornblihtt. Coordination between transcription and pre-mRNA processing. *FEBS Lett.*, 498:179–182, 2001.

[16] Thomas Cremer, Marion Cremer, Steffen Dietzel, Stefan Müler, Irina Solovei, and Stanislav Fakan. Chromosome territories — a functional nuclear landscape. *Current Opinion in Cell Biology*, 18:307 – 316, 2006.

[17] Bertil Daneholt. Assembly and transport of a premessenger RNP particle. *Proc. Natl. Acad. Sci. U.S.A.*, 98:7012–7017, 2001.

[18] Sergio U. Dani, Akira Hoir, and Gerhard F. Walter. *Principles of Neural Aging*. Elsevier Science Pub Co, first edition, 1997.

[19] Albert D. G. de Roos. The origin of the eukaryotic cell based on conservation of existing interfaces. *Artificial Life*, 12:513–523, 2006.

[20] C. Dingwall, J. Robbins, S. M. Dilworth, B. Roberts, and W. D. Richardson. The nucleoplasmin nuclear location sequence is larger and more complex than that of SV-40 large T antigen. *The Journal of Cell Biology*, 107:841–849, 1988.

[21] Albert Einstein. Eine neue bestimmung der molekuldimentionen. *Ann. Physik*, 324, 1906.

[22] D. W. Fawcett. *An Atlas of Fine Structure: The Cell, Its Organelles and Inclusions*. W. B. Saunders Co., Philadelphia, 1966.

[23] Carl M. Feldherr and Debra Akin. The location of the transport gate in the nuclear pore complex. *J. Cell Sci.*, 110:3065–3070, 1997.

[24] Beatriz M. A. Fontoura, Samuel Dales, Gunter Blobel, and Hualin Zhong. The nucleoporin Nup98 associates with the intranuclear filamentous protein network of TPR. *Proc. Natl. Acad. Sci. U.S.A.*, 98:3208–3213, 2001.

[25] W. W. Franke. On the universility of the nuclear pore complex structure. *Z. Zellforsch*, 105:405–429, 1970.

[26] C. W. Gardiner. *Handbook of Stochastic Methods*. Springer, Berlin, second edition, 1985.

[27] M. A. Gibson and J. Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The Journal of Physical Chemistry A*, 104:1876–1889, 2000.

[28] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.

[29] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2240–2361, 1977.

[30] J. Han and J. Herzfeld. Macromolecular diffusion in crowded solutions. *Biophys. J.*, 65:1155–1161, 1993.

[31] Jeff Hasty and James J. Collins. Translating the noise. *Nat. Genet.*, 31:13–14, 2002.

[32] Vishwanath Iyer and Kevin Struhl. Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.*, 93:5208–5212, 1996.

[33] O. A. Y. Jackson. An analysis of departures from the exponential distribution. *J. Roy. Statist. Soc. Ser. B*, 29:540–549, 1967.

[34] C. W. Jones, M. W. Dalton, and L. H. Townley. Interspecific comparisons of the structure and regulation of the drosophila ecdysone-inducible gene E74. *Genetics*, 127:535–543, 1991.

[35] Arie Kaffman and Erin K. O'Shea. Regulation of nuclear localization: A key to a door. *Annu. Rev. Cell Dev. Biol.*, 15(1):291–339, 1999.

[36] Daniel Kalderon, Bruce L. Roberts, William D. Richardson, and Alan E. Smith. A short amino acid sequence able to specify nuclear location. *Cell*, 39:499 – 509, 1984.

[37] Elena Kiseleva, Sheona P. Drummond, Martin W. Goldberg, Sandra A. Rutherford, Terence D. Allen, and Katherine L. Wilson. Actin- and protein-4.1-containing filaments link nuclear pore complexes to subnuclear organelles in *Xenopus* oocyte nuclei. *J. Cell Sci.*, 117:2481–2490, 2004.

[38] Tomasz J. Kozubowski, Anna K. Panorska, Fares Qeadan, Alexander Gershunov, and Debra Rominger. Testing exponentiality versus pareto distribution via likelihood ratio. *Commun. Stat. Simulat.*, 38:118–139, 2009.

[39] Hartmut Kuthan. A mathematical model of single target site location by Brownian movement in subcellular compartments. *J. Theor. Biol.*, 221:79–87, 2003.

[40] Benjamin Lewin. *Genes IX*. Jones & Bartlett Publishers, ninth edition, 2007.

[41] M. Lorenz, D. Popp, and K. C. Holmes. Refinement of the F-actin model against X-ray fiber diffraction data by the use of a directed mutation algorithm. *J. Mol. Biol.*, 234:826–836, 1993.

[42] Katherine Luby-Phelps. Effect of cytoarchitecture on the transport and localization of protein synthetic machinery. *J. Cell. Biochem.*, 52:140–147, 1993.

[43] Ian G. Macara. Transport into and out of the nucleus. *Microbiol. Mol. Biol. Rev.*, 65:570–594, 2001.

[44] Lynn Margulis. *Symbiosis in Cell Evolution*. W. H. Freeman and Company, San Francisco, 1981.

[45] P. B. Mason and K. Struhl. Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. *Mol. Cell*, 17:831–840, 2005.

[46] G. G. Maul and L. Deaven. Quantitative determination of nuclear pore complexes in cycling cells with differing DNA content. *J. Cell Biol.*, 73:748–760, 1977.

[47] O. L. Miller Jr., Barbara A. Hamkalo, and C. A. Thomas Jr. Visualization of bacterial genes in action. *Science*, 169:392–395, 1970.

[48] Chris Molenaar, Abadir Abdulle, Aarti Gena, Hans J Tanke, and Roeland W Dirks. Poly(A)+ RNAs roam the cell nucleus and pass through speckle domains in transcriptionally active and inactive cells. *J Cell Biol*, 165(2):191–202, 2004.

[49] Zarmik Moqtaderi and Yu Bai. TBP-associated factors are not generally required for transcriptional activation in yeast. *Nature*, 383:188, 1996.

[50] D. Muhlrad, C. J. Decker, and R. Parker. Turnover mechanism of the stable yeast PGK1 mRNA. *Mol. Cell. Biol.*, 15:2145–2156, 1995.

[51] Nelly Panté and Ueli Aebi. Sequential binding of import ligands to distinct nucleopore regions during their nuclear import. *Science*, 273:1729–1732, 1996.

[52] Joan C. Politz, Elizabeth S. Browne, David E. Wolf, and Thoru Pederson. Intranuclear diffusion and hybridization state of oligonucleotides measured by fluorescence correlation spectroscopy in living cells. *Proc. Natl. Acad. Sci. U.S.A.*, 95:6043–6048, 1998.

[53] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135:216–226, 2008.

[54] Stephen Ramsey, David Orrell, and Hamid Bolouri. Dizzy: Stochastic simulation of large-scale genetic regulatory networks. *J. Bioinform. and Computat. Biol.*, 3:415–436, 2005.

[55] Uta Raue, Stefan Oellerer, and Sabine Rospert. Association of protein biogenesis factors at the yeast ribosomal tunnel exit is affected by the translational status and nascent polypeptide sequence. *Journal of Biological Chemistry*, 282(11):7809–7816, 2007.

[56] Andre S. Ribeiro. Stochastic and delayed stochastic models of gene expression and regulation. *Math. Biosci.*, 223:1–11, 2010.

[57] Marc R. Roussel. The use of delay differential equations in chemical kinetics. *J. Phys. Chem.*, 100:8323–8330, 1996.

[58] Marc R. Roussel and Rui Zhu. Stochastic kinetics description of a simple transcription model. *Bull. Math. Biol.*, 68:1681–1713, 2006.

[59] Marc R. Roussel and Rui Zhu. Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. *Phys. Biol.*, 3:274–284, 2006.

[60] Steven L. Sanders, Krassimira A. Garbett, and P. Anthony Weil. Molecular characterization of *Saccharomyces cerevisiae* TFIID. *Mol. Cell. Biol.*, 22:6000–6013, 2002.

[61] D. C. Savage. Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.*, 31:107–133, 1977.

[62] K. Smetana, W. J. Steele, and H. Busch. A nuclear ribonucleoprotein network. *Exp. Cell Res.*, 31:198–201, 1963.

[63] M. Takemura. Poxviruses and the origin of the eukaryotic nucleus. *J. Mol. Evol.*, 52:419–425, 2001.

[64] Carl S. Thummel. Mechanisms of transcriptional timing in *Drosophila*. *Science*, 255:39–40, 1992.

[65] Helene Tourriere, Karim Chebli, and Jamal Tazi. mRNA degradation machines in eukaryotic cells. *Biochimie*, 84:821 – 837, 2002.

[66] Kimiko M. Tsutsui, Kuniaki Sano, and Ken Tsutsui. Dynamic view of the nuclear matrix. *Acta. Med. Okayama*, 59:113–120, 2005.

[67] Diana Y. Vargas, Arjun Raj, Salvatore A. E. Marras, Fred Russel Kramer, and Sanjay Tyagi. Mechanism of mRNA transport in the nucleus. *Proc. Natl. Acad. Sci. U.S.A*, pages 17008–17013, 2005.

[68] Donald Voet, Judith G. Voet, and Charlotte W. Pratt. *Fundamentals of Biochemistry*. John Wiley and Sons, Inc, New York, 1 edition, 1999.

[69] Tobias von der Haar. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Systems Biology*, 2, 2008.

[70] Andreas Wagner. Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution*, 22:1365–1374, 2005.

[71] James Watson and Francis Crick. A structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738, 1953.

[72] Mark Winey, Defne Yarar, Thomas H. Giddings Jr., and David N. Mastronarde. Nuclear pore complex number and distribution throughout the *Saccharomyces cerevisiae* cell cycle by three-dimensional reconstruction from electron micrographs of nuclear envelopes. *Mol. Biol. Cell*, 8:2119–2132, 1997.

[73] Woj M. Wojtowicz, John J. Flanagan, S. Sean Millard, S. Lawrence Zipursky, and James C. Clenmens. Alternative splicing of *Drosophila* Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, 118:619–633, 2004.

[74] Hugo Würtele and Pierre Chartrand. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Res.*, 14:477–495, 2006.

[75] Ji Yu, Jie Xiao, Xiaojia Ren, Kaiqin Lao, and X. Sunney Xie. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311:1600–1603, 2006.

[76] D. Zenklusen, D. R. Larson, and R. H. Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.*, 15:1263–1270, 2008.