ANALYTIC SOLUTIONS FOR STOCHASTIC MODELS OF TRANSCRIPTION

SEYED HOSSEIN HOSSEINI
Master of Science, Institute for Advanced Studies in Basic Science, Zanjan, Iran, 2011

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

MASTER OF SCIENCE

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

ANALYTIC SOLUTIONS FOR STOCHASTIC MODELS OF TRANSCRIPTION

SEYED HOSSEIN HOSSEINI

Date of Defense: December 19, 2016

Dr. Marc R. Roussel Supervisor	Professor	Ph.D.
Dr. H. J. Wieden Committee Member	Professor	Ph.D.
Dr. Mark Walton Committee Member	Professor	Ph.D.
Dr. René Boeré Chair, Thesis Examination Committee	Professor	Ph.D.

Dedication

To my lovely niece,

Esma

Abstract

The spatio-temporal organization of one RNA polymerase (RNAP) along a DNA strand is explored by studying a kinetic stochastic model for transcription process, the first step in gene expression. An explicit expression for the probability density of one RNAP was found and compared with stochastic simulation results from the corresponding detailed stochastic model. The explicit solution predicts that the movement of RNAP in large genes (genes of a few hundred nucleotides or more) is advective. It provides a justification for the use of delays in gene expression modeling, especially in delay-stochastic models. The kinetic model for the elongation stage of transcription was extended to two bodies, and the related Fokker-Planck equation was developed. This equation describes the evolution of the joint probability density for two RNAPs along the DNA track.

Acknowledgments

I would like to thank Prof. Marc R. Roussel, my supervisor, for the past two and half years. I really appreciate his guidance, patience and great support. Thanks for all encouragements and insightful discussions. The valuable guidance of my committee members, Prof. Hans-Joachim Wieden and Prof. Mark Walton has influenced the course of my project and is greatly appreciated. A huge thank you to my parents and my brothers as well, whose love and reassurance kept me going through the years. Thanks to Prof. Roussel for providing a stochastic simulation code for transcription by a single polymerase. Finally, I would like to acknowledge the financial support provided by the University of Lethbridge.

Contents

List of Figures

List of Abbreviations

ASEP ..... Asymmetric Simple Exclusion Process
ATP ..... Adenosine Triphosphate
CF ..... Cleavage Factor
CME ..... Chemical Master Equations
CPF ..... Cleavage and Poly-Adenylation Factor
CTD ..... C-Terminal Domain
EC ..... Elongation Complex
FPE ..... Fokker-Planck Equation
GTF ..... General Transcription Factor
NTP ..... Nucleoside Triphosphate
ODE ..... Ordinary Differential Equation
PDE ..... Partial Differential Equation
PIC ..... Pre-Initiation Complex
PPi ..... Pyrophosphate
RNAP ..... RNA polymerase
RNAPII ..... RNA polymerase II
TBP ..... TATA-Binding Protein
TC ..... Termination Complex
TFIIB ..... Transcription Factor II B
TFIIE ..... Transcription Factor II E
TFIID ..... Transcription Factor II D
TFIIH ..... Transcription Factor II H
TSS ..... Transcription Start Site

Chapter 1

Introduction

Two critical cellular processes are the transcription of the DNA to RNA and the translation of messenger RNA (mRNA) to protein. One can think of each of these processes as a movement of a complex machine on a template strand. In the case of transcription, an RNA polymerase moves along the DNA strand to produce the RNA whereas in translation, a ribosome moves along the RNA to senthesize a protein. Both of these processes can be considered as bio-polymerization processes in which new macromolecules are polymerized [17]. Such processes can be divided into three phases: initiation, elongation and termination.

Transcription is the focus of this thesis. Analytical solutions to a simplified kinetic model of transcription of protein-coding genes in eukaryotic cells [95] are obtained. To this purpose, we use stochastic kinetic methods (the chemical master equation and stochastic simulations) since transcription is considered as a stochastic process.

This chapter starts with a brief review of the transcription process. Then we talk about the stochastic methods used in this work. A brief review of kinetic modeling of the transcription process follows. Finally the objectives of this thesis are laid out.

1.1 Transcription

Transcription is a complex process during which the RNA biosynthesis occurs. Although transcription is very similar in both prokaryotic and eukaryotic cells, there are significant differences in proteins and factors involved in the transcription process of both types of cells. The enzyme responsible for transcription processes is *RNA polymerase*. Even though there is a single type of RNA polymerase in prokaryotes (responsible for synthesizing all types of RNAs), there is a set of three basic nuclear RNA polymerases shared by all eukaryotes [74]: *RNA polymerase I*, *RNA polymerase II*, and *RNA polymerase III* [94]. Each of these three polymerases is in charge of transcribing specific types of RNAs. While RNA polymerase II is utilized to synthesize the protein-coding RNAs, RNA polymerase I transcribes the 45S pre-rRNA, from which the 18S, 5.8S and 28S rRNAs are obtained. RNA polymerase III is in charge of synthesizing a variety of small RNAs including tRNAs and the 5S rRNA (reviewed by Arimbasseri and Maraia [3]).

In eukaryotic cells, transcription is only one of the steps to produce a mature messenger RNA (mRNA), and other co-transcriptional and post-transcriptional processes are also crucial. These processes include the modification of both ends of the RNA, cleaving the introns from the RNA, and export of the mRNA out of the nucleus into the cytoplasm for translation to proteins.

While the length of genes in prokaryotic cells is comparatively similar and typically around 1000 base pairs (1 kb), there is huge variation in gene size of eukaryotic cells because of the large number of introns in eukaryotic genes [87]. The smallest human protein-coding genes are a few hundred base pairs long and the largest one belongs to dystrophin with 2.4 Mb [88]. The average length of human genes is 53.6 kb [87].

In this thesis, since I consider the kinetics of the transcription of protein-coding RNAs in

eukaryotic cells, I just discuss the case of RNA polymerase II transcription. Three stages of transcription (initiation, elongation and termination) are discussed briefly in the following paragraphs.

Initiation

One of the essential steps at which transcription is regulated is the binding of the RNA polymerase (RNAP) to a DNA strand and formation of a pre-initiation complex (PIC). The PIC is comprised of multiple proteins (general transcription factors) that recruit an RNAP to the promoter region of DNA and catalyses melting of the DNA around the transcription start site (TSS), then starts transcription by formation of a small number of phosphodiester bonds of RNA. The minimal number of general transcription factors for formation of the PIC includes: TFIIB, TFIID (which includes TATA-binding protein (TBP)), TFIIE, TFIIF, TFIIH, and the RNAPII itself [57]. Besides, the effects of other complexes such as mediators, activators and nucleosome remodeling and modifying complexes are also very crucial (reviewed by Roeder [75] and Li et al. [46]) though they won't be included in the generic model studied here.

The canonical transcription initiation process involves assembly of the PIC at a TATA element [82]. First, the TBP (a subunit of TFIID) binds to the TATA region, bending the DNA [35, 36]. Next, the general transcription factors (GTFs), each of which has a specific function, and RNAPII bind step-wise to the promoter region. The most complicated GTF is TFIIH which is responsible for hydrolyzing ATP in order to provide the energy required to unwind the DNA (by a DNA helicase, one of its subunits) and enabling RNAPII to have access to the template strand around the TSS. In order to disengage the GTFs from the RNAPII, another subunit of TFIIH (a protein kinase) phosphorylates Ser5 from the C-terminal domain (CTD) of RNAPII [1].

Elongation

Elongation has been divided into two stages: early elongation and processive elongation [85]. After binding of RNAP to the promoter region and forming a closed promoter complex, a sequence of conformational changes occurs on DNA that unzip the two strands of the DNA to form an open promoter complex. Afterwards, RNAPII starts transcription of the DNA template producing a complementary RNA. (Here, I consider this stage as early elongation though there is a difficulty distinguishing the boundary between initiation and early elongation [85].) During this period, RNA polymerase enters into abortive synthesis and release of short RNA transcripts (shorter than $9-11$ nucleotides). This process is known as abortive initiation [59]. By formation of a stable RNA-DNA hybrid following transcription of an RNA length longer that 7 nucleotides, the elongation complex (EC) becomes capable of escaping from the promoter [55, 60, 70]. Pal and Luse also showed that after formation of a transcript length of 23 nucleotides, the EC becomes more stable [59]. It is worth mentioning that in prokaryotic transcription, 75 percent of the initiation events end up with the synthesis of short transcripts through abortive initiation [52]. Also, an *in vitro* study suggests that more than half of the initiation events in eukaryotes lead to abortive initiation [48].

When the stable EC reaches approximately 50 nucleotides downstream of the DNA, it pauses in a process known as "promoter-proximal pausing" [53, 68]. This process is one of the regulatory events during transcription and is controlled by elongation factors such as DSIF [98] and NELF [103]. From this point forward, the EC enters the stage of processive elongation in which the gene is completely transcribed.

Termination

For termination of protein-coding genes in eukaryotes, which all have a poly(A) tail at the end of the gene, two different scenarios or models have been proposed for the disso-

ciation of RNAPII from the gene [76]: allosteric and torpedo. The poly(A) signal, which has a significant role in the termination of transcription, is responsible for processing and formation of the 3′-end of the nascent RNA [67]. Two protein complexes, cleavage and poly-adenylation factor (CPF) and cleavage factor (CF) interact with the poly(A) signal and are in charge of cleavage of the RNA and subsequent poly-adenylation of the 3′-end [23, 34].

After 3′-end formation, the RNA is exported to the cytoplasm ready for translation [31]. However, the RNAPII is still on the gene transcribing the DNA template after the poly(A) region and the polymerase should be dissociated from the gene [28]. According to the allosteric model, dissociation of the RNAPII from the DNA is because of the loss of factors stimulating the processive elongation of the EC [47]. This leads to dissociation of RNAPII in two steps, where first the EC pauses and then the RNAPII dissociates from the DNA [61]. According to the torpedo model [72], upon the cleavage of the RNA and subsequent polyadenylation, the EC still transcribes the gene and produces a transcript. Then an exonuclease/helicase (Rat1 in yeast [37] and Xrn2 in human [100]) binds to the 5′-end of this transcript and moves along the transcript and catches the RNAPII. This process destabilizes the RNAPII leading to removal of RNAPII from the DNA [12].

RNA polymerase pausing

Single-molecule optical trapping experiments show that the movement of RNA polymerase along the DNA strand is not continuous and is punctuated many times by pauses [58]. There are two main classes of pauses: (1) backtrack pausing, in which RNAPII slides backward reversibly along both the DNA and the RNA, (2) non-backtrack pausing, in which conformational changes in the RNA polymerase active site stop the nucleotide addition cycle [44, 58]. Backtracking pauses can be affected by the presence of a trailing RNA polymerase that can restrict how far an RNA polymerase can backtrack [22], and also

it can be stopped when there is a dense traffic of polymerases on the gene, whereas non-backtracking pauses can have a dominating influence on the transcription rate [39]. At high transcription initiation rates where there is a large number of RNA polymerases on the same gene, paused polymerases act as an obstacle for movement of the other RNA polymerases on the gene causing a traffic jam [40].

Traffic of RNA Polymerases

The movement of RNAPII along a DNA strand affects and/or is affected by other biochemical processes on DNA. A well-studied one is the interaction of replication and transcription events [65, 66]. Another one is the interaction between DNA repair machinery and transcription [81]. The case of interest in this thesis is the interaction of RNAPIIs with each other, where they have started initiation of transcription from the same promoter.

The speed of the RNAPII elongation together with possible pausing and arrest events can lead to collisions of the RNAPIIs with each other. In the case of T7 bacteriophage RNA polymerase, the leading RNA polymerase is removed by the action of the trailing polymerase [105]. However, in the case of bacterial RNA polymerases and RNAPII, the collision can have a positive effect on progression of the RNAPs [21, 80].

An interesting case is the simultaneous transcription of ribosomal RNA genes where multiple RNA polymerases are transcribing the same gene resulting in the production of a great number of rRNAs. This phenomenon was visualized by Miller and Beatty using the electron micrograph of a segment of nucleolar core from *Triturus viridescens* [30, 56] (figure 1.1).

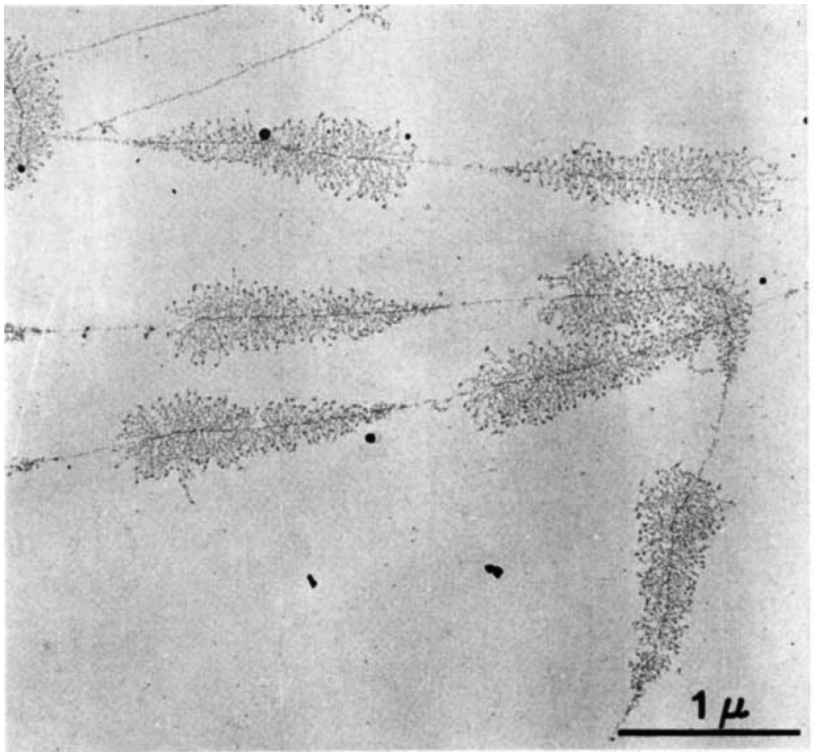

Figure 1.1: Electron micrograph image of a segment of unwound nucleolar core from *Triturus viridescens* (reproduced from [30]).

1.2 Stochastic modeling

To describe the kinetics of different steps in gene expression, stochastic methods are necessary because the classical continuous mass-action kinetics is not valid when there is a small number of molecules involved in gene expression.

In order to model a stochastic process, we need some mathematical tools. The evolution of a stochastic process can be described by the chemical master equations (CME). Below, I will describe this tool and give an example of how it works in kinetic modeling of gene expression. Then I describe the Fokker-Planck equation which is a partial differential equation for describing the evolution of a stochastic system in space and time. This equation is of interest here since I derive it from the chemical master equations and solve it for the transcription elongation process later in this thesis.

1.2.1 Compartmental modeling

Compartmental modeling is one of important tools used for analysing biological systems [2]. It is based on separating the system into a finite number of components, called compartments. The compartments cannot be described independently of each other. The material (here the probabilities) can either flow from one compartment to another, be added into a compartment from a source, or removed from a compartment (a sink). We divide the transcription process into three "compartments" corresponding to the three phases of transcription, and develop a mathematical representation of this phenomenon using the chemical master equation (CME) approach (see next subsection).

1.2.2 Chemical master equation

When a low copy number of molecules is involved in a chemical system, we cannot use the classical kinetic equations. Instead, since the fluctuations in the sequence and timing of chemical reactions and the number of final products are significant, one has to consider the corresponding master equation. The chemical master equation (CME) describes the time evolution of the probability distribution for the state of the chemical system [93].

Assume that $P(\mathbf{x};t)$ is the probability that the chemical system is in state $\mathbf{x}$, where $\mathbf{x}$ represents a vector of states $\mathbf{x} = x_1, x_2, ..., x_N$, and x_i represents the number of molecules of type i. Then we have

$$\frac{dP(\mathbf{x};t)}{dt} = \sum_{\mathbf{x}' \neq \mathbf{x}} \left[W_{\mathbf{x}\mathbf{x}'} P(\mathbf{x}';t) - W_{\mathbf{x}'\mathbf{x}} P(\mathbf{x};t) \right] \tag{1.1}$$

where $W_{\mathbf{x}\mathbf{x}'}$ is the rate of transition probability from state $\mathbf{x}'$ to state $\mathbf{x}$.

The CME is a gain-loss equation for the probability of being in each state of the system. The first term on the right side is the gain of the state $\mathbf{x}$ because of the transition from state

$\mathbf{x}'$ to $\mathbf{x}$ and the second term is the loss of the state $\mathbf{x}$ because of transition from $\mathbf{x}$ to $\mathbf{x}'$ [93].

Here, I am going to show how to develop the CMEs and outline the mathematical steps to derive a solution for a stochastic kinetic problem. I consider a part of a signaling cascade where a ligand X binds an inactive enzyme, causing a conformational change that activates the catalytic activity of the enzyme (This was one of assignments in the Stochastic Processes in Biochemistry course taught by Marc R. Roussel in Spring 2015 at the University of Lethbridge). The number of product molecules (P) triggers the next step in the cascade when it exceeds a specific level. The reactions in this mechanism are as follows:

$$\mathrm{E_i} + \mathrm{X} \underset{k_{-\mathrm{a}}}{\overset{k_{\mathrm{a}}}{\rightleftharpoons}} \mathrm{E} \tag{1.2}$$

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{C} \xrightarrow{k_{-2}} \mathrm{E} + \mathrm{P} \tag{1.3}$$

where k_{a}, $k_{-\mathrm{a}}$, k_1, k_{-1} and k_{-2} are the rate constants for individual steps in each reaction. Let's assume that while X is synthesized, it is also subject to degradation with a rate constant of k_{d}:

$$\xrightarrow{k_{\mathrm{x}}} \mathrm{X} \xrightarrow{k_{\mathrm{d}}} \tag{1.4}$$

where k_x is the rate of production of X.

The goal is to obtain the mean first passage time for p (the number of molecules of P) reaching a particular level p_m.

As a first step towards solving this problem, we derive the chemical master equation (CME) for these model reactions. To write down a solvable CME, we need to make some assumptions—the general CME is unsolvable without making assumptions. We assume that the amounts of X (X_0) and S (S_0) are much greater than the total amount of enzyme

(E_0). We suppose that the production and degradation of X are fast (faster than the other reactions) and also the rate constants (k_x and k_d) are such that X accumulates to a high level, much higher than E_0. (It should be indicated that this example is here for illustration purposes, and that the assumptions would have to be verified for any particular system.) In such conditions, the level of X molecules will reach a steady-state. Thus, we have the following mass conservation equations: $s+p+c=S_0$, $e_i+e+c=E_0$ and $x+e+c=X_0$, where s, p, c, e_i, e and x denote integer values that the random variables $S(t)$, $P(t)$, $C(t)$, $E_i(t)$, $E(t)$ and $X(t)$ (representing the number of molecules of each reactant at time t) can take respectively.

The conservation equations enable us to describe the state of the system by just three populations. Here, I take e, p and c. From the assumptions and the conservation equations, we can assume that $e_i = E_0 - e - c$, $s \approx S_0$ and $x \approx X_0$. Now, we can write the chemical master equation for this process:

$$\begin{aligned}\frac{dP_{e,p,c}}{dt} = {} & \kappa_a(E_0-e-c+1)(X_0)P_{e-1,p,c} + \kappa_{-a}(e+1)P_{e+1,p,c} + \kappa_1(e+1)(S_0+1)P_{e+1,p,c-1} \\ & + \kappa_{-1}(c+1)P_{e-1,p,c+1} + \kappa_{-2}(c+1)P_{e-1,p-1,c+1} \\ & - [\kappa_a(E_0-e-c)(X_0) + \kappa_{-a}(e) + \kappa_1(e)(S_0) + \kappa_{-1}(c) + \kappa_{-2}(c)]P_{e,p,c}\end{aligned} \tag{1.5}$$

The next step is introducing the joint generating function [93].

$$F(u,v,w,t) = \sum_{e=0}^{E_0}\sum_{p=0}^{S_0}\sum_{c=0}^{E_0} u^e v^p w^c P_{e,p,c} \tag{1.6}$$

Multiplying (1.5) by $u^e v^p w^c$ and summing over possible values of e, p and c will give us a partial differential equation (PDE) for the generating function $F(u,v,w,t)$.

$$\frac{\partial F(u,v,w,t)}{\partial t} = f(u,v,w,\frac{dF}{du},\frac{dF}{dw},\frac{dF}{du}) \tag{1.7}$$

Now, we can obtain the distribution function for the number of each component of the reaction. We just need to put $u=1$ and $w=1$ in the joint generating function. This way,

the generating function will be just a function of v and time ($F(v,t)$). By expanding $F(v,t)$ in v, and using (1.6), we can obtain the distribution for the number of P molecules ($P(p,t)$). We can also get the mean and the variance easily from the generating function.

The next step is deriving the 'propagator' probability $P_{p|p'}(t)$, which is the probability of having p number of P molecules at time t given p' molecules initially. We can obtain that from the generating function since we have this relationship between $F(v,t)$ and $P_{p|p'}(t)$ as $F(v,t) = \sum_p P_{p|p'}(t)v^p$.

Using $P_{p|p'}(t)$ and $P(p,t)$, we can obtain the distribution of the first passage time for p reaching a particular level p_m (let us call it $f_{p_m}(t)$). We then have [93]

$$P(p_m,t) = \int_0^t dt' f_{p_m}(t') P_{p_m|p_m}(t-t') \tag{1.8}$$

This integral is a Volterra integral equation of the first kind and can be solved numerically. This example was meant to illustrate a standard technique for solving the CME that isn't used in this work. There is another approach to these problems presented in the following subsection that is directly useful to this work.

1.2.3 Fokker-Planck equation

The Fokker-Planck equation is a particular partial differential equation (PDE) which offers a great tool to describe stochastic systems involving fluctuations and noise. It was first introduced by Fokker [24] and Planck [63] to describe the Brownian motion of a small particle immersed in a fluid. The general form of this equation for a one-variable system is [73]:

$$\frac{\partial W(x,t)}{\partial t} = \left[-\frac{\partial}{\partial x} D^{(1)}(x) + \frac{\partial^2}{\partial x^2} D^{(2)}(x) \right] W(x,t) \tag{1.9}$$

and for N variables:

$$\frac{\partial W(\mathbf{x},t)}{\partial t} = \left[-\sum_{i=1}^{N} \frac{\partial}{\partial x} D_i^{(1)}(\mathbf{x}) + \sum_{i,j=1}^{N} \frac{\partial^2}{\partial x_i \partial x_j} D_{ij}^{(2)}(\mathbf{x}) \right] W(\mathbf{x},t) \tag{1.10}$$

where W is the distribution function, $\mathbf{x} = x_1, x_2, ..., x_N$, $D^{(1)}$ and $D_i^{(1)}$ are the drift coefficients (a vector in the later case) and $D^{(2)}$ and $D_{ij}^{(2)}$ are the diffusion coefficients (a tensor).

In fact, this equation is a diffusion equation with an extra first-order derivative (drift term). By solving the Fokker-Planck equation, one can obtain the distribution function of the stochastic system from which you can obtain the average properties for macroscopic variables. Obtaining an exact solution for this equation has always been a big challenge, specially for systems with finite boundary conditions and it is necessary to use some approximation methods.

It is also worth mentioning that when the fluctuations in a system are negligible, the diffusion term can be insignificant (so removable), and the PDE becomes approximately an advection equation.

1.3 Modeling of transcription

In this section, I will discuss recent developments in the mathematical modeling of gene expression, especially the transcription part. The modeling of gene expression can be classified into two categories. In the first category, the models are more coarse-grained and deal with fewer components to describe the whole system of gene expression. For instance, the degradation and production of both RNAs and proteins are usually considered as single-step processes. Here are a few papers in this category: [8, 41, 42, 45, 62, 69, 83, 89, 96, 106]. In the second category, the models are describing the details of both cases of transcription and translation. For example, the transcription process will be considered as a pro-

cess including thousands of chemical reactions. Here are some papers in this category: [5, 13, 33, 71, 77, 78, 92, 97, 101].

Gene expression is controlled by the concentrations, states and the locations of molecules such as DNA, RNA polymerases and the regulatory molecules. Fluctuations in the amount or activity of these macromolecules lead to corresponding fluctuations in the production of a gene product [20, 54]. In chemical systems, the fluctuations scale as $\sqrt{N}$, where N is the number of molecules in gene expression. As the system grows, the fluctuations become relatively less important since $\frac{\sqrt{N}}{N} = \frac{1}{\sqrt{N}}$ decreases with growing N [64]. So, the stochastic nature of chemical reactions is very important at small copy numbers. Therefore, since there is a low copy number of molecules in gene expression, stochasticity plays a significant role in this system.

1.3.1 Modeling of gene expression

Here, I consider an example of the first kind of gene expression model mentioned in the first paragraph of this section. Suppose we have a promoter that can be in one of two states, active or inactive, depending on the binding of an effector. This system can be considered in a three-stage model of gene expression, including activation of the gene, transcription and translation (see Fig. 1.2 A). Shahrezaei and Swain [83] obtained an expression for the stationary protein distribution by solving the master equation developed for this model. It should be mentioned that traditionally, people focus on stationary processes because they are typically easier to solve than non-stationary processes. Here, I discuss some of their results.

If $P^{(0)}_{n,m}$ and $P^{(1)}_{n,m}$ are the probabilities of having n proteins and m RNAs in inactive and active states respectively at time t, the master equation of this process can be written as:

$$\begin{aligned}\frac{\partial P^{(0)}_{m,n}}{\partial t} =& \kappa_1 P^{(1)}_{m,n} - \kappa_0 P^{(0)}_{m,n} + (n+1)P^{(0)}_{m,n+1} - nP^{(0)}_{m,n} \\ &+\gamma\Big[(m+1)P^{(0)}_{m+1,n} - mP^{(0)}_{m,n} + bm(P^{(0)}_{m,n-1} - P^{(0)}_{m,n})\Big]\end{aligned} \tag{1.11}$$

$$\begin{aligned}\frac{\partial P^{(1)}_{m,n}}{\partial t} =& -\kappa_1 P^{(1)}_{m,n} + \kappa_0 P^{(0)}_{m,n} + (n+1)P^{(1)}_{m,n+1} - nP^{(1)}_{m,n} \\ &+a(P^{(1)}_{m-1,n} - P^{(1)}_{m,n}) \\ &+\gamma\Big[(m+1)P^{(1)}_{m+1,n} - mP^{(1)}_{m,n} + bm(P^{(1)}_{m,n-1} - P^{(1)}_{m,n})\Big]\end{aligned} \tag{1.12}$$

where $a = \frac{\nu_0}{d_1}$, in which ν_0 is the probability of transcription per unit time and d_1 is the probability of degradation of a protein per unit time, $b = \frac{\nu_1}{d_0}$, in which ν_1 is the probability of translation per unit time and d_0 is the probability of degradation of an RNA per unit time, $\gamma = \frac{d_1}{d_2}$ represents the ratio of protein lifetime to the RNA lifetime, $\tau = d_1 t$ is the dimensionless time, $\kappa_0 = \frac{k_0}{d_1}$ is and $\kappa_1 = \frac{k_1}{d_1}$ (figure 1.2 A).

The path for dealing with this set of coupled ODEs is the same as for the signaling cascade problem described on page 8. They used the method of generating functions introduced in section 1.2.1 and obtained the probability distributions for the number of proteins. Since mRNA lifetimes are much smaller than protein lifetimes ($\gamma \ll 1$), the stationary solution for the protein distribution is:

$$\begin{aligned}P_n =& \frac{\Gamma(\alpha+n)\Gamma(\beta+n)\Gamma(\kappa_0+\kappa_1)}{\Gamma(n+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\kappa_0+\kappa_1+n)} \times \left(\frac{b}{1+b}\right)^n \left(1-\frac{b}{1+b}\right)^\alpha \\ &\times {}_2F_1\left(\alpha+n, \kappa_0+\kappa_1-\beta, \kappa_0+\kappa_1+n; \frac{b}{1+b}\right)\end{aligned} \tag{1.13}$$

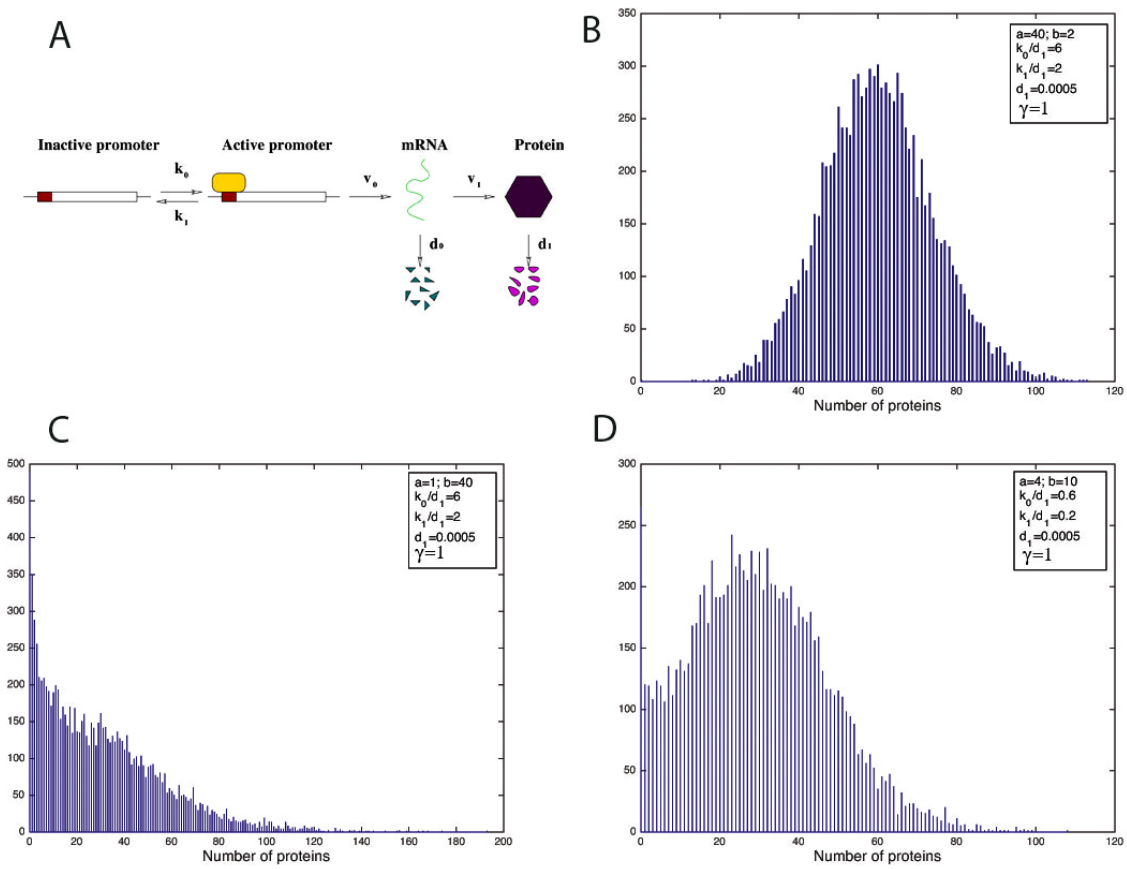

Figure 1.2: A) Schematic model of the three-stage gene expression model [83]. B,C,D) distribution of proteins for different values of the parameters.

where ${}_2F_1$ is a hypergeometric function and

$$\alpha = \frac{1}{2}(a + \kappa_0 + \kappa_1 + \phi)$$

$$\beta = \frac{1}{2}(a + \kappa_0 + \kappa_1 - \phi)$$

This analytical expression properly predicts the probability distribution of proteins for the three-stage model. Figure 1.2 just shows my own stochastic simulation results of the protein distribution for three different sets of parameter values. This figure doesn't show the anlaytical results (refer to the paper [83]).

One can obtain statistical properties such as the mean and the variance of the distribution among other things from this model. Figures 1.2 C and D show a bi-modality (though less than obvious in my own simulation results) which is because of a slow transition between inactive and active states of the promoter.

Here, an example of the coarse-grained type of gene expression models was considered to illustrate how you can use the CME approach to deal with this type of modeling.

1.3.2 Modeling a single RNA polymerase elongation

There have been a couple of kinetic models regarding the transcription process. Since transcription has been divided into three stages of initiation, elongation, and termination, most models have concentrated mainly on one of these stages. Different approaches have been developed regarding each one of the stages. Here I briefly review work on transcription elongation.

Jülicher and Bruinsma [33] developed a detailed model to describe the polymerization kinetics during the elongation stage. They considered the effect of internal deformation (strain) of RNA polymerase and also sequence sensitivity for the motion of RNA polymerase. Wang et al. [99] also proposed a model that takes into account a *multi-step* kinetics for translocation of RNA polymerase. Essentially, most of the models have been proposed to simulate the optical trapping experiments of transcription elongation which has been extensively explored [5, 6, 99].

von Hippel's group is one of the pioneering groups in modeling the transcription kinetics. They started by a thermodynamic analysis of transcription elongation and considered the stability of the elongation complex nucleotide-by-nucleotide along a DNA strand [102]. They found that the elongation complex is stable except at the end of the gene where because of the hairpin structure and effects of regulatory factors it becomes unstable. This

study of sequence dependence was extended by the Wang group, incorporating the energy landscape for translocation of RNA polymerase and providing a full kinetic analysis [6, 7]. Their modeling approach was enough to predict and explain the observations of the single-molecule experiments.

Roussel and Zhu also proposed a single-gene transcription model to explore the stochastic kinetics of this process [78]. They obtained analytical and numerical results for the distribution of the transcriptional delay and also the distribution of elongation rate. They also studied the case where there are many RNAPs transcribing the same gene and obtained stochastic simulation results [77, 78]. Voliotis et al. [97] studied the effect of transcriptional pauses on the distribution of transcriptional time and suggested that these pauses play a significant role in the variability of transcription rates.

1.3.3 Modeling of the interaction of RNA polymerases

The models discussed by now have considered just the properties of a single RNA polymerase and are mostly for prokaryotic cells. Here, I discuss the effect of the interaction of RNA polymerases initiating from the same promoter on the transcription kinetics. Since the transcription process is a non-stationary system, it needs more sophisticated methods to obtain analytical results than the foregoing examples.

Transcription interference refers to the interaction of RNAPs initiating from two different promoters (reviewed by Shearwin et al. [84]). Even though it is not treated in this thesis, it is worthy of a brief discussion. Essentially, due to the direction of the two RNAPs (initiated from two convergent, divergent or overlapping promoters) the transcription of an RNAP from a promoter can have a suppressive (and therefore regulatory) effect on the transcription of another promoter. There are three possible mechanisms: occlusion (where

the initiation of the RNAP is occluded by the passing RNAPs), collision (of RNAPs between two promoters) and ”sitting-duck” collisions (a mechanism where a passing RNAP detaches the RNAP already initiated and ready to elongate) [11]. Transcription interference in prokaryotic cells was explored by Sneppen et al. [86] where they used three different approaches (analytical, stochastic simulation and mean-field treatment) to determine the predominant mechanisms in different situations.

For the case of RNAPs initiated from the same promoter like ribosomal RNA genes, the interaction of RNAPs in the elongation phase can also have significant effects on the transcription kinetics. The stochasticity of this process and the sequence-dependent translocation of RNAPs plus other possible reasons such as roadblocks, pauses and so on can lead to the traffic of RNAPs along the gene, which has a resemblance to vehicle and pedestrian traffic [9, 15].

The active transport of multiple RNAPs along a gene can be considered as an asymmetric simple exclusion process (ASEP) in which each one of the RNAPs translocates nucleotide by nucleotide based on an exclusion rule (because of the steric repulsion effect of the RNAPs). The first model of this kind was introduced by McDonald and Gibbs [49, 50] for the translation process. However, their model was very simplified and they assumed a ribosome to be as a solid rod occupying a single site. Later on, the model was extended [18, 43] and the ribosomes were considered as multi-site solid rods. The exact analytical solution of the ASEP-type models is unachievable and only the steady-state solution of this model was obtained by some analytical methods [14]. Chowdhury’s group explored the RNAP traffic in this framework and obtained some statistical properties such as the average rate of transcription and the level of the transcription noise and its dependence on the model parameters [92].

The traffic of RNAPs is a rate-limiting process for highly transcribed genes. Klumpp and Hwa [40] considered the effect of pausing on ribosomal RNA transcription and they found that RNAP traffic and pauses can cause an intense traffic jam. However, they showed that to keep a high rate of transcription, pauses are suppressed by the antitermination complexes, and stalled polymerases are removed by the termination factor rho.

1.4 Objectives

This thesis focuses on obtaining non-stationary analytical solutions for an existing stochastic model of transcription [95] in eukaryotic cells. In chapter two, I first present a simplified version of Vashishtha's model [95] and describe the chemical reactions occurring in each compartment of the transcription process. Then I utilize this model to write the corresponding chemical master equations for each compartment. For both initiation and termination phases, since there are few ordinary differential equations (ODEs), the probability functions for being in each state of the model can be easily obtained. For the elongation compartment, we again write down the chemical master equation. However, as there is a large number of ODEs in this compartment, I derive a continuum limit of these ODEs and find the corresponding Fokker-Planck equation. In genes of a few hundred nucleotides or more, where the diffusion term in the Fokker-Planck equation will be shown to be negligible, we solve the equation and obtain an explicit expression for the probability distribution of an RNA polymerase as a function of time and space. The mean velocity and the diffusion coefficient of the elongation compartment are also obtained in section 2.2. In addition, in this chapter, we consider the effect of two types of pauses in the elongation compartment: ubiquitous pauses and a strong pause. For the first one, we obtain a Fokker-Planck equation which gives us useful information regarding the mean velocity and diffusion coefficient of the system (see equations 2.26 and 2.27). For the second one, we find the effect of a strong pause on the probability density function in the elongation compartment. In chapter 3, I consider how the interaction of two tandem head-to-tail RNA polymerases can affect the

kinetics of each of the RNAPs at each compartment. We obtain the probability distributions for the initiation and termination time of each RNAP. For the elongation compartment, we obtain a Fokker-Planck equation describing the two elongating RNAPs.

Chapter 2

Spatiotemporal organization of a single RNA polymerase on a DNA strand

In this chapter, I have developed a kinetic model to explore the spatio-temporal organization of an RNA polymerase along the DNA strand in eukaryotic cells. I obtain the probability density profiles, the mean velocity and other statistical properties related to one RNA polymerase. It is found that the transcription elongation is a purely advective process in the case of large genes (genes of at a few hundred nucleotides or more). However, when the length of a gene is small, the effect of diffusion becomes significant. I have also considered the effect of ubiquitous pauses and also a strong pause along a gene on the kinetics of transcription.

2.1 Eukaryotic transcription model

I start with a simplified version of Vashishtha's kinetic model for the eukaryotic transcription process [95], removing the abortive initiation, early elongation/promoter escape and promoter-proximal pausing from Vashishtha's model. The model is divided into three compartments: initiation, elongation, and termination.

For the initiation compartment, first the initiation factors, including the TATA-binding protein (TBP), bind to the TATA box of the promoter region (pro) of the DNA (reaction 2.1), and then the RNA polymerase (RNAP) locates the promoter site and binds to it forming the pre-initiation complex (PIC) (reaction 2.2).

$$\mathrm{TBP} + \mathrm{pro} \xrightarrow{k_{\mathrm{bind}}} \mathrm{TBP}\cdot\mathrm{pro} \tag{2.1}$$

$$\mathrm{TBP}\cdot\mathrm{pro} + \mathrm{RNAP} \xrightarrow{k_{\mathrm{PIC}}} \mathrm{PIC} \tag{2.2}$$

After forming a stable PIC, the RNA polymerase moves to the first site of the DNA template. Reaction (2.3) represents the translocation of the active site of the RNA polymerase to the first unoccupied nucleotide U_1 and the conversion of this unoccupied nucleotide into the occupied nucleotide O_1,

$$\mathrm{PIC} + \mathrm{U}_1 \xrightarrow{k_{\mathrm{init}}} \mathrm{O}_1 \tag{2.3}$$

After occupying the first site, the RNA polymerase is activated by binding two nucleotide triphosphate molecules, changing the occupied state (O_1) to the activated state (A_1). Then the RNA polymerase is translocated to the adjacent site after formation of a phosphodiester bond between two nucleotides, which provides the required energy for translocation [77].

The translocation of RNA polymerase (RNAP) along the DNA during the elongation part of the transcription process can be described by two types of mechanisms. The first one is the power stroke mechanism, suggesting that RNAP translocation is a result of chemical changes, phosphodiester bond formation and then release of pyrophosphate (PPi). In this mechanism, thermal fluctuations are not considered essential [104]. The second one is the Brownian ratchet mechanism. In this case, thermal fluctuations inherent in the system cause random transitions between pre- and post-translocated enzyme states, and such thermally-driven motions are rectified to the post-translocated state by the incorporation of the incoming NTP (nucleoside triphosphate) [29]. In the kinetic model of Roussel and Zhu [78], for the elongation part, there are two states, occupied (O) and activated (A) states

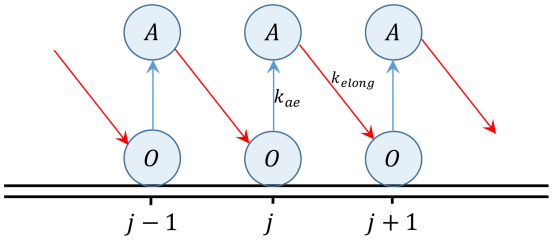

Figure 2.1: Two internal states for each site of the elongation compartment: occupied (O) and activated (A)

(figure 2.1), which can be considered as post-translocated and pre-translocated states respectively.

The process of activation and translocation of RNA polymerase repeats itself as RNAP translocates nucleotide by nucleotide along the gene. This repetition can be considered as the elongation compartment of transcription (reactions (2.4) and (2.5)).

$$\mathrm{O_i} \xrightarrow{k_{\mathrm{ae}}} \mathrm{A_i} \qquad i = 1,2,...,N-1 \tag{2.4}$$

$$\mathrm{A_i} + \mathrm{U_{i+1}} \xrightarrow{k_{\mathrm{elong}}} \mathrm{O_{i+1}} + \mathrm{U_i} \qquad i = 1,2,...,N-1 \tag{2.5}$$

According to allosteric model, transcription of the poly (A) site at the end of the gene can cause a termination-inducing conformational change in the elongation complex [76]. After the activation step at site N (reaction (2.6)), the conversion of the elongation complex into a termination complex (TC) happens (which is represented by reaction (2.7)). Finally, the termination complex dissociates and releases a complete pre-mRNA of the transcribed gene (reaction (2.8)).

$$\mathrm{O}_N \xrightarrow{k_{\mathrm{ae}}} \mathrm{A}_N \tag{2.6}$$

$$\mathrm{A_N} \xrightarrow{k_{\mathrm{TC}}} \mathrm{TC} \tag{2.7}$$

$$\mathrm{TC} \xrightarrow{k_{\mathrm{term}}} \mathrm{RNA} + \mathrm{RNAP} + \mathrm{U_N} \tag{2.8}$$

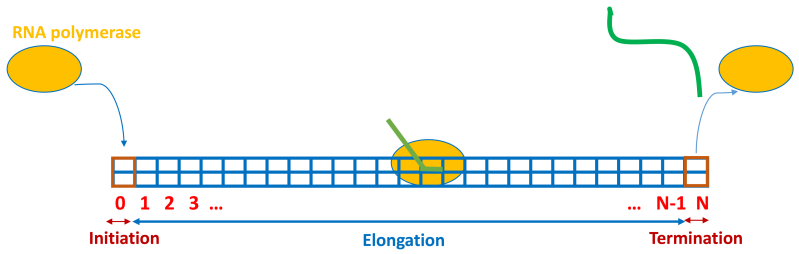

Figure 2.2: A schematic representation of RNA polymerase on the DNA template strand

2.2 Probability distribution for a single RNA Polymerase

2.2.1 Initiation

We assume that all initiation processes are happening at $x = 0$ and all the termination processes are happening at $x = N$ (the nucleotide at the end of the transcribed sequence) (figure 2.2). By solving the chemical master equation (CME) for each state in these compartments, we can get probability distributions.

Chemical reactions involved in this model are first-order processes because we assume that there is just a single RNA polymerase active on the DNA template strand, so for now we neglect the interaction of one RNA polymerase with other ones. However, in the next chapter, where we deal with the two-polymerase theory, we cannot neglect the interaction of RNA polymerases.

The governing master equations are

$$\frac{dP_{\text{pro}}(t)}{dt} = -k_{\text{bind}}P_{\text{pro}}(t) \tag{2.9}$$

$$\frac{dP_{\text{TBP·pro}}(t)}{dt} = k_{\text{bind}}P_{\text{pro}}(t) - k_{\text{PIC}}P_{\text{TBP·pro}}(t) \tag{2.10}$$

$$\frac{dP_{\text{PIC}}(t)}{dt} = k_{\text{PIC}}P_{\text{TBP·pro}}(t) - k_{\text{init}}P_{\text{PIC}}(t) \tag{2.11}$$

$$\frac{\mathrm{dP_O(t)}}{\mathrm{d}t} = k_{\text{init}}P_{\text{PIC}}(t) \tag{2.12}$$

with initial conditions $P_{\text{pro}}(0) = 1$, $P_{\text{TBP·pro}}(0) = 0$, $P_{\text{PIC}}(0) = 0$ and $P_{\text{O}}(0) = 0$.

These linear ordinary differential equations can be solved easily to find the probability distribution of being in each state. I should indicate that P_{pro} is the probability that the promoter has never been bound. The solutions for $P_{\text{PIC}}(t)$, $P_{\text{pro}}(t)$ and $P_{\text{TBP·pro}}(t)$ are

$$P_{pro}(t) = e^{-k_{\text{bind}}t} \tag{2.13}$$

$$P_{TBP\cdot pro}(t) = \frac{k_{\text{bind}}}{k_{\text{PIC}} - k_{\text{bind}}}\left(e^{-k_{\text{bind}}t} - e^{-k_{\text{PIC}}t}\right) \tag{2.14}$$

$$P_{PIC}(t) = \frac{k_{PIC}\,k_{bind}\,\left((k_{PIC} - k_{init})\,\mathrm{e}^{-k_{bind}\,t} + (k_{init} - k_{bind})\,\mathrm{e}^{-k_{PIC}\,t} - (k_{PIC} - k_{bind})\,\mathrm{e}^{-k_{init}\,t}\right)}{(k_{PIC} - k_{bind})\,(k_{bind} - k_{init})\,(k_{init} - k_{PIC})} \tag{2.15}$$

We can obtain the initiation rate ($k_{\text{init}}P_{\text{PIC}}(t)$) from (2.15). We also carried out a Gillespie stochastic simulation [25, 26] of the model (using a program supplied by M. Roussel) in which the number of realizations (or, equivalently, transcriptions) was 10^6. Figure 3.3 shows the analytical and stochastic simulation results for $P_{\text{PIC}}(t)$ where we can see a good agreement between them.

2.2.2 Elongation

The stochastic process of transcription is modelled with the chemical master equation (CME). We write the CME for a single RNA polymerase. The CME is a set of ordinary differential equations (ODEs), one for each state of the model. For any template strand of reasonable length, the number of states is large, so solving the CME analytically is not feasible. The biggest challenge in dealing with the elongation part is having a large number of sites (or nucleotides). If we write the master equation for every state at all sites, we will have a set of about $2N$ (N is number of sites) ODEs which makes it difficult to be solved since N is typically of the order of 70 (for the tRNAs) to 2.3×10^6 (the largest human gene is the dystrophin gene with 2300 kilobases [88]) among human genes.

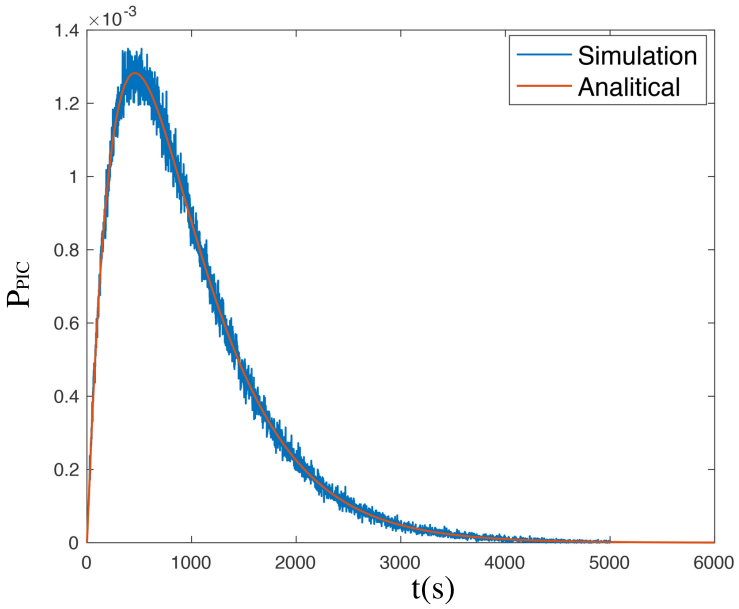

Figure 2.3: A comparison between stochastic simulation and analytic solution for $P_{PIC}(t)$. The values of the parameters: $k_{PIC} = 0.0029s^{-1}$, $k_{bind} = 0.0016s^{-1}$ and $k_{init} = 0.6s^{-1}$ [95, p. 48].

2.2.2.1 Governing equations

The master equations in the elongation compartment are

$$\frac{dP_{i,O}(t)}{dt} = k_{\text{elong}}P_{i-1,\text{A}}(t) - k_{\text{ae}}P_{i,\text{O}}(t) \tag{2.16}$$

$$\frac{dP_{i,A}(t)}{dt} = k_{\text{ae}}P_{i,\text{O}}(t) - k_{\text{elong}}P_{i,\text{A}}(t) \tag{2.17}$$

where $i = 1,2,3,...,N-1$.

The probability of being at site i is the probability of being in state A at site i plus the probability of being in state O at site i, that is

$$P_i(t) = P_{i,\text{O}}(t) + P_{i,\text{A}}(t) \tag{2.18}$$

Using equations 2.16, 2.17 and 2.18, we obtain

$$\frac{dP_i}{dt} = k_{elong}(P_{i-1,\mathrm{A}} - P_{i,\mathrm{A}}) \tag{2.19}$$

In the CMEs, the variable i, representing the position of polymerase along the DNA template strand, is discrete-valued since the polymerase occupies discrete sites (nucleotides). A practical way to obtain results in this part is by making a continuum assumption for the position i. The reason is that, given that most genes are relatively long, one nucleotide represents a small change in position relative to the length of the gene. This idea will be formalized in the scaling later in this section. Hence, first we obtain master equations for states in site i, and then make some approximations to obtain the related partial differential equation (PDE) for $p(x,t)$.

At this point, we need a relationship between $P_{i,\mathrm{A}}$ and P_i since 2.19 is not a closed equation. To address this issue, we can consider the residence time in each state. The mean residence time in the O state is $1/k_{\mathrm{ae}}$, and the mean residence time in the A state is, similarly, $1/k_{\mathrm{elong}}$. The overall mean residence time at a particular site is therefore the sum of these two quantities $t_{\mathrm{residence}} = 1/k_{\mathrm{ae}} + 1/k_{\mathrm{elong}}$, from which we can get the proportion of the time spent in each state at a particular site. Now if you assume that only a short time is spent at each site, i.e. that the residence times are short, then these are essentially instantaneous estimates of the proportion of the probability in each state.

$$\frac{P_{i,\mathrm{A}}(t)}{P_i(t)} = \frac{\text{The residence time in state A at site } i}{\text{The overall residence time at site } i} = \frac{1/k_{\mathrm{elong}}}{1/k_{\mathrm{ae}} + 1/k_{\mathrm{elong}}} = \frac{k_{\mathrm{ae}}}{k_{\mathrm{ae}} + k_{\mathrm{elong}}} \tag{2.20}$$

In addition, we can obtain the same results by using the steady state assumption for equation 2.17, $\frac{dP_{i,A}(t)}{dt} = 0$. So we have $P_{i,O}(t) = \frac{k_{elong}}{k_{ae}} P_{i,A}(t)$. Substituting this into equation 2.18, we have

$$P_i(t) = (1 + \frac{k_{elong}}{k_{ae}}) P_{i,A}(t) \tag{2.21}$$

which is the same as 2.20.

So from equation 2.19, we have

$$\frac{dP_i(t)}{dt} = \frac{k_{elong}k_{ae}}{k_{elong}+k_{ae}}(P_{i-1}(t) - P_i(t)) \tag{2.22}$$

Now, let's define a dimensionless variable, $X = i\Delta X$ in which $\Delta X = \frac{1}{N}$). If we assume that the length of a typical human gene (N) is long enough, we can make a continuum assumption. So, in a continuum limit, when ΔX approaches to zero, using the Taylor series for P_{i-1} up to second order, we have

$$P_{i-1} = P_i - \Delta X\frac{\partial P_i}{\partial X} + \frac{1}{2}\Delta X^2\frac{\partial^2 P_i}{\partial X^2} \tag{2.23}$$

In the continuum limit: $P_i(t) \to p(X,t)$. If we substitute equation 2.23 for P_{i-1} in equation 2.19, we obtain the Fokker-Planck equation for the elongation part.

$$\frac{\partial p(X,t)}{\partial t} = -\frac{1}{N}\frac{k_{elong}k_{ae}}{k_{elong}+k_{ae}}\frac{\partial p(X,t)}{\partial X} + \frac{1}{2N^2}\frac{k_{elong}k_{ae}}{k_{elong}+k_{ae}}\frac{\partial^2 p(X,t)}{\partial X^2} \tag{2.24}$$

The resultant PDE is of the form of a Fokker-Planck equation. This treatment gives us a mean velocity, and also tells us how the distribution spreads out as it advects through the elongation region.

The variable X in equation 2.24 is dimensionless. If we give its dimension back ($X = \frac{x}{N}$), we will have

$$\frac{\partial p(x,t)}{\partial t} = -a\frac{\partial p(x,t)}{\partial x} + \frac{D}{2}\frac{\partial^2 p(x,t)}{\partial x^2} \tag{2.25}$$

where we define

$$a = \frac{k_{elong}k_{ae}}{k_{elong}+k_{ae}}(1\,nuc.) \tag{2.26}$$

and

$$D = \frac{k_{elong} k_{ae}}{k_{elong} + k_{ae}} (1\, nuc.^2) \tag{2.27}$$

as the mean velocity and the diffusion coefficient respectively.

Equation 2.25 can also be written in the form of a continuity equation as follows:

$$\frac{\partial p(x,t)}{\partial t} + \frac{\partial J(x,t)}{\partial x} = 0 \tag{2.28}$$

where we can define probability density flux as

$$J(x,t) = a\, p(x,t) - \frac{D}{2} \frac{\partial p(x,t)}{\partial x} \tag{2.29}$$

Since $p(x,t)$ is a probability density along the x axis, it has units of $[x]^{-1}$. The advection velocity a has units of $[x][t]^{-1}$ and the diffusion coefficient D has the units of $[x][t]^{-2}$. Accordingly, the flux has units of $[t]^{-1}$. The interpretation of this flux is that, at any given x at time t, J units of probability transit through x per unit time. Thinking of equation 2.11 in the same way, we see that the term $k_{init} P_{PIC}$, which has units of $[t]^{-1}$, corresponds to the rate of probability loss from the PIC compartment, i.e. the amount of probability leaving the compartment per unit time. This probability transits through $x_0 = 0$, the left-hand edge of the elongation compartment. Thus, the boundary condition at $x = x_0 = 0$ is

$$J(x,t) = a\, p(x,t) - \frac{D}{2} \frac{\partial p(x,t)}{\partial x} \bigg|_{x=0} = k_{init} P_{PIC}(t) \tag{2.30}$$

At the end of the elongation compartment, $x = N$, probability will exit into the termination compartment and not come back out again. This means that the diffusion term in the flux should be zero. So we have

$$\frac{\partial p(x,t)}{\partial x} \bigg|_{x=N} = 0 \tag{2.31}$$

The PDE is also subject to the following initial condition at $t = t_0 = 0$:

$$p(x,t)\Big|_{t=0} = 0 \tag{2.32}$$

2.2.2.2 Derivation of analytical solution

By defining dimensionless variables, $X = \frac{x}{N}$ and $\tau = \frac{t\,a}{N} = t\,\overline{a}$ where $\overline{a} = \frac{a}{N}$, $q(X,\tau) = N\,p(x,t)$ and inserting them into equations 2.24, 2.30, 2.31 and 2.32, we can obtain the PDE and the initial and boundary conditions in a dimensionless form.

$$\frac{\partial q(X,\tau)}{\partial \tau} = -\frac{\partial q(X,\tau)}{\partial X} + \frac{1}{\mathrm{Pe}}\frac{\partial^2 q(X,\tau)}{\partial X^2}, \quad 0 < X < 1, \quad \tau > 0 \tag{2.33}$$

$$q(X,\tau) = 0 \quad \text{at} \quad \tau = 0 \tag{2.34}$$

$$q(X,\tau) - \frac{1}{\mathrm{Pe}}\frac{\partial q(X,\tau)}{\partial X} = \frac{k_{init}}{\overline{a}} P_{PIC}(\tau/\overline{a}) \quad \text{at} \quad X = 0 \tag{2.35}$$

$$\frac{\partial q(X,\tau)}{\partial X} = 0 \quad \text{at} \quad X = 1 \tag{2.36}$$

where $\mathrm{Pe}(= 2N)$ is the Peclet number, and since N (the number of nucleotides) is a large number, $\varepsilon = \frac{1}{\mathrm{Pe}}$ will be a very small number. We can take advantage of this small parameter and use the perturbation methods to get an approximate analytic solution for the PDE. By assuming a boundary layer at the right side of elongation compartment ($X \to 1^-$), we find an inner solution for the boundary layer region and an outer solution for the region away from the boundary layer. What we are doing here is replacing our single partial differential equation by a set of simpler differential equations in each of the outer and the inner regions [10].

We start by defining the approximate solutions for inner and outer regions $q_{\mathrm{outer}}(X,\tau) = \sum_{n=0}^{\infty} \varepsilon^n q_{\mathrm{out},n}(X,\tau)$ and $q_{\mathrm{inner}}(Y,\tau) = \sum_{n=0}^{\infty} \varepsilon^n q_{\mathrm{in},n}(Y,\tau)$ where $Y = \frac{1-X}{\varepsilon}$ is an inner variable. Since ε is very small in the case of large genes, the zero-order approximation would be a

good approximation for each of the outer and inner regions. So we just need to get $q_{\text{out},0}$ and $q_{\text{in},0}$.

First, we obtain the outer solution. By substituting $q_{\text{outer}} = q_{\text{out},0}$ in the PDE and the boundary conditions and setting $\varepsilon = 0$, we have

$$\frac{\partial}{\partial \tau} q_{\text{out},0}(X,\tau) = -\frac{\partial}{\partial X} q_{\text{out},0}(X,\tau) \tag{2.37}$$

$$q_{\text{out},0}(X,\tau) = 0 \quad at \quad \tau = 0 \tag{2.38}$$

$$q_{\text{out},0}(X,\tau) = \frac{k_{\text{init}}}{\bar{a}} P_{\text{PIC}}(\tau/\bar{a}) \quad at \quad X = 0 \tag{2.39}$$

We can solve this PDE by using the method of Laplace transform in Maple 2016 [51]. We can write the solution simply as

$$q_{\text{out},0}(X,\tau) = \frac{k_{init}}{\bar{a}} \left(1 - H(-\tau + X)\right) P_{PIC}(\tau - X) \tag{2.40}$$

where $H(-\tau + X)$ is a Heaviside function (its value is zero when $X < \tau$ and it is 1 when $X > \tau$).

The outer solution just satisfies the left boundary condition. Now, we find the inner solution for the boundary-layer region. By substituting the inner variable, $Y = \frac{1-X}{\varepsilon}$, into the PDE and boundary conditions, we have

For $\varepsilon = 0$

$$\frac{\partial}{\partial Y} q_{\text{in},0}(Y,\tau) + \frac{\partial^2}{\partial Y^2} q_{\text{in},0}(Y,\tau) = 0 \tag{2.41}$$

and the right boundary condition becomes

$$\frac{\partial q_{\text{in},0}(Y,\tau)}{\partial Y} = 0 \quad at \quad Y = 0 \quad (X = 1) \tag{2.42}$$

The solution for this differential equation is

$$q_{\text{in},0}(Y,\tau) = F(t) \tag{2.43}$$

where $F(t)$ is an unknown function of time which will be determined by matching the inner and outer solutions.

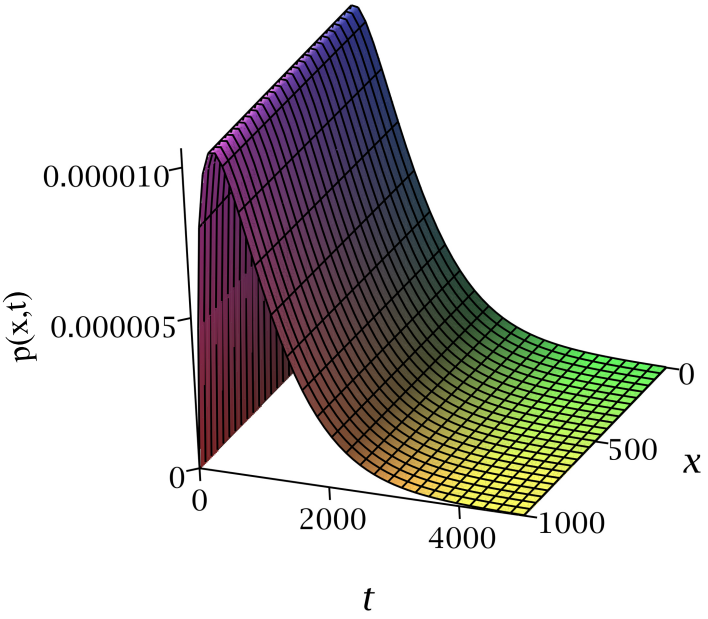

Figure 2.4: The probability density distribution for the elongating RNAP as a function of x and t (equation 2.45). The parameter values: $k_{ae} = 144s^{-1}, k_{elong} = 144s^{-1}, k_{PIC} = 0.0029s^{-1}$, $k_{bind} = 0.0016s^{-1}$ and $k_{init} = 0.6s^{-1}$, $N = 1000$. The values of the parameters in the model are from Vashishtha's thesis [95, p. 48].

By expanding both the inner and outer solutions and matching them, we find the unknown function $F(t)$. Finally we can get the zero-order approximations for the inner solution as follows

$$q_{\text{in},0}(Y,\tau) = \frac{k_{init}}{\bar{a}} \left(1 - H(-\tau+1)\right) P_{PIC}(\tau-1) \tag{2.44}$$

Finally, the zero order solution for the PDE can be written as a piecewise function:

$$q(X,\tau) = \begin{cases} \frac{k_{init}}{a}\,(1-H(-\tau+X))\,P_{PIC}(\tau-X) & 0<X<1-\varepsilon \\ \frac{k_{init}}{a}\,(1-H(-\tau+1))\,P_{PIC}(\tau-1) & 1-\varepsilon<X<1 \end{cases} \tag{2.45}$$

The advection velocity in the PDE is the average velocity of the polymerase. Here, we just have the advective process with a finite velocity. The solution has to be zero where the time has not progressed sufficiently to actually let the solution reach that point. In other words, because of finite propagation speed, you cannot have effects infinitely far away. However, if you have the diffusive part in the PDE, then we no longer have a finite propagation speed. Figure 2.4 shows this probability density distribution as a function of x and t. As seen, the solution is a wave travelling without deformation through the elongation compartment (x direction).

I also solved the PDE numerically by using a discretization method called a "first-order upwind scheme" [16, 32] in Matlab to test our analytical results. Figure 2.5 compares the approximate solution of the PDE which has just an advective term to the numerical solution of the full PDE. The gradual decrease in the tail of the solution to the full PDE in figure 2.5a, which comes from the diffusive term, is an approximation to a random process associated with each step.

2.2.3 Termination

The flux out of the elongation compartment is $J(x=N,t) = a\,p(x=N,t)$. Using equation 2.44, it gives us

$$J(x=N,t) = k_{init}\,(1-H(-\tau+1))\,P_{PIC}(\tau-1) \tag{2.46}$$

where $\tau = t\,\bar{a}$.

This flux enters into the termination compartment, so we should have the following

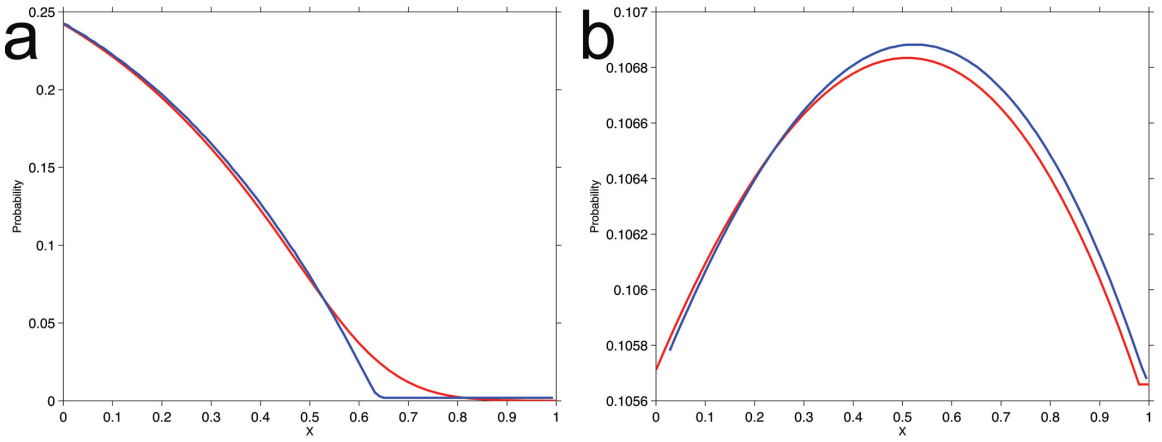

Figure 2.5: A comparison between numerical (red line) and analytical (blue line) solutions of equation 2.33 with boundary conditions (2.35) and (2.36) and initial conditions (2.34). The numerical solution includes the diffusion term, but the analytical solution (equation 2.45) doesn't. The parameter values: a) $k_{ae} = 250s^{-1}, k_{elong} = 30s^{-1}, k_{PIC} = 0.0029s^{-1}$, $k_{bind} = 0.0016s^{-1}$ and $k_{init} = 0.6s^{-1}, \tau = 0.64$, $N = 1000$. b) $k_{ae} = 144s^{-1}, k_{elong} = 144s^{-1}, k_{PIC} = 0.0029s^{-1}$, $k_{bind} = 0.0016s^{-1}$ and $k_{init} = 0.6s^{-1}$, $\tau = 3.83$, $N = 1000$. The values of the parameters in the model are from Vashishtha's thesis [95, p. 48].

ODEs for the termination part

$$\frac{dP_{O,N}}{dt} = J(x = N, t) - k_{ae}P_{O,N}(t) \tag{2.47}$$

$$\frac{dP_{A,N}}{dt} = k_{ae}P_{O,N}(t) - k_{TC}P_{A,N} \tag{2.48}$$

$$\frac{dP_{TC}}{dt} = k_{TC}P_{A,N} - k_{term}P_{TC} \tag{2.49}$$

$$\frac{dP_T}{dt} = k_{term}P_{TC} \tag{2.50}$$

where $P_{TC}(t)$ is the probability of termination complex (TC) formation and $P_T(t)$ represents the probability for release of a transcript (T).

We can solve the ODEs based on the initial conditions of $P_{O,N}(t = 0) = P_{A,N}(t = 0) = P_{TC}(t = 0) = P_T(t = 0) = 0$ to derive an expression for $P_T(t)$. This was computed with Maple, however, it was too complicated to write down here.

Figure 2.6 shows a comparison between the analytical solution and stochastic simu-

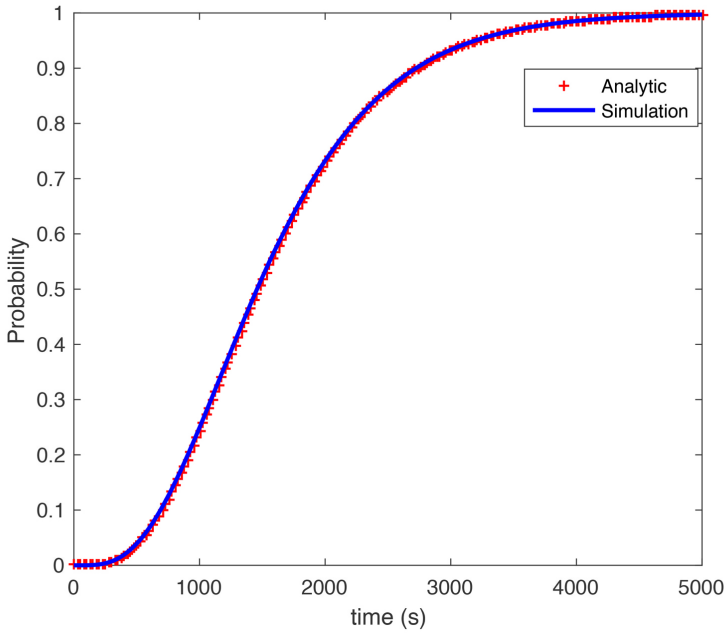

Figure 2.6: A comparison between stochastic simulation (blue line) and analytical (+ symbols) solutions for $P_T(t)$. The parameter values: $k_{ae} = 250s^{-1}, k_{elong} = 30s^{-1}, k_{PIC} = 0.0029s^{-1}$, $k_{bind} = 0.0016s^{-1}$, $k_{init} = 0.6s^{-1}$, $k_{term} = 0.0032s^{-1}$ and $N = 1000$. The values of the parameters in the model are from Vashishtha's thesis [95, p. 48]

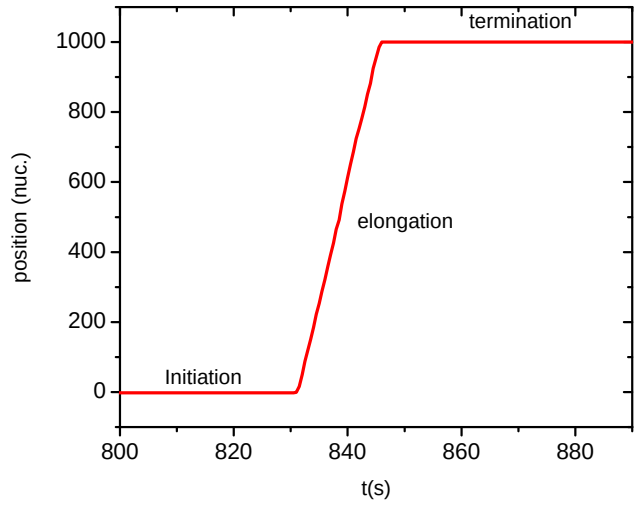

Figure 2.7: Position of the RNA polymerase vs. time obtained from the stochastic simulation of a single RNAP. The parameter values: $k_{ae} = 144s^{-1}, k_{elong} = 144s^{-1}, k_{PIC} = 0.0029s^{-1}$, $k_{bind} = 0.0016s^{-1}$ and $k_{init} = 0.6s^{-1}$, $N = 1000$. The values of the parameters in the model are from Vashishtha's thesis [95, p. 48]

lations of $P_T(t)$. Here, one can see a good match in spite of the very simple advective treatment of the elongation compartment. For the elongation compartment, we obtained a Fokker-Planck equation 2.25 with a diffusive term, which was neglected. However, the diffusive term obtained from the Fokker-Planck treatment has probably overestimated the diffusivity associated with the elongation compartment, as shown by the outstanding agreement between the analytic and simulation results in the end.

Figure 2.7 shows the position of the RNAP against time obtained from the stochastic simulation for a single RNAP. It indicates that when there are no pauses along a large gene, the elongation process is advective and the RNAP moves with a constant average velocity.

2.3 Ubiquitous pausing and its effect on transcription rate

If there is a pausing state at each site of the transcription elongation, we need to add the following chemical reactions for each site in the elongation compartment of the model (see section 2.1). Here, in this model, we assume that pausing occurs from the O state; that is just is an assumption although we could assume having a pause from the A state or from both of the states.

$$\mathrm{O_i} \xrightarrow{k_{\mathrm{pause}}} \mathrm{P_i} \tag{2.51}$$

$$\mathrm{P_i} \xrightarrow{k_{\mathrm{release}}} \mathrm{O_i} \tag{2.52}$$

We can obtain master equations for states in site i, and then will make some approximations to obtain the related partial differential equation for $p(x,t)$.

The master equations in the elongation compartment are

$$\frac{dP_{i,O}}{dt} = k_{\mathrm{elong}}P_{i-1,A} - k_{\mathrm{ae}}P_{i,O} + k_{\mathrm{release}}P_{i,P} - k_{\mathrm{pause}}P_{i,O} \tag{2.53}$$

$$\frac{dP_{i,A}}{dt} = k_{\text{ae}}P_{i,O} - k_{\text{elong}}P_{i,A} \tag{2.54}$$

$$\frac{dP_{i,P}}{dt} = k_{\text{pause}}P_{i,O} - k_{\text{release}}P_{i,P} \tag{2.55}$$

The probability of being at site i is the probability of being in state A at site i plus the probability of being in state O at site i plus the probability of being in state P at site i, that is

$$P_i = P_{i.O} + P_{i,A} + P_{i,P} \tag{2.56}$$

Using equations 2.53 to 2.56, we can obtain the following relationship between $p_{i,A}$ and p_i.

$$\frac{dP_i}{dt} = k_{elong}(P_{i-1,A} - P_{i,A}) \tag{2.57}$$

As explained in the previous section, the residence time argument gives us the relationship between $P_{i,A}$ and P_i, which is the same result that we can get from the steady state assumption. Using equations 2.54, 2.55 in the steady assumption, $\frac{dP_{i,A}}{dt} = 0$, $\frac{dP_{i,P}}{dt} = 0$ and equation 2.56, we have

$$P_{i,A} = \frac{1}{1 + \frac{k_{\text{elong}}}{k_{ae}}\left(1 + \frac{k_{\text{pause}}}{k_{\text{release}}}\right)} P_i \tag{2.58}$$

Substituting equation 2.58 into equation 2.57 and taking a Taylor series expansion over P_{i-1} up to second order, we get the following differential equation.

$$\frac{\partial p(x,t)}{\partial t} = -\tilde{a}\frac{\partial p(x,t)}{\partial x} + \frac{\tilde{D}}{2}\frac{\partial^2 p(x,t)}{\partial x^2} \tag{2.59}$$

where

$$\tilde{a} = (1nuc.)\frac{k_{\text{elong}}}{1+\frac{k_{elong}}{k_{ae}}\left(1+\frac{k_{\text{pause}}}{k_{\text{release}}}\right)} \tag{2.60}$$

is the mean velocity and

$$\tilde{D} = (1nuc.)^2\frac{k_{\text{elong}}}{1+\frac{k_{\text{elong}}}{k_{ae}}\left(1+\frac{k_{\text{pause}}}{k_{\text{release}}}\right)} \tag{2.61}$$

is the diffusion coefficient respectively.

Here I obtained a Fokker-Planck equation for the case of ubiquitous pausing. The only difference from our earlier results without pausing (see equation 2.25) is the difference in the mean velocity and the diffusion coefficient. Comparing equation 2.60 with equation 2.26, we that slowing down elongation is the *only* effect of ubiquitous pausing.

2.4 Effect of a strong pause along a gene

In this section, I consider the elongation process where there is a strong pause site on a gene. In this case, there is a probability flux coming into the pause region, and in the pause region the properties are altered and then it gets back to normal behavior. As discussed in section 2.3, pausing can have a direct effect on the mean velocity 2.60 and the diffusion coefficient 2.61. The boundary conditions before and after the pause should be matched up with the pause region. To apply this reasoning to the model, we could have a pause-free velocity in the left and right side of the pause region and a reduced velocity through a short pause region.

In the case of a large gene, we can ignore the diffusion term as seen in section 2.2. Thus, the PDE describing the left and right side of the pause location will be

$$\frac{\partial p(x,t)}{\partial t} = -a\frac{\partial p(x,t)}{\partial x} \tag{2.62}$$

Pausing is considered as an extended region containing more that one nucleotide. Pauses

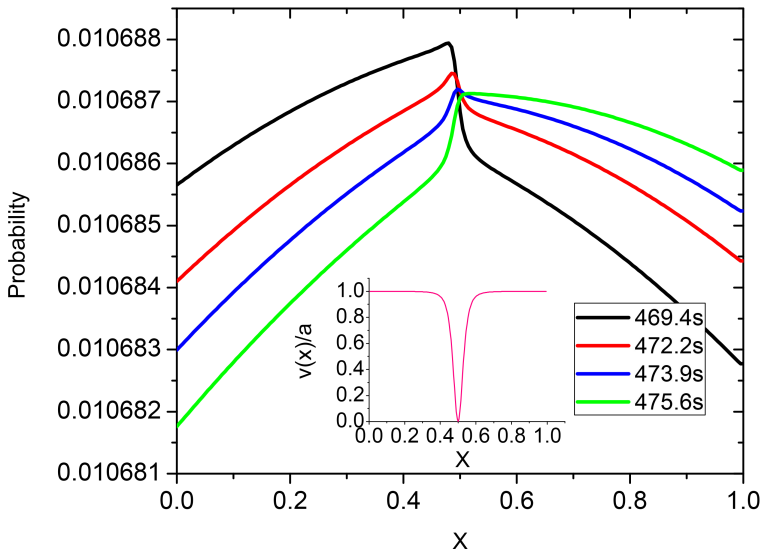

Figure 2.8: Effect of a strong pause in the middle of a gene on the probability distribution. The parameter values: $k_{ae} = 144s^{-1}, k_{elong} = 144s^{-1}, k_{PIC} = 0.0029s^{-1}$, $k_{bind} = 0.0016s^{-1}$ and $k_{init} = 0.6s^{-1}$, $n = m = 2$, $A = B = 0.05$, $X_P = \frac{1}{2}$ and $N = 1000$. The values of the parameters in the model are from Vashishtha's thesis [95, p. 48]

are often caused by pausing hairpins [4, 90, 91], secondary structures formed 10–12 nucleotides upstream from the 3' end of the RNA causing pauses (a temporary delay in the nucleotide addition to the growing transcript) as a result of their interaction with the RNAP. So we should consider a space-dependent velocity for this region ($v(x)$) since a pausing event reduces the effective elongation velocity (see the advection coefficient in equation 2.59). The velocity profile might possibly have a minimum along this region to satisfy the reduced velocity of the RNA polymerase.

A function of the form 2.63 , where A, B, m and n can describe the shape of the velocity profile in the pausing region, would be a good choice. The shape of this function is shown as an inset to figure 2.8.

$$v(x) = a\left(1 - \frac{A^n}{(x - x_P)^n + A^n}\right)\left(1 - \frac{B^m}{(x - x_P)^m + B^m}\right) \tag{2.63}$$

The PDE (2.62) was solved numerically by using a "first-order upwind scheme" [16, 32]

in Matlab. Figure 2.8 shows the probability density as a function of position of the RNA polymerase (both are dimensionless) for an example representing the effect of a strong pause on the probability density of the RNAP. As time increases from 469.4 s to 475.6 s, we can see a qualitative change of behaviour. At 469.4s and 472.2s, the probability density of the RNAP to be downstream of the pause region (after departing the pause region) is smaller that the probability density of being upstream (before entering the pause region). However, at time 473.9 s and 475.6 s, it is opposite and the probability density is larger downstream of the pause that upstream.

2.5 Summary

In this chapter, we studied a simplified version of Vashishtha's model [95] for transcription of a single RNA polymerase in eukaryotic cells to characterize the stochastic properties of the transcription process. We obtained an explicit expression for the probability density function (equation 2.45) describing the spatio-temporal organization of elongating RNA polymerase valid in the case of large genes (genes of a few hundred nucleotides or more). From this, we concluded that the entering probability flux from the elongation into the termination compartment is the same as entering flux from the initiation into the elongation compartment but with a delay in time ($\tau = N/a$). Comparing the analytic result with the stochastic simulation result (figure 2.6) confirmed it. Therefore, we found that the elongation compartment can be considered as a delay which is useful for understanding the temporal properties of transcription. It is also a justification for the use of delays in gene expression models, and in particular in delay-stochastic models [71, 79]. We then obtained the probability distributions of initiation times and of total transcription time. In addition, solving the one-body problem gave us vital clues as to how to tackle the two-body problem.

We also considered the effect of ubiquitous pausing in the elongation compartment and showed it slows down the elongation speed of RNA polymerase. The effect of strong pauses

was also considered and we showed that how it can affect the probability density function in the elongation compartment, causing a distortion in the travelling probability wave. We may argue that the strong pause can affect the delay time in the elongation compartment. Having the Fokker-Planck equation in the case of a strong pause, we can obtain the first passage time of the RNAP reaching the end of the elongation compartment. It will give us the delay time in this case.

Chapter 3

Two-body effects on gene expression

In this chapter, the theory to explore the effect of interaction of two elongating RNA polymerases on the transcription kinetics has been investigated. When there is a trailing RNA polymerase behind the leading RNA polymerase, the leading RNAP can act as a moving wall and the probability distribution of the trailing RNA polymerase may pile up behind that. We would therefore expect that the leading RNAP would affect the shape of the probability distribution for the trailing RNAP.

3.1 Joint probability distribution for two RNA polymerases

3.1.1 Initiation

To develop the model for the two-body problem, we assume that each RNA polymerase (in the case of RNA polymerase II) occupies sites from $n = 20$ bp downstream of the transcription start site (TSS) and $m = 50$ bp behind the TSS [19, 38]. So the total length of the RNA polymerase ($\Delta = m + n$) is 70 bp. The trailing RNAP cannot bind the promoter region until the leading RNAP has moved far enough downstream. To include this mutual exclusion effect, we introduce

$$Q(t) = 1 - \left(\int_{x=0}^{\Delta} p(x,t)dx + P_{pro}^{(1)}(t) + P_{TBP\cdot pro}^{(1)}(t) + P_{PIC}^{(1)}(t) \right) \tag{3.1}$$

as the probability that the promoter has been cleared by the first polymerase. In equation 3.1, the integral term, in which the integrand, $p(x,t)$, is the probability density function

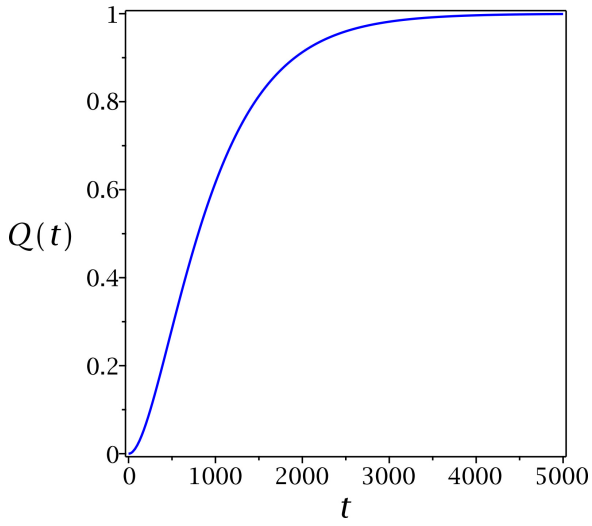

Figure 3.1: The probability that there is no part of the first RNAP in the region required for binding of initiation factors. The parameters are the same as in figure 2.6.

for the elongation compartment (the dimensional form of equation 2.45), represents the probability that the first RNAP overlaps the region of -50 to $+20$ positions downstream of the transcription start site. $P^{(1)}_{pro}(t)$ (equation 2.13), $P^{(1)}_{TBP\cdot pro}(t)$ (equation 2.14) and $P^{(1)}_{PIC}(t)$ (equation 2.15) are the probabilities of the first RNAP to be in any of the initial states. This region needs to be clear in order for the trailing RNAP to position itself at the TSS, and that we assume that this needs to be clear in order for other initiation factors to bind. Figure 3.1 shows how $Q(t)$ changes with time.

We need to define the probability flux feeding the elongation compartment. For the first RNA polymerase binding to the promoter to initiate transcription, as explained in subsection 2.2.2, we have the following at $x = x_0 = 0$:

$$a\,p(x,t) - \frac{D}{2}\frac{\partial p(x,t)}{\partial x}\bigg|_{x=0} = k_{init} P^{(1)}_{PIC}(t) \tag{3.2}$$

For the second RNAP, the entering probability flux from the initiation part should be conditional upon clearance of the promoter region from the first one. Obviously, there are two dependent events. One has to be completed before the other one can be initiated. First, we need to solve the initiation problem for the second polymerase. If we assume that the

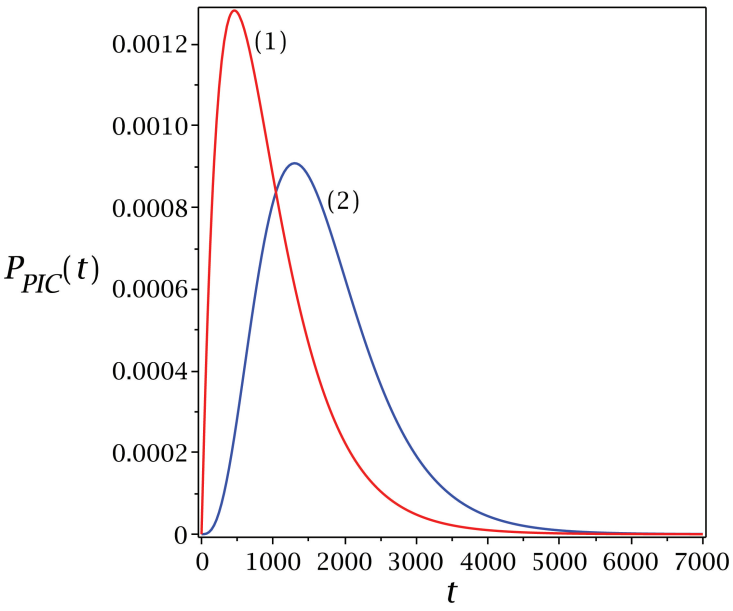

Figure 3.2: The probabilities of being in the PIC state for the first and the second RNAPs. The parameters are the same as in figure 2.6.

concentrations of the promoter, TBP and other initiation factors are constant, we can write the following equations for initiation of the second RNA polymerase.

$$\frac{dP^{(2)}_{pro}(t)}{dt} = -k_{bind}\, Q(t) P^{(2)}_{pro}(t) \tag{3.3}$$

$$\frac{dP^{(2)}_{TBP\cdot pro}(t)}{dt} = k_{bind}\, Q(t) P^{(2)}_{pro}(t) - k_{PIC} P^{(2)}_{TBP\cdot pro}(t) \tag{3.4}$$

$$\frac{dP^{(2)}_{PIC}(t)}{dt} = k_{PIC} P^{(2)}_{TBP\cdot pro}(t) - k_{init} P^{(2)}_{PIC}(t) \tag{3.5}$$

$$\frac{dP^{(2)}_{O(t)}}{dt} = k_{init} P^{(2)}_{PIC}(t) \tag{3.6}$$

where $Q(t)$ has been defined in equation 3.1. Here, $P^{(2)}_{pro}$ is the probability that the second RNAP has not yet bound to the promoter.

So we have a system of ODEs with the initial conditions of $P^{(2)}_{pro}(t) = 1$, $P^{(2)}_{TBP\cdot pro}(t) = 0$, $P^{(2)}_{PIC}(t) = 0$, $P^{(2)}_{O}(t) = 0$. Maple was unsuccessful at solving this system of ODEs so the system was solved numerically. Here, figure 3.2 shows the numerical results obtained for

$P^{(2)}_{PIC}(t)$ compared with $P^{(1)}_{PIC}$ of the first RNA polymerase. As seen in this figure, the mean initiation time for the second RNAP is larger than that for the first RNAP as expected. In addition, the distribution for the second RNAP is broader than the distribution for the first RNAP that indicates having larger fluctuation or standard deviation for the second RNAP.

From these equations, we can get a conditional probability for initiation of the second RNA polymerase. It will give us the probability flux entering the elongation part:

$$a\,p(y,t) - \frac{D}{2}\frac{\partial p(y,t)}{\partial y}\bigg|_{y=0} = k_{init} P^{(2)}_{PIC}(t) \tag{3.7}$$

3.1.2 Elongation

3.1.2.1 Governing equation

We introduce $P(j,k;t)$ to be the joint probability of having the active site of the first RNA polymerase, RNAP$^{(1)}$, to be at site j and the trailing RNA polymerase, RNAP$^{(2)}$, to be at site k. In the model, we have just two states (occupied (O) and activated (A)) associated with the active site of the elongating RNAPs. $P(j,\mathrm{O};k,\mathrm{A};t)$ is, for example, the probability of RNAP$^{(1)}$ to be in the occupied state at site j and RNAP$^{(2)}$ to be in the activated state at site k.

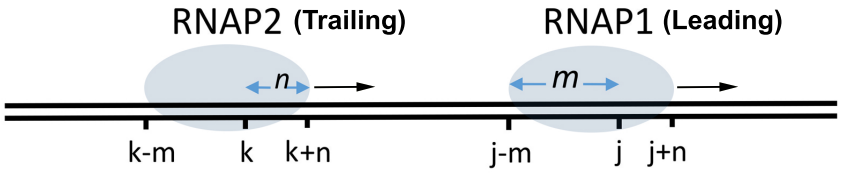

Figure 3.3: Two RNA polymerases on a DNA strand. RNAP$^{(1)}$ is the leading and RNAP$^{(2)}$ is the trailing one moving from left to right. The length of both RNAPs are the same and it is $\Delta = m+n$. The integer indices j and k give the locations of the active sites of the two RNAPs. We assume that the RNAPs are rigid objects and they cannot occupy the same active site and also they cannot cross each other. This means that there is always this condition between two RNAPs' active sites: $j-k \geq \Delta$.

We write master equations for the joint probability distributions for the four possible combinations of states for two RNAPs. The combinations are: $P(j,O;k,O;t)$, $P(j,O;k,A;t)$,

$P(j,A;k,O;t)$ and $P(j,A;k,A;t)$. These probabilities are not independent of each other because they are mutually exclusive, and we have the following for $P(j,k,t)$:

$$P(j,k,t) = P(j,O;k,O;t) + P(j,O;k,A;t) + P(j,A;k,O;t) + P(j,A;k,A;t) \tag{3.8}$$

There are four possible transitions for each combination of states that includes both transitions into and out of each state. For example, if RNAP$^{(1)}$ is in state A and RNAP$^{(2)}$ is in state O, the possibilities are as shown in figure 3.4. The master equation for the time evolution of this state has been given in equation 3.9.

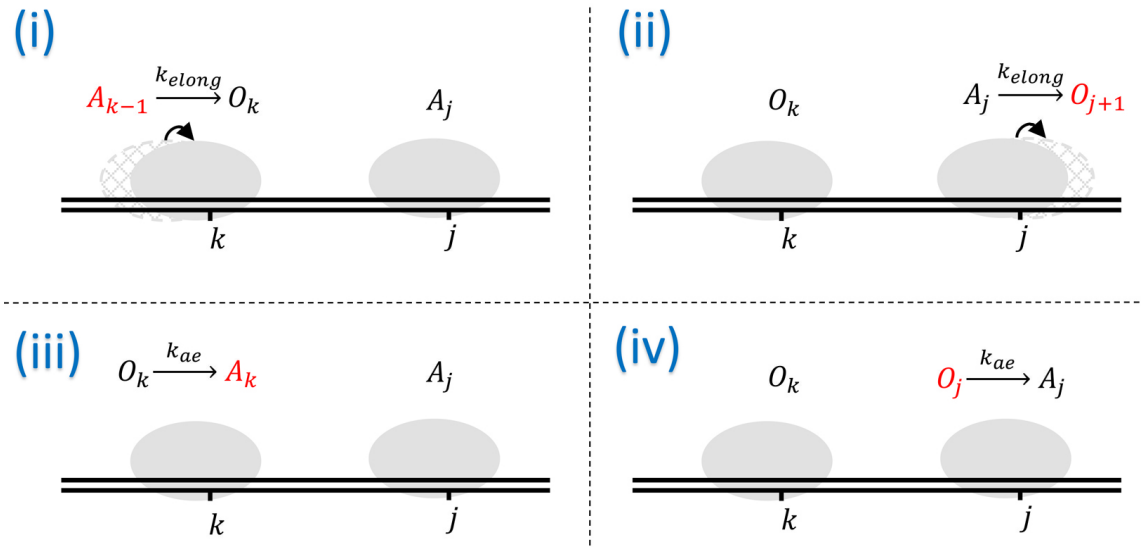

Figure 3.4: Four possible transitions into and out of state (j,A,k,A). Panels (i) and (iv) represent transitions into the (j,A;k,O) state and panels (ii) and (iii) represent transitions out of this state. (i) the translocation of RNAP$^{(2)}$ from site $k-1$ to k. (ii) the translocation RNAP$^{(1)}$ from site j to $j+1$. (iii) the activation of RNAP$^{(2)}$ at site k. (iv) the activation of RNAP$^{(1)}$ at site j.

The master equations are as follows:

$$\begin{aligned}\frac{dP(j,A;k,O;t)}{dt} &= k_{elong}P(j,A,t;k-1,A,t)\,H(j-k-\Delta) \\ &\quad -k_{ae}P(j,A;k,O;t)\,H(j-k-\Delta) \\ &\quad +k_{ae}P(j,O;k,O;t)\,H(j-k-\Delta) \\ &\quad -k_{elong}P(j,A;k,O;t)\,H(j-k-\Delta)\end{aligned} \tag{3.9}$$

$$\begin{aligned}\frac{dP(j,A;k,A;t)}{dt} &= k_{ae}P(j,A;k,O;t)\,H(j-k-\Delta) \\ &\quad -k_{elong}P(j,A;k,A;t)\,H(j-(k+1)-\Delta) \\ &\quad +k_{ae}P(j,O;k,A;t)H(j-k-\Delta) \\ &\quad -k_{elong}P(j,A;k,A;t)H(j-k-\Delta)\end{aligned} \tag{3.10}$$

$$\begin{aligned}\frac{dP(j,O;k,O;t)}{dt} &= k_{elong}P(j,O,t;k-1,A,t)\,H(j-k-\Delta) \\ &\quad -k_{ae}P(j,O;k,O;t)H(j-k-\Delta) \\ &\quad +k_{elong}P(j-1,A,t;k,O,t)\,H((j-1)-k-\Delta) \\ &\quad -k_{ae}P(j,O;k,O;t)H(j-k-\Delta)\end{aligned} \tag{3.11}$$

$$\begin{aligned}\frac{dP(j,O;k,A;t)}{dt} &= k_{ae}P(j,O;k,O;t)H(j-k-\Delta) \\ &\quad -k_{elong}P(j,O;k,A;t)\,H(j-(k+1)-\Delta) \\ &\quad +k_{elong}P(j-1,A,t;k,A,t)\,H((j-1)-k-\Delta) \\ &\quad -k_{ae}P(j,O;k,A;t)H(j-k-\Delta)\end{aligned} \tag{3.12}$$

$H(\xi)$ is a Heaviside function, which is zero if $\xi < 0$ and 1 if $\xi \geq 0$. Note that I am using the convention in which $H(0) = 1$. So here $H(j-k-\Delta)$ is zero if $j-k < \Delta$ and 1 if

$j-k \geq \Delta$, representing the condition that the distance between j and k cannot be less than Δ (nucleotides).

By summation over both sides of equations 3.12 to 3.9, and using equation 3.8, we can obtain

$$\begin{aligned}\frac{dP(j,k,t)}{dt} &= k_{elong}P(j,O;k-1,A,t)\,H(j-k-\Delta) \\ &+k_{elong}P(j,A;k-1,A,t)\,H(j-k-\Delta) \\ &-k_{elong}P(j,O;k,A;t)\,H(j-(k+1)-\Delta) \\ &-k_{elong}P(j,A;k,A;t)\,H(j-(k+1)-\Delta) \\ &+k_{elong}P(j-1,A;k,A,t)\,H((j-1)-k-\Delta) \\ &+k_{elong}P(j-1,A;k,O,t)\,H((j-1)-k-\Delta) \\ &-k_{elong}P(j,A;k,A;t)\,H(j-k-\Delta) \\ &-k_{elong}P(j,A;k,O;t)\,H(j-k-\Delta)\end{aligned} \tag{3.13}$$

Let us define

$$P(j,O;k,A;t)+P(j,A;k,A;t)=P(j;k,A,t) \tag{3.14}$$

$$P(j,A;k,O;t)+P(j,A;k,A;t)=P(j,A;k,t) \tag{3.15}$$

Then, using equations 3.14 and 3.15

$$\begin{aligned}\frac{dP(j,k,t)}{dt} &= k_{elong}P(j;k-1,A,t)\,H(j-k-\Delta) \\ &-k_{elong}P(j;k,A,t)\,H(j-(k+1)-\Delta) \\ &+k_{elong}P(j-1,A;k,t)\,H((j-1)-k-\Delta) \\ &-k_{elong}P(j,A;k,t)\,H(j-k-\Delta)\end{aligned} \tag{3.16}$$

Assuming a steady-state probability current as in 2.21, we get $P(j,k,t)=(1+\frac{k_{elong}}{k_{ae}})P(j,k,A,t)$

and $P(j,k,t) = (1+\frac{k_{elong}}{k_{ae}})P(j,A,k,t)$, and using a Taylor expansion

$$P(j,k-1,t) = P(j,k,t) - \Delta Y \frac{\partial P(j,k,t)}{\partial Y} + \frac{1}{2}\Delta Y^2 \frac{\partial^2 P(j,k,t)}{\partial Y^2} - ... \tag{3.17}$$

And the same expansion for $P(j-1,t;k,t)$ gives

$$P(j-1,k,t) = P(j,k,t) - \Delta X \frac{\partial P(j,k,t)}{\partial X} + \frac{1}{2}\Delta X^2 \frac{\partial^2 P(j,k,t)}{\partial X^2} - ... \tag{3.18}$$

If we substitute these approximations into equation 3.16, we have

$$\begin{aligned}
\frac{dP(j,k,t)}{dt} = \frac{k_{elong}k_{ae}}{k_{elong}+k_{ae}}\Big[P(j,k,t) - \Delta Y \frac{\partial P(j,k,t)}{\partial Y} + \frac{1}{2}\Delta Y^2 \frac{\partial^2 P(j,k,t)}{\partial Y^2}\Big] H(j-k-\Delta) \\
- \frac{k_{elong}k_{ae}}{k_{elong}+k_{ae}} P(j,k,t)\, H(j-(k+1)-\Delta) \\
+ \frac{k_{elong}k_{ae}}{k_{elong}+k_{ae}}\Big[P(j,k,t) - \Delta X \frac{\partial P(j,k,t)}{\partial X} + \frac{1}{2}\Delta X^2 \frac{\partial^2 P(j,k,t)}{\partial X^2}\Big] H((j-1)-k-\Delta) \\
- \frac{k_{elong}k_{ae}}{k_{elong}+k_{ae}} P(j,k,t)\, H(j-k-\Delta)
\end{aligned}$$

Finally, in the continuum limit, when the length of a gene (N) is large enough, we can define ($X = j\Delta X$, $Y = k\Delta Y$ where ΔX and ΔY are dimensionless increments defined as $\frac{1}{N}$).

If we give the dimension of X back ($X = \frac{x}{N}$), we can obtain the following Fokker-Planck equation for two polymerases:

$$\begin{aligned}
\frac{\partial p(x,y,t)}{\partial t} = & -a\,H(x-y-\Delta-1)\,p(x,y,t) + a\,H(x-y-\Delta)\,p(x,y,t) \\
& -a\,H(x-y-\Delta)\frac{\partial p(x,y,t)}{\partial y} + \frac{D}{2}\,H(x-y-\Delta)\frac{\partial^2 p(x,y,t)}{\partial y^2} \\
& -a\,H(x-y-\Delta)p(x,y,t) + a\,H(x-y-\Delta-1)\,p(x,y,t) \\
& -a\,H(x-y-\Delta-1)\frac{\partial p(x,y,t)}{\partial x} + \frac{D}{2}\,H(x-y-\Delta-1)\frac{\partial^2 p(x,y,t)}{\partial x^2}
\end{aligned} \tag{3.19}$$

where a and D (see 2.25) are the average velocity and the diffusion coefficient respectively.

3.1.2.2 Derivation of analytical solution

Let us define the following dimensionless variables:

$$X = \frac{x}{N}$$

$$Y = \frac{y}{N}$$

$$\tau = \frac{t\,a}{N}$$

$$q(X,Y,\tau) = N\,p(X,Y,t)$$

$$\alpha = \frac{\Delta}{N}$$

$$\beta = \frac{1}{N}$$

$$Pe = 2N$$

Inserting these variables into equation 3.19 will give us a dimensionless partial differential equation that describes the evolution of the joint probability density distribution of the two RNA polymerases.

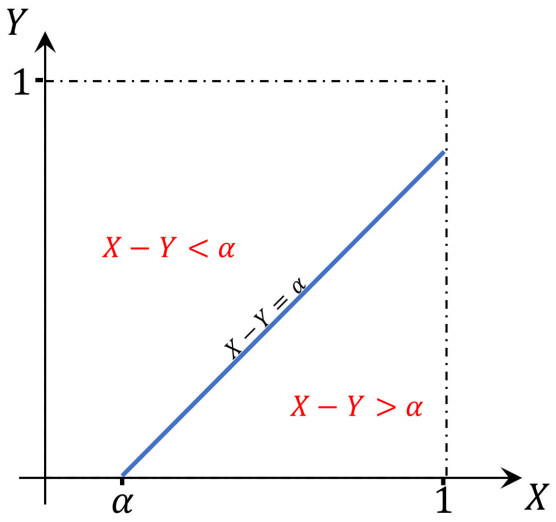

Figure 3.5: Domain and boundaries for the two-polymerase Fokker-Planck equation.

Two cases:

1. $X - Y \geq \alpha + \beta$

   The PDE becomes

$$\frac{\partial q(X,Y,\tau)}{\partial \tau} = -\frac{\partial q(X,Y,\tau)}{\partial Y} + \frac{1}{\mathrm{Pe}}\frac{\partial^2 q(X,Y,\tau)}{\partial Y^2} - \frac{\partial q(X,Y,\tau)}{\partial X} + \frac{1}{\mathrm{Pe}}\frac{\partial^2 q(X,Y,\tau)}{\partial X^2}$$

2. $X - Y = \alpha$

   This case is happening when two RNA polymerases are bumper-to-bumper (see figure 3.5). So having a no-flux condition across a boundary representing "bumper-to-bumper" contact is reasonable.

$$\hat{n} \cdot \vec{J}\Big|_{X=Y-\alpha} = 0 \tag{3.20}$$

   where

$$\hat{n} = \frac{\hat{X} - \hat{Y}}{\sqrt{2}}$$

$$\vec{J} = \Big(- q(X,Y,\tau) + \frac{1}{Pe}\frac{\partial q(X,Y,\tau)}{\partial X}\Big)\hat{X} + \Big(- q(X,Y,\tau) + \frac{1}{Pe}\frac{\partial q(X,Y,\tau)}{\partial Y}\Big)\hat{Y}$$

Eventually, we can describe the kinetics of two RNA polymerases by the following Fokker-Planck equation and initial/boundary equations

$$\frac{\partial q(X,Y,\tau)}{\partial \tau} = -\frac{\partial q(X,Y,\tau)}{\partial Y} + \frac{1}{\mathrm{Pe}}\frac{\partial^2 q(X,Y,\tau)}{\partial Y^2} - \frac{\partial q(X,Y,\tau)}{\partial X} + \frac{1}{\mathrm{Pe}}\frac{\partial^2 q(X,Y,\tau)}{\partial X^2} \tag{3.21}$$

where $0 < X, Y < 1, \quad \tau > 0$

$$q(X,Y,\tau) = 0 \quad \text{at} \quad \tau = 0 \tag{3.22}$$

$$q(X,\tau) - \frac{1}{\text{Pe}}\frac{\partial q(X,\tau)}{\partial X} = \frac{k_{init}}{\overline{a}} P^{(2)}_{PIC}(\tau/\overline{a}) \quad \text{at} \quad X = 0 \tag{3.23}$$

$$q(Y,\tau) - \frac{1}{\text{Pe}}\frac{\partial q(Y,\tau)}{\partial Y} = \frac{k_{init}}{\overline{a}} P^{(2)}_{PIC}(\tau/\overline{a}) \quad \text{at} \quad Y = 0 \tag{3.24}$$

$$\frac{\partial q(X,\tau)}{\partial X} = 0 \quad \text{at} \quad X = 1 \tag{3.25}$$

$$\frac{\partial q(Y,\tau)}{\partial Y} = 0 \quad \text{at} \quad Y = 1 \tag{3.26}$$

$$\hat{n} \cdot \vec{J} = 0 \quad \text{at} \quad X = Y - \alpha \tag{3.27}$$

3.1.3 Termination

The termination of the leading RNA polymerase wouldn't be a problem since there are no restrictions for it. However, the termination of the trailing RNA polymerase would be affected by the existence of the leading one. As explained in the initiation section of this chapter, we have to make sure that the leading RNAP is completely dissociated from the end of the gene so that the trailing RNAP gets into the termination region. So we need a probability that guarantees that there is no part of the first RNAP at the termination region. We define

$$W(t) = P^{(1)}_T(t) \tag{3.28}$$

where $P^{(1)}_T(t)$ is the probability that the leading RNA polymerase is dissociated from the gene.

The probability flux for the trailing RNAP, $J^{(2)}(y,t)$, enters into the termination compartment from the elongation compartment. In order to take the steric exclusion effect into account, we multiply the k_{ae} terms in equations 3.29 and 3.30 by $W(t)$. This is the simplest way we can include the exclusion effect in the master equations. If you multiply $J^{(2)}(y = N,t)$ in equation 3.29 by $W(t)$, you will end up having a probability leak (since

$W(t) < 1$). In, addition, there isn't a steric repulsion problem in the context of this model in the elongation compartment up to site $N-\Delta$. We also neglect the part of elongation compartment in the interval $k \in [N-\Delta+1, N]$ Since $\Delta \ll N$ for large genes. So we can replace $P^{(2)}_{O,N-\Delta}$ with $P^{(2)}_{O,N}$. Hence, the master equations for the termination of the trailing RNAP are as follows:

$$\frac{dP^{(2)}_{O,N}}{dt} = J^{(2)}(y=N,t) - k_{ae}W(t)P^{(2)}_{O,N}(t) \tag{3.29}$$

$$\frac{dP^{(2)}_{A,N}}{dt} = k_{ae}W(t)P^{(2)}_{O,N}(t) - k_{TC}P^{(2)}_{A,N}(t) \tag{3.30}$$

$$\frac{dP^{(2)}_{TC}}{dt} = k_{TC}P^{(2)}_{A,N}(t) - k_{\text{term}}P^{(2)}_{TC}(t) \tag{3.31}$$

$$\frac{dP^{(2)}_{T}(t)}{dt} = k_{\text{term}}P^{(2)}_{TC}(t) \tag{3.32}$$

where $J^{(2)}(y=N,t)$ is the flux of the probability for the trailing RNAP. If we assume that the elongation of the trailing RNAP is advective, the probability density for the second RNAP would be as follows:

$$J^{(2)}(y=N,t) = k_{\text{init}}\,H(t-t_0)\,P^{(2)}_{PIC}(t-t_0) \tag{3.33}$$

where $t_0 = \frac{N}{a}$ is the elongation time.

In section 3.1.1, we mentioned that we could only get numerical solutions for $P^{(2)}_{PIC}$. If we had an analytical solution for $P^{(2)}_{PIC}$, we could evaluate $J^{(2)}(y=N,t)$ using equation 3.33 explicitly, and equation 3.29 becomes a simple linear ordinary differential equation with an inhomogeneous term. Then, solving this system of ODEs would be possible and would give us explicit expressions for the probability distributions of being in each state.

3.2 Extending to a many-body problem

In the previous section, we found a Fokker-Plank equation (FPE) for two elongating RNA polymerases. In this section, I generalize our two-polymerase FPE to the case of N polymerases and find the corresponding FPE and boundary/initial conditions.

The PDE becomes

$$\frac{\partial q(\vec{X},\tau)}{\partial \tau} = \sum_{n=1}^{N} \Big[- \frac{\partial q(\vec{X},\tau)}{\partial X_n} + \frac{1}{\mathrm{Pe}} \frac{\partial^2 q(\vec{X},\tau)}{\partial X_n^2} \Big]$$

We can write the following equations for initiation of each of RNA polymerases.

$$\frac{dP^{(n)}_{TBP\cdot pro}(t)}{dt} = k_{bind}\, P^{(n-1)}(x_{n-1} > \Delta, t) - k_{PIC} P^{(n)}_{TBP\cdot pro}(t) \tag{3.34}$$

where the probability that $x > \Delta$ is the conditional probability as follows:

$$P^{(n-1)}(x_{n-1} > \Delta, t) = 1 - \left(\int_{x=0}^{\Delta} p(x_{n-1}, t)dx + P^{(n-1)}_{pro}(t) + P^{(n-1)}_{TBP\cdot pro}(t) + P^{(n-1)}_{PIC}(t) \right) \tag{3.35}$$

$$\frac{dP^{(n)}_{PIC}(t)}{dt} = k_{PIC} P^{(n)}_{TBP\cdot pro}(t) - k_{init} P^{(n)}_{PIC}(t) \tag{3.36}$$

So the entering flux for each RNAP is

$$a\, p(x_n, t) - \frac{D}{2} \frac{\partial p(x_n, t)}{\partial x_n} = k_{init} P^{(n)}_{PIC}(t) \tag{3.37}$$

For the right boundary conditions or the termination part, we have $\left.\frac{\partial p(x_n,t)}{\partial x_n}\right|_{x_n = x_f} = 0$ and the fact that the RNA polymerases cannot pass each other, can be described by the following condition.

$$\hat{n} \cdot \vec{J}\Big|_{X_n - X_{n-1} = \alpha} = 0 \tag{3.38}$$

where

$$\hat{n} = \frac{\hat{X}_n - \hat{X}_{n-1}}{\sqrt{2}}$$

$$\vec{J} = \Big(- q(\vec{X},\tau) + \frac{1}{\mathrm{Pe}} \frac{\partial q(\vec{X},\tau)}{\partial X_n} \Big) \hat{X}_n + \Big((- q(\vec{X},\tau) + \frac{1}{\mathrm{Pe}} \frac{\partial q(\vec{X},\tau)}{\partial X_{n-1}}) \Big) \hat{X}_{n-1}$$

3.3 Summary

In this chapter, we considered the transcription of two RNAPs initiating from the same promoter of a gene in eukaryotic cells. The probability distributions in the initiation compartment were obtained for both polymerases and compared (figure 3.2). For the elongation compartment, the Fokker-Planck equation was obtained for the joint probability distribution of two elongating RNA polymerases (3.21). The related non-trivial boundary conditions are also obtained. However, there is more work to be done to obtain either exact or approximate solutions of this equation. For the termination compartment, we obtained the master equations describing the evolution of different states for each one of the RNAPs in this compartment. Here we assumed the elongation compartment for both RNAPs to be like a delay in time between initiation and termination. This way we easily obtained the entering probability flux into the termination compartment for both RNAPs. The mutual exclusion effect was also taken into account to consider the effect of the leading RNAP on the termination of the trailing one. However, there is again more work to be done to obtain exact solutions for the probability distributions of each state for both RNAPs.

Chapter 4

Conclusions and future direction

Summary and conclusions

In chapter 2 of this work, a simplified version of the transcription model proposed by Vashishtha [95] for eukaryotic cells was utilized to describe the kinetics of a single RNAP during transcription. By writing the master equations for the initiation compartment and solving them, an analytical expression was obtained for the probability distribution of the initiation time which gives the probability flux entering into the elongation compartment. By writing the master equations for the elongation compartment and using some approximations, I obtained a corresponding Fokker-Planck equation. The Fokker-Planck equation is just an equation of motion for the distribution function of fluctuating macroscopic variables (here, the position of RNAP along the gene (x)) and time (t)). The equation was solved analytically for the limit of a long gene (genes of at a few hundred nucleotides or more) to obtain an approximate solution for the probability density distribution describing the spatio-temporal organization of the elongating RNAP. The master equations for the termination compartment were solved to give an expression for the probability distribution of the total transcription time. It is worth mentioning that all three compartments are connected through boundary conditions which are necessary to know in order to solve the differential equations in each compartment. The mean velocity and the diffusion coefficient for the elongating RNAP are also obtained from the Fokker-Planck equation.

One of the important results of this thesis was that transcription is an advective process

when the length of a gene is large enough. If we look at the Fokker-Planck equation, we see that the coefficient of the diffusion term is dependent on the length of the gene and can be ignored for large genes. As a result, it is found that the elongation compartment can be considered as a delay which can be a justification for the use of delays in gene expression models, and in particular in delay-stochastic models [71, 79].

The effect of two specific types of pausing during elongation was also considered. First, I explored the effect of ubiquitous pausing on the transcription process, especially on the mean velocity and the diffusion coefficient of an RNAP. As expected, ubiquitous pausing has no other effect than to reduce the velocity and diffusion coefficient. Second, the effect of a strong pause was considered. Here, since the velocity of the RNAP wouldn't be constant, a Hill-type function for the velocity profile was introduced to replace the advection coefficient in our Fokker-Planck equation. The equation was solved numerically giving us the probability density distribution for the elongating RNAP.

In chapter 3, a mathematical model was developed to consider the transcription of two RNAPs initiating from the same gene. The second RNA polymerase cannot bind to the promoter of the gene if the first RNAP occupies the promoter region. I obtain the probability distribution for the initiation rate for each one of the RNAPs. Next, for the elongation compartment, starting from our kinetic model, both the Fokker-Planck equation for the case of two RNAPs and a nontrivial boundary conditions for this equation were obtained. The boundary condition arises by setting the probability flux to zero when the two RNAPs come in contact with each other. Having the entering flux from the initiation into the elongation compartment for both RNAPs as the left boundary condition and the point that there is no diffusivity on the right side which gives us the right boundary condition, will help us to solve this PDE (future work). We also studied the termination part and found the probability distribution for the transcription time for each of the RNAPs.

It should be indicated that having analytical solutions for this simple model helps us to explore the parameter space rapidly. It leads to exact results which will be approximately correct when the assumptions of the model are slightly relaxed, and which may form the basis for, e.g., perturbative solutions as we move away from the conditions of strict validity. Fitting experimental results to analytic models, and therefore extracting estimates of model parameters, is also another possibility.

Future direction

Future work would be to make the model more realistic and add more details to it. In this thesis, we used a simplified version of Vashishtha's model [95] where we excluded abortive initiation, promoter escape and promoter-proximal pausing from the model. For transcription catalyzed by RNA polymerase II, over half of the initiations are abortive [48]. For the next step, we can consider adding abortive initiation to our model since it is an important process. Effectively in this case, you will get a probability leak early in the elongation compartment. So, one can imagine that we have a pipe representing the elongation compartment that is fed by initiation and in turn feeds termination. The leak-out of the pipe leads to a re-initialization to a state in the initiation compartment. In addition, we have neglected the sequence dependence that could possibly be added into our model in future work.

Another interesting thing would be determining which of two rate constants in the elongation part is bigger, k_{elong} or k_{ae}. The rate constant, k_{ae}, depends on the availability of nucleotides, and k_{elong} doesn't. So most possibly one of those might be rate-limiting. We can consider this issue in our model to get a better understanding of transcription elongation. It would also be worthwhile to get something more rigorous than just assuming the average residence time or the steady state, by actually writing down a formal solution for the ODEs in the elongation compartment. Therefore, we may be able to get a rigorous

relationship between $P_{O,i}$ and $P_{\mathrm{A},i}$ using a scaling argument.

Another thing we couldn't accomplish in this work was finding an analytical solution for the case of two elongating RNA polymerases. It would be the next priority for our future work. This will help us to explore the effect of the interaction of RNAPs, specially the effect of the leading RNAP movement on the elongation of the trailing one.

Eventually, we would consider the many-body problem so that it will enable us to say something about collective motion of RNAPs and to predict the results of related experimental data if available. We can also consider the interaction of the RNA polymerase with other machines. For example, in prokaryotes the polymerase can interact with the ribosome. The ribosome starts translation before the polymerase gets finished [27]. So it would be interesting if we could model and explore the interaction of two different types of molecular machines.

Bibliography

[1] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. (2014). *Molecular Biology of the Cell. 6th edition*. Garland Science.

[2] Anderson, D. H. (2013). *Compartmental modeling and tracer kinetics*, volume 50. Springer Science & Business Media.

[3] Arimbasseri, A. G. and Maraia, R. J. (2016). RNA polymerase III advances: Structural and tRNA functional views. *Trends in Biochemical Sciences*, 41(6):546 – 559.

[4] Artsimovitch, I. and Landick, R. (2000). Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. *Proceedings of the National Academy of Sciences*, 97(13):7090–7095.

[5] Bai, L., Fulbright, R. M., and Wang, M. D. (2007). Mechanochemical kinetics of transcription elongation. *Physical Review Letters*, 98(6):068103.

[6] Bai, L., Santangelo, T. J., and Wang, M. D. (2006). Single-molecule analysis of RNA polymerase transcription. *Annual Review of Biophysics and Biomolecular Structure*, 35:343–360.

[7] Bai, L., Shundrovsky, A., and Wang, M. D. (2004). Sequence-dependent kinetic model for transcription elongation by RNA polymerase. *Journal of Molecular Biology*, 344(2):335–349.

[8] Bar-Yam, Y., Harmon, D., and de Bivort, B. (2009). Attractors and democratic dynamics. *Science*, 323(5917):1016–1017.

[9] Bellomo, N. and Dogbe, C. (2011). On the modeling of traffic and crowds: A survey of models, speculations, and perspectives. *SIAM Review*, 53(3):409–463.

[10] Bender, C. M. and Orszag, S. A. (1999). *Advanced Mathematical Methods for Scientists and Engineers I*. Springer Science & Business Media.

[11] Callen, B. P., Shearwin, K. E., and Egan, J. B. (2004). Transcriptional interference between convergent promoters caused by elongation over the promoter. *Molecular cell*, 14(5):647–656.

[12] Chinchilla, K., Rodriguez-Molina, J. B., Ursic, D., Finkel, J. S., Ansari, A. Z., and Culbertson, M. R. (2012). Interactions of sen1, nrd1, and nab3 with multiple phosphorylated forms of the rpb1 c-terminal domain in saccharomyces cerevisiae. *Eukaryotic Cell*, 11(4):417–429.

[13] Chou, T. and Lakatos, G. (2004). Clustered bottlenecks in mRNA translation and protein synthesis. *Physical Review Letters*, 93(19):198101.

[14] Chou, T., Mallick, K., and Zia, R. (2011). Non-equilibrium statistical mechanics: from a paradigmatic model to biological transport. *Reports on Progress in Physics*, 74(11):116601.

[15] Chowdhury, D., Schadschneider, A., and Nishinari, K. (2005). Physics of transport and traffic phenomena in biology: from molecular motors and cells to organisms. *Physics of Life Reviews*, 2(4):318–352.

[16] Courant, R., Isaacson, E., and Rees, M. (1952). On the solution of nonlinear hyperbolic differential equations by finite differences. *Communications on Pure and Applied Mathematics*, 5(3):243–255.

[17] Davis, L., Gedeon, T., Gedeon, J., and Thorenson, J. (2014). A traffic flow model for bio-polymerization processes. *Journal of Mathematical Biology*, 68(3):667–700.

[18] Dong, J., Schmittmann, B., and Zia, R. K. (2007). Towards a model for protein production rates. *Journal of Statistical Physics*, 128(1-2):21–34.

[19] Dvir, A., Conaway, J. W., and Conaway, R. C. (2001). Mechanism of transcription initiation and promoter escape by RNA polymerase II. *Current Opinion in Genetics and Development*, 11(2):209 – 214.

[20] Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186.

[21] Epshtein, V. and Nudler, E. (2003). Cooperation between RNA polymerase molecules in transcription elongation. *Science*, 300(5620):801–805.

[22] Epshtein, V., Toulmé, F., Rahmouni, A. R., Borukhov, S., and Nudler, E. (2003). Transcription through the roadblocks: the role of RNA polymerase cooperation. *The EMBO Journal*, 22(18):4719–4727.

[23] Ezeokonkwo, C., Ghazy, M. A., Zhelkovsky, A., Yeh, P.-C., and Moore, C. (2012). Novel interactions at the essential n-terminus of poly(a) polymerase that could regulate poly(a) addition in saccharomyces cerevisiae. *FEBS Letters*, 586(8):1173–1178.

[24] Fokker, A. (1914). Fokker-planck equation. *Ann. Physik*, 43:810.

[25] Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434.

[26] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

[27] Gowrishankar, J. and Harinarayanan, R. (2004). Why is transcription coupled to translation in bacteria? *Molecular Microbiology*, 54(3):598–603.

[28] Greger, I. H., Aranda, A., and Proudfoot, N. (2000). Balancing transcriptional interference and initiation on the gal7 promoter of saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 97(15):8415–8420.

[29] Guajardo, R. and Sousa, R. (1997). A model for the mechanism of polymerase translocation. *Journal of Molecular Biology*, 265(1):8–19.

[30] Hamkalo, B. and Miller Jr, O. (1973). Electron microscopy of genetic activity. *Annual Review of Biochemistry*, 42(1):379–396.

[31] Hocine, S., Singer, R. H., and Grünwald, D. (2010). RNA processing and export. *Cold Spring Harbor perspectives in biology*, 2(12):a000752.

[32] Hoffmann, K. A. and Chiang, S. T. (2000). Computational fluid dynamics, Vol. 1. *Wichita, KS: Engineering Education System*.

[33] Jülicher, F. and Bruinsma, R. (1998). Motion of RNA polymerase along DNA: a stochastic model. *Biophysical Journal*, 74(3):1169–1185.

[34] Keller, W. and Minvielle-Sebastia, L. (1997). A comparison of mammalian and yeast pre-mRNA 3'-end processing. *Current Opinion in Cell Biology*, 9(3):329 – 336.

[35] Kim, J. L., Nikolov, D. B., and Burley, S. K. (1993a). Co-crystal structure of tbp recognizing the minor groove of a tata element. *Nature*, 365(6446):520–527.

[36] Kim, J. L., Nikolov, D. B., and Burley, S. K. (1993b). Co-crystal structure of tbp recognizing the minor groove of a tata element. *Nature*, 365(6446):520–527.

[37] Kim, M., Krogan, N. J., Vasiljeva, L., Rando, O. J., Nedea, E., Greenblatt, J. F., and Buratowski, S. (2004). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature*, 432(7016):517–522.

[38] Kim, T.-K., Lagrange, T., Wang, Y.-H., Griffith, J. D., Reinberg, D., and Ebright, R. H. (1997). Trajectory of DNA in the RNA polymerase II transcription preinitiation complex. *Proceedings of the National Academy of Sciences*, 94(23):12268–12273.

[39] Klumpp, S. (2011). Pausing and backtracking in transcription under dense traffic conditions. *Journal of Statistical Physics*, 142(6):1252–1267.

[40] Klumpp, S. and Hwa, T. (2008). Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination. *Proceedings of the National Academy of Sciences*, 105(47):18159–18164.

[41] Kulasiri, D., Nguyen, L. K., Samarasinghe, S., and Xie, Z. (2008). A review of systems biology perspective on genetic regulatory networks with examples. *Current Bioinformatics*, 3(3):197–225.

[42] Lagomarsino, M. C., Bassetti, B., Castellani, G., and Remondini, D. (2009). Functional models for large-scale gene regulation networks: realism and fiction. *Molecular BioSystems*, 5(4):335–344.

[43] Lakatos, G. and Chou, T. (2003). Totally asymmetric exclusion processes with particles of arbitrary size. *Journal of Physics A: Mathematical and General*, 36(8):2027.

[44] Landick, R. (2006). The regulatory roles and mechanism of transcriptional pausing. *Biochemical Society Transactions*, 34(6):1062–1066.

[45] Lepzelter, D., Feng, H., and Wang, J. (2010). Oscillation, cooperativity, and intermediates in the self-repressing gene. *Chemical Physics Letters*, 490(4):216–220.

[46] Li, B., Carey, M., and Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, 128(4):707 – 719.

[47] Logan, J., Falck-Pedersen, E., Darnell, J. E., and Shenk, T. (1987). A poly (a) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase ii in the mouse beta maj-globin gene. *Proceedings of the National Academy of Sciences*, 84(23):8306–8310.

[48] Luse, D. S. and Jacob, G. A. (1987). Abortive initiation by RNA polymerase II in vitro at the adenovirus 2 major late promoter. *Journal of Biological Chemistry*, 262(31):14990–14997.

[49] MacDonald, C. T. and Gibbs, J. H. (1969). Concerning the kinetics of polypeptide synthesis on polyribosomes. *Biopolymers*, 7(5):707–725.

[50] MacDonald, C. T., Gibbs, J. H., and Pipkin, A. C. (1968). Kinetics of biopolymerization on nucleic acid templates. *Biopolymers*, 6(1):1–25.

[51] Maple (2016). The Maplesoft Inc., Waterloo, ON, http://www.maplesoft.com/products/Maple.

[52] Margeat, E., Kapanidis, A. N., Tinnefeld, P., Wang, Y., Mukhopadhyay, J., Ebright, R. H., and Weiss, S. (2006). Direct observation of abortive initiation and promoter escape within single immobilized transcription complexes. *Biophysical Journal*, 90(4):1419–1431.

[53] Marshall, N. F. and Price, D. H. (1992). Control of formation of two distinct classes of RNA polymerase ii elongation complexes. *Molecular and Cellular Biology*, 12(5):2078–2090.

[54] McAdams, H. and Arkin, A. (1997). Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819.

[55] McClure, W. R. (1985). Mechanism and control of transcription initiation in prokaryotes. *Annual Review of Biochemistry*, 54(1):171–204.

[56] Miller, O. and Beatty, B. R. (1969). Visualization of nucleolar genes. *Science*, 164(3882):955–957.

[57] Nechaev, S. and Adelman, K. (2011). Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1809(1):34 – 45.

[58] Neuman, K. C., Abbondanzieri, E. A., Landick, R., Gelles, J., and Block, S. M. (2003). Ubiquitous transcriptional pausing is independent of RNA polymerase backtracking. *Cell*, 115(4):437 – 447.

[59] Pal, M. and Luse, D. S. (2003). The initiationelongation transition: Lateral mobility of RNA in RNA polymerase II complexes is greatly reduced at +8/+9 and absent by +23. *Proceedings of the National Academy of Sciences*, 100(10):5700–5705.

[60] Pal, M., Ponticelli, A. S., and Luse, D. S. (2005). The Role of the Transcription Bubble and TFIIB in Promoter Clearance by RNA Polymerase II. *Molecular Cell*, 19(1):101 – 110.

[61] Park, N. J., Tsao, D. C., and Martinson, H. G. (2004). The two steps of poly (a)-dependent termination, pausing and release, can be uncoupled by truncation of the RNA polymerase ii carboxyl-terminal repeat domain. *Molecular and Cellular Biology*, 24(10):4092–4103.

[62] Paulsson, J. (2005). Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175.

[63] Planck, M. (1917). Sitzungsber. *Preuss. Akad. Wiss. Phys. Math. Kl*, 325:3.

[64] Plischke, M. and Bergersen, B. (2006). *Equilibrium statistical physics*. World Scientific.

[65] Pomerantz, R. T. and ODonnell, M. (2008). The replisome uses mRNA as a primer after colliding with RNA polymerase. *Nature*, 456(7223):762–766.

[66] Pomerantz, R. T. and ODonnell, M. (2010). Direct restart of a replication fork stalled by a head-on RNA polymerase. *Science*, 327(5965):590–592.

[67] Proudfoot, N. J. (2011). Ending the message: poly(a) signals then and now. *Genes and Development*, 25(17):1770–1782.

[68] Rahl, P. B., Lin, C. Y., Seila, A. C., Flynn, R. A., McCuine, S., Burge, C. B., Sharp, P. A., and Young, R. A. (2010). c-myc regulates transcriptional pause release. *Cell*, 141(3):432 – 445.

[69] Rausenberger, J., Fleck, C., Timmer, J., and Kollmann, M. (2009). Signatures of gene expression noise in cellular systems. *Progress in Biophysics and Molecular Biology*, 100(1):57–66.

[70] Revyakin, A., Liu, C., Ebright, R. H., and Strick, T. R. (2006). Abortive Initiation and Productive Initiation by RNA Polymerase Involve DNA Scrunching. *Science*, 314(5802):1139–1143.

[71] Ribeiro, A. S., Smolander, O.-P., Rajala, T., Häkkinen, A., and Yli-Harja, O. (2009). Delayed stochastic model of transcription at the single nucleotide level. *Journal of Computational Biology*, 16(4):539–553.

[72] Richard, P. and Manley, J. L. (2009). Transcription termination by nuclear RNA polymerases. *Genes and Development*, 23(11):1247–1269.

[73] Risken, H. and Frank, T. (1996). *The Fokker-Planck Equation: Methods of Solution and Applications*, volume 18. Springer Science & Business Media.

[74] Robert, G. and William, J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*, 224:235.

[75] Roeder, R. G. (2005). Transcriptional regulation and the role of diverse coactivators in animal cells. $\{FEBS\}$ *Letters*, 579(4):909 – 915. Molecular Mechanisms of Biological Systems. 130th Nobel Symposium.

[76] Rosonina, E., Kaneko, S., and Manley, J. L. (2006). Terminating the transcript: breaking up is hard to do. *Genes and Development*, 20(9):1050–1056.

[77] Roussel, M. R. (2013). On the distribution of transcription times. *BIOMATH*, 2(1):1307247.

[78] Roussel, M. R. and Zhu, R. (2006a). Stochastic kinetics description of a simple transcription model. *Bulletin of Mathematical Biology*, 68(7):1681–1713.

[79] Roussel, M. R. and Zhu, R. (2006b). Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. *Physical Biology*, 3(4):274.

[80] Saeki, H. and Svejstrup, J. Q. (2009). Stability, flexibility, and dynamic interactions of colliding RNA polymerase II elongation complexes. *Molecular cell*, 35(2):191–205.

[81] Selth, L. A., Sigurdsson, S., and Svejstrup, J. Q. (2010). Transcript elongation by RNA polymerase II. *Annual Review of Biochemistry*, 79:271–293.

[82] Sequeira-Mendes, J. and Gómez, M. (2012). On the opportunistic nature of transcription and replication initiation in the metazoan genome. *Bioessays*, 34(2):119–125.

[83] Shahrezaei, V. and Swain, P. S. (2008). Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261.

[84] Shearwin, K. E., Callen, B. P., and Egan, J. B. (2005). Transcriptional interference–a crash course. *Trends in Genetics*, 21(6):339–345.

[85] Sims, R. J., Belotserkovskaya, R., and Reinberg, D. (2004). Elongation by RNA polymerase II: the short and long of it. *Genes and Development*, 18(20):2437–2468.

[86] Sneppen, K., Dodd, I. B., Shearwin, K. E., Palmer, A. C., Schubert, R. A., Callen, B. P., and Egan, J. B. (2005). A mathematical model for transcriptional interference by RNA polymerase traffic in *Escherichia coli*. *Journal of Molecular Biology*, 346(2):399–409.

[87] Strachan, T., Read, A. P., and Strachan, T. (2011). *Human Molecular Genetics*. Garland Science, New York.

[88] Tennyson, C. N., Klamut, H. J., and Worton, R. G. (1995). The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nature Genetics*, 9(2):184–190.

[89] Tiana, G., Krishna, S., Pigolotti, S., Jensen, M. H., and Sneppen, K. (2007). Oscillations and temporal signalling in cells. *Physical Biology*, 4(2):R1.

[90] Toulokhonov, I., Artsimovitch, I., and Landick, R. (2001). Allosteric control of RNA polymerase by a site that contacts nascent RNA hairpins. *Science*, 292(5517):730–733.

[91] Toulokhonov, I. and Landick, R. (2003). The flap domain is required for pause RNA hairpin inhibition of catalysis by RNA polymerase and can modulate intrinsic termination. *Molecular Cell*, 12(5):1125–1136.

[92] Tripathi, T. and Chowdhury, D. (2008). Interacting RNA polymerase motors on a DNA track: Effects of traffic congestion and intrinsic noise on RNA synthesis. *Physical Review E*, 77(1):011921.

[93] Van Kampen, N. G. (1992). *Stochastic Processes in Physics and Chemistry*, volume 1. Elsevier.

[94] Vannini, A. and Cramer, P. (2012). Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Molecular Cell*, 45(4):439 – 446.

[95] Vashishtha, S. (2011). Stochastic modeling of eukaryotic transcription at the single nucleotide level. Master's thesis, University of Lethbridge.

[96] Veiga, D. F., Dutta, B., and Balázsi, G. (2010). Network inference and network response identification: moving genome-scale data to the next level of biological discovery. *Molecular BioSystems*, 6(3):469–480.

[97] Voliotis, M., Cohen, N., Molina-París, C., and Liverpool, T. B. (2008). Fluctuations, pauses, and backtracking in DNA transcription. *Biophysical Journal*, 94(2):334–348.

[98] Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G. A., Winston, F., et al. (1998). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes and Development*, 12(3):343–356.

[99] Wang, H.-Y., Elston, T., Mogilner, A., and Oster, G. (1998). Force generation in RNA polymerase. *Biophysical Journal*, 74(3):1186–1202.

[100] West, S., Gromak, N., and Proudfoot, N. J. (2004). Human $5' \rightarrow 3'$ exonuclease xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature*, 432(7016):522–525.

[101] Woo, H.-J. (2006). Analytical theory of the nonequilibrium spatial distribution of RNA polymerase translocations. *Physical Review E*, 74(1):011907.

[102] Yager, T. D. and Von Hippel, P. H. (1991). A thermodynamic analysis of RNA transcript elongation and termination in *Escherichia coli*. *Biochemistry*, 30(4):1097–1118.

[103] Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J., and Handa, H. (1999). NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*, 97(1):41–51.

[104] Yin, Y. W. and Steitz, T. A. (2004). The structural mechanism of translocation and helicase activity in T7 RNA polymerase. *Cell*, 116(3):393–404.

[105] Zhou, Y. and Martin, C. T. (2006). Observed instability of T7 RNA polymerase elongation complexes can be dominated by collision-induced bumping. *Journal of Biological Chemistry*, 281(34):24441–24448.

[106] Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes and Development*, 21(9):1010–1024.