STOCHASTIC MODELING OF THE TORPEDO MECHANISM OF EUKARYOTIC TRANSCRIPTION TERMINATION

REBA-JEAN MURPHY
Bachelor of Science, Augustana Faculty (University of Alberta), 2010

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

STOCHASTIC MODELING OF THE TORPEDO MECHANISM OF EUKARYOTIC TRANSCRIPTION TERMINATION

REBA-JEAN MURPHY

Date of Defense: June 27, 2017

Dr. Marc Roussel Supervisor	Professor	Ph.D.
Dr. Ute Wieden-Kothe Committee Member	Associate Professor	Ph.D.
Dr. John Sheriff Committee Member	Assistant Professor	Ph.D.
Dr. Ovidiu Radulescu External Examiner Université de Montpellier Montpellier, France	Professor	Ph.D.
Dr. Paul Hayes Chair, Thesis Examination Committee	Professor	Ph.D.

Dedication

One M.Sc. Thesis, in completion, dedicated to Jesus Christ.

Abstract

The torpedo mechanism is a type of eukaryotic transcription termination completed by RNAPII and consists of a chase-down of RNAPII by the exonuclease Xrn2 on the 3' flanking region of the RNA transcript. A biochemical model of the mechanism is detailed. By applying Gillespie simulations and using a one-dimensional biased random walk, the effects of the relative speeds of RNAPII to Xrn2 and the size of the intergenic region on termination via the torpedo mechanism are explored. The success rates of termination in terms of the likelihood of downstream gene interference, the distribution of successful termination times, and the distribution of nucleotides synthesized by RNAPII in the termination region are discussed.

Acknowledgments

My supervisor, Dr. Marc Roussel, has been a huge inspiration and dedicated mentor, so thank you. In addition, I extend my thanks to the Roussel research lab and the Alberta RNA Research and Training Institute (ARRTI) and its members for the fantastic inspiration. Thank you to the members of my thesis examination committee. As for the personal end, it is hard to find some catch-all groupings for the vast numbers of fantastic people who have walked into my life and changed things for the better in the last few years. To name some, the members of the Gate Church and other brothers and sisters in Christ, the Lethbridge climbing community, members of the object manipulation club, and the LGTBQ++ community all help ground me, but I know that of course I've missed some. Friends and family, you are my home.

Contents

List of Tables

List of Figures

List of Abbreviations

bp: nucleotide base pairs

CF1A: cleavage factor 1A

CoTC element: cotranscriptional cleavage element

CPF: cleavage and polyadenylation factor

CTD: carboxy-terminal domain of the Rpd1 subunit in RNA polymerase II

CUT: cryptic unstable transcript

IRS: intergenic region size

NAC: nucleotide addition cycle

nt: nucleotide

PAS: poly(A) signal

PSF: protein-associated splicing factor

RNAP: prokaryotic RNA polymerase

RNAPI, RNAPII, and RNAPIII: eukaryotic RNA polymerases I, II, and III

Chapter 1

Introduction

1.1 On transcription termination

Transcription termination is the means by which an RNA polymerase (RNAP) is dislodged from double-stranded DNA after the production of an RNA molecule [2]. The ways that termination is achieved are numerous and diverse across the prokaryotic and eukaryotic domains [3–5]. During transcription, bacterial RNAP and eukaryotic RNAPI, II, and III all operate on a transcription bubble [2]. The transcription bubble is an open, or melted, piece of double stranded DNA about 10 to 20 nucleotides (nt) long [2]. Within this transcription bubble is an RNA-DNA hybrid which is approximately 8 nt long [2, 3]. This RNA polymerase combined with the transcription bubble and RNA-DNA hybrid is typically called the elongation complex. In general, termination is induced by various methods which disrupt the RNA-DNA hybrid, causing instability of the elongation complex and disengaging RNA polymerase [3, 4]. The types of termination mechanisms have been grouped according to the type of RNA polymerase that uses it.

Bacterial RNAP is destabilized using hairpins, a secondary structure which is formed from the nascent RNA being transcribed. The hairpin then pushes against RNAP [3] or interacts with the trigger loop (a structure of RNAP which is essential to the nucleotide addition cycle) [6] to induce termination. Additionally, RNAP termination can be induced through destabilization by the Rho complex [3, 5]. A translocase called Mfd was also found to use ATP in order to push RNAP from behind to destabilize it [3]. RNAPI, which primarily transcribes ribosomal RNA (rRNA) uses a protein called Reb1, which acts like

a roadblock to forcibly halt RNAPI on a DNA sequence for which the RNA-DNA hybrid is unstable [3, 5]. RNAPIII is responsible for rRNA, as well as tRNA and other non-coding RNA (ncRNA). Its primary termination mechanisms are RNA sequence dependent, requiring no additional factors for termination [3].

Of particular interest is the termination mechanism for RNAPII, the polymerase responsible for the transcription of mRNA products and a number of snRNAs and cryptic unstable transcripts (CUTs) [4, 7, 8]. RNAPII is composed of twelve subunits named Rpb1, Rpb2,... Rpb12 [9]. There are ten core subunits which are mostly conserved amongst eukaryotic RNA polymerases, and an additional subcomplex made from Rpb4/Rpb7 [9]. RNAPII is unique in its structure among RNA polymerases, possessing a carboxy-terminal domain (CTD) on the Rpb1 subunit that gets phosphorylated in specified patterns during the transcription process [4, 10]. The CTD is built of heptad units which most often follow a canonical sequence Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7. The CTD is tail-like and has a flexible structure, and a number of NMR and X-ray studies have been done to determine its structure in complex with various proteins [9]. The general location of the CTD's connection to RNAPII matters, but not the specific connection to the polymerase [11]. The CTD was removed from Rpb1, and attached to the carboxy-terminal ends of the subunits Rpb4 and 6, and 9 [11]. The mutants with CTD connections to subunits Rpb4 and Rpb6 (in close proximity to the WT connection point) were viable whereas the Rpb9 mutant wasn't. Heidemann *et al.* [12] warn that different species with the same CTD phosphorylation patterns react differently to those alterations, and that strict interpretation of phosphorylation patterns as a universal code for transcriptional processing is not recommended. With this understanding, there are some general phosphorylation patterns observed on the CTD of RNAPII as it approaches the end of a gene and prepares for termination that have been shown to recruit termination related proteins [4, 10, 12, 13]. Ser5P tends to be dephosphorylated into Ser5 further and further along the transcript, and its availability at the end of the transcript seems to depend on the gene length [10]. Meanwhile, Ser2P levels begin to

rise during transcription and peak in the termination region [4, 10, 12, 13]. Ser2P levels drop after the polymerase has moved about 100 nt downstream of the poly(A) signal (PAS) [10]. Tyr1P tends to gradually increase during transcription but abruptly drops right before the PAS [12]. Finally, phosphorylation of Tyr4 abruptly occurs just after the PAS [12].

There are three models of transcription termination for RNAPII: the Nrd1-Nab3-Sen1 dependent pathway, the allosteric model, and the torpedo model. Nrd1-Nab3-Sen1 dependent termination is typically utilized for snRNAs and some CUTs, but has also been used in short mRNA genes [4, 7]. Sen1 is recruited by Ser2 phosphorylation (Ser2P) on the CTD, whilst Nrd1 is recruited by Ser5P [10], which might explain why this pathway is sometimes used for short mRNAs where Ser5P hasn't had a chance to be dephosphorylated. It is hypothesized that the helicase Sen1 unwinds the RNA-DNA hybrid in RNAPII in order to induce termination [4]. As for most mRNA transcripts, a mixture of two models, the allosteric model and torpedo model, seems to be the leading hypothesis for RNAPII transcription termination [3, 4, 7, 14, 15]. In allosteric termination, it is hypothesized that termination factors associate with RNAPII in order to produce a conformational change that destabilizes the polymerase and causes termination [7]. Meanwhile, the torpedo model relies on cleavage machinery at the PAS to provide an exposed RNA for an exonuclease, which then breaks down RNA faster than RNAPII transcribes it. When the exonuclease catches up with RNAPII, it "torpedoes" into it, destabilizing the polymerase [7]. This mixed allosteric-torpedo model has been recently coined cleavage and polyadenylation factor (CPF) termination by Porrua *et al.* [5], for the reason that while a number of proteins have been identified as having an effect on termination, the underlying mechanisms still have ambiguity.

In *Saccharomyces cerevisiae*, Cleavage Factor 1A (CF1A) contains the subunit Pcf11 upon which both the allosteric and torpedo mechanism rely [4, 7]. The allosteric model typically refers to Pcf11 related activities that destabilize RNAPII without the use of an exonuclease. Zhang *et al.* showed that Pcf11 binds to both the CTD of RNAPII and to

the RNA, and in the same study, that it is capable of terminating a paused RNAPII [16]. A cleavage incapable mutant of Pcf11 was still helpful in termination [7], while another Pcf11 mutant that is cleavage capable but unable to bind to the CTD results in defective termination (as reviewed in [10]). Overall, the ability for Pcf11 to bridge between RNA and the CTD of RNAPII is vital in this particular termination mechanism [16]. Phosphorylation of Ser2 on RNAPII's CTD heptad units recruits Pcf11, while parallel phosphorylation of Tyr1 inhibits that recruitment [12]. Finally, Pcf11 is only termination capable when there are PAS or cotranscriptional cleavage (CoTC) elements available [17]. West and Proudfoot [17] found that RNA degradation was reduced when Pcf11 was reduced. The cleavage activity of Pcf11 makes it a requirement for the torpedo model.

The torpedo model requires the cleavage of the RNA transcript in order to allow the entry of the 5' to 3' exonuclease Xrn2 (Rat1 in yeast), which can occur at the PAS or CoTC elements [17, 18]. The torpedo model was shown to be linked to 3' end processing through the recruitment of Xrn2 to the PAS region by its ability to associate with protein-associated splicing factor (PSF) in HeLa cells [19] and the aforementioned requirement of Pcf11 for proper Xrn2 recruitment in yeast [20]. The torpedo model requires that Xrn2 (Rat1 in yeast) is recruited to the 3' flanking region that is still being transcribed by RNAPII after 3' processing begins on the pre-mRNA. Once Xrn2 is attached, it begins to degrade the RNA at a faster rate than RNAPII synthesizes it. This leads to Xrn2 being likened to a torpedo hitting its target RNAPII as the degradation of the RNA brings it closer to RNAPII. Once Xrn2 catches up to the polymerase, an event occurs that destabilizes the polymerase and induces termination. The mechanism of this event presently unknown. An overview of transcription termination by RNAPII can be seen in Figure 1.1. Overall, an allosteric-torpedo mechanism provides redundancy which in turn increases the chances of efficient termination. Experiments which inhibit the allosteric and torpedo mechanisms through knockdown and mutation of termination factors seem to show that neither mechanism can completely account for the success rates of termination [14, 15].

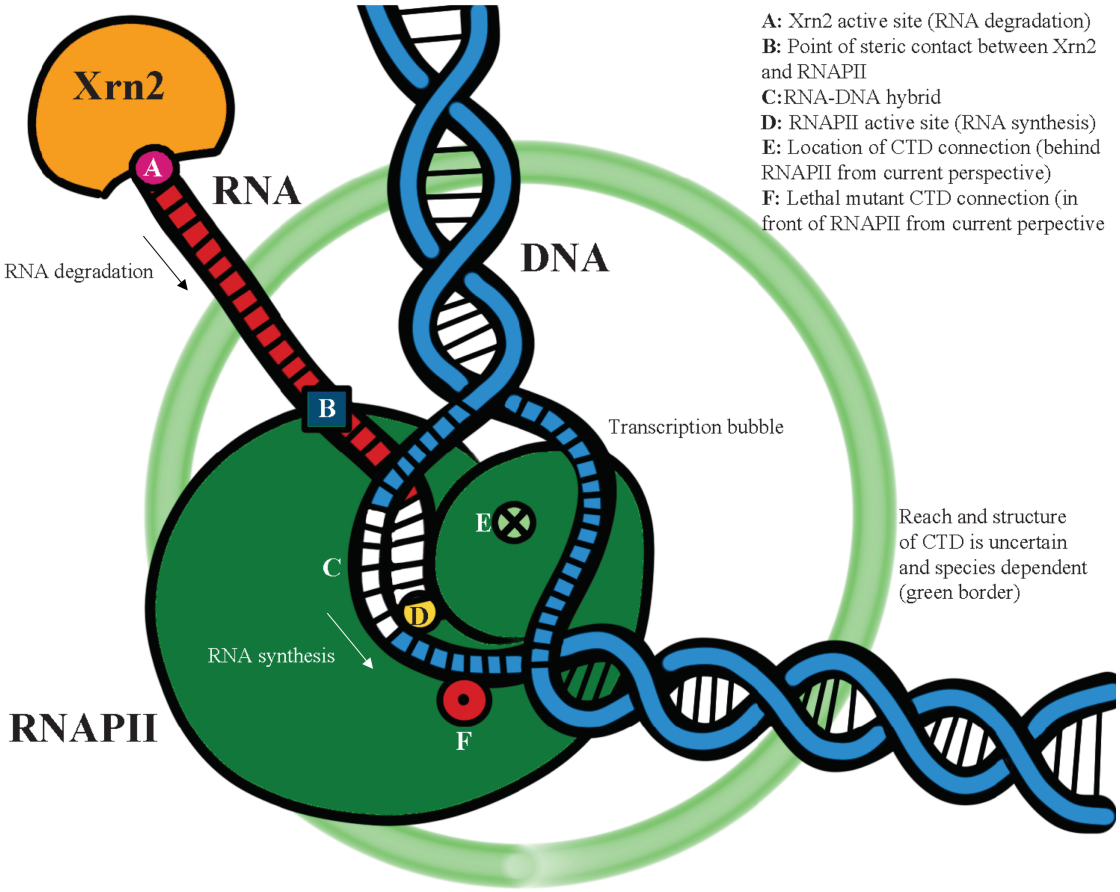


Figure 1.1: Transcription termination of RNAPII by the torpedo mechanism. An idealized model of RNAPII is used to highlight key aspects of the torpedo mechanism. The elongation complex composed of RNAPII, double stranded DNA with an open transcription bubble, the RNA-DNA hybrid (position C) within the bubble, and newly transcribed RNA shows the general configuration of RNAPII during the nucleotide addition cycle (occurs at position D). The RNA is cleaved by CF1A (not shown), allowing Xrn2 to be recruited to the 5' end of the RNA (position A) and to begin degrading the RNA transcript. Xrn2 degrades RNA faster than RNAPII can synthesize it and reaches position B, where an unknown interaction occurs between Xrn2 and RNAPII to disengage RNAPII from the transcript. The reach of the CTD shown by the green circle is difficult to assess due to variable length of the CTD from species to species [12] and flexible structure [9, 12]. Position E shows the CTD connection point of successful mutants on the surface furthest from current point of view (Rpb4 and Rpb6). Lethality is observed when the base of the CTD is connected to Rpb9 at position F, on the surface closest to the current point of view.

As more information on transcription termination becomes available, more emphasis has been put on identifying key unifying mechanisms across domains [3, 5]. Porrua *et al.* [5] recently collected the key mechanisms found in the literature: sequence dependent termination, road blocks, allosteric, and RNA-DNA hybrid shortening through hypertranslocation and shearing forces. They identified that there is a problem with trying to sort the termination mechanisms into these categories. The information we have about termination mechanisms is often limited to identifying the structure and function of proteins along with observing termination defects upon inducing specific mutations or knocking down particular proteins. While these experiments give us many hypotheses about the underlying mechanics, there is often a gap in knowledge based on whether or not these proposals are mechanically viable. For example, how Xrn2 induces termination upon catching up to RNAPII is still debated. If kinetic modeling is to reveal whether or not a particular mechanism is viable, we need to be able to account for stochasticity in the torpedo model in relation to RNAPII behaviour on the end of the gene.

1.2 Objectives and theoretical background

Efficient termination is important. When an RNA polymerase is not stopped properly and instead runs through a termination region, it results in downstream gene interference, which would result in genes not required being expressed [4, 14]. Additionally, efficiency reduces the amount of the energy wasted by the cell from building RNAs doomed for degradation [14]. Finally, efficient termination can increase RNAPII recycling [20, 21]. Modeling transcription termination therefore becomes of interest. Previous models of transcription in its entirety included a simple termination mechanism that can be likened to the allosteric model [22–24]. An analytic approach by Zhu and Roussel [23], revisited by Roussel in a later work [24], used Markov chains to model a simple prokaryotic version of transcription. In it, termination was a single reaction at the end of a long chain of independent reactions. The thesis written by Vashishtha in 2011 used the Gillespie algorithm

to model the more complex eukaryotic transcription; however, termination was represented the same way as in the prokaryotic works [22]. These stochastic models depend on every reaction in the system modeled having to happen in a particular order, where there is only one possible sequence of events. Greive *et al.* [25] built a general transcription model that can apply to both eukaryotic and prokaryotic transcription. Generalized pause and termination events were defined in such a way that allows them to be inserted into the chemical master equation describing a model of transcription as required. This lets them mimic any specific gene of interest. Solutions depicting the polymerase position on the transcript as a function of time were numerically obtained with an ODE solver. In their model, the observation that termination events are associated with pauses is used to build a two step approach to termination: entry into a reversible pause state at a specified nucleotide in the sequence and then entry into an irreversible pause event. This definition of the termination mechanism, while allowing for the possibility of read-through, relates most to allosteric termination.

To my knowledge, the chase-down aspect of RNAPII by Xrn2 in the torpedo mechanism has not yet been modeled and is mathematically different from the mechanisms implemented in these previous works. The total number of reactions required for completion of the mechanism is not constant, and the kinetic competition between Xrn2 and RNAPII means that there is no specific order in which the reactions must occur. This thesis is going to address modeling the torpedo mechanism in two separate ways, Gillespie algorithms and random walks. The Gillespie algorithm is a stochastic algorithm which uses a defined set of reactions to alter the reactant populations of a system one reaction at a time [26]. When the populations of reactants in a system are available in low quantities, Gillespie simulations help capture how the system behaves in response to noise in the population levels. Random walks are the mathematical problem of a "walker's" movement through space being defined through the assignment of probabilities for moving in different directions in space [27]. Random walks can be multidimensional and in discrete or continuous time and space. The random walk is a commonly discussed problem, and a variety of questions can be asked

about random walks, such as the time of return to a position and first passage time problems. The first passage time problem can be used to represent the chase-down of RNAPII by Xrn2.

1.3 Outline of thesis

The details of the torpedo mechanism are explored in detail in Chapter 2 and two variants of the biochemical model are built with the intent of investigating what effect simplifying the nucleotide addition cycle and exonuclease degradation cycle has on the results obtained. Looking into this issue allows us to explore whether the more detailed model could be properly mimicked by a simpler one. A brief overview of literature that provides parameters for the model made can be found in Chapter 3. In Chapter 4, Gillespie simulations of the chase-down mechanism described in Chapter 2 are run for a variety of parameters, and the predicted success rates of termination, distribution of termination times, and the number of nucleotides synthesized by RNAPII during the chase-down are explored. Chapter 4 also contains the results of comparing the two model variants to each other. In Chapter 5, random walk theory is applied. Instead of taking DNA as the reference frame, the relative position of Xrn2 to RNAPII is counted instead, and reactions associated with the two proteins are assigned to the behaviour of a single random walker on a number line. The problem then becomes a first passage problem to a value that represents the Xrn2 closing in on RNAPII. Overall, Gillespie simulations and the first passage time equations of the random walk are used to understand how the polymerase and exonuclease speeds affect the torpedo model's success rate, amount of time required for termination, and the typical number of nucleotides transcribed (sometimes known as the termination region length [28]). This effort is made in the interests of illuminating the characteristics of termination which would be predicted by the present understanding of the torpedo mechanism.

Chapter 2

The Biochemical Model

The torpedo mechanism needs to be simplified into a few reactions for the purpose of biochemical modeling. These reactions need to express that RNAPII is transcribing RNA in the elongation complex throughout the process of 3'-end cleavage, exonuclease attachment and the subsequent degradation. Finally, the conditions under which termination occurs also need definition. This section aims to clearly define the characteristics of these events and to make simplifications to produce a comprehensive biochemical model.

The elongation complex works within a 10-20 nt opened part of the double stranded DNA called the transcription bubble [2]. Within it, there is an approximately 8 base-pairs (bp) long DNA-RNA hybrid [3]. Destabilizing this hybrid is thought to induce termination [3, 4]. At the front of this hybrid is the polymerase's active site where the nucleotide addition cycle (NAC) occurs. Correct NTP incorporation is largely guided by the trigger loop of RNAPII, a structure that is functionally conserved across most RNA polymerases [1]. NTP diffuses into the open frame of the transcription bubble. Typically, if the NTP is the proper conjugate base to the DNA base in the polymerase active site, the trigger loop assists to position the NTP for bond formation with the transcript [1, 29, 30]. If the NTP that was in the open frame does not match the DNA template, the trigger loop sterically clears it from the active site to allow another NTP in [1, 30]. Translocation of the polymerase can happen simultaneously with trigger loop NTP selection and is not a rate limiting step, but both translocation and NTP selection need to occur before the final bond formation of NTP with the 3' end of the RNA through pyrophosphate release can occur [1].

A rough first biochemical elongation model disregards the detailed dynamics of the NAC and simplifies the elongation process to a single reaction per nucleotide. This model is used in the one-step variant of the mechanism:

$$\text{RNAPII}_i \xrightarrow{p} \text{RNAPII}_{i+1}, \tag{2.1}$$

where i represents the location of the polymerase active site on the DNA transcript in nucleotides.

A more complex biochemical representation of elongation by RNAPII is a two-step process. Previous works on transcription models [22, 23] used a two-step elongation model composed of an activated and an occupied state. NTP positioning is the alignment of incoming cognate NTP to the RNA in the active site in preparation for bond formation and includes non-productive nucleotide addition cycles due to the diffusion of non-cognate NTP into the active site [1]. Instead of counting non-cognate NTP positioning events, we count only successful NTP positioning. As NTP diffuses into the active site, the polymerase moves back and forth between two nucleotide positions in a reversible translocation reaction. The translocation step can occur before or after NTP positioning and is not rate-limiting [1]; therefore we combined the two into one reaction. When NTP is aligned with the transcript and RNAPII is in a forward translocated position, bond formation between the NTP and RNA follows, which stops RNAPII from translocating backwards. Two states of RNAPII are defined. State one (P1) describes a polymerase which has NTP properly positioned and is in a forward translocated state, such that the bond formation of NTP with the RNA is ready to occur. P1 catalyzes the NTP bond formation and releases pyrophosphate, solidifying the polymerase's position on the DNA template. State two (P2) describes the polymerase as being available for NTP positioning and translocation. P2 undergoes a reaction which describes the successful positioning of NTP and translocation to the next

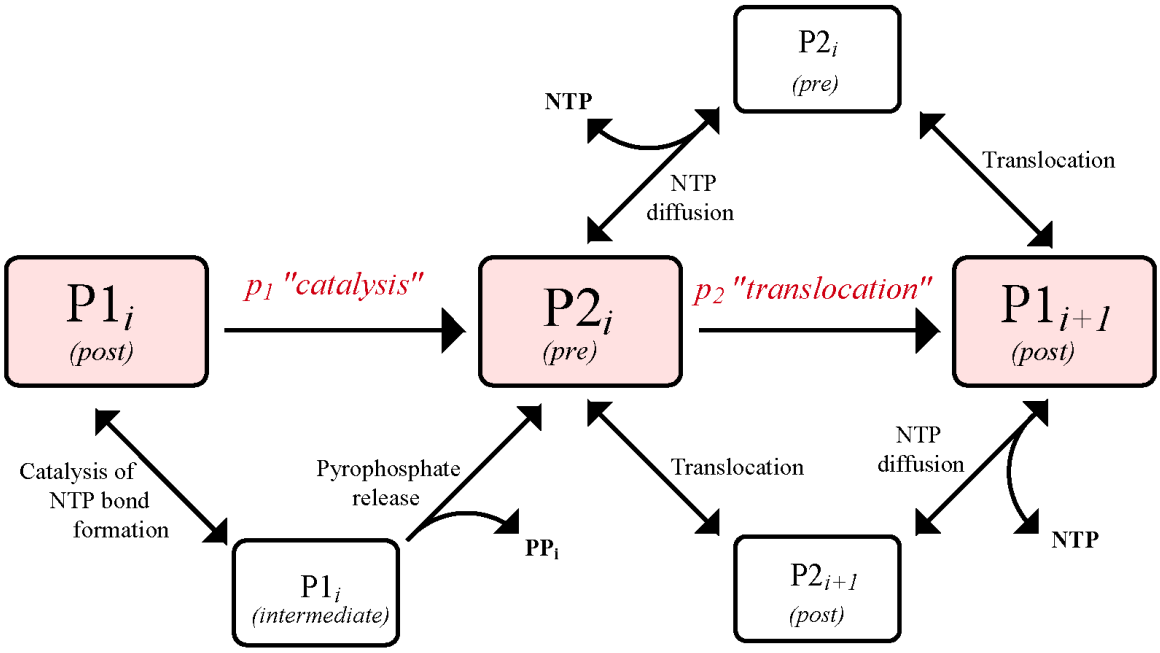


Figure 2.1: Adaptation of Figure 2C from Larson *et al.* [1] to highlight the simplifications made to the nucleotide addition cycle. Pink boxes represent the states of the two-step variant of the model, where reactions in red are our simplified reaction rates. White boxes contain the true intermediates of the reactions, and black-text reactions are those predicted by Larson *et al.* Note that the reversible reactions of translocation or NTP placement through diffusion can happen in two different orders. Traditional pre- and post-translocation states are labeled in the boxes.

base pair on the DNA to return to the P1 state.

$$\mathrm{P1}_i \xrightarrow{p_1} \mathrm{P2}_i \tag{2.2}$$

$$\mathrm{P2}_i \xrightarrow{p_2} \mathrm{P1}_{i+1} \tag{2.3}$$

Overall, p_1 represents the rate of NTP bond formation and pyrophosphate release (the catalyzed reaction), while p_2 describes the rate of NTP positioning and translocation (the translocation reaction). The simplifications made to the NAC can be found in Figure 2.1. Once the polymerase has transcribed the poly(A) signal (PAS), it becomes available to the 3'-end processing machinery. Of particular interest to the torpedo mechanism is the means by which the PAS signals endonucleolytic cleavage of the RNA in order to make a

5' monophosphate on a nucleic acid available for exonuclease Xrn2 recognition. The number of proteins interacting with RNAPII and the RNA during the closely coupled 3'-end processing events of splicing, cleavage and polyadenylation is extensive [31]. Sometimes coined pre-termination cleavage [21], the endonucleolytic event can be induced 10–35 nt downstream of the PAS [2]. Sequence homologs of the RNA cleavage machinery in the eukaryotic kingdom are not always functionally equivalent, making the identification of the endonuclease responsible for cleaving the pre-mRNA from the 3' flanking region difficult [31]. Additionally, co-transcriptional cleavage elements (CoTC) downstream of the PAS exist where self-cleaving activity of the RNA has been previously observed, like in human β-globin [17, 18, 32]. In regards to the present biochemical model, wherever and however cleavage is determined, the initial conditions of the system are set such that the first nucleotide in the 3'-flanking region that is exposed by pre-termination cleavage is $j_0 = 0$, and that this location is already available for cleavage at time $t = 0$. An additional parameter α is set to describe how far behind the polymerase active site this cleavage site needs to be in order to be sterically accessible by the cleavage machinery. The steric bulk upstream of the polymerase is estimated to be 17 nt long [33], assuming that the minimum distance between Xrn2 and RNAPII active sites measures the steric bulk of the polymerase in general. This quantity, denoted q, defines the distance Xrn2's active site is behind RNAPII's active site when termination occurs. α is estimated to be the same as q, with the acknowledgement that it may need to change when more information on pretermination cleavage, the machinery involved, and the primary mechanisms used becomes available. For now, the cleavage mechanism is represented by a simplified reaction, where PAS is the poly(A) signal and ERNA_0 is the exposed 3'-flanking region of the cleaved RNA. Cleavage can only occur when RNAPII is at location $i = \alpha$ or beyond.

$$\mathrm{PAS} + \mathrm{RNAPII}_{i\geq\alpha} \xrightarrow{a_1} \mathrm{ERNA}_0 + \mathrm{RNAPII}_{i\geq\alpha} \tag{2.4}$$

Since the PAS and RNAPII are part of a single transcription complex, this reaction

behaves like an intramolecular reaction and is therefore a first order reaction. After the endonucleolytic event, the enzyme Rtt103 associates with both the CTD of the polymerase when Ser2 is phosphorylated and Xrn2 (Rat1) [10, 12, 34]. Hypothetically, this allows the exonuclease Xrn2 to also be readily available for association to the 5' end of the cleaved RNA. Let $X1_0$ represent Xrn2 attached to the 5' flanking region in state one, defined below, at nucleotide position 0 (the point of cleavage).

$$\mathrm{ERNA}_0 \xrightarrow{a_2} \mathrm{X1}_0 \tag{2.5}$$

Xrn2 can now begin the degradation of the RNA being produced from RNAPII. The amino acid residues of the active site of the XRN nuclease family are highly conserved; as a result, while studies to confirm the similarity in catalytic mechanism of Xrn1 and Xrn2 are sparse, it is generally suspected that the mechanism is the same across the family [35]. The XRN family nucleases have a structural pocket in common which restricts target RNA to those having exposed 5' monophosphates [36, 37]. Xrn2's cytoplasmic homolog Xrn1 requires the RNA target to have at least 4 nucleotides available for the protein to associate with the RNA [36]. *S. cerevisiae* Xrn2 was shown to be able to degrade an exposed RNA transcript that was 21 nt long from RNAPII's active site [33]. Since steric interference between the two proteins occurs 17 to 18 nt behind the active site of the polymerase, we can put together that for Xrn2, a 4 to 5 nt lag behind the steric bulk is enough to associate with the RNA. This puts an upper limit on the minimum amount of lag behind the polymerase required for Xrn2 association and is in accordance with Xrn1 data.

The XRN family's processive degradation requires two steps per nucleotide: the hydrolysis of the target phosphate group on the 5' nucleic acid base, followed by a translocation step when the nucleotide is released [38]. One-step and two-step variants of degradation of RNA by Xrn2 are constructed. The one-step variant is

$$\mathrm{Xrn2}_j \xrightarrow{v} \mathrm{Xrn2}_{j+1} \tag{2.6}$$

The two-step variant's state one, X1, represents the exonuclease having been properly positioned for the hydrolysis step and state two, X2, represents an exonuclease that has cleaved the RNA substrate and is ready to move to the next nucleotide.

$$\mathrm{X1}_j \xrightarrow{v_1} \mathrm{X2}_j \tag{2.7}$$

$$\mathrm{X2}_j \xrightarrow{v_2} \mathrm{X1}_{j+1} \tag{2.8}$$

It has been hypothesized that the tower domain of Xrn2 interacts with RNAPII to induce termination [37], but the specific mechanism by which Xrn2 disengages RNAPII is unknown [7, 14]. Rai1 is a termination factor which binds to a complementary surface of Xrn2 and is responsible for converting 5' triphosphate RNA ends into 5' monophosphate RNA ends [37]. In one study, it was found that Xrn2, the Xrn2/Rai1 and Xrn2/Rai1/Rtt103 complexes were unable to disengage an elongation capable RNAPII which was paused through not providing NTP *in vitro* [39]. A follow-up study found that the Xrn2/Rai1 complex was capable of inducing termination when NTP units that did not match the DNA template were provided and misincorporation events occurred [33]. In this experiment, it was both demonstrated that the misincorporation successfully paused the RNAPII, interrupting read-through when NTP was added in further stages of the experiment, and that adding non-canonical NTP during Xrn2 addition resulted in termination of RNAPII. However, true NTP misincorporation events are extremely rare, happening on the order of 10^{-3} to 10^{-5} times as often as normal incorporation events [30]. This reduces the likelihood that a regular event such as the termination mechanism would rely on such an event. These results make it likely that at least some sort of more frequent pause event which changes the conformation of the DNA-RNA hybrid is required for termination to be induced by Rat1; however, a pause event wasn't incorporated in our model. Instead, the only condition for termination was set such that Xrn2 degrades the nucleotide 17 bases behind RNAPII's active site, and then the exonuclease pushes itself into the polymerase and destabilizes the

complex (one- and two-step models):

$$\mathrm{Xrn2}_{i-18} \xrightarrow{v} \mathrm{Xrn2}_{i-17} + \mathrm{RNAPII} + \mathrm{DNA} \tag{2.9}$$

$$\mathrm{X2}_{i-18} \xrightarrow{v_2} \mathrm{X1}_{i-17} + \mathrm{RNAPII} + \mathrm{DNA} \tag{2.10}$$

The final step of the termination mechanism results in RNAPII being freed from the DNA and becoming available for another transcription event, whilst Xrn2 degrades the remaining RNA fragment as indicated from results of successful termination events in [33].

As an alternative to a successful termination event, it is possible that Xrn2 never catches up to the polymerase, and RNAPII runs through to transcribe the next gene in the DNA. An intergenic region size (IRS) is defined. A failed termination, or read-through, event is defined as RNAPII transcribing to the end of the IRS for the one-step and two step model variants, respectively:

$$\mathrm{RNAPII}_{\mathrm{IRS}-1} \xrightarrow{p} \mathrm{RNAPII}_{rt} \tag{2.11}$$

$$\mathrm{P2}_{\mathrm{IRS}-1} \xrightarrow{p_1} \mathrm{P1}_{rt} \tag{2.12}$$

RNAPII_{rt} and $\mathrm{P1}_{rt}$ represent the polymerase which has read through the intergenic region.

Chapter 3

Parameters of the Model and Relevant Data

Biologically valid ranges for some parameters were found in the literature. The following chapter outlines the nature and sources of those parameters.

Larson *et al.* found by surveying the literature that transcription by RNAPII during the elongation phase of transcription occurs at a rate of 10–70 nt/s (nucleotides per second) [1]. They measured the kinetics of the nucleotide addition cycle by building a single-molecule assay that allowed them to apply a physical force on the transcribing polymerase. Based on these experiments they measured transcription rates ranging from 20–60 nt/s depending on the force applied, concentration of NTP, and concentration of ammonium ion[1] for polymerases in the elongation phase. In comparison, Fong *et al.* [28] used WT RNAPII and two mutants with known transcription rates from elongation phase experiments. WT RNAPII typically synthesizes RNA at rate of 28 nt/s, whereas the mutants R749H and E1126G have rates of 8 nt/s and 32 nt/s respectively. While these mutants were used in the study to investigate the effects of RNAPII speeds on polymerase occupancy in the termination region (which provided supportive evidence for the torpedo model), their transcription rates within the termination region were not measured.

Recalling from Chapter 2, the nucleotide addition cycle of RNAPII for our two-step model was defined such that the bond formation between NTP and the RNA with pyrophosphate release was called catalysis by RNAPII. Catalysis changes the polymerase's two-step

[1] *In vitro* elongation rates tend to be slower than *in vivo* rates; ammonium ion increases *in vitro* rates to comparable levels [1].

state from P1 to P2, and in our model the catalysis reaction solidifies the polymerase's positioning on the DNA, preventing backwards translocation to the original polymerase position. Successful NTP positioning and translocation were grouped into one reaction. For convenience this reaction is called the translocation reaction, taking RNAPII from P2 to P1 and advancing the polymerase one nucleotide. In previous simulation work on the entire transcription process [22], there was no data for the detailed kinetics of RNAPII elongation. Vashishtha assumed that his activation (catalysis) and translocation reactions would have an equal share of the overall transcription rate, so he evenly divided the rate of 72 nt/s to obtain activation and translocation rates of 144 nt/s each. In the work by Larson *et al.*, they calculated a rate of catalysis with and without ammonia as $77 \pm 3s^{-1}$ and $34 \pm 2s^{-1}$ [1]. Let those values be p_1. We can therefore estimate rates of our translocation reaction using

$$\frac{1}{p} = \frac{1}{p_1} + \frac{1}{p_2}. \tag{3.1}$$

When $p = 55$ nt/s and $p_1 = 77$ nt/s as calculated from the Larson *et al.* experiments at 1 mM NTP, 100mM NH_4^+ [1], then $p_2 = 193$ nt/s. If we use their data from elongating RNAPII at 1 mM NTP with no ammonium instead, where $p = 23$ nt/s and $p_1 = 34$ nt/s, then $p_2 = 71$ nt/s. Either set of values can be used to calculate $p/p_1 \approx 0.7$. This ratio is the average percentage of time spent waiting for catalysis to occur: RNAPII is in state P1 70% of the time.

Any data that was found concerning the average Xrn2 degradation rate was mixed with Xrn2 recruitment rates. In [28], the effect of recruitment rate was indirectly explored through two cell lines. In the first, WT Xrn2 was used and the RNAPII occupancy of the termination region of a multitude of genes was measured. In the second, exonucleolytically dead Xrn2 competed with wildtype Xrn2 for the cleaved RNA. RNAPII was found to terminate further downstream for this WT/mutant strain. While it demonstrates the competition between WT and mutant Xrn2 to be recruited to the 5' end of the transcript, a multitude of other reactions which contribute to RNAPII's extended occupancy down the

termination region are unaccounted for. Some of these reactions are the disassociation rate of Xrn2 as well as the additional rounds of the NAC RNAPII partakes in. These additional reactions negate the chances of calculating an Xrn2 recruitment rate. Meanwhile, Kaneko *et al.* attempted to kinetically describe Xrn2 in accordance to the availability of protein associated splicing factor (PSF) in HeLa cells [19]. They use RNAPII termination locations on the gene as a measure of this reaction; therefore, Xrn2 recruitment rate a_2 is not separated from the degradation reaction v in the one-step variant or v_1 and v_2 in the two-step variant. In fact, the entire chase-down mechanism is included in the resultant times. The torpedo model does not have a set number of reactions with regards to RNAPII's NAC and Xrn2's number of nucleotides to be excised. Additionally, Xrn2 recruitment rates are not separated from PSF cleavage rates in the experiment, making it difficult to glean any meaningful parameters for Xrn2.

The next parameter is the intergenic region size (IRS), which determines how much space RNAPII has before it interferes with downstream genes. The human median intergenic region size was suggested to be 3949 base pairs long, down from previous estimates of 14 000 base pairs [40]. As the discovery of novel genes continues, it is expected that intergenic regions are currently overestimated [40]. Meanwhile, the intergenic regions of *S. cerevisiae* average 536 bp, though there were measurement limitations due to the experiment's requirement of a 300 bp minimum length for detection [41]. Therefore, we suppose that biologically valid intergenic region lengths may range anywhere from 100 to 10 000 base pairs for this study.

Finally, there exists some data related to the number of nucleotides synthesized by RNAPII before termination occurs. RNAPII is reported to travel more than 2000 kilobases after the PAS before termination occurs for human mRNA genes [8]. ChIP experiments can show the occupancy of RNAPII on the DNA after the PAS to give an idea of where in the intergenic region termination typically occurs. RNAPII populations rise after the PAS and tend to peak around 2000 bp downstream, then drop off to background noise around 8000

bp [28]. The rise in RNAPII populations is attributed to induced pausing and slowing of the polymerase, and the drop off of RNAPII populations to termination events [28]. This suggests that termination is most likely to occur 2000–8000 bp downstream of the PAS. While termination region data does not explicitly go into the simulation work, it can help determine whether the results are reasonable.

Chapter 4

Comparison of Simulation Results for the One- and Two-Step Chase-Down Variants

Stochastic simulations were produced for the one-step and two-step model variants via the method outlined by Gillespie [26] for the biochemical reaction sets described in Chapter 2. All simulations were performed using MATLAB R2016b [42]. Performance differences for termination success rate, distribution of termination times, and waste RNA production were investigated for just the chase-down process. While it is supposed that the two-step model is a more accurate depiction of the true system, the purpose of comparing the model simulations is to check the feasibility of simplifying the overall torpedo model to the one-step variant. The distribution of the chase-down time and the overall number of nucleotides synthesized by RNAPII for each model variant is investigated as a function of the ratio of polymerase synthesis rates to Xrn2 degradation rates (p:v) as described in Section 4.1.

4.1 Simulation set up

This section is meant to establish some choices that were made in terms of parameter exploration and to introduce the language used to talk about certain simulation conditions.

Searches through the research literature did not produce kinetic data for Xrn2. This includes the reaction rates of hydrolysis of the nucleotide by Xrn2 and the translocation of Xrn2 to the next nucleotide. An overall RNA degradation rate by the exonuclease was not obtained. The rate of Xrn2 recruitment to the 5' end of RNA wasn't found either.

As such, the rates of RNAPII and Xrn2 are talked about in relation to each other in this work. RNAPII's rate of reaction is highly subject to modification according to the situations arising *in vivo*. RNAPII accumulation after the PAS is typically attributed to polymerase pausing, and is observed with higher frequency in short genes, when RNAPII navigates histone rich regions, and in genes that had higher transcription speeds during the elongation phase [43]. In comparison, it is not clear if the degradation rate of Xrn2 is altered *in vivo*. The Xrn protein family share a structural pocket that the exonucleolytic active site resides in [36], and it is believed that Xrn2 handles RNA secondary structure with the assistence of Rai1 [37]. No information was found on whether secondary structures in the RNA slow Xrn2 degradation rates. The approach chosen was to set the overall degradation rate v of Xrn2 as constant and arbitrarily 1, and the overall rate of RNAPII nucleotide addition p is then relative to the speed of Xrn2. Therefore, all results reported as p:v ought to be taken as events due to relative speeds between the two proteins.

The simulations make comparisons between one-step and two-step variant models. The average RNAPII and Xrn2 rates between the two model variants is kept constant. For every comparison between one-step and two-step simulations, the following relations describing reaction rates are always kept:

$$\frac{1}{p} = \frac{1}{p_1} + \frac{1}{p_2} \tag{4.1}$$

$$\frac{1}{v} = \frac{1}{v_1} + \frac{1}{v_2} \tag{4.2}$$

The following paragraph intends to address the language used to talk about the catalysis/activation and translocation rates of RNAPII/Xrn2, henceforth referred to as the two-step rates. For both Xrn2 and RNAPII, the two-step rates can be set such that there is one very fast and another comparatively slow reaction, which is going to be called having asymmetric two-step rates. For example, if the value p/p_1 is very small, it means that p_1 is large and is a fast reaction, and the polymerase in state P1 is quickly converted to P2. If p/p_1 is

very large, it means that p_1 is a slow reaction and the conversion of reactant into product takes more time. Both of these are examples of having asymmetric two-step rates. The symmetric two-step rate selection for RNAPII is then $p_1 = p_2$. As previously described in Chapter 3, the average amount of time the polymerase or exonuclease spends in a particular state of the two-step model (P1 and P2 for RNAPII, X1 and X2 for Xrn2) is relatable to the rate of the reaction that governs the enzyme's exit from that state. The value p/p_1 can be interpreted as the proportion of time the polymerase spends in state P1 waiting for catalysis to occur. v/v_1 represents the proportion of time Xrn2 spends in state X1 waiting for the hydrolysis step that excises a nucleotide from the RNA transcript.

During the initial exploratory simulation work, it was noticed that symmetric two-step rate selection tended to yield the largest differences in results between the one and two-step variants. Thus many comparisons of one- and two-step simulations are done using $p_1 = p_2$ and $v_1 = v_2$ as the two-step variant rates. This trend is discussed in Section 4.5. A variety of other parameters are kept constant during the simulations. The number of nucleotides Xrn2 must catch up is determined by the initial positions of RNAPII (i_0) and Xrn2 (j_0). Termination activity has been shown to be possible by Xrn2 when the polymerase is given a 20–40 nt head start *in vitro*, and the closest Xrn2 can get to RNAPII before steric interference between the two proteins occurs is 17/18 nt behind RNAPII's active site [33]. That study attempted to discern the nature of the interaction between RNAPII and Xrn2 using a 31 nt long RNA sequence for the majority of tests, making this value desirable for potential experimental comparison; therefore, $j_0 = 0$ and $i_0 = 31$ throughout this study. For the model, it is assumed that steric interference between Xrn2 and RNAPII is the mechanism of termination, as suggested might be the case in previous publications [33, 37, 44]. The steric interference event occurs upon Xrn2's translocation step in the model, where the exonuclease's movement closes the gap between the two proteins. The distance between the active sites at the moment of steric interference is chosen to be $q = 17$ nt based on [33]. For the two-step model, RNAPII and Xrn2 begin in states P1 and X1 respectively, maximizing

Table 4.1: Default values for one- and two-step variant simulations

Parameter	Description	Value	Units
j_0	Initial position of Xrn2	0	nucleotide position
i_0	Initial position of RNAPII	31	nucleotide position
q	Distance Xrn2 is behind RNAPII's active site when termination occurs	17	nucleotides
p	Rate of RNAPII	varied	arbitrary units
v	Rate of Xrn2	1	arbitrary units
IRS	Intergenic region size	5000	nucleotides

the time for either to undergo translocation for the first time in the simulation. This was chosen in order to try and match the one-step model, where translocation is the observable product of the NAC and exonucleolytic activity. An investigation revealed that the state the two enzymes start in can affect results in simulations with asymmetrical two-step cases, which will be discussed in Section 4.7. Finally, unless otherwise stated, a mid-range intergenic region size of 5000 bp is used in the simulations. 10 000 simulations are run for every parameter set. The parameters $i_0 = 31$, $j_0 = 0$, $q = 17$, IRS $= 5000$ and sample size 10 000 will be referred to as the default conditions, summarized in Table 4.1.

4.2 Success rates of termination

In every simulation, there exist two possible outcomes: either the exonuclease catches up to RNAPII and termination occurs, or termination is unsuccessful and the polymerase interferes with the gene located downstream. Both outcomes are of interest due to their potential roles in gene expression. With the following simulations, we gain an understanding of the conditions required to guarantee termination via the torpedo chase-down mechanism as well as the conditions that may favor multiple genes being transcribed at a time instead of termination taking place.

The one- and two-step variants of the chase-down mechanism are simulated for the values of $p{:}v = [0.95, 1.00, 1.05]$ in Figure 4.1. In the two-step variant simulations, the rates

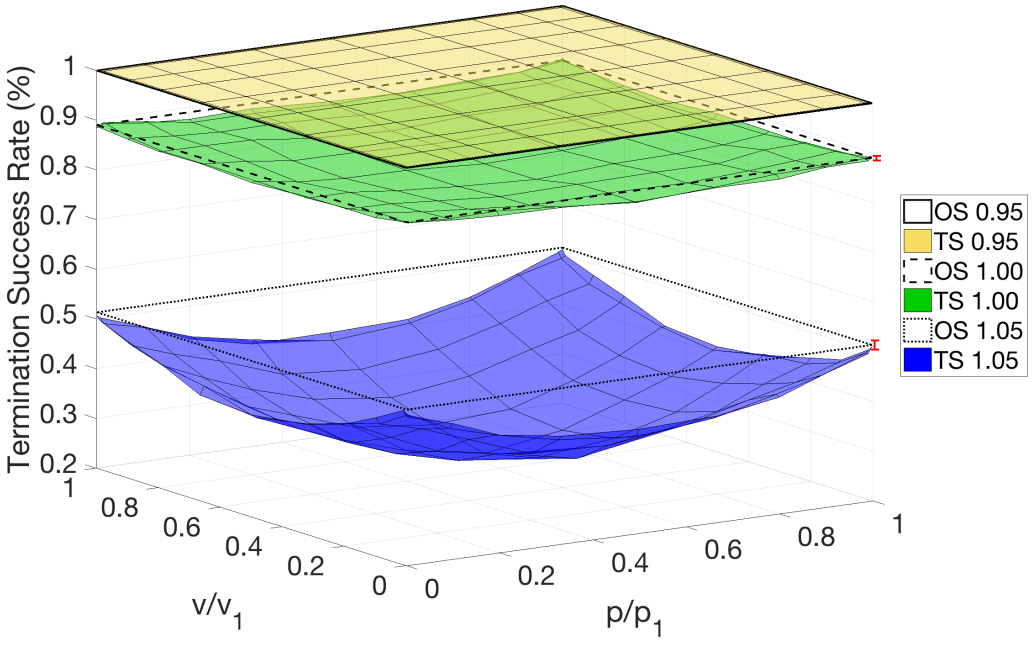


Figure 4.1: A comparison of the success rate of the one-step (OS) and two-step (TS) chase-down variants for default simulation conditions. Three p:v ratios are depicted for each variant. The OS variants do not depend on p_1, p_2, v_1 or v_2, and are depicted as bold outlines. The change in success rate of the TS variant can be seen as a function of p/p_1 and v/v_1 and are shown as surfaces. For each surface, the typical range of sample points predicted by the binomial distribution are given to the right as red error bars $[\hat{x}-2\hat{\sigma}/N, \hat{x}+2\hat{\sigma}/N]$ when it is large enough to visualize.

governing the change from state to state of the polymerase, p_1 and p_2, and exonuclease, v_1 and v_2, are varied as described within Section 4.1. Success rate was calculated by dividing the number of successful trials by the total number of trials. The three surfaces together show that the success rate is sustained at fairly high rates until $p = v$ at a 'mid-range' intergenic region size of 5000 nucleotides (nt). Success rates drop very quickly after polymerase speed exceeds exonuclease speed, and it can be seen that success rate is sensitive to p/p_1 and v/v_1. The surfaces for the two-step variant success rate appear to approach the calculated success rate of the one-step variant as the two-step rates of Xrn2 and RNAPII become increasingly asymmetric. This is a sensible result: if one of the two-step rates becomes fast enough, reaction time may be so short as to be negligible in comparison to time it takes to complete the other reaction. However, the success rate of the two-step variant sometimes exceeded that of the calculated success rate of the one-step variant at rates with the highest asymmetry, particularly when $p{:}v = 1$. For example, when $p{:}v = 1$ and the two-step rates chosen such that $\frac{p}{p_1} = \frac{v}{v_1} = 0.9901$, then the two-step simulations calculated a success rate of 0.8947, whereas the one-step model for $p{:}v = 1$ calculated the success rate to be 0.8902. If it is true that the two-step model's success rates have an upper bound described by the one-step variant we have to explain why this happened. As stated beforehand, each trial representing the outcome of a single Gillespie simulation has only two outcomes: transcription termination success or failure. Each trial, starting with the same parameters and same code, are independent of each other and have the same probability of success or failure. Thus, each simulation is a Bernoulli trial with success probability x and failure probability $(1-x) = y$. From there we can identify that running $N = 10\,000$ simulations and observing the number of successful termination events equates to obtaining a sample from a binomial distribution. The mean number of successful trials of a binomial distribution is $\mu = Nx$. Assuming that the probability of success calculated from these simulations is accurate, we estimate the success probability as $\hat{x}$. An estimate of the standard deviation $\hat{\sigma}$ of the binomial distribution at each point on the two-step variant surface was estimated

through using the standard deviation of a binomial distribution $\sigma = (Nxy)^{1/2}$. An error bar for each surface in Figure 4.1 depicting a typical $[\hat{x} - 2\hat{\sigma}/N, \hat{x} + 2\hat{\sigma}/N]$ for that surface was drawn in order to demonstrate that small variations in sampling can be explained by the natural variance of binomial distributions.

It is observed from Figure 4.1 that the differences in termination success rates between the one- and two-step variants of the model are the most extreme when $p_1 = p_2$ and $v_1 = v_2$. Therefore under this parameter set, the dependence of the success rate on the ratio between p and v is calculated at an intergenic region size of 5000 nt and the results are compared in Figure 4.2. In general, both variants of the chase-down model have high success rates until $p{:}v = 0.95$, and both have nearly 0% success rate by the time $p{:}v = 1.5$. For $p{:}v \in (0.95, 1.5)$, the two-step variant of the model predicts lower success rates, dropping quickly after $p{:}v = 1$. It is possible that when $p{:}v > 1$, termination relies on the unlikely event that the polymerase progresses very little despite its higher rate of reaction. This unusual amount of time spent at a given nucleotide has no underlying mechanism assigned to it other than the random fluctuations in the system. The exonuclease advances more rapidly due to random chance and termination occurs. The involvement of slow rounds of nucleotide addition by RNAPII in successful termination events when $p{:}v > 1$ is touched on again in Section 4.6.

Finally, the termination success rate was calculated for a range of intergenic region sizes and $p{:}v$ values for the one-step variant. It can be seen in Figure 4.3 that success rates drop faster for genes with small intergenic regions, as could be expected. However, it can be seen that the difference in the success rate between 100 nt and 1000 nt is much larger than for the 1000 nt and 10 000 nt cases, suggesting that increasing the intergenic region size can only improve the success rate so much for a given $p{:}v$ ratio. Success rate is always near zero by the time $p{:}v = 1.5$. Intriguingly, these results show that the stochastic nature of the system has the ability to deliver an unlikely termination event at times even when the polymerase speed slightly exceeds that of the exonuclease, something that would have been

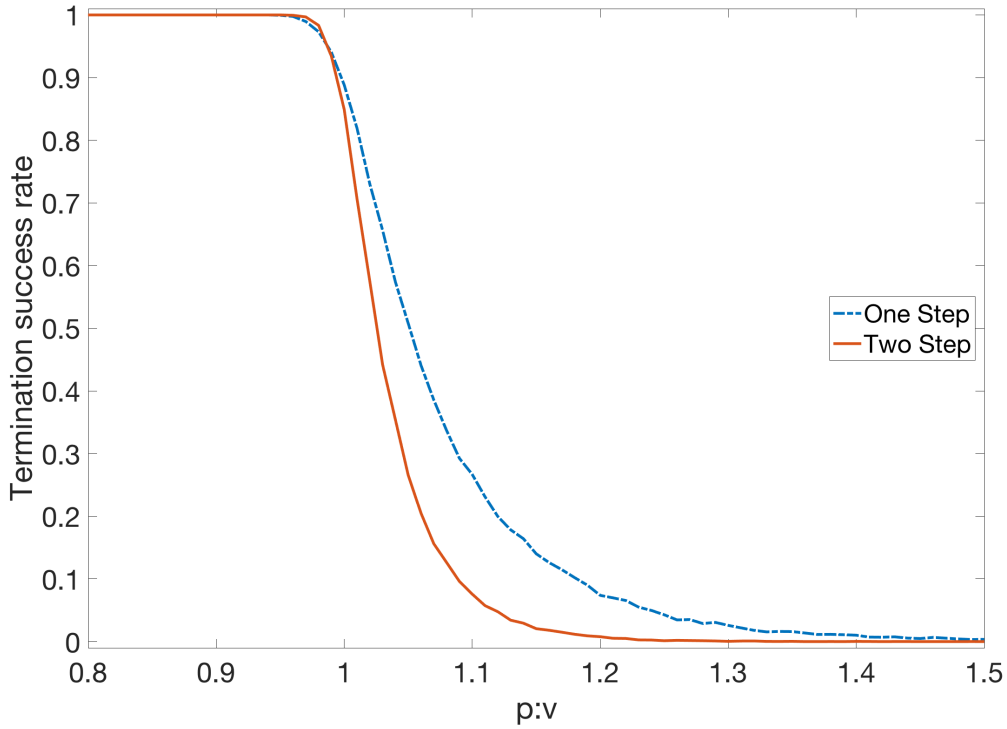


Figure 4.2: A comparison of the success rate of transcription termination of the one- and two-step chase-down variants as a function of the ratio between polymerase and exonuclease rates (p:v) under default conditions. Symmetrical two-step rates were chosen for the two-step variant.

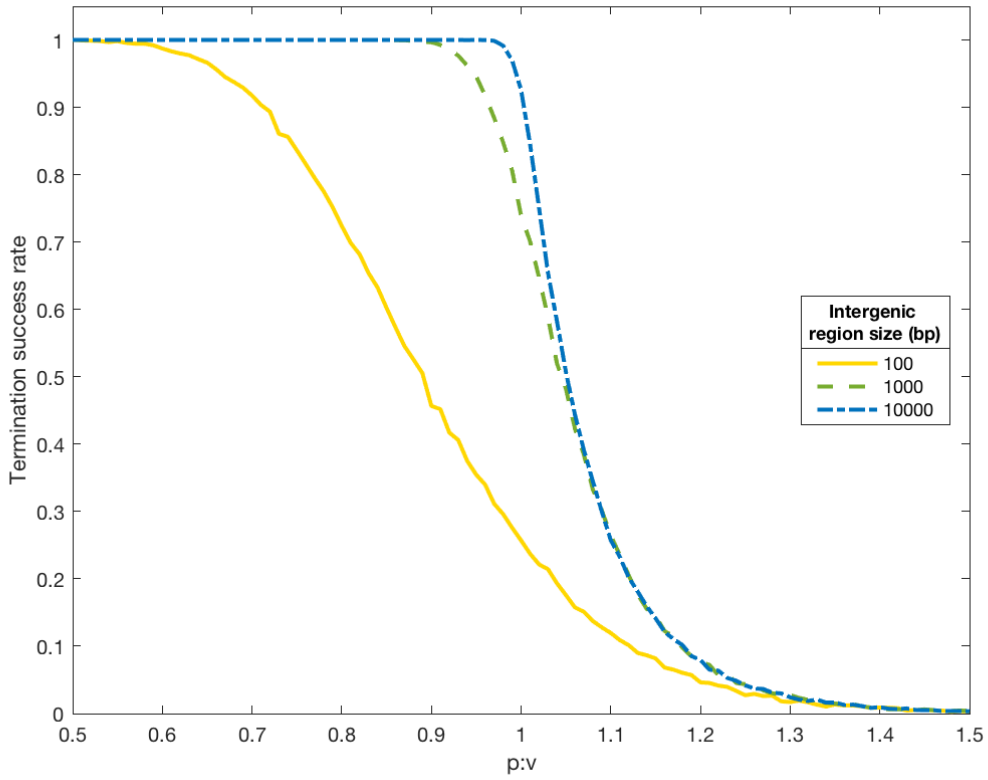


Figure 4.3: Termination success rate of the chase-down of RNAPII by Xrn2 in the one-step variant of the model with respect to relative RNAPII:Xrn2 speeds (p:v). Three different intergenic region sizes are used to demonstrate that the success rate begins to drop faster for low p:v values for small intergenic region sizes. With the exception of intergenic region size, simulations are run at default conditions.

an impossibility in a deterministic system.

4.3 Distribution of successful termination times

Now that some estimates of the probability of successful termination in the chase-down have been obtained, the amount of time this mechanism might take to terminate transcription is investigated. How the p:v ratio affects the distribution of successful termination times of each model variant is explored. Normalized distributions of termination times for the one-step and two-step variations of the chase-down model are displayed for a number of p:v ratios in Figure 4.4. Separate graphs were used to display the $p < v$ and $p > v$ cases for

clarity. All graphs use the $p = v$ (p:$v = 1$) case as a reference. All distributions in Figure 4.4 are unimodal distributions with positive skewness and show that both the one- and two-step variants are subject to similar changes. By comparing Figures 4.4a and 4.4b to each other for the one-step model, and Figures 4.4c and 4.4d for the two-step, we can see that the distribution becomes less sharply peaked as $p \to v$ from either $+$ or $-$ directions. The average successful termination time dropping again is nonintuitave and discussed in detail in Section 4.6. The tail of the distribution (amount of skew) is the longest when p:$v = 1$, and becomes shorter as the difference between the polymerase and exonuclease speeds becomes greater. The sharpness of the peaks becomes greater more quickly for $p > v$ values than for $p < v$ values. Additionally, although the normalized graphs do not show this explicitly, the sample sizes get smaller and smaller as p:v grows due to the decrease in termination success rates. Overall, there is a qualitative similarity of trends between the one- and two-step variants in response to changes in the p:v ratio.

Direct comparisons of the two variants of the chase-down mechanism can be seen for select p:v values in Figure 4.5. For each subfigure, a different p:v value is chosen and distributions are compared. For all p:v values investigated, the peak of the two-step distribution is both smaller and located at a greater time value than in the one-step distribution. Meanwhile, the tail of the two-step distribution is always shorter than that of the one-step distribution. This means that the two-step variant will predict that most successful termination events occur at a later time than the one-step variant, with the trade-off that the less frequent successful termination events with unusually long termination times usually happen in a shorter time frame in the two-step variant than in the one-step variant.

4.4 Correlations between nucleotide synthesis and termination time

The amount of excess RNA produced by RNAPII after the PAS is sometimes called the size of the termination zone [28]. To investigate predicted excess RNA lengths after the PAS of the two model variants, the number of nucleotides added to the RNA by RNAPII

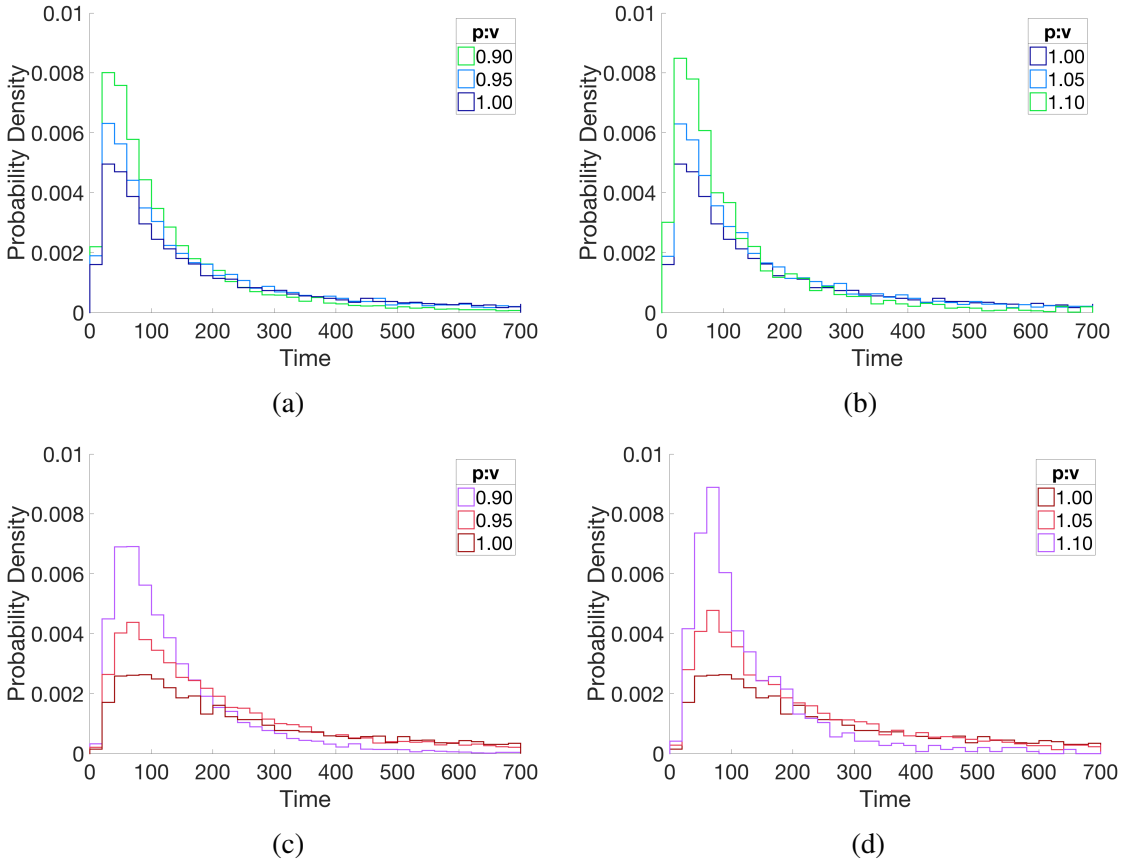


Figure 4.4: Estimates of the probability density function of successful termination time for (a) the one-step variant when $p \leq v$, (b) the one-step variant when $p \geq v$, (c) the two-step variant when $p \leq v$, and (d) the two-step variant when $p \geq v$,under default simulation conditions. Two-step reaction rates are set to symmetric values. For all distributions such that $p{:}v = 1$, the tail of the distribution extends to ≈ 5000 time units, and when $p \neq v$, the tail of the distribution typically ends before 5000 time units (not shown).

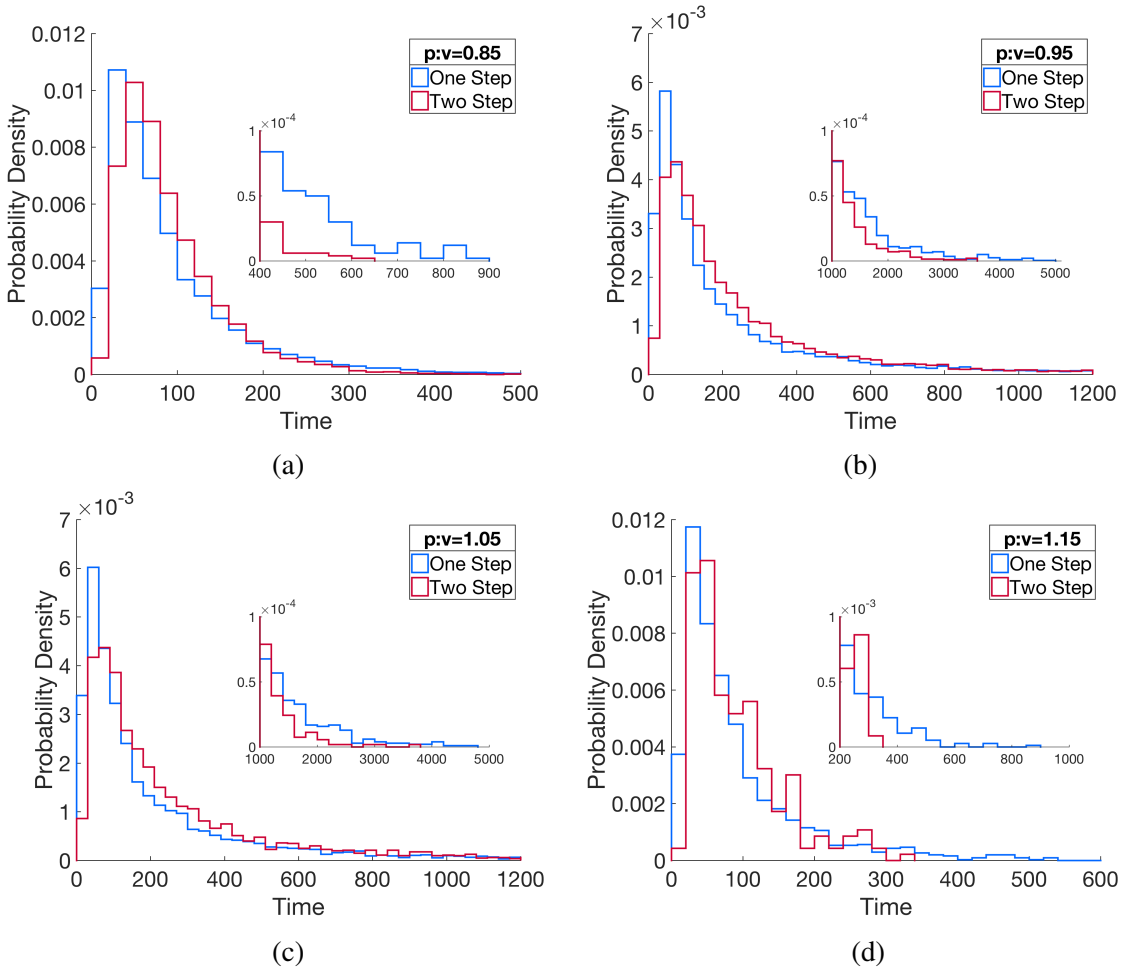


Figure 4.5: Estimates of the probability density function comparing the distributions of successful termination time for the one- and two-step chase-down variants for various p:v ratios under default simulation conditions. Two-step reaction rates of the two-step variants are set to symmetric. Insets depict the tails of each distribution.

for successful termination events was plotted in Figure 4.6 for various values of p:v. Additionally a scatter plot of time versus nucleotides added for successful termination events was created in Figure 4.7. The trends of the number of nucleotides added to the RNA product in Figure 4.6 closely resemble the time probability distributions found in Figure 4.4. Why the similarities exist becomes apparent when Figure 4.7 is viewed; a positive linear relationship between the time taken and the size of the termination zone is observed. This relationship was observed for both one- and two-step variants of the model and for all p:v ratios investigated. The simulation set up makes it so that Xrn2 or RNAPII progression are the only options for each time step taken until the termination step. Each successful termination event meets the condition of a select number of additional Xrn2 reactions occurring over and above those of the polymerase, and thus, steps by RNAPII are directly related to the amount of time needed to finish the simulation. Such a result is more of a confirmation of effective simulation set up than novelty; however, the acknowledgement of a linear relationship between the number of nucleotides synthesized by RNAPII and the amount of time termination took now gives us a tool to explain more complex phenomena later in Section 4.6.

4.5 Two-sample Kolmorogov-Smirnov tests on one- and two-step variants

Thus far, the qualitative differences between the one- and two-step variants of the chase-down model were investigated using the two-step symmetric conditions, meaning $p_1 = p_2$ and $v_1 = v_2$, where $p_1 = 2p$ and $v_1 = 2v$. This choice that was made based on preliminary results which showed that this arrangement indeed gave differences in the two distributions. The literature supports the two-step variant being a more accurate depiction of the true biological system (discussed in Chapter 2), but the one-step variant was included in hopes that a simplification could be made. How close to symmetric do the rates of the two-step model need to be before the differences in its distribution become ‘too much to simplify’

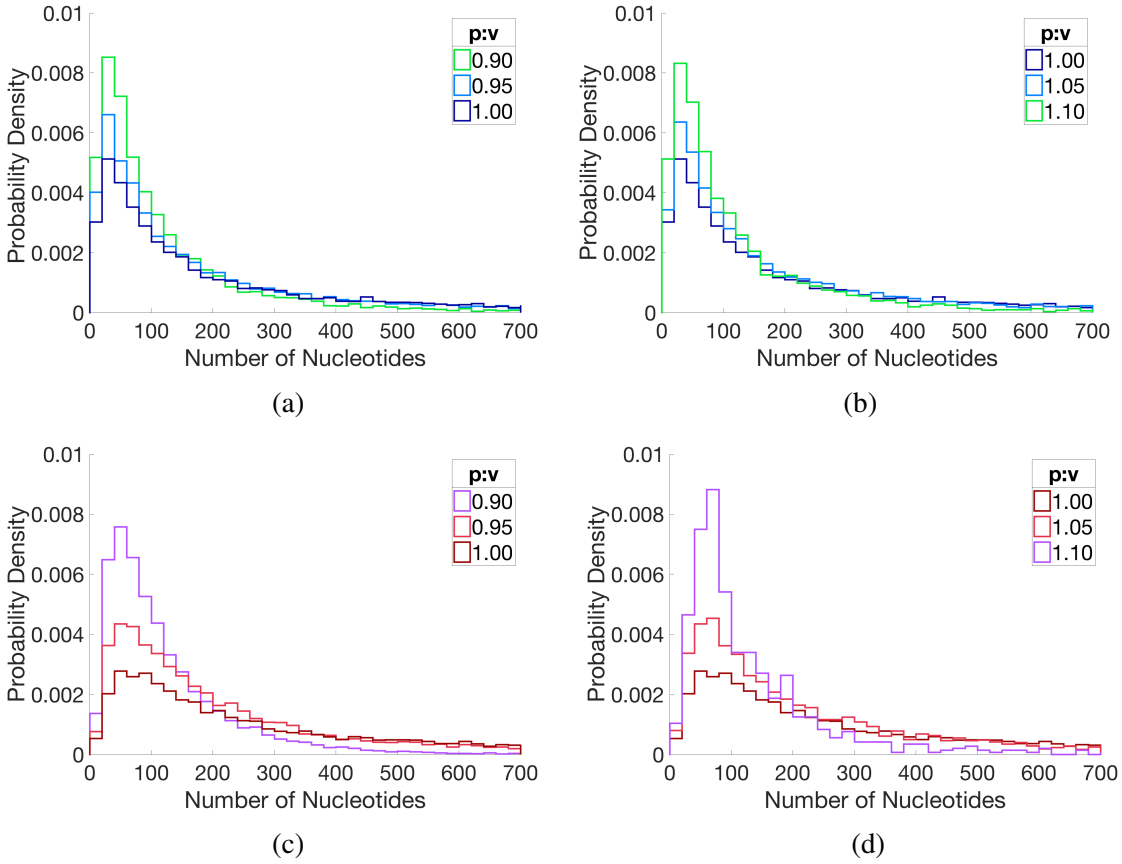


Figure 4.6: The distribution of the number of nucleotides added by RNAPII after the PAS in successful termination events at different p:v ratios for (a) the one-step variant when $p \leq v$, (b) the one-step variant when $p \geq v$, (c) the two-step variant when $p \leq v$, and (d) the two-step variant when $p \geq v$, under default simulation conditions. The two-step variant's reaction rates are set to symmetric.

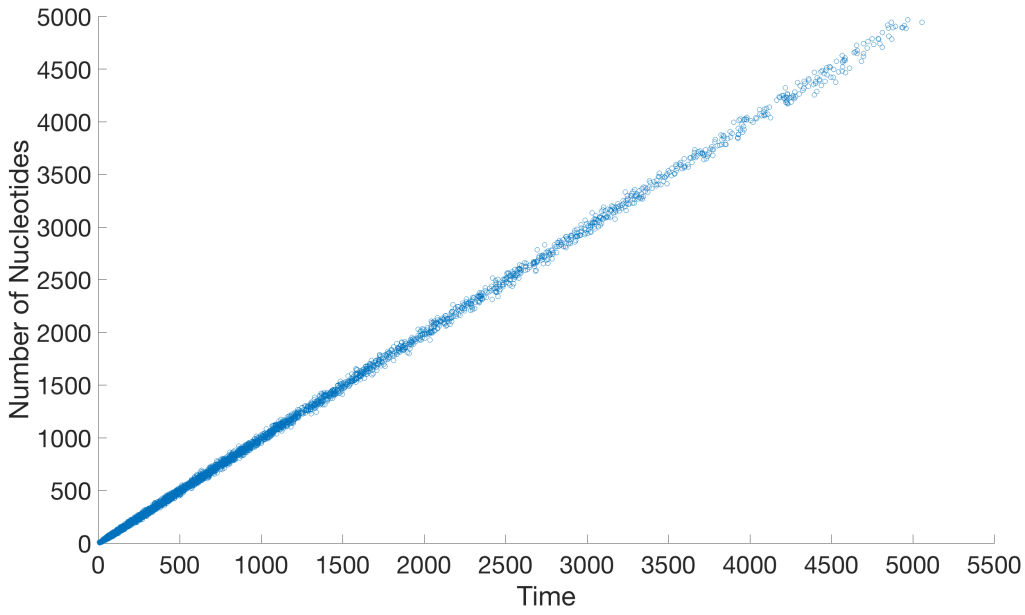


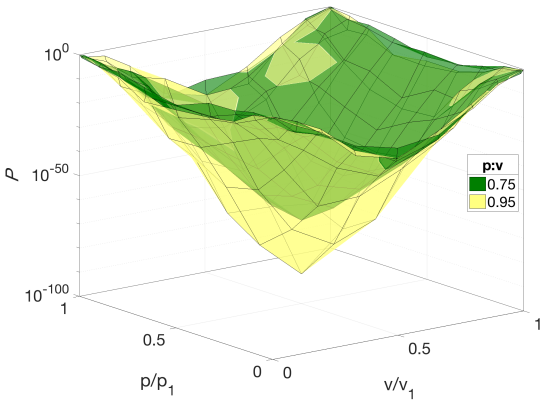
Figure 4.7: Linear relationship between the number of nucleotides added and the time required for a successful termination event. Each successful termination event's number of excess nucleotides synthesized by RNAPII is plotted as a function of the time taken to complete the termination event. The simulation set is for the one-step variant when $p = v$ for default simulation conditions.

using the one-step model?

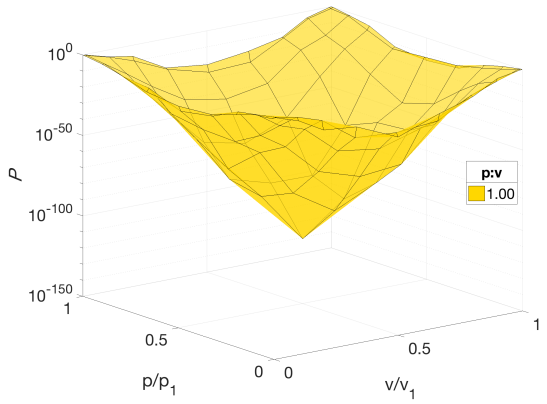
We use the two-sample Kolmogorov-Smirnov (KS) test to investigate the effect of two-step rates on the two-step variant's similarity to the one-step variant. The KS test statistic is the maximum difference between the two samples' cumulative distribution functions and is thus non-parametric, making it applicable to any distribution including those without a known formula [45]. The two-sample KS test uses the test statistic to calculate the p-value P, which is the probability that two experimental samples were drawn from underlying probability density functions which are the same [45]. In our context, we know that the underlying mechanisms of the one- and two-step variants generating the samples are in fact different, so P takes on a slightly different interpretation. P ought to give a rough indication of when the underlying distributions of the two model variants are similar enough that the one-step variant can be used to mimic the two-step variant. The KS test has some limitations. For example, the KS test can make type II errors (two samples are mistaken to be from the same population when they are not) when sample sizes are low enough that the difference in shape of the two distributions can be missed [45]. Therefore, P is used in conjunction with general trends and observations made in previous sections to qualify the differences between the two model variants.

Two-sample Kolmogorov-Smirnov tests were performed using MATLAB R2016b [42] comparing the distributions of successful termination time of one- and two-step model variants of the chase-down mechanism for the p:v ratios of $[0.75, 0.95, 1.00, 1.05, 1.25]$. The trends in the similarity of the one-step and two-step samples as a function of p/p_1 and v/v_1 are depicted in Figure 4.8. It was found through some small resampling experiments that P can range over several orders of magnitude. The reason for this is that the sample size of 10 000 for each parameter set did not thoroughly sample the distribution due to the skewness in the distributions (tails on distributions are hard to sample properly due to the frequency of their occurrence in a simulation). In addition, as p:v increases, sample sizes drop because the success rate of termination drops. This results in the previously mentioned

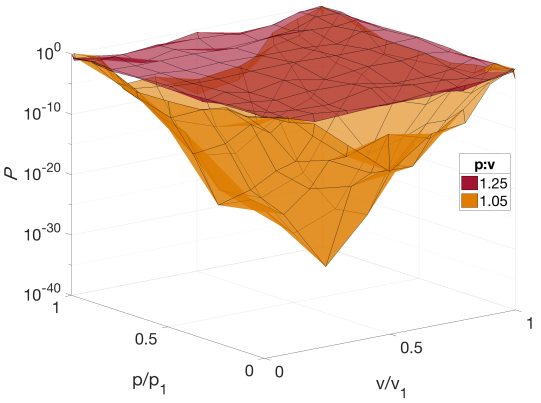
Type II error produced by inadequate sampling. The consequences of shrinking sample size can be seen in Figure 4.8c for $p{:}v = 1.25$, where P trends have been lost. Throughout almost all tests, the smallest P was obtained when $p/p_1 = 0.5$ and $v/v_1 = 0.5$ (i.e., when $p_1 = p_2$ and $v_1 = v_2$). Resampling experiments showed that the different minimum in $p{:}v = 1.05$ seen in Figure 4.8c can be accounted for by the variability of P. At the other extreme, P trends towards 1 when p/p_1 and v/v_1 each approach either 0 or 1. If both p/p_1 and v/v_1 have values within (0, 0.05) and (0.95, 1), then $P > 0.05$. The P drops very quickly as p/p_1 and v/v_1 change to mid range values. By the time both p/p_1 and v/v_1 are in (0.1, 0.9), $P < 0.0001$ for all $p{:}v$ ratios except 1.25. We can interpret this roughly with a few examples. If both RNAPII and Xrn2 occupy one of their two-step states for less than 1% of the time ($0.00 < p/p_1 < 0.01$ and $0.99 < v/v_1 < 1.00$ for example), using the one-step variant to model the system might be considered an acceptable simplification. If however the smallest of either two-step rate of both enzymes takes on average 10% of the time at the average nucleotide, then replacement of the two-step variant with the one-step is insufficient. The limitations on the P calculations discourage further investigation of the region $p/p_1, v/v_1 \in (0.05, 0.10) \bigcup (0.90, 0.95)$. In Chapter 3, it was calculated that RNAPII would spend about 70% of the time in state P1 and 30% of the time in state P2. For $p/p_1 = 0.7$, $P \in (10^{-80}, 10^{-3})$, where the range is seen based on the value of $p{:}v$ (excluding $p{:}v = 1.25$) and v/v_1. As a result, simplifying the overall torpedo model from two-step to one-step chase-down mechanisms would lead to observable distribution differences for biologically valid ranges of RNAPII transcription. Interestingly, a scrutiny of Figure 4.8 reveals that the KS tests respond differently to p/p_1 and v/v_1. Increasing the symmetry in Xrn2 two-step states seems to create more dissimilar distributions faster than changing the ratios of the RNAPII two-step states. The reason for this is not immediately obvious, and may be a topic for future research.



(a) p:$v < 1$



(b) p:$v = 1$



(c) p:$v > 1$

Figure 4.8: P-values P of the two-sample Kolmogorov-Smirnov test results used to estimate ranges of p/p_1 and v/v_1 for which the one-step variant is capable of mimicking two-step results, using various p:v ratios. The axis labels p/p_1 and v/v_1 represent the average proportion of time spent in states P1 and X1 for RNAPII and Xrn2 respectively. Different surfaces depict different p:v ratios. The results are separated into three graphs, a-c, for clarity. The simulations were run under default conditions.

4.6 Statistical parameters of successful termination time

In Section 4.3, qualitative observations of the differences between the distribution of successful termination times for the one- and two-step variants of the chase-down model were made. Some interesting characteristics of those differences arise when various statistical parameters of those distributions are observed. Simulations are used to calculate the mean, coefficient of variation, skewness, and excess kurtosis of the two chase-down variants (Figure 4.9) with default parameter conditions. The simulations of the two-step variant were run with symmetric two-step rates.

It was first observed that the mean successful termination time of the one- and two-step models do not always match, as can be seen in Figure 4.9a. It was not immediately clear why this occurs when the average rates for the polymerase and exonuclease were intentionally set to the same values in the one- and two-step variants. The next few paragraphs are devoted to explaining the phenomenon.

The mean termination time changes for both the one-step and two-step variants based on the size of the intergenic region and the p:v ratio. This observation is shown in Figure 4.10. To investigate, comparisons of the distribution of successful termination times for each intergenic region are shown against each other for select p:v ratios in Figure 4.11. Referencing Figure 4.10, the p:v ratios of 0.8 and 1.2 both yielded equivalent mean successful termination times for the intergenic region sizes of 500, 1000, 5000, and 10 000. The distributions of successful termination time for these cases were overlaid in Figures 4.11a, 4.11b, 4.11e and 4.11f for both model variants. These distributions match up well for these intergenic region sizes. Conversely, the greatest differences in mean successful termination time can be seen when p:$v = 1$. The distributions for these cases are shown in Figures 4.11c and 4.11d. It appears that there exists an underlying distribution governed by the average speeds of the polymerase and exonuclease, and this distribution gets truncated depending on the intergenic region size.

Thus far, it is established that the mean successful termination time matches up for low

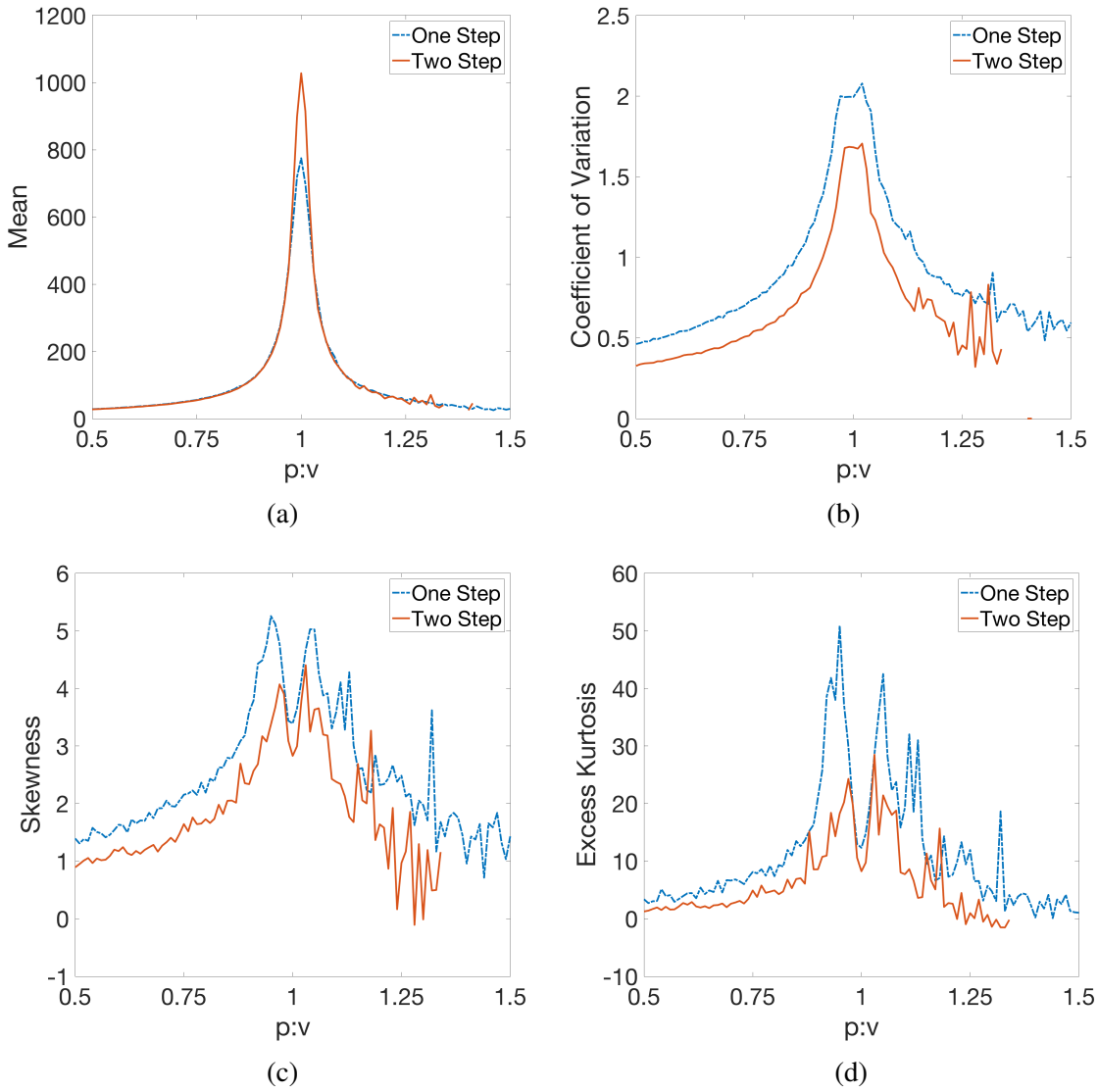
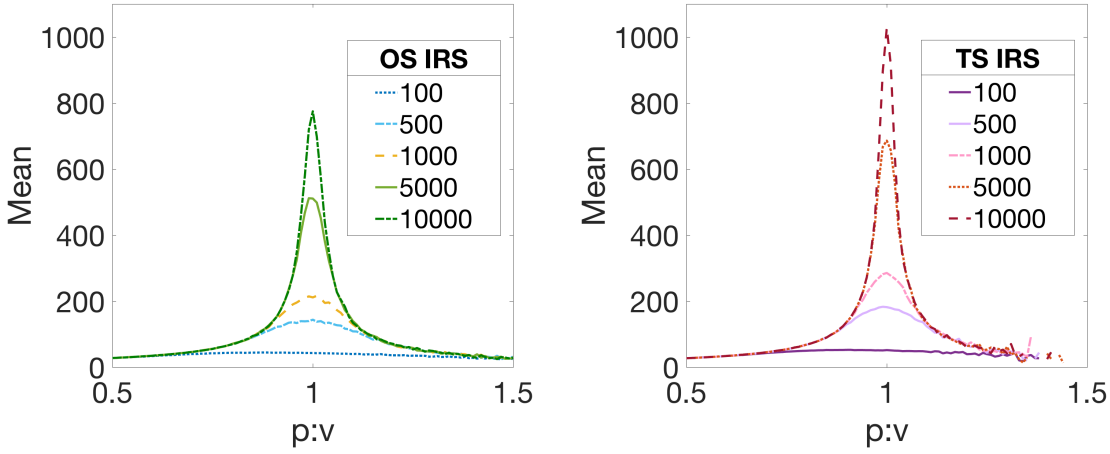


Figure 4.9: The (a) mean, (b) coefficient of variation, (c) skewness, and (d) excess kurtosis of the distribution of successful termination times of the one- and two-step variants (OS and TS respectively) as a function of the ratio of p:v for default parameters. Symmetric two-step rates are chosen for the two-step variant models.



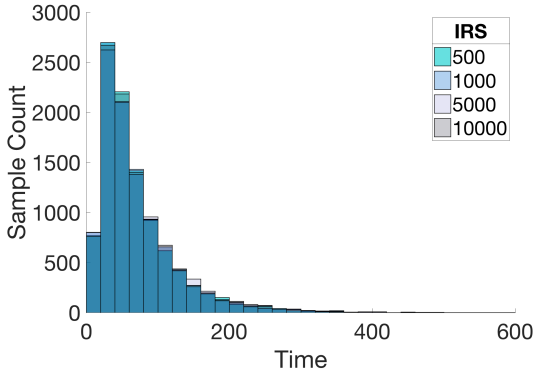
(a) (b)

Figure 4.10: Mean successful termination time for various intergenic region sizes (IRS) for (a) the one-step variant and (b) the two-step variant of the chase-down model, as a function of the polymerase rate to exonuclease rate p:v ratio. Aside from the intergenic region, default simulation parameters are used, and the two-step rates of the two-step variant are symmetric.

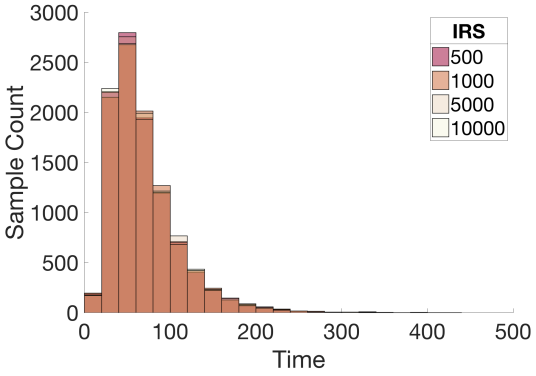
p:v ratios and also for high p:v ratios, with a region of p:v ratios close to 1 that do not match, but it turns out that all three loosely defined cases have different physical reasons for why they turn out the way they do. The following describes the events that actually lead to the phenomena seen in the mean successful termination time for the one-step and two-step chase-down variants.

When p:v is low enough, the success rate for both the one-step and two-step variants is 100%. When the success rate is 100%, the intergenic region size does not become a contributing factor to the distribution of successful termination times. Referencing Figures 4.6a and 4.6c, it becomes clear that if p:v is small enough, the polymerase never makes it to the end of the intergenic region because termination is so efficient. There exists a linear relationship between nucleotide synthesis and time for termination to occur (Figure 4.7), so mean successful termination times are also small.

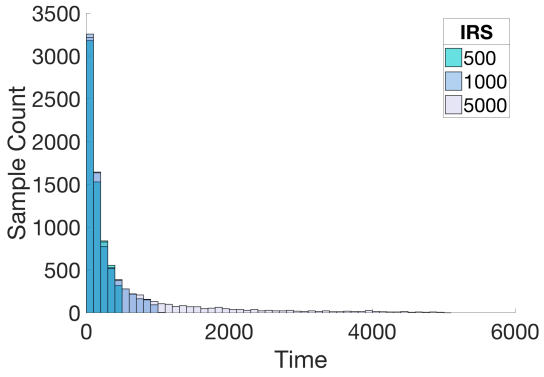
Alternatively, p:v can become sufficiently large that the success rate is extremely small. When termination does succeed, it does so with very little synthesis by the polymerase. Termination occurs sporadically in this region, and both Figures 4.6b and 4.6d show that the



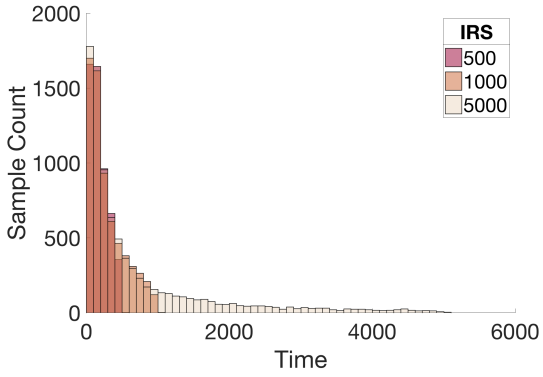
(a) One-step variant for p:$v = 0.8$



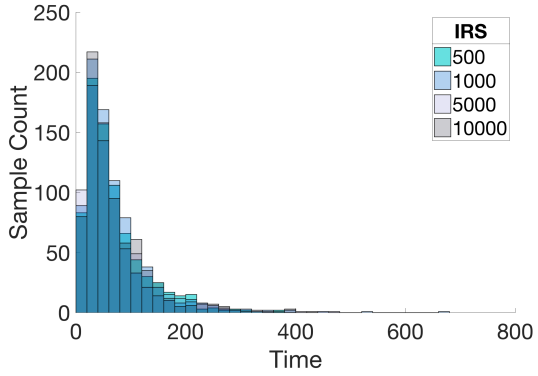
(b) Two-step variant for p:$v = 0.8$



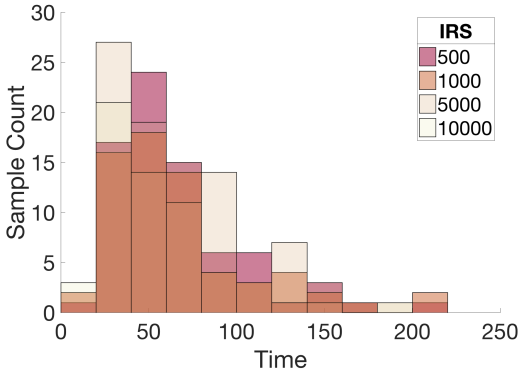
(c) One-step variant for p:$v = 1.0$



(d) Two-step variant for p:$v = 1.0$



(e) One-step variant for p:$v = 1.2$



(f) Two-step variant for p:$v = 1.2$

Figure 4.11: Demonstration of an underlying distribution which depicts the distribution of successful termination times and is modified by intergenic region size.This underlying distribution is based on p:v and is truncated by the restrictions put on by the intergenic region size for p:v values close to 1. Sample counts are used to display the data instead of normalized probability distributions because the graphs demonstrate how the intergenic region size determines the success rate by discriminating against simulations in the tail of the underlying distribution. Simulations were run at default conditions and two-step variants run with symmetric two-step rates.

instances of successful termination never involve the majority of the intergenic region. The number of nucleotides synthesized, and therefore the amount of time taken, between the two outcome groups (successful versus unsuccessful termination) are distinct from one another. The intergenic region size does not truncate the underlying distribution of successful termination times when p is sufficiently larger than v either.

In the $p = v$ case, we see that the underlying distribution is truncated by the restriction of the intergenic region. The distribution of the number of nucleotides synthesized runs right to the end of the intergenic region (not shown) for both the one- and two-step models. It is possible for $p = v$ that the tails of the underlying distribution might extend to infinity, but that the two-step models tail approaches zero faster than that of the one-step model, given the observation of the tail dropping off faster under other p:v conditions. The mean successful termination time for a hypothetical gene with an infinite intergenic region would be infinity. Some validation for this hypothesis for the one-step model is to follow in Chapter 5. If the tail of the one-step model for the distribution of nucleotide synthesis (and therefore termination times) is always longer than that of the two-step model, it will always be truncated by the intergenic region first. Therefore, while the underlying distributions of termination time might be governed by the average termination speeds, and would have equal mean termination times as can be seen for the $p < v$ and $p > v$ cases, the truncation of the one-step tail of that underlying distribution results in a lower observed successful termination time as seen in Figure 4.10a.

A number of other statistical parameters were calculated and displayed in Figure 4.9. Oftentimes, measuring the growth of the standard deviation in relation to the changes in the mean of a probability distribution helps identify changes in overall shape. The coefficient of variation CV is the standard deviation measured in relation to the size of the mean,

$$CV = \frac{\sigma}{\mu}, \tag{4.3}$$

where σ is the standard deviation of the sample and μ is the mean/first moment.

The CV of the distribution of successful termination times is displayed in Figure 4.9b. The one-step variant always predicts a higher coefficient of variation than the two-step variant. This means that successful termination events happen in a more predictable time range for the two-step variant than that of the one-step variant. For both variants, the coefficient of variation is largest near values p:$v \approx 1$; that is, when RNAPII and Xrn2 have comparable overall speeds, the distribution of termination times spreads out faster than the mean grows. There is noise in the CV estimates for high p:v values due to the drop in sample sizes; this noise propagates through all statistics calculated from higher moments.

The next statistic observed is the skewness. The skewness is a measure of asymmetry for a probability distribution function and is calculated by [46]

$$\gamma_1 = \frac{\mu_3}{\sigma^3}, \tag{4.4}$$

where μ_3 is the third central moment. Figure 4.9c shows the calculated values of skewness for various $p : v$ ratios of the one- and two-step variants. All skews calculated from the distribution of successful termination time are positive, a sensible result as a time $t < 0$ has no physical meaning in the system. Meanwhile, the values of t which are allowed by the system extend to ∞. The two-step model typically has a smaller skewness than that of the one-step model.

Finally, the excess kurtosis is a measure of outliers in a distribution [47, 48] calculated through [46]

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3. \tag{4.5}$$

For reference, a normal distribution has an excess kurtosis of 0 and an exponential distribution has an excess kurtosis of 6 [49]. For all p:v ratios sampled, the excess kurtosis is larger than that of a normal distribution, meaning that when the chase-down mechanism produces outliers, they are going to be found at more extreme values than those found in a normal distribution. Large excess kurtosis values are seen in Figure 4.9d, with the curiosity that it

appears to drop at p:$v = 1$ and be peaked on both sides of this value. While kurtosis can be used as a direct measure of tailedness in symmetrical distributions [50], we see high skewness in the distributions and thus have to be more careful of our interpretation of it. Ali [50] demonstrated that the act of summing distributions with different variances can make very large excess kurtosis values. The kurtosis of the chase-down system may be subject to this phenomenon. When p:v is very small one can suppose that there is very little polymerase activity at all, leading to any particular successful outcome being the summation of primarily exonuclease steps, all governed by the same reaction rate v. The two-step model shown uses symmetric two-step rates, $v_1 = v_2$, and thus, each step is still drawn from the same exponential distribution. In the case that p:v is large, successful termination times still rely almost exclusively on exonuclease activity to avoid the event of a runaway polymerase, so the same logic applies, and the kurtosis remains low. Meanwhile, both the skewness and excess kurtosis have a sudden drop in the area of $p = v$ due to the truncation of the chase-down distribution created by RNAPII reaching the end of the intergenic region.

4.7 Two-sample Kolmogorov-Smirnov tests for the relevance of the order of fast and slow reactions governing two-step states of Xrn2 and RNAPII

Finally, whether the length of time the polymerase and exonuclease spend in each of their respective two-step states affects the distribution of termination times is investigated. Does it matter whether or not the rate of reaction of catalysis of RNAPII or the rate of reaction of translocation is faster, or is it only the relative difference that matters? The two-sample Kolmogorov-Smirnov test is used again to compare pairs of different two-step variant distributions A and B which satisfy the following relations: $p_A = p_B$, $v_A = v_B$, $p_{1A} = p_{2B}$, $p_{1B} = p_{2A}$, $v_{1A} = v_{2B}$, and $v_{1B} = v_{2A}$, for a range of parameters. Three different ways of swapping the two-step rates were investigated. The first way is by changing only the polymerase rates with each other (the values of rates p_1 and p_2 are swapped).

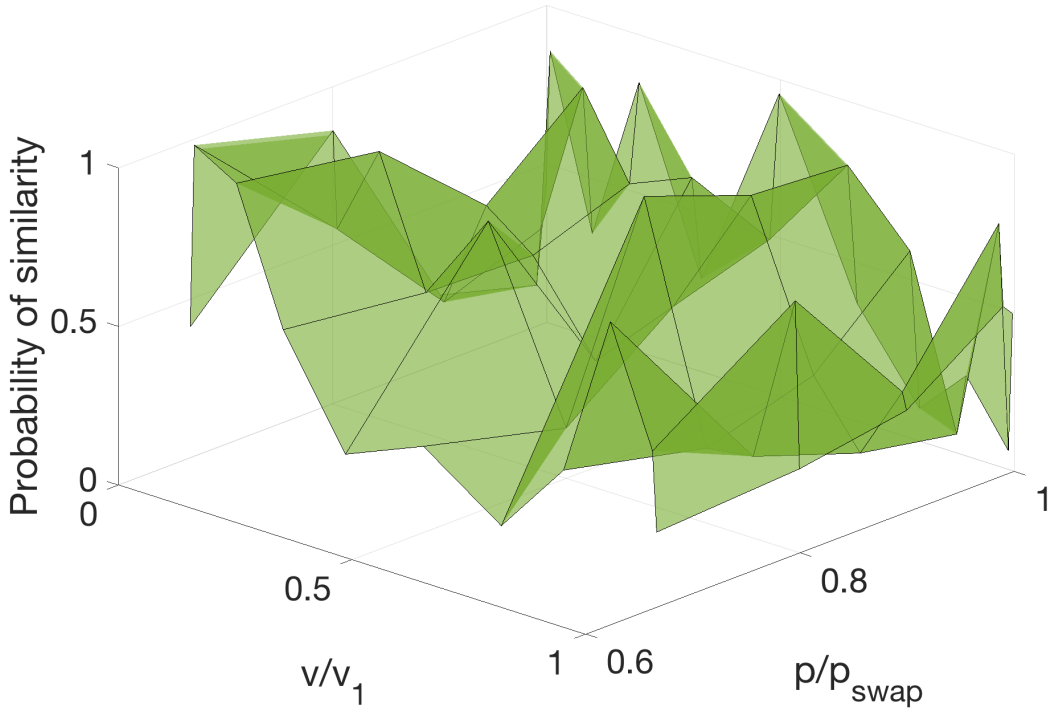


Figure 4.12: Results (P) of two-sample KS tests for switching only polymerase two-step states as an example of relevance of the order of fast and slow reactions in the two-step variant. Rate constants set such that overall polymerase rate $1/p = 1/p_1 + 1/p_2$ and $1/v = 1/v_1 + 1/v_2$. The distributions A and B being compared satisfy the following conditions: $p_{swap} = p_{1A} = p_{2B}$ and $p_{2A} = p_{1B}$, while the exonuclease two-step rates are not swapped but a range of v/v_1 values is explored. The simulations were run at default conditions and p:$v = 0.75$.

Secondly, the values of exonuclease rates v_1 and v_2 were swapped and the RNAPII rates left alone. Finally, both polymerase and exonuclease two-step rates were swapped at the same time. It was found that while P became low on occasion throughout the data, for the three different types of swaps, for all two-step rates investigated, and for the five p:v ratios $[0.75, 0.95, 1.00, 1.05, 1.25]$, no discernible pattern of consistently low probabilities was observed that would suggest that a rejection of the null hypothesis would be merited. Typical results can be seen in Figure 4.12. For the default parameters used the order of fast and slow reactions do not matter for the two-step chase-down variant.

Upon further scrutiny it was found that this result was a characteristic of the initial states

originally chosen for RNAPII and Xrn2 in the two-step variant. All previous simulations were chosen to have the polymerase and exonuclease start in states P1 and X1 respectively. Next, a small number of simulations were run to explore the effects of initial two-step state on success rates and the distribution of termination times. The asymmetric two-step rates described previously in this section were run for $p{:}v = 1$ for initial conditions of $[P1, X1]$, $[P2, X1]$, $[P1, X2]$, and $[P2, X2]$. In addition, the initial nucleotide position for RNAPII was altered and simulations run for $i_0 = [21, 31, 41]$. The success rates were not affected by the initial two-step states. All discrepancies in the results could be explained by the variance inherent to Bernoulli trials as described in Section 4.2. As for the distribution of successful termination times, two-sample KS tests were performed to compare the distributions that satisfy the conditions of distribution A and B as described before. Both the initial conditions and the initial nucleotide position change the total number of additional reactions Xrn2 has to perform above and beyond RNAPII in order to induce termination. For example, simulations that start with the initial position of RNAPII at 31, Xrn2 at 0, and the initial states P1 and X1 will require Xrn2 to undergo 62 more reactions than RNAPII does in order to induce termination. The results of the KS tests are reported in Figure 4.13. Overall, the effects of initial two-step states are reduced as the initial number of nucleotides between the enzyme increases. This suggests that the differences are tied into the number of reactions required. By comparing the P1X1 and P2X2 data, it also becomes apparent that the distributions are affected by the number of slow reactions in particular. As the total number of additional reactions required by Xrn2 increases, the contributions of a single missing or additional reaction have a smaller effect on the overall distribution of termination times.

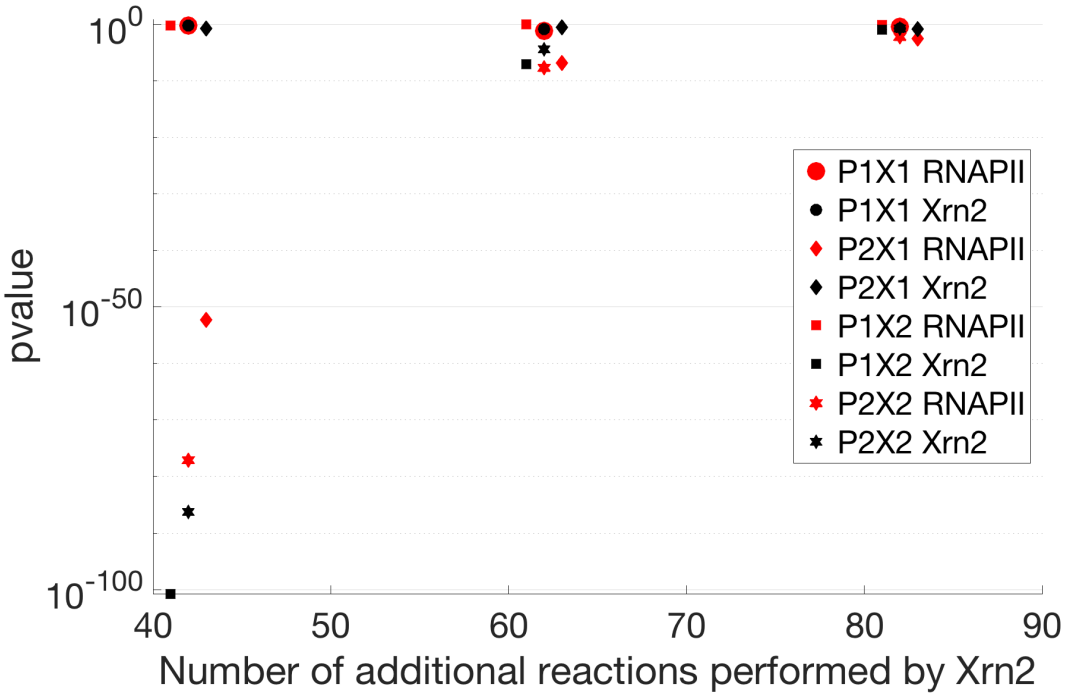


Figure 4.13: KS test P on the distribution of successful termination times of asymmetric two-state rates as a function of the number of additional reactions required by Xrn2 to overcome RNAPII. Three different initial RNAPII positions were considered $[21,31,41]$. The four possible combinations of initial states are considered. Red data is comparing distributions A and B with asymmetric RNAPII two-state rates such that $p_{1A} = p_{2B} = 1.001p$ and $p_{1B} = p_{2A} = 1001p$ and symmetric Xrn2 two-state rates. Black data is comparing distributions with symmetric RNAPII two-state rates and asymmetric Xrn2 two-state rates such that $v_{1A} = v_{2B} = 1.001v$ and $v_{1B} = v_{2A} = 1001v$. Default simulation parameters are used.

Chapter 5

Analytic Solution of the One-Step Model

In the one-step model, the movement of the exonuclease Xrn2 and the polymerase RNAPII are each modeled as a single reaction per protein body as described in the biochemical model section. In this chapter, the one-step chase-down mechanism is transferred into a one-dimensional, discrete space, continuous time random walk and the analytic results are compared to the simulation results from Chapter 4.

5.1 Model setup

RNAPII and Xrn2 are said to be at positions i and j respectively. Every reaction that occurs to a given protein yields a physical translocation to the next nucleotide in the sequence. For RNAPII, this means the protein moves one nucleotide ahead on the DNA, yielding a lengthening of the RNA transcript by one nucleotide. For Xrn2, the translocation of the exonuclease occurs on the RNA and results in the release of the cleaved nucleotide from the transcript and shortening of the RNA.

$$\text{RNAPII}_i \xrightarrow{p} \text{RNAPII}_{i+1}, \tag{5.1}$$

$$\text{Xrn2}_j \xrightarrow{v} \text{Xrn2}_{j+1} \tag{5.2}$$

The progression of RNAPII and Xrn2 are controlled by the rates of reaction p and v respectively. The initial positions of RNAPII and Xrn2 at time $t = 0$ are i_0 and j_0, respectively. Termination occurs when the distance in nucleotides between the exonuclease and

the polymerase closes to the minimum distance at which steric interference occurs, which is said to be $q = 17$ or 18 bases downstream of the polymerase active site i on the RNA [33]. The problem can be transferred into a first passage time problem with the proper modifications. We transfer RNAPII and Xrn2 onto a single discrete number line. Due to the 1-to-1 nature of DNA and RNA, if a nucleotide on the RNA is set to the same integer value as the DNA base that it is transcribed from, then the numbering system of the RNA and DNA is equivalent. This legitimizes putting RNAPII and Xrn2 on the same number line. Given that the nucleotide position of each protein is represented by an integer value, we assume that every reaction that RNAPII or Xrn2 undergoes results in a single discrete step on the number line. The information that tells us about whether a termination event will occur is the distance between the active sites of RNAPII and Xrn2 when Xrn2 makes steric contact with RNAPII, and in the world of math we are allowed to count that distance in whichever way we like so long as the information of the system is preserved. The original position of RNAPII's steric bulk behind its active site on the DNA is set to N, while the original position of the active site of Xrn2 on the RNA is set to m, where $m < N$. Next, we fix RNAPII's position permanently at N on the number line by transferring the results of its reaction to the position of Xrn2, which becomes our walker W. The distance between RNAPII and Xrn2 is now represented by the distance between the walker and position N. A step by W in the positive (right) direction on the number line represents the advancement of Xrn2, while a step by W in the negative (left) direction represents the advancement of RNAPII. Let the position of the walker W on the number line at a given time t be denoted n. The newly remodeled reaction set that describes the system is now

$$\mathrm{W}_n \xrightarrow{v} \mathrm{W}_{n+1}, \textit{and} \tag{5.3}$$

$$\mathrm{W}_n \xrightarrow{\mathrm{p}} \mathrm{W}_{n-1}, \tag{5.4}$$

The problem of the chase-down of RNAPII by Xrn2 has now been transferred into that of a

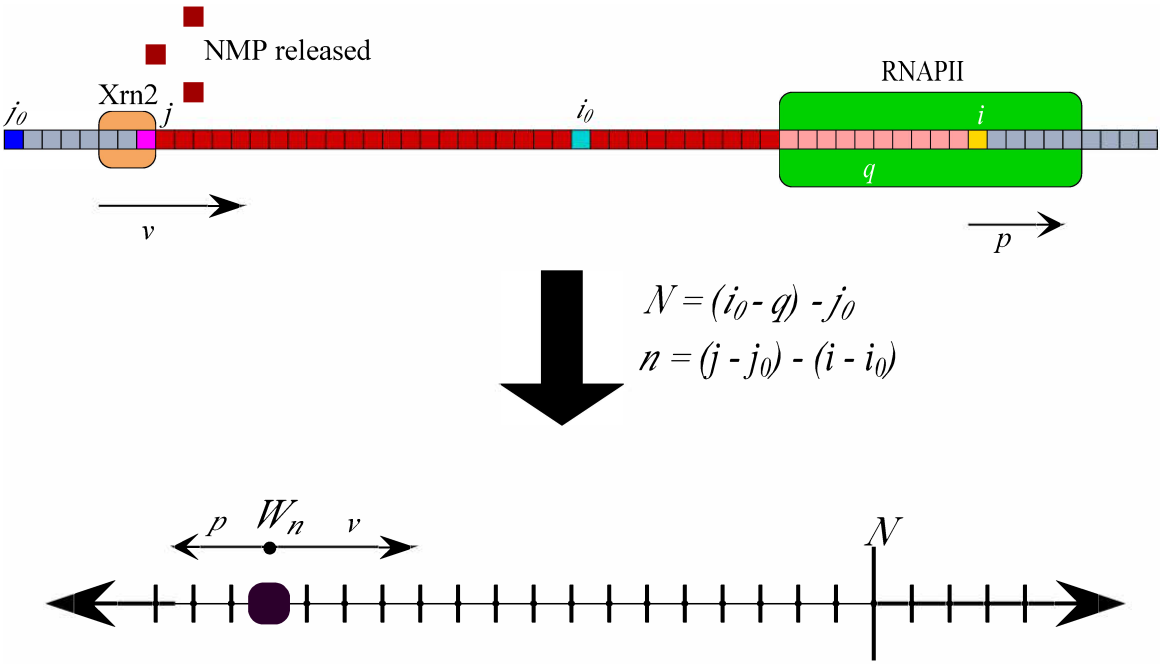


Figure 5.1: Converting the chase-down of RNAPII by Xrn2 into the random walk. An RNA strand's nucleotides are numbered according to the position of nucleotides in the template DNA they are transcribed from. Grey RNA does not exist, but has been degraded by Xrn2 (left) or is to be transcribed by RNAPII (right). Xrn2 was at position j_0 (blue) and RNAPII was at position i_0 (turquoise) at time $t = 0$. RNAPII builds the RNA at site i (yellow), at speed p. Pink RNA is protected from Xrn2 activity by the steric bulk of RNAPII, for a total of q nucleotides protected at any given time. Meanwhile, Xrn2 excises NMP from the RNA at site j (magenta) at rate v. N is the initial relative distance in nucleotides between Xrn2 and the RNAPII, but in the first diagram both RNAPII and Xrn2 move. Both the polymerase speed p and exonuclease speed v are transferred to a single walker W at position n. $N-n$ represents the relative distance between the two enzymes.

random walker in continuous time on a discrete number line. The walker starts at position m at time $t = 0$, moves in a positive direction towards N with a probability per unit time of v, and moves in a negative direction away from N with a probability per unit time of p. $N-n$ represents the relative distance between the two enzymes. When the walker reaches N, the termination event occurs. A summary of changing the chase-down of RNAPII by Xrn2 into a discrete random walk on a one-dimensional lattice can be found in Figure 5.1. Let $P(n,t|m,0)$ represent the probability of the walker being at point n at time t, given that its initial position was m at time $t = 0$. The walk described is spatially homogeneous—that

is, the rates of movement to the left or right are not dependent on lattice position [27], as we have not described alternative reaction rates dependent on the type of nucleotide being transcribed by RNAPII or excised by Xrn2. The function $P(n,t|m,0)$ is well known under these circumstances and is [51]

$$P(n,t|m,0) = \left(\frac{v}{p}\right)^{\frac{n-m}{2}} e^{-(p+v)t} I_{n-m}\big(2t(pv)^{1/2}\big), \tag{5.5}$$

where $I_{n-m}(x)$ is the modified Bessel function of order $n-m$. To put this into a more useful form,

$$P(n,t|m,0) = \left(\frac{v}{p}\right)^{\frac{n-m}{2}} \sum_{k=0}^{\infty} \left[\frac{t^{2k+n-m}(pv)^{\frac{2k+n-m}{2}}}{k!(k+n-m)!}\right] e^{-(p+v)t} \tag{5.6}$$

The distribution of termination times becomes the distribution of the first passage time of the walker from m to N. Let $F(N,t|m,0)$ be the probability of arriving at point N for the first time at time t given that the walker began at position m at time $t = 0$.

Suppose $m < N < n$. Then the first passage time distribution to point N can be found through the following relation [52]:

$$P(n,t|m,0) = \int_0^t F(N,t-\tau|m,0)P(n,\tau|N,0)d\tau \tag{5.7}$$

Equation 5.7 is a convolution. Using the properties of the Laplace transform on convolutions, the Laplace transform of the first passage time is

$$\mathcal{L}\big(F(N,t|m,0)\big) = \frac{\mathcal{L}\big(P(n,t|m,0)\big)}{\mathcal{L}\big(P(n,t|N,0)\big)} \tag{5.8}$$

The moments about zero of a probability distribution can be calculated from its Laplace transform through the following formula [53].

$$\mu_b' = (-1)^b \frac{d^b \mathcal{L}\big(F(N,t|m,0)\big)}{ds^b}\Big|_{s=0} \tag{5.9}$$

The first four moments about zero of the first passage time are calculated explicitly and

are given in Equations 5.10 to 5.13:

$$\mu_1' = \frac{2^{N-m}(N-m)(v)^{N-m}}{|p-v|(|p-v|+p+v)^{N-m}} \tag{5.10}$$

$$\mu_2' = \mu_1' \left(\frac{(N-m)|p-v|+p+v}{(p-v)^2} \right) \tag{5.11}$$

$$\mu_3' = \left(\frac{(N-m)|p-v|+2p+2v}{(p-v)^2} \right) \mu_2' + \left(\frac{4pv}{(p-v)^4} \right) \mu_1' \tag{5.12}$$

$$\begin{aligned} \mu_4' = & \frac{\mu_1'(N-m)\left((p-v)^2(N-m)^2+11p^2+38pv+11v^2\right)|p-v|}{(p-v)^4} \\ & + \frac{\mu_1'\left(6(p+v)((p-v)^2(N-m)^2+p^2+8pv+v^2\right)}{(p-v)^4} \end{aligned} \tag{5.13}$$

Now that the moments about zero are calculated, a number of statistical values describing the underlying probability distribution function can be computed. The central moments $\sigma^2 = \mu_2$, μ_3, and μ_4 are calculated by [46]

$$\mu_1 = \mu_1' \tag{5.14}$$

$$\mu_2 = \mu_2' - {\mu_1'}^2 \tag{5.15}$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2{\mu_1'}^3 \tag{5.16}$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6{\mu_1'}^2\mu_2' - 3{\mu_1'}^4 \tag{5.17}$$

From these central moments, the CV, skewness, and excess kurtosis can be calculated via Equations 4.3, 4.4, and 4.5 respectively.

5.2 Comparing the random-walk model to one-step simulation results

The computations of the moments about zero, central moments, and statistical measures of the random walk solution were performed by using Maple 2015 [54]. Simulation data of the one-step model for a range of intergenic region sizes was then imported from MATLAB R2016b [42] for comparison. Figure 5.2 depicts the analytic and simulation data for the mean (5.2a), coefficient of variation (5.2b), skewness (5.2c), and excess kurtosis (5.2d). Two points can be generally made from the results in Figure 5.2: a fit is observed only for the conditions $p < v$, and the analytic solution is best matched by simulations with large intergenic regions. One property of the unbiased random walk on a 1D discrete infinite lattice is that the probability of visiting any point on the line is 1, at the expense that the mean first passage time becomes infinite; this walk is said to be recurrent [27]. This may seem counterintuitive, but taking the limit of Equation 5.2a as $p \to v$ demonstrates that even if m and N are only one unit apart, the number of walks to the left and away from N are infinite, and the longer those walks which contributed to the mean first passage time are, the longer the mean first passage time gets. Alternatively, transient walks are defined as those which have less than unit probability of revisiting a point [27]. If the walk is biased though, eventual visitation to a point is only guaranteed if the walk is biased towards the point. Therefore, when $p > v$, the probability of reaching the endpoint eventually is no longer 1, and the computed first passage time for all possible walks no is longer meaningful. Therefore, the analytic model does not address the circumstance of $p > v$. Proceeding from this, the fit of the analytic model to simulation data for the other statistical parameters was restricted to $p < v$. To address why simulation results with large intergenic regions best fit the analytical solution, we remember that the intergenic region sets an additional way for the simulation to end. The success rate of termination is maximized for large intergenic region sizes for the conditions $p < v$ because it gives the exonuclease more and more time to catch up. This in turn lowers the likelihood of downstream gene interference. The analytic solution possesses no such alternate boundary condition. Since the analytic first

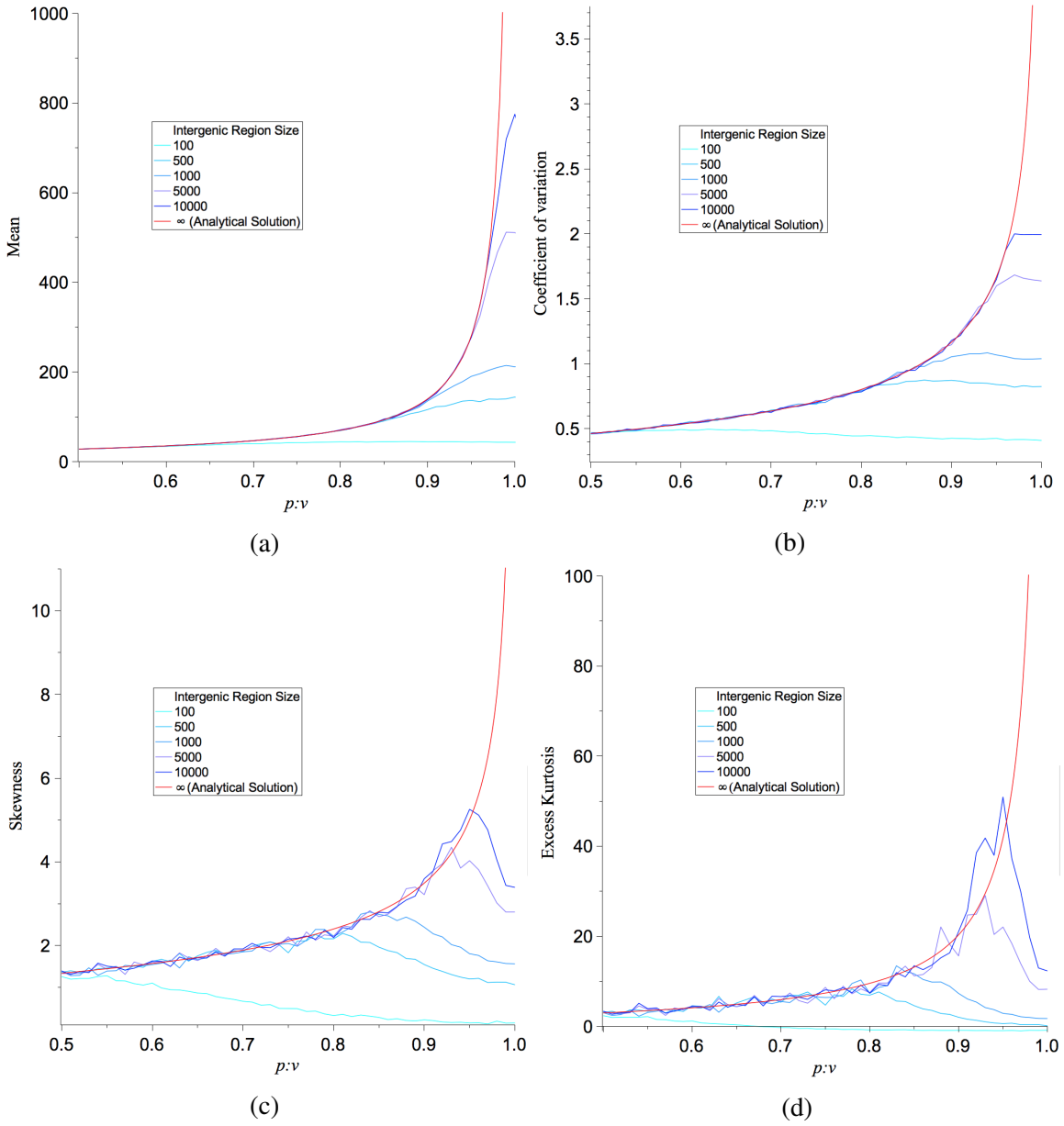


Figure 5.2: Statistical outcomes of the random-walk solution (red) in comparison to various simulation results of the one-step model using varying intergenic regions as a function of the ratio of *p:v*. Depicted are (a) mean, (b) coefficient of variation, (c) the skewness, and (d) excess kurtosis

passage time and one-step simulation models are better matched the longer the intergenic region gets, it suggests that the analytic solution correctly gives the distribution for the one-step simulation variant for an infinite intergenic region. It appears that for all statistical values, infinity is approached as $p \to v$ in the infinite intergenic region case. The change in skewness based on the p:v ratio can be seen in Figure 5.2c. A positive skew makes sense as there is a boundary at $t = 0$; physically there is no such thing as a negative time and the results depict that accurately. As for the excess kurtosis in Figure 5.2d, we see that the analytical solution does not experience the decrease that the simulations do as $p \to v$. The excess kurtosis might become large when there is a mixture of both RNAPII and Xrn2 activity in the one-step variant simulations as discussed in Section 4.6. The drop of excess kurtosis in simulation results is primarily a result of the cutoff of the distribution due to intergenic region size. The smaller the intergenic region, the sooner we see the excess kurtosis drop. Overall, we see that the analytic solution derived from the random walk is a good fit for when $p < v$, and that the larger the intergenic region in the simulations, the better the match to this analytic solution.

Chapter 6

Discussion and Conclusions

6.1 Summary

A biochemical model of the torpedo termination mechanism was built to characterize key aspects of the mechanism. The chase-down of RNAPII by Xrn2 is the primary feature of the model. It was assumed that termination occurs as a result of steric contact between the two enzymes when Xrn2 reaches RNAPII on the transcript. Gillespie simulations were built from the biochemical model to discern termination success rates, the distribution of successful termination times, and the distribution of the number of nucleotides synthesized by RNAPII during the termination process. Two variants of the chase-down mechanism were created. The one-step variant depicts the nucleotide addition cycle as one reaction per nucleotide. The two-step variant models translocation and NTP bond formation with RNA by RNAPII as two separate reactions. The exonuclease's RNA degradation cycle was similarly depicted with either one (one-step variant) or two (two-step variant) reactions per nucleotide excised from the 5' end of the RNA. The two-step variant of the model was considered to be a closer representation of the proposed torpedo kinetics from experimental data than the one-step model, but the simplicity of the one-step model allowed for the creation of an analytic solution equivalent to a one-dimensional random walk.

A variety of parameters needed to describe the mechanism were defined and sought out from the literature. Due to the lack of Xrn2 kinetic data, conclusions from the study were expressed in terms of the relative difference in speeds between the RNAPII NAC and Xrn2 RNA degradation rates instead of measured rates. Additionally, RNAPII was given a

constant number of nucleotides headstart on Xrn2 for the duration of the study.

Eventually, *in vitro* experiments with similar set ups to those produced by Dengl and Cramer [39], Park *et al.* [33] or Larson *et al.* [1] may be developed in order to measure a torpedo termination success rate consisting of just Xrn2 recruitment, chase-down, and termination. In the first two experiments, the DNA is biotinylated and then biotin beads are used to tether the DNA along with the associated RNAPII and RNA segment. The RNA segment size can be chosen, which means that the initial positions of Xrn2 and RNAPII can be selected. A small RNAPII head start of 31 nt was chosen for our study because these experiments are being done with RNA segments of similar length, in the hopes of making comparisons of our model to future *in vitro* experiments. These modest head starts could also represent an *in vivo* termination event with very fast RNA cleavage and Xrn2 recruitment reactions or polymerase pausing coinciding with those reactions. Meanwhile, the experiment by Larson *et al.* uses polystyrene beads to tether the DNA upstream of the biotinylated RNAPII elongation complex, and optical tweezers are used to measure RNAPII position on the DNA as a function of time. It allows a much longer DNA segment to be used than the other experiments, which would help emulate an intergenic region. It may be possible to measure a time of termination from simultaneous addition of Xrn2 and NTP with this set up through measurements of the position of RNAPII.

In the simulations, Xrn2 either catches the polymerase, or the polymerase escapes and interferes with the downstream gene. The effect of the ratio of the polymerase's speed to the exonuclease's speed (p:v) on termination success rates was explored. For large intergenic region sizes, the success rate of termination drops very quickly around $p{:}v = 1$. Small intergenic regions result in lower success rates with respect to p:v, but success rates can be optimized only so much by expanding the intergenic region size. This suggests that little to no improvements in termination success rates would be observed by increasing the intergenic region size beyond 10 000 nucleotides. Success rate drops faster for the two-step model than for the one-step model. Non-zero success rates in the one- and two-step

variants of the model near the region $p{:}v = [1, 1.25]$ are a result of the stochasticity of the system and would not be seen in a deterministic model. It is likely that this result depends on the initial positioning conditions of Xrn2 and RNAPII, which would vary due to RNA cleavage and Xrn2 recruitment rates, as well as RNAPII speeds. It becomes particularly relevant in cases where fast RNA cleavage and exonuclease recruitment are observed, or if temporary pausing of RNAPII occurs during cleavage and recruitment events. Regardless of intergenic region size and model variant, the success rate is near zero by the time $p{:}v = 1.5$, demonstrating that this stochastic quirk can only go so far to produce occasional termination.

Different features of the chase-down model are put on display depending on the comparative speeds of Xrn2 and RNAPII. If RNAPII is much slower than the exonuclease, termination success rate is high and the mechanism is time efficient. As RNAPII becomes comparatively faster, success rates drop as expected. Intriguingly, as RNAPII becomes slightly faster than the exonuclease, the distribution of successful termination times has a lowered coefficient of variation, skewness and excess kurtosis, suggesting that successful termination events are efficient. If polymerase and exonuclease speeds are comparable, it raises the possibility that the torpedo mechanism could have evolved in eukaryotes despite dependability flaws simply because it works really well occasionally. The redundancy provided by the allosteric model in the overall termination mechanism could cover for some unreliability in the torpedo model.

The analytic model using the random walk successfully mimics one-step simulations for the mean, coefficient of variation, skewness and excess kurtosis of the distribution of termination times for $p < v$. The analytic model is the one-step solution for an infinite intergenic region. The one- and two-step variants of the model were compared with each other for when the variant's use the same average speeds of RNAPII and Xrn2. Both models would have the same average transcription termination time, except the finite intergenic region has a censoring effect on both distributions, leading to the average successful termi-

nation time to be higher for the two-step model. The one- and two-step variant distributions differ in that the two-step variant has a lower coefficient of variation, lower skew, and lower excess kurtosis. The peak of the termination times occurs at a later time for the two-step variant than in the one-step variant. The one-step variant's viability to predict the two-step transcription termination events was investigated. Two-sample Kolmorogov-Smirnov tests revealed that the dissimilarity between the one- and two-step models became distinct when either the exonuclease or the polymerase spent at least 10% of the time in the state which was more quickly transitioned from. Estimations made in Chapter 3 revealed that RNAPII spends approximately 70% of its time in state one and 30% in state two, suggesting that for biologically observed ranges, the one-step model oversimplifies the system. Therefore, the analytic model obtained would predict a higher coefficient of variation, skew and excess kurtosis than would be observed in the two-step variant.

6.2 Discussion

The torpedo termination mechanism stands out from other termination mechanisms because of the kinetic competition between RNAPII and Xrn2. This competition made the total number of reactions required for completion variable and therefore required a different modeling approach from termination mechanisms used in previous works [22–25]. It has been demonstrated that the kinetic competition between RNAPII and Xrn2 occurs and determines the final location of transcription termination downstream of the PAS in both *S. cerevisiae* and human cells [28, 34]. The results of our model support these experimental results, also predicting termination further downstream (relating to the number of nucleotides synthesized by RNAPII) as the ratio of polymerase to exonuclease speeds p:v increases. What is striking is that in many cases, transcription beyond the PAS can stretch from a few hundred (in *S. cerevisiae*) to thousands of nucleotides (in human cell lines) [14, 28, 34, 55]. Our model demonstrates that the effects of stochasticity alone can account for similar termination region sizes.

There are a number of details of the torpedo model which were not addressed in this study which would affect termination region size. If RNA cleavage and Xrn2 recruitment rates are slow, we could expect RNAPII to get further ahead during those reactions, which would in turn make it more difficult for Xrn2 to catch RNAPII and lengthen the termination region. Additionally, we modeled the termination reaction as an automatic consequence of Xrn2 catching RNAPII. If instead there is an additional condition required during proper Xrn2 positioning for the termination mechanism to occur, we could expect some instances of Xrn2 reaching RNAPII to not produce termination, thereby lengthening the termination region. On the other hand, RNAPII is subject to pausing in the termination region [16, 20, 28, 43], giving RNAPII a smaller head start which makes the probability of the Xrn2 catching up to the polymerase even greater and shortening the termination region size.

In terms of relating our model's data to experiments, our model most closely resembles the set up of the *in vitro* experiment of Park *et al.* [33]. However, in the experiment, Rai1 was required in addition to Xrn2, and the key aspect of the termination mechanism they highlighted as being different from previous studies [39] was pausing of RNAPII caused by the incorporation of noncanonical NTP during exonuclease activity. In addition, the experiment included the Xrn2 association reaction, and our model implicitly had all four NTPs required for transcription by RNAPII available. Pausing events due to misincorporation seen in [33] occur rarely, as misincorporation events themselves are rare [1]. Furthermore, it was found that the Xrn2/Rai1 complex increases the number of successful termination events when it starts with a longer cleaved RNA transcript [33], something that our model cannot replicate. The termination event was modeled such that there was a steric interaction between Xrn2 and RNAPII which could only occur when the two proteins were a set number of nucleotides apart and which was so efficient as to be assumed to occur when the required distance between RNAPII and Xrn2 active sites was achieved. This set up has a narrow termination opportunity and does not distinguish the type of mechanism that occurs once Xrn2 reaches the polymerase. Park *et al.* proposed that the increased termina-

tion efficiency upon lengthening the RNA to be degraded was due to Xrn2 accumulating momentum as it degrades the transcript, thereby inducing more force on RNAPII upon contact. These results warrant a closer investigation, as it is unclear how the speed of a previous hydrolysis reaction and translocation event could affect the next. An experiment determining whether Xrn2 somehow increases efficiency upon several rounds of NMP excision could be performed by comparing rates of RNA degradation based on RNA length. Since the Xrn2/Rai1 complex could not terminate bacterial RNAP, Park *et al.* also proposed that Xrn2 and RNAPII have a specific interaction with each other that is not merely Xrn2 torpedoing the polymerase. Xrn2 possesses a structure called the tower domain which it shares with its cytoplasmic homolog Xrn1 [38]. The tower domain was proposed to interact with RNAPII during transcription termination [37]. Xrn1 tagged to be directed to nucleus cannot rescue termination defects observed by using the Xrn2/Rat1 mutant Rat1-1 even though it rescues the phenotype overall [20, 36]. One possible explanation is if Xrn2 could terminate RNAPII at a small range of nucleotides behind the polymerase, and that it is merely the proximity of Xrn2 to RNAPII that counts. Their research covered RNA transcript sizes of 21 to 42 nt long, so it is feasible that a proximity effect would have dominated those results and left out the overall effects of an extended chase-down which is seen *in vivo* [28].

6.3 Future work

Overall, this work characterizes some of the effects we would see from the sole contribution of the kinetic competition between RNAPII and Xrn2 as hypothesized for the torpedo model. An important part of future work will consist of expanding this model to include more aspects of the termination mechanism, incorporating the chase-down of RNAPII into the torpedo mechanism in its entirety as described in Chapter 2. This includes both simulations and attempting to extend the two-step variant of the model analytically. The expansion of the model would logically start with an investigation of the effects of RNAPII's start

position in the chase-down process. The more nucleotides downstream RNAPII is at the beginning of Xrn2 association, the longer it will take for the exonuclease to catch up. Whether larger distances between the exonuclease and polymerase start positions can account for the large termination regions observed *in vivo* remains to be seen. Once the effects of the initial polymerase position on the chase-down are understood, incorporating RNA cleavage and Xrn2 recruitment events into the model will dictate what the initial conditions of the chase-down are. Two types of pausing may be required to be incorporated into the torpedo model. The polymerase occupancy on the termination region *in vivo* suggests that pausing or slowing by RNAPII during termination events do occur proximal to the PAS [28]. Since experiments demonstrate kinetic competition between Xrn2 and RNAPII beyond this region [28, 34], this type of pausing is not an indefinite event, nor does it appear to immediately induce termination. It might however play a role in limiting the head start RNAPII can get during Xrn2 association. The other type of pausing is allegedly required for the termination mechanism via Xrn2 [33], which would introduce an additional condition on termination apart from positioning of RNAPII and Xrn2 on the DNA and RNA.

The torpedo model is part of the PAS or cleavage-dependent termination model, and combining this model with the allosteric termination mechanism would provide a more complete view of termination. In a complete PAS-dependent termination model, the percent reliance on either termination pathway *in vivo* could be explored. Given that Pcf11 and Xrn2 recruitment are not independent events [20], modelling PAS-dependent termination may provide insight into their relationship to each other. For instance, Xrn2 is recruited by Rtt103 to the CTD, and whether Rtt103 and Pcf11 recruitement to RNAPII via Ser2 phosphorylation of the CTD [12] is competitive or cooperative could be addressed. Given the complexity of this model, it would be imperative to advance the model from the chase-down in logical steps based on available kinetic data. Building the model piece-wise would also help to identify parameters of the model with larger contributions to the outcomes with respect to nucleotide synthesis by RNAPII, the distribution of termination times, and

success rates of termination.

Some recommendations for future work also involve reporting results in a manner which is more directly comparable to experimental results. This work focused on reporting results as termination success rates, the distribution of successful termination times, and RNAPII's RNA synthesis. It became clear near the end of this work that the results could have been used to calculate an RNAPII occupancy on the termination region similar to those seen in recent experiments. Instead of calculating the number of nucleotides synthesized by RNAPII it would be recommended to transfer this value into RNAPII occupancy levels on the termination region for the purposes of comparison to experimental results. Should the flux of NMP and pyrophosphate into the region due to Xrn2's exonucleolytic activity become significant in some way to biological function, future work should also include a number of nucleotides excised from the 3' flanking region. Eventually, the distribution of termination times may later become useful to understand polymerase recycling rates once relevant kinetic data becomes available.

Bibliography

[1] Larson, M. H.; Zhou, J.; Kaplan, C. D.; Palangat, M.; Kornberg, R. D.; Landick, R.; Block, S. M. *Proceedings of the National Academy of Sciences of the United States of America* **2012**, *109*, 6555–6560.

[2] Reece, J. B.; Urry, L. A.; Cain, M. L.; Wasserman, S. A.; Minorsky, P. V.; Jackson, R. B.; Campbell, N. A. *Campbell Biology*, 9th ed.; Pearson Benjamin Cummings: 1301 Sansome St., San Francisco, CA 94111, 2011; Chapter 17, pp 325–350.

[3] Gilmour, D. S.; Fan, R. *Journal of Biological Chemistry* **2008**, *283*, 661–664.

[4] Kuehner, J.; Pearson, E.; Moore, C. *Nature Reviews Molecular Cell Biology* **2011**, *12*, 283–294.

[5] Porrua, O.; Boudvillain, M.; Libri, D. *Trends in Genetics* **2016**, *32*, 508–522.

[6] Epshtein, V.; Cardinale, C. J.; Ruckenstein, A. E.; Borukhov, S.; Nudler, E. *Molecular Cell* **2007**, *28*, 991–1001.

[7] Wu, J. *Posttranscriptional Gene Regulation: RNA Processing in Eukaryotes*; Wiley-VCH Verlag GmbH & Co. KGaA: Boschstr. 12, 69469 Weinheim, Germany, 2013; Chapter 2, pp 9–40.

[8] O'Reilly, D.; Kuznetsova, O. V.; Laitem, C.; Zaborowska, J.; Dienstbier, M.; Murphy, S. *Nucleic Acids Research* **2014**, *42*, 264–275.

[9] Cramer, P.; et al. *Annual Review of Biophysics* **2008**, *37*, 337–352.

[10] Hsin, J.-P.; Manley, J. L. *Genes & Development* **2012**, *26*, 2119–2137.

[11] Suh, H.; Hazelbaker, D. Z.; Soares, L. M.; Buratowski, S. *Molecular Cell* **2013**, *51*, 850–858.

[12] Heidemann, M.; Hintermair, C.; Voβ, K.; Eick, D. *Biochimica et Biophysica Acta* **2013**, *1829*, 55–62.

[13] Davidson, S.; Macpherson, N.; Mitchell, J. A. *Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire* **2013**, *91*, 22–30.

[14] Banerjee, A.; Sammarco, M.; Ditch, S.; Wang, J.; Grabczyk, E. *PloS ONE* **2009**, *4*, e6193.

[15] Rosonina, E.; Kaneko, S.; Manley, J. L. *Genes & Development* **2006**, *20*, 1050–1056.

[16] Zhang, Z.; Gilmour, D. S. *Molecular Cell* **2006**, *21*, 65–74.

[17] West, S.; Proudfoot, N. J. *Nucleic Acids Research* **2008**, *36*, 905–914.

[18] White, E.; Kamieniarz-Gdula, K.; Dye, M.; Proudfoot, N. *Nucleic Acids Research* **2013**, *41*, 1797–1806.

[19] Kaneko, S.; Rozenblatt-Rosen, O.; Meyerson, M.; Manley, J. L. *Genes & Development* **2007**, *21*, 1779–1789.

[20] Luo, W.; Johnson, A. W.; Bentley, D. L. *Genes & Development* **2006**, *20*, 954–965.

[21] Dye, M. J.; Proudfoot, N. J. *Cell* **2001**, *105*, 669–681.

[22] Vashishtha, S.; Master's thesis; University of Lethbridge; 2011.

[23] Roussel, M. R.; Zhu, R. *Bulletin of Mathematical Biology* **2006**, *68*, 1681–1713.

[24] Roussel, M. R. *Biomath* **2013**, *2*.

[25] Greive, S. J.; Goodarzi, J. P.; Weitzel, S. E.; von Hippel, P. H. *Biophysical Journal* **2011**, *101*, 1155–1165.

[26] Gillespie, D. T. *The Journal of Physical Chemistry* **1977**, *81*, 2340–2361.

[27] Spitzer, F. In [56]; Chapter 1, pp 1–13.

[28] Fong, N.; Brannan, K.; Erickson, B.; Kim, H.; Cortazar, M. A.; Sheridan, R. M.; Nguyen, T.; Karp, S.; Bentley, D. L. *Molecular Cell* **2015**, *60*, 256–267.

[29] Brueckner, F.; Armache, K.-J.; Cheung, A.; Damsma, G. E.; Kettenberger, H.; Lehmann, E.; Sydow, J.; Cramer, P. *Acta Crystallographica Section D* **2009**, *65*, 112–120.

[30] Yuzenkova, Y.; Bochkareva, A.; Tadigotla, V. R.; Roghanian, M.; Zorov, S.; Severinov, K.; Zenkin, N. *BMC Biology* **2010**, *8*, 54.

[31] de Vries, H.; Rüegsegger, U.; Hübner, W.; Friedlein, A.; Langen, H.; Keller, W. *EMBO Journal* **2000**, *19*, 5895–5904.

[32] Teixeira, A.; Tahiri-Alaoui, A.; West, S.; Thomas, B.; Ramadass, A.; Martianov, I.; Dye, M.; James, W.; Proudfoot, N. J.; Akoulitchev, A. *Nature* **2004**, *432*, 526–530.

[33] Park, J.; Kang, M.; Kim, M. *Nucleic Acids Research* **2015**, *43*, 2625–2637.

[34] Kim, M.; Krogan, N. J.; Vasiljeva, L.; Rando, O. J.; Nedea, E.; Greenblatt, J. F.; Buratowski, S. *Nature* **2004**, *432*, 517.

[35] Yang, W. *Quarterly Reviews of Biophysics* **2011**, *44*, 1–93.

[36] Nagarajan, V.; Jones, C. I.; Newbury, S. F.; Green, P. J. *Biochimica et Biophysica Acta* **2013**, *1829*, 590–603.

[37] Xiang, S.; Cooper-Morgan, A.; Jiao, X.; Kiledjian, M.; Manley, J. L.; Tong, L. *Nature* **2009**, *458*, 784–788.

[38] Jinek, M.; Coyle, S. M.; Doudna, J. A. *Molecular Cell* **2011**, *41*, 600–608.

[39] Dengl, S.; Cramer, P. *Journal of Biological Chemistry* **2009**, *284*, 21270–21279.

[40] Djebali, S.; et al. *Nature* **2012**, *489*, 101–108.

[41] Hurowitz, E. H.; Brown, P. O. *Genome Biology* **2003**, *5*, 5:R2.

[42] *MATLAB R2016b*; The MathWorks Inc., Natick, MA, 2016.

[43] Grosso, A. R.; de Almeida, S. F.; Braga, J.; Carmo-Fonseca, M. *Genome Research* **2012**, *22*, 1447–1456.

[44] Pearson, E. L.; Moore, C. L. *Journal of Biological Chemistry* **2013**, *288*, 19750–19759.

[45] Salkind, N. J. *Encyclopedia of Research Design*; Sage Publications, 2010; Vol. 2, Chapter Kolmogorov-Smirnov Test, pp 663–667.

[46] Larsen, R. J.; Marx, M. L. *An Introduction to Mathematical Statistics and its Applications*, 4th ed.; Yagan, S., Ed.; Pearson Education, Inc., 2006; Chapter 3, pp 128–273.

[47] MathWorks; *Kurtosis*; 2017; Retrieved March 15, 2017. `https://www.mathworks.com/help/stats/kurtosis.html?s_tid=srchtitle`.

[48] Westfall, P. H. *The American Statistician* **2014**, *68*, 191–195.

[49] Weisstein, E. W.; *Kurtosis*; 2017; Retrieved January 12, 2017. `http://mathworld.wolfram.com/Kurtosis.html`.

[50] Ali, M. M. *Journal of the American Statistical Association* **1974**, *69*, 543–545.

[51] Goel, N. S.; Richter-Dyn, N. *Stochastic Models in Biology*; Academic Press, Inc., 1974; ISBN 0-12-287460-9.

[52] Siegert, A. J. F. *Physical Review* **1951**, *81*, 617–623.

[53] Feller, W. *An Introduction to Probability Theory and Its Applications*; John Wiley & Sons, Inc, 1966; Vol. II.

[54] *Maple 2015*; Maplesoft, a division of Waterloo Maple Inc., Waterloo Ontario.

[55] Mischo, H. E.; Proudfoot, N. J. *Biochimica et Biophysica Acta* **2013**, *1829*, 174–185.

[56] Spitzer, F. *Principles of Random Walk*; University series in higher mathematics; D. Van Nostrand Company, Inc., 1964.