STOCHASTIC MODELING OF EUKARYOTIC TRANSCRIPTION AT THE SINGLE NUCLEOTIDE LEVEL

SAURABH VASHISHTHA

M.Sc. Bioinformatics, Ch. Charan Singh University, Meerut, 2005

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

M.Sc. BIOCHEMISTRY

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

Dedication

To my parents.

Abstract

DNA is the genetic material of a cell and is copied in the form of pre-mRNA through transcription in eukaryotes. RNA polymerase II is responsible for the transcription of all genes that express proteins. Transcription is a significant source of the stochasticity in gene expression. In this thesis, I discuss the development of a biochemically detailed model of eukaryotic transcription, which includes pre-initiation complex (PIC) assembly, abortive initiation, promoter-proximal pausing and termination as the points that can be slow steps for transcription. The stochastic properties of this model are studied in detail by stochastic simulations with some preliminary mathematical analysis. The results of this model suggest that PIC assembly can play the most significant role in affecting the transcription dynamics. In addition, promoter-proximal pausing has been identified as a potential noise regulatory step in eukaryotic transcription. These results show excellent agreement with many experimental studies.

Acknowledgements

I would like to thank my supervisor Dr. Marc R. Roussel for giving me this wonderful opportunity to pursue a Master's under his supervision, for his constant support and guidance throughout my graduate studies. Coming in without a mathematical background, this thesis would not have been possible without his patience, guidance and support. I especially appreciate his openness to my ideas and for encouraging me to incorporate those ideas in my research.

My other committee members, Dr. Ute Kothe and Dr. John Sheriff, played an indispensable part in my Master's by helping and encouraging me continuously. I am thankful for their substantial feedback, guidance and patience.

I would like to extend a special thanks to Dr. Terry Tang for the useful discussions in the beginning of my research journey. Last but not the least, I would like to thank all the members of the Roussel lab for making this period unforgettable.

Table of Contents

List of Tables

List of Figures

Abbreviations

AdMLP: Adenovirus major late promoter
BRE: TFIIB recognizing element
CME: Chemical master equation
CTD: carboxy-terminal domain
DM: Direct method
DPE: Downstream core promoter element
DSIF: DRB (5,6-dichloro-1-β-D-ribofuranosylbenzimidazole) sensitivity inducing factor
EEC: Early elongation complex
FRM: First reaction method
GTF: General transcription factors
IL-2: Interleukin-2 promoter
ncRNA: Non-coding RNA
NELF: Negative elongation factor
NRM: Next reaction method
nt: Nucleotide
ODE: Ordinary differential equations
ODM: Optimized direct method
PIC: Pre-initiation complex
Pol I, II, III, IV, V: RNA Polymerase I, II, III, IV and V
PTEFb: Positive transcription elongation factor b
SDM: Sorting direct method
siRNA: small interfering RNA
SSA: Stochastic simulation algorithm
TATA-box: TATAAAA consensus sequence
TAF: TBP associated factors
TBP: TATA binding protein
TEC: Transcription elongation complex
TFIIA, B, D, E and F: Transcription factors II A, B, D, E and F
TSS: Transcription start site
URN: Uniform random number

Chapter 1

Introduction

According to the central dogma of molecular biology, genetic information can be transferred from nucleic acid to nucleic acid and from nucleic acid to protein, i.e. from DNA to RNA (transcription) and from RNA to protein (translation) [1]. DNA is the genetic material of the cell that gives an organism a unique identity. DNA stores genetic information in the form of a nucleotide sequence. The transfer of this genetic information from DNA to messenger RNA (mRNA) to proteins is termed gene expression. Eukaryotic gene expression is a complex process, which is completed in several stages including chromatin remodeling, transcription, mRNA processing (including splicing), mRNA transport to cytoplasm and translation [1]. The regulation of gene expression is a fundamental process for the development, growth and survival of an organism. Gene expression can be regulated at any of the individual steps involved, but it is mostly regulated at the level of transcription [2].

Stochasticity (noise) is part of the inherent nature of eukaryotic as well as prokaryotic gene expression. Stochasticity occurs due to the intrinsic randomness of individual biochemical reactions involved in the transcription and translation processes, the small number of molecules involved, random binding of regulatory factors and random occurrence of events in various steps, in spite of similar environmental conditions [3, 4]. This noise (randomness) can affect cellular differentiation, cellular functions, adaptability of organisms to their environment and phenotypic variation [3, 5] Therefore, studying sources of stochasticity is an important prerequisite to understand

how cells cope with noise. Stochastic models help us to get better insights into the effects of the noise on the biological phenomena described above as well as into the sources of this noise. In this thesis, a biochemically detailed stochastic model of eukaryotic transcription at the single nucleotide level has been developed. This stochastic model will help to get better insights into the possible sources of noise, the slow steps of transcription and their effects on the biological phenomena described above.

This chapter will review the biochemical details of the steps involved in eukaryotic transcription. In addition, various possible rate-limiting steps will be discussed that can be significant causes of stochasticity. At last, stochasticity in gene expression and in transcription especially will be reviewed, including previous modeling studies of transcription and gene expression.

1.1 Transcription

Transcription is the copying of the genetic information stored in the DNA into RNA. In prokaryotes, a single RNA polymerase is responsible for all transcription whereas several types of RNA polymerases are known in eukaryotes. Each RNA polymerase encodes a different type of RNA: RNA polymerase I (Pol I) helps in the synthesis of ribosomal RNA (rRNA), RNA polymerase II (Pol II) synthesizes the precursor of mRNA (pre-mRNA), RNA polymerase III (Pol III) helps in the formation of transfer RNA (tRNA), RNA polymerase IV (Pol IV) synthesizes small interfering RNA (siRNA) in plants [6] and RNA polymerase V (Pol V) generates non-coding RNA (ncRNA) in plants [7]. Of these polymerases, Pol II is responsible for the transcription of all genes that express proteins. All general transcription factors (GTFs) including Pol II bind to the core

promoter of the gene and Pol II transcribes the gene to form a complementary nonfunctional pre-mRNA transcript, which is processed either co-transcriptionally or later to form functional mRNA. Eukaryotic transcription has been extensively studied experimentally. The real-time dynamics of the eukaryotic transcription has been studied in [8-11]. In addition, several recent reviews cover different aspects of eukaryotic transcription such as the mechanism of transcription [12, 13], and the structural aspects of different transcription complexes and their roles in the transcription mechanism [14, 15].

For transcription, binding of transcription factors (TFs) is necessary. This binding depends on the state of the gene. Broadly, a gene can be in one of three states: repressed (off), basal and induced (on) [16]. A gene remains 'off' when it is packed in the chromatin and the promoter region is not available to the transcription factors; a gene in the basal state resides in an open chromatin region but is expressed at low levels [16]; whereas, if the gene is in the 'on' state, its expression reaches high levels due to the binding of the transcription factors with the help of activators.

More precisely, eukaryotic transcription starts with the binding of general transcription factors at the core promoter region of a gene in the 'on' state [14, 17]. A core promoter is a part of typical eukaryotic promoter region, which consists of numerous binding sites for sequence-specific activator proteins [16, 18]. The core promoter region covers a region from ~35 base pairs upstream to 30-35 base pairs (bp) downstream of the transcription start site (TSS) [17]. It consists of a TATAAAA consensus sequence (TATA box), approximately 25-30 base pairs (bp) upstream of the TSS, which facilitates the binding of transcription factors [17, 19, 20]. It has been found that only 30-32% of genes in *Drosophila* have a functional TATA box ([21], reviewed in [14, 17]). The core

promoter usually has other regions such as BRE (TFIIB recognizing element), DPE (Downstream core promoter element) and Inr (Initiator), which may facilitate the binding of the transcription factors on an AT-rich region in the absence of a TATA box [17]. The distance of the TSS from the TATA box varies from organism to organism. In metazoans, the TSS remains 25-30 bp downstream of the TATA box whereas in *Saccharomyces cerevisiae* multiple TSSs are present 40-120 bp downstream of the TATA box [17, 19, 22]. Transcription can be broadly divided into four phases: (a) pre-initiation complex (PIC) assembly, during which establishment of the transcription machinery through the binding of transcription factors occurs, (b) promoter escape, during which initiation and escape commitment occur and abortive initiation may also occur, (c) elongation, in which Pol II moves along the DNA template elongating the nascent transcript, and (d) termination, in which dissociation of the transcription complex takes place releasing a new pre-mRNA transcript.

1.1.1 Pre-initiation complex (PIC) assembly

There are several mechanisms known for PIC assembly. Regardless of the differences of mechanisms of PIC assembly, the first step on TATA-containing promoters is similar: TATA binding protein (TBP) monomer along with 13 other TBP associated factors (TAFs) binds to the TATA box with the help of gene-specific activators [23]. Structurally, TBP and 13 other TBP associated factors (TAFs) together form TFIID [16, 20, 23-26]. In other words, TBP represents the central DNA-binding subunit of TFIID [27]. TBP alone is capable of a basal level of transcription in some genes, but for a high level of transcription, i.e. activator dependent transcription, recruitment of TAFs with

other general transcription factors is required [16, 28]. TBP is also required on TATA-less promoters and binds at a position near -30 in an AT-rich region, where the TATA box normally resides [14, 16, 25]. Even in the absence of a TATA box or of an AT-rich region, the binding of TBP on DNA remains conserved [14, 29].

Being required in both TATA-containing and TATA-less promoters, indicates some important function for TBP. There are contradictory views about the function of TBP [16]. It acts as a scaffold with TAFs (TFIID) for the other general transcription factors and may also play a role in locating the TSS. Various structural studies of TBP suggest that the carboxyl-terminal domain of this protein is conserved through the phylogeny of this protein whereas the amino-terminal domain is not conserved [16, 23, 25, 28]. The carboxyl-terminal domain (CTD) of TBP helps in its binding to the minor groove of DNA. The carboxyl-terminal domain of saddle-shaped TBP binds to the DNA at a very oblique angle of 22° and makes the DNA bend towards its major groove [28, 30]. Most of these structural studies suggest that only monomeric TBP can bind to the TATA box.

In the 1990s, B. F. Pugh and his colleagues extensively studied the kinetics of TBP/TFIID binding to DNA as well as TBP/TFIID dimerization [16, 25, 31-35]. It is not known to this date whether TBP binds to TFIID (TAFs) before or after binding to the DNA. However, it is clear from these studies that TBP/TFIID normally tends to form homodimers, when not bound to the DNA [31, 33, 35]. Therefore, TBP and TFIID are largely found in the dimer form under physiological conditions [32]. It is important to note that all the studies that mentioned TFIID dimerization considered the TBP as a part of the 14 subunits of TFIID. Many studies also suggest that TBP/TFIID monomer

binding to the DNA is in competition with the TBP dimer formation, which is a form of auto repression [32, 33]. Once TBP/TFIID is bound to DNA, the TBP/TFIID/TATA-box complex provides a platform for the binding of other general transcription factors (GTFs). However, there are a number of contradictions about TBP and the mechanisms associated with it. B. F. Pugh reviewed these contradictions about TBP and associated factors in detail [16].

There are two commonly known mechanisms for the recruitment of general transcription factors at the TATA box. The first mechanism suggests an ordered binding of general transcription factors (GTFs) including Pol II on the TBP/TFIID complex prebound to the TATA box. The detailed mechanism of the ordered recruitment of GTFs has been reviewed in [36, 37]. According to this mechanism, the general transcription factors are recruited in a certain order, i.e. after the binding of TBP/TFIID to the TATA box, TFIIB binds to the TBP/TFIID complex, which helps in recruiting the Pol II/TFIIF complex to the transcription machinery. On binding of Pol II, TFIIE and TFIIH get recruited to the transcription machinery. According to this model of GTF recruitment, TFIIA can bind to the transcription machinery at any time after the recruitment of TFIID and TFIIB [36]. This model was supported by the finding that other than GTFs, the chromatin modifying factors, are also recruited in a specific order [38, 39]. The finding of several holoenzymes with different GTFs challenged the idea of ordered recruitment [40-42]. However, some specific transcription factors were missing in these holoenzymes. Further investigation on the yeast Pol II transcription complex established the holoenzyme binding model of GTF recruitment: TBP/TFIID and TFIIA bind together in

the first step, followed by the binding of the Pol II holoenzyme which includes other GTFs in the second step (Figure 1.1) [43].

The most important component of the transcription machinery is RNA Polymerase II (Pol II). Professor Roger D. Kornberg received the Nobel Prize for Chemistry, in 2006, for his phenomenal contribution to the structural studies of different polymerases along with various transcription factors. Dr. Kornberg and his colleagues studied the structures of Pol II during various steps of transcription with and without several transcription factors at various resolutions [44-48].

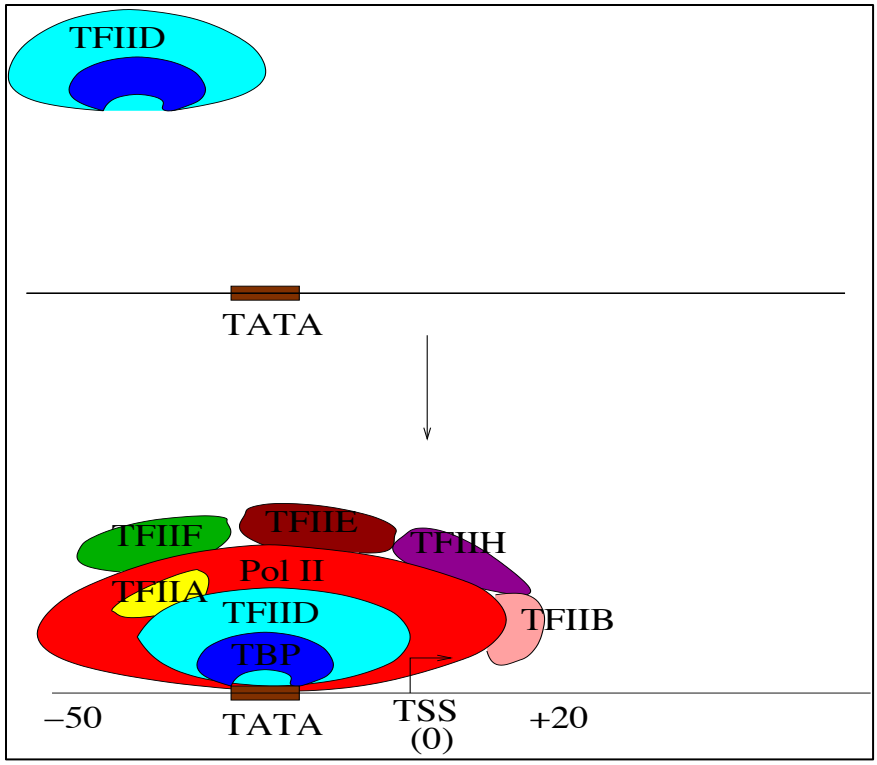

Figure 1.1: Schematic diagram of the holoenzyme binding mechanism of pre-initiation complex assembly: TFIID with TATA binding protein (TBP, blue) binds first on the TATA box and directs the RNA Polymerase II holoenzyme to form the pre-initiation complex (PIC).

Other than Dr. Kornberg's group, there are several other research groups involved in the structural studies of RNA Pol II and its complexes with transcription factors. Structurally, Pol II is a 12-subunit complex, in which ten subunits represent the core enzyme, attached with the heterodimer of subunits Rpb4 and Rpb7 [14, 49]. Out of these 12 subunits, only three subunits, Rpb1, Rpb3 and Rpb5, remain in contact with the DNA template after cross-linking [50]. Every transcription factor of the pre-initiation complex has a specialized function. For example, Pol II catalyzes the transcription of all the genes synthesizing mRNA; TFIID, which is composed of TBP and TAFs, directs PIC assembly on the TATA box [20, 51, 52]. All the general transcription factors and their functions are summarized in Table 1.1.

The kinetics for the assembly of the pre-initiation complex has not been studied extensively. However, in some studies, the early steps of transcription on different promoters and different transcription systems have been studied [53, 54]. These two studies provide some insights into the kinetics of pre-initiation complex assembly. One of these studies [53] has been performed on adenovirus major late promoter (AdMLP) with a minimal transcription system, whereas the other study [54] was performed on the interleukin-2 (IL-2) promoter with the fully active transcription system.

The rate of pre-initiation complex assembly on the pre-bound TATA binding protein (TBP) was found to be extremely fast i.e. $\geq 0.1\ s^{-1}$ with the minimal transcription system at AdMLP [53], whereas the rate of PIC assembly with a fully active transcription system at the IL-2 promoter was found to be one of the major slow steps of transcription [54]. These kinetic studies provided data to choose the rate constants for PIC assembly in this study (chapter 2).

Table 1.1: General transcription factors contained in the pre-initiation complex and their functions

TFIID	Includes TBP and 13 TAFs, binds to TATA box and shows co-activator activity through interactions with activators. Also promotes TFIIB binding
TFIIA	Acts as a co-activator as well as playing an important part in stabilizing the TBP/TFIID/TATA complex
TFIIB	Helps in recruiting Pol II and TFIIF, and finds the accurate TSS
RNA Polymerase II	Catalyzes transcription
TFIIF	Remains strongly bound with Pol II and helps in directing it to TBP/TFIID/TFIIB/TATA complex, necessary for TFIIH and TFIIE binding
TFIIE	Helps in recruiting TFIIH and stimulates kinase and ATPase activity of TFIIH
TFIIH	Plays a key role in promoter clearance and opening. Shows kinase activity and helps in phosphorylation of C-terminal domain of Pol II, aids in the transition from initiation to elongation

1.1.2 Initiation/abortive initiation/promoter escape

The early movement of Pol II has only been studied in the last 20 years or so. Therefore, relatively little information is available in comparison to the pre-initiation complex assembly. However, this early movement of Pol II has been studied sufficiently to get insights into the mechanistic and kinetic aspects [53-64]. Some studies divide this initial movement of Pol II on the first 15 positions into several parts: initiation, escape commitment and promoter escape, including the possibility of abortive initiation [53, 54,

56, 57]. On the other hand, some studies termed promoter escape as the movement on the first 14-15 positions and subdivided promoter escape into initiation, escape commitment and abortive initiation [61-64]. Promoter escape is also misinterpreted as promoter clearance in some studies ([59, 60] and references within [61]). However, promoter clearance is completely different from promoter escape. Promoter clearance completes when the active site of Pol II reaches 60-70 bp downstream of the TSS, whereas promoter escape completes with the synthesis of a 14-15 nt long transcript [61]. Regardless of the division of the phases in the initial movement of Pol II, we know that all these steps except promoter clearance occur during the movement from positions 1-15.

For the successful initiation of transcription, the accurate positioning of the active site of Pol II at the transcription start site (TSS) is necessary, which occurs with the help of TFIIB. Once PIC assembly is complete, nucleoside triphosphate (NTP) molecules interact with Pol II and consequently DNA melts around the TSS and forms the transcription bubble to provide an open complex for the movement of Pol II. The size of the transcription bubble remains irregular during initiation and promoter escape. In an *in vitro* study of early transcription, the transcription bubble has been observed from position -9 to +2 at the start of transcription, and after the formation of the first phosphodiester bond, the downstream edge of transcription bubble was observed to expand to position +8 [59]. In another study, the transcription bubble has been found to cover ~17 bp, with an 8 nt long nascent transcript, whereas the bubble size has been found to be only 10 bp with a transcript length of 9 nt due to bubble collapse [60].

The formation of the first phosphodiester bond involves the phosphorylation of the C-terminal domain of the large subunit of Pol II through the kinase activity of TFIIH,

causing the movement of Pol II to the first nucleotide position [51]. The initial DNA-RNA hybrid remains very unstable on the first 3 nt positions. In between the formation of the 3rd and 4th phosphodiester bonds, a special transition, 'escape commitment', has been observed on the AdMLP as well as on the IL-2 promoters [53, 54, 56, 57]. As a result of this transition, the ternary complex becomes relatively more stable. A similar kind of transition has been observed in another study [59]. In this study, formation of a promoter open complex was tested with ATPγS, a reversible inhibitor of transcription. It has been observed that this open region can be reversed to closed complex by the addition of ATPγS to the pre-mRNA product until it reaches a 4 nt length. As soon as it reaches the 4 nt length, the open region is insensitive to ATPγS [59]. Kinetically, the movement of Pol II on the first 3 positions on IL-2 or AdMLP promoter is very fast. The rate constant for this movement has been found to be $\geq 0.1 s^{-1}$ [53-56].

Promoter escape starts with the synthesis of the fourth nucleotide of the pre-mRNA transcript [53] and ends with the synthesis of the 14th or 15th nucleotide of the nascent transcript (reviewed in [61]). In this phase of transcription, the DNA-RNA hybrid is considered to be relatively strong and is committed to complete promoter escape. *In vitro* kinetic studies of promoter escape show contradictory observations: In studies on AdMLP with a minimal as well as a full transcription system, the movement of Pol II from positions 4-28 was found to be a slow step of the transcription with a very low rate constant of $\sim 2 \times 10^{-3} s^{-1}$ [54, 56], and has thus been suggested to be the rate-limiting step in early transcription [53, 57]. Further *in vitro* investigation of this observation suggested that the translocation on the 8th position is responsible for the rate-limiting movement of

Pol II from positions 4-28 [55]. On the other hand, in another *in vitro* study on the IL-2 promoter with a fully functional transcription system, the movement from positions 4-28 has been found to be slow with a rate constant of 0.025 (±0.002) s^{-1} [54]. This rate constant was slow but not rate limiting as in the minimal transcription system. These observations suggest a role for promoter escape in controling the rate of transcription.

Another complexity that has been observed in this early movement of Pol II is abortive initiation. In general, the transcription machinery becomes less prone to abort after passing the 4^{th} nucleotide. However, abortive transcripts of 3 -15 nt have been found in experiments. Abortive initiation in eukaryotes has been studied in less detail compared to prokaryotes, perhaps because of the more complex transcription system of eukaryotes [61]. In prokaryotes, typical *in vitro* lengths of abortive transcripts have been found to be between 2-16 nt [61, 65], whereas abortive transcripts have been found to be 11-15 nt long *in vivo* [66]. In eukaryotes, 2-10 nt long abortive transcripts have been found *in vitro* [59, 67] whereas, one review suggests the length of abortive pre-mRNA to be 3-15 nt long [61]. The detailed mechanism of abortive initiation in eukaryotic transcription is still unknown, however, a single molecule study of prokaryotic transcription suggests that abortive initiation occurs when the polymerase is actively translocating [68]. In addition, there is no precise information available regarding the fraction of abortive initiation events. However, a biochemical study of prokaryotic transcription suggests that ~75% of the initiation events end up with the synthesis of short abortive transcripts through abortive initiation [69]. Another *in vitro* study suggests that more than half of the initiation events lead to abortive initiation in eukaryotic transcription [67]. If the

transcript does not abort in the abortive initiation prone area, then it enters into the elongation phase of transcription.

1.1.3 Elongation/promoter-proximal pausing

In the past, the events occurring during polymerase recruitment were considered responsible for the regulation of gene expression, but recent studies of elongation emerged with the contradictory view that post-recruitment events can also play a major role in regulation, and can be sufficiently slow to affect the transcription dynamics [2, 70]. The distribution of Pol II on the DNA template in some *in vivo* studies suggests that the processes in between the recruitment of Pol II and productive elongation can be the rate-limiting steps in eukaryotic transcription [12]. Elongation has been one of the most studied steps of the transcription process in recent years. Elongation starts with the completion of promoter escape. However, there are some contradictory views about the elongation phase starting point. Some studies include promoter escape as a part of the elongation phase [2, 12, 13], whereas others viewed it as a distinct phase of transcription [61, 71]. In the present study, promoter escape and elongation have been considered as separate phases of transcription. Elongation factors add to the transcription complex forming the elongation complex. The elongation complex moves on the DNA template to synthesize the transcript. The complex before or during promoter-proximal pausing is called the early elongation complex (EEC), and after promoter-proximal pausing, it is called the transcription elongation complex (TEC). The EEC, despite stable binding of RNA to the transcription complex, is more prone to transcription arrest and backtracking than the mature TEC. Promoter-proximal pausing plays a significant role as a checkpoint,

i.e. in checking the DNA-RNA hybrid and that the transcription was accurately done. Therefore, promoter-proximal pausing reduces the chances of transcription arrest and backtracking that happen due to wrong transcription or a weak DNA-RNA hybrid [2].

Promoter-proximal pausing has been extensively studied in the *hsp70* gene of *Drosophila* [72, 73] and the *c-Myc* and *c-FOS* genes of human [74, 75]. Promoter-proximal pausing has been found to occur between positions 20-50 in most eukaryotes (reviewed in [2, 75-79]). It is considered as a point for the regulation of eukaryotic transcription [75]. Several studies of promoter-proximal pausing have shown that promoter-proximal pausing represents a common phenomenon of eukaryotic transcription, which is not dependent on the promoters and other transcription factors [75, 80]. It has also been studied as a possible rate limiting step of the early eukaryotic transcription [80]. Biochemical studies of promoter-proximal pausing suggest that Pol II enters the paused state rapidly but escapes from the paused state in a slow manner [75]. The average percentage of polymerases that pause in the promoter-proximal region is not known for all eukaryotic genes. An *in vivo* study on mutated Chinese hamster cells showed that ~75% of the polymerases moved rapidly, but ~25 % of the polymerases were immobile [8]. It is believed that promoter-proximal pausing is a common phenomenon in eukaryotic gene expression [75], and perhaps occurs in most genes. Similarly, the duration of pausing is unknown for most eukaryotic genes, but the kinetic studies of the *hsp70* gene of *Drosophila* suggest that a pause can last as long as 25-30 minutes [72, 73]. However, the duration of a pause may depend on several factors such as type of gene, promoter strength, etc. [76].

The mechanism of promoter-proximal pausing was not known until the remarkable discovery of two negative elongation factors DSIF [DRB (5,6-dichloro-1-β-D-ribofuranosylbenzimidazole) Sensitivity Inducing Factor] and NELF (Negative Elongation Factor) [81, 82]. DSIF and NELF together can inhibit the movement of Pol II on the DNA template [76, 77, 79, 82]. In a recent study of these elongation factors, it has been established that a transcript at least 18 nt long is necessary for the binding of DSIF and NELF together on the elongation complex [83]. However, earlier, a promoter-proximal pause had been detected before position 18 in transcription of the human *hsp70* gene [84], which suggests that a pause can occur before position 18. Other than the involvement of negative elongation factors for promoter-proximal pausing, some studies support the existence of nucleosomes ahead of Pol II. According to these studies, nucleosomes are responsible for the interruption of the transcript elongation in the promoter-proximal region [76, 84, 85]. The exact function of promoter-proximal pausing is thought to be the rapid induction of transcription (reviewed in [76]). In *Drosophila*, transcription was found to be induced rapidly with a deoxycycline-regulated activator, in a few minutes with a paused Pol II, but in absence of paused Pol II, transcription had not been induced even in 90 minutes [86]. Other than rapid induction, promoter-proximal pausing works as the checkpoint for early transcription, required because of the complexity of the process [77]. It is believed that during the pause, the DNA-RNA hybrid is checked and several DNA associated pathways occur such as capping, DNA methylation etc. [77, 87].

During the movement of Pol II in the promoter-proximal region, newly synthesized pre-mRNA can undergo co-transcriptional processing events. However, these

events are not in the scope of this thesis. A detailed review about these processes including the process of elongation is available in [71].

Several elongation factors help Pol II for escaping from the paused state and to enter into productive elongation. P-TEFb (Positive transcription elongation factor b) is one important factor known to phosphorylate negative transcription factors DSIF and NELF. P-TEFb also phosphorylates the Ser2 (Serine 2) amino acid residue present in the CTD of Pol II [88, 89]. In addition, FACT is another factor characterized in a reconstituted *in vitro* transcription system where it has been shown to assist P-TEFb in allowing the polymerase to escape from the paused state [90]. Other than these factors, TFIIF has also been studied as an important factor to increase the elongation rate [88]. TFIIS, a transcript cleavage factor has also been shown to promote elongation in regions of a gene that are susceptible to backtracking and arrest proximal to the promoter [91]. The above-mentioned factors with some other factors help Pol II to overcome the paused state and to enter into productive elongation to transcribe the remainder of the gene.

The polymerases that do not abort or arrest during the early elongation of the transcript, move rapidly on the DNA template until the termination site is reached. However, movement of Pol II in the productive elongation phase is not simply the synthesis of nascent transcript with fast movement on the DNA template. There are important mechanisms acting during elongation such as the remodeling of chromatin, histone acetylation, covalent histone modifications, etc. [87]. In addition, a high level of transcription elongation gives rise to several DNA-related abnormalities such as DNA recombination and mutagenesis, therefore DNA repair mechanisms are going on with elongation [87]. It is important to note that this thesis will only concentrate on

transcription, excluding these other processes. Many other mechanisms related to transcription elongation and its interplay with other DNA-related processes are reviewed in [87]. There is a lack of information for the regulation of transcription in positions downstream to the promoter-proximal pausing region, but the loss of nucleosomes for elongation indicates a modest level of regulation in post-proximal movement of Pol II [12].

The kinetics of the elongation process has been studied extensively with different genes and promoters in eukaryotes. The rate constant for the complete transcription process has been found to be 1.1-2.5 kilobases (kb)/min using various techniques such as RT-PCR, nuclear run-on assays and fluorescence *in situ* hybridization [92-94]. However, there was no information available in these studies regarding the time spent in initiating and terminating the transcripts. The average length of a human gene is ~14 kilobases (kb) [95]. According to these rate constants, a polymerase should take 6-13 minutes for the completion of half of the human gene transcription cycle, whereas Hiroshi Kimura and colleagues found that the polymerase takes 14-20 min to complete half of the transcription cycle *in vivo* [8], which is significantly higher than the previously observed time. Kimura et al. considered the time taken in initiation and termination as a part of this time. However, they did not come up with the rate constants for the different phases of transcription.

Recently, the transcription dynamics have been studied or reviewed in many papers [9, 11, 96, 97]. Out of these studies, an excellent *in vivo* study, using photobleaching, photoactivation as well as mathematical modeling techniques, reported the fastest elongation rate constant ever [9]. The elongation rate constant reported in this

study was 4.3 kb min^{-1}. It is notable that this rate constant was obtained by modeling calculations to get the rate constant for the rapid elongation, removing the effect of pausing [11]. This study is one of the most detailed studies of transcription dynamics *in vivo* for a mammalian Pol II, reporting the rate constants for most of the steps of transcription.

Two years later, another excellent study by Jarnail Singh and Richard A Padgett reported an average elongation rate constant of 3.8 kb min^{-1} [97]. This rate constant was obtained by averaging the elongation rate constants obtained for 15 different human genes [97]. They used DRB (5,6-dichloro-1-β-D-ribofuranosylbenzimidazole) combined with quantitative RT-PCR for analyzing the transcription process. This rate constant was relatively close to the earlier reported fastest rate constant of 4.3 kb min^{-1}. The small difference between these rate constants is because the rate constant reported by Singh & Padgett includes the possibility of pausing in the proximal region [97]. Therefore, it can be concluded after reviewing these studies that the movement of Pol II is very fast at the positions beyond the promoter-proximal region with no other possibility of pausing, arrest and backtracking. The elongation phase of transcription ends when Pol II reaches the termination site.

1.1.4 Termination

Termination is the last step of transcription, in which the dissociation of the transcription machinery and pre-mRNA transcript takes place [98, 99]. Termination of Pol II transcription in eukaryotes is one of the most poorly understood steps of eukaryotic transcription [100]. However, it is well known that the polyadenylation signal is required

for Pol II termination [98-101]. Several studies established a possible connection between polyadenylation and termination [98, 99, 102]. It has also been shown that termination takes place in the distal region downstream of the poly (A) signal sites (reviewed in [102]). However, there is no information about the precise site of termination *in vivo*, rather it is considered to be stochastic in nature [99].

There are two well-known models for the termination of transcription: the torpedo and allosteric models [98, 103]. Both models explain the connection of polyadenylation signals with termination in the downstream region. According to the torpedo model, during the transcription of the polyadenylation signal site, one of the co-transcriptional processing events, i.e. co-transcriptional cleavage of the transcript, provides an entry point for an exonuclease factor such as Rat1 in yeast [103] or Xrn2 in metazoans [99]. The exonuclease activity of this factor can increase the probability of termination at the pause sites downstream of the poly (A) signal site. It has also been shown in another study on a mammalian promoter that these downstream pause sites may promote transcriptional termination [101]. Alternatively, the allosteric model states that transcription of the poly (A) site can cause a termination-inducing conformational change in the elongation complex [98]. This conformational change may involve the removal of elongation factors opposing termination or the addition of termination factors assisting termination at a termination site [99]. In recent times, this model of termination has become more popular than the torpedo model.

During termination, the transcription machinery dissociates from the DNA template and a nascent inactive pre-mRNA transcript is released. This pre-mRNA transcript can undergo processing to synthesize functional mRNA immediately after

termination or during the transcription process even before termination, as mentioned earlier. However, mRNA processing is out of the scope of this thesis.

1.2 Stochasticity in gene expression/transcription

It is well known that the unique identity of an organism is due to its DNA. However, it was not clear how two organisms could differ in their appearance and behavior even with identical genetic backgrounds e.g. identical twins. Researchers found that the stochasticity or noise in the expression of genes is one of the reasons for the uniqueness of individuals. Noise has been characterized as extrinsic or intrinsic through various experimental studies [104, 105]. The variations that affect two simultaneously expressing genes in the same way have been characterized as extrinsic noise, such as variations in the number of polymerases, ribosomes, etc., whereas the variability due to the stochastic nature of biochemical reactions involved, the small number of molecules involved, random binding of regulatory factors and random occurrence of events in various steps are known as intrinsic noise in gene expression [3, 104-107]. In other words, stochasticity or noise is the inherent nature of eukaryotic as well as prokaryotic gene expression. Several experimental [105, 108] and theoretical [109, 110] studies quantified noise in prokaryotic gene expression and concluded that prokaryotic gene expression is very noisy [106]. This noise can be quantified using the coefficient of variation (CV), which can be obtained by dividing the standard deviation (σ), e.g. in the protein expression level, by the mean expression level (μ), i.e. $CV = \sigma / \mu$. The coefficient of variation is the most frequently used measure for the noise in gene expression. However, some other factors

have also been used to quantify the noise such as noise strength (φ), which is the variance over the mean ($\varphi = \sigma^2 / \mu$) [3], also known as the Fano factor.

Further, efforts have been made to quantify the noise in eukaryotes through joint experimental and theoretical studies [111, 112]. These studies suggested that the noise in eukaryotic gene expression is different from that in prokaryotic gene expression. The rates of transition between different promoter states and promoter accessibility might play an important role in affecting the stochasticity in eukaryotic gene expression [3, 104, 111-113]. All of these studies of eukaryotic gene expression investigated the relation between the mean expression level and the variability in expression from this mean value [106]. Blake et al. [112, 113] found that the stochasticity in gene expression is strongly dependent on the transcriptional rate. This finding suggests a significant role of transcription in creating the noise in eukaryotic gene expression.

Some interesting points are noticeable in all of the studies discussed above: (a) most of the studies described are concentrated on the gene expression as a whole, not on transcription, (b) all the experimental studies quantify the protein expression level to determine the noise in the system, and (c) all the theoretical studies present an on-off model of gene expression and ignore the biochemical details involved in the steps of gene expression. These studies gave insights into the stochasticity in gene expression but could not identify the steps responsible for this stochasticity. At most, one can only identify transcription as a major source for the stochasticity in gene expression. Further, it was not evident from these studies which of the steps of transcription or translation are most significantly responsible for the stochasticity in gene expression.

Some efforts have been made to study the stochastic kinetics of transcription through simple models of a single gene [114, 115]. The model in [115] was able to capture the stochastic kinetics of a fairly simple prokaryotic transcription model including initiation, elongation and termination steps by using the stochastic simulation algorithm (SSA) [116] and chemical master equation (CME) [117, 118]. However, this model ignored various biochemical details of the transcription to keep the model analytically solvable through CME [115]. In the very same year, Roussel and Zhu introduced a new version of SSA, the delayed SSA, to study prokaryotic gene expression [119]. This algorithm introduces a delayed output instead of including all the steps of transcription and translation [119]. The delayed SSA emerged as a useful algorithm to approximate the stochastic kinetics of large gene regulatory networks involving many biochemical reactions because of the low computational cost [119, 120]. However, it is an approximation to the exact SSA [119].

A number of theoretical studies investigated the stochastic aspects of prokaryotic transcription and gene expression in a more detailed manner [121-127]. Most of these stochastic studies examined the dynamics of prokaryotic transcription or a complete gene expression pathway [121-124, 127], using the delayed SSA [119] and used the framework of the pioneering model of transcription, developed by Roussel and Zhu [115]. Of the studies mentioned above, the most interesting one presents the delayed model of transcription at the single nucleotide level [121], in which the prokaryotic transcription process has been modeled in the most detailed form. This model accounts for the promoter open complex formation with other alternative pathways including pausing, arrest, misincorporation and editing, pyrophosphorolysis and premature

termination. The transcription dynamics was studied and, later on, this model was extended to study gene expression dynamics [122]. However, the central theme of the many of these studies was to study the effects of pausing on prokaryotic transcription or gene expression [121, 122, 124]. Very recently, another study considered the coupled transcription and translation as an extension of these models of gene expression [127].

Most of the eukaryotic studies until recently were concentrated on the kinetic or mechanistic aspects of eukaryotic transcription. Recently, some experimental studies concentrated specifically on the stochastic nature of transcription in metazoans [128, 129]. These single-cell studies suggested that transcription pulses arise due to the intrinsic randomness of the activation and inactivation of a gene rather than to extrinsic sources of variability such as number of transcription factors, etc. [129]. In addition, a recent review on transcriptional bursting identified the events occurring before promoter escape, i.e. preinitiation complex assembly with initiation and promoter-proximal pausing (early elongation), as the possible reasons for the transcriptional bursting [130].

On the theoretical front, until now, no efforts have been made towards the study of transcription dynamics in eukaryotes, similar to those in prokaryotes. To my knowledge, there is not a single model of eukaryotic transcription or gene expression available that can efficiently study the roles of individual steps involved in eukaryotic transcription or translation in causing the stochasticity. However, as explained before, some efforts have been made to quantify the total noise (extrinsic and intrinsic) with combined application of experiments and theory [104, 111-113].

In this thesis, a sufficiently detailed biochemical model of eukaryotic transcription will be developed, which includes PIC assembly, abortive initiation, promoter-proximal

pausing and termination as the points that can be slow steps for transcription. In addition, the stochastic properties of this process at the single nucleotide level using a variant of the classic SSA will be studied. The classic SSA is considered as one of the most suitable methods for the detailed simulations of a biochemical process such as transcription. Two different variants of the transcription model are presented for SSA analysis: Simulations have been performed in single and multiple polymerase cases to study the effects of included slow steps on the eukaryotic transcription dynamics and to distinguish effects due to the dynamics of a single polymerase from those that might arise from interactions between polymerases. In addition, analytic probability distributions for the single-step movement of Pol II in various stages of eukaryotic transcription are obtained for the single polymerase case. This mathematical analysis can be further used to develop expressions for the complete probability distribution of transcription times.

Chapter 2

Model description

Eukaryotic transcription is a complex process. For the development of this model, the biochemistry involved in transcription has been considered in detail. All the steps of eukaryotic transcription described in chapter 1 are included in this model. However, some mechanistic assumptions have been made due to the lack of available information. These assumptions are aimed to keeping the model analytically solvable with reduced complexity without losing important biochemical details. As described earlier, transcription is divided into five different phases: pre-initiation complex formation, initiation, promoter escape including abortive initiation, elongation with promoter-proximal pausing, and termination. This chapter will describe the development of this model. The biochemical reactions for the steps of transcription will be discussed with the criteria for the selection of kinetic parameters for these reactions. In addition, the different assumptions made to model these biochemical reactions will also be discussed in this chapter.

2.1 Model reactions

The first phase of eukaryotic transcription is PIC assembly. PIC assembly itself is a very complex process, which involves binding of many general transcription factors (GTFs) on the TATA box [20, 23]. Out of the two known mechanisms for the assembly of the pre-initiation complex (section 1.1.1), the mechanism that involves the binding of Pol II holoenzyme is modeled here.

2.1.1 PIC assembly reactions

On the basis of the information from the reviewed literature in Chapter 1, PIC assembly is modeled as a two-step process in this model: (1) the binding of TBP/TFIID/TFIIA to the TATA box or to an A-T rich region in the promoter and, (2) the binding of Pol II holoenzyme on the TBP/TFIID/TFIIA/TATA-box complex.

Step A of Figure 2.1 shows TBP dimerization, TBP dimer dissociation, and TBP/TFIID monomer complex formation. Step B shows the binding of the TBP/TFIID complex on the TATA box. The binding of the TBP monomer to DNA and the dimerization of TBP monomers are two competing processes. TBP remains largely in the form of dimer in the nucleus when not bound to the DNA because of the slow dimer dissociation [32, 33]. Therefore, the small rate constant of the TBP dimer dissociation can dictate the rate constant for TBP/TFIID binding to the TATA box. It is important to note that the formation of TBP/TFIID monomer complex before binding to DNA is a mechanistic assumption made due to the lack of information about the sequence of events in this first step of PIC assembly. However, this assumption probably wouldn't make much difference since the TBP dissociation is so slow. Another notable thing for this mechanism is the role of TFIIA. TFIIA has been found to play an important role in loading the TBP/TFIID complex on the DNA [34, 35]. TFIID dimers have been found to dissociate with a half-life of similar magnitude as TBP dimers ($t_{1/2} \approx 10$ min) due to the interaction with TFIIA [32, 35].

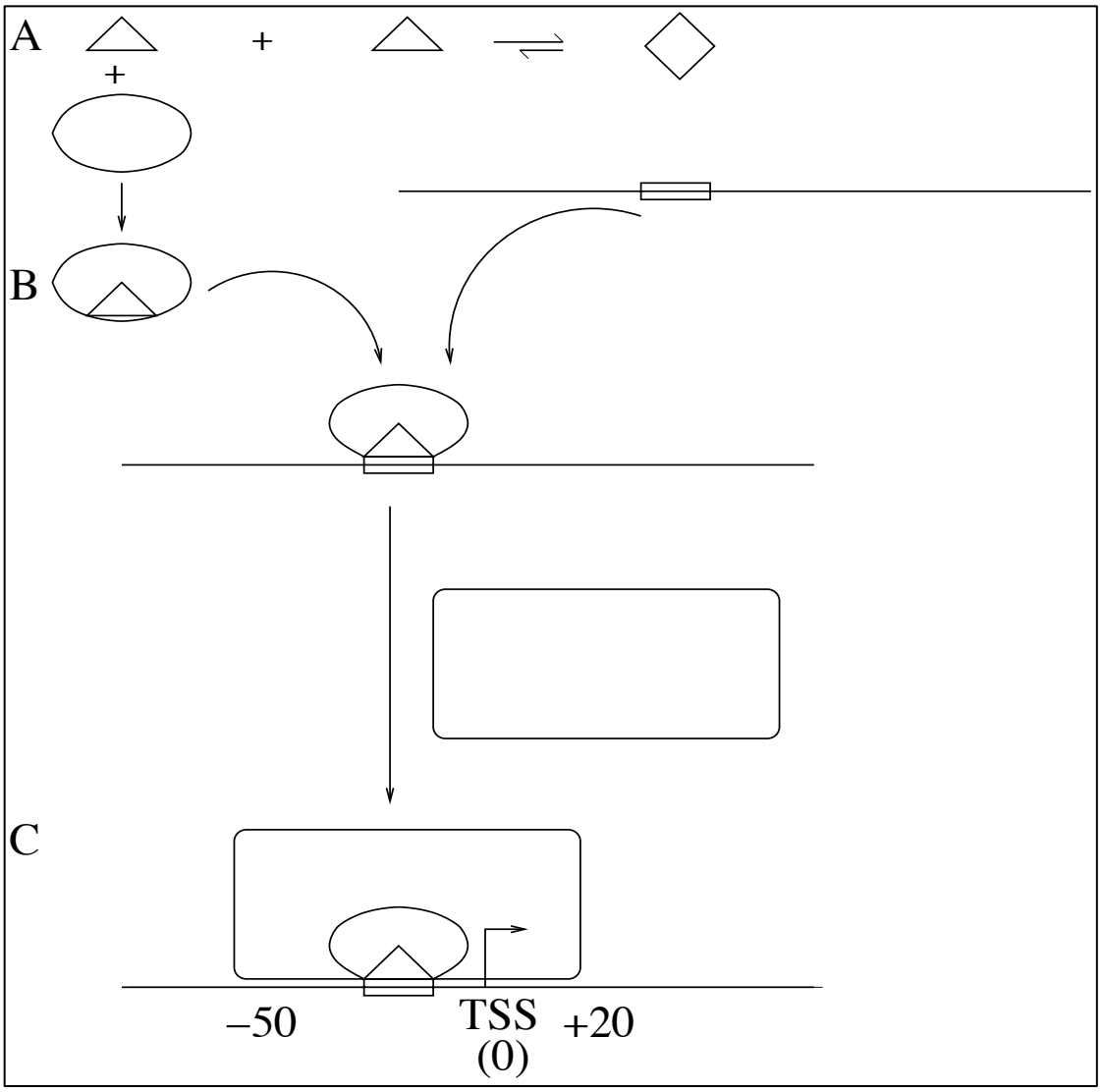

Figure 2.1: Schematic diagram of pre-initiation complex assembly in the model: In A, the TBP monomer (triangle) binding to TFIID (oval) is in competition with the formation of the TBP dimer (diamond) and forms the TBP/TFIID complex prior to binding to the TATA box (rectangle). In B, the TBP/TFIID complex binds to the TATA box (rectangle) present on the DNA template (solid line). In C, the RNAP holoenzyme composed of Pol II, TFIIF, TFIIE, TFIIH and TFIIB (large box) binds to the pre-bound TBP/TFIID complex and forms the pre-initiation complex at the transcription start site (TSS). The TSS coincides with the position of the active site of Pol II (bent arrow) at the end of PIC assembly. The front end of Pol II is 20 nt downstream (+20) and the back end is 50 nt upstream (-50) of the TSS.

In the absence of TFIIA, dimer dissociation of TBP and TFIID has been found to be extremely slow ($t_{1/2}$ >20 and ~28 min for TBP and TFIID respectively) [32, 35]. Therefore, TFIIA can be a major player in regulation of TBP/TFIID binding on the DNA and in stabilizing the TBP/TFIID/TATA-box complex. However, TFIIA binding does not represent a slow step. It is notable that the model presented here does not consider TFIID dissociation because in the studies of TFIID dissociation, TBP has been considered as the part of TFIID. Therefore, both factors have been considered as a single entity.

In this model, TBP and/or TFIID dimer dissociation represent major slow steps in pre-initiation complex assembly. Therefore, the TBP dimer dissociation rate constant has been explicitly used as the rate constant for TBP/TFIID binding to the DNA template, replacing the rate constants for the TBP/TFIID monomer complex formation, TBP/TFIID complex binding to DNA and the binding of TFIIA on the TBP/TFIID complex pre-bound to the TATA box.

In the model reaction (1), TBP represents the complex of TBP, TFIID and TFIIA, pro is the TATA box or an AT-rich region of the core promoter, TBP.pro represents the TBP/TFIID/TFIIA/TATA-box complex and k_{bind} represents the stochastic rate constant for the dissociation of the TBP dimer into TBP monomer and the binding of the TBP/TFIID/TFIIA complex to the TATA box together, the former being known to be slow.

$$TBP + pro \xrightarrow{k_{bind}} TBP.pro \qquad (1)$$

After the formation of TBP/TFIID/TFIIA/TATA-box complex $(TBP.pro)$, Pol II holoenzyme binds to this stable complex. Pol II holoenzyme consists of the transcription factors TFIIB, TFIIE, TFIIF and TFIIH. According to the ordered recruitment mechanism

for the binding of the transcription factors, the binding of Pol II/TFIIF represents the only slow step in the PIC assembly after the formation of the TBP/TFIID/TFIIA/TATA-box complex [54]. Therefore, the rate constant for the binding of Pol II/TFIIF has been used as the rate constant for the binding of Pol II holoenzyme in this model. In reaction (2), $RNAP_{holo}$ represents the Pol II holoenzyme with other GTFs, PIC represents the pre-initiation complex and k_{PIC} represents the rate constant for the binding of Pol II holoenzyme on the TBP/TFIID/TFIIA/TATA-box complex.

$$TBP.pro + RNAP_{holo} \xrightarrow{k_{PIC}} PIC \tag{2}$$

To study the consequences of the rate constant of PIC assembly, it has also been considered as a fast process in the simulations of this model. For the fast version of PIC assembly, reactions (1) and (2) have been replaced by the following reaction with a single rate constant, k_{FPIC}, for fast PIC assembly:

$$pro + RNAP_{holo} \xrightarrow{k_{FPIC}} PIC \tag{2a}$$

2.1.2 Initiation, abortive initiation and promoter escape reactions

The initiation phase of transcription involves the early movement of Pol II on the DNA template strand. In this model, the front end of RNA Pol II is ~20 nucleotides downstream of the transcription start site (TSS) and its back end is ~50 nucleotides upstream of the TSS at the start of transcription (Figure 2.1), which satisfies the cross-linking contacts of RNA Pol II subunits [50, 64]. Therefore, during the movement of Pol II, active sites of consecutive polymerases remain at least $\Delta = 70$ bp apart. The position of Pol II is given by the position of the active site in the model.

After the establishment of a stable pre-initiation complex, DNA melts and forms a transcription bubble, and Pol II moves on the DNA template with the formation of the first phosphodiester bond. Reaction (3) represents the translocation of the active site of the Pol II to the first unoccupied (U_1) nucleotide and the conversion of this unoccupied nucleotide into the occupied (O_1) nucleotide, which is possible, only if the nucleotide located 70 bp downstream from site 1 is unoccupied (U_Δ).

$$PIC + U_\Delta \xrightarrow{k_{ini}} O_1 \tag{3}$$

In this model, three states of nucleotides have been used: unoccupied (U), occupied (O) and activated (A). These nucleotide states occur during the movement of Pol II on the DNA template strand. Movement of Pol II from a position i to $i+1$ completes in two steps: First, the active site of the Pol II moves to a particular unoccupied position (U_i) and changes it to the occupied state (O_i), similar to the reaction (3) for the very first movement of Pol II. Second, Pol II is activated by the binding of NTP molecules to the Pol II, changing the occupied site (O_i) to the activated state (A_i) and making Pol II ready for translocation to the next unoccupied nucleotide (U_{i+1}) (Figure 2.2). Splitting of the translocation process into two steps like this can be arranged so that any desired average translocation time is obtained, but it affects the distribution of completion times, which is exponential for a single step, but has a nonzero maximum for two steps. Even if we don't know how the rate constants should be partitioned over the two processes, if we know that something happens in two or more steps, then we have studied an important aspect of the statistics by splitting the overall process into two steps.

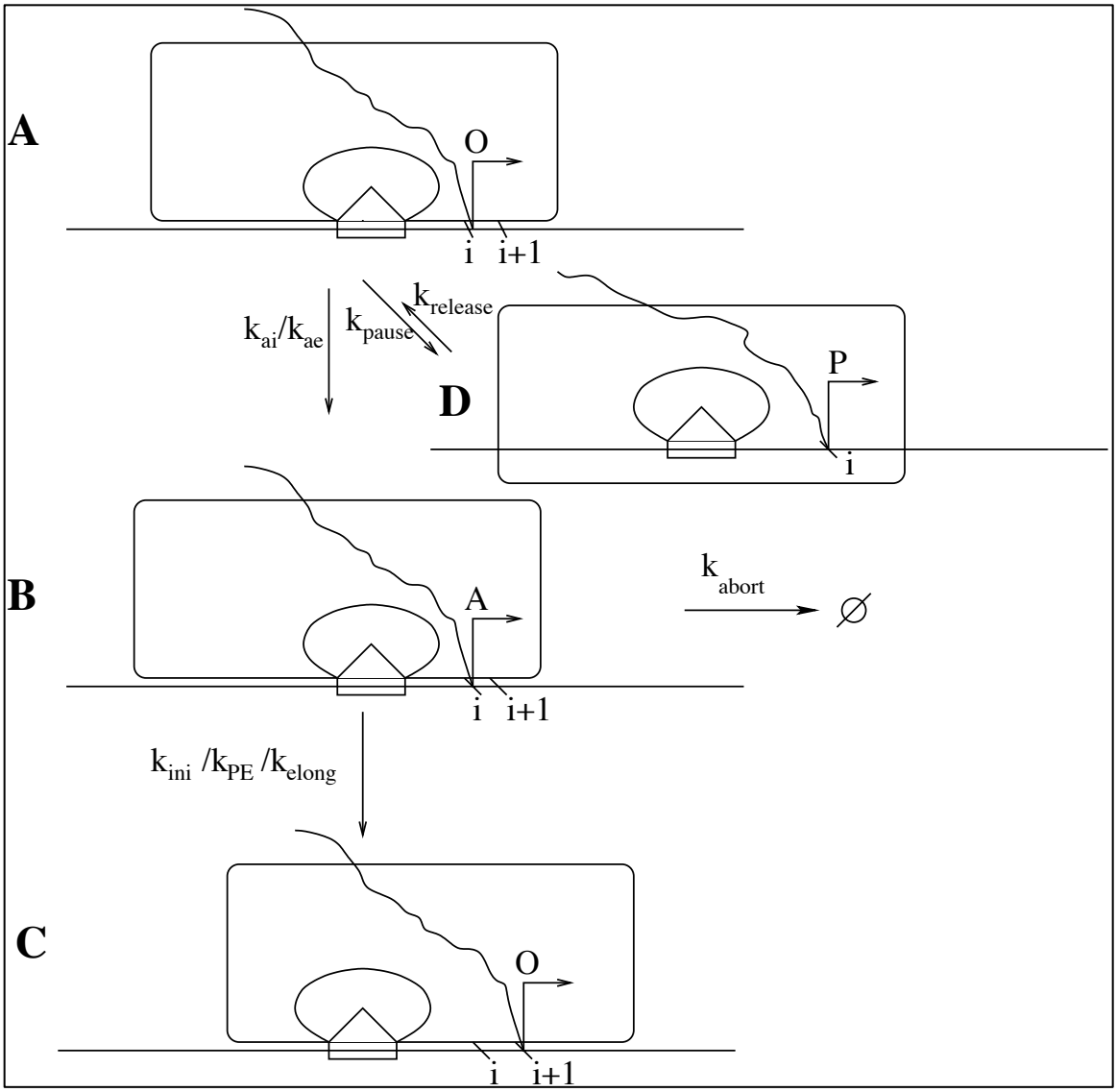

Figure 2.2: Schematic diagram of the states of nucleotides representing the movement of Pol II: In these diagrams, the large box represents Pol II holoenzyme, with its active site represented by the bent arrow. The small rectangle represents the TATA box, the triangle represents TBP, and the small oval shape represents TFIID. In A, PIC assembly is complete and the nucleotide i to be transcribed in the early transcription region has become occupied. In B, NTP binding has occurred, and the polymerase has become activated (A_i) for translocation (represented by *). This transcription machinery can abort from the activated state in positions 4-15. In C, the active site has moved to position $i+1$ and position i becomes unoccupied (U_i). D represents the paused state in which the polymerases pause at a particular position for some time in the promoter-proximal region. Note that the polymerase is not drawn to scale

Translocation of Pol II (reaction 5) is only possible if there is not another polymerase active site 70 nt downstream, the minimum possible distance between the two consecutive active sites [50, 64]. Reaction (4) represents the activation of Pol II during motion over the first 15 positions during the promoter escape phase of transcription. Reaction (5) of the model represents the translocation of Pol II over nucleotide positions 2-15.

$$O_i \xrightarrow{k_{ai}} A_i,\ i = 1,2,...,15 \tag{4}$$

$$A_i + U_{i+\Delta} \xrightarrow{k_{PE}} O_{i+1} + U_i,\ i = 1,2,...,15 \tag{5}$$

As described in section 1.1.2, the initial transcription events are prone to abort on the first few positions. The escape commitment transition, which occurs after 4 nucleotides have been added to the nascent transcript, indicates a commitment to complete the promoter escape phase by Pol II [53, 54, 56, 57]. However, 4-15 nt long abortive transcripts have been found in biochemical studies. Reaction (6) represents the abortive initiation in this model:

$$A_i \xrightarrow{k_{abort}} TBP + RNAP_{holo} + RNA_{abort} + U_i,\ i = 4,5,...15 \tag{6}$$

In reaction (6), a mechanistic assumption of the break down of the transcription machinery in abortive initiation has been made. Specifically, it has been assumed that the RNA polymerase immediately dissociates from the DNA template with other transcription factors. This may not be the case *in vivo* but there is a lack of experimental information on the mechanism of abortive initiation in eukaryotes. Another important thing to note in the reaction (6) is the hypothesis that abortive initiation occurs from the

activated state (A_i). It has been found in prokaryotic transcription through a single molecule study that abortive initiation takes place during the forward translocation of the front end of the polymerase [68]. In this model, the activated state of the nucleotide is the state when Pol II is ready for the forward translocation. Therefore, the RNA-DNA hybrid may become weaker during the activated state in comparison to the occupied state. However, the exact mechanism for this process is known neither for prokaryotes nor for eukaryotes.

Reactions (3)-(5) of this model also include the promoter escape step of transcription that completes with the formation of a 14-15 nt long transcript [61]. Promoter escape has also been studied as a possible slow step in eukaryotic transcription in many biochemical studies [53, 54, 56, 64].

2.1.3 Promoter-proximal pausing and elongation reactions

In early elongation, promoter-proximal pausing has been studied as a possible slow step as well as a checkpoint for elongation of the nascent transcript. The possible mechanism and the available kinetics have been discussed in detail in section 1.1.3. Promoter-proximal pausing in this model has been defined as a state of the nucleotide (Figure 2.2). In addition, promoter-proximal pausing is considered to occur in this model between positions 16-50 instead of the conventionally quoted range of 20-50. The reason for choosing the 16-50 region for promoter-proximal pausing is the detection of a pause before position 20 in the human *hsp70* gene [84]. In this model, the human transcription data is the major focus for the parameters. Reaction (7) represents the entrance into the paused state and reaction (8) represents the exit from the paused state.

$$O_i \xrightarrow{k_{pause}} P_i,\ i = 16,17,...,50 \tag{7}$$

$$P_i \xrightarrow{k_{release}} O_i,\ i = 16,17,...,50 \tag{8}$$

It is important to note that in this model of eukaryotic transcription, pausing proceeds from the occupied state (O_i) of the nucleotide. This is an assumption of this model, made necessary by the lack of information on the mechanism of pausing. In these model reactions, pausing is treated as a Poisson process, potentially allowing a polymerase to pause many times in the promoter-proximal region. This is also a mechanistic assumption of our model.

The transcription events that do not abort or arrest undergo productive elongation of the transcript. Productive elongation is a rapid phase of transcription (section 1.1.3). In the model, the mechanism of movement of Pol II on the DNA template is similar in the initiation and elongation phases (Figure 2.2). It is important to note that the elongation phase of this model does not consider other alternative pathways such as premature termination, backtracking and transcription arrest. These pathways have been found to have a very small effect on the transcription dynamics in prokaryotic transcription [121]. Therefore, an even smaller effect can be expected in eukaryotic transcription, the latter being more regulated. Reactions (9) and (10) represent the productive elongation phase of this model.

$$A_i + U_{i+\Delta} \xrightarrow{k_{elong}} O_{i+1} + U_i,\ i = 16,17,...,n-1 \tag{9}$$

$$O_i \xrightarrow{k_{ae}} A_i,\ i = 16,17,...,n \tag{10}$$

2.1.4 Termination reactions

Little is known about the termination of eukaryotic transcription. However, two models for the termination of eukaryotic transcription are commonly proposed: the torpedo and allosteric models (section 1.1.4). Of these models, the latter has become more widely accepted in recent times [99]. Therefore, the allosteric model has been followed to model the termination reactions. For this purpose, termination has been divided into two steps: first, the formation of a termination complex and second, the dissociation of transcription machinery from the DNA template. It is important to note that in the model, the conformational change of the elongation complex to a termination complex (reaction 11) takes place at the last nucleotide of the gene to be transcribed. Polyadenylation and other mRNA processing steps have not been included. Processing steps such as splicing and mRNA transport to cytoplasm are interesting phases of gene expression to model in themselves. The other reason for not including these processes is that we wanted to capture the dynamics of transcription alone, which will help us to get insights into the effects of various steps on the transcription times.

Reaction (11) represents the conversion of an elongation complex into a termination complex, which is similar to the conformational change step of the allosteric model, and reaction (12) represents the dissociation of termination complex, releasing a complete pre-mRNA of the transcribed gene.

$$A_n \xrightarrow{k_{TC}} TC \tag{11}$$

$$TC \xrightarrow{k_{term}} RNA + RNAP_{holo} + U_n \tag{12}$$

Note that in reaction (11) the conformational change occurs from the activated state of the last nucleotide (A_n). It is an assumption made due to the lack of experimental information about the state from which termination takes place.

It is important to note that in reaction (12), RNA represents the pre-mRNA transcript**.** This pre-mRNA transcript may undergo co-transcriptional mRNA processing and splicing. Alternatively, it can enter these processes right after the synthesis of the transcript. However, the synthesis of the pre-mRNA transcript through transcription, neglecting processing of the primary transcript, is the focus of this study.

2.2 Kinetic parameters

Stochastic modeling of a dynamic process such as transcription depends on the stochastic rate constants involved in the various phases of the process. These stochastic rate constants can be derived from the experimentally measured rate constants. Here, the kinetic parameters with the selection criteria for those parameters are presented.

Stochastic models involve a reaction propensity, which is analogous to the rate of reaction in a deterministic model [116]. The propensity in turn involves a stochastic rate constant, which is the specific rate at which probability is transferred from one state to another. In reactions (1) and (2), the states would be a bare promoter (pro), the TBP.pro complex, and the pre-initiation complex (PIC). Since we deal with a single gene, and given the rate-limiting processes in reactions (1) and (2) discussed in detail in chapter 1, these stochastic rate constants are equivalent to the pseudo-first-order rate constants measured *in vivo* [32, 54].

The biological significance of the rate constants for reactions (1) and (2) can be understood by considering the PIC assembly process in detail. Only a couple of rate-limiting steps can replace several steps involved in the PIC assembly without losing any kinetics of the process. Several kinetic studies of the interaction of TBP with DNA have represented the binding of the TBP monomer to DNA by a fast single second-order rate constant of approximately 5×10^5 $M^{-1}s^{-1}$ [30, 131, 132]. In addition, the rate constant for TBP dimerization, calculated from the dissociation constants and concentrations provided in [32] is 10^5 $M^{-1}s^{-1}$ (reviewed in [23]). Many studies suggest that TBP dimers cannot bind to DNA [31, 33]. On the other hand, the dissociation rate constant for the TBP dimer has been found to be 1.6×10^{-3} s^{-1} [32]. The kinetic details of the binding of TBP to TFIID and the binding of TFIIA to TBP/TFIID are not known to date but these processes do not represent the slow steps in PIC assembly. Therefore, the dissociation rate constant for TBP dimer can be used as the rate constant for the assembly of the TBP/TFIID/TFIIA complex on the TATA box. A kinetic study by Coleman and Pugh [32] directly supports this modeling decision.

On the other hand, the rate constant of the binding of Pol II holoenzyme to the pre-bound TBP/TFIID/TFIIA complex can be interpreted as the rate constant for the binding of the Pol II/TFIIF complex. The binding of the Pol II/TFIIF complex has been found as the only possible slow step in the ordered recruitment mechanism of transcription factors (discussed in section 1.1.1) [54]. Taking these facts into consideration, I determined the stochastic rate constant for the Pol II holoenzyme binding as $k_{PIC} = 0.0029 s^{-1}$, which is similar to the pseudo-first order rate constant of the binding of Pol II/TFIIF complex on the TBP/TFIID/TFIIA complex [54]. The rate constants for

the binding of TBP and Pol II holoenzyme are the rate constants measured *in vitro* that might be representative of slow PIC assembly *in vivo*. In addition, the case where PIC assembly is a fast process is also studied in this thesis. For this purpose, the reactions (1) and (2) are replaced by reaction (2a) and k_{bind} and k_{PIC} are replaced by a single fast rate constant for PIC assembly i.e. $k_{FPIC} = 0.1s^{-1}$. Fast PIC assembly has been explored on the basis of two kinetic studies [53, 56]. In these studies, early movement of Pol II i.e. promoter escape has been studied on a minimal transcription system with fast PIC assembly. In this thesis, a full range of PIC assembly rate constants implied by these data has been studied. All the kinetic parameters used are summarized in Table 3.1.

Further, reactions (3)-(6) represent the movement of Pol II on the first 15 positions that includes initiation, abortive initiation and promoter escape. Stochastic rate constants k_{ini}, k_{ai} and k_{PE} of reactions (3)-(5) have been calculated with the help of two-step calculation process. First, the stochastic rate constant for the translocation of Pol II at each of the first 15 positions i.e. $k_{trans} = 0.3s^{-1}$ has been calculated from the observed rate constant $\left(k_{obs} = 0.025s^{-1}\right)$ obtained in a biochemical study [54] (*vide infra*). This observed rate constant $\left(k_{obs}\right)$ is the rate constant for the translocation of polymerase through the first 28 positions of the interleukin-2 promoter [54]. Second, two equal rate constants k_{ini} or k_{PE} and $k_{ai} = 0.6s^{-1}$, representing the translocation of Pol II and the state change from occupied to activated respectively, have been calculated from k_{trans}. In this study, the expression $k_{trans} = \left(1/k_{ini} + 1/k_{ai} \right)^{-1}$ has been used to calculate the values of k_{ini} and k_{ai}, which together represent the movement of Pol II on the first nucleotide.

Similarly, k_{PE} and k_{ai} have been calculated by the expression $k_{trans} = \left(1/k_{PE} + 1/k_{ai} \right)^{-1}$ and collectively represent the movement of Pol II on positions 2-15.

It is important to know how an observed rate constant for movement through many sites, i.e. $k_{obs} = 0.025 s^{-1}$, was converted to a stochastic rate constant for translocation through a single site, i.e. $k_{trans} = 0.3 s^{-1}$, and then how this stochastic rate constant was divided into $k_{ini} = k_{PE} = k_{ai} = 0.6 s^{-1}$. As described above, k_{obs} represents the rate constant for the movement of Pol II through the first 28 positions. The movement of Pol II on the first 4 nucleotide positions (before escape commitment) has been found to be very fast in comparison to the movement after escape commitment on the first 28 nt positions [54]. Out of these remaining 24 positions, Pol II is supposed to move faster with the productive elongation rate constant $\left(k_{move} = 72 s^{-1} \right)$ after the completion of promoter escape, i.e. on positions 15-28. Therefore, the slow movement of Pol II in the promoter escape phase, i.e. in positions 4-15, is the main contributor to the observed rate constant. The observed rate constant for the movement of Pol II in positions 4-28 represents the effective rate constant obtained by the slow movement of Pol II in positions 4-15 and the fast movement in positions 15-28. Therefore, we calculated the stochastic rate constant for the movement in positions 4-15 $\left(k_{trans} = 0.3 s^{-1} \right)$ by using the expression $1/k_{obs} = n_{slow} \cdot 1/k_{trans} + \left(24 - n_{slow} \right) \cdot 1/k_{move}$. Here, $n_{slow} = 11$ represents the number of nucleotide positions for slow movement of Pol II i.e positions 4-15, and $k_{move} = 72 s^{-1}$ is the rate constant for productive elongation obtained on a human gene [9]. Given the lack of data on the relative sizes of k_{ini}, k_{PE} and k_{ai}, I chose to make them all the same in the

default parameter set of the model i.e. $k_{ini} = k_{PE} = k_{ai} = 0.6s^{-1}$, which is consistent with the observed rate constant $k_{obs} = 0.025s^{-1}$ for positions 4-28 [54]. The consequences of this parameter selection are explored in chapter 3.

Reaction (6) of the model represents abortive initiation. There is no information available on the kinetic rate constants for abortive initiation. Therefore, the stochastic rate constant for abortive initiation has been estimated on the basis of the information gathered from the literature in section 1.1.2 of this thesis. We estimate that 60-70% of the initiation events abort. This estimation has been made on the basis of two different *in vitro* observations. In eukaryotes, more than half of the initiation events have been found to lead to abortive initiation [67] and ~75% of the initiation events have been found to be aborted in prokaryotes [69]. We set the stochastic rate constant for abortive initiation to $k_{abort} = 0.05s^{-1}$. With this rate constant and the other parameters of the model, 60-70% of the transcription initiations abort. The abortive initiation rate constant may not be accurate for the conditions in reality but it will help to explore the effects of abortive initiation on transcription dynamics. Moreover, these rate constants are always ready for fine-tuning in this model.

Next, reactions (7) and (8) represent the entrance in the paused state and the exit from the paused state, respectively, in early elongation. As explained earlier, the possibility of pausing has been considered in the region of positions 16-50 in this model. There are contradictory views about the percentage of polymerases that pause in the promoter-proximal region (for details, see section 1.1.3). In this study, promoter-proximal pausing has been considered as a common phenomenon in eukaryotic

transcription, assuming that more than 50% of the polymerases pause at least once in the promoter-proximal region. For the default parameters of this model, $k_{pause} = 3.8s^{-1}$ has been estimated as the stochastic rate constant to enter the paused state and $k_{release} = 0.002s^{-1}$, as the stochastic rate constant to exit from the paused state. It is noticeable that these estimated stochastic rate constants agree with the biochemical observation of fast entry into the paused state and slow escape from this paused state [75].

Reactions (9) and (10) represent the elongation phase of transcription. The elongation phase starts just after the completion of promoter escape and ends at the termination site. In other words, the fast movement of Pol II continues from position 16-*n*. In this model, $n = 14$ kb i.e. the average length of a human gene [95]. It is important to mention that promoter-proximal pausing is not a separate phase of transcription. It may occur during the early elongation, i.e the fast movement of Pol II in positions 16-50. After position 50, elongation proceeds without pausing. The rate constant for elongation has been calculated through the solution of a mathematical model as 4.3 kb/min [9], which is equivalent to a rate constant $k_{move} = 72s^{-1}$. k_{move} represents the rate constant for the translocation of Pol II from one position to other. Therefore, it can also be broken into two equal parts with the expression $k_{move} = \left(1/k_{elong} + 1/k_{ae} \right)^{-1}$ to relate with the two-step movement of Pol II assumed in this study. The values of $k_{elong} = k_{ae} = 144s^{-1}$ are arbitrary choices given the lack of the data enabling us to distinguish these two rate constants.

However, the consequences of altering these rate constants have been studied in this thesis (Chapter 3).

Reactions (11) and (12) represent termination in this model. All the biochemical details to consider termination as a two-step process have been provided earlier (sections 1.1.4 and 2.1.4). There is a lack of literature describing kinetics of termination. However, a recent *in vivo* study of transcription dynamics [9] provided the rate constants involved in various steps of transcription through *in vivo* measurements as well as the solution of a mathematical model. The first-order rate constant for termination $k_T = 0.0016 s^{-1}$, provided in the above-mentioned study, has been used for calculating k_{TC} and k_{term}. $k_{TC} = k_{term} = 0.0032 s^{-1}$ has been calculated by the expression $k_T = \left(1/k_{TC} + 1/k_{term} \right)^{-1}$, similarly to the initiation and elongation rate constants. Again, due to the lack of experimental data, the two different steps have been assumed to take the same amount of time with similar rate constants.

Chapter 3

Stochastic simulations

Small molecular populations are important sources of stochasticity in chemical or biochemical systems. Systems with small numbers of molecules can be understood through the simple example of a eukaryotic cell. A typical eukaryotic cell contains zero, one or two (depending on promoter strength) copies of an active gene at a time, which can be transcribed into a few mRNAs through a process involving the random collisions of many transcription factors with DNA. Each mRNA can be translated into 100 or so proteins on average [1, 4]. Therefore, the steps of gene expression typically involve small numbers of molecules. This kind of system can be modeled through stochastic modeling techniques such as the chemical master equation (CME) and the stochastic simulations algorithm (SSA), treating the system as spatially homogenous.

The chemical master equation (CME) and stochastic simulation algorithm (SSA) are the most popular ways to study the time evolution of chemical and biochemical systems having small numbers of molecules. The CME is a potentially infinite set of ordinary differential equations (ODEs) for the probabilities of the various states of the system (number of molecules of each kind, states of nucleotides, etc.), and most of the time, it is hard (or impossible) to solve analytically. The SSA generates realizations of the stochastic process whose probability distribution evolves according to the CME [131]. Stochastic simulations of these models are helpful to study the role of individual biochemical reactions in the creation of stochasticity. In addition, simulations allow us to study the interdependence of these reactions on each other and their combined effects on

the dynamics of the system. This chapter will review the classic SSA with an overview of some other available versions of SSA, followed by the description of a new version of the SSA used for studying this model. At last, the results obtained by the simulations will be discussed for the single polymerase and multiple polymerase variants of the model.

3.1 Classic stochastic simulation algorithm (SSA) and its variants

The stochastic simulation algorithm was proposed by Gillespie as an exact method for generating realizations of the time evolution of a well-mixed chemical system [116, 131]. The SSA represents an alternative to the traditional procedure of solving ODEs for ordinary chemical systems where fluctuations are insignificant [131]. It is also a method for numerically examining the predictions of the master equation. The SSA is a Monte-Carlo simulation algorithm that generates a sequence of reaction events and of time intervals between the events in a chemical or biochemical system [131].

The classic SSA is also known as the direct method (DM). High computational cost is one of the major drawbacks of the classic SSA [116]. It may take an enormous amount of time for a large system. Many efforts have been made to improve the classic SSA such as the first reaction method [131], sorting direct method (SDM) [132], next reaction method (NRM) [133], optimized direct method (ODM) [134], etc. Out of these methods, NRM and ODM were the most effective to reduce the computational cost. However, the SDM is a further improvement of the ODM in terms of computation cost for some specialized gene regulatory network models. Most of these alternative methods to improve the performance of the classic SSA have been discussed in [134, 135]. Other than these, some approximation methods have also been developed to increase the

computational efficiency at the cost of accuracy such as the tau leaping method [136] and the delay SSA [119, 137].

In the classic SSA, each reaction R_υ, has a reaction propensity function (a_υ). The propensity function (a_υ) gives the probability of the occurrence of a reaction (υ) in a small time period $(\tau + \partial\tau)$. This reaction propensity function can be calculated with the help of the stochastic rate constant (c_υ) and the combinatorial function (h_υ) obtained through the stoichiometry of the reaction R_υ.

$$a_\upsilon = c_\upsilon \cdot h_\upsilon,\ \upsilon = 1,...,M \tag{13}$$

The theory of stochastic kinetics defines the stochastic rate constant (c_υ) as the probability per unit time for the occurrence of that reaction for a given set of reactant molecules, and the combinatorial function (h_υ) is the number of different reactant sets that can be made at given composition for example, for a reaction $A + B \rightarrow C$, if there are N_A molecules of A and N_B molecules of B, Then $h_\upsilon = N_A \cdot N_B$.

The total propensity of all the possible reactions can be calculated by summing up all the propensities of these reactions, as follows:

$$a_0 = \sum_{\upsilon=1}^{M} a_\upsilon,\ \upsilon = 1,...,M \tag{14}$$

where M is the number of reaction channels. The next reaction to occur and the reaction time τ can be chosen through a random trial, which depends on the propensity functions of the reactions. The probability that reaction υ occurs is proportional to a_υ. Then the number of molecules and time can be updated in the classic SSA.

In simulations of this model with multiple polymerases, the classic SSA would be awkward because of the variable number of reactions occurring in each time step. To avoid these difficulties, the SSA has been applied in a slightly different manner to choose the next polymerase to react and the reaction that the selected polymerase undergoes. The total propensity for a polymerase has been defined as the sum of the propensities of the possible reactions (a_x), it can do in a particular step. Therefore,

$$a_x = \sum_{\upsilon=1}^{M} a_\upsilon \,,\ \upsilon = 1,...,M \,. \tag{15}$$

Here, x represents the number of the polymerase ($x = 1,...,n$) and M represents the number of reaction channels available to a given polymerase. The total propensity for all polymerases on the strand has been calculated by adding all the propensities of the polymerases, similarly to the propensities of reactions.

$$a_{sum} = \sum_{x=1}^{n} a_x \,,\ x = 1,...,n \tag{16}$$

The next polymerase to react can be determined by applying an SSA step, i.e. by using a pseudo-random number r_1 generated from a uniform random number (URN) generator as follows:

$$a_x = \sum_{x=1}^{\mu-1} a_x < r_1 \cdot a_{sum} \le \sum_{x=1}^{\mu} a_x \tag{17}$$

where μ is the number of the polymerase to be chosen.

At last, the subdivision of the interval of a_x according to the propensities of individual reactions determines the next possible reaction on the chosen polymerase. Similarly to the classic SSA, the reaction time τ is calculated as follows:

$$\tau = \left(1 / a_{sum} \right) \cdot \ln\left(1 / r_2 \right) \tag{18}$$

where r_2 is a second, independently drawn, uniform variate.

All the simulations of single polymerase and multiple polymerase versions of this model have been written in the MATLAB 7.0 language (MathWorks). The major alteration in this variant of the SSA is the two-step selection method: The first step selects a polymerase and the second selects a reaction, instead of selecting the reaction directly. This alteration in the SSA can be useful for the selection of events in a system whose elements can undergo a variable number of reactions depending on the state of the system. In particular, this adjustment to the classic SSA helps to make the coding for these simulations easier and more efficient by reducing the complexity of the program in terms of length of the program. However, further tests are required to prove the efficiency of this SSA variant in terms of computational cost.

3.2 Single polymerase simulations

The model described in Chapter 2 has been simulated with the help of our SSA variant. In this thesis, two versions of the model have been studied through stochastic simulations: the single polymerase and the multiple polymerase version of the model. In this section, the analysis of stochastic simulations for the single polymerase cases will be

discussed. All kinetic parameters used in the simulations have been provided in Table 3.1. Single polymerase case simulations have been performed to get insights into the time taken for the transcription of a single gene. Simulations were started with an unoccupied strand.

Table 3.1: Parameters used in the model.

Parameter	Value	Source	Comments
k_{bind}	$0.0016 s^{-1}$	[32]	
k_{PIC}	$0.0029 s^{-1}$	[54]	
k_{FPIC}	$0.1 s^{-1}$	[53, 56]	Replaces both k_{bind} and k_{PIC}
k_{ini}	$0.6 s^{-1}$	[54]	Calculated from k_{trans}
k_{PE}	$0.6 s^{-1}$	[54]	Calculated from k_{trans}
k_{ai}	$0.6 s^{-1}$	[54]	Calculated from k_{trans}
k_{trans}	$0.3 s^{-1}$	[54]	Calculated from [54]
k_{abort}	$0.05 s^{-1}$		Estimated
k_{elong}	$144 s^{-1}$	[9]	Calculated from k_{move}
k_{ae}	$144 s^{-1}$	[9]	Calculated from k_{move}
k_{move}	$72 s^{-1}$	[9]	
k_{pause}	$3.8 s^{-1}$		Estimated
$k_{release}$	$0.002 s^{-1}$		Estimated
k_{TC}	$0.0032 s^{-1}$	[9]	Calculated from k_T
k_{term}	$0.0032 s^{-1}$	[9]	Calculated from k_T
k_T	$0.0016 s^{-1}$	[9]	
DNA template length	14 *kb*	[95]	

To get insights into the trajectory of Pol II during transcription, the movement of a single Pol II on the 14 kb long DNA sequence has been simulated. In these simulations, abortive initiation has been neglected because the transcription machinery has been assumed to dissociate from the DNA strand in abortive initiation. Therefore, consideration of abortive initiation in single polymerase cases will not fulfill the objective of these simulations to get the probability distributions for transcription time. Ten different simulations of transcription are shown in Figure 3.1. It is clearly visible in the curves obtained through these simulations that Pol II stays for a considerable time at the beginning and at the end of the DNA sequence. However, Pol II moves rapidly with an almost constant rate through the middle of the sequence, i.e. in the elongation phase. The fast and regular movement of Pol II in the elongation phase in this model is due to the lack of other sources of variability such as pyrophosphorolysis and backtracking. These alternative pathways have been found to have a very small effect in prokaryotic transcription [121]. Therefore, these pathways are expected to have even smaller effects on the movement of Pol II in eukaryotes, being more regulated.

One can conclude from these curves that PIC assembly, promoter-proximal pausing and termination are the steps that will typically take the longest to complete. By contrast, productive elongation is a fast and continuous process even for a relatively long sequence like the one modeled here. It is important to note that the duration for PIC assembly, promoter-proximal pausing and termination are different in every curve. These processes are largely responsible for the noise (stochasticity) in transcription.

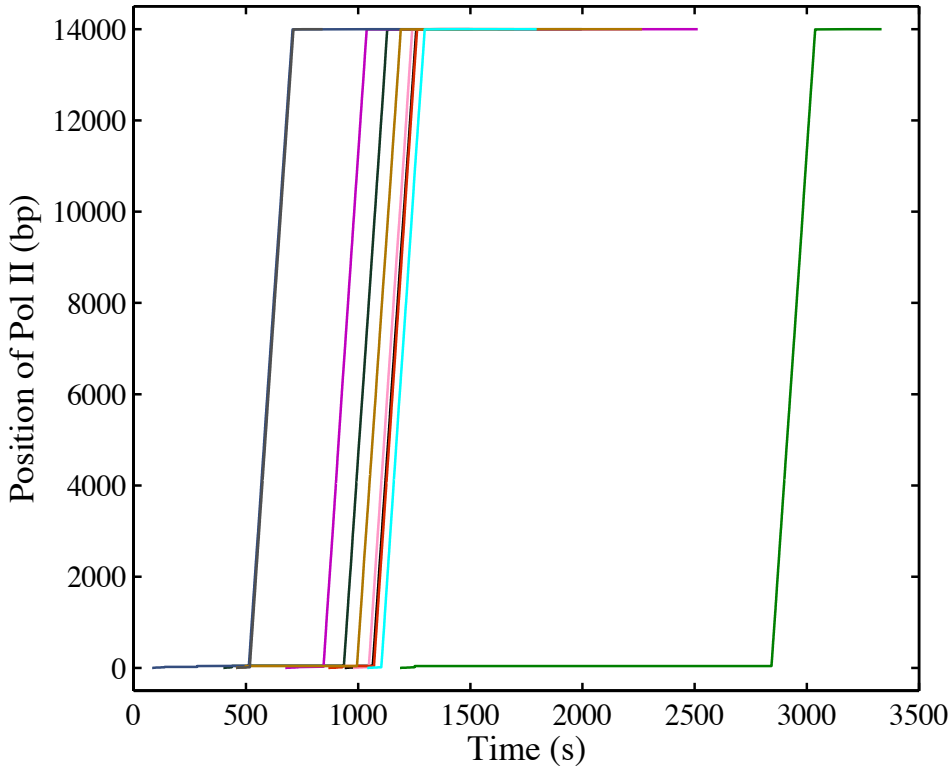

Figure 3.1: Movement of Pol II on the DNA template strand: Pol II movement in 10 different runs of transcription. The region without a line at the beginning represents the PIC assembly time. The lag phases at the beginning and at the end of each curve represent promoter-proximal pausing and termination.

The curves in Figure 3.1 demonstrate the possible sources of stochasticity in eukaryotic transcription. One of the conclusions is that the initial movement of Pol II can play a crucial role in affecting the transcription dynamics. The trajectory of Pol II in the initial positions cannot be clearly visualized from these curves. Therefore, the movement on the initial 50 positions has been more closely inspected for a typical simulation (Figure 3.2).

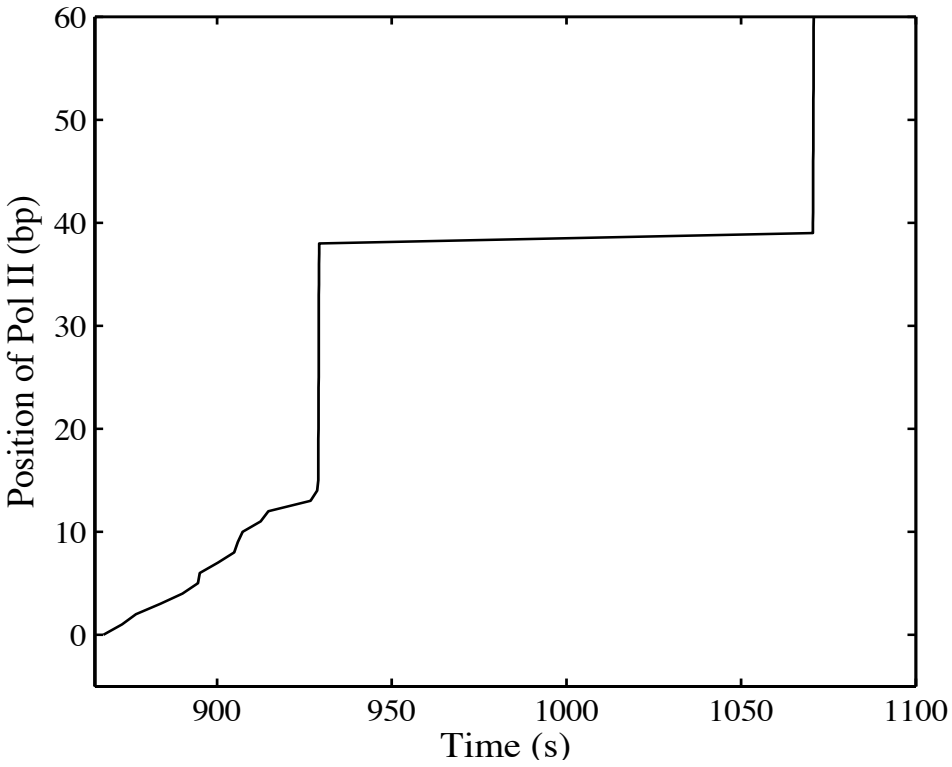

Figure 3.2: Initial movement of RNA Pol II and promoter-proximal pause in one particular simulation (dark red curve of Figure 3.1)

The choice of first 50 positions has been made because all of the slow steps of transcription considered in this model can occur only before position 50 on the DNA template. In addition, this curve will provide a way to compare the movement of Pol II in the promoter escape phase i.e. positions 1-15 with the movement in the elongation phase i.e. the movement after position 15.

The curve in Figure 3.2 shows that Pol II has not started to move until approximately 860 s in this particular transcription run, and then it moves slowly in the first 15 positions. The difference of the transcription rate in promoter escape and elongation phase is clearly visible by comparing the movement over the first 15 positions

with the movement after position 15, i.e. the steep curve. In this particular simulation, a 141 s pause appears in the proximal region.

Simulation results are examples for a particular set of kinetic parameters of a model. Therefore, probability distributions are the better way of representing these results. Probability distributions of transcription times have been obtained for the different single polymerase simulation cases.

3.2.1 Probability distributions

In stochastic simulations of genetic regulatory systems, the events following PIC assembly are often replaced by a delay before the transcript appears [115, 122, 138]. The corresponding delay in this model would be the time from addition of TBP to the production of a pre-mRNA transcript with slow PIC assembly, otherwise it would describe the time from addition of Pol II to the production of a transcript with fast PIC assembly. PIC assembly, promoter proximal pausing and termination can all play key roles for the transcription delay. Four different cases of the single polymerase version of the model have been studied through simulations: (A) with slow PIC assembly and the possibility of promoter-proximal pausing, (B) with slow PIC assembly and no possibility of pausing, (C) with fast PIC assembly and the possibility of promoter-proximal pausing, and (D) with fast PIC assembly and no pausing. It is notable in the single polymerase cases that abortive initiation has not been considered because of the assumption of the breakdown of the transcription machinery in abortive initiation.

The probability distributions of the cases mentioned above have been obtained. To obtain these probability densities, data for the 50,000 transcriptional times were obtained through simulations and divided into equal sized bins. The probability density function (*pdf*) was calculated by dividing the number of samples per bin by the product of the total number of samples and width of a bin. Suppose that n is the number of samples in a bin, N is the total number of samples and L is the width of the bin. Then,

$$pdf = \frac{n}{(N \times L)} \tag{19}$$

The statistical properties of the curves shown in Figure 3.3 have been studied and compared to each other to get insights into the effects of PIC assembly and promoter-proximal pausing. The effect of slow PIC assembly on the delay distribution is clearly visible by comparing the distributions with the slow and fast PIC assembly in the absence of promoter-proximal pausing, i.e. the distributions in Figures 3.3 (B) and (D).

The change of PIC assembly from slow to fast shifted the mean transcription completion time from ~31 min (1840 s) to ~15 min (883 s). The difference in the averages can be understood from the mean PIC assembly times in the two model variants. For slow PIC assembly, the average assembly time should be $k_{bind}^{-1} + k_{PIC}^{-1} = 970$ s. For fast PIC assembly, the average time is $k_{FPIC}^{-1} = 10$ s. The difference between these two assembly times, 960 s, closely matches the difference in the mean transcriptional delays, 957 s. The standard deviation in transcription delay for the transcription by a single Pol II with slow PIC assembly is 13.9 min (835 s), whereas with fast PIC assembly the standard deviation in transcription time is 7.4 min (444 s).

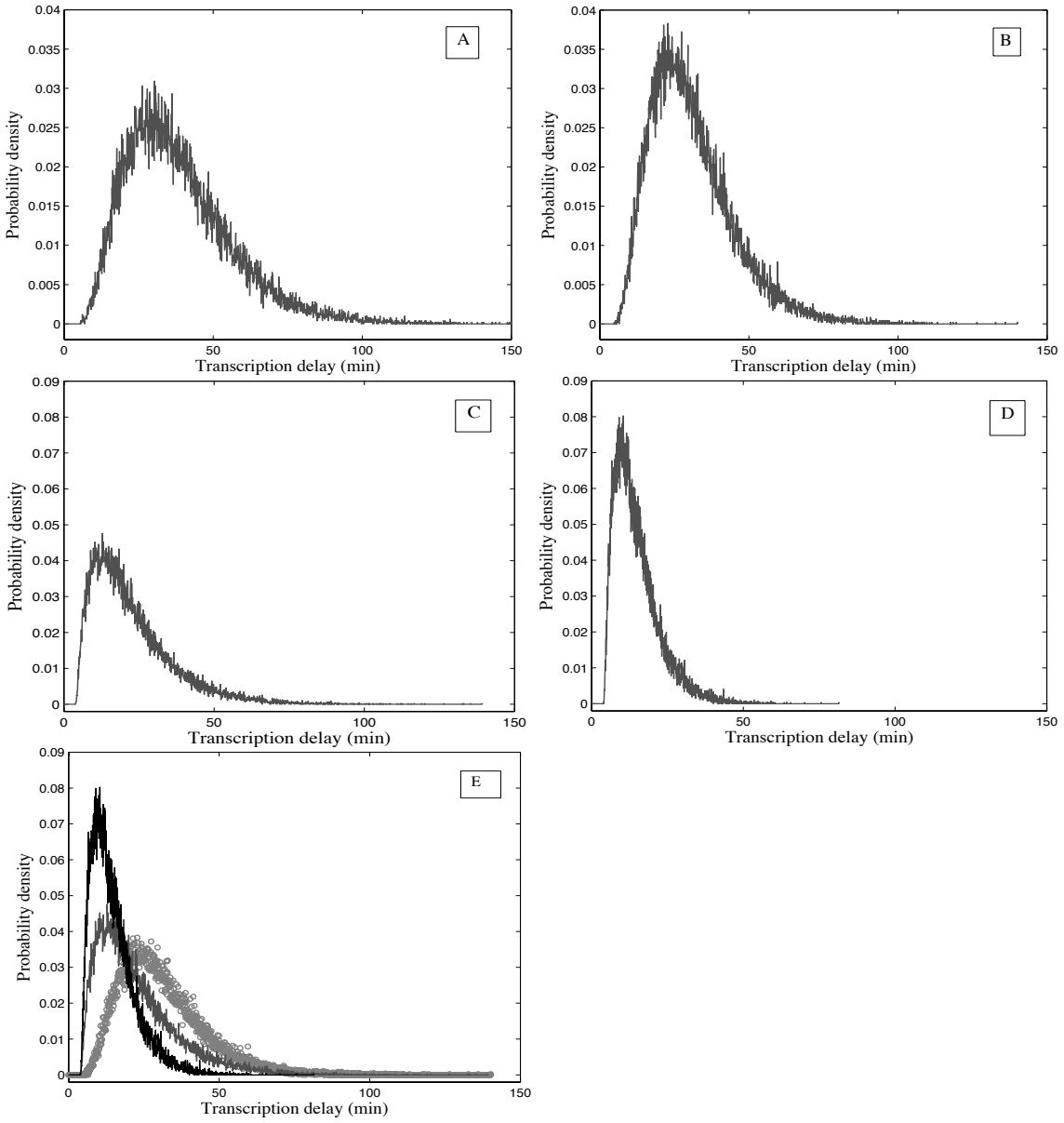

Figure 3.3: Probability distributions of the transcription delay obtained by SSA for 50,000 simulations of transcription with (A) slow PIC assembly and promoter-proximal pausing, (B) slow PIC assembly in the absence of promoter-proximal pausing, (C) fast PIC assembly and promoter-proximal pausing and, (D) fast PIC assembly and no promoter-proximal pausing. (E) shows the comparison of the distributions in (B) (open circles), (C) (grey line) and (D) (black line).

These results suggest that PIC assembly is a major source of the variability in the transcription of a gene with a weak promoter (slow PIC assembly). Even with strong promoters (fast PIC assembly), it can be a source of variability in transcription. However, further tests on these results show some unintuitive results. These results are discussed in section 3.2.2 below.

To further validate these effects of PIC assembly on the mean transcription time, the slow and fast PIC assembly distributions in the presence of promoter-proximal pausing (Figure 3.3 A and C) have been compared and a similar difference of 16 min has been found in the mean transcription delay (not shown separately). This is a consequence of the independence of promoter-proximal pausing and PIC assembly, and is a test of our software. The mean transcription time obtained in the presence of slow PIC assembly and promoter-proximal pausing is 38.3 min. The transcription time obtained with the consideration of all steps of transcription in the single polymerase case is in excellent agreement with an *in vivo* study performed on Chinese hamster cell lines [8].

Next, the effects of promoter-proximal pausing on the probability distributions of the transcription delay were investigated. For this purpose, the delay distributions obtained through simulation of the transcription with a single Pol II with fast PIC assembly and promoter-proximal pausing (Figure 3.3 C) and with fast PIC assembly but with no promoter-proximal pause (Figure 3.3 D) have been compared (grey and black line in Figure 3.3 E). Promoter-proximal pausing has only a small effect on the mode of the distribution, as can be seen in Figure 3.3 C and E, but pauses do make the distribution broader. Accordingly, pauses increase the average transcription time moderately, from

14.7 to 22.4 minutes (52% increase), but increase the standard deviation quite a bit more, from 7.4 to 13.6 minutes (84% increase).

For studying the effects of termination on the statistics of transcription, the distribution with slow PIC assembly and promoter-proximal pausing in the presence of fast termination has been compared with the distribution of Figure 3.3 A, where termination is a slow two-step process (Figure 3.4). The grey distribution in Figure 3.4 was obtained with an arbitrary assumption of one-step fast termination with a stochastic rate constant $k_{term} = 0.1\,s^{-1}$ similarly to the assumption made for PIC assembly where reactions (1) and (2) were replaced by a single reaction (2a).

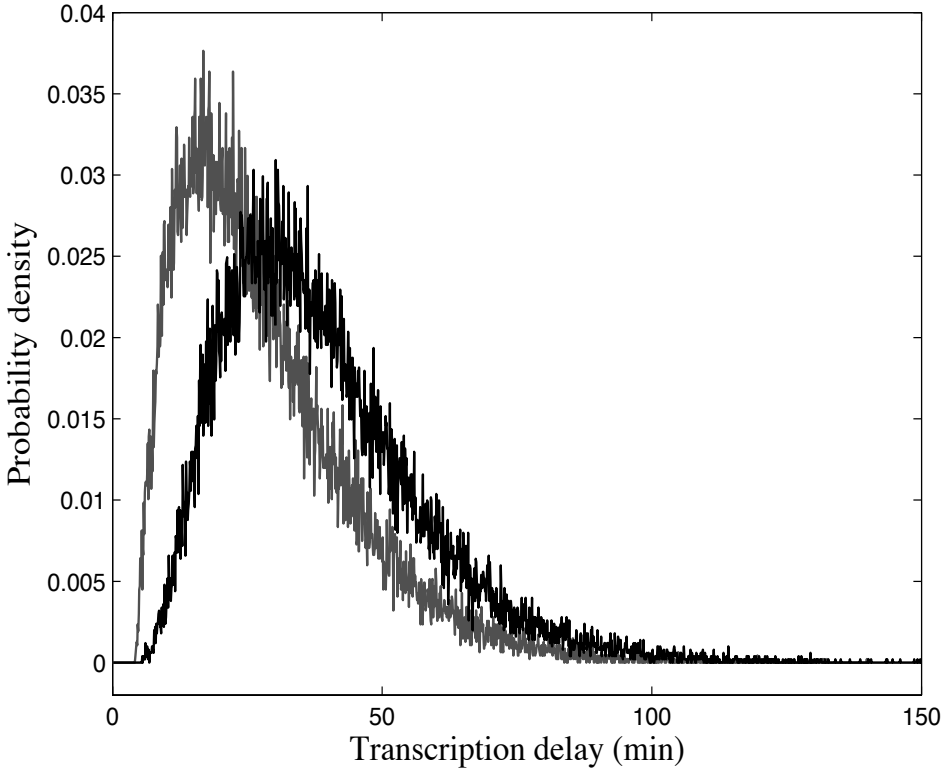

Figure 3.4: Comparison of the delay distributions obtained from 50,000 simulations of transcription with slow PIC assembly, promoter-proximal pausing and fast termination (grey), with the probability distribution obtained for the transcription with slow PIC assembly, promoter-proximal pausing and slow termination (black).

The mean transcription time with fast termination is reduced to 28.1 min from 38.3 min with slow two-step termination (27% reduction). However, the standard deviation shows a small reduction of 1.9 min i.e. 18.1 min to 16.2 min (11% reduction). These results show that it is easy to predict the general change in the standard deviation and mean transcription time change in predictable ways when the rate constants are changed, but they do not change in proportion to each other. Thus, the behavior of their ratio, the CV is difficult to predict. Therefore, the CV for these distributions was further investigated.

3.2.2 Coefficient of variation (CV)

The statistical properties studied above give some insights into the statistical behavior of the transcription system in the presence and absence of different alternative pathways, but the main objectives of this thesis are not fulfilled yet. As mentioned earlier, CV is the ratio of the standard deviation to the mean. Moreover, CV is a dimensionless quantity. This makes it possible to directly compare CVs in different situations, making the CV one of the most popular measures of noise.

CVs have been calculated for the distributions shown in Figure 3.4. The CV for the distribution with slow PIC assembly, promoter-proximal pausing and slow termination, i.e. two step termination, has been calculated as 0.47 whereas the CV with fast termination in the presence of the other two processes has been found to be 0.58. Intuitively, slow termination is expected to add variability to the system but the results

contradict this intuition. This behavior of the CV suggests that slow termination plays a role in noise regulation for a single polymerase.

Figure 3.3 E shows the comparison of the distributions of transcription times with (i) slow PIC assembly (Figure 3.3 B), (ii) fast PIC assembly with promoter-proximal pausing (Figure 3.3 C), and (iii) fast PIC assembly without promoter-proximal pausing (Figure 3.3 D). Coefficients of variation (CV) have been calculated for these distributions to find out the relative variability in the system. The CV for the distribution with promoter-proximal pausing (grey line) has been found to be higher (0.61) than the CV for the transcription time distribution without promoter-proximal pausing (black line) (0.50), which is the expected behavior of the CV.

Further analysis of the CV of the distributions with slow and fast PIC assembly has also been performed. It is perhaps surprising that the CV for transcription with slow PIC assembly (open circles) is lower (0.45) than the CV for transcription with fast PIC assembly (black line) (0.50). This case is also similar to the case of slow and fast termination. To find out the reason for this behavior, further mathematical analysis is required, but the behavior of the CV can be further tested. The pattern of CVs in all the cases with a single polymerase indicates that slow events in the process are reducing the variability in the transcription time, contrary to expectations. The results for all the combinations of slow/fast PIC assembly and inclusion/exclusion of promoter-proximal pausing are summarized in Table 3.2.

The CVs have been computed for a large range of single-step PIC assembly rate constants, i.e. $k_{FPIC} = 10^{-5} - 1\ \mathrm{s}^{-1}$. Figure 3.5 is clearly indicating an unintuitive situation in which the CV goes down with increasing rate constant values until $5 \times 10^{-3}\ \mathrm{s}^{-1}$ and

then the CV starts to increase again and goes towards saturation with the further increase in the PIC assembly rate constant. Figure 3.5 shows that the minimum CV obtained is 0.45 with a rate constant of $k_{FPIC} = 5 \times 10^{-3}$ s^{-1}, which is close to the net rate constant implied by the two rate constants of PIC assembly, i.e. $k_{FPIC} = 1 \times 10^{-3}$ s^{-1}. However, the CV obtained for the latter value of k_{FPIC} is higher than for the fast PIC assembly rate constant, i.e. $k_{FPIC} = 0.1$ s^{-1}, unlike the comparison of fast PIC assembly to two-step PIC assembly shown in Table 3.2. It is interesting that a subtle change in one part of the model can have such a large effect on model statistics, and highlights the importance of having the correct number of steps to model key processes. In Figure 3.5, the part of the curve where the CV goes down is expected because an increased rate constant is expected to decrease the variability, but the part of the curve where the CV increases again is an unintuitive behavior of CV. This abnormal behavior of CV may occur due to the combined effect of the other steps of transcription such as promoter escape, productive elongation and termination in the case of Figure 3.5.

Table 3.2: Mean transcription time, coefficient of variation and 75th percentile to mean ratio in different simulations for the single polymerase cases. Abbreviations: PIC = pre-initiation complex, AI = abortive initiation, PPP = promoter-proximal pausing.

Simulation case	Mean transcription time (s)	Coefficient of variation (CV)	75th percentile to mean ratio
Fast PIC	883.4	0.50	1.25
Slow PIC	1840	0.45	1.24
Slow PIC and PPP	2303	0.47	1.25
Fast PIC and PPP	1342	0.61	1.28
Slow PIC, PPP and fast termination	1684	0.58	1.30

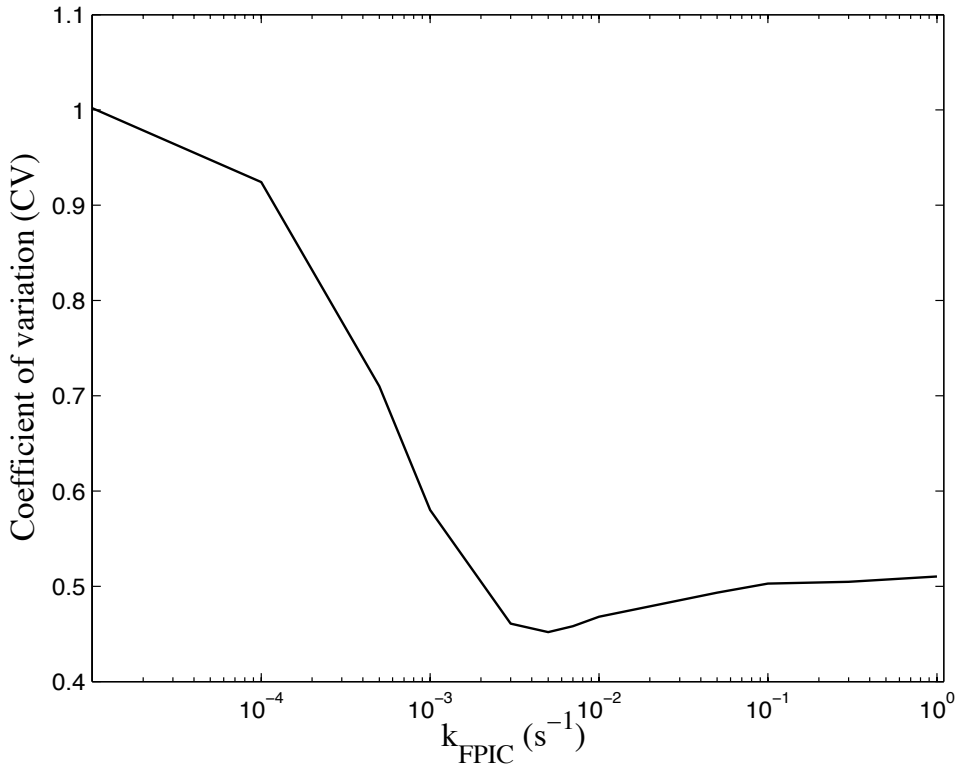

Figure 3.5: Coefficient of variation (CV) for different values of fast PIC assembly rate constants. Note that the x-axis is on a logarithmic scale.

3.2.3 Test for the tails

One more question which arose in the examination of the distributions in Figure 3.3 was whether these distributions were similar in shape or not, i.e. if they could be seen as being related by a coordinate transformation or not. In particular, we wanted to know whether these distributions had similarly shaped tails. In order to study this question, the ratio of the 75^{th} percentile of the transcription time to the mean, which is a quick indicator of the rate of the decay of the tail, has been computed. In each case with a single polymerase, this ratio has been found to be ~1.25, indicating that they have similarly shaped tails. The

values of 75th percentile to mean ratios for all single polymerase cases have been listed in Table 3.2.

3.2.4 Pause distribution

Promoter-proximal pause duration in eukaryotes can range from a few seconds to several minutes [76], as in the case of the Drosophila *hsp70* heat shock gene [72]. This wide range of pause durations may have regulatory functions, either for the timing of the appearance of transcripts, or for the generation of fluctuations in gene expression. In this model, Pol II can pause at any site in the promoter-proximal region (positions 15-50) or it may not halt at any position at all. These pauses occur randomly due to the competition between reactions (4) and (7) of chapter 2. How often do these pauses occur in the promoter-proximal region? To answer this question, the number of pauses in the transcription process with a single Pol II has been counted. The probability that Pol II will not pause at all in the promoter-proximal region is ~40% (Figure 3.6). At the other extreme, Pol II was found to pause 4 or more times in the promoter-proximal region 1.6% of the time.

On the basis of the results described above, one can conclude that PIC assembly is the major contributor to affect the dynamics of transcription in terms of transcription times for a wide range of PIC assembly rate constants. However, for a certain range of PIC assembly rate constants, it may act in the regulation of the variability in transcription with a single polymerase due to the combined effects of other rate constants. Promoter-proximal pausing was modeled as a Poisson process. It can significantly affect the variability of transcription by a single polymerase. However, it does not play as

significant role as PIC assembly in affecting the transcription times. Termination has never been considered to have a significant role in the transcription dynamics, but the results in the single polymerase case suggest that termination plays a significant role in slowing down the process. It may also play a role in the regulation of noise for the rate constants used in this model for a single polymerase case.

It is important to note that these results are valid if only a single polymerase is considered to transcribe the gene. To reach a more general conclusion, these alternative pathways need to be tested in the multiple polymerase case, which will be closer to biological reality. Therefore, a detailed computational analysis of multi-polymerase cases of the model has been performed.

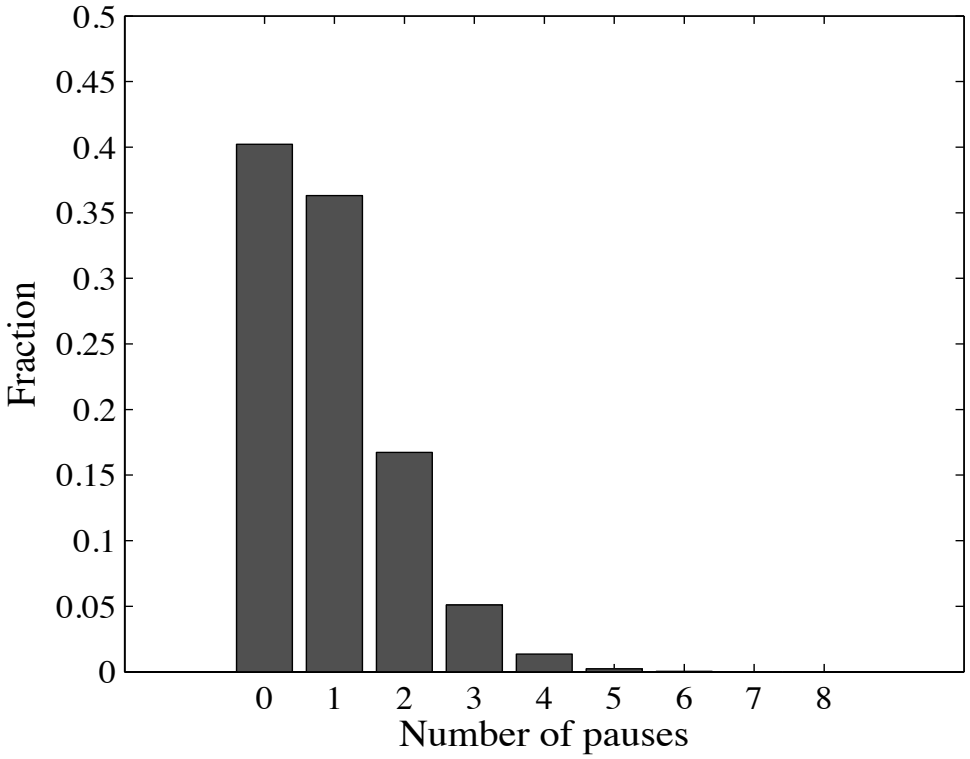

Figure 3.6: Pause distribution in 50,000 simulations of the transcription process.

3.3 Multiple polymerase simulations

Transcription is a dynamic process in which several polymerases can move on the DNA template at a particular time, synthesizing different transcripts. Single polymerase cases give insights into the effect of slow PIC assembly, promoter-proximal pausing and termination on the transcription times. In addition, the single polymerase case models the first transcription event after a gene has been activated, i.e. the speed of response. To imitate the transcription process in a biologically relevant manner, the effects of abortive initiation with the other two alternative pathways discussed above have been studied in the multiple polymerase case. The movement of various polymerases on a DNA strand is a process similar to the movement of traffic on a narrow road. This traffic can be a major issue affecting the transcriptional dynamics; therefore, first of all, the traffic density has been studied in the presence or absence of possible slow steps of transcription. The traffic situation will also be helpful to get insights into the level of regulation by PIC assembly, abortive initiation and termination in eukaryotic transcription.

3.3.1 Traffic in eukaryotic transcription

In prokaryotic transcription, traffic has been recognized as a major source of transcriptional noise [3, 121]. This traffic may occur due to the large number of RNA polymerases on the DNA template and the high speed of the various stages of transcription in prokaryotes. Several alternative pathways such as short and long pauses in the distal region of the gene, pyrophosphorolysis, transcription arrest, etc., may contribute to this traffic in prokaryotes [121]. In prokaryotes, pausing may force the following polymerases to wait for the movement of the paused polymerase, or may cause

collisions between polymerases. These collisions and the consequent increased number of polymerases queued on the DNA template can act as a significant cause of transcriptional bursts, leading to noise in gene expression.

In this model, PIC assembly, abortive initiation and termination are the possible steps that can contribute to reduce the traffic. How often polymerases have to wait for the downstream polymerase to move or how often collisions occur in eukaryotic transcription is still unknown. To study this question, the number of Pol II on the DNA template has been examined under different conditions.

First, the positions of the polymerases over time have been studied in the presence of slow PIC assembly, abortive initiation and promoter-proximal pausing. It has been observed that in the presence of all of these processes, at most two polymerases remain on the strand at any time and their active centers remain on average 7000 nt apart during this movement. Therefore, one can conclude that traffic is not a source of noise in eukaryotic transcription under normal circumstances. Next, the traffic situation has been studied in the absence of abortive initiation at different time points and a similar situation has been observed as above. Therefore, abortive initiation is not a significant factor spacing out the polymerases. Next, the traffic situation has been examined in the absence of both promoter-proximal pausing and abortive initiation. There was still not significant traffic in this case.

Finally, the effect of PIC assembly has been studied on the traffic on a DNA strand. To study this effect, the two stochastic rate constants of PIC assembly, i.e. k_{bind} and k_{PIC}, have been replaced with one single rate constant, as in the case of fast PIC

assembly, i.e. k_{FPIC}. For large values of k_{FPIC}, the polymerases were present on the strand at very high density and all the other polymerases were lined up waiting for the previous polymerases in the queue to move, except for the first and last polymerases. To study the effect of PIC assembly on traffic in more detail, the density of polymerases $\left(N/N_{\max}\right)$ on the DNA strand has been plotted against the rate constant for PIC assembly $\left(k_{FPIC}\right)$ (Figure 3.8). The polymerase density increases with the increase in the rate constant for PIC assembly, but it does not reach the highest possible level. The largest possible number of polymerases in these simulations is 200 $\left(14\text{kb}/70\right)$ but the largest average number of polymerases obtained for a very high rate constant of PIC assembly is 180. This difference in the number of polymerases on the strand may be due to the presence of other processes such as abortive initiation and promoter-proximal pausing. A large distance has been found between the two polymerases added last. This large difference between the last two polymerases added is due to the large acceleration when a polymerase passes the promoter-escape region. The part of the curve up to about $k_{FPIC} = 0.03s^{-1}$ shown in Figure 3.7 is a low traffic regime and the part of the curve above $k_{FPIC} = 0.05s^{-1}$ shows the high traffic regime with a phase transition in between. It is also notable that the realistic value of $k_{FPIC} = 0.001s^{-1}$ lies in low traffic regime. This value of k_{FPIC} is derived from the two experimentally measured rate constant of PIC assembly used as the default parameters of the model.

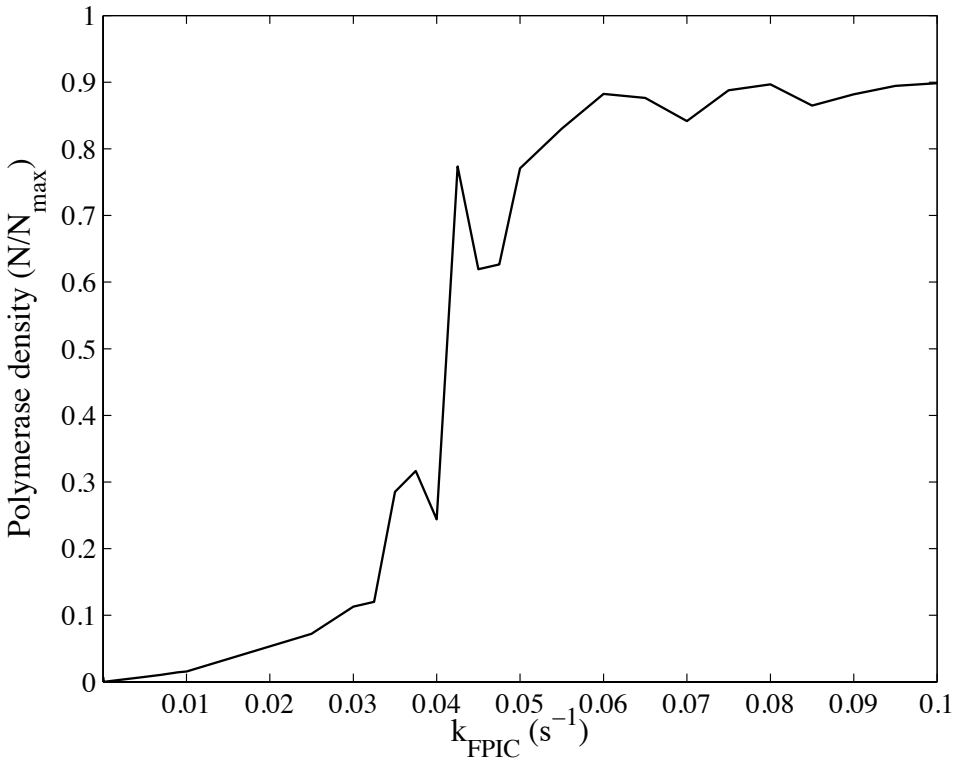

Figure 3.7: Polymerase density on the DNA template strand for different PIC assembly rate constants. Here, N_{max}=200 is the maximum possible number of polymerases on the strand and N is the average number of polymerases present on the DNA strand obtained from the simulations performed for the accumulation of 50,000 pre-mRNAs.

The analysis of the traffic situation performed in Figure 3.7 and the probability distributions in Figure 3.2 give an idea of the slowness of transcription in eukaryotes. Therefore, the accumulation of pre-mRNAs in a period of 24 h due to the transcription of a human gene in many different circumstances has been studied (Figure 3.8). The 24 h time frame is relevant because it is a typical division time for many adult cell lines. The distributions in Figure 3.8 demonstrate that transcription in eukaryotes is extremely slow. The transcription has been found extremely slow when the slow PIC assembly with other two slow steps i.e. abortive initiation and promoter-proximal pausing has been considered

in the process (Figure 3.8 A). The average number of accumulated pre-mRNAs has been found to be 32 in 24 h time period with a standard deviation of 5 pre-mRNAs (Figure 3.8A). On the other hand, 69 and 42 pre-mRNAs have been accumulated on average in a 24 h time period in the absence of abortive initiation and promoter-proximal pausing respectively (Figure 3.8 B and C).

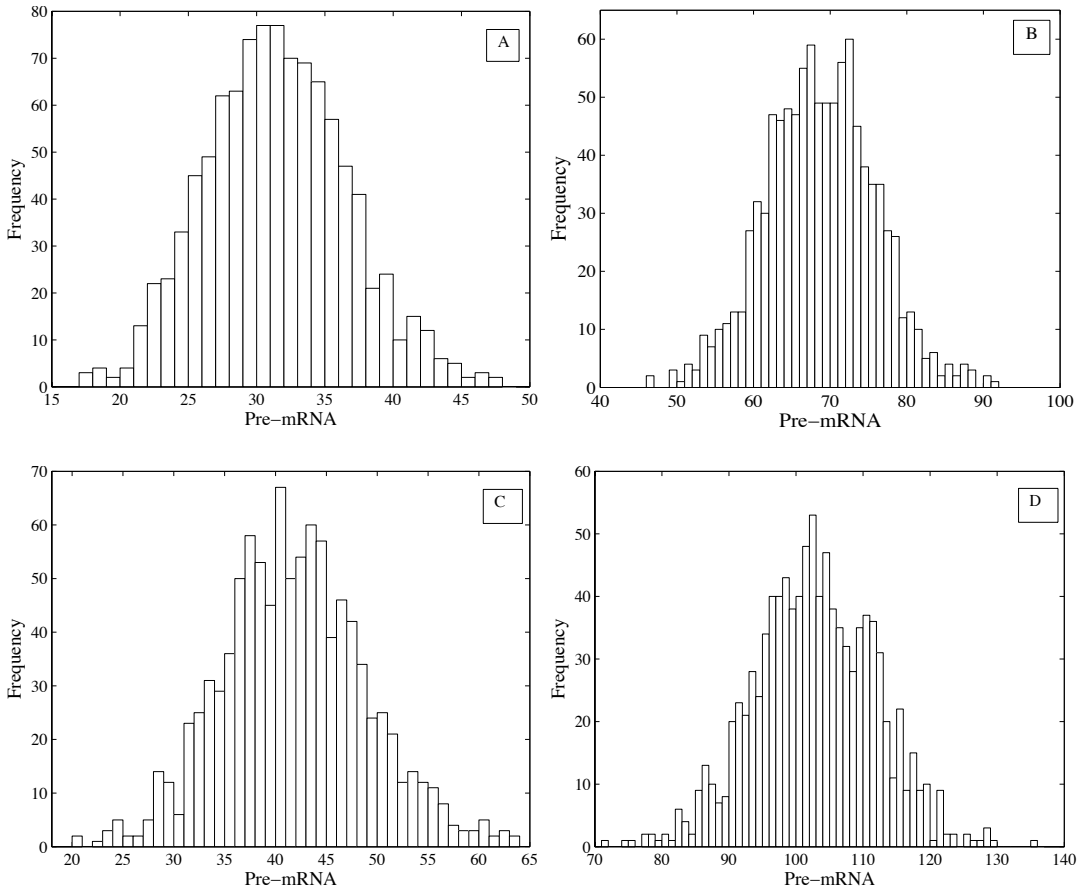

Figure 3.8: Distributions of pre-mRNAs synthesized by the transcription of a single gene in 24 hours under different circumstances: (A) with slow PIC assembly, promoter-proximal pausing, and abortive initiation, (B) with slow PIC assembly and promoter-proximal pausing in absence of abortive initiation, (C) with slow PIC assembly and abortive initiation with no promoter-proximal pausing and, (D) fast PIC assembly with promoter-proximal pausing and abortive initiation. Note that the simulations show the distributions of 1000 runs of 24 h of transcription.

The standard deviation in both cases has been found to be of 7 pre-mRNAs. Other than these cases, the fast PIC assembly has also been studied in the presence of abortive initiation and promoter-proximal pausing (Figure 3.8 D). In this case, the average number of pre-mRNAs synthesized has been found to be 103 in a period of 24 h with a standard deviation of 9 pre-mRNAs.

These results are in excellent agreement with a recent study of mammalian gene expression [139] where 5,000 mammalian genes were studied through a pulse labeling technique to study the correlation between mRNA and protein populations, and then a simple mathematical model was used with the quantitative data of the experiments to calculate a median transcription rate of two transcripts per hour. Comparatively, the model presented here shows a mean and median rate of 1.3-2.9 pre-mRNAs/h for different cases of slow PIC assembly, the realistic case for PIC assembly rate constants. It is important to note that no efforts have been made to achieve the transcription times of the mentioned studies. All the rate constants have been calculated or estimated individually. It is clearly evident from the results of single and multiple polymerase cases of transcription times that the model developed here, is showing a good performance.

3.3.2 Probability distributions

It is important to note that in the multiple polymerase case, the distributions are for the time difference between syntheses of consecutive RNAs instead of the transcription time. The reason for using the time interval is that we can get steady-state transcription rates from the interval between the additions of polymerases. These distributions have been compared to each other to get insights into the individual and combined effects of the

steps mentioned above (Figure 3.9). It is also important to note that in all the distributions in the multiple polymerase case, termination has been considered as a slow two-step process.

In this model, the newly added Pol II can abort easily during its early movement on the DNA template. To study the effect of abortive initiation on transcription dynamics, the distribution of the time difference between the syntheses of two consecutive RNAs for the case with the presence of all three slow steps has been compared with the distribution lacking abortive initiation (Figure 3.9 A and D). The average time between the syntheses of two consecutive RNAs is 47.4 min, if 62% of initiation events get aborted, whereas if nascent transcripts do not abort, the average time-difference between the syntheses of two consecutive RNAs is 20.6 min, less than half of the interval when abortive initiation is considered. There was no significant difference found in the shapes of the distributions using the ratio of the 75^{th} percentile to the mean.

The behavior of the CVs was found to be as expected since abortive initiation will cause a variable number of initiation events to be required before one proceeds to productive elongation, which increases the variability in the system and so, the CV. The CV for the time difference between the syntheses of two consecutive RNAs has been calculated as 0.78 in the absence of abortive initiation whereas, the CV was found to increase to 0.80 in the presence of abortive initiation (Table 3.3). However, it is a small increase.

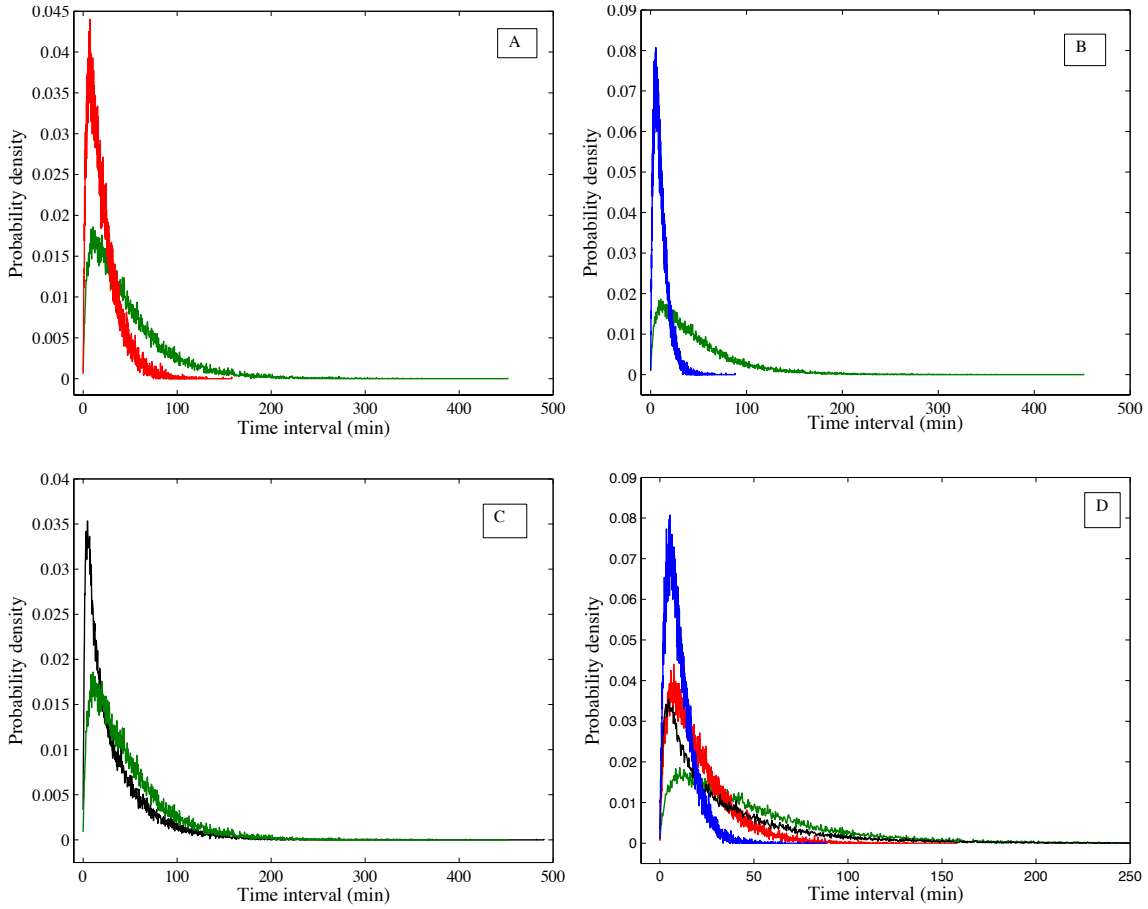

Figure 3.9: Comparisons of the distributions of time difference between the syntheses of consecutive RNAs: In (A), the delay distribution for the slow PIC assembly with abortive initiation (green) is compared with the delay distribution in the absence of abortive initiation (red). In (B), the distribution with slow PIC assembly (green) is compared with the fast PIC assembly distribution (blue). In (C), the delay distribution for slow PIC assembly with promoter-proximal pausing (green) is compared with the delay distribution in the absence of promoter-proximal pausing (black) and in (D), all four cases are compared to each other. The simulations were run until 50,000 RNA transcripts had accumulated.

Next, the effect of PIC assembly on the transcription dynamics has been investigated in the multiple polymerase case. The distributions were obtained for the time interval between the syntheses of two consecutive RNAs with slow PIC assembly (green) and with fast PIC assembly (blue) (Figure 3.9 B and D). The effect of PIC assembly rate

on the time interval has been found to be even larger than the effect of abortive initiation. The average time difference between the syntheses of two consecutive pre-mRNAs dropped down from 47.4 min to 10.5 min with fast PIC assembly. This effect is evident just by looking at the tail of the distribution. The 75^{th} percentile to mean ratio has also been calculated for both of these distributions and no material difference in the shapes of the distributions has been found, similarly to the earlier distributions.

In addition to PIC assembly and abortive initiation, promoter-proximal pausing has also been investigated as a significant point of regulation potentially limiting the rate of eukaryotic transcription [72, 73, 75-77, 80]. In this study, promoter-proximal pausing has also been studied as a significant factor affecting the transcription dynamics in the single polymerase case (Figure 3.3). To further investigate the effects of promoter-proximal pausing on transcription dynamics, its effect on the probability distribution for the time difference in syntheses of consecutive RNAs has been studied in the multiple polymerase case.

The distributions were obtained for the transcription process having slow PIC assembly, abortive initiation and slow termination in the presence (green) and absence (black) of promoter-proximal pausing (Figure 3.9 C and D). Absence of promoter-proximal pausing reduces the average time difference from 47.4 min to 33.3 min but the shapes of both distributions are similar in the presence or absence of promoter-proximal pausing. The standard deviations of the time differences in both cases have been calculated and the standard deviation with promoter-proximal pausing was found to be higher (41 min) than without promoter-proximal pausing (36 min). Further, a higher CV (1.08) in the absence of promoter-proximal pausing has been found than in its presence

(0.86) with slow PIC assembly (Table 3.3). This reduced CV indicates a potential noise regulatory role of promoter-proximal pausing.

In addition to the comparisons described above, the effect of termination on the time interval between syntheses of consecutive RNAs has also been studied in the multiple polymerase case with an arbitrary assumption of fast termination similar to that made in the single polymerase case (not shown). A rate constant similar to the fast termination case studied in the single polymerase case ($k_{term} = 0.1\ s^{-1}$) was used. It has been found that when multiple polymerases move on the DNA template, termination does not make any significant difference to the time interval between syntheses. The reason for termination having no effect is that the different termination rate constants affect transcription times by the same amount, on average, so that the time difference is maintained. Moreover, the fact that fast termination doesn't have any effect on the distributions of time differences reinforces the previous conclusion that traffic is not significant in eukaryotic transcription.

Table 3.3: Mean time interval, coefficient of variation and 75th percentile to mean ratio in different simulations for multiple polymerase cases. Abbreviations: PIC = pre-initiation complex, AI = abortive initiation, PPP = promoter-proximal pausing.

Simulation case	Mean time interval (s)	Coefficient of variation (CV)	75th percentile to mean ratio
Slow PIC, AI and PPP	2843	0.86	1.36
Fast PIC, AI and PPP	631	0.70	1.34
Slow PIC and PPP	1233	0.78	1.37
Slow PIC and AI	1998	1.08	1.36

The slowness of eukaryotic transcription and the low traffic situation can be understood by the fact the termination rate constants of the model are capable to terminate only 5.8 to 8.7 transcription events per hour [9]. There is still no traffic or colliding polymerases found near the termination site. The reason for this low traffic near the termination site is that the combined rate constant of PIC assembly i.e. $k_{FPIC} = 0.001\ s^{-1}$ is lower than the combined termination rate constant i.e. $k_T = 0.0016\ s^{-1}$. The mean time difference between syntheses of consecutive RNAs reduces to 22% in the absence of slow PIC assembly, to 44.5 % in the absence of abortive initiation and to 70% in the absence of promoter-proximal pausing from the time difference with all the other slow processes (Table 3.3). These results verify again the largest role of PIC assembly in controlling the transcription time.

Therefore, it is very much evident that, for the parameters of our model, PIC assembly plays the most significant role in affecting the eukaryotic transcription dynamics, abortive initiation emerged as the second most prominent source to affect the transcription dynamics, and at last promoter-proximal pausing also affects transcription dynamics. Slow termination does not significantly change the time difference in the multiple polymerase case although it changes the transcription time significantly in single polymerase case.

In another way, promoter-proximal pausing affects the system most significantly as it reduces the CV from 1.08 to 0.86, which shows that promoter-proximal pausing reduces the variability in the eukaryotic transcription in the multiple polymerase case. In this model, promoter-proximal pausing occurs from positions 16-50. It acts as a step for regulation by stabilizing the DNA-RNA hybrid and by providing the time to check the

synthesized nascent RNA [77]. It also controls the speed of the movement of Pol II on the DNA template, which reduces the traffic downstream. These regulatory functions of promoter-proximal pausing may lead to the reduced variability in the system, which can be identified through the CV.

3.3.3 Altering initiation and elongation rate constants

Experimental results give information about the average rate constant for movement. As mentioned in section 2.2, the rate constants used for the state change and the translocation in the initiation and elongation steps of this model have been set to the same values as each other due to the lack of experimental data. What if these rate constants are not equal? To answer this question, the consequences of altering the stochastic rate constants of these steps of eukaryotic transcription have been studied. The net rate constant for the movement of the polymerase on the first 15 positions was kept fixed at $k_{trans} = 0.3s^{-1}$. For a given value of $k_{ini} = k_{PE}$, k_{ai} was calculated from equation $k_{trans} = \left(1/k_{PE} + 1/k_{ai} \right)^{-1}$. Increasing k_{ini} and k_{PE} decreases the percentage of abortive initiation and so the mean time difference between the syntheses of consecutive pre-mRNA transcripts (Figure 3.10). In this model, reactions (5) and (6) are in competition in the abortive-initiation prone region, i.e. on positions 4-15; therefore, increasing k_{PE} in reaction (5) increases the probability of movement of Pol II while reducing the probability of aborting. If our assumption about the state from which aborting happens is wrong, it may still be possible for evolution to tune parameters not directly connected to abortion (i.e. other than k_{abort} in the model) to optimize the rate of abortion.

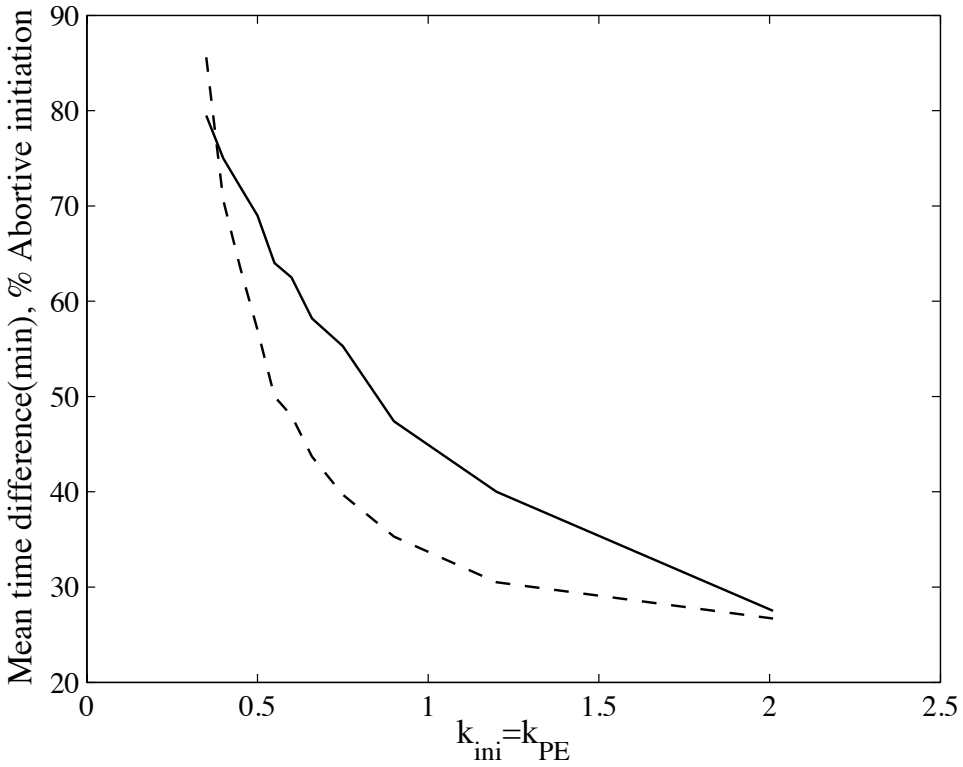

Figure 3.10: Consequences of altering k_{ini} and k_{PE} on the percentage of abortive initiation (solid line) and mean time difference between the syntheses of two consecutive RNAs (dashed line). The net rate constant for translocation was kept fixed at $k_{trans} = 0.3s^{-1}$. Simulations of transcription were performed until 1,000 transcripts had accumulated.

In this model, elongation has also been modeled in a similar fashion to initiation. The stochastic rate constants for the movement in positions 16-*n* have been set to $k_{elong} = k_{ae} = 144s^{-1}$ to match up to the net rate constant $k_{move} = 72s^{-1}$ for the movement of Pol II during productive elongation [9]. In a similar fashion as above, the stochastic rate constant for the elongation events (k_{elong},k_{ae}) have been altered while keeping k_{move} constant. When k_{elong} was increased and k_{ae} correspondingly decreased, an increase in the percentage of the promoter-proximal pauses was noted (Figure 3.11).

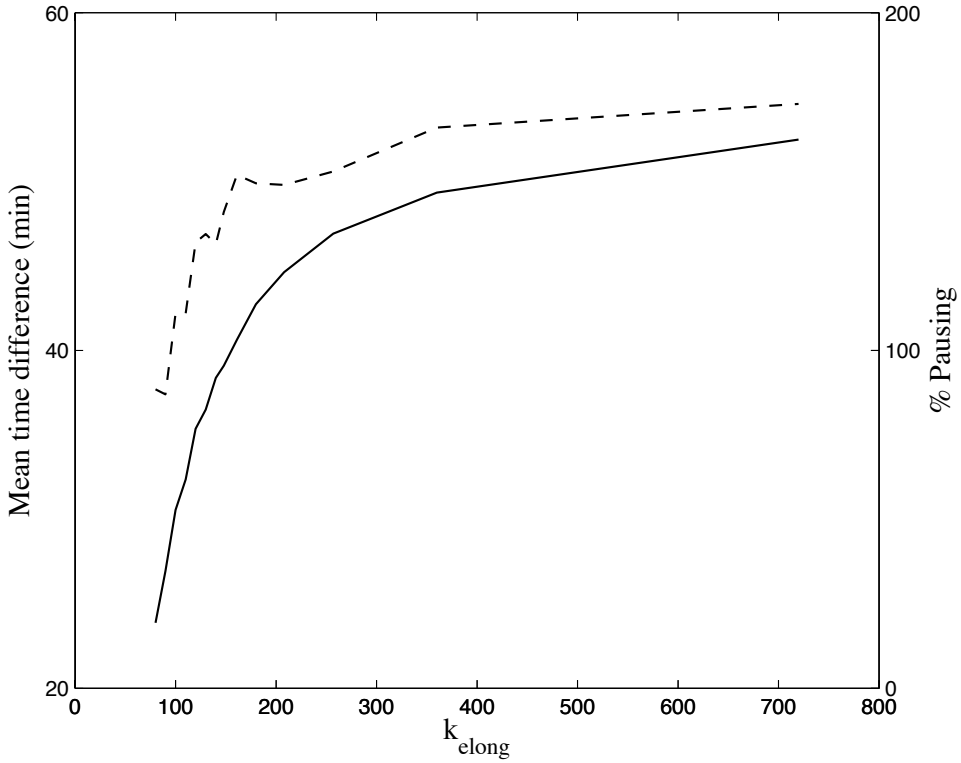

Figure 3.11: Consequences of altering k_{elong}, on the percentage of pausing (solid line) and mean time interval between the syntheses of two consecutive RNAs (dashed line); Simulations of transcription were performed until 1,000 transcripts had accumulated. Note that k_{ae} was also tuned to maintain the net elongation rate constant equal to k_{move}.

The mean time difference between mRNA syntheses was also affected in this case but the effect was smaller than the effect of altering the initiation rate constants. The main reason for the effect on the percentage of polymerases pausing is the competition between reactions (7) and (10). Therefore, a decrease in the rate constant for activation $\left(k_{ae}\right)$ or, in other words, an increase in the rate constant for translocation of Pol II $\left(k_{elong}\right)$ increases the probability of pausing. In turn, the greater percentage of pauses leads to an increase in the mean time difference.

These results suggest two model predictions: first, in the case of abortive initiation, that sequences where translocation is fast once the polymerase has been

activated may be less prone to abort, all other things being equal; and second, that the polymerase is more prone to pause on sequences where translocation is fast if the polymerase enters the paused state from the occupied state, i.e. prior to activation.

Chapter 4

Mathematical analysis

The results obtained in chapter 3 are approximations obtained from stochastic simulations of the transcription model, which depend on the kinetic parameters of the model. These results provide a good approximation for the transcription of some genes in humans but they do not represent a generalized form valid for every eukaryotic gene. To obtain generally valid results, it is necessary to derive the mathematical expression for the probability distributions of transcription time. Solution of the chemical master equation (CME) is an exact approach to get an expression for the transcription time considering transcription as occurring in a homogenous system. The chemical master equation is a set of more than 3^n ordinary differential equations in this model. In the model presented by Roussel and Zhu, there were approximately 3^n variables because of three states defined in the model, where n is the number of nucleotides in the gene [115]. In the model developed in this thesis, there are a few more states, but the number of states still grows exponentially with n. Since $n = 14{,}000$ for a typical human gene, it is impossible to solve the CME directly for this model. However, Roussel and Zhu used an alternative clever approach to get the moments of distribution for the transcription time [115]. In addition, an alternative non-linear master equation having $2n$ variables known as the site-oriented Markov model has been derived to approximate the distributions obtained by the CME in the same study ([115] and references therein).

In this study, we succeed in getting the distributions for the single step movement of the polymerase in different regions by using the CME approach, as a first step towards

the calculation of the moments of the distribution of the transcription times for this eukaryotic transcription system. This chapter will provide a detailed description of the derivations of the distributions of the single step movement of the polymerase in various phases of transcription.

4.1 Probability densities

It is important to mention that the transcription of a single gene by a single polymerase has been considered to calculate the probability densities of single-jump movements. Single-jump movement of the polymerase can be defined as the movement of Pol II from the occupied state on site i, i.e. O_i, to the occupied state on site $i+1$, i.e. O_{i+1}. In normal elongation, the transition from O_i to O_{i+1} completes with the scheme $O_i \rightarrow A_i \rightarrow O_{i+1}$. In the productive elongation region, three states of nucleotides have been defined: unoccupied (U), occupied (O) and activated (A). Additional states are needed to describe initiation, pausing and termination. These states are mutually exclusive. In addition to that, this model implies some simplifications to the calculation of the combinatorial functions (h), being a model of a single gene (section 3.1), i.e. that $h = 1$ for all relevant steps.

We calculated the single-jump probability densities separately for the different phases of transcription where the polymerase moves with different speeds and different processes may be involved. These different phases are the movement to the first nucleotide from the TSS including PIC assembly; initiation (movement from position 1-3); promoter escape (movement on positions 3-15 including the probability of abortive

initiation); movement in the promoter-proximal pausing prone region, i.e. positions 16-50; productive elongation i.e. positions 16-n; and termination. In this section, the method to calculate the single-jump probability density in complex processes such as the first translocation and promoter escape will be described in detail. Other single-jump distributions will be stated as results without detailed derivation.

4.1.1 Single step probability density for first translocation

The single-step probability density $\rho_i(\tau_i)$ describes the relative likelihood of the occurrence of the movement of the polymerase from the position i to $i+1$ at time τ_i. In this model, PIC assembly completes in two steps, i.e. binding of TBP followed by the binding of Pol II, and then the movement to the first site occurs. I also considered models where PIC assembly was a one-step process. I begin here by analyzing the case where PIC assembly is a one-step process, i.e. $B \to A_0 \to O_1$. Here, B and A_0 represent promoter and PIC of this model, respectively. For these transitions, conditional probability densities were calculated similarly to Roussel and Zhu [115]. In this case, transition starts from the state *B*, therefore the probability densities obtained will be conditioned to be in state *B* at time zero. The following equations were obtained that govern the conditional probability of transition to position 1:

$$\frac{\partial \rho_B(t)}{\partial t} = -k_{FPIC} \cdot \rho_B(t), \text{ with initial condition } \rho_B(0) = 1; \tag{19a}$$

$$\frac{\partial \rho_A^{(0)}(t)}{\partial t} = k_{FPIC} \cdot \rho_B(t) - k_{ini} \cdot \rho_A^{(0)}(t), \text{ with initial condition } \rho_A^{(0)}(0) = 0 ; \quad (19b)$$

$$\frac{\partial \rho_O^{(1)}(t)}{\partial t} = k_{ini} \cdot \rho_A^{(0)}(t), \text{ with initial condition } \rho_O^{(1)}(0) = 1 . \quad (19c)$$

In these equations, $\rho_B(t)$ represents the probability of being in state *B* at time *t*, i.e. the bare promoter, $\rho_A^{(0)}(t)$ represents the probability of being in state *A* at position 0 at time *t*, i.e. in the form of *PIC*, whereas $\rho_O^{(1)}(t)$ represents the probability of being in state *O* at time *t* at position 1. These equations are a special case of the chemical master equation and therefore are exact for the calculation of the single-jump probability density $\rho_0(\tau_0)$. The conditional probability density obtained in equation (19c) corresponds to the cumulative probability density for the completion of this step. The linear differential equations provided in (19) can be easily solved to get the cumulative probability distributions for the completion time of this step,

$$\rho(t) = \rho_O^{(1)}(t) = \frac{k_{ini}\left(1 - e^{-k_{FPIC} \cdot t}\right) - k_{FPIC}\left(1 - e^{-k_{ini} \cdot t}\right)}{k_{ini} - k_{FPIC}} . \quad (20)$$

The probability distribution for the single-jump movement can be obtained from the cumulative probability distribution by taking its derivative with respect to *t* and then replacing $t = \tau_0$.

$$\rho_0(\tau_0) = \left.\frac{\partial \rho}{\partial \tau}\right|_{t=\tau_0} = \frac{k_{FPIC} \cdot k_{ini}\left(e^{-k_{ini} \cdot \tau_0} - e^{-k_{FPIC} \cdot \tau_0}\right)}{k_{FPIC} - k_{ini}} . \quad (21)$$

These calculations have been performed using Maple (version 14.00, Maplesoft). The distribution obtained in equation (21) can be plotted against the values of τ_0. Figure

4.1(A) shows the probability distribution for the initiation time, assuming one-step PIC assembly.

Similarly, single-jump probability densities can be obtained for two-step PIC assembly by using one more intermediate state in the transition mentioned above, i.e. $B \to C \to A_0 \to O_1$. Here, B and C are representing the bare promoter and TBP.pro respectively. With the consideration of this transition, the conditional probabilities have been obtained similarly to the case mentioned above starting from the following equations:

$$\frac{\partial \rho_B(t)}{\partial t} = -k_{bind} \cdot \rho_B(t), \text{ with initial condition } \rho_B(0) = 1; \tag{22a}$$

$$\frac{\partial \rho_C(t)}{\partial t} = k_{bind} \cdot \rho_B(t) - k_{PIC} \cdot \rho_C(t), \text{ with initial condition } \rho_C(0) = 0; \tag{22b}$$

$$\frac{\partial \rho_A^{(0)}(t)}{\partial t} = k_{PIC} \cdot \rho_C(t) - k_{ini} \cdot \rho_A^{(0)}(t), \text{ with initial condition } \rho_A^{(0)}(0) = 0; \tag{22c}$$

$$\frac{\partial \rho_O^{(1)}}{\partial t} = k_{ini} \cdot \rho_A^{(0)}(t), \text{ with initial condition } \rho_O^{(1)}(0) = 0. \tag{22d}$$

From these equations for the conditional probability densities, the probability distribution for the single-jump movement have been obtained by a process similar to that which produced equation (21):

$$\rho_0(\tau_0) = \frac{k_{bind} \cdot k_{ini} \cdot k_{PIC} \left\{ e^{-k_{bind} \cdot \tau_0} \cdot (k_{ini} - k_{PIC}) + e^{-k_{PIC} \cdot \tau_0} \cdot (k_{bind} - k_{ini}) + e^{-k_{ini} \cdot \tau_0} \cdot (k_{PIC} - k_{bind}) \right\}}{(k_{PIC} - k_{bind})(k_{bind} - k_{ini})(k_{PIC} - k_{ini})} \tag{23}$$

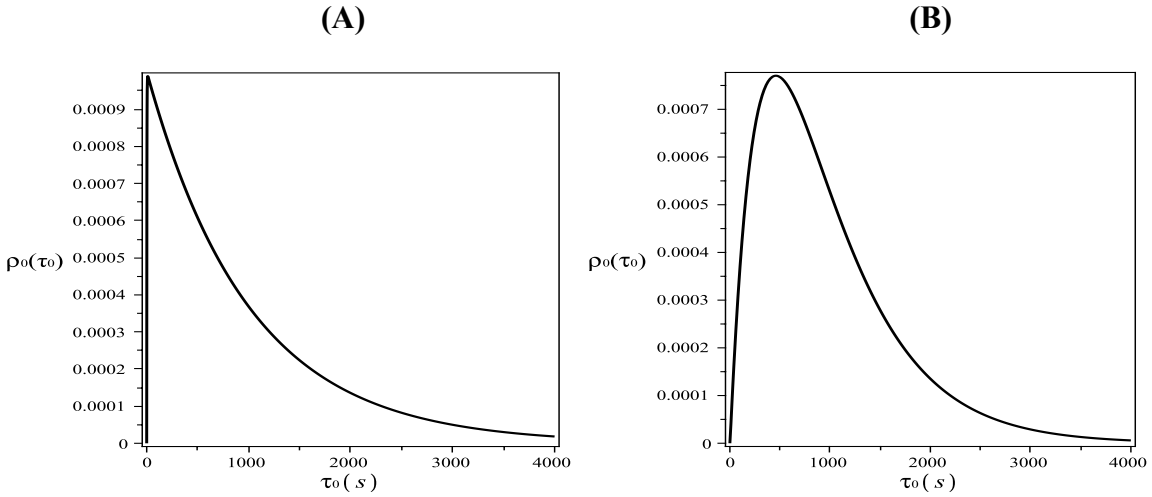

Figure 4.1: Probability distributions obtained for the single-jump movement of Pol II on the first position, considering one-step PIC assembly (A), and two-step PIC assembly (B). In (A), k_{FPIC} = 0.001s^{-1} and in (B), k_{bind} = 0.0016s^{-1}and k_{PIC} = 0.0029s^{-1} as in the stochastic simulations (Table 3.1).

Equation (23) is plotted in Figure 4.1 (B). The average time obtained for the movement of polymerase to the first nucleotide after PIC assembly from this figure is ~971.50 s whereas the simulation results show an average time for (slow) PIC assembly of 970.64 s (Figure 3.3 B). This figure reinforces the conclusion of simulations that initiation in eukaryotic transcription is slow.

4.1.2 Single-step probability density for promoter escape

Similarly to equations (19) and (22), the equations governing the conditional probabilities of promoter escape can be obtained. These equations are

$$\frac{\partial \rho_O^{(i)}(t)}{\partial t} = -k_{ai} \cdot \rho_O^{(i)}(t), \text{ conditioned by } \rho_O^{(i)}(0) = 1; \tag{24a}$$

$$\frac{\partial \rho_A^{(i)}(t)}{\partial t} = k_{ai} \cdot \rho_O^{(i)}(t) - k_{PE} \cdot \rho_A^{(i)}(t) - k_{abort} \cdot \rho_A^{(i)}(t), \text{ conditioned by } \rho_A^{(i)}(0) = 0; \tag{24b}$$

$$\frac{\partial \rho_O^{(i+1)}(t)}{\partial t} = k_{PE} \cdot \rho_A^{(i)}(t), \text{ conditioned by } \rho_O^{(i+1)}(0) = 0. \quad (24c)$$

Here, $i = 3,4,...,15$ for all of three parts of equation (24).

The probability distribution for the movement of Pol II from position i to $i+1$ can be calculated similarly to the first translocation distribution,

$$\hat{\rho}_i(\tau_i) = \frac{k_{PE} \cdot k_{ai}\left(e^{-k_{ai}\cdot\tau_i} - e^{-(k_{PE}+k_{abort})\cdot\tau_i}\right)}{k_{PE} + k_{abort} - k_{ai}} . \quad (25)$$

This probability distribution includes the possibility of abortive initiation. Therefore, it is not normalized to 1. In fact,

$$\int_0^\infty \hat{\rho}_i(\tau_i) d\tau_i = \frac{k_{PE}}{k_{PE} + k_{abort}} = P \text{ (step successful)} \quad (26)$$

In order to eventually get the distribution of transcription times, we need the conditional probability distribution for the case where the step from site i to $i+1$ is successful. This is,

$$\rho_i(\tau_i) = \frac{\hat{\rho}(\tau_i)}{P \text{ (step successful)}} = \frac{k_{ai}\left(e^{-k_{ai}\cdot\tau_i} - e^{-(k_{PE}+k_{abort})\cdot\tau_i}\right)(k_{PE} + k_{abort})}{(k_{PE} + k_{abort} - k_{ai})}, i = 3,4,...,15 \quad (27)$$

The distribution is illustrated in Figure 4.2. It is very clear from Figure 4.2 that the movement of the polymerase in the promoter escape region takes place faster than PIC assembly (Figure 4.1). However, this movement is comparatively slower than the movement of the polymerase in the productive elongation phase (Figure 3.1 and 4.3). This result also agrees with the simulations where the movement of the polymerase in the promoter escape region has been found to be slower than the productive elongation phase (Figure 3.2).

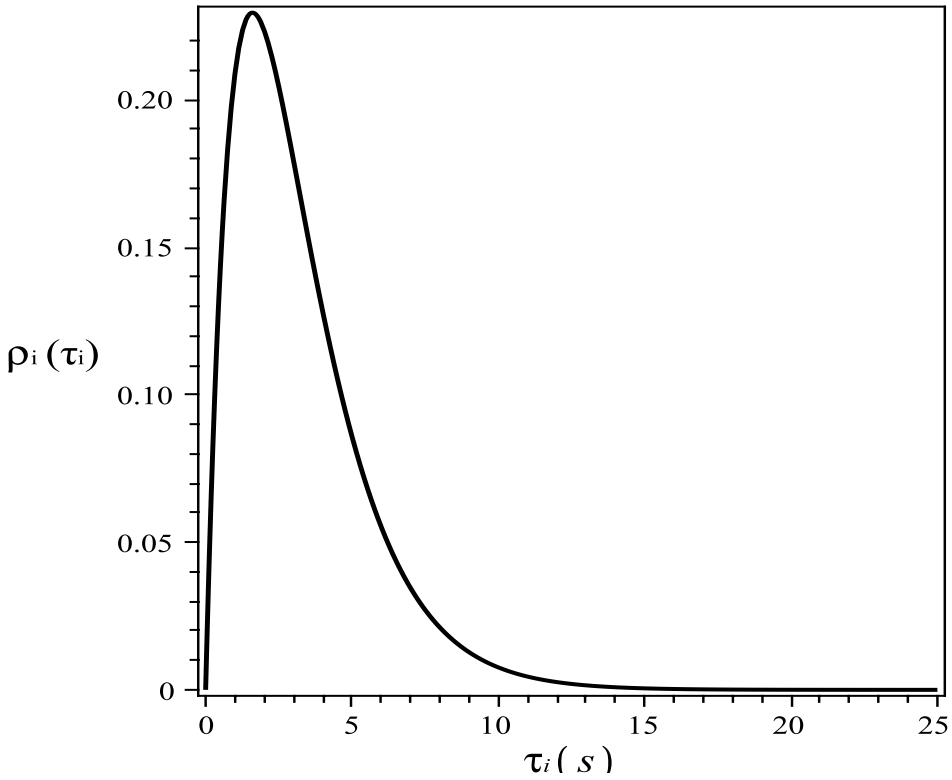

Figure 4.2: Probability distribution for the single-jump movement in the promoter escape phase of transcription. The rate constants used here are those used in the stochastic simulations of the promoter escape phase (Table 3.1).

4.1.3 Probability distributions for other phases

Probability distributions for other phases such as productive elongation and termination can be obtained similarly to the distributions of the first translocation and abortive initiation. The probability distribution for the single-jump movement of Pol II in the promoter-proximal region, where pausing can occur, was obtained first. It is a little more complex to obtain the distribution with promoter-proximal pausing because of the competition between pausing and activation. It is important to note again that pausing has been considered to occur from the occupied state in this model. Therefore, the probability of being in the occupied state is always joined with the probability of the polymerase

being in the paused state. We calculated these conditional probabilities by the following equation,

$$\frac{\partial \rho_O^{(i)}(t)}{\partial t} = -k_{ae} \cdot \rho_O^{(i)}(t) - k_{pause} \cdot \rho_O^{(i)}(t) + k_{release} \cdot \rho_P^{(i)}(t), \text{ with } \rho_O^{(i)}(0) = 1;$$

$$\frac{\partial \rho_P^{(i)}(t)}{\partial t} = k_{pause} \cdot \rho_O^{(i)}(t) - k_{release} \cdot \rho_P^{(i)}(t), \text{ with } \rho_P^{(i)}(0) = 0. \quad (28a)$$

Here, $\rho_P^{(i)}(t)$ represents the probability of the polymerase being in the paused state at time *t*. The solution $\rho_O^{(i)}(t)$ of equation (28a) can be used to calculate the conditional probabilities for the activated state at position i and the occupied state at position $i+1$, similarly to the other processes described before:

$$\frac{\partial \rho_A^{(i)}(t)}{\partial t} = k_{ae} \cdot \rho_O^{(i)}(t) - k_{elong} \cdot \rho_A^{(i)}(t), \text{ with } \rho_A^{(i)}(0) = 0; \quad (28b)$$

$$\frac{\partial \rho_O^{(i+1)}(t)}{\partial t} = k_{elong} \cdot \rho_A^{(i)}(t), \text{ with } \rho_O^{(i+1)}(0) = 0 \quad (28c)$$

Maple was used to calculate the single-jump probability density similarly to the other processes. The expressions obtained through these calculations are extremely complex and impractical to type into this thesis. However, the distributions can be plotted against different values of τ_i (Figure 4.3).

It is comparatively very easy to calculate the single-jump probability density for productive elongation because no other process interferes with productive elongation. Conditional probability densities for productive elongation are similar to the promoter-proximal pausing phase except for equation (28a) because the promoter-proximal pausing phase is part of early elongation and is controlled by the same rate constants except for the possibility of pausing.

The equations governing the conditional probabilities of the single-jump movement in the productive elongation phase are

$$\frac{\partial \rho_O^{(i)}(t)}{\partial t} = -k_{ae} \cdot \rho_O^{(i)}(t), \text{ with } \rho_O^{(i)}(0) = 1; \tag{29a}$$

$$\frac{\partial \rho_A^{(i)}(t)}{\partial t} = k_{ae} \cdot \rho_O^{(i)}(t) - k_{elong} \cdot \rho_A^{(i)}(t), \text{ with } \rho_A^{(i)}(0) = 1; \tag{29b}$$

$$\frac{\partial \rho_O^{(i+1)}(t)}{\partial t} = k_{elong} \cdot \rho_A^{(i)}(t), \text{ with } \rho_O^{(i+1)}(0) = 1. \tag{29c}$$

With these equations, it is possible to obtain the probability distribution for the single-jump movement from site i to site $i+1$ by obtaining the cumulative probability distributions and then by taking the derivative of these cumulative probabilities with respect to τ_i, in a very similar way to the previously calculated probability distributions for the different phases of transcription.

The probability distribution for the single-jump movement in the productive elongation phase has been obtained as

$$\rho_i(\tau_i) = \frac{k_{ae} \cdot k_{elong}\left(e^{-k_{elong} \cdot \tau_i} - e^{-k_{ae} \cdot \tau_i}\right)}{k_{ae} - k_{elong}}, \; i = 16,17,...,n-1. \tag{30}$$

The probability distribution in eq. (30) has been plotted against τ_i but it is not shown here, as it is similar in appearance to Figure 4.3.

The movement of Pol II in positions 1-3 follows a similar mechanism as productive elongation with no obstacles to movement. Therefore, the distribution for the translocation of polymerase on positions 1-3 can be obtained similarly to the distribution for elongation.

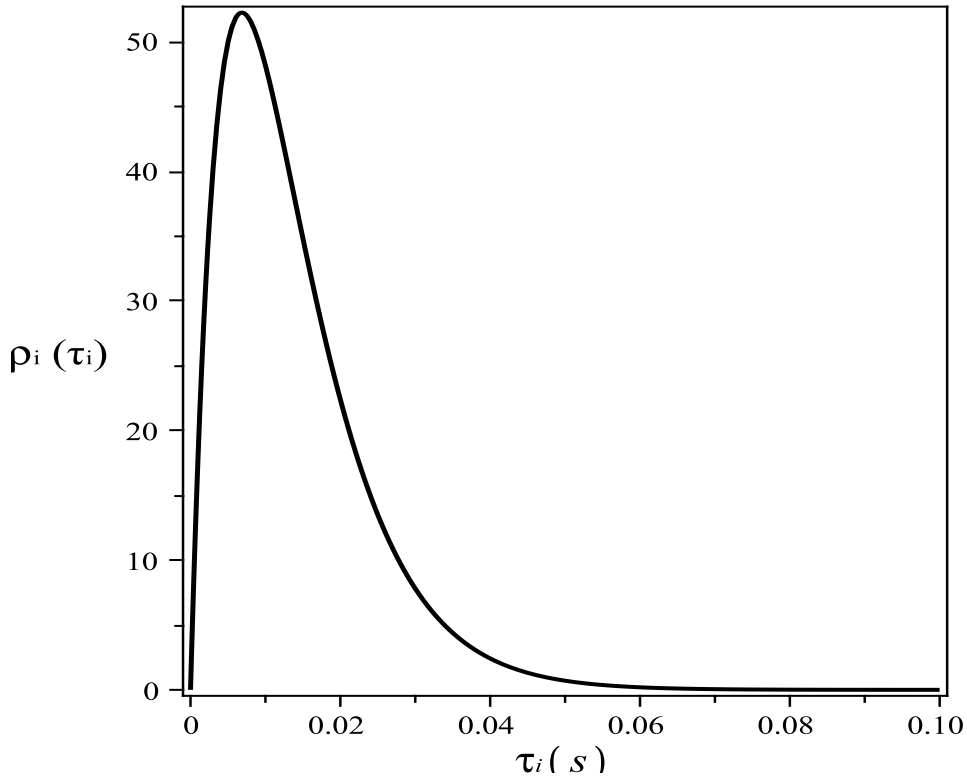

Figure 4.3: Probability distribution for the single-jump movement in the promoter-proximal phase.

This distribution is governed by the rate constants k_{ini} and k_{ai} instead of elongation rate constants. The equations for the conditional probabilities for the time required to take a single step can be obtained in similar fashion. These equations are

$$\frac{\partial \rho_O^{(i)}(t)}{\partial t} = -k_{ai} \cdot \rho_O^{(i)}(t), \text{ with } \rho_O^{(i)}(0) = 1; \tag{31a}$$

$$\frac{\partial \rho_A^{(i)}(t)}{\partial t} = k_{ai} \cdot \rho_O^{(i)}(t) - k_{PE} \cdot \rho_A^{(i)}(t), \text{ with } \rho_A^{(i)}(0) = 1; \tag{31b}$$

$$\frac{\partial \rho_O^{(i+1)}(t)}{\partial t} = k_{PE} \cdot \rho_A^{(i)}(t), \text{ with } \rho_O^{(i+1)}(0) = 1. \tag{31c}$$

These equations can be used to get the cumulative probability density and then the probability distribution for the single-jump movement in positions 1-3, which reads as follows,

$$\rho_i(\tau_i) = \frac{k_{ai} \cdot k_{PE}\left(e^{-k_{PE}\cdot\tau_i} - e^{-k_{ai}\cdot\tau_i}\right)}{k_{ai} - k_{PE}},\ i = 1,...,3. \tag{32}$$

On plotting this distribution (not shown), it is very similar to the distribution shown for the promoter escape phase (Figure 4.2). However, the promoter escape phase has a possibility of abortive initiation.

Termination is the last step of transcription. In this model, termination has been considered as a two-step process, similar to PIC assembly. In stochastic simulations, the consequences of single-step termination have been explored similarly to PIC assembly. Here, both mechanisms for termination have been explored in obtaining the probability distributions for the single-jump movement of Pol II at the termination site. This single-jump movement distribution is actually the probability distribution of the termination from the last position of the transcribed sequence.

Similarly to PIC assembly, single-step termination can be expressed as an $O_n \to A_n \to T$ transition. Here, the T state represents the termination of transcription. For two-step termination, this transition will be $O_n \to A_n \to TC \to T$. Obviously, these transitions are governed by the stochastic rate constant of state transition k_{ae} and the rate constants for the termination k_{TC} and k_{term}. The latter two termination rate constants are replaced by a single equivalent rate constant k_T in single-step termination. The first step towards obtaining the single-jump probability density of termination is to calculate the

conditional probabilities for the state transitions similarly to the processes mentioned above. These equations are

$$\frac{\partial \rho_O^{(n)}(t)}{\partial t} = -k_{ae} \cdot \rho_O^{(n)}(t), \text{ with } \rho_O^{(n)}(0) = 1; \tag{33a}$$

$$\frac{\partial \rho_A^{(n)}(t)}{\partial t} = k_{ae} \cdot \rho_O^{(n)}(t) - k_T \cdot \rho_A^{(n)}(t); \text{ with } \rho_A^{(n)}(0) = 0; \tag{33b}$$

$$\frac{\partial \rho_T(t)}{\partial t} = k_T \cdot \rho_A^{(n)}(t), \text{ with } \rho_T(0) = 0 \tag{33c}$$

for single-step termination, and

$$\frac{\partial \rho_O^{(n)}(t)}{\partial t} = -k_{ae} \cdot \rho_O^{(n)}(t), \text{ with } \rho_O^{(n)}(0) = 1; \tag{34a}$$

$$\frac{\partial \rho_A^{(n)}(t)}{\partial t} = k_{ae} \cdot \rho_O^{(n)}(t) - k_{TC} \cdot \rho_A^{(n)}(t), \text{ with } \rho_A^{(n)}(0) = 0; \tag{34b}$$

$$\frac{\partial \rho_{TC}(t)}{\partial t} = k_{TC} \cdot \rho_A^{(n)}(t) - k_{term} \cdot \rho_{TC}(t), \text{ with } \rho_{TC}(0) = 0; \tag{34c}$$

$$\frac{\partial \rho_T(t)}{\partial t} = k_{term} \cdot \rho_{TC}(t), \text{ with } \rho_T(0) = 0 \tag{34d}$$

for two-step termination. These equations are similar in form to the equations obtained in PIC assembly (equations 19a-c and 22a-d). The single-jump probability density for both the mechanisms can be obtained by following similar procedures as for PIC assembly. The probability distributions obtained for these mechanisms are respectively,

$$\rho_n(\tau_n) = \frac{k_{ae} \cdot k_T \left(e^{-k_{ae} \cdot \tau_n} - e^{-k_T \cdot \tau_n} \right)}{k_T - k_{ae}} \tag{35}$$

$$\rho_n(\tau_n) = \frac{k_{ae} \cdot k_{term} \cdot k_{TC} \left\{ e^{-k_{ae} \cdot \tau_n} \cdot \left(k_{term} - k_{TC} \right) + e^{-k_{TC} \cdot \tau_n} \cdot \left(k_{ae} - k_{term} \right) + e^{-k_{term} \cdot \tau_n} \cdot \left(k_{TC} - k_{ae} \right) \right\}}{\left(k_{TC} - k_{ae} \right) \left(k_{ae} - k_{term} \right) \left(k_{TC} - k_{term} \right)} \tag{36}$$

Equations (35) and (36) are the single-jump probability distributions for the one and two-step termination mechanisms, respectively. Both of these distributions are plotted in Figure 4.4 A and B at parameters that give the same average termination time. Consideration of one-step or two-step termination affects the shape of the distribution similarly to the case of PIC assembly. In Figure 4.4 (B), the rate constants used for the distribution are slightly different from those used in the stochastic simulations. However, the net rate constant for termination is the same. These different rate constants were used due to the presence of the $\left(k_{TC} - k_{term}\right)$ term in equation (36). It is clearly visible from Figures 4.4 (A) and (B) that termination is slow in eukaryotic transcription.

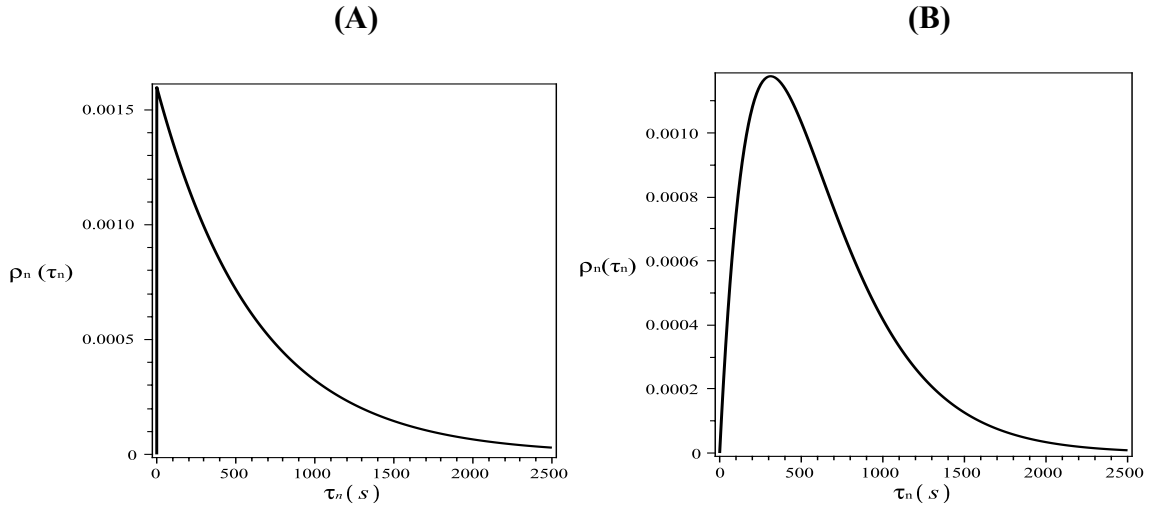

Figure 4.4: Probability distributions for the single-jump movement of Pol II leading to the termination following one (A) and two-step (B) termination mechanisms; the rate constants used here are k_T = 0.0016 $s^{-1}$for the one-step termination mechanism, and k_{TC} = 0.003201s^{-1} and k_{term} = 0.003199 s^{-1} for the two-step mechanism.

I have verified that the average movement time obtained from these distributions by using the rate constants used in the simulations agrees with the stochastic simulations, as was shown for the example of PIC assembly. These distributions can be used to get the moments of the probability distributions for the transcription times using the method of Roussel and Zhu [115], or possibly the complete distribution (Theodore J. Perkins, personal communication). The mathematics will be more difficult because of the added processes in this model. Completing these calculations is left for future work.

Chapter 5

Conclusions

5.1 Summary

In this thesis, a biochemically detailed model of eukaryotic transcription has been developed that specially focuses on the possible slow steps, i.e. pre-initiation complex assembly, abortive initiation, promoter-proximal pausing and termination. Stochastic simulations have been used to study the transcription dynamics. Further, analytical distributions have been obtained for the movement of Pol II in different phases of transcription. The objectives of studying the effects of these possible slow and stochastic steps on the transcription time and the intrinsic stochasticity of transcription have been fulfilled. Another objective of this study, i.e. to identify the major contributors to the noise in eukaryotic transcription, has also been attained. In addition, a partial success has been achieved by obtaining the analytical probability density functions for the single-jump movement of Pol II in various phases of eukaryotic transcription. To my knowledge, this is the first model that studies the stochastic kinetics of eukaryotic transcription in detail. In this chapter, the conclusions for the results obtained through the simulations will be described for the possible slow steps of transcription in their order of occurrence, along with future directions for this point. Preliminary results of theoretical analysis are promising but not sufficient to conclude anything at this point. However, the single-jump probability densities indicate that the analysis is proceeding in the right direction because the single-jump probability density expressions for every step of transcription agree with the simulation results. This verifies the validity of the

simulations. The future target to get the probability distributions remains challenging because of the complexity of the model.

5.2 Conclusions

This model approximates the kinetics of the Pol II holoenzyme as well as an ordered recruitment mechanisms of PIC assembly. On the basis of the results obtained through the stochastic simulations of the single polymerase variant of the model, it is not hard to conclude that the PIC assembly is the most significant factor for affecting the transcription time. Interestingly, the dependence of the coefficient of variation (CV) on the kinetic parameters of the model is contrary to expectations for PIC assembly in the single polymerase case. PIC assembly has been found to be the most significant contributor to the time interval of the synthesis of consecutive pre-mRNAs in the multiple polymerase case. This verifies the result of the single polymerase case that identifies a similar role of PIC assembly in the transcription time. PIC assembly has also been found as a significant source of the variability in the multiple polymerase case. In addition, results of this model show that the rate constant for PIC assembly can affect the traffic situation most significantly.

Therefore, the results of the single and multiple polymerase cases establish the most significant role of PIC assembly in affecting the transcription delay and time intervals between the syntheses of consecutive pre-mRNAs. These results also identify the major role of PIC assembly in creation of the variability in eukaryotic transcription. There is a lack of literature to verify the results for the effects of PIC assembly on transcription dynamics obtained through this model. However, the stochastic rate

constants for PIC assembly used in this study are calculated from the rate constants provided in experimental studies for human genes.

Abortive initiation has been investigated only in the multiple polymerase cases due to the model assumption of the breakdown of the transcription machinery on abortive initiation. Abortive initiation emerged as the second most significant factor affecting the time interval between the syntheses of consecutive pre-mRNA transcripts amongst other possible slow steps considered in this model. Abortive initiation adds to the variability in transcription, but its role is not significant in comparison to the other slow steps.

Promoter-proximal pausing has been identified as one of the sources of transcription delay and the delay in the time interval between the syntheses of two consecutive pre-mRNAs, but its role is less effective in comparison to the role of PIC assembly and abortive initiation discussed above. However, observations regarding time delays cannot deny its role in the regulation of variability. In the multiple polymerase case, this role is made evident by a sudden increase of CV when promoter-proximal pausing is omitted from the process, which suggests the noise regulatory role of promoter-proximal pausing.

Termination in eukaryotic transcription has been found to have a significant contribution to the delay in transcription time (more so than promoter-proximal pausing) in the single polymerase case. It can also play a role in noise regulation. It reduces the noise in eukaryotic transcription for a single polymerase. However, in the multiple polymerase case, termination plays no significant role in affecting the time difference between the syntheses of two consecutive pre-mRNAs.

In addition to the roles of the steps of transcription described above, the consideration of kinetic parameters can also play a role in the transcription dynamics. It has been shown in this thesis that the alterations in the rate constants for the state transition and translocation of Pol II in the promoter escape and promoter-proximal pausing phases can significantly affect the percentage of abortive initiation and promoter-proximal pausing. Consequently, these parameters cannot be independently fixed, but must be co-varied to match particular experimental delay and pause distributions. On the basis of these results, two model predictions have been introduced (Figures 3.11 and 3.12). The first prediction states that in the promoter escape phase of transcription, an activated polymerase may be less prone to abort if the movement of the polymerase is fast, and according to the second prediction, the polymerase may be more prone to pause in the promoter-proximal region if it spends less time in the activated state.

Our results are in excellent agreement with a very recent joint experimental and modeling study of mammalian gene expression [139], where a median of two transcripts synthesized per hour was calculated with a mathematical model of their experimental data for 5,000 mammalian genes. It is notable that most of the kinetic parameters used in the model presented in this thesis are from human gene transcription data and have been calculated or estimated for every step individually. These parameters were not tuned to obtain the experimental synthesis rate.

In addition, these results also show good agreement with the *in vivo* study performed on Chinese hamster cell lines [8], where the time taken for the completion of half of the transcription cycle with slow initiation has been found to be in the range of 14-20 min. Our model predicts a time period of 19.2 min to complete half of the

transcription cycle by Pol II, which is towards the high-end limit of the experimental measurements. The good performance of this model indicates that it is ready for fine-tuning for specific genes as experimental data become available, or that it could even be fit to transcription data to estimate some of the rate constants.

5.3 Future perspectives

This model is the beginning of this kind of modeling in eukaryotic transcription. There is still a lot of room to study the dynamics of eukaryotic transcription. The immediate future perspective for this model is to get the probability distributions for the complete transcription times from the single-jump probability distributions provided in this thesis. These probability distributions will provide a more generalized form to the model results, which can be useful for experimentalists as well as theoretical researchers. At the level of the steps of transcription, a model of chromatin remodeling with PIC assembly can itself be an interesting subject to study because these are complex processes [140]. Some efforts have been already started towards the investigation of fluctuations in chromatin remodeling [141]. Further, on the step level, it should not be difficult to include other slow steps such as transcriptional arrest, backtracking, premature termination, etc., to refine this model for the dynamics of eukaryotic transcription. However, these mechanisms have not been found to have significant effects in prokaryotes [121]. Therefore, a huge effect cannot be expected in eukaryotes because of the higher level of regulation than in prokaryotes. This model has some estimated rate constants, as for example the promoter-proximal pausing and abortive initiation rate constants. These steps can be revisited on the availability of further experimental data. The same applies

for assumptions such as that, the transcription machinery dissociates on abortive initiation.

Another interesting research direction for this model can be to study the dynamics with the inclusion of splicing in this model. Splicing is a complex process and is relatively slow, taking approximately 10 minutes to complete [97]. At the mechanism level, it is itself a very complex research subject. There is endless scope for research in this field.

REFERENCES

1. Bruce A, Bray D, Lewis J, Raff M, Roberts K & Watson J (1989) Molecular biology of the cell. *Garland Publishing Inc, New York and London, Second Edn*.
2. Saunders A, Core LJ & Lis JT (2006) Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol* **7**, 557-567.
3. Kaern M, Elston TC, Blake WJ & Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* **6**, 451-464.
4. Paulsson J (2005) Models of stochastic gene expression. *Phys Life Rev* **2**, 157-175.
5. Jia T & Kulkarni RV (2011) Intrinsic noise in stochastic models of gene expression with molecular memory and bursting. *Phys Rev Lett* **106**, 058102.
6. Herr AJ, Jensen MB, Dalmay T, Baulcombe DC (2005) RNA polymerase IV directs silencing of endogenous DNA. *Science* **308**, 118–120.
7. Wierzbicki AT, Ream TS, Haag JR & Pikaard CS (2009) RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat Genet* **41**, 630 - 634.
8. Kimura H, Sugaya K & Cook PR (2002) The transcription cycle of RNA polymerase II in living cells. *J Cell Biol* **159**, 777-782.
9. Darzacq X, Shav-Tal Y, de Turris V, Brody Y, Shenoy SM, Phair RD & Singer RH (2007) In vivo dynamics of RNA polymerase II transcription. *Nat Struct Mol Biol* **14**, 796-806.
10. Brody Y & Shav-Tal Y (2008) Visualizing transcription in real-time. *Cent Eur J Biol* **3**, 11 - 18.
11. Ardehali MB & Lis JT (2009) Tracking rates of transcription and splicing in vivo. *Nat Struct Mol Biol* **16**, 1123-1124.
12. Fuda NJ, Ardehali MB & Lis JT (2009) Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186-192.
13. Venters BJ & Pugh BF (2009) How eukaryotic genes are transcribed. *Crit Rev Biochem Mol Biol* **44**, 117-141.
14. Hahn S (2004) Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* **11**, 394 - 403.
15. Kornberg RD (2007) The molecular basis of eukaryotic transcription. *Proc Natl Acad Sci U S A* **104**, 12955-12961.
16. Pugh BF (2000) Control of gene expression through regulation of the TATA-binding protein. *Gene* **255**, 1-14.
17. Smale ST & Kadonaga JT (2003) The RNA polymerase II core promoter. *Annu Rev Biochem* **72**, 449-479.

18. Green MR (2005) Eukaryotic transcription activation: Right on target. *Mol Cell* **18**, 399-402.

19. Corden JL (2008) Yeast Pol II start-site selection: the long and the short of it. *EMBO Rep* **9**, 1084 - 1086.

20. Sikorski TW, Buratowski, S. (2009) The basal initiation machinery: beyond the general transcription factors. *Curr Opin Cell Biol* **21**, 344-351.

21. Ohler U, Liao, G-C., Niemann, H., Rubin, G. (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biology* **3**, 1-12.

22. Kuehner JN & Brow DA (2006) Quantitative analysis of in vivo initiator selection by Yeast RNA polymerase II supports a scanning model. *J Biol Chem* **281**, 14119-14128.

23. Cox JM, Kays AR, Sanchez JF & Schepartz A (1998) Preinitiation complex assembly: potentially a bumpy path. *Curr Opin Chem Biol* **2**, 11-17.

24. Papai G, Weil P & Schultz P (2011) New insights into the function of transcription factor TFIID from recent structural studies. *Curr Opin Gen Dev* **21**, 219-224.

25. Coleman RA & Pugh BF (1995) Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA. *J Biol Chem* **270**, 13850-13859.

26. Tjian R (1996) The biochemistry of transcription in eukaryotes: A paradigm for multisubunit regulatory complexes. *Phil Trans R Soc Lond B* **351**, 491-499.

27. Tora L & Timmers HTM (2010) The TATA box regulates TATA-binding protein (TBP) dynamics in vivo. *Trends Biochem Sci* **35**, 309-314.

28. Kim JL, Nikolov DB & Burley SK (1993) Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365**, 520 - 527.

29. Patikoglou GA, Kim JL, Sun L, Yang S-H, Kodadek T & Burley SK (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev* **13**, 3217-3230.

30. Perez-Howard GM, Weil PA & Beechem JM (1995) Yeast TATA binding protein interaction with DNA: fluorescence determination of oligomeric state, equilibrium binding, on-rate, and dissociation kinetics. *Biochemistry* **34**, 8005-8017.

31. Taggart AKP & Pugh BF (1996) Dimerization of TFIID when not bound to DNA. *Science* **272**, 1331-1333.

32. Coleman RA & Pugh BF (1997) Slow dimer dissociation of the TATA binding protein dictates the kinetics of DNA binding. *Proc Natl Acad Sci U S A* **94**, 7221-7226.

33. Coleman RA, Taggart AKP, Benjamin LR & Pugh BF (1995) Dimerization of the TATA binding protein. *J Biol Chem* **270**, 13842-13849.

34. Jackson-Fisher AJ, Chitikila C, Mitra M & Pugh BF (1999) A role for TBP dimerization in preventing unregulated gene expression. *Mol Cell* **3**, 717-727.

35. Coleman RA, Taggart AKP, Burma S, Chicca JJ & Pugh BF (1999) TFIIA regulates TBP and TFIID dimers. *Molecular Cell* **4**, 451-457.

36. Orphanides G, Lagrange T & Reinberg D (1996) The general transcription factors of RNA polymerase II. *Gene Dev* **10**, 2657-2683.

37. Roeder RG (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci* **21**, 327-335.

38. Agalioti T, Lomvardas S, Parekh B, Yie J, Maniatis T & Thanos D (2000) Ordered recruitment of chromatin modifying and general transcription factors to the IFN-β promoter. *Cell* **103**, 667-678.

39. Cosma MP (2002) Ordered recruitment: Gene-specific mechanism of transcription activation. *Mol Cell* **10**, 227-236.

40. Myer VE & Young RA (1998) RNA polymerase II holoenzymes and subcomplexes. *J Biol Chem* **273**, 27757-27760.

41. Chang M & Jaehning JA (1997) A multiplicity of mediators: Alternative forms of transcription complexes communicate with transcriptional regulators. *Nucl Acids Res* **25**, 4861-4865.

42. Koleske AJ & Young RA (1995) The RNA polymerase II holoenzyme and its implications for gene regulation. *Trends Biochem Sci* **20**, 113-116.

43. Ranish JA, Yudkovsky N & Hahn S (1999) Intermediates in formation and activity of the RNA polymerase II preinitiation complex: holoenzyme recruitment and a postrecruitment role for the TATA box and TFIIB. *Genes Dev* **13**, 49-63.

44. Cramer P, Bushnell, D. A., Kornberg, R. D. (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292**, 1863-1876.

45. Bushnell DA & Kornberg RD (2003) Complete, 12-subunit RNA polymerase II at 4.1-A resolution: Implications for the initiation of transcription. *Proc Natl Acad Sci U S A* **100**, 6969-6973.

46. Bushnell DA, Westover KD, Davis RE & Kornberg RD (2004) Structural basis of transcription: An RNA polymerase II-TFIIB cocrystal at 4.5 angstroms. *Science* **303**, 983-988.

47. Liu X, Bushnell DA, Wang D, Calero G & Kornberg RD (2010) Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. *Science* **327**, 206-209.

48. Westover KD, Bushnell DA & Kornberg RD (2004) Structural basis of transcription: Separation of RNA from DNA by RNA Polymerase II. *Science* **303**, 1014-1016.

49. Cramer P, Armache K J, Baumli S, Benkert S, Brueckner F, Buchen C, Damsma G E, Dengl S, Geiger, S R, Jasiak A J, Jawhari A, , Jennebach S, Kamenski T, Kettenberger H, Kuhn CD, Lehmann E, Leike K, Sydow JF & Vannini A (2008) Structure of eukaryotic RNA polymerases. *Annu Rev Biophys* **37**, 337 - 352.

50. Kim T-K, Lagrange T, Wang Y-H, Griffith JD, Reinberg D & Ebright RH (1997) Trajectory of DNA in the RNA polymerase II transcription preinitiation complex. *Proc Natl Acad Sci USA* **94**, 12268-12273.

51. Nikolov DB & Burley SK (1997) RNA polymerase II transcription initiation: A structural view. *Proc Natl Acad Sci USA* **94**, 15-22.

52. Woychik NA & Hampsey M (2002) The RNA polymerase II machinery: Structure illuminates function. *Cell* **108**, 453-463.

53. Kugel JF & Goodrich JA (1998) Promoter escape limits the rate of RNA polymerase II transcription and is enhanced by TFIIE, TFIIH, and ATP on negatively supercoiled DNA. *Proc Natl Acad Sci U S A* **95**, 9232-9237.

54. Ferguson HA, Kugel JF & Goodrich JA (2001) Kinetic and mechanistic analysis of the RNA polymerase II transcription reaction at the human interleukin-2 promoter. *J Mol Biol* **314**, 993-1006.

55. Hieb AR, Baran S, Goodrich JA & Kugel JF (2006) An 8 nt RNA triggers a rate-limiting shift of RNA polymerase II complexes into elongation. *EMBO J* **25**, 3100-3109.

56. Kugel JF & Goodrich JA (2000) A kinetic model for the early steps of RNA synthesis by human RNA polymerase II. *J Biol Chem* **275**, 40483-40491.

57. Kugel JF & Goodrich JA (2002) Translocation after synthesis of a four-nucleotide RNA commits RNA polymerase II to promoter escape. *Mol Cell Biol* **22**, 762-773.

58. Weaver JR, Kugel JF & Goodrich JA (2005) The sequence at specific positions in the early transcribed region sets the rate of transcript synthesis by RNA polymerase II in vitro. *J Biol Chem* **280**, 39860-39869.

59. Holstege FCP, Fiedler U & Timmers HTM (1997) Three transitions in the RNA polymerase II transcription complex during initiation. *EMBO J* **16**, 7468 - 7480.

60. Pal M, Ponticelli AS & Luse DS (2005) The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II. *Mol Cell* **19**, 101-110.

61. Dvir A (2002) Promoter escape by RNA polymerase II. *Biochim Biophys Acta (BBA)* **1577**, 208-223.

62. Dvir A, Tan S, Conaway JW & Conaway RC (1997) Promoter escape by RNA polymerase II. *J Biol Chem* **272**, 28175-28178.

63. Dvir A, Tan S, Conaway JW & Conaway RC (1997) Promoter escape by RNA polymerase II. Formation of an escape-competent transcriptional intermediate is a prerequisite for exit of polymerase from the promoter. *J Biol Chem* **272**, 28175-28178.

64. Dvir A, Conaway JW & Conaway RC (2001) Mechanism of transcription initiation and promoter escape by RNA polymerase II. *Curr Opin Genetics Dev* **11**, 209-214.

65. Hsu LM (2002) Promoter clearance and escape in prokaryotes. *Biochim Biophys Acta (BBA)* **1577**, 191-207.

66. Goldman SR, Ebright RH & Nickels BE (2009) Direct detection of abortive RNA transcripts in vivo. *Science* **324**, 927-928.

67. Luse DS & Jacob GA (1987) Abortive initiation by RNA polymerase II in vitro at the adenovirus 2 major late promoter. *J Biol Chem* **262**, 14990-14997.

68. Margeat E, Kapanidis AN, Tinnefeld P, Wang Y, Mukhopadhyay J, Ebright RH & Weiss S (2006) Direct observation of abortive initiation and promoter escape within single immobilized transcription complexes. *Biophys j* **90**, 1419-1431.

69. Kubori T & Shimamoto N (1996) A branched pathway in the early stage of transcription by Escherichia coli RNA polymerase. *J Mol Biol* **256**, 449-457.

70. Margaritis T & Holstege FCP (2008) Poised RNA Polymerase II gives pause for thought. *Cell* **133**, 581-584.

71. Sims RJ, Belotserkovskaya R & Reinberg D (2004) Elongation by RNA polymerase II: the short and long of it. *Genes Dev* **18**, 2437-2468.

72. Li B, Weber J, Chen Y, Greenleaf A & Gilmour DS (1996) Analyses of promoter-proximal pausing by RNA polymerase II on the hsp70 heat shock gene promoter in a Drosophila nuclear extract. *Mol Cell Biol* **16**, 5433-5443.

73. Tang H, Liu Y, Madabusi L & Gilmour DS (2000) Promoter-proximal pausing on the hsp70 promoter in Drosophila melanogaster depends on the upstream regulator. *Mol Cell Biol* **20**, 2569-2580.

74. Krumm A, Meulia T, Brunvand M & Groudine M (1992) The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region. *Genes Dev* **6**, 2201-2213.

75. Core LJ & Lis JT (2008) Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* **319**, 1791-1792.

76. Gilmour DS (2009) Promoter proximal pausing on genes in metazoans. *Chromosoma* **118**, 1-10.

77. Fujita T & Schlegel W (2010) Promoter-proximal pausing of RNA polymerase II: an opportunity to regulate gene transcription. *J Recept Sig Trans* **30**, 31-42.

78. Li J & Gilmour DS (2011) Promoter proximal pausing and the control of gene expression. *Curr Opin Gen Dev* **21**, 231-235.

79. Wade JT & Struhl K (2008) The transition from transcriptional initiation to elongation. *Curr Opin Genetics Dev* **18**, 130-136.

80. Krumm A, Hickey LB & Groudine M (1995) Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev* **9**, 559-572.

81. Wada T, Takagi T, Yamaguchi Y, Ferdous A, Imai T, Hirose S, Sugimoto S, Yano K, Hartzog GA, Winston F, Buratowski S & Handa H (1998) DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev* **12**, 343-356.

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

83. Missra A & Gilmour DS (2010) Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. *Proc Natl Acad Sci USA* **107**, 11301-11306.

84. Brown SA, Imbalzano AN & Kingston RE (1996) Activator-dependent regulation of transcriptional pausing on nucleosomal templates. *Genes Dev* **10**, 1479-1490.

85. Orphanides G, LeRoy G, Chang C-H, Luse DS & Reinberg D (1998) FACT, a factor that facilitates transcript elongation through nucleosomes. *Cell* **92**, 105-116.

86. Wang YV, Tang H & Gilmour DS (2005) Identification in vivo of different rate-limiting steps associated with transcriptional activators in the presence and absence of a GAGA element. *Mol Cell Biol* **25**, 3543-3552.

87. Selth LA, Sigurdsson S & Svejstrup JQ (2010) Transcript elongation by RNA polymerase II. *Ann Rev Biochem* **79**, 271-293.

88. Renner DB, Yamaguchi Y, Wada T, Handa H & Price DH (2001) A highly purified RNA Polymerase II elongation control system. *J Biol Chem* **276**, 42601-42609.

89. Cheng B & Price DH (2007) Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *J Biol Chem* **282**, 21901-21912.

90. Wada T, Orphanides G, Hasegawa J, Kim D-K, Shima D, Yamaguchi Y, Fukuda A, Hisatake K, Oh S, Reinberg D & Handa H (2000) FACT relieves DSIF/NELF-mediated inhibition of transcriptional elongation and reveals functional differences between P-TEFb and TFIIH. *Mol Cell* **5**, 1067-1072.

91. Adelman K, Marr MT, Werner J, Saunders A, Ni Z, Andrulis ED & Lis JT (2005) Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. *Mol Cell* **17**, 103-112.

92. O'Brien T & Lis JT (1993) Rapid changes in Drosophila transcription after an instantaneous heat shock. *Mol Cell Biol* **13**, 3456-3463.

93. Femino AM, Fay FS, Fogarty K & Singer RH (1998) Visualization of single RNA transcripts in situ. *Science* **280**, 585-590.

94. Tennyson CN, Klamut HJ & Worton RG (1995) The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat Genet* **9**, 184-190.

95. Consortium IHGS (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

96. Darzacq X, Singer RH & Shav-Tal Y (2005) Dynamics of transcription and mRNA export. *Curr Opin Cell Biol* **17**, 332-339.

97. Singh J & Padgett RA (2009) Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* **16**, 1128-1133.

98. Rosonina E, Kaneko S & Manley JL (2006) Terminating the transcript: breaking up is hard to do. *Genes Dev* **20**, 1050-1056.

99. Gilmour DS & Fan R (2008) Derailing the Locomotive: Transcription termination. *J Biol Chem* **283**, 661-664.

100. West S, Proudfoot NJ & Dye MJ (2008) Molecular dissection of mammalian RNA polymerase II transcriptional termination. *Mol Cell* **29**, 600-610.

101. Gromak N, West S & Proudfoot NJ (2006) Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol Cell Biol* **26**, 3986-3996.

102. Proudfoot NJ (1989) How RNA polymerase II terminates transcription in higher eukaryotes. *Trends Biochem Sci* **14**, 105-110.

103. Luo W, Johnson AW & Bentley DL (2006) The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: Implications for a unified allosteric-torpedo model. *Genes Dev* **20**, 954-965.

104. Raser JM & O'Shea EK (2005) Noise in gene expression: Origins, consequences, and control. *Science* **309**, 2010-2013.

105. Elowitz MB, Levine AJ, Siggia ED & Swain PS (2002) Stochastic gene expression in a single cell. *Science* **297**, 1183-1186.

106. Raj A & van Oudenaarden A (2008) Nature, nurture, or chance: Stochastic gene expression and Its consequences. *Cell* **135**, 216-226.

107. Becskei A, Kaufmann BB & van Oudenaarden A (2005) Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat Genet* **37**, 937-944.

108. Ozbudak EM, Thattai M, Kurtser I, Grossman AD & van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. *Nat Genet* **31**, 69 - 73.

109. McAdams HH & Arkin A (1997) Stochastic mechanisms in gene expression. *Proc Natl Acad Sci USA* **94**, 814-819.

110. Kepler TB & Elston TC (2001) Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophys j* **81**, 3116-3136.

111. Raser JM & O'Shea EK (2004) Control of stochasticity in eukaryotic gene expression. *Science* **304**, 1811-1814.

112. Blake WJ, Kaern M, Cantor CR & Collins JJ (2003) Noise in eukaryotic gene expression. *Nature* **422**, 633 - 637.

113. Blake WJ, Balázsi G, Kohanski MA, Isaacs FJ, Murphy KF, Kuang Y, Cantor CR, Walt DR & Collins JJ (2006) Phenotypic consequences of promoter-mediated transcriptional noise. *Mol Cell* **24**, 853-865.

114. Lewis J (2003) Autoinhibition with transcriptional delay: A simple mechanism for the Zebrafish somitogenesis oscillator. *Curr Biol* **13**, 1398-1408.

115. Roussel M & Zhu R (2006) Stochastic kinetics description of a simple transcription model. *Bull Math Biol* **68**, 1681-1713.

116. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* **81**, 2340 - 2361.

117. Oppenheim I, Shuler K & Weiss G (1977) Stochastic processes in chemical physics: The master equation. *MIT Press, Cambridge, MA*.

118. Gillespie DT (1992) A rigorous derivation of the chemical master equation. *Physica A* **188**, 404-425.

119. Roussel MR & Zhu R (2006) Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. *Phys Biol* **3**, 274 - 284.

120. Zhu R, Ribeiro AS, Salahub D & Kauffman SA (2007) Studying genetic regulatory networks at the molecular level: Delayed reaction stochastic models. *J Theor Biol* **246**, 725-745.

121. Ribeiro AS, Smolander OP, Rajala T, Häkkinen A & Yli-Harja O (2009) Delayed stochastic model of transcription at the single nucleotide level. *J Comp Biol* **16**, 539 - 553.

122. Rajala T, Häkkinen A, Healy S, Yli-Harja O & Ribeiro AS (2010) Effects of transcriptional pausing on gene expression dynamics. *PLoS Comput Biol* **6**, e1000704.

123. Ribeiro AS, Häkkinen A, Mannerström H, Lloyd-Price J & Yli-Harja O (2010) Effects of the promoter open complex formation on gene expression dynamics. *Phys Rev E* **81**, 011912.

124. Ribeiro AS, Häkkinen A, Healy S & Yli-Harja O (2010) Dynamical effects of transcriptional pause-prone sites. *Comp Biol Chem* **34**, 143-148.

125. Voliotis M, Cohen N, Molina-ParÌs C & Liverpool TB (2008) Fluctuations, pauses, and backtracking in DNA transcription. *Biophys J* **94**, 334-348.

126. Voliotis M, Cohen N, Molina-Paris C & Liverpool TB (2009) Backtracking and proofreading in DNA transcription. *Phys Rev Lett* **102**, 258101.

127. Mäkelä J, Lloyd-Price J, Yli-Harja O & Ribeiro A (2011) Stochastic sequence-level model of coupled transcription and translation in prokaryotes. *BMC Bioinformatics* **12**, 121.

128. Raj A, Peskin CS, Tranchina D, Vargas DY & Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* **4**, e309.

129. Harper CV, Finkenstädt B, Woodcock D J, Friedrichsen S , Semprini S, Ashall L, Spiller DG, Mullins JJ, Rand DA, Davis JRE & White MRH (2011) Dynamic analysis of stochastic transcription cycles. *PLoS Biol* **9**, e1000607.

130. Chubb JR & Liverpool TB (2010) Bursts and pulses: insights from single cell studies into transcriptional mechanisms. *Curr Opin Gen Dev* **20**, 478-484.

131. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comp Phys* **22**, 403-434.

132. McCollum JM, Peterson GD, Cox CD, Simpson ML & Samatova NF (2006) The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. *Comp Biol Chem* **30**, 39-49.

133. Gibson MA & Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. *J Phys Chem A* **104**, 1876-1889.

134. Cao Y, Li H & Petzold L (2004) Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *J Chem Phys* **121**, 4059-4067.

135. Gillespie DT (2007) Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem* **58**, 35-55.

136. Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. *J Chem Phys* **115**, 1716-1733.

137. Bratsun D, Volfson D, Tsimring LS & Hasty J (2005) Delay-induced stochastic oscillations in gene regulation. *Proc Natl Acad Sci USA* **102**, 14593-14598.

138. Ribeiro AS (2010) Stochastic and delayed stochastic models of gene expression and regulation. *Math Biosci* **223**, 1-11.

139. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W & Selbach M (2011) Global quantification of mammalian gene expression control. *Nature* **473**, 337-342.

140. Boeger H, Griesenbeck J & Kornberg RD (2008) Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. *Cell* **133**, 716-726.

141. Mazloom AR, Basu K, Mandal SS & Das SK (2010) Chromatin remodeling in silico: A stochastic model for SWI/SNF. *Biosystems* **99**, 179-191.