Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(6): 6225–6262.
DOI: 10.3934/mbe.2024273
Received: 19 November 2023
Revised: 01 March 2024
Accepted: 24 April 2024
Published: 13 June 2024

Research article

Analytic delay distributions for a family of gene transcription models

S. Hossein Hosseini and Marc R. Roussel*

Alberta RNA Research and Training Institute, Department of Chemistry and Biochemistry, University of Lethbridge, Lethbridge, AB T1K 3M4, Canada

* **Correspondence:** Email: roussel@uleth.ca; Tel: +1 403 329 2326.

Abstract: Models intended to describe the time evolution of a gene network must somehow include transcription, the DNA-templated synthesis of RNA, and translation, the RNA-templated synthesis of proteins. In eukaryotes, the DNA template for transcription can be very long, often consisting of tens of thousands of nucleotides, and lengthy pauses may punctuate this process. Accordingly, transcription can last for many minutes, in some cases hours. There is a long history of introducing delays in gene expression models to take the transcription and translation times into account. Here we study a family of detailed transcription models that includes initiation, elongation, and termination reactions. We establish a framework for computing the distribution of transcription times, and work out these distributions for some typical cases. For elongation, a fixed delay is a good model provided elongation is fast compared to initiation and termination, and there are no sites where long pauses occur. The initiation and termination phases of the model then generate a nontrivial delay distribution, and elongation shifts this distribution by an amount corresponding to the elongation delay. When initiation and termination are relatively fast, the distribution of elongation times can be approximated by a Gaussian. A convolution of this Gaussian with the initiation and termination time distributions gives another analytic approximation to the transcription time distribution. If there are long pauses during elongation, because of the modularity of the family of models considered, the elongation phase can be partitioned into reactions generating a simple delay (elongation through regions where there are no long pauses), and reactions whose distribution of waiting times must be considered explicitly (initiation, termination, and motion through regions where long pauses are likely). In these cases, the distribution of transcription times again involves a nontrivial part and a shift due to fast elongation processes.

Keywords: gene transcription; delay distribution; transcriptional pausing

1. Introduction

Transcription is a complex process for the biosynthesis of RNA from a DNA template. Roughly, this process can be divided into three phases: initiation, elongation and termination [1]. During initiation, a set of protein factors bind the DNA template upstream of the start site and facilitate the binding and positioning of RNA polymerase [2,3]. The first few nucleotides are added to the nascent RNA, and the polymerase transitions into productive elongation thereafter. Elongation consists of a sequence of rapid nucleotide addition steps [4]. While each individual step is fast (over $70\,\mathrm{nt\,s^{-1}}$ [4]; nt = nucleotide), genes can be very long. The average human gene, which contains many introns that must be spliced out to form the final transcript, has a length of 28 000 nt [5], and much longer genes are known [6], so the elongation time can be significant. Finally, transcription must be terminated. In eukaryotes, redundant mechanisms cooperate to ensure termination [7].

Transcription is only the first step in gene expression. In eukaryotes, the primary transcript is subjected to processing [8], which includes splicing, 5' capping, and 3' polyadenylation; the completed messenger RNA (mRNA) is packaged with proteins into a messenger ribonucleoprotein (mRNP) complex that is required for export from the nucleus [9]; it must make its way to the nuclear pore and be exported to the cytoplasm [10]; it may need to be shuttled to a specific part of the cell [11]; and finally it is translated to a protein. Since splicing [12], 5' capping [13] and mRNP assembly [14] are generally co-transcriptional, a minimal model of gene expression in eukaryotes might consider just transcription, nuclear export, and translation. All three of these major processes may be separately regulated, so it is not, in general, possible to collapse them into a single protein-synthesis step as is often done in gene expression models. If we want to model a gene network however, we almost certainly do not want to model transcription, nuclear export and translation in detail. Rather, we would prefer to treat each of these processes as effective reactions; one reaction for transcription, one for export, and one for translation. If we intend to model the temporal evolution of a gene network, it will be necessary to consider the time required for each of these processes. Human genes are very large with, as noted above, an average size of 28 kb (kilobases) [5].* In the absence of complicating factors (pauses, co-transcriptional splicing), elongation proceeds at $4.3\,\mathrm{kb\,min^{-1}}$ (approximately $70\,\mathrm{nt\,s^{-1}}$) [4], so the transcription elongation phase for a typical human gene would take approximately 6.5 min. Some genes are larger, and pausing for splicing can add substantially to the transcription time [6]. Moreover, some events in transcription initiation can be similarly slow [15, 16]. Nuclear export can take 10 to 20 minutes [17]. Because the coding sequence is much shorter than the original gene—in humans, the coding sequence is only about 4% of the gene, the rest being intronic sequences removed by splicing [5]—translation will usually be much faster. Accordingly, transcription, including pauses associated with splicing, and nuclear export will tend to dominate the gene expression time. One way to avoid detailed models of these processes while still considering the expression time of a gene is to incorporate delays [18–22], which may be fixed or distributed [23–25], into a model containing effective reactions. But this of course means that we need to know something about the distributions of these delays.

*For practical purposes, nucleotides and bases are interchangeable in the discussion of rates of templated polynucleotide synthesis (transcription, DNA replication, etc.) and related statistics. Typically, nucleotides are used to count the units being added to the polynucleotide (RNA, DNA), while bases (usually kb) are used to measure the template. Since one nucleotide is added for each complementary base in the template, the rate of synthesis can equivalently be expressed in nt or in bases per unit time. However, 'knt' is, to our knowledge, never used, so when SI prefixes are to be used, distances along the template are always measured in bases, hence rates measured in $\mathrm{nt\,s^{-1}}$ or $\mathrm{kb\,min^{-1}}$.

In this paper, we will compute the transcriptional delay distribution, the transcriptional delay being defined here as the delay before the appearance of the product RNA *given that* the gene is active at time zero, in a class of models of varying degrees of complexity. Although the transcription model studied here is similar to prokaryotic models presented elsewhere [26, 27], we focus on a model and on parameters appropriate to eukaryotes [28]. We further narrow the focus to protein-coding genes transcribed by RNA polymerase II (RNAPII). In eukaryotes, most protein-coding genes are rarely transcribed so that it will be rare for two polymerases to interact [29]. Even if we take into account that RNAs are often produced in bursts [30–32], polymerases tend to accelerate as they travel down the gene [33,34], which should minimize polymerase-polymerase interactions, at least for polymerases that have translocated sufficiently far along the gene. Accordingly, we study a single-polymerase model.

There have been a number of models of the kinetics of transcription. Several models have focused on the mechanics and statistical thermodynamics of elongation [35–39], and some of these models have yielded insights into the factors responsible for the transcriptional delay. In prior work, we proposed a single-gene prokaryotic transcription model from which we obtained analytical and numerical results for the distribution of the transcriptional delay [27]. We also studied the case where many polymerases transcribe the same gene and obtained stochastic simulation results [26–28]. A similar approach was followed by von Hippel's group, obtaining time-dependent solutions from which a delay distribution could be inferred [40]. Filatova and coworkers have obtained distributions of elongation times as well as means and variances of various stages of transcription for a somewhat simpler transcription model than is considered here, but also considering random activation and deactivation of the gene [41]. Boettiger and coworkers also obtained distributions of transcription times focusing on initiation and treating elongation and termination as a single Poisson process [42]. Xu and colleagues have examined the dwell time in paused states [43]. There have been a number of other modeling studies of transcription of varying complexity and with a variety of objectives [44–47] (e.g.). Of course, there is a formal similarity between transcription and translation, which are both templated processes, so we find similar work also in translation modeling. Garai and coworkers have computed both dwell times at a codon, and distributions of translation times [48, 49]. Perhaps most closely related to the present contribution is the work of Mier-y-Terán-Romero and coworkers, who showed that for a long mRNA, an advection equation is obtained in the continuum limit such that a fixed elongation delay emerges for the process of translation [50].

Single-molecule optical trapping experiments show that the movement of RNA polymerase along the DNA strand is not continuous and is punctuated many times by pauses [51, 52]. Mechanistically, there are two main classes of pauses: (1) backtracking pauses, in which RNAPII slides backward reversibly along both the DNA and the RNA, and (2) non-backtracking pauses, in which conformational changes in the RNA polymerase active site stop the nucleotide addition cycle [52, 53]. Backtracking pauses can be affected by the presence of a trailing RNA polymerase that can restrict how far an RNA polymerase can backtrack [54], leading to density dependence of the transcription rate [55]. Voliotis et al. [56] studied a model of backtracking transcriptional pauses, for which they obtained a distribution of transcription times. They found that backtracking leads to a heavy-tailed distribution, a conclusion also reached by Klumpp [55].

Pauses can occur in many parts of a gene, and for many different reasons. In eukaryotes, the polymerase almost always pauses just downstream of the start site in an event known as promoter-proximal pausing [57–59]. These pauses are thought to have a regulatory role. Pausing can be associated with

features of the sequence [36, 60, 61]. Splicing can also cause pausing [62]. Moreover, pausing facilitates termination of transcription [63].

We treat transcription both with and without pauses, and we consider both short ubiquitous pauses and specific pause sites where lengthy pauses occur. In Section 2, we present a simplified version of Vashishtha's model of eukaryotic transcription [28] that includes assembly of the pre-initiation complex, initiation, elongation and termination. Vashishtha additionally considered abortive initiation, and distinguished early elongation (a promoter escape phase with the possibility of promoter-proximal pausing) from productive elongation. With the exception of abortive initiation, the neglected kinetic processes are captured in later sections of this work, albeit in somewhat different form. In Section 3, we write the chemical master equations corresponding to each phase of transcription. For both initiation and termination phases, since there are few ordinary differential equations (ODEs), the time-dependent probability for being in each state of the model can easily be obtained. For the elongation phase, we again write down the chemical master equation. We find that there are two distinct cases. When the elongation phase is short in duration compared to initiation or termination, which will be the case for many genes for which initiation is slow, the polymerase's motion can be treated as advective, and the elongation phase can then be treated as introducing a fixed delay between initiation and termination. As noted earlier, the advective treatment developed here is similar in spirit, if not in detail, to the work of Mier-y-Terán-Romero and coworkers on translation [50]. On the other hand, when the duration of the elongation phase becomes significant compared to the slower of initiation or termination, given that most genes are relatively long, an application of the central-limit theorem predicts a Gaussian distribution of elongation times. In both cases, it is possible to obtain an analytic expression for the distribution of the total transcription time. Because of the modularity of the class of models considered [40] such that the total delay is just a sum of delays at each nucleotide, we can treat long pause sites separately, which we do in Section 4. For simplicity, we go back to the case where elongation is fast, and we consider a single strong pause site, obtaining the distribution of the total transcriptional delay in this case. In all our calculations to this point, we assume that binding of the RNA polymerase to the promoter is irreversible. In Section 5, we relax this assumption and discover that the distribution of transcription times can be sensitive to the reversibility of RNAP binding when the transition to elongation is slow. Finally, in Section 6, we discuss a number of possible extensions and refinements, notably the treatment of pause exit due to the action of a release factor, and the possibility of using our results as a basis for simulating genes expressed in bursts.

This work is situated in a larger literature on the use of delays in chemical and biochemical models. We can think of the states appearing between initiation and the completion of transcription as intermediates in a synthesis pathway. The problem of replacing repeated steps generating a sequence of intermediates by distributed delays has been intensely studied, notably by MacDonald [64–66], and by Barrio and coworkers [67, 68]. Epstein has additionally considered the issue of "bottleneck intermediates" [69], of which strong pause sites would be an example. We have already mentioned the contribution of Mier-y-Terán-Romero and coworkers [50], which provided a justification for the use of delays in modeling translation. The distinction made in this paper between cases where elongation is fast and those where elongation is relatively slow adds to our understanding of the proper use of fixed and distributed delays in gene expression models. Moreover, the availability of analytic expressions for the distribution of transcription times in a variety of scenarios will be useful for delay-stochastic simulations of gene expression networks [70, 71].

2. Eukaryotic transcription model

We start with a simplified version of Vashishtha's kinetic model for the eukaryotic transcription process [28], removing abortive initiation, early elongation/promoter escape and promoter-proximal pausing. Promoter-proximal pausing [72, 73] can, at least to a first approximation, be modeled as a distinct, long pause, as we will do in section 4. Similarly, the early elongation phase, during which the polymerase adds nucleotides more slowly than it does in the productive elongations phase [74, 75], could be treated analogously to productive elongation, but with different kinetic parameters. Abortive initiation [76] represents an alternative pathway that can be included in a transcription model [28]. The effect of abortive initiation would be to reduce the frequency of initiation. Since we focus here on events that lead to the production of a protein, we leave this out. We similarly leave the other refinements mentioned in this paragraph to later work. Models of a similar degree of complexity are sufficient to analyze some experimental datasets [77]. Moreover, note that the simplest model considered in this contribution is similar to previously studied models of bacterial transcription [26,27]. With some reinterpretation of symbols, the results obtained here could thus be applied to bacterial models, and perhaps also to structurally similar translation models [50, 78, 79].

Before describing the transcription model, it is useful to think about the context in which the delay distributions derived in later sections are likely to be used. In delay-stochastic simulations [70, 71], we might model transcription as the simple delayed mass-action [80] process

$$\mathrm{pro}(t) + \mathrm{RNAPII}(t) \xrightarrow{k_{\mathrm{ini}}} \mathrm{pro}(t) + \mathrm{RNAPII}(t+\tau_{\mathrm{tr}}) + \mathrm{RNA}(t+\tau_{\mathrm{tr}}), \tag{1}$$

where pro(t) is understood to be a promoter that is able to initiate transcription, i.e., one that is in the 'on' state. The activity of a gene is regulated by transcription factors acting either as repressors or as activators. These interactions would be modeled by separate mass-action processes, e.g.,

$$\mathrm{pro} + \mathrm{R} \rightleftharpoons \mathrm{pro}\cdot\mathrm{R}, \tag{2a}$$

$$\mathrm{R} + \mathrm{L} \rightleftharpoons \mathrm{R}\cdot\mathrm{L} \tag{2b}$$

for a repressor R that is inactivated by binding to a ligand L. Thus, regulatory processes that turn a gene on and off would be modeled by their own reactions and would not be involved in determining the distribution of the delay τ_{tr}, which is the focus of this contribution. In other words, we study the 'pure' transcriptional process on the assumption that binding or conformational transition events that turn a gene on or off are handled separately.

Having narrowed our focus as noted above, the transcription model is divided into the usual three phases: initiation, elongation, and termination. This section is correspondingly organized.

2.1. Initiation

The initiation phase starts with binding of initiation factors, including the TATA-binding protein (TBP), to the TATA box of the promoter region (pro) of a gene [reaction (3)], following which RNAPII can bind to the promoter, forming the pre-initiation complex (PIC) [reaction (4)] [2]:

$$\mathrm{TBP} + \mathrm{pro} \xrightarrow{k^{(2)}_{\mathrm{bind}}} \mathrm{TBP}\cdot\mathrm{pro}, \tag{3}$$

$$\mathrm{TBP}\cdot\mathrm{pro} + \mathrm{RNAPII} \xrightarrow{k^{(2)}_{\mathrm{PIC}}} \mathrm{PIC}. \tag{4}$$

The rate constants $k^{(2)}_{\mathrm{bind}}$ and $k^{(2)}_{\mathrm{PIC}}$ are second-order rate constants; the superscripts appear for emphasis. The reverse of reaction (3) is neglected because TBP dissociation from DNA is extremely slow, with a half-life of several minutes [81, 82]. TBP probably dissociates rapidly following transcription initiation [83], but we need not consider the recycling of this factor given that we focus here on a single polymerase. Dissociation of RNAPII from the promoter is, on the other hand, a frequent event [4], so $k^{(2)}_{\mathrm{PIC}}$ is best thought of as an effective rate constant that takes into account the repeated binding and dissociation of the polymerase prior to forming a stable PIC, at least for now. We consider the effect of reversible binding of the polymerase explicitly in Section 5.

After forming a stable PIC, the RNA polymerase moves to the first nucleotide of the DNA template. Reaction (5) represents the translocation of the active site of the RNA polymerase to the first nucleotide (nucleotide 1) if this nucleotide is unoccupied (U state), thus converting U_1 into an occupied nucleotide (O state):

$$\mathrm{PIC} + \mathrm{U}_1 \xrightarrow{k_{\mathrm{init}}} \mathrm{O}_1. \tag{5}$$

Since we only consider a single polymerase here, the forward site will always be unoccupied. We maintain this notation for consistency with our earlier work [26–28]. Note that the engagement of the polymerase represented by reaction (5) is likely a complex process. However, reaction (5) will be a reasonable approximation if this process has a single rate-limiting step.

The minimal initiation model considered here captures the slowest steps in a single round of initiation at promoters that include a TATA box. Initiation models can be much more complicated [42]. We could, for example, allow for reinitiation by having either the promoter or $\mathrm{TBP}\cdot\mathrm{pro}$ become available again after promoter clearance (i.e., after the polymerase has traveled sufficiently far downstream).[†] This is easily done in stochastic simulations [28]. However, here we want to focus on analytic distributions obtained for a single polymerase. Some analytic progress can be made for a model with multiple polymerases [84]. Some of these results however depend on the single-polymerase distributions, so the work presented here is an important starting point. Moreover, not all genes display transcriptional bursting [77], so it is worth considering initiation of a single polymerase.

From here and until we reach the termination phase, transcription consists of cycling through binding and recognition of a nucleoside triphosphate complementary to the template nucleotide, translocation to the next template position, and formation of a phosphodiester bond, which adds the selected nucleotide to the growing RNA chain. The early steps are slower than the corresponding steps in productive elongation [74] but, for simplicity, we ignore these kinetic differences in this contribution. Thus, the binding of the first nucleotide is treated as indistinguishable from binding of any other nucleotide and is included in the elongation part of the model.

2.2. Elongation

Although conceptually there are at least three distinct processes in elongation (binding of a complementary nucleoside triphosphate, translocation, catalysis of phosphodiester bond formation), the simplest models of transcription describe the process in terms of two states [27, 85], roughly corresponding to the pre- and post-translocated states. More [35, 36, 86] or fewer [40] states can be used to

[†]*In vivo*, $\mathrm{TBP}\cdot\mathrm{pro}$ is rapidly removed after initiation by Mot1 [83]. Including a reaction to represent the action of Mot1 following clearance of the promoter would likely lead to transcriptional bursts, a topic to which we will return in the discussion.

model elongation. In our models [26–28], the two states involved in productive elongation are called occupied (O) and activated (A) (Figure 1), corresponding respectively to the pre- and post-translocated states.

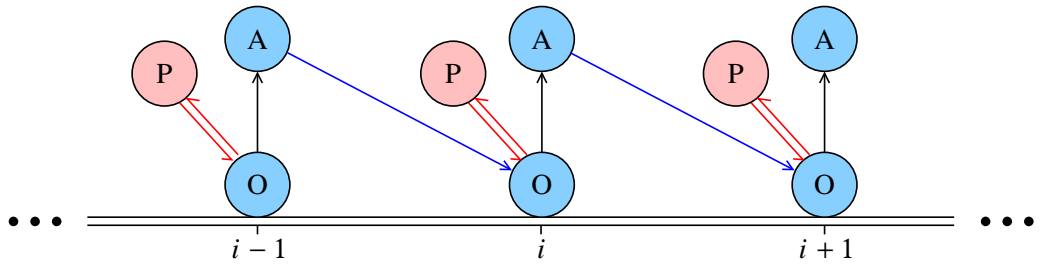

Figure 1. Schematic of the elongation phase showing the occupied (O), activated (A) and paused (P) states at three consecutive nucleotides.

The process of activation and translocation of RNA polymerase repeats itself as RNAPII translocates nucleotide by nucleotide along the gene. These repeated reactions are responsible for productive elongation [reactions (6) and (7)].

$$\mathrm{O}_i \xrightarrow{k_{\mathrm{ae}}} \mathrm{A}_i, \qquad i = 1, 2, \ldots, N; \tag{6}$$

$$\mathrm{A}_i + \mathrm{U}_{i+1} \xrightarrow{k_{\mathrm{add}}} \mathrm{O}_{i+1} + \mathrm{U}_i, \quad i = 1, 2, \ldots, N-1. \tag{7}$$

Here, N is the length of the template, i.e., the transcribed sequence.

During elongation, the polymerase may pause for any of a number of reasons: There are so-called "ubiquitous" pauses that may occur in any part of the gene [51], promoter-proximal pauses associated with a change in state of the polymerase as it transitions from initiation to elongation [87], pauses associated with splicing [88], and pauses associated with termination [63]. Pausing models can be more or less complex and incorporate backtracking [56] or arrest [89]. Here we opt for a simple model of a reversible pause. We neglect backtracking, which is likely appropriate for short, ubiquitous pauses, but may be less so for long pauses. It appears that pausing proceeds from the pre-translocated (O) state [89, 90], so we write

$$\mathrm{O}_i \underset{k_{r,i}}{\overset{k_{p,i}}{\rightleftharpoons}} \mathrm{P}_i, \qquad i = 1, 2, \ldots, N. \tag{8}$$

We allow here for the possibility that the rate constants for pausing, $k_{p,i}$, and for the release from the pause, $k_{r,i}$, might be template-dependent. On the other hand, we need make no such allowance for the rate constants associated with productive elongation, k_{ae} and k_{add}, which are believed to be constant over the length of a gene and across genes within a particular organism [91]. Figure 1 shows the off-pathway P state along with the O and A states.

2.3. Termination

According to the allosteric model of termination, transcription of the poly(A) signal at the end of the gene can cause a termination-inducing conformational change in the elongation complex [92]. After the last reaction of elongation [activation at nucleotide N, reaction (6)], the elongation complex is converted into a termination complex (TC) [reaction (9)]. Finally, the termination complex dissociates

and releases the RNA [reaction (10)].

$$A_N \xrightarrow{k_{TC}} TC, \tag{9}$$

$$TC \xrightarrow{k_{term}} RNA + RNAPII + U_N. \tag{10}$$

The above reactions assume 100% efficiency of allosteric termination. In reality, there is at least one, and possibly two failsafe mechanisms backing up allosteric termination [7]. We assume a single mechanism here for simplicity. In terms of the completion of the transcript, the torpedo model will behave similarly to allosteric termination since its first step is to cleave the transcript at the poly(A) signaling sequence, releasing the former from the transcriptional machinery [7].

3. Time-dependent probability distribution for transcription by RNAPII

3.1. Initiation

Figure 2 shows a schematic of the model. RNA polymerases move one nucleotide at a time along the template strand [93], so each of the numbered sites in the diagram represents one nucleotide. The polymerase initiates at the promoter, represented by a single site labeled '0' in the diagram, and terminates at the end of the transcribed sequence, represented here as the N'th nucleotide. By solving the chemical master equation (CME) for each state in the three phases of transcription, we can obtain time-dependent probability distributions for the progress of transcription, from which we can recover the transcriptional delay distribution.

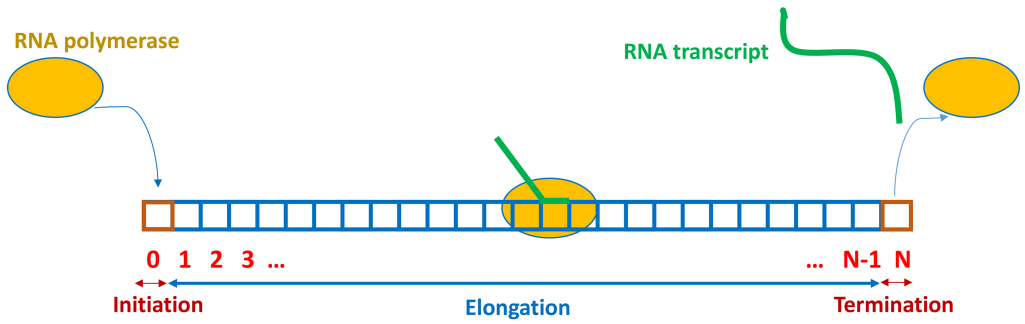

Figure 2. A schematic representation of RNA polymerase on the DNA template strand.

In principle, the propensities of reactions (3) and (4) depend on the populations of available TBP and RNAPII, respectively. Here, we assume that there are constant pools of these two species in the nucleus so that we can treat reactions (3) and (4) as pseudo-first order reactions, with rate constants $k_{bind} = k^{(2)}_{bind}[TBP]$, and $k_{PIC} = k^{(2)}_{PIC}[RNAPII]$, respectively. The governing master equations are

$$\frac{dP_{pro}(t)}{dt} = -k_{bind}P_{pro}(t), \tag{11a}$$

$$\frac{dP_{TBP\cdot pro}(t)}{dt} = k_{bind}P_{pro}(t) - k_{PIC}P_{TBP\cdot pro}(t), \tag{11b}$$

$$\frac{dP_{\mathrm{PIC}}(t)}{dt} = k_{\mathrm{PIC}} P_{\mathrm{TBP\cdot pro}}(t) - k_{\mathrm{init}} P_{\mathrm{PIC}}(t), \tag{11c}$$

with initial condition $P_{\mathrm{pro}}(0) = 1$, $P_{\mathrm{TBP\cdot pro}}(0) = P_{\mathrm{PIC}}(0) = 0$. In these equations, $P_{\mathrm{pro}}(t)$ is the probability that the promoter has never been bound up to time t, i.e., it is the survival probability of the empty promoter. Since we treat the single-polymerase case, we do not need to consider promoter clearance [94]. $P_{\mathrm{TBP\cdot pro}}(t)$ is the probability that, at time t, TBP is bound to the promoter. Similarly, $P_{\mathrm{PIC}}(t)$ is the probability that a complete PIC is assembled at the promoter at time t. The initial condition corresponds to a gene that has been turned on at $t = 0$, prior to which TBP would not have been able to bind the promoter. It is possible to consider other initial conditions. For example, for some modes of regulation, it might be possible for TBP to bind the promoter while the polymerase is unable to bind until the gene is turned on. In this case, we would choose an initial condition such that $P_{\mathrm{pro}}(0) + P_{\mathrm{TBP\cdot pro}}(0) = 1$ and $P_{\mathrm{pro}}(0)/P_{\mathrm{TBP\cdot pro}}(0) = K_d^{(\mathrm{eff})}$, where $K_d^{(\mathrm{eff})}$ is the effective dissociation constant for the TBP $\cdot$ pro complex at cellular TBP concentrations.

The time-dependent rate at which probability exits initiation and enters elongation is, from reaction (5),

$$v_{\mathrm{init}}(t) \equiv \rho_{\mathrm{init}}(t) = k_{\mathrm{init}} P_{\mathrm{PIC}}(t), \tag{12}$$

where $\rho_{\mathrm{init}}(t)$ is the distribution of initiation times. The equality of v_{init} and ρ_{init} follows because the integral of the rate at which initiation is completed is the cumulative distribution of initiation times.

The linear ordinary differential equations (11a) to (11c) can be solved to find the probability distribution of being in each state. Equation (12) can then be used to obtain the distribution of initiation times. The latter is (assuming all of the rate constants are distinct)

$$\rho_{\mathrm{init}}(t) = \frac{k_{\mathrm{bind}}\, k_{\mathrm{PIC}}\, k_{\mathrm{init}}}{(k_{\mathrm{PIC}} - k_{\mathrm{bind}})(k_{\mathrm{bind}} - k_{\mathrm{init}})(k_{\mathrm{init}} - k_{\mathrm{PIC}})}\Big[(k_{\mathrm{PIC}} - k_{\mathrm{init}})e^{-k_{\mathrm{bind}}t} + (k_{\mathrm{init}} - k_{\mathrm{bind}})e^{-k_{\mathrm{PIC}}t} + (k_{\mathrm{bind}} - k_{\mathrm{PIC}})e^{-k_{\mathrm{init}}t}\Big]. \tag{13}$$

Figure 3 shows $\rho_{\mathrm{init}}(t)$. The parameters used in this study were estimated by Vashishtha [28] and are given in Table 1. They are likely representative of genes expressed at lower rates, although Vashishtha showed that they were consistent with the average behavior of transcription in mammalian cells [95]. Briefly, the rate of TBP binding to the TATA box is assumed not to be limited by reaction (3), but by dissociation of TBP dimers, so the effective rate constant for this non-elementary step is the rate constant for dimer dissociation, $k_{\mathrm{bind}} = 0.0016\,\mathrm{s}^{-1}$ [15]. The rate constant for PIC assembly, $k_{\mathrm{PIC}} = 0.0029\,\mathrm{s}^{-1}$, was taken from an *in vitro* study [96]. Finally, based on experimental data for the progress of the polymerase through the first few nucleotides [96], Vashishtha inferred a lower bound for k_{init} of $0.3\,\mathrm{s}^{-1}$ [28]. The value of k_{init} eventually adopted is somewhat arbitrary, but in any event, it is clear that binding of TBP and PIC assembly are the rate limiting steps. Although the parameters chosen for this study may represent a gene at which initiation is particularly inefficient, typical eukaryotic initiation times are relatively long [1, 76, 96], resulting in well-spaced polymerases [97], which supports our decision to focus on a single-polymerase model.

3.2. Elongation

3.2.1. Continuum approximation

The distribution of transcription times is an N-fold convolution of passage times through each nucleotide. For very simple models, this convolution is tractable, and is the basis of the well-known

Table 1. Default values of the rate constants used in this study.

Rate constant	**Value/**s^{-1}
k_{bind}	0.0016
k_{PIC}	0.0029
k_{init}	0.6
k_{ae}	144
k_{add}	144
$k_{p,i}$	100
$k_{r,i}$	3.6
k_{TC}	0.0032
k_{term}	0.003

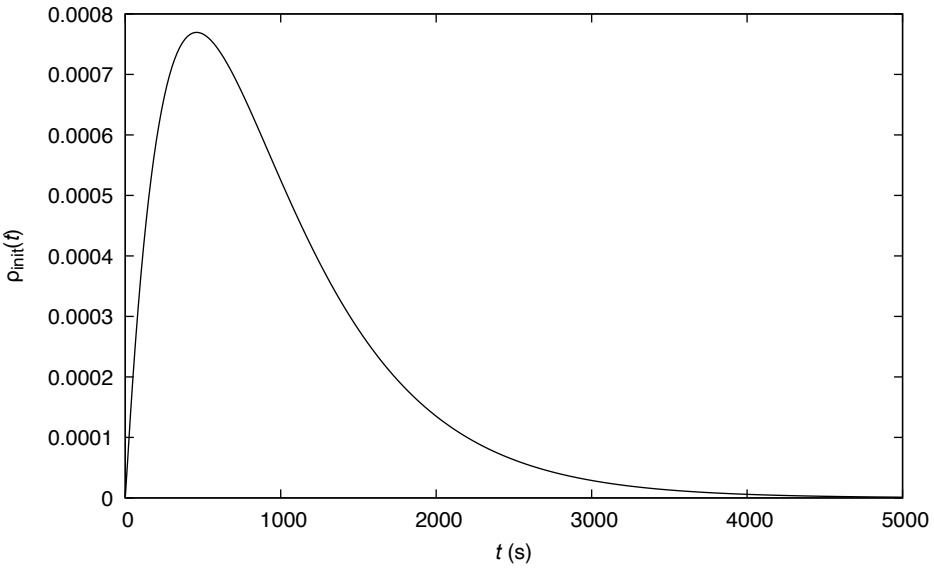

Figure 3. $\rho_{\text{init}}(t)$ from Eq (13) with the parameters of Table 1.

"linear chain trick" [64–66, 98]. This problem is less straightforward if we take into account deviations from the sequential addition of nucleotides such as pausing. For typical protein-coding genes in eukaryotes, N might range from several hundred nucleotides to tens of thousands [5]. The large value of N for typical genes and the high rates of elongation [4] afford an opportunity to make a continuum approximation. The methods applied here are somewhat analogous to those used to obtain a wave equation in a chain of coupled oscillators [99] and have a long history in mathematical biology [100, 101]. Note also the close analogy to the derivation of delays in age-structured models of populations [102].

Denote by $P_{i,\sigma}$ the probability that the polymerase has its active site at nucleotide i and that it is in state σ, where σ can take on the values O, P or A. The master equations governing the time evolution of the elongation phase are

$$\frac{dP_{i,\mathrm{O}}(t)}{dt} = k_{\mathrm{add}}P_{i-1,\mathrm{A}}(t) - k_{\mathrm{ae}}P_{i,\mathrm{O}}(t) - k_{p,i}P_{i,\mathrm{O}}(t) + k_{r,i}P_{i,\mathrm{P}}(t), \tag{14a}$$

$$\frac{dP_{i,\mathrm{A}}(t)}{dt} = k_{\mathrm{ae}}P_{i,\mathrm{O}}(t) - k_{\mathrm{add}}P_{i,\mathrm{A}}(t), \tag{14b}$$

$$\frac{dP_{i,\mathrm{P}}(t)}{dt} = k_{p,i}P_{i,\mathrm{O}}(t) - k_{r,i}P_{i,\mathrm{P}}(t), \tag{14c}$$

where $i = 2, 3, ..., N-1$. (Equation (14a) is slightly modified for $i = 1$ since the O state is reached from the PIC state according to reaction (5). This does not alter the argument below in any substantial way.)

Because the states are mutually exclusive, the probability that the polymerase is at nucleotide i in any state, P_i, is the sum of the probabilities $P_{i,\sigma}$, that is

$$P_i(t) = P_{i,\mathrm{O}}(t) + P_{i,\mathrm{A}}(t) + P_{i,\mathrm{P}}(t). \tag{15}$$

Using Eqs (14)–(15), we obtain

$$\frac{dP_i}{dt} = k_{\mathrm{add}}\left(P_{i-1,\mathrm{A}} - P_{i,\mathrm{A}}\right). \tag{16}$$

We want to treat the motion on the time scale of elongation. We therefore introduce the rescaled time

$$\theta = t/\tau_e, \tag{17}$$

where τ_e is the mean total elongation time. This quantity will be computed below. Introducing this scaling in Eq (16), we get

$$\frac{dP_i}{d\theta} = k_{\mathrm{add}}\tau_e\left(P_{i-1,\mathrm{A}} - P_{i,\mathrm{A}}\right). \tag{18}$$

For long genes, an increment of one nucleotide is a small change in position measured relative to the length of the gene. Therefore define

$$x = i/N, \tag{19}$$

$$\delta = 1/N. \tag{20}$$

In a continuous representation, the $P_{i,\mathrm{A}}(\theta)$, which are supported on the natural numbers, will become continuous functions of x, $\rho_{\mathrm{A}}(x,\theta)$, with the correspondence $P_{i,\mathrm{A}}(\theta) = \delta\rho_{\mathrm{A}}(i\delta,\theta)$. Similarly, we define

the probability density for all states combined by $P_i(\theta) = \delta\rho_e(i\delta, \theta)$. Because δ is small, from the definition of the derivative, we have

$$\begin{aligned} P_{i-1,\mathrm{A}}(\theta) - P_{i,\mathrm{A}}(\theta) &= \delta\left(\rho_A(x-\delta,\theta) - \rho_A(x,\theta)\right) \\ &\approx -\delta^2 \frac{\partial \rho_{\mathrm{A}}(x,\theta)}{\partial x}. \end{aligned} \tag{21}$$

This will hold provided the $P_{i,\mathrm{A}}$ vary sufficiently slowly. Otherwise, the estimate of the derivative on which Eq (21) is based is meaningless. This approach will therefore fail if the gene contains a specific site at which probability accumulates for a time during elongation due to significant slowing of elongation over a short stretch of the sequence. Take the extreme case where the residence time at one particular nucleotide, i, is several orders of magnitude larger than at adjacent nucleotides: then there will be a period of time of $O(\tau_i)$ over which $P_i(t) \sim O(1)$ with $P_j(t) \approx 0$ at adjacent nucleotides and where the derivative estimate of Eq (21) is therefore invalid. Slowing could be due to any of the elongation parameters, but will most commonly occur because of a strong pause site. We treat the latter in Section 4. On the other hand, slowly varying parameters, including those associated with pausing, do not in principle pose any particular difficulty for this approach. We note in passing that the treatment of translation by Mier-y-Terán-Romero and coworkers also required sufficiently slowly varying parameters.

Substituting (21) and the definitions of the densities into (18), we get the PDE

$$\frac{\partial \rho_e}{\partial \theta} = -k_{\mathrm{add}}\tau_e \delta \frac{\partial \rho_{\mathrm{A}}}{\partial x}. \tag{22}$$

Equation (22) is not closed as it relates the time derivative of ρ_e to the gradient of ρ_A. If we examine the state of a specified nucleotide from an ensemble of similarly prepared systems at an arbitrary moment in time,[‡] the conditional probability of catching it in a particular state given that the polymerase has reached this nucleotide is proportional to its mean dwell time in that state. In particular,

$$P_{i,\mathrm{A}} = P_i P\left(\mathrm{A}|i\right) = P_i \frac{\tau_{i,\mathrm{A}}}{\tau_i}, \tag{23}$$

where $\tau_{i,\mathrm{A}} = k_{\mathrm{add}}^{-1}$ is the mean dwell time in the A state at nucleotide i, and τ_i is the total dwell time at nucleotide i. Converting to densities, Eq (22) becomes

$$\frac{\partial \rho_e}{\partial \theta} = -\delta\tau_e \frac{\partial}{\partial x}\left(\frac{\rho_e}{\tau(x)}\right), \tag{24}$$

where $\tau(i\delta) = \tau_i$, and $\tau(x)$ interpolates between these values for $x \neq i\delta$. A similar equation has been obtained by Mier-y-Terán-Romero and coworkers for translation by the ribosome [50]. Following these authors, provided the velocity is sufficiently slowly varying, we obtain a mean elongation delay of, converting time back to dimensional units,

$$\tau_e = N \int_0^1 \tau(x)\, dx. \tag{25}$$

[‡]Because the τ_i are short, Eq (23) should hold true provided observations are not made at the earliest time that some molecules in the ensemble reach nucleotide i. In other words, (23) should hold except in a brief window of duration $O(\tau_i)$ when $P_i(t)$ is beginning to rise.

We see from this equation that $\tau(x)$ has the interpretation of the time required for the polymerase to move the distance of one nucleotide (δ in the x scaling), which is consistent with the original description of $\tau(x)$ as interpolating the τ_i.

We will focus for now on the case in which the kinetic parameters are identical at every nucleotide. Then, $\tau(x) = \tau_i$ is a constant, and

$$\tau_e = (N-1)\tau_i + k_{\text{ae}}^{-1} \approx N\tau_i \tag{26}$$

for large N. Taking into account (20), (24) reduces to

$$\frac{\partial \rho_e}{\partial \theta} = -\frac{\partial \rho_e}{\partial x}, \tag{27}$$

i.e., a simple advective equation with unit velocity. (An alternative derivation of this equation is to be found in [84].) At the left boundary ($x = 0$), probability flows in at a rate $k_{\text{init}} P_{\text{PIC}}(t)$ [reaction (5)] or, in the scaling (17), $\tau_e k_{\text{init}} P_{\text{PIC}}(\tau_e \theta)$. The appropriate boundary condition for (27) is therefore

$$\rho_e(0, \theta) = \tau_e k_{\text{init}} P_{\text{PIC}}(\tau_e \theta). \tag{28}$$

(Technically, since the nucleotides are numbered from 1, the boundary condition should be applied at $x = \delta$ rather than $x = 0$. However, since δ is assumed small, there is no significant error in the boundary condition given above.) Since we deal with just one polymerase, and since this polymerase cannot start transcription before $t = 0$, we also have the initial condition $\rho_e(x, 0) = 0$.

The PDE (27) with the boundary condition (28) simply transports probability through the elongation compartment at unit velocity. Accordingly, the solution is

$$\rho_e(x, \theta) = \tau_e k_{\text{init}} P_{\text{PIC}}\left(\tau_e(\theta - x)\right) H(\theta - x), \tag{29}$$

where $H(\cdot)$ is the Heaviside function. The flux exiting elongation, at $x = 1$, transformed back to the original time scale, is therefore

$$J_e(t) = k_{\text{init}} P_{\text{PIC}}(t - \tau_e) H(t - \tau_e). \tag{30}$$

This solution holds for any fixed kinetic parameters, with or without pausing. We conclude that for transcription, as was the case for translation [50], elongation normally acts as a simple delay. A similar result was obtained elsewhere using a Laplace transform technique [26].

Figure 4 shows the position of the polymerase against time obtained from a stochastic simulation for a single polymerase in the absence of pausing, while Figure 5 shows representative trajectories from similar simulations with frequent short pauses. These simulations confirm that for long genes without long pause sites, the elongation process is essentially advective, with RNAPII moving at a roughly constant velocity. From an ensemble of 40 000 realizations, we found a mean and standard deviation of the elongation rate of $72.0 \pm 1.8\,\text{nt}\,\text{s}^{-1}$. The situation with frequent short pauses is not so different. Over 40 000 realizations, we found that the velocities during the elongation phase varied from 3.7 to $6.5\,\text{nt}\,\text{s}^{-1}$, with a mean of $4.8\,\text{nt}\,\text{s}^{-1}$. While the observed range of elongation rates is relatively large, the variability in the elongation times is tiny compared to the variability in the initiation and termination times, as is evident from Figure 5. Interestingly, the relatively constant time required to get through elongation has been observed experimentally [1]. Bel and coworkers have also provided theoretical arguments demonstrating that the distribution of a process consisting of sequential steps, even if the individual steps have some complexity, would be expected to converge to a narrow distribution when the number of steps is large [103].

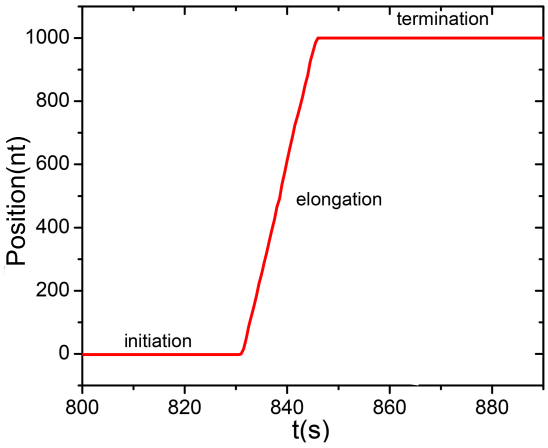

Figure 4. Position of the RNA polymerase vs. time obtained from the stochastic simulation of a single polymerase in the absence of pausing. Parameter values are as in Table 1 except $k_{p,i} = k_{r,i} = 0$ with $N = 1000$.

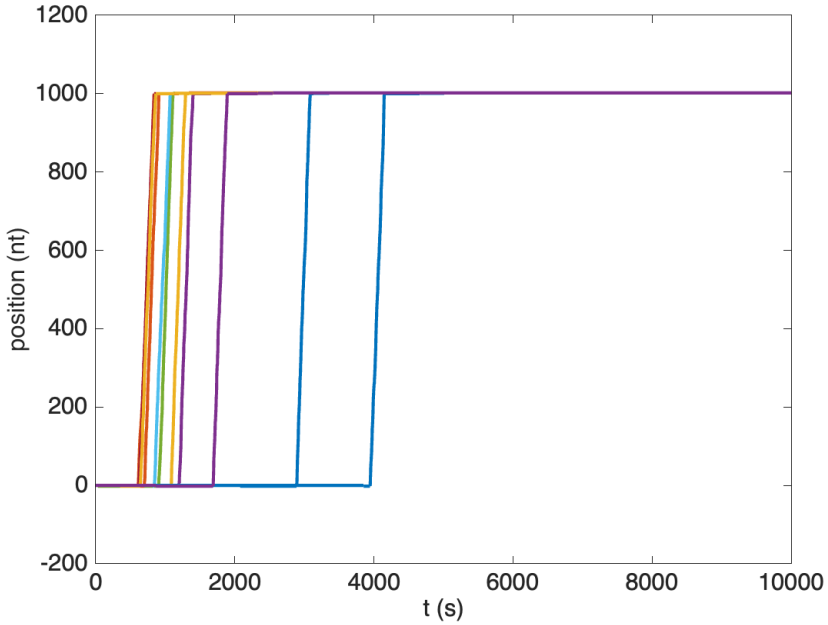

Figure 5. Several realizations of the RNA polymerase position vs. time obtained from stochastic simulations of single polymerases with ubiquitous pausing. The parameter values are as in Table 1 with $N = 1000$.

3.2.2. The elongation delay

Equation (30) shows that, subject to the underlying assumption that elongation is a fast process, elongation only contributes a delay. Mier-y-Terán-Romero and coworkers have shown that the same holds true for slowly varying delays [50]. In this model, which leaves out backtracking that would require hydrolysis of the 3' end of the transcript, addition of nucleotides is strictly sequential. Accordingly,

$$\tau_e = \sum_{i=1}^{N-1} \tau_i + k_{\text{ae}}^{-1}. \tag{31}$$

(Note that reaction (6) includes N steps, but reaction (7) only includes $N-1$ steps, where N is the length of the transcribed sequence.) Similarly, the transition from the pre- to the post-translocated state is irreversible, so

$$\tau_i = \tau_{i,\text{O}\vee\text{P}} + \tau_{i,\text{A}}, \tag{32}$$

where $\tau_{i,\text{O}\vee\text{P}}$ is the mean dwell time in the aggregated pre-translocation O and P states. As noted previously, $\tau_{i,\text{A}} = k_{\text{add}}^{-1}$. We now turn to the calculation of $\tau_{i,\text{O}\vee\text{P}}$.

We focus on the steps

$$\text{P}_i \underset{k_{p,i}}{\overset{k_{r,i}}{\rightleftharpoons}} \text{O}_i \xrightarrow{k_{\text{ae}}} \tag{33}$$

which define the escape process from the O $\vee$ P aggregated state. The master equation for this escape process is

$$\frac{dP_{i,\text{P}}}{dt} = k_{p,i}P_{i,\text{O}} - k_{r,i}P_{i,\text{P}}, \tag{34a}$$

$$\frac{dP_{i,\text{O}}}{dt} = -k_{p,i}P_{i,\text{O}} + k_{r,i}P_{i,\text{P}} - k_{\text{ae}}P_{i,\text{O}}, \tag{34b}$$

with initial conditions $P_{i,\text{O}}(0) = 1$, $P_{i,\text{P}}(0) = 0$. The probability density for the dwell time in the aggregated O and P states at nucleotide i, i.e., for making a transition from state O to state A, is $\rho_{\text{OA},i} = k_{\text{ae}}P_{i,\text{O}}(t)$. After some tedious algebra, we find

$$\rho_{\text{OA},i}(t) = \frac{k_{\text{ae}}}{2\sqrt{D_i}}\Big[e^{-(S_i+\sqrt{D_i})t/2}\left(\sqrt{D_i} + k_{\text{ae}} + k_{p,i} - k_{r,i}\right) + e^{-(S_i-\sqrt{D_i})t/2}\left(\sqrt{D_i} - k_{\text{ae}} - k_{p,i} + k_{r,i}\right)\Big], \tag{35a}$$

where

$$S_i = k_{\text{ae}} + k_{p,i} + k_{r,i}, \tag{35b}$$

$$D_i = S_i^2 - 4k_{\text{ae}}k_{r,i}. \tag{35c}$$

As usual, we compute the mean time spent in the O and P states by

$$\tau_{i,\text{O}\vee\text{P}} = \int_0^\infty t\rho_{\text{OA},i}dt = \frac{k_{p,i} + k_{r,i}}{k_{\text{ae}}\,k_{r,i}}. \tag{36}$$

Equation (32) now implies the dwell time at a particular nucleotide:

$$\tau_i = \frac{k_{p,i} + k_{r,i}}{k_{\text{ae}}k_{r,i}} + \frac{1}{k_{\text{add}}} = \frac{1}{k_{\text{ae}}}\left(1 + \frac{k_{p,i}}{k_{r,i}}\right) + \frac{1}{k_{\text{add}}}. \tag{37}$$

Note that Eqs (35a) and (37) were derived without any approximations. Thus, they apply to any nucleotide, including those where long pauses occur, discussed in Section 4.

It is instructive to consider a few special limits of Eq (37). First, if there is no pausing, $k_{p,i} = 0$, and we get

$$\tau_i|_{k_{p,i}=0} = \frac{1}{k_{\text{ae}}} + \frac{1}{k_{\text{add}}} \tag{38}$$

which is, as expected, the sum of the mean dwell times in the O and A states. On the other hand, for very strong pause sites with $k_{p,i}/k_{r,i} \gg \max(1, k_{\text{ae}}/k_{\text{add}})$, we get

$$\tau_i \approx \frac{k_{p,i}}{k_{r,i}} \frac{1}{k_{\text{ae}}}. \tag{39}$$

In this case, the dwell time in the O and P states dominates, and is computed from the pause-free O-state dwell time stretched by the proportion of the time spent in the paused state.

τ_i is the mean time required to get from nucleotide i to nucleotide $i + 1$. If the kinetic parameters are identical at each nucleotide, then τ_i^{-1} is the mean velocity of the polymerase. For the parameters of Figure 4, we can calculate a velocity $v = 72\,\text{nt}\,\text{s}^{-1}$, which is in perfect agreement with the mean velocity obtained from the numerical experiments mentioned earlier. For the case with pausing (Figure 5), Eq (37) leads to a velocity of $v = 4.8\,\text{nt}\,\text{s}^{-1}$, which again agrees with the mean over many realizations of the stochastic simulations.

It will also be useful to consider the variance in the total elongation time τ_e. According to the central limit theorem [104], assuming N is large, the distribution of τ_e approaches a Gaussian with variance

$$\text{var}(\tau_e) = \sum_{i=1}^{N} \text{var}(\tau_i), \tag{40}$$

making the obvious large-N approximations. The τ_i need not be identically distributed, but there should not be a small number of nucleotides where passage is particularly slow. This case will be considered in Section 4. The variance of τ_i is easily calculated:

$$\text{var}(\tau_i) = \text{var}(\tau_{i,\text{O}\vee\text{P}}) + \text{var}(\tau_{i,\text{A}}) \tag{41}$$

and $\text{var}(\tau_{i,\text{A}}) = k_{\text{add}}^{-2}$. To calculate $\text{var}(\tau_{i,\text{O}\vee\text{P}})$, we first compute $\text{E}(\tau_{i,\text{O}\vee\text{P}}^2)$, where $\text{E}(\cdot)$ denotes the expectation value:

$$\text{E}(\tau_{i,\text{O}\vee\text{P}}^2) = \int_0^\infty t^2 \rho_{\text{OA},i} dt = \frac{2(k_{p,i} + k_{r,i})^2 + 2k_{\text{ae}}k_{p,i}}{k_{\text{ae}}^2 k_{r,i}^2}. \tag{42}$$

Then,

$$\text{var}(\tau_{i,\text{O}\vee\text{P}}) = \text{E}(\tau_{i,\text{O}\vee\text{P}}^2) - \tau_{i,\text{O}\vee\text{P}}^2 = \frac{(k_{p,i} + k_{r,i})^2 + 2k_{\text{ae}}k_{p,i}}{k_{\text{ae}}^2 k_{r,i}^2} \tag{43}$$

so that

$$\text{var}(\tau_i) = \frac{(k_{p,i} + k_{r,i})^2 + 2k_{\text{ae}}k_{p,i}}{k_{\text{ae}}^2 k_{r,i}^2} + \frac{1}{k_{\text{add}}^2}. \tag{44}$$

In the special case where all sites are identical, the distribution of τ_e therefore converges to a Gaussian with mean given by Eqs (31) and (37) and variance

$$\mathrm{var}(\tau_e) = N\,\mathrm{var}(\tau_i) = N\left(\frac{(k_{p,i}+k_{r,i})^2 + 2k_{\mathrm{ae}}k_{p,i}}{k_{\mathrm{ae}}^2 k_{r,i}^2} + \frac{1}{k_{\mathrm{add}}^2}\right). \tag{45}$$

If we let $\sigma_e^2 \equiv \mathrm{var}(\tau_e)$, then the distribution of elongation times can be written

$$\rho_e(t) = \frac{e^{-(t-\tau_e)^2/2\sigma_e^2}}{\sqrt{2\pi\sigma_e^2}}. \tag{46}$$

Several comments about this equation are in order:

1. Provided we compute τ_e from Eq (31) and σ_e^2 from Eq (40), this equation is applicable to any situation where there are no sites where transit is unusually slow, i.e., no long pausing sites. It is not necessary for the distributions of the dwell sites to be identical at each nucleotide.
2. In cases like those of Figures 4 and 5 where the distribution of elongation times is narrow relative to the distribution of at least one of elongation or termination, $\rho_e(t)$ can be replaced by a Dirac δ distribution. Then elongation contributes a simple fixed delay to the overall distribution of transcription times. This is the situation to which Eq (30) applies.
3. Technically, the expression for $\rho_e(t)$ is supported on $t \in (-\infty, \infty)$. Clearly, the negative values of t are meaningless. Thus, Eq (46) will only apply when $\tau_e \gg \sigma_e$ such that $\rho_e(0)$ is extremely small and the $t < 0$ tail of the distribution contributes negligibly to the total probability. In these cases, we can treat $\rho_e(t)$ as if it is supported on the physically meaningful interval $t \in [0, \infty)$ with negligible error. Since σ_e grows as $\sqrt{N}$ and τ_e grows as N, this will always be the case for sufficiently large N.

3.3. Termination

Assume that we are in the normal situation where the flux out of elongation and into termination is given by Eq (30). Thus, the following ODEs govern the evolution during termination:

$$\frac{dP_{N,\mathrm{A}}(t)}{dt} = J_e(t) - k_{\mathrm{TC}}P_{N,\mathrm{A}}(t), \tag{47a}$$

$$\frac{dP_{\mathrm{TC}}(t)}{dt} = k_{\mathrm{TC}}P_{N,\mathrm{A}}(t) - k_{\mathrm{term}}P_{\mathrm{TC}}(t), \tag{47b}$$

$$\frac{dP_{\mathrm{T}}(t)}{dt} = k_{\mathrm{term}}P_{\mathrm{TC}}(t), \tag{47c}$$

where $P_{\mathrm{TC}}(t)$ is the probability that there is a termination complex (TC) at time t, and $P_{\mathrm{T}}(t)$ represents the probability that the transcript (T) has been released, with $J_e(t)$ given by Eq (30).

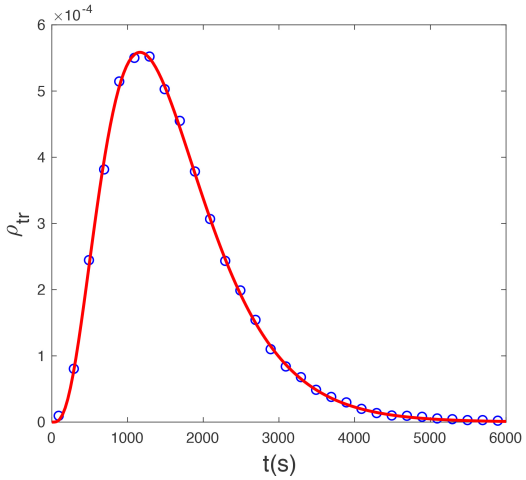

Figure 6. A comparison between stochastic simulation (circles) and analytical (red line) solutions for $\rho_{\text{tr}}(t)$ in the absence of pausing. Parameter values are as in Table 1 except $k_{p,i} = k_{r,i} = 0$ with $N = 1000$. The numerical distribution was computed from 40 000 realizations of the transcription process.

3.4. Distributions of transcription times

We can solve the ODEs (47) with the initial conditions of $P_{N,\text{A}}(0) = P_{\text{TC}}(0) = P_{\text{T}}(0) = 0$ to derive an expression for probability density of the entire transcription process, $\rho_{\text{tr}}(t)$ $(= dP_{\text{T}}/dt = k_{\text{term}}P_{\text{TC}}(t))$:[§]

$$\begin{aligned}\rho_{\text{tr}}(t) = H(t-\tau_e)\, k_{\text{bind}}\, k_{\text{PIC}}\, k_{\text{init}}\, k_{\text{TC}}\, k_{\text{term}} \\ \times\Big\{&\left[(k_{\text{bind}}-k_{\text{PIC}})\,(k_{\text{bind}}-k_{\text{init}})\,(k_{\text{bind}}-k_{\text{TC}})\,(k_{\text{bind}}-k_{\text{term}})\right]^{-1} e^{-k_{\text{bind}}\,(t-\tau_e)} \\ &+\left[(k_{\text{PIC}}-k_{\text{bind}})\,(k_{\text{PIC}}-k_{\text{init}})\,(k_{\text{PIC}}-k_{\text{TC}})\,(k_{\text{PIC}}-k_{\text{term}})\right]^{-1} e^{-k_{\text{PIC}}\,(t-\tau_e)} \\ &+\left[(k_{\text{init}}-k_{\text{bind}})\,(k_{\text{init}}-k_{\text{PIC}})\,(k_{\text{init}}-k_{\text{TC}})\,(k_{\text{init}}-k_{\text{term}})\right]^{-1} e^{-k_{\text{init}}\,(t-\tau_e)} \\ &+\left[(k_{\text{TC}}-k_{\text{bind}})\,(k_{\text{TC}}-k_{\text{PIC}})\,(k_{\text{TC}}-k_{\text{init}})\,(k_{\text{TC}}-k_{\text{term}})\right]^{-1} e^{-k_{\text{TC}}\,(t-\tau_e)} \\ &+\left[(k_{\text{term}}-k_{\text{bind}})\,(k_{\text{term}}-k_{\text{PIC}})\,(k_{\text{term}}-k_{\text{init}})\,(k_{\text{term}}-k_{\text{TC}})\right]^{-1} e^{-k_{\text{term}}\,(t-\tau_e)}\Big\}.\end{aligned} \tag{48}$$

Figures 6 and 7 show comparisons between the analytical solution and stochastic simulations of $\rho_{\text{tr}}(t)$, showing excellent agreement. Note the effect of ubiquitous pausing in Figure 7: the shape of the distribution is not changed, but the distribution is shifted to the right with increasing pausing probability. In other words, ubiquitous pauses only contribute to the constant elongation delay under normal conditions.

An alternative approach to termination, which will shortly prove useful, is to solve directly for the distribution of termination times. We do this by solving the equations

$$\frac{dP_{N,\text{A}}(t)}{dt} = -k_{\text{TC}}P_{N,\text{A}}(t), \tag{49a}$$

$$\frac{dP_{\text{TC}}(t)}{dt} = k_{\text{TC}}P_{N,\text{A}}(t) - k_{\text{term}}P_{\text{TC}}(t), \tag{49b}$$

[§]There is at least one other way to arrive at Eq (48): Below, we will derive the distribution of termination times, Eq (50). We can compute the convolution of the distributions of initiation times, Eq (13), with the distributions of termination times and with a δ distribution representing the distribution of elongation times.

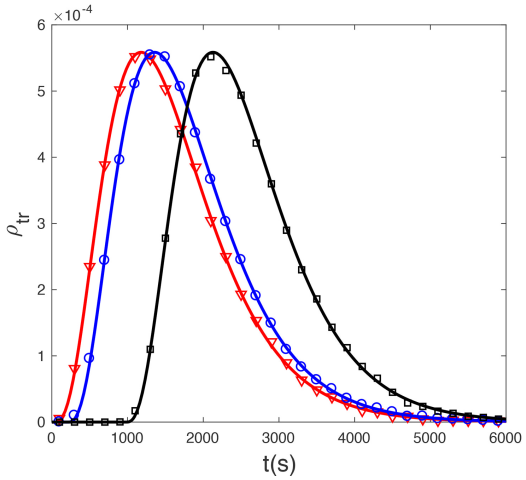

Figure 7. Probability density for transcription time with ubiquitous pausing. We compare stochastic simulation results with analytical results. Parameter values are as in Table 1 with $N = 1000$ except for the $k_{p,i}$ values. The pausing rate constant, $k_{p,i}$, is $7.6\,\mathrm{s}^{-1}$ (red curve with triangles), $100\,\mathrm{s}^{-1}$ (blue curve with circles) and $500\,\mathrm{s}^{-1}$ (black curve with squares). The numerical distribution was computed from 40 000 realizations of the transcription process.

subject to the initial conditions $P_{N,\mathrm{A}}(0) = 1$, $P_{\mathrm{TC}}(0) = 0$. Some elementary mathematics shows the distribution of termination times to be

$$\rho_{\mathrm{T}}(t) = k_{\mathrm{term}} P_{\mathrm{TC}} = \frac{k_{\mathrm{term}}\, k_{\mathrm{TC}}}{k_{\mathrm{TC}} - k_{\mathrm{term}}} \left(e^{-k_{\mathrm{term}} t} - e^{-k_{\mathrm{TC}} t} \right). \tag{50}$$

Equation (48) will provide an accurate approximation to the distribution of transcription times provided elongation is must faster than initiation and termination combined. Otherwise, we must consider the Gaussian distribution of elongation times. There are various ways in which we can imagine globally slowing elongation. Numerical experiments reveal that they all have the same effect since the distribution of elongation times is Gaussian, regardless of the details of the kinetics at any particular nucleotide, again with the proviso that there there are not narrow regions of a few nucleotides where transit is unusually slow. Consider in particular a situation in which the supply of nucleotides is strongly limited. In this model, this corresponds to decreasing the value of k_{ae}. Because of the role of nucleotides in the Brownian ratchet mechanism of RNA polymerases [86, 105], low occupancy of the nucleotide addition site would be expected to cause increased backtracking and pausing as well [89, 106], further reducing the polymerase's average elongation rate.

We can derive an equation for the distribution of transcription times of a slow polymerase by taking a convolution of the initiation, termination and elongation time distributions under the assumption that Eq (46) accurately describes the distribution of elongation times. This is most conveniently done by first taking the convolution of the initiation and termination time distributions, which gives an equation similar to Eq (48) but with $\tau_e = 0$. We then take the convolution of this initiation-termination distribution, ρ_{IT}, with the elongation time distribution as follows:

$$\rho_{\mathrm{tr}}(t) = \int_{\tau=0}^{t} \rho_{\mathrm{IT}}(t-\tau)\rho_e(\tau)\, d\tau. \tag{51}$$

The result of this calculation is

$$\begin{aligned}\rho_{\mathrm{tr}}(t) = \frac{1}{2}k_{\mathrm{bind}}\,k_{\mathrm{PIC}}\,k_{\mathrm{init}}\,k_{\mathrm{TC}}\,k_{\mathrm{term}}\\ \times\Bigg\{&\frac{e^{-k_{\mathrm{bind}}(t-\tau_e-k_{\mathrm{bind}}\sigma_e^2/2)}\left[\mathrm{erf}\left(\frac{k_{\mathrm{bind}}\sigma_e^2+\tau_e}{\sqrt{2\sigma_e^2}}\right)-\mathrm{erf}\left(\frac{k_{\mathrm{bind}}\sigma_e^2-t+\tau_e}{\sqrt{2\sigma_e^2}}\right)\right]}{(k_{\mathrm{bind}}-k_{\mathrm{PIC}})(k_{\mathrm{bind}}-k_{\mathrm{init}})(k_{\mathrm{bind}}-k_{\mathrm{TC}})(k_{\mathrm{bind}}-k_{\mathrm{term}})}\\ &+\frac{e^{-k_{\mathrm{PIC}}(t-\tau_e-k_{\mathrm{PIC}}\sigma_e^2/2)}\left[\mathrm{erf}\left(\frac{k_{\mathrm{PIC}}\sigma_e^2+\tau_e}{\sqrt{2\sigma_e^2}}\right)-\mathrm{erf}\left(\frac{k_{\mathrm{PIC}}\sigma_e^2-t+\tau_e}{\sqrt{2\sigma_e^2}}\right)\right]}{(k_{\mathrm{PIC}}-k_{\mathrm{bind}})(k_{\mathrm{PIC}}-k_{\mathrm{init}})(k_{\mathrm{PIC}}-k_{\mathrm{TC}})(k_{\mathrm{PIC}}-k_{\mathrm{term}})}\\ &+\frac{e^{-k_{\mathrm{init}}(t-\tau_e-k_{\mathrm{init}}\sigma_e^2/2)}\left[\mathrm{erf}\left(\frac{k_{\mathrm{init}}\sigma_e^2+\tau_e}{\sqrt{2\sigma_e^2}}\right)-\mathrm{erf}\left(\frac{k_{\mathrm{init}}\sigma_e^2-t+\tau_e}{\sqrt{2\sigma_e^2}}\right)\right]}{(k_{\mathrm{init}}-k_{\mathrm{bind}})(k_{\mathrm{init}}-k_{\mathrm{PIC}})(k_{\mathrm{init}}-k_{\mathrm{TC}})(k_{\mathrm{init}}-k_{\mathrm{term}})}\\ &+\frac{e^{-k_{\mathrm{TC}}(t-\tau_e-k_{\mathrm{TC}}\sigma_e^2/2)}\left[\mathrm{erf}\left(\frac{k_{\mathrm{TC}}\sigma_e^2+\tau_e}{\sqrt{2\sigma_e^2}}\right)-\mathrm{erf}\left(\frac{k_{\mathrm{TC}}\sigma_e^2-t+\tau_e}{\sqrt{2\sigma_e^2}}\right)\right]}{(k_{\mathrm{TC}}-k_{\mathrm{bind}})(k_{\mathrm{TC}}-k_{\mathrm{PIC}})(k_{\mathrm{TC}}-k_{\mathrm{init}})(k_{\mathrm{TC}}-k_{\mathrm{term}})}\\ &+\frac{e^{-k_{\mathrm{term}}(t-\tau_e-k_{\mathrm{term}}\sigma_e^2/2)}\left[\mathrm{erf}\left(\frac{k_{\mathrm{term}}\sigma_e^2+\tau_e}{\sqrt{2\sigma_e^2}}\right)-\mathrm{erf}\left(\frac{k_{\mathrm{term}}\sigma_e^2-t+\tau_e}{\sqrt{2\sigma_e^2}}\right)\right]}{(k_{\mathrm{term}}-k_{\mathrm{bind}})(k_{\mathrm{term}}-k_{\mathrm{PIC}})(k_{\mathrm{term}}-k_{\mathrm{init}})(k_{\mathrm{term}}-k_{\mathrm{TC}})}\Bigg\}\end{aligned} \tag{52}$$

We now consider a polymerase whose rate constant for the O to A transition (i.e., binding of the appropriate nucleotide) has been reduced by a factor of 100 while the pausing rate constant has been set to $10\,\mathrm{s}^{-1}$. The results of stochastic simulations are plotted along with Eqs (48) and (52) in Figure 8(a). Because, at these high relative rates of pausing, much of the elongation time is spent in pauses, this reduction in k_{ae} and $k_{p,i}$ from the default values has the effect of increasing τ_i by a factor of more than 10, from 0.21 to 2.63 s. Even at this extremely slow rate of translation (about $0.4\,\mathrm{nt\,s}^{-1}$), the advective approximation to elongation still holds and either expression for the distribution of elongation times fits the simulation data equally well. It is only at even lower elongation rates that the more complex Eq (52) is necessary. For example, Figure 8(b) shows the distribution at $k_{p,i} = 200\,\mathrm{s}^{-1}$ with the same value of k_{ae} as in panel (a). At these parameters, τ_i has increased to 39 s, corresponding to a glacial rate of transcription. The distribution of transcription times is now strongly affected by the Gaussian distribution of elongation times. Equation (48) no longer applies, and Eq (52) now gives the correct distribution.

While Figure 8 was produced at parameters that generate unrealistically slow elongation rates, the same issue would arise if both initiation and termination were fast processes, as they will sometimes be. Figure 9 shows the simulated distribution along with the approximations (48) and (52) for such a case. We see again that the advective approximation is a poor fit while Eq (52) provides an excellent approximation to the distribution of transcription times. The reason is exactly the same as it was for the parameters of Figure 8(b): when initiation or termination are slow compared to the duration of the elongation process, then Eq (48) applies; otherwise, we must consider the distribution of elongation times as well as the distributions of initiation and termination times, and Eq (52) must be used, as is the case in Figures 8(b) and 9.

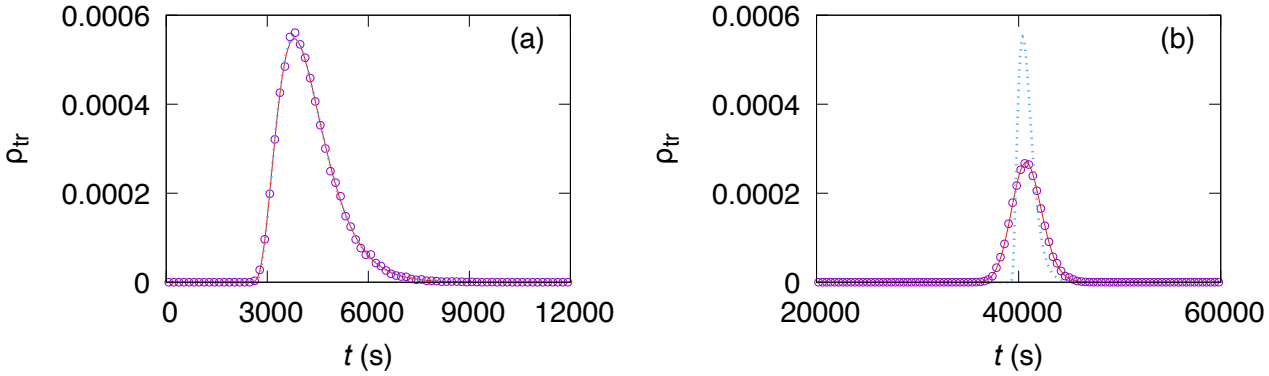

Figure 8. Transcription time distributions for polymerases that are transcribing slowly. In both panels, $k_{ae} = 1.44\,s^{-1}$ and $N = 1000$. In panel (a), $k_{p,i} = 10\,s^{-1}$ while in panel (b), $k_{p,i} = 200\,s^{-1}$. All other parameters are as in Table 1. Circles show simulation results, the dotted blue lines show the transcription time distribution in the advective approximation, Eq (48), while the solid red line is the quasi-exact Eq (52).

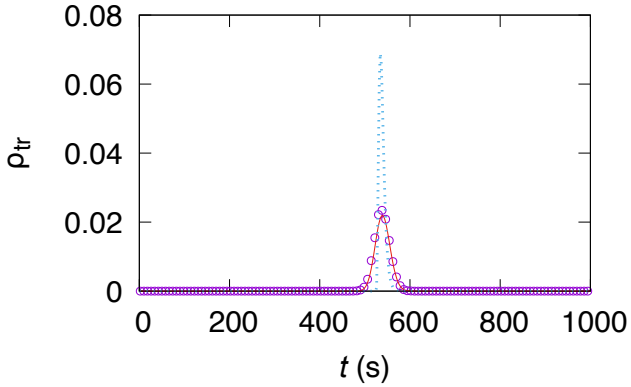

Figure 9. Transcription time distributions for the case where both initiation and termination are fast processes, with parameters $k_{bind} = 1$, $k_{PIC} = 0.2$, $k_{init} = 0.6$, $k_{ae} = 144$, $k_{add} = 144$, $k_{p,i} = 20$, $k_{r,i} = 3.6$, $k_{TC} = 0.3$ and $k_{term} = 0.4\,s^{-1}$ for a 10 000 nt gene. Circles show the distribution computed from 40 000 stochastic simulations, the dotted blue line is Eq (48), and the solid red line is Eq (52).

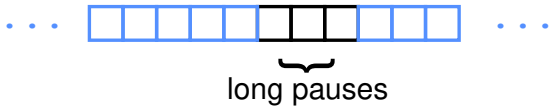

Figure 10. Schematic illustration of a long pause region, here consisting of two nucleotides. Outside the region where long pauses may occur, elongation is treated as contributing a delay. In the region of the long pause, the dwell time distribution is explicitly computed.

We emphasize that the simulation data are always well fit by Eq (52), while Eq (48) is only valid if elongation is fast compared to initiation and termination. Thus, Eq (52) is more general. The only difficulty with using Eq (52) is that its evaluation generally requires higher precision than standard double-precision floating-point arithmetic. We used Maple to evaluate this expression, setting the number of digits of precision to values up to 35 (adjusted upward when anomalies were noted), roughly corresponding to quadruple-precision floating-point numbers.

4. Effect of a long pause along a gene

The case of a site at which a lengthy pause may occur has to be treated separately, for the reasons outlined in Section 3.2.1. We take advantage here of the modular nature [40] of our model. Specifically, we can break up the elongation region into segments where only short ubiquitous pauses occur, dealt with using the theory of Section 3.2, and segments where long pause sites are found, as illustrated in Figure 10. At any given long pause site, the distribution of the dwell time in the combined O and P states is given by Eq (35a). We can take a convolution of this dwell time distribution with the other relevant distributions to obtain the overall transcriptional delay distribution. For simplicity, we will focus in this section on the case where elongation is advective, although it should equally be possible to carry out these calculations for the case of elongation times with a Gaussian distribution.

The convolution of probability densities for initiation and termination times ($\rho_{\mathrm{I,T}}$) is calculated by

$$\rho_{\mathrm{I,T}}(\tau) = \int_0^\tau \rho_{\mathrm{PIC}}(\tau - t)\,\rho_{\mathrm{T}}(t)\,dt. \tag{53}$$

This equation gives us an alternative path to Eq (48), by simply shifting the distribution $\rho_{\mathrm{I,T}}(t)$ by the elongation delay, τ_e.

For a long pause at a single nucleotide at $i = i_p$, the case on which we focus here, we can obtain the convolution of $\rho_{\mathrm{I,T}}(t)$ and $\rho_{OA,i_p}(t)$, Eqs (53) and (35a), respectively:

$$\rho_{\mathrm{I,T,P}}(\tau) = \int_0^\tau \rho_{OA,i_p}(\tau - t)\,\rho_{\mathrm{I,T}}(t)\,dt. \tag{54}$$

In the event that long pauses can occur at multiple nucleotides, this last step is iterated, i.e., we take the q-fold convolution with ρ_{OA,i_p}, where q is the number of nucleotides at which long pauses may occur and i_p represents each of the positions where a long pause occurs, potentially with different kinetic parameters at each of these sites. Once the convolution has been calculated, the final result can be obtained by shifting the distribution by τ_e', the delay due to all processes in elongation not involving the O $\vee$ P aggregated state at the long pause site. The distribution for the transcription time is therefore

$$\rho_{\mathrm{tr}}(t) = H(t - \tau_e')\,\rho_{I,T,P}(t - \tau_e'). \tag{55}$$

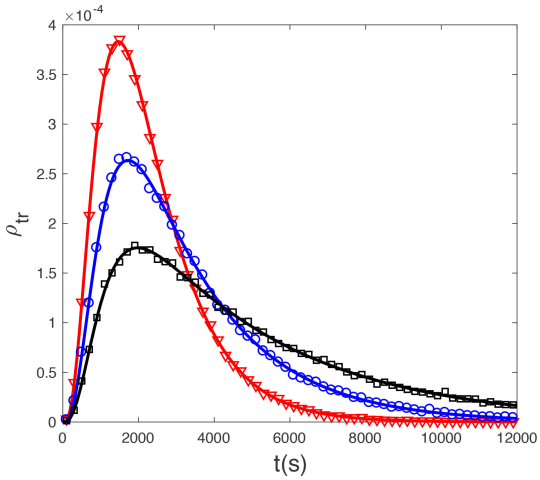

Figure 11. Effect of a strong pause in the gene on the probability density of transcription comparing stochastic simulation results (symbols) and the analytical solution (lines). Parameter values are as in Table 1 with $N = 1000$ except that $k_{p,i} = k_{r,i} = 0$ everywhere except at nucleotide i_p. At this nucleotide, $k_{r,i_p} = 0.002\,\mathrm{s}^{-1}$ with each curve representing a different value of k_{p,i_p}: $k_{p,i_p} = 1000\,\mathrm{s}^{-1}$ for the black line and squares, $500\,\mathrm{s}^{-1}$ for the blue line and circles, and $200\,\mathrm{s}^{-1}$ for the red line and triangles. Each numerical distribution was computed from 10^5 realizations of the transcription process.

The expression that results from this calculation is awkward and is not reproduced here, but is available in a Maple worksheet provided in the Supplemental Materials.

Let $\mathcal{P}$ be the set of nucleotides at which a long pause may occur, and $\mathcal{R}$ be the set of nucleotides where "regular" elongation occurs. $\mathcal{R} \cup \mathcal{P}$ is the entire transcribed sequence, i.e., $\mathcal{R}$ is the complement of $\mathcal{P}$ in the transcribed sequence. The set $\mathcal{P}$ consists of q nucleotides. τ_e' includes $N-1$ extension steps [reaction (7)], and $N-q$ steps with short pauses for which (36) applies. Thus, τ_e' is, in general,

$$\tau_e' = \frac{N-1}{k_{\mathrm{add}}} + \sum_{i\in\mathcal{R}} \frac{k_{p,i} + k_{r,i}}{k_{\mathrm{ae}}\, k_{r,i}}. \tag{56}$$

For the specific case where long pauses only occur at one nucleotide, this reduces to $\tau_e' = (N-1)\tau_i$, with τ_i given by Eq (37).

Figure 11 shows the probability density of the transcription time for three different values of pausing rate constant in a case with a very long pause (mean pause time of 500 s). Pauses of this duration might be associated with splicing, which is quite a slow process [91].

5. Reversible RNAPII binding

It is possible to imagine many variations on the model presented in section 2. Because RNA polymerase binding is likely reversible on time scales relevant to transcription initiation [4], we consider a model in which reaction (4) is replaced by

$$\mathrm{TBP}\cdot\mathrm{pro} + \mathrm{RNAPII} \underset{k_{-\mathrm{PIC}}}{\overset{k_{\mathrm{PIC}}^{(2)}}{\rightleftharpoons}} \mathrm{PIC}. \tag{57}$$

The master equation for initiation becomes

$$\frac{dP_{\text{pro}}}{dt} = -k_{\text{bind}} P_{\text{pro}}, \tag{58a}$$

$$\frac{dP_{\text{TBP·pro}}}{dt} = k_{\text{bind}} P_{\text{pro}} - k_{\text{PIC}} P_{\text{TBP·pro}} + k_{-\text{PIC}} P_{\text{PIC}}, \tag{58b}$$

$$\frac{dP_{\text{PIC}}}{dt} = k_{\text{PIC}} P_{\text{TBP·pro}} - k_{-\text{PIC}} P_{\text{PIC}} - k_{\text{init}} P_{\text{PIC}} \tag{58c}$$

in the same pseudo-first-order regime as was previously assumed to treat RNAPII binding to the promoter. Also as before, we use the initial condition $P_{\text{pro}}(0) = 1$, $P_{\text{TBP·pro}}(0) = P_{\text{PIC}}(0) = 0$. Define

$$s = \sqrt{k_{\text{init}}^2 + 2k_{\text{init}}\,(k_{-\text{PIC}} - k_{\text{PIC}}) + (k_{\text{PIC}} + k_{-\text{PIC}})^2}, \tag{59a}$$

$$\lambda_1 = \frac{1}{2}\,(k_{\text{PIC}} + k_{-\text{PIC}} + k_{\text{init}} + s)\,, \tag{59b}$$

$$\lambda_2 = \frac{1}{2}\,(k_{\text{PIC}} + k_{-\text{PIC}} + k_{\text{init}} - s)\,. \tag{59c}$$

The resulting distribution of initiation times can then be written

$$\rho_{\text{init}}(t) = \frac{k_{\text{bind}}\,k_{\text{PIC}}\,k_{\text{init}}}{s(\lambda_1 - k_{\text{bind}})(\lambda_2 - k_{\text{bind}})}\left[e^{-\lambda_2 t}(k_{\text{bind}} - \lambda_1) - e^{-\lambda_1 t}(k_{\text{bind}} - \lambda_2) + se^{-k_{\text{bind}} t}\right]. \tag{60}$$

It is easy to verify that (60) reduces to (13) when $k_{-\text{PIC}} = 0$.

For the parameters used in our previous figures, the distribution of initiation times is not very sensitive to the value of $k_{-\text{PIC}}$. Figure 12 shows the distribution of initiation times at different values of $k_{-\text{PIC}}$ for a set of parameters inspired by the data of Darzacq and coworkers [4]. The value of k_{init} suggested by their data is much smaller than the one used by Vashishtha [28]. As a result, dissociation of the polymerase and entry into the elongation phase compete directly, which explains why ρ_{init} is much more sensitive to $k_{-\text{PIC}}$ in this regime.

In Section 2.1, we suggested that we could treat binding of RNA polymerase as an irreversible process with an effective rate constant k_{PIC}, i.e., that we could use a reduced value of this rate constant instead of taking reversible binding into account. To test this idea, we calculated the average initiation time in the model with reversible RNAPII binding, and then chose k_{PIC} in the irreversible model to match the average initiation time. We carried out this calculation for the parameters of Figure 12 where the distribution is more sensitive to the value of $k_{-\text{PIC}}$. The result of this calculation is shown in Figure 13. Differences in the shapes of the two distributions can clearly be seen. These differences persist to surprisingly small values of $k_{-\text{PIC}}$, but eventually vanish (results not shown), as they must. Thus, there appears to be a wide range of parameters where reversibility has a noted effect on the distribution of initiation times. Whether these differences are sufficient to be distinguishable based on noisy experimental data is an open question.

6. Conclusions

In this work, a simplified version of the transcription model proposed by Vashishtha [28] for eukaryotic cells was utilized to describe the kinetics of a single RNAPII during transcription. The master

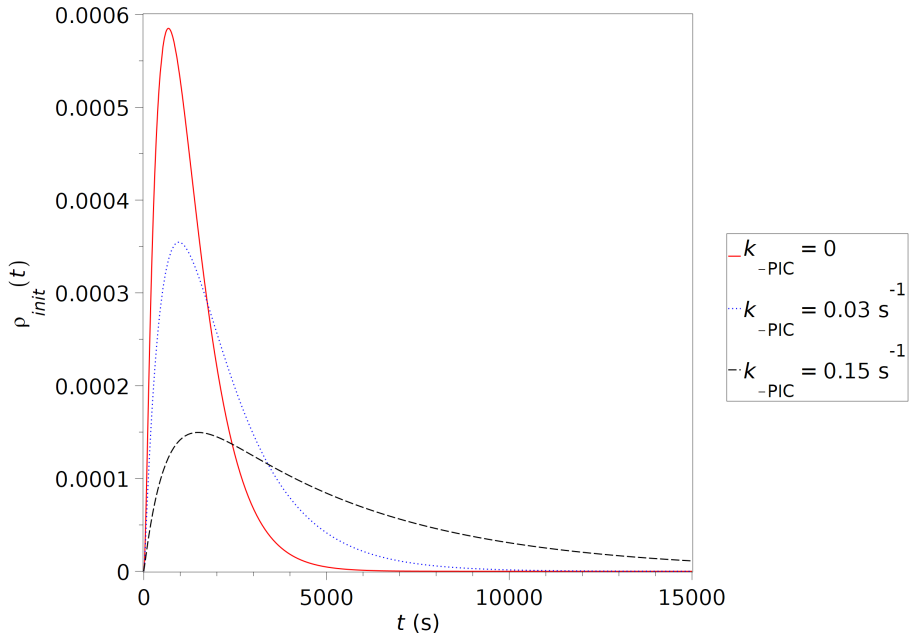

Figure 12. Distribution of initiation times for $k_{\text{bind}} = 1.6 \times 10^{-3}\,\text{s}^{-1}$, $k_{\text{PIC}} = 0.022\,\text{s}^{-1}$, $k_{\text{init}} = 1.59 \times 10^{-3}\,\text{s}^{-1}$, and $k_{-\text{PIC}} = 0$ (red solid line), 0.03 (blue dotted line) or $0.15\,\text{s}^{-1}$ (black dashed line).

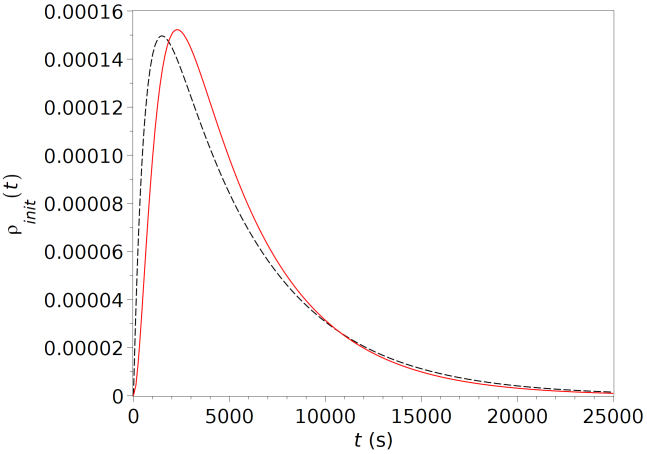

Figure 13. Distribution of initiation times for the model with reversible RNAPII binding (dashed black curve) and for the model with irreversible RNAPII binding (solid red curve). The value of k_{PIC} was adjusted in the latter model to match the average initiation time of the former. For the model with reversible binding (dashed black curve), $k_{\text{PIC}} = 0.022\,\text{s}^{-1}$ and $k_{-\text{PIC}} = 0.15\,\text{s}^{-1}$. For the model with irreversible binding (solid red curve), $k_{\text{PIC}} = 2.31 \times 10^{-4}\,\text{s}^{-1}$. Common parameters: $k_{\text{bind}} = 1.6 \times 10^{-3}\,\text{s}^{-1}$, $k_{\text{init}} = 1.59 \times 10^{-3}\,\text{s}^{-1}$.

equations for initiation and termination were solved exactly. As in the work of Mier-y-Terán-Romero et al. [50], for the case where elongation is fast, the motion during elongation led to an advection equation so that the elongation phase has the effect of a simple delay connecting initiation and termination. There are other ways of showing that simple elongation will often be adequately modeled by a fixed delay [26]. Ultimately, all of these methods depend, directly or indirectly, on the linear growth of both the mean elongation time and the variance with the number of nucleotides such that the width of the distribution becomes relatively narrower as N grows [103]. Contrastingly, in the case where elongation is relatively slow, we found that the distribution of transcription times was poorly reproduced unless we took into account the Gaussian distribution of elongation times. It then becomes an interesting question whether the resulting distribution of transcription times is dynamically important or if, given the relatively narrow and roughly symmetric distribution typically obtained in these cases, a simple fixed delay would yield the same phenomenology in most gene expression models. These questions can be explored using delay-stochastic simulations [70, 107], where it is easy to vary the delay distribution [71] and thus to determine the sensitivity of the dynamics to these distributions. The few studies available so far suggest that the delay distribution is sometimes important whether the dynamics are treated stochastically [108] or deterministically [24, 25], but sometimes not [109].

In addition to delay distributions for homogeneous or near-homogeneous elongation processes, we considered the possibility of long pause sites, which are extreme cases of systems with variable transcription rate. Elsewhere, one of us (SHH) found that variable velocity along the template can lead to some non-trivial phenomenology [84]. Interestingly, elongation in regions without long pauses just contributes a fixed delay according to Eq (56), regardless of where the pause occurs relative to the rapid-elongation regions. This result is reminiscent of Epstein's identification of "bottleneck intermediates" in his treatment of delays arising from consecutive first-order reactions [69].

The delay distributions presented here are likely to be particularly useful for delay-stochastic simulations [70, 71, 110–112]. The delay-stochastic simulation algorithm (DSSA) is an extension of the Gillespie algorithm [113] that allows for delayed product formation. Thus, transcription can, in the simplest case, be written as the single delayed reaction (1). The transcriptional delay may be fixed or distributed. In the latter case, a new value of the delay is generated from the appropriate distribution each time this reaction "fires" [71]. Using the DSSA therefore requires that we know the distribution of transcription times, $\rho_{\mathrm{tr}}(t)$, if only to determine whether or not a fixed delay is appropriate. The distributions derived here can be compared to facilitate these decisions once the parameters of a model have been selected. When Eqs (48) and (52) agree, elongation contributes a fixed delay; otherwise, it will likely be necessary to consider a distribution of elongation delays.

While we have focused on the total transcription time, including the events of initiation, there are a few other ways our results could be used. We could explicitly model the initiation steps, and use a (distributed) delay to represent only elongation and termination. This would be the simplest approach to modeling the expression of a gene whose transcription is regulated at initiation, and again, modularity plays a role in allowing us to use parts of the calculations presented here in more complex situations. For a gene regulated at a pause site [114], elongation could be broken up into pieces, very much as we did to obtain the distributions shown in Figure 11, treating regulation at the pause site explicitly. For example, if a release factor (RF) such as P-TEFb [72] is required to release a paused

complex (PC) from an obligatory pause, we could write, neglecting promoter clearance,[¶]

$$\mathrm{pro}(t) + \mathrm{RNAPII}(t) \xrightarrow{k_{\mathrm{ini}}} \mathrm{pro}(t) + \mathrm{PC}(t + \tau_p), \tag{61a}$$

$$\mathrm{PC}(t) + \mathrm{RF}(t) \xrightarrow{k_{\mathrm{rel}}} \mathrm{RNAPII}(t + \tau_f) + \mathrm{RNA}(t + \tau_f) + \mathrm{RF}(t + \tau_{\mathrm{RFC}}) \tag{61b}$$

where τ_p is the time required to reach the pause site, τ_f is the time required to finish transcription after release from the pause and τ_{RFC} is the lifetime of the polymerase–release-factor complex. If there are no other sites in the gene where long pauses occur and if the sequences preceding and following the pause site are sufficiently long, τ_p and τ_f could be treated as fixed delays as seen in Section 3.2. The interesting dynamics would then lie in the production and/or destruction of RF, which would itself be described by appropriate reactions, including possibly delayed reactions representing its transcription and translation.

In the case of a gene expressed in bursts, the results obtained here could be used to describe the transcription time distribution of the first polymerase in a burst, on the assumption that interactions between polymerases do not affect the behavior of the pioneer polymerase.[‖] We could write a delayed mass-action reaction for a burst as follows:

$$\mathrm{pro}(t) \xrightarrow{k_{\mathrm{ini}}} \mathrm{pro}_i(t + \tau_b) + \mathrm{RNA}(t + \tau_{\mathrm{tr}}) + \sum_{i=2}^{b} \mathrm{RNA}(t + \tau_i), \tag{62}$$

where pro_i is an inactive form of the promoter. This reaction assumes that the nuclear concentration of RNAPII is roughly constant in time so that this concentration can be included in the pseudo-first-order rate constant k_{ini}. The delay τ_b is the duration of a burst generating b RNA molecules. This delay is required to ensure that a burst does not start at a particular copy of a gene before the previous burst has ended. The delay τ_{tr} is the transcription time of the pioneer polymerase which, as we argued above, could be randomly generated according to a distribution from a single-polymerase model. The τ_i are the completion times of the subsequent transcripts measured from the beginning of the burst. These completion times could be computed sequentially as follows:

$$\tau_i = \tau_{\mathrm{tr}} + \sum_{j=1}^{i-1} \Delta\tau_i, \tag{63}$$

where $\Delta\tau_i = \tau_{i+1} - \tau_i$ (identifying $\tau_1 \equiv \tau_{\mathrm{tr}}$) is the time between completions of subsequent transcripts. Although we have not tried this, the distribution of this quantity should be derivable from our model in the low-traffic limit by considering the time required for promoter clearance, initiation of a subsequent transcript (possibly considering retention of TBP for a time after the pioneer polymerase has initiated transcription) and differences in termination times between consecutive polymerases. Alternatively,

[¶]If we wanted to deal with promoter clearance, we would replace pro(t) on the right-hand side of Eq (61a) by pro($t+\tau_c$), where τ_c is the time required for RNAPII to clear the promoter and therefore make the latter available for binding another molecule of polymerase [115].

[‖]It is likely that polymerases do affect each other's motions [33, 54, 116]. However, the nature of these interactions is, to our knowledge, not well characterized at this time, which makes them difficult to incorporate into models. Treating the first polymerase in a burst as traveling freely along the gene may be the best we can do right now, give or take the possibility of considering the prevention of backtracking by trailing polymerases. Moreover, the approximation that the pioneer polymerase travels freely is likely to run into trouble in the "Kurtz" limit [117] where frequent bursts mean that the leading polymerase in a burst may well interact with the trailing polymerase of the previous burst. Thus, Eq (62) is likely more appropriate in Jia et al.'s "Lévy" limit where bursts are well separated in time.

Szavits-Nossan and Grima have recently obtained some results on the distribution of $\Delta\tau_i$ [118]. Both the distributions of burst duration τ_b and of burst size b could be generated from a model [119, 120], or experimental distributions could be used [32, 121–123].** The model would be completed by including one or more reactions that return the promoter to the active state. For telegrapher's noise, this would be the simple Poisson process $\text{pro}_i \to \text{pro}$. More complicated reactivation mechanisms can of course be contemplated [124]. Given the distributions described above, genes displaying bursting could be studied using delay-stochastic simulations, with the single-polymerase transcription time distribution as a key element.

As we have emphasized, the construction of the delay-stochastic models described above is facilitated by the modularity of the transcription model studied here. As noted by Greive and coworkers for a related model [40], modularity means that we can mix and match different descriptions of initiation, elongation, pausing and termination to suit any given gene. The mathematical analysis of these models is similarly modular: we can take the pieces derived here and combine them as needed by convolution.

As explained in Section 2, we have substantially simplified Vashishtha's model. Studying the effects of various omitted details on transcription dynamics is of course a potential avenue for extending this work. The experimental observation of, as Larson and coworkers put it, "deterministic" kinetics during the elongation phase of transcription [1] suggests that a more refined model of productive elongation would not behave differently than the simple model studied here. This is a consequence of the law of large numbers (LLN) [125] and of the large sizes of many eukaryotic genes: If the kinetics are identical at every nucleotide, then $\tau_e = \sum_i \tau_i = N\bar{\tau}_i$, where $\bar{\tau}_i$ is the sample mean of the residence times at each nucleotide. The LLN says that $\bar{\tau}_i$ converges to the mean of the underlying distribution. Thus, τ_e converges to a fixed value for long genes. If the residence times are not identically distributed, then there are generalizations of the law of large numbers leading to the same conclusion provided the τ_i satisfy some additional (physically reasonable) conditions [126]: τ_e converges to a constant for sufficiently large N. Pausing is a different matter, as Voliotis and coworkers have shown [56]: backtracked pausing can contribute a heavy tail to the distribution of transcription times. To model genes that are prone to backtracking pauses in specific regions of the sequence, it would be possible to combine the Voliotis *et al.* pausing module with our model to obtain the corresponding distribution of elongation times.

Shortly after initiation, RNAPII tends to pause 20–40 nt downstream of the start site in an event known as promoter-proximal pausing [127]. A promoter-proximal pause that is not subject to specific regulation is relatively easy to model since it can be dealt with using the splitting trick illustrated in Figure 10. The modularity of the model in this case means that the exact location of the pause site is irrelevant. The situation is different if the polymerase is released from its promoter-proximal pause by a specific release factor with its own dynamics. Consider again model (61) assuming that the concentration of the release factor cannot be treated as being constant in time. This will be the case if (e.g.) a signaling cascade modulates the activity of the release factor. In a dynamic model of such a system, we would have to explicitly include reaction (61b) so that the time dependence of the release from the pause could be properly captured. Thus, we would need to know the distribution of τ_p since this interval of time necessarily precedes the action of the release factor. In particular, note that the concentration of the release factor could change during the time interval τ_p. However,

**Note however that the value of τ_b is implied by the values of the random variables b and $\Delta\tau_i$, so experimental distributions of the burst time could be used to validate the models for the distributions of b and/or $\Delta\tau_i$.

we probably cannot treat the initial elongation leading to the pause using the theory of Section 3.2 because only a short run of nucleotides precedes promoter-proximal pausing, thus violating the central assumption in the development of the advection equation, namely that N is large. We could solve the first-passage time for reaching the pause site, but even for the moderate number of nucleotides separating the start and promoter-proximal pause sites, this is cumbersome unless the kinetics are particularly simple. An additional difficulty is that the promoter-proximal pause probably occurs in a region covering several nucleotides, and not at a specific nucleotide in any given gene. Fortunately, compared to typical initiation times (Figure 3), the contribution of the elongation time to τ_p may well be negligible despite the slower elongation rate in this region of a gene [96]. Even if the time required to reach the promoter-proximal pausing site is not negligible, it is likely possible to develop useful approximations to the distribution of τ_p.

Termination is another area ripe for investigation in terms of its effect on transcription dynamics as it is quite slow in eukaryotes [4]. Indeed, we have initiated such studies in our laboratory [128]. It is also interesting in that multiple independent mechanisms cooperate to ensure termination [7], introducing an additional element of stochasticity right at the end of the process of transcription.

Even within the context of the model studied here, we have not exhausted the possibilities for further study. In particular, while we have developed the relevant equation of motion, we have not studied in much detail the effects of variable kinetic parameters on the density of polymerases. At higher initiation rates, this could have an effect on polymerase-polymerase interactions. Some preliminary calculations on a system in which the polymerase slows down dramatically in one region of the gene have been carried out [84], but this hardly scratches the surface of what could be done.

The delays that appear in models of gene expression are measurable quantities. There has been recent progress towards inferring these delays from experimental data [129, 130], with some caveats: depending on what is measured, not all kinetic parameters, indeed not all delays, are identifiable. This creates an opportunity for a two-way discussion with experiment: Clearly, these experimental measurements can provide data to be used in modeling. In addition, the current contribution and similar efforts in other groups can provide parameterizable forms for the delay distributions that can be used to fit experimental results, and thereby to obtain estimates of some kinetic parameters that determine these distributions.

As noted in the introduction, the eventual goal of this research is to provide delay distributions that can be used in models of gene expression. In addition to the distribution of transcription times which is the topic of this contribution, we will also need distributions for the nuclear export time, on which some progress has been made [131–133] but which remains very much an open problem, and for translation [48–50]. We can look forward to soon having all the pieces necessary either for the direct inclusion in a model of these three major steps in eukaryotic gene expression, or at the very least to enable rational decisions about whether they need to be explicitly included in a given model.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. D. R. Larson, D. Zenklusen, B. Wu, J. A. Chao, R. H. Singer, Real-time observation of transcription initiation and elongation on an endogeneous yeast gene, *Science*, **332** (2011), 475–478. https://doi.org/10.1126/science.1202142
2. S. Buratowski, S. Hahn, L. Guarente, P. A. Sharp, Five intermediate complexes in transcription initiation by RNA polymerase II, *Cell*, **56** (1989), 549–561. https://doi.org/10.1016/0092-8674(89)90578-3
3. A. Dvir, J. W. Conaway, R. C. Conaway, Mechanism of transcription initiation and promoter escape by RNA polymerase II, *Curr. Opin. Genet. Dev.*, **11** (2001), 209–214. https://doi.org/10.1016/S0959-437X(00)00181-7
4. X. Darzacq, Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, R. H. Singer, *In vivo* dynamics of RNA polymerase II transcription, *Nat. Struct. Mol. Biol.*, **14** (2007), 796–806. https://doi.org/10.1038/nsmb1280
5. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, et al., The sequence of the human genome, *Science*, **291** (2001), 1304–1351. https://doi.org/10.1126/science.1058040
6. C. N. Tennyson, H. J. Klamut, R. G. Worton, The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced, *Nat. Genet.*, **9** (1995), 184–190. https://doi.org/10.1038/ng0295-184
7. J.-F. Lemay, F. Bachand, Fail-safe transcription termination: Because one is never enough, *RNA Biol.*, **12** (2015), 927–932. https://doi.org/10.1080/15476286.2015.1073433
8. R. Ben-Yishay, Y. Shav-Tal, The dynamic lifecycle of mRNA in the nucleus, *Curr. Opin. Cell Biol.*, **58** (2019), 69–75. https://doi.org/10.1016/j.ceb.2019.02.007
9. B. Daneholt, Assembly and transport of a premessenger RNP particle, *Proc. Natl. Acad. Sci. U.S.A.*, **98** (2001), 7012–7017. https://doi.org/10.1073/pnas.111145498
10. J. Sheinberger, Y. Shav-Tal, The dynamic pathway of nuclear RNA in eukaryotes, *Nucleus*, **4** (2013), 195–205. https://doi.org/10.4161/nucl.24434
11. A. Chaudhuri, S. Das, B. Das, Localization elements and zip codes in the intracellular transport and localization of messenger RNAs in *Saccharomyces cerevisiae*, *WIREs RNA*, **11** (2020), e1591. https://doi.org/10.1002/wrna.1591
12. U. Schmidt, E. Basyuk, M.-C. Robert, M. Yoshida, J.-P. Villemin, D. Auboeuf, et al., Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: Implications for alternative splicing regulation, *J. Cell Biol.*, **193** (2011), 819–829. https://doi.org/10.1083/jcb.201009012

13. P. Cramer, A. Srebrow, S. Kadener, S. Werbajh, M. de la Mata, G. Melen, et al., Coordination between transcription and pre-mRNA processing, *FEBS Lett.*, **498** (2001), 179–182. https://doi.org/10.1016/S0014-5793(01)02485-1

14. A. Babour, C. Dargemont, F. Stutz, Ubiquitin and assembly of export competent mRNP, *Biochim. Biophys. Acta*, **1819** (2012), 521–530. https://doi.org/10.1016/j.bbagrm.2011.12.006

15. R. A. Coleman, B. F. Pugh, Slow dimer dissociation of the TATA binding protein dictates the kinetics of DNA binding, *Proc. Natl. Acad. Sci. U.S.A.*, **94** (1997), 7221–7226. https://doi.org/10.1073/pnas.94.14.7221

16. J. F. Kugel, J. A. Goodrich, A kinetic model for the early steps of RNA synthesis by human RNA polymerase II, *J. Biol. Chem.*, **275** (2000), 40483–40491. https://doi.org/10.1074/jbc.M006401200

17. A. Kalo, I. Kanter, A. Shraga, J. Sheinberger, H. Tzemach, N. Kinor, et al., Cellular levels of signaling factors are sensed by β-actin alleles to modulate transcriptional pulse intensity, *Cell Rep.*, **11** (2015), 419–432. https://doi.org/10.1016/j.celrep.2015.03.039

18. R. D. Bliss, P. R. Painter, A. G. Marr, Role of feedback inhibition in stabilizing the classical operon, *J. Theor. Biol.*, **97** (1982), 177–193. https://doi.org/10.1016/0022-5193(82)90098-4

19. F. Buchholtz, F. W. Schneider, Computer simulation of T3/T7 phage infection using lag times, *Biophys. Chem.*, **26** (1987), 171–179. https://doi.org/10.1016/0301-4622(87)80020-0

20. S. N. Busenberg, J. M. Mahaffy, The effects of dimension and size for a compartmental model of repression, *SIAM J. Appl. Math.*, **48** (1988), 882–903. https://doi.org/10.1137/0148049

21. J. Lewis, Autoinhibition with transcriptional delay: A simple mechanism for the zebrafish somitogenesis oscillator, *Curr. Biol.*, **13** (2003), 1398–1408. https://doi.org/10.1016/S0960-9822(03)00534-7

22. N. A. M. Monk, Oscillatory expression of Hes1, p53, and NF-κB driven by transcriptional time delays, *Curr. Biol.*, **13** (2003), 1409–1413. https://doi.org/10.1016/S0960-9822(03)00494-9

23. L.-J. Chiu, M.-Y. Ling, E.-H. Wu, C.-X. You, S.-T. Lin, C.-C. Shu, The distributed delay rearranges the bimodal distribution at protein level, *J. Taiwan Inst. Chem. Eng.*, **137** (2022), 104436. https://doi.org/10.1016/j.jtice.2022.104436

24. M. Jansen, P. Pfaffelhuber, Stochastic gene expression with delay, *J. Theor. Biol.*, **364** (2015), 355–363. https://doi.org/10.1016/j.jtbi.2014.09.031

25. K. Rateitschak, O. Wolkenhauer, Intracellular delay limits cyclic changes in gene expression, *Math. Biosci.*, **205** (2007), 163–179. https://doi.org/10.1016/j.mbs.2006.08.010

26. M. R. Roussel, On the distribution of transcription times, *BIOMATH*, **2** (2013), 1307247. https://doi.org/10.11145/j.biomath.2013.07.247

27. M. R. Roussel, R. Zhu, Stochastic kinetics description of a simple transcription model, *Bull. Math. Biol.*, **68** (2006), 1681–1713. https://doi.org/10.1007/s11538-005-9048-6

28. S. Vashishtha, *Stochastic modeling of eukaryotic transcription at the single nucleotide level*, M.Sc. thesis, University of Lethbridge, 2011, URL `https://www.uleth.ca/dspace/handle/10133/3190`.

29. V. Pelechano, S. Chávez, J. E. Pérez-Ortín, A complete set of nascent transcription rates for yeast genes, *PLoS One*, **5** (2010), e15442. https://doi.org/10.1371/journal.pone.0015442

30. T. Muramoto, D. Cannon, M. Gierliński, A. Corrigan, G. J. Barton, J. R. Chubb, Live imaging of nascent RNA dynamics reveals distinct types of transcriptional pulse regulation, *Proc. Natl. Acad. Sci. U.S.A.*, **109** (2012), 7350–7355. https://doi.org/10.1073/pnas.1117603109
31. A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, S. Tyagi, Stochastic mRNA synthesis in mammalian cells, *PLoS Biol.*, **4** (2006), e309. https://doi.org/10.1371/journal.pbio.0040309
32. D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, F. Naef, Mammalian genes are transcribed with widely different bursting kinetics, *Science*, **332** (2011), 472–474. https://doi.org/10.1126/science.1198817
33. I. Jonkers, H. Kwak, J. T. Lis, Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons, *eLife*, **3** (2014), e02407. https://doi.org/10.7554/eLife.02407
34. P. K. Parua, G. T. Booth, M. Sansó, B. Benjamin, J. C. Tanny, J. T. Lis, R. P. Fisher, A Cdk9-PP1 switch regulates the elongation-termination transition of RNA polymerase II, *Nature*, **558** (2018), 460–464. https://doi.org/10.1038/s41586-018-0214-z
35. L. Bai, R. M. Fulbright, M. D. Wang, Mechanochemical kinetics of transcription elongation, *Phys. Rev. Lett.*, **98** (2007), 068103. https://doi.org/10.1103/PhysRevLett.98.068103
36. L. Bai, A. Shundrovsky, M. D. Wang, Sequence-dependent kinetic model for transcription elongation by RNA polymerase, *J. Mol. Biol.*, **344** (2004), 335–349. https://doi.org/10.1016/j.jmb.2004.08.107
37. F. Jülicher, R. Bruinsma, Motion of RNA polymerase along DNA: a stochastic model, *Biophys. J.*, **74** (1998), 1169–1185. https://doi.org/10.1016/S0006-3495(98)77833-6
38. H.-Y. Wang, T. Elston, A. Mogilner, G. Oster, Force generation in RNA polymerase, *Biophys. J.*, **74** (1998), 1186–1202. https://doi.org/10.1016/S0006-3495(98)77834-8
39. T. D. Yager, P. H. Von Hippel, A thermodynamic analysis of RNA transcript elongation and termination in *Escherichia coli*, *Biochemistry*, **30** (1991), 1097–1118. https://doi.org/10.1021/bi00218a032
40. S. J. Greive, J. P. Goodarzi, S. E. Weitzel, P. H. von Hippel, Development of a "modular" scheme to describe the kinetics of transcript elongation by RNA polymerase, *Biophys. J.*, **101** (2011), 1155–1165. https://doi.org/10.1016/j.bpj.2011.07.042
41. T. Filatova, N. Popovic, R. Grima, Statistics of nascent and mature RNA fluctuations in a stochastic model of transcriptional initiation, elongation, pausing, and termination, *Bull. Math. Biol.*, **83** (2021), 3. https://doi.org/10.1007/s11538-020-00827-7
42. A. N. Boettiger, P. L. Ralph, S. N. Evans, Transcriptional regulation: Effects of promoter proximal pausing on speed, synchrony and reliability, *PLoS Comput. Biol.*, **7** (2011), e1001136. https://doi.org/10.1371/journal.pcbi.1001136
43. X. Xu, N. Kumar, A. Krishnan, R. V. Kulkarni, Stochastic modeling of dwell-time distributions during transcriptional pausing and initiation, in *52nd IEEE Conference on Decision and Control*, 2013, 4068–4073.
44. M. Hamano, Stochastic transcription elongation via rule based modelling, *Electron. Notes Theor. Comput. Sci.*, **326** (2016), 73–88. https://doi.org/10.1016/j.entcs.2016.09.019
45. S. Klumpp, T. Hwa, Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination, *Proc. Natl. Acad. Sci. U.S.A.*, **105** (2008), 18159–18164. https://doi.org/10.1073/pnas.0806084105

46. A. S. Ribeiro, O.-P. Smolander, T. Rajala, A. Häkkinen, O. Yli-Harja, Delayed stochastic model of transcription at the single nucleotide level, *J. Comput. Biol.*, **16** (2009), 539–553. https://doi.org/10.1089/cmb.2008.0153
47. M. J. Schilstra, C. L. Nehaniv, Stochastic model of template-directed elongation processes in biology, *BioSystems*, **102** (2010), 55–60. https://doi.org/10.1016/j.biosystems.2010.07.006
48. A. Garai, D. Chowdhury, D. Chowdhury, T. V. Ramakrishnan, Stochastic kinetics of ribosomes: Single motor properties and collective behavior, *Phys. Rev. E*, **80** (2009), 011908. https://doi.org/10.1103/PhysRevE.80.011908
49. A. Garai, D. Chowdhury, T. V. Ramakrishnan, Fluctuations in protein synthesis from a single RNA template: Stochastic kinetics of ribosomes, *Phys. Rev. E*, **79** (2009), 011916. https://doi.org/10.1103/PhysRevE.79.011916
50. L. Mier-y-Terán-Romero, M. Silber, V. Hatzimanikatis, The origins of time-delay in template biopolymerization processes, *PLoS Comput. Biol.*, **6** (2010), e1000726. https://doi.org/10.1371/journal.pcbi.1000726
51. L. S. Churchman, J. S. Weissman, Nascent transcript sequencing visualizes transcription at nucleotide resolution, *Nature*, **469** (2011), 368–373. https://doi.org/10.1038/nature09652
52. K. C. Neuman, E. A. Abbondanzieri, R. Landick, J. Gelles, S. M. Block, Ubiquitous transcriptional pausing is independent of RNA polymerase backtracking, *Cell*, **115** (2003), 437 – 447. https://doi.org/10.1016/S0092-8674(03)00845-6
53. R. Landick, The regulatory roles and mechanism of transcriptional pausing, *Biochem. Soc. Trans.*, **34** (2006), 1062–1066. https://doi.org/10.1042/BST0341062
54. V. Epshtein, F. Toulmé, A. R. Rahmouni, S. Borukhov, E. Nudler, Transcription through the roadblocks: the role of RNA polymerase cooperation, *EMBO J.*, **22** (2003), 4719–4727. https://doi.org/10.1093/emboj/cdg452
55. S. Klumpp, Pausing and backtracking in transcription under dense traffic conditions, *J. Stat. Phys.*, **142** (2011), 1252–1267. https://doi.org/10.1007/s10955-011-0120-3
56. M. Voliotis, N. Cohen, C. Molina-París, T. B. Liverpool, Fluctuations, pauses, and backtracking in DNA transcription, *Biophys. J.*, **94** (2008), 334–348. https://doi.org/10.1529/biophysj.107.105767
57. J. Li, D. S. Gilmour, Promoter proximal pausing and the control of gene expression, *Curr. Opin. Genet. Dev.*, **21** (2011), 231–235. https://doi.org/10.1016/j.gde.2011.01.010
58. S. Nechaev, K. Adelman, Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation, *Biochim. Biophys. Acta*, **1809** (2011), 34–45. https://doi.org/10.1016/j.bbagrm.2010.11.001
59. P. B. Rahl, C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine, C. B. Burge, et al., c-Myc regulates transcriptional pause release, *Cell*, **141** (2010), 432–445. https://doi.org/10.1016/j.cell.2010.03.030
60. P. Feng, A. Xiao, M. Fang, F. Wan, S. Li, P. Lang, et al., A machine learning-based framework for modeling transcription elongation, *Proc. Natl. Acad. Sci. U.S.A.*, **118** (2021), e2007450118. https://doi.org/10.1073/pnas.2007450118
61. B. Zamft, L. Bintu, T. Ishibashi, C. Bustamante, Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases, *Proc. Natl. Acad. Sci. U.S.A.*, **109** (2012), 8948–8953. https://doi.org/10.1073/pnas.1205063109

62. R. D. Alexander, S. A. Innocente, J. D. Barrass, J. D. Beggs, Splicing-dependent RNA polymerase pausing in yeast, *Mol. Cell*, **40** (2010), 582–593. https://doi.org/10.1016/j.molcel.2010.11.005

63. N. Gromak, S. West, N. J. Proudfoot, Pause sites promote transcriptional termination of mammalian RNA polymerase II, *Mol. Cell. Biol.*, **26** (2006), 3986–3996. https://doi.org/10.1128/MCB.26.10.3986-3996.2006

64. N. MacDonald, Time delay in prey-predator models, *Math. Biosci.*, **28** (1976), 321–330. https://doi.org/10.1016/0025-5564(76)90130-9

65. N. MacDonald, Time lag in a model of a biochemical reaction sequence with end product inhibition, *J. Theor. Biol.*, **67** (1977), 549–556. https://doi.org/10.1016/0022-5193(77)90056-X

66. N. MacDonald, *Biological Delay Systems: Linear Stability Theory*, Cambridge, Cambridge, 1989.

67. M. Barrio, A. Leier, T. T. Marquez-Lago, Reduction of chemical reaction networks through delay distributions, *J. Chem. Phys.*, **138** (2013), 104114. https://doi.org/10.1063/1.4793982

68. A. Leier, M. Barrio, T. T. Marquez-Lago, Exact model reduction with delays: Closed-form distributions and extensions to fully bi-directional monomolecular reactions, *J. R. Soc. Interface*, **11** (2014), 20140108. https://doi.org/10.1098/rsif.2014.0108

69. I. R. Epstein, Differential delay equations in chemical kinetics: Some simple linear model systems, *J. Chem. Phys.*, **92** (1990), 1702–1712. https://doi.org/10.1063/1.458052

70. D. Bratsun, D. Volfson, L. S. Tsimring, J. Hasty, Delay-induced stochastic oscillations in gene regulation, *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005), 14593–14598. https://doi.org/10.1073/pnas.0503858102

71. M. R. Roussel, R. Zhu, Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression, *Phys. Biol.*, **3** (2006), 274. https://doi.org/10.1088/1478-3975/3/4/005

72. B. H. Jennings, Pausing for thought: Disrupting the early transcription elongation checkpoint leads to developmental defects and tumourigenesis, *BioEssays*, **35** (2013), 553–560. https://doi.org/10.1002/bies.201200179

73. H. Kwak, N. J. Fuda, L. J. Core, J. T. Lis, Precise maps of RNA polymerase reveal how promoters direct initiation and pausing, *Science*, **339** (2013), 950–953. https://doi.org/10.1126/science.1229386

74. A. R. Hieb, S. Baran, J. A. Goodrich, J. F. Kugel, An 8nt RNA triggers a rate-limiting shift of RNA polymerase II complexes into elongation, *EMBO J.*, **25** (2006), 3100–3109. https://doi.org/10.1038/sj.emboj.7601197

75. T. J. Stasevich, Y. Hayashi-Takanaka, Y. Sato, K. Maehara, Y. Ohkawa, K. Sakata-Sogawa, et al., Regulation of RNA polymerase II activation by histone acetylation in single living cells, *Nature*, **516** (2014), 272–275. https://doi.org/10.1038/nature13714

76. B. Steurer, R. C. Janssens, B. Geverts, M. E. Geijer, F. Wienholz, A. F. Theil, et al., Live-cell analysis of endogeneous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA polymerase II, *Proc. Natl. Acad. Sci. U.S.A.*, **115** (2018), E4368–E4376. https://doi.org/10.1073/pnas.1717920115

77. J. Liu, D. Hansen, E. Eck, Y. J. Kim, M. Turner, S. Alamos, H. G. Garcia, Real-time single-cell characterization of the eukaryotic transcription cycle reveals correlations between RNA initiation, elongation, and cleavage, *PLoS Comput. Biol.*, **17** (2021), e1008999. https://doi.org/10.1371/journal.pcbi.1008999

78. A. Kremling, Comment on mathematical models which describe transcription and calculate the relationship between mRNA and protein expression ratio, *Biotech. Bioeng.*, **96** (2007), 815–819. https://doi.org/10.1002/bit.21065

79. N. Mitarai, S. Pedersen, Control of ribosome traffic by position-dependent choice of synonymous codons, *Phys. Biol.*, **10** (2013), 056011. https://doi.org/10.1088/1478-3975/10/5/056011

80. M. R. Roussel, The use of delay differential equations in chemical kinetics, *J. Phys. Chem.*, **100** (1996), 8323–8330. https://doi.org/10.1021/jp9600672

81. R. A. Coleman, B. F. Pugh, Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA, *J. Biol. Chem.*, **270** (1995), 13850–13859. https://doi.org/10.1074/jbc.270.23.13850

82. A. Dasgupta, S. A. Juedes, R. O. Sprouse, D. T. Auble, Mot1-mediated control of transcription complex assembly and activity, *EMBO J.*, **24** (2005), 1717–1729. https://doi.org/10.1038/sj.emboj.7600646

83. R. O. Sprouse, T. S. Karpova, F. Mueller, A. Dasgupta, J. G. McNally, D. T. Auble, Regulation of TATA-binding protein dynamics in living yeast cells, *Proc. Natl. Acad. Sci. U.S.A.*, **105** (2008), 13304–13308. https://doi.org/10.1073/pnas.0801901105

84. S. H. Hosseini, *Analytic Solutions for Stochastic Models of Transcription*, Master's thesis, University of Lethbridge, 2016, URL `https://www.uleth.ca/dspace/handle/10133/4791`.

85. H.-J. Woo, Analytical theory of the nonequilibrium spatial distribution of RNA polymerase translocations, *Phys. Rev. E*, **74** (2006), 011907. https://doi.org/10.1103/PhysRevE.74.011907

86. M. H. Larson, J. Zhou, C. D. Kaplan, M. Palangat, R. D. Kornberg, R. Landick, S. M. Block, Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II, *Proc. Natl. Acad. Sci. U.S.A.*, **109** (2012), 6555–6560. https://doi.org/10.1073/pnas.1200939109

87. A. C. M. Cheung, P. Cramer, Structural basis of RNA polymerase II backtracking, arrest and reactivation, *Nature*, **471** (2011), 249–253. https://doi.org/10.1038/nature09785

88. G. Brzyżek, S. Świeżewski, Mutual interdependence of splicing and transcription elongation, *Transcription*, **6** (2015), 37–39. https://doi.org/10.1080/21541264.2015.1040146

89. M. Imashimizu, M. L. Kireeva, L. Lubkowska, D. Gotte, A. R. Parks, J. N. Strathem, M. Kashlev, Intrinsic translocation barrier as an initial step in pausing by RNA polymerase II, *J. Mol. Biol.*, **425** (2013), 697–712. https://doi.org/10.1016/j.jmb.2012.12.002

90. J. W. Roberts, Molecular basis of transcription pausing, *Science*, **344** (2014), 1226–1227. https://doi.org/10.1126/science.1255712

91. J. Singh, R. A. Padgett, Rates of *in situ* transcription and splicing in large human genes, *Nat. Struct. Mol. Biol.*, **16** (2009), 1128–1133. https://doi.org/10.1038/nsmb.1666

92. E. Rosonina, S. Kaneko, J. L. Manley, Terminating the transcript: breaking up is hard to do, *Genes Dev.*, **20** (2006), 1050–1056. https://doi.org/10.1101/gad.1431606

93. E. A. Abbondanzieri, W. J. Greenleaf, J. W. Shaevitz, R. Landick, S. M. Block, Direct observation of base-pair stepping by RNA polymerase, *Nature*, **438** (2005), 460–465. https://doi.org/10.1038/nature04268

94. L. M. Hsu, Promoter clearance and escape in prokaryotes, *Biochim. Biophys. Acta*, **1577** (2002), 191–207. https://doi.org/10.1016/S0167-4781(02)00452-9

95. H. Kimura, K. Sugaya, P. R. Cook, The transcription cycle of RNA polymerase II in living cells, *J. Cell Biol.*, **159** (2002), 777–782. https://doi.org/10.1083/jcb.200206019

96. H. A. Ferguson, J. F. Kugel, J. A. Goodrich, Kinetic and mechanistic analysis of the RNA polymerase II transcription reaction at the human interleukin-2 promoter, *J. Mol. Biol.*, **314** (2001), 993–1006. https://doi.org/10.1006/jmbi.2000.5215

97. D. A. Jackson, F. J. Iborra, E. M. M. Manders, P. R. Cook, Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei, *Mol. Biol. Cell*, **9** (1998), 1523–1536. https://doi.org/10.1091/mbc.9.6.1523

98. P. J. Hurtado, A. S. Kirosingh, Generalizations of the 'linear chain trick': Incorporating more flexible dwell time distributions into mean field ODE models, *J. Math. Biol.*, **79** (2019), 1831–1883. https://doi.org/10.1007/s00285-019-01412-w

99. H. Golstein, *Classical Mechanics*, chapter 12, Addison-Wesley, Reading, Massachusetts, 1980.

100. H. G. Othmer, A continuum model for coupled cells, *J. Math. Biol.*, **17** (1983), 351–369. https://doi.org/10.1007/BF00276521

101. C. J. Roussel, M. R. Roussel, Reaction-diffusion models of development with state-dependent chemical diffusion coefficients, *Prog. Biophys. Mol. Biol.*, **86** (2004), 113–160. https://doi.org/10.1016/j.pbiomolbio.2004.03.001

102. D. Sulsky, R. R. Vance, W. I. Newman, Time delays in age-structured populations, *J. Theor. Biol.*, **141** (1989), 403–422. https://doi.org/10.1016/S0022-5193(89)80122-5

103. G. Bel, B. Munsky, I. Nemenman, The simplicity of completion time distributions for common complex biochemical processes, *Phys. Biol.*, **7** (2010), 016003. https://doi.org/10.1088/1478-3975/7/1/016003

104. P. Billingsley, *Probability and Measure*, Wiley, New York, 1995.

105. G. Bar-Nahum, V. Epshtein, A. E. Ruckenstein, R. Rafikov, A. Mustaev, E. Nudler, A ratchet mechanism of transcription elongation and its control, *Cell*, **120** (2005), 183–193. https://doi.org/10.1016/j.cell.2004.11.045

106. J. W. Shaevitz, E. A. Abbondanzieri, R. Landick, S. M. Block, Backtracking by single RNA polymerase molecules observed at near-base-pair resolution, *Nature*, **426** (2003), 684–687. https://doi.org/10.1038/nature02191

107. M. A. Gibson, J. Bruck, Efficient exact stochastic simulation of chemical systems with many species and many channels, *J. Phys. Chem. A*, **104** (2000), 1876–1889. https://doi.org/10.1021/jp993732q

108. H. T. Banks, J. Catenacci, S. Hu, A comparison of stochastic systems with different types of delays, *Stoch. Anal. Appl.*, **31** (2013), 913–955. https://doi.org/10.1080/07362994.2013.806217

109. Y.-L. Feng, J.-M. Dong, X.-L. Tang, Non-Markovian effect on gene transcriptional systems, *Chin. Phys. Lett.*, **33** (2016), 108701. https://doi.org/10.1088/0256-307X/33/10/108701

110. J. Lloyd-Price, A. Gupta, A. S. Ribeiro, SGNS2: A compartmentalized stochastic chemical kinetics simulator for dynamic cell populations, *Bioinformatics*, **28** (2012), 3004–3005. https://doi.org/10.1093/bioinformatics/bts556

111. T. Maarleveld, *StochPy User Guide, Release 2.3.0*, 2015, URL `https://sourceforge.net/projects/stochpy/files/stochpy_userguide_2.3.pdf/download`.

112. A. S. Ribeiro, J. Lloyd-Price, SGN Sim, a stochastic genetic networks simulator, *Bioinformatics*, **23** (2007), 777–779. https://doi.org/10.1093/bioinformatics/btm004

113. D. T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, *J. Comput. Phys.*, **22** (1976), 403–434. https://doi.org/10.1016/0021-9991(76)90041-3

114. R. J. Sims III, R. Belotserkovskaya, D. Reinberg, Elongation by RNA polymerase II: the short and long of it, *Genes Dev.*, **18** (2004), 2437–2468. https://doi.org/10.1101/gad.1235904

115. E. A. M. Trofimenkoff, M. R. Roussel, Small binding-site clearance delays are not negligible in gene expression modeling, *Math. Biosci.*, **325** (2020), 108376. https://doi.org/10.1016/j.mbs.2020.108376

116. V. Epshtein, E. Nudler, Cooperation between RNA polymerase molecules in transcription elongation, *Science*, **300** (2003), 801–805. https://doi.org/10.1126/science.1083219

117. C. Jia, L. Y. Wang, G. G. Yin, M. Q. Zhang, Single-cell stochastic gene expression kinetics with coupled positive-plus-negative feedback, *Phys. Rev. E*, **100** (2019), 052406. https://doi.org/10.1103/PhysRevE.100.052406

118. J. Szavits-Nossan, R. Grima, Uncovering the effect of RNA polymerase steric interactions on gene expression noise: Analytical distributions of nascent and mature RNA numbers, *Phys. Rev. E*, **108** (2023), 034405. https://doi.org/10.1103/PhysRevE.108.034405

119. P. Bokes, J. R. King, A. T. A. Wood, M. Loose, Transcriptional bursting diversifies the behaviour of a toggle switch: Hybrid simulation of stochastic gene expression, *Bull. Math. Biol.*, **75** (2013), 351–371. https://doi.org/10.1007/s11538-013-9811-z

120. M. Dobrzyński, F. J. Bruggeman, Elongation dynamics shape bursty transcription and translation, *Proc. Natl. Acad. Sci. U.S.A.*, **106** (2009), 2583–2588.https://doi.org/10.1073/pnas.0803507106

121. L. Cai, N. Friedman, X. S. Xie, Stochastic protein expression in individual cells at the single molecule level, *Nature*, **440** (2006), 358–362. https://doi.org/10.1038/nature04599

122. A. J. M. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, et al., Genomic encoding of transcriptional burst kinetics, *Nature*, **565** (2019), 251–254. https://doi.org/10.1038/s41586-018-0836-1

123. Y. Wan, D. G. Anastasakis, J. Rodriguez, M. Palangat, P. Gudla, G. Zaki, et al., Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection, *Cell*, **184** (2021), 2878–2895. https://doi.org/10.1016/j.cell.2021.04.012

124. C. Jia, Y. Li, Analytical time-dependent distributions for gene expression models with complex promoter switching mechanisms, *SIAM J. Appl. Math.*, **83** (2023), 1572–1602. https://doi.org/10.1137/22M147219X

125. B. W. Lindgren, G. W. McElrath, D. A. Berry, *Introduction to Probability and Statistics*, 154–156, 4th edition, Macmillan, New York, 1978.

126. M. Janisch, Kolmogorov's strong law of large numbers holds for pairwise uncorrelated random variables, *Theory Probab. Appl.*, **66** (2021), 263–275. https://doi.org/10.4213/tvp5459

127. B. Li, J. A. Weber, Y. Chen, A. L. Greenleaf, D. S. Gilmour, 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** (1996), 5433–5443. https://doi.org/10.1128/MCB.16.10.5433

128. R.-J. Murphy, *Stochastic Modeling of the Torpedo Mechanism of Eukaryotic Transcription Termination*, Master's thesis, University of Lethbridge, 2017, URL `https://www.uleth.ca/dspace/handle/10133/4906`.

129. B. Choi, Y.-Y. Cheng, S. Cinar, W. Ott, M. R. Bennett, K. Josić, J. K. Kim, Bayesian inference of distributed time delay in transcriptional and translational regulation, *Bioinformatics*, **36** (2020), 586–593. https://doi.org/10.1093/bioinformatics/btz574

130. H. Hong, M. J. Cortez, Y.-Y. Cheng, H. J. Kim, B. Choi, K. Josić, J. K. Kim, Inferring delays in partially observed gene regulation processes, *Bioinformatics*, **39** (2023), btad670. https://doi.org/10.1093/bioinformatics/btad670

131. D. Holcman, Z. Schuss, The narrow escape problem, *SIAM Rev.*, **56** (2014), 213–257. https://doi.org/10.1137/120898395

132. M. R. Roussel, T. Tang, Simulation of mRNA diffusion in the nuclear environment, *IET Syst. Biol.*, **6** (2012), 125–133. https://doi.org/10.1049/iet-syb.2011.0032

133. S. Tang, *Mathematical Modeling of Eukaryotic Gene Expression*, PhD thesis, University of Lethbridge, 2010, available at: `https://www.uleth.ca/dspace/handle/10133/2567`.

AIMS Press