Roussel, Marc R.

2013

On the distribution of transcription times

Department of Chemistry and Biochemistry

Original article Biomath 2 (2013), 1307247, 1–14

Biomath Forum

BIOMATH

http://www.biomathforum.org/biomath/index.php/biomath/

On the Distribution of Transcription Times

Marc R. Roussel
Alberta RNA Research and Training Institute
Department of Chemistry and Biochemistry, University of Lethbridge
Lethbridge, Alberta, Canada
Email: roussel@uleth.ca

Received: 22 April 2013, accepted: 24 July 2013, published: 15 September 2013

***Abstract*—A previously studied model of prokaryotic transcription [Roussel and Zhu, Bull. Math. Biol. 68 (2006) 1681–1713] is revisited. The first four moments of the distribution of transcription times are obtained analytically and analyzed. A Gaussian is found to be a poor approximation to this distribution for short transcription units at typical values of the rate constants, but a good approximation for long transcription units. An approximate form of the distribution is obtained in which the slow steps are treated exactly and the fast steps are lumped together into a single lag term. This approximate form might be particularly useful as a function to be fit to experimental transcription time distributions. Multi-polymerase effects are also studied by simulation. We find that the analytic model generally predicts the behavior of the multi-polymerase simulations, often quantitatively, provided termination is not rate-limiting.**

***Keywords*-gene transcription; stochastic model**

I. Introduction

For many years, proteins were regarded as the "hardware" of the cell, with DNA as the "software" [1, p. 111]. The central dogma of molecular biology relegated RNA to a secondary role, that of a messenger between the software and hardware layers. Ribosomal and transfer RNA (rRNA and tRNA) have of course been known for a long time but, being involved in the translation machinery, they were still considered, implicitly at least, to be a means to an end, the end being proteins. In fact, "gene expression" was often explicitly defined to mean the expression of proteins as a result of transcription and translation [2, p. 327]. The roles of messenger RNA (mRNA), rRNA and tRNA in translation of course continue to be studied, and with just cause, but our appreciation for the versatility of RNA has expanded tremendously in recent years with the discovery of the diverse cellular roles of RNAs, including catalysis [3], [4], sensing [5], and regulation [6], [7], [8], [9], all roles traditionally believed to be the exclusive province of proteins.

With this greater appreciation of RNA has come an increased interest in RNA synthesis and processing processes. Gene transcription has always been of interest because of its role in the expression of protein-coding genes, but now we are equally interested in the transcription of non-coding RNAs (ncRNAs) [10], [11]. Transcription thus claims for itself a larger portion of the stage, with gene expression now including processes in which an RNA, rather than a protein, is the "gene product".

In the last few years, several models of the transcription process have appeared [12], [13], [14], [15], [16], [17], [18]. The unifying theme of these models has been a stochastic formulation designed to facilitate a study of the statistical properties of transcription. Why should we think of transcription as a stochastic process? There are a number of reasons. First and foremost is that stochasticity is an inescapable property of transcription. If we consider a typical gene in a diploid cell, there are either zero, one or two copies of the gene active at any given time. Small molecular populations inevitably lead to large fluctuations, which in this case manifest themselves as transcriptional bursts [19], [20]. The inherent stochasticity of transcription leads us to a second reason to consider stochastic models of this process, namely that

the stochasticity at this level is an important determinant of fluctuations in protein expression levels [21], [22], [23], and these fluctuations can have functional significance [24]. If we want to understand fluctuations in RNA or protein levels, or if we want to model a genetic control system in which these fluctuations are likely to be important, then we need a strong understanding of the statistical properties of transcription. If we can, for example, obtain an easily evaluated formula for the transcription time, we can apply delay-stochastic modeling methods [25], [26], [27], [28], [29] to simulate a gene expression system [30], [31].

One of the benefits of a mathematical model, and particularly of one with analytic solutions, is that it is easy to vary parameters. Among other benefits, this sometimes allows us to rapidly discover unexpected behaviors, or to provide specific criteria for the observation of some particular phenomenon. In this spirit, we have pursued models that both provide a reasonable cartoon of a biological system, and that are, in some useful limit, analytically solvable [13], [18]. While the analytic solutions are correct only under particular conditions, they often provide vital clues for interesting regimes to be studied by simulation and, ultimately, in the experimental laboratory.

In the next section, our model, which was originally presented elsewhere [13], is described, and its first four moments are obtained analytically. In section III, the moments are analyzed using the analytic expressions obtained in section II, with particular emphasis on shorter transcripts in which some interesting statistical effects are observed. An approximate form of the distribution of transcription times is also obtained, which involves an analytic expression for the distribution of the slow steps and an empirical lag phase. In section IV, the analytic predictions of this model are compared to stochastic simulations. The concluding section reviews some of the findings and offers some perspectives on this area of research.

II. A PROKARYOTIC TRANSCRIPTION MODEL

A. Model description

In our models, the nucleotides of the DNA template strand are numbered 1 to n, where n is the length of the transcription unit. Conceptually, we track the position of the active site of the RNA polymerase. In multi-polymerase simulations, a rigid polymerase is assumed (consistent with the lack of "inchworm" movements in transcription [32]), and the leading and trailing edges of the polymerase are located relative to the active site [18]. Polymerases can be prevented from overlapping by imposing a minimum distance constraint between the active sites of adjacent polymerases [18], or by adopting a more sophisticated labeling system than the one used here [16].

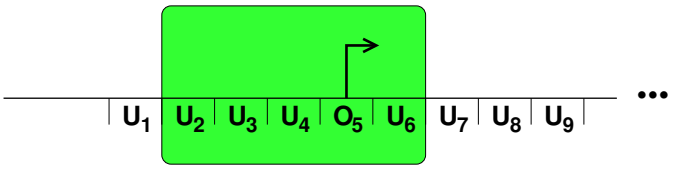

Fig. 1. Schematic diagram illustrating the relationship between the polymerase active site and the template strand states. The polymerase is represented by the green rounded rectangle, and its active site by the hooked arrow. Here, the active site occupies site 5, while all other sites are "unoccupied" in the sense discussed in the text. This illustration is not intended to realistically portray the geometry of the transcription complex. In particular, a much longer stretch of DNA passes through the polymerase than is shown here.

Each site on the DNA template strand can be labeled U (unoccupied), O (occupied by the active site of the polymerase), or A (activated for translocation). Figure 1 illustrates the relationship between the polymerase and the U and O nucleotide states. The A state can be thought of as a variant of the O state, in the sense that it represents a state in which the polymerase active site occupies a template site *and* the polymerase has been activated for translocation. We use subscripts to indicate the site to which a given state applies. Thus, U_i indicates an unoccupied site i, O_i indicates an occupied site i, and so on. In the independent polymerase case, we only need to track the location of the active site. Any site not occupied by the active site of a polymerase is labeled U_i, even though the polymerase covers many sites on the template. As noted above, we can maintain this notation and deal with the multi-polymerase case by adding distance constraints between active sites to our model, as we do in the stochastic simulations in section IV.

In the first step of our model, the polymerase locates the transcription start site (TSS):

$$\mathrm{RNAP} + \mathrm{U}_1 \xrightarrow{k_0} \mathrm{O}_1. \tag{1}$$

This of course is not a single-step process. Minimally, it would involve binding of appropriate initiation factors to the promoter, loading of the polymerase, and positioning of the polymerase at the TSS, which would typically also involve conformational changes of the DNA (e.g. unwinding) [33], [34], [35]. As an initial model however, we assume that this multi-step process has a single rate-

limiting step, represented as shown above. Note also that in a delay stochastic model of gene expression [26], [27], [28], we might separate out the binding of initiation factors and the initial binding of the polymerase to the promoter as steps to be explicitly modeled, particularly since these steps are often regulated [34]. In the present case, we further assume that the local RNA polymerase concentration is constant, so that this process reduces to a pseudo-first-order process. These assumptions can easily be relaxed, as discussed later.

Once the polymerase is positioned, the initiation complex is activated by binding the two nucleotide triphosphates complementary to the first two nucleotides of the expressed sequence. Assuming a constant pool of free nucleotides, this process can be represented as the pseudo-first-order reaction

$$\mathrm{O}_1 \xrightarrow{k_1} \mathrm{A}_1. \tag{2}$$

Note that the first two steps are being assigned a slightly different interpretation than in our original paper [13]. This is a conscious choice which will enable us to more easily use the results of our analysis in other contexts, particularly in delay stochastic simulations. Consequently, we expect k_0 to be smaller than k_1, since the former includes the rate-limiting transition from a closed to an open promoter complex [36].

Once the polymerase has been activated, formation of the phosphodiester bond between the two bound nucleotides provides the free energy required to drive the polymerase forward, i.e. to translocate:

$$\mathrm{A}_1 + \mathrm{U}_2 \xrightarrow{k_2'} \mathrm{O}_2 + \mathrm{U}_1. \tag{3}$$

Although shown as involving two "reactants", the reaction is represented this way only to maintain the logic of the labeling of sites. It is in fact a first-order process if polymerases are sufficiently widely spaced along the template strand. Otherwise, it is still a first-order process, but one whose rate constant is contingent on the position of a downstream polymerase (if any). In the multi-polymerase case, we number the polymerases in the order in which they loaded onto the DNA, so polymerase 1 is the one furthest downstream, and the polymerase with the highest index is the last one to have loaded. Let x_j be the current position of the active site of polymerase j. Then,

$$k_2' = \begin{cases} k_2 & \text{if } x_{j-1} - x_j > \Delta, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Here, k_2 is the value of the rate constant for a polymerase that is free to move, and Δ is the minimum distance between polymerase active sites.

After the first translocation, nucleotides are added one at a time in a process that alternates between activation by a nucleotide and translocation driven by phosphodiester bond formation:

$$\mathrm{O}_i \xrightarrow{k_3} \mathrm{A}_i, \quad i = 2, 3, \ldots, n; \tag{5}$$

$$\mathrm{A}_i + \mathrm{U}_{i+1} \xrightarrow{k_2'} \mathrm{O}_{i+1} + \mathrm{U}_i, \quad i = 2, 3, \ldots, n-1. \tag{6}$$

These two steps together model the elongation phase of transcription. Note that we assume here that translocation, whether from the first site (reaction 3) or from subsequent sites (reaction 6), occurs with a common rate constant, an assumption which can easily be relaxed as we discuss below.

Finally, we model termination as a single-step process although, as with initiation, we could contemplate much more complex models. In our simplified model, we assume that a polymerase becomes activated for termination in much the same way and with a similar rate as in activation for translocation. However, once the polymerase has been activated and the last phosphodiester bond formed, the polymerase, RNA and template strand dissociate from each other:

$$\mathrm{A}_n \xrightarrow{k_4} \mathrm{RNAP} + \mathrm{RNA} + \mathrm{U}_n. \tag{7}$$

Many criticisms could be made of this model. However, as a minimal model, it allows us to explore the effects of various steps on the overall statistical properties of transcription.

B. Model analysis methodology

The single-polymerase model is solvable, in the sense that the moments of the distribution of transcription times can be obtained analytically. This case arises when transcription initiation is sufficiently infrequent that polymerases only rarely interact with each other [13]. As pointed out by Greive et al. [17], our current model belongs to the class of Brownian ratchets, undergoing a set of irreversible transitions from one state to the next. We can obtain the distribution of "jump" times from one site to the next, and then using some straightforward mathematical tricks, obtain the moments of the distribution of total transcription time. We have refined this procedure somewhat since our original publication [13], so we work through the details here.

Let $\rho_i(\tau_i)$ be the distribution of jump times (τ_i) from site i to site $i+1$. We adopt the convention that $\rho_0(\tau_0)$

is the distribution of initiation times, and $\rho_n(\tau_n)$ is the distribution of termination times. The jump from site i to site $i+1$ is independent of all the other jumps, so the joint probability distribution for the jump times is just the product

$$\rho(\tau_0, \tau_1, \ldots, \tau_n) = \prod_{i=0}^{n} \rho_i(\tau_i). \tag{8}$$

The total transcription time is defined by

$$\tau = \sum_{i=0}^{n} \tau_i. \tag{9}$$

By a standard theorem of statistics, the distribution of the total transcription time is given by the convolution

$$\begin{aligned}\rho(\tau) &= \int_{\sum \tau_i = \tau} \cdots \int \rho(\tau_0, \tau_1, \ldots, \tau_n)\, d\tau_1 \ldots d\tau_n \\ &= \int_{\sum \tau_i = \tau} \cdots \int \prod_{i=0}^{n} \rho_i(\tau_i)\, d\tau_1 \ldots d\tau_n. \end{aligned} \tag{10}$$

For the class of models considered here, this convolution is not straightforwardly computable except in some special cases. However, the convolution theorem for Laplace transforms [37] allows us to convert this problem into a tractable form. Let $\tilde{f}(s) \equiv \mathcal{L}_s[f(t)]$, the Laplace transform of $f(t)$. Then,

$$\tilde{\rho}(s) = \prod_{i=0}^{n} \tilde{\rho}_i(s). \tag{11}$$

We therefore only need the Laplace transforms of the distributions of jump times from one site to the next in order to obtain the Laplace transform of the overall distribution of transcription times. The jump distribution Laplace transforms are easy to obtain from the Laplace transforms of the master equations for the survival of the occupation of a site, as illustrated in the next section. From here, we have two options:

1) The Laplace transform can sometimes be numerically (or semi-numerically) inverted to obtain the distribution of transcription times.
2) The moments of the distribution can always be obtained by differentiation of $\tilde{\rho}(s)$ [13]. From the definition of the Laplace transform, we have

$$\tilde{\rho}(s) = \int_0^\infty e^{-s\tau} \rho(\tau)\, d\tau. \tag{12}$$

By definition, the moments of the distribution are given by

$$\langle \tau^m \rangle = \int_0^\infty \tau^m \rho(\tau)\, d\tau. \tag{13}$$

Taking successive derivatives of the Laplace transform, we get

$$\langle \tau^m \rangle = (-1)^m \left. \frac{d^m \tilde{\rho}}{ds^m} \right|_{s=0}. \tag{14}$$

Differentiation is a simple mechanical operation which, in the worst case, can be carried out reliably by a symbolic algebra system. Arbitrarily high moments can be obtained by this method.

Both approaches are illustrated below.

We note here that our models are modular in precisely the same sense used by Greive et al. [17] and that, moreover, the analysis has a corresponding modular structure: We can replace any part of the model with one describing different chemistry, and the only effect is to replace the corresponding jump distribution Laplace transform in equation (11). Thus, studying model variations requires adjustments only to those parts of the analysis directly concerned with the parts of the model that have been changed. For example, rather than assuming that all translocation steps [reactions (3) and (6) in the current model] have a common rate constant, we could assume, for example, that there were different rate constants at each of the first several sites, until the polymerase-DNA-RNA complex had stabilized, after which these rate constants could reach a constant value for the remainder of the transcription process. The cost of such an assumption would be a number of added parameters, and some additional mathematical derivations for the $\tilde{\rho}_i$ in the region of variable translocation rate constant.

C. Jump distributions

In this section, we work out the Laplace transforms of the jump distributions for our model. Formally, each jump distribution is the solution of a survival problem [38] for the occupation of a given site by the polymerase active site.

In the following work, we denote by $p_{i,\sigma}$ the probability that site i is in state $\sigma \in \{\mathrm{U}, \mathrm{O}, \mathrm{A}\}$. Because we treat the case of a single polymerase, we do not need to consider joint probabilities (e.g. the probability that site i is in state A while the site to which the leading edge of the polymerase will move is unoccupied). The construction of the master equations for the joint probabilities was discussed in our previous work [13], although extending the methods used here to that case is still very much an open problem.

We start with $\tilde{\rho}_0$, which is associated with initiation or, in our model, reaction (1). Here, the relevant survival problem is the survival time for an unoccupied site 1.

Thus we assume that at $t = 0$, site 1 is unoccupied and monitor this site until it becomes occupied. Under the pseudo-first-order (constant polymerase pool) assumption, the corresponding master equation is

$$\frac{dp_{1,\mathrm{U}}}{dt} = -k_0 p_{1,\mathrm{U}}, \tag{15a}$$

$$\frac{dp_{1,\mathrm{O}}}{dt} = k_0 p_{1,\mathrm{U}}, \tag{15b}$$

with initial conditions $p_{1,\mathrm{U}}(0) = 1$, $p_{1,\mathrm{O}}(0) = 0$. While equation (15b) is apparently redundant, it plays a very important role in the survival problem: Since O_1 is a sink state in this simplified master equation, $p_{1,\mathrm{O}}(t)$ is the cumulative probability distribution for initiation. Thus,

$$p_{1,\mathrm{O}}(t) = \int_0^t \rho_0(\tau_0)\, d\tau_0, \tag{16}$$

or, using the fundamental theorem of calculus,

$$\rho_0(t) = \frac{dp_{1,\mathrm{O}}}{dt}. \tag{17}$$

In the space of Laplace transforms, and using the identity [37]

$$\mathcal{L}_s\left[df/dt\right] = s\tilde{f}(s) - f(0), \tag{18}$$

equation (17) becomes

$$\tilde{\rho}_0(s) = s\tilde{p}_{1,\mathrm{O}}(s). \tag{19}$$

The Laplace transform of equations (15) subject to the appropriate initial conditions is

$$s\tilde{p}_{1,\mathrm{U}} - 1 = -k_0\tilde{p}_{1,\mathrm{U}}, \tag{20a}$$

$$\tilde{\rho}_0(s) = s\tilde{p}_{1,\mathrm{O}} = k_0\tilde{p}_{1,\mathrm{U}}. \tag{20b}$$

Solving these equations, we get

$$\tilde{\rho}_0(s) = \frac{k_0}{s + k_0}. \tag{21}$$

This is, not surprisingly, the Laplace transform of an exponential probability distribution [37].

Once the polymerase has reached site 1, the sequence of steps (2) and (3) is required to reach site 2. The relevant master equation is

$$\frac{dp_{1,\mathrm{O}}}{dt} = -k_1 p_{1,\mathrm{O}}, \tag{22a}$$

$$\frac{dp_{1,\mathrm{A}}}{dt} = k_1 p_{1,\mathrm{O}} - k_2 p_{1,\mathrm{A}}, \tag{22b}$$

$$\frac{dp_{2,\mathrm{O}}}{dt} = k_2 p_{1,\mathrm{A}}. \tag{22c}$$

Here, $p_{2,\mathrm{O}}$ is the cumulative probability distribution for the transcription time. The initial conditions used to determine the jump time distribution are $p_{1,\mathrm{O}}(0) = 1$, $p_{1,\mathrm{A}}(0) = p_{2,\mathrm{O}}(0) = 0$. Taking the Laplace transform of these equations, we get

$$s\tilde{p}_{1,\mathrm{O}} - 1 = -k_1\tilde{p}_{1,\mathrm{O}}, \tag{23a}$$

$$s\tilde{p}_{1,\mathrm{A}} = k_1\tilde{p}_{1,\mathrm{O}} - k_2\tilde{p}_{1,\mathrm{A}}, \tag{23b}$$

$$\tilde{\rho}_1(s) = s\tilde{p}_{2,\mathrm{O}} = k_2\tilde{p}_{1,\mathrm{A}}. \tag{23c}$$

The last of these equations follows from the interpretation of $p_{2,\mathrm{O}}$ as a cumulative probability distribution for the jump times. Solving these equations, we get

$$\tilde{\rho}_1(s) = \frac{k_1 k_2}{(s + k_1)(s + k_2)}. \tag{24}$$

The elongation phase, which in this model extends from nucleotides 2 to $n - 1$, consists of reactions (5) and (6). The algebra required to derive the equation for $\tilde{\rho}_i(s)$ in this region is of course essentially identical to that required to obtain $\tilde{\rho}_1(s)$, with the result

$$\tilde{\rho}_i(s) = \frac{k_2 k_3}{(s + k_2)(s + k_3)}, \qquad i = 2, 3, \ldots, n - 1. \tag{25}$$

Termination consists of steps (5) and (7). Again, the algebra to be carried out is not too different from the foregoing. We get

$$\tilde{\rho}_n(s) = \frac{k_3 k_4}{(s + k_3)(s + k_4)}. \tag{26}$$

If we assemble the Laplace transforms of the jump time distributions according to equation (11), we get

$$\tilde{\rho}(s) = k_0 k_1 k_2^{n-1} k_3^{n-1} k_4 (s + k_0)^{-1}(s + k_1)^{-1} \times (s + k_2)^{-(n-1)}(s + k_3)^{-(n-1)}(s + k_4)^{-1}. \tag{27}$$

This equation differs from one presented in [13] by the term (21) due to the different interpretation of the two initial reactions mentioned above.

D. Moments of the distribution

Using equation (14), for the Laplace transform (27), the first two moments about zero work out to

$$\langle\tau\rangle = \frac{1}{k_0} + \frac{1}{k_1} + \frac{n-1}{k_2} + \frac{n-1}{k_3} + \frac{1}{k_4}; \tag{28}$$

$$\langle\tau^2\rangle = \langle\tau\rangle^2 + \frac{1}{k_0^2} + \frac{1}{k_1^2} + \frac{n-1}{k_2^2} + \frac{n-1}{k_3^2} + \frac{1}{k_4^2}. \tag{29}$$

Higher moments can easily be computed, although their algebraic forms are much more complicated.

The central moments actually have somewhat simpler forms than the moments about zero:

$$\begin{aligned}\sigma^2 &= \langle\tau^2\rangle - \langle\tau\rangle^2 \\ &= \frac{1}{k_0^2} + \frac{1}{k_1^2} + \frac{n-1}{k_2^2} + \frac{n-1}{k_3^2} + \frac{1}{k_4^2}; \end{aligned} \tag{30}$$

$$\begin{aligned}\mu_3 &= \langle\tau^3\rangle - 3\sigma^2\langle\tau\rangle - \langle\tau\rangle^3 \\ &= 2\left(\frac{1}{k_0^3} + \frac{1}{k_1^3} + \frac{n-1}{k_2^3} + \frac{n-1}{k_3^3} + \frac{1}{k_4^3}\right); \end{aligned} \tag{31}$$

$$\begin{aligned}\mu_4 &= \langle\tau^4\rangle - 4\mu_3\langle\tau\rangle - 6\sigma^2\langle\tau\rangle^2 - \langle\tau\rangle^4 \\ &= 3\sigma^4 + 6\left(\frac{1}{k_0^4} + \frac{1}{k_1^4} + \frac{n-1}{k_2^4} + \frac{n-1}{k_3^4} + \frac{1}{k_4^4}\right). \end{aligned} \tag{32}$$

III. Analytic results

A. Noise minimization

Analytic expressions are of course amenable to deeper analysis than simulation results. Accordingly, despite the approximations made to obtain these expressions, analysis plays a prominent role in the work of my laboratory on these problems. We can then carry out computational experiments to verify whether the properties discovered by analysis are robust, in particular to interactions between polymerases which are neglected in the foregoing work.

One interesting observation we have made repeatedly in our work is that there typically exist combinations of parameters in these models that minimize the variability in the transcription time [18], in the following sense: The coefficient of variation (CV) is defined by

$$\mathrm{CV} = \sigma/\langle\tau\rangle. \tag{33}$$

This is a measure of the relative variability in the transcription time, and is therefore one of many measures of transcriptional noise. From equations (28) and (30), we have

$$\mathrm{CV} = \frac{\sqrt{\frac{1}{k_0^2} + \frac{1}{k_1^2} + \frac{n-1}{k_2^2} + \frac{n-1}{k_3^2} + \frac{1}{k_4^2}}}{\frac{1}{k_0} + \frac{1}{k_1} + \frac{n-1}{k_2} + \frac{n-1}{k_3} + \frac{1}{k_4}}. \tag{34}$$

We have observed that the CV frequently displays a minimum when plotted against one or the other of the rate constants [18]. In fact, there is a global minimum value for the CV in the single-polymerase case, as we shall now see.

To minimize the CV, we differentiate with respect to each of the k_i and set the derivatives equal to zero. In general, from (33), we have

$$\frac{\partial(\mathrm{CV})}{\partial k_i} = \frac{1}{\langle\tau\rangle^2}\left(\frac{\partial\sigma}{\partial k_i}\langle\tau\rangle - \sigma\frac{\partial\langle\tau\rangle}{\partial k_i}\right), \tag{35}$$

which is obviously zero when

$$\frac{\partial\sigma}{\partial k_i}\langle\tau\rangle = \sigma\frac{\partial\langle\tau\rangle}{\partial k_i}. \tag{36}$$

If we evaluate equation (36) for each k_i using equations (28) and (30), we get, in each case and after some simplification,

$$\frac{\sigma^2}{\langle\tau\rangle} = \frac{1}{k_i} \tag{37}$$

which, intriguingly, is a condition on the Fano factor, another commonly used measure of noise [21], [39]. Since each of the k_i is equal to $\langle\tau\rangle/\sigma^2$ at the critical point, then they must be equal to each other. At this critical point (which is in fact a degenerate line in the space of rate constants, parameterized by the value of the common rate constant, corresponding to the choice of time scale), the CV becomes

$$\mathrm{CV}_{\min} = (2n+1)^{-1/2}. \tag{38}$$

An easy if somewhat tedious calculation shows this critical point to be a minimum, hence the label in equation (38). Thus we reach the interesting conclusion that the CV is minimized when the rate constants for all the steps of transcription are identical.

Note that we need not simultaneously optimize with respect to all of the rate constants. For example, if we suppose that the rate constants k_0 and k_4 are most directly subject to evolutionary pressures, we could minimize with respect to just those two rate constants. Equation (37) still holds for $i = 0$ and 4, so we must have $k_0 = k_4$, i.e. the CV reaches a local minimum when the initiation and termination rates are matched. In this case, we calculate

$$k_0 = k_4 = \frac{\frac{1}{k_1} + \frac{n-1}{k_2} + \frac{n-1}{k_3}}{\frac{1}{k_1^2} + \frac{n-1}{k_2^2} + \frac{n-1}{k_3^2}}. \tag{39}$$

In figure 2, the CV is plotted vs k_0 and k_4. From equation (39), we calculate that the minimum CV should occur at $k_0 = k_4 = 5.24\,\mathrm{s}^{-1}$ for the parameters used to generate the figure, and this is indeed what we observe. The value of the CV at this minimum is 0.067, which is somewhat larger than the global minimum CV of 0.050 given by equation (38) for the value of n used here. Note that the CV surface shown in the figure increases only slowly away from the minimum. Thus, CVs approaching the theoretical constrained minimum can be obtained for a wide range of values of the initiation and termination rate constants, provided neither of these is too small. The value of k_0 at which the minimum occurs is perhaps a little large [40], but noting the scale of the figure, we see

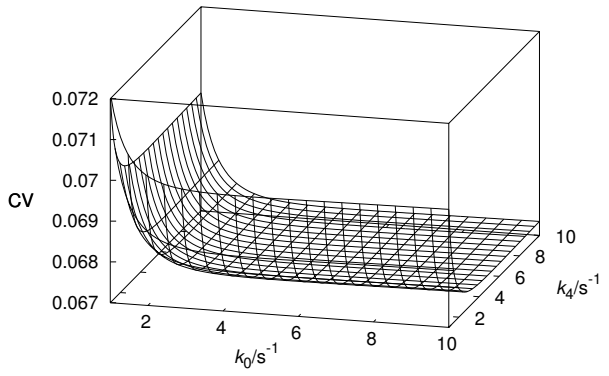

Fig. 2. CV plotted vs k_0 and k_4 with $k_1 = 0.5\,\mathrm{s}^{-1}$, $k_2 = 10\,\mathrm{s}^{-1}$, $k_3 = 10\,\mathrm{s}^{-1}$ and $n = 200$ nt.

TABLE I
BIOLOGICALLY REASONABLE RANGES FOR THE RATE CONSTANTS

k_0	10^{-2}–$10^{0}\,\mathrm{s}^{-1}$	[36], [40]
k_1	$\frac{1}{20}k_3$–$\frac{1}{10}k_3$	[36]
k_2	5–800 s^{-1}	[13]
k_3	5–800 s^{-1}	[13]
k_4	$\geq k_0$	[13]

that CVs of the same order of magnitude as the minimum are available within the biologically plausible range (k_0 up to about $1\,\mathrm{s}^{-1}$ [40]).

If we numerically minimize the CV (using the `Optimization[Minimize]` routine in Maple 15 [41]) while constraining the rate constants to lie within biologically plausible ranges (table I), we find the minimum value to be CV=0.055 for our small 200-nucleotide transcription unit when $k_0 = 1.00\,\mathrm{s}^{-1}$, $k_1 = 0.65\,\mathrm{s}^{-1}$, $k_2 = 5.00\,\mathrm{s}^{-1}$, $k_3 = 6.51\,\mathrm{s}^{-1}$ and $k_4 = 4.55\,\mathrm{s}^{-1}$, which is very close to the global minimum. Selective pressures favoring reasonably constant intervals between RNA syntheses would therefore tend to favor rapid initiation (i.e. efficient promoters) but slow elongation. The question then occurs of whether such pressures exist for any transcription units.

In the foregoing and in much of the rest of this paper, we focus on small values of n. The main reason for doing so is that many of the statistical properties studied here assume interesting extreme values in this regime. Small sequences are of biological interest since regulatory RNAs and other ncRNAs are often transcribed from small transcription units [8]. However, we will examine the effect of varying n as appropriate. For example, we can consider the effect of n on the CV of the transcription time. At large n, equation (34) tends to

$$\mathrm{CV} \to \frac{\sqrt{k_2^2 + k_3^2}}{k_2 + k_3}\,\frac{1}{\sqrt{n}}. \tag{40}$$

Thus, we see that the CV must eventually decrease with n. Moreover, the coefficient of $n^{-1/2}$ is strictly bounded between $2^{-1/2}$ (when both rate constants are equal) and 1 (when one of the coefficients is much smaller than the other, a situation that can be closely approached by taking the extreme values from table I). This means that larger transcription units have smaller CVs, although the CV does not decay quickly with transcription unit length.

B. Shape of the distribution

When faced with the problem of choosing a distribution with which to model a random process, a theoretician will generally, absent specific reasons to the contrary, choose a Gaussian. The central limit theorem certainly suggests that this is a sensible default choice [38]. However, when we have a detailed, solvable statistical model, we can explore the shape of the distribution in detail and give explicit conditions under which a Gaussian is not expected to be a good model.

To study these questions, it is useful to introduce two additional statistical parameters. The first of these is the skewness, defined by

$$\gamma_1 = \mu_3/\sigma^3, \tag{41}$$

where μ_3 is the third central moment [equation (31)]. The name of this quantity suggests its interpretation: it is a measure of asymmetry of the distribution. A symmetric distribution such as a Gaussian has a skewness of zero, while the exponential distribution has a skewness of two [42]. Because there is a sharp cut-off in the distribution of transcription times at zero, but there is no such cut-off for the long-time tail, the skewness of this distribution will always be positive, as is evident from equation (31). However, the skewness can be large or small, and if reasonably small, then a Gaussian approximation may be sufficiently accurate for generating transcription times in a delay stochastic simulation.

The second statistical parameter of interest is the excess kurtosis defined by

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3, \tag{42}$$

where μ_4 is the fourth central moment [equation (32)]. The excess kurtosis measures the heaviness of the tails of a distribution [43]. The excess kurtosis of a Gaussian

is zero, a Laplace (double exponential) distribution has an excess kurtosis of three, but the simple exponential distribution has an excess kurtosis of six, showing that the kurtosis is a subtle quantity whose precise value depends both on the tail and on the overall shape of the distribution. Combining equations (30) and (32), we find that the distribution of transcription times has an excess kurtosis of

$$\gamma_2 = \frac{6}{\sigma^4}\left(\frac{1}{k_0^4} + \frac{1}{k_1^4} + \frac{n-1}{k_2^4} + \frac{n-1}{k_3^4} + \frac{1}{k_4^4}\right) > 0. \tag{43}$$

A positive γ_2 implies that the distribution of transcription times always has heavier tails than a Gaussian. But how much heavier? In other words, under what conditions will γ_2 be small, so that a Gaussian approximation is appropriate, and when do we expect it to be large, so that a more careful choice of statistical models will be necessary?

From table I, we see that k_0 will typically be the smallest rate constant of the model. Suppose that $k_0 \ll \min\left(k_1, k_2(n-1)^{-1/4}, k_3(n-1)^{-1/4}, k_4\right)$. (The conditions on k_2 and k_3 are more restrictive than they need be, but will be needed to obtain results for the kurtosis below. The point of this demonstration is that there is a common set of conditions that makes the distribution distinctly non-Gaussian.) Then $\sigma \approx k_0^{-1}$ and $\mu_3 \approx 2k_0^{-3}$, from which we obtain $\gamma_1 \approx 2$, a value representing a highly skewed distribution.

Under the same conditions as above, $\gamma_2 \approx 6$, a value consistent with an exponential distribution. We therefore conclude that when k_0 is sufficiently small and n is not too large, the distribution is significantly non-Gaussian, with large skewness and excess kurtosis. Given the ranges in table I, these conditions will typically be realized for short transcription units provided k_4 is not too similar to k_0. Accordingly, it will typically be the case that the distribution of transcription times of short transcription units is poorly modeled by a Gaussian.

So what do these highly skewed, large-kurtosis distributions look like? Figure 3 shows an example of one of these distributions, for values of the parameters that give $\gamma_1 = 1.90$ and $\gamma_2 = 5.56$, with a CV of 0.75. This distribution was computed by the semi-numerical inversion of the Laplace transform using Maple 15 [41]. (For a Laplace transform of the form (27), Maple is able to obtain the inverse Laplace transform exactly. The only difficulty is that the coefficients of the final expression involve complicated combinations of the original parameters, which must be evaluated carefully to get accurate results. It is therefore necessary to use extra precision,

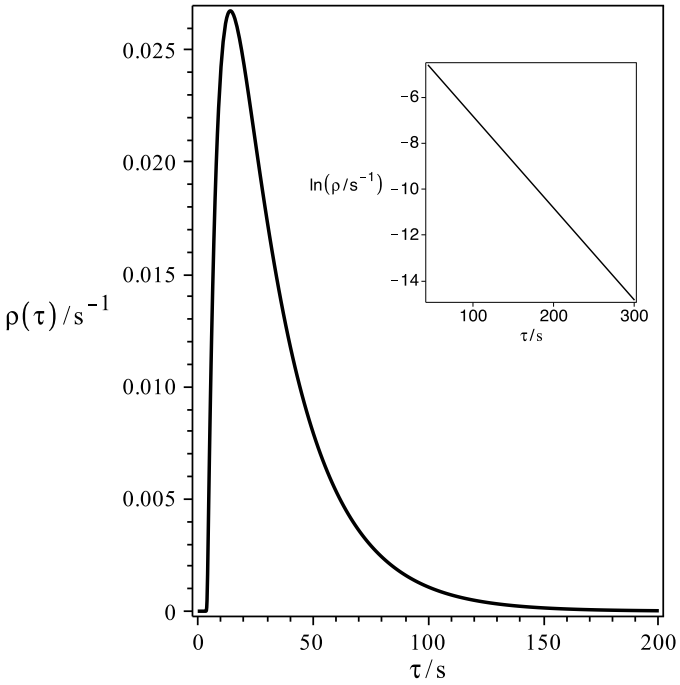

Fig. 3. Distribution of transcription times obtained by semi-numerical Laplace inversion of equation (27) for $k_0 = 0.04\,\mathrm{s}^{-1}$, $k_1 = 8.0\,\mathrm{s}^{-1}$, $k_2 = k_3 = 100.0\,\mathrm{s}^{-1}$, $k_4 = 0.2\,\mathrm{s}^{-1}$ and $n = 200\,\mathrm{nt}$. The inverse Laplace transform was computed in Maple [41], and stability of the result with respect to the number of floating-point digits used was verified. The inset shows a semi-log plot of the tail (the upper quartile of the distribution).

and to verify that the coefficients are stable to variation in the number of digits used in the calculation.) It is clear that no Gaussian could correctly capture the behavior of this distribution, both because of the strong asymmetry and because of the exponential decay of the tail. The excess kurtosis calculated for these parameters is almost as large as that of an exponential distribution, which is related to the slow, exponential decay of the tail. Indeed, in the limit of large τ, $d\ln\rho/d\tau \to -0.04\,\mathrm{s}^{-1}$ for the distribution shown in figure 3 (evaluated numerically), which is $-k_0$. This is not a coincidence: the tail in this sequential model of transcription is dominated by the slowest step, which in the case studied here is the initial location of the TSS by the polymerase.

We can also ask ourselves what happens to the skewness and excess kurtosis in the limit of large n. It is easy to see from equations (30), (31) and (41) that

$$\gamma_1 \to \frac{2(k_2^3 + k_3^3)}{\left(k_2^2 + k_3^2\right)^{3/2}}\frac{1}{\sqrt{n}}, \tag{44}$$

and from equations (30) and (43) that

$$\gamma_2 \to \frac{6(k_2^4 + k_3^4)}{(k_2^2 + k_3^2)^2}\frac{1}{n}. \tag{45}$$

Taking these two results together, we see that in the limit of large n, both the skewness and excess kurtosis tend to zero, so that a Gaussian approximation becomes increasingly accurate. We also see that the skewness goes to zero much more slowly than the excess kurtosis. There will therefore be a regime where it might be important to take the skewness of the distribution into account, but where the slowly decaying tail might not be so significant.

C. A limiting case

While we cannot, in general, analytically invert the Laplace transform (27), we can obtain approximate distributions valid in some special cases. Suppose for example that k_0 and k_4 are substantially smaller than all of the other rate constants (the case considered in figure 3). Then, for $s \ll \min(k_1, k_2, k_3)$, $\tilde{\rho}$ is approximately

$$\tilde{\rho} \approx \frac{k_0 k_4}{(s + k_0)(s + k_4)}. \tag{46}$$

This approximation would cease to hold at larger values of s. From the definition of the Laplace transform (12), we see that, because of the exponential term, the Laplace transform at large s is only significantly dependent on the value of the function at small τ. Flipping this observation on its head, the breakdown of the approximation given above at large s will only affect the distribution at small transcription times.

If we invert the Laplace transform (46), we get the two-exponential distribution

$$\rho(\tau) \approx \frac{k_0 k_4}{k_4 - k_0} \left(e^{-k_0 \tau} - e^{-k_4 \tau} \right) \equiv \rho_a(\tau) \tag{47}$$

if $k_0 \neq k_4$. (When $k_0 = k_4$, we get a gamma distribution.) The exact and approximate distributions are compared in figure 4. As expected, the two distributions differ at small τ: While the shapes of the distributions are very similar, the exact distribution has a region of negligible probability density up to about $\tau = 4\,\mathrm{s}$. This discrepancy is due to elongation, which we neglect completely in the approximate model. Although any single elongation step is fast, the large number of elongation steps results in a lag, i.e. an essentially nil probability that transcription will terminate before a certain time has elapsed, even if, by luck, the initiation and termination steps happen quickly. A much better fit to the exact distribution can be obtained by modifying the naive approximation (47) as follows:

$$\rho(\tau) \approx H(\tau - \tau_{\min}) \rho_a(\tau - \tau_{\min}), \tag{48}$$

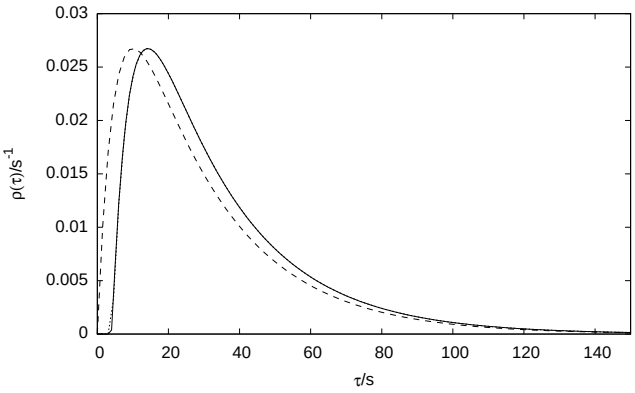

Fig. 4. Exact distribution of transcription times (solid curve replotted from figure 3), two-exponential approximate distribution (47) (dashed curve), and lag-corrected approximate distribution (48) (dotted, almost coincident with the exact distribution). The two-exponential distribution uses the exact values of k_0 and k_4, i.e. $k_0 = 0.04\,\mathrm{s}^{-1}$ and $k_4 = 0.2\,\mathrm{s}^{-1}$. For the lag-corrected distribution, k_0, k_4 and $\tau_{\min}$ are treated as fitting parameters, with least-squares estimates $k_0 = 0.039\,987 \pm 0.000\,021\,\mathrm{s}^{-1}$, $k_4 = 0.2001 \pm 0.0003\,\mathrm{s}^{-1}$, and $\tau_{\min} = 4.111 \pm 0.004\,\mathrm{s}^{-1}$.

where $\tau_{\min}$ is an additional fitting parameter and $H(\cdot)$ is the Heaviside function (0 for negative arguments, 1 for positive arguments). We call this the lag-corrected distribution.

Equation (48) was fit to the exact distribution using the Marquardt-Levenberg algorithm as implemented in Gnuplot 4.6 [44]. Both of the rate constants as well as the lag time were used as fitting parameters, resulting in the following estimates: $k_0 = 0.039\,987 \pm 0.000\,021\,\mathrm{s}^{-1}$, $k_4 = 0.2001 \pm 0.0003\,\mathrm{s}^{-1}$, and $\tau_{\min} = 4.111 \pm 0.004\,\mathrm{s}$. Note the excellent agreement between the least-squares values of k_0 and k_4 and the values used to generate the exact distribution. Moreover, from figure 4, we see that the lag-corrected distribution closely reproduces the exact distribution. Our simple ansatz involving the two small rate constants and a lag therefore gives an excellent account of the overall shape of the distribution. By fitting, we can recover k_0 and k_4, as well as the empirical parameter $\tau_{\min}$.

What is the physical meaning of $\tau_{\min}$? Several consecutive rapid steps have a narrow distribution, converging to a Dirac delta distribution as $n \to \infty$ [45], [46], [47]. The mean time for the fast steps is therefore the value of $\tau_{\min}$, at least to the extent that the lag-corrected distribution represents the exact distribution. Here, the consecutive rapid steps are the binding of the initial pair of nucleotide triphosphates to the polymerase (reaction 2) and the elongation steps (reactions 3/6 and 5).

Thus,

$$\tau_{\min} = \frac{1}{k_1} + \frac{n-1}{k_2} + \frac{n-1}{k_3}. \tag{49}$$

For the parameters of figures 3 and 4, this analytic estimate of $\tau_{\min}$ is 4.105 s, which is in excellent agreement with the value obtained by fitting equation (48) to the exact distribution.

IV. Stochastic simulations

A number of criticisms could be leveled at our model and at its analysis thus far. Perhaps the most serious criticism might be that our analysis assumes a single polymerase not interacting with other molecular machines. The case where many polymerases may transcribe the same transcription unit at the same time can easily be dealt with by stochastic (Gillespie) simulation [48]. We can then compare the predictions of our analytic theory with those of the stochastic simulations.

The stochastic model requires an additional parameter, namely the minimum distance between polymerase active sites, Δ. The bacterial RNA polymerase protects approximately 35 bp (base pairs) of the DNA duplex from cleavage by nucleases [49]. This is the length of DNA that is inaccessible to other macromolecular machines while RNAP is transcribing a gene. Moreover, the DNA takes a 90° bend on its way through the polymerase [50] (figure 5). The minimum distance between polymerase active sites is clearly 35 bp. However, there may be additional steric factors limiting the distance of closest approach, such as DNA conformational requirements, or the need to accommodate the RNA exiting the leading polymerase. For the sake of argument, we suppose that polymerases must be spaced by at least $\Delta = 40$ bp. The computed distributions of transcription times are not greatly sensitive to this choice (figure 6). In the simulations, the transcription time was taken to be the time from clearance of the TSS by the previous polymerase to completion of the transcript. This is a direct analog of the transcription time considered in the single-polymerase analytic theory.

Figure 6 also compares the distributions from Gillespie simulations to the analytic distribution (replotted from figure 3). Despite the inclusion of interactions between polymerases in the simulations, which are absent from the analytic theory, the distributions obtained are not greatly different. The statistics of the distributions are compared in Table II. The statistics confirm our visual assessment that the distribution of transcription times is not greatly affected by Δ, nor is the distribution obtained

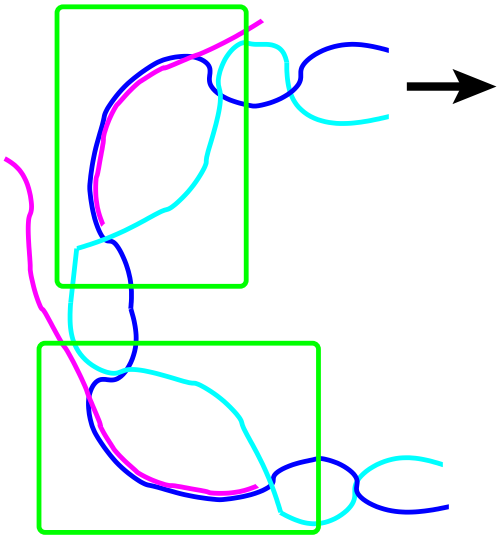

Fig. 5. Schematic diagram of a pair of polymerases simultaneously transcribing a gene. The green boxes represent the polymerases, the blue curve is the template strand, the cyan curve is the coding strand, and the magenta curve is the nascent RNA. The arrow indicates the direction in which the DNA is pulled through the polymerases. Approximately 35 bp are protected from cleavage, i.e. this length of RNA is sufficiently inside the polymerase to be inaccessible to other macromolecular machines.

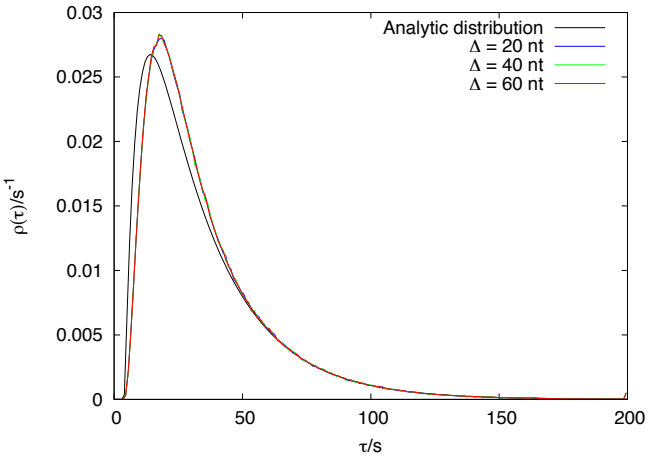

Fig. 6. Analytic distribution of transcription times (replotted from figure 3), and simulated distributions for three different values of Δ. Each simulated distribution was obtained from stochastic simulations continued until 10^6 RNAs had been synthesized.

TABLE II
Statistics for the distributions at various values of Δ compared to the statistics of the analytic distribution for the parameters of figure 6.

Δ/nt	$\langle\tau\rangle$	σ	γ_1	γ_2
20	35.37	24.82	1.98	5.97
40	35.36	24.84	1.98	6.03
60	35.34	24.84	1.98	6.02
Analytic	34.11	25.50	1.90	5.56

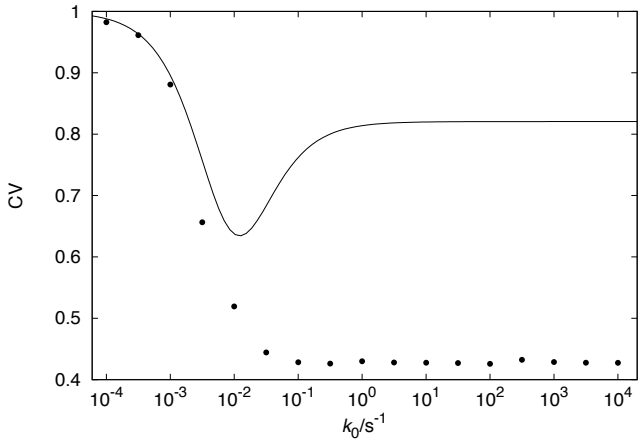

Fig. 7. CV [equation (33)] vs k_0 for $k_1 = 5\,\mathrm{s}^{-1}$, $k_2 = 10\,\mathrm{s}^{-1}$, $k_3 = 100\,\mathrm{s}^{-1}$, $k_4 = 0.01\,\mathrm{s}^{-1}$, $n = 200\,\mathrm{nt}$ and $\Delta = 40\,\mathrm{nt}$. The solid curve is the CV computed from the analytic theory, i.e. equation (34). The points are from stochastic simulations. For each value of k_0, the model was simulated until 100 000 RNAs had been synthesized.

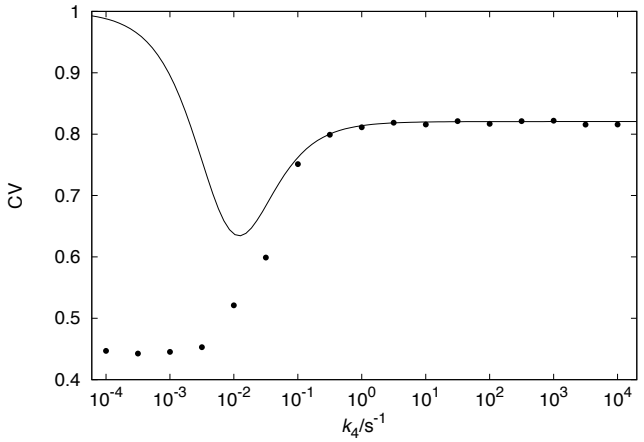

Fig. 8. CV vs k_4. All parameters and simulation conditions are as in figure 7, except $k_0 = 0.01\,\mathrm{s}^{-1}$. The solid curve is the CV computed from the analytic theory, i.e. equation (34). The points are from stochastic simulations.

from the analytic theory dramatically different from the distributions computed by stochastic simulation.

As in the analytic theory, in some stochastic simulations of related models, a minimum CV has been observed as one of the rate constants is varied [18]. Figure 7 shows the CV plotted vs k_0 both for the analytic model and from stochastic simulations. You will note that the curve computed from the simulations does not have the pronounced minimum of the analytic result. At low initiation frequencies, the density of polymerases on the template strand is small and the single-polymerase analytic theory predicts many properties of the model, including the CV, reasonably well. However, as the initiation frequency increases, interactions between polymerases become more frequent, and the single-polymerase equations become less and less accurate. As k_0 increases beyond k_4, the initiation frequency exceeds the termination frequency, and a traffic jam ensues in the simulations. Interestingly, the traffic jam condition results in a lower minimum CV than is observed at these parameters in the single-polymerase case. This is perhaps not surprising. Once a traffic jam has formed, most of the time in the queue is spent waiting, with motion limited to short bursts when the leading motor (in this case, a polymerase) exits the jam (terminates transcription).

Note that some pairs of rate constants, notably k_0 and k_4, appear symmetrically in the single-polymerase equation (34). Thus, if we set k_0 to the value of k_4 we used in figure 7, and vary k_4, the analytic theory predicts an identical curve to that obtained by varying k_0. (Compare the solid curves in figures 7 and 8.) When there are many polymerases however, varying k_0 or k_4 is not the same because it matters whether $k_0 < k_4$, in which case termination is faster than initiation and no traffic jam occurs, or $k_0 > k_4$, which leads to a traffic jam. When varying k_4, the single-polymerase theory is therefore accurate for large values of k_4.

As outlined above, the reason that the single-polymerase theory fails when either k_0 is large or k_4 is small is that termination becomes rate limiting, which causes the polymerases to pile up along the strand. On the other hand, if we vary one of the other rate constants under conditions in which one of the two initiation steps [reactions (1) and (2)] is rate limiting, we get good agreement between the single-polymerase theory and the stochastic simulations. In particular, the predicted minimum in the CV as we vary the parameters becomes a robust feature of the model, as seen in figure 9.

V. Discussion and conclusions

If we look at figures 7, 8 and 9, we see that the single-polymerase analytic theory gives excellent results provided initiation [reactions (1) and (2)] is rate limiting. Miller's classic electron micrograph snapshots of bacterial transcription in action always show the polymerases well spaced [51]. Thus, for typical transcription units, we do not observe the slow termination processes that would cause the traffic jams shown here to lead to significant deviations from the single-polymerase theory. Of course, transcriptional pauses, particularly if they occur late in a

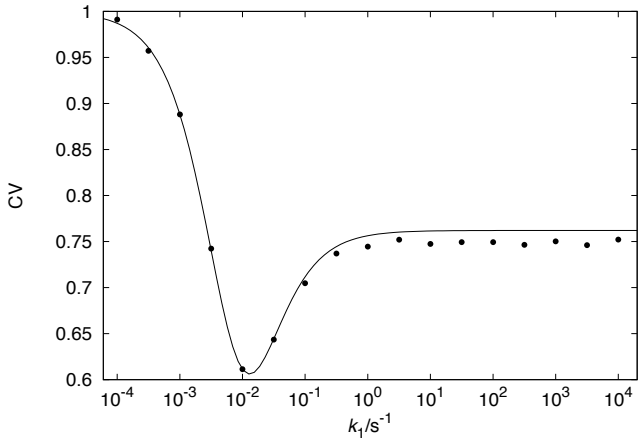

Fig. 9. CV vs k_1 computed from the analytic theory (equation (34), solid line) and from stochastic simulations (dots). The parameters were as follows: $k_0 = 0.01\,\mathrm{s}^{-1}$, $k_2 = 10\,\mathrm{s}^{-1}$, $k_3 = 100\,\mathrm{s}^{-1}$, $k_4 = 0.1\,\mathrm{s}^{-1}$, $n = 200$ and $\Delta = 40$. In the stochastic simulations, statistics were collected until 100 000 RNAs had been synthesized.

transcription unit, could have a similar effect. Pauses are known to occur during transcription in prokaryotes [52], [53], and there are specific sites on the template that are pause-prone [54]. The effect of pausing on single-polymerase transcription statistics has been studied by Voliotis and coworkers [15], where pausing was found to cause a heavy-tailed distribution of transcription times. Ribeiro's group studied pausing in a gene expression model and found, among other things, that a pause in the middle of a gene can cause a trimodal distribution of intervals between transcript completions, with the middle mode corresponding to "normal" spacing of the polymerase, one mode corresponding to microbursts (two or three transcriptions completing in an unusually short interval), a less extreme version of the traffic jams observed here, and another corresponding to the long intervals occasioned when one polymerase runs through the pause site without pausing, while the next one does pause [55]. Clearly, it would be interesting to look at the similarities and differences between the effects of pausing and of simple traffic jams caused by slow termination.

Traffic is a particular problem for the ribosomal RNA genes, which are transcribed at very high rates [56]. Klumpp and Hwa have studied traffic in a model of rRNA transcription, where they found that short pauses, which are common events during transcription, would cause traffic jams were it not for the action of the antiterminator (AT) complex which, among other things, inhibits pausing [14]. To maintain high rates, it is not enough to inhibit pausing in most elongation complexes; it is also necessary to remove complexes that have not assembled with AT, lest they cause traffic jams. Here, Klumpp and Hwa found that the termination factor Rho can have precisely the desired effect by removing slow, pause-prone polymerases from the template, thus allowing elongation complexes properly complexed with AT to work at the optimal rate. For rapidly transcribed genes, of which the rRNA genes are an extreme, traffic may thus prove to be a rate-limiting process.

In the large n regime, the CV, skewness and excess kurtosis all go to zero [equations (40), (44) and (45)]. At large n then, the distribution is at once relatively narrow, minimally skewed, and roughly as heavy-tailed as a Gaussian. Long transcription units therefore pose few modeling difficulties. Their transcription time distributions are reasonably Gaussian, so two parameters, the mean and variance, are sufficient to describe these distributions. It might even be tolerable, because of the small CV, to use a fixed delay in these cases, just as can be done for ordinary differential equation subsystems consisting of a linear decay chain [47].

The transcription time distributions of short transcription units on the other hand may be highly skewed and have large excess kurtoses. In such distributions, the mode, median and mean are quite different, so that there is no uniquely defined "typical" behavior. Accordingly, modeling the expression of short transcription units requires a careful approach. We may in some cases be able to obtain the exact distribution by inversion of the Laplace transform (27) (or the equivalent expression for a model involving additional biochemical steps) by a semi-numerical method, as was done for figure 3 using Maple [41], or by a fully numerical method [57]. Alternatively, we could use the lag-corrected ansatz (48) as an approximation to the exact distribution. Any of the above representations of the distribution can then be used to generate random deviates in a delay-stochastic code like SGN Sim [29]. Of course, all approaches based on the computation of a single-polymerase transcription time distribution assume that perturbations due to interactions between polymerases, or between polymerases and ribosomes [58], are insignificant. If this is not the case, then eventually we will want to develop a many-body theory that yields approximate analytic distributions or moments thereof.

The approach of section III-C can be generalized to any small number of slow steps since it is easy to work out the inverse Laplace transform in these cases with the assistance of a computer algebra system like Maple [41].

If we have a mechanistic model in which the slow steps are identified, fitting a lag-corrected distribution to an experimental distribution will give estimates of the slow rate constants and of the lag time, which combines all the fast processes. These lag-corrected distributions are much simpler to handle than their exact counterparts, and summarize all of the information that can reliably be extracted from a sequential set of reaction processes with a few slow steps and many fast steps. Note that experimental distributions of transcription times are beginning to appear in the literature [59], [60].

As mentioned earlier, our model is modular. More complicated modules can be substituted for some of the modules in the current, highly simplified model, leading to more complex survival time problems. Our eukaryotic transcription model [18] contains some modules not present in our current prokaryotic model, such as abortive initiation and pausing modules. Because of the product form of the Laplace transform (11), replacing or adding modules to the model is easy. In the model studied here, the various moments and derived quantities of interest (CV, skewness and kurtosis) adopt particularly simple forms. However, if we insert more complex modules into the model, especially ones that represent alternative pathways rather than simply additions to the sequential chain of events, then the expressions obtained for the moments become less straightforward. Many interesting modules can still be treated in the framework presented here. Future publications will describe this work, with the current publication, as well as our original paper [13], serving as an important baseline against which new models can be compared.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. D. Watson, Molecular Biology of the Cell, Garland, 1983.

[2] W. K. Purves, G. H. Orians, Life: The Science of Biology, Sinauer, Sunderland, MA, 1983.

[3] G. Talini, S. Branciamore, E. Gallori, Ribozymes: Flexible molecular devices at work, Biochimie 93 (2011) 1998–2005. http://dx.doi.org/{10.1016/j.biochi.2011.06.026}

[4] D. M. J. Lilley, Mechanisms of RNA catalysis, Phil. Trans. R. Soc., Ser. B 366 (2011) 2910–2917. http://dx.doi.org/{10.1098/rstb.2011.0132}

[5] A. Serganov, E. Nudler, A decade of riboswitches, Cell 152 (2013) 17–24. http://dx.doi.org/10.1016/j.cell.2012.12.024

[6] K. B. Massirer, A. E. Pasquinelli, The evolving role of microRNAs in animal gene expression, BioEssays 28 (2006) 449–452. http://dx.doi.org/10.1002/bies.20406

[7] K. Kim, Y. S. Lee, R. W. Carthew, Conversion of pre-RISC to holo-RISC by Ago2 during assembly of RNAi complexes, RNA 13 (2007) 22–29. http://dx.doi.org/10.1261/rna.283207

[8] L. S. Waters, G. Storz, Regulatory RNAs in bacteria, Cell 136 (2009) 615–628. http://dx.doi.org/10.1016/j.cell.2009.01.043

[9] M. Falaleeva, S. Stamm, Processing of snoRNAs as a new source of regulatory non-coding RNAs, BioEssays 35 (2013) 46–54. http://dx.doi.org/10.1002/bies.201200117

[10] J. Ponjavic, C. P. Ponting, The long and the short of RNA maps, BioEssays 29 (2007) 1077–1080. http://dx.doi.org/10.1002/bies.20669

[11] P. Carninci, The long and short of RNAs, Nature 457 (2009) 974–975. http://dx.doi.org/10.1038/457974b

[12] F. Jülicher, R. Bruinsma, Motion of RNA polymerase along DNA: A stochastic model, Biophys. J. 74 (1998) 1169–1185. http://dx.doi.org/10.1016/S0006-3495(98)77833-6

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

[14] 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. http://dx.doi.org/10.1073/pnas.0806084105

[15] M. Voliotis, N. Cohen, C. Molina-París, T. B. Liverpool, Fluctuations, pauses, and backtracking in DNA transcription, Biophys. J. 94 (2008) 334–348. http://dx.doi.org/10.1529/biophysj.107.105767

[16] 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. http://dx.doi.org/10.1089/cmb.2008.0153

[17] 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. http://dx.doi.org/10.1016/j.bpj.2011.07.042

[18] S. Vashishtha, Stochastic modeling of eukaryotic transcription at the single nucleotide level, Master's thesis, University of Lethbridge (2011). https://www.uleth.ca/dspace/handle/10133/3190

[19] J. R. Chubb, T. B. Liverpool, Bursts and pulses: Insights from single cell studies into transcriptional regulation, Curr. Opin. Genet. Dev. 20 (2010) 478–484. http://dx.doi.org/10.1016/j.gde.2010.06.009

[20] 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. http://dx.doi.org/10.1126/science.1198817

[21] M. Thattai, A. van Oudenaarden, Intrinsic noise in gene regulatory networks, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 8614–8619. http://dx.doi.org/10.1073/pnas.151588598

[22] J. Hasty, J. J. Collins, Translating the noise, Nat. Genet. 31 (2002) 13–14. http://dx.doi.org/10.1038/ng0502-13

[23] R. Zhu, D. Salahub, Delay stochastic simulation of single-gene expression reveals a detailed relationship between protein noise and mean abundance, FEBS Lett. 582 (2008) 2905–2910. http://dx.doi.org/10.1016/j.febslet.2008.07.028

[24] T.-L. To, N. Maheshri, Noise can induce bimodality in positive transcriptional feedback loops without bistability, Science 327 (2010) 1142–1145. http://dx.doi.org/10.1126/science.1178962

[25] 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. http://dx.doi.org/10.1021/jp993732q

[26] 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. http://dx.doi.org/10.1073/pnas.0503858102

[27] M. Barrio, K. Burrage, A. Leier, T. Tian, Oscillatory regulation of Hes1: Discrete stochastic delay modelling and simulation, PLoS Comput. Biol. 2 (2006) e117. http://dx.doi.org/10.1371/journal.pcbi.0020117

[28] 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) 274–284. http://dx.doi.org/10.1088/1478-3975/3/4/005

[29] A. S. Ribeiro, J. Lloyd-Price, SGN Sim, a stochastic genetic networks simulator, Bioinformatics 23 (2007) 777–779. http://dx.doi.org/10.1093/bioinformatics/btm004

[30] A. S. Ribeiro, Stochastic and delayed stochastic models of gene expression and regulation, Math. Biosci. 223 (2010) 1–11. http://dx.doi.org/10.1016/j.mbs.2009.10.007

[31] I. Potapov, J. Lloyd-Price, O. Yli-Harja, A. S. Ribeiro, Dynamics of a genetic toggle switch at the nucleotide and codon levels, Phys. Rev. E 84 (2011) 031903. http://dx.doi.org/10.1103/PhysRevE.84.031903

[32] E. Nudler, A. Goldfarb, M. Kashlev, Discontinuous mechanism of transcription elongation, Science 265 (1994) 793–796. http://dx.doi.org/10.1126/science.8047884

[33] D. A. Bushnell, K. D. Westover, R. E. Davis, R. D. Kornberg, Structural basis of transcription: An RNA polymerase II-TFIIB cocrystal at 4.5 angstroms, Science 303 (2004) 983–988. http://dx.doi.org/10.1126/science.1090838

[34] N. J. Fuda, M. B. Ardehali, J. T. Lis, Defining mechanisms that regulate RNA polymerase II transcription *in vivo*, Nature 461 (2009) 186–192. http://dx.doi.org/10.1038/nature08449

[35] A. Golubev, Genes at work in random bouts, BioEssays 34 (2012) 311–319. http://dx.doi.org/10.1002/bies.201100119

[36] W. R. McClure, Rate-limiting steps in RNA chain initiation, Proc. Natl. Acad. Sci. U.S.A. 77 (1980) 5634–5638.

[37] E. Butkov, Mathematical Physics, Addison-Wesley, Reading, MA, 1968.

[38] N. G. van Kampen, Stochastic Processes in Physics and Chemistry, North-Holland, Amsterdam, 1981.

[39] J. W. Shuai, S. Zeng, P. Jung, Coherence resonance: On the use and abuse of the Fano factor, Fluct. Noise Lett. 2 (2002) L139–L146. http://dx.doi.org/10.1142/S0219477502000749

[40] A. Drennan, M. Kraemer, M. Capp, T. Gries, E. Ruff, C. Sheppard, S. Wigneshweraraj, I. Artsimovitch, M. T. Record, Jr., Key roles of the downstream mobile jaw of *Escherichia coli* RNA polymerase in transcription initiation, Biochemistry 51 (2012) 9447–9459. http://dx.doi.org/10.1021/bi301260u

[41] Maplesoft, Waterloo, ON, Maple 15 (2011). http://www.maplesoft.com

[42] M. G. Bulmer, Principles of Statistics, Dover, Mineola, N.Y., 1979.

[43] M. M. Ali, Stochastic ordering and kurtosis measure, J. Amer. Statist. Assoc. 69 (1974) 543–545. http://dx.doi.org/10.1080/01621459.1974.10482990

[44] T. Williams, C. Kelley, Gnuplot 4.6 (2012). http://www.gnuplot.info

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

[46] K. L. Cooke, Z. Grossman, Discrete delay, distributed delay and stability switches, J. Math. Anal. Appl. 86 (1982) 592–627. http://dx.doi.org/10.1016/0022-247X(82)90243-8

[47] C. J. Roussel, M. R. Roussel, Delay-differential equations and the model equivalence problem in chemical kinetics, Phys. Can. 57 (2001) 114–120.

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

[49] S. J. Greive, P. H. von Hippel, Thinking quantitatively about transcriptional regulation, Nat. Rev. Mol. Cell Biol. 6 (2005) 221–232. http://dx.doi.org/10.1038/nrm1588

[50] N. Korzheva, A. Mustaev, M. Kozlov, A. Malhotra, V. Nikiforov, A. Goldfarb, S. A. Darst, A structural model of transcription elongation, Science 289 (2000) 619–625. http://dx.doi.org/10.1126/science.289.5479.619

[51] O. L. Miller, Jr., B. A. Hamkalo, C. A. Thomas, Jr., Visualization of bacterial genes in action, Science 169 (1970) 392–395. http://dx.doi.org/10.1126/science.169.3943.392

[52] K. Adelman, A. La Porta, T. J. Santangelo, J. T. Lis, J. W. Roberts, M. D. Wang, Single molecule analysis of RNA polymerase elongation reveals uniform kinetic behavior, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 13538–13543. http://dx.doi.org/10.1073/pnas.212358999

[53] 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. http://dx.doi.org/10.1038/nature02191

[54] R. J. Davenport, G. J. L. Wuite, R. Landick, C. Bustamante, Single-molecule study of transcriptional pausing and arrest by *E. coli* RNA polymerase, Science 287 (2000) 2497–2500. http://dx.doi.org/10.1126/science.287.5462.2497

[55] T. Rajala, A. Häkkinen, S. Healy, O. Yli-Harja, A. S. Ribeiro, Effects of transcriptional pausing on gene expression dynamics, PLoS Comput. Biol. 6 (2010) e1000704. http://dx.doi.org/10.1371/journal.pcbi.1000704

[56] S. Klumpp, T. Hwa, Traffic patrol in the transcription of ribosomal RNA, RNA Biol. 6 (2009) 392–394. http://dx.doi.org/10.4161/rna.6.4.8952

[57] J. Lee, D. Sheen, An accurate numerical inversion of Laplace transforms based on the location of their poles, Comput. Math. Appl. 48 (2004) 1415–1423. http://dx.doi.org/10.1016/j.camwa.2004.08.003

[58] S. Proshkin, A. R. Rahmouni, A. Mironov, E. Nudler, Cooperation between translating ribosomes and RNA polymerase in transcription elongation, Science 328 (2010) 504–508. http://dx.doi.org/10.1126/science.1184939

[59] M. Kandhavelu, A. Häkkinen, O. Yli-Harja, A. S. Ribeiro, Single-molecule dynamics of transcription of the lar promoter, Phys. Biol. 9 (2012) 026004. http://dx.doi.org/10.1088/1478-3975/9/2/026004

[60] M. Kandhavelu, J. Lloyd-Price, A. Gupta, A.-B. Muthukrishnan, O. Yli-Harja, A. S. Ribeiro, Regulation of mean and noise of the in vivo kinetics of transcription under the control of the *lac/ara-1* promoter, FEBS Lett. 586 (2012) 3870–3875. http://dx.doi.org/10.1016/j.febslet.2012.09.014