Bulletin of Mathematical Biology (2006) 68: 1681–1713
DOI 10.1007/s11538-005-9048-6

ORIGINAL ARTICLE

Stochastic Kinetics Description of a Simple Transcription Model

Marc R. Roussel*, Rui Zhu

Department of Chemistry and Biochemistry, University of Lethbridge, Lethbridge, Alberta T1K 3M4, Canada

Received: 21 April 2005 / Accepted: 8 August 2005 / Published online: 20 June 2006

Abstract We study a stochastic model of transcription kinetics in order to characterize the distributions of transcriptional delay and of elongation rates. Transcriptional delay is the time which elapses between the binding of RNA polymerase to a promoter sequence and its dissociation from the DNA template strand with consequent release of the transcript. Transcription elongation is the process by which the RNA polymerase slides along the template strand. The model considers a DNA template strand with one promoter site and n nucleotide sites, and five types of reaction processes, which we think are key ones in transcription. The chemical master equation is a set of ordinary differential equations in 3^n variables, where n is the number of bases in the template. This model is too huge to be handled if n is large. We manage to get a reduced Markov model which has only $2n$ independent variables and can well approximate the original dynamics. We obtain a number of analytical and numerical results for this model, including delay and transcript elongation rate distributions. Recent studies of single-RNA polymerase transcription by using optical-trapping techniques raise an issue of whether the elongation rates measured in a population are heterogeneous or not. Our model implies that in the cases studied, different RNA polymerase molecules move at different characteristic rates along the template strand. We also discuss the implications of this work for the mathematical modeling of genetic regulatory circuits.

Keywords Transcription · Elongation · Delay · Stochastic kinetics · Master equation

1. Introduction

The law of mass-action, which treats concentrations as continuous variables, is not suitable for systems consisting of small numbers of molecules in which individual reaction events dominate the behavior. Rather, stochastic kinetics methods are

*Corresponding author.
E-mail address: roussel@uleth.ca (Marc R. Roussel).

necessary for the proper description of small systems. It is well known that transcription, the first phase of gene expression, involves only very small numbers of reacting molecules in cells. The aim of this paper was to study a transcription model system by using two stochastic kinetics methods, namely the chemical master equation and stochastic simulations, both of which are based on the same microphysical premise, namely that of a well-mixed system (Gillespie, 1992). Although cells are not by any means well-mixed compartments, studying such models is a sensible first step toward understanding stochasticity and its effects in biochemical reactions. Analysis of the chemical master equation, when it is feasible, leads to analytic expressions for the statistical distributions of random variables of interest, while stochastic simulations help us develop intuition by showing us how instances of a process unfold.

Transcription is the process that occurs as DNA is used as a template to create RNA. A single gene may contain hundreds or thousands of base pairs. A similar number of nucleoside triphosphate molecules (NTPs) will have to be added to the transcript through the action of RNA polymerase (RNAP) molecules and transcription regulatory factors. The process consists of three main stages: initiation, elongation, and termination. Each of the three stages is composed of many reaction processes. The model presented in this paper considers only five basic reaction processes: RNAP binding to promoters, an active transcription complex head forming, the transcription complex stepping forward by one nucleotide on the DNA template strand, a complementary NTP binding to the complex, and the complex dissociating. The first two processes are involved in initiation, the next two repeat in elongation, and the last one occurs in termination.

In reality, the transcription process is much more complex than the model presented here (von Hippel, 1998). Moreover, transcription differs in detail in prokaryotes and eukaryotes. It is not difficult to add more kinds of processes into the model, making it more complex. However, if too much more detail is added to the model, the master equation analysis will become inaccessible. The model analyzed here describes the key reactions involved in all transcription processes, yet is sufficiently simple that we can derive analytic expressions for statistical quantities from the master equation analysis. We use this simple model as our starting point, intending to obtain some insights into transcription in view of the intrinsically stochastic nature of chemical reactions. Specifically, we have two main purposes in this work. One is to get the distribution of transcriptional delay, and the other is the distribution of transcription elongation rate.

In gene expression, some biological processes are obviously time consuming, such as the transcript birth and modification, mature mRNA transport to the cytoplasm, and mRNA translation to a functional protein molecule (Ota et al., 2003). These delays often play important roles in biochemical dynamics and so must be incorporated in mathematical models which include a genetic regulatory component (Bliss et al., 1982; Buchholtz and Schneider, 1987; Busenberg and Mahaffy, 1988; Mahaffy et al., 1992; Smolen et al., 1998). Recent modeling studies by Lewis and Monk show that time delays in eukaryotes play a key role in the production of the oscillations of some expressed proteins, called delay-driven oscillations (Lewis, 2003; Monk, 2003). Besides delays, another important factor in gene expression is stochastic fluctuation because these systems involve small numbers of molecules.

Most of the studies which have considered transcriptional and translational delays have been based on differential equations and have thus neglected the stochastic aspect of genetic regulation. Lewis (2003) gives some stochastic kinetics descriptions, but the time delay is still assumed to be constant in the modeling. Monk (2003) considers distributed delays, but not stochastic kinetics. Smolen et al. (1999, 2001) considered both stochastic kinetics and distributed delays (separately), but their simulation methods were not directly based on the underlying microscopic physics. Similar comments could apply to a number of other studies. The fundamental problem, which we begin to address in this paper is that we do not know the true distributions of the delays in gene transcription and translation. We focus here on transcription. The modeling issues in treating translation are very similar, and there is no reason to think that the approach taken in this paper would not extend naturally to protein synthesis.

Transcriptional delay and transcription elongation are two closely related terms in transcription kinetics. Transcription elongation is the process by which the RNAP slides along the template strand, adding bases to the transcript as it goes. As with the time delay, the transcription elongation rate is also a random variable. A full understanding of gene expression requires knowledge of the statistical character of the elongation rate in different situations. Modern techniques can allow us to monitor transcription processes by individual RNAP molecules, and much statistical information about elongation rates has been obtained. In some single-molecule experiments that measure elongation rates of single *Escherichia coli* (*E. coli*) RNAP molecules in vitro, each RNAP is observed to move at a single characteristic rate over a distance of 1000 bp of template DNA, but the elongation rates measured in a population are heterogeneous (Davenport et al., 2000; Tolić-Nørrelykke et al., 2004). However, it is sometimes difficult to distinguish experimentally between population heterogeneity and reaction stochasticity. Accordingly, the above interpretation has been controversial (Adelman et al., 2002). In the second part of this work, we aim to discover, by studying the distribution of elongation rates, how much of the observed fluctuations in elongation rates are due to intrinsic noise. This provides a basis for discussing the apparently paradoxical experimental results.

In Section 2, we give our transcription model. In Section 3, a nonlinear master equation for the stochastic kinetics of the model is derived, and several analytic results are obtained. In Section 4, we explore the distributions of time delay and transcription elongation rate in various cases. In the final section, Section 5, we summarize the main features of the model, and comment on the implications of our findings.

2. A single-gene transcription model

It is known that transcription takes place in three stages: initiation, elongation and termination. In each stage, there are many reaction processes. We propose the following simple single-gene transcription mechanism involving only five kinds of reactions which we think are basic ones in most transcriptions both in prokaryotes and in eukaryotes (Stryer, 1988; von Hippel and Pasman, 2002).

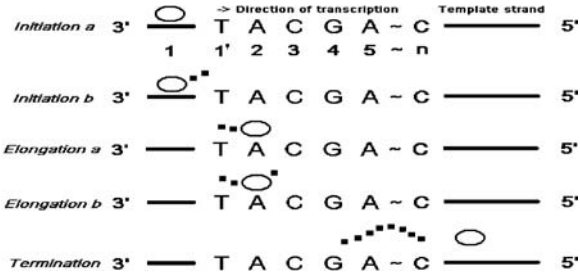

Fig. 1 Illustration of the transcription mechanism proposed in this paper. The ellipse denotes an RNAP molecule, and the black squares denote NTPs. See text for details.

Figure 1 schematically illustrates these five kinds of key reactions for this mechanism. The strand of the double stranded DNA that the RNAP binds to as it moves along is called the template strand. The promoter is numbered 1, and nucleotides that are after the promoter where transcription actually starts are numbered $1'$, 2, 3, 4, ..., n, respectively. Here, n is the number of nucleotides in the gene.

Firstly, initiation begins as RNAP recognizes and binds to the promoter. This is the first step of initiation, which is illustrated in Fig. 1 as *initiation a*. Next, the two strands of DNA in that binding region begin to unwind so that the RNAP can get closer to the individual nucleotides. An unwound segment of DNA forms and stays unwound throughout the transcription process. Meanwhile, two nucleoside triphosphate molecules (NTPs) that can pair with the nucleotides in the $1'$ and 2 positions of the template strand bind to the RNAP. These processes are involved in the second step of initiation called *initiation b*. After the active head is formed, the elongation process begins, which is divided into two steps. *Elongation a*: A phosphodiester bond forms between these two ribonucleotides, generating some chemical energy; using the energy, the RNAP steps forward to the position 2. *Elongation b*: One NTP paired with the nucleotide in the position 3 binds to the RNAP. Later on, a second phosphodiester bond forms (*elongation a*), driving RNAP a second forward step to the position 3 (*elongation b*). The two elongation steps repeat until the last nucleotide in the n position is reached. During transcription, the RNAP molecule, the unwound DNA segment, and the nascent RNA molecule compose an important complex called the transcription complex. In the elongation phase more NTPs are added as the transcription complex slides along the DNA template strand to form a transcript chain. Finally, *termination* comes as the formation of phosphodiester bonds ceases, the transcription complex dissociates, and the melted region of DNA rewinds. Just as the template strand of DNA being transcribed has promoter regions indicating where transcription should start, it also has stop signals to end the process.

In the above description of the mechanism, we note that there exist three states for each site i, $i = 1, 2, 3, \ldots, n$. We denote site i by U_i (unoccupied site) if there is no RNAP on it, by O_i (occupied site) when an RNAP has moved to it, and by A_i (activated site) if the RNAP at the site has bound an NTP (or two for $i = 1$) and

is ready to perform its catalytic ratchet action. Based on the mechanism, we write down the corresponding chemical processes:
Initiation

a. A RNAP molecule, denoted by P, binds to the promoters, forming a closed-promoter complex; the first occupied site forms.

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

b. The two strands of DNA unwind in the region binding to the RNAP, forming an open-promoter complex; at the same time, two NTPs bind to the RNAP. The first activated site forms.

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

Elongation

a. A phosphodiester bond forms, and the generated energy drives the RNAP to step forward; an activated site becomes an unoccupied one while the closest downstream site makes a transition from the unoccupied to the occupied state.

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

b. An NTP binds to the RNAP; the occupied site in step 3 becomes active.

$$\mathrm{O}_i \xrightarrow{k_3} \mathrm{A}_i \quad \text{for} \quad i = 2, 3, \ldots, n \tag{4}$$

Termination

The transcription complex dissociates: A mature transcript is produced and the RNAP releases the DNA. The last activated site becomes unoccupied.

$$\mathrm{A}_n \xrightarrow{k_4} \mathrm{P} + \mathrm{R} + \mathrm{U}_n \tag{5}$$

Here, R is an RNA transcript.

In steps (2) and (4), we omitted the NTPs. If there is a large pool of NTPs maintained at constant concentrations by homeostatic mechanisms, then step (4) will certainly reduce to a quasi-first-order reaction, as shown. Since cell-cycle checkpoints inhibit DNA synthesis when the cell is in a biochemically unfavorable state (Elledge, 1996), NTP homeostasis is in fact a sensible hypothesis for our model. The argument is a bit trickier for step (2), which requires the binding of a pair of NTPs and thus arguably the breaking up of this step into several component biochemical processes. However, our model is sufficient to gain an understanding of the factors which affect transcription so that we were reluctant to add much detail

to our description of initiation. We also note that in vitro studies of transcription kinetics are sometimes carried out in flow cells where the NTP concentrations are rigorously constant (Davenport et al., 2000). In other cases, a large excess of NTPs is used, which has the same effect. The use of the pool chemical approximation for the NTPs is thus reasonable for in vivo kinetics, and essentially exact for most in vitro studies.

It should be noted that the values of the probability rate constants for the two repeated elongation events are the same for all of the sites involved in elongation. Implicitly, this involves two further assumptions, neither of which is strictly true in vivo, but both of which can be made at least approximately true in vitro: First, that site-to-site motion is sequence independent, and second, that the NTPs are all equally abundant.

Finally, our model assumes that polymerases occupy just one site (one nucleotide) on the DNA template. This is perhaps the most questionable assumption of our model. The justification for this approximation is that initiation rates are such that RNAPs typically do not crowd together during transcription (Miller and Beatty, 1969; Miller et al., 1970). Accordingly, it will be a relatively rare event for two RNAPs to come sufficiently close together for this model simplification to matter. Both this and the other assumptions outlined above were motivated by a desire to start from a simple, analytically tractable model which retains at least the key features of transcription. In later work, we can add various complicating factors and see their effects, using the current model as a base. For now however it seemed unwise to create an overly complex model which could not be properly understood.

Reasonable ranges for the reaction probability rate constants (k_i) can all be estimated from experimental data. Strong promoters cause frequent binding of RNAP, as often as every 2 s in *E. coli*, while genes with weak promoters are transcribed about once in 10 min (Stryer, 1988). In our stochastic model, this rate corresponds to the number of available polymerases multiplied by k_0, a quantity which we call $\tilde{k}_0$ in the formal development below. A reasonable range for $\tilde{k}_0$ would therefore be from 0.0017 to 0.5 s^{-1}. The transition from the closed promoter complex to the open promoter complex is usually slower than the binding of RNAP to the promoters, and it is a rate-limiting step in the initiation (McClure, 1980). Thus, in our model, $k_1 << \tilde{k}_0$. The rate of transcription elongation varies dramatically among RNAPs, ranging from 5 to 400 nt/s (nucleotides per second) (Uptain et al., 1997). Eukaryotic RNAPs are estimated to elongate at 20–30 nt/s in vivo, and the elongation rate in *E. coli* is higher, about 50 nt/s. Thus, k_2 and k_3 should both be between 5 and 800 s^{-1}.

The magnitude relationship between the initiation and elongation rates strongly implies that in reality the RNAP molecules, until they move to the end of the strand, are typically separated from each other by a considerable distance. This has been clearly shown in the electron micrographs of transcription (Miller and Beatty, 1969; Miller et al., 1970). Also, by observing those electron micrographs, we can see that the RNAP molecules are almost distributed evenly on the template strand. This indicates that the termination rate should not be lower than the initiation rate; otherwise, RNAPs would accumulate at the end of the strand. We thus get $k_1 \leq k_4$.

3. Stochastic kinetics of the transcription model

3.1. Chemical master equation

Consider a reaction system of m molecular species that chemically interact through r reaction channels. The state of this system can be specified by the numbers of each of the m species, i.e. by the vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, where x_i is the number of molecules of type i. The state vector $\mathbf{x}$ is a discrete random variable. There exists for each reaction channel u, a function a_u called the propensity function such that $a_u \mathrm{d}t$ gives the probability of the corresponding reaction u occurring in time $\mathrm{d}t$. The propensity function has the form of $a_u = k_u h_u(\mathbf{x})$, where $h_u(\mathbf{x})$ is the number of combinations of reactant molecules which could participate in reaction u, and k_u is the reaction probability rate constant for reaction u.

The chemical master equation describes the time evolution of the probability distribution. Let $P(\mathbf{x}; t)$ be the probability that the system is in state $\mathbf{x}$ at any time t. The rate of change of $P(\mathbf{x}; t)$ is given by (Oppenheim et al., 1977)

$$\frac{\partial P(\mathbf{x};t)}{\partial t} = \sum_{u=1}^{r} [-P(\mathbf{x};t)k_u h_u(\mathbf{x}) + P(\mathbf{x} - \mathbf{v}_u; t)k_u h_u(\mathbf{x} - \mathbf{v}_u)]. \tag{6}$$

In this equation, $\mathbf{v}_u$ is the state-change vector, whose ith component is the change of x_i produced by one reaction event of type u.

To get the master equation of our transcription model, we firstly should know what species can determine the state of the system and how many reaction channels are concerned in the model.

Since we treat a model involving a single gene in a solution containing RNAP molecules, the total number of RNAP molecules N_p^0 is conserved, and is equal to $N_p^0 = N_p + \sum_i [\chi(i, \mathrm{A}) + \chi(i, \mathrm{O})]$, where N_p is the number of free polymerase molecules, and $\chi(i, \sigma)$ is an indicator function which takes the value 1 if site i is in state σ and 0 otherwise. We can therefore calculate N_p at any time from the total number of RNAP molecules and the state of the DNA strand. Thus, the model can consist of just the states of the DNA sites provided we also assume a large and effectively constant pool of NTPs. For a DNA template strand with $n + 1$ sites (one promoter site and n nucleotide sites, see Fig. 1), there are three possible states (U, O or A) for each site (except $1'$), so there are 3^n states. The state vector in this model is $(\chi(1, \mathrm{U}), \chi(1, \mathrm{O}), \chi(1, \mathrm{A}), \ldots, \chi(n, \mathrm{U}), \chi(n, \mathrm{O}), \chi(n, \mathrm{A}))$. Note that since the three states are mutually exclusive, only one of $\chi(i, \mathrm{U})$, $\chi(i, \mathrm{O})$ or $\chi(i, \mathrm{A})$ has unit value at any given time, the other two being null. As for the number of the reaction channels involved in the model, according to the elementary reactions proposed in the previous section, we have $2n + 1$ of them.

Suppose that we wanted to write down the chemical master equation for the full system. This will describe the evolution in time of the probability distribution $P(\chi(1, \mathrm{U}), \chi(1, \mathrm{O}), \chi(1, \mathrm{A}), \ldots, \chi(n, \mathrm{U}), \chi(n, \mathrm{O}), \chi(n, \mathrm{A}))$ for all possible combinations of the $\chi(i, \sigma)$, i.e. for all allowed states of the model. The problem in solving the chemical master equation is evident: There are 3^n states. Even for small values of n, this is a colossal number. Direct study of the chemical master equation is therefore not practical.

As a result of some of the simplifications used in our model, the functions $h_u(\mathbf{x})$ which appear in the master equation (6) have particularly simple forms. Consider for instance $h_0(1, 0, 0, \chi(2, \mathrm{U}), \chi(2, \mathrm{O}), \chi(2, \mathrm{A}), \ldots, \chi(n, \mathrm{U}), \chi(n, \mathrm{O}), \chi(n, \mathrm{A}))$ (associated with reaction (1), numbering the h_u like the rate constants). Since we treat the case where there is just one copy of the gene, the number of different combinations of the "reactants" P and U_1 corresponding to this state is just $h_0 = N_p$. For any state in which $\chi(1, \mathrm{U}) = 0$, $h_0 = 0$. The h_u for reactions (2)–(5) are even simpler. When the template strand is in the appropriate state to permit reaction, $h_u = 1$.

It is possible to calculate the transcriptional delay distribution from the chemical master equation for the case in which a single polymerase is active on the DNA template or, equivalently, cases where the polymerases are sufficiently widely spaced that they do not interfere with each other. Let τ_i be the time required for the RNAP to move from site i to site $i + 1$, i.e., the time it takes, given that site i is in state O at time t, for site $i + 1$ to reach state O. Each τ_i is a random variable with a distribution which we want to compute. The overall delay distribution is the distribution of the variable

$$\tau = \sum_{i=1}^{n} \tau_i. \tag{7}$$

Our first target is the single-jump probability density $\rho_i(\tau_i)$, a quantity such that $\rho_i(\tau_i)d\tau_i$ is the probability that the jump occurs between time τ_i and $\tau_i + d\tau_i$. We start out in state O at site i and treat $\{i+1, \mathrm{O}\}$, or equivalently $\{i, \mathrm{U}\}$, as an absorbing state for the purpose of this calculation. We will thus obtain conditional probabilities $p^{(i)}(\chi(i, \mathrm{O}), \chi(i, \mathrm{A}), \chi(i, \mathrm{U}))$ conditioned on the arrival of the single-polymerase molecule at site i at time zero. The governing equations for these conditional probabilities are obtained directly as special cases of the master equation (6) and are thus exact for the purpose of determining single-jump properties like $\rho_i(\tau_i)$. These equations are

$$\begin{aligned}
\mathrm{d}p^{(1)}(1, 0, 0)/\mathrm{d}t &= -k_1 p^{(1)}(1, 0, 0), \\
\mathrm{d}p^{(1)}(0, 1, 0)/\mathrm{d}t &= k_1 p^{(1)}(1, 0, 0) - k_2 p^{(1)}(0, 1, 0), \\
\mathrm{d}p^{(1)}(0, 0, 1)/\mathrm{d}t &= k_2 p^{(1)}(0, 1, 0),
\end{aligned} \tag{8}$$

for the first site;

$$\begin{aligned}
\mathrm{d}p^{(i)}(1, 0, 0)/\mathrm{d}t &= -k_3 p^{(i)}(1, 0, 0), \\
\mathrm{d}p^{(i)}(0, 1, 0)/\mathrm{d}t &= k_3 p^{(i)}(1, 0, 0) - k_2 p^{(i)}(0, 1, 0), \\
\mathrm{d}p^{(i)}(0, 0, 1)/\mathrm{d}t &= k_2 p^{(i)}(0, 1, 0),
\end{aligned} \tag{9}$$

for the next $n - 2$ sites ($i = 2$ to $n-1$); and

$$\begin{aligned}
\mathrm{d}p^{(n)}(1, 0, 0)/\mathrm{d}t &= -k_3 p^{(n)}(1, 0, 0), \\
\mathrm{d}p^{(n)}(0, 1, 0)/\mathrm{d}t &= k_3 p^{(n)}(1, 0, 0) - k_4 p^{(n)}(0, 1, 0), \\
\mathrm{d}p^{(n)}(0, 0, 1)/\mathrm{d}t &= k_4 p^{(n)}(0, 1, 0),
\end{aligned} \tag{10}$$

for the last site. The appropriate initial conditions for solving Eqs. (8)–(10) are $p^{(i)}(1,0,0) = 1$ and $p^{(i)}(0,1,0) = p^{(i)}(0,0,1) = 0$ for $i = 1, 2, \ldots, n$. The linear differential Eqs. (8)–(10) are easily solved to yield, at $t = \tau_i$,

$$\begin{aligned} p^{(1)}(0,0,1) &= [k_1(1-e^{-k_2\tau_1}) - k_2(1-e^{-k_1\tau_1})]/(k_1-k_2); \\ p^{(i)}(0,0,1) &= [k_3(1-e^{-k_2\tau_i}) - k_2(1-e^{-k_3\tau_i})]/(k_3-k_2), \\ & \qquad i = 2,3,\ldots,n-1; \\ p^{(n)}(0,0,1) &= [k_3(1-e^{-k_4\tau_n}) - k_4(1-e^{-k_3\tau_n})]/(k_3-k_4). \end{aligned} \tag{11}$$

These quantities are the cumulative probability distributions for the τ_i. Thus,

$$\begin{aligned} \rho_1(\tau_1) &= \mathrm{d}p^{(1)}(0,0,1)/\mathrm{d}\tau_1 = k_1k_2(e^{-k_2\tau_1} - e^{-k_1\tau_1})/(k_1-k_2); \\ \rho_i(\tau_i) &= \mathrm{d}p^{(i)}(0,0,1)/\mathrm{d}\tau_i = k_2k_3(e^{-k_2\tau_i} - e^{-k_3\tau_i})/(k_3-k_2), \qquad i = 2,3,\ldots,n-1; \\ \rho_n(\tau_n) &= \mathrm{d}p^{(n)}(0,0,1)/\mathrm{d}\tau_n = k_3k_4(e^{-k_4\tau_n} - e^{-k_3\tau_n})/(k_3-k_4). \end{aligned} \tag{12}$$

These equations were derived under the conditions $k_1 \neq k_2$, $k_2 \neq k_3$, and $k_3 \neq k_4$.

Now we want the distribution of the total delay defined by Eq. (7). Suppose that $\rho(\tau_1, \tau_2, \ldots, \tau_n)$ is the joint distribution of the variables τ_i. In this case, because the jumps are independently distributed, $\rho(\tau_1, \tau_2, \ldots, \tau_n) = \prod_{i=1}^{n} \rho_i(\tau_i)$. In terms of the probability addition law, the probability that a cumulated random variable has a particular value τ is the sum of the probabilities over all sequences of jumps such that $\tau_1 + \tau_2 + \ldots + \tau_n = \tau$. Therefore, we get (Feller, 1968)

$$\rho(\tau) = \int_{\sum \tau_i = \tau} \cdots \int \rho(\tau_1, \tau_2, \ldots, \tau_n)\mathrm{d}\tau_1\mathrm{d}\tau_2 \ldots \mathrm{d}\tau_n. \tag{13}$$

This is a multi-dimensional convolution. In order to get the integration, we can follow Schnitzer and Block's (1995) method of working with the Laplace transforms, $\tilde{\rho}(s) = \int_0^\infty e^{-s\tau}\rho(\tau)\mathrm{d}\tau$, $\tilde{\rho}_i(s) = \int_0^\infty e^{-s\tau_i}\rho_i(\tau_i)\mathrm{d}\tau_i$ for $i = 1, 2, \ldots, n$. The convolutions become multiplications in the Laplace transform domain (Butkov, 1968):

$$\tilde{\rho}(s) = \prod_{i=1}^{n} \tilde{\rho}_i(s). \tag{14}$$

Using Eqs. (12), we can get the Laplace transforms of $\rho_i(\tau_i)$ for $i = 1, 2, \ldots, n$, which have the following simple forms,

$$\begin{aligned} \tilde{\rho}_1(s) &= k_1k_2/(s+k_1)(s+k_2); \\ \tilde{\rho}_i(s) &= k_2k_3/(s+k_2)(s+k_3), \qquad i = 2,3,\ldots,n-1; \\ \tilde{\rho}_n(s) &= k_3k_4/(s+k_3)(s+k_4). \end{aligned} \tag{15}$$

Substituting them into Eq. (14), we get the Laplace transforms of $\rho(\tau)$,

$$\tilde{\rho}(s) = \frac{k_1(k_2k_3)^{n-1}k_4}{(s+k_1)(s+k_2)^{n-1}(s+k_3)^{n-1}(s+k_4)}. \tag{16}$$

Taking the inverse Laplace transform of Eq. (16), we should have obtained $\rho(\tau)$. However, it seems difficult to do this in general. We have obtained the inverse Laplace transform for $n = 2$, 3, and 4, finding that they have the following general form:

$$\rho(\tau) = \sum_{i=1}^{4} a_i e^{-k_i\tau} + \sum_{i=1}^{n-2} (b_i e^{-k_2\tau} + c_i e^{-k_3\tau})\tau^i, \tag{17}$$

where a_i, b_i, and c_i are constants determined by the number of sites as well as the four probability rate constants. We believe that the general form of the delay distribution is given by Eq. (17) even though we have not been able to prove this for $n > 4$. Note that there are $2n$ undetermined constants in Eq. (17) so that for larger values of n, it is not even feasible to recover these constants by fitting to experimental or simulation results.

Although we can not get the analytic form of the distribution except in some special cases, the mean value and standard deviation of the distribution can be easily obtained. Since the mth moment of $\rho(\tau)$ is,

$$\langle\tau^m\rangle = \int_0^\infty \tau^m\rho(\tau)\mathrm{d}\tau = (-1)^m \left.\frac{d^m\tilde{\rho}(s)}{\mathrm{d}s^m}\right|_{s=0}, \tag{18}$$

the mean value $\langle\tau\rangle$ and standard deviation $\Delta\tau = \sqrt{\langle\tau^2\rangle - \langle\tau\rangle^2}$ are, respectively,

$$\langle\tau\rangle = \alpha_0 + \alpha_1(n-1) \tag{19}$$

and

$$\Delta\tau = \sqrt{\beta_0 + \beta_1(n-1)}, \tag{20}$$

where $\alpha_0 = 1/k_1 + 1/k_4$, $\alpha_1 = 1/k_2 + 1/k_3$, $\beta_0 = 1/k_1^2 + 1/k_4^2$, and $\beta_1 = 1/k_2^2 + 1/k_3^2$.

Besides the mean values and standard deviations, the first few moments of the delay may be useful as well. The expressions for the higher moments are too unwieldy to be given here, but are easily computed from Eq. (18) using a symbolic algebra manipulator like Maple. With the first four moments, it should be possible to fit experimental data to recover the four rate constants. One attempt has already been made to measure transcriptional delays based on published DNA microarray data (Ota et al., 2003, based on the data of Spellman et al., 1998 and of Lee et al., 2002). There are two related problems with this technique, both related to the very great effort required to collect gene expression profiles: First, the sampling rate is too low to obtain more than rough estimates of the mean and standard deviation. Moreover, it is doubtful that sufficient data of this kind could be collected to measure very long delays (e.g. 16 h in the extreme case of the human dystrophin

gene; Tennyson et al., 1995). In principle, transcriptional delay distributions could be measured using similar experiments to those used to measure elongation rates (see, e.g., Adelman et al., 2002; Tolić-Nørrelykke et al., 2004). In fact, as we mention in Section 4.3, the delay and elongation rate distributions are related. It is however not clear whether the data from such experiments would be of sufficient quality to estimate four moments of the distribution. The best alternative would be to specifically design experiments to measure transcriptional delays. For instance, it should be possible to measure these delays under experimental conditions where the polymerase binds rapidly to the promoter sequence (e.g. large excess of polymerase), measuring the lag between mixing of the reagents and the appearance of RNA transcripts. Because these lags are relatively long (minutes to hours), routine kinetics techniques would be adequate. The main difficulty would be in devising means of detecting the complete transcript. One possibility would be to use a fluorescent probe designed to bind to the terminal section of the transcript.

As an alternative to analysis, the chemical master equation can also be solved numerically using Gillespie's stochastic simulation algorithm (Gillespie, 1976, 1977). This is a Monte Carlo method which generates trajectories consistent with the probability distribution specified by Eq. (6). A large ensemble of stochastic trajectories computed using this algorithm thus solves the chemical master equation, while individual trajectories allow us to see possible realizations of the underlying stochastic process.

3.2. Site-oriented Markov model

We derive here a nonlinear master equation for the model described in Section 2 assuming that the sites on the DNA template are statistically independent. In appropriate limits, this model should be equivalent to the chemical master equation, but involves only $2n$ independent variables.

Statistical independence of the sites is a strong condition. Even if there is just one polymerase molecule, the sequential occupation of the sites in this model means that the probabilities of occupation of various sites will not be statistically independent. We however expect the sites to be approximately statistically independent in at least one case, namely $k_2 \gg k_3$: In this case, the rate-limiting process is (4), a reaction which does not involve interaction between sites.

Because the site states U, O, and A are mutually exclusive, the probability of a strand state $(\chi(1,\mathrm{U}),\chi(1,\mathrm{O}),\chi(1,\mathrm{A}),\ldots,\chi(n,\mathrm{U}),\chi(n,\mathrm{O}),\chi(n,\mathrm{A}))$ is zero unless exactly one of the set $\{\chi(i,\mathrm{U}),\chi(i,\mathrm{O}),\chi(i,\mathrm{A})\}$ is unity for each $i \in (1,2,\ldots,n)$. Let $p_{i,\sigma}$ be the probability that site i is in state σ, the element of the set $\{\mathrm{U},\mathrm{O},\mathrm{A}\}$ such that $\chi(i,\sigma)=1$. If the sites are assumed to be independent, we should be able to write $P(\chi(1,\mathrm{U}),\chi(1,\mathrm{O}),\chi(1,\mathrm{A}),\ldots,\chi(n,\mathrm{U}),\chi(n,\mathrm{O}),\chi(n,\mathrm{A})) = \prod_{i=1}^{n} p_{i,\sigma}$. Under the independent-site assumption, it should thus be possible to use the $3n$ variables $p_{i,\sigma}$ instead of the 3^n strand state probabilities $P(\chi(1,\mathrm{U}),\chi(1,\mathrm{O}),\chi(1,\mathrm{A}),\ldots,\chi(n,\mathrm{U}),\chi(n,\mathrm{O}),\chi(n,\mathrm{A}))$. Furthermore, since the three states are mutually exclusive, we have $\sum_\sigma p_{i,\sigma}=1$. We therefore only need two independent variables per site. This makes the problem formulated in these variables tractable, even if n is large.

We can write down evolution equations for the $p_{i,\sigma}$ from basic statistical arguments. Let us start with the equation for $p_{1,\mathrm{O}}$. The rate of transitions to state O_1 from state U_1 given that there are exactly N_p polymerase molecules and that site 1 is unoccupied ($\chi(1, \mathrm{U}) = 1$) is $k_0 N_p \chi(1, \mathrm{U}) = k_0 N_p$. This last expression is a conditional state-to-state transition rate. The net rate of transitions to this state will therefore be given by the sum over all possible values of N_p of the products of the conditional rates $k_0 N_p$ by the joint probabilities that there are exactly N_p molecules of polymerase and that site 1 is unoccupied at time t, denoted by $P(N_p, \{1, \mathrm{U}\})$. The rate at which site 1 makes transitions from O to A is easier to write down, depending only on the state of site 1 at time t. This gives us the rate equation

$$\frac{\mathrm{d}p_{1,\mathrm{O}}}{\mathrm{d}t} = \sum_{N_P} k_0 N_p P\left(N_p, \{1, \mathrm{U}\}\right) - k_1 p_{1,\mathrm{O}}. \tag{21}$$

For $i > 1$, we reach state O_i from states A_{i-1} and U_i (Eq. 3) at a rate $k_2 \chi(i{-}1, \mathrm{A})\chi(i, \mathrm{U})$. The conditional rate thus reduces to k_2 which, multiplied by the probability of the precursor state, gives us the rate of change of $p_{i,\mathrm{O}}$. Thus, the rate equations for these variables are

$$\frac{\mathrm{d}p_{i,\mathrm{O}}}{\mathrm{d}t} = k_2 P\left(\{i-1, \mathrm{A}\}, \{i, \mathrm{U}\}\right) - k_3 p_{i,\mathrm{O}}. \tag{22}$$

Arguing similarly, the full set of rate equations is as follows:

$$\begin{aligned}
&\mathrm{d}p_{1,\mathrm{O}}/\mathrm{d}t = \sum_{N_p} k_0 N_p P(N_p, \{1, \mathrm{U}\}) - k_1 p_{1,\mathrm{O}};\\
&\mathrm{d}p_{1,A}/\mathrm{d}t = k_1 p_{1,\mathrm{O}} - k_2 P(\{1, \mathrm{A}\}, \{2, \mathrm{U}\});\\
&\mathrm{d}p_{i,O}/\mathrm{d}t = k_2 P(\{i-1, \mathrm{A}\}, \{i, \mathrm{U}\}) - k_3 p_{i,\mathrm{O}}, \qquad i = 2, 3, \ldots, n\\
&\mathrm{d}p_{i,A}/\mathrm{d}t = k_3 p_{i,O} - k_2 P(\{i, \mathrm{A}\}, \{i+1, \mathrm{U}\}), \qquad i = 2, 3, \ldots, n-1;\\
&\mathrm{d}p_{n,A}/\mathrm{d}t = k_3 p_{n,\mathrm{O}} - k_4 p_{n,A};\\
&\quad p_{i,U} = 1 - p_{i,O} - p_{i,A}, \qquad i = 1, 2, \ldots, n.
\end{aligned} \tag{23}$$

While these equations are rigorous, they do not constitute a closed set since they involve the joint probabilities $P\left(N_p, \{1, \mathrm{U}\}\right)$ and $P\left(\{i, \mathrm{A}\}, \{i+1, \mathrm{U}\}\right)$. If we assume a large equilibrated pool of RNAP so that N_p is essentially fixed, then we have

$$\begin{aligned}
\mathrm{d}p_{1,\mathrm{O}}/\mathrm{d}t &= \sum_{N_\mathrm{p}} k_0 N_p P(N_p, \{1, \mathrm{U}\}) - k_1 p_{1,\mathrm{O}}\\
&\approx k_0 \sum_{N_\mathrm{p}} N_p \delta_{N_p, N_p^{\mathrm{eq}}} p_{1,\mathrm{U}} - k_1 p_{1,\mathrm{O}}\\
&= k_0 N_p^{\mathrm{eq}} p_{1,\mathrm{U}} - k_1 p_{1,\mathrm{O}}\\
&= \tilde{k}_0 p_{1,\mathrm{U}} - k_1 p_{1,\mathrm{O}},
\end{aligned} \tag{24}$$

where $\delta_{N_p,N_p^{\mathrm{eq}}}$ is a Kronecker delta symbol, and $\tilde{k}_0 = k_0 N_p^{\mathrm{eq}}$. Moreover, the joint probability $P(\{i,\mathrm{A}\},\{i+1,\mathrm{U}\})$ is easy to evaluate according to the independent-site assumption: $P(\{i,\mathrm{A}\},\{i+1,\mathrm{U}\}) = p_{i,\mathrm{A}}\,p_{i+1,\mathrm{U}}$. We are now ready to write down the governing equations for our model, namely the following set of $2n$ ODEs with n added conservation relations:

$$
\begin{aligned}
\mathrm{d}p_{1,\mathrm{O}}/\mathrm{d}t &= \tilde{k}_0 p_{1,\mathrm{U}} - k_1 p_{1,\mathrm{O}};\\
\mathrm{d}p_{1,\mathrm{A}}/\mathrm{d}t &= k_1 p_{1,\mathrm{O}} - k_2 p_{1,\mathrm{A}} p_{2,\mathrm{U}};\\
\mathrm{d}p_{i,\mathrm{O}}/\mathrm{d}t &= k_2 p_{i-1,\mathrm{A}} p_{i,\mathrm{U}} - k_3 p_{i,\mathrm{O}}, \qquad i = 2,3,\ldots,n;\\
\mathrm{d}p_{i,\mathrm{A}}/\mathrm{d}t &= k_3 p_{i,\mathrm{O}} - k_2 p_{i,\mathrm{A}} p_{i+1,\mathrm{U}}, \qquad i = 2,3,\ldots,n-1;\\
\mathrm{d}p_{n,\mathrm{A}}/\mathrm{d}t &= k_3 p_{n,\mathrm{O}} - k_4 p_{n,\mathrm{A}};\\
p_{i,\mathrm{U}} &= 1 - p_{i,\mathrm{O}} - p_{i,\mathrm{A}}, \qquad i = 1,2,\ldots,n.
\end{aligned}
\tag{25}
$$

We refer to Eqs. (25) as the *site-oriented Markov model*. Models of this type are also sometimes known as nonlinear master equations in the chemical literature (see, e.g., Davis and Skodje, 2001). The transformation from the chemical master equation for the 3^n possible states to the site-oriented equations involving $2n$ differential equations has a cost: The chemical master equation (6) is linear, while the site-oriented model is not. The nonlinear terms in Eqs. (25) appear as a result of the factoring of joint probabilities as described above and are only correct insofar as the sites can be treated as being statistically independent. However, the tremendous reduction in the number of variables which we need to consider makes this transformation worthwhile whenever it is at least approximately valid.

Having obtained the site-oriented Markov model, we extract from it the single-jump probability density. We only show the equations for the general case ($i = 2,3,\ldots,n-1$). If we again consider a polymerase occupying site i at $t = 0$, and treat site $i+1$ as an absorbing state, then $p_{i+1,\mathrm{U}} = p_{i,\mathrm{O}} + p_{i,\mathrm{A}}$, since site $i+1$ is empty as long as site i is occupied. The counterparts of Eqs. (9) based on Eqs. (25) are then

$$
\begin{aligned}
l\mathrm{d}p_{i,\mathrm{O}}/\mathrm{d}t &= -k_3 p_{i,\mathrm{O}},\\
\mathrm{d}p_{i,\mathrm{A}}/\mathrm{d}t &= k_3 p_{i,\mathrm{O}} - k_2 p_{i,\mathrm{A}}\left(p_{i,\mathrm{O}} + p_{i,\mathrm{A}}\right),\\
p_{i+1,\mathrm{O}} &= 1 - \left(p_{i,\mathrm{O}} + p_{i,\mathrm{A}}\right).
\end{aligned}
\tag{26}
$$

(Note that we could also have written the last member of Eqs. (9) as a probability conservation equation as we did here.) The solution, derived using the symbolic algebra system Maple, can be written in the following form:

$$
p_{i+1,\mathrm{O}}(\tau_i) = 1 - \frac{\exp(-\gamma e^{-k_3\tau_i})}{e^{-\gamma} + \gamma\,\mathrm{Ei}(1,\gamma e^{-k_3\tau_i}) - \gamma\,\mathrm{Ei}(1,\gamma)},
\tag{27}
$$

where $\gamma = k_2/k_3$ and $\mathrm{Ei}(u, z)$ is an exponential integral (Abramowitz and Stegun, 1965). Differentiating this expression with respect to τ_i, we get, after some rearrangement,

$$\hat{\rho}_i(\tau_i) = k_2 \left(1 - p_{i+1,\mathrm{O}}(\tau_i)\right) \left(1 - p_{i+1,\mathrm{O}}(\tau_i) - e^{-k_3\tau_i}\right), \tag{28}$$

where the circumflex indicates that this is an approximation derived from Eqs. (25). Unfortunately, it does not seem to be possible to obtain an analytic expression for the distribution of the total delay τ from Eq. (28) as we did for the exact chemical master equation. Nevertheless, a comparison of Eqs. (12) and (28) is instructive as it will help bring out some of the consequences of the transformation to the nonlinear site-oriented Markov model (25).

In accord with our earlier physical argument, if $k_2 >> k_3$, then after a short transient, a quasi-steady-state is reached in which $dp^{(i)}(0, 1, 0)/\mathrm{d}t$ in Eqs. (9) and $\mathrm{d}p_{i,\mathrm{A}}/\mathrm{d}t$ in Eqs. (26) are approximately zero. Accordingly, the rate of transitions to the absorbing state is controlled by the first equation in each set, which are the same within a change of notation, and we find that both distributions are then nearly indistinguishable. If on the other hand $k_2 << k_3$, then $p_{i,\mathrm{O}} + p_{i,\mathrm{A}} \approx 1$ during the rise in $p_{i,A}$. Thus, the portion of the delay distribution up to the maximum is about the same in the two models. The tails do differ significantly however since they are governed by different kinetics. The greatest differences between the two distributions occur when $k_2 \approx k_3$. Figure 2 shows the two distributions in such a case. Note that the maximum occurs at slightly smaller values of τ_i and is of smaller amplitude in the site-oriented model. Moreover, $\hat{\rho}_i$ has a longer tail than ρ_i. Thus, we may expect some differences in results derived from the chemical master equation and from the site-oriented Markov model, particularly in this parameter regime. We will see later however that the differences in the distribution of the total delay (Eq. (7)) are less than might be expected from the differences in the single-jump delay distributions seen in Fig. 2.

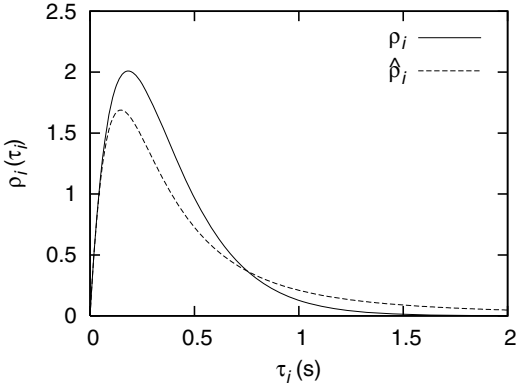

Fig. 2 Single-jump probability densities determined both from the exact master equation and from the approximate site-oriented Markov model for $k_2 = 6\,\mathrm{s}^{-1}$ and $k_3 = 5\,\mathrm{s}^{-1}$.

It is possible to obtain the distribution of the total delay for a single polymerase (or well-spaced polymerases) by integration of the site-oriented Markov model. With initial conditions

$$\begin{array}{lll} p_{1,\mathrm{O}}(0)=1, & p_{1,\mathrm{U}}(0)=p_{1,\mathrm{A}}(0)=0, & \text{and} \\ p_{i,\mathrm{U}}(0)=1, & p_{i,\mathrm{O}}(0)=p_{i,\mathrm{A}}(0)=0, & \text{for } i=2,3,\ldots,n, \end{array} \tag{29}$$

Eqs. (25) describe an ensemble such that each sample is followed from the moment of binding of an RNAP to the promoter site of the template strand. The transcriptional delay is the time that it takes for the RNAP to leave the DNA template strand. The term $k_4 p_{n,\mathrm{A}}$ in Eqs. (25) gives the rate at which the probability of completing the transcription process increases. Thus, the cumulative probability that transcription is complete at time t, $P_{\mathrm{complete,cum}}(t)$, is given by $P_{\mathrm{complete,cum}}(t) = \int_0^t k_4 p_{n,\mathrm{A}}(\tau)\mathrm{d}\tau$. Let $\rho(\tau)$ be the transcriptional delay probability density. Of course, it is also the case that $P_{\mathrm{complete,cum}}(t) = \int_0^t \rho(\tau)\mathrm{d}\tau$. Comparing these two equations implies that $\rho(\tau) = k_4 p_{n,\mathrm{A}}(\tau)$. This gives us a direct way to calculate the transcriptional delay probability distribution from a solution of Eqs. (25) with initial conditions given by Eqs. (29). In what follows, we refer to the calculation of the delay distribution from numerical solutions of the site-oriented Markov model as the *integration method*.

In the next section, we will test the limits of various treatments described earlier. We will then go on to study the stochastic kinetics of the transcription model in some specified cases using these tools.

4. Results and discussion

4.1. Validation of modeling methods

We begin by focusing on the single-RNAP case in order to establish the validity of the site-oriented Markov model. Figure 3 shows the delay distributions calculated from Eq. (17) and by integration of the site-oriented Markov model for $n = 4$, the largest value for which we have the parameters of the analytic distribution. By an argument analogous to that presented in the last section, if k_1 is much smaller than the other rate constants, particularly for small values of n, then the kinetics of the formation of the open-promoter complex is rate-limiting and the distributions will tend to be similar. For a value of k_1 which is a little larger than the high end of the estimates presented in Section 2 and using the smallest reasonable values of k_2, k_3, and k_4, there are noticeable, but not dramatic, differences between the two distributions (solid line and filled symbols in Fig. 3). Note in fact that the values of k_2 and k_3 used to generate these curves are identical to those used for the single-jump delay distributions shown in Fig. 2, the latter clearly differing more from each other than the overall delay distribution in Fig. 3. If we choose slightly larger values of k_2, k_3, and k_4, the two distributions become nearly indistinguishable (dashed curve and open symbols).

In order to validate the site-oriented Markov model for larger systems, we must resort to stochastic simulation. As shown in Fig. 4, results of stochastic simulation and of the integration method are perfectly overlapped. This confirms that in

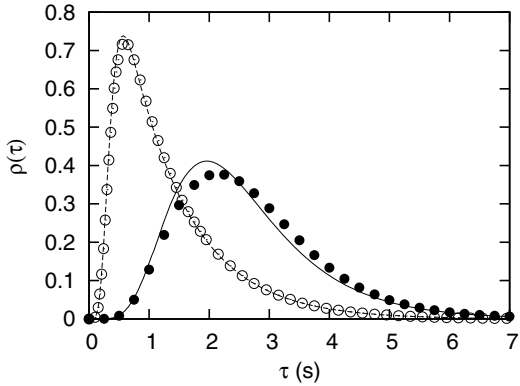

Fig. 3 Delay probability densities computed from Eq. (17) (*curves*) and by the integration method from the site-oriented Markov model (*symbols*). For all four sets of results, $n = 4$ and $k_1 = 1\,\mathrm{s}^{-1}$. Solid curve and filled symbols: $k_2 = 6\,\mathrm{s}^{-1}$, $k_3 = 5\,\mathrm{s}^{-1}$ and $k_4 = 2\,\mathrm{s}^{-1}$. Dashed curve and open symbols: $k_2 = 20\,\mathrm{s}^{-1}$, $k_3 = 19\,\mathrm{s}^{-1}$ and $k_4 = 21\,\mathrm{s}^{-1}$.

our studies of delay distributions the chemical master equation can generally be replaced by the site-oriented model without losing any important detail.

Since most of our theoretical development is based on a single polymerase moving along the template, we also examine the effect of allowing multiple RNAPs to function simultaneously. We use a DNA template strand with $n = 100$ as an example, corresponding to a very small protein containing about 30 amino acids. The probability rate constants are set as follows: $\tilde{k}_0 = 5\,\mathrm{s}^{-1}$, $k_1 = 1\,\mathrm{s}^{-1}$, $k_2 = k_3 = k_4 = 100\,\mathrm{s}^{-1}$. Note that the parameter $\tilde{k}_0$ is only required in the multiple-RNAP cases where it controls how often RNAPs bind to the promoter site, i.e. the spacing between the RNAPs. The average initiation rate is thus $(1/\tilde{k}_0 + 1/k_1)^{-1} = 0.83$ RNAPs/s, the average elongation rate is $(1/k_2 + 1/k_3)^{-1} = 50$ nt/s, and the

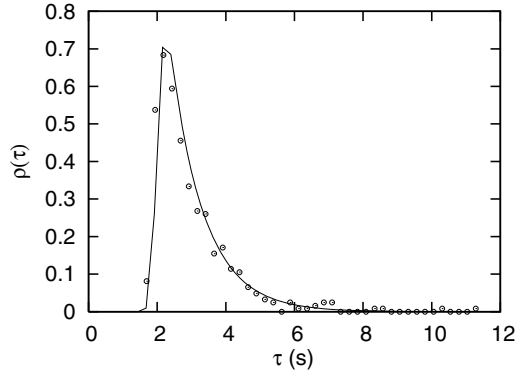

Fig. 4 Comparison of the delay probability distributions obtained by the stochastic simulation algorithm (*open circles*) and from the site-oriented Markov model using the integration method (*solid line*) for $n = 100$, $k_1 = 1\,\mathrm{s}^{-1}$, $k_2 = k_3 = k_4 = 100\,\mathrm{s}^{-1}$.

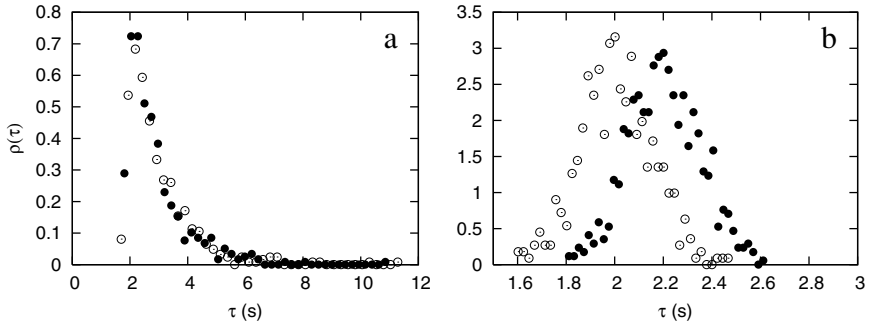

Fig. 5 Probability density of the transcriptional delay obtained by stochastic simulation. (a) $n = 100, \tilde{k}_0 = 5\,\mathrm{s}^{-1}, k_1 = 1\,\mathrm{s}^{-1}, k_2 = k_3 = k_4 = 100\,\mathrm{s}^{-1}$; (b) $n = 100, \tilde{k}_0 = 10\,\mathrm{s}^{-1}, k_1 = 100\,\mathrm{s}^{-1}, k_2 = k_3 = k_4 = 100\,\mathrm{s}^{-1}$. Filled circles come from the multiple-RNAP case with initial conditions of $\chi(i,\mathrm{U}) = 1$, $i = 1, 2, \ldots, n$, and open circles from the single-RNAP case with initial conditions of $\chi(1,\mathrm{O}) = 1$, $\chi(i,\mathrm{U}) = 1$, $i = 2, 3, \ldots, n$. Note that $\tilde{k}_0$ is only used in the multiple-RNAP simulations.

average termination rate is $(1/k_3 + 1/k_4)^{-1} = 50$ RNAPs/s. Note that we use a value of $\tilde{k}_0$ which is larger than the maximum estimated above. Most of our results are not very sensitive to this parameter because process (2) is rate-limiting for initiation. We get the probability distribution of the transcriptional delay by stochastic simulation using 506 samples with initial conditions $\chi(i,\mathrm{U}) = 1$, $i = 1, 2, \ldots, n$. The result is shown with filled circles in Fig. 5a for the multiple-RNAP case. As mentioned earlier, the transcriptional delay is measured from time of binding to completion of the process. Note that this means that the value of $\tilde{k}_0$ only affects the delay distribution through its effect on the spacing between the polymerases. For the corresponding single-RNAP case, we can use the initial conditions of $\chi(1,\mathrm{O}) = 1$, $\chi(i,\mathrm{U}) = 1$, $i = 2, 3, \ldots, n$, in the stochastic simulation algorithm. The obtained probability density from 500 samples is shown with open circles in Fig. 5a. We can see that the distributions for single-RNAP and multiple-RNAP cases agree well with each other. This confirms that the delay distribution in the multiple-RNAP case can be replaced by that of the corresponding single-RNAP case. On the contrary, in another comparison where the values of the probability rate constants are the same as in the first case except that $\tilde{k}_0 = 10\,\mathrm{s}^{-1}$ and $k_1 = 100\,\mathrm{s}^{-1}$, i.e., the initiation rate is much faster than that in the first case, the two distributions don't agree any more (see Fig. 5b), indicating the independent-site assumption is void. However, even in this case with parameters which are well outside physically reasonable ranges, the two distributions are not dramatically different. Thus, the analytic treatment based on the independent-site assumption should give results which are at least qualitatively correct over a wide range of parameters and reaction conditions.

After the above comparisons, we focus below on using the integration method to get the delay probability distributions from the site-oriented Markov model. Furthermore, since the average length of an mRNA molecule is about 1200 bases in *E. coli* (Stryer, 1988), we focus on DNA template strands of 1000–2000 bases.

4.2. Time delay distribution

In this part, we explore the delay distributions, focusing mainly on the single-RNAP distributions. However, as illustrated above, these distributions could be viewed as those for the corresponding multiple-RNAP cases if the independent-site assumption is valid. We vary, respectively, the number of sites n, the initiation probability rate constant k_1, the elongation probability rate constants, k_2 and k_3, and the termination probability rate constant k_4. In total, we looked at 13 cases, which are listed in Table 1. The mean values, standard deviations, and variation coefficients (the ratio of standard deviation to the mean value) reported in Table 1 were calculated both by the integration method, and using Eqs. (19) and (20). Note that the means and standard deviations computed by the two methods agree remarkably closely, the agreement being much better than the single-jump distributions in Fig. 2 would lead us to believe. This again supports the use of results from the site-oriented Markov model (25) to study this system.

We first vary n from 100 to 2000 (cases 1–4 in Table 1). As seen in Fig. 6, the mean of the delay distribution obviously increases with increasing n, as predicted by Eq. (19). The half width of the delay distribution increases more slowly with n, which is again in accord with the theoretical result (20). Consequently, the relative sharpness of the peak increases significantly from $n = 100$–2000. This can be seen by the change of the variation coefficients from 0.332 to 0.028. This means that the relative fluctuations of the transcriptional delay get increasingly weaker as the template strand length increases. It is worth stressing that the magnitude of relative fluctuations of delay indicated by the variation coefficient is an important indicator in stochastic kinetic studies of transcription: If it is small, the spread in transcript completion times is small compared to a typical completion time itself, and so the polymerase behaves in a highly regular, clock-like fashion. The variation coefficient can be calculated from Eqs. (19) and (20). We plot this coefficient

Table 1. Thirteen cases presented in Section 4.2 studying the transcriptional delay distribution, together with the corresponding mean values (MV), standard deviations (SD), and variation coefficients (VC) of the delay, which are calculated by the integration method

Case	n	k_1 (s^{-1})	k_2 (s^{-1})	k_3 (s^{-1})	k_4 (s^{-1})	MV (s)	SD (s)	VC
1	100	1	100	100	100	3.00 (2.99)	1.00 (1.01)	0.33
2	500	1	100	100	100	11.03 (10.99)	1.03 (1.05)	0.094
3	1000	1	100	100	100	21.06 (20.99)	1.07 (1.10)	0.051
4	2000	1	100	100	100	41.12 (40.99)	1.15 (1.18)	0.028
5	1000	0.5	100	100	100	22.03 (21.99)	2.03 (2.05)	0.092
6	1000	0.1	100	100	100	29.47 (29.99)	9.09 (10.01)	0.31
7	1000	1	100	100	50	21.08 (21.00)	1.07 (1.10)	0.051
8	1000	1	100	100	10	21.16 (21.08)	1.08 (1.10)	0.051
9	1000	1	100	100	1	22.06 (21.98)	1.47 (1.48)	0.066
10	1000	1	100	150	100	17.73 (17.66)	1.05 (1.07)	0.059
11	1000	1	150	100	100	17.70 (17.66)	1.05 (1.07)	0.059
12	1000	1	150	150	100	14.37 (14.33)	1.03 (1.04)	0.072
13	1000	1	200	200	100	11.02 (11.00)	1.02 (1.02)	0.092

Note. The MVs and SDs in parentheses are calculated by Eqs. (19) and (20), respectively.

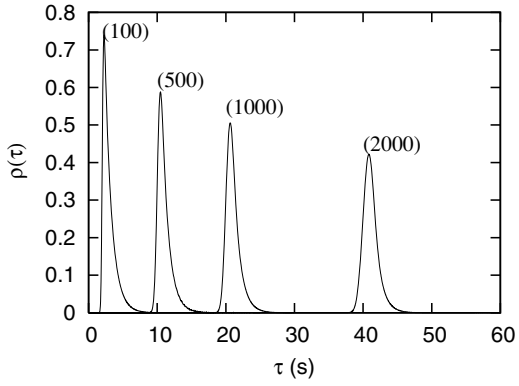

Fig. 6 Delay probability distributions for four cases with $k_1 = 1\,\mathrm{s}^{-1}$, $k_2 = k_3 = k_4 = 100\,\mathrm{s}^{-1}$ with different number of sites. The labels (*i*) in the figure indicate the number of sites in each case.

in Fig. 7. Note the steep drop in this quantity as n increases toward the range of typical gene sizes.

Next, we modify the initiation rate by varying k_1, which involves cases 3, 5 and 6 in Table 1. As shown in Fig. 8, the initiation rate mainly influences the standard deviation of the distribution, i.e., the lower the initiation rate, the larger the standard deviation. This means that the fluctuations of delay get increasingly larger with the decrease of the initiation rate. Increasing k_1 on the other hand only sharpens the distribution up to a limit. For the parameters of Fig. 8, we can calculate from Eq. (20) that $\Delta\tau \to 0.63\,\mathrm{s}$ as $k_1 \to \infty$, a value which may be compared to the standard deviation of 1.10 s calculated for $k_1 = 1\,\mathrm{s}^{-1}$. Even for templates of moderate length and at relatively large values of k_1, the formation of the open-promoter complex is thus responsible for a significant part of the stochastic fluctuations in the transcription time.

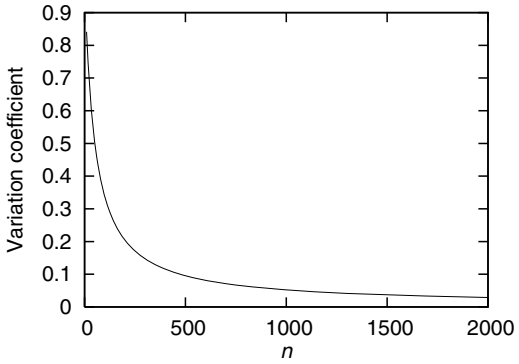

Fig. 7 The variation coefficient against the number of sites, n, calculated from Eqs. (19) and (20) with the parameters of Fig. 6.

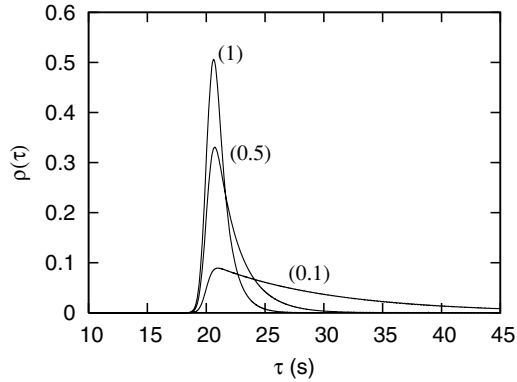

Fig. 8 Delay probability distributions for three cases with $n = 1000$, $k_2 = k_3 = k_4 = 100\,\mathrm{s}^{-1}$ with different initiation probability rate constants k_1. The quantities in parentheses are the values of k_1 for each curve.

The initiation rate controls the rate of input of RNAP to the template strand, while the termination rate controls the rate of exit of RNAP from the strand. We now study how the termination rate affects the delay distribution. We decrease k_4 from $100\,\mathrm{s}^{-1}$ to $1\,\mathrm{s}^{-1}$, giving four cases numbered 3 and 7–9 in Table 1. The corresponding delay probability distributions are shown in Fig. 9. It is interesting that the distribution does not change much as k_4 changes from 100 to $10\,\mathrm{s}^{-1}$, while the distribution for case 9 in which $k_4 = 1\,\mathrm{s}^{-1}$ is quite different from those for the other cases. Note that the curves in Fig. 9 are for the single-RNAP case. This sudden change is associated with an even more dramatic alteration of the distribution obtained from the multiple-RNAP simulations (symbols in Fig. 9). To see why, Fig. 10 illustrates three transcription snapshots, respectively, corresponding

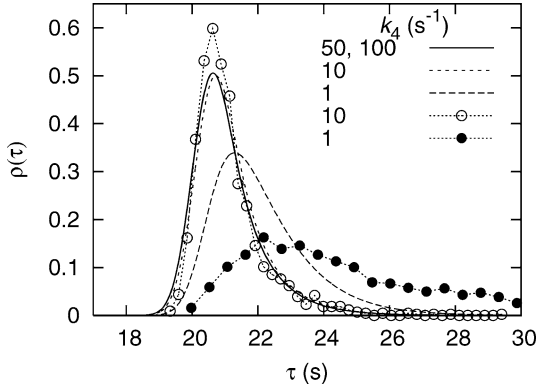

Fig. 9 Delay probability distributions for $n = 1000$, $k_1 = 1\,\mathrm{s}^{-1}$, $k_2 = k_3 = 100\,\mathrm{s}^{-1}$ and different termination probability rate constants k_4. The solid and dashed curves were computed using the integration method. The points connected by dotted curves were obtained by collecting statistics from stochastic simulations of multiple-RNAP simulations with $\tilde{k}_0 = 5\,\mathrm{s}^{-1}$.

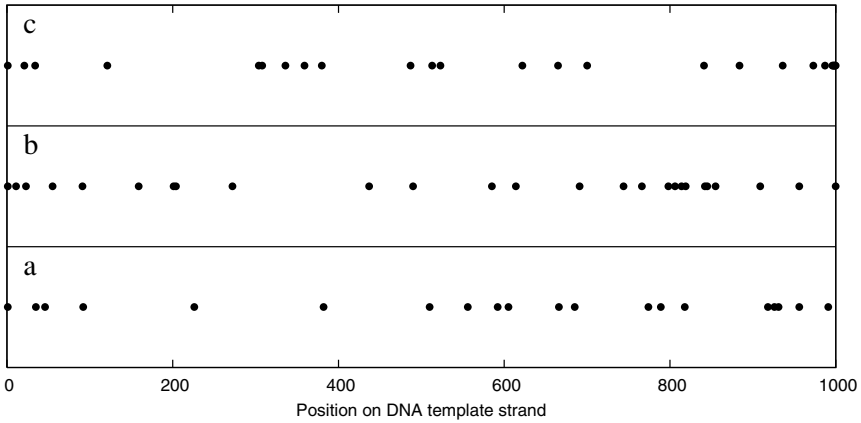

Fig. 10 Snapshots of the DNA template strand during transcription. (a), (b), and (c) correspond to the three cases $k_4 = 100\,\mathrm{s}^{-1}$, $10\,\mathrm{s}^{-1}$, and $1\,\mathrm{s}^{-1}$, respectively. The other parameters are set as $n = 1000$, $\tilde{k}_0 = 5\,\mathrm{s}^{-1}$, $k_1 = 1\,\mathrm{s}^{-1}$, $k_2 = k_3 = 100\,\mathrm{s}^{-1}$. The site is occupied by an RNAP if there is a black point on that site; otherwise it is empty.

to $k_4 = 100$, 10, and $1\,\mathrm{s}^{-1}$. For the lowest-termination rate case, we can see that some RNAP molecules accumulate at the end of the strand. This phenomenon is seldom observed at the higher values of k_4. We should note that for the case of $k_4 = 1\,\mathrm{s}^{-1}$, the independent-site assumption does not work any more because RNAPs often accumulate at the end of the strand. Thus, we should use stochastic simulation to get the correct delay distribution for these multiple-RNAP cases, as was done in Fig. 9. For $k_4 = 1\,\mathrm{s}^{-1}$, the multiple-RNAP distribution spreads much more than the single-RNAP distribution obtained from the site-oriented Markov model. The mean delay and standard deviation are, respectively, about 25.66 and 4.1 s, compared with 22.06 and 1.47 s for the distribution obtained from the site-oriented Markov model. Even at $k_4 = 10\,\mathrm{s}^{-1}$, the stochastic simulation and site-oriented Markov model treatments start to diverge.

This sudden change is analogous to phase transition in road traffic (Nagatani, 2002), if viewing RNAPs as vehicles. It is a rather simple case which concerns only single direction, one-dimensional road traffic. To see the phase transition more clearly, we choose an order parameter and observe its change against k_4. The order parameter here is defined as the average number of the last 10 template sites occupied by RNAP after a transient of 50 s. For the parameters of Fig. 9 except $n = 100$, we show in Fig. 11 the plot of the order parameter versus k_4. We see three different regimes. For $k_4 > 2\,\mathrm{s}^{-1}$, the order parameter is nearly zero and its fluctuation strength is low, indicating that RNAPs seldom accumulate at the end of the strand. We call this mode 1. For $2\,\mathrm{s}^{-1} > k_4 > 0.2\,\mathrm{s}^{-1}$, the order parameter suddenly changes from $\sim$0 to $\sim$10 and the fluctuations in the order parameter become very large, indicating that RNAPs now and then accumulate at the end of the strand. We call this mode 2. Finally for $k_4 < 0.2\,\mathrm{s}^{-1}$, the order parameter is fixed around 10 and the fluctuation strength abruptly drops, indicating that RNAPs finally occupy almost all the sites on the strand, leading to a real traffic jam. We call this mode 3. Traffic

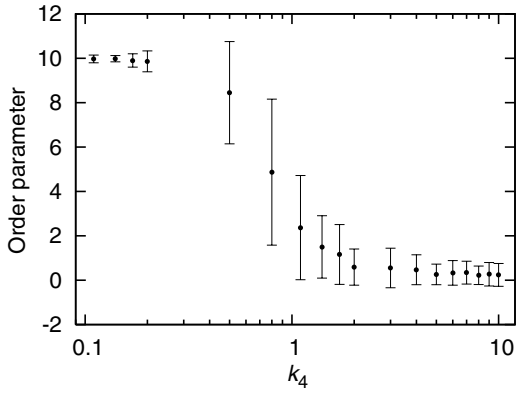

Fig. 11 Plot of the order parameter against k_4. The other parameters are $n = 100$, $\tilde{k}_0 = 5\,\mathrm{s}^{-1}$, $k_1 = 1\,\mathrm{s}^{-1}$, $k_2 = k_3 = 100\,\mathrm{s}^{-1}$. The order parameter is defined as the average number of the last 10 template sites occupied by RNAP 50 s after the initiation of transcription. For each value of k_4, 100 samples are used to get the order parameter and standard deviation, the latter shown as error bars. Note the logarithmic scale for the abscissa.

mode 2 is just the sudden phase transition between traffic modes 1 and 3. Note the large increase in the size of the fluctuations in the order parameter in mode 2, a property which is commonly observed near phase transitions (Stanley, 1971). Accordingly, we conclude that there exist two critical termination rates. Beyond the higher one, the delay distribution does not change much on varying the termination rate (mode 1). However, below it, RNAP molecules tend to accumulate at the end of the strand (mode 2), resulting in a broadening of the delay distribution and in an increase in the mean delay. As the termination rate continues to decrease to the second critical value, RNAPs will finally occupy almost all the sites (mode 3). Interestingly, it seems that RNAPs seldom, if ever, accumulate at the end of the template strand in experimental observations (Miller and Beatty, 1969; Miller et al., 1970). Thus, this phase transition probably would not be observable under normal conditions. It might however be possible to observe it in vitro using antibodies of differing binding affinities for the sequence in the termination region of the template, thus slowing the effective termination rate.

Finally, we study the influence of elongation rate. There are two parameters, k_2 and k_3, associated to the elongation rate. According to the exact theory developed in Section 3.1, the delay distribution is symmetric with respect to interchange of these two constants. In Fig. 12, we verify that the site-oriented Markov model approximately reproduces this symmetry using cases 10: $k_2 = 100\,\mathrm{s}^{-1}$, $k_3 = 150\,\mathrm{s}^{-1}$; and 11: $k_2 = 150\,\mathrm{s}^{-1}$, $k_3 = 100\,\mathrm{s}^{-1}$. The delay distributions are almost the same, further validating the use of the site-oriented model in this parameter range. Now, we increase both k_2 and k_3 from 100 to $200\,\mathrm{s}^{-1}$ (cases 3, 12, 13). As expected, the mean value of delay distribution decreases obviously with increasing elongation rate. The standard deviation doesn't change much, however. In addition, the variation coefficient, which measures the magnitude of the relative fluctuations, increases a little with increased elongation rate.

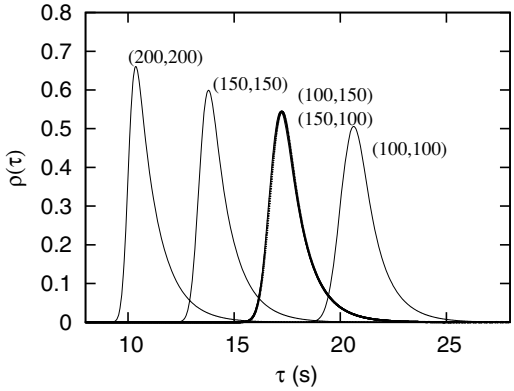

Fig. 12 Delay probability distributions for five cases with $n = 1000$, $k_1 = 1\,\mathrm{s}^{-1}$, $k_4 = 100\,\mathrm{s}^{-1}$ with different probability elongation rate constants k_2 and k_3 given for each curve as an ordered pair.

In summary, the magnitude of relative fluctuations of transcriptional delay obviously decreases as either n or k_1 increases, and increases a little with increasing k_2 and k_3, but is not much affected by k_4 as long as RNAPs do not accumulate at the end of the strand.

4.3. Transcript elongation rate

Two kinds of transcription elongation rates are measured in experimental studies: long-term and short-term average rates. The long-term average rate is defined as the ratio of the length of a mature transcript to its elongation time. The short-term average rate is calculated in this way: The length of a small part of the transcript is divided by the time it takes for the elongation of that part. The short-term average rate is also called instantaneous rate. In fact, the two kinds of average rates can have the same definition in our modeling: the ratio of the length of a transcript to its elongation time. The transcript could be either long or short. Modifying the length of a transcript is easily realized in our model by varying the number of sites, n. In what follows, we denote the elongation rate by ω, and define the average rate by $\omega = n/\tau$.

Given the relationship between the elongation rate and the transcriptional delay, according to standard statistical theory, we can obtain the probability density for the elongation rate, $\bar{\rho}(\omega)$, from the delay probability density, $\rho(\tau)$, by (DeGroot, 1975)

$$\bar{\rho}(\omega) = \rho\left(\tau(\omega)\right)\left|\frac{\mathrm{d}\tau}{\mathrm{d}\omega}\right| = \rho(n/\omega)\frac{n}{\omega^2}. \tag{30}$$

Thus we can use all the machinery developed to compute delay probability densities for the elongation rate distributions. In what follows, when there is no possibility of ambiguity, we denote the distribution of elongation rates simply by $\rho(\omega)$.

For the single-RNAP case studied above, we can convert the full transcription model to a transcription elongation model by setting $k_1 = k_3$ and $k_4 = k_2$, reducing the model to the two steps involved in elongation, namely activation and translocation. In this part, we use the converted elongation model to study the elongation process only, focusing especially on the probability distribution of elongation rate, $\rho(\omega)$. Since k_1 is now equal to k_3, the former cannot be small enough to guarantee the validity of the independent-site assumption. Thus, the elongation rate probability distributions from the site-oriented Markov model and stochastic simulation are not in agreement any more, as illustrated in Fig. 13b and d. However, they are not completely different. In particular, the shapes of the distributions, including the standard deviations, are quite similar, and there is only a small shift in the mean value. Therefore, we still use the site-oriented Markov model below to obtain analytic insight into the elongation rate distribution.

While we have been unable to derive equations analogous to Eqs. (19) and (20) for the elongation rate, we can obtain useful approximate equations from these relationships. We think of $\langle\tau\rangle$ as the measurement of a mean with uncertainty $\Delta\tau$. Using the usual rules for computing uncertainties of functionally related quantities, we obtain

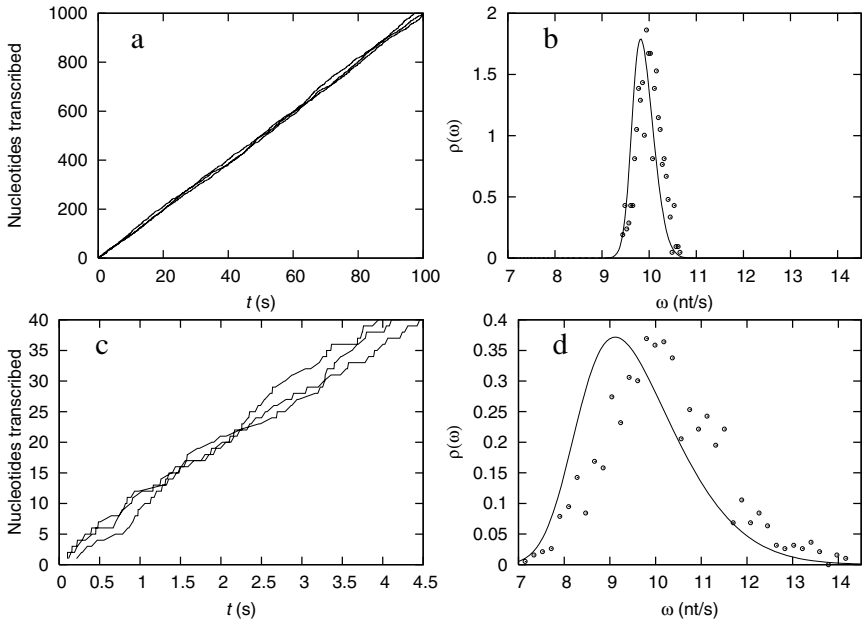

Fig. 13 Several stochastic simulation samples of transcript elongation for (a) $n = 1000$ and (c) 40, with probability rate constants $k_1 = k_2 = k_3 = k_4 = 20\,\mathrm{s}^{-1}$. The corresponding probability distributions of elongation rate are shown in (b) and (d), respectively. Solid lines were obtained from solutions of the site-oriented Markov model, and open circles by the stochastic simulation algorithm.

$$\omega = \frac{n}{\langle\tau\rangle}\left(1 \pm \frac{\Delta\tau}{\langle\tau\rangle}\right). \tag{31}$$

Since $k_1 = k_3$ and $k_4 = k_2$, Eqs. (19) and (20) reduce to $\langle\tau\rangle = \alpha_1 n$ and $\Delta\tau = \sqrt{\beta_1 n}$. Thus,

$$\langle\omega\rangle = n/\langle\tau\rangle = 1/\alpha_1 = (1/k_2 + 1/k_3)^{-1}, \tag{32}$$

and

$$\Delta\omega = n\Delta\tau/\langle\tau\rangle^2 = \frac{\beta_1^{1/2}}{\alpha_1^2} n^{-1/2} = \frac{k_2 k_3 \sqrt{k_2^2 + k_3^2}}{(k_2 + k_3)^2} n^{-1/2}. \tag{33}$$

(The estimate (32) can be shown to be accurate to the extent that the difference between $\langle 1/\tau\rangle$ and $1/\langle\tau\rangle$ is negligible.)

We studied seven cases this time, which are listed in Table 2, together with the corresponding mean values, standard deviations, and variation coefficients calculated by the integration method. For comparison, the mean rates and standard deviations calculated by Eqs. (32) and (33) are also given in Table 2. We first present cases 14 and 15, which have $n = 1000$ and 40, respectively. Comparing the rate probability distributions of the two cases, which are shown with solid lines in Fig. 13b and d, respectively, we can tell that the magnitude of the rate fluctuations decreases obviously with increasing template strand length. To compare the two cases further, we give Fig. 13a and c, respectively, showing several stochastic simulation samples of transcript elongation. The fluctuations of each sample for case 15 are relatively stronger than those of each sample for case 14. For the case 14 with $n = 1000$, each of those RNAP molecules almost move at a constant rate from beginning to end of the transcription. Additionally, according to the rate distribution, the mean value of their elongation rates is 9.88 nt/s and the standard deviation of the rate distribution is only 0.23 nt/s, indicating their transcripts experience almost the same characteristic elongation rate. This is consistent with Eq. (33), which predicts that the standard deviation of the distribution of elongation rates should decrease as n increases, as observed in the Figure.

Table 2. Seven cases presented in Section 4.3 studying the transcription elongation rate distribution, together with the corresponding MVs, SDs, and VCs, which are calculated by the integration method

Case	n	k_1 (s^{-1})	k_2 (s^{-1})	k_3 (s^{-1})	k_4 (s^{-1})	MV (nt/s)	SD (nt/s)	VC
14	1000	20	20	20	20	9.88 (10.00)	0.23 (0.22)	0.023
15	40	20	20	20	20	9.60 (10.00)	1.16 (1.12)	0.12
16	40	60	20	60	20	14.22 (15.00)	1.95 (1.88)	0.14
17	40	20	60	20	60	14.82 (15.00)	1.89 (1.88)	0.13
18	40	40	40	40	40	19.20 (20.00)	2.31 (2.24)	0.12
19	40	200	20	200	20	17.26 (18.18)	2.71 (2.63)	0.16
20	40	20	200	20	200	18.40 (18.18)	2.70 (2.63)	0.15

Note. The MVs and SDs in parentheses are calculated by Eqs. (32) and (33), respectively.

Briefly, in our single-molecule simulation experiments, every RNAP should move at an almost constant rate as long as they follow the same kinetics, i.e. have the same set of probability rate constants. Tolić-Nørrelykke et al. (2004) reported in their single-molecule experiments of *E. coli* transcription that those characteristic rates measured in a population of highly purified RNAP molecules are different for different molecules. They used a ∼2650-bp DNA template in the experiments. If the polymerases followed the same kinetics, the elongation rates for different polymerases should have been almost the same according to our model. Thus, the explanation of the experimental phenomenon is that each of the RNAP molecules follows different kinetics, indicating that the elongation rates are heterogeneous.

Interestingly, another study reports a quite paradoxical result, i.e., that the elongation kinetics of *E. coli* RNAP molecules is remarkably homogeneous (Adelman et al., 2002). We should note that the instantaneous rate was used in the latter study, i.e., a sample rate value is extracted from a 3 s averaging window of the elongation. The average elongation rate obtained was about 12.0 nt/s. So, the average number of nucleotides transcribed in those 3 s averaging windows is about 36. To model this situation, we use a DNA template strand with $n = 40$. We study 6 cases (cases 15–20), which have different values of the elongation probability rate constants. Their probability distributions of elongation rate are shown in Fig. 14. See cases 15–18 first. It is obvious that the standard deviation of the rate distribution increases with the increment of the mean rate. Note that the site-oriented Markov model artifactually predicts a slightly different dependence of the mean rate on k_2 and k_3 (cases 16 and 17). The availability of theoretical expressions such as Eq. (32) is revealed to be an important feature of our model, to avoid the over-interpretation of results derived from the site-oriented Markov model. To further study this issue, we also present two extreme cases 19 and 20, where k_2 is either much smaller or larger than k_3. We can see a significant difference in the mean rates, while the predicted standard deviations and variation coefficients are quite similar. Thus, even in cases where the site-oriented model makes

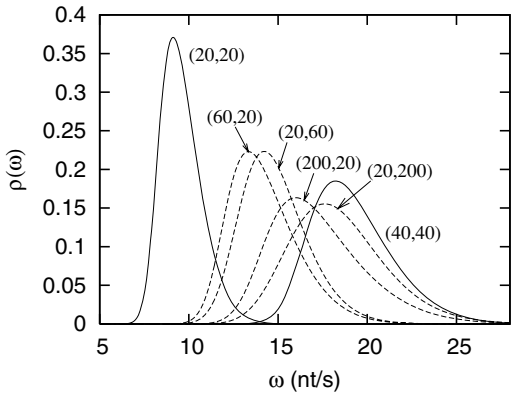

Fig. 14 Probability distributions of elongation rate for six cases with $n = 40$ with different probability elongation rate constants. The ordered pairs (i, j) indicate the values of $k_1 = k_3$ and of $k_2 = k_4$.

qualitatively incorrect predictions for the rates, it still predicts measures of dispersion rather accurately.

Adelman et al. (2002) also studied the statistics of elongation kinetics, and found substantial heterogeneity, which they ascribed to differences in the kinetics of pausing. We note that their Fig. 2 also suggests substantial differences in the shape of the distribution for different polymerases at high transcription rates, again supporting the interpretation that different polymerases have different kinetic parameters. Adelman et al. (2002) also show some typical time courses for transcription (Figs. 1B and 2A of their paper) for a DNA template with $n \approx 2000$. In Fig. 13a, we show three simulation samples for a template with $n \approx 1000$. Obviously, the three experimental elongation processes from the study of Adelman et al. (2002) differ from one another much more than the three simulations. It is doubtful that adding pausing to our model would result in the degree of apparent heterogeneity seen in the results of Adelman et al. (2002), although clearly this question requires further investigation. As seen in the analysis of the experimental results, pausing causes an added mode in the probability density peaking near zero. We tentatively conclude, in agreement with Tolić-Nørrelykke et al. (2004), that RNAPs display considerable population heterogeneity in their kinetic constants.

The experimental elongation rate distributions reported by Adelman et al. (2002) for individual polymerase molecules are much broader than ours. There could be many reasons for this difference. One factor which we neglected in our model and which would tend to substantially broaden the distribution of instantaneous rates is sequence dependence. Other factors such as extrinsic sources of noise may also contribute. However, the three distributions given in Fig. 2 of Adelman et al. (2002) do agree with our simulations and analytical results in one important way: As the mean rate increases, so does the standard deviation. (See Eqs. (32) and (33), and imagine increasing either k_2 or k_3.)

The above modeling indicates that in a population of RNAP molecules following the same kinetics, the strength of intrinsic fluctuations of elongation rate is so small that individual RNAP molecules behave in a highly regular fashion for larger numbers of base pairs of template DNA, say over 1000. Experimental observations of heterogeneity in elongation rates must then be due to differences in the probability rate constants between transcription complexes.

5. Conclusions

Using an independent-site assumption, we obtained a reduced site-oriented model whose predictions in many cases agree closely with results from the chemical master equation or equivalent stochastic simulations. This reduction was not without consequences. In particular, it destroyed the symmetry between k_2 and k_3 in several theoretical expressions. (See for instance the exact single-jump delay distribution (12) and the corresponding Eq. (28) derived from the site-oriented Markov model.) The differences between the two sets of results are typically small, but not altogether negligible. These differences arise from the assumption of statistical independence among the sites used to close the exact site-oriented master equation (23). In principle, more sophisticated closures could be used which would alleviate some of the minor inaccuracies in our treatment.

Most previous modeling work on transcription was based on relatively detailed physical modeling of the mechanics (Uptain et al., 1997; Jülicher and Bruinsma, 1998; Wang et al., 1998) or thermodynamics (Bai et al., 2004) of elongation. By contrast, our model treats transcription as a set of chemical rate processes with characteristic stochastic rate constants. This level of description is appropriate for understanding the results of experiments such as the transcript elongation rate measurements discussed above. Moreover, our model contains a relatively small number of parameters which could in principle be estimated by fitting model predictions to experimental data.

Our model omits many details of transcription. It does not explicitly consider regulatory factors involved in the transcription process. Although our model can account for short pauses (see Fig. 13c), no explicit mechanisms of blocking events such as pausing, backtracking, and arrest are included in modeling of the elongation process. The formation of an activated site is treated as a pseudo-first-order process, when it in fact obviously involves the arrival of NTPs. Also, in the elongation process, the probability rate constants for the two repeated events are independent of the sequence. This is an idealization since, e.g., certain sequences promote pausing (Davenport et al., 2000). In addition, we view the RNAP molecules moving on the DNA strand as points with no volume, i.e., they are connected to the strand at just one nucleotide. The fact that, for realistic parameters, the RNAPs are typically well separated makes this a reasonable starting point for modeling. However, there is no doubt that allowing RNAPs to occupy multiple nucleotide sites would make the effects of crowding on the template (e.g. the phase transition in Fig. 11) appear sooner.

It would not be difficult to obtain a model comparable to our site-oriented Markov model (25) which incorporated all of the factors mentioned above. For instance, steric hindrance could easily be added by allowing each polymerase to occupy more than one site. Using appropriate initiation rules, motion along the template could be represented by keeping track of the front and rear ends of the region occupied by the polymerase, as well as the location of the catalytic site:

$$\mathrm{A}_i + \mathrm{U}_{i+\delta+1} \to \mathrm{U}_{i-\eta} + \mathrm{O}_i + \mathrm{C}_{i+1} + \mathrm{O}_{i+d+1}. \tag{34}$$

At the beginning of this step, the polymerase occupies sites $i-\eta$ to $i+\delta$, so all sites in this range are marked as occupied (O), except for site i which is marked active (A) to indicate that the polymerase's active site is here and that the complex is ready to step forward. After the step, the states of the end points and of the former position of the active site have been updated to reflect the motion, and the new position of the catalytic site is marked with state C, the latter being a new feature of this extended model. While this model is no doubt a great deal more realistic, the opportunities to derive analytic expressions for the delay and elongation rate statistics are correspondingly reduced. The same is true of other possible model modifications. However, stochastic simulations carried out on modern high-speed computers will always work. Our current model, which allows at least some analysis, is thus an excellent base from which to determine the relative importance of one or another complicating factor.

We mentioned in the introduction that Lewis's (2003) and Monk's (2003) models can show delay-driven oscillations in the quantities of some expressed proteins. Both of these models involve time delays caused by transcription and translation. The above results show that the length of the DNA template strand strongly controls the magnitude of relative fluctuations of the transcriptional delay. Considering the transcriptional delay only, it would be of interest to investigate whether the oscillations can still be retained and how the delay fluctuations affect the oscillatory dynamics in a stochastic framework, especially for short template strands for which the relative magnitude of the delay fluctuations is greater.

Our results also indicate that the initiation rate is a highly significant factor in determining the magnitude of relative fluctuations in the transcriptional delay. Here, the initiation rate is controlled by k_1, i.e., the probability rate constant for the transition from closed- to open- promoter complex. At low initiation rates, the rate of RNA synthesis shows large fluctuations, the delay distribution sharpening substantially at high initiation rates (Fig. 8). Cells might vary the rhythm of gene expression by adjusting the initiation rate through feedback mechanisms. A similar effect has also been seen in a recent stochastic simulation study of the mammalian circadian clock (Forger and Peskin, 2005). This detailed biochemical model includes terms corresponding to transcription, translation, protein feedback, and dimerization. It does not explicitly treat RNAP binding to promoters and sliding along the DNA strand. It does however consider the activation time of genes, i.e., how long it takes for transcription to start, which corresponds to the inverse of the initiation rate in our studies. Interestingly, irregular rhythms are obtained when activation takes a long time, while regular and robust rhythms are obtained when activation becomes quick.

In order to study the distribution of transcription elongation rate, we used a transcription elongation model which was obtained directly from the full transcription model. By analyzing the calculated rate distributions, we concluded that in a population of RNAP molecules following the same kinetics, each of them almost moves at a constant rate if the DNA template strand is sufficiently long. In a population of highly purified RNAP molecules, with rare exceptions, individual RNAP molecules were observed to move at a constant characteristic rate on a $\sim$2650-bp DNA template (Tolić-Nørrelykke et al., 2004). This is consistent with our simulation result. Additionally, Tolić-Nørrelykke et al. (2004) found that the characteristic velocities are different for different molecules, displaying a broad, unimodal distribution across the molecular population. Though our model does not delve into the molecular details which would allow us to discuss why the RNAPs behave differently, it does provide the means to quantify how much of the widths of the observed distributions are due to stochastic kinetics, and thus how much must be due to population heterogeneity.

Some of the above results are obtained by analyzing the magnitude of relative fluctuations of delay, which is measured by the variation coefficient of the delay distribution. The magnitude of relative fluctuations is a significant indicator in stochastic kinetics studies. In the early single molecule experiments, Svoboda et al. (1994) studied the nanometer-sized steps taken by single molecules of the

motor protein kinesin, by studying the magnitude of relative fluctuations in the displacement of silica beads driven by them. Later, Schnitzer and Block (1995) extended the work of Svoboda and coworkers, and presented theoretical methods for analyzing processive enzyme behavior. This is referred to as fluctuation analysis, and is the basis for our analysis of the chemical master equation. Note that they referred to the magnitude of relative fluctuations as the randomness parameter, defined by $r = (< \tau^2 > - < \tau >^2)/< \tau >^2$, which is the square of variation coefficient of the distribution of τ. For a simple one-step elongation model, they showed that the value of r decreases with increasing length of the DNA template strand, which is consistent with the result obtained from our two-step elongation model.

A transcription model, perhaps with some added controls, is adequate to study the dynamics of functionally significant RNAs. However, most RNA transcripts are in fact messenger RNAs which are further translated to protein. Accordingly, a joint transcription-translation model would be of great interest. In outline, translation is not all that different from transcription. Indeed, models of translation are in many ways analogous to those we studied here, and many of the issues to be faced are the same (Drew, 2001; Heyd and Drew, 2003). We would therefore expect that the translation process itself would generate delay distributions similar to those seen in this study. The overall transcriptional-translational delay distribution would therefore be a convolution of two distributions, each with a characteristic shape. In eukaryotes, these two processes occur independently in separate compartments. A model for nuclear translocation of RNA from the nucleus and of regulatory proteins back into the nucleus will thus be required to treat this case. In prokaryotes, translation starts before transcription is complete. Accordingly, these two processes are not independent. There are of course other differences, such as the splicing of exons and other transcript processing events which occur exclusively in eukaryotes. It is not clear how these various factors affect the distributions of protein synthesis delays after the receipt of initiation signals. In modeling studies, we can not only compare the very different end results of the evolution of the genetic apparatuses of prokaryotes and eukaryotes, but we may also consider independently the various factors which affect the dynamics of transcription and translation in these organisms. Accordingly, we may be able to say something about the evolutionary pressures which act on these genetic systems, particularly as these relate to the management of the inevitable fluctuations which arise in the course of synthesizing RNA or proteins (McAdams and Arkin, 1999; Smolen et al., 1999, 2000; Elowitz et al., 2002; Hasty and Collins, 2002).

Most models which include transcription-translation delays use a single-fixed delay. As we have seen, for small peptides (many of which have regulatory roles) or at low initiation rates, the delay distribution is relatively broader (Figs. 7 and 8), and fixed delays may not be a very good approximation. It remains to be seen however whether the use of fixed rather than distributed delays causes much more than minor quantitative differences in models. When the overall delay is sufficiently small, it is known that even dramatic changes in the shape of the delay distribution have very little effect on the qualitative behavior (Roussel, 1996; Roussel and Roussel, 2001). Related observations in the specific context of genetic regulatory models have been made by Smolen et al. (2001) and by Monk (2003). However, genetic

regulatory delays are not obviously small relative to the other time scales governing cellular function, such that we are on uncertain ground. Future investigations of these issues should prove highly revealing.

Modeling studies which include distributed delays have often used gamma-distributed delays, since these are reducible to a linear chain, i.e. to an ODE model (Cooke and Grossman, 1982; MacDonald, 1989). We note briefly that Eq. (17) is *not* a gamma distribution. Indeed, we have attempted to fit gamma distributions to some of the data from our simulations. Clear deviations from a gamma distribution were seen. Following our comments above, it of course remains to be seen whether this makes any practical difference in simulations.

Finally, the heterogeneity in RNAP kinetics raises some difficult issues in the modeling of operons in particular. The kinetic parameters of a single polymerase could be treated as random variables. However, if a set of genes is transcribed serially by the same polymerase, their transcriptional delays will clearly be correlated. It would be worthwhile to engage in some simple stochastic simulation studies of a well-characterized operon in order to determine whether these correlations have any effect on model dynamics.

Acknowledgments

The authors thank Terry Tang for useful discussions. This work was supported by grants to R. Z. from an Alberta Ingenuity Fund Fellowship and to M. R. R. from the Natural Sciences and Engineering Research Council of Canada.

References

Abramowitz, M., Stegun, I.A., 1965. Handbook of Mathematical Functions. Dover, New York.

Adelman, K., La Porta, A., Santangelo, T.J., Lis, J.T., Roberts, J.W., Wang, M.D., 2002. Single molecule analysis of RNA polymerase elongation reveals uniform kinetic behavior. Proc. Natl. Acad. Sci. U.S.A. 99, 13538–13543.

Bai, L., Shundrovsky, A., Wang, M.D., 2004. Sequence-dependent kinetic model for transcription elongation by RNA polymerase. J. Mol. Biol. 344, 335–349.

Bliss, R.D., Painter, P.R., Marr, A.G., 1982. Role of feedback inhibition in stabilizing the classical operon. J. Theor. Biol. 97, 177–193.

Buchholtz, F., Schneider, F.W., 1987. Computer simulation of T3/T7 phage infection using lag times. Biophys. Chem. 26, 171–179.

Busenberg, S.N., Mahaffy, J.M., 1988. The effects of dimension and size for a compartmental model of repression. SIAM. J. Appl. Math. 48, 882–903.

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

Cooke, K.L., Grossman, Z., 1982. Discrete delay, distributed delay and stability switches. J. Math. Anal. Appl. 86, 592–627.

Davenport, R.J., Wuite, G.J.L., Landick, R., Bustamante, C., 2000. Single-molecule study of transcriptional pausing and arrest by *E. coli* RNA polymerase. Science 287, 2497–2500.

Davis, M.J., Skodje, R.T., 2001. Geometric approach to multiple-time-scale kinetics: A nonlinear master equation describing vibration-to-vibration relaxation. Z. Phys. Chem. 215, 233–252.

DeGroot, M.H., 1975. Probability and Statistics. Addison-Wesley, Reading, MA.

Drew, D.A., 2001. A mathematical model for prokaryotic protein synthesis. Bull. Math. Biol. 63, 329–351.

Elledge, S.J., 1996. Cell cycle checkpoints: Preventing an identity crisis. Science 274, 1664–1672.

Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S., 2002. Stochastic gene expression in a single cell. Science 297, 1183–1186.

Feller, W., 1968. An Introduction to Probability Theory and Its Applications, Vols. 1 and 2. Wiley, New York.
Forger, D.B., Peskin, C.S., 2005. Stochastic simulation of the mammalian circadian clock. Proc. Natl. Acad. Sci. U.S.A. 102, 321–324.
Gillespie, D.T., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comp. Phys. 22, 403–434.
Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81, 2340–2361.
Gillespie, D.T., 1992. A rigorous derivation of the chemical master equation. Physica A 188, 404–425.
Hasty, J., Collins, J.J., 2002. Translating the noise. Nat. Gen. 31, 13–14.
Heyd, A., Drew, D.A., 2003. A mathematical model for elongation of a peptide chain. Bull. Math. Biol. 65, 1095–1109.
Jülicher, F., Bruinsma, R., 1998. Motion of RNA polymerase along DNA: A stochastic model. Biophys. J. 74, 1169–1185.
Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Z. Bar-Joseph, Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A., 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298, 799–804.
Lewis, J., 2003. Autoinhibition with transcriptional delay: A simple mechanism for the zebrafish somitogenesis oscillator. Curr. Biol. 13, 1398–1408.
MacDonald, N., 1989. Biological Delay Systems: Linear Stability Theory. Cambridge University Press, Cambridge, UK.
Mahaffy, J.M., Jorgensen, D.A., Vanderheyden, R.L., 1992. Oscillations in a model of repression with external control. J. Math. Biol. 30, 669–691.
McAdams, H.H., Arkin, A., 1999. It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet. 15, 65–69.
McClure, W.R., 1980. Rate-limiting steps in RNA chain initiation. Proc. Natl. Acad. Sci. U.S.A. 77, 5634–5638.
Miller, O.L., Beatty, B.R., 1969. Portrait of a gene. J. Cell. Physiol. 74 (Suppl. 1), 225–232.
Miller, O.L., Beatty, B.R., Hamkalo, B.A., Thomas, C.A., 1970. Electron microscopic visualization of transcription. Cold Spring Harb. Symp. Quant. Biol. 35, 505–512.
Monk, N.A.M., 2003. Oscillatory expression of Hes1, p53, and NF-κB driven by transcriptional time delays. Curr. Biol. 13, 1409–1413.
Nagatani, T., 2002. The physics of traffic jams. Rep. Prog. Phys. 65, 1331–1386.
Oppenheim, I., Shuler, K.E., Weiss, G.H., 1977. Stochastic Processes in Chemical Physics: The Master Equation. MIT Press, Cambridge, MA.
Ota, K., Yamada, T., Yamanishi, Y., Goto, S., Kanehisa, M., 2003. Comprehensive analysis of delay in transcriptional regulation using expression profiles. Genome Inform. 14, 302–303.
Roussel, C.J., Roussel, M.R., 2001. Delay-differential equations and the model equivalence problem in chemical kinetics. Phys. Can. 57, 114–120.
Roussel, M.R., 1996. The use of delay differential equations in chemical kinetics. J. Phys. Chem. 100, 8323–8330.
Schnitzer, M.J., Block, S.M., 1995. Statistical kinetics of processive enzymes. Cold Spring Harb. Symp. Quant. Biol. 60, 793–802.
Smolen, P., Baxter, D.A., Byrne, J.H., 1998. Frequency selectivity, multistability, and oscillations emerge from models of genetic regulatory systems. Am. J. Physiol. 274, C531–C542.
Smolen, P., Baxter, D.A., Byrne, J.H., 1999. Effects of macromolecular transport and stochastic fluctuations on dynamics of genetic regulatory systems. Am. J. Physiol. 277, C777–C790.
Smolen, P., Baxter, D.A., Byrne, J.H., 2000. Modeling transcriptional control in gene networks—Methods, recent results, and future directions. Bull. Math. Biol. 62, 247–292.
Smolen, P., Baxter, D.A., Byrne, J.H., 2001. Modeling circadian oscillations with interlocking positive and negative feedback loops. J. Neurosci. 21, 6644–6656.
Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell 9, 3273–3297.
Stanley, H.E., 1971. Introduction to Phase Transitions and Critical Phenomena. Oxford University Press, New York.

Stryer, L., 1988. Biochemistry, 3rd ed. W. H. Freeman, New York.
Svoboda, K., Mitra, P.P., Block, S.M., 1994. Fluctuation analysis of motor protein movement and single enzyme kinetics. Proc. Natl. Acad. Sci. U.S.A. 91, 11782–11786.
Tennyson, C.N., Klamut, H.J., Worton, R.G., 1995. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. Nat. Genet. 9, 184–190.
Tolić-Nørrelykke, S.F., Engh, A.M., Landick, R., Gelles, J., 2004. Diversity in the rates of transcript elongation by single RNA polymerase molecules. J. Biol. Chem. 279, 3292–3299.
Uptain, S.M., Kane, C.M., Chamberlin, M.J., 1997. Basic mechanisms of transcript elongation and its regulation. Annu. Rev. Biochem. 66, 117–172.
von Hippel, P.H., 1998. An integrated model of the transcription complex in elongation, termination, and editing. Science 281, 660–665.
von Hippel, P.H., Pasman, Z., 2002. Reaction pathways in transcript elongation. Biophys. Chem. 101–102, 401–423.
Wang, H.-Y., Elston, T., Mogilner, A., Oster, G., 1998. Force generation in RNA polymerase. Biophys. J. 74, 1186–1202.