A Look-Ahead Model for the Elongation Dynamics of Transcription

Yujiro Richard Yamada[†*] and Charles S. Peskin[‡]
[†]Department of Mathematics, University of Michigan, Ann Arbor, Michigan; and [‡]Courant Institute of Mathematical Sciences, New York University, New York, New York

ABSTRACT This article introduces a chemical kinetic model of the transcriptional elongation dynamics of RNA polymerase. The model's novel concept is a look-ahead feature, in which nucleotides bind reversibly to the DNA before being incorporated covalently into the nascent RNA chain. Analytical and computational methods for studying the behavior of the look-ahead model are introduced, and several approaches to parameter estimation are tested on synthetic and also on actual experimental data. Two types of experimental data are considered: 1), the mean velocity of RNA polymerase as a function of the ambient concentrations of the ribonucleoside triphosphates; and 2), the distribution of time intervals between the forward steps of RNA polymerase. By separately fitting the look-ahead model to these two types of data, we obtain estimates of the model parameters. The most difficult parameter to estimate is the width of the look-ahead window. Both types of data suggest a small window size, but the second type does a better job of distinguishing the different window sizes. These latter data rule out a window size of 1, and they strongly suggest a look-ahead window that is approximately four bases in width. Additional experiments to determine the window size are proposed.

INTRODUCTION

RNA polymerase is the key enzyme of transcription, the step at which most regulation of gene expression occurs. Transcription consists of three distinct processes: initiation, elongation, and termination. Of these processes, elongation has been until recently the least studied, but this situation has fortunately changed with the advent and extensive use of single-molecule force microscopy (1–8).

From a modeling perspective, elongation is the transcriptional step most amenable to a quantitative description. The motion of RNA polymerase during transcription can be viewed as a stochastic process, more specifically as a random walk along the DNA. The goal of modeling is to characterize this random walk. Previous models of this kind (1–8) have all been mechanical in nature, i.e., they have considered, in one way or another, the elastic forces that arise within the RNA polymerase molecule during transcriptional elongation.

In this article (see also preliminary reports (9) and (10)), we introduce a formal chemical kinetic model for the dynamics of the movement of RNA polymerase along DNA. In our proposed model, we focus on the discrete events of reversible binding and unbinding of nucleotides to the DNA, and on the covalent linkage of nucleotides into the nascent RNA chain. In this sense, our model is formal, because it only considers the stepwise motion of the RNA polymerase, not the physics of how that motion is generated. The model proposed herein is most easily visualized in terms of the power-stroke mechanism for the forward motion of RNA polymerase (11,12), since we assume that covalent linkage of nucleotide to the nascent RNA chain is synchronous (at least on the timescales of interest) with forward translocation of the RNA polymerase by one basepair along the DNA. Our model could also be consistent with a Brownian ratchet mechanism (13) in which covalent linkage of nucleotide to the nascent RNA chain locks in diffusive forward motion of the RNA polymerase, which is provided that the overall time elapsed during a forward move would be short in comparison to the time intervals among the chemical events of binding, unbinding, and covalent linkage. We are concerned here with a sequence of chemical events, not with the physical mechanism that propels the enzyme forward.

The emphasis of this article is on parameter estimation. We first describe a stochastic simulation method that can be used to generate synthetic data on which parameter estimation procedures can be tested, and then we discuss a master-equation analysis that yields noise-free predictions for comparison with experimental data during parameter fitting. Two sets of published experimental data are considered in this article as targets for parameter estimation, and additional experiments are proposed. The first set of published data is that of Adelman et al. (14). It involves measurements of the mean velocity of transcription as a function of the ambient concentrations of the four ribonucleoside triphosphates. Velocity histograms are also reported in this work. The second set of published data (15) employs fixed concentrations of the ribonucleoside triphosphates, which are chosen to be equally rate-limiting. These concentrations are also chosen to be much lower than the values that are typically used, thus slowing the process of transcription to the point that individual forward steps of RNA polymerase are easily resolved. Such an experiment reveals the statistical distribution of the time intervals between successive forward steps of RNA polymerase, and this is valuable information for parameter estimation.

Submitted February 28, 2008, and accepted for publication December 29, 2008.

*Correspondence: yryamada@umich.edu

Editor: Taekjip Ha.

0006-3495/09/04/3015/17 $2.00

doi: 10.1016/j.bpj.2008.12.3955

The most difficult parameter to estimate turns out to be the size of the look-ahead window. This is an integer parameter, denoted w, which is equal to the number of sites within the transcription bubble at which ribonucleoside triphosphates may be reversibly bound to the DNA template strand, before their covalent linkage to the nascent RNA chain. Although $w = 1$ may be regarded as a special case of the look-ahead model (as we do in this article), it should be kept in mind that only when $w > 1$ does the look-ahead model deserve its name, since it is only if $w > 1$ that there is any parallel processing of the ribonucleoside triphosphates, with selection of the correct base being done at several DNA template-strand sites simultaneously.

Our approach to the determination of the integer parameter w is simply to try different values of w and for each such value to fit the model to the experimental data by adjusting the rate constants of the model. We then compare the quality of the fit that can be achieved for each of the different hypothesized values of w. This is a fair comparison, since the model is formulated in such a way that the total number of parameters is independent of w.

When this fitting procedure is applied to the experimental data of Adelman et al. (14), the best fit to the mean transcription velocity as a function of the ribonucleoside triphosphate concentrations seems to be obtained with $w = 1$ or with $w = 2$, and the fit seems to become gradually worse as the window size increases from there. One might hope that the velocity histograms would help to choose between $w = 1$ and $w = 2$, but in fact these two cases predict nearly identical velocity histograms, both of which underestimate the spread in the experimental velocity histogram by roughly a factor of two (although this may well be explained by experimental variability not taken into account by the theory).

The fit of the model to the statistical distribution of the time intervals (waiting times) between successive forward moves, as reported in Abbodanzieri et al. (15), is much more successful at resolving the window size. Here, it turns out that there is a qualitative distinction between the predictions of the model with $w = 1$ and corresponding predictions with $w > 1$. Specifically, the predicted waiting time distribution in the case $w = 1$ is nonmonotonic: it rises to a peak and then decays. The waiting time distributions for $w > 1$ are monotone decreasing, as are the experimental data. An excellent fit is obtained for $w = 4$. We regard this as evidence in favor of the look-ahead hypothesis.

Additional experiments specifically designed to determine the window size are proposed, and the procedures for extracting the window size from the proposed experiments are tested on synthetic data.

THE MODEL

During elongation, the double-stranded DNA is locally melted by the RNA polymerase over a distance of ~14–17 basepairs. This locally melted region is known as the transcription bubble. Within the transcription bubble, one strand of the DNA acts as a template, upon which complementary ribonucleoside triphosphates (ATP, GTP, CTP, and UTP) can reversibly bind and unbind to/from the DNA template strand. It has been hypothesized, however, that only a part of the transcription bubble is actually used for transcription. The size of this window of activity within the transcription bubble formed by the RNA polymerase is an integer parameter of our model. The binding of ribonucleoside triphosphates within the window of activity is assumed to be reversible.

An irreversible reaction, however, is the incorporation of a nucleotide into the nascent RNA chain. This can occur only when that nucleotide is reversibly bound at the first site of the window of activity, i.e., the site at the 3′ end of the nascent RNA chain. When such incorporation of a nucleotide into the nascent RNA chain occurs, we assume that the RNA polymerase (and hence the transcription bubble and the window of activity) translocates forward one basepair. If the window of activity has a size of more than one basepair, it is quite likely that when the polymerase molecule, and hence the window, moves forward, it will already find the correct nucleotide bound at what has just become the site where that nucleotide can be incorporated into the growing RNA chain. This is the look-ahead feature of the model, a kind of parallel processing: placement of the correct ribonucleoside triphosphate at each site on the template strand of the DNA can occur before that site has been reached by the nascent RNA molecule.

The model is completely specified, then, by the following parameters:

- w is the length (in bases) of the look-ahead window.
- $(k_{on})_{ij}$ is the rate constant for reversible binding of ribonucleoside triphosphate of type i (ATP,CTP, GTP, or UTP) to deoxyribonucleotide of type j (A, C, G, T) in the template strand within the window of activity.
- $(k_{off})_{ij}$ is the rate constant for unbinding of reversibly bound ribonucleoside triphosphate of type i from deoxyribonucleotide of type j.
- $(k_f)_{ij}$ is the rate constant for covalent incorporation of nucleotide of type i into the nascent RNA chain, provided that there is a ribonucleoside triphosphate of type i reversibly bound to a deoxyribonucleotide of type j at the first site or the window of activity.

Note that we consider not only correct Watson-Crick basepairings, but also the possibility of errors. The parameter $(k_{on})_{ij}$ is of course, much larger, and $(k_{off})_{ij}$ much smaller, when (i,j) is a correct Watson-Crick basepair than otherwise. This mechanism protects against errors in transcription. Further error protection could be obtained by making $(k_f)_{ij}$ larger when (i,j) is a correct Watson-Crick basepair then when it is not. In our simulations, however, we have assumed that k_f is constant, independent of (i,j).

Fig. 1 shows the look-ahead window of RNA polymerase. Since the first site (*left end* of *box*, indicated by *vertical tick mark*) is unoccupied, the polymerase cannot move forward. Possible events are the unbinding of C, G, or U, or the binding of any ribonucleoside triphosphate (rNTP) to any of the five unoccupied sites. Fig. 2 (*top*) is the same as Fig. 1 except that the first site within the look-ahead window is also occupied. Possible events still include the unbinding of any of the reversibly bound rNTPs or the binding of any rNTPs (including incorrect Watson-Crick basepairing) to any of the unoccupied sites. In this case, however, there is an additional possible event because the first site is occupied, namely, the forward motion of RNA polymerase, as depicted by the arrow in the figure. Note, in particular, that after this motion the new first site in the window may again be occupied (as shown), leading to the possibility of another forward step as a subsequent event.

Simulation and analysis of the look-ahead model

A stochastic approach

One approach in studying the proposed model is to use stochastic computational methods. We model the movement of RNA polymerase along DNA using the Gillespie algorithm (16,17). For every possible transition, a suitable rate constant is assigned: for each unoccupied site within the window of activity, there are four binding rate constants, one for each of the ribonucleoside triphosphates that can possibly occupy that site. Note that if a site is occupied within the window of activity, then there is a rate constant for

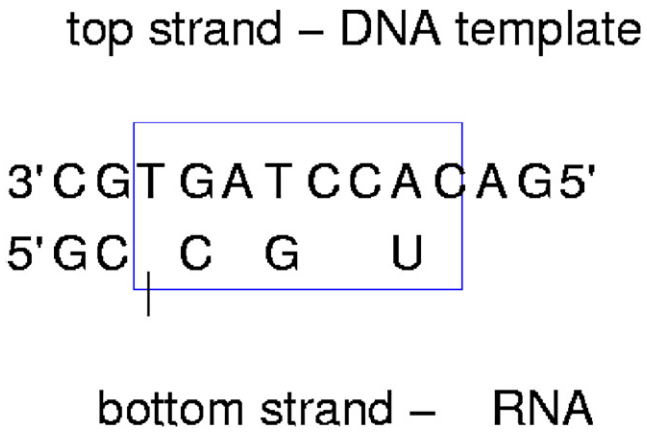

FIGURE 1 Schematic of the look-ahead model. Enclosed within the box is the window of activity. The top row of letters represents the DNA template strand, which is the strand complementary to the RNA molecule that is being synthesized. The nontemplate (coding) strand of DNA is not shown. The lower row, to the left of the window of activity, represents the nascent RNA strand. Within the window of activity, if a position is empty, a ribonucleoside triphosphate (rNTP) can bind reversibly at that position; and conversely if a position within the window of activity is occupied, then the rNTP at that position can dissociate, leaving that site of the window empty again. In its general form, the look-ahead model allows for incorrect (i.e., non-Watson-Crick) basepairing within the window of activity (for example, the *G* at the fourth position of the window), and also for the incorporation of incorrect bases into the nascent RNA chain (not shown here). The first position of the window of activity, known as the active center, is special and is indicated by a vertical mark in the figure. If that site is occupied, the rNTP that is located there can be covalently and irreversibly linked to the nascent RNA chain. When this happens, the whole RNA polymerase molecule moves one basepair forward along the DNA (see Fig. 2).

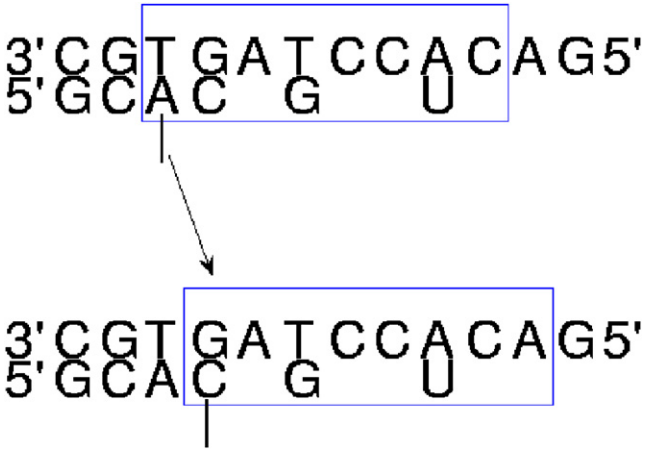

FIGURE 2 Forward motion of RNA polymerase in the look-ahead model. If the first site (active center, *vertical mark* in the figure) of the window of activity is occupied, the rNTP that is located there can be covalently and irreversibly linked to the nascent RNA chain. When this happens, the RNA polymerase simultaneously moves one basepair forward along the DNA. Thus, the whole window of activity moves one step to the right, as shown. In the example shown here, not only the first site but also the second site is occupied before the move. The result is that after the move the active center is again occupied, so another forward move can happen without waiting for the active center to fill. Thus, two (or more, depending on how many adjacent sites are filled starting from the active center) forward moves are likely to happen in rapid succession (but not simultaneously, since each is regarded as a separate step with its own exponentially distributed waiting time). When a forward move results in an empty active center, a longer delay is likely, since the active center has to fill before the next forward move can occur. In its most general form, the look-ahead model allows for the incorporation of incorrect bases (i.e., those that are not Watson-Crick basepaired with the corresponding bases on the DNA template strand) into the nascent RNA chain, although the rate constant for doing so is presumably smaller than that for correctly paired bases.

the ribonucleoside triphosphates on that site to dissociate, and if the first site within the look-ahead window is occupied, then there is a rate constant for the RNA polymerase to translocate forward one basepair along the DNA, incorporating the rNTP at the first window site into the nascent RNA chain while so doing.

The Gillespie algorithm jumps from event to event. Let $K = (k_1 + \ldots + k_m)$ be the sum of the individual reaction rates of those reactions that are possible given the current state, where each of the k_n values is selected from one of the $(k_{on})_{ij}$, $(k_{off})_{ij}$, and, $(k_f)_{ij}$ (if appropriate). Note that the number of possible reactions at any given time is given by $m = 4u + (w - u) + b$, where w is the window size, u is the number of unoccupied sites, and $b = 1$ if the first site is occupied and $b = 0$ otherwise. At each step, choose the time T to the next event from the probability density function

$$K\, e^{-Kt},$$

and then, independently of the above, choose which event occurred so that event j is chosen with probability,

$$\frac{k_j}{K}.$$

A master-equation formulation

Another approach to studying the look-ahead model is to formulate and solve the master equation that describes the time evolution of the probabilities of the different possible states of the model. Although the master equation describes an underlying stochastic process, the evolution of probabilities that it describes is deterministic, since these probabilities refer to a large ensemble of similar systems. Thus, the master-equation solution is noise-free, even though the underlying dynamics of the look-ahead model are stochastic. The same parameters that were used above when introducing the look-ahead model also appear in the master-equation formulation. We simplify the problem, however, by considering only correct Watson-Crick basepairing. Another simplification made here is that the DNA sequence is generated by a random process in which the choice of base at each location is made independently for the different locations on the DNA. Thus, we assume that the DNA sequence is fully characterized by the four base frequencies, whose sum must be one.

A master equation is a first-order differential equation describing the time-evolution of the probability of a system to occupy each one of a discrete set of states,

$$\frac{dP(l)}{dt} = \sum_{k:k\neq l} (P(k)R(k,l) - P(l)R(l,k)),$$

where $P(k)$, which is a function of time although we do not write that explicitly, is the probability that the system is in the state k at any particular time, and where $R(k, l)$, which in our case will be independent of time, is the probability per unit time that the system in state k will make a transition to state l. Once the master equation has been formulated, we study its steady state by setting each of the time derivatives $dP(l)/dt$ equal to zero, along with an additional constraint that the probabilities of all states add up to one.

The formulation of the master equation for the look-ahead model proceeds as follows.

Let w = window size. Possible states of the window are

$$\begin{pmatrix} a_1 & a_2 & \dots & a_{w-1} & a_w \\ b_1 & b_2 & \dots & b_{w-1} & b_w \end{pmatrix}, \tag{1}$$

where $a_i \in \{1, 2, 3, 4\}$ and $b_i \in \{0, 1\}$.

Here a_i indicates which DNA base on the template strand is located at site i within the window, and b_i indicates whether a complementary RNA base is present ($b_i = 1$) or absent ($b_i = 0$).

Possible reactions and corresponding rate constants are described below. For reversible binding events, we have the set of reactions

$$\begin{pmatrix} a \\ b \end{pmatrix} \xrightarrow{(b_i = 0)k_{on}(a_i)} \begin{pmatrix} a \\ b + \delta^i \end{pmatrix}, \tag{2}$$

where $i = 1,\dots,w$ and $k_{on}(a_i)$ is the probability per unit time of binding an rNTP to site i of the window of activity when base a_i is present at the corresponding site on the DNA template strand, given that site i is currently empty, i.e., that it does not currently have an rNTP bound. The notation $(b_i = 0)$ is a Boolean expression that evaluates to 1 when it is true and 0 when it is false, and similarly for other such expressions that appear below. Recall that the values of b_i are 1 or 0, depending on whether site i is occupied by an rNTP or not. The factor $(b_i = 0)$ in the probability per unit time for filling site i therefore makes that probability per unit time equal to zero if site i is already filled. The notation δ^i represents a vector of length w with 1 in the i^{th} position and all other elements equal to zero, so that

$$\delta^i_j = \begin{cases} 1 \text{, if } i = j \\ 0 \text{, otherwise} \end{cases}. \tag{3}$$

Thus, if b denotes a state in which site i is empty, $b + \delta^i$ denotes a state in which all sites other than i are the same as in state b, but site i is filled.

For unbinding events, we have

$$\begin{pmatrix} a \\ b \end{pmatrix} \xrightarrow{(b_i = 1)k_{off}(a_i)} \begin{pmatrix} a \\ b - \delta^i \end{pmatrix}, \tag{4}$$

where $k_{off}(a_i)$ is the probability per unit time of the unbinding of an rNTP from site i of the window of activity, given that the base a_i is present at the corresponding site on the DNA template strand, and also that there is currently an rNTP (reversibly) bound at site i. The latter condition is enforced by the Boolean factor $(b_i = 1)$ in the unbinding rate.

If the first site of the window of activity is occupied, then we must also allow for the incorporation reaction in which the RNA base located in position 1 of the window is covalently incorporated into the nascent RNA chain; the window then shifts forward by one basepair along the DNA. Recall that, in our model formulation, covalent linkage and forward motion are simultaneous.

When the window steps forward (to the right in our notation), all the a_i and b_i values shift one step to the left relative to the window. In this shift, the values that were originally stored as a_1 and b_1 are discarded, and we have to decide what values to put in a_w and b_w. Immediately after the shift (forward movement of the RNA polymerase) it is clear that we should set $b_w = 0$, since there has not been time for an rNTP to bind to the newly created last site that has just been introduced into the window of activity. It is also clear that a_w should be set equal to the value that represents the base on the DNA template strand that has just been drawn into the window of activity. Recall the assumption, stated above, which we make in this section, that the DNA sequence is random, with bases drawn independently from specified base frequencies for the DNA template strand. Let the probability of choosing base j for any particular position be $\alpha(j)$, where $j = 1, 2, 3, 4$, $\alpha(j) > 0$, and $\sum_{j=1}^{4} \alpha(j) = 1$. Then, immediately after the shift, we may set $a_w = j$ with probability $\alpha(j)$.

It is now clear that the possible reactions and corresponding probabilities per unit time associated with incorporation of a base into the nascent RNA chain, together with the associated forward movement of the RNA polymerase molecule, are

$$\begin{pmatrix} a \\ b \end{pmatrix} \xrightarrow{(b_i = 1)k_f\alpha(a'_w)} \begin{pmatrix} a_2 & \dots & a_w & a'_w \\ b_2 & \dots & b_w & 0 \end{pmatrix}, \tag{5}$$

where $a'_w = 1, 2, 3, 4$.

For a given starting state (a, b), there are, at most, w possible binding reactions (Eq. 2 with $i = 1, 2, \dots, w$); at most, w possible unbinding reactions (Eq. 4 with $i = 1, 2, \dots, w$); and at most, four possible incorporation/forward-stepping reactions (Eq. 5 with $a'_w = 1, 2, 3, 4$). In all three cases, only some of these possible reactions have nonzero rates, as indicated by the Boolean factors $(b_i = 0)$, $(b_i = 1)$, and $(b_1 = 1)$ in their rate constants (probabilities per unit time).

These reactions were written in terms of the state of origin. We also need to express them in terms of the destination state. In that case, the same reactions as above will appear but they, and their rates, will be expressed slightly differently. For reversible binding events, we have

$$\begin{pmatrix} a \\ b - \delta^i \end{pmatrix} \xrightarrow{(b_i = 1)k_{on}(a_i)} \begin{pmatrix} a \\ b \end{pmatrix}. \tag{6}$$

Note that the rate constant now has the factor $(b_i = 1)$, instead of $(b_i = 0)$. The reason is that b now refers to the destination state. For unbinding events, we have

$$\begin{pmatrix} a \\ b + \delta^i \end{pmatrix} \xrightarrow{(b_i = 0)k_{off}(a_i)} \begin{pmatrix} a \\ b \end{pmatrix}. \tag{7}$$

Finally, we have for the forward step of the RNA polymerase molecule,

$$\begin{pmatrix} a'_1 & a_1 & \dots & a_{w-2} & a_{w-1} \\ 1 & b_1 & \dots & b_{w-2} & b_{w-1} \end{pmatrix} \times \xrightarrow{(b_w = 0)k_f\alpha(a_w)} \begin{pmatrix} a_1 & a_2 & \dots & a_{w-1} & a_w \\ b_1 & b_2 & \dots & b_{w-1} & b_w \end{pmatrix}. \tag{8}$$

Note that the condition $(b_1 = 1)$ is no longer needed here, since that requirement is built into the origin state. It is replaced by $(b_w = 0)$, since the destination state cannot have anything bound to the last site in the window immediately after the forward move of the RNA polymerase.

The master equation may now be written as

$$\frac{d}{dt}P\begin{pmatrix} a \\ b \end{pmatrix} = \sum_{i=1}^{w}(b_i = 1)k_{on}(a_i)P\begin{pmatrix} a \\ b - \delta^i \end{pmatrix} - \sum_{i=1}^{w}(b_i = 0)k_{on}(a_i)P\begin{pmatrix} a \\ b \end{pmatrix} + \sum_{i=1}^{w}(b_i = 0)k_{off}(a_i)P\begin{pmatrix} a \\ b + \delta^i \end{pmatrix} - \sum_{i=1}^{w}(b_i = 1)k_{off}(a_i)P\begin{pmatrix} a \\ b \end{pmatrix} + (b_w = 0)k_f\alpha(a_w)\sum_{a'_1=1}^{4}P\begin{pmatrix} a'_1, a_1, \dots, a_{w-1} \\ 1, b_1, \dots, b_{w-1} \end{pmatrix} - (b_1 = 1)k_fP\begin{pmatrix} a \\ b \end{pmatrix}.$$

There is one such equation for each of the 8^w choices of $\begin{pmatrix} a \\ b \end{pmatrix}$. The steady-state equations are of course found by setting $\frac{d}{dt}P\begin{pmatrix} a \\ b \end{pmatrix} = 0$ and imposing the normalization

$$\sum_{\binom{a}{b}} P\binom{a}{b} = 1. \quad (9)$$

Once the steady-state equations have been solved, the mean forward velocity of the RNA polymerase in basepairs per second may be evaluated as

$$\bar{v} = \sum_{\binom{a}{b}} (b_1 = 1) k_f P\binom{a}{b}. \quad (10)$$

Note that $\bar{v}$ is just the product of k_f and the probability that $b_1 = 1$.

The master-equation and stochastic approaches are consistent

To verify that the stochastic simulation and the steady-state master-equation solution give the same mean velocity results, we consider a sequence of template DNA generated by the following simple stochastic process: each base is chosen independently with equal probabilities for the four possible outcomes. Note that the particular sequence chosen is only used in the stochastic simulation; the master equation only involves the base frequencies. We found that the only difference between the two results was the statistical error of the stochastic simulation, which can be reduced by increasing the length of the run. Such results are shown in Table 1.

If an actual DNA sequence is used in the stochastic simulation, then the best we can do to match it in the master-equation formulation is to input the four base frequencies from that DNA sequence. In this situation, we no longer expect perfect agreement in the computed mean velocities, even in the limit of infinitely long stochastic simulations, since the stochastic simulation result may depend on correlations in the given base sequence to which the master-equation formulation is blind. Our simulations found a small but persistent discrepancy between the mean velocity computed by the stochastic simulation when an actual DNA sequence was used and that predicted by the steady-state master-equation solution. Because the discrepancy is small, in practice, we can justify using the master-equation formulation for real DNA sequences by reflecting its base frequencies.

PARAMETER ESTIMATION

Interpretation of experimental data

We first discuss the type of experimental data that are shown in Fig. 2 of Bai et al. (7). In the experiments reported there, a particular rNTP concentration was varied (with the other three rNTP concentrations held constant at 1000 μM) to determine the influence of the varied rNTP concentration on the mean velocity of the RNA polymerase molecule. This was done for all four rNTP concentrations separately.

In our interpretation of these experimental data, we assume that the reversible binding of an rNTP to its complementary base on the template DNA strand is governed by the law of mass action. Thus,

$$(k_{on})_i = \left(k'_{on}\right)_i \left(\frac{[rNTP]_i}{[rNTP]_0}\right), \quad (11)$$

where $i = 1, 2, 3, 4$ specifies a particular ribonucleoside triphosphate and where $[rNTP]_i$ is the ambient concentration of that rNTP. In the above equation, $[rNTP]_0 = 1$ mM $= 1000$ μM is an arbitrarily chosen reference concentration that is introduced so that the units of $(k_{on})_i$ and $(k'_{on})_i$ are the same, namely s^{-1}. The particular value chosen for $[rNTP]_0$ has no significance at all.

It is important to note that the above mass action equation only holds for direct simple binding with no intervening binding events, such as the rNTP binding to another site in the RNA polymerase before binding to its complementary base on the DNA template strand. Note that $(k'_{on})_i$ is, by the mass action hypothesis made above, independent of concentration and is the actual parameter that we wish to find by comparing the model's results with the experimental data. No matter how many different combinations of rNTP concentrations were used in the experiment, there are only four distinct values of $(k'_{on})_i$. This type of experimental data is useful because each additional combination of rNTP concentrations enriches the data set without increasing the number of model parameters, provided that the mass action assumption is made as described above.

Model calibration to noise-free synthetic data

Before considering actual experimental data (for which the true parameters are unknown), we test our approach to parameter estimation by generating synthetic data for an arbitrarily chosen set of parameter values, to see whether those parameters can be recovered by fitting the model to the synthetic data. The synthetic data that we generate will be of the type discussed above, i.e., they will describe the mean velocity of transcription as a function of the different rNTP concentrations.

TABLE 1 Comparison of elongation velocities computed by stochastic Gillespie simulation with elongation velocities obtained by solving the master equation of the look-ahead model

Length of DNA strand	w=1	w=2	w=3	w=4	w=5
100 kbp	145.1926	214.1003	273.4290	323.7073	371.2490
200 kbp	144.8008	214.8499	271.0661	323.8148	369.3665
600 kbp	145.2393	214.9488	271.4428	323.2738	370.6913
1 Mbp	145.1523	214.7566	271.6649	322.9719	369.4076
2 Mbp	145.0068	214.6831	272.1620	322.9694	369.7292
Master equation	145.0777	214.7221	272.0295	322.9051	369.5776

Off-rates have been set equal to zero for this comparison. $(k_{on})_{ATP} = 100.0$, $(k_{on})_{CTP} = 150.0$, $(k_{on})_{GTP} = 200.0$, $(k_{on})_{UTP} = 250.0$, and $(k_f) = 2100.0$. These constants were arbitrarily chosen; any choice of constants will result in the same conclusion.

There are two fundamentally different ways that such synthetic data can be generated. One is to use the master-equation formulation, which generates noise-free synthetic data, and the other is to use stochastic simulation, which generates noisy data with a noise level that can be adjusted (as in an actual experiment) by varying the amount of data that is collected. These two kinds of synthetic data will be used in this subsection and in the next, respectively. In both cases, though, regardless of which method was used to generate the synthetic data, we use the master-equation formulation in the parameter fitting process itself.

To make the parameter fitting procedure more robust by reducing the dimension of the parameter space, we make certain a priori assumptions that reduce the number of unknown parameters. In this article, we only do parameter fitting under the following simplifying assumptions: First, only correct Watson-Crick basepairing is considered. Next, we assume that all of the off-rates are negligible, and that the forward rate is independent of which nucleotide is being incorporated into the nascent RNA chain. Finally, we treat the base frequencies of the DNA template strand as known parameters, since these can be independently measured in any particular case. With these assumptions, we have six unknown parameters to consider: the window size w, and the five rate constants $(k'_{on})_A$, $(k'_{on})_C$, $(k'_{on})_G$, $(k'_{on})_U$, and k_f. Of course the window size is restricted to positive integer values (and in practice we only consider the values 1, 2, 3, or 4), and the rate constants are not allowed to be negative. There are no other constraints.

The objective function that we seek to minimize during parameter estimation is simply the squares' sum of the differences of the computed mean velocities from the experimental mean velocities (which are synthetic data in this section and the next, but which then will be taken as the actual experimental data of (7)). The way that we deal with the discrete parameter w is simply exhaustive search, i.e., we do a separate minimization of the objective function for each value of w and see which gives the smallest value of the objective function (which will be called the residual in the following). For each fixed w, we use the nonlinear least-squares package of MATLAB (The MathWorks, Natick, MA) to do the minimization of the objective function with respect to the five rate constants listed above. To construct an initial guess we choose each of these rate constants randomly and independently from an exponential distribution.

Noise-free synthetic data were generated for window sizes 1, 2, 3, and 4, with rate constants chosen arbitrarily, and then the true parameters were forgotten, so to speak, and parameter fitting was done as described above to see whether the true parameters, including the window size, could be recovered. The results, shown in Table 2, indicate not only that a reasonable residual value can be returned, but also that the original set of parameters can indeed be reliably recovered.

Model calibration to stochastic synthetic data

In the previous subsection, we studied parameter estimation of a noise-free model to noise-free synthetic data. Here, we study parameter estimation of the same noise-free model in the context of stochastic synthetic data. The reason for doing this, of course, is that stochastic synthetic data are more representative of the kind of data that would actually be available from a real experiment. Our approach to parameter estimation here is exactly the same as in the previous subsection; the only difference is that stochastic simulations are used to generate the synthetic data. This introduces an additional consideration, however, which is the amount of data that is collected in any particular simulated experiment. As in real experiments, we regard each synthetic experiment as being comprised of some number of runs. Each run involves the synthesis of an RNA chain containing ~1800 bases. Recall that the output of interest is the mean velocity of transcription, which is obtained by averaging over all of the runs. Clearly, this mean velocity will be increasingly noise-free as the number of runs increases, and this should facilitate recovery of the true parameters. What we seek to determine, then, is the number of runs that will be needed for successful parameter recovery.

The results of the parameter estimation of the look-ahead model to stochastic synthetic data for different numbers of runs and for window size $w = 3$ can be found in Table 3. We observe that as the number of runs increases, the residual values get smaller, for the correct window size case. The residuals for the wrong window sizes are much larger than the residuals for the correct window size (see Table 4).

TABLE 2 Parameter estimation to non-noisy synthetic data

Window size	$(k_{on})'_{ATP}$	$(k_{on})'_{CTP}$	$(k_{on})'_{GTP}$	$(k_{on})'_{UTP}$	(k_f)	Residual value
1	34.8820	447.2032	19.7825	13.3369	75987.4229	5.1393
2	25.0187	250.1942	15.0130	10.0144	2026.0413	2.8637e-04
2	25.0	250.0	15.0	10.0	2500.0	0.0000
3	21.3233	197.3928	13.1974	8.9647	126.9686	0.3382
4	18.7992	167.8213	11.8758	8.1697	73.9755	0.9203

Parameter values obtained after fitting master-equation solutions to synthetic data (also generated by solving the master equation) are presented in this table. The version of the look-ahead model used here involved six parameters that are regarded as unknown during the fitting procedure: the window size, w; the four k'_{on} rates; and the k_f rate. In the case shown here, the actual window size is $w = 2$, and the rate constants used to generate the synthetic data are shown in the highlighted line of the table. Best-fit rate constants are shown for hypothesized window sizes $w = 1, 2, 3, 4$.

TABLE 3 Parameter estimation to stochastic synthetic data

Number of runs	$(k_{on})'_{ATP}$	$(k_{on})'_{CTP}$	$(k_{on})'_{GTP}$	$(k_{on})'_{UTP}$	(k_f)	Residual value
1	144.1399	292.3984	18.0066	258.0734	25.5338	2.4289
2	136.1948	276.5599	18.3828	19204.9089	25.1818	0.5733
5	150.2007	298.6737	18.6721	10505.9542	25.0187	0.6421
10	146.1004	281.6812	18.5434	538.3728	25.2058	0.2443
20	147.8141	285.7769	18.2752	628.0292	25.3852	0.0793
30	154.1701	291.5470	18.4290	28782.2366	25.2366	0.0514
100	156.3131	288.5127	18.5213	1122.4812	25.1661	0.0291
Actual	150.0	300.0	20.0	2000.0	25.0	0.0000

This table summarizes the estimated parameter values obtained by best fit to stochastic synthetic data when the assumed window size used in the parameter fitting matches the window size that was used to generate the synthetic data (in this case, $w = 3$). The number of runs indicates how many times an elongation experiment was performed to produce the synthetic data used in the parameter estimation.

The story is essentially the same (data not shown) when the true window size is different from 3.

The conclusion of these studies with stochastic synthetic data is that 30 runs (at each set of rNTP concentrations) suffice for the reliable recovery of the true parameters. This is a feasible number of runs for an actual experiment (see (14)).

Parameter estimation to experimental data

In the previous subsections, we calibrated our model to synthetic data; we concluded that the methodology outlined above for parameter estimation reasonably recovers the original (i.e., true) parameters. This was demonstrated both for noise-free synthetic data and also for noisy synthetic data generated by stochastic simulation. In the latter case, it was necessary to control the noise by doing sufficiently many runs (30 runs) to obtain each data point.

We now estimate parameters that give the best fit of the look-ahead model to the actual experimental data found in Bai et al. (7). As in the synthetic data case, the fit is based on mean velocity as a function of concentrations of the various rNTPs, and the master-equation formulation of the model is used in the parameter-estimation procedure. The results are summarized in Table 5. The magnitudes of the residuals indicate that the best window sizes are 1 and 2. These two best fits are visualized in Figs. 3 and 4.

TABLE 4 Residual values from parameter estimation to stochastic synthetic data

Number of runs	1-Bp window	2-Bp window	3-Bp window
1	3.6678	2.6918	2.4289
2	2.1468	0.7738	0.5733
5	2.1024	0.8715	0.6421
10	1.5291	0.3480	0.2443
20	1.8996	0.3734	0.0793
30	1.8153	0.3340	0.0514
100	1.7086	0.3042	0.0292

In this table, the window size $w = 3$ is used to generate the stochastic synthetic data, with varying numbers of runs. Parameter fitting is done for hypothesized window sizes $w = 1, 2, 3$ (of which only the last is correct). As the number of runs used to generate the stochastic synthetic data increases, the residuals of the parameter fit gets very small when the correct window size ($w = 3$) is used, but levels off at considerably larger values when an incorrect window size ($w = 1, 2$) is used in the parameter fitting.

As an additional check on the model, the estimated parameter values are used to generate velocity histograms, and these are compared to the corresponding velocity histograms that are found experimentally (see Fig. 5). One might hope that the velocity histograms would help distinguish between the window sizes 1 and 2, but this is not the case. Indeed the predicted velocity histograms for those two cases are virtually indistinguishable from each other, and have approximately half the width of the corresponding experimental histogram. Although this discrepancy may point to deficiencies in the look-ahead model (and in particular to the special case of the look-ahead model that was used in doing the parameter fitting), it is also possible that there are sources of noise in the experimental procedure and data collection that are not taken into account in our simulations.

Waiting time distribution

A more detailed approach to study the statistics of the motion of RNA polymerase is to analyze the distribution of the waiting times between successive base incorporations into the nascent RNA. In a recent publication (15), an experiment is described to measure this waiting time distribution under very low rNTP concentrations, concentrations which, besides being low, were chosen to be equally rate-limiting. (Note that the phrase ''equally rate-limiting'' is not intended to imply that the binding of rNTP is the rate-limiting step in the forward progress of RNA polymerase during transcription elongation. Instead, it refers to a condition in which the ambient concentrations of the different rNTP have been adjusted so that the mean time required for each DNA base to be transcribed is the same for all four of the DNA bases.) We now compare the results of these published experiments to the predictions of a special case of the look-ahead model: 1), only correct Watson-Crick basepairing is allowed; 2), all four binding rates k_{on} are equal, and all of the unbinding rates k_{off} are zero; and 3), the forward (incorporation) rate, k_f, which is relevant only when the first site of the look-ahead window is occupied, is the same, regardless of which base is being incorporated.

Note in particular the assumption that all four of the binding rates k_{on} are equal. Within the framework of the look-ahead

TABLE 5 Parameter estimation results to experimental data

Window size	$(k_{on})'_{ATP}$	$(k_{on})'_{CTP}$	$(k_{on})'_{GTP}$	$(k_{on})'_{UTP}$	(k_f)	Residual value
1	74.2694	480.7675	87.0047	53.6504	26.7532	6.303
2	56.6867	308.2255	64.7741	42.6405	21.6249	6.420
3	46.2350	239.3362	52.2671	35.2516	20.3795	7.764
4	39.4171	199.2069	44.2863	30.1923	19.8513	9.190
5	34.6143	172.2444	38.7318	26.5381	19.5634	10.507

Parameter values obtained fitting simulation results to actual experimental data are shown in this table.

model with negligible off-rates and a single forward rate, this is the parameter choice that realizes the condition used in the experiment that all four of the rNTP concentrations are equally rate-limiting. This is an important simplification, since it reduces the number of parameters that need to be determined, and even more so since it, together with the assumption that the forward (incorporation) rate is independent of which base is being incorporated, makes the statistics of the motion of RNA polymerase completely independent of the DNA sequence, thus simplifying the analysis of the model.

Under these simplifying assumptions, it is straightforward to show that the waiting time distribution of the look-ahead model is always of the form

$$\rho_T(t) = \frac{1-\theta}{1+\theta} k_f e^{-k_f t} + \frac{2\theta}{1+\theta} \frac{k_f k_{on}}{k_f - k_{on}} \left(e^{-k_{on} t} - e^{-k_f t}\right), \tag{12}$$

where $\rho_T(t)$ is the probability density for the time T of the next forward move after the forward move that occurred at $t = 0$, and where θ is a parameter that depends on the window size, w, in a manner that is detailed below for the particular cases $w = 1, 2, 3, 4$.

The explanation of this general formula for the waiting time distribution is very simple. Immediately after a forward move, the first site of the window of activity may be occupied or unoccupied. If it is occupied, then the time to wait until the next forward move is simply an exponentially distributed random

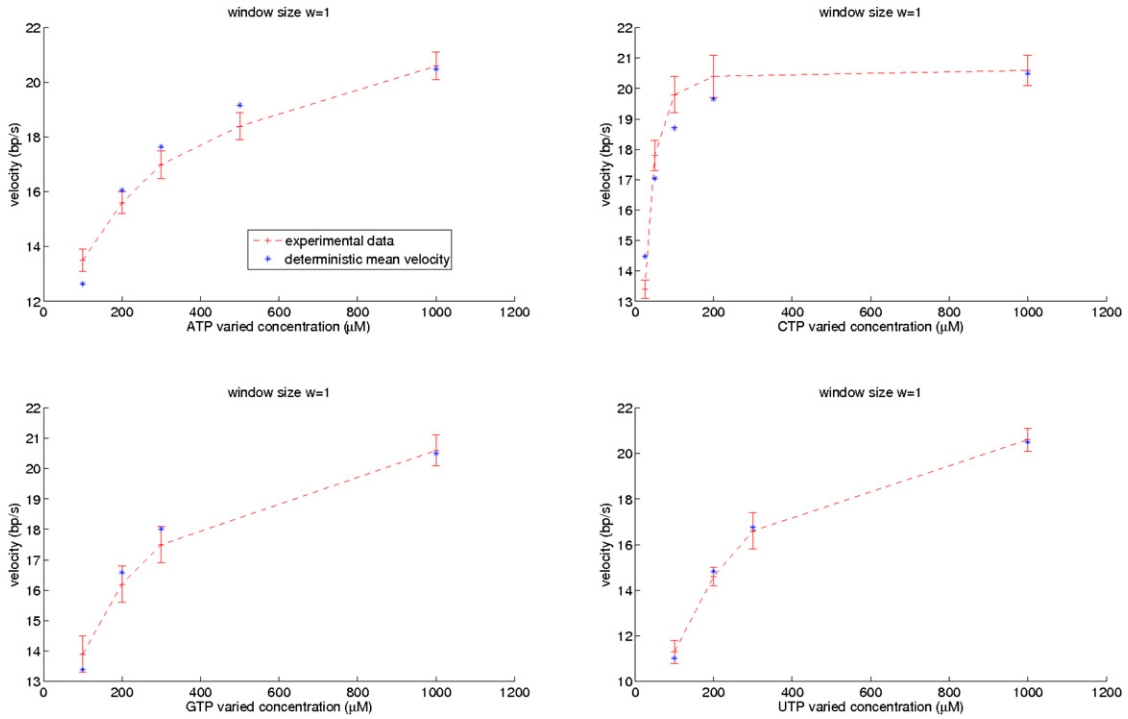

FIGURE 3 Fit of the look-ahead model with window size $w = 1$ to experimental data. The red error bars connected by dashed lines show the experimental velocity of RNA polymerase (in basepairs per second) as a function of rNTP concentrations, as reported in Bai et al. (7). In each of the four plots, one of the rNTP concentrations is varied while the others are held constant at 1000 μM. The form of the look-ahead model that is fit to these data allows only correct Watson-Crick basepairing, sets all k_{off} rates equal to zero, and assumes that k_f is independent of which base is being incorporated into the nascent RNA chain. Blue stars show the mean velocities of this version of the look-ahead model computed by solving the steady-state master equation with base frequencies of the DNA template strand chosen to match those of the template strand of the DNA tether used in the experiments, with window size $w = 1$, and with the five unknown rate constants k_f and $(k'_{on})_i$ of the model chosen to give best fit to the experimental data shown in the figure. The computed results fall within or very near the experimental error bars in all cases.

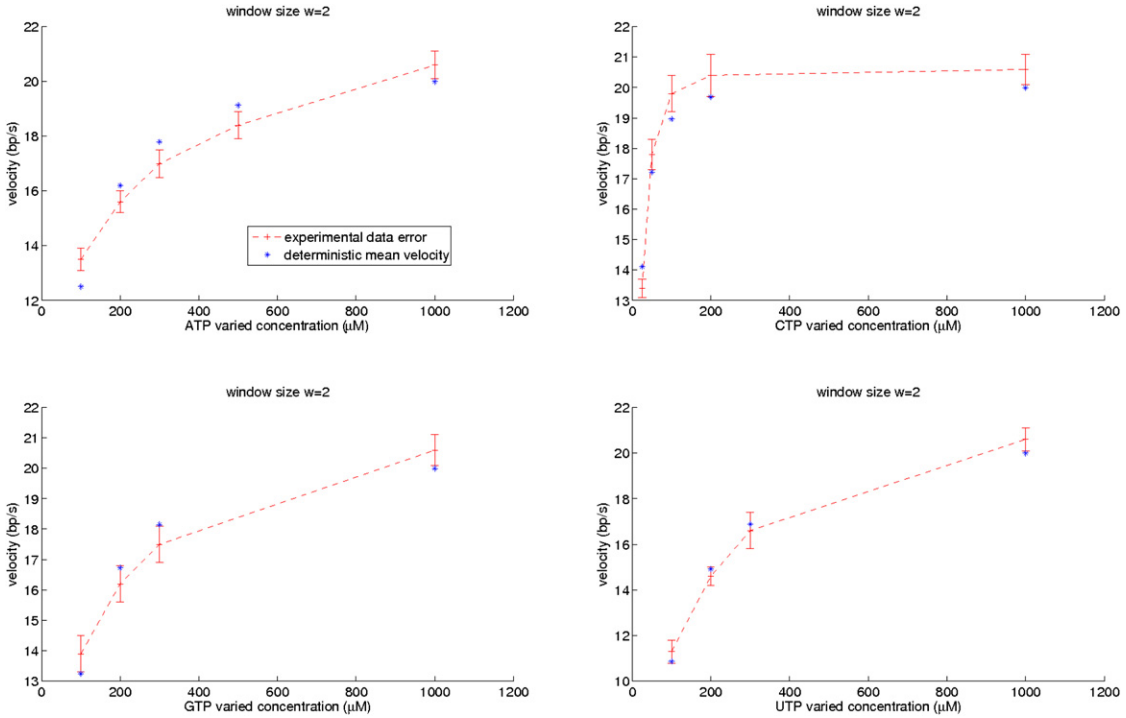

FIGURE 4 Fit of the look-ahead model with window size $w = 2$ to experimental data (7). See Fig. 3 legend. The only change here in comparison to that figure is that the window size $w = 2$ was used, and best-fit rate constants for that window size were found and used to obtain the computed velocities (*blue stars*). Even though the best-fit rate constants are different here from the ones found with $w = 1$, the quality of the fit is similar.

variable with mean $1/k_f$, as in the first term on the right-hand side of Eq. 12. In the opposite case, in which the first site is unoccupied immediately after a forward move, then the next forward move cannot occur until that site fills, an event that has probability per unit time k_{on}. In these circumstances, the waiting time until the next forward is the sum of two independent exponentially distributed random variables, the first with mean $1/k_{on}$ and the second with mean $1/k_f$. The probability density for the sum has the form of a difference of exponentials, as in the second term on the right-hand side of Eq. 12. The factor $(1 - \theta)/(1 + \theta)$ is the probability that the first site is occupied immediately after a forward move, and the factor $2\theta/(1 + \theta)$ is the probability that the first site is unoccupied immediately after a forward move. Note that these two factors add up to one. Different window sizes have different waiting time distributions only because these probabilities depend upon the window size. In particular, for window size 1 it is always the case that the first site is empty immediately after a forward move, so $\theta = 1$ when $w = 1$. Clearly, with k_{on} and k_f held constant, increasing the window size can only decrease the probability that the first site is empty immediately after a forward move, thus we expect that θ will decrease as the window size increases.

The problem of determining the value of θ as a function of the ratio $\gamma = k_{on}/k_f$ for any particular window size is a challenge for which the difficulty seems to grow rapidly with the window size. We have managed to solve this problem for $w = 1, 2, 3, 4$, and have verified the results by computer simulation. The formulae we have found are

$$\theta_{w=1} = 1$$

$$\theta_{w=2} = \frac{1}{2}\left(\frac{1}{1+\gamma}\right)$$

$$\theta_{w=3} = \frac{1}{2}\left(\frac{1}{1+\gamma}\right)^2\left(\frac{7+12\gamma}{10+12\gamma}\right)$$

$$\theta_{w=4} = \frac{1}{2}\left(\frac{1}{1+\gamma}\right)^3\left(\frac{1-\frac{1}{2}K\left(1+\frac{1}{18}\left(\frac{1}{1+\gamma}\right)\left(\frac{1}{1+2\gamma}\right)\right)}{1+\frac{1}{2}\left(\frac{1}{1+\gamma}\right)\left(\frac{1}{1+2\gamma}\right)\left(1-\frac{8}{9}K\right)}\right),$$

where $\gamma = k_{on}/k_f$, and where

$$K = \left(\frac{1+\gamma}{(1+2\gamma)^2}\right)\left(\frac{13+36\gamma}{21+36\gamma}\right). \tag{13}$$

We have fit the above formula for the waiting time distribution to the experimental data reported in Abbodanzieri et al. (15). For each window size separately, we have found the parameters k_{on} and k_f that give the best fit of the model to the data, in a least-squares sense. The data are reported in Abbodanzieri et al. (15) on a semilogarithmic plot; that is, the logarithm of the probability density is plotted against the waiting time, and we have done the fit with the data in that format as well (see Fig. 6). Since the logarithmic scale emphasizes rare events, however, we have also replotted (but not refit) both the data and the best-fit theoretical curves on an ordinary linear plot for comparison (see Fig. 7).

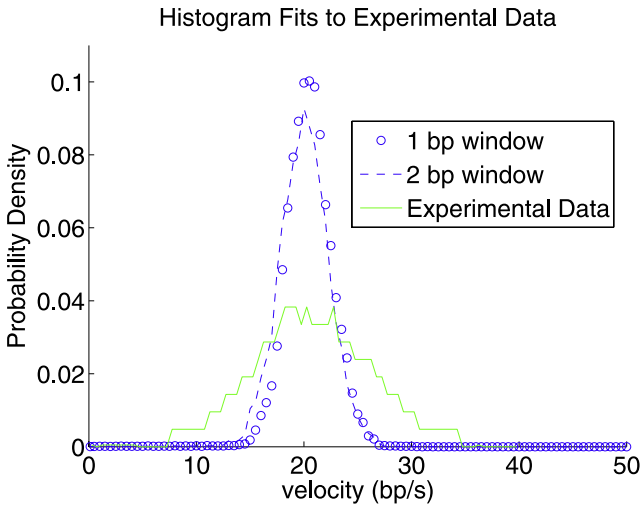

FIGURE 5 Comparison of computed and experimental velocity histograms. Since velocity histograms were not used in the parameter fitting, they provide an independent check on the validity of the model. Computed histograms obtained by stochastic simulation with the best-fit parameters found above for window sizes $w = 1$ and $w = 2$ (*blue open circles* and *blue dashed line*, respectively) are compared to the experimental velocity histogram (*green solid line*) from Bai et al. (7). The two window sizes (with best-fit rate constants determined separately in each case) give nearly identical results, which are narrower than the experimental histogram by roughly a factor of 2. This suggests that there is an additional source of variability not taken into account by the particular form of the look-ahead model used in fitting the mean velocity data. (Recall in particular that off-rates were neglected, that incorrect Watson-Crick basepairing was not allowed, and that the forward rate was assumed independent of which base was being incorporated into the nascent RNA chain.) This additional source of noise may be attributed to the instrumental noise in single molecule experiments; specifically, the unidirectional drift and the heterogeneity in the RNA polymerase enzymes may cause a large variance in the experimental population elongation rate (29). This noise may explain why the simulation velocities have smaller variance in the average velocity in comparison to the experimental measurements.

The waiting time distribution clearly distinguishes the different window sizes. The most important result here is that the window size $w = 1$, in which there is no look-ahead at all, is clearly ruled out by the data. The theoretical probability density of the waiting time in that case has the form given by the second term only on the right-hand side of Eq. 12. This term, which is a difference of two exponentials, describes a curve that rises from zero to a peak value before decaying, unlike the data, which are monotone decreasing. The fact that the shortest waiting times have the highest probability densities in the data is strong qualitative evidence in favor of the look-ahead concept, since this observation implies that the first site of the window of activity is quite likely to be occupied immediately after a forward move, and this requires the kind of parallel processing that is implied by the look-ahead model.

The fit of the model prediction to the data is particularly good for the window size $w = 4$, the largest window size for which we currently have a theoretical result available for comparison. Although the fit for this case on the logarithmic scale (Fig. 6) still shows some error for the longer waiting times (which occur only rarely in the data), that error becomes invisible when the data and the theoretical results

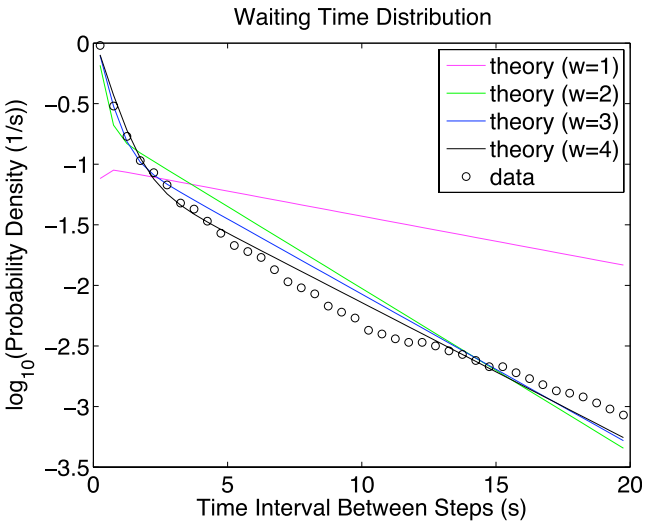

FIGURE 6 Fit of the theoretical predictions (Eq. 13) of the look-ahead model to the experimental distributions of the waiting times between forward moves of the RNA polymerase molecule. Four window sizes ($w = 1, 2, 3, 4$) are considered, and each of these cases (*solid lines*) has been separately best-fit to the experimental data (*open circles*). Horizontal axis is the waiting time between forward moves of the RNA polymerase molecule, in seconds, and the vertical axis is the base 10 logarithm of the probability density for the occurrence of each waiting time. The experimental data are replotted from Fig. S3(c) of Abbodanzieri et al. (15). Note the poor character of the fit for window size 1 and the dramatic improvement with increasing window size up to window size 4, for which the fit is remarkably good; see also linear plot of the same data in Fig. 7.

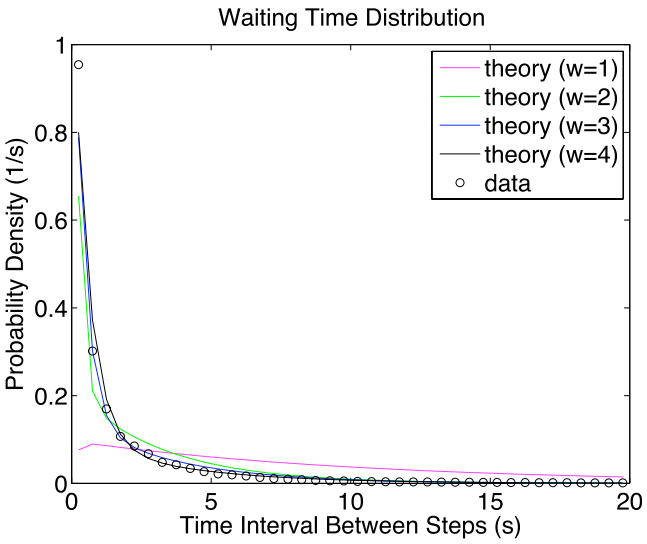

FIGURE 7 The experimental data and theoretical curves of Fig. 6 are here replotted with a linear scale for the vertical (probability density) axis. Here as on the logarithmic plot, the fit for window size 1 is poor, and the fit for window size 4 is excellent. The logarithmic plot visually emphasized the infrequently occurring long waiting times, which are here relatively suppressed, with the consequence that the fit for window size 4 looks essentially perfect, even though the fitting was done to the logarithmic form of the data.

are replotted on a linear scale, where the visual impression is of an essentially perfect fit (Fig. 7).

The conclusion that the non-lookahead case $w = 1$ has a waiting time distribution that rises from zero to a peak before decaying (contrary to the monotone decay of the experimental data) is very general and not dependent on specific modeling assumptions. Any transcription model with a single site for binding of rNTP has the feature that at least two steps are needed per forward step of the enzyme, namely the binding of rNTP and its covalent linkage to the nascent RNA chain. Any such model will therefore have a nonmonotone waiting time distribution qualitatively like that derived above for the case $w = 1$. The only escape from this conclusion is that the rising phase of the waiting time distribution may be so fast that it is not resolved by the experimental measurement. This would be the case, for example, if the binding/unbinding of rNTP were a process of rapid equilibrium. We consider this possibility below.

Rapid equilibrium limit

A potential criticism of the parameter fitting procedures considered in this article is that all of the unbinding rates have arbitrarily been set equal to zero. In this section, therefore, we briefly discuss the opposite limit, in which the reversible binding/unbinding within the window of activity is regarded as a rapid equilibrium process. In this limit, the size of the window of activity makes no difference, so the look-ahead feature of our model becomes irrelevant, and we might as well consider only the case $w = 1$. This obviously implies that it is futile to try to determine the window size by parameter fitting if reversible binding is a rapid equilibrium process.

Let us consider the form of the waiting time distribution under the rapid equilibrium assumption with the rNTP concentrations chosen to be equally rate-limiting. What ''equally rate-limiting'' means in the context of rapid equilibrium is that the product $k_f p_{occupied}$ is the same for all four of the rNTP, where $p_{occupied}$ is the probability that a site which can bind that particular rNTP is occupied. Note that k_f and $p_{occupied}$ may separately differ for the different rNTPs, provided that their product is the same for all four rNTPs. This can always be achieved by adjusting the ambient rNTP concentrations, since each of the $p_{occupied}$ can be adjusted within the interval (0, 1) by varying the corresponding rNTP concentration.

Under the conditions described in the previous paragraph, it is easy to see that the waiting-time distribution is a simple exponential of the form $k \exp(-kt)$, where $k = k_f p_{occupied}$ This would give a straight line on a semilogarithmic plot and is inconsistent with the experimental data reported in Abbodanzieri et al. (15).

One can always argue, however, that the equally-rate-limiting condition as described above may not have been perfectly achieved in the experiment. In that case, the waiting time distribution under the assumption of rapid equilibrium would be mixture of several exponentials, and it might indeed be possible to fit the experimental data with such a model. Further experimental work may be needed to clarify this issue. If the rapid equilibrium assumption is correct, then it should be possible to find a combination of ambient rNTP concentrations that fulfill the above conditions and make the waiting time distribution into a single exponential. This would disprove (or at least make irrelevant) the look-ahead model, since rapid equilibrium makes the first site of the look-ahead window be the only one that matters.

A further prediction of the rapid-equilibrium assumption, in common with all models that have $w = 1$, is that the waiting times for the individual forward moves of the RNA polymerase model should be statistically independent of each other. This will be discussed more fully below; see Proposed Experimental Test to Rule Out a Large Class of Models in which Look-Ahead Does Not Occur.

Proposed experiments to determine the window size

The foregoing results leave some ambiguity about the size of the look-ahead window. Experimental data on mean velocity as a function of concentration (7) are best fit by the look-ahead model with $w = 1$ or $w = 2$. Velocity histograms obtained in those same experiments do not help to determine the window size (and indeed have widths that are approximately twice that predicted by the model, regardless of the window size), but experimental data on waiting time distributions obtained with low, equally rate-limiting ambient rNTP concentrations (15) are well fit by the look-ahead model with $w = 4$. To help resolve this ambiguity, we now propose two additional experiments that may help to determine the window size.

The experiments we propose are both of the type in which ambient rNTP concentrations are varied and the mean velocity of transcription is measured. In the first proposed experiment, we again exploit the notion of equally rate-limiting concentrations (15) to obtain what we call universal curves. There is one such curve for each window size, with no adjustable parameters. In the second case, we propose experiments with saturating rNTP concentrations for more direct determination of the parameter k_f, after which it should be straightforward to determine the window size.

Throughout this section, we employ the six-parameter version of the look-ahead model that was considered previously. Recall that the unknown parameters of this version of the model are the window size, w, the forward rate, k_f, and the four concentration-independent on-rates, $(k'_{on})_i$, from which the on-rates themselves, $(k_{on})_i$, can be determined once the rNTP concentrations are known. It is the ability to manipulate the on-rates by varying the concentrations that motivates the experimental protocols proposed here.

Universal curves

This proposed method of determining the window size is based on the observation that the look-ahead model simplifies enormously in the special case that all four of the rates $(k_{on})_i$ are equal. In that case, the DNA sequence becomes irrelevant, and the unknown parameters are reduced to three: w, k_f, and k_{on}. Let $\overline{v}$ be the mean velocity of the RNA polymerase along the DNA in basepairs per second. From dimensional considerations, it is clear that $\overline{v}/k_f$ is determined by k_{on}/k_f for any particular window size, w. It is intuitively clear that $\overline{v}/k_f$ is a monotonically increasing function of k_{on}/k_f. This function starts at zero when its argument is zero and asymptotically approaches one as its argument approaches infinity. There is one such function for each window size w. These functions involve dimensionless variables only and have no adjustable parameters. In that sense, they are universal, and their graphs are universal curves. The universal curves may be obtained by solving the master equation in the appropriate special cases and then plotting the results as $\overline{v}/k_f$ versus k_{on}/k_f.

Examples of the universal curves are plotted in Fig. 8. As w increases, the curves shift up and to the left, i.e., the velocity is an increasing function of w when the other parameters are held fixed. This is a reflection of the parallel-processing feature of the look-ahead model. The RNA polymerase moves faster when w is larger because of the opportunity to bind more rNTP in advance of their covalent incorporation into the growing RNA chain.

For purposes of parameter estimation, however, the most important feature of the universal curves is that each of them is unique. If one could make a plot of experimental data of $\overline{v}/k_f$ as a function of k_{on}/k_f, that plot would presumably fall on one and only one of the universal curves. The one with which it agreed would reveal the correct value of w.

To make use of this idea, though, we have to make all of the on-rates equal, and also we then need to be able to vary that common on-rate and plot the results in terms of the dimensionless variables stated above, namely $\overline{v}/k_f$ as a function of k_{on}/k_f. It is not immediately obvious how we can do any of this, since we do not know any of the parameters of the model a priori.

To overcome this difficulty, we make use of the parameter-fitting procedure described above, in which the data take the form of plots of the mean velocity of transcription as a function of the concentrations of each of the rNTP, varied one at a time. Even though that parameter fitting procedure is not very effective for determining the value of w, it does determine the best-fit rate constants, k_f and $(k'_{on})_i$, for any hypothesized value of w. We can therefore check whether any particular guess for w is correct in the following way:

Step 1. Given the experimental data on the mean velocity of transcription as a function of each of the rNTP concentrations, together with a hypothesized value of w, determine the best-fit rate constants k_f and $(k'_{on})_i$.

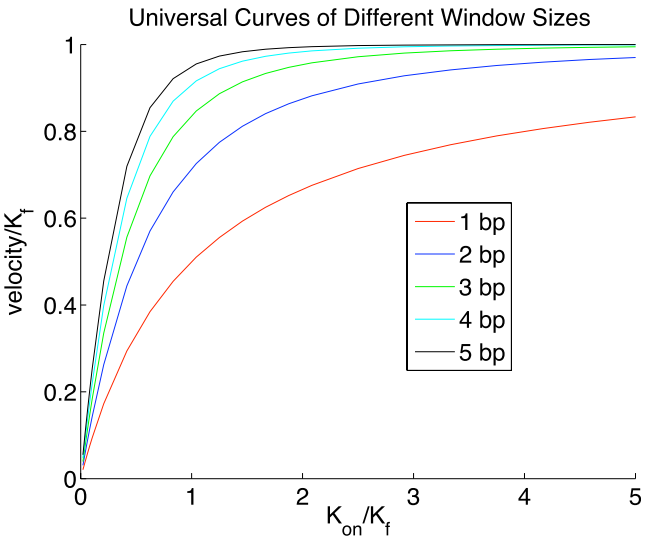

FIGURE 8 Universal curves that give $\overline{v}/k_f$ as a function of k_{on}/k_f when the rNTP concentrations have been adjusted, so that all four of the on-rates are equal (and then varied in fixed proportions to vary the common value of k_{on}). Results computed by solving the steady-state master equation for various window sizes $w = 1, 2, 3, 4, 5$ are shown. The mean velocity increases with increasing window size for fixed values of the rate constants of the model because of the look-ahead feature that rNTP molecules can be bound and held in readiness within the window of activity in advance of their being incorporated into the nascent RNA chain. Note that this look-ahead effect is very substantial, as w increases for small window sizes, but that it tends to saturate (diminishing returns) as the window size grows, suggesting convergence to a limiting universal curve for large window sizes. Each of the universal curves is expressed in terms of dimensionless variables and has an absolute significance, with no adjustable parameters. The look-ahead model used here is restricted by the conditions that off-rates are neglected, incorrect Watson-Crick basepairing is forbidden, and the forward rate is assumed independent of which base is being incorporated into the nascent RNA chain.

Step 2. Use the fitted values of $(k'_{on})_i$ to determine the rNTP concentrations that make all four of the $(k_{on})_i = k_{on}$, independent of i. Since $(k_{on})_i = (k'_{on})_i[\text{rNTP}]_i/[\text{rNTP}]_0$, the correct choice of $[\text{rNTP}]_i$ to yield any particular common on-rate k_{on} is given by

$$[\text{rNTP}]_i = [\text{rNTP}]_0 \frac{k_{on}}{(k'_{on})_i}. \tag{14}$$

Step 3. Now do an experiment with the rNTP concentrations set according to the above formula, and plot a single point with coordinates $k_{on}/k_f, \overline{v}/k_f$. In these ratios, the value of k_f that should be used is the one that was obtained during the parameter fitting for the hypothesized value of w.

Step 4. Repeat this procedure for enough values of k_{on}/k_f to get a picture of the graph of $\overline{v}/k_f$ versus k_{on}/k_f.

Step 5. Plot the data points of this graph on the same axes as the family of universal curves. If the result fits the universal curve for the hypothesized value of w, then that value of w is correct, or at least self-consistent.

What is less clear, perhaps, is what will happen when the hypothesized value of w is incorrect. In that case, the parameters obtained by best fit will be wrong, the concentrations used will not actually yield equal values of k_{on}, and the results will not, typically, fall on any of the universal curves.

Fortunately, we can test the proposed experiment by computer simulation using synthetic data. In this test, we consider only window sizes $w = 1, 2, 3$ for clarity of illustration, but the method can be extended without difficulty to larger window sizes. The results of the proposed experiment by computer simulation can be seen in Fig. 9. The correct window size can be inferred from the calculated curve that most closely matches to one of the three universal curves.

RNA polymerase velocity at high rNTP concentrations

When parameter fitting is done for a hypothesized window size, the best-fit value of k_f depends on the window size in a systematic way. This is again because of the parallel-processing feature of the look-ahead model as discussed above.

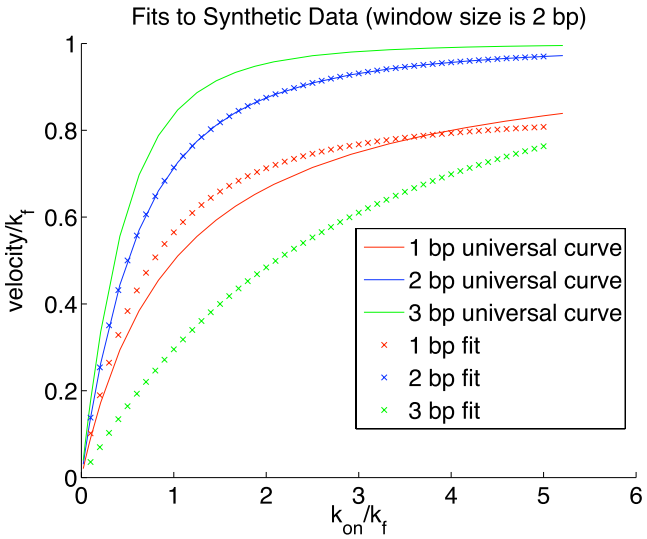

FIGURE 9 Example of the use of the universal curves to determine the window size. Synthetic data like those shown in Figs. 3 and 4 were generated with the following true parameters: $w = 2$, $k_f = 25.0$/s, $(k'_{on})_A = 150.0$/s, $(k'_{on})_C = 300.0$/s, $(k'_{on})_G = 20.0$/s, and $(k'_{on})_U = 2000.0$/s. The look-ahead model was then fit to the synthetic data with hypothesized window sizes $\tilde{w} = 1, 2, 3$. For each hypothesized window size, additional synthetic data were generated using the true rate constants and with a simulated experimental protocol involving rNTP concentrations adjusted in an attempt to achieve equal on-rates. Note, however, that this attempt is only successful to the extent that the rate constants have been correctly identified, which is only the case when the hypothesized window size is correct. The new synthetic data are plotted in the manner that should produce one of the universal curves if the parameters have been correctly identified. The plotting procedure uses the parameters known to the investigator, which are the best-fit parameters for each hypothesized window size, not the true parameters. Results are compared to the corresponding universal curve in each case. The result that fits its universal curve (in this case, $\tilde{w} = 2$) determines the true window size. Note that the synthetic data obtained with $\tilde{w} = 1, 3$ do not match their own (or indeed any other) universal curve.

Since a larger window size produces faster motion for any given set of rate constants, the fitting procedure necessarily adjusts the rate constants to compensate for the window size in an attempt to match the observed mean velocity. The result is that the best-fit value of k_f will be a decreasing function of the hypothesized window size. This means that if we have an independent way to measure k_f, we can use that independent measurement to determine the window size, simply by seeing which of the hypothesized window sizes led to the most accurate prediction of k_f.

Within the framework of the look-ahead model, the most obvious way to measure k_f is to employ saturating concentrations of all four rNTPs, so that the window is always fully occupied, and the RNA polymerase simply moves forward with probability per unit time equal to k_f. Indeed, experimentalists seem to be not too far from this condition when they set all of the rNTP concentrations equal to 1000 μM. From Figs. 3 and 4, however, it is clear that this does not quite produce the limiting velocity of forward movement, and that higher concentrations would be needed for that purpose.

As before, we test this proposed experiment by computer simulation. Table 6 summarizes the results. The table shows that the velocities computed at saturating rNTP concentrations do indeed match the values of k_f that were obtained by parameter fitting with the correct window size.

Proposed experimental test to rule out a large class of models in which look-ahead does not occur

The original formulation of the model proposed in this article is very general. Besides the window size w, it involves $3 \times 4 \times 4$ parameters $(k_f)_{ij}$, $(k_{ON})_{ij}$, and $(k_{OFF})_{ij}$, where i denotes one of the four possible DNA bases and j denotes one of the four possible rNTPs. In particular, this general formulation allows for the possibility of non-Watson-Crick basepairing and for errors in transcription. One can make the model even more general than this in the case $w > 1$ by allowing $(k_{ON})_{ij}$ and $(k_{OFF})_{ij}$ to depend not only on i and j but also on position within the window of activity. Also, one can generalize even further by including the limiting case of rapid

TABLE 6 The following table shows how the true window size can be found once the forward rate constant k_f of the look-ahead model has been independently measured

Actual window size	$(k_f)_{\tilde{w}=1}$	$(k_f)_{\tilde{w}=2}$	$(k_f)_{\tilde{w}=3}$	Saturating velocity
$w = 1$	24.9990	20.7479	19.4603	24.9990
$w = 2$	29.7632	25.0000	23.4987	24.9999
$w = 3$	30.8796	26.4439	25.0000	24.9999

The quantity $(k_f)_{\tilde{w}=1,2,3}$ refer to the forward rate constants obtained by parameter estimation, where $\tilde{w}$ is the hypothesized window size used during the parameter estimation process. The saturating velocity is defined as the limiting velocity at high rNTP concentrations, which we computationally simulated. Since the saturating velocity is equal to the true value of k_f it should match one of the $(k_f)_{\tilde{w}=1,2,3}$ values. The specific $\tilde{w}$ for which the match occurs is the true window size.

equilibrium in which $(k_{ON})_{ij} \to \infty$ and $(k_{OFF})_{ij} \to \infty$, but in such a way that $\frac{(k_{ON})_{ij}}{(k_{OFF})_{ij}}$ has a finite limit.

Within the framework of this large class of models, we seek an experimental test that can potentially rule out all of the non-lookahead models. These are the models with $w = 1$, and also all of the rapid equilibrium models (regardless of w). The models with $w = 1$ have no look-ahead feature, and the rapid equilibrium models may as well have $w = 1$, since the activity at window sites other than the first (if any) has no effect on the dynamics of transcription in the case of rapid equilibrium.

The experiment that we propose involves the transcription of a random DNA sequence, more specifically one in which the bases at the different sites along the DNA are chosen independently according to prescribed base frequencies. The ambient concentrations of the various rNTP should be chosen sufficiently low that the times of the individual forward moves of the RNA polymerase can be resolved, as in Abbodanzieri et al. (15). There is no requirement here that these concentrations should be equally rate-limiting, however.

The experiment that we have just described defines a stationary stochastic process of which the output is the sequence of times at which the RNA polymerase makes its forward moves. For all of the models that we have classified above as non-lookahead models, it is easy to see that the stochastic process in question is a renewal process, in which the time intervals between successive forward moves are independent random variables. We may regard this universal prediction of the non-lookahead case as a null hypothesis, and use randomization tests for serial correlation such as those discussed in Manly (18) to see whether the null hypothesis may be rejected. Rejection of the null hypothesis would not prove the validity of the look-ahead model, but it would rule out a large number of non-lookahead alternatives.

DISCUSSION AND CONCLUSIONS

Because our chemical kinetic model assumes the simultaneous incorporation of nucleotides along with unidirectional forward translocation of the RNA polymerase, our model is most easily visualized in terms of powerstroke mechanisms such as those of Yin and Steitz (11) and Gong et al. (12). We emphasize, however, that our model is agnostic as to physical mechanism, and deals only with chemical kinetic events such as binding, unbinding, and covalent linkage of bases to the nascent RNA chain (which we regard as being synchronous with forward motion of the RNA polymerase enzyme).

We argue that backward translocation is uncommon for several reasons: 1), the breaking of a covalent bond of the nascent RNA chain is energetically unfavorable; 2), at certain sites, the folding of the nascent RNA chain into a hairpin provides a backstop that prevents the nascent RNA chain from moving backward; and 3), backward translocation occurs only under special circumstances, namely during transcriptional arrest, transcriptional termination, or a complete absence of rNTPs (19,20). Our proposed model is best supported by the experimental work of Gong et al. (12), which disputes backward translocation and supports the idea of presorting rNTPs on template DNA sites upstream of the active site.

The nature of pauses in the motion of RNA polymerase has been much debated. Pausing is important to understand because it enables synchronization of enzymatic events and regulates the overall speed of transcription. Recent single molecule experiments on transcriptional elongation (14,19,21,22) have all reached different results and conclusions concerning the nature of pausing. Forde et al. (21) has hypothesized that elongation is a bipartite mechanism, in which the RNA polymerase backtracks followed by a conformational change of the polymerase complex, which results in an arrested molecule incapable of being rescued by an assisted mechanical force. Bai et al. (7,23) have hypothesized that pausing is the result of backward translocations along the DNA. Neuman et al. (19) and Shaevitz et al. (24) have hypothesized that a structural rearrangement within the RNA polymerase enzyme is the cause of short pausing. Based on the latter experiments (19,24), the majority of pausing has been shown to be short and ubiquitous, and is not the result of backtracking along the DNA; instead, it is thought that the polymerase enters an off-pathway state of pause (25). Longer pauses (those >20 s), on the other hand, occur much less frequently and are hypothesized to occur by an entirely different mechanism.

In the look-ahead model, the statistics of the motion of RNA polymerase may be described as follows. Consider the limit in which the forward rate constant is very fast. Then RNA polymerase moves forward every time that the first site within the look-ahead window becomes occupied. The distribution of the waiting time for this to occur will be exponential with a rate constant that may be sequence-dependent. Once a forward step does occur, it may be immediately followed by one or several additional forward steps, depending on how many adjacent sites within the look-ahead window happened to be filled at the moment when the first site is filled. Put another way, the RNA polymerase slides the length of the adjacently filled sites within the window of activity. Such sliding is consistent with the inchworm model (26) of transcriptional elongation that was popular during the 1980s. The inchworm model has never been formally ruled out (19).

An interesting property of the look-ahead model that we have not yet fully explored is the potential role of the look-ahead feature in preventing transcription errors. Assuming that there is a nonzero probability of incorporating an incorrect nucleotide covalently into the nascent RNA chain, it becomes important to reduce the probability of such an incorrect base being present at the site where it would be incorporated. This may be accomplished by having a high off-rate for incorrect basepairings, and by allowing sufficient time for this off-rate to be effective. The look-ahead model provides this possibility (in contrast to a model that only involves binding followed by a covalent linkage).

Using the master-equation formulation of our model, we performed parameter estimation to both synthetic and actual data. Our computational experiments involving parameter fitting to synthetic data show that original parameters can be recovered, even when the synthetic data, generated using the Gillespie method, are noisy. The amount of noise that is introduced in this way decreases inversely as the square-root of the number of runs that are used to generate the synthetic data. By varying the number of runs, we are able to assess the influence of this type of noise on the parameter estimation process. The scenario considered here, in which the synthetic experimental data are corrupted by noise, is more realistic than the noise-free case. Note, in particular, that we are not simply adding arbitrary noise to the data, but instead are considering a type of noise that is intrinsic to the physical process under consideration. Moreover, our computational experiments show that the number of individuals runs necessary to recover the original parameters from noisy synthetic data is not prohibitive, but instead is a feasible number to do in an actual experiment.

We have also performed parameter estimation studies based on two different types of actual experimental data. The first kind of data that we employed concerns the mean velocity of RNA polymerase as a function of the ambient rNTP concentrations, varied one at a time (7). The best fits of the predictions of the look-ahead model to such data are achieved with the window sizes $w = 1$ and $w = 2$. The second kind of data that we used is the statistical distribution of the waiting times between forward moves of the RNA polymerase enzyme (15). These data were obtained with the ambient rNTP concentrations chosen to be equally rate-limiting, an important condition which simplifies the analysis of the look-ahead model. The fit of the predictions of the model to these data clearly rules out the window size $w = 1$ and is excellent for the window size $w = 4$. In this connection, it should be noted that Abbodanzieri et al. interpret their own data as being consistent with a secondary site for rNTP binding, a suggestion that seems to be in accord with the look-ahead concept.

All of the parameter fitting in this article has been done under the assumption that the unbinding rates from the sites within the look-ahead window are negligible. This assumption was made primarily to avoid the proliferation of parameters that would otherwise result. We have, however, briefly considered the opposite assumption, i.e., that the binding/unbinding of rNTP to sites within the look-ahead window are in rapid equilibrium. The rapid equilibrium assumption makes the size of the look-ahead window irrelevant, so one may as well consider $w = 1$, and theoretical results are relatively easy to derive. In particular, it is easy to predict the form of the waiting time distribution for comparison with the experimental data of Abbodanzieri et al. (15). We have done this for the special case in which the ambient rNTP concentrations have been adjusted to make $k_f p_{\text{occupied}}$ the same for all of the different rNTP. Note that these assumptions imply that each of the rNTP concentrations is equally rate-limiting, as in the experiment reported in Abbodanzieri et al. (15). In this special case of rapid equilibrium, the theoretical waiting time distribution is a simple exponential, which is inconsistent with the experimental data (15).

Because our parameter fitting results give different answers for the window size, we have proposed two additional experiments to help resolve this issue. In both cases, we have shown that the proposed method of determining the window size is effective when applied to synthetic data. Since these proposed experiments have not yet been done, actual data are not available.

The first of the proposed experiments is based on the observation that when the rNTP concentrations are manipulated in a specific way, all four of the on-rates become equal and the relationship between the mean velocity of RNA polymerase and the common on-rate can be expressed in terms of certain universal curves, a different one for each window size. These curves relate dimensionless variables and do not involve any adjustable parameters, so it should be possible to determine the window size by seeing which of the universal curves best fits the data.

In a second proposed experiment, we suggest using saturating concentrations of all four rNTPs so that the mean velocity of the RNA polymerase, expressed in basepairs per second, will be equal to the parameter k_f of the model. The reason this should determine the window size is that different hypothesized window sizes lead to different predictions of k_f, so an independent determination of k_f will tell which of these predictions is correct.

A limitation of the parameter fitting done in this article is that it has involved only a special case of the look-ahead model. This special case is characterized by the following additional assumptions, as well as those of the look-ahead model itself. 1), Only correct Watson-Crick basepairing is allowed. 2), The forward rate is assumed to be independent of which nucleotide is being incorporated into the growing RNA chain. 3), We assume that the off-rates can all be neglected.

Quite possibly, one or more of these limitations is responsible for the discrepancy that remains between model predictions and experimental results, even when we have made a best fit of the parameters of the model. Note, for example, that our velocity histograms computed with best-fit parameters are narrower than those obtained experimentally (see Fig. 5). This might not be the case if incorrect Watson-Crick basepairing were allowed, for example. Such issues will be the subject of future research.

The above described limitations are to some extent overcome, however, by our proposed experiment to rule out a large class of non-lookahead models. The discussion of this proposed experiment is based upon the full model of this article, without the simplifying assumptions that were made to facilitate parameter fitting, and also without relying on the use of equally rate-limiting rNTP concentrations.

Within this large class of models, we identify a subset that we refer to as non-lookahead models, and we note how all of these can potentially be ruled out by a statistical test involving rejection of the null hypothesis that the time intervals between successive forward moves on RNA polymerase are independent random variables.

An important question not considered in this article is the structural basis for our proposed look-ahead model. As discussed in Vassylyev et al. (27), there is structural evidence for a preinsertion site that is distinct from the catalytic site of RNA polymerase. Noncovalent binding and selection of the correct rNTP occurs at the preinsertion site, and hydrolysis and linkage to the nascent RNA chain occurs at the catalytic site. This hypothesis, described in Vassylyev et al. (27), is similar, but not identical to the look-ahead model with a window size $w = 2$. The differences are that in the look-ahead model of this article it is possible for an rNTP to bind directly to the catalytic site (if that site should happen to be empty) as well as to the preinsertion site. Another difference, perhaps consistent with Vassylyev et al. (27) but not discussed there, is the possibility of parallel processing that exists in the look-ahead model: the preinsertion site can fill while the catalytic site is occupied. Strong qualitative evidence in favor of such parallel processing comes from the experimental fact that the measured probability density of the waiting time for a forward move is monotone decreasing (15), so that the most likely waiting time is zero. This cannot be the case if two (or more) kinetic steps must occur serially for each forward step of the RNA polymerase molecule. As reported herein, our best fit to the data in Abbodanzieri et al. (15) occurs with a look-ahead model whose window size is 4. We are not aware of any structural data that would support a window size >2, however, so this leaves a discrepancy between kinetic and structural evidence that needs to be resolved.

Finally, it is important to keep in mind that the data to which our proposed model predictions are compared in this article come from experiments on prokaryotic RNA polymerase. There is no reason why the look-ahead model should not be applicable to eukaryotic RNA polymerases; indeed, one might reasonably expect larger window sizes in the eukaryotic case. It is therefore exciting to note that single-force microscopy has recently been applied to the study of transcriptional elongation by eukaryotic RNA polymerase (28). This opens up the possibility that the model described here will have a new domain of applicability. Indeed, by fitting the model both to the prokaryotic and to the eukaryotic RNA polymerases, one should be able to learn more about the differences between these two classes of related enzymes.

We thank the Center for Applied Mathematics at Cornell University for the generous use of computing resources during this project. We thank Arthur LaPorta, Lu Bai, and Dan S. Johnson at LASSP/Cornell University for information concerning their single molecular experiments. In addition, we thank Udo Wehmeier for preliminary work on the model of this article, and Darren J. Wilkinson for discussion with Y.R.Y. about parameter estimation approaches. Andrew Matteson was instrumental in pointing out an error in an earlier version of Eq. 13, which is now correct, thanks to his vigilance. We thank the two anonymous reviewers of this article for their insightful comments and suggestions, especially for calling our attention to the waiting time distributions reported in Abbodanzieri et al. (15) that seem to be the best available data for determining the window size of the look-ahead model. Finally, we thank Daniel B. Forger for his advice and support.

Y.R.Y was supported on an National Science Foundation-Integrative Graduate Education and Research Traineeship grant No. DGE-033366 and C.S.P. was supported in part by National Institutes of Health grant No. 1P50GM071558-01A2 to the Systems Biology Center in New York.

Soli deo gloria.

REFERENCES

1. Bustamante, C., J. Macosko, and G. Wuite. 2000. Grabbing the cat by the tail: manipulated molecules one by one. *Nat. Rev. Mol. Cell Biol.* 1:130–136.
2. Strick, T., J. Allemand, V. Croquette, and D. Bensimon. 2001. The manipulation of single biomolecules. *Phys. Today.* 54:46–51.
3. Bustamante, C., Z. Bryant, and S. Smith. 2003. Ten years of tension: single-molecule DNA mechanics. *Nature.* 421:423–427.
4. Julicher, F., and R. Bruinsma. 1998. Motion of RNA polymerase along DNA: a stochastic model. *Biophys. J.* 74:1169–1185.
5. Guajardo, R., and R. Sousa. 1997. A model for the mechanism of polymerase translocation. *J. Mol. Biol.* 265:8–19.
6. Tadigotla, V., D. Maoileigidh, A. Sengupta, V. Epshtein, R. Ebright, et al. 2006. Thermodynamic and kinetic modeling of transcriptional pausing. *Proc. Natl. Acad. Sci. USA.* 103:8–19.
7. Bai, L., R. Fulbright, and M. D. Wang. 2007. Mechanochemical kinetics of transcription elongation. *Phys. Rev. Lett.* 98:068103.
8. Wang, H., T. Elston, A. Mogilner, and G. Oster. 1998. Force generation in RNA polymerase. *Biophys. J.* 74:1186–1202.
9. Yamada, Y.R., and C.S. Peskin. 2006. A chemical kinetic model of transcriptional elongation. arXiv:q-bio.BM/0603012 v2.
10. Yamada, Y.R. 2007. Quantitative models of transcriptional elongation. (Thesis). Cornell University, Ithaca, NY.
11. Yin, Y., and T. Steitz. 2004. The structural mechanism of translocation and helicase activity in T7 RNA polymerase. *Cell.* 116:393–404.
12. Gong, X., C. Zhang, M. Feig, and Z. Burton. 2005. Dynamic error correction and regulation of downstream bubble opening by human RNA polymerase II. *Mol. Cell.* 18:461–470.
13. Transcription elongation. *In* Pausing, Arrest, and Termination: Structure and Mechanism: Lectures at the KITP 2003 Program on Bio-Molecular Networks. 2003. http://online.itp.ucsb.edu/online/bionet03/ebright/.
14. Adelman, K., A. LaPorta, T. Santangelo, J. Lis, J. Roberts, et al. 2002. Single molecule analysis of RNA polymerase elongation reveals uniform kinetic behavior. *Proc. Natl. Acad. Sci. USA.* 99:13538–13543.
15. Abbodanzieri, E., W. Greenleaf, J. Shaevitz, R. Landick, and S. Block. 2005. Direct observation of base-pair stepping by RNA polymerase. *Nature.* 438:460–465.
16. Gillespie, G. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* 22:403–434.
17. Gillespie, G. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340–2361.
18. Manly, B. 1998. Randomization, Bootstrap and Monte Carlo Methods in Biology Chapman and Hall, London, UK.
19. Neuman, K., E. Abbondanzieri, R. Landick, J. Gelles, and S. Block. 2003. Ubiquitous transcriptional pausing is independent of RNA polymerase pausing. *Cell.* 115:437–447.

20. Nudler, E., A. Mustaev, E. Lukhtanov, and A. Goldfarb. 1997. The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell*. 89:33–41.

21. Forde, N., D. Izhaky, G. Woodcock, G. Wuite, and C. Bustamante. 2002. Using mechanical force to probe the mechanism of pausing and arrest during continuous elongation by *E. coli* RNA polymerase. *Proc. Natl. Acad. Sci. USA*. 99:11682–11687.

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

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

24. Shaevitz, J., E. Abbodanzieri, R. Landick, and S. Block. 2003. Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. *Nature*. 426:684–687.

25. Herbert, K., A. LaPorta, B. Wong, R. Mooney, K. Neuman, et al. 2006. Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell*. 125:1083–1094.

26. Chamberlin, M. 1995. New models for the mechanism of transcription elongation and its regulation. *In* The Harvey Lectures.. Wiley, New York.

27. Vassylyev, D., M. Vassylyeva, A. Perederina, T. Tahirov, and I. Artsimovitch. 2007. Structural basis for transcription elongation by bacterial RNA polymerase. *Nature*. 448:157–162.

28. Galburt, E., S. Grill, A. Wiedmann, L. Lubkowska, J. Choy, et al. 2007. Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner. *Nature*. 446:820–823.

29. Tolic-Norrelykke, S., A. Engh, R. Landick, and J. Gelles. 2004. Diversity in the rates of transcript elongation by single RNA polymerase molecules. *J. Biol. Chem*. 279:3292–3299.