Math. Model. Nat. Phenom.
Vol. 5, No. 3, 2010, pp. 206-227
DOI: 10.1051/mmnp/20105313

The Influence of Look-Ahead on the Error Rate of Transcription

Y. R. Yamada[a] [1] **and C. S. Peskin** [b]

[a] Department of Mathematics, University of Michigan, 48109-1043 Ann Arbor, MI, USA
[b] Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

Abstract. In this paper we study the error rate of RNA synthesis in the look-ahead model for the random walk of RNA polymerase along DNA during transcription. The model's central assumption is the existence of a *window of activity* in which ribonucleoside triphosphates (rNTPs) bind reversibly to the template DNA strand before being hydrolyzed and linked covalently to the nascent RNA chain. An unknown, but important, integer parameter of this model is the window size w. Here, we use mathematical analysis and computer simulation to study the rate at which transcriptional errors occur as a function of w. We find dramatic reduction in the error rate of transcription as w increases, especially for small values of w. The error reduction method provided by look-ahead occurs *before* hydrolysis and covalent linkage of rNTP to the nascent RNA chain, and is therefore distinct from error correction mechanisms that have previously been considered.

Key words: transcription modeling; elongation dynamics of transcription; error-correcting mechanisms; Gillespie simulation; chemical master equation
AMS subject classification: 92C40

1. Introduction

The elongation phase of transcription is the stage in which RNA polymerase incorporates ribonucleoside triphosphates (rNTPs) into the nascent RNA chain. In bacteria, the rate of chain elongation ranges from $\sim 30-80$ bases/second, and may depend on the rate at which the cell is growing [36]. The error rate of transcription in vitro is ~ 1 to 2 per 100,000 basepairs [30, 5, 18], but the error

[1] Corresponding author. E-mail: yryamada@umich.edu

Article published by EDP Sciences and available at http://www.mmnp-journal.org or http://dx.doi.org/10.1051/mmnp/20105313

rate of transcription in vivo is much debated. Although RNA polymerase is highly conserved in all living organisms, the two most studied polymerases are bacterial polymerase from Escherichia coli and eukaryotic RNA polymerase II (Pol II). Almost all current data for studying errors in transcription are from studies of these two polymerases.

Of all the steps in transcription, the elongation stage is most amenable to a physical description [10]. Recent experimental advances in single molecule microscopy have produced high quality dynamic data [2, 11, 28] on the elongation dynamics of transcription. These advances have aided in developing accurate quantitative models of elongation. These models have mainly focused on transcription pausing and backtracking [3, 4, 27, 33, 37, 39, 40, 41]. The look-ahead model of transcription elongation dynamics proposed by Yamada and Peskin [40, 41] assumes that there exists a window of activity in which rNTPs bind reversibly to the template DNA strand before being hydrolyzed and linked covalently to the nascent RNA chain. An important parameter of the model is the window size, in bases, denoted by w. In the present paper, we study the influence of the window size w on the error rate of transcription.

Transcriptional fidelity is clearly important for the survival of cells and organisms. Several error-correcting mechanisms involving proofreading to maintain fidelity in transcription have been proposed [7, 13, 20, 42]. It has also been hypothesized that transcription elongation factors play an important role in enhancing transcriptional fidelity, including factors such as GreA, GreB and NusA [26, 31] in bacteria and TFIIS in eukaryotes [16, 29]. Several investigators have studied mutations of Pol II and their effects on transcriptional elongation rate, control and fidelity [6, 15, 17, 21, 22, 23]. Here we investigate an error *reduction* mechanism that is inherent in the look-ahead model.

The distinction between error reduction and error correction is that error reduction occurs before an incorrect rNTP has been hydrolyzed and incorporated into the nascent RNA chain, whereas error correction occurs afterwards. Error reduction is thus inherently more economical than error correction. The two mechanisms are by no means mutually exclusive, and nature may well employ both of them. In this paper, however, we limit our considerations to the study of an error reduction mechanism that is inherent in the look-ahead concept.

2. The Model

This section contains a complete statement of the look-ahead model [41], as well as a description of the special case of the look-ahead model that is used in the present paper to assess the influence of look-ahead on the error rate of transcription. The model description here is in verbal form; the equations that govern the model appear in the following sections as needed, see also [41].

The look-ahead model assumes that there is a window of activity (subset of the transcription bubble formed by RNA polymerase) within which ribonucleoside triphosphates (rNTPs) can bind reversibly to the template strand of the DNA prior to being linked covalently to the nascent RNA chain. The sites within the window of activity where such reversible binding may occur are numbered $j = 1, 2, \ldots, w$. We assume that the binding and unbinding of rNTP to these different sites

occur independently of whether any of the other sites are occupied or unoccupied, and also that the rate constants for these reactions depend only upon the chemical identity of the DNA base that is present at a site and upon the chemical identity of the rNTP that is binding or unbinding there, but not upon the location of the site within the window of activity.

Site #1 of the window of activity is special. We assume that it is the only site at which covalent linkage of an rNTP to the nascent RNA chain may occur. Therefore, covalent linkage of an rNTP can only occur when site #1 of the window of activity is occupied. The rate constant for covalent linkage may depend on which DNA base is present at site #1 and also on which rNTP is reversibly bound there.

When such covalent linkage does occur, we assume that this causes the RNA polymerase, and with it the transcription bubble and the window of activity, to move forward a distance of one basepair along the DNA. The result from the point of view of the window of activity is downward shift in the state of its sites. Let $s(j)$ denote the state of site j immediately before the covalent linkage event, and let $s'(j)$ denote the state of site j immediately afterwards. By "state" of a site we mean the following: (1) the identity of the template strand DNA base that is located at that site, (2) whether or not that DNA base has an rNTP reversibly bound to it, and finally (3) the identity of that rNTP if there is one present. The downward shift of the states of the sites within the window of activity that accompanies a forward move of the RNA polymerase is described as follows: $s'(j) = s(j+1)$ for $j = 1, 2, \ldots, w-1$. Note that $s(1)$ plays no role here, since the prior content of site #1 leaves the window of activity during a forward move of the RNA polymerase. In fact, the rNTP at site #1 is the one that is incorporated into the nascent RNA chain during (or immediately prior to) that forward move. Also, note that the above formula does not specify $s'(w)$. Since site w is a new one that has just been drawn into the window of activity by the forward move, its DNA base is the one that was immediately downstream of the window of activity on the template strand immediately prior to the forward move, and we know that there cannot be any rNTP bound to the DNA base at site #w immediately following a forward move. This is because there has not been any time for such binding to occur.

In general, the look-ahead model as defined above has a large number of parameters. Specifically, there are 16 rate constants for reversible binding of an rNTP to a DNA base within the window of activity, 16 rate constants for the corresponding unbinding reactions, and 16 rate constants for the covalent linkage to the nascent RNA chain of an rNTP that is reversibly bound to the DNA base at site #1 of the window of activity. (The number $16 = 4 \times 4$ arises in each case because the rate constant in question depends on the chemical identity of the DNA base and also on the chemical identity of the rNTP, with 4 choices for each.) The window size w, which of course is a positive integer, is yet another parameter of the look-ahead model.

Since our purpose here is not to make realistic parameter choices, but rather to study the potential of look-ahead as an error-reduction mechanism, we restrict the parameters in such a way as to highlight the issue of fidelity of transcription in its simplest possible form. This is accomplished by treating all Watson-Crick base pairs as equivalent to each other, and likewise all non-Watson-Crick base pairs as equivalent to each other, while of course maintaining the distinction between a Watson-Crick and a non-Watson-Crick base pair. We regard any Watson-Crick base pair as "correct," and therefore we use the subscript "C" on any rate constant that pertains to a Watson-Crick

base pair. Similarly, we regard a non-Watson-Crick base pair as "incorrect," and use the subscript "I" to denote any rate constant pertaining to such a miss-matched pair.

The parameters of the model are as follows:

w = window size in bases (a positive integer)

α_C,α_I = rate constants for binding (association) of the correct (C) or any particular incorrect (I) rNTP to an available DNA base at any particular site within the window of activity

β_C,β_I = rate constants for unbinding (dissociation) of a correct (C) or incorrect (I) rNTP that is reversibly bound within the window of activity

k_C,k_I = rate constants for hydrolysis and covalent linkage to the nascent RNA chain of an rNTP that is correctly (C) or incorrectly (I) bound at the site of incorporation (site #1) in the window of activity.

Note that for any given DNA base, there is only one correct rNTP that can bind to it, but there are three incorrect choices of rNTP. Therefore, the overall rate constant for incorrectly filling an empty site is $3\alpha_I$. This factor of 3 is significant, since it biases the system in favor of errors and makes the problem of achieving high-fidelity transcription all the more difficult.

The six rate constants defined above are all first-order rate constants, with units of reciprocal time. By the law of mass action, the binding rate constants α_C and α_I are each proportional to the concentrations of their respective rNTPs. We assume here that the concentrations of the four rNTPs are equal, so that the symmetry stated above in which all Watson-Crick pairs are equivalent, and all non-Watson-Crick pairs are equivalent, is not disturbed by unequal ambient concentrations of the different rNTPs.

Although w may be any positive integer, it should be kept in mind that the look-ahead feature of the model is only present when $w > 1$. When there is more than one site within the window of activity, the model has a parallel-processing or assembly-line character, in which rNTPs can be lined up along the template strand of the DNA, so that with high probability an rNTP is already present when the RNA polymerase is ready to incorporate it into the nascent RNA chain. This parallel-processing feature obviously accelerates transcription. Its impact on the fidelity of transcription, if any, is not so readily apparent. We study this issue by comparing the error rate of the model with $w = 1$, which may be called the non-look-ahead case, to the error rate when $w > 1$.

3. Analysis of the Model

Analysis of the Error Rate when $w = 1$

The case in which $w = 1$ is the special case of the look-ahead model in which there is no look-ahead feature. In this case, the window of activity contains a single site which may be in one of three states: it may have no rNTP bound, it may have the correct rNTP bound, or it may have an

incorrect rNTP bound to the DNA base that is located within the window. The transitions among these possible states, together with their rate constants, are depicted in Figure 1. Note in particular the arrows with rate constants k_I and k_C in Figure 1. These depict the hydrolysis and irreversible covalent linkage to the nascent RNA chain of the incorrect or correct rNTP that is reversibly bound at the site. According to our assumptions, such covalent linkage is accompanied by a forward shift of the RNA polymerase molecule by one basepair along the DNA. Immediately after this forward move, there cannot be any rNTP bound to the DNA base that has just been brought into the window of activity. This explains why the arrows labeled by k_I and k_C point back to the empty site in Figure 1.

We analyze the steady state of the kinetic scheme depicted in Figure 1. Let

p_I = probability that an incorrect rNTP is reversibly bound

p_C = probability that a correct rNTP is reversibly bound

$1-(p_I+p_C)$ = probability no rNTP is bound.

The steady state equations can be read directly from Figure 1. They are:

$$\begin{aligned} 3\alpha_I(1-(p_I+p_C)) &= (\beta_I+k_I)p_I \\ \alpha_C(1-(p_I+p_C)) &= (\beta_C+k_C)p_C. \end{aligned}$$

Putting these equations in standard form, we obtain:

$$\begin{aligned} (\beta_I+k_I+3\alpha_I)p_I+3\alpha_I p_C &= 3\alpha_I \\ \alpha_C p_I+(\beta_C+k_C+\alpha_C)p_C &= \alpha_C, \end{aligned}$$

The determinant is:

$$\begin{aligned} \Delta_1 &= (\beta_I+k_I+3\alpha_I)(\beta_C+k_C+\alpha_C)-3\alpha_I\alpha_C \\ &= (\beta_I+k_I)(\beta_C+k_C)+3\alpha_I(\beta_C+k_C)+\alpha_C(\beta_I+k_I), \end{aligned}$$

and we have:

$$\begin{aligned} p_I &= \frac{3\alpha_I(\beta_C+k_C+\alpha_C)-\alpha_C 3\alpha_I}{\Delta_1} \\ &= \frac{3\alpha_I(\beta_C+k_C)}{\Delta_1} \\ p_C &= \frac{(\beta_I+k_I+3\alpha_I)\alpha_C-3\alpha_I\alpha_C}{\Delta_1} \\ &= \frac{\alpha_C(\beta_I+k_I)}{\Delta_1}. \end{aligned}$$

Having solved for p_I and p_C, we can evaluate the error rate in the following way. By definition, the error rate E is given by

$$E=\frac{v_I}{v}, \tag{3.1}$$

where v is the number of bases transcribed per second, and v_I is the number of bases transcribed *incorrectly* per second. From the kinetic scheme of Figure 1, it is clear that

$$\begin{aligned} v_I &= p_I k_I \\ v &= p_I k_I + p_C k_C. \end{aligned}$$

For window size $w = 1$, we therefore have the following results:

$$\begin{aligned} v &= k_C p_C + k_I p_I = \frac{k_C \alpha_C(\beta_I + k_I) + k_I 3\alpha_I(\beta_C + k_C)}{(\beta_I + k_I)(\beta_C + k_C) + 3\alpha_I(\beta_C + k_C) + \alpha_C(\beta_I + k_I)} \\ E_1 &= \frac{k_I p_I}{k_C p_C + k_I p_I} = \frac{\theta_1}{1 + \theta_1}, \end{aligned}$$

where

$$\theta_1 = \frac{k_I 3\alpha_I(\beta_C + k_C)}{k_C \alpha_C(\beta_I + k_I)}.$$

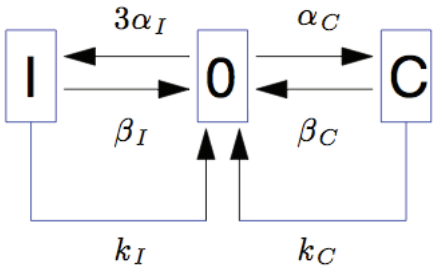

Figure 1: The kinetic scheme of the look-ahead model in the special case $w = 1$. This is the case in which the model has no look-ahead feature, since there is only one site in the window of activity. The parameters α_C and α_I are the rate constants for binding a correct (C) or incorrect (I) rNTP to an empty site. The parameters β_C and β_I are the rate constants for unbinding of a correct (C) or incorrect (I) rNTP that is reversibly bound at that site, thus leaving the site empty again (state 0). The parameters k_C and k_I are the rate constants for the hydrolysis and covalent linkage to the nascent RNA chain of a correct (C) or an incorrect (I) rNTP that was reversibly bound at the site in question. Like an unbinding reaction, covalent linkage also results in an empty site, since it involves a forward move of the RNA polymerase to the next location along the DNA. Note that the rate constant for incorrectly filling an empty site is $3\alpha_I$ since there are always 3 possible incorrect choices for an rNTP.

Analysis of the Error Rate when $w \longrightarrow \infty$

Of course, the window size cannot be infinite, but it is instructive to consider this case, as it presents the look-ahead effect in its most pure form. Later on we shall use numerical methods to study the error rate as a function of w.

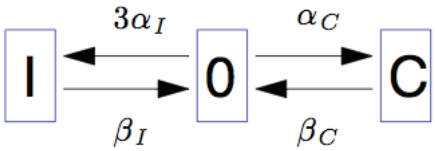

Figure 2: The kinetic scheme for any site *other* than site #1 of the window of activity. Here the only reactions are binding (α) and unbinding (β) of correct (C) or incorrect (I) rNTP. The state labeled 0 has no rNTP bound. In the limiting case of infinite window size, it may be assumed that the kinetic scheme shown here is in its steady state.

To analyze the limit of infinite window size, we adopt a different point of view from the one used in the previous section, and we think about the transient from one incorporation event to the next. Suppose that such an event has occurred at $t = 0$. We focus attention on first site of the window of activity. It has three possible states which we will denote by I, 0, and C where 0 denotes an empty site, i.e., a site to which no rNTP is bound. Because we are considering the limit $w \longrightarrow \infty$, the site in question has had plenty of time to equilibrate during its journey through the window. Throughout this journey, for all $t < 0$, the site in question has been at some location *other* than location #1. Because of this, the only reactions that could have occurred at this site during $t < 0$ are those of binding or unbinding of correct or incorrect rNTP. Therefore, the probabilities of the different states at the site which has just become site #1 immediately after a forward move are given by:

$$\begin{aligned} p_I(0) &= \frac{3\alpha_I\beta_C}{3\alpha_I\beta_C + \alpha_C\beta_I + \beta_I\beta_C} \\ p_0(0) &= \frac{\beta_I\beta_C}{3\alpha_I\beta_C + \alpha_C\beta_I + \beta_I\beta_C} \\ p_C(0) &= \frac{\alpha_C\beta_I}{3\alpha_I\beta_C + \alpha_C\beta_I + \beta_I\beta_C}. \end{aligned}$$

It is easy to check that the above probabilities are the normalized steady state of the kinetic scheme found in Figure 2. It is a special case of the one we considered above for $w = 1$; we can just set $k_C = k_I = 0$ and obtain the needed results.

Now that we have the probabilities for the different states of the first site immediately after a forward move, we have to consider the time-dependent evolution of these probabilities. This evolution is governed by the transient kinetic scheme illustrated in Figure 3.

The differential equations of this scheme are given by:

$$\begin{aligned} \frac{dp_I}{dt} &= 3\alpha_I p_0 - (\beta_I + k_I)p_I \\ \frac{dp_0}{dt} &= \beta_I p_I + \beta_C p_C - (3\alpha_I + \alpha_C)p_0 \\ \frac{dp_C}{dt} &= \alpha_C p_0 - (\beta_C + k_C)p_C. \end{aligned}$$

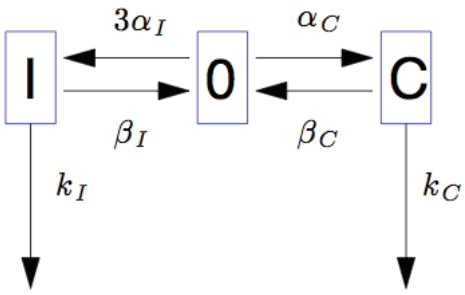

Figure 3: The kinetic scheme of site #1 of the look-ahead model. The process shown here terminates when either of the vertical arrows is traversed, since this describes the hydrolysis and covalent linkage of a correct (C) or incorrect (I) rNTP to the nascent RNA chain. This causes the RNA polymerase to move forward along the DNA, and the state (I, 0, or C) of the site that was previously site #2 becomes that of site #1.

The initial data have been given above. There are two ways to exit, i.e., by the covalent-linkage reactions with rate constants k_I and k_C, and the error rate E_∞ is exactly equal to the probability of exit from state I, by the route with rate constant k_I. It follows that

$$E_\infty = \int_0^\infty k_I p_I(t) dt.$$

We do not actually need to solve the full initial-value problem in order to evaluate E_∞. Instead, we can integrate all three equations from 0 to ∞ with respect to time. Since exit by one route or the other eventually occurs, $p_I(\infty) = p_0(\infty) = p_C(\infty) = 0$, and we get:

$$\begin{aligned} -p_I(0) &= 3\alpha_I T_0 - (\beta_I + k_I) T_I \\ -p_0(0) &= \beta_I T_I + \beta_C T_C - (3\alpha_I + \alpha_C) T_0 \\ -p_C(0) &= \alpha_C T_0 - (\beta_C + k_C) T_C, \end{aligned}$$

where $T_0 = \int_0^\infty p_0(t)dt$, $T_I = \int_0^\infty p_I(t)dt$, and $T_C = \int_0^\infty p_C(t)dt$. In matrix form, the above equations are as follows:

$$\begin{pmatrix} (\beta_I + k_I) & -3\alpha_I & 0 \\ -\beta_I & (3\alpha_I + \alpha_C) & -\beta_C \\ 0 & -\alpha_C & (\beta_C + k_C) \end{pmatrix} \begin{pmatrix} T_I \\ T_0 \\ T_C \end{pmatrix} = \begin{pmatrix} p_I(0) \\ p_0(0) \\ p_C(0) \end{pmatrix}.$$

The determinant of this system is:

$$\Delta_\infty = 3\alpha_I k_I (\beta_C + k_C) + \alpha_C k_C (\beta_I + k_I).$$

Of special interest is:

$$\begin{aligned} T_I &= \frac{\begin{vmatrix} p_I(0) & -3\alpha_I & 0 \\ p_0(0) & (3\alpha_I+\alpha_C) & -\beta_C \\ p_C(0) & -\alpha_C & (\beta_C+k_C) \end{vmatrix}}{\Delta_\infty} \\ &= \frac{p_I(0)[(3\alpha_I+\alpha_C)(\beta_C+k_C)-\alpha_C\beta_C]+3\alpha_I[p_0(0)(\beta_C+k_C)+p_C(0)\beta_C]}{\Delta_\infty} \\ &= \frac{p_I(0)[3\alpha_I(\beta_C+k_C)+\alpha_C k_C]+p_0(0)(3\alpha_I(\beta_C+k_C))+p_C(0)(3\alpha_I(\beta_C+k_C)-3\alpha_I k_C)}{\Delta_\infty}. \end{aligned}$$

Since $p_I(0)+p_0(0)+p_C(0)=1$, the above result reduces to:

$$T_I = \frac{3\alpha_I(\beta_C+k_C)+p_I(0)\alpha_C k_C - p_C(0)3\alpha_I k_C}{\Delta_\infty}.$$

The error rate for $w=\infty$ is then given by:

$$E_\infty = k_I T_I = \frac{3\alpha_I k_I(\beta_C+k_C)}{\Delta_\infty} - \frac{p_C(0)3\alpha_I k_I k_C - p_I(0)\alpha_C k_C k_I}{\Delta_\infty}.$$

Comparison with the $w=1$ case shows that the first term is E_1. Factoring this, we get:

$$E_\infty = E_1\left(1 - \frac{k_C}{k_C+\beta_C}\left(p_C(0) - p_I(0)\frac{\alpha_C}{3\alpha_I}\right)\right). \tag{3.2}$$

Substituting our previous results for $p_C(0)$ and $p_I(0)$, this equation becomes:

$$\begin{aligned} E_\infty &= E_1\left(1 - \frac{k_C\alpha_C}{k_C+\beta_C}\frac{\beta_I-\beta_C}{3\alpha_I\beta_C+\alpha_C\beta_I+\beta_I\beta_C}\right) \\ &= E_1\left(1 - \frac{\alpha_C}{\beta_C}\frac{k_C}{k_C+\beta_C}\frac{1-(\beta_C/\beta_I)}{(3\alpha_I/\beta_I)+(\alpha_C/\beta_C)+1}\right) \\ &= E_1\left(1 - \frac{(\alpha_C/\beta_C)}{1+(\alpha_C/\beta_C)+(3\alpha_I/\beta_I)}\frac{1-(\beta_C/\beta_I)}{1+(\beta_C/k_C)}\right). \end{aligned}$$

Note that

$$\beta_C < \beta_I \quad \Rightarrow \quad E_\infty < E_1.$$

We note in passing that the formula for the error rate in Eq. 3.2 is actually correct for any window size, provided that one substitutes into it the appropriate values of $p_C(0)$ and $p_I(0)$. These are the probabilities that, immediately after a forward move of the RNA polymerase molecule, the site that has just become site #1 of the window of activity already contains a correct or an incorrect rNTP, respectively. (These probabilities do not, in general, add up to 1, since the site may also be empty.) Unfortunately, the probabilities $p_C(0)$ and $p_I(0)$ are not easy to evaluate analytically for an arbitrary window size w. In the special case $w=1$, we have $p_C(0)=p_I(0)=0$, and the

right-hand side of Eq. 3.2 reduces to E_1, thus confirming the consistency of our analysis. In the special case $w = \infty$, we can also evaluate these probabilities from steady-state considerations, as has been done above.

Before turning to the case of general window size, for which numerical methods are needed, we summarize what can be learned about the effect of look-ahead on error rate by comparing the cases $w = 1$ and $w = \infty$. Recall that $w = 1$ is the non-look-ahead case, whereas $w = \infty$ is the case of maximal look-ahead.

An instructive special case is obtained by taking the limit $\beta_C \longrightarrow 0$. This is the case in which a correct Watson-Crick pair is very stable and for practical purposes does not dissociate (within the time-scales that are relevant for transcription). Taking this limit in the formulae for E_1 and E_∞, we find that E_1 has a nonzero limit but that $E_\infty \longrightarrow 0$. In this special case, then, the limiting error rate at infinite window size is zero, even though the error rate remains non-zero with the same parameters in the absence of look-ahead. The reason for this effect is clear. When the window size is 1, there is always the possibility that an incorrect rNTP will bind and be incorporated before the correct rNTP happens to bind. With look-ahead, however, there is more time for the incorrect rNTP to dissociate and be replaced by a correct rNTP. Once the correct rNTP binds, it remains in place until it is incorporated (under the assumption that $\beta_C = 0$). As the window size increases, the probability that any particular DNA base will already have the correct rNTP bound by the time it enters site #1 of the window of activity becomes overwhelming, since the binding of the correct rNTP is irreversible (under the assumption that $\beta_C = 0$).

More generally, suppose that the six rate constants of the model fall into two groups that we shall call "fast" and "slow". Let the fast rate constants be k_C, α_C, and β_I, and let the slow rate constants be k_I, α_I, and β_C. Thus, we assume that the fast reactions are the binding and incorporation of correct rNTP and the unbinding of incorrect rNTP, whereas the slow reactions are the binding and incorporation of incorrect rNTP and the unbinding of correct rNTP. These are certainly plausible assumptions. Let us assume, moreover, that there is a significant gap in the speeds of the fast and slow reactions, such that any reaction in the fast group is much faster than any reaction in the slow group. Under these conditions we can derive approximate expressions for the error rates when $w = 1$ and when $w = \infty$, as follows:

$$E_1 = \frac{3\alpha_I}{\alpha_C}\frac{k_I}{\beta_I}$$

$$E_\infty = E_1\left(\frac{\beta_C}{k_C} + \frac{\beta_C}{\alpha_C} + \frac{\beta_C}{\beta_I}\right). \tag{3.3}$$

According to our assumptions about fast and slow reactions, the error rate is already small in the case $w = 1$, but it is further reduced by look-ahead as $w \longrightarrow \infty$ by the small factor the multiplies E_1 on the right-hand side of the above equation.

Numerical Evaluation of the Error Rate of Transcription for Arbitrary Window Size w

Because of the limitations of mathematical analysis of the model for an arbitrary window size, we employed numerical methods to investigate the relationship between window size and error rate. We used two independent methods so that each would provide a check on the other.

Stochastic Simulation

The dynamics of the look-ahead model are described by a discrete-state continuous-time stochastic process, which can be simulated directly by an event-driven methodology that is often called the Gillespie method [8, 9]. The details of the application of this methodology to the look-ahead model are described in [41].

Event-driven simulation jumps from one event to the next, where an event is the occurrence of one of the reactions of the model (binding, unbinding, or covalent linkage of an rNTP). Immediately after an event has occurred, the state of the model is ascertained, and a list of reactions that are possible in that state is constructed. The sum, K, of the rate constants of the possible reactions is then used to in the determination of the time interval until the next event by choosing that time interval at random with probability density $K\exp(-Kt)$. The particular event that next occurs is chosen (independently of the inter-event time interval) according to the rule that a reaction with rate constant k will be chosen with probability k/K.

In the computational experiments reported here, we simulated the transcription of a DNA strand 300,000 base pairs in length. The DNA sequence to be transcribed was chosen randomly, but note that the sequence actually has no significance when the rate constants of the model are chosen in the manner described above, such that all Watson-Crick base pairs are equivalent, all non-Watson-Crick base pairs are equivalent, and all rNTPs are present in equal concentrations. Results will be presented below, along with those obtained from the master-equation formulation, which is described next.

Master-Equation Formulation

The state of the system at any given time is given by a vector

$$s = (s_0, s_1, \cdots s_w),$$

where

$$s_j = \begin{cases} 0 \text{ , if site j is empty} \\ 1 \text{ , if site j is correctly filled} \\ 2 \text{ , if site j is incorrectly filled} \end{cases}$$

Let δ^j be a vector of length w with components

$$\delta^j_m = \begin{cases} 1 \text{ , if } m = j \\ 0 \text{ , otherwise} \end{cases} .$$

Then the possible transitions of the system starting from state s are as follows. For binding events we have:

$$\begin{aligned} s &\xrightarrow{(s_j=0)\ \alpha_C} s+\delta^j \\ s &\xrightarrow{(s_j=0)\ 3\alpha_I} s+2\delta^j. \end{aligned}$$

For unbinding events we have:

$$\begin{aligned} s &\xrightarrow{(s_j=1)\ \beta_C} s-\delta^j \\ s &\xrightarrow{(s_j=2)\ \beta_I} s-2\delta^j. \end{aligned}$$

For forward movement (i.e. covalent linkage) we have:

$$\begin{aligned} s &\xrightarrow{(s_1=1)\ k_C} (s_2,\cdots,s_w,0) \\ s &\xrightarrow{(s_1=2)\ k_I} (s_2,\cdots,s_w,0). \end{aligned}$$

In the expressions for the rate constants, factors like $(s_j = 0)$ are Boolean expressions that evaluate to 1 if they are true, and 0 if they are false. They express the condition under which the transition may occur.

Next we define:

$$\begin{aligned} R(s,s') = & \sum_{j=1}^{w}(s_j=0)(s'=s+\delta^j)\alpha_C \\ + & \sum_{j=1}^{w}(s_j=0)(s'=s+2\delta^j)3\alpha_I \\ + & \sum_{j=1}^{w}(s_j=1)(s'=s-\delta^j)\beta_C \\ + & \sum_{j=1}^{w}(s_j=2)(s'=s-2\delta^j)\beta_I \\ + & (s_1=1)(s'=Ts)k_C \\ + & (s_1=2)(s'=Ts)k_I, \end{aligned}$$

where

$$T(s_1,\cdots,s_w)=(s_2,\cdots,s_w,0).$$

It is important to define R in this additive manner (instead of defining individual elements separately) since there may be more than one transition connect a given pair of states. For example,

$$(1,0,\cdots,0) \xrightarrow{k_C} (0,0,\cdots,0)$$
$$(1,0,\cdots,0) \xrightarrow{\beta_C} (0,0,\cdots,0).$$

In such cases, we want R to contain the sum of the rates of the different possible transitions from $s \longrightarrow s'$. This situation will happen automatically if we initialize R to 0 and then update R by *adding* the relevant element for each possible transition.

The size of R is $3^w \times 3^w$. In practice, we must assign an integer from $1,\cdots,3^w$ to each state. This is done according to the rule:

$$i = \mathrm{index}(s) = 1 + \sum_{j=1}^{w} s_j 3^{j-1}.$$

Thus $s_1,\cdots,s_w$ are the ternary digits of $i-1$, but in the opposite of the usual order, since s_1 is the least significant digit.

Note that the diagonal elements of R are zero. The rows of R give the rate constants for leaving a given state, and the columns give the rate constants for entering a given state.

In terms of R, the master equation is:

$$\frac{d}{dt}p(s',t) = \sum_s p(s,t)R(s,s') - \sum_{s''} p(s',t)R(s',s'').$$

Let

$$A(s,s') = \begin{cases} R(s,s') \text{ , if } s \neq s' \\ -\sum_{s''} R(s',s'') \text{ , if } s = s' \end{cases}$$

Then,

$$\frac{d}{dt}p(s',t) = \sum_s p(s,t)A(s,s'),$$

and the steady-state solution, which we wish to obtain, is the normalized solution of

$$0 = \sum_s p(s)A(s,s'),$$

where the normalization is given by the following condition:

$$1 = \sum_s p(s).$$

Once we have the normalized solution, we may evaluate:

$$v_C = \sum_s (s_1 = 1)k_C p(s)$$
$$v_I = \sum_s (s_1 = 2)k_I p(s).$$

Here v_C is the rate at which bases are correctly incorporated into the nascent RNA chain, and v_I is the rate at which bases are incorrectly incorporated into the nascent RNA chain. The forward velocity of the RNA polymerase (in bases transcribed per unit time) is:

$$v = v_C + v_I,$$

and the error rate is:

$$E = \frac{v_I}{v_C + v_I} = \frac{v_I}{v}.$$

.

Numerical Results

The parameters used to obtain the results reported here are stated in Table 1. These parameters are not meant to be realistic; they were chosen to illustrate the error-reduction capability of look-ahead. In particular β_C is chosen much smaller than β_I. This choice is motivated by the idea that a Watson-Crick base pair should be more stable (and hence, slower to dissociate) than a non-Watson-Crick base pair. The quantitative strength of this effect may be influenced by the local environment within the RNA polymerase molecule, and cannot therefore be estimated from physical-chemical measurements on the unbinding of rNTP from single stranded DNA in solution. The assumed smallness of β_C is what makes look-ahead particularly effective as an error-reduction mechanism.

It should also be noted that we have chosen k_I an order of magnitude smaller than k_C. This is based on the assumption that the RNA polymerase molecule can detect whether the rNTP at site #1 of the window is correctly matched to the DNA base at that site, and will proceed to link the rNTP covalently to the nascent RNA chain at a higher rate if that rNTP is correct as opposed to incorrect. The assumed disparity between k_I and k_C is an error-reduction mechanism that operates independently of look-ahead.

We have used the two computational methodologies described above to obtain the overall transcription rate and also the error rate of transcription, each as a function of the window size w, with all of the other parameters of the look-ahead model held fixed. The results concerning the overall rate of transcription are presented in Table 2 and Figure 4. They show, as expected, that transcription proceeds faster as the size of the look-ahead window increases. This is an expression of the parallel processing aspect of look-ahead. The larger the size of the look-ahead window, the higher the probability that an rNTP will already be present at site #1 of the window of activity immediately following a forward move of the RNA polymerase molecule. When this occurs, incorporation of that rNTP can proceed without waiting for the site to fill.

In many processes of everyday life, speed and accuracy are inversely related. One might expect on this basis that the increase in speed described above would be accompanied by a decrease in fidelity, but this is not the case. The results in Table 3 and Figure 5 show that the error rate of transcription decreases dramatically as the size of the look-ahead window increases. Note that most of this improvement in fidelity occurs for small window sizes, with saturation of the effect as the window size becomes large. It is quite remarkable that no price in terms of speed is paid for

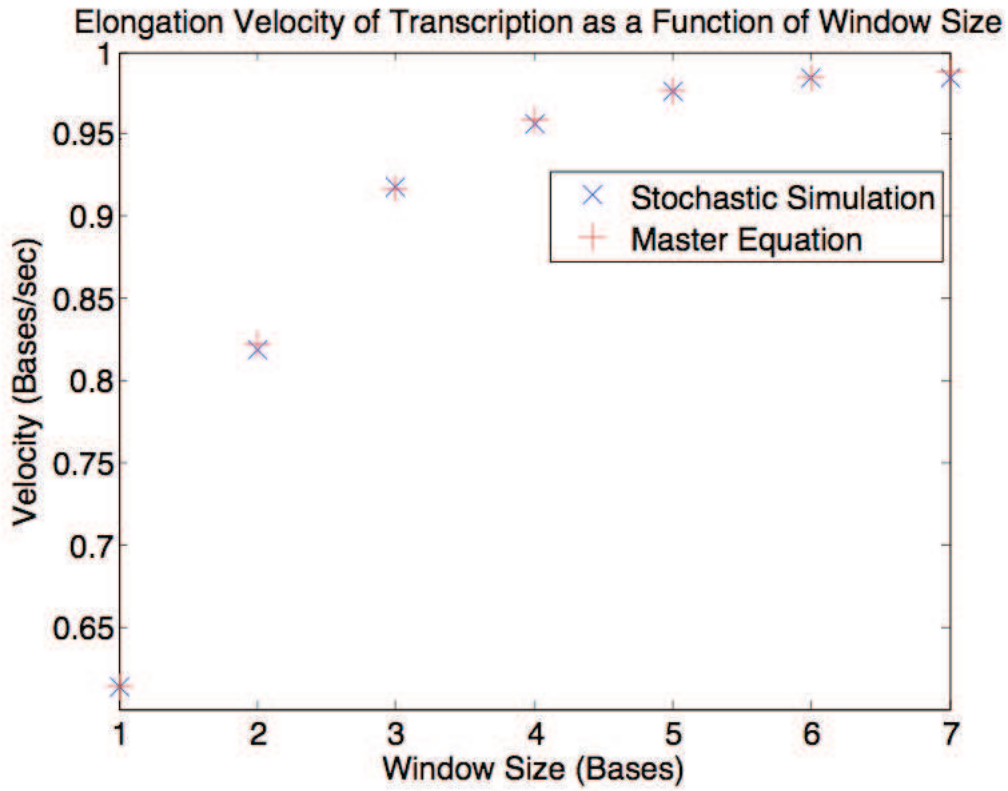

Figure 4: The elongation velocity of transcription is an increasing function of the size of the look-ahead window. This is an expression of the parallel processing feature of the look-ahead model [41]. Note the agreement of the results of the two methods of computation.

this improvement in fidelity. On the contrary, as discussed above, look-ahead increases the velocity of transcription.

The two drastically different methods of calculation employed here (stochastic simulation and solution of the steady-state master equation) give essentially the same results. This strongly suggests that the speed and error rate of transcription are being calculated correctly (within the framework of the assumptions of the look-ahead model).

Discussion, Conclusions, and Future Work

In this paper, we have studied the influence of look-ahead [41] on the error rate of transcription, and have shown that dramatic reduction in the error rate can be achieved by making the number of sites within the look-ahead window larger than one. Large look-ahead windows are not needed for this purpose. Indeed, the improvement as the window size increases is greatest for small window sizes and the fidelity gradually saturates (i.e., stops improving) as the window size increases further. The

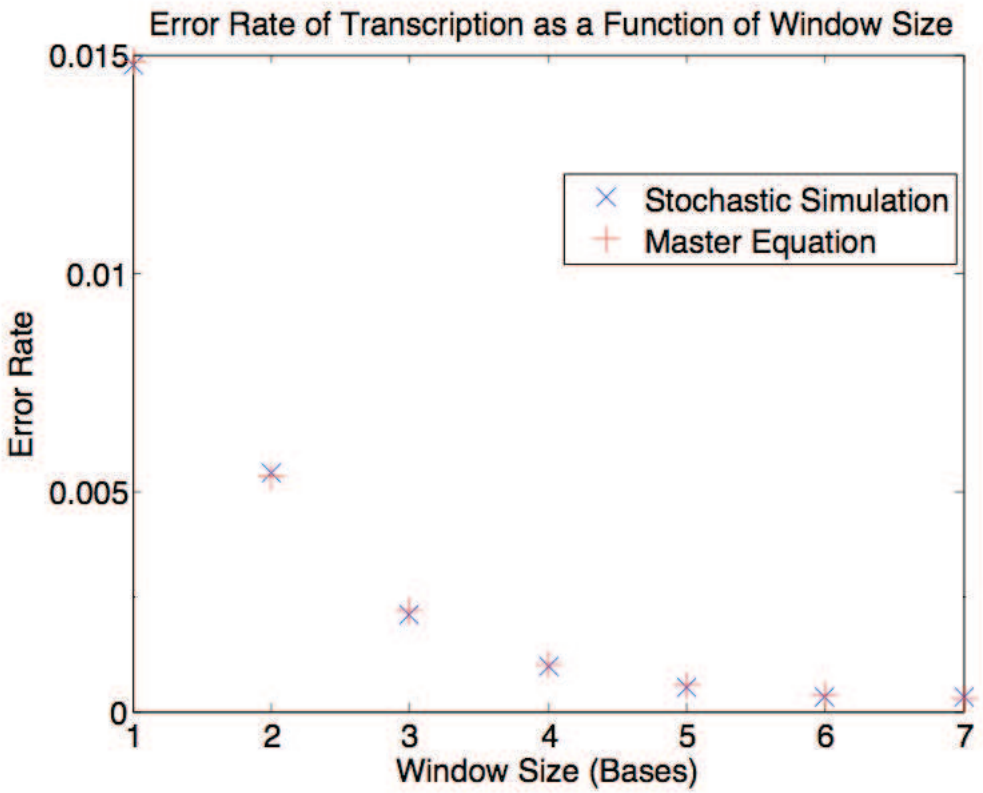

Figure 5: The error rate of transcription is a strongly decreasing function of the size of the look-ahead window, especially for small window sizes. This is because the look-ahead mechanism allows time for the correct complementary rNTP to be selected by a DNA base before that DNA base reaches site #1 of the window of activity, where hydrolysis and covalent linkage of the rNTP to the nascent RNA chain occur. Note the agreement of the results of the two methods of computation.

predicted improvement in fidelity is achieved by the look-ahead mechanism with no reduction in speed; on the contrary, the speed of transcription is also enhanced by look-ahead.

The error-*reduction* mechanism enabled by look-ahead is different from error-*correcting* mechanisms that have been considered previously. The fundamental difference is that the look-ahead mechanism, as its name implies, acts *before* the metabolically costly steps of hydrolysis and covalent linkage of an rNTP to the nascent RNA chain. In this sense, look-ahead is an economical mechanism that prevents errors so that they do not need to be corrected after the fact. Error reduction (e.g., via look-ahead) and error-correction are by no means incompatible, and one would guess that nature would exploit both possibilities in order to make the error rate of transcription as low as possible.

The look-ahead model is fundamentally a chemical-kinetic scheme, and the error-reduction mechanism of look-ahead accordingly resembles kinetic proofreading, which was independently proposed by Hopfield and Ninio [14, 25]. The concept of kinetic proofreading has been invoked

Table 1: Parameter values used in the model simulations. These parameter values were chosen arbitrarily to illustrate the possible influence of look-ahead on the error rate of transcription.

Parameter Symbol	Description of Parameter	Value
α_C	Binding of Correct rNTP to empty site	2
α_I	Binding Incorrect rNTP to empty site	2
β_C	Unbinding of Correct rNTP from occupied site	0.01
β_I	Unbinding of Incorrect rNTP from occupied site	20
k_C	Hydrolysis and Covalent Linkage of Correct rNTP	1
k_I	Hydrolysis and Covalent Linkage of Incorrect rNTP	0.10

to explain the low error rate of translation as well as other physiological processes such as T-cell receptor signal transduction, signal transduction specificity and RecA protein binding dynamics [1, 24, 12, 32, 35]. A distinctive feature of look-ahead, however, is the use of several sites, assembly-line fashion, to allow more time for discrimination to occur.

Concerning the elongation dynamics of transcription, recent papers by Voliotis et. al. [37, 38] have explored the implications of a hypothesized kinetic proofreading mechanism. Specifically, the backtracking of the RNA polymerase can reverse the covalent linkage of an incorrect base in the nascent chain. Another variation of this mechanism was proposed by [18, 19, 20] in which error correction occurs after the hydrolysis but before the covalent linkage of the rNTP into the nascent RNA chain. For comparison, note that no distinction is made in the current form of the look-ahead model between the hydrolysis step and the covalent linkage step. A summary of various mechanisms that have been proposed either to *correct* or to *reduce* errors during transcription can be found in Table 4.

An important project for future work is to devise experiments and mathematical/computational methods to determine realistic parameters for the look-ahead model. This will almost certainly require departure from the idealized case considered in the present paper, in which the only distinction made was that between a Watson-Crick base pair and a non-Watson-Crick base pair. In reality it is more likely that each of the 16 possible base pairs (one choice of DNA base and one choice of rNTP) has its own particular binding rate constant, unbinding rate constant, and covalent linkage rate constant. Although parameter fitting was done in [41], that paper considered only a special case in which unbinding was neglected and only correct Watson-Crick pairs were ever allowed to form. The problem of parameter fitting in the general case is of course much more difficult. The task may be somewhat eased, however, by the experimental ability to control the DNA sequence that is being transcribed as well as the ambient concentrations of the different rNTP, and to count the errors of transcription that actually occur.

Table 2: Transcription velocity (bases transcribed per unit time) as a function of window size. Comparison of the results of master-equation computations and stochastic simulations using a random DNA sequence comprised of 300,000 basepairs. In both cases, the parameters are those shown in Table 1. Because of the arbitrariness of these parameters, the absolute velocities are not significant. Note the agreement of the results obtained by the two methods, and the upward trend of the velocity as the window size increases.

Window Size	Master-Equation Velocity	Stochastic Velocity
1	0.613061	0.613510
2	0.822285	0.818836
3	0.916041	0.917765
4	0.958165	0.956782
5	0.976547	0.975784
6	0.984337	0.986362
7	0.987563	0.984162

Table 3: Error rate of transcription as a function of window size. Comparison of the results of master-equation computations and stochastic simulations using a random DNA sequence comprised of 300,000 basepairs. In both cases, the parameters are those shown in Table 1. Note the agreement of the results obtained by the two methods, and the strongly downward trend of the error rate as the window size increases, especially for small window sizes.

Window Size	Master-Equation Error Rate	Stochastic Error Rate
1	0.014850	0.014700
2	0.005378	0.005423
3	0.002284	0.002193
4	0.001083	0.001023
5	0.000591	0.000536
6	0.000388	0.000393
7	0.000305	0.000316

Table 4: Possible fidelity mechanisms during transcriptional elongation. This table summarizes mechanisms that have been proposed either to reduce or to correct errors during transcriptional elongation. The look-ahead model is an example of an error reduction mechanism, in which errors are avoided before the incorrect ribonucleotide triphosphate (rNTP) is hydrolyzed and incorporated into the nascent RNA chain. This mechanism is in contrast to error-correcting mechanisms such as kinetic proofreading [14, 25] and hydrolysis rejection. In hydrolysis rejection, the rejection of an incorrect rNTP is done after the hydrolysis step, but before the covalent linkage. Finally, there are proposed error-correcting mechanism that operate after the incorrect rNTP has already been incorporated into the nascent RNA chain. The polymerase expends energy in the form of ATP to actively correct for such an error. The precise mechanism for this kind of correction remains unknown but is thought to involve additional enzymes [28].

Type of Mechanism	Distinguishing Feature	References
Error Correcting	Post-Covalent Linkage, Post-Hydrolysis	[37, 38, 2]
Hydrolysis Rejection	Pre-Covalent Linkage, Post-Hydrolysis	[18, 19, 20]
Error Reduction	Pre-Covalent Linkage, Pre-Hydrolysis	[40, 41]

Acknowledgements

We thank the organizers of the 2009 Shanks Conference on Mathematical Sciences in Biology and Biomedicine at Vanderbilt University for the opportunity to present this research. We would like to acknowledge the support and advice of Professor Daniel B. Forger. The first author was supported on an NSF-IGERT Grant DGE-033366, and the second author was supported in part by NIH Grant 1P50GM071558-01A2 to the Systems Biology Center in New York. Soli Deo Gloria.

References

[1] A. Alon. An introduction to systems biology: design principles of biological circuits. Chapman and Hall, Boca Raton, 2007.

[2] E. Abbodanzieri, W. Greenleaf, J. Shaevitz, R. Landick, S. Block. *Direct observation of base-pair stepping by RNA polymerase*. Nature, 438 (2005), 460-465.

[3] L. Bai, R. Fulbright, M. Wang. *Mechanochemical kinetics of transcription elongation*. Phys. Rev. Lett., 98 (2007), No. 6, 068103.

[4] G. Bar-Nahum, V. Epshtein, A. Ruckenstein, R. Rafikov, A. Mustaev, E. Nudler. *A ratchet mechanism of transcription elongation and its control*. Cell, 120 (2005), No. 2, 183-193.

[5] A. Blank, J. Gallant, R. Burgess, L. Loeb. *An RNA polymerase mutant with reduced accuracy of chain elongation*. Biochemistry, 25 (1986), No. 20, 5920-5928.

[6] Y. Chen, D. Chafin, D. Price, A. Greenleaf. *Drosophila RNA polymerase II mutants that affect transcription elongation*. Jour. Biol. Chem., 271 (1996), No. 11, 5993-5999.

[7] G. Eichhorn, P. Chuknyisky, J. Butzow, R. Beal, C. Garland, C. Janzen, P. Clark, E. Tarien. *A structural model for fidelity in transcription*. Proc. Natl. Acad. Sci., 91 (1994), No. 16, 7613-7617.

[8] D. Gillespie. *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*. J. Comp. Phys., 22 (1976), No. 4, 403-434.

[9] D. Gillespie. *Exact stochastic simulation of coupled chemical reactions*. J. Phys. Chem., 81 (1977), No. 25, 2340-2361.

[10] S. Greive, P. von Hippel. *Thinking quantitatively about transcriptional regulation*. Nat. Rev. Mol. Cell Biol., 6 (2005), 221-232.

[11] K. Herbert, W. Greenleaf, S. Block. *Single-molecule studies of RNA polymerase: motoring along*. Annu. Rev. Biochem., 77 (2008), 149-176.

[12] W. Hlavacek, A. Redondo, H. Metzger, C. Wofsy, B. Goldstein. *Kinetic proofreading models for cell signaling predict ways to escape kinetic proofreading*. Proc. Natl. Acad. Sci., 98 (2001), No. 13, 7295-7300.

[13] S. Holmes, T. Santangelo, C. Cunningham, J. Roberts, D. Erie. *Kinetic investigation of Escherichia coli RNA polymerase mutants that influence nucleotide discrimination and transcription fidelity*. Jour. Biol. Chem., 281(2006), No. 27, 18677-18683.

[14] J. Hopfield. *Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity*. Proc. Natl. Acad. Sci., 71 (1974), No. 10, 4135-4139.

[15] K. Howe, C. Kane, A. Ares. *Perturbation of transcription elongation influences the fidelity of internal exon inclusion in saccharomyces cerevisiae*. RNA, 9 (2003), No. 8, 993-1006.

[16] C. Jeon, K. Agarwal. *Fidelity of RNA polymerase II transcription controlled by elongation factor TFIIS*. Proc. Natl. Acad. Sci., 93 (1996), No. 24, 13677-13682.

[17] M. Kireeva, Y. Nedlialkov, G. Cremona, Y. Purtov, L. Lubkowska, F. Malagon, Z. Burton, J. Strathern, M. Kashlev. *Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation*. Mol. Cell, 30 (2008), No. 5, 557-566.

[18] R. Libby, J. Gallant. *The role of RNA polymerase in transcriptional fidelity*. Mol. Microbiol., 5 (1991), No. 5, 999-1004.

[19] R. Libby, J Gallant. *Phosphorolytic error correction during transcription*. Mol. Microbiol., 12 (1994), No. 1, 121-129.

[20] R. Libby, L. Nelson, J. Calvo, J. Gallant. *Transcriptional proofreading in escherichia coli*. EMBO Jour., 8 (1989), No. 10, 3153-3158.

[21] F. Malagon, M. Kireeva, B. Shafer, L. Lubkowska, M. Kashlev, J. Strathern. *Mutations in the saccharomyces cerevisiae RPB1 gene conferring hypersensitivity to 6-Azauracil*. Genetics, 172 (2006), No. 4, 2201-2209.

[22] P. Mason, K. Struhl. *Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo*. Mol. Cell, 17 (2005), No. 6, 831-840.

[23] M. de la Mata, C. Alonso, S. Kadener, J. Fededa, M. Blaustein, J. Pelisch, P. Cramer, D. Bentley, A. Kornblihtt. *A Slow RNA Polymerase II Affects Alternative Splicing in Vivo*. Mol. Cell, 12 (2003), No. 2, 525-532.

[24] T. McKeithan. *Kinetic proofreading in T-cell receptor signal transduction*. Proc. Natl. Acad. Sci., 92 (1995), No. 11, 5042-5046.

[25] J. Ninio. *Kinetic amplification of enzyme discrimination*. Biochimie, 57 (1975), No. 5, 587-595.

[26] J. Roberts, S. Shankar, J. Filter. *RNA polymerase elongation Ffactors*. Annu. Rev. Microbiol., 62 (2008), 211-233.

[27] J. Roussel, R. Zhu. *Stochastic kinetics description of a simple transcription model*. Bull. Math. Biol., 68 (2006), No. 7, 1681-1713.

[28] J. Shaevitz, E. Abbondanzieri, R. Landick, S. Block . *Backtracking by single RNA polymerase molecules observed at near-base-pair resolution*. Nature, 426 (2003), 684-687.

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

[30] C. Springgate, L. Loeb. *On the fidelity of transcription by escherichia coli ribonucleic acid polymerase*. J. Mol. Biol., 97 (1975), No. 4, 577-591.

[31] E. Stepanova, J. Lee, M. Ozerova, E. Semenova, K. Datsenko, B. Wanner, K. Severinov, S. Borukhov. *Analysis of promoter targets for Escheichia coli transcription elongation factor GreA in vivo and in vitro*. J. Bateriol., 189 (2007), No. 24, 8772-8785.

[32] P. Swain, E. Siggia. *The role of proofreading in signal transduction specifity*. Biophys. J., 82 (2007), No. 6, 2928-2933.

[33] V. Tadigotla, D. O'Maoileidigh, A. Sengupta, V. Epshtein, R. Ebright, E. Nudler, A. Ruckenstein. *Thermodynamic and kinetic modeling of transcriptional pausing*. Prof. Natl. Acad. Sci.,103 (2006), No. 12, 4439-4444.

[34] J. Thomas, A. Platas, D. Hawley. *Transcriptional fidelity and proofreading by RNA polymerase II* Cell, 93 (1998), No. 4, 627-637.

[35] T. Tlusty, R. Bar-Ziv, A. Libchaber. *High-fidelity DNA sensing by protein binding fluctuations*. Phys. Rev. Lett., 93 (2004), No. 25, 2581031.

[36] U. Vogel, K. Jensen. *The RNA chain elongation rate in escherichia coli depends on the growth rate*. J. Bacteriol., 176 (1994), No. 10, 2807-2813.

[37] M. Voliotis, N. Cohen, C. Molina-Paris, T. Liverpool. *Fluctuations, pauses, and backtracking in DNA transcription*. Biophys. J., 94 (2008), No. 2, 334-348.

[38] M. Voliotis, N. Cohen, C. Molina-Paris, T. Liverpool. *Backtracking and error correction in DNA transcription* in *The Art and Science of Statistical Bioinformatics*. 104-107, Leeds University Press, Leeds, 2008.

[39] P. Xie. *A dynamic model for transcription elongation and sequence-dependent short pauses by RNA polymerase*. BioSystems, 93 (2008), 199-210.

[40] Y. Yamada, C. Peskin. *A chemical kinetic model of transcriptional elongation*. LANL ArXiv (2006), q-bio.BM/0603012.

[41] Y. Yamada, C. Peskin. *A look-ahead model for the elongation dynamics of transcription*. Biophys. J., 96 (2009), No. 8, 3015-3031.

[42] N. Zenkin, Y. Yuzenkova, K. Severinov. *Transcript-assisted transcriptional proofreading*. Science, 313 (2006), No. 5786, 518-520.