

Backtracking and Proofreading in DNA Transcription

Margaritis Voliotis,^{1,2} Netta Cohen,¹ Carmen Molina-París,² and Tanniemola B. Liverpool^{3,*}

¹*School of Computing, University of Leeds, Leeds, LS2 9JT, United Kingdom*

²*Department of Applied Mathematics, University of Leeds, Leeds, LS2 9JT, United Kingdom*

³*Department of Mathematics, University of Bristol, Bristol, BS8 1TW, United Kingdom*

(Received 13 June 2008; published 22 June 2009)

Biological cell function crucially relies on the accuracy of RNA sequences, transcribed from the DNA genetic code. To ensure sufficiently high fidelity in the face of high spontaneous error rates during transcription, error correction mechanisms must play an important role. A particular mechanism of transcriptional error correction involves backtracking of the RNA polymerase and RNA cleavage. Motivated by recent single molecule experiments characterizing the dynamics of backtracking, we present a microscopic model of this editing process. We show that such a mechanism can yield error frequencies that are in agreement with *in vivo* observations.

DOI: 10.1103/PhysRevLett.102.258101

PACS numbers: 87.15.rp, 82.39.Fk, 87.10.Mn, 36.20.Fz

The accuracy with which genetic information is processed is an essential factor in the survival and perpetuation of life. Efficient error correction mechanisms are therefore necessary for countering the frequent errors introduced by thermal fluctuations. For example, simple thermodynamic considerations suggest that during DNA transcription passive errors should occur with high propensity [10^{-3} – 10^{-2} errors/nucleotide (nt)]. Nevertheless, transcriptional error rates appear significantly lower (10^{-5} errors/nt) [1]. Kinetic proofreading (KP) [2] provides a general phenomenological framework for understanding mechanisms that ensure low error rates and increased specificity in life processes [2]. To complement this general level of description, quantitative and predictive models that incorporate detailed information about specific biological processes are needed [3].

A particularly important example is the transcription of DNA into RNA. However, a comprehensive understanding of the mechanisms involved in transcriptional error correction is still lacking. Classical KP postulates the existence of a high energy intermediate along the polymerization pathway [2], acting as a fidelity checkpoint and enhancing the discriminatory power of the RNA polymerase (RNAP). Such an intermediate has indeed been suggested by recent structural studies of DNA transcription [4]. In addition, the RNAP's ability to induce cleavage of the RNA (or its so-called nuclease activity) suggests an alternative mode of transcriptional error correction, hereafter referred to as *nucleolytic proofreading*. This involves the backward sliding (*backtracking*) of the RNAP on the DNA template followed by *cleavage* of the nascent transcript [5]. In this manner previously misincorporated nucleotides can be discarded and repolymerized. The existence of these different proofreading mechanisms raises interesting questions regarding their relative roles in enhancing transcriptional fidelity. These can be answered by the construction of predictive models able to discriminate between the different processes.

During backtracking, the active site of the RNAP disengages from the 3' end of the transcript, and the transcription elongation complex (TEC), consisting of the RNAP and the DNA-RNA hybrid, steps backwards along the DNA [5]. The subsequent cleavage of the RNA chain is catalyzed by the active site of the polymerase and in certain cases accessory proteins are necessary to stimulate the reaction [6,7]. Recent single molecule experiments [8] provide support for nucleolytic proofreading by showing that (i) artificially induced misincorporation increases backtracking and (ii) cleavage factors reduce backtracking lifetimes.

In this Letter, we propose a stochastic, nonequilibrium model of transcription elongation involving polymerization of correct and incorrect nucleotides, backtracking, and RNA cleavage. We use the model to assess the role of nucleolytic proofreading in terms of the *error fraction*, defined as the ratio of probabilities of incorporating an incorrect as compared to a correct nucleotide at a given position of the transcript [2]. We study the problem both analytically, in different limits, and numerically, using stochastic simulations. Our results indicate that transcriptional error correction, involving backtracking by multiple nucleotides [8] and RNA cleavage, yields results consistent with multistep KP in the limit of high backtracking rates. More importantly, our results offer a quantitative understanding of nucleolytic proofreading by linking the observed error rate directly to the microscopic rates of the process. Finally, we suggest a number of experiments to test our model and clarify the role of nucleolytic proofreading in transcription.

Transcription elongation can be described in terms of two variables [9]. Let $n = 0, \dots, N$ denote the length of the transcript or equivalently the template position of the last transcribed nucleotide [10]. Let $m = 0, \dots, M$ denote the position of the TEC (specifically the RNAP's active site) relative to n (i.e., the corresponding position of the active site along the DNA template is $n - m$). State $m = 0$ cor-

responds to a TEC in an active state, where polymerization of the next nucleotide can occur, while $m > 0$ corresponds to a TEC in a backtracked state [see Fig. 1(a)]. Extensive backtracking is often blocked by RNA secondary structures (e.g., hairpins) that are formed in the portion of the transcript outside the TEC [5]. Therefore, we assume that backtracking is restricted to a fixed distance $m = M$, which we take to be independent of n [11]. The process starts with the TEC at $(n = 0, m = 0)$ and terminates at $(n = N, m = 0)$.

A schematic diagram of state transitions for the model is given in Fig. 1(b). Given a TEC in an active state $(n, m = 0)$, the TEC can either backtrack to state $(n, m = 1)$ with rate k_b or polymerize the next nucleotide $(n + 1, m = 0)$. Polymerization of correct and incorrect nucleotides proceeds with effective rates k_p and ϵk_p , respectively, yielding a spontaneous error fraction ϵ . Once backtracked the TEC hops randomly between adjacent backtracked states $(n, 0 < m \leq M)$ at rate c . However, given an error at some position $n - l$ ($l \geq 0$) transition of the TEC from state $(n, m = l + 1)$ to $(n, m = l)$ occurs at a slower rate \bar{c} . Finally, from each backtracked state cleavage can occur with rate k_c . Cleavage from any state $(n, m > l)$ ensures removal of the error.

The distinct hopping rate at an error site ($\bar{c} \ll c$) is the key ingredient of this error correction process since it increases the likelihood of cleavage at states $(n, m > l)$. The ratio of the two hopping rates is given by $\bar{c}/c \approx e^{-\Delta G/kT}$ [12], where ΔG is the free energy increase due to the incorporation of an incorrect nucleotide in the RNA-DNA hybrid. The ratio of the polymerization rates for correct and incorrect nucleotides can also be approximated by ΔG , i.e., $\epsilon \approx e^{-\Delta G/kT} \approx \bar{c}/c$ [2].

For the analytic treatment of the model we first consider the dynamics of the process at a fixed template position n

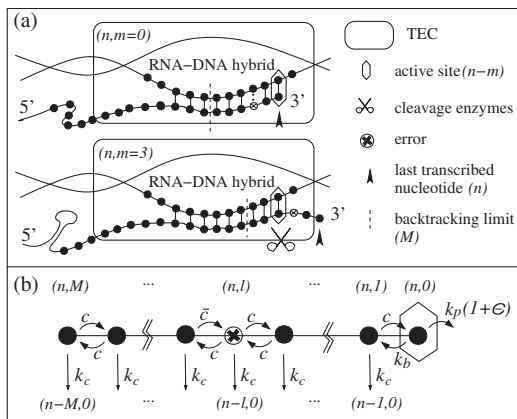


FIG. 1. (a) Schematic illustration of the model. The RNA is marked by 3' and 5'. The transcription elongation complex (TEC) is depicted in the active $(n, m = 0)$ (top) and in a backtracked $(n, m = 3)$ (bottom) state, both with $M = 5$. (b) Schematic illustration of the TEC dynamics at a given position n . The TEC will eventually polymerize forward or cleave from one of the backtracked states.

which allows us to construct an effective model of the full elongation process. The master equation

$$\dot{\mathbf{P}}(t) = \mathbf{W}^{(s)} \cdot \mathbf{P}(t) \quad (1)$$

defines the stochastic dynamics of the TEC at a fixed position n . \mathbf{P} is a column vector of size $(M + 1)$ with entries $P_m(t)$, the probabilities of finding the TEC at translocation state m at time t , having started from $m = 0$ at $t = 0$. $\mathbf{W}^{(s)}$ is the $(M + 1) \times (M + 1)$ transition matrix. The transcription index s is a binary list of 0's and 1's representing the sequence of correct (0) and incorrect (1) nucleotides along the entire transcript. In particular, $s \in S^n$ with $S = \{0, 1\}$ (i.e., for an error at position $n - l$, $s_{n-l} = 1$). The general tridiagonal structure of $\mathbf{W}^{(s)}$ is given below. Along the main diagonal: $W_{j,j}^{(s)} = -[2c + s_{n-j+2}(\bar{c} - c) + k_c]$ except for $W_{1,1}^{(s)} = -[(1 + \epsilon)k_p + k_b]$ and $W_{M+1,M+1}^{(s)} = -[c + s_{n-M+1}(\bar{c} - c) + k_c]$. Along the first diagonal below the main: $W_{j+1,j}^{(s)} = c$, except for $W_{2,1}^{(s)} = k_b$. Along the first diagonal above the main: $W_{j,j+1}^{(s)} = c + s_{n-j+1}(\bar{c} - c)$. All other components are zero. Note that the form of the matrix depends only on the last M elements of s .

The above formulation of $\mathbf{W}^{(s)}$ implies $M + 1$ absorbing boundaries, corresponding either to polymerization from state $m = 0$ or cleavage from each possible backtracked state ($1 \leq m \leq M$). By applying the Laplace transform $\tilde{\mathbf{P}}(z) = \int_0^\infty e^{-zt} \mathbf{P}(t) dt$ to Eq. (1), we obtain a system of algebraic difference equations, which can be used to derive the splitting probabilities p_m for eventually hitting boundary m ($0 \leq m \leq M$) and the corresponding conditional mean exit times, t_m [13]. Note that both p_m and t_m depend on the sequence s .

We now use the splitting probabilities p_m to construct an effective model for the elongation dynamics. Let $\Pi_n^{(s)}(t)$ be the probability of finding a transcript of length n and index s at time t . The transcript can either be extended by one nucleotide (through polymerization) or get shortened by up to M nucleotides (through backtracking and cleavage). These transitions occur with rates r_m , proportional to the splitting probabilities obtained above, i.e., $r_m = p_m/\tau$ ($0 \leq m \leq M$), where τ defines a sufficiently long time scale (i.e., $\tau \gg t_m$, $0 \leq m \leq M$). We note that all results given below depend only on the relative rates and hence do not depend on the exact definition of τ . Summing over s , one obtains $\Pi_n(t) = \sum_{s \in S^n} \Pi_n^{(s)}(t)$, the probability of finding a transcript of length n irrespective of its composition. The dynamics of $\Pi_n(t)$ can be expressed as

$$\frac{d\Pi_n}{dt} = \mathcal{J}_{n-1|0} - \mathcal{J}_{n|0} + \sum_{m=1}^M (\mathcal{J}_{n+m|m} - \mathcal{J}_{n|m}), \quad (2)$$

where $\mathcal{J}_{n|m} = \sum_{s \in S^n} r_m^{(s)} \Pi_n^{(s)}(t)$. For any specific M , Eq. (2) can be used to obtain an expression for \mathcal{P}_n ($\tilde{\mathcal{P}}_n$), the probability of reaching the terminal position N , having

incorporated a correct (incorrect) nucleotide at position n . The error fraction for position n is defined as $\mathcal{E} \equiv \bar{\mathcal{P}}_n/\mathcal{P}_n$. Given a large ensemble of completed transcripts, \mathcal{E} gives the ratio of the number of transcripts with correct nucleotides to those with incorrect nucleotides at position n .

For simplicity, in most of the analysis below, we treat the case $M = 1$, where the TEC can backtrack by only one nucleotide. We introduce the following dimensionless quantities to characterize the competing processes in the model: $\alpha_1 \equiv k_c/c$ and $\alpha_2 \equiv k_c/\bar{c} = \alpha_1/\epsilon$ capture the efficiency of cleavage of correct and incorrect nucleotides, respectively, and $K \equiv k_p/k_b$ the tendency of the TEC to backtrack. The splitting probabilities, obtained from Eq. (1), are determined completely by the identity of the last incorporated nucleotide, s_n . We denote these splitting probabilities when $s_n = 0$ or 1 with p_i and \bar{p}_i , respectively, where $i = 0$ corresponds to polymerization of s_n and $i = 1$ to cleavage. The splitting probabilities take the form $p_0 = \kappa(\epsilon, \alpha_1)/[\kappa(\epsilon, \alpha_1) + \alpha_1]$, $p_1 = 1 - p_0$, $\bar{p}_0 = \kappa(\epsilon, \alpha_2)/[\kappa(\epsilon, \alpha_2) + \alpha_2]$, and $\bar{p}_1 = 1 - \bar{p}_0$, where $\kappa(\epsilon, a) = K(1 + \epsilon)(1 + a)$.

Given the above splitting probabilities, Eq. (2) can now be written for $M = 1$. Laplace transform techniques [13] then yield the termination probabilities $\mathcal{P}_n = \mathcal{N}p_0/(1 - A_n p_0)$ and $\bar{\mathcal{P}}_n = \mathcal{N}\epsilon\bar{p}_0/(1 - A_n\bar{p}_0)$. Here, \mathcal{N} is the normalization constant (such that $\mathcal{P}_n + \bar{\mathcal{P}}_n = 1$), and in the limit $\epsilon \rightarrow 0$, one has $A_n \approx \beta(\beta^{N-n} - 1)/(\beta^{N-n+1} - 1)$, where $\beta = p_1/p_0$ [14]. Thus, the error fraction for $M = 1$ is

$$\mathcal{E} = \frac{\epsilon\bar{p}_0(1 - A_n p_0)}{p_0(1 - A_n\bar{p}_0)}. \quad (3)$$

Figure 2 (top panel) shows the error fraction \mathcal{E} for different positions n as a function of K .

We next consider two limits where \mathcal{E} attains a constant value independent of position n . In the limit $K \gg 1$, one expects that the rare backtracking can hardly improve the error fraction. Indeed, in this limit Eq. (3) reduces to $\mathcal{E} \approx \epsilon$. On the other hand, in the limit $K \ll \alpha_1 \ll \epsilon$, cleavage events dominate the process, and Eq. (3) reduces to $\mathcal{E} \approx \epsilon\bar{p}_0/p_0$, or, in terms of the microscopic rate parameters, $\mathcal{E} \approx \epsilon\bar{c}/c$. Hence, the error fraction depends only on ϵ and the ratio of hopping rates. Since we take these two quantities to be approximately equal, we obtain the limiting error fraction for $M = 1$ to be $\mathcal{E} \approx \epsilon^2$. These two limits are illustrated in Fig. 2 (bottom panel). Numerical data were generated using stochastic simulations [15] of the full elongation model.

In the more general case of $1 \leq M \ll 1/\epsilon$ (i.e., with at most one error occurring in a region of M nucleotides), it can similarly be shown that in the same limit ($K \ll \alpha_1 \ll \epsilon$) the error fraction is

$$\mathcal{E} \approx \epsilon^{M+1} \frac{M^M}{\Gamma(M+1)}, \quad (4)$$

where Γ denotes the Gamma function. Thus, nucleolytic

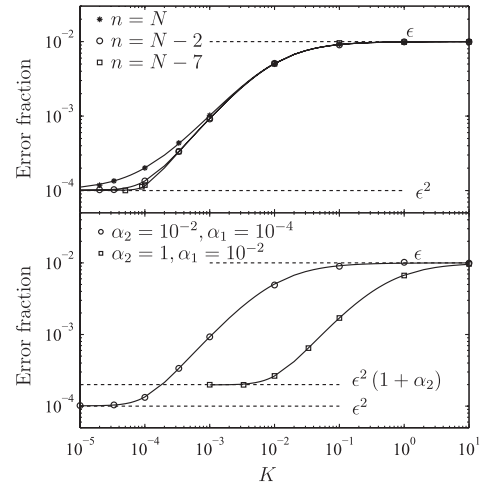


FIG. 2. The error fraction as a function of K ($M = 1$ case). Analytic results [Eq. (3)] are plotted as solid lines, while markers show results obtained from stochastic simulations of the elongation model. Top: The error fraction for different positions with $\alpha_1 = 10^{-4}$, $\alpha_2 = 10^{-2}$, $\epsilon = 10^{-2}$, and $N = 9$. Bottom: The error fraction for different cleavage efficiencies with $\epsilon = 10^{-2}$, $n = N - 2$, and $N = 4$. Dashed lines show limits discussed in text.

proofreading can result in error fractions that scale exponentially with the maximum backtracking distance M . We note that the error fraction attained by KP has a similar dependence on the number of intermediate states [2].

So far we have assumed a constant backtracking rate. However, the presence of an error in the RNA-DNA hybrid could destabilize the TEC, causing more frequent backtracks. A simple model capturing this has backtracking rate \bar{k}_b if an error is within M nucleotides from the 3' RNA end, and k_b otherwise ($k_b < \bar{k}_b$). This can be approximated by an effective backtracking rate $k_b^* = M\epsilon\bar{k}_b + k_b$, giving rise to an effective $K^* = k_p/k_b^* = K/[\epsilon/\epsilon^* + 1]$, where $K = k_p/k_b$ and $\epsilon^* = k_b/(\bar{k}_b M)$. Furthermore, a reasonable assumption is that the TEC rarely backtracks when no errors are present, i.e., $K \gg 1$. Parameter ϵ^* is an intrinsic error scale: When $\epsilon/\epsilon^* \ll 1$ the high K^* regime is obtained, whereas for $\epsilon/\epsilon^* \gg 1$ the behavior of the model is shifted towards the low K^* regime [16].

Let us now estimate the error fractions implied by our model taking into account information from experimental studies. The spontaneous error fraction ϵ can be calculated from the free energy difference due to a misincorporated nucleotide ($\Delta G \approx 4-7kT$), i.e., $\epsilon \approx e^{-\Delta G/kT} \approx 10^{-3}-10^{-2}$ [1]. An estimate of the cleavage rate (for bacterial RNAP in the presence of saturating concentrations of accessory cleavage factors) based on biochemical experiments is $k_c \approx 0.1-1 \text{ s}^{-1}$ [17]. Finally, single molecule experiments have suggested that the TEC hops between backtracked states with rate $c \approx 1-10 \text{ s}^{-1}$ [8]. Using estimates of the maximum error $\epsilon \approx 0.01$, slowest cleavage rates $k_c \approx 0.1 \text{ s}^{-1}$ and fastest hopping rate $c \approx 10 \text{ s}^{-1}$ we can obtain estimates of the lower bounds on the

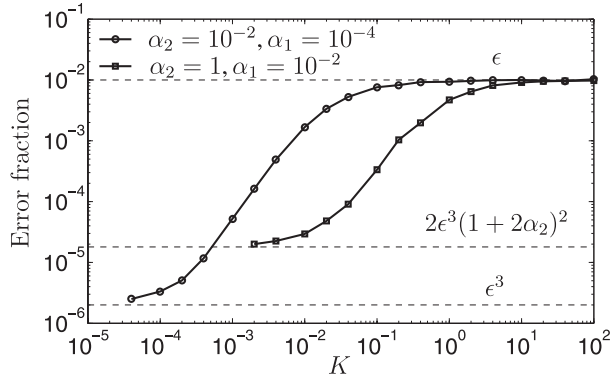


FIG. 3. Error fraction as a function of K ($M = 2$ case). Results were obtained using stochastic simulations of the model for $N = 4$, $\epsilon = 10^{-2}$, and $\alpha_1 = 10^{-2}, 10^{-4}$.

“cleavage efficiencies” $\alpha_1 \sim 0.01$ and $\alpha_2 \sim 1$. These estimates yield error fractions comparable to the ones observed *in vivo*, even for $M = 1$ but sufficiently low values of K (see Fig. 2, bottom panel). Most importantly, however, low error fractions can be obtained in our model even well away from the limiting regime with small M (see Fig. 3 for the $M = 2$ case).

In summary, we have presented a microscopic model of a transcription editing mechanism, involving backtracking and RNA cleavage. Our work extends the existing qualitative description of the process by linking the observed error rates directly to microscopic rate parameters. Backtracking by more than one nucleotide provides a multiple-checking reaction, which probes the fidelity of the last few nucleotides before the next polymerization step. We find, in accordance with the KP scheme, that the greater the delay introduced by this step, the greater the accuracy of the process [2]. Consistent with this picture, the minimum error fraction is obtained in the limit where backtracking and cleavage dynamics dominate the process. In this limit, the error fraction scales exponentially with the maximum backtracking distance M .

Recent experiments have provided support for at least two mechanisms of transcriptional error correction [4,8,18,19]. The first one involves a fidelity checkpoint during the nucleotide addition cycle [20], whereas the second involves backtracking of the RNAP and RNA cleavage. Our model suggests experiments that would provide the quantitative details required to discriminate between these mechanisms and elucidate their relative roles in transcriptional proofreading.

A particular prediction of our model is the strong dependence of transcriptional fidelity on backtracking rates. For example, guanine-cytosine-rich domains that lead to lower backtracking rates (due to the increased stability of the RNA-DNA hybrid) [21] should reduce the efficiency of error correction. More importantly, single molecule manipulation techniques can be used to vary backtracking rates in a controlled manner and validate our model. In particular, applying a load is expected to strongly affect

nucleolytic proofreading since the TEC moves a distance $\sim M\delta x$ (where $\delta x = 3.4 \text{ \AA}$) during the backtracking phase. In contrast, minor effects are expected for proofreading mechanisms along the polymerization pathway, since they should only involve small movements ($\ll \delta x$) of the enzyme. Finally, experimental studies have already revealed that specific mutations in the sequence of RNAP can have a profound effect on transcriptional fidelity [22]. By precisely studying the effects of the mutations on backtracking rates, single molecule experiments with such mutant RNAPs can be used to assess whether nucleolytic proofreading can compensate for such deficiencies.

T. B. L. acknowledges the hospitality of the Curie Institute in Paris.

*t.liverpool@bristol.ac.uk

- [1] A. Blank *et al.*, *Biochemistry* **25**, 5920 (1986).
- [2] J. J. Hopfield, *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4135 (1974); J. Ninio, *Biochimie* **57**, 587 (1975).
- [3] T. McKeithan, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 5042 (1995); J. Yan, M. O. Magnasco, and J. F. Marko, *Nature (London)* **401**, 932 (1999); P. S. Swain and E. D. Siggia, *Biophys. J.* **82**, 2928 (2002).
- [4] D. G. Vassilyev *et al.*, *Nature (London)* **448**, 163 (2007).
- [5] S. J. Greive and P. H. von Hippel, *Nat. Rev. Mol. Cell Biol.* **6**, 221 (2005).
- [6] M. J. Thomas, A. A. Platas, and D. K. Hawley, *Cell* **93**, 627 (1998).
- [7] R. N. Fish and C. M. Kane, *Biochim. Biophys. Acta, Gene Struct. Expr.* **1577**, 287 (2002).
- [8] J. W. Shaevitz *et al.*, *Nature (London)* **426**, 684 (2003); E. A. Galburt *et al.*, *Nature (London)* **446**, 820 (2007).
- [9] M. Voliotis *et al.*, *Biophys. J.* **94**, 334 (2008).
- [10] We define $n = 0$ to be the position at which the elongation phase is entered, a few (8–10) nucleotides downstream of the actual transcriptional starting point.
- [11] For positions $n < M$, backtracking is restricted to $m = n$.
- [12] J. Howard, *Mechanics of Motor Proteins and the Cytoskeleton* (Sinauer, Sunderland, MA, 2001).
- [13] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, Cambridge, England, 2001).
- [14] A_n can be understood as the probability that starting from position $n + 1$ cleavage to position n will occur (before the terminal position N has been reached).
- [15] D. T. Gillespie, *J. Phys. Chem.* **81**, 2340 (1977).
- [16] Error correction is attempted more often when the spontaneous error rate is high.
- [17] E. Sosunova *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15 469 (2003).
- [18] N. Zenkin, Y. Yuzenkova, and K. Severinov, *Science* **313**, 518 (2006).
- [19] N. Alic *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 10 400 (2007).
- [20] Such a mechanism can be described by a model similar to our model with $M = 1$ (with a minimum error fraction limited to ϵ^2).
- [21] T. Ambjörnsson *et al.*, *Phys. Rev. Lett.* **97**, 128105 (2006).
- [22] S. F. Holmes *et al.*, *J. Biol. Chem.* **281**, 18 677 (2006).