Theoretical Study of the Fidelity of Transcription

Yao-Gen Shu[1,2], Ming Li[3] and Zhong-Can Ou-Yang[2]
[1]*Bioinformatics Laboratory of Yishang Innovation Technology Co., Ltd, Beijing 100081, China*
[2]*Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China*
[3]*School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*

Keywords: Gene Transcription Fidelity, First-passage Approach, First-order Neighbor Effects.

Abstract: This year we celebrate the 50th anniversary of the discovery of the three eukaryotic RNA polymerases. Ever since this seminal event was uncovered by Robert Roeder in 1969(Roeder and Rutter, 1969), researchers have investigated the intricate mechanisms of gene transcription with great dedication. However, there is not breakthrough in study of the fidelity of transcription still. Here, we propose a simplest model with first-order neighbor effects, a first-passage approach, to theoretically investigate the gene transcription fidelity.

1 INTRODUCTION

Transcription is the process in which a gene's DNA sequence is transcribed by an RNA polymerase (RNAp). RNAp uses one of the DNA strands as a template to make a new, complementary RNA molecule. The transcription cycle includes three phases: initiation, elongation, and termination. The initiation phase involves recognition of promoter DNA, DNA opening, and synthesis of a short initial RNA oligomer. During the elongation phase, the polymerase uses the DNA template to extend the growing RNA chain in a processive manner. Finally, DNA and RNA are released during termination, and the polymerase can then be recycled and re-initiate transcription(Cramer, 2019).

The probability of mismatch during transcription is more than that during replication of the DNA. It is because errors in transcription may not be fatal due to its non-inherited. When any mismatched nucleotide is added to the template DNA, the RNAp will halt and then it will either proofread the nascent chain or continue without correcting. There are two major ways of proofreading: pyrophosphorolytic editing and hydrolytic one. Once the incorrect nucleotide is added, the very first way to rectify that mistake is by pyrophosphorolytic editing. In this, a pyrophosphate (PPi) will enter the active cleft and attack the wrong nucleotide added, i.e. NMP (such as AMP, GMP, CMP or UMP). It would lead to the conversion of NMP to NTP. Thus the incorrect nucleotide is removed, and the RNAp will add a correct nucleotide. If pyrophosphorolytic editing can not rectify the mRNA sequence, then another method called hydrolytic editing is activated. In this proofreading, certain proteins belonging to Gre and Nus protein family gets activated. If the correct base pair is not added, the RNAp will halt, and these proteins will enter the active cleft of the core enzyme. These proteins would remove 4 Nucleotides from the nascent RNA chain and then the RNAp will add the correct nucleotides again(Libby and Gallant, 1991; von Hippel, 1998).

It's now widely acknowledged that match (A→U, T→A and G↔C, denoted as Right(R) pairs) play a dominate role in the transcription, while the mismatch (denoted as Wrong(W) pairs) occur with very low probability. This is not due to the difference between the free energy of R and W pairs in the double chains: in fact, this free energy difference is only about $2 \sim 4k_BT$(where k_B and T are Boltzmann constant and absolute temperature respectively.), which cannot account for such low mismatch probability if it is estimated by Boltzmann factor. As pointed out by J.Hopfield(Hopfield, 1974) and J.Ninio(Ninio, 1975) in study of the fidelity DNA polymerases, the low mismatch probability may originate from the huge difference of replication kinetics between R and W(I. R. Lehman and Kornberg, 1958; Kunkel and Bebenek, 2000).

The fidelity of DNA replication has been investigated in kinetics recently(Y. G. Shu and Li, 2015; Y. S. Song and Li, 2017; Gaspard, 2017; Q. S. Li and Li, 2019; M. Li and Shu, 2018). However, the fidelity of transcription is rarely studied. Here, we propose a

Shu, Y., Li, M. and Ou-Yang, Z.
Theoretical Study of the Fidelity of Transcription.
DOI: 10.5220/0009181002390241
In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) - Volume 3: BIOINFORMATICS*, pages 239-241
ISBN: 978-989-758-398-8; ISSN: 2184-4305

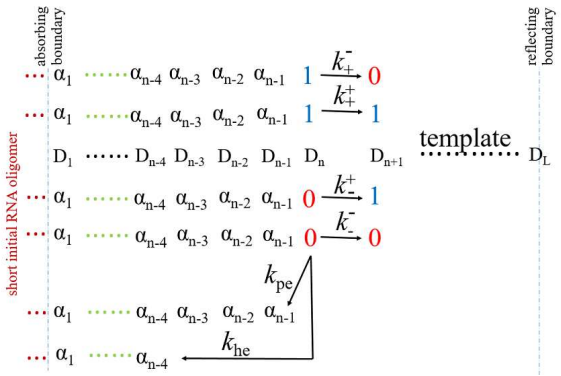

Figure 1: Kinetics of transcription in our model. At the end of initiation phase (after the synthesis of a short initial RNA oligomer, that is the correct nucleation), there is a reflecting boundary at starting of elongation, which corresponds to Eqs.(1) or (2). At $i \in [2,4]$, there is no hydrolytic editing if incorrect nucleotide is added, which corresponds to Eqs.(3) or (4). At $i \in [5, L-6]$, a normal elongation phase, which can be described by Eqs.(5) or (6). There is no hydrolytic editing for $i \in [L-5, L-2]$ due to absorbing boundary at termination as presenting in Eqs.(7) or (8). At absorbing boundary (termination, $i = L-1$), the kinetics is described by Eqs.(9) or (10).

simplest model to theoretically try the fidelity of transcription.

2 FIRST-PASSAGE MODEL

For brevity and not losing generality, we suppose that the gene consists of two kinds of units A (denotes A or T) and G (denotes G or C), and correspondingly two kinds of monomers U (denotes U or A) and C (denotes C or G) are to be added to the end of the growing chain and paired with A or G to transcript the gene with L bases. Since U pairs with A much more probably than with G, we denote $\binom{\mathrm{A}}{\mathrm{U}}$ as the match 1 pair and $\binom{\mathrm{A}}{\mathrm{G}}$ as the mismatch 0 pair. Similarly, we denote $\binom{\mathrm{G}}{\mathrm{C}}$ as 1 and $\binom{\mathrm{G}}{\mathrm{U}}$ as 0. Besides, we only consider the first-order copolymeriztion processes as below: the adding rate of a correct nucleotide at match end is k_+^+; that of a correct nucleotide at mismatch end is k_-^+, while the adding rate of an incorrect nucleotide at match end is k_+^-; that of a incorrect nucleotide at mismatch end is k_-^-.

Since transcription proceeds unidirectionally, we assume that the nascent chain initiates from a pre-existing seed (a short initial RNA oligomer) after initiation phase, then elongates as a template-directed binary copolymer, and terminates at when its length reaches L. In the elongation phase, the monomer U or C can be added to or deleted from the growing end due to proofreading, where k_{pe} and k_{he} denote the rate of pyrophosphorolytic editing and that of hydrolytic one with removing 4 bases respectively. In contrast, the initial seed and the lastly-added monomer can not be deleted. In other words, this is a first-passage process from a reflecting boundary at the first position to an absorbing boundary at the last position. It's worth to note that the initiation and termination here are purely imaginary to simplify the mathematical treatments and do not imply the fact of initiation and termination of transcription. The first-passage treatment largely simplifies the calculations by introducing a closed set of kinetic equations, so it is more convenient to be chosen for approximate calculations. Above kinetic framework is shown in Fig.1.

The probability of the growing RNA sequence $\alpha_1 \ldots \alpha_i$ (where $1 \le i \le L$, and $\alpha_i = 1$ denotes pairs of $\binom{\mathrm{A}}{\mathrm{U}}$, $\binom{\mathrm{T}}{\mathrm{A}}$, $\binom{\mathrm{G}}{\mathrm{C}}$, or $\binom{\mathrm{C}}{\mathrm{G}}$. Otherwise, $\alpha_i = 0$.) appearing in the transcription process along gene $\mathrm{D}_1 \cdots \mathrm{D}_L$ (D_i denotes A/T/C/G) at time t is denoted as $\rho_{\alpha_1 \cdots \alpha_i}^{\mathrm{D}_1 \cdots \mathrm{D}_L}(t)$. Now we have the following master equations.

At the end of initiation phase (after the synthesis of a short initial RNA oligomer, that is the correct nucleation), there is a reflecting boundary at starting of elongation.

$$\dot{\rho}_1^{\mathrm{D}_1 \cdots \mathrm{D}_L} = k_+^+ + k_{\mathrm{pe}} \rho_{10}^{\mathrm{D}_1 \cdots \mathrm{D}_L} + k_{\mathrm{he}} \rho_{1\alpha_2\alpha_3\alpha_4 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \quad (1)$$

$$\dot{\rho}_0^{\mathrm{D}_1 \cdots \mathrm{D}_L} = k_+^- + k_{\mathrm{pe}} \rho_{00}^{\mathrm{D}_1 \cdots \mathrm{D}_L} + k_{\mathrm{he}} \rho_{0\alpha_2\alpha_3\alpha_4 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \quad (2)$$

There is no hydrolytic editing if incorrect nucleotide is added at $2 \le i \le 4$,

$$\begin{aligned} \dot{\rho}_{\alpha_1 \cdots \alpha_i 1}^{\mathrm{D}_1 \cdots \mathrm{D}_L} &= k_{\mathrm{pe}} \rho_{\alpha_1 \cdots \alpha_i 10}^{\mathrm{D}_1 \cdots \mathrm{D}_L} + k_{\mathrm{he}} \rho_{\alpha_1 \cdots \alpha_i 1 \alpha_{i+2} \cdots \alpha_{i+4} 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \\ &\quad + k_+^+ \rho_{\alpha_1 \cdots \alpha_{i-1} 1}^{\mathrm{D}_1 \cdots \mathrm{D}_L} + k_-^+ \rho_{\alpha_1 \cdots \alpha_{i-1} 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \\ &\quad - \left(k_+^+ + k_+^-\right) \rho_{\alpha_1 \cdots \alpha_i 1}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \quad (3) \\ \dot{\rho}_{\alpha_1 \cdots \alpha_i 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} &= k_{\mathrm{pe}} \rho_{\alpha_1 \cdots \alpha_i 00}^{\mathrm{D}_1 \cdots \mathrm{D}_L} + k_{\mathrm{he}} \rho_{\alpha_1 \cdots \alpha_i 0 \alpha_{i+2} \cdots \alpha_{i+4} 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \\ &\quad + k_-^- \rho_{\alpha_1 \cdots \alpha_{i-1} 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} + k_+^- \rho_{\alpha_1 \cdots \alpha_{i-1} 1}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \\ &\quad - \left(k_{\mathrm{pe}} + k_-^+ + k_-^-\right) \rho_{\alpha_1 \cdots \alpha_i 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \quad (4) \end{aligned}$$

At elongation phase for $5 \le i \le L-5$

$$\begin{aligned} \dot{\rho}_{\alpha_1 \cdots \alpha_i 1}^{\mathrm{D}_1 \cdots \mathrm{D}_L} &= k_{\mathrm{pe}} \rho_{\alpha_1 \cdots \alpha_i 10}^{\mathrm{D}_1 \cdots \mathrm{D}_L} + k_{\mathrm{he}} \rho_{\alpha_1 \cdots \alpha_i 1 \alpha_{i+2} \cdots \alpha_{i+4} 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \\ &\quad + k_+^+ \rho_{\alpha_1 \cdots \alpha_{i-1} 1}^{\mathrm{D}_1 \cdots \mathrm{D}_L} + k_-^+ \rho_{\alpha_1 \cdots \alpha_{i-1} 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \\ &\quad - \left(k_+^+ + k_+^-\right) \rho_{\alpha_1 \cdots \alpha_i 1}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \quad (5) \\ \dot{\rho}_{\alpha_1 \cdots \alpha_i 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} &= k_{\mathrm{pe}} \rho_{\alpha_1 \cdots \alpha_i 00}^{\mathrm{D}_1 \cdots \mathrm{D}_L} + k_{\mathrm{he}} \rho_{\alpha_1 \cdots \alpha_i 0 \alpha_{i+2} \cdots \alpha_{i+4} 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \\ &\quad + k_-^- \rho_{\alpha_1 \cdots \alpha_{i-1} 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} + k_+^- \rho_{\alpha_1 \cdots \alpha_{i-1} 1}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \\ &\quad - \left(k_{\mathrm{he}} + k_{\mathrm{pe}} + k_-^+ + k_-^-\right) \rho_{\alpha_1 \cdots \alpha_i 0}^{\mathrm{D}_1 \cdots \mathrm{D}_L} \quad (6) \end{aligned}$$

There is not item of hydrolytic editing for $L-4 \le$

$i \leq L-2$ due to absorbing boundary at termination.

$$\begin{aligned}\dot{\rho}^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_i 1} &= k_{\rm pe}\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_i 10} + k^+_+\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_{i-1}1} \\ &+ k^+_-\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_{i-1}0} - \left(k^+_+ + k^-_+\right)\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_i 1} \quad (7)\\ \dot{\rho}^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_i 0} &= k_{\rm pe}\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_i 00} \\ &+ k^-_-\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_{i-1}0} + k^-_+\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_{i-1}1} \\ &- \left(k_{\rm he} + k_{\rm pe} + k^+_- + k^-_-\right)\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_i 0} \quad (8)\end{aligned}$$

At absorbing boundary (termination, $i = L-1$),

$$\begin{aligned}\dot{\rho}^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_i 1} &= k^+_+\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_{i-1}1} + k^+_-\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_{i-1}0} \quad (9)\\ \dot{\rho}^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_i 0} &= k^-_-\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_{i-1}0} + k^-_+\rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_{i-1}1} \quad (10)\end{aligned}$$

One of our major concerns is the final distribution of the nascent RNA sequence, i.e, the long-time limit $P^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_L} = \rho^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_L}(t\to\infty)$. To calculate it, we assume the initial conditions $\rho^{D_1\cdots D_L}_{\alpha_1}(t=0) \equiv q_{\alpha_1}$, where $q_1 + q_0 = 1$. q_{α_1} can be arbitrarily chosen because it has negligible impacts on the fidelity profile except few positions near the reflecting boundary. Thus, the $P^{D_1\cdots D_L}_{\alpha_1\cdots\alpha_L}$ can be calculated by simulation with known parameters L, k^+_-, k^-_-, k^+_+, k^-_+, $k_{\rm pe}$ and $k_{\rm he}$.

We define the fidelity of limited length transcription in percent as

$$\text{fidelity of transcription} \equiv \frac{\sum_{i=1}^{L}\alpha_i}{L}\times 100\%, \quad (11)$$

which is different from the definition of that of DNA replication.

3 DISCUSSION

Though mismatch in transcription is not fatal, it is tightly related to cancer. In this manuscript, we proposed a general approach, based on the first-passage description of the transcription process, to calculate the positional fidelity for any given gene. The mathematical treatment is shown in Eqs.(1)-(10) with 6 parameters. Although we only consider the first-order copolymeriztion processes, it can be extended to higher-order copolymeriztion as Ref(Q. S. Li and Li, 2019).

We neither do the simulation nor solve analytically Eqs. by iteration as Refs.(Gaspard, 2017; Q. S. Li and Li, 2019) because of the lack of experimental data such as parameters and fidelity. Nevertheless, this is a preliminary theoretical investigation. We hope it will motivates experimenter to quantitatively measure parameters and the fidelity defined above.

ACKNOWLEDGEMENTS

The authors thank the financial support by Key Research Program of Frontier Sciences of CAS (No. Y7Y1472Y61), National Natural Science Foundation of China (No.11574329, 11774358, 11675180), CAS Strategic Priority Research Program (No. XDA17010504), and CAS Biophysics Interdisciplinary Innovation Team Project (No.2060299).

REFERENCES

Cramer, P. (2019). Eukaryotic transcription turns 50. *Cell*, 179:808.

Gaspard, P. (2017). Iterated function systems for DNA replication. *Phys. Rev. E*, 96:042403.

Hopfield, J. J. (1974). Kinetic proofreading: A new mechanism for reducing errors in biosynthetic processes requiring high specificity. *PNAS*, 71:4135.

I. R. Lehman, M. J. Bessman, E. S. S. and Kornberg, A. (1958). Enzymatic synthesis of deoxyribonucleic acid: I. preparation of substrates and partial purification of an enzyme from escherichia coli. *J. Biol. Chem.*, 233:163.

Kunkel, T. A. and Bebenek, K. (2000). DNA replication fidelity. *Ann. Rev. Biochem.*, 69:497.

Libby, R. T. and Gallant, J. A. (1991). The role of rna polymerase in transcriptional fidelity. *Mol. Micro.*, 5:999.

M. Li, Z. C. O.-Y. and Shu, Y. G. (2018). Study on the fidelity of biodevice T7 DNA polymerase. *BIOSTEC2018*, Volume3: BIOINFORMATICS:135.

Ninio, J. (1975). Kinetic amplification of enzyme discrimination. *Biochimie*, 57:587.

Q. S. Li, P. D. Zheng, Y. G. S. Z. C. O.-Y. and Li, M. (2019). Template-specific fidelity of DNA replication with high-order neighbor effects: A first-passage approach. *Phys. Rev. E*, 100:012131.

Roeder, R. G. and Rutter, W. J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*, 224:234.

von Hippel, P. H. (1998). An integrated model of the transcription complex in elongation, termination, and editing. *Science*, 281:660.

Y. G. Shu, Y. S. Song, Z. C. O.-Y. and Li, M. (2015). A general theory of kinetics and thermodynamics of steady-state copolymerization. *J. Phys.:Condens. Matter*, 27:235105.

Y. S. Song, Y. G. Shu, X. Z. Z. C. O.-Y. and Li, M. (2017). Proofreading of DNA polymerase: a new kinetic model with higher-order terminal effects. *J. Phys.:Condens. Matter*, 29:025101.