# Template-specific fidelity of DNA replication with high-order neighbor effects: A first-passage approach

Qiu-Shi Li,[1] Pei-Dong Zheng,[1] Yao-Gen Shu,[2] Zhong-Can Ou-Yang,[2] and Ming Li[1,*]

[1]*School of Physical Sciences, University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, People's Republic of China*

[2]*Institute of Theoretical Physics, Chinese Academy of Sciences, Zhong Guan Cun East Street 55, PO Box 2735, Beijing 100190, People's Republic of China*

(Received 8 April 2019; revised manuscript received 11 June 2019; published 22 July 2019)

DNA replication fidelity is a critical issue in molecular biology. Biochemical experiments have provided key insights on the mechanism of fidelity control by DNA polymerases in the past decades, whereas systematic theoretical studies on this issue began only recently. Because of the underlying difficulties of mathematical treatment, comprehensive surveys on the template-specific replication kinetics are still rare. Here we propose a first-passage approach to address this problem, in particular the positional fidelity, for complicated processes with high-order neighbor effects. Under biologically relevant conditions, we derived approximate analytical expressions of the positional fidelity which show intuitively how some key kinetic pathways are coordinated to guarantee the high fidelity, as well as the high velocity, of the replication processes. It is also shown that the fidelity at any template position is dominantly determined by the nearest-neighbor template sequences, which is consistent with the idea that replication mutations are randomly distributed in the genome.

DOI: 10.1103/PhysRevE.100.012131

## I. INTRODUCTION

Since the Watson-Crick (WC) base-pairing rules of double-strand DNA were first discovered [1], template-directed DNA replication has became a critical research subject to understand genetic variations and evolution. It's now widely acknowledged that WC pairings [A-T and G-C, denoted as Right ($R$) pairs] play a dominant role in the replication process to maintain the genome stability, while the non-WC pairings [denoted as Wrong ($W$) pairs] occur with very low probability (about $10^{-4}$ to $10^{-10}$, dependent on species). This is not due to the difference between the free energy of $R$ and $W$ pairs in the double-helical DNA: in fact, this free energy difference is only about 2–4$k_BT$, which cannot account for such low error rates if estimated by the Boltzmann factor. As pointed out by Hopfield [2] and Ninio [3], the low error rates originate from the huge difference between the replication kinetics of $R$ and that of $W$, which is realized by high-fidelity DNA polymerases (DNAPs) [4,5].

DNAP often consists of a polymerase domain and a proofreading domain. The former catalyzes the template-dependent synthesis of the nascent chain. The latter excises the terminal unit of the growing chain, with a higher excision probability for $W$ than for $R$. While experiments have revealed for a long time that the replication fidelity is determined by both the polymerization kinetics and the proofreading kinetics, related problems were not solved, e.g., how to estimate the positional fidelity (reciprocal of the error rate at each template position), if all the template-specific kinetic parameters are experimentally measured. Because of the mathematical difficulties of handling the kinetic equations of such complex copolymerization processes, systematic theoretical studies on these issues appeared quite recently. So far there are two categories of models.

One assumes that the kinetic parameters of all $R$ (or $W$) pairs are of the same order of magnitude and thus describe the replication approximately as a $R/W$ binary copolymerization process(i.e., the specific template sequence is not considered explicitly). This simplification has long been used in biochemistry for theoretical modeling (e.g., see the historical literature [2,3] or more recent publications like [6–9]). However, thorough studies on such processes appeared only recently, especially for cases in which the rates of monomer addition or deletion at the end of the growing chain depend on the preceding one or more units. Such higher-order neighbor effects may be significant if the terminus of the growing chain contains one or a few $W$s which can destabilize the terminus and hence affect the monomer addition or deletion. These effects have been treated recently by theories under steady-state assumptions, and the overall replication fidelity and growth velocity were calculated numerically or analytically [10–12]. In these theories, the copolymerization process was described as a homogenous Markov chain. This is, however, not appropriate for real cases in which the template DNA sequence is highly inhomogeneous and the kinetic parameters of $R/W$ are highly sequence-dependent.

These template-sequence specificities have not received much attention until very recently. In a series of works, Gaspard has considered all 16 types of base pairs in the kinetic models and handled the high-order neighbor effects successfully [13–17]. By assuming that the probability of any possible sequence of the growing chain can be approximated as a backward (i.e., opposite to the growing direction) inhomogeneous Markov chain in the long-time limit, he succeeded in

*liming@ucas.ac.cn

 ©2019 American Physical Society

proposing an iteration algorithm to numerically compute the positional fidelity or velocity for any given template sequence (i.e., the fidelity or velocity profile). However, there are still many questions to be further addressed. For instance, long-range correlations in real genomic DNA sequences have been reported [18], which seems to imply that the template-specific replication fidelity at different positions in a large range may be correlated. This is doubtful, for it's hard to conceive that replication mutations at different positions have long-range correlations rather than be randomly distributed as widely believed. To what a range do the positional quantities depend on the surrounding template sequence? Do the correlations in the template sequence (if any) have any influence on the fidelity or velocity profile?

In this paper, we propose a different approach to address these template-specific problems. Our method is based on a first-passage description of the replication process. This leads to exact expressions of the probability of the nascent chain sequence as forward inhomogeneous Markov chains. In contrast to the backward Markov chain assumed in the iteration algorithm [17], the forward form is more convenient for approximate numerical or analytical calculations, which offers intuitive insights on how DNAP achieves high fidelity by proofreading while maintaining high velocity. Below we introduce this method, starting from simple binary copolymerization processes with first-order nearest-neighbor effects. We will also show how to generalize this method to more complicated systems.

## II. THE BASIC THEORY: THE FIRST-ORDER REPLICATION PROCESSES

For brevity and not losing generality, we suppose that the template sequence consists of two types of units $A$ and $B$, and correspondingly two types of monomers $a$ and $b$ are added to the active end of the growing chain (i.e., the 3′-end of the nascent DNA chain) and paired with $A$ or $B$ to form a double-strand structure. If $a$ pairs with $A$ much more probably than with $B$, we denote $\binom{A}{a}$ as $R$ and $\binom{A}{b}$ as $W$. Similarly, we denote $\binom{B}{b}$ as $R$ and $\binom{B}{a}$ as $W$.

Given any template sequence of length $L$ (e.g., a region of interests in a real genome), since DNA replication proceeds unidirectionally from the 3′ end to the 5′ end of the template, we assume that the nascent chain initiates from a preexisting seed (either $a$ or $b$) paired with the 3′ end unit of the template, then grows and terminates at the 5′ end of the template. In the growing stage, the monomer $a$ or $b$ can be added to the end by the polymerase domain of DNAP or deleted from the end by the proofreading domain. In contrast, the initial seed and the last added monomer cannot 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 noting that the initiation and termination here are purely imaginary to simplify the mathematical treatments and do not correspond to the real initiation and termination events in biological DNA replication processes. We will show later that different choices of the boundary conditions do not change our major results and conclusions.

For the first-order processes, we assume that the rates of addition or deletion of any monomer $a$ or $b$ depend on the preceding neighbor, denoted as $k^{XY}_{\alpha\beta}$ and $r^{XY}_{\alpha\beta}$, respectively. $\binom{X}{\alpha}$ presents the preceding base pair, and $Y$ is the template unit to which the monomer $\beta$ is paired, $X, Y = A, B$ and $\alpha, \beta = a, b$. The termination step occurs with the addition rate of $k^{XY}_{\alpha\beta}$. It should be noted that all the kinetic parameters here are effective rates. For instance, $k^{XY}_{\alpha\beta}$ is in fact the effective polymerization rate, which is contributed by several reaction substeps and dependent on the monomer concentrations. $r^{XY}_{\alpha\beta}$ is also the effective excision rate contributed by two substeps: the terminus of the growing chain being transferred from the polymerase to the exonuclease and then excised by the exonuclease. It can also be taken as the effective depolymerization rate contributed by several reaction substeps (particularly the PPi attacking the phosphodiester bond of the DNA backbone) in the polymerase, if the exonuclease activity of DNAP is not considered (e.g., the exonuclease activity is inhibited). There have been several methods to obtain such effective rates, e.g., methods based on steady-state approximation [11,12] and quasi-equilibrium approximation [14,15]. The fidelity of DNAP can then be computed in the framework of the minimal reaction schemes with such effective rate parameters. For brevity to illustrate the basic logic of our method, we will not go into such details in this paper.

The probability of the growing chain sequence $\alpha_1\alpha_2\cdots\alpha_i$ $(1 \leqslant i \leqslant L)$ at time $t$ is denoted as $p^{X_1X_2\cdots X_i\cdots X_L}_{\alpha_1\alpha_2\cdots\alpha_i}(t)$. Now we have the following master equations:

$$
\begin{aligned}
\dot{p}^{X_1\cdots X_L}_{\alpha_1} &= r^{X_1X_2}_{\alpha_1a}\, p^{X_1X_2\cdots X_L}_{\alpha_1a} + r^{X_1X_2}_{\alpha_1b}\, p^{X_1X_2\cdots X_L}_{\alpha_1b} \\
&\quad - \left(k^{X_1X_2}_{\alpha_1a} + k^{X_1X_2}_{\alpha_1b}\right) p^{X_1\cdots X_L}_{\alpha_1}, \\
\dot{p}^{X_1\cdots X_i\cdots X_L}_{\alpha_1\cdots\alpha_i} &= k^{X_{i-1}X_i}_{\alpha_{i-1}\alpha_i}\, p^{X_1\cdots X_{i-1}\cdots X_L}_{\alpha_1\cdots\alpha_{i-1}} \\
&\quad + r^{X_iX_{i+1}}_{\alpha_ia}\, p^{X_1\cdots X_iX_{i+1}\cdots X_L}_{\alpha_1\cdots\alpha_ia} \\
&\quad + r^{X_iX_{i+1}}_{\alpha_ib}\, p^{X_1\cdots X_iX_{i+1}\cdots X_L}_{\alpha_1\cdots\alpha_ib} \\
&\quad - \left(r^{X_{i-1}X_i}_{\alpha_{i-1}\alpha_i} + k^{X_iX_{i+1}}_{\alpha_ia} + k^{X_iX_{i+1}}_{\alpha_ib}\right) \\
&\quad \times p^{X_1\cdots X_i\cdots X_L}_{\alpha_1\cdots\alpha_i}, \quad 2 \leqslant i \leqslant L-2, \\
\dot{p}^{X_1\cdots X_{L-1}X_L}_{\alpha_1\cdots\alpha_{L-1}} &= k^{X_{L-2}X_{L-1}}_{\alpha_{L-2}\alpha_{L-1}}\, p^{X_1\cdots X_{L-2}\cdots X_L}_{\alpha_1\cdots\alpha_{L-2}} \\
&\quad - \left(r^{X_{L-2}X_{L-1}}_{\alpha_{L-2}\alpha_{L-1}} + k^{X_{L-1}X_L}_{\alpha_{L-1}a} + k^{X_{L-1}X_L}_{\alpha_{L-1}b}\right) \\
&\quad \times p^{X_1\cdots X_{L-1}X_L}_{\alpha_1\cdots\alpha_{L-1}}, \\
\dot{p}^{X_1\cdots X_L}_{\alpha_1\cdots\alpha_L} &= k^{X_{L-1}X_L}_{\alpha_{L-1}\alpha_L}\, p^{X_1\cdots X_{L-1}X_L}_{\alpha_1\cdots\alpha_{L-1}}.
\end{aligned} \tag{1}
$$

One of our major concerns is the final sequence distribution of the nascent chain, i.e., the long-time limit $P^{X_1\cdots X_L}_{\alpha_1\cdots\alpha_L} = p^{X_1\cdots X_L}_{\alpha_1\cdots\alpha_L}(t\to\infty)$. To calculate it, we assume the initial conditions $p^{X_1\cdots X_L}_{\alpha_1}(t=0) = q^{X_1}_{\alpha_1}$, $q^{X_1}_a + q^{X_1}_b = 1$ ($q^{X_1}_{\alpha_1}$ can be arbitrarily chosen; it has negligible impacts on the fidelity profile except a few positions near the reflecting boundary), $p^{X_1\cdots X_i\cdots X_L}_{\alpha_1\cdots\alpha_i}(t=0) = 0$ $(i \geqslant 2)$, and the long-time limits $p^{X_1\cdots X_i\cdots X_L}_{\alpha_1\cdots\alpha_i}(t\to\infty) = 0$ $(1 \leqslant i < L)$. We integrate [denoting $\Gamma^{X_1\cdots X_i\cdots X_L}_{\alpha_1\cdots\alpha_i} \equiv \int_0^\infty p^{X_1\cdots X_i\cdots X_L}_{\alpha_1\cdots\alpha_i}(t)\,dt$] and solve the above equations to obtain the following iteration relations:

$$
\begin{aligned}
P^{X_1X_2\cdots X_L}_{\alpha_1\alpha_2\cdots\alpha_L} &= \left(q^{X_1}_{\alpha_1}/g^{X_1\cdots X_L}_{\alpha_1}\right)\Pi^{X_1X_2\cdots X_L}_{\alpha_1\alpha_2} \\
&\quad \times \Pi^{X_2X_3\cdots X_L}_{\alpha_2\alpha_3}\cdots\Pi^{X_{L-2}X_{L-1}X_L}_{\alpha_{L-2}\alpha_{L-1}} k^{X_{L-1}X_L}_{\alpha_{L-1}\alpha_L},
\end{aligned}
$$

$$\begin{aligned}\Pi^{X_iX_{i+1}\cdots X_L}_{\alpha_i\alpha_{i+1}} &= k^{X_iX_{i+1}}_{\alpha_i\alpha_{i+1}}\big/\big(r^{X_iX_{i+1}}_{\alpha_i\alpha_{i+1}} + g^{X_{i+1}\cdots X_L}_{\alpha_{i+1}}\big),\\ g^{X_{i+1}\cdots X_L}_{\alpha_{i+1}} &= \Pi^{X_{i+1}X_{i+2}\cdots X_L}_{\alpha_{i+1}a} g^{X_{i+2}\cdots X_L}_{a}\\ &\quad + \Pi^{X_{i+1}X_{i+2}\cdots X_L}_{\alpha_{i+1}b} g^{X_{i+2}\cdots X_L}_{b},\\ g^{X_{L-1}X_L}_{\alpha_{L-1}} &\equiv k^{X_{L-1}X_L}_{\alpha_{L-1}a} + k^{X_{L-1}X_L}_{\alpha_{L-1}b}. \end{aligned} \tag{2}$$

Equation (2) can be transformed into a more intuitive form, a forward inhomogeneous Markov chain:

$$\begin{aligned}P^{X_1X_2\cdots X_L}_{\alpha_1\alpha_2\cdots\alpha_L} &= q^{X_1}_{\alpha_1} M^{X_1X_2\cdots X_L}_{\alpha_1\alpha_2} M^{X_2X_3\cdots X_L}_{\alpha_2\alpha_3}\\ &\quad \cdots M^{X_{L-2}X_{L-1}X_L}_{\alpha_{L-2}\alpha_{L-2}} M^{X_{L-1}X_L}_{\alpha_{L-1}\alpha_L},\\ M^{X_iX_{i+1}\cdots X_L}_{\alpha_i\alpha_{i+1}} &= \Pi^{X_iX_{i+1}\cdots X_L}_{\alpha_i\alpha_{i+1}} g^{X_{i+1}\cdots X_L}_{\alpha_{i+1}}\big/ g^{X_i\cdots X_L}_{\alpha_i},\\ M^{X_{L-1}X_L}_{\alpha_{L-1}\alpha_L} &\equiv k^{X_{L-1}X_L}_{\alpha_{L-1}\alpha_L}\big/ g^{X_{L-1}X_L}_{\alpha_{L-1}}. \end{aligned} \tag{3}$$

Here $M$ is the stochastic transfer matrix with each row sum equal to 1, i.e., $M^{X_iX_{i+1}\cdots X_L}_{\alpha_i a} + M^{X_iX_{i+1}\cdots X_L}_{\alpha_i b} = 1$. By Eq. (3) one can calculate any positional quantities of interest, e.g., the positional probability, $P^{X_m}_{\alpha_m} = \sum_{\{\alpha_i, i\neq m\}} P^{X_1\cdots X_L}_{\alpha_1\cdots\alpha_L}$, or equivalently $(P^{X_m}_a, P^{X_m}_b) = (q^{X_1}_a, q^{X_1}_b)\cdot M^{X_1\cdots X_L}\cdots M^{X_{m-1}X_m\cdots X_L}$. This forward form of the Markov chain is more convenient for approximate analytical calculations (see Sec. IV) than the backward Markov chain assumed in the iteration algorithm [17]. For instance, in the extreme case when all deletion rates are neglected, the forward transfer matrix $M$ can be intuitively and precisely written as $M^{X_{L-1}X_L}_{\alpha_{L-1}\alpha_L} = k^{X_{L-1}X_L}_{\alpha_{L-1}\alpha_L}/(k^{X_{L-1}X_L}_{\alpha_{L-1}a} + k^{X_{L-1}X_L}_{\alpha_{L-1}b})$, whereas the backward transfer matrix in the iteration algorithm can only be numerically computed.

Similarly, we also have

$$\begin{aligned}\Gamma^{X_1X_2\cdots X_m\cdots X_L}_{\alpha_1\alpha_2\cdots\alpha_m} &= q^{X_1}_{\alpha_1} M^{X_1X_2\cdots X_L}_{\alpha_1\alpha_2}\\ &\quad \cdots M^{X_{m-1}X_m\cdots X_L}_{\alpha_{m-1}\alpha_m}\big/ g^{X_m\cdots X_L}_{\alpha_m}. \end{aligned} \tag{4}$$

Note that the first-passage time (from position 1 to $L$) distribution $F(t)$ is determined by the equation $F(t) = -\frac{d}{dt}\sum_{m=1}^{L-1}\sum_{\{\alpha_1\cdots\alpha_m\}} p^{X_1X_2\cdots X_m\cdots X_L}_{\alpha_1\alpha_2\cdots\alpha_m}(t)$, then it's easy to show that the mean first-passage time $\langle T\rangle = \int_0^{+\infty} tF(t)\,dt = \sum_{m=1}^{L-1}\Gamma_m$. Here $\Gamma_m$ is defined as

$$\begin{aligned}\Gamma_m &= \sum_{\{\alpha_1\cdots\alpha_m\}} \Gamma^{X_1X_2\cdots X_m\cdots X_L}_{\alpha_1\alpha_2\cdots\alpha_m}\\ &= \sum_{\{\alpha_1\cdots\alpha_m\}} \int_0^{+\infty} p^{X_1X_2\cdots X_m\cdots X_L}_{\alpha_1\alpha_2\cdots\alpha_m}(t)\,dt\\ &= \int_0^{+\infty} p^{X_m}(t)\,dt. \end{aligned} \tag{5}$$

According to this definition, $\Gamma_m$ is exactly the mean cumulative dwell time of the growing chain of length $m$ during the first-passage process (a detailed explanation is given in Appendix A). In other words, $1/\Gamma_m$ can be regarded as the local growth velocity at position $m$. $\Gamma_m$ can also be cast in another form:

$$\begin{aligned}\Gamma_m &= \sum_{\{\alpha_1\cdots\alpha_m\}} \Gamma^{X_1X_2\cdots X_m\cdots X_L}_{\alpha_1\alpha_2\cdots\alpha_m}\\ &= P^{X_m}_a\big/ g^{X_m\cdots X_L}_a + P^{X_m}_b\big/ g^{X_m\cdots X_L}_b, \end{aligned} \tag{6}$$

which is equivalent to Eqs. (21) and (22) (the mean cumulative dwell time at position $m$ calculated by the iteration algorithm) in Ref. [17], and $g^{X_m\cdots X_L}_\alpha$ is equivalent to $v_{m_l}$ given by Eq. (18) in that paper.

Now the probability profile $P^{X_m}_{a,b}$ and the velocity profile $v_m = 1/\Gamma_m$ can be computed respectively. Figure 1 shows that the numerical results agrees perfectly well with the simulation results given by Gillespie algorithm [19].

Our first-passage (FP) calculations are also in perfect agreement with the numerical results given by the iteration algorithm (denoted as IFS in Ref. [17]), as shown by the illustrative example in Fig. 2. It should be pointed out that since the two algorithms assume different boundary conditions, the numerical results are somewhat different near the two boundaries. However, by expanding the template sequence from both ends in our FP algorithm, the difference can be largely decreased or even eliminated. For instance, the original template sequence is repeated three times to get an expanded new template, and the computed profiles of the middle copy show no difference with the results of the IFS algorithm (Fig. 2). This treatment and the expanded template are also used to obtain Fig. 3, Fig. 4, and Fig. 5.

## III. CORRELATIONS IN THE PROBABILITY PROFILE

Correlations could be present in the probability profile due to the nearest or higher-order neighbor effects. To see if there are long-range correlations in the first-order processes, we calculate the correlation function between any two template positions, say, $i$, $j$. The function is defined as $C^{X_i.X_j}_{\alpha_i.\alpha_j} = \sum_{\{\alpha_k, k\neq i,j\}} P^{X_1..X_k..X_L}_{\alpha_1..\alpha_k..\alpha_L} - P^{X_i}_{\alpha_i} P^{X_j}_{\alpha_j}$. There are four types of correlation functions. To quantify the maximal correlations, we define $C_{\max}(i,d) = \max_{\alpha_i,\alpha_{i+d}}(|C^{X_i,X_{i+d}}_{\alpha_i,\alpha_{i+d}}|)$, $C_{\max}(i,0) = 0$, for any position $i$, and correspondingly the relative correlation function $\widetilde{C}^{X_i.X_j}_{\alpha_i.\alpha_j} = C^{X_i.X_j}_{\alpha_i.\alpha_j}/(P^{X_i}_{\alpha_i} P^{X_j}_{\alpha_j})$ and $\widetilde{C}_{\max}(i,d) = \max_{\alpha_i,\alpha_{i+d}}(|\widetilde{C}^{X_i,X_{i+d}}_{\alpha_i,\alpha_{i+d}}|)$.

Under some conditions (e.g., the bio-relevant conditions such as Parameter 2, which is explained in detail in Sec. IV), either $C_{\max}(i,d)$ or $\widetilde{C}_{\max}(i,d)$ decays abruptly with the correlation length 1 [illustrated by Fig. 3(a)], implying that the positional probability is determined by its nearest neighbors. This does not hold in general, of course. For instance, the correlation length becomes much larger under some extreme conditions [e.g., Parameter 3 in Fig. 3(b)] where one can no longer identify consistent pairing rules for each type of template units. For instance, $\binom{A}{a}$ is the dominant pairing (say, with a probability larger than 0.9) only for a part of the template $A$s while $\binom{A}{b}$ is dominant for the rest, so no WC-like pairing rules ($R$ or $W$) can be universally assigned to $A$. Therefore, DNA synthesized in these cases can no longer fulfill its fundamental role as genetic material. Such extreme conditions are out of the scope of this paper and will not be discussed below. Appendix B gives a detailed explanation of how such long-range correlations may occur.

## IV. THE NEAREST-NEIGHBOR APPROXIMATION UNDER BIO-RELEVANT CONDITIONS

The nearest-neighbor correlations can be observed under the so-called biologically relevant conditions, which are

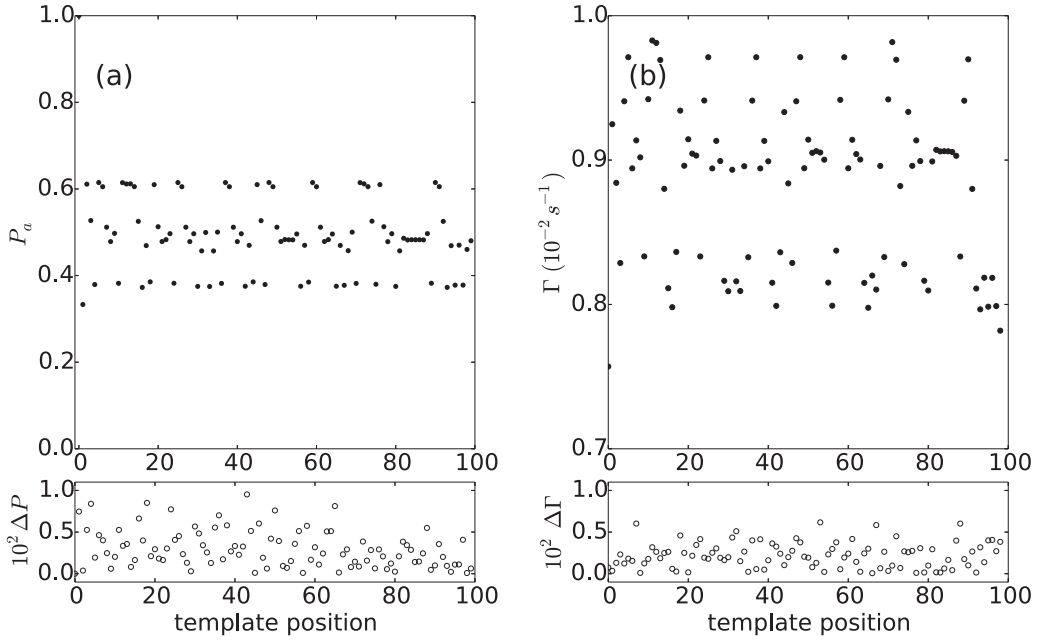


FIG. 1. The comparison between numerical and simulation results, with given kinetic parameters (Parameter 1; see Appendix D) and the random template sequence of length 100 (see Appendix D). The statistics are made over $10^5$ simulations. (a) (top) Numerical results of $P_a$ for each template position; (bottom) the relative difference $\Delta P = \max_{\alpha_m=a,b}(|P^{\rm num}_{\alpha_m} - P^{\rm sim}_{\alpha_m}|/P^{\rm num}_{\alpha_m})$. (b) (top) Numerical results of the mean cumulative dwell time $\Gamma$ for each location; (bottom) the relative difference $\Delta\Gamma = |\Gamma^{\rm num} - \Gamma^{\rm sim}|/\Gamma^{\rm num}$.

inspired by the measured kinetic parameters of real DNAPs. These conditions ensure that, compared with the replication catalyzed only by the polymerase domain of DNAP, the introduction of proofreading domain can significantly enhance the replication fidelity while still maintaining the high overall velocity.

The bio-relevant conditions for the first-order process are intuitive, as below:

(a) $k^{XY}_{RR} \gg k^{XY}_{RW}$, which means that the addition of $R$ is always much faster than that of $W$.

(b) $k^{XY}_{WR}/k^{XY}_{WW} \gg k^{XY}_{RR}/k^{XY}_{RW}$, which means that the successive additions of $W$ are almost inhibited. In fact, $k^{XY}_{WW}$ are hard to measure in experiments, so $k^{XY}_{WW} \sim 0$ are always assumed.

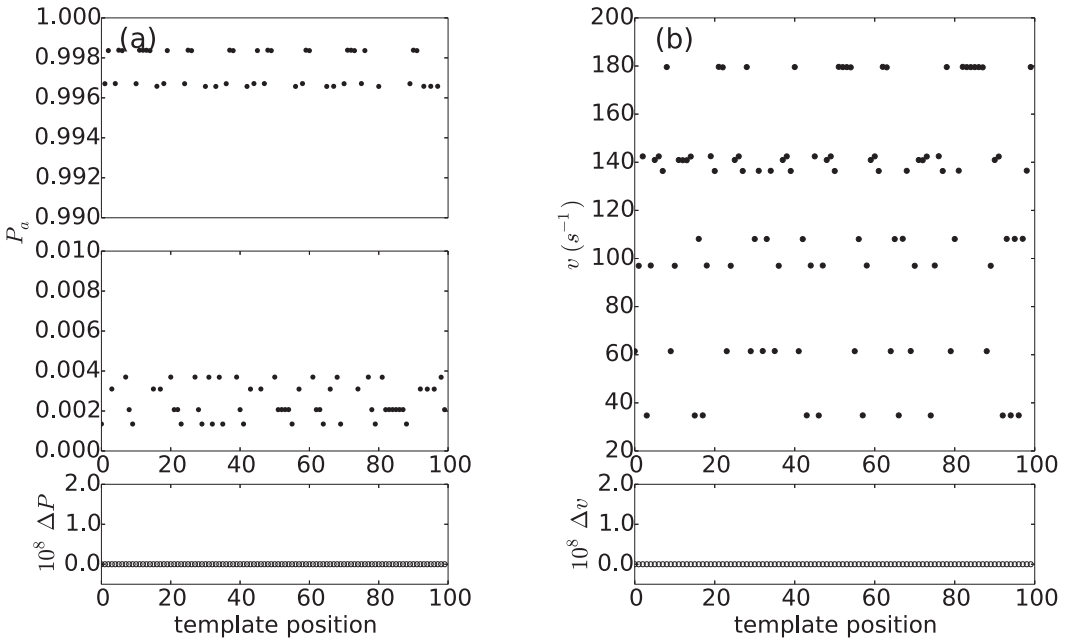


FIG. 2. Comparison between the numerical results given by FP and IFS under Parameter 2 (Appendix D). The expanded random template is used in the computations. (a) (top) Pa for each position given by the FP algorithm; (bottom) the relative difference $\Delta P = \max_{\alpha_m=a,b}|P^{FP}_{\alpha_m} - P^{IFS}_{\alpha_m}|/P^{FP}_{\alpha_m}$. (b) (top) $v$ for each position given by FP algorithm; (bottom) the relative difference $\Delta v = |v^{FP} - v^{IFS}|/v^{FP}$.

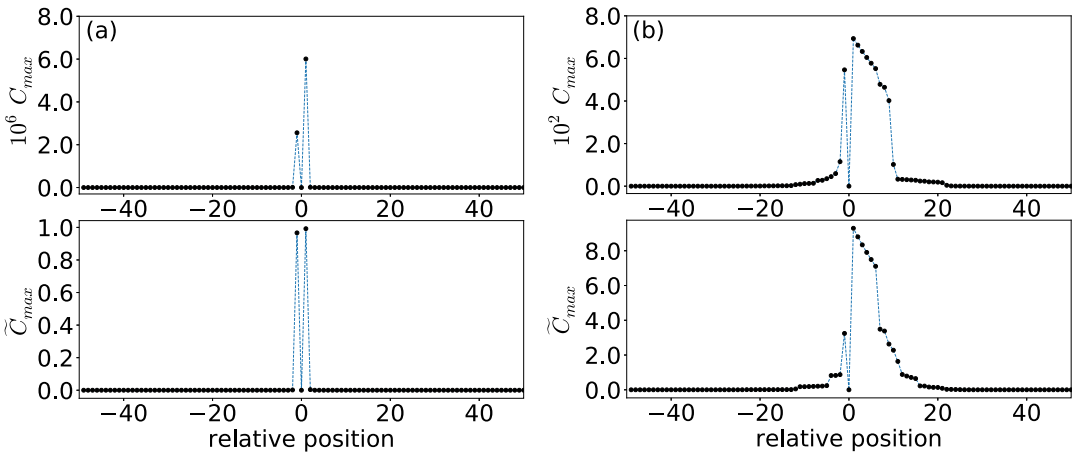


FIG. 3. The correlation $C_{\max}$ and the relative correlation $\widetilde{C}_{\max}$ between the position 50 and the rest positions of the random template. (a) Under Parameter 2 (bio-relevant conditions). (b) Under Parameter 3.

(c) $k_{RR}^{YZ} \gg r_{RR}^{XY}, r_{WR}^{XY}$, which means that the successive additions of $R$ always dominate the replication process in order to guarantee the high replication velocity (i.e., the introduction of proofreading almost does not decrease the overall velocity), at the cost that a buried $W$ is hard to proofread.

(d) $r_{WW}^{XY} > r_{RW}^{XY}$, which means that the terminus containing more $W$s is more likely to be proofread.

Here $\gg$ means that the term on the left side is more than 10 times bigger than that on the right side. These conditions are consistent with experimental observations of real DNAPs (see Sec. 3.2 in Ref. [12] for the data) and in fact are much more general (for comparison, e.g., $k_{RR}^{XY}/k_{RW}^{XY} > 10^5$ and $k_{RR}^{XY} \gg k_{WR}^{XY}$ are always observed in real DNAPs). Under such general conditions, the exact method introduced in Sec. II can be well approximated by the following method. We start from the iteration relations

$$g_{\alpha_i}^{X_i\cdots X_L} = \frac{k_{\alpha_i a}^{X_i X_{i+1}}}{1 + r_{\alpha_i a}^{X_i X_{i+1}} / g_a^{X_{i+1}\cdots X_L}} + \frac{k_{\alpha_i b}^{X_i X_{i+1}}}{1 + r_{\alpha_i b}^{X_i X_{i+1}} / g_b^{X_{i+1}\cdots X_L}},$$
$$g_{\alpha_{L-1}}^{X_{L-1}X_L} \equiv k_{\alpha_{L-1}a}^{X_{L-1}X_L} + k_{\alpha_{L-1}b}^{X_{L-1}X_L}. \tag{7}$$

Under bio-relevant conditions, one has $k_{\alpha_{L-1}R}^{X_{L-1}X_L} \gg k_{\alpha_{L-1}W}^{X_{L-1}X_L}$, so $g_{\alpha_{L-1}}^{X_{L-1}X_L} \simeq k_{\alpha_{L-1}R}^{X_{L-1}X_L}$.

The next iteration is

$$g_{\alpha_{L-2}}^{X_{L-2}X_{L-1}X_L} = \frac{k_{\alpha_{L-2}a}^{X_{L-2}X_{L-1}}}{1 + r_{\alpha_{L-2}a}^{X_{L-2}X_{L-1}} / g_a^{X_{L-1}X_L}} + \frac{k_{\alpha_{L-2}b}^{X_{L-2}X_{L-1}}}{1 + r_{\alpha_{L-2}b}^{X_{L-2}X_{L-11}} / g_b^{X_{L-1}X_L}}. \tag{8}$$

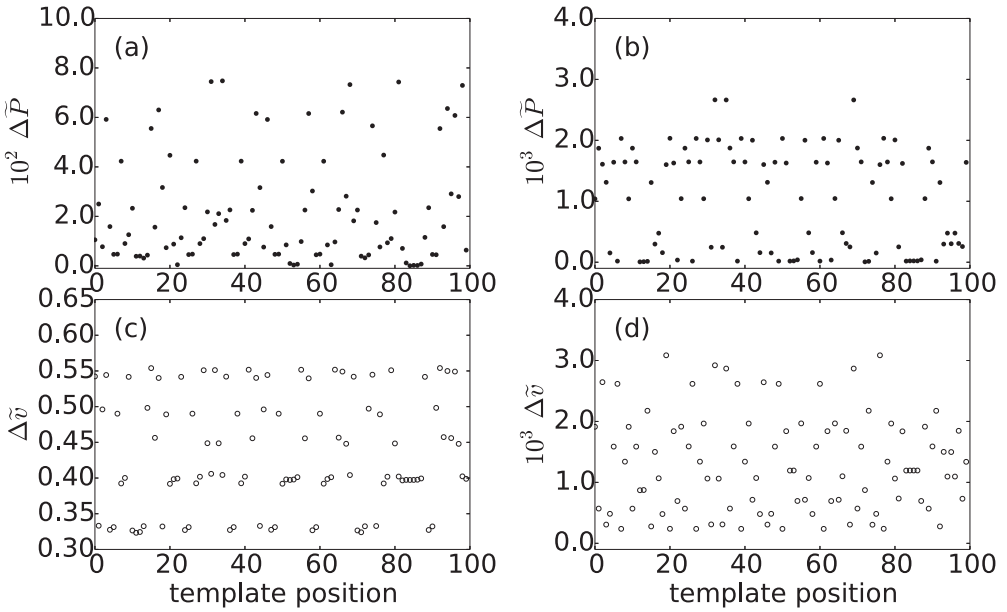


FIG. 4. Comparison between the precise (pre) and approximate (app) numerical results. $\Delta\widetilde{P}_i = \max_{\alpha_i=a,b}(|P_{\alpha_i}^{\text{app}} - P_{\alpha_i}^{\text{pre}}|/P_{\alpha_i}^{\text{pre}})$ and $\Delta\widetilde{v} = |v^{\text{app}} - v^{\text{pre}}|/v^{\text{pre}}$. (a, c) $\Delta\widetilde{P}$, $\Delta\widetilde{v}$, under Parameter 1. (b, d) $\Delta\widetilde{P}$, $\Delta\widetilde{v}$, under Parameter 2.

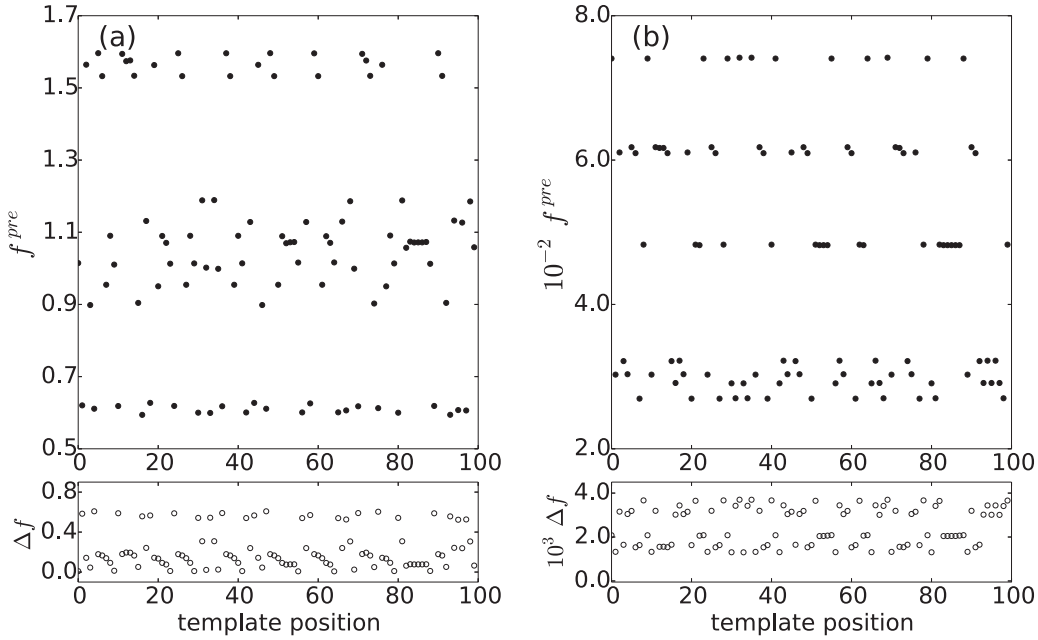


FIG. 5. Comparison between the precise (pre) and approximated (app) fidelity profile. $\Delta f = |f^{\rm pre} - f^{\rm app}|/f^{\rm pre}$. (a) $f^{\rm pre}$ (top) and $\Delta f$ (bottom), under Parameter 1. (b) $f^{\rm pre}$ (top) and $\Delta f$ (bottom), under Parameter 2.

If $X_{L-1} = A$, then $r^{X_{L-2}A}_{\alpha_{L-2}a} \ll k^{AX_L}_{aR} \simeq g^{AX_L}_{a}$ and $k^{X_{L-2}A}_{\alpha_{L-2}a} \gg k^{X_{L-2}A}_{\alpha_{L-2}b}$. This leads to $g^{X_{L-2}AX_L}_{\alpha_{L-2}} \simeq k^{X_{L-2}A}_{\alpha_{L-2}a}$. If $X_{L-1} = B$, then $r^{X_{L-2}B}_{\alpha_{L-2}b} \ll k^{BX_L}_{bR} \simeq g^{BX_L}_{b}$ and $k^{X_{L-2}B}_{\alpha_{L-2}b} \gg k^{X_{L-2}B}_{\alpha_{L-2}a}$. This leads to $g^{X_{L-2}BX_L}_{\alpha_{L-2}} \simeq k^{X_{L-2}B}_{\alpha_{L-2}b}$. Combining these two results, we get $g^{X_{L-2}X_{L-1}X_L}_{\alpha_{L-2}} \simeq k^{X_{L-2}X_{L-1}}_{\alpha_{L-2}R}$.

Following the same logic, we obtain $g^{X_iX_{i+1}\cdots X_L}_{\alpha_i} \simeq k^{X_iX_{i+1}}_{\alpha_iR}$ (denoted as $g^{X_iX_{i+1}}_{\alpha_i}$) and $\prod^{X_iX_{i+1}\cdots X_L}_{\alpha_i\alpha_{i+1}} \simeq k^{X_iX_{i+1}}_{\alpha_i\alpha_{i+1}}/(r^{X_iX_{i+1}}_{\alpha_i\alpha_{i+1}} + k^{X_{i+1}X_{i+2}}_{\alpha_{i+1}R})$ (denoted as $\prod^{X_iX_{i+1}X_{i+2}}_{\alpha_i\alpha_{i+1}}$). Correspondingly, the stochastic transfer matrix is approximated as $M^{X_iX_{i+1}X_{i+2}}_{\alpha_i\alpha_{i+1}}$ [the nearest-neighbor (NN) approximation] which can be transformed by row or column exchange into the equivalent form

$$\begin{bmatrix} M_{RR} & M_{RW} \\ M_{WR} & M_{WW} \end{bmatrix}$$

and correspondingly,

$$\begin{aligned} P^{X_1X_2\cdots X_L}_{s_1s_2\cdots s_L} &\simeq q^{X_1}_{s_1} M^{X_1X_2X_3}_{s_1s_2} \cdots M^{X_{i-1}X_iX_{i+1}}_{s_{i-1}s_i} \\ &\quad \cdots M^{X_{L-1}X_L}_{s_{L-1}s_L}, \\ s_i &= R, W \quad (i = 1, \ldots, L). \end{aligned} \tag{9}$$

Now we get the approximate expressions of the elements of $M$. For instance,

$$\begin{aligned} M^{X_iX_{i+1}X_{i+2}}_{RW} &\simeq \frac{k^{X_iX_{i+1}}_{RW}}{r^{X_iX_{i+1}}_{RW} + g^{X_{i+1}X_{i+2}}_{W}} \frac{g^{X_{i+1}X_{i+2}}_{W}}{g^{X_iX_{i+1}}_{R}} \\ &\simeq \frac{k^{X_iX_{i+1}}_{RW}}{r^{X_iX_{i+1}}_{RW} + k^{X_{i+1}X_{i+2}}_{WR}} \frac{k^{X_{i+1}X_{i+2}}_{WR}}{k^{X_iX_{i+1}}_{RR}} \\ &= \frac{k^{X_iX_{i+1}}_{RW}}{k^{X_iX_{i+1}}_{RR}} \Bigg/ \left(1 + \frac{r^{X_iX_{i+1}}_{RW}}{k^{X_{i+1}X_{i+2}}_{WR}}\right). \end{aligned} \tag{10}$$

It can be shown that $M_{RR} \gg M_{RW}$, $M_{WR} \gg M_{WW}$, and $M_{RW} \gg M_{WW}$ always hold under bio-relevant conditions. For stochastic matrices like $M$,

$$\begin{bmatrix} 1-y & y \\ 1-z & z \end{bmatrix}$$

with $z \ll y \ll 1$, one can verify that its left eigenvector associated with the largest eigenvalue 1 is approximately $P = (1-y, y)$ ($\lim_{n\to\infty} M^n$ converges to a matrix in which each row is the eigenvector $P$; for more details of the heuristic analysis see Appendix C). $P$ is a good approximation of the precise probability distribution at position $i+1$, which can be verified numerically [see Figs. 4(b) and 4(d)]. Even under some conditions different from bio-relevant conditions (Parameters 1), the NN approximation can also give results of the same orders of magnitude with the precise numerical results [Figs. 4(a) and 4(c)]. This approximation, however, fails under conditions far different from bio-relevant conditions (data not shown here; see the Supplemental Material for more details [20]).

One can also obtain the analytical expressions of the probability profile $(P^{i+1}_R, P^{i+1}_W) \simeq (M^{X_iX_{i+1}X_{i+2}}_{RR}, M^{X_iX_{i+1}X_{i+2}}_{RW})$ which thus gives the approximate fidelity profile

$$f^{X_{i+1}} \simeq \frac{M^{X_iX_{i+1}X_{i+2}}_{RR}}{M^{X_iX_{i+1}X_{i+2}}_{RW}} \simeq \frac{k^{X_iX_{i+1}}_{RR}}{k^{X_iX_{i+1}}_{RW}} \left(1 + \frac{r^{X_iX_{i+1}}_{RW}}{k^{X_{i+1}X_{i+2}}_{WR}}\right), \tag{11}$$

which includes the analytical expression of fidelity Eq. (15) in Ref. [12] (template sequence specificity ignored), as a limiting case. Such expressions show clearly how the polymerase and the exonuclease coordinate kinetically to contribute to the overall fidelity. The first factor is solely contributed by the polymerase even if no NN effects are explicitly considered. The second term, which accounts for the proofreading efficiency of the last $W$, can be contributed by the exonuclease only when NN effects do exist.

The approximate profile shows perfect agreement with the precise profile under bio-relevant conditions [Fig. 5(b)] and also provides a good estimate under some other conditions [Fig. 5(a) where the approximate and the precise velocity is of the same order of magnitude].

The NN approximation immediately leads to the conclusion that any kind of correlations in the template sequence, if one exists (e.g., the possible long-range correlations in the noncoding DNA sequences [18]), has no substantial impact on the positional quantities (data are shown in Supplemental al [20]). This is consistent with the widely acknowledged idea that DNA replication mutations are randomly distributed in the genome.

## V. GENERALIZATION TO MULTICOMPONENT SYSTEMS

The above methods can be readily generalized to more realistic cases; e.g., in real DNA replication there are four types of monomers (A, G, T, C) being added or deleted. Below we consider a general multicomponent system which consists of $n$ types of units $A_i$ $(i = 1, \ldots, n)$ in the template and $n$ types of monomers $a_i$ $(i = 1, \ldots, n)$ in the solution, and each $A_i$ forms the right pair ($R$) with only one monomer $a_i$ and forms wrong pairs with the rest monomers (denoted as $W_i, i = 1, 2, \ldots, n-1$). The corresponding bio-relevant conditions are just a simple generalization of those in the preceding section:

(a) $k_{RR}^{XY} \gg k_{RW_i}^{XY}$,

(b) $k_{W_iR}^{XY}/k_{W_iW_j}^{XY} \gg k_{RR}^{XY}/k_{RW_k}^{XY}$,

(c) $k_{RR}^{YZ} \gg r_{RR}^{XY}, r_{W_iR}^{XY}$,

(d) $r_{W_iW_j}^{XY} > r_{RW_j}^{XY}$.

Similarly we rearrange the transfer matrix $M^{X_{i-1}X_iX_{i+2}}$ to a standard form:

$$\begin{bmatrix} M_{RR} & M_{RW_1} & \cdots & M_{RW_{(n-1)}} \\ M_{WR}^{(1)} & M_{WW_1}^{(1)} & \cdots & M_{WW_{(n-1)}}^{(1)} \\ & & \cdots & \\ M_{WR}^{(n-1)} & M_{WW_1}^{(n-1)} & \cdots & M_{WW_{(n-1)}}^{(n-1)} \end{bmatrix}.$$

It can be shown that $M_{WW} \ll M_{RW} \ll 1$ in general under bio-relevant conditions, so the eigenvector $V_1$ of this matrix is approximately $(M_{RR}, M_{RW_1}, \ldots, M_{RW_{(n-1)}})$, which is a good approximation of the probability vector $P^{X_i}$.

Simple calculations give results almost the same as Eq. (10),

$$M_{RW_j}^{X_{i-1}X_iX_{i+1}} = \frac{k_{RW_j}^{X_{i-1}X_i}}{k_{RR}^{X_{i-1}X_i}} \Big/ \left(1 + \frac{r_{RW_j}^{X_{i-1}X_i}}{k_{W_jR}^{X_iX_{i+1}}}\right), \tag{12}$$

where $j = 1, \ldots, n-1$.

Now the positional fidelity at $i$ is $f^{X_i} = M_{RR}^{X_{i-1}X_iX_{i+1}}/(\sum_{j=1}^{n-1} M_{RW_j}^{X_{i-1}X_iX_{i+1}})$.

## VI. GENERALIZATION TO HIGHER ORDER PROCESSES

For $h$-order processes, we set the initial seed as a given distribution $q_{\alpha_1\cdots\alpha_h}^{X_1\cdots X_h}$. One can follow the logic of Sec. II to obtain

$$\begin{aligned} P_{\alpha_1\cdots\alpha_L}^{X_1\cdots X_L} &= \left(q_{\alpha_1\cdots\alpha_h}^{X_1\cdots X_h} \Big/ g_{\alpha_1\cdots\alpha_h}^{X_1\cdots X_h\cdots X_L}\right) \\ &\quad \times \Pi_{\alpha_1\cdots\alpha_{h+1}}^{X_1\cdots X_{h+1}\cdots X_L} \cdots \Pi_{\alpha_i\cdots\alpha_{i+h}}^{X_i\cdots X_{i+h}\cdots X_L} \\ &\quad \cdots \Pi_{\alpha_{L-h-1}\cdots\alpha_{L-1}}^{X_{L-h-1}\cdots X_{L-1}X_L} k_{\alpha_{L-h}\cdots\alpha_L}^{X_{L-h}\cdots X_L}, \\ \Pi_{\alpha_i\cdots\alpha_{i+h}}^{X_i\cdots X_{i+h}\cdots X_L} &= k_{\alpha_i\cdots\alpha_{i+h}}^{X_i\cdots X_{i+h}} \Big/ \left(r_{\alpha_i\cdots\alpha_{i+h}}^{X_i\cdots X_{i+h}} + g_{\alpha_{i+1}\cdots\alpha_{i+h}}^{X_{i+1}\cdots X_{i+h}\cdots X_L}\right), \\ g_{\alpha_{i+1}\cdots\alpha_{i+h}}^{X_{i+1}\cdots X_{i+h}\cdots X_L} &= \Pi_{\alpha_{i+1}\cdots\alpha_{i+h}a}^{X_{i+1}\cdots X_{i+h}X_{i+h+1}\cdots X_L} g_{\alpha_{i+2}\cdots\alpha_{i+h}a}^{X_{i+2}\cdots X_{i+h}X_{i+h+1}\cdots X_L} \\ &\quad + \Pi_{\alpha_{i+1}\cdots\alpha_{i+h}b}^{X_{i+1}\cdots X_{i+h}X_{i+h+1}\cdots X_L} g_{\alpha_{i+2}\cdots\alpha_{i+h}b}^{X_{i+2}\cdots X_{i+h}X_{i+h+1}\cdots X_L}, \\ g_{\alpha_{L-h}\cdots\alpha_{L-1}}^{X_{L-h}\cdots X_{L-1}X_L} &\equiv k_{\alpha_{L-h}\cdots a}^{X_{L-h}\cdots X_L} + k_{\alpha_{L-h}\cdots b}^{X_{L-h}\cdots X_L} \end{aligned} \tag{13}$$

or equivalently,

$$\begin{aligned} P_{\alpha_1\alpha_2\cdots\alpha_L}^{X_1X_2\cdots X_L} &= q_{\alpha_1\cdots\alpha_h}^{X_1\cdots X_h} M_{\alpha_1\cdots\alpha_{h+1}}^{X_1\cdots X_{h+1}\cdots X_L} \\ &\quad \cdots M_{\alpha_i\cdots\alpha_{i+h}}^{X_i\cdots X_{i+h}\cdots X_L} \cdots M_{\alpha_{L-h}\cdots\alpha_L}^{X_{L-h}\cdots X_L}, \\ M_{\alpha_i\cdots\alpha_{i+h}}^{X_i\cdots X_{i+h}\cdots X_L} &= \Pi_{\alpha_i\cdots\alpha_{i+h}}^{X_i\cdots X_{i+h}\cdots X_L} g_{\alpha_{i+1}\cdots\alpha_{i+h}}^{X_{i+1}\cdots X_{i+h}\cdots X_L} \Big/ g_{\alpha_i\cdots\alpha_{i+h-1}}^{X_i\cdots X_{i+h-1}\cdots X_L}, \\ M_{\alpha_{L-h}\cdots\alpha_L}^{X_{L-h}\cdots X_L} &\equiv k_{\alpha_{L-h}\cdots\alpha_L}^{X_{L-h}\cdots X_L} \Big/ g_{\alpha_{L-h}\cdots\alpha_{L-1}}^{X_{L-h}\cdots X_{L-1}X_L}. \end{aligned} \tag{14}$$

The NN approximations can also applied to these processes under the corresponding bio-relevant conditions. For illustration, we give only a brief introduction to the second-order processes of binary systems. The bio-relevant conditions similar to those in Sec. IV are proposed as below:

(a) $k_{\alpha\beta R}^{XYZ} \gg k_{\alpha\beta W}^{XYZ}$, $\alpha, \beta = R, W$, which means that the addition rates of $R$ are always much larger than that of $W$.

(b) $\tilde{k}_{\alpha\beta R}^{XYZ}/\tilde{k}_{\alpha\beta W}^{XYZ} \gg \tilde{k}_{RRR}^{XYZ}/\tilde{k}_{RRW}^{XYZ}$, $\alpha\beta = RW, WR, WW$. $\tilde{k}_{\alpha\beta\gamma}^{XYZ} \equiv k_{\alpha\beta\gamma}^{XYZ}/(1 + r_{\alpha\beta\gamma}^{XYZ}/k_{\beta\gamma R}^{YZU})$ is approximately the renormalized addition rates. In fact, here $k_{\alpha\beta W}^{XYZ} \simeq 0$ are always assumed since successive additions of $W$ are almost inhibited. So these conditions are naturally satisfied.

(c) $k_{RRR}^{YZU} \gg r_{RRR}^{XYZ}, r_{WRR}^{XYZ}$, which means that the successive additions of $R$ always dominate the overall replication process.

(d) $r_{WWR}^{XYZ} > r_{RWR}^{XYZ}$, which means that the terminus containing more $W$s is more likely to be proofread.

(e) $r_{\alpha\beta W}^{XYZ}/k_{\beta WR}^{YZU} > r_{\alpha\beta R}^{XYZ}/k_{\beta RR}^{YZU}$, $\alpha, \beta = R, W$, which mean that the terminal $W$ is always more probable to be deleted than the terminal $R$.

To calculate the positional quantities at position $i$, we first obtain the transfer matrix $M^{X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}}$ by the following two iterations, starting from $g_{\alpha_i\alpha_{i+1}}^{X_iX_{i+1}X_{i+2}} \simeq k_{\alpha_i\alpha_{i+1}a}^{X_iX_{i+1}X_{i+2}} + k_{\alpha_i\alpha_{i+1}b}^{X_iX_{i+1}X_{i+2}} \simeq k_{\alpha_i\alpha_{i+1}R}^{X_iX_{i+1}X_{i+2}}$:

$$\begin{aligned} \textbf{(I)}\ \Pi_{\alpha_{i-1}\alpha_i\alpha_{i+1}}^{X_{i-1}X_iX_{i+1}X_{i+2}} &= k_{\alpha_{i-1}\alpha_i\alpha_{i+1}}^{X_{i-1}X_iX_{i+1}} \Big/ \left(r_{\alpha_{i-1}\alpha_i\alpha_{i+1}}^{X_{i-1}X_iX_{i+1}} + g_{\alpha_i\alpha_{i+1}}^{X_iX_{i+1}X_{i+2}}\right), \\ g_{\alpha_{i-1}\alpha_i}^{X_{i-1}X_iX_{i+1}X_{i+2}} &= \Pi_{\alpha_{i-1}\alpha_ia}^{X_{i-1}X_iX_{i+1}X_{i+2}} g_{\alpha_ia}^{X_iX_{i+1}X_{i+2}} \\ &\quad + \Pi_{\alpha_{i-1}\alpha_ib}^{X_{i-1}X_iX_{i+1}X_{i+2}} g_{\alpha_ib}^{X_iX_{i+1}X_{i+2}}, \\ \textbf{(II)}\ \Pi_{\alpha_{i-2}\alpha_{i-1}\alpha_i}^{X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}} &= k_{\alpha_{i-2}\alpha_{i-1}\alpha_i}^{X_{i-2}X_{i-1}X_i} \Big/ \left(r_{\alpha_{i-2}\alpha_{i-1}\alpha_i}^{X_{i-2}X_{i-1}X_i} + g_{\alpha_{i-1}\alpha_i}^{X_{i-1}X_iX_{i+1}X_{i+2}}\right), \\ g_{\alpha_{i-2}\alpha_{i-1}}^{X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}} &= \Pi_{\alpha_{i-2}\alpha_{i-1}a}^{X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}} g_{\alpha_{i-1}a}^{X_{i-1}X_iX_{i+1}X_{i+2}} \\ &\quad + \Pi_{\alpha_{i-2}\alpha_{i-1}a}^{X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}} g_{\alpha_{i-1}b}^{X_{i-1}X_iX_{i+1}X_{i+2}}, \\ M_{\alpha_{i-2}\alpha_{i-1}\alpha_i}^{X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}} &= \Pi_{\alpha_{i-2}\alpha_{i-1}\alpha_i}^{X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}} \\ &\quad \times g_{\alpha_{i-1}\alpha_i}^{X_{i-1}X_iX_{i+1}X_{i+2}} \Big/ g_{\alpha_{i-2}\alpha_{i-1}}^{X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}}. \end{aligned}$$

$M$ can be rewritten as a first-order Markov transfer matrix, with four rows indexed as $^{X_{i-2}X_{i-1}}_{s_{i-2}s_{i-1}}$ ($RR, RW, WR, WW$ from up to bottom) and four columns indexed as $^{X_{i-1}X_i}_{s_{i-1}s_i}$ ($RR, RW, WR, WW$ from left to right), $s = R, W$:

$$\begin{bmatrix} M_{RRR} & M_{RRW} & 0 & 0 \\ 0 & 0 & M_{RWR} & M_{RWW} \\ M_{WRR} & M_{WRW} & 0 & 0 \\ 0 & 0 & M_{WWR} & M_{WWW} \end{bmatrix}.$$

It can be shown that $M_{WWW}, M_{RWW}, M_{WRW} \ll M_{RRW} \ll 1$. The eigenvector $V_1$ of this matrix is approximately $(1 - 2M_{RRW}, M_{RRW}, M_{RRW}, 0)$, which can be regarded as the positional probability $P^{X_{i-1}X_i} = (P^{X_{i-1}X_i}_{RR}, P^{X_{i-1}X_i}_{RW}, P^{X_{i-1}X_i}_{WR}, P^{X_{i-1}X_i}_{WW})$.

To be specific,

$$M^{X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}}_{R_{i-2}R_{i-1}W_i} \simeq \frac{k^{X_{i-2}X_{i-1}X_i}_{R_{i-2}R_{i-1}W_i}}{k^{X_{i-2}X_{i-1}X_i}_{R_{i-2}R_{i-1}R_i}} \Bigg/ \left[1 + \frac{r^{X_{i-2}X_{i-1}X_i}_{R_{i-2}R_{i-1}W_i}}{k^{X_{i-1}X_iX_{i+1}}_{R_{i-1}W_iR_{i+1}}}\left(1 + \frac{r^{X_{i-1}X_iX_{i+1}}_{R_{i-1}W_iR_{i+1}}}{k^{X_iX_{i+1}X_{i+2}}_{W_iR_{i+1}R_{i+2}}}\right)\right]. \tag{15}$$

The positional probability at position $i$ can be calculated by $P^{X_i}_R = P^{X_{i-1}X_i}_{RR} + P^{X_{i-1}X_i}_{WR} = 1 - M_{RRW} = M_{RRR}(\simeq 1)$, $P^{X_i}_W = P^{X_{i-1}X_i}_{RW} + P^{X_{i-1}X_i}_{WW} = M_{RRW}$. So the fidelity at position $i$ is

$$f^{X_i} \simeq 1/M^{X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}}_{R_{i-2}R_{i-1}W_i} \simeq \frac{k^{X_{i-2}X_{i-1}X_i}_{R_{i-2}R_{i-1}R_i}}{k^{X_{i-2}X_{i-1}X_i}_{R_{i-2}R_{i-1}W_i}} \times \left[1 + \frac{r^{X_{i-2}X_{i-1}X_i}_{R_{i-2}R_{i-1}W_i}}{k^{X_{i-1}X_iX_{i+1}}_{R_{i-1}W_iR_{i+1}}}\left(1 + \frac{r^{X_{i-1}X_iX_{i+1}}_{R_{i-1}W_iR_{i+1}}}{k^{X_iX_{i+1}X_{i+2}}_{W_iR_{i+1}R_{i+2}}}\right)\right], \tag{16}$$

which agrees with Eq. (15) in Ref. [12]. Such expressions show intuitively the contributions of the polymerase and the exonuclease to the overall fidelity: the first factor is contributed solely by the polymerase even if no neighbor effects are explicitly considered, but the second factor can only be contributed by the exonuclease when the second-order neighbor effects do exist. In particular, the second factor consists of two terms corresponding to the proofreading of the last $W$ and the penultimate $W$, respectively, which means higher proofreading efficiency can be obtained by high-order neighbor effects [compared with Eq. (11)]. This may explain the extreme high fidelity of some eukaryotic DNAPs.

The above logic can be directly extended to $h$-order processes. Under the corresponding bio-relevant conditions, one can show that the $(2h+1)$-neighbors $X_{i-h}\cdots X_{i-1}(X_i)X_{i+1}\cdots X_{i+h}$ contribute overwhelmingly to $\Pi^{X_{i-h}\cdots X_i\cdots X_L}_{\alpha_{i-h}\cdots\alpha_i}$. With this generalized NN approximation, we can readily calculate any positional quantities at $i$ by assuming $g^{X_i\cdots X_{i+h-1}\cdots X_L}_{\alpha_i\cdots\alpha_{i+h-1}} \simeq k^{X_i\cdots X_{i+h-1}X_{i+h}\cdots X_L}_{\alpha_i\cdots\alpha_{i+h-1}R}$.

## VII. SUMMARY

Studies on the template-specific fidelity of DNAPs are important to understand how genetic mutations are generated and controlled. While biochemical experiments have offered many insights on this issue, systematic theoretical studies are still rare. The only work appeared two years ago [17], which dealt with the long-time limit of the replication kinetics and proposed an iteration algorithm to numerically compute the fidelity or velocity profile. In this paper, we proposed a different method, based on the first-passage description of the replication process, to address these problems for complicated processes with high-order neighbor effects. Although the boundary conditions in our method are different from the periodic boundary condition in the iteration algorithm, it was verified numerically that these two choices always give the same results.

Our method, however, largely simplifies the calculations by introducing a closed set of kinetic equations and is more convenient for approximate analytical calculations. We showed that the positional fidelity and velocity can be reliably estimated by the nearest-neighbor approximations under bio-relevant conditions. The analytical expressions of the positional fidelity were derived, which show intuitively how the template-dependent proofreading pathways could be coordinated with the polymerization pathways to achieve high fidelity. These results also indicate that the positional quantities are dependent only on the closely surrounding template sequence and irrelevant to the statistical features (e.g., long-range correlations) of the template sequence, which is consistent with the widely held belief that replication mutations are randomly distributed among the genome. This is also a justification of the somewhat arbitrary choices of the template sequence (e.g., any expanded sequence containing the sequence under study can be chosen as the template) and the initial condition (at the reflecting boundary) in our method.

Our method can also be applied to more realistic cases in which either the addition or the deletion of monomers consists of multiple substeps. The widely used models in biochemistry are nonsuccessive excision models (Model II in Ref. [12]; also see the biochemical references therein), and the template-specific fidelity has been investigated in previous works [14,15] by using the IFS algorithm and additional steady-state assumptions which are usually adopted to handle multistep processes in biochemistry (e.g., the well-known Michaelis-Menten kinetics). Our FP method, however, can be directly applied to such models without appealing for any steady-state assumptions. Moreover, successive excision models are also possible in principle (e.g., Model I in Ref. [12], or equivalently the model in Ref. [21]), though they are rarely considered in the literature and the template effects have never been discussed. Our method can be slightly modified to handle such models. Comprehensive discussions on the above issues are too much lengthy to be presented here and will be published elsewhere.

## ACKNOWLEDGMENTS

The authors acknowledge the financial support by National Natural Science Foundation of China (Grants No.11675180, No. 11574329, and No. 11774358), Key Research Program of Frontier Sciences of CAS (Grant No. Y7Y1472Y61), the CAS Biophysics Interdisciplinary Innovation Team Project (Grant No.2060299), the CAS Strategic Priority Research Program (Grant No. XDA17010504), and the Joint NSFC-ISF Research Program (Grant No. 51561145002).

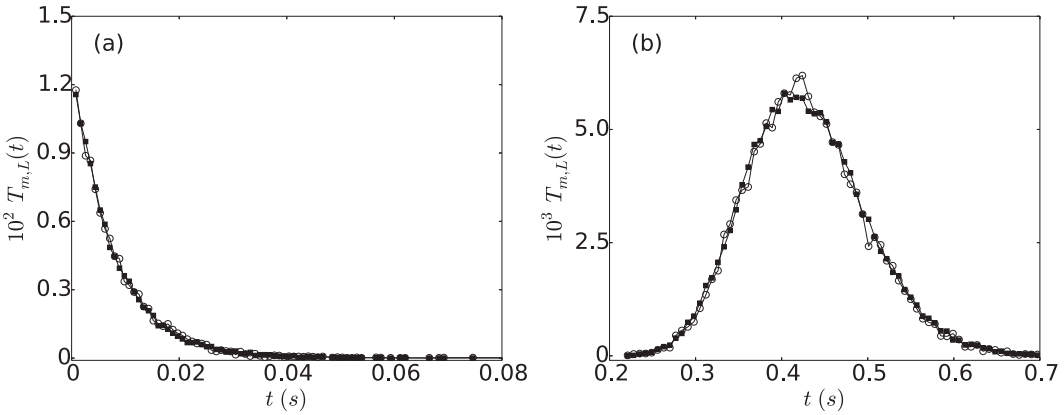


FIG. 6. Verification of Eq. (A1) by simulations for the random template. The statistics are made over $10^5$ simulations under Parameter 1. Open circles: Cumulative dwell time distribution reconstructed from simulations of the original FP process (from the template position 1 to L). Filled squares: Cumulative dwell time distribution reconstructed according to Eq. (A1), from simulations of the new FP processes (from the position $m$ to L with $a$ or $b$ at $m$). (a) $T_{99,100}(t)$; (b) $T_{51,100}(t)$.

## APPENDIX A: THE CUMULATIVE DWELL TIME DISTRIBUTION AT POSITION *m*

We mentioned in the main text that $\Gamma_m$ is the mean cumulative dwell time of the growing chain at template position $m$ (which may be revisited many times in a simulation), according to its definition $\Gamma_m = \int_0^{+\infty} p^{X_m}(t)\,dt$. To understand this, one can imagine $N$ simulation trajectories generated by the Gillespie algorithm (the first-passage process is divided into infinitesimal intervals $dt$) and select the $N_m(t)$ trajectories in which the growing end stays at position $m$ at time $t$, to get a statistics of $p^{X_m}(t) = N_m(t)/N$ as well as the infinitesimal dwell time $dt$ at $m$. As the copolymerization proceeds, the total dwell time at $m$ contributed from all the $N$ trajectories should be $\int_0^{+\infty} N_m(t)\,dt$, hence the mean cumulative dwell time per trajectory is given by $\int_0^{+\infty} p^{X_m}(t)\,dt$.

One can further investigate the cumulative dwell time distribution at $m$. Denote the total cumulative dwell time that the growing chain spends when its length $n$ reaches gion $m \leqslant n < L$ as $t$, and define the corresponding time distribution as $T_{m,L}(t)$, then we have $\int_0^{\infty} T_{m,L}(t)\,dt = 1$, and it's also known from above that $\int_0^{\infty} t T_{m,L}(t)\,dt = \sum_{n=m}^{L} \Gamma_n$. From the simulation results, we found that $T_{m,L}(t)$ can be precisely expressed as

$$T_{m,L}(t) = \sum_{\alpha_m=a,b} P_{\alpha_m}^{X_m} T_{m,L}^{\alpha_m}(t). \tag{A1}$$

$P_{\alpha_m}^{X_m}$ is the final probability distribution at $m$, as calculated in the main text. $T_{m,L}^{\alpha_m}(t)$ is defined as the first-passage time distribution of a new imaginary replication process which initiates at the template position $m$ with initial conditions $q_{\alpha_m}^{X_m} = 1$ ($\alpha_m = a$ or $b$) and again terminates at position $L$. This equation can be precisely verified by numerical calculations (Fig. 6).

One can also calculate the positional cumulative dwell time distribution $T_m(t)$, by using the convolution relation

$$T_{m,L}(t) = \int_0^t T_m(\tau) T_{m+1,L}(t-\tau)\,d\tau$$

with

$$\begin{aligned} T_{L-1}(t) &\equiv T_{L-1,L}(t) \\ &= \sum_{\alpha_{L-1}=a,b} P_{\alpha_{L-1}}^{X_{L-1}} \left(k_{\alpha_{L-1}a}^{X_{L-1}X_L} + k_{\alpha_{L-1}b}^{X_{L-1}X_L}\right) \\ &\quad \times e^{-(k_{\alpha_{L-1}a}^{X_{L-1}X_L} + k_{\alpha_{L-1}b}^{X_{L-1}X_L})t}. \end{aligned}$$

## APPENDIX B: THE CORRELATION FUNCTION

Defining $2 \times 2$ matrix $B^{i,j} \equiv M^{X_i X_{i+1} \cdots L} \cdots M^{X_{j-1} X_j \cdots L}$, we can rewrite the correlation function (defined in Sec. III) as

$$\begin{aligned} C_{\alpha_i,\alpha_j}^{X_i,X_j} &= P_{\alpha_i}^{X_i} \left( B_{\alpha_i,\alpha_j}^{i,j} - \sum_{\alpha_i} P_{\alpha_i}^{X_i} B_{\alpha_i,\alpha_j}^{i,j} \right) \\ &= P_{\alpha_i}^{X_i} P_{\tilde{\alpha}_i}^{X_i} \left( B_{\alpha_i,\alpha_j}^{i,j} - B_{\tilde{\alpha}_i,\alpha_j}^{i,j} \right). \end{aligned}$$

$\alpha_i = a, b$, and $\tilde{\alpha}_i$ denotes the monomer different from $\alpha_i$.

The stochastic matrix $M$ satisfies row normalization, and the multiplication of two such matrices results in a new stochastic matrix. For $j = i + 2$,

$$B^{i,i+2} = \begin{bmatrix} x & 1-x \\ y & 1-y \end{bmatrix} \begin{bmatrix} x' & 1-x' \\ y' & 1-y' \end{bmatrix} = \begin{bmatrix} c & 1-c \\ d & 1-d \end{bmatrix}$$

TABLE I. Random template.

| 1–10 | 11–20 | 21–30 | 31–40 | 41–50 |
|---|---|---|---|---|
| BAABAAABBB | AAAAABABAA | BBBBAAABBB | ABBABBAAAB | BBABAABAAA |
| 51–60 | 61–70 | 71–80 | 81–90 | 91–100 |
| BBBBBBABAA | ABBBBABABB | AAAABAABBB | ABBBBBBBBA | AABABABABB |

TABLE II. Kinetic parameters ($s^{-1}$, simulation time unit).

| Template | Parameters | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | | | 2 | | | | 3 | | | |
| Di-unit | AA | AB | BA | BB | AA | AB | BA | BB | AA | AB | BA | BB |
| $k_{aa}$ | 65.0 | 45.0 | 76.0 | 45.0 | 250.0 | 0.42 | 0.52 | 0.0001 | 12 344.0 | 55 325.0 | 43.0 | 5436.0 |
| $k_{ab}$ | 68.0 | 45.0 | 64.0 | 97.0 | 0.77 | 200.0 | 0.0001 | 0.8 | 34.0 | 6325.0 | 2456.0 | 54.0 |
| $k_{ba}$ | 54.0 | 95.0 | 56.0 | 78.0 | 0.14 | 0.0001 | 150.0 | 0.56 | 3432.0 | 342.0 | 243.0 | 5456.0 |
| $k_{bb}$ | 45.0 | 66.0 | 80.0 | 67.0 | 0.0001 | 0.92 | 0.69 | 300.0 | 657 890.0 | 3424.0 | 54.0 | 1324.0 |
| $r_{aa}$ | 12.0 | 23.0 | 7.0 | 4.0 | 0.0065 | 0.018 | 0.026 | 2.0 | 314.0 | 3244.0 | 543.0 | 32.0 |
| $r_{ab}$ | 16.0 | 24.0 | 16.0 | 4.0 | 0.033 | 0.0007 | 3.0 | 0.011 | 2.0 | 3.0 | 434.0 | 2.0 |
| $r_{ba}$ | 22.0 | 9.0 | 17.0 | 28.0 | 0.036 | 5.0 | 0.0018 | 0.067 | 3.0 | 4.0 | 543.0 | 234.0 |
| $r_{bb}$ | 14.0 | 23.0 | 12.0 | 19.0 | 2.0 | 0.046 | 0.098 | 0.0015 | 43.0 | 5.0 | 73.0 | 12.0 |

Parameter 1: the addition rates and deletions rates are of the same orders of magnitude, which is different from the bio-relevant conditions.
Parameter 2: bio-relevant conditions in which $R$ and $W$ (base pairs) can be uniquely defined for each template unit (say, $A$-$a$, $B$-$b$).
Parameter 3: all the rates are randomly assigned, which strongly violates the bio-relevant conditions: no $R$ or $W$ can be properly defined for each template unit.

in which $c-d=(x-y)(x'-y')$, and $0<x,y<1$. Similarly,

$$B^{i,i+3}=\begin{bmatrix} c & 1-c \\ d & 1-d \end{bmatrix}\begin{bmatrix} x'' & 1-x'' \\ y'' & 1-y'' \end{bmatrix}=\begin{bmatrix} e & 1-e \\ f & 1-f \end{bmatrix}$$

in which $e-f=(c-d)(x''-y'')$. By the same logic, we finally get

$$\left|C^{X_i,X_j}_{\alpha_i,\alpha_j}\right|=P^{X_i}_a P^{X_i}_b \prod_{m=i}^{j-1}\left|M^{X_m X_{m+1}\cdots X_L}_{aa}-M^{X_m X_{m+1}\cdots X_L}_{ba}\right|.$$

This formula shows that correlation decays abruptly when the transfer matrix is close to

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad \text{(M1)}$$

and decays slowly if the transfer matrix is close to

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \qquad \text{(M2)}$$

Under the bio-relevant conditions like Parameter 2, the transfer matrix $M$ at each template position is like M1, resulting in the abrupt decay of correlation and allowing us to make the nearest-neighbor assumptions at each position. For Parameter 3, however, the transfer matrix $M$ at some position is like M2, and thus the correlation decays slowly [e.g., Fig. 3(b)].

## APPENDIX C: THE EIGENVECTOR APPROXIMATION

Aperiodic and irreducible stochastic matrices like $M$ have an important property according to the Perron-Frobenius theorem, i.e., their largest eigenvalue is $\lambda_1=1$, which always associates with one and only one positive eigenvector $V_1$ being properly normalized to 1. Other eigenvalues and eigenvectors are denoted as $\lambda_i$ and $V_i$, $i=2,3,\ldots,n$ is the dimension of the matrix. For stochastic matrices like $M$ under bio-relevant conditions, $\lambda_i (i\geqslant 2)$ are always far less than 1. Any probability distribution vector $P$ can be decomposed as $P=V_1+\sum_{i>1}s_i V_i$, so $PM=V_1+\sum_{i>1}\lambda_i s_i V_i$. If $PM$ does not differ much from $P$, the second summation in the above equality is always far less than $V_1$, so $PM$ can be approximated by $V_1$.

On the other hand, we also know that $(P^{X_i}_R, P^{X_i}_W)=(P^{X_{i-1}}_R, P^{X_{i-1}}_W)\cdot M^{X_{i-1}X_iX_i+1}$, and $(P^{X_i}_R, P^{X_i}_W)$ is indeed not too different from $(P^{X_{i-1}}_R, P^{X_{i-1}}_W)$ [they both are around (1,0)]. So we can safely approximate $(P^{X_i}_R, P^{X_i}_W)$ by the eigenvector $V_1$ of the matrix $M^{X_{i-1}X_iX_i+1}$.

## APPENDIX D: THE TEMPLATE SEQUENCES AND KINETIC PARAMETERS

The DNA template and kinetic parameters used in the numerical computations and simulations in the main text are shown in Tables I and II.

In Sec. IV it has been shown that our first-passage approach and nearest-neighbor approximation can reliably reproduce the fidelity and velocity profile under bio-relevant conditions, which means that these positional quantities are irrelevant to the long-range properties of the template sequence. To better illustrate this, we have carried out numerical computations for a Markov chain template sequence (Table III) in which the probability of consecutive $A$s (or $B$s) is taken as 0.8. Our results (see the Supplemental Material [20]) clearly show that the NN approximation still holds for such strongly correlated template sequences.

TABLE III. Markov template.

| 1–10 | 11–20 | 21–30 | 31–40 | 41–50 |
|---|---|---|---|---|
| AAAAAAAAAA | AAAAAAAAAA | BBBBBBBBBB | BBBBBBBBBB | BBBBBBBBBA |
| 51–60 | 61–70 | 71–80 | 81–90 | 91–100 |
| AAAAABBAAA | AAAAAAAAAA | AAABBBBBBB | BBBBBBBBBB | BBBAAAAAAA |

[1] J. D. Watson and F. H. C. Crick, Nature (London) **171**, 737 (1953).
[2] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **71**, 4135 (1974).
[3] J. Ninio, Biochimie **57**, 587 (1975).
[4] I. R. Lehman, M. J. Bessman, E. S. Simms, and A. Kornberg, J. Biol. Chem. **233**, 163 (1958).
[5] T. A. Kunkel and K. Bebenek, Ann. Rev. Biochem. **69**, 497 (2000).
[6] S. S. Patel, I. Wong, and K. A. Johnson, Biochemistry **30**, 511 (1991).
[7] I. Wong, S. S. Patel, and K. A. Johnson, Biochemistry **30**, 526 (1991).
[8] Y.-C. Tsai and K. A. Johnson, Biochemistry **45**, 9675 (2006).
[9] K. A. Johnson, BBA-Proteins Proteomics **1804**, 1041 (2010).
[10] P. Gaspard and D. Andrieux, J. Chem. Phys. **141**, 044908 (2014).
[11] Y.-G. Shu, Y.-S. Song, Zhong-Can, Ou-Yang, and M. Li, J. Phys.: Condens. Matter **27**, 235105 (2015).
[12] Y.-S. Song, Y.-G. Shu, X. Zhou, Z.-C. Ou-Yang, and M. Li, J. Phys.: Condens. Matter **29**, 025101 (2017).
[13] P. Gaspard, J. Stat. Phys. **164**, 17 (2016).
[14] P. Gaspard, Phys. Rev. E **93**, 042419 (2016).
[15] P. Gaspard, Phys. Rev. E **93**, 042420 (2016).
[16] P. Gaspard, Phys. Rev. Lett. **117**, 238101 (2016).
[17] P. Gaspard, Phys. Rev. E **96**, 042403 (2017).
[18] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[19] D. T. Gillespie, J. Phys. Chem. **81**, 2340 (1977).
[20] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.100.012131 for data with different templates under different conditions.
[21] A. K. Sharma and D. Chowdhury, J. Phys.: Condens. Matter **25**, 374105 (2013).