The 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 Li1∗

1School of Physical Science, University of Chinese Academy of Sciences
2Institute of Theoretical Physics, Chinese Academy of Sciences
(Dated: April 9, 2019)

DNA replication fidelity is a critical issue in molecular biology. Biochemical experiments have provided key insights on the mechanism of fidelity control by DNAP 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 proposed 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 shows intuitively how some key kinetic pathways are coordinated to guarantee the high fidelity, as well as the high velocity, of the replication processes. It was 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.

PACS numbers: 87.10.Ed, 82.39.-k, 87.15.R-

I. INTRODUCTION

Since the Watson-Crick (WC) base-pairing rules of double-strand DNA were discovered [1], template-directed DNA replication has become 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 dominate 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 10⁻¹⁰ to 10⁻¹⁴, 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 – 4kBT which cannot account for such low error rates if estimated by Boltzmann factor. As pointed out by J.Hopfield[2] and J.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 (DNAP) [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 describes 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 modelling (e.g, see the historical literatures [2,3] or more recent publications like [4,5]). 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 [6,11]. 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, P.Gaspard has considered all the 16 types of base pairs in the kinetic models and handled the high-order neighbor effects successfully [13,14,15]. 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 to propose 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, the iteration algorithm goes through the entire template sequence cyclically for numerical convergence, which indicates that the fidelity at any position may depend on the entire sequence. This is doubtful, for it’s a hard to conceive that replication mutations at different positions have long-range correlations rather than 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 DNA replication 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 \((A)\) as \(R\) and \((B)\) as \(W\). Similarly, we denote \((B)\) as \(R\) and \((A)\) 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 pre-existing 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 DNP 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 to note 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 \(k_{XY}^{a}\) and \(r_{XY}^{a}\) respectively. \((X)\) presents the preceding base pair and \((Y)\) is the template unit to which the monomer \(a\) is paired, \(X, Y = A, B\) and \(\alpha, \beta = a, b\). The termination step occurs with the addition rate of \(k_{XY}^{a}\). It should be noted that all the kinetic parameters here are effective rates.

For instance, \(k_{XY}^{a}\) is in fact the effective polymerization rate which is controlled by several sub-steps and dependent on the monomer concentrations. In this manuscript, we will not go into such details.

The probability of the growing chain sequence \(\alpha_{1}\alpha_{2}...\alpha_{i}(1 \leq i \leq L)\) at time \(t\) is denoted as \(p^{\alpha_{1}\alpha_{2}...\alpha_{i}}_{\alpha_{i}+1}(t)\). Now we have the following master equations.

\[
\begin{align*}
\dot{p}^{X_{1}...X_{L}}_{\alpha_{1}} &= r_{X_{1}}^{a} X_{1} p^{X_{1}...X_{L}}_{\alpha_{1}^{a}} + r_{X_{1}}^{b} X_{1} p^{X_{1}...X_{L}}_{\alpha_{1}^{b}} \\
&- \left( k_{X_{1}X_{2}}^{a} + k_{X_{1}X_{2}}^{b} \right) p^{X_{1}...X_{L}}_{\alpha_{1}^{a}} \\
\dot{p}^{X_{1}...X_{L}}_{\alpha_{1}} &= k_{X_{1}X_{2}}^{a} X_{1} p^{X_{1}...X_{L}}_{\alpha_{1}^{a}} + r_{X_{1}X_{2}}^{b} X_{1} p^{X_{1}...X_{L}}_{\alpha_{1}^{b}} \\
&+ r_{X_{1}X_{2}}^{a} X_{1} p^{X_{1}...X_{L}}_{\alpha_{1}^{a}} \\
&- \left( r_{X_{1}X_{2}}^{a} + r_{X_{1}X_{2}}^{b} \right) p^{X_{1}...X_{L}}_{\alpha_{1}^{a}} \\
\dot{p}^{X_{1}...X_{L-1}X_{L}}_{\alpha_{1}...\alpha_{i-1}} &= k_{X_{L-1}X_{L}}^{a} X_{1} p^{X_{1}...X_{L-1}X_{L}}_{\alpha_{1}...\alpha_{i-1}} \\
&- \left( r_{X_{L-1}X_{L}}^{a} + r_{X_{L-1}X_{L}}^{b} \right) p^{X_{1}...X_{L-1}X_{L}}_{\alpha_{1}...\alpha_{i-1}} \\
\dot{p}^{X_{1}...X_{L}}_{\alpha_{1}...\alpha_{i-1}} &= k_{X_{L-1}X_{L}}^{a} X_{1} p^{X_{1}...X_{L-1}X_{L}}_{\alpha_{1}...\alpha_{i-1}} \\
&- \left( r_{X_{L-1}X_{L}}^{a} + r_{X_{L-1}X_{L}}^{b} \right) p^{X_{1}...X_{L-1}X_{L}}_{\alpha_{1}...\alpha_{i-1}} \\
\end{align*}
\]

One of our major concerns is the final sequence distribution of the nascent chain, i.e., the long-time limit \(p^{X_{1}...X_{L}}_{\alpha_{1}...\alpha_{L}} = p^{X_{1}...X_{L}}(t \to \infty)\). To calculate it, we assume the initial conditions \(p^{X_{1}...X_{L}}_{\alpha_{1}...\alpha_{i}}(t = 0) = q^{\alpha_{1}}, q^{a_{1}} + q^{b_{1}} = 1 (q^{a_{1}}\text{ can be arbitrarily chosen. It has negligible impacts on the fidelity profile except few positions near the reflecting boundary})\) and the long-time limits \(p^{X_{1}...X_{L}}(t \to \infty) = 0 \text{ (} i \geq 2 \text{)}\) and \(p^{X_{1}...X_{L}}(t \to \infty) = 0 \text{ (} 1 \leq i < L \text{)}\). We integrate (denoting \(\Gamma^{X_{1}...X_{L}}_{\alpha_{1}...\alpha_{i}} = \int_{0}^{\infty} p^{X_{1}...X_{L}}_{\alpha_{1}...\alpha_{i}}(t)dt\)) and solve the above equations to obtain the following iteration relations.

\[
\begin{align*}
\Pi^{X_{1}X_{2}...X_{L}}_{\alpha_{1}...\alpha_{i-1}} &= \left( q^{\alpha_{1}} \right)^{X_{1}}_{\alpha_{1}} \Pi^{X_{1}X_{2}...X_{L}}_{\alpha_{1}...\alpha_{i-1}} \\
&- \Pi^{X_{1}X_{2}...X_{L}}_{\alpha_{1}...\alpha_{i-1}} \left( k_{X_{1}X_{2}}^{a} + k_{X_{1}X_{2}}^{b} \right) \Pi^{X_{1}X_{2}...X_{L}}_{\alpha_{1}...\alpha_{i-1}} \\
&- \Pi^{X_{1}X_{2}...X_{L}}_{\alpha_{1}...\alpha_{i-1}} \left( k_{X_{1}X_{2}}^{a} + k_{X_{1}X_{2}}^{b} \right) \Pi^{X_{1}X_{2}...X_{L}}_{\alpha_{1}...\alpha_{i-1}} \left( k_{X_{1}X_{2}}^{a} + k_{X_{1}X_{2}}^{b} \right) \\
&+ \Pi^{X_{1}X_{2}...X_{L}}_{\alpha_{1}...\alpha_{i-1}} \left( k_{X_{1}X_{2}}^{a} + k_{X_{1}X_{2}}^{b} \right) \Pi^{X_{1}X_{2}...X_{L}}_{\alpha_{1}...\alpha_{i-1}} \left( k_{X_{1}X_{2}}^{a} + k_{X_{1}X_{2}}^{b} \right) \\
\end{align*}
\]
Now the probability profile $P_{a,b}^{X_m}$ and the velocity profile $v_m = 1/\Gamma_m$ can be computed respectively. Fig. 11 shows that the numerical results agree perfectly well with the simulation results given by Gillespie algorithm 15.

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 point out that since the two algorithm 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 shows no difference with the results of IFS algorithm (Fig. 2). This treatment and the expanded template are also used to obtain Fig. 3, Fig. 11 and Fig. 3.

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 $C_{\alpha_i, \alpha_j}^{X_i, X_j} = \sum_{\{\alpha_k, k \neq i, j\}} P_{\alpha_1, \alpha_k, \ldots, \alpha_L}^{X_i, X_k, \ldots, X_L} - P_{\alpha_i}^{X_i} \cdot P_{\alpha_j}^{X_j}$ for any position $i$, and correspondingly the relative correlation function $\tilde{C}_{\alpha_i, \alpha_j}^{X_i, X_j} = C_{\alpha_i, \alpha_j}^{X_i, X_j} / P_{\alpha_i}^{X_i} \cdot P_{\alpha_j}^{X_j}$ and $\tilde{C}_{\text{max}}(i, d)$ decays abruptly with the correlation length 1 (illustrated by Fig. 9), implying that the positional probability is determined by its nearest neighbors. This does not hold in general, of course. For instance, the correlation length will become much larger under some extreme conditions (e.g, Parameter 3 in Fig. 3b) where one can not identify consistent pairing rules for each type of template units. For instance, $(A^\dagger A)$ is the dominant pairing (say, with a probability larger than 0.9) only for a part of the template $A$ while $(A^\dagger A)$ is dominant for the rest, so no Watson-Crick 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 and long-range correlations are out of the scope of this manuscript and will not be discussed in later sections.
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 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 maintain the high overall velocity.

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

(a) \( k_{RR}^{XY} \gg k_{RW}^{XY} \), which mean that the addition of \( R \) are always much faster than that of \( W \).
(b) $k_{XY}^{W} / k_{WW}^{XY} \gg k_{RR}^{XY} / k_{WW}^{XY}$, which mean that the successive additions of W are almost inhibited. In fact, $k_{WW}^{XY}$ are hard to measure in experiments, so $k_{WW}^{XY} \sim 0$ are always assumed.

(c) $k_{RR}^{XY} \gg r_{RR}^{XY}, r_{WW}^{XY}$, which mean 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 of that a buried W is hard to be proofread.

(d) $k_{WW}^{XY} \gg r_{WW}^{XY}$, which mean that the terminus containing more Ws 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., $k_{RR}^{XY} / k_{WW}^{XY} > 10^5$ and $k_{XY}^{XY} \gg r_{XY}^{XY}$ are always observed in real DNAPs). Under such general conditions, the exact method introduced in Sec.11 can be well approximated by the following method. We start from the iteration relations

$$g_{\alpha_1}^{X_1 \ldots X_L} = \frac{k_{\alpha_1}^{X_1 \ldots X_L}}{1 + r_{\alpha_1}^{X_1 \ldots X_L} / g_{\alpha_1}^{X_1 \ldots X_L}}$$

$$g_{\alpha_1}^{X_1 \ldots X_L} = \frac{k_{\alpha_1}^{X_1 \ldots X_L}}{1 + r_{\alpha_1}^{X_1 \ldots X_L} / g_{\alpha_1}^{X_1 \ldots X_L}}$$

$$g_{\alpha_1}^{X_1 \ldots X_L} = \frac{k_{\alpha_1}^{X_1 \ldots X_L}}{1 + r_{\alpha_1}^{X_1 \ldots X_L} / g_{\alpha_1}^{X_1 \ldots X_L}}$$

Under bio-relevant conditions, one has $k_{\alpha_1}^{X_1 \ldots X_L} \gg k_{\alpha_1}^{X_1 \ldots X_L}$, so $g_{\alpha_1}^{X_1 \ldots X_L} \sim k_{\alpha_1}^{X_1 \ldots X_L}$. The next iteration is

$$g_{\alpha_2}^{X_1 \ldots X_L} = \frac{k_{\alpha_1}^{X_1 \ldots X_L}}{1 + r_{\alpha_1}^{X_1 \ldots X_L} / g_{\alpha_1}^{X_1 \ldots X_L}}$$

Combining these two results, we get $g_{\alpha_2}^{X_1 \ldots X_L} \sim k_{\alpha_2}^{X_1 \ldots X_L}$.

Following the same logic, we obtain $g_{\alpha_i}^{X_1 \ldots X_L}$ (denoted as $g_{\alpha_i}^{X_1 \ldots X_L}$) and $\prod_{\alpha_i}^{X_1 \ldots X_L} \sim \prod_{\alpha_i}^{X_1 \ldots X_L}$ (denoted as $\prod_{\alpha_i}^{X_1 \ldots X_L}$). Correspondingly, the stochastic transfer matrix is approximated as $M_{\alpha_i}^{X_1 \ldots X_L}$ (the nearest-neighbor or 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,

$$p_{s_1 \ldots s_L}^{X_1 \ldots X_L} \sim q_{s_1}^{X_1} M_{s_1}^{X_2 \ldots X_L} \ldots M_{s_1 s_2}^{X_1 \ldots X_L}$$

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

$$M_{RW}^{X_1 \ldots X_L} \sim \frac{k_{RW}^{X_1 \ldots X_L}}{r_{WW}^{X_1 \ldots X_L} + g_{WW}^{X_1 \ldots X_L}}$$

$$= \frac{k_{RW}^{X_1 \ldots X_L}}{k_{RR}^{X_1 \ldots X_L}} \left( 1 + \frac{g_{WW}^{X_1 \ldots X_L}}{k_{WW}^{X_1 \ldots X_L}} \right)$$

FIG. 3: The correlation $C_{max}$ and the relative correlation $\tilde{C}_{max}$ between the position 50 and the rest positions of the random template. (a) under Parameter 2 (bio-relevant conditions). (b) under Parameter 3.
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 \( z << y << 1 \), one can verify that its left eigenvector associated with the largest eigenvalue 1 is approximately \( P = (1 - y, y) \) (\( \lim n \rightarrow \infty M^n \) converges to a matrix in which each row is the eigenvector \( P \)). For more details of heuristic analysis, see Appendix 2. \( P \) is a good approximation of the precise probability distribution at position \( i+1 \), which can be verified numerically (see Figs. 4(b,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,c)). This approximation, however, fails under conditions far different from bio-relevant conditions (data not shown here. see Supplementary Materials). This is consistent with the widely acknowledged idea that DNA replication mutations are randomly distributed in the genome.

**V. GENERALIZATION TO MULTI-COMPONENT 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 multi-component system which consist of \( n \) types of units \( A_i \) (\( i = 1, .., n \)) in the template and \( n \) types of monomers \( a_i \) (\( i = 1, .., n \)) 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 \( W_i, i = 1, 2, .., n-1 \)). The corresponding bio-relevant conditions are just a simple generalization of those in the preceding section.

(a) \( k_{RR}^{X+1} \gg k_{RW}^{X+1} \)

(b) \( r_{W_i}^{XY} \gg r_{W_j}^{XY} \gg k_{RR}^{XY} \)

(c) \( k_{RR}^{XY} \gg r_{W_i}^{XY} \)

(d) \( r_{W_i}^{XY} > r_{W_j}^{XY} \)

Similarly we rearrange the transfer matrix.
It can be shown that the distribution under bio-relevant conditions, so the eigenvector

where, for instance, we set the initial seed as a given value. One can follow the logic of Sec.III to obtain

Now the positional fidelity at \( i \) is

\[
M_{X_{i-1}X_{i+1}X_i} = \frac{M_{X_{i-1}X_{i+1}X_i}}{M_{X_{i-1}X_{i+1}X_i}} \left( 1 + \frac{M_{X_{i-1}X_{i+1}X_i}}{M_{X_{i-1}X_{i+1}X_i}} \right), \tag{12}
\]

where \( j = 1, \ldots, n - 1 \).

VI. GENERALIZATION TO HIGHER ORDER PROCESSES

For \( h \)-order processes, we set the initial seed as a given distribution \( \rho_{X_{i-1}X_i} \). One can follow the logic of Sec.III to obtain

\[
P_{X_{i-1}X_i} = \frac{X_{i-1}X_i}{\rho_{X_{i-1}X_i}} \left( 1 + \frac{X_{i-1}X_i}{\rho_{X_{i-1}X_i}} \right)
\]

\[
\Pi_{X_{i-1}X_i} = \frac{X_{i-1}X_i}{\rho_{X_{i-1}X_i}} \left( 1 + \frac{X_{i-1}X_i}{\rho_{X_{i-1}X_i}} \right), \tag{13}
\]

and equivalently,

\[
P_{X_{i-1}X_i} = \frac{X_{i-1}X_i}{\rho_{X_{i-1}X_i}} \left( 1 + \frac{X_{i-1}X_i}{\rho_{X_{i-1}X_i}} \right), \tag{14}
\]

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

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

(b) \( k_{XYZ} / k_{YZW} \gg k_{XYR} / k_{XWR} \), \( \alpha \beta = RW, WR, WW \). \( k_{XYZ} \equiv k_{XYZ} / \left( 1 + r_{XYZ} / k_{YZW} \right) \) is approximately the renormalized addition rates.
The positional probability at position $i$ can be calculated by $P_i^X = P_{RR}^X + P_{WW}^X = 1 - M_{RR} = M_{RRR}$. It is more likely to be proofread intuitively that how the template-dependent proofreading pathways could be coordinated with the polymerization pathways to achieve high fidelity. These results also indicates that the positional quantities are only dependent on the closely surrounding template sequence and relevant 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 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 sub-steps (e.g., see Ref. [8]). The basic theory in Sec. II can be slightly modified to handle...

\[ M = \begin{bmatrix} M_{RRR} & M_{RRW} & M_{RW} \\ M_{WR} & M_{RR} & M_{RW} \\ M_{WW} & M_{WR} & M_{WW} \end{bmatrix} \]

It can be shown that $M_{WW} = M_{RR} = M_{WWW} << 1 - M_{RR}$, which is approximated by $M_{RR} = 1 - M_{RR}$. The eigenvalue $\lambda_1$ of this matrix is approximately $1 - 2M_{RR}$, which implies that $M_{RR} = 1 - M_{RR}$. The eigenvalue $\lambda_1$ of this matrix is approximately $1 - 2M_{RR}$, which implies that $M_{RR} = 1 - M_{RR}$.

\[ M = \begin{bmatrix} \lambda_1 & \lambda_2 & \lambda_3 \\ \lambda_2 & \lambda_1 & \lambda_3 \\ \lambda_3 & \lambda_2 & \lambda_1 \end{bmatrix} \]

Studies on the template-specific fidelity of DNAPs are important to understand how genetic mutations are generated and controlled. While biochemical experiments have offered much 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 manuscript, we proposed a different method, based on the first-passage description of the replication process, to address these problem for complicated processes with high-order neighbor effects. Although the boundary conditions in our method is 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 shows intuitively that how the template-dependent proofreading pathways could be coordinated with the polymerization pathways to achieve high fidelity. These results also indicates that the positional quantities are only dependent on the closely surrounding template sequence and relevant 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 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 sub-steps (e.g., see Ref. [8]). The basic theory in Sec. II can be slightly modified to handle...
these problems without appealing for additional steady-state assumptions which are usually adopted to model multi-step processes in biochemistry (e.g., the well known Michaelis-Menten kinetics which is also used to study the DNA replication in Ref. [14]). Comprehensive discussions on this issue, as well as applications to real DNA replication systems, will be presented elsewhere.

VIII. ACKNOWLEDGMENTS

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

Appendices

1. The dwell time distribution at position m

It’s mentioned in the main text that $\Gamma_m$ is the mean dwell time of the growing chain at template position $m$, 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 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 $t$, i.e., 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 dwell time per trajectory is given by $\int_0^{+\infty} p^X_m(t)dt$.

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

$$T_{m,L}(t) = \sum_{\alpha_m=a,b} P_{\alpha_m}^{X_m} T^\alpha_{m,L}(t)$$

with $P_{\alpha_m}^{X_m}$ is the final probability distribution at $m$, as calculated in the main text. $T^\alpha_{m,L}(t)$ is defined as the first-passage time distribution of a new imaginary replication process which initiates at the template position $m$ with initial conditions $q^X_{\alpha_m} = 1$ ($\alpha_m = a$ 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 dwell time distribution $T_{m,L}(t)$, by using the convolution relation

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

with

$$T_{L-1}(t) = T_{L-1,L}(t)$$

$$= \sum_{\alpha_{L-1}=a,b} P_{\alpha_{L-1}}^{X_{L-1}} \left(k_{\alpha_{L-1}a} X_{L-1} + k_{\alpha_{L-1}b} X_{L-1}\right)$$

$$\cdot e^{-k_{\alpha_{L-1}a} X_{L-1} + k_{\alpha_{L-1}b} X_{L-1}t}$$

2. The eigenvector approximation

Aperiodic and irreducible stochastic matrices like $M$ have an important property according to 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, $\lambda_i (i \geq 2)$ are always far less than 1. Any probability distribution vector $P$ can be decomposed as $P = V_1 + \sum_{i>1} \lambda_i s_i V_i$, so $PM = V_1 + \sum_{i>1} \lambda_i s_i V_i$. If $PM$ differs not 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 $(P_X^R, P_X^W) = (P_X^{R-1}, P_X^{W-1}) \cdot M X_{i-1} X_i + 1$, and $(P_X^R, P_X^W)$ is indeed not too different from $(P_X^{R-1}, P_X^{W-1})$ (they both are around $(1,0)$). So we can safely approximate $(P_X^R, P_X^W)$ by the eigenvector $V_1$ of the matrix $M X_{i-1} X_i + 1$.

3. 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 Table II and Table III.

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-rang 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 Supplementary Materials) clearly show that the NN approximation still holds for such strongly-correlated template sequences.
FIG. 6: Verification of Eq. 16 by simulations for the random template. The statistics are made over $10^5$ simulations under Parameter 1. **Red**: dwell time distribution reconstructed from simulations of the original FP process (from the template position 1 to L). **Blue**: dwell time distribution reconstructed according to Eq. 16 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)$T_{51,100}(t)$.

### TABLE I: Random template

| 1-10       | 11-20      | 21-30      | 31-40      | 41-50      |
|------------|------------|------------|------------|------------|
| BAABAAAAABB| AAAAAABAAA| BBBBBAAABB| ABBABAAABBB| BBABAABAAA|
| 51-60      | 61-70      | 71-80      | 81-90      | 91-100     |
| BBBBBAABAA| ABBBBABABB| AAAABAABBB| ABBBBBBBBBBA| AABABABABB|

### TABLE II: kinetic parameters ($s^{-1}$, simulation time unit)

| Parameters | 1     | 2     | 3     |
|------------|-------|-------|-------|
| Template   |       |       |       |
| di-unit    |       |       |       |
| $k_{aa}$   | 65.0  | 50.0  | 65.0  |
| $k_{ab}$   | 68.0  | 50.0  | 64.0  |
| $k_{ba}$   | 54.0  | 95.0  | 56.0  |
| $k_{bb}$   | 45.0  | 66.0  | 80.0  |
| $r_{aa}$   | 12.0  | 23.0  | 7.0   |
| $r_{ab}$   | 16.0  | 24.0  | 16.0  |
| $r_{ba}$   | 22.0  | 9.0   | 17.0  |
| $r_{bb}$   | 14.0  | 23.0  | 12.0  |

Parameters 1: the addition rates and deletions rates are of the same orders of magnitude, which is different from the bio-relevant conditions.

Parameters 2: bio-relevant conditions in which $R$ and $W$ (base pairs) can be uniquely defined for each template unit (say $A-a$, $B-b$).

Parameters 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.

### TABLE III: Markov template

| 1-10       | 11-20      | 21-30      | 31-40      | 41-50      |
|------------|------------|------------|------------|------------|
| AAAAAAAAAA| AAAAAAAAAA| BBBBBBBBBBB| BBBBBBBBBBB| BBBBBBBBBBA|
| 51-60      | 61-70      | 71-80      | 81-90      | 91-100     |
| AAAAABBBAA| AAAAAAAAAA| AABBBBBBBB| BBBBBBBBBBB| BBAAAAAAA  |
[1] J. D. Watson, F. H. C. Crick, *Nature*, 171 (1953) 737.
[2] J. J. Hopfield, *PNAS*, 71 (1974) 4135.
[3] J. Ninio, *Biochimie*, 57 (1975) 587.
[4] I. R. Lehman, M. J. Bessman, E. S. Simms, A. Kornberg, *J. biol. Chem.*, 233 (1958) 163.
[5] T. A. Kunkel, K. Bebenek, *Ann. Rev. Biochem.*, 69 (2000) 497.
[6] S. S. Patel, I. Wong, K. A. Johnson, *Biochemistry*, 30 (1991) 511.
[7] I. Wong, S. S. Patel, K. A. Johnson, *Biochemistry*, 30 (1991) 526.
[8] Yu-Chih Tsai, K. A. Johnson, *Biochemistry*, 45 (2006) 9675.
[9] K. A. Johnson, *BBA-Proteins and Proteomics*, 1804 (2010) 1041.
[10] P. Gaspard, D. Andrieux, *J. Chem. Phys.*, 141 (2014) 044908.
[11] Yao-Gen Shu, Yong-Shun Song, Zhong-Can Ou-Yang, Ming Li, *J. Phys.: Condensed Matter*, 27 (2015) 235105.
[12] Yong-Shun Song, Yao-Gen Shu, Xin Zhou, Zhong-Can Ou-Yang, Ming Li, *J. Phys.: Condensed Matter*, 29 (2017) 025101.
[13] P. Gaspard, *J. Stat. Phys.*, 164 (2016) 17.
[14] P. Gaspard, *Phys. Rev. E.*, 93 (2016) 042419.
[15] P. Gaspard, *Phys. Rev. E.*, 93 (2016) 042420.
[16] P. Gaspard, *Phys. Rev. Lett.*, 117 (2016) 238101.
[17] P. Gaspard, *Phys. Rev. E.*, 96 (2017) 042403.
[18] D. T. Gillespie, *J. Phys. Chem.*, 81 (1977) 2340.
[19] W. Li, K. Kaneko, *Europhysics Letters*, 17 (1992) 655.