Iterated function systems for DNA replication

Pierre Gaspard
Center for Nonlinear Phenomena and Complex Systems, Université libre de Bruxelles (ULB), Code Postal 231, Campus Plaine, B-1050 Brussels, Belgium
(Received 17 July 2017; published 4 October 2017)

The kinetic equations of DNA replication are shown to be exactly solved in terms of *iterated function systems*, running along the template sequence and giving the statistical properties of the copy sequences, as well as the kinetic and thermodynamic properties of the replication process. With this method, different effects due to sequence heterogeneity can be studied, in particular, a transition between linear and sublinear growths in time of the copies, and a transition between continuous and fractal distributions of the local velocities of the DNA polymerase along the template. The method is applied to the human mitochondrial DNA polymerase γ without and with exonuclease proofreading.

DOI: 10.1103/PhysRevE.96.042403

I. INTRODUCTION

During DNA replication, the copolymerization of new nucleic acid chains is catalyzed by enzymes called DNA polymerases [1]. These molecular machines perform the synthesis of new DNA strands by moving along templates formed by the DNA strands of previous generations. Since DNA codes genetic information, the template sequences are heterogeneous and constitute one-dimensional disordered media for the motion of polymerases [2–10]. This motion is determined by the enzymatic kinetics of polymerases, which has been studied experimentally in biochemistry. The main kinetic parameters have been measured for several DNA polymerases [11–26]. The knowledge of kinetics is probably most complete for human mitochondrial DNA polymerase γ, in which case the kinetic parameters are known for the 16 possible nucleotide pairings between the DNA template and its copy [18–21]. This knowledge allows us to set up detailed kinetic models for template-directed DNA copolymerization and thus to investigate the effects of sequence heterogeneity. These effects are overlooked if the kinetic parameters are supposed to depend only on whether nucleotide pairing is correct or incorrect [27–34]. However, discrepancies have been observed between so simplified models and the Monte Carlo simulations of the complete kinetics [32,33]. As recently shown in Ref. [10], these difficulties can be overcome by using *iterated function systems* (IFS) [35,36]. These mathematical recurrences are used for constructing fractals [35], for image compression [37], as well as in the theory of diffusion in one-dimensional disordered lattices [38,39]. In the context of template-directed copolymerization, IFS provide the exact long-time solution of the kinetic equations [10].

The purpose of this paper is to describe in detail the application of the IFS method to DNA replication. With the IFS method, we can now investigate the effects of sequence heterogeneity. In particular, the IFS method can determine how the pairing probabilities, the replication errors, and the local velocity of the DNA polymerase vary along the template sequence, allowing the study of sequence-specific properties as a function of nucleotide concentrations. It turns out that the variations of the local velocity may have continuous or fractal distributions, depending on the nucleotide concentrations. Besides, the mean growth velocity may vanish over a whole range of nucleotide concentrations where the growth of the copy can be sublinear in time instead of linear, as for other processes of random drift in disordered media [40–45]. The effects of sequence heterogeneity manifest themselves more strongly if the nucleotide concentrations are imbalanced, in which case the replication errors become more frequent. It is shown that such imbalances in the nucleotide concentrations have important consequences on the composition of DNA sequences. Furthermore, the effects of sequence heterogeneity may influence the nonequilibrium thermodynamics of DNA replication.

The vehicle of our study is the human mitochondrial DNA polymerase γ, for which the kinetic parameters of the 16 pairings are known [18–21] and the aforementioned issue thus arises. The same kinetic model as in Refs. [32,33] will be used, but with the exact IFS method of Ref. [10], instead of the approximation of Refs. [32,33] simplifying the kinetics to correct or incorrect pairings. Moreover, the IFS method will be applied to the DNA polymerases without and with the dedicated proofreading mechanism provided by the exonuclease activity.

The plan of the paper is as follows. The IFS method to solve exactly the kinetic equations of DNA replication is presented in Sec. II. In Sec. III, the IFS method is applied to exonuclease-deficient (exo^-) human mitochondrial DNA polymerase moving along templates composed of the four nucleotides. In Sec. IV, the template is supposed to be composed of only two different species of nucleotides in order to analyze more closely the fractal distributions of local velocities. In Sec. V, the IFS method is applied to (exo^+) human mitochondrial DNA polymerase with exonuclease proofreading. Conclusion and perspectives are drawn in Sec. VI.

II. IFS METHOD

A. Kinetic equations

The Michaelis-Menten kinetics of DNA polymerases can be described as successive random events of nucleotide attachment and detachment, which is schematically represented as

follows:

$$\begin{array}{lll} \omega = m_1 m_2 \ldots m_{l-1} & + \; m_l & \rightleftharpoons \quad \omega' = m_1 m_2 \ldots m_{l-1} m_l \\ \alpha = n_1 \, n_2 \; \ldots \; n_{l-1} \, n_l \, n_{l+1} \; \ldots & & \alpha \; = n_1 \, n_2 \; \ldots \; n_{l-1} \, n_l \, n_{l+1} \; \ldots , \end{array} \tag{1}$$

where α denotes the template composed of the nucleotides $n_l \in \{\mathrm{A,C,G,T}\}$ and ω the copy composed of $m_l \in \{\mathrm{A,C,G,T}\}$. The template sequence can be periodic, a Bernoulli chain of probabilities $\{\nu_\mathrm{A}, \nu_\mathrm{C}, \nu_\mathrm{G}, \nu_\mathrm{T}\}$, or else.

The attachment and detachment rates $W_{\pm m_l m_{l-1}, l}$ depend not only on the nucleotide m_l that is attached or detached, but also on the previously incorporated nucleotide m_{l-1}, which plays an important role in the case of exonuclease proofreading [13,33]. The rates also depend on the template sequence and vary from site to site, which is denoted by the subscript l after the comma. Moreover, the rates depend on the nucleotide concentrations as will be explained in Sec. III. The solution surrounding the macromolecular chains is supposed to be dilute and to contain nucleotide species at invariant concentrations so that the rates do not change in time.

The kinetics is described at the level of a single DNA polymerase in terms of the probability that the copy has the sequence $\omega = m_1 \ldots m_l$ at time t, provided that the template has the sequence $\alpha = n_1 \ldots n_l n_{l+1} \ldots$. The kinetic equations ruling the process are given by

$$\begin{aligned} \frac{d}{dt} P_t \begin{pmatrix} m_1 \ldots m_l \\ n_1 \; \ldots \; n_l \, n_{l+1} \cdots \end{pmatrix} &= W_{+m_l m_{l-1}, l} \, P_t \begin{pmatrix} m_1 \ldots m_{l-1} \\ n_1 \; \ldots \; n_{l-1} \, n_l \ldots \end{pmatrix} + \sum_{m_{l+1}} W_{-m_{l+1} m_l, l+1} \, P_t \begin{pmatrix} m_1 \ldots m_l m_{l+1} \\ n_1 \; \ldots \; n_l \, n_{l+1} \, n_{l+2} \ldots \end{pmatrix} \\ &\quad - \left(W_{-m_l m_{l-1}, l} + \sum_{m_{l+1}} W_{+m_{l+1} m_l, l+1} \right) P_t \begin{pmatrix} m_1 \ldots m_l \\ n_1 \; \ldots \; n_l \, n_{l+1} \ldots \end{pmatrix}, \end{aligned} \tag{2}$$

where the first gain term describes the attachment of the nucleotide m_l to a chain of length $l-1$, the second ones the detachments of the different possible nucleotides m_{l+1} from a chain of length $l+1$, while the loss terms are, respectively, due to the detachment of m_l and the attachments of m_{l+1} [32,33].

B. Factorization of the probabilities during steady growth

After a long enough time, the probabilities ruled by Eq. (2) will factorize as for a Markov chain because the rates depend on the previously incorporated nucleotide [10,46]. As a consequence, these probabilities take the following form:

$$P_t \begin{pmatrix} m_1 \ldots m_l \\ n_1 \; \ldots \; n_l \, n_{l+1} \ldots \end{pmatrix} \simeq p_t(l) \prod_{j=1}^{l-1} \mu(m_j | m_{j+1}, j) \, \mu(m_l, l), \tag{3}$$

in terms of the conditional probabilities $\mu(m_j|m_{j+1}, j)$ of the Markov chain, the tip probability $\mu(m_l, l)$ of the last nucleotide m_l that is incorporated at the growing end of the chain, and the probability that the copy has the length l at time t:

$$p_t(l) \equiv \sum_{m_1 \ldots m_l} P_t \begin{pmatrix} m_1 \ldots m_l \\ n_1 \; \ldots \; n_l \, n_{l+1} \ldots \end{pmatrix}. \tag{4}$$

The conditional and tip probabilities are normalized according to

$$\sum_{m_j} \mu(m_j | m_{j+1}, j) = 1 \tag{5}$$

for $j = 1,2,\ldots,l-1$, and

$$\sum_{m_l} \mu(m_l, l) = 1 \, . \tag{6}$$

The mean length of the copy is defined by

$$\langle l \rangle_t \equiv \sum_{l=1}^{\infty} l \; p_t(l) \tag{7}$$

and the mean growth velocity by

$$v \equiv \frac{d}{dt} \langle l \rangle_t . \tag{8}$$

In the regimes of steady growth, this velocity is invariant and positive $v > 0$. However, the mean growth velocity may vanish $v = 0$, while the chain is still growing sublinearly in time as $\langle l \rangle_t \sim t^\gamma$ with $0 < \gamma < 1$ [10,46]. The growth stops at equilibrium where $v = 0$ and $\gamma = 0$.

In order to determine the growth velocity, we may consider a periodic template with a long period L. If we introduce the waiting time τ_l of the polymerase at the location l of the template, the velocity should take the value

$$v = \frac{L}{\sum_{l=1}^{L} \tau_l} . \tag{9}$$

The length distribution $p_t(l)$ is expected to drift towards longer and longer lengths at the velocity v, while becoming broader and broader as time increases. The probability $p_t(l)$ of the length l should be proportional to the corresponding waiting time τ_l. If this probability distribution was normalized according to $\sum_{l=1}^{L} p_t(l) = 1$ over the period L, we would have that

$$p_t(l) \simeq \frac{v}{L} \, \tau_l = \frac{v}{L \, x_l}, \tag{10}$$

where

$$x_l \equiv \frac{1}{\tau_l} \tag{11}$$

is the local velocity of the polymerase at the location l of the template. Taking the limit $L \to \infty$ in Eq. (9), the mean growth velocity is given by

$$\frac{1}{v} = \langle \tau_l \rangle = \left\langle \frac{1}{x_l} \right\rangle, \tag{12}$$

where $\langle \cdot \rangle = \lim_{L\to\infty} L^{-1} \sum_{l=1}^{L} (\cdot)_l$ denotes the average over an arbitrarily long template sequence [42,43].

C. Backward and forward recurrences

The kinetic equations (2) being linear, their solutions can be expressed as linear superpositions of the modes $P_t \sim \exp(s_q t + iql)$ where s_q is vanishing with q, either as $s_q \simeq -ivq$ if $v > 0$ or more slowly if $v = 0$. Since the probability distribution becomes broader and broader in the long-time limit, the solutions of Eqs. (2) are dominated by modes with smaller and smaller values of $|q|$ so that the left-hand sides of Eqs. (2) will thus also vanish [10,46]. An alternative method is to use the Laplace transform $\tilde{P}_z$ of P_t, in which case the left-hand sides of Eqs. (2) become $z\tilde{P}_z$ if no chain is already formed at time zero. In the long-time limit, smaller and smaller values of z should be considered and, again, the asymptotic solutions will be given by taking the left-hand sides of Eqs. (2) equal to zero. On this ground, the asymptotic solutions of Eqs. (2) can be deduced as follows.

Here, we present an extended version of the deduction summarized in the Supplemental Material of Ref. [10] by first introducing the local partial waiting times

$$u_{m_l,l} \equiv \mu(m_l,l)\, \tau_l. \tag{13}$$

Using Eq. (10), the factorization (3) becomes

$$P_t\begin{pmatrix} m_1 \dots m_l \\ n_1 \dots n_l\, n_{l+1} \cdots \end{pmatrix} \simeq \frac{v}{L}\, \mu(m_1|m_2,1) \dots \mu(m_{l-2}|m_{l-1},l-2)\, \mu(m_{l-1}|m_l,l-1)\, u_{m_l,l}. \tag{14}$$

Inserting this expression into the kinetic equation (2), we obtain after simplification the following equation:

$$0 \simeq W_{+m_l m_{l-1},l}\, u_{m_{l-1},l-1} + \mu(m_{l-1}|m_l,l-1)\left[\sum_{m_{l+1}} W_{-m_{l+1}m_l,l+1}\, \mu(m_l|m_{l+1},l)\, u_{m_{l+1},l+1} - \left(W_{-m_l m_{l-1},l} + \sum_{m_{l+1}} W_{+m_{l+1}m_l,l+1}\right) u_{m_l,l}\right]. \tag{15}$$

Defining the local partial velocities as

$$v_{m_l,l} \equiv \sum_{m_{l+1}} W_{+m_{l+1}m_l,l+1} - \sum_{m_{l+1}} W_{-m_{l+1}m_l,l+1}\, \mu(m_l|m_{l+1},l)\, \frac{u_{m_{l+1},l+1}}{u_{m_l,l}}, \tag{16}$$

Eq. (15) gives the following expression for the conditional probabilities:

$$\mu(m_{l-1}|m_l,l-1) = \frac{W_{+m_l m_{l-1},l}}{W_{-m_l m_{l-1},l} + v_{m_l,l}}\, \frac{u_{m_{l-1},l-1}}{u_{m_l,l}}. \tag{17}$$

If we replace the conditional probabilities in Eq. (16) by their expression (17), we deduce the backward recurrence [10]

$$v_{m_{l-1},l-1} = \sum_{m_l} \frac{W_{+m_l m_{l-1},l}}{W_{-m_l m_{l-1},l} + v_{m_l,l}}\, v_{m_l,l}, \tag{18}$$

forming an IFS in the four-dimensional space of local partial velocities $\mathbf{v}_l = \{v_{m_l,l}\}$ [35,36]. An important result is that the functions $\mathbf{v}_{l-1} = \mathbf{f}_l(\mathbf{v}_l)$ of the IFS are explicitly defined in terms of the attachment and detachment rates of the kinetics. The recurrence proceeds by picking up the function $\mathbf{f}_l$ corresponding to the local subsequence of nucleotides at the location l of the template. The function $\mathbf{f}_l$ is chosen at random if the template sequence is random. There are as many possible functions $\mathbf{f}_l$ as nucleotide subsequences determining the rates of the polymerase, as further explained in Sec. III.

Now, summing the conditional probability (17) over m_{l-1} and using the normalization condition (5), we get the forward recurrence [10]

$$u_{m_l,l} = \sum_{m_{l-1}} \frac{W_{+m_l m_{l-1},l}}{W_{-m_l m_{l-1},l} + v_{m_l,l}}\, u_{m_{l-1},l-1}. \tag{19}$$

The backward and forward recurrences (18) and (19) determine all the statistical properties of the copy [10,46]. In particular, the stationary probability $\bar{\mu}(m_l,l)$ of the Markov chain, called bulk probabilities in the following, should satisfy

$$\sum_{m_l} \mu(m_{l-1}|m_l,l-1)\, \bar{\mu}(m_l,l) = \bar{\mu}(m_{l-1},l-1) \quad \text{for every } l. \tag{20}$$

Replacing therein the conditional probabilities by their expression (17), comparing with Eq. (18), and using the normalization condition $\sum_{m_l} \bar{\mu}(m_l,l) = 1$, we see that the bulk probabilities can be written as

$$\bar{\mu}(m_l,l) = u_{m_l,l}\, v_{m_l,l}. \tag{21}$$

On the one hand, applying the normalization condition of the tip probabilities to Eq. (13), we find that the waiting times of the polymerase are given by

$$\tau_l = \sum_{m_l} u_{m_l,l} = \frac{1}{x_l}. \tag{22}$$

On the other hand, replacing the quantities $u_{m_l,l}$ by their expression (13) in Eq. (21) and using the normalization condition of the bulk probabilities, the local velocities x_l of the polymerase at every location along the template are

obtained as

$$x_l = \sum_{m_l} v_{m_l,l}\,\mu(m_l,l) = \frac{1}{\tau_l}. \tag{23}$$

Finally, the mean growth velocity is evaluated by Eq. (12). The quantities (13) and (16) are said to be partial because they represent the partial contribution of each nucleotide species to the corresponding local quantity, as expressed by Eqs. (22) and (23).

The IFS method is numerically implemented by considering a template sequence of long length L forming a loop. The backward recurrence (18) is run several times on the loop to achieve numerical convergence for the local partial velocities $\{v_{m_l,l}\}_{l=1}^{L}$. Thereafter, the forward recurrence is also run several times on the loop to obtain the partial waiting times $\{u_{m_l,l}\}_{l=1}^{L}$, from which the waiting times $\{\tau_l\}_{l=1}^{L}$ are deduced with Eq. (22) and finally the mean growth velocity by Eq. (12).

Moreover, the replication error probability is defined by averaging the bulk probabilities of incorrect pairing along the sequence:

$$\eta \equiv \lim_{L\to\infty} \frac{1}{L}\sum_{l=1}^{L} [1-\bar{\mu}(\tilde{n}_l,l)], \tag{24}$$

where $\tilde{n}_l$ denotes the nucleotide in the copy that is complementary to the corresponding nucleotide n_l of the template.

D. Thermodynamics

For enzymes with both the polymerase and exonuclease activities, the two corresponding reactions ρ contribute to the attachment and detachment rates [33]

$$W_{\pm m_l m_{l-1},l} = \sum_{\rho} W^{\rho}_{\pm m_l m_{l-1},l}, \tag{25}$$

as well as to the thermodynamic entropy production rate

$$\frac{1}{R}\frac{d_{\rm i}S}{dt} = \sum_{l}\sum_{m_1\ldots m_l}\sum_{\rho}\left[W^{\rho}_{+m_l m_{l-1},l}\,P_t\begin{pmatrix} m_1\ldots m_{l-1} \\ n_1\ \ldots\ n_{l-1}\,n_l\ldots\end{pmatrix} - W^{\rho}_{-m_l m_{l-1},l}\,P_t\begin{pmatrix} m_1\ldots m_{l} \\ n_1\ \ldots\ n_{l}\,n_{l+1}\ldots\end{pmatrix}\right] \times \ln\frac{W^{\rho}_{+m_l m_{l-1},l}\,P_t\begin{pmatrix} m_1\ldots m_{l-1} \\ n_1\ldots n_{l-1}\,n_l\ldots\end{pmatrix}}{W^{\rho}_{-m_l m_{l-1},l}\,P_t\begin{pmatrix} m_1\ldots m_{l} \\ n_1\ldots n_{l}\,n_{l+1}\ldots\end{pmatrix}} \geqslant 0, \tag{26}$$

where $R = 8.31451\,\mathrm{J\,K^{-1}\,mol^{-1}}$ is the molar gas constant [47–52]. The entropy production rate is always non-negative in agreement with the second law of thermodynamics.

Replacing the probability P_t by its asymptotic expression (14) into Eq. (26), the entropy production rate in steady growth regimes can be expressed as [27,53]

$$\Sigma \equiv \frac{1}{R}\frac{d_{\rm i}S}{dt} = v\,[\epsilon + D(\omega|\alpha)] \equiv v\,A \geqslant 0, \tag{27}$$

in terms of the mean growth velocity (12), the free-energy driving force

$$\epsilon \equiv \lim_{L\to\infty}\frac{1}{L}\sum_{l=1}^{L}\sum_{m_{l-1}m_l}\sum_{\rho}\mu(m_{l-1}|m_l,l-1)\,\bar{\mu}(m_l,l) \times \ln\frac{W^{\rho}_{+m_l m_{l-1},l}}{W^{\rho}_{-m_l m_{l-1},l}}, \tag{28}$$

and the conditional Shannon disorder of the copy with respect to the template

$$D(\omega|\alpha) \equiv \lim_{L\to\infty} -\frac{1}{L}\sum_{l=1}^{L}\sum_{m_{l-1}m_l}\mu(m_{l-1}|m_l,l-1)\,\bar{\mu}(m_l,l) \times \ln\mu(m_{l-1}|m_l,l-1) \geqslant 0. \tag{29}$$

We notice that the latter two quantities are given by averaging over the template sequence. Since the conditional and bulk probabilities (17) and (21) vary along the sequence, the free-energy driving force and the conditional Shannon disorder are sensitive to the sequence heterogeneity. Their sum defines the entropy production per incorporated monomeric unit, called affinity [53]:

$$A = \epsilon + D(\omega|\alpha). \tag{30}$$

The conditional Shannon disorder characterizes the amount of replication errors. The higher the replication fidelity, the lower the conditional Shannon disorder, which is vanishing if the copying process is error free.

In Monte Carlo simulations, the numerical evaluation of the conditional Shannon disorder can be quite demanding if the replication errors are very rare. Indeed, before averaging over the template sequence, the conditional and bulk probabilities should have been computed at every location of the sequence with a good enough numerical precision. The lower the error probability (24), the larger the statistical ensemble of copies used to compute these probabilities.

Thermodynamic equilibrium is identified as the state where the affinity is vanishing, $A = 0$. Because of sequence heterogeneity, the equilibrium state does not coincide with the state where the mean growth velocity is vanishing, $v = 0$, which is the threshold for a growth that is linear in time. Between equilibrium and this threshold, the growth of the copy is sublinear in time as $\langle l\rangle_t \sim t^{\gamma}$ with an exponent $0 < \gamma < 1$ so that the mean growth velocity remains equal to zero. In this sublinear growth regime, the entropy produced during the time interval t is also growing sublinearly in time as $\langle \Delta_{\rm i}S\rangle_t \sim t^{\gamma}$, so that the entropy production rate is also equal to zero although the process is still out of equilibrium.

In view of these results, sequence heterogeneity may significantly influence the nonequilibrium thermodynamics of template-directed copolymerization.

III. APPLICATION TO exo$^-$ HUMAN MITOCHONDRIAL DNA POLYMERASE

In this section, we consider the same kinetic model of exo$^-$ human mitochondrial DNA polymerase as in Ref. [32]. Here, this model is analyzed with the IFS method of Ref. [10], instead of the simplifying approximation used in Ref. [32] where the

kinetics was assumed to discriminate only between correct and incorrect pairings, although the kinetic parameters are known to be different for the 16 nucleotide pairings [21]. The results of Ref. [32] are thus revisited and the IFS method is shown to provide the sequence-specific properties and their dependencies on the four different nucleotide concentrations, which has not been possible with the simplifying approximation of Ref. [32].

A. Kinetics of exo$^-$ DNA polymerases

Exonuclease-deficient DNA polymerases are enzymes (E) catalyzing the polymerization of DNA strands by the incorporation of deoxyribonucleoside triphosphate dNTP (i.e., dATP, dCTP, dGTP, or dTTP) and the release of inorganic pyrophosphate PP_i:

polymerase activity:

$$E \cdot DNA_{l-1} + dNTP \rightleftharpoons E \cdot DNA_l + PP_i, \quad (31)$$

where l denotes the length of the DNA strand counted by the number of incorporated nucleotides.

The structure and function of DNA polymerases have been studied experimentally [11–26]. Their kinetics proceeds in several steps [13]. The two main steps are (1) the formation of a base pair between a nucleotide dNTP $= m_l PP_i$ coming from the surrounding solution and the nucleotide n_l of the template; (2) the release of pyrophosphate PP_i and the formation of a phosphodiester bond between the nucleotide m_l and the previously incorporated nucleotide m_{l-1} of the growing DNA strand. Since the first step is typically faster than the second, the binding of dNTP and its dissociation are in quasiequilibrium on the time scale of the second step, so that the kinetics is of Michaelis-Menten type [54,55].

We use the following notations [32]. Given that the previously formed base pair is $m_{l-1}{:}n_{l-1}$, $K_{\substack{m_l m_{l-1}\\ n_l n_{l-1}}}$ is the dissociation constant of m_l from n_l; $k^{\rm p}_{\substack{\pm m_l m_{l-1}\\ n_l n_{l-1}}}$ are the rate constants for the polymerization and depolymerization of m_l after its pairing with n_l. Furthermore, $[m_l P]$ denotes the concentration of the deoxyribonucleoside triphosphate $m_l PP_i$ and [P] the concentration of pyrophosphate PP_i. The concentration unit is the *mole per liter* (M). With these notations, the attachment rate of m_l is given by

$$W^{\rm p}_{\substack{+m_l m_{l-1}\\ n_l n_{l-1}}} \equiv \frac{k^{\rm p}_{\substack{+m_l m_{l-1}\\ n_l n_{l-1}}}[m_l P]}{K_{\substack{m_l m_{l-1}\\ n_l n_{l-1}}} Q_{\substack{m_{l-1}\\ n_l n_{l-1}}}} \quad (32)$$

and the detachment rate of m_l by

$$W^{\rm p}_{\substack{-m_l m_{l-1}\\ n_{l+1} n_l n_{l-1}}} \equiv \frac{k^{\rm p}_{\substack{-m_l m_{l-1}\\ n_l n_{l-1}}}[P]}{Q_{\substack{m_l\\ n_{l+1} n_l}}}, \quad (33)$$

where

$$Q_{\substack{m_{l-1}\\ n_l n_{l-1}}} \equiv 1 + \sum_{m_l} \frac{[m_l P]}{K_{\substack{m_l m_{l-1}\\ n_l n_{l-1}}}} \quad (34)$$

is the Michaelis-Menten denominator. Note that the denominator in the depolymerization rate (33) is shifted by one unit with respect to the denominator in the polymerization rate (32). In the kinetic Eqs. (2), the rates are thus

$$W_{+m_l m_{l-1},l} = W^{\rm p}_{\substack{+m_l m_{l-1}\\ n_l n_{l-1}}}, \quad (35)$$

$$W_{-m_l m_{l-1},l} = W^{\rm p}_{\substack{-m_l m_{l-1}\\ n_{l+1} n_l n_{l-1}}}, \quad (36)$$

where the symbol l after the comma in the subscripts on the left-hand side stands for the subsequence $n_{l+1}\, n_l\, n_{l-1}$ at the location l of the template.

Experimental data being rare for the depolymerization rate constants, it is assumed for simplicity that they are proportional to the corresponding polymerization rate constant according to

$$k^{\rm p}_{\substack{-m_{l+1} m_l\\ n_{l+1} n_l}} = \frac{1}{K_{\rm P}} k^{\rm p}_{\substack{+m_{l+1} m_l\\ n_{l+1} n_l}}, \quad (37)$$

where the constant associated with pyrophosphorolysis is here supposed to take the value $K_{\rm P} = 200$ mM [32].

For exo$^-$ human mitochondrial DNA polymerase γ at 37 °C, the rate constants have been measured experimentally for the 16 different pairings following a correct incorporation and they are given in Table I [21]. After an incorrect incorporation, the polymerase is known to slow down [13]. In this case, the polymerization rate constants for the formation of a correct (c) or incorrect (i) pair are, respectively, given by

$$k^{\rm p}_{+{\rm c}|{\rm i}} = 0.52\ {\rm s}^{-1}, \quad (38)$$

$$k^{\rm p}_{+{\rm i}|{\rm i}} = 0.154\ {\rm s}^{-1}, \quad (39)$$

and the corresponding dissociation constants by

$$K_{{\rm c}|{\rm i}} = K_{{\rm i}|{\rm i}} = 404\ \mu{\rm M}, \quad (40)$$

according to Ref. [19]. Moreover, the dissociation of the enzyme from DNA is neglected as in Ref. [32].

TABLE I. Exo$^-$ human mitochondrial DNA polymerase γ at 37 °C: the polymerization rate constants and dissociation constants used in the numerical simulations for a nucleotide attachment following a correct (c) incorporation. The data are from Ref. [21].

$m{:}n$ Pair	$k^{\rm p}_{+\substack{m\\ n}\,|{\rm c}}$ (s^{-1})	$K_{\substack{m\\ n}\,|{\rm c}}$ (μM)
A:T	45	0.8
A:G	0.042	250
A:C	0.1	160
A:A	0.0036	25
C:T	0.038	360
C:G	43	0.9
C:C	0.003	140
C:A	0.1	540
G:T	1.16	70
G:G	0.066	150
G:C	37	0.8
G:A	0.1	1000
T:T	0.013	57
T:G	0.16	200
T:C	0.012	180
T:A	25	0.6

For normal cells, the physiological nucleotide concentrations take the following mean values [56]:

$$[\mathrm{dATP}] = 3.2\ \mu\mathrm{M}, \tag{41}$$

$$[\mathrm{dCTP}] = 2.1\ \mu\mathrm{M}, \tag{42}$$

$$[\mathrm{dGTP}] = 1.5\ \mu\mathrm{M}, \tag{43}$$

$$[\mathrm{dTTP}] = 5.4\ \mu\mathrm{M}, \tag{44}$$

while the physiological pyrophosphate concentration is typically $[\mathrm{PP_i}] = 10^{-4}$ M [57].

B. IFS for the exo$^-$ DNA polymerase kinetics

Since the rates depend on the nucleotide subsequence $n_{l-1}\,n_l\,n_{l+1}$ at every location l of the template, the kinetic equations (2) are solved by considering the IFS

$$\mathbf{v}_{l-1} = \mathbf{f}_{n_{l-1}\,n_l\,n_{l+1}}(\mathbf{v}_l) \tag{45}$$

given by Eq. (18) and running backward along the template. This backward recurrence determines the four local partial velocities $\mathbf{v}_l = \{v_{m_l,l}\} \in \mathbb{R}^4$. Next, these latter are used to run the forward recurrence (19) giving the local partial waiting times $\mathbf{u}_l = \{u_{m_l,l}\} \in \mathbb{R}^4$ and thus the conditional and bulk probabilities, as shown in Sec. II C.

If the template is composed of a single nucleotide species, the recurrence (45) converges toward a fixed point $\mathbf{v}_*$ in the four-dimensional space of local partial velocities. If the template was periodic for instance as $\alpha = \mathrm{ACACAC}\ldots$, the recurrence would converge to an orbit $\{\mathbf{v}_*, \mathbf{v}'_*\}$ of period two, such that $\mathbf{v}'_* = \mathbf{f}_{\mathrm{ACA}}(\mathbf{v}_*)$ and $\mathbf{v}_* = \mathbf{f}_{\mathrm{CAC}}(\mathbf{v}'_*)$.

Here, the template is chosen as a Bernoulli chain with equal probabilities for the four types of nucleotides: $\nu_\mathrm{A} = \nu_\mathrm{C} = \nu_\mathrm{G} = \nu_\mathrm{T} = \frac{1}{4}$. In this case, one out of 4^3 functions is selected at every location l of the template according to the random subsequence of nucleotides $n_{l-1}\,n_l\,n_{l+1}$ in the template. In this way, the recurrence will typically generate an erratic orbit. The set of orbits corresponding to all the possible template sequences forms an attractor $F_\mathbf{v} \subset \mathbb{R}^4$ in the four-dimensional space of local partial velocities [35,36]. This attractor is bounded because the functions (18) of the IFS (45) are non-negative and each has a supremum. As shown here below, this attractor may be fractal.

In order to test the predictions of the IFS method, the kinetics of the DNA polymerase is numerically simulated with Gillespie's algorithm [58,59] explained in Appendix C of Ref. [32]. For a given template, the growth of many copies is simulated in order to obtain the conditional and bulk probabilities at every location of the template, as well as the other quantities of interest.

C. Sequence-specific properties

Figure 1 depicts the bulk probabilities and the waiting times of the DNA polymerase under physiological conditions along the template sequence shown above. In the Monte Carlo simulations with Gillespie's algorithm, the statistics has been carried out over 10^8 copies and the results are depicted as open symbols. There is a very nice agreement with the results of the IFS method shown as crosses, which confirms at the numerical accuracy of the computation that the IFS provides the exact long-time solution of the kinetic equations (2). Furthermore, the IFS method is about 10^6 faster than the Monte Carlo simulations. We see in Fig. 1 that the bulk probabilities are very close to the unit value for the Watson-Crick (correct) pairings and very low around 10^{-4} for incorrect pairings, which corresponds to the error probability $\eta \simeq 1.13 \times 10^{-4}$. Moreover, we observe the sequence-specific variations of the pairing probabilities and the polymerase waiting times. For instance, there is a subsequence TTT near the beginning of the template where the probability for the Watson-Crick pairing A:T drops below the unit value, while the incorrect pairing G:T has the largest probability among incorrect pairings. This is explained by the fact that, for incorrect pairings, the polymerization rate constant of G:T is the largest in Table I, while the corresponding dissociation constant remains relatively low. For this reason, the substitution of A by G is the most prominent in Fig. 1.

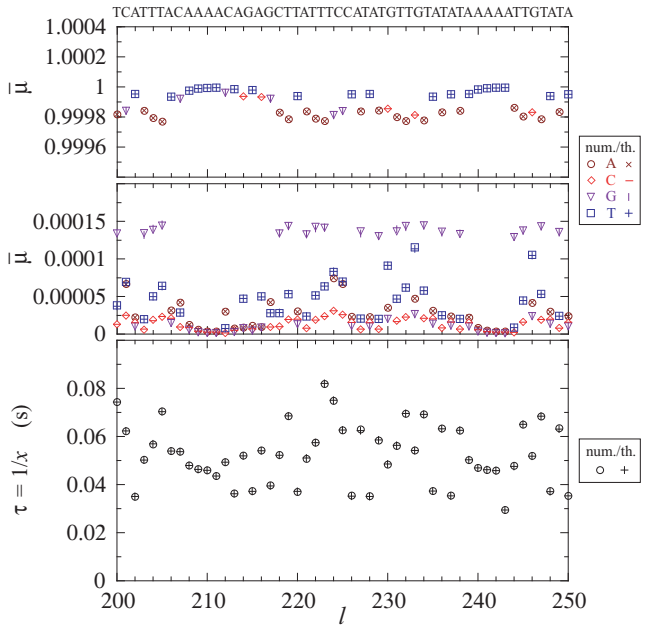

FIG. 1. Exo$^-$ human mitochondrial DNA polymerase: bulk probabilities and waiting times of the polymerase moving at the physiological concentrations (41)–(44) and $[\mathrm{PP_i}] = 10^{-4}$ M, along the Bernoulli template shown above. The theoretical results are depicted as crosses and those of Monte Carlo simulations with a statistics over 10^8 copies as open symbols. Here, the mean growth velocity and the error probability take the values $v \simeq 18.6\ \mathrm{s}^{-1}$ and $\eta \simeq 1.13 \times 10^{-4}$, computed with the IFS method over a template of length $L = 10^5$.

If the pool of nucleotides is no longer balanced as in Fig. 2 where the dGTP concentration takes a much lower value than for the three other nucleotides, the replication errors become more frequent, especially at the locations where the template has the complementary nucleotide C. At these locations, a substitution with the nucleotide A in place of G has a probability of about $\bar{\mu} \simeq 0.04$–0.05. Such events do not occur for the locations $226 \leqslant l \leqslant 250$ where the template sequence does not contain the nucleotide C in the example

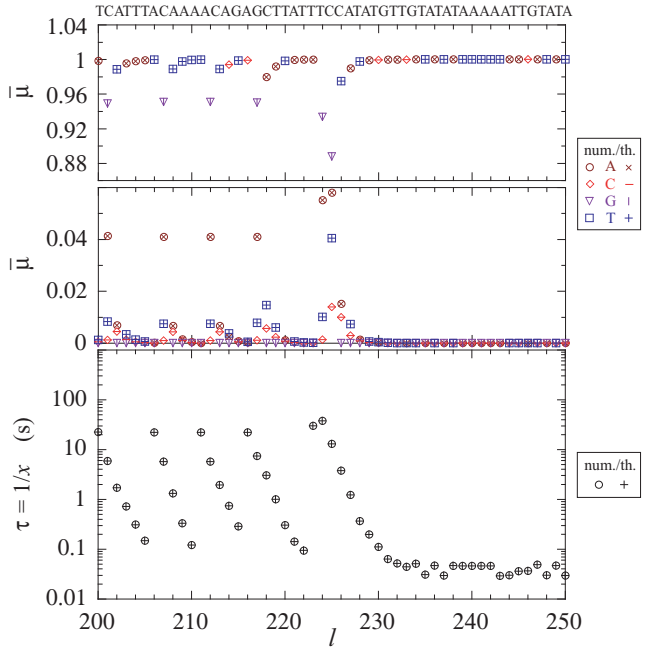

FIG. 2. Exo$^-$ human mitochondrial DNA polymerase: bulk probabilities and waiting times of the polymerase moving at the physiological concentrations (41), (42), (44), and $[\mathrm{PP_i}] = 10^{-4}$ M, but $[\mathrm{dGTP_i}] = 10^{-9}$ M, along the Bernoulli template shown above. The theoretical results are depicted as crosses and those of Monte Carlo simulations with a statistics over 10^8 copies as open symbols. Here, the mean growth velocity and the error probability take the values $v \simeq 9.02 \times 10^{-2}$ s^{-1} and $\eta \simeq 2.98 \times 10^{-2}$, computed with the IFS method over a template of length $L = 10^5$.

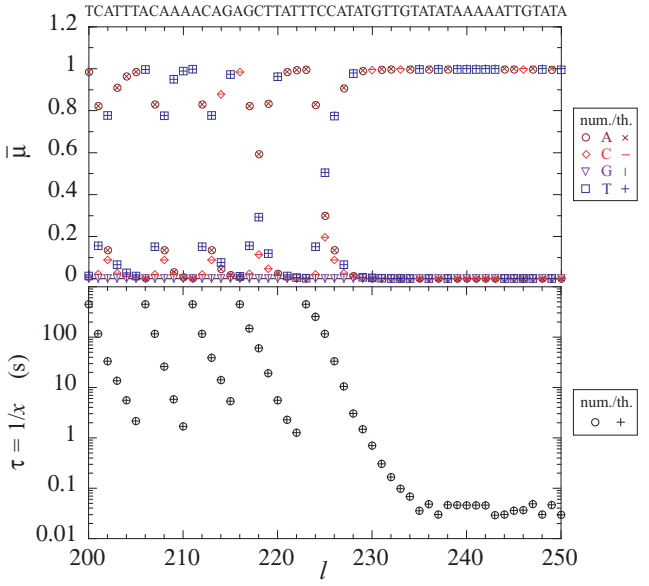

FIG. 3. Exo$^-$ human mitochondrial DNA polymerase: bulk probabilities and waiting times of the polymerase moving at the physiological concentrations (41), (42), (44), and $[\mathrm{PP_i}] = 10^{-4}$ M, but $[\mathrm{dGTP_i}] = 0$, along the Bernoulli template shown above. The theoretical results are depicted as crosses and those of Monte Carlo simulations with a statistics over 10^8 copies as open symbols. Here, the mean growth velocity and the error probability take the values $v \simeq 7.50 \times 10^{-3}$ s^{-1} and $\eta \simeq 0.355$, computed with the IFS method over a template of length $L = 10^5$.

of Fig. 2. In the lower panel of Fig. 2, the waiting time is seen to become quite large when a nucleotide C occurs in the template sequence. Thereafter, the waiting time tends to decrease progressively to the normal level.

If dGTP is no more present in the surrounding solution, the probabilities of replication errors and the waiting times are even larger at the template locations with a nucleotide C, as seen in Fig. 3, in which case the error probability is very large $\eta \simeq 0.355$. Moreover, substitutions by G are no longer occurring, although they were still possible in the conditions of Fig. 2.

Instead, if dGTP is supplied at a concentration larger than the physiological one, the patterns of bulk probabilities and waiting times are different, as observed in Fig. 4. Under these conditions, the replication errors tend to occur especially around the locations where T is present. The reason is again that the polymerization rate constant of G:T is the largest among those for incorrect pairings while the corresponding dissociation constant remains relatively low, as seen in Table I. As in the previous conditions, the waiting time of the polymerase is larger where substitutions with incorrect pairs occur.

All these sequence-specific effects due to template heterogeneity are obtained with the IFS method much faster than with Monte Carlo simulations.

D. Sequence-averaged properties

Aside from knowing how the replication properties vary along the template, we can also consider the global properties

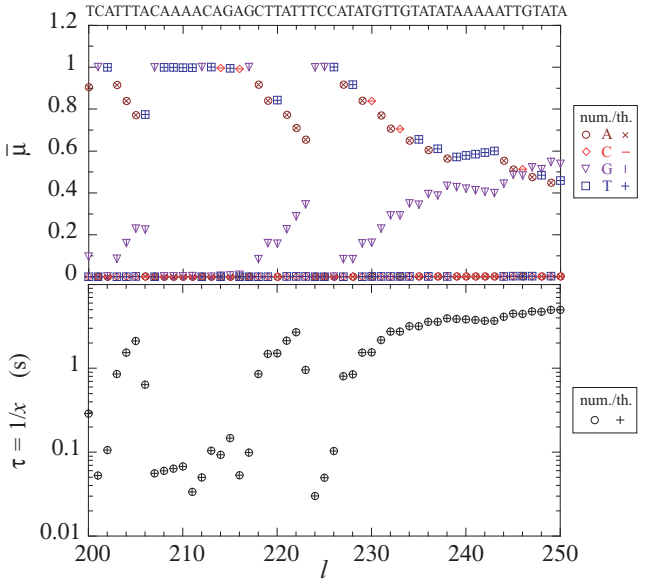

FIG. 4. Exo$^-$ human mitochondrial DNA polymerase: bulk probabilities and waiting times of the polymerase moving at the physiological concentrations (41), (42), (44), and $[\mathrm{PP_i}] = 10^{-4}$ M, but $[\mathrm{dGTP}] = 10^{-3}$ M, along the Bernoulli template shown above. The theoretical results are depicted as crosses and those of Monte Carlo simulations with a statistics over 10^8 copies as open symbols. Here, the mean growth velocity and the error probability take the values $v \simeq 1.54$ s^{-1} and $\eta \simeq 7.88 \times 10^{-2}$, computed with the IFS method over a template of length $L = 10^5$.

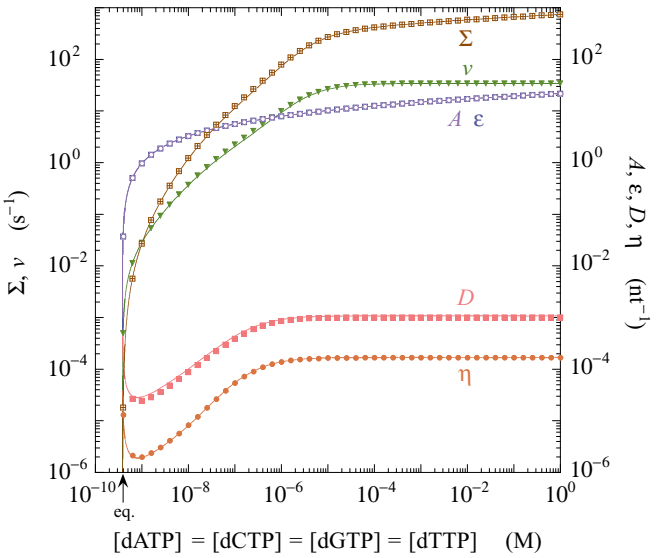

FIG. 5. Exo$^-$ human mitochondrial DNA polymerase at the pyrophosphate concentration $[PP_i] = 10^{-4}$ M: velocity v, affinity $A = \epsilon + D$, free-energy driving force ϵ, conditional Shannon disorder D, error probability η, and entropy production rate $\Sigma = Av$ versus equal nucleotide concentrations. The dots are obtained by Monte Carlo simulations generating 10^5 copies of maximum length 10^5. The lines depict the results of the IFS run over sequences of length 10^5.

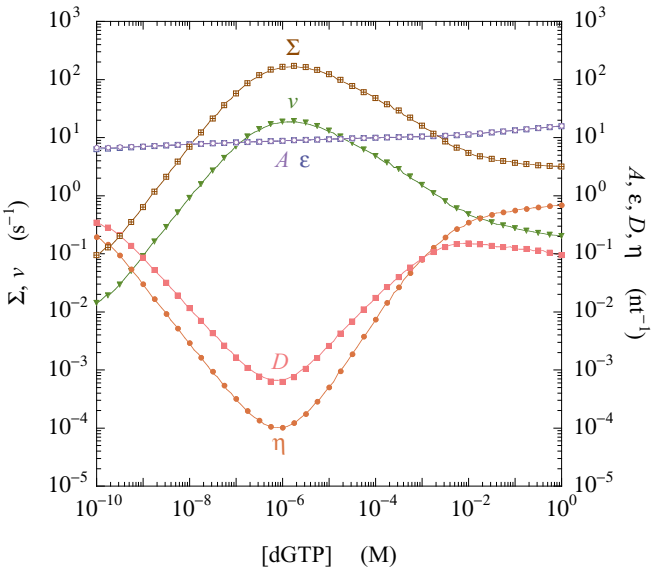

FIG. 6. Exo$^-$ human mitochondrial DNA polymerase at the physiological concentrations (41), (42), (44) and pyrophosphate concentration $[PP_i] = 10^{-4}$ M: velocity v, affinity $A = \epsilon + D$, free-energy driving force ϵ, conditional Shannon disorder D, error probability η, and entropy production rate $\Sigma = Av$ versus the concentration [dGTP]. The dots are obtained by Monte Carlo simulations generating 10^5 copies of maximum length 10^5. The lines depict the results of the IFS run over sequences of length 10^4.

obtained by averaging over the whole template sequence, such as the mean growth velocity (12), the error probability (24), the free-energy driving force (28), the conditional Shannon disorder (29), the affinity (30), and the entropy production rate (27).

In Fig. 5, these quantities are shown as a function of equal nucleotide concentrations. For these global properties as well, there is a nice agreement between the results of the Monte Carlo simulations (dots) and those of the IFS method (lines). The affinity $A = \epsilon + D$ is observed to vanish at the

equilibrium concentration:

$$[\text{dATP}] = [\text{dCTP}] = [\text{dGTP}] = [\text{dTTP}] \simeq 3.83 \times 10^{-10}\ \text{M}, \tag{46}$$

while the mean growth velocity v is zero at the

linear growth threshold:

$$[\text{dATP}] = [\text{dCTP}] = [\text{dGTP}] = [\text{dTTP}] \simeq 3.88 \times 10^{-10}\ \text{M}. \tag{47}$$

Since both values are very close to each other, the interval of sublinear growth in time is not apparent in Fig. 5. At large concentrations away from equilibrium, the mean growth velocity, the error probability, and the conditional Shannon disorder reach plateaus at the following values:

$$v_\infty \simeq 34\ \text{s}^{-1}, \tag{48}$$

$$\eta_\infty \simeq 1.7 \times 10^{-4}, \tag{49}$$

$$D_\infty \simeq 1.0 \times 10^{-3}, \tag{50}$$

while the free-energy driving force behaves as

$$\epsilon_\infty \simeq \ln \frac{[\text{dNTP}]}{3.84 \times 10^{-10}\ \text{M}}, \tag{51}$$

in agreement with the values obtained at full speed in Ref. [32]. The mean growth velocity manifests the key features of Michaelis-Menten kinetics, namely, the plateau at large nucleotide concentrations after the increase at lower concentrations. Between equilibrium and the full speed regime, we notice that the error probability η and the disorder D decrease to lower values than at full speed. Figure 5 is obtained under the same conditions as in Figs. 5 and 6 of Ref. [32]. We see that the IFS method solves the discrepancy that existed between the results of Monte Carlo simulations and the theoretical expectations of the Bernoulli- and Markov-chain models ignoring the effects of sequence heterogeneity [32]. Another difference is that the conditional Shannon disorder is here calculated by its exact expression (29) while it was estimated in Ref. [32] by the approximation $D \simeq \eta \ln(3e/\eta)$ using the error probability η.

In order to understand the consequences of imbalance in the nucleotide pool, the quantities of interest are depicted in Fig. 6 as a function of the dGTP concentration with physiological concentrations for the three other nucleotides. The minimal value of the error probability is observed at $[\text{dGTP}] \simeq 0.8\ \mu\text{M}$ while the mean growth velocity reaches its maximal value at $[\text{dGTP}] \simeq 1.57\ \mu\text{M}$, which are of the same order of magnitude as the three other concentrations (41), (42), and (44). Otherwise, Fig. 6 shows that the replication errors become more frequent as the imbalance increases between the four nucleotide concentrations.

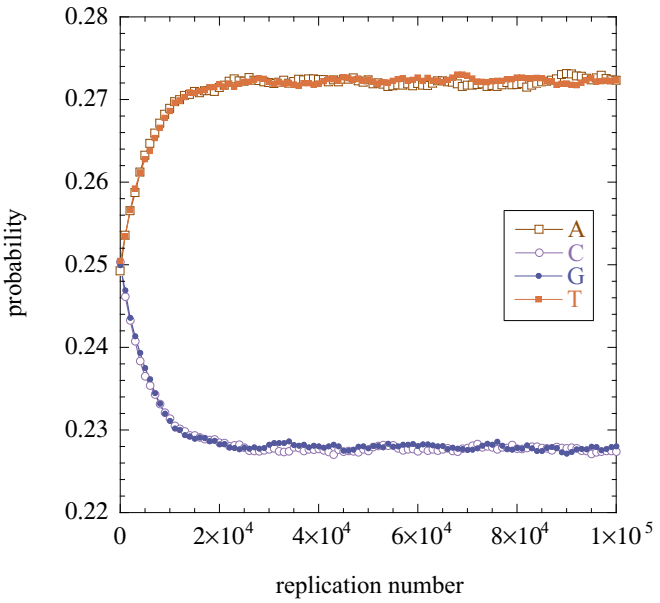

FIG. 7. Exo$^-$ human mitochondrial DNA polymerase at the physiological concentrations (41)–(44) and $[PP_i] = 10^{-4}$ M: probabilities of the four bases composing DNA strands of length 10^6 versus the number of replications during the Monte Carlo simulation of successive replications. The evolution starts from a Bernoulli sequence with the equal probabilities of the four bases. The mean growth velocity and error probability are, respectively, $v \simeq 18.8$ and $\eta \simeq 1.1 \times 10^{-4}$.

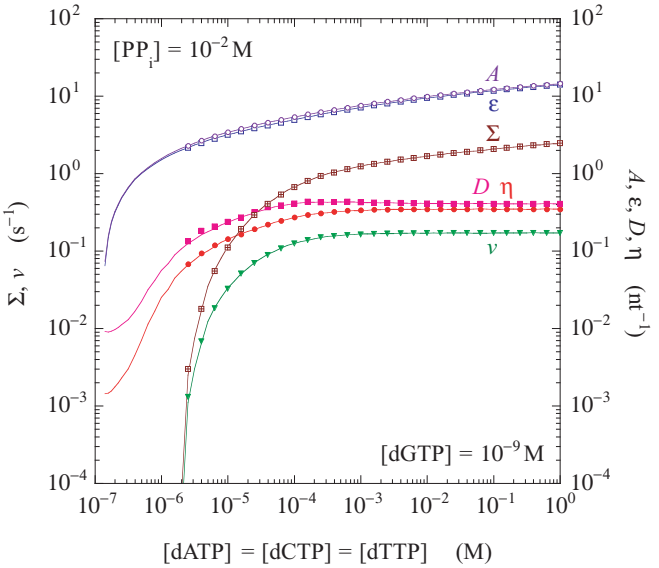

FIG. 8. Exo$^-$ human mitochondrial DNA polymerase at the concentrations $[dGTP] = 10^{-9}$ M and $[PP_i] = 10^{-2}$ M: velocity v, affinity $A = \epsilon + D$, free-energy driving force ϵ, conditional Shannon disorder D, error probability η, and entropy production rate $\Sigma = Av$ versus the equal nucleotide concentrations [dATP] = [dCTP] = [dTTP]. The dots are obtained by Monte Carlo simulations generating 10^5 copies of maximum length 10^5. The lines depict the results of the IFS run over sequences of length 10^5.

E. Molecular evolution after successive replications

Any imbalance in the nucleotide pool generates replication errors, which may affect the genome composition during evolution. As revealed by Eqs. (41)–(44), the cellular nucleotide metabolism does not precisely maintain equal nucleotide concentrations. The total concentration of the complementary nucleotides dATP and dTTP is typically larger than the total concentration of dCTP and dGTP [56]. As a consequence, substitutions with A and T are expected to be more frequent than by C or G upon replication errors. This phenomenon is indeed observed in Fig. 7, showing the evolution of DNA composition after many successive replications. The evolution is generated by a Monte Carlo simulation starting with a template given by a Bernoulli chain with equal probabilities $\tilde{\mu}_0(\mathrm{A}) = \tilde{\mu}_0(\mathrm{C}) = \tilde{\mu}_0(\mathrm{G}) = \tilde{\mu}_0(\mathrm{T}) = \frac{1}{4}$. At each replication, a copy is generated by Monte Carlo simulation and its reversal is used as template for the next replication. The copy sequence is reversed before using it as template because the synthesis of the copy proceeds in the 5'-to-3' direction along a template with the opposite direction [1]. In Fig. 7, the evolution of the four nucleotide frequencies in the DNA sequence can be fitted by

$$\tilde{\mu}_k(\mathrm{A}) \simeq \tilde{\mu}_k(\mathrm{T}) \simeq 0.272 - 0.022\ \exp(-\Gamma k), \quad (52)$$

$$\tilde{\mu}_k(\mathrm{C}) \simeq \tilde{\mu}_k(\mathrm{G}) \simeq 0.228 + 0.022\ \exp(-\Gamma k), \quad (53)$$

as a function of the replication number k. The relaxation rate towards the mean asymptotic value is evaluated to be

$$\Gamma \simeq 0.000\,19 \quad \text{per replication}, \quad (54)$$

which is of the same order of magnitude as the mean error probability $\eta \simeq 0.000\,11$, as expected. A similar drift is also observed [60] for the average nucleotide concentrations [56] that are about 10 times larger than the values (41)–(44) for normal cells, which are here used.

A remarkable result is that the bases A and T are known to be more frequent than C and T in the DNA sequences of many organisms [61,62]. Therefore, the observed nucleotide frequencies could be mainly determined by the average intracellular level of nucleotide concentrations, which is typically lower in dCTP and dGTP than in dATP and dTTP [56].

F. Regime of anomalous drift

Anomalous drift with a sublinear growth in time manifests itself in the presence of strong disorder if the growth is slow enough. In order to enhance this effect due to sequence heterogeneity, the nucleotide concentrations are imbalanced with equal concentrations for the nucleotides dATP, dCTP, and dTTP, but the very small concentration $[dGTP] = 10^{-9}$ M. The imbalance in the nucleotide concentrations generates a large variability of the rates along the template, hence disorder in the landscape where the DNA polymerase is moving. Moreover, the pyrophosphate concentration here takes the larger value $[PP_i] = 10^{-2}$ M, which tends to shift equilibrium towards larger nucleotide concentrations, as seen in Fig. 8. Under these conditions, the affinity $A = \epsilon + D$ is zero at the equilibrium concentration:

$$[\mathrm{dATP}] = [\mathrm{dCTP}] = [\mathrm{dTTP}] \simeq 1.28 \times 10^{-7}\ \mathrm{M}, \quad (55)$$

and the mean growth velocity v at the

linear growth threshold:

$$[\mathrm{dATP}] = [\mathrm{dCTP}] = [\mathrm{dTTP}] \simeq 2.33 \times 10^{-6}\ \mathrm{M}, \quad (56)$$

as calculated by the IFS method with a template sequence of length $L = 10^6$. In-between these two concentration values, the growth is sublinear in time so that the mean growth velocity and the entropy production rate remain equal to zero. However, the quantities defined per incorporated nucleotide, i.e., the affinity A, the free-energy driving force ϵ, the disorder D, and the error probability η, continue to take nonvanishing values. In this regime, the computational time required to have long enough copies in Monte Carlo simulations becomes prohibitive, although the IFS method continues to provide accurate estimations of the different quantities. This explains that, in Fig. 8, the dots giving the results of Monte Carlo simulations cannot be obtained below the concentration of 2×10^{-6} M, although the results of the IFS method are still plotted as lines. Otherwise, there is again good agreement between dots and lines.

At large nucleotide concentrations, plateaus are observed as before for the mean growth velocity, the error probability, and the conditional Shannon disorder:

$$v_\infty \simeq 0.17\ \mathrm{s}^{-1}, \quad (57)$$

$$\eta_\infty \simeq 0.35, \quad (58)$$

$$D_\infty \simeq 0.41, \quad (59)$$

while the free-energy driving force behaves as

$$\epsilon_\infty \simeq \ln \frac{[\mathrm{dNTP}]}{8.0 \times 10^{-7}\ \mathrm{M}}, \quad (60)$$

but the velocity is smaller while the error probability and the disorder are much larger because of the replication errors due to the nucleotide imbalance. The free-energy driving force takes the smaller value (60) than (51) because the pyrophosphate concentration is here larger and, moreover, the more frequent incorrect pairings tend to decrease the free energy by their larger dissociation constant $K^{m}_{n|c}$ given in Table I.

IV. Exo$^-$ HUMAN MITOCHONDRIAL DNA POLYMERASE MOVING ALONG BINARY TEMPLATES

The purpose of this section is to demonstrate another effect of sequence heterogeneity, which manifests itself generically at any nucleotide concentrations. This effect is the sequence dependence of the local velocities $\{x_l\}_{l=1}^{L}$ of the DNA polymerase or, equivalently, of its waiting times $\{\tau_l = 1/x_l\}_{l=1}^{L}$. As a consequence, the local velocities vary along the template sequence and they form a distribution. Remarkably, this distribution may be continuous or fractal depending on nucleotide concentrations. In order to get numerical evidence for this phenomenon, we consider the motion of the DNA polymerase along a binary sequence taken as a Bernoulli chain with the probabilities $\nu_\mathrm{A} = \nu_\mathrm{T} = \frac{1}{2}$, but $\nu_\mathrm{C} = \nu_\mathrm{G} = 0$. The surrounding solution is supposed to contain only the nucleotides dATP and dTTP in order to avoid substitutions with the bases C and G, i.e., $[\mathrm{dCTP}] = [\mathrm{dGTP}] = 0$. Moreover, the pyrophosphate concentration is again taken at the large value $[\mathrm{PP_i}] = 10^{-2}$ M.

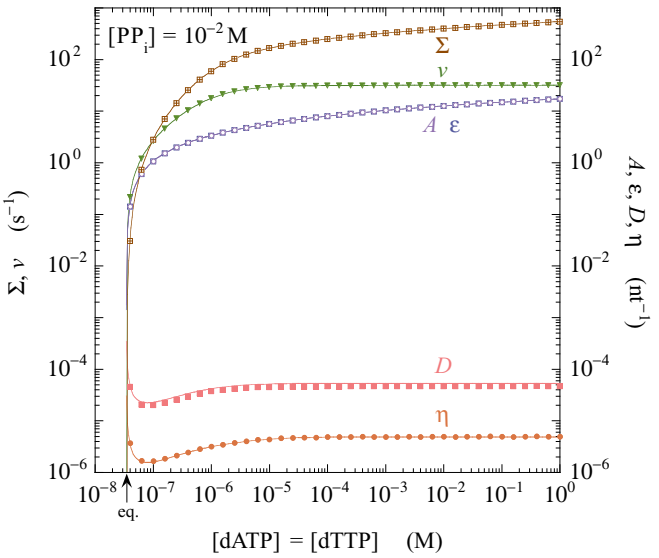

FIG. 9. Exo$^-$ human mitochondrial DNA polymerase at the pyrophosphate concentration $[\mathrm{PP_i}] = 10^{-2}$ M, moving along a binary Bernoulli template composed of the two bases A and T: velocity v, affinity $A = \epsilon + D$, free-energy driving force ϵ, conditional Shannon disorder D, error probability η, and entropy production rate $\Sigma = Av$ versus the equal nucleotide concentrations $[\mathrm{dATP}] = [\mathrm{dTTP}]$. The dots are obtained by Monte Carlo simulations generating 10^5 copies of maximum length 10^5. The lines depict the results of the IFS run over sequences of length 10^5.

Here, the IFS is also given by Eq. (45), now acting in the two-dimensional space of the local partial velocities $\mathbf{v}_l = \{v_{\mathrm{A},l}, v_{\mathrm{T},l}\} \in \mathbb{R}^2$. Because of Eq. (23), the local velocities are related to the partial ones according to

$$x_l = v_{\mathrm{A},l}\,\mu(\mathrm{A},l) + v_{\mathrm{T},l}\,\mu(\mathrm{T},l), \quad (61)$$

where $\mu(\mathrm{A},l)$ and $\mu(\mathrm{T},l)$ are the corresponding tip probabilities.

A. Sequence-averaged properties

Figure 9 shows the sequence-averaged properties as a function of equal nucleotide concentrations $[\mathrm{dATP}] = [\mathrm{dTTP}]$. Here, the error probability η is much smaller than in the presence of the four nucleotides. The reason is that, in the presence of the four nucleotides, substitutions with the other bases C and G occur at a relatively larger rate than with the bases A and T.

At full speed, the plateau values of the mean growth velocity, the error probability, and the conditional Shannon disorder are given by

$$v_\infty \simeq 31.5\ \mathrm{s}^{-1}, \quad (62)$$

$$\eta_\infty \simeq 4.9 \times 10^{-6}, \quad (63)$$

$$D_\infty \simeq 5.3 \times 10^{-5}, \quad (64)$$

while the free-energy driving force behaves as

$$\epsilon_\infty \simeq \ln \frac{[\mathrm{dNTP}]}{3.465 \times 10^{-8}\,\mathrm{M}}, \tag{65}$$

according to the IFS method with a template sequence of length $L = 10^5$. The numerical evaluation of the conditional Shannon disorder is here made difficult by the low error probability. Even with a Monte Carlo statistics using 10^5 copies, the disorder is still underestimated at the value $D_\infty^{(\mathrm{MC})} \simeq 4.7 \times 10^{-5}$, which is lower than the value (64) predicted by the IFS method. Numerical convergence is observed if the statistics is increased, which becomes computationally expensive in the present conditions, although the IFS method quickly provides an accurate value.

In the conditions of Fig. 9, we have the

equilibrium concentration:

$$[\mathrm{dATP}] = [\mathrm{dTTP}] \simeq 3.46 \times 10^{-8}\ \mathrm{M}, \tag{66}$$

where $A = 0$ and the

linear growth threshold:

$$[\mathrm{dATP}] = [\mathrm{dTTP}] \simeq 3.50 \times 10^{-8}\ \mathrm{M}, \tag{67}$$

where $v = 0$, as calculated by the IFS method with a template sequence of length $L = 10^6$. Therefore, the regime of sublinear growth in time extends over a very small interval of concentrations because the nucleotide concentrations [dATP] and [dTTP] are well balanced since they are equal. However, the purpose is here to investigate the distribution of local velocities along the binary template.

B. Fractal distribution of velocities

Figure 10 shows the distribution of the local partial velocities $\{v_{\mathrm{A},l}\}_{l=1}^{L}$ as a function of equal nucleotide concentrations [dATP] = [dTTP] in the same conditions as in Fig. 9. We observe in Fig. 10 that the local partial velocities $v_{\mathrm{A},l}$ are distributed around either large values of order one or small values of the order $v_\mathrm{A} \simeq 1300 \times [\mathrm{dATP}]\ \mathrm{s}^{-1}$. The large values occur with a bulk probability very close to one, although the low values occur with a tiny probability of the same order as the error probability (63). Therefore, the distribution is dominated by the large values due to Watson-Crick pairings, which manifest a self-similar structure as the nucleotide concentration increases. The local partial velocities $\{v_{\mathrm{T},l}\}_{l=1}^{L}$ have a similar distribution as in Fig. 10.

The corresponding distribution of the local velocities (61) is shown in Fig. 11 versus the nucleotide concentrations as in Fig. 10. As aforementioned, the term corresponding to incorrect pairings is very small so that x_l is essentially determined by the term due to Watson-Crick correct pairings $\tilde{n}_l{:}n_l$ where $\tilde{n}_l$ denotes the base that is complementary to the base n_l of the template: $x_l \simeq v_{\tilde{n}_l,l}\,\mu(\tilde{n}_l,l)$. However, this latter expression is given by a partial velocity $v_{\tilde{n}_l,l}$ multiplied by the associated tip probability $\mu(\tilde{n}_l,l)$. The partial velocity has the distribution depicted in Fig. 10, but the tip probability is also distributed. Consequently, the product of both quantities (giving x_l) has a broader distribution than the local partial velocity. This explains the fact that the self-similar structures are thicker in Fig. 11 than in Fig. 10.

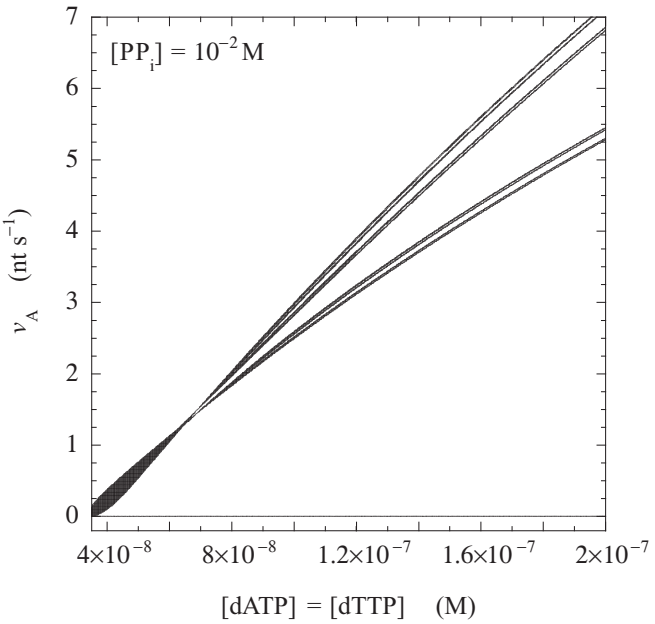

FIG. 10. Exo$^-$ human mitochondrial DNA polymerase at the pyrophosphate concentration $[\mathrm{PP_i}] = 10^{-2}$ M, moving along a binary Bernoulli template composed of the two bases A and T: the distribution of the local partial velocity v_A versus the equal concentrations [dATP] = [dTTP]. The computation is carried out with the IFS method along a sequence of length 10^4.

This is confirmed by the calculation of the fractal dimensions in Fig. 12 for both distributions. The dimension of the distribution F_v of local partial velocities is computed from the number of points of the distribution in a ball of radius r around

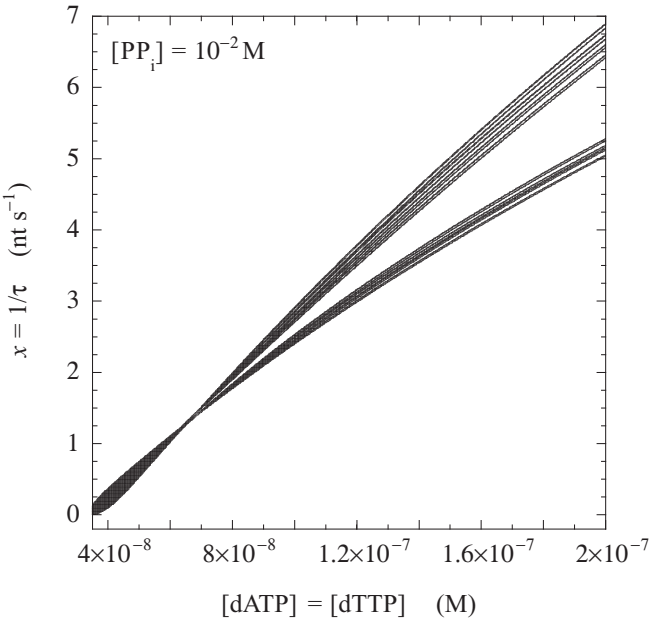

FIG. 11. The distribution of the local velocity x versus the equal concentrations [dATP] = [dTTP] under the same conditions as in Fig. 10.

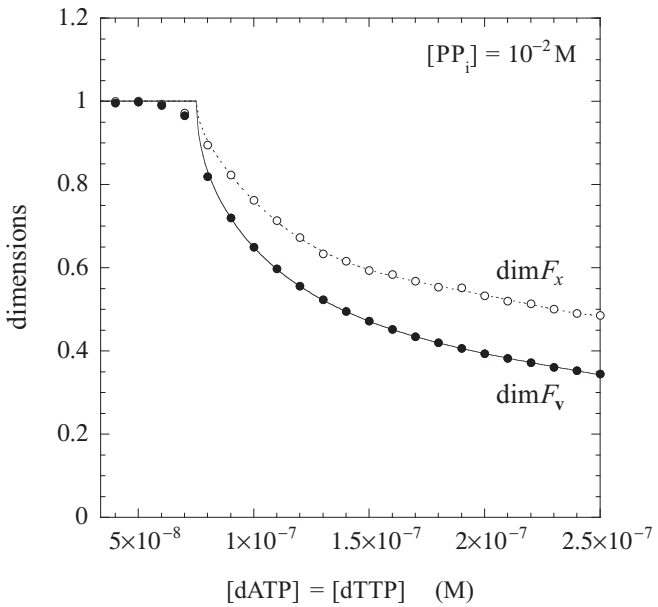

FIG. 12. The fractal dimensions of the support of the distributions of the local partial velocities $\mathbf{v}$ (dim$F_\mathbf{v}$) and the local velocity x (dimF_x) versus the equal concentrations [dATP] = [dTTP] under the same conditions as in Fig. 10.

a set of $N_{\rm ref}$ reference points:

$$N_{\rm v}(r) \equiv \sum_{j=1}^{N_{\rm ref}} \sum_{l=1}^{L} \theta(r - \|\mathbf{v}_l - \mathbf{v}_{jL/N_{\rm ref}}\|) \tag{68}$$

in the space of partial velocities $\mathbf{v}_l$, where $\theta(\cdot)$ denotes the Heaviside function equal to one if its argument is positive and zero otherwise, while $\|\cdot\|$ is the Euclidean distance. As the radius r decreases, this number scales as

$$N_{\rm v}(r) \sim r^{{\rm dim}F_\mathbf{v}}, \tag{69}$$

with an exponent giving the dimension dim$F_\mathbf{v}$ of the distribution in the space of partial velocities $\mathbf{v}_l$ [63]. A similar expression holds for the dimension dimF_x of the distribution F_x of local velocities x_l. In Fig. 12, the dimensions are computed for the distributions obtained with the IFS method running along a template sequence of length $L = 10^5$ and with $N_{\rm ref} = 5000$ reference points. We observe that the dimensions are smaller than the unit value for the concentrations

$$[{\rm dATP}] = [{\rm dTTP}] \gtrsim 7.5 \times 10^{-8}\ {\rm M}, \tag{70}$$

where the distributions are fractal, manifesting self-similar structures. For lower concentrations, the dimensions reach the unit value, in which case the distribution of the local velocities x_l has a continuous support in the one-dimensional space of x. In the fractal regime (70), the dimension is smaller in the space of partial velocities $\mathbf{v}$ than in the space of local velocities x for the reason explained here above that the local velocities x are given by Eq. (61) where the tip probabilities are distributed as well as the partial velocities $\mathbf{v}$, which has for consequence to broaden the distribution of the local velocities (61).

TABLE II. Exo$^-$ human mitochondrial DNA polymerase γ at 37 °C: the rates of the error-free model (71) at the concentrations [dATP] = [dTTP] = 10^{-7} M, [dCTP] = [dGTP] = 0, and [PP$_i$] = 10^{-2} M.

n_l	n_{l+1}	$W^{\rm p}_{+\tilde{n}_l\ n_l}$ (s^{-1})	$W^{\rm p}_{n_{l+1}\ -\tilde{n}_l\ n_l}$ (s^{-1})
A	A	3.56	1.07
A	T	3.56	1.11
T	A	4.99	1.92
T	T	4.99	2.00

C. Error-free models

The self-similar structures of the distribution of local partial velocities can already be understood by neglecting replication errors. Indeed, the error probability is very small under the present conditions so that the replication process can be described by error-free models, in which only Watson-Crick pairs are possible. The rates of incorrect pairings being very small, they may be assumed to be negligible, so that the copy sequence is precisely the complementary sequence of the template, $\omega = m_1 m_2 \ldots m_l \ldots = \tilde{n}_1 \tilde{n}_2 \ldots \tilde{n}_l \ldots$. Since there is no error, the tip and bulk probabilities are equal to one for Watson-Crick pairs and zero otherwise, so that the relevant local partial velocities are those of Watson-Crick pairs, which coincide with the local velocities x_l according to Eq. (23). Moreover, since the previously formed pair $\tilde{n}_{l-1}{:}n_{l-1}$ is always correct, the error-free model of the human mitochondrial DNA polymerase only uses the kinetic parameters of Table I, so that there is no dependence on n_{l-1}. For error-free kinetics, the IFS (18) thus becomes

$$x_{l-1} = f_{n_l n_{l+1}}(x_l) \quad \text{with} \quad f_{n_l n_{l+1}}(x) = \frac{W^{\rm p}_{+\tilde{n}_l\ n_l}\ x}{W^{\rm p}_{n_{l+1}\ -\tilde{n}_l\ n_l} + x} \tag{71}$$

in terms of the rates (32) and (33) for Watson-Crick correct pairings.

For the equal concentrations [dATP] = [dTTP] = 10^{-7} M, these rates are given in Table II. We note that the attachment rates do not depend on n_{l+1} while the detachment rates have a weak dependence on n_{l+1}. Taking the median values of the detachment rates for n_{l+1} = A and T, the IFS (71) can be approximated by

$$x_{l-1} = f_{n_l}(x_l) \quad \text{with} \quad \begin{cases} f_{\rm A}(x) = \frac{3.56\,x}{1.09+x} & \text{if} \quad n = {\rm A}, \\ f_{\rm T}(x) = \frac{4.99\,x}{1.96+x} & \text{if} \quad n = {\rm T}, \end{cases} \tag{72}$$

with the probabilities $\nu_{\rm A} = \nu_{\rm T} = \frac{1}{2}$. Successive iterations of this simplified IFS are depicted in Fig. 13, showing how the fractal distribution (in red) is generated by iterating random functions $f_n(x)$ backward along the template sequence. We see in Fig. 13 that the fractal distribution extends between the two fixed points of the functions, which are located at the intersections of the functions with the diagonal: $x_n = f_n(x_n)$ for n = A, T. We notice that the IFS (72) satisfies the condition of hyperbolicity, which guarantees the existence of a compact attractor [36].

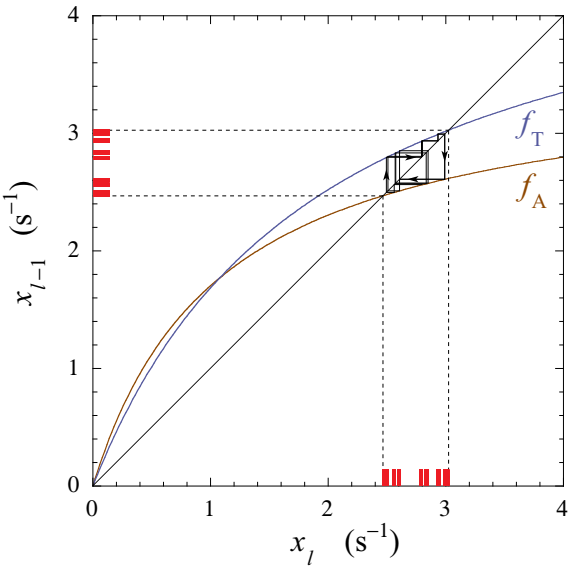

FIG. 13. The IFS of the simple error-free model (72) at the concentrations [dATP] = [dTTP] = 10^{-7} M, [dCTP] = [dGTP] = 0, and [PP$_i$] = 10^{-2} M. The iterations are shown giving the local velocities over a sequence containing 15 successive nucleotides. The fractal is shown in red along the horizontal and vertical axes.

As the concentration increases, the attachment rates also increase because they are essentially proportional to the concentrations. Consequently, the functions $f_n(x)$ take larger values and the interval containing the fractal distribution shifts towards larger values of the velocities, as observed in Figs. 10 and 11.

At the concentration [dATP] = [dTTP] = 10^{-7} M, the fractal dimension can be estimated by finding the root d of $\sum_n f_n'(x_n)^d = 1$, where $f_n' = df_n/dx$ is the derivative and x_n the fixed point of the function $f_n(x)$ [46]. This calculation gives the estimation $d \simeq 0.66$ in good agreement with the numerical computation in the presence of replication errors: $\dim F_{\mathbf{v}} \simeq 0.65$ at this concentration. For the IFS given by Eq. (71), the numerical calculation of the fractal dimension also gives $d \simeq 0.65$.

Therefore, replication errors are not required to generate the fractal distribution of local velocities. We should notice that the regime of anomalous drift also exists in the error-free models if the attachment rates are small enough with respect to the detachment rates [46].

In the presence of replication errors at the concentration [dATP] = [dTTP] = 10^{-7} M, the dominant term corresponding to Watson-Crick pairings in Eq. (61) ranges in the interval $v_{\tilde{n}_l,l}\,\mu(\tilde{n}_l,l) \simeq 2.4$-$3$ s^{-1}, as seen in Fig. 11. The second term corresponding to incorrect pairings $m_l \neq \tilde{n}_l$ is in the range $v_{m_l,l}\,\mu(m_l,l) \simeq 3$–$6 \times 10^{-6}$ s^{-1}. In this regard, the approximation $x_l \simeq v_{\tilde{n}_l,l}\,\mu(\tilde{n}_l,l)$ is well justified. Nevertheless, the tip probabilities of correct pairings are distributed in the range $\mu(\tilde{n}_l,l) \simeq 0.96$–$0.98$, which represents variations of the order of 2%–4%. Since the tip probabilities multiply the local partial velocities to give the local velocities, these latter are more distributed by this amount than the partial velocities, which explains the difference of self-similar structures between Figs. 10 and 11, as well as the larger value $\dim F_x \simeq 0.76$ for the local velocities than the value $\dim F_{\mathbf{v}} \simeq 0.65$ for the partial velocities. In the presence of replication errors, the error-free models provide good approximations for the fractal dimension of the partial velocities because these latter are directly given by the IFS (18). However, the fractal dimension of the local velocities cannot be directly obtained from the sole IFS (18) since they are further broadened by the tip probabilities according to Eq. (23).

V. APPLICATION TO exo$^+$ HUMAN MITOCHONDRIAL DNA POLYMERASE

In this section, the same kinetic model as in Ref. [33] is adopted for the exo$^+$ human mitochondrial DNA polymerase. Here, the analysis is carried out with the exact IFS method of Ref. [10], showing that this method also applies in the presence of a dedicated proofreading mechanism.

A. Kinetics of exo$^+$ DNA polymerases

The IFS method can also be applied to DNA polymerases with exonuclease proofreading. Such enzymes have both the polymerase activity (31) described in previous sections and the exonuclease activity performed by specific subunits that cleave the lastly incorporated nucleotide and release a nucleoside monophosphate to the surrounding solution:

exonuclease activity:

$$\mathrm{E}\cdot\mathrm{DNA}_l + \mathrm{H_2O} \rightleftharpoons \mathrm{E}\cdot\mathrm{DNA}_{l-1} + \mathrm{dNMP}. \tag{73}$$

Proofreading is performed because the polymerase activity is slowed down upon incorrect pairings so that the DNA strand has the time to switch to the exonuclease subunit where the error correction happens [13]. The rates of the exonuclease activity have been obtained in Ref. [33]. The rate of dNMP dissociation by the exonuclease is given by

$$W^{\mathrm{x}}{}_{\substack{n_{l+1}\,n_l\,n_{l-1}}}^{\;-m_l m_{l-1}} \equiv \frac{k^{\mathrm{x}}{}_{\substack{n_l\,n_{l-1}}}^{\;-m_l m_{l-1}}}{Q_{\substack{n_{l+1}\,n_l}}^{\;m_l}} \tag{74}$$

and the rate of the reversed reaction that is dNMP binding by

$$W^{\mathrm{x}}_{\substack{n_l\,n_{l-1}}}{}^{+m_l m_{l-1}} \equiv \frac{k^{\mathrm{x}}_{\substack{n_l\,n_{l-1}}}{}^{+m_l m_{l-1}}[m_l]}{Q_{\substack{n_l\,n_{l-1}}}^{\;m_{l-1}}} \tag{75}$$

with the same denominators (34) as in Sec. III. If the rate constants of dNMP dissociation were known, the rate constants of dNMP binding would be given by

$$k^{\mathrm{x}}_{\substack{n\,n'}}{}^{+mm'} = k^{\mathrm{x}}_{\substack{n\,n'}}{}^{-mm'} \frac{K_{\mathrm{P}}}{K_{\substack{n\,n'}}^{\;mm'}\,c^0} \exp\left(\frac{\Delta G^0}{RT}\right), \tag{76}$$

where $\Delta G^0 = -45.6$ kJ/mol is the standard free enthalpy of hydrolysis into pyrophosphate [28,64,65], R the molar gas constant, T the temperature, and $c^0 = 1$ M the standard concentration [33].

The rates of the exonuclease activity should be added to those of the polymerase activity so that the transition rates in

the kinetic Eqs. (2) here have the following expressions:

$$W_{+m_l m_{l-1},l} = W^{\rm p}_{\substack{+m_l m_{l-1}\\ n_l\, n_{l-1}}} + W^{\rm x}_{\substack{+m_l m_{l-1}\\ n_l\, n_{l-1}}}\,, \tag{77}$$

$$W_{-m_l m_{l-1},l} = W^{\rm p}_{\substack{-m_l m_{l-1}\\ n_{l+1}\, n_l\, n_{l-1}}} + W^{\rm x}_{\substack{-m_l m_{l-1}\\ n_{l+1}\, n_l\, n_{l-1}}}\,. \tag{78}$$

In the regime of steady growth, the mean growth velocity v is equal to the difference

$$v = r^{\rm p} - r^{\rm x} \tag{79}$$

between the polymerase and exonuclease rates [33]. Consequently, the growth velocity is vanishing $v = 0$ if the rates of the polymerase and exonuclease activities are equal to each other: $r^{\rm p} = r^{\rm x}$. Therefore, dNTP continues to be consumed and the system here remains out of thermodynamic equilibrium, even if the growth velocity is zero.

The polymerase and exonuclease rates can be expressed in terms of the mean growth velocity (12), the conditional probabilities (17), and local partial waiting times (13) as

$$r^\rho = \nu^\rho\, v \lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} \sum_{m_{l-1} m_l} \Bigg[W^{\rho}_{\substack{+m_l m_{l-1}\\ n_l\, n_{l-1}}}\, u_{m_{l-1},l-1} - W^{\rho}_{\substack{-m_l m_{l-1}\\ n_{l+1}\, n_l\, n_{l-1}}}\, \mu(m_{l-1}|m_l, l-1)\, u_{m_l,l} \Bigg], \tag{80}$$

where $\nu^{\rm p} = +1$ is the stoichiometric coefficient of the polymerase activity $\rho = {\rm p}$, and $\nu^{\rm x} = -1$ the one of the exonuclease activity $\rho = {\rm x}$. The free-energy driving force is given by Eq. (28) with the sum extended to $\rho = {\rm p,x}$, while the other quantities are as before.

The rate constants of the exonuclease activity have been measured experimentally for the exo$^+$ human mitochondrial DNA polymerase γ at 37 °C and they take the values

$$k^{\rm x}_{-{\rm c}} = 0.05\ {\rm s}^{-1}, \tag{81}$$

$$k^{\rm x}_{-{\rm i}} = 0.4\ {\rm s}^{-1} \tag{82}$$

for the cleavage of correct and incorrect nucleotides [20]. The concentration of monophosphate nucleoside is taken equal to [dNMP] $= 10^{-5}$ M in the range of physiological values [56]. At such concentrations, the rate of dNMP binding is extremely small and negligible.

The IFS is given by Eq. (18) with the rates (77) and (78) and the method is otherwise similar.

B. Sequence-averaged properties

The calculation of the sequence-averaged properties in Ref. [33] is revisited with the IFS method. In Figs. 5 and 6 of Ref. [33], a discrepancy was observed between the results of Monte Carlo simulations and the theoretical expectations with the Bernoulli- and Markov-chain models, especially for the exonuclease rate $r^{\rm x}$. Here, the different quantities of interest are recalculated under the same conditions, but with the IFS method. We see in Fig. 14 that there is now agreement between the results of the Monte Carlo simulations and the IFS method, which provides a complete understanding of the effects of sequence heterogeneity, even in the presence of the exonuclease proofreading mechanism.

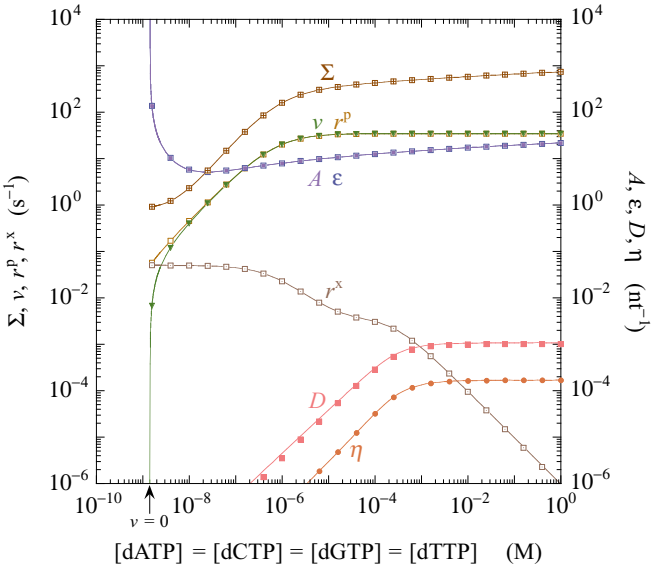

FIG. 14. Exo$^+$ human mitochondrial DNA polymerase at the concentrations [PP$_{\rm i}$] $= 10^{-4}$ M and [dNMP] $= 10^{-5}$ M: velocity v, affinity $A = \epsilon + D$, free-energy driving force ϵ, conditional Shannon disorder D, error probability η, and entropy production rate $\Sigma = Av$ versus the equal nucleotide concentrations [dATP] = [dCTP] = [dGTP] = [dTTP]. The dots are obtained by Monte Carlo simulations generating 10^5 copies of maximum length 10^5. The lines depict the results of the IFS run over sequences of length 10^5.

Otherwise, Fig. 14 confirms the observations already made in Ref. [33]. At zero velocity $v = 0$, the entropy production rate Σ does not vanish because of the ongoing antagonistic exonuclease and polymerase activities at the concentration value:

$$v = 0: \quad [{\rm dATP}] = [{\rm dCTP}] = [{\rm dGTP}] = [{\rm dTTP}] \simeq 1.44 \times 10^{-9}\ {\rm M}, \tag{83}$$

given by the IFS method with a template sequence of total length $L = 10^5$. Furthermore, we see in Fig. 14 the implication of the exonuclease activity for the error probability and the conditional Shannon disorder, i.e., these quantities become significantly smaller than with the sole polymerase activity at nucleotide concentrations lower than about 5×10^{-4} M. The exonuclease proofreading thus improves the replication fidelity by a factor of 100 under physiological nucleotide concentrations, confirming the results of Ref. [33].

Thanks to the IFS method, the effects of imbalance in the nucleotide pool can now be investigated for the polymerase with its exonuclease activity, which has not been possible with the simplifying assumptions of Ref. [33]. Figure 15 shows the sequence-averaged quantities versus equal concentrations for dATP, dCTP, and dTTP, but a much lower concentration for dGTP. Here also, there is agreement between the results of Monte Carlo simulations (dots) and those of the IFS method (lines). The velocity is here vanishing at the concentrations

$$v = 0: \quad [{\rm dATP}] = [{\rm dCTP}] = [{\rm dTTP}] \simeq 1.69 \times 10^{-9}\ {\rm M}, \tag{84}$$

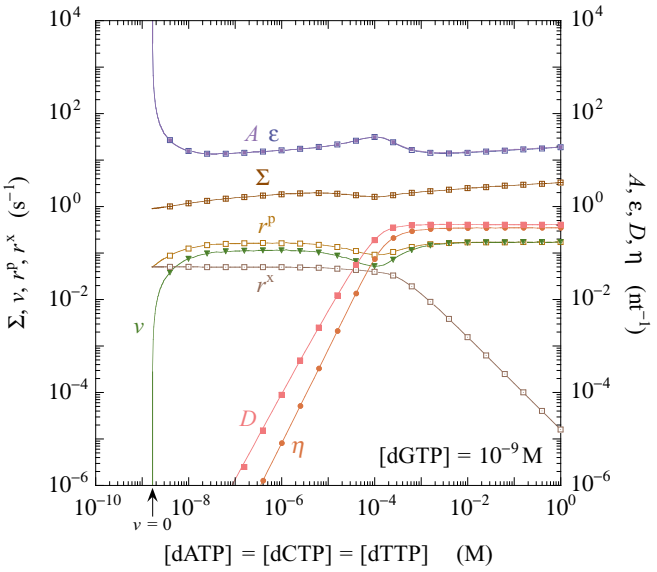

FIG. 15. Exo$^+$ human mitochondrial DNA polymerase at the concentrations [dGTP] $= 10^{-9}$ M, [PP$_i$] $= 10^{-4}$ M, and [dNMP] $= 10^{-5}$ M: velocity v, affinity $A = \epsilon + D$, free-energy driving force ϵ, conditional Shannon disorder D, error probability η, and entropy production rate $\Sigma = Av$ versus the equal nucleotide concentrations [dATP] $=$ [dCTP] $=$ [dTTP]. The dots are obtained by Monte Carlo simulations generating 10^5 copies of maximum length 10^5. The lines depict the results of the IFS run over sequences of length 10^5.

as calculated by the IFS method with a template sequence of $L = 10^5$ bases. The sequence heterogeneity effects manifest themselves by the higher error probability and conditional Shannon disorder than in the conditions of Fig. 14. On the one hand, the polymerase rate $r^{\rm p}$, as well as the mean growth velocity and the entropy production rate, are here lower than in Fig. 14. On the other hand, the exonuclease rate reaches its maximum value $r^{\rm x} \simeq k^{\rm x}_{-{\rm c}}$ given by Eq. (81) at the low nucleotide concentration where the velocity is vanishing, as in Fig. 14. At high concentrations, the exonuclease rate decreases as $r^{\rm x} \sim [{\rm dNTP}]^{-1}$, playing a negligible role in error correction. Therefore, the exonuclease activity achieves proofreading at nucleotide concentrations lower than about 10^{-4} M. The mechanism is that the polymerase is slowed down if an error occurs, allowing the transition to the exonuclease where the incorrect nucleotide is cleaved and the error thus corrected [13,33]. With imbalance in the nucleotide concentrations, the exonuclease proofreading is still very efficient since the error probability is reduced by a factor of 1000 with respect to its full speed value $\eta_\infty \simeq 0.35$, as seen in Fig. 15 at physiological concentrations for dATP, dCTP, and dTTP.

VI. CONCLUSION AND PERSPECTIVES

This paper develops the *iterated function system* (IFS) method for DNA replication. In Ref. [10], the IFS method has been shown to provide the exact long-time solution of the kinetic equations for template-directed copolymerization. For the case of DNA replication, this method determines the probability of any sequence for a copy grown on some template. This probability factorizes into the conditional probabilities to find some nucleotide given that the next one is known. Inserting this factorization in the kinetic equations and taking the long-time limit, we deduce the recurrence (18) running backward along the template and forming an IFS for the local partial velocities. Besides, the local waiting times of the polymerase are calculated by the complementary forward recurrence (19). Together, these two recurrences determine all the statistical properties of the copy sequences, including the local pairing and error probabilities, as well as the kinetic and thermodynamic properties of the growth process, such as the entropy production.

Remarkably, the IFS method allows us to study the effects of sequence heterogeneity, which are (1) the sequence dependence of the pairing probabilities, replication errors, and local velocities of the DNA polymerase along the template; (2) the existence of a regime of anomalous drift of the polymerase close to equilibrium; (3) the continuous or fractal distribution of the local velocities; (4) a modification of the conditional Shannon disorder due to sequence heterogeneity; among others. Moreover, we can now investigate these effects as a function of different concentrations for the four species of nucleotides. Therefore, the IFS method overcomes the difficulties encountered in simplified models where the rates are assumed to depend only on whether pairing is correct or incorrect. For this simplifying assumption to hold, the concentrations of the four nucleotide species should be equal and the template is considered as homogeneous.

Since DNA template sequences are typically heterogeneous, DNA polymerases perform random drifts along the one-dimensional disorder tracks constituted by the templates. Accordingly, the pairing probability, the replication errors, as well as the local velocity of the polymerase undergo sequence-specific variations. In the presence of strong heterogeneity, e.g., caused by some imbalance in the nucleotide concentrations, the random drift may become sublinear instead of linear in time and the mean growth velocity vanishes over a whole interval of nucleotide concentrations between the thermodynamic equilibrium where the growth stops and the threshold of growth with a positive velocity. If some nucleotide imbalance manifests itself, the replication errors become more frequent so that the error probability increases and the mean growth velocity decreases. The imbalance thus enhances the replication errors and leads to base substitutions with the most concentrated nucleotides. As a consequence, the DNA composition may undergo an evolution after many successive replications. The present results suggest that there is a link between the nucleotide metabolism, which is known to be poorer in dCTP and dGTP than in dATP and dTTP [56], and the DNA composition where the bases A and T are observed to be more frequent than C and G [61,62].

Another important effect of sequence heterogeneity explained by the IFS method is that the distribution of the polymerase local velocities along the template is either continuous or fractal. In the latter case, the distribution presents self-similar structures that are here studied in detail for a polymerase drifting along a binary template only composed of the bases A and T. These structures are characterized by their fractal dimension. It is shown that the fractal distribution is not caused by the replication errors because this effect persists in error-free models, to which the IFS method also applies.

Since the IFS method shows that the probabilities of the copy sequences factorize into the conditional probabilities to find successive bases along the copy, the conditional Shannon disorder can be determined with precision taking into account the sequence heterogeneity, which is not possible with theoretical models reducing the template to a homogeneous medium.

The IFS method has been here applied to the kinetics of the human mitochondrial DNA polymerase γ without and with its exonuclease activity. By the way, the results of Refs. [32,33] have been revisited, solving discrepancies that remained between Monte Carlo simulations and theoretical models ignoring the heterogeneity of template sequences. As here shown, the statistical, kinetic, and thermodynamic properties of DNA replication can now be determined with full accuracy, even in the presence of the exonuclease dedicated proofreading mechanism. In this latter case, the exonuclease proofreading activity is shown to be robust with respect to some imbalance in the nucleotide concentrations.

The IFS method opens broad perspectives in our understanding of the kinetics and thermodynamics of DNA replication, transcription, and translation because this method provides the exact solution of the kinetic equations giving in detail the sequence-specific effects [10]. Such effects also concern transcription by RNA polymerases [6–9,66,67], as well as translation by ribosomes [68,69]. In the future, it is expected that the IFS method will be applied as well to these other processes of fundamental importance in biology.

ACKNOWLEDGMENTS

This research is financially supported by the Université libre de Bruxelles (ULB), the Fonds de la Recherche Scientifique-FNRS under the Grant No. PDR T.0094.16 for the project "SYMSTATPHYS," and the Belgian Federal Government under the Interuniversity Attraction Pole project P7/18 "DYGEST".

[1] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology* (Garland, New York, 1998).
[2] T. Harms and R. Lipowsky, Phys. Rev. Lett. **79**, 2895 (1997).
[3] F. Jülicher and R. Bruinsma, Biophys. J. **74**, 1169 (1998).
[4] H.-Y. Wang, T. Elston, A. Mogilner, and G. Oster, Biophys. J. **74**, 1186 (1998).
[5] Y. Kafri, D. K. Lubensky, and D. R. Nelson, Biophys. J. **86**, 3373 (2004).
[6] W. J. Greenleaf and S. M. Block, Science **313**, 801 (2006).
[7] K. M. Herbert, A. La Porta, B. J. Wong, R. A. Mooney, K. C. Neuman, R. Landick, and S. M. Block, Cell **125**, 1083 (2006).
[8] L. Bai, A. Shundrovsky, and M. D. Wang, J. Mol. Biol. **344**, 335 (2004).
[9] L. Bai, R. M. Fulbright, and M. D. Wang, Phys. Rev. Lett. **98**, 068103 (2007).
[10] P. Gaspard, Phys. Rev. Lett. **117**, 238101 (2016).
[11] L. A. Loeb and T. A. Kunkel, Annu. Rev. Biochem. **51**, 429 (1982).
[12] H. Echols and M. F. Goodman, Annu. Rev. Biochem. **60**, 477 (1991).
[13] K. A. Johnson, Annu. Rev. Biochem. **62**, 685 (1993).
[14] T. A. Kunkel and K. Bebenek, Annu. Rev. Biochem. **69**, 497 (2000).
[15] S. S. Patel, I. Wong, and K. A. Johnson, Biochemistry **30**, 511 (1991).
[16] I. Wong, S. S. Patel, and K. A. Johnson, Biochemistry **30**, 526 (1991).
[17] M. J. Donlin, S. S. Patel, and K. A. Johnson, Biochemistry **30**, 538 (1991).
[18] M. J. Longley, D. Nguyen, T. A. Kunkel, and W. C. Copeland, J. Biol. Chem. **276**, 38555 (2001).
[19] A. A. Johnson and K. A. Johnson, J. Biol. Chem. **276**, 38090 (2001).
[20] A. A. Johnson and K. A. Johnson, J. Biol. Chem. **276**, 38097 (2001).
[21] H. R. Lee and K. A. Johnson, J. Biol. Chem. **281**, 36236 (2006).
[22] C. A. Sucato, T. G. Upton, B. A. Kashemirov, J. Osuna, K. Oertell, W. A. Beard, S. H. Wilson, J. Florián, A. Warshel, C. E. McKenna, and M. F. Goodman, Biochemistry **47**, 870 (2008).
[23] M. P. Roettger, M. Bakhtina, and M.-D. Tsai, Biochemistry **47**, 9718 (2008).
[24] L. Zhang, J. A. Brown, S. A. Newmister, and Z. Suo, Biochemistry **48**, 7492 (2009).
[25] L. M. Dieckman, R. E. Johnson, S. Prakash, and M. T. Washington, Biochemistry **49**, 7344 (2010).
[26] R. J. Bauer, M. T. Begley, and M. A. Trakselis, Biochemistry **51**, 1996 (2012).
[27] C. H. Bennett, Biosystems **11**, 85 (1979).
[28] F. Cady and H. Qian, Phys. Biol. **6**, 036011 (2009).
[29] A. K. Sharma and D. Chowdhury, Phys. Rev. E **86**, 011913 (2012).
[30] P. Sartori and S. Pigolotti, Phys. Rev. Lett. **110**, 188101 (2013).
[31] T. Saito, Phys. Rev. E **89**, 062716 (2014).
[32] P. Gaspard, Phys. Rev. E **93**, 042419 (2016).
[33] P. Gaspard, Phys. Rev. E **93**, 042420 (2016).
[34] Y.-S. Song, Y.-G. Shu, X. Zhou, Z.-C. Ou-Yang, and M. Li, J. Phys.: Condens. Matter **29**, 025101 (2017).
[35] M. F. Barnsley and S. Demko, Proc. R. Soc. London, Ser. A **399**, 243 (1985).
[36] A. Lasota and M. C. Mackey, *Chaos, Fractals, and Noise*, 2nd ed. (Springer, New York, 1994).
[37] M. F. Barnsley, Not. Am. Math. Soc. **43**, 657 (1996).
[38] C. Van den Broeck and T. Tél, in *From Phase Transitions to Chaos*, edited by G. Györgyi, I. Kondor, L. Sasvári, and T. Tél (World Scientific, Singapore, 1992), pp. 227–236.
[39] T. Wichmann, A. Giacometti, and K. P. N. Murthy, Phys. Rev. E **52**, 481 (1995).
[40] J. Bernasconi, S. Alexander, and R. Orbach, Phys. Rev. Lett. **41**, 185 (1978).
[41] J. Bernasconi and W. R. Schneider, J. Phys. A: Math. Gen. **15**, L729 (1982).
[42] B. Derrida and Y. Pomeau, Phys. Rev. Lett. **48**, 627 (1982).
[43] B. Derrida, J. Stat. Phys. **31**, 433 (1983).

[44] C. Aslangul, M. Barthelemy, N. Pottier, and D. Saint-James, J. Stat. Phys. **59**, 11 (1990).
[45] J. P. Bouchaud, A. Comtet, A. Georges, and P. Le Doussal, Ann. Phys. (NY) **201**, 285 (1990).
[46] P. Gaspard, J. Stat. Mech.: Theory Exp. (2017) 024003.
[47] I. Prigogine, *Introduction to Thermodynamics of Irreversible Processes* (Charles C. Thomas, Springfield, IL, 1955).
[48] J. Schnakenberg, Rev. Mod. Phys. **48**, 571 (1976).
[49] G. Nicolis, Rep. Prog. Phys. **42**, 225 (1979).
[50] Luo Jiu-Li, C. Van den Broeck, and G. Nicolis, Z. Phys. B **56**, 165 (1984).
[51] D.-Q. Jiang, M. Qian, and M.-P. Qian, *Mathematical Theory of Nonequilibrium Steady States* (Springer, Berlin, 2004).
[52] P. Gaspard, J. Chem. Phys. **120**, 8898 (2004).
[53] D. Andrieux and P. Gaspard, Proc. Natl. Acad. Sci. USA **105**, 9516 (2008).
[54] L. Michaelis and M. L. Menten, Biochem. Z. **49**, 333 (1913).
[55] K. A. Johnson and R. S. Goody, Biochemistry **50**, 8264 (2011).
[56] T. W. Traut, Mol. Cell. Biochem. **140**, 1 (1994).
[57] J. K. Heinonen, *Biological Role of Inorganic Pyrophosphate* (Springer, New York, 2001).
[58] D. T. Gillespie, J. Comput. Phys. **22**, 403 (1976).
[59] D. T. Gillespie, J. Phys. Chem. **81**, 2340 (1977).
[60] P. Gaspard, Philos. Trans. R. Soc. A **374**, 20160147 (2016); **375**, 20170053(E) (2017).
[61] A. Provata, C. Nicolis, and G. Nicolis, Phys. Rev. E **89**, 052105 (2014).
[62] A. Provata, C. Nicolis, and G. Nicolis, Comput. Biol. Chem. **53**, 5 (2014).
[63] K. Falconer, *Fractal Geometry* (Wiley, Chichester, 1990).
[64] C. A. S. A. Minetti, D. P. Remeta, H. Miller, C. A. Gelfand, G. E. Plum, A. P. Grollman, and K. J. Breslauer, Proc. Natl. Acad. Sci. USA **100**, 14719 (2003).
[65] P. A. Frey and A. Arabshahi, Biochemistry **34**, 11307 (1995).
[66] E. A. Galburt, S. W. Grill, A. Wiedmann, L. Lubkowska, J. Choy, E. Nogales, M. Kashlev, and C. Bustamante, Nature (London) **446**, 820 (2007).
[67] M. Dangkulwanich, T. Ishibashi, S. Liu, M. L. Kireeva, L. Lubkowska, M. Kashlev, and C. J. Bustamante, eLife **2**, e00971 (2013).
[68] A. Garai, D. Chowdhury, D. Chowdhury, and T. V. Ramakrishnan, Phys. Rev. E **80**, 011908 (2009).
[69] D. Chowdhury, Phys. Rep. **529**, 1 (2013).