Eur. Phys. J. Special Topics **224**, 825-838 (2015)

Thermodynamics of information processing at the molecular scale

Pierre Gaspard
Center for Nonlinear Phenomena and Complex Systems,
Université Libre de Bruxelles, Code Postal 231, Campus Plaine, B-1050 Brussels, Belgium

The thermodynamics of copolymerization processes leading to the formation and replication of copolymers is presented. Copolymers are natural one-dimensional supports of information, which may be coded in the sequence of monomeric units randomly attached or detached during their synthesis. Multivariate fluctuation relations are here established for copolymerization processes synthesizing Bernoulli and first-order Markov chains. Thereof, the thermodynamic entropy production is deduced and shown to depend on quantities characterizing information possibly stored in the copolymer sequence.

I. INTRODUCTION

DNA provides biological cells with the information storage having the highest known density of two bits coded in a base pair, which is composed of about sixty atoms. This high capacity is interesting to design artificial information storage systems. Synthetic copolymers such as DNA [1, 2], as well as other macromolecules [3, 4], are envisaged for such systems. As a matter of fact, encoding and retrieving information involve physico-chemical processes that are ruled by thermodynamics and the microscopic mechanics of atoms. However, fundamental differences exist between the processes at the macroscopic scale where a strict control of motion can be achieved, and the processes at the molecular scale where the atoms undergo ceaseless irregular motions due to thermal fluctuations. In this regard, the question arises whether we may conceive information storage systems at the molecular scale on the basis of our macroscopic intuition.

It is only at the temperature of a few kelvins that the artificial atomic constructions built in the nineties by scanning tunneling microscopes [5] are not destructed by surface diffusion or desorption. If information should be encoded into atomic structures that remain stable at ambient temperature, they must be made of strong enough chemical bonds. Yet, information should be easily accessible, which is not the case for three-dimensional structures. Reading or writing information in two-dimensional structures needs a strict bidirectional control, which can be achieved by macroscopic mechanical devices (e.g., magnetic- or optical-disk drives). At the molecular scale, such a strict control becomes difficult because of thermal fluctuations. This difficulty is alleviated for one-dimensional molecular structures such as copolymers where information is encoded in a sequence of covalent bonds. Information coding requires that the copolymer is composed of at least two species of monomeric units. In these one-dimensional structures, information can be processed by running a molecular machine along the copolymer.

At this scale, a key role is thus played by the kinetic processes synthesizing and replicating copolymers. Since kinetic processes dissipate energy, thermodynamics is expected to introduce limitations on the rates at which information can be processed.

Recently, important advances have been carried out in the thermodynamics of copolymerization processes without or with a template [6–10]. These processes are described by the mass action law of chemical kinetics. The monomers are randomly attached or detached to the growing copolymer. In this way, a sequence of monomeric units is generated that may contain information transmitted from a template. The remarkable result is that the thermodynamic entropy production depends on the Shannon disorder per monomer in the sequence or the mutual information between the template and the copy, thus establishing a missing link between thermodynamics and information theory [6–10]. These advances shed a new light on the emergence of information at the molecular scale, emphasizing the fact that information processing dissipates energy and is thus ruled by thermodynamics. Even if the kinetic processes are driven by energy consumption, they are fluctuating at the molecular scale so that errors always happen at some rate during information transmission. In particular, these errors are at the origin of genetic mutations, the properties of which depend on the distance from thermodynamic equilibrium.

The purpose of the present paper is to explain how nonequilibrium thermodynamics can be formulated for the formation of copolymers, which may store information at the molecular scale. Different processes should be considered whether the attachment or detachment of monomers depends on the previously incorporated monomeric units or not, leading to sequences behaving as Bernoulli or Markov chains. In this paper, we establish multivariate fluctuation relations for copolymerization processes. These relations are fundamental to deduce the thermodynamic entropy production for nonequilibrium stochastic processes [11]. We also show how to resolve a paradox reminiscent of

Gibbs' paradox [12] in the limit of identical and indistinguishable monomers. Furthermore, the thermodynamics of information transmission is developed to understand how the distance from equilibrium may influence the error probability in the replication of information-containing molecular sequences.

The paper is organized as follows. The multivariate fluctuation relations and the deduction of entropy production are established for the copolymerization of Bernoulli chains in Section II and first-order Markov chains in Section III. The limit of identical and indistinguishable monomers is analyzed for Bernoulli chains in Section II. Section IV is devoted to information transmission processes. Conclusions are drawn in Section V.

II. THE THERMODYNAMICS OF BERNOULLI CHAIN COPOLYMERIZATION

A. Free copolymerization of Bernoulli chains

The description of a copolymer in solution is coarse grained into the sequence of monomeric units m_j composing its chain: $\boldsymbol{\omega} = m_1 m_2 \cdots m_l$. These units are of different species $m_j \in \{1, 2, ..., M\}$. The length of the copolymer is equal to the number of monomeric units in the chain: $l = |\boldsymbol{\omega}|$. The sequence $\boldsymbol{\omega}$ evolves in time because of the random attachments and detachments of monomers m coming or returning to the surrounding solution where they are diluted at some concentrations:

$$m_1 m_2 \cdots m_{l-1} \; + \; m_l \; \rightleftharpoons \; m_1 m_2 \cdots m_{l-1} m_l \, . \tag{1}$$

The solution is supposed to be large enough so that the concentrations of monomers are kept constant during the whole copolymerization process. The process is stochastic and assumed to be ruled by a Markovian master equation [6, 8].

Remarkably, if the attachment rates do not depend on the previously incorporated monomeric unit and the detachment rates only depend on the monomeric unit returning to the solution, i.e.,

$$W(\boldsymbol{\omega} \to \boldsymbol{\omega} m) = w_{+m} \qquad \text{and} \qquad W(\boldsymbol{\omega} m \to \boldsymbol{\omega}) = w_{-m} \qquad \text{for} \quad m = 1, 2, ..., M \, , \tag{2}$$

the growing copolymer forms a Bernoulli chain, as proved in Refs. [8, 9]. The stationary probability of a monomeric unit of species m is given by

$$\mu(m) = \frac{w_{+m}}{w_{-m} + v} \tag{3}$$

where v is the mean growth speed. This latter is determined from the normalization condition $\sum_{m=1}^{M} \mu(m) = 1$ [8, 9].

B. Multivariate fluctuation relation

Here, we establish a symmetry relation for the fluctuations in the numbers of monomers incorporated in the chain during copolymerization. We denote by N_m the number of monomeric units of species m in the chain and by $\mathbf{N} = \{N_1, N_2, ..., N_M\}$ the set of numbers for every monomeric species. The length of the chain grown since the start of copolymerization is given by $l = \sum_{m=1}^{M} N_m$.

The time evolution of the probability that, at time t, the chain has the sequence $m_1 m_2 \cdots m_l$ and contains the numbers $\mathbf{N}$ of monomeric units is ruled by the master equation:

$$\begin{aligned} \frac{d}{dt} P_t(m_1 \cdots m_{l-1} m_l, \mathbf{N}) &= w_{+m_l} P_t(m_1 \cdots m_{l-1}, \mathbf{N} - \mathbf{1}_{m_l}) \\ &+ \sum_{m_{l+1}=1}^{M} w_{-m_{l+1}} P_t(m_1 \cdots m_{l-1} m_l m_{l+1}, \mathbf{N} + \mathbf{1}_{m_{l+1}}) \\ &- \left(w_{-m_l} + \sum_{m_{l+1}=1}^{M} w_{+m_{l+1}} \right) P_t(m_1 \cdots m_{l-1} m_l, \mathbf{N}) \end{aligned} \tag{4}$$

with the notation $\mathbf{N} \pm \mathbf{1}_m = \{N_1, ..., N_{m-1}, N_m \pm 1, N_{m+1}, ..., N_M\}$. In the long-time limit, we can show [8, 9] that the probability factorizes as

$$P_t(m_1 m_2 \cdots m_l, \mathbf{N}) \simeq p_t(\mathbf{N}) \, \mu(m_1) \, \mu(m_2) \, \cdots \, \mu(m_l) \tag{5}$$

in terms of the probability $p_t(\mathbf{N})$ that the chain contains the numbers $\mathbf{N}$ of monomeric units and the probabilities (3). Inserting Eq. (5) in the master equation (4) and summing over all the possible sequences $m_1 \cdots m_l$, we obtain an equation for the time evolution of the probability $p_t(\mathbf{N})$:

$$\frac{d}{dt} p_t(\mathbf{N}) = \hat{L}\, p_t(\mathbf{N}) \tag{6}$$

with the linear operator:

$$\hat{L} = \sum_{m=1}^{M} \left[w_{+m} \left(\hat{E}_m^- - 1 \right) + w_{-m}\, \mu(m) \left(\hat{E}_m^+ - 1 \right) \right] , \tag{7}$$

which is expressed in terms of the creation-annihilation operators

$$\hat{E}_m^{\pm} f(\mathbf{N}) = \exp\left(\pm \frac{\partial}{\partial N_m} \right) f(\mathbf{N}) = f(\mathbf{N} \pm \mathbf{1}_m) . \tag{8}$$

The cumulant generating function of the numbers of monomers incorporated in the chain is defined as

$$Q(\boldsymbol{\lambda}) \equiv \lim_{t\to\infty} -\frac{1}{t} \ln \left\langle \mathrm{e}^{-\boldsymbol{\lambda}\cdot\mathbf{N}} \right\rangle_t \tag{9}$$

in terms of the counting parameters $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, ..., \lambda_M\}$. As shown elsewhere [13], this generating function is given by solving $\hat{L}_{\boldsymbol{\lambda}} \phi = -Q(\boldsymbol{\lambda}) \phi$ for the leading eigenvalue of the modified operator

$$\hat{L}_{\boldsymbol{\lambda}} \equiv \mathrm{e}^{-\boldsymbol{\lambda}\cdot\mathbf{N}}\, \hat{L}\, \mathrm{e}^{+\boldsymbol{\lambda}\cdot\mathbf{N}} . \tag{10}$$

The eigenfunctions of this operator take the form $\phi = \mathrm{e}^{i\mathbf{k}\cdot\mathbf{N}}$ and the leading eigenvalue is obtained for arbitrarily small values of the wave vector $\mathbf{k}$ as

$$Q(\boldsymbol{\lambda}) = \sum_{m=1}^{M} \left[w_{+m} \left(1 - \mathrm{e}^{-\lambda_m} \right) + w_{-m}\, \mu(m) \left(1 - \mathrm{e}^{+\lambda_m} \right) \right] . \tag{11}$$

A fundamental result is that this generating function obeys the symmetry relation

$$Q(\boldsymbol{\lambda}) = Q(\mathbf{A} - \boldsymbol{\lambda}) \tag{12}$$

in terms of the affinities $\mathbf{A} = \{A_1, A_2, ..., A_M\}$ associated with every species of monomeric units:

$$A_m \equiv \ln \frac{w_{+m}}{w_{-m}\, \mu(m)} \qquad (m = 1, 2, ..., M) . \tag{13}$$

As a corollary of the symmetry relation (12), the probabilities of opposite fluctuations in the numbers of monomers incorporated in the copolymer satisfy the multivariate fluctuation relation:

$$\frac{p_t(\mathbf{N})}{p_t(-\mathbf{N})} \simeq_{t\to\infty} \mathrm{e}^{\mathbf{A}\cdot\mathbf{N}} , \tag{14}$$

in terms of the affinities (13).

At equilibrium, the affinities (13) vanish and the principle of detailed balancing is recovered according to which opposite fluctuations are equiprobable: $p_t^{(\mathrm{eq})}(\mathbf{N}) \simeq p_t^{(\mathrm{eq})}(-\mathbf{N})$. Away from equilibrium, the fluctuations manifest a directionality driven by the affinities (13).

C. Entropy production

The thermodynamic entropy production and its non-negativity can be deduced from the multivariate fluctuation relation (14). Indeed, the Kullback-Leibler divergence between the probability distributions of opposite fluctuations is non negative and equal to the statistical average $\langle \mathbf{A} \cdot \mathbf{N} \rangle_t$ [11]. The average values of the currents of monomers from the solution to the copolymer are given by

$$J_m = \lim_{t\to\infty} \frac{1}{t} \langle N_m \rangle_t = \frac{\partial Q}{\partial \lambda_m}(\mathbf{0}) = w_{+m} - w_{-m}\, \mu(m) \qquad (m = 1, 2, ..., M) . \tag{15}$$

Therefore, we obtain the thermodynamic entropy production as

$$\frac{1}{k_{\rm B}}\frac{d_{\rm i}S}{dt} = \mathbf{A}\cdot\mathbf{J} = \sum_{m=1}^{M}\left[w_{+m} - w_{-m}\,\mu(m)\right]\,\ln\frac{w_{+m}}{w_{-m}\,\mu(m)} \geq 0\,, \tag{16}$$

which is non negative as a consequence of the multivariate fluctuation relation (14) and in agreement with the second law of thermodynamics.

Since the currents (15) are the incorporation rates of each type of monomeric units in the chain, the growth speed of the chain is found to be equal to their sum:

$$v = \sum_{m=1}^{M} J_m = \sum_{m=1}^{M}\left[w_{+m} - w_{-m}\,\mu(m)\right]\,. \tag{17}$$

Using Eq. (3), the currents can be expressed as $J_m = v\,\mu(m)$. Consequently, the entropy production can be written as

$$\frac{1}{k_{\rm B}}\frac{d_{\rm i}S}{dt} = A\,v \geq 0 \tag{18}$$

in terms of the average affinity

$$A = \sum_{m=1}^{M}\mu(m)\,A_m = \epsilon + D(\boldsymbol{\omega})\,, \tag{19}$$

which can be decomposed into the free-energy driving force [14]

$$\epsilon = \sum_{m=1}^{M}\mu(m)\,\ln\frac{w_{+m}}{w_{-m}} \tag{20}$$

and the Shannon disorder per monomer

$$D(\boldsymbol{\omega}) = -\sum_{m=1}^{M}\mu(m)\,\ln\mu(m) \geq 0\,, \tag{21}$$

as expected for a Bernoulli chain [6, 8, 9].

The thermodynamics of copolymerization thus follows from the multivariate fluctuation relation in the case of the growth of Bernoulli chains.

D. The limit of identical and indistinguishable monomers

If all the monomers belong to the same species, there is only to consider one attachment rate w_+ and one detachment rate w_-. The growing polymer sequence is fully characterized by its length l. The probability $p_t(l)$ that the length takes the value l at the time t obeys the master equation

$$\frac{dp_t(l)}{dt} = w_+\,p_t(l-1) + w_-\,p_t(l+1) - (w_+ + w_-)\,p_t(l)\,. \tag{22}$$

The solution starting from the initial condition $p_0(l) = \delta_{l,0}$ is given by

$$p_t(l) = {\rm e}^{-(w_+ + w_-)t}\,I_l\left(2\sqrt{w_+ w_-}\,t\right)\,\left(\frac{w_+}{w_-}\right)^{l/2}\,, \tag{23}$$

in terms of the Bessel function $I_l(z) = I_{-l}(z)$ [15, 16]. The proof of the fluctuation relation is thus straightforward:

$$\frac{p_t(l)}{p_t(-l)} = {\rm e}^{A\,l}\,, \tag{24}$$

which holds for all times with the affinity

$$A = \ln \frac{w_+}{w_-} \,. \tag{25}$$

The affinity is here equal to the driving force $A = \epsilon$ and the Shannon disorder is of course vanishing $D = 0$. The growth velocity is given by

$$v = w_+ - w_- \,. \tag{26}$$

The puzzle is to understand how the entropy production of copolymerization (19)-(21) reduces to the one of simple polymerization. Let us consider the copolymerization of two monomeric species $M = 2$ in the limit where they are identical, but still distinguishable. Since they are identical, their attachment and detachment rates tend to

$$w_{+1} = w_{+2} \equiv \tilde{w}_+ \,, \tag{27}$$
$$w_{-1} = w_{-2} \equiv \tilde{w}_- \,. \tag{28}$$

In this limit, the probabilities (3) of both monomers are equal $\mu(1) = \mu(2) = 1/2$, so that the Shannon disorder per monomer (21) is not vanishing, $D = \ln 2$, because they are still distinguishable. The driving force (20) is given by

$$\epsilon = \ln \frac{\tilde{w}_+}{\tilde{w}_-} \tag{29}$$

and the affinity (19) by

$$A = \epsilon + D = \ln \frac{2\,\tilde{w}_+}{\tilde{w}_-} \,. \tag{30}$$

Moreover, the growth speed takes the value

$$v = 2\,\tilde{w}_+ - \tilde{w}_- \,. \tag{31}$$

The comparison with Eqs. (25) and (26) shows that the attachment rate of identical but distinguishable monomers is twice the attachment rate of identical and indistinguishable monomers, while they have the same detachment rate:

$$w_+ = 2\,\tilde{w}_+ \,, \tag{32}$$
$$w_- = \tilde{w}_- \,, \tag{33}$$

as it should. This correspondence between the rates solves the puzzle, which is reminiscent of well-known Gibbs' paradox of equilibrium statistical mechanics [12].

III. THE COPOLYMERIZATION OF MARKOV CHAINS

A. Free copolymerization of first-order Markov chains

Remarkably, the kinetics of free copolymerization can also be exactly solved in the long-time limit if the attachment and detachment rates depend on the previously incorporated monomeric unit as

$$W(\cdots m_{l-1} \to \cdots m_{l-1} m_l) = w_{+m_l|m_{l-1}} \qquad \text{and}$$
$$W(\cdots m_{l-1} m_l \to \cdots m_{l-1}) = w_{-m_l|m_{l-1}} \tag{34}$$

for $m_{l-1}, m_l = 1, 2, ..., M$. In this case, the growing copolymer forms a first-order Markov chain [9]. The exact solution can be expressed in terms of partial speeds given by solving the following equations

$$v_m = \sum_{n=1}^{M} \frac{w_{+n|m}\, v_n}{w_{-n|m} + v_n} \,, \tag{35}$$

for positive roots. The tip probabilities to find some type of monomeric units at the tip of the growing chain are then obtained as

$$\sum_{m=1}^{M} \frac{w_{+n|m}}{w_{-n|m} + v_n}\, \mu(m) = \mu(n) \,, \tag{36}$$

and the conditional probabilities of the monomeric units in the Markov chain as

$$\mu(m|n) = \frac{w_{+n|m}\,\mu(m)}{(w_{-n|m} + v_n)\,\mu(n)}\,, \tag{37}$$

so that the probability of a sequence factorizes as

$$\mu_l(m_1 m_2 \cdots m_l) = \prod_{j=1}^{l-1} \mu(m_j|m_{j+1})\,\mu(m_l)\,. \tag{38}$$

The mean growth speed is found to be equal to

$$v = \sum_{m=1}^{M} v_m\,\mu(m)\,. \tag{39}$$

The bulk probabilities are defined as the stationary probabilities of the monomeric units in the Markov chain: $\bar{\mu}(m) = \sum_{n=1}^{M} \mu(m|n)\,\bar{\mu}(n)$. They are related to the tip probabilities by $\bar{\mu}(m) = \mu(m)\,v_m/v$ [9].

B. Multivariate fluctuation relation

A multivariate fluctuation relation can also be established for the copolymerization of first-order Markov chains by considering the probability that the chain forms the sequence $m_1 m_2 \cdots m_l$ and contains the numbers $\mathbf{N} = \{N_{mn}\}$ of monomeric doublets mn $(m, n = 1, 2, ..., M)$. This probability is ruled by the master equation:

$$\begin{aligned}\frac{d}{dt}\,P_t(m_1 \cdots m_{l-1} m_l, \mathbf{N}) &= w_{+m_l|m_{l-1}}\,P_t(m_1 \cdots m_{l-1}, \mathbf{N} - \mathbf{1}_{m_{l-1}m_l}) \\ &+ \sum_{m_{l+1}=1}^{M} w_{-m_{l+1}|m_l}\,P_t(m_1 \cdots m_{l-1} m_l m_{l+1}, \mathbf{N} + \mathbf{1}_{m_l m_{l+1}}) \\ &- \left(w_{-m_l|m_{l-1}} + \sum_{m_{l+1}=1}^{M} w_{+m_{l+1}|m_l}\right) P_t(m_1 \cdots m_{l-1} m_l, \mathbf{N})\end{aligned} \tag{40}$$

with a similar notation as in Eq. (4). As for the Bernoulli case, the probability factorizes in the long-time limit and we obtain the equation $dp_t/dt = \hat{L}p_t$ for the probabilities $p_t(\mathbf{N})$ that the chain contains the numbers $\mathbf{N}$ of monomeric doublets with the linear operator:

$$\hat{L} = \sum_{m,n=1}^{M} \left[w_{+n|m}\,\mu(m)\left(\hat{E}^{-}_{mn} - 1\right) + w_{-n|m}\,\mu(m|n)\,\mu(n)\left(\hat{E}^{+}_{mn} - 1\right)\right]. \tag{41}$$

Following the same reasoning as before, we get the cumulant generating function:

$$Q(\boldsymbol{\lambda}) = \sum_{m,n=1}^{M} \left[w_{+n|m}\,\mu(m)\left(1 - \mathrm{e}^{-\lambda_{mn}}\right) + w_{-n|m}\,\mu(m|n)\,\mu(n)\left(1 - \mathrm{e}^{+\lambda_{mn}}\right)\right], \tag{42}$$

which obeys the same symmetry relation as Eq. (12), but with the affinities $\mathbf{A} = \{A_{mn}\}$ associated with the monomeric doublets:

$$A_{mn} \equiv \ln \frac{w_{+n|m}\,\mu(m)}{w_{-n|m}\,\mu(m|n)\,\mu(n)} \qquad (m, n = 1, 2, ..., M)\,. \tag{43}$$

Hence, the multivariate fluctuation relation $p_t(\mathbf{N}) \simeq \mathrm{e}^{\mathbf{A}\cdot\mathbf{N}}\,p_t(-\mathbf{N})$ is satisfied for $t \to \infty$ in terms of the affinities (43).

C. Entropy production

The average values of the currents are here given by

$$J_{mn} = \lim_{t\to\infty} \frac{1}{t} \langle N_{mn} \rangle_t = \frac{\partial Q}{\partial \lambda_{mn}}(\mathbf{0}) = w_{+n|m}\, \mu(m) - w_{-n|m}\, \mu(m|n)\, \mu(n) \tag{44}$$

and the thermodynamic entropy production is obtained as

$$\frac{1}{k_\mathrm{B}} \frac{d_\mathrm{i} S}{dt} = \sum_{m,n=1}^{M} \left[w_{+n|m}\, \mu(m) - w_{-n|m}\, \mu(m|n)\, \mu(n) \right] \ln \frac{w_{+n|m}\, \mu(m)}{w_{-n|m}\, \mu(m|n)\, \mu(n)} \geq 0 \,. \tag{45}$$

Now, the currents can be related to the mean growth speed v and the bulk probabilities $\bar{\mu}(n)$ by using Eq. (37) to get $J_{mn} = v\, \bar{\mu}(n)\, \mu(m|n)$. Replacing in the entropy production (45), this latter takes the general form (18) with the average affinity

$$A = \sum_{m,n=1}^{M} \bar{\mu}(n)\, \mu(m|n)\, A_{mn} = \epsilon + D(\boldsymbol{\omega}) \,, \tag{46}$$

which can be split in terms of the free-energy driving force

$$\epsilon = \sum_{m,n=1}^{M} \bar{\mu}(n)\, \mu(m|n)\, \ln \frac{w_{+n|m}}{w_{-n|m}} \tag{47}$$

and the Shannon disorder per monomer

$$D(\boldsymbol{\omega}) = -\sum_{m,n=1}^{M} \bar{\mu}(n)\, \mu(m|n)\, \ln \mu(m|n) \geq 0 \,. \tag{48}$$

For first-order Markov chains, the thermodynamics of copolymerization can thus be deduced from the multivariate fluctuation relation ruling the incorporation of monomeric doublets.

IV. THE THERMODYNAMICS OF INFORMATION TRANSMISSION

The kinetics of copolymerization with a template is complicated to solve in general, except if the attachment and detachment rates only depend on whether the pairing between the monomeric units of the copy and the template is correct or incorrect. In the case of DNA, the (correct) Watson-Crick base pairs are A-T, T-A, C-G, and G-C, among the $M^2 = 16$ possible pairs. Under the assumption that the only different rates are for correct pairing, $w_{\pm\mathrm{c}}$, and incorrect pairing, $w_{\pm\mathrm{i}}$, the kinetics reduces to the free copolymerization of a Bernoulli chain composed of correct and incorrect base pairs [14, 17]. Introducing the error probability η such that

$$\mu(\mathrm{c}) \;=\; \frac{w_{+\mathrm{c}}}{w_{-\mathrm{c}} + v} = 1 - \eta \,, \tag{49}$$

$$\mu(\mathrm{i}) \;=\; \frac{w_{+\mathrm{i}}}{w_{-\mathrm{i}} + v} = \frac{\eta}{M-1} \,, \tag{50}$$

the mean growth speed is given by

$$v = \frac{w_{+\mathrm{c}}}{1-\eta} - w_{-\mathrm{c}} = (M-1)\, \frac{w_{+\mathrm{i}}}{\eta} - w_{-\mathrm{i}} \,, \tag{51}$$

and the conditional Shannon disorder per monomer by

$$D(\boldsymbol{\omega}|\boldsymbol{\alpha}) = -(1-\eta)\, \ln(1-\eta) - \eta\, \ln \frac{\eta}{M-1} \,, \tag{52}$$

which is depicted in Fig. 1. In the limit of low error probability ($\eta \ll 1$), this conditional disorder vanishes as $D(\boldsymbol{\omega}|\boldsymbol{\alpha}) \simeq \eta\, \ln[\mathrm{e}(M-1)/\eta]$. Moreover, the free-energy driving force (20) here takes the form

$$\epsilon = (1-\eta)\, \ln \frac{w_{+\mathrm{c}}}{w_{-\mathrm{c}}} + \eta\, \ln \frac{w_{+\mathrm{i}}}{w_{-\mathrm{i}}} \,, \tag{53}$$

and the entropy production is given by

$$\frac{1}{k_{\rm B}}\frac{d_{\rm i}S}{dt} = v\,[\epsilon + D(\boldsymbol{\omega}|\boldsymbol{\alpha})] \geq 0\,. \tag{54}$$

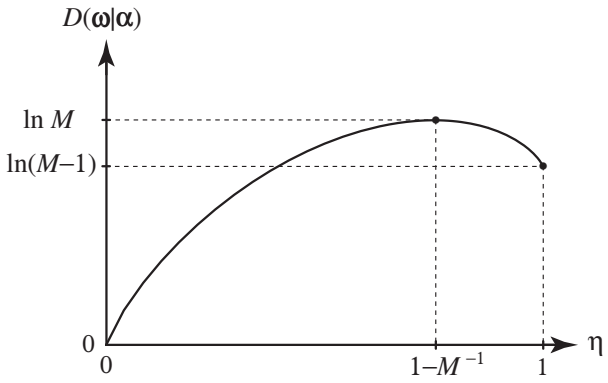

FIG. 1: The conditional Shannon disorder per monomer (52) versus the error probability η for the copolymerization with a template of Bernoulli chains composed of M monomeric species.

If the template is itself a Bernoulli chain, the probabilities of its monomeric units are equal to $\nu(m) = 1/M$ for all $m = 1, 2, ..., M$. This would be the case if the template contained the maximum possible information. In this case, the probabilities of the monomeric units in the copy take the same values $\mu(m) = \sum_{n=1}^{N} \nu(n)\,\mu(m|n) = 1/M$. Therefore, the overall Shannon disorder per monomer in the copy is given by $D(\boldsymbol{\omega}) = \ln M$ and the mutual information per monomer between the template and the copy by

$$I(\boldsymbol{\omega},\boldsymbol{\alpha}) = D(\boldsymbol{\omega}) - D(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \ln M + (1-\eta)\,\ln(1-\eta) + \eta\,\ln\frac{\eta}{M-1}\,. \tag{55}$$

According to Fig. 1, the mutual information reaches its maximal possible value $I(\boldsymbol{\omega},\boldsymbol{\alpha})_{\max} = \ln M$, if the error probability is the smallest ($\eta = 0$). Under such circumstances, the fidelity of information transmission is the highest. This could be approached in the fully irreversible regime where the detachment rates are zero, $w_{-{\rm c}} = w_{-{\rm i}} = 0$. In this regime, the driving force (53) and the entropy production (54) are infinite, while the mean growth speed is equal to $v_\infty = w_{+{\rm c}} + (M-1)w_{+{\rm i}}$ and the error probability to

$$\eta_\infty = \left(1 + \frac{1}{M-1}\frac{w_{+{\rm c}}}{w_{+{\rm i}}}\right)^{-1}\,. \tag{56}$$

If the attachment rate is higher for correct monomers than for incorrect ones, $w_{+{\rm c}} \gg w_{+{\rm i}}$, the error probability is small ($\eta_\infty \ll 1$) and the mutual information can be estimated as

$$I(\boldsymbol{\omega},\boldsymbol{\alpha})_\infty \simeq \ln M - (M-1)\frac{w_{+{\rm i}}}{w_{+{\rm c}}}\ln\frac{{\rm e}\,w_{+{\rm c}}}{w_{+{\rm i}}}\,. \tag{57}$$

DNA polymerases may have attachment rates in the ratio $w_{+{\rm c}}/w_{+{\rm i}} \simeq 3\times 10^6$ so that the error probability can be as low as $\eta_\infty \simeq 10^{-6}$, giving the conditional Shannon disorder per nucleotide $D(\boldsymbol{\omega}|\boldsymbol{\alpha}) \simeq 1.5\times 10^{-6}$ at a speed of about 300 nucleotides per second [18]. Since the disorder is very low, the driving force does not need to be very high for DNA polymerases to be effectively functioning in this regime.

V. CONCLUSIONS

Whether some molecular structure may be supposed to contain information is a matter of its involvement in a physico-chemical process, the outcome of which critically depends on this information. In this respect, information is an emergent property. A sequence of monomeric units is just a particular molecular structure among others. This sequence is said to contain information by its ability to trigger a specific action that would not happen for a different sequence. In the present framework, free copolymerization generates sequences with various degrees of disorder. It is

only with the process of depolymerization, or for copolymerization with a template, that the initial sequence may be supposed to contain some meaningful information.

In particular, depolymerization starts from a copolymer that has been previously synthesized and that may thus carry some information that is erased. Landauer's principle is satisfied during depolymerization because the thermodynamic entropy production (in units of Boltzmann's constant) is always greater than or equal to the maximum possible Shannon information that could be contained in the copolymer sequence [10]. During copolymerization with a template, this latter may contain meaningful information (such as genetic information) that would be transmitted to the copy. Remarkably, the entropy production depends on the mutual information between the template and the copy [6].

At a more fundamental level, we have here shown that the thermodynamics of copolymerization can be deduced from multivariate fluctuation relations, which are valid arbitrarily far from equilibrium and find their origin in microreversibility [11]. Multivariate fluctuation relations are here proved for the multiple chemical currents of monomeric units or doublets incorporated into Bernoulli or Markov chains. In this way, the thermodynamic entropy production can be deduced from the corresponding multivariate fluctuation relation. During the steady growth of the copolymer, the entropy production can be decomposed into a contribution from free energy and another one due to the disorder in the sequence of monomeric units. Analytic expressions are obtained for the free-energy driving force and the disorder per monomer in the copolymerization of Bernoulli or first-order Markov chains, as a consequence of the multivariate fluctuation relation.

We have also seen how copolymerization reduces to simple polymerization in the limit where the monomers are identical and indistinguishable, which was a puzzle reminiscent of Gibbs' paradox of equilibrium statistical mechanics [12].

Furthermore, the thermodynamics of information transmission during copolymerization with a template can be exactly analyzed in the particular case where the attachment and detachment rates only differ between correct and incorrect pairing. In this case, the analysis reduces to the case of Bernoulli chains. The free-energy driving force and the mutual information between the template and the copy can thus be expressed in terms of the error probability. These results open the way to the quantitative study of thermodynamics during DNA replication, upon which we hope to report in future.

Acknowledgments

This research is financially supported by the Université Libre de Bruxelles and the Belgian Science Policy Office under the Interuniversity Attraction Pole project P7/18 "DYGEST".

[1] G. M. Church, Y. Gao, and S. Kosuri, Science **337**, 1628 (2012).
[2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, Nature **494**, 77 (2013).
[3] J.-F. Lutz, M. Ouchi, D. R. Liu, and M. Sawamoto, Science **341**, 628 (2013).
[4] H. Colquhoun and J.-F. Lutz, Nat. Chem. **6**, 455 (2014).
[5] M. F. Crommie, C. P. Lutz, and D. M. Eigler, Science **262**, 218 (1993).
[6] D. Andrieux and P. Gaspard, Proc. Natl. Acad. Sci. USA **105**, 9516 (2008).
[7] C. Jarzynski, Proc. Natl. Acad. Sci. USA **105**, 9451 (2008).
[8] D. Andrieux and P. Gaspard, J. Chem. Phys. **130**, 014901 (2009).
[9] P. Gaspard and D. Andrieux, J. Chem. Phys. **141**, 044908 (2014).
[10] D. Andrieux and P. Gaspard, EPL **103**, 30004 (2013).
[11] P. Gaspard, New J. Phys. **15**, 115014 (2013).
[12] R. K. Pathria, *Statistical Mechanics* (Pergamon, Oxford, 1972).
[13] D. Andrieux, *Nonequilibrium Statistical Thermodynamics at the Nanoscale* (Thèse de doctorat, Université Libre de Bruxelles, 2008).
[14] C. H. Bennett, Biosystems **11**, 85 (1979).
[15] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981).
[16] D. Andrieux and P. Gaspard, Phys. Rev. E **77**, 031137 (2008).
[17] P. Sartori and S. Pigolotti, Phys. Rev. Lett. **110**, 188101 (2013).
[18] K. A. Johnson, Annu. Rev. Biochem. **62**, 685 (1993).