Information management in DNA replication modeled by directional, stochastic chains with memory

J. Ricardo Arias-Gonzalez[a)]
Instituto Madrileño de Estudios Avanzados en Nanociencia, CNB-CSIC-IMDEA Nanociencia Associated Unit "Unidad de Nanobiotecnología," C/Faraday 9, Cantoblanco, 28049 Madrid, Spain

(Received 16 June 2016; accepted 26 October 2016; published online 14 November 2016)

Stochastic chains represent a key variety of phenomena in many branches of science within the context of information theory and thermodynamics. They are typically approached by a sequence of independent events or by a memoryless Markov process. Stochastic chains are of special significance to molecular biology, where genes are conveyed by linear polymers made up of molecular subunits and transferred from DNA to proteins by specialized molecular motors in the presence of errors. Here, we demonstrate that when memory is introduced, the statistics of the chain depends on the mechanism by which objects or symbols are assembled, even in the slow dynamics limit wherein friction can be neglected. To analyze these systems, we introduce a sequence-dependent partition function, investigate its properties, and compare it to the standard normalization defined by the statistical physics of ensembles. We then apply this theory to characterize the enzyme-mediated information transfer involved in DNA replication under the real, non-equilibrium conditions, reproducing measured error rates and explaining the typical 100-fold increase in fidelity that is experimentally found when proofreading and edition take place. Our model further predicts that approximately 1 kT has to be consumed to elevate fidelity in one order of magnitude. We anticipate that our results are necessary to interpret configurational order and information management in many molecular systems within biophysics, materials science, communication, and engineering. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4967335]

I. INTRODUCTION

Many processes in nature are directional. For example, replication, transcription, and translation in biology involve molecular machines that polymerize individual molecular subunits into linear chains in only one direction out of two.[1,2] These systems are paradigmatic of the relation between thermodynamic and information entropy because the chaining of nucleotides or aminoacids according to a template inherently carries genetic information from which species propagate.[3,4] More in depth, an entropic configuration of molecular subunits involves a symbol sequence that conveys information from DNA to proteins. Specific sequences are recognized by specialized proteins that can not only correlate present subunit insertion but also proofread and edit errors, a process that is possible, thanks to the existence of memory. Language writing, copying, reading, and editing are also directional, linear processes in which symbols are arranged and correlated with previous ones to make meaningful sentences.[5] The existence of memory effects and directionality in physics and information theory, not only regarding natural but also for the development of artificial systems, is likewise an essential matter with profound roots.[6–8]

Nanoscale chains are subjected to fluctuations and therefore, a statistical treatment is necessary to understand their physics. In equilibrium, linear systems are normally described by a density operator, as defined by the Boltzmann exponential of the Hamiltonian normalized to the trace.[9,10] Linear chains that involve an assembly dynamics are directional arrangements with history, and therefore these systems implicitly comprise spatial and temporal aspects. Memoryless stochastic processes, namely, those whose evolution only depends on present events, are normally treated by Markovian conditional probabilities and general non-Markovian processes are treated by memory kernel functions and dynamical maps,[12–15] which involve phenomenological ansatz. However, to our knowledge, the full memory in a linear stochastic chain has not been considered, this problem being especially important to relate non-Markovianity to stochastic processes that can be characterized by a partition function.

This problem is the key to understand the information transfer in molecular biology, where there only exists a weak interplay between formal information theory and DNA replication, DNA transcription into RNA, or RNA translation into proteins. With regards to DNA replication, there is a wealth of fidelity data based on biochemical assays[16–19] but there is to date no formal (*ab initio*) deduction of the order of magnitude of error rates or of the energy required by a general DNA polymerase (DNAp) to increase fidelity in the real, non-equilibrium process from a purely thermodynamic viewpoint.

In the following, we show that stochastic chains with memory that are fueled in one direction in the absence of friction cannot be treated by the standard statistical ensemble normalization but by a sequence-dependent partition function. To do this, we consider conditional probabilities of full extent

a) Electronic mail: ricardo.arias@imdea.org

to previous neighbors for the growing chain, i.e., we consider the full memory of the chain, thus including correlations over present and all the past events. In the second part of the paper, we apply this analysis to information theory and characterize the copy and edition of genetic information. The fidelity of these mechano-chemical processes, which we previously studied in the quasistatic limit,[20] was shown to increase at the cost of energy dissipation under non-equilibrium pathways.[21–25] Our model reveals the energy consumed by DNA polymerases to achieve real error rates in replication.

II. THEORY

The system, either classic or quantum (Appendix A), is built by ordering objects on a linear chain.[10,26] We assume that the interaction of one object, i, with the rest, $1,\ldots,i-1,i+1,\ldots,n$, is restricted to its previous neighbors, $1,2,\ldots,i-1$, see Fig. 1. Although the interaction with forward members can exist, we suppose that it does not affect the sequence construction.

A pure state, ν, of the system is specified by a sequence of objects, $x_1,\ldots,x_n$, which stem from a multivariate random variable $\mathbf{X}$, as denoted by

$$\nu = \{x_1, x_2, \ldots, x_i, \ldots, x_{n-1}, x_n\}. \tag{1}$$

Values x_i represent symbols from an alphabet, particles from a set of fixed types, or physical variables (position, momentum, spin, vibrational frequency, occupation number, etc.). Their alphabet or domain will be denoted by χ. Each random variable is in turn a function of several variables of the microscopic, local environment, as, for example, pressure, temperature, ionic conditions, or pH.

The most general Hamiltonian is[11]

$$H(\mathbf{X}) \equiv H(X_1,\ldots,X_n) = \sum_{i=1}^{n} H(X_i; X_{i-1},\ldots,X_1) \tag{2}$$

and the energy of state ν is

$$E_\nu \equiv E(\mathbf{x}) = E(x_1,\ldots,x_n) = \sum_{i=1}^{n} E(x_i; x_{i-1},\ldots,x_1), \tag{3}$$

where the *partial energy* $E(x_i; x_{i-1},\ldots,x_1)$ is the energy of object x_i provided that the previous objects (which constitute a *partial sequence*) are $(x_1,\ldots x_{i-1})$.

A. Directionality and memory impose a sequence-dependent partition function

The probability of a microstate for the general Hamiltonian of Eq. (2) can be expressed as follows:

$$\begin{aligned} p_\nu &\equiv \Pr\{X_1 = x_1,\ldots,X_n = x_n\} = p(\mathbf{x}) = p(x_1,\ldots,x_n) \\ &= p(x_1)p(x_2|x_1)\cdots p(x_n|x_{n-1},\ldots,x_1), \end{aligned} \tag{4}$$

where the last part of the equation is the general expansion of the joint probability as a product of conditional probabilities.[27] It can be factorized since the probability at each step, i, depends only on the previous neighbors.

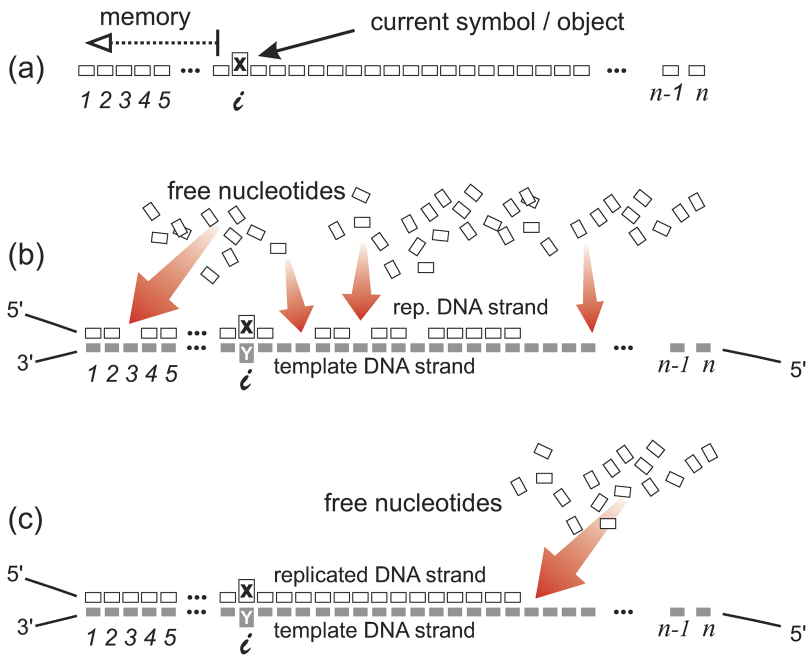

FIG. 1. Construction of a stochastic chain with memory. (a) Sketch of a stochastic sequence of n symbols/objects for which the memory at each position i is constituted by the $i-1$ previous symbols/objects. The chain may be constructed by assembling the symbols/objects removably and without constraints of order or direction, as in (b), or by a directional, stepwise mechanism, as in (c). Schemes (b) and (c) further represent the example of DNA replication in which nucleotides are branched on a template strand according to Watson-Crick complementarity (double-helix structure not represented); (b) represents the spontaneous replication and (c) the enzyme-mediated replication, which is performed by the so-called DNA polymerase (not shown in the scheme), from the 3'-end to the 5'-end of the template strand.

The Gibbs entropy of the system is

$$S(X_1,\ldots,X_n) = -k\langle \ln p\rangle = -k\sum_{\nu=1}^{N} p_\nu \ln p_\nu = -k\sum_{x_1,\ldots,x_n} p(x_1,\ldots,x_n)\ln p(x_1,\ldots,x_n), \tag{5}$$

where k is the Boltzmann constant, "ln" the natural logarithm and N the number of microstates. The mean (internal) energy of the system is

$$\langle E\rangle = \sum_{\nu=1}^{N} p_\nu E_\nu = \sum_{x_1,\ldots,x_n} p(x_1,\ldots,x_n)\,E(x_1,\ldots,x_n). \tag{6}$$

In equilibrium with balanced flows of matter or energy, probabilities can be normalized to a partition function in the form

$$Z(\beta,n) \equiv \sum_{\nu=1}^{N}\exp\left[-\beta E_\nu\right] = \sum_{x_1,\ldots,x_n}\exp\left[-\beta E(x_1,\ldots,x_n)\right] = \sum_{x_1,\ldots,x_n}\exp\left[-\beta\sum_{i=1}^{n}E(x_i;x_{i-1},\ldots,x_1)\right], \tag{7}$$

where $\beta = 1/kT$ and T is the absolute temperature. The probability of a configuration is thus

$$p_\nu = \frac{\exp(-\beta E_\nu)}{Z}. \tag{8}$$

This equation represents the well-known equilibrium probability for a system that has thermalized into a particular configuration. No driving forces other than those that guide the spontaneous process are present. Then, neither a temporal directionality is imposed (i.e., object at position i could have been incorporated or taken its final value, x_i, before or after that at position $i+1$, x_{i+1}) nor a one-by-one incorporation is assumed. What is more, this formalism involves that objects can fluctuate non-sequentially among their different values before a final configuration is reached. This equilibrium, which involves interacting objects, is susceptible to be solved paralleling the 1-dimensional Ising model.

Many processes are, however, directional, for example, DNA replication, in which a polymerase protein copies a template DNA strand from the 5' to the 3' end,[20–22] or English writing, which takes place from left to right. Directionality can be imposed by physicochemical constraints or rules, irrespectively of whether the process is quasistatic (i.e., with such a slow dynamics that it approaches equilibrium conditions at each step) or non-equilibrium. For such directional processes, the probability has to be calculated at each step as

$$p(x_i|x_{i-1},\ldots,x_1) = \frac{e^{-\beta E(x_i;x_{i-1},\ldots,x_1)}}{\sum_{x_i'} e^{-\beta E(x_i';x_{i-1},\ldots,x_1)}}. \tag{9}$$

This mechanism constrains the sequence in which the different configurations are accessible at each step. Therefore, even though each sequence may need an infinite time to be configured, this mechanism does not involve a global thermalization of the system. The states so accessed comprise all the possible ones, as in Eqs. (7) and (8), but the total probability of a configuration is strictly different from that given by Eq. (8) and can be expressed as

$$p_\nu^{(D)} = \frac{\exp(-\beta E_\nu)}{Z_\nu}, \tag{10}$$

where superindex "D" denotes a directional mechanism and Z_ν is given by

$$Z_\nu(\beta,n) \equiv \sum_{x_1',\ldots,x_n'}\exp\left[-\beta\sum_{i=1}^{n}E\left(x_i';x_{i-1},\ldots,x_1\right)\right]. \tag{11}$$

We can name Z_ν *sequence-dependent partition function* since it depends on previous events (not primed variables in Eq. (11)). The fact that $\sum_\nu p_\nu^{(D)} = 1$ follows from the application of the fact that $\sum_{x_i} p(x_i|x_{i-1},\ldots,x_1) = 1$, see Eq. (9), into Eq. (4). In contrast to the directional case, the sums in the denominator of the equilibrium probabilities, Eq. (8), (i.e., in the partition function, Eq. (7)) are nested and therefore they cannot be factorized as independent sums.

The partition function is not therefore unique for directional processes and this represents the main message of this part of the article, the consequences and physical meaning of which will be analyzed below.

If we neglect interactions with previous neighbors, Eqs. (7) and (11) converge to the same value and probabilities, Eqs. (8) and (10), collapse into the same expression. In other words, in the absence of previous-neighbor interactions, directionality and global thermalization into equilibrium are equivalent. This was shown for DNA replication elsewhere[20] and will be rigorously demonstrated later on this article.

It is easy to demonstrate the following properties for the sequence-dependent partition function:

$$\left\langle\frac{1}{Z_\nu}\right\rangle = \frac{1}{Z}, \quad \langle Z_\nu\rangle_D = Z, \tag{12}$$

$$\sum_\nu Z_\nu \geq Z, \quad \exp(-\beta E_\nu) \leq Z_\nu, \tag{13}$$

where subindex "D" in Eq. (12) indicates that the mean has been taken by using the directional probability given by Eq. (10) whereas the absence of this subindex indicates the use of the probability given by Eq. (8). The proof of, for example, Eq. (12) is $1 = \sum_\nu p_\nu = (1/Z)\sum_\nu Z_\nu p_\nu^{(D)} = \langle Z_\nu\rangle_D/Z$. From Eq. (12), it is easy to understand that Z_ν/Z can be greater, equal, or smaller than 1; likewise, that $\left\langle p_\nu^{(D)}/p_\nu\right\rangle = 1$ and that $\left\langle p_\nu/p_\nu^{(D)}\right\rangle_D = 1$.

For the sake of completeness, an extension of this formalism to quantum statistics is presented in Appendix A.

B. Equilibrium and directional statistics approach each other for weak memory effects

To gain physical insight, we study next the case in which the chain construction is sufficiently smoothly dependent on its history. First, we formulate the following theorem, the proof of which is in Appendix B:

Theorem (Independence limit). *Let ν be a stochastic, linear chain with memory (Eq. (1)) sequentially constructed $i : 1 \to n$. Let Z and Z_ν be the normal (equilibrium) and the sequence-dependent partition functions (Eqs. (7) and (11), (A4), respectively), and let $E_i = E(x_i; x_{i-1}, \ldots, x_1)$ and $E'_i = E\left(x_i; x'_{i-1}, \ldots, x'_1\right)$ be the energies of object x_i relative to two different partial sequences (Eq. (3)). If the normalized energy difference $|E_i - E'_i|/kT \to 0$, $\forall i$, then $Z/Z_\nu \to 1$.*

This theorem provides the adequate link between directionally driven processes and equilibrium thermodynamics in the absence of friction. It states that when memory effects are weak enough, the construction of a stochastic chain by a defined mechanism approaches an equilibrium thermalization of independent (non-interacting) objects. *Weak enough* here means that the energetic cost for incorporating any new object in the growing chain is affected by a sufficiently low quantity with respect to the thermal energy level (kT) due to the sequence of previous incorporations.

This theorem also applies when sequence-dependent energies, E_i and E'_i, are much lower than the thermal level, kT, at every step i. Certainly, $E_i, E'_i \ll kT \Rightarrow \left|E_i - E'_i\right| \ll kT$. This is the high-temperature limit in which the thermal energy is so large compared to the energies associated with the neighboring interactions that the system dynamics is dominated by random fluctuations. This is the case of the ideal gas.

As previously stated and further reflected in the demonstration (Appendix B), the limit in the theorem is reached when interactions with previous neighbors are neglected; in such a case, X_i can be approximated by independent random variables.

The stochastic chain described as a thermal equilibrium process, which includes a Boltzmann, scalar partition function, does not involve a mechanism in the chain construction, and therefore it comprises all the possible pathways. Directional chain construction, in contrast, independently of whether it can be approached by an equilibrium stepping dynamics or not, involves a mechanism that drives the building of the chain through a certain microscopic pathway. In both cases, all the possible configurations are accessible, but in the latter, directionality dictates how the chain must be constructed. In the following, we apply this framework to information theory, where these interpretations become clearer and are later on exploited to calculate error rates in the copy and edition of information in cells.

III. DNA REPLICATION

We envision a system that computes information by assembling objects, like atoms or molecules, in the same way as a so-called Turing machine, i.e., by manipulating the objects one at a time in a row according to fixed rules. The alphabet, χ, is conformed by a number $|\chi|$ of atoms/molecules or states of the same atom/molecule. We can think of, for example, a two-state atom used to engrave information on a surface in binary code or trapped atomic ions for quantum computation, correlated by either existing physical interactions or symbolic operational rules.[28] We will focus our analysis on a natural system with intrinsic physical interactions which has evolved to encode information: the genomes in living beings, which entail an alphabet of four symbols (the four nucleobases: adenine, cytosine, guanine and thymine/uracil or A, C, G, and T/U, respectively). Of the three main processes that transfer information in the cell, namely, replication → transcription → translation,[2] we will next work out a minimal model that addresses the main features of DNA replication with and without error correction.

A. Toy model

The physical interactions of the objects in the chain give rise to the memory. To describe these neighboring interactions, we will use simple energy functions with explicit dependence on the random variables and the position i: $E(x_i; x_{i-1}, \ldots, x_1) = f(E_1(x_1), \ldots, E_i(x_i); i)$, where $E_i(x_i)$ represents the energy of an individual object in the chain at position i when its interactions with its previous neighbors are negligible (independent variables). To address the neighboring-interaction strength, we introduce a real parameter, α, which increases when interactions monotonously become weaker and fulfills that $\alpha \to +\infty$ in the limit of no interactions. We propose then partial energy functions with linear dependence on the independent energies,

$$E_i(x_i; x_{i-1}, \ldots, x_1) = \sum_{j=1}^{i} \kappa(i-j; \alpha) E_j(x_j), \tag{14}$$

where κ can be considered as a kernel function that, in contrast to probability kernels in, for example, Markov processes, is used in the energy domain. This kernel can be therefore positive or negative, which will represent positive or negative feedbacks, respectively. To illustrate representative physical systems, we will further provide κ with the next properties,

$$\lim_{j \to i} \kappa(i-j; \alpha) = 1, \tag{15}$$

$$\lim_{\alpha \to +\infty} \kappa(i-j; \alpha) = \delta_{ij}, \tag{16}$$

$$\lim_{i-j \to +\infty} \kappa(i-j; \alpha) = 0. \tag{17}$$

The first condition sets the energy of the current object in the absence of previous neighbor interactions, the second one allows to recover the case of independent variables, and the third one establishes that the influence of previous neighbors decays with the distance. Two examples of kernel functions are shown in Appendix C, one with hyperbolic and the other with exponential attenuation in the influence of the previous neighbors over the present.

The total energy of a sequence is (see Eq. (3))

$$E_\nu = \sum_{i=1}^{n} \sum_{j=1}^{i} \kappa(i-j; \alpha) E_j(x_j). \tag{18}$$

This statistical simplicity is sufficient to strengthen the differences between equilibrium and directional statistics and, furthermore, to deduce experimental error rates and non-equilibrium energies in DNA replication, as we will consider in Subsections III B and III C.

The standard and sequence-dependent partition functions are, respectively,

$$Z = \prod_{i=1}^{n} Z_i' = \prod_{i=1}^{n} \sum_{x_i \in \mathcal{X}} \exp\left(-\beta E_i(x_i) \sum_{j=i}^{n} \kappa(n-j;\alpha)\right), \tag{19}$$

$$Z_\nu = \prod_{i=1}^{n} Z_i = \prod_{i=1}^{n} \exp\left(-\beta \sum_{j=1}^{i-1} \kappa(i-j;\alpha) E_j(x_j)\right) \sum_{x_i' \in \mathcal{X}} e^{-\beta E_i(x_i')}. \tag{20}$$

Another practical consequence of the use of memory functions linear in the energy as in Eq. (14) is that directional probabilities reduce to the case of independent variables,

$$p_\nu^{(D)} = \prod_{i=1}^{n} \frac{e^{-\beta E_i(x_i)}}{\sum_{x_i' \in \mathcal{X}} e^{-\beta E_i(x_i')}} = \prod_{i=1}^{n} p_i(x_i) = p_\nu^{(id)}, \tag{21}$$

where $p_\nu^{(id)} \equiv e^{-\beta E_\nu^{(id)}}/Z^{(id)}$, $E_\nu^{(id)} \equiv \sum_{i=1}^{n} E_i(x_i)$, and $Z^{(id)} \equiv \prod_{i=1}^{n} \sum_{x_i \in \mathcal{X}} e^{-\beta E_i(x_i)}$ are the probability, the energy, and the partition function for a sequence of independently distributed (*id*) symbols. In these conditions, the directional entropy, $S^{(D)}$, as defined by the Shannon expression $S^{(D)}(X_1, \ldots, X_n) = -k\left\langle \ln p^{(D)} \right\rangle_D = -k \sum_{\nu=1}^{N} p_\nu^{(D)} \ln p_\nu^{(D)}$, and that for independent variables, $S^{(id)}$, is the same, namely $S^{(D)} = S^{(id)}$, but not the standard equilibrium entropy, S. Other thermodynamic potentials of the chain, like the internal or the free energies, are different for the standard, sequence-dependent, and *id* cases because their dependence on the variables x_i does not only come through compositions with the probability functions p_i (see Eq. (21)).

B. Replication of genetic information

The copy of information in this cellular process is mainly determined by the complementarity between the replicated and template strands, which is given by the so-called Watson-Crick base-pairing, namely, A–T, T–A, G–C, and C–G (nucleobase U corresponds to RNA, and therefore does not appear in replication). A base-pair is held stable by H-bonds and constitutes a subunit of a DNA double-helix, made up of two parallel strands from the point of view of the information but chemically antiparallel (i.e., with opposite 3' and 5' chemical ends): a template and a replicate (Fig. 1). Individual nucleobases plus a carbon sugar and a phosphate group (monophosphate nucleotides) are successively attached by a phosphodiester bond on each strand. The replicated strand is built by the enzyme DNA polymerase (DNAp), which catalyzes the addition of each new nucleotide in the replicated strand in a Turing-like fashion. Memory is determined by the structural fitting of the DNAp to the resulting double-stranded DNA helix.[16,20] Since a DNAp typically covers one helical turn, comprising 10-11 base pairs,[1,29,30] it is a good approximation to consider that the memory extends to these range of previous base pairs. Taking into account the behavior of the memory kernels proposed in Appendix C, such situation can be approached in our toy model by a coupling strength $1 < \alpha \leq 2$.

Errors appear when non-Watson and Crick base-pairs are assembled: A G–T base-pair, for example, is a common mismatch that does not abruptly change the DNA helical structure. In addition, rare tautomeric forms of the four bases occur spontaneously that give rise to wobble structures at the level of the mispair without the distortion of the helix geometry.[31]

Replication takes place through a directional mechanism like that represented in Fig. 1(c), and therefore, its statistics are those of a directional, stochastic chain with memory. There exists in the cell an additional process generally called *proofreading* in which errors are recognized by a DNAp and removed by exonucleolysis, hence allowing the edition of the information.[23,25,32] This process typically increases fidelity by a 100-fold (and more generally from a few-fold to 1000-fold), depending on the DNAp type.[16,33] Proofreading comprises defined action mechanisms but its effect can be approximately understood as in Fig. 1(b), where inserted nucleotides can be replaced by new ones, that is, the state of the chain at position i can change in an Ising-like fashion, and therefore, probabilities can be calculated by using the standard partition function taking into account the neighboring interactions.[20]

The incorporation of a nucleotide in the replicated strand involves a free energy release in the range of -1.5 to -3 kT for Watson-Crick unions (correct nucleotide) and a typical energy absorption in the range of +1 to +2 kT for most non-Watson-Crick unions (wrong nucleotide or error).[34] This variability for both correct incorporations and errors, beyond sequence dependence, is due to the catalytic nature of the nucleotide addition reaction, which makes free energies pH, temperature, and ionic-strength dependent.[35] We will approximate this energy spectrum by the following energy functions:

$$E_i(x_i) = \begin{cases} -E, & x_i \text{ correct} \\ +E, & x_i \text{ error}, \end{cases} \tag{22}$$

where E is a real, positive number. Additionally, we propose that at each position i there is only one so-called correct x_i, being the rest errors.

This energy spectrum is a simplification of four main facts:[34,36] (i) G–C and C–G unions, on one hand, and A–T and T–A ones, on the other hand, are energetically similar; (ii) G–C or C–G unions are energetically different from A–T or T–A ones; (iii) there are some error types, which are frequent, that involve near zero energies, both positive and negative; (iv) base-pairing energies are also dependent on the previously formed base-pair.

The last fact, which involves a Markov chain, was used to study DNA replication in equilibrium:[20] the DNAp was assumed, first, to passively let each new nucleotide incorporate on the template strand according to its free energy difference and, second, to recognize the secondary structure of the resultant double-stranded DNA. This latter assumption is based on the fact that the DNA helical conformation appears after every newly formed base-pair stacks on the previous one with a defined geometry.[37]

The energies given by Eq. (22) thus represent the energetic level at which correct symbols and errors are placed with respect to the thermal level (kT) without addressing the fine structure of this spectrum, which would be necessary, for instance, to characterize error rates for a specific error type and/or with respect to a particular template. In addition, the toy model does not take into account the incidence of kinetically controlled tautomerization of the base-pairs, which are non-dissociative and may give rise to spontaneous point errors.[38–42] We will show, however, that the energies given by Eq. (22) constitute a very good approximation to describe the

important features of fidelity in DNA replication with and without error correction.

The standard, sequence-dependent, and *id* partition functions reduce to

$$Z = \prod_{i=1}^{n} Z'_i = \prod_{i=1}^{n} \left(e^{\beta E a_i(\alpha)} + (|\chi| - 1) e^{-\beta E a_i(\alpha)} \right), \quad (23)$$

$$Z_\nu = \prod_{i=1}^{n} Z_i = \prod_{i=1}^{n} \left(e^{\beta E} + (|\chi| - 1) e^{-\beta E} \right) \times \exp\left(-\beta \sum_{j=1}^{i-1} \kappa(i-j;\alpha) E_j(x_j) \right), \quad (24)$$

$$Z^{(id)} = \left(e^{\beta E} + (|\chi| - 1) e^{-\beta E} \right)^n, \quad (25)$$

where

$$a_i(\alpha) \equiv \sum_{j=1}^{i} \kappa(i-j;\alpha). \quad (26)$$

It is important to note that while index i runs over the ordered sequence positions, $1, \ldots, n$, in Eq. (24), this index is strictly not related to sequence positions in Eq. (23).

In the limit of no interactions, the standard and sequence-dependent partition functions converge to the case of independent variables,

$$\lim_{\alpha \to +\infty} Z = \lim_{\alpha \to +\infty} Z_\nu = Z^{(id)}. \quad (27)$$

Likewise, in the limit of low energies with respect to the thermal level, the three partition functions are equivalent,

$$\lim_{\beta E \to 0} Z = \lim_{\beta E \to 0} Z_\nu = \lim_{\beta E \to 0} Z^{(id)} = |\chi|^n. \quad (28)$$

The equality of the standard and sequence-dependent partition functions in these two limits is expected according to the *independence limit* theorem.

We can define the perfect and imperfect sequences, ν_P and ν_I, respectively, as those with the lowest and highest energies, E_{ν_P} and E_{ν_I}, respectively. These sequences correspond to monotonous series of either correct or incorrect symbols according to Eq. (22), and their energies are $E_{\nu_P} = -EA$ and $E_{\nu_I} = +EA$, being

$$A(\alpha) \equiv \sum_{i=1}^{n} a_i(\alpha). \quad (29)$$

The energy of a sequence can always be expressed in terms of the energy of the perfect sequence. Namely, it can be proved that for a sequence ν_m with m errors at positions $i = k_h$, $h = 1, \ldots, m$, the energies can be expressed as

$$E_{\nu_m} = E_{\nu_P} + 2E \sum_{h=1}^{m} \left(1 + \sum_{i=k_h+1}^{n} \kappa(i-k_h;\alpha) \right). \quad (30)$$

It is clear that E_{ν_m} takes different values depending on where these errors are located and that there exist degenerate levels.

For the kernel functions defined in Appendix C, with both positive and negative feedbacks, and for coupling strengths $\alpha \geq 1.75$, since $\sum_{i=k_h+1}^{n} \kappa(i-k_h;\alpha) = \sum_{j=1}^{n-k_h} \kappa(n-k_h-j+1;\alpha)$, it is fulfilled that $|\sum_{j=1}^{n-k_h} \kappa(n-k_h-j+1;\alpha)| \leq |\sum_{j=1}^{n-1} \kappa(n-j;\alpha)| = |a_n(\alpha)-1| < 1$, then $E_{\nu_P} < E_{\nu_m}$ in Eq. (30). Besides, in these conditions, the energy grows at each position i in big steps of size E interspersed by small steps of size $E \sum_{j=1}^{i-1} \kappa(i-j,\alpha)$, either positive or negative depending on the feedback sign, from E_{ν_P} to E_{ν_I}. Namely, for the all-error case, $m \to n$ and $k_h \to h$ and $\sum_{h=1}^{m}(1 + \sum_{j=1}^{n-k_h} \kappa(n-k_h-j+1;\alpha)) \to \sum_{i=1}^{n}(1 + \sum_{j=1}^{n-i} \kappa(n-i-j+1;\alpha)) = \sum_{i=1}^{n}(1 + \sum_{j=1}^{i-1} \kappa(i-j;\alpha)) = A(\alpha)$, then $E_{\nu_n} = E_{\nu_I}$.

The entropy for the standard and directional processes is

$$S(\beta E, |\chi|; \alpha) = k \ln Z(\beta E, |\chi|; \alpha) - k\beta E \sum_{i=1}^{n} a_i(\alpha) \Lambda_i(\beta E, |\chi|; \alpha), \quad (31)$$

$$S^{(D)}(\beta E, |\chi|) = k \ln Z^{(id)}(\beta E, |\chi|) - kn\beta E \Lambda^{(id)}(\beta E, |\chi|), \quad (32)$$

where Z and $Z^{(id)}$ are given by Eqs. (23) and (25), respectively, and Λ_i and $\Lambda^{(id)}$ are

$$\Lambda_i(\beta E, |\chi|; \alpha) = \frac{e^{\beta E a_i(\alpha)} - (|\chi| - 1) e^{-\beta E a_i(\alpha)}}{e^{\beta E a_i(\alpha)} + (|\chi| - 1) e^{-\beta E a_i(\alpha)}}, \quad (33)$$

$$\Lambda^{(id)}(\beta E, |\chi|) = \frac{e^{\beta E} - (|\chi| - 1) e^{-\beta E}}{e^{\beta E} + (|\chi| - 1) e^{-\beta E}}. \quad (34)$$

Using the limits shown in Eqs. (27) and (28), it is easy to see that

$$\lim_{\alpha \to +\infty} S = \lim_{\alpha \to +\infty} S^{(D)} = S^{(id)} \quad (35)$$

and that

$$\lim_{\beta E \to 0} S = \lim_{\beta E \to 0} S^{(D)} = \lim_{\beta E \to 0} S^{(id)} = kn \ln |\chi|, \quad (36)$$

which represents the maximum uncertainty. Besides, it is fulfilled that

$$\lim_{\beta E \to +\infty} S = \lim_{\beta E \to +\infty} S^{(D)} = \lim_{\beta E \to +\infty} S^{(id)} = 0, \quad (37)$$

which is expected for $\alpha > 1$ since the high contrast between correct insertions and errors lead asymptotically to a total certainty.

Figure 2 shows the behavior of the entropy as a function of the energy spent in building the replicated strand. For comparison, it is shown the case of a binary alphabet, which could be observed when symbols were limited to either *A* and *T* or *G* and *C* nucleotides. The limit for low energy with respect to the thermal level, Eq. (36), as well as the asymptotic behavior for high energies, Eq. (37), is fulfilled. Since the exponential energy kernel decays faster with the number of nearest neighbors than the hyperbolic one, the energy range for which S strongly differs from $S^{(D)}$ is shorter for the former. This behavior is also observed for a fixed energy kernel at increasing α. Likewise, when the coupling between the past and the present exhibits a positive feedback, there is a shorter energy range for which S strongly differs from $S^{(D)}$ than when the feedback is negative.

As expected positive and negative feedbacks have opposite effects: The former, Figs. 2(a) and 2(c), makes revision an effective process since it decreases the entropy with respect to the directional mechanism. It is observed for the latter, Figs. 2(b) and 2(d), that the entropy of the directional mechanism is

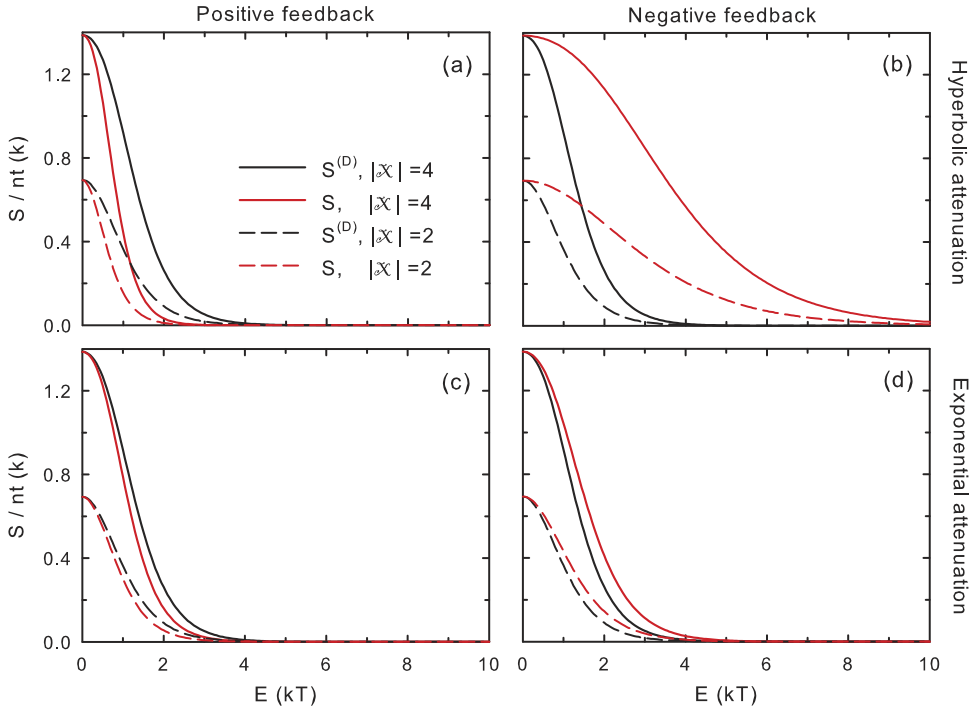

FIG. 2. Entropy per nucleotide (nt) as a function of the energy contrast between correct and wrong insertions. The curves in panels (a) and (b) have been calculated by imposing a hyperbolic attenuation for the influence of the past over the present and those in panels (c) and (d) by using an exponential attenuation (see Appendix C), both of which with coupling strength $\alpha = 2$. Left panels, (a) and (c), positive feedback; right panels, (b) and (d), negative feedback. Results for the standard, S, and directional, $S^{(D)}$, entropies are shown for binary and quaternary alphabets.

the lowest, thus making proofreading worsen the initial replicate. In the real process, proofreading is able to decrease the entropy of the initial replicate, which is reflected in a lower number of errors in the final replicate. Therefore, the negative feedback case does not overall comply with the natural DNA replication process.

C. Error rates in real (non-equilibrium) DNA replication and edition

DNA replication and exonucleolytic proofreading are non-equilibrium: the energy spent in incorporating nucleotides is higher than those arising from the free energy differences discussed above.[34,36] It is known that fidelity increases at the cost of such extra energy consumption.[21,22] The mechanism of the DNAp with proofreading is thus active: it eventually increases the number of correct nucleotides (decreases the number of errors) in the replicated strand with respect to the passive processes represented in Fig. 1 and studied previously.[20] The DNAp improves the discrimination between correct and wrong incorporations based on geometric selection of base-pairs and induced fit of the enzyme to the resulting structure.[16] In fact, we previously found numerically that the qualitative error types in polymerization are mainly determined by the thermodynamic affinity of nucleobases.[20]

We then propose that, from a thermodynamic point of view (i.e., independently of further mechano-chemical, kinetic, and electrostatic assumptions),[31,33] the non-equilibrium action of the DNAp in both the absence and presence of exonucleolytic activity is to increase the energy contrast between correct nucleotides and errors with respect to the free energy difference involved in their passive incorporation. We assume that this action extends to the length of the double-stranded DNA that the DNAp structurally fits, which, as previously mentioned, corresponds to one helix turn, and that it can be modeled by an energy kernel function.

The available energy that (both DNA and RNA) polymerases can use at each step i is obtained from an enzymatic cycle, which basically comprises a triphosphate nucleotide (NTP)-binding step followed by the NTP hydrolysis reaction with phosphodiester bond formation and release of pyrophosphate (PPi), with a turnover energy of $\sim 12 - 13\ kT$, depending on the process (either replication or transcription) and environmental conditions.[35,43,44]

The probability of error, p_{error} or $p^{(D)}_{error}$, depending on whether the process is in equilibrium or directional, respectively, can be obtained through the general Asymptotic Equipartition Property (AEP).[26] For large n and low probability of error per incorporated symbol, it follows that (see Appendix D)

$$p_{error} \simeq \frac{s}{k} \quad \text{and} \quad p^{(D)}_{error} \simeq \frac{s^{(D)}}{k}, \tag{38}$$

where s and $s^{(D)}$ are the entropy rates (entropies per incorporated nucleotide) for the equilibrium and the directional processes, respectively. As explained in Appendix D, the use of the general AEP allows the calculation of error rates in typical replicated strands, i.e., those which are mostly the representative of a DNA replication process.

Figure 3 shows the logarithm to base ten of the entropy per nucleotide, which can be approximately considered the order of magnitude of the probability of error for

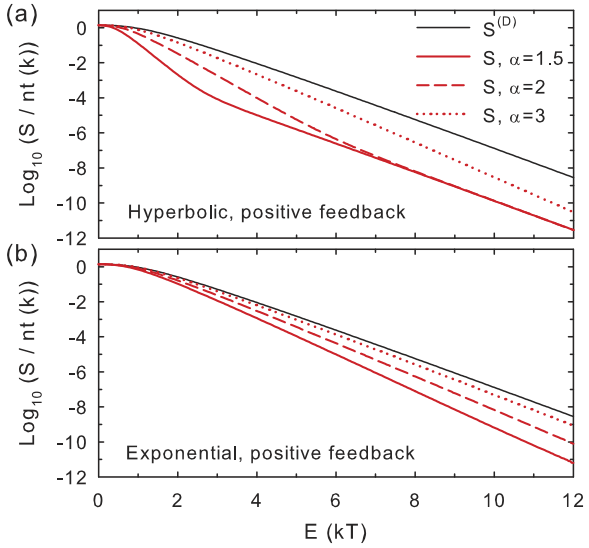

FIG. 3. Fidelity in DNA replication with and without proofreading as a function of the average energy spent in copying each nucleotide. The graph shows the logarithm to base ten of the entropy per nucleotide (nt), which is approximately the order of magnitude of the error rate for high enough E (see the text for details). In panel (a), a hyperbolic, positive feedback decay coupling between the previous neighbors and the present has been used and in (b), an exponential, positive feedback decay coupling. As observed, the toy model predicts a 10 to 1000-fold difference between error rates in replication with (red curves) and without (black curve) proofreading, depending on the memory coupling strength, α.

$\beta E > 2$ (s/k, $s^{(D)}/k \lesssim 10^{-1}$), see Eq. (38) and Appendix D. For energies comparable to the real free energies for nucleotide incorporation,[34,36] namely, $E \sim 2\ kT$, it is observed that the error rates for replication, Eq. (32) (directional statistics), and proofreading, Eq. (31) (standard statistics), provide error rates $p^{(D)}_{error} \sim 10^{-1}$ and $p_{error} \sim 10^{-2}$, in agreement with former simulations[20] with the complete set of experimental energies,[34] hence validating the toy model with the energy functions of Eq. (22).

This model makes it possible to explain another important feature that takes place in the real process (i.e., out of equilibrium): the typical 100-fold, Fig. 3(a), and 10-fold, Fig. 3(b), differences are observed between replication with and without proofreading, which take place for energies $\beta E > 4$. Finally, it reveals the typical order of magnitude of replication with and without proofreading as a function of the energy used for each nucleotide incorporation ($1 < \alpha < 2$). For example, a DNAp needs to spend approximately 3 extra kT (over the $\sim 2\ kT$ free energy difference in equilibrium) to achieve a fidelity of $p_{error} \sim 10^{-5}$ and $p^{(D)}_{error} \sim 10^{-3}$, with and without proofreading, respectively. This is the case of the Φ29-DNAp, able to act as both a polymerase and an exonuclease.[32,45] In order to achieve fidelities of more accurate polymerases,[17,18] namely, $p_{error} \sim 10^{-7}$ and $p^{(D)}_{error} \sim 10^{-5}$ with and without proofreading, respectively, our model predicts an extra energy of $\sim 5\ kT$.

The energy required to increase one order of magnitude the fidelity in both replication and proofreading is $\sim 1\ kT$ per nucleotide according to the slopes shown in Fig. 3. The model further predicts that if all the energy coming from the NTP hydrolysis, after phosphodiester bond formation and PPi release, $\sim 12-13\ kT$, was used to increase fidelity, the probabilities of error would be $p_{error} \sim 10^{-12} - 10^{-11}$ and $p^{(D)}_{error} \sim 10^{-9}$ with and without proofreading, respectively. This limit is not reachable since part of this energy is used by the DNAp to mechanically translocate along the DNA template. In practise, the mechanical work to power the translocation of the DNAp is $W = F \times d \approx 3\ kT$/step (where, according to single-molecule experiments, $F \approx 34\ pN$ is the typical force exerted by the DNAp[32,46–48] and $d \approx 0.34\ nm$ is the separation between base pairs in B-DNA[1]), leaving a maximum of $\sim 10\ kT$ for *informational work*. Then, the maximum fidelities (i.e., minimum error rates) are $p_{error} \sim 10^{-10} - 10^{-9}$ and $p^{(D)}_{error} \sim 10^{-7}$, with and without proofreading, respectively.

According to these results and considering those found previously in equilibrium,[20] we conclude that error rates for replication with proofreading, p_{error}, range from $\sim 10^{-10} - 10^{-9}$ to $\sim 10^{-3} - 10^{-2}$ and without proofreading, $p^{(D)}_{error}$, from $\sim 10^{-7}$ to $\sim 10^{-1}$. These universal boundaries are confirmed by the experimental work to date.[16–18]

Our analysis could be extended to DNA *strand-directed mismatch repair*, in which replication errors that escape the proofreading process are treated by other proteins that work through auxiliary pathways.[49] This process can improve fidelity typically by an extra 100-fold. All in all, DNA replication plus proofreading and subsequent mismatch repair mechanisms yield a probability of error $\sim 10^{-10} - 10^{-9}$, which would involve a total of $\sim 10\ kT$ spent by different proteins on each nucleotide incorporation.

IV. CONCLUSIONS

We have shown that the physics of a stochastic chain construction depends on the existence of directionality in the presence of memory. We have demonstrated that configurational probabilities for thermal equilibrium collapse are strictly different from those for the directional construction. Then, we have established their thermodynamic relations, thus interpreting their physical meaning in terms of the internal energy and entropy. Directionality correlates with a time arrow, in contrast to thermal equilibrium, which is timeless. Besides, the fact that the partition function depends on the history is a consequence of the fact that a thermodynamic limit cannot be defined for driven stochastic processes.

Stochastic chains with memory can be found in diverse, nanoscale scenarios. In this regard, we have worked out a minimal model in information theory and have applied it to quantitatively describe fidelity in DNA replication. We have obtained the order of magnitude of the error rates in the real, non-equilibrium process. In addition, we have explained how proofreading can increase fidelity by a 10-fold to a 1000-fold, in agreement with former biochemical experiments. These results comply with an interpretation in which the use of the standard partition function in the statistical treatment of a stochastic chain of symbols with memory involves that the information is proofread and edited for error correction, whereas the sequence-dependent partition function herein introduced applies to the directional growth of the chain without revision. Finally, we have revealed that the energy required to increase fidelity, i.e., to decrease the error rate, in both replication and proofreading is approximately $1\ kT$ per order of magnitude.

We believe that our results are not only important to molecular biology but also to other more general small systems, as well as to computer science and communication theory, because the generation, edition, and transfer of information rely on the use of stochastic chains of interacting objects, either physical or symbolic. For example, the composition of human language or the generation of genomes throughout evolution could be treated under the general scheme described here.

ACKNOWLEDGMENTS

It is a pleasure to thank J. M. R. Parrondo and D. G. Aleja for fruitful discussion. This work was supported the Spanish Ministry of Economy and Competitiveness (Grant Nos. MAT2013-49455-EXP and MAT2015-71806-R).

APPENDIX A: QUANTUM STATISTICS FORMULATION

A pure state (see Eq. (1)) in quantum notation is expressed as $|\nu\rangle$, which corresponds to an eigenvector in a Hilbert space, $|\nu\rangle \in \mathscr{H}$, and is a stationary solution of $H|\nu\rangle = E_\nu|\nu\rangle$, with $\langle\nu|\nu'\rangle = \delta_{\nu\nu'}$. Each vector $|\nu\rangle$ is constructed on the basis of the distinguishability and symmetry of the quantum objects that comprise the chain. Energies, E_ν, are formally obtained as eigenvalues of H, $E_\nu = \langle\nu|H(\mathbf{X})|\nu\rangle$.

We introduce the two-sequence Hamiltonian, $\hat{H}$, as

$$\hat{H}(\mathbf{X}',\mathbf{X}) \equiv \sum_{i=1}^{n} H(X'_i; X_{i-1},\ldots,X_1), \tag{A1}$$

whose mathematical expression fulfills $\hat{H}(\mathbf{X},\mathbf{X}) \sim H(\mathbf{X})$. This Hamiltonian can be introduced in quantum statistics as a sesquilinear form, $\hat{H} : \mathscr{H}\times\mathscr{H} \to \mathbb{C}$, namely, $(\langle\mu'| + \langle\nu'|)\hat{H}(|\nu\rangle + |\mu\rangle) = \langle\mu'|\hat{H}|\nu\rangle + \langle\mu'|\hat{H}|\mu\rangle + \langle\nu'|\hat{H}|\nu\rangle + \langle\nu'|\hat{H}|\mu\rangle$, and $\langle(a)\nu'|\hat{H}|(b)\nu\rangle = \overline{a}b\langle\nu'|\hat{H}|\nu\rangle$, for all $|\mu\rangle, |\mu'\rangle, |\nu\rangle, |\nu'\rangle \in \mathscr{H}$ and all a and $b \in \mathbb{C}$, being $\overline{a}$ the complex conjugate of a, in contrast to H, which is a linear operator $H : \mathscr{H} \to \mathscr{H}$. The complex bilinear form $\hat{H}$ does not represent an observable; the energy operator was actually given in Eq. (2). The matrix elements of $\hat{H}$ are

$$\langle\nu'|\hat{H}(\mathbf{X}',\mathbf{X})|\nu\rangle = E_{\nu'\nu}, \tag{A2}$$

where $E_{\nu'\nu}$ is the two-sequence energy,

$$E_{\nu'\nu} \equiv \sum_{i=1}^{n} E(x'_i; x_{i-1},\ldots,x_1), \tag{A3}$$

and $\nu' = \{x'_1, x'_2, \ldots, x'_i, \ldots, x'_{n-1}, x'_n\}$. Then, the sequence-dependent partition function, Eq. (11), can be spelled as

$$Z_\nu \equiv \sum_{\nu'=1}^{N} \exp(-\beta E_{\nu'\nu}). \tag{A4}$$

To fully understand the physical meaning and relationship between the two partition functions, we represent their terms in matrix form

$$\mathbf{Z} = (Z_{\nu'\nu}) = \left(e^{-\beta E_{\nu'\nu}}\right) = \begin{pmatrix} e^{-\beta E_1} & e^{-\beta E_{12}} & \ldots & e^{-\beta E_{1N}} \\ e^{-\beta E_{21}} & e^{-\beta E_2} & \ldots & e^{-\beta E_{2N}} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-\beta E_{N1}} & e^{-\beta E_{N2}} & \ldots & e^{-\beta E_N} \end{pmatrix}, \tag{A5}$$

where we have used that $E_{\nu\nu} = E_\nu$. This matrix is easily defined in the Quantum Statistics formalism as

$$Z_{\nu'\nu} = \langle\nu'|e^{-\beta\hat{H}(\mathbf{X}',\mathbf{X})}|\nu\rangle, \tag{A6}$$

with which the standard partition function conserves its quantum definition,

$$\begin{aligned} Z = \mathrm{Tr}\mathbf{Z} &= \mathrm{Tr}\left\{\exp\left[-\beta\hat{H}(\mathbf{X}',\mathbf{X})\right]\right\} \\ &= \mathrm{Tr}\left\{\exp\left[-\beta H(\mathbf{X})\right]\right\}. \end{aligned} \tag{A7}$$

The formalism introduced in Eqs. (9) and (A7) allows the sequence-dependent partition function, Eq. (11), to be integrated into the statistical physics framework. In short, a *directional* mechanism in a quasistatic linear chain construction is introduced by defining probabilities (Eq. (10)) whose normalization (Eq. (11)) depends on the history of the system. Such normalization is obtained by summing up over the column elements on the matrix representation of the exponential of the two-sequence Hamiltonian (Eq. (A1)). It is noticeable that the density operator, ρ, for quantum directional, stochastic chains with memory does not conserve the trace, $\mathrm{Tr}\,\rho = 1$, but the sum over the elements of each individual column in its matrix representation, which is a consequence of the saturation of the conditional probabilities, $\sum_{x_i} p(x_i|x_{i-1},\ldots,x_1) = 1$, that follows after Eq. (9).

APPENDIX B: PROOF OF THE INDEPENDENCE LIMIT THEOREM

When memory effects are sufficiently mild, we can express the partial energies as

$$\begin{aligned} E(x'_i; x'_{i-1},\ldots,x'_1) &= E(x'_i; x_{i-1},\ldots,x_1) \\ &\quad + \epsilon(x_{i-1},\ldots,x_1), \end{aligned} \tag{B1}$$

where $\epsilon_i \equiv \epsilon(x_{i-1},\ldots,x_1)$ is the energy difference between energies associated with different, partial sequences $(x'_{i-1},\ldots,x'_1)$ and $(x_{i-1},\ldots,x_1)$. For all i, it fulfills

$$\begin{aligned} &\epsilon(x_{i-1},\ldots,x_1) = 0, \\ &\text{if } (x_{i-1},\ldots,x_1) = (x'_{i-1},\ldots,x'_1), \text{ and} \end{aligned} \tag{B2}$$

$$|\epsilon(x_{i-1},\ldots,x_1)| \ll kT, \quad \text{otherwise.} \tag{B3}$$

Then,

$$\begin{aligned} E_{\nu'} &= \sum_{i=1}^{n} E(x'_i; x'_{i-1},\ldots,x'_1) \\ &= \sum_{i=1}^{n} E(x'_i; x_{i-1},\ldots,x_1) + \sum_{i=1}^{n} \epsilon(x_{i-1},\ldots,x_1) \\ &= E_{\nu'\nu} + \epsilon_\nu, \end{aligned} \tag{B4}$$

where we have used the definition of the two-sequence energy, Eq. (A3), and that $\epsilon_\nu \equiv \sum_{i=1}^{n}\epsilon_i$ is a sequence-dependent small energy. Consequently,

$$\begin{aligned} Z - Z_\nu &= \sum_{\nu'=1}^{N} \exp(-\beta E_{\nu'}) - \sum_{\nu'=1}^{N} \exp(-\beta E_{\nu'\nu}) \\ &= \sum_{\nu'=1}^{N} \exp(-\beta E_{\nu'\nu})\left[\exp(-\beta\epsilon_\nu) - 1\right]. \end{aligned} \tag{B5}$$

We expand $\exp(-\beta\epsilon_\nu)$ in the limit $\beta\epsilon_i \to 0, \forall i$,

$$\exp(-\beta\epsilon_\nu) = \exp\left(-\beta\sum_{i=1}^{n}\epsilon_i\right) = \prod_{i=1}^{n}\exp(-\beta\epsilon_i) \approx 1 - \beta\sum_{i=1}^{n}\epsilon_i = 1 - \beta\epsilon_\nu. \quad \text{(B6)}$$

From Eq. (B5), it follows that

$$Z - Z_\nu \approx Z_\nu(-\beta\epsilon_\nu) = \varepsilon_\nu Z_\nu, \quad \text{(B7)}$$

where $-\beta\epsilon_\nu \equiv \varepsilon_\nu$ is a small number. Then, $Z/Z_\nu \approx 1 + \varepsilon_\nu$ with $\varepsilon_\nu \sim 0$.

APPENDIX C: ENERGY KERNELS

The following functions

$$\kappa(i-j;\alpha) = \begin{cases} 1, & j = i \\ +(-)1/(i-j+1)^\alpha, & j < i, \end{cases} \quad \text{(C1)}$$

and

$$\kappa(i-j;\alpha) = \begin{cases} 1, & j = i \\ +(-)e^{-\alpha(i-j)}, & j < i, \end{cases} \quad \text{(C2)}$$

with $\alpha > 1$ and $i \geq j$, satisfy conditions (15)-(17) and therefore they can be used for the positive (negative) feedback coupling regimes. In addition, they are bounded for $\alpha > 1$,

$$\sum_{j=1}^{i-1}\frac{1}{(i-j+1)^\alpha} \leq \sum_{j=1}^{n-1}\frac{1}{(n-j+1)^\alpha} = \sum_{j=1}^{n}\frac{1}{j^\alpha} - 1 \leq \sum_{j=1}^{\infty}\frac{1}{j^\alpha} - 1 = \zeta(\alpha) - 1 < +\infty, \quad \text{(C3)}$$

where we have used that $n \geq i$ and that the harmonic series of order α (or Riemann zeta function of α), $\sum_{j=1}^{\infty} 1/j^\alpha$, with sum $\zeta(\alpha)$, is convergent for $\alpha > 1$. Since $1/(i-j+1)^\alpha \geq e^{-\alpha(i-j)}$ ($\alpha > 0, i \geq j, i, j = 1, \ldots, n$) is fulfilled, the exponential kernel is also bounded for at least the same range of coupling strengths ($\alpha > 1$). The convergence of the kernel functions makes the model independent of the size of the chain for n sufficiently large. For $\alpha \geq 1.75$, the last part of Eq. (C3) is < 1, implying that for typical coupling strengths, the influence of the previous symbols over the energy of the present one is always lower than the energy of the present symbol individually.

For $\alpha \geq 1.5$, the influence of the 12th neighbor from the current position ($i - j = 12$) has attenuated more than a 97% and the accumulated influence of all the neighbors beyond the 12th one has attenuated more than an 80% in both kernel types.

APPENDIX D: ERROR RATES

The Shannon-McMillan-Breiman theorem, or general AEP, states[26]

$$-\frac{1}{n}k\ln p(X_1, \ldots X_n) \to s(\chi), \quad \text{(D1)}$$

with probability 1, where $\{X_i\}$ is a finite-valued stationary ergodic process and $s(\chi)$ is its entropy rate, which is defined from the entropy, $S(X_1, \ldots, X_n)$, as[26]

$$s(\chi) = \lim_{n\to\infty}\frac{1}{n}S(X_1, \ldots, X_n). \quad \text{(D2)}$$

This theorem relates the probability of the sequences that majorly contribute to the entropy, i.e., sequences in the so-called typical set,[26] which are nearly equiprobable, to the entropy rate. This probability can in turn be expressed as a product of *typical* probabilities per incorporated symbol, $\overline{p(X_i)}$,

$$p(X_1, \ldots X_n) = \overline{p(X_i)} \times \underbrace{\cdots}_{(n \text{ times})} \times \overline{p(X_i)}. \quad \text{(D3)}$$

In a system with positive feedback in which the probabilities for correct symbol incorporations are the highest, the sequences of the typical set are those with the lowest number of errors. This is the case of DNA replication, in which Watson-Crick rules apply and correlations over previous neighbors enhance these rules. Then, the typical probability of error per symbol, or error rate, p_{error}, is obtained from $p_{error} = 1 - \overline{p(X_i)} = 1 - (p(X_1, \ldots X_n))^{1/n}$. When the probability of error per symbol is very low (high certainty), it follows that the entropy is also very low and therefore

$$p_{error} \to 1 - \exp\left(-\frac{s(\chi)}{k}\right) \simeq \frac{s(\chi)}{k}, \quad \text{(D4)}$$

In conclusion, the error rate for a positive-feedback system for which the correctly incorporated symbols correspond to the highest probabilities can be estimated from the geometric mean of the probability of the sequences in the typical set, which is in turn obtained from the general AEP. Then, for a sufficiently low probability of error per incorporated symbol, the error rate is proportional to the entropy rate.

In the directional chain construction, there is a one-to-one correspondence between time and objects/symbols. Since the memory of the system comprises a finite number of previous objects, only if the total number objects/symbols is infinite, the system will visit all the possible configurations, thus preserving ergodicity. For the case of finite but very large n, the calculation of error rates through the AEP for the directional chain is therefore approximate.

[1] J. R. Arias-Gonzalez, Integr. Biol. **6**, 904 (2014).
[2] C. Bustamante, C. Cheng, and Y. X. Mejia, Cell **144**, 480 (2011).
[3] A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz, Nature **483**, 187 (2012).
[4] R. Landauer, IBM J. Res. Develop. **5**, 261–269 (1961).
[5] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[6] C. H. Bennett, Int. J. Theor. Phys. **21**, 905 (1982).
[7] F. G. S. L. Brandão and M. B. Plenio, Nat. Phys. **4**, 873 (2008).
[8] B.-H. Liu, L. Li, Y.-G. Huang, C.-F. Li, G.-C. Guo, E.-M. Laine, H.-P. Breuer, and J. Piilo, Nat. Phys. **7**, 931 (2011).
[9] R. K. Pathria and P. D. Beale, *Statistical Mechanics*, 3rd ed. (Academic Press, Boston, 2011).
[10] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, 1987).
[11] J. R. Arias-Gonzalez, e-print arXiv:1511.06139 [cond-mat.stat-mech] (2015).
[12] M. C. Wang and G. E. Uhlenbeck, Rev. Mod. Phys. **17**, 323 (1945).
[13] S. Pressé, J. Lee, and K. A. Dill, J. Phys. Chem. B **117**, 495 (2013).
[14] H.-P. Breuer, J. Phys. B: At., Mol. Opt. Phys. **45**, 154001 (2012).
[15] A. Rivas, S. F. Huelga, and M. B. Plenio, Rep. Prog. Phys. **77**, 094001 (2014).
[16] T. A. Kunkel and K. Bebenek, Annu. Rev. Biochem. **69**, 497 (2000).
[17] L. A. Loeb and T. A. Kunkel, Annu. Rev. Biochem. **51**, 429 (1982).
[18] H. R. Lee and K. A. Johnson, J. Biol. Chem. **281**, 36236 (2006).
[19] F. Bernardi and J. Ninio, Biochimie **60**, 1083 (1978).

[20]J. R. Arias-Gonzalez, PLoS One **7**, e42272 (2012).
[21]D. Andrieux and P. Gaspard, Proc. Natl. Acad. Sci. U. S. A. **105**, 9516 (2008).
[22]D. Andrieux and P. Gaspard, J. Chem. Phys. **130**, 014901 (2009).
[23]C. H. Bennett, Biosystems **11**, 85 (1979).
[24]J. Ninio, Biochimie **57**, 587 (1975).
[25]J. J. Hopfield, Proc. Natl. Acad. Sci. U. S. A. **71**, 4135 (1974).
[26]T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, 1991).
[27]M. Fisz, *Probability Theory and Mathematical Statistics* (Krieger Publishing Company, 1980).
[28]P. Schindler, D. Nigg, T. Monz, J. T. Barreiro, E. Martinez, S. X. Wang, S. Quint, M. F. Brandl, V. Nebendahl, C. F. Roos, M. Chwalla, M. Hennrich, and R. Blatt, New J. Phys. **15**, 123012 (2013).
[29]S. Kamtekar, A. J. Berman, J. Wang, J. M. Lázaro, M. de Vega, L. Blanco, M. Salas, and T. A. Steitz, Mol. Cell **16**, 609 (2004).
[30]S. J. Johnson and L. S. Beese, Cell **116**, 803 (2004).
[31]O. O. Brovarets' and D. M. Hovorun, RSC Adv. **5**, 99594 (2015).
[32]B. Ibarra, Y. R. Chemla, S. Plyasunov, S. B. Smith, J. M. Lázaro, M. Salas, and C. Bustamante, EMBO J. **28**, 2794 (2009).
[33]H. Echols and M. F. Goodman, Annu. Rev. Biochem. **60**, 477 (1991).
[34]J. SantaLucia, Jr. and D. Hicks, Annu. Rev. Biophys. Biomol. Struct. **33**, 415 (2004).
[35]D. A. Erie, T. D. Yager, and P. H. von Hippel, Annu. Rev. Biophys. Biomol. Struct. **21**, 379 (1992).
[36]J. SantaLucia, Jr., Proc. Natl. Acad. Sci. U. S. A. **95**, 1460 (1998).
[37]C. R. Calladine, H. R. Drew, B. F. Luisi, and A. A. Travers, *Understanding DNA: The Molecule and How It Works*, 3rd ed. (Elsevier, 2004).
[38]O. O. Brovarets' and D. M. Hovorun, Phys. Chem. Chem. Phys. **17**, 15103 (2015).
[39]O. O. Brovarets' and D. M. Hovorun, J. Biomol. Struct. Dyn. **33**, 2297 (2015).
[40]O. O. Brovarets' and D. M. Hovorun, J. Biomol. Struct. Dyn. **33**, 2710 (2015).
[41]O. O. Brovarets' and D. M. Hovorun, RSC Adv. **5**, 66318 (2015).
[42]O. O. Brovarets' and D. M. Hovorun, Phys. Chem. Chem. Phys. **17**, 21381 (2015).
[43]R. Guajardo and R. Sousa, J. Mol. Biol. **265**, 8 (1997).
[44]H. Yin, M. D. Wang, K. Svoboda, R. Landick, S. M. Block, and J. Gelles, Science **270**, 1653 (1995).
[45]J. Saturno, L. Blanco, M. Salas, and J. A. Esteban, J. Biol. Chem. **270**, 31235 (1995).
[46]G. J. L. Wuite, S. B. Smith, M. Young, D. Keller, and C. Bustamante, Nature **404**, 103 (2000).
[47]J. A. Morin, F. J. Cao, J. M. Lázaro, J. R. Arias-Gonzalez, J. M. Valpuesta, J. L. Carrascosa, M. Salas, and B. Ibarra, Proc. Natl. Acad. Sci. U. S. A. **109**, 8115 (2012).
[48]J. A. Morin, F. J. Cao, J. M. Lázaro, J. R. Arias-Gonzalez, J. M. Valpuesta, J. R. Carrascosa, M. Salas, and B. Ibarra, Nucleic Acids Res. **43**, 3643 (2015).
[49]R. R. Iyer, A. Pluciennik, V. Burdett, and P. L. Modrich, Chem. Rev. **106**, 302 (2006).