Inverted and mirror repeats in model nucleotide sequences

Fabrizio Lillo[1,2] and Marco Spanò[1]
[1]*Dipartimento di Fisica e Tecnologie Relative, Università di Palermo, Viale delle Scienze, I-90128, Palermo, Italy*
[2]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*
(Received 16 May 2007; published 23 October 2007)

We analytically and numerically study the probabilistic properties of inverted and mirror repeats in model sequences of nucleic acids. We consider both perfect and nonperfect repeats, i.e., repeats with mismatches and gaps. The considered sequence models are independent identically distributed (i.i.d.) sequences, Markov processes and long-range sequences. We show that the number of repeats in correlated sequences is significantly larger than in i.i.d. sequences and that this discrepancy increases exponentially with the repeat length for long-range sequences.

DOI: 10.1103/PhysRevE.76.041914 PACS number(s): 87.10.+e, 02.50.−r, 05.40.−a

I. INTRODUCTION

The complete sequencing of large genomes has lead us to reconsider the importance of noncoding DNA or RNA in the regulation of the activity of the cell [1]. Many different types of sequences able to have a regulatory role have been discovered. Among these sequences inverted and mirror repeats play an important role. For example, inverted repeats provide the necessary condition for the potential existence of a hairpin structure in the transcribed messenger RNA or cruciform structures in DNA [2]. Inverted repeats play also an important role for regulation of transcription and translation. In bacteria, inverted repeats and the associated hairpin structures are often part of ρ-independent transcription terminators [3,4]. In recent years there has been a growing interest for these structures triggered by the discovery of new classes of regulatory elements. Prominent examples of these new regulatory RNA families are microRNA (miRNA) [5–7] and small interference RNA (siRNA) [8,9]. Most of these structures share the property of being associated with a hairpin secondary structure. DNA or RNA short sequences that may be associated to RNA secondary structures are present in genomes of different species of phages, viruses, bacteria, and eukaryotes. Indications about the potential existence of RNA secondary structures can be inferred throughout the detection of short pair sequences having the characteristic of inverted repeats in the investigated genomes [10–13]. Also mirror repeats may have multiple biological roles. For example, perfect or near-perfect homopurine or homopyrimidine mirror repeats can adopt triple-helical H conformations [14]. Several computer programs have been developed to detect repeats and/or the associated secondary structure in DNA or RNA sequences [15,16]. Few studies have considered the problem of the expected number of repeats in model sequences [17,18], mainly investigating the clustering of repeats.

The purpose of this paper is to derive analytical and numerical expressions for the expected number of two specific, yet very important, types of repeats under the assumption that the investigated sequence can be modeled with a given family of stochastic process. In this paper we consider inverted and mirror repeats and we investigate four different types of sequence models. Specifically, we consider independent and identically distributed sequences, first-order Markov chains, higher order Markov processes, and long-memory sequences. For the first two types of models we are able to derive analytically expressions for the number of repeats, while for the last two classes of models we use numerical simulations to infer phenomenological expressions for the expected number of repeats.

The outline of the paper is the following. In Sec. II we introduce the investigated repeats and in Sec. III we introduce the sequence models discussed in the paper. In Sec. IV we consider independent and identically distributed sequences and we derive several analytical expressions for repeats. In Sec. V we consider first-order Markov chains and in Sec. VI we compute numerically the expected number of repeats for higher order Markov processes. In Sec. VII we consider long-memory sequences and Sec. VIII concludes.

II. INVERTED AND MIRROR REPEATS

In this paper we consider two types of repeats, i.e., inverted and mirror repeats. These repeats are composed by two nonoverlapping segments of nucleotide sequence that can be separated by another nucleotide subsequence. A mirror repeat is, for example, 5′GATTCGAacgAGCTTAG3′ where the sequence GATTCGA is repeated in an inverted way after the spacer acg. An inverted repeat is, for example, given by the sequence 5′GATTCGAacgTCGAATC3′ where the sequence GATTCGA is repeated and complemented after the spacer acg. One of the problems in counting repeats is the fact that a single repeat can be counted many times if one does not define in some way a maximal repeat. Consider, for example, the sequence 5′aggaatcgatcttaacgaagatcgattcca3′. This sequence contains many different inverted repeats, for example, 5′aggAATCGatcttaacgaagatCGATTcca3′ or 5′aggaaTCGATCttaacgaaGATCGAttcca3′. If one does not consider inverted with mismatches, there is one *maximal* inverted repeat, i.e., 5′aGGAATCGATCTTaacgAAGATCGATTCCa3′, in which the first base before and after the structure are not complementary and also the first and the last base of the spacer aacg are not complementary. When one considers inverted or mirror repeats with mismatches the definition of maximal repeat is less clear and must be clearly defined (see

Sec. IV B). In this paper we are interested in finding the expected number of maximal inverted and mirror repeats in model genome sequences.

A repeat is characterized by the assignment of a matching rule between couples of nucleotides. For RNA sequences the matching rule is defined by a 4×4 matrix whose rows and columns correspond to nucleotides A, C, G, and U. A matrix entry is 1 if the matching between the nucleotides in the row and in the column is allowed and zero elsewhere. For example, the characteristic matrix for inverted repeats in which only Watson-Crick base pair (i.e., A-U and C-G) are allowed is

$$\mathbf{M}_{(\mathrm{inv})} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{1}$$

If the pairing G-U (or GU wobble) is allowed the matrix becomes

$$\mathbf{M}'_{(\mathrm{inv})} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}. \tag{2}$$

Finally, for mirror repeats the characteristic matrix is

$$\mathbf{M}_{(\mathrm{mir})} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3}$$

Inverted and mirror repeats can be formed both in DNA and in RNA. Our results are the same for both nucleic acids provided one replaces T with U.

Given a matching rule, a *perfect repeat* of stem length ℓ exists at point x when, for a loop value m, every base $x+1-i$ matches every base $x+m+i$ for $1\leq i\leq \ell$. The sequence from $x+1-\ell$ to x will be called left arm of the stem, whereas the sequence from $x+m+1$ to $x+m+\ell$ will be called right arm of the stem. Since we are interested in maximal repeats, the repeats are defined also by requiring that base $x+1$ does not match base $x+m$ and base $x-\ell$ does not match with base $x+m+\ell+1$. We will call these repeats perfect because there are no bulges or mismatches. A *gap* or *one-base bulge* is present in the left arm of the stem if there exists an index j such that the above relation is true for $i\leq j$, whereas for $i>j$ every base $x-i$ matches every base $x+m+i$. The extension to bulge in the right arm of the stem is straightforward. Finally, a one nucleotide *mismatch* (or internal loop) is present in the stem if for some i' between 1 and $\ell-2$, the base $x-i'$ does not match with base $x+m+i'+1$. More mismatches or a mismatch composed of more than one base can be present in a stem. Inverted repeats are known to be able to create hairpin structures in single strand nucleic acids. Figure 1 shows an example of hairpin structure formed by an inverted repeat with a bulge and a three base mismatch. The caption should help the reader in understanding the terminology used in this paper.

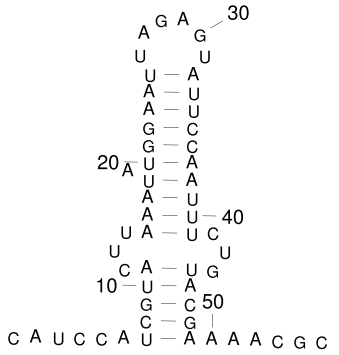

FIG. 1. Secondary structure formed by an inverted repeat in single stranded RNA. Bases from 7 to 25 constitute the left arm of the stem and bases from 32 to 49 constitute the right arm. The loop is made by bases 26–31. At base 19 there is a gap (or one-base bulge), and bases 12–14 and 42–44 constitute a three base mismatch (or internal loop). Note that base 6 is not complementary to base 50 and base 26 is not complementary to base 31 in order to have maximal repeats. According to the terminology used in the paper we have $\ell=18$, $m=6$, $k=3$, $g=1$.

The purpose of this paper is to derive the expected number of repeats of a given type for simple models of nucleotide sequences. We shall indicate with $N(\ell,m,k,g)$ the expected number of repeats of stem of length ℓ, loop of length m, k one-nucleotide mismatches, and g gaps. The calculation of the expected number of repeats is complex for two main reasons. The first problem is to compute the probability $P(\ell,m,k,g)$ that a given short sequence generated according to a sequence model can host a repeat with given characteristics. Once this probability is known the next problem is to estimate the expected number of repeats observed in a long sequence (genome) composed by N nucleotides. If the occurrence of different structures were independent one from the other the expected number would be simply $N(\ell,m,k,g)=NP(\ell,m,k,g)$. Unfortunately, in general the occurrence of a given structure is not independent of the presence of another structure. In the statistical search of simple words in genomes this is a known problem (see, for example, [19]). However, since we search for maximal repeats and the structures we are interested in are long and complex, we neglect the problem of nonindependence. In all the cases considered below we have performed extensive numerical simulations to test our formulas and, indirectly, the independence assumption. By performing careful statistical tests (usually χ^2 tests) we cannot reject the hypothesis that our formulas are correct. For this reason in the following we present the formulas both for $N(\ell,m,k,g)$ and for $P(\ell,m,k,g)$.

III. MODELS FOR NUCLEOTIDE SEQUENCES

A. Independent identically distributed sequences

The simplest model for nucleotide sequences is the independent identically distributed (i.i.d.) model. In this model one assumes independent nucleotides with probabilities p_a,

p_c, p_g, and p_u, such that $p_a+p_c+p_g+p_u=1$. Although it is known that correlation between nucleotides are significant, this model allows exact analytical calculations and can be used as a useful starting point.

It is useful to define the probability vector $\mathbf{p}^T \equiv (p_a,p_c,p_g,p_u)$, where the elements are the nucleotide probabilities. Given a type of structure characterized by the matrix $\mathbf{M}$ we introduce the scalar quantity

$$q = \mathbf{p}^T \mathbf{M} \mathbf{p}. \tag{4}$$

For example, inverted repeats have $q=2p_ap_u+2p_cp_g$, whereas for mirror repeats $q=p_a^2+p_c^2+p_g^2+p_u^2$.

B. Markov models

A better class of models for nucleotide sequences is the class of Markov processes. Let us consider for convenience the infinite sequence X_i, where $i \in Z$ and Z is the set of integers. An ergodic stationary mth-order Markov chain is characterized by the transition matrix

$$p(a_{m+1}|a_1, \ldots, a_m) = P(X_i = a_{m+1}|X_{i-1} = a_m, \ldots, X_{i-m} = a_1). \tag{5}$$

The simplest Markov chain we shall consider extensively in the following is the first-order Markov chain. This type of process is characterized by the 4×4 transition matrix $p(a_2|a_1)$. By taking powers of this matrix one can also define the k-step transition matrix whose elements are $p_k(b|a) = P(X_i=b|X_{i-k}=a)$. In this notation $p(a_2|a_1)=p_1(a_2|a_1)$.

The model parameters, i.e., the order of the Markov chain and the transition probabilities, of a real sequence can be estimated by the maximum-likelihood method (see, for example, [19]).

C. Long-memory models

In recent years it has been proposed that parts of real genomes are not well described by Markovian models, but rather that a long-memory (or long-range) process describes better the correlation properties of nucleotide sequences [20–24]. There are several ways of detecting and modeling correlation properties of nucleotide sequences. The approach we will follow is called "DNA walk" [20] and consists in mapping the nucleotide sequence in a one-dimensional random walk x. Since there are four different residues in a RNA sequence while the random walk has two possible directions ($\Delta x=\pm1$), one needs to choose a mapping rule from the four residues to the two directions. Several different mapping rules have been introduced [24]. In the present paper we consider two important rules: (i) the purine-pyrimidine rule (or RY rule) which assigns $\Delta x=+1$ if the residue is a purine (A or G) and $\Delta x=-1$ if the residue is a pyrimidine (C or U) and (ii) the hydrogen bond energy rule (or SW rule) which assigns $\Delta x=+1$ for strongly bonded residues (C or G) and assigns $\Delta x=-1$ for weakly bonded residues (A or U). This second rule can be useful to take into account the isochore structure of the genome [25]. By using these rules it has been observed that in most cases noncoding DNA sequences, i.e., DNA sequences not coding for proteins, display long-memory properties of the corresponding DNA walk. We remind that a long-memory process is a process whose autocorrelation function of Δx_i decays in time as $\mathrm{Corr}(\Delta x_{i+\tau}\Delta x_i) \sim \tau^{-\gamma}$, where $0<\gamma<1$. Long-memory processes are an important class of stochastic process that have found application in many different fields [26]. The autocorrelation function of a long-memory process is not integrable in τ between 0 and $+\infty$ and, as a consequence, the process does not have a typical time scale. Long-memory processes are better characterized by the Hurst exponent H that, for long-memory processes, is $H=1-\gamma/2$. Thus for long-memory processes $1/2<H<1$.

Long-memory properties of nucleotide sequences has been associated to different genome characteristics including nucleosomal structure in eukaryotes [27], to the presence of isochores [25] and to the presence of tandem repeats [28]. More recently it has been suggested that in some genomes (for example, human) the correlation properties of DNA cannot be captured by a single Hurst exponent, but rather that the Hurst exponent may depend on the observation scale [29,30]. Different scales can be associated with different biological structures (genes, transposable elements, isochores).

In Sec. VII we use simulations of long-memory nucleotide sequences to obtain phenomenological expressions for the expected number of inverted or mirror repeats. In order to simulate long-memory nucleotide sequences we generate a fractional Gaussian noise (FGN) signal [26] with Hurst exponent H_0 by using the R package. The unconditional distribution of the FGN is Gaussian and in order to obtain binary sequences we simply take the sign of the FGN. Extensive numerical simulations have shown that the sign of a FGN has a Hurst exponent given by $H_{\mathrm{sign}} \simeq H_0-0.02$. Thus the algorithm to generate, for example, a long-memory nucleotide sequence with Hurst exponent H and according to rule SW is the following: (i) generate a FGN with Hurst exponent equal to $H+0.02$, (ii) take the sign, (iii) if the sign is positive with probability 1/2 the nucleotide is C and with probability 1/2 the nucleotide is G, (iv) similarly, if the sign is negative with probability 1/2 the nucleotide is A and with probability 1/2 the nucleotide is T. Note that in this way we obtain a sequence with equal nucleotide frequencies. It is possible to modify this algorithm to have a variable CG content by replacing the sign function with the Heavyside step function $\Theta(x-q)$ where x is the outcome of the FGN and q is an appropriate quantile value ($q=0$ for a CG content of 50%). For a generic value of q, the correct amount to add to the Hurst exponent has been estimated by careful numerical simulations. It is worth noting that we have also used other methods to generate the long-memory sequences obtaining similar results. The investigated methods are the fARIMA model [26] and the patch model [31]. Finally it is important to stress that this simulation method under specifies the full correlation structure of the nucleotide sequence. To give a specific example consider a SW long-memory sequence with a CG content close to one. In this limit the sequence is in fact i.i.d. and the reason is that the model does not specify the correlation of C and G.

IV. INVERTED AND MIRROR REPEATS IN i.i.d. SEQUENCES

A. Perfect repeats

The expected number of perfect repeats of stem length ℓ and loop length m in a i.i.d. genome of length N characterized by the parameter q is

$$N(\ell,m) = N(1-q)^{\alpha}q^{\ell}, \tag{6}$$

where the exponent α is equal to 1 for $m \leq 1$ and is equal to 2 for $m \geq 2$. In other words we need to impose that the ℓ bases of the left arm of the stem match with the corresponding bases in the right arm. Moreover, we need to impose that the first couple of bases in the loop do not match, such as the first couple of bases at the end of the stem. When the loop is shorter than two nucleotides one cannot impose that the first couple of bases in the loop do not match and this explains the different value of the exponent α. Since in an i.i.d. sequence the occurrences of nucleotides are independent, probabilities factorize and Eq. (6) is obtained. This expression has been used, for example, in Ref. [12] to investigate the number of perfect inverted repeats in bacterial genomes.

B. Inverted with mismatches

A mismatch in a repeat is the presence of a pair of nucleotides in the stem that do not match. We indicate with k the number of mismatches in the stem and we look for an expression for $N(\ell,m,k)$. We prove that the expected number is

$$N(\ell,m,k) = N\binom{\ell-2}{k}(1-q)^{\alpha+k}q^{\ell-k}, \tag{7}$$

where the exponent α assumes the same values as in Eq. (6). In fact a mismatch can be present only in one of the $\ell-2$ internal nucleotides of the stem [i.e., from the second to the $(\ell-1)$th nucleotide]. There are $\binom{\ell-2}{k}$ ways of placing k mismatches in $\ell-2$ internal bases of the stem.

One of the problems with Eq. (7) is the fact that, for example, a repeat with one mismatch can also be seen as a repeat with zero mismatches and a shorter stem. We shall denote these two repeats as *embedded*. One is usually interested in counting more embedded repeats only once. Moreover, programs designed for the search of inverted repeats, such as *palindrome* of the EMBOSS package [15], effectively count embedded inverted repeats only once. Therefore, we need a formula for nonembedded repeats. Clearly any repeat with, say, zero mismatches can be thought of as part of a longer repeat with a large number of mismatches. In other words, we need to introduce an upper value of the number of mismatches, in order to find an expression for nonembedded repeats up to a chosen value of the number of possible mismatches. For example, we can ask for the expected number of inverted repeats with zero mismatches that cannot be seen as part of longer inverted repeats with one mismatch. This of course does not guarantee that the found repeats cannot be part of repeats with two mismatches. From an operative point of view, this corresponds with the run of the search program (for example, *palindrome*) with a maximal number of mismatches equal to $\bar{k}$. Therefore, a quantity more meaningful than Eq. (7) is $N^{(\bar{k})}(\ell,m,k)$, which is the expected number of repeats of stem length ℓ, loop length m, and k mismatches, that cannot be part of a longer repeat of the same type with at most $\bar{k}$ mismatches. By definition $\bar{k} \geq k$. The two expressions of Eqs. (6) and (7) correspond to $N^{(0)}(\ell,m,0)$ and $N^{(k)}(\ell,m,k)$, respectively.

When $\bar{k}=1$, we have

$$N^{(1)}(\ell,m,0) = N(1-q)^{\alpha}q^{\ell},$$

$$\alpha = \begin{cases} 2 & \text{for } 0 \leq m \leq 1, \\ 3 & \text{for } 2 \leq m \leq 3, \\ 4 & \text{for } m \geq 4. \end{cases} \tag{8}$$

When $\bar{k}=2$, we have

$$N^{(2)}(\ell,m,1) = N(\ell-2)(1-q)^{\alpha+1}q^{\ell-1},$$

$$\alpha = \begin{cases} 2 & \text{for } 0 \leq m \leq 1, \\ 3 & \text{for } 2 \leq m \leq 3, \\ 4 & \text{for } m \geq 4, \end{cases} \tag{9}$$

and

$$N^{(2)}(\ell,m,0) = N(1-q)^{\alpha}q^{\ell},$$

$$\alpha = \begin{cases} 3 & \text{for } 0 \leq m \leq 1, \\ 4 & \text{for } 2 \leq m \leq 3, \\ 5 & \text{for } 4 \leq m \leq 5, \\ 6 & \text{for } m \geq 6. \end{cases} \tag{10}$$

The general formula is

$$N^{(\bar{k})}(\ell,m,k) = N\binom{\ell-2}{k}(1-q)^{\alpha+\beta}q^{\ell-k},$$

$$\alpha = \begin{cases} 1 & \text{for } 0 \leq m \leq 1, \\ 2 & \text{for } m \geq 2, \end{cases}$$

$$\beta = (\bar{k}-k) + \max\left\{0, \min\left[\left(\frac{m}{2}\right)-1, \bar{k}-k\right]\right\}, \tag{11}$$

where $[x]$ indicates the integer part of x.

We have performed extensive numerical simulations of artificial genomes and we have verified that these expressions are correct. Specifically, we have written computer programs able to detect inverted or mirror repeats with the required characteristics (stem and loop length, mismatches, etc.). Then we have performed a χ^2 test between the frequency of observed repeats and the frequency expected by our theory. In all cases we cannot reject the hypothesis that our formulas are correct.

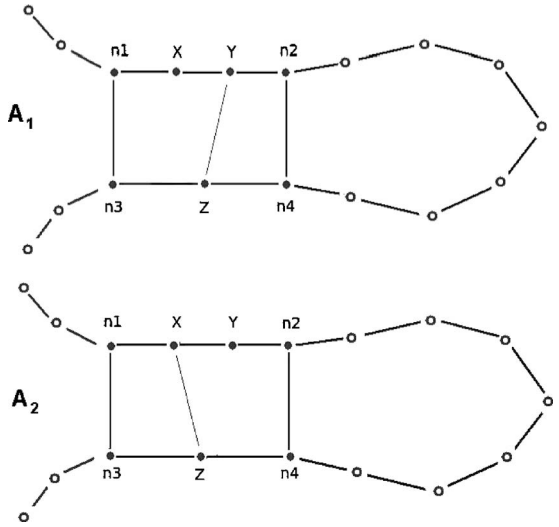

FIG. 2. Schematic representation of the two secondary structures that can be formed by an inverted repeat with stem length $\ell=3$, loop length $m=7$, and one gap in the case when both base X and base Y are complementary to base Z (and therefore $X=Y$). The continuous thin lines indicate Watson-Crick base pairs, whereas continuous thick lines indicate the nucleic acid backbone.

C. Repeats with one gap

We consider now the case of inverted and mirror repeats with one gap in the stem and no mismatches. We shall indicate with ℓ the number of links in the stem, since in such a structure there will be ℓ nucleotides in one branch of the stem and $\ell+1$ in the other. The expected number of repeats with the gap in one *specific* position is the same as for perfect repeats [see Eq. (6)], i.e.,

$$N(\ell,m,k=0,g=1)=N(1-q)^{\alpha}q^{\ell}, \tag{12}$$

where the exponent α is equal to 1 for $m\leq 1$ and is equal to 2 for $m\geq 2$. One could think that, since there are $\ell-1$ possible positions for the gap (on one arm), the expected number of repeats with one gap in any position of one arm is simply $\ell-1$ times the value in Eq. (12). This is wrong because the probability of observing the gap in one position is not independent from the probability of observing the gap in another position. To understand why, let us consider an inverted repeat with $\ell=3$ and one gap. As shown in Fig. 2 there are two positions for the gap, and the corresponding structures are indicated as A_1 and A_2 in the figure. The probability of observing either A_1 or A_2 or both is

$$P(A_1\cup A_2)=P(A_1)+P(A_2)-P(A_1\cap A_2). \tag{13}$$

$P(A_1)$ and $P(A_2)$ are equal to the quantity in Eq. (12), whereas $P(A_1\cap A_2)$ is the joint probability that the sequence can form both structures A_1 and A_2. By looking at the figure we note that the sequence can form both structures if $X=Y=\bar{Z}$, where the bar indicates complementarity. Thus, the joint probability is

$$\begin{aligned}P(A_1\cap A_2)&=(1-q)^{\alpha}q^2(p_a^2p_t+p_ap_t^2+p_c^2p_g+p_cp_g^2)\\&\equiv(1-q)^{\alpha}q^2\tilde{q}.\end{aligned} \tag{14}$$

For inverted repeats the quantity $\tilde{q}$ is the probability that $X=Y=\bar{Z}$ and it is equal to $p_a^2p_t+p_ap_t^2+p_c^2p_g+p_cp_g^2$. Analogously for mirror repeats $\tilde{q}$ is the probability that $X=Y=Z$ and it is equal to $p_a^3+p_t^3+p_c^3+p_t^3$. In conclusion, the expected number of repeats with $\ell=3$ and one gap is

$$N(\ell=3,m,k=0,g=1)=N(1-q)^{\alpha}q^2(2q-\tilde{q}), \tag{15}$$

which is of course different from the naive (and wrong) answer given by 2 times the equation (12). The generalization of this last formula to a generic value of ℓ is not straightforward and the derivation is reported in Appendix A. The result is

$$N(\ell,m,k=0,1)=2Nq^{\ell-1}(1-q)^{\alpha}[(\ell-1)q-(\ell-2)\tilde{q}],$$

$$\alpha=\begin{cases}1 & \text{for } 0\leq m\leq 1,\\ 2 & \text{for } m\geq 2,\end{cases} \tag{16}$$

where the factor 2 in front of N is due to the fact that the gap can be found in one of the two arms. It is worth noting that for large ℓ the correct answer of Eq. (16) is 3/4 of the naive and wrong answer given by $\ell-1$ times the expression of Eq. (12).

V. INVERTED AND MIRROR REPEATS IN FIRST-ORDER MARKOV CHAINS

We now give the expression for the expected number of repeats for a model sequence described by a first-order Markov chain. We consider the simplest case of the expected number of perfect repeats with a given stem (of length ℓ, as before) and a generic loop of length $m>2$.

The calculation is performed in Appendix B and the result is

$$\begin{aligned}P_{\text{Markov}}(\ell,m)=&\sum_{n_1,n_2,\ldots,n_\ell=1}^{4}p(n_1n_2\cdots n_\ell)p(\bar{n}_\ell\cdots\bar{n}_2\bar{n}_1)\\&\times\frac{\left(p(n_1)-\sum_{x=1}^{4}p(n_1|x)p(\bar{x}|\bar{n}_1)\right)\left(p_{m+1}(\bar{n}_\ell|n_\ell)-\sum_{y=1}^{4}p(\bar{n}_\ell|y)p_{m-1}(y|\bar{y})p(\bar{y}|n_\ell)\right)}{p(n_1)p(\bar{n}_\ell)},\end{aligned} \tag{17}$$

where $\bar{n}_i$ indicates a base matching with base n_i, i.e., the complementary of n_i for inverted repeats and $\bar{n}_i=n_i$ for mirror repeats. In Eq. (17) $p(n_i)$ is the probability of the occurrence of base i and $p(n_1 n_2 \cdots n_\ell)$ is the probability of the occurrence of the word $n_1 n_2 \cdots n_\ell$, that for Markov chain is easily computable (see also Appendix B). Even if the expression (17) looks complex, the numerical summation is easily and quickly performed, for example, with simple programs in Mathematica. It is worth noting that the summation is over 4^ℓ terms, whereas a direct calculation taking into account all the possible repeats would require to sum $4^{2\ell+2+m}$ terms.

The functional dependence of $P_{\text{Markov}}(\ell,m)$ from ℓ and m are not evident by eye, such as the relative magnitude of $P_{\text{Markov}}(\ell,m)$ and $P_{i.i.d.}(\ell,m)=(1-q)^\alpha q^\ell$ for an i.i.d. genome [see Eq. (6)]. Thus, we discuss here these issues by considering Markov models with parameters equal to the ones obtained by real genomes of model organisms. Specifically, we consider four complete genomes: (i) the Hepatitis B virus (accession NC_003977, length=3215 bp), (ii) the *Escherichia coli* K12 genome (accession NC_000913, length =4 639 675 bp), (iii) the *Drosophila melanogaster* mitochondrion (accession NC_001709, length=19 517 bp), and (iv) the *Homo sapiens* mitochondrion (accession NC_001807, length=16 571 bp). Moreover, we consider inverted repeats.

We first discuss the dependence of $P_{\text{Markov}}(\ell,m)$ from the loop length m. To this end we compute the ratio $P_{\text{Markov}}(\ell,m)/P_{i.i.d.}(\ell,m)$ for the stem length fixed at $\ell=4$, 5, and 6. Figure 3 shows this quantity for the four model genomes. We see that $P_{\text{Markov}}(\ell,m)$ has a small dependence from m. More precisely for m larger than few units, $P_{\text{Markov}}(\ell,m)/P_{i.i.d.}(\ell,m)$ becomes independent on m. The loop length dependence for small values of m can be positive [panels (a), (c), and (d)] or negative [panel (b)] with respect to the value for large m. In all cases the ratio

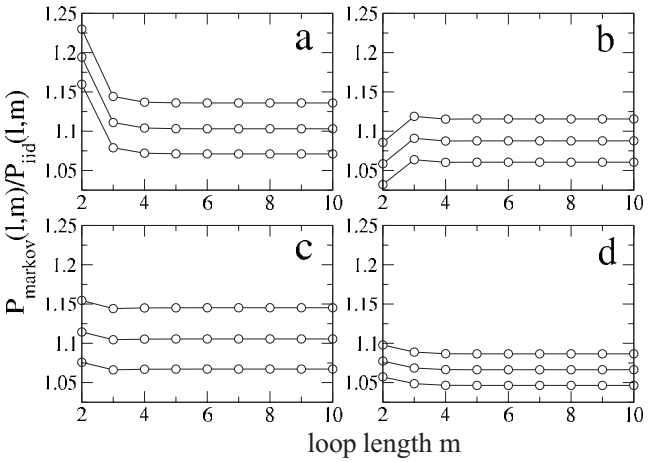

FIG. 3. Plots of the ratio $P_{\text{Markov}}(\ell,m)/P_{i.i.d.}(\ell,m)$ between the probability of observing an inverted repeat with stem length ℓ and loop length m in a Markov and in an i.i.d. genome as a function of the loop length m. The parameters characterizing the models are estimated by four model genomes, i.e., Hepatitis B virus (a), *E. coli* (b), *Drosophila* mitochondrion (c), and *Homo* mitochondrion (d). In each panel the curves refer to $\ell=4$, $\ell=5$, and $\ell=6$ (from bottom to top).

$P_{\text{Markov}}(\ell,m)/P_{i.i.d.}(\ell,m)$ is significantly larger than one and it increases with the stem length ℓ.

Because of the small dependence on m we can consider $P_{\text{Markov}}(\ell,m)$ for large values of m as a good approximation of the probability of observing repeats. This approximation leads to a simplification of Eq. (17). In fact, when m is large one can approximate the conditional probabilities in Eq. (17), $p_{m+1}(\bar{n}_\ell|n_\ell)\simeq p(\bar{n}_\ell)$ and $p_{m-1}(y|\bar{y})\simeq p(y)$. Thus, the probability $P_{\text{Markov}}(\ell,m)$ becomes independent from m and equal to

$$P_{\text{Markov}}(\ell,m)=\sum_{n_1,n_2,\ldots,n_\ell=1}^{4} p(n_1 n_2 \cdots n_\ell)p(\bar{n}_\ell \cdots \bar{n}_2 \bar{n}_1)\frac{\left(p(n_1)-\sum_{x=1}^{4} p(n_1|x)p(\bar{x}|\bar{n}_1)\right)\left(p(\bar{n}_\ell)-\sum_{y=1}^{4} p(\bar{n}_\ell|y)p(y)p(\bar{y}|n_\ell)\right)}{p(n_1)p(\bar{n}_\ell)}. \tag{18}$$

We can now study the dependence of $P_{\text{Markov}}(\ell,m)$ from the stem length ℓ, by considering the cases when m is larger than 4 bp. Figure 4 shows the ratio $P_{\text{Markov}}(\ell,m)/P_{i.i.d.}(\ell,m)$ as a function of ℓ for the four genomes. In all cases the ratio $P_{\text{Markov}}(\ell,m)/P_{i.i.d.}(\ell,m)$ increases almost linearly with the stem length ℓ. For $\ell\leq 10$ the order of magnitude of the error made by the i.i.d. model in predicting the number of repeats of a Markov sequence ranges between a few percent and 30%.

A. A simplified model

The fact that even for large values of m the number of inverted repeats expected in a Markovian genome is significantly larger than the number expected in an i.i.d. genome can be explained in a simplified model of genome sequence. We assume that the nucleotide alphabet is composed only by two symbols (instead of four), that the transition matrix is parametrized as

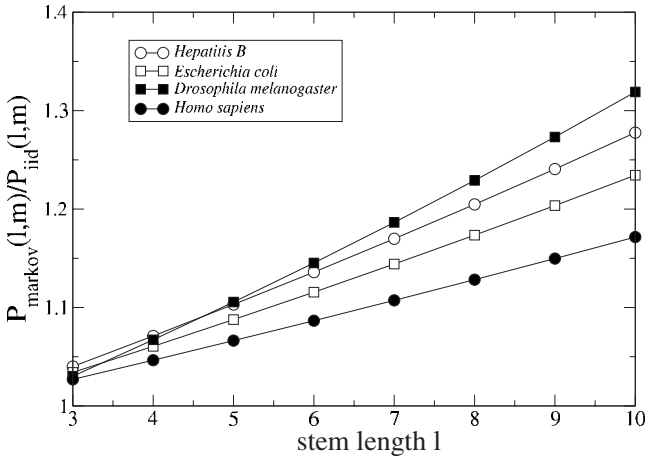

FIG. 4. Plots of the ratio $P_{\mathrm{Markov}}(\ell,m)/P_{i.i.d.}(\ell,m)$ between the probability of observing an inverted repeat with stem length ℓ and loop length $m>5$ in a Markov and in an i.i.d. genome as a function of the stem length ℓ. The parameters characterizing the models are estimated by four model genomes, i.e., Hepatitis B virus (empty circles), *E. coli* (empty squares), *Drosophila* mitochondrion (filled squares), and *Homo* mitochondrion (filled circles).

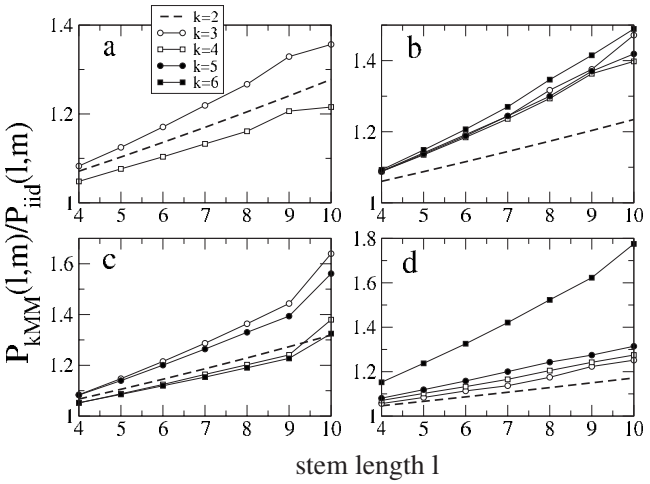

FIG. 5. Plots of the ratio $P_{kMM}(\ell,m)/P_{i.i.d.}(\ell,m)$ between the probability of observing an inverted repeat with stem length ℓ and loop length $m>5$ in a kth-order Markov and in an i.i.d. genome as a function of the stem length ℓ. The parameters characterizing the models are estimated by four model genomes, i.e., Hepatitis B virus [panel (a)], *E. coli* [panel (b)], *Drosophila* mitochondrion [panel (c)], and *Homo* mitochondrion [panel (d)].

$$\begin{pmatrix} \frac{1}{2}+\delta & \frac{1}{2}-\delta \\ \frac{1}{2}-\delta & \frac{1}{2}+\delta \end{pmatrix}, \tag{19}$$

and that the process is stationary, so that the probability for the two symbols are equal to 1/2. The parameter δ is a measure of the distance from the i.i.d. model. With this transition matrix, the conditional probability $p(n_2|n_1)$ is equal to $1/2+\delta$ if $n_1=n_2$ and to $1/2-\delta$ if $n_1\neq n_2$. We shall call permanence the first case and change the second one. We simplify further the original model by removing the constraint that the repeat is maximal, i.e., the condition that the two bases before and after the repeat are not complementary and that the first and last base in the loop are not complementary. The probability of an inverted repeat of stem length ℓ and loop length $m\gg 1$ is given by the product of the probability of the left-hand part of the stem times probability of the right-hand part of the stem. The probabilities factorize because we have assumed that the loop is large. Now the probability for a given word in the left-hand part of the stem is $2^{-1}(1/2-\delta)^{d_1}(1/2+\delta)^{d_2}$, where d_1 is the number permanencies, whereas d_2 is the number of changes. Clearly it is $d_1+d_2=\ell-1$. The probability for the inverted and complemented word in the right arm of the stem is equal, so the probability for a given inverted word is $[2^{-1}(1/2-\delta)^{d_1}(1/2+\delta)^{d_2}]^2$. We must sum this quantity over all possible words, i.e.,

$$P(\ell)=\frac{2}{4}\sum_{d_1=0}^{\ell-1}\binom{\ell-1}{d_1}\left(\frac{1}{2}+\delta\right)^{2d_1}\left(\frac{1}{2}-\delta\right)^{2(\ell-1-d_1)}$$
$$=\frac{1}{2}\left(\frac{1}{2}+2\delta^2\right)^{\ell-1}, \tag{20}$$

where the factor 2 in front of the sum comes from the fact that there are two possible words with the same position of the permanencies and of the changes obtained by exchanging one symbol with the other. For an i.i.d. sequence the probability for an inverted stem length ℓ is $P_{i.i.d.}(\ell)=2^{-\ell}$, thus the ratio is

$$\frac{P(\ell)}{P_{i.i.d.}(\ell)}=\frac{\frac{1}{2}\left(\frac{1}{2}+2\delta^2\right)^{\ell-1}}{\frac{1}{2^\ell}}=(1+4\delta^2)^{\ell-1}. \tag{21}$$

For small values of δ, i.e., for Markovian sequences not too different from i.i.d. ones, the binomial expansion gives

$$\frac{P(\ell)}{P_{i.i.d.}(\ell)}\simeq 1+(\ell-1)4\delta^2,\quad \delta\ll 1, \tag{22}$$

which is the almost linear behavior observed in Fig. 4. Thus, we expect that the linear behavior observed in Fig. 4 for the more complete model is valid for a moderate value of the stem.

VI. HIGHER ORDER MARKOV MODELS

In the case of higher order Markov processes the analytical computation of the expected number of inverted and mirror repeats becomes considerably more complex. Instead of trying to obtain complicated expression with difficult interpretation, we perform numerical simulations of higher order Markov chains and we compare the observed number of repeats with the number expected from the i.i.d. theory. The results of our simulations are shown in Fig. 5 and indicate that the error made in using an i.i.d. model to estimate the expected number of inverted repeats in a Markov chain increases with (i) the stem length ℓ and (ii) the order of the Markov process. Nevertheless it is worth pointing out that for moderate values of the stem length the ratio

$P_{kMM}(\ell,m)/P_{i.i.d.}(\ell,m)$ increases approximately linearly with ℓ. This implies that in the considered range the number of inverted repeats in a Markovian genome is given by

$$P_{kMM}(\ell,m) \sim A_k \ell q^{\ell}, \tag{23}$$

where A_k is a parameter which slowly increases with the order k of the Markov process.

VII. LONG-MEMORY PROCESSES

Finally we consider the problem of estimating numerically the probability of occurrence of an inverted or a mirror repeat in a long-range nucleotide sequence. Since most of the repeats with biological role are likely to be found in noncoding regions of the genome which are often composed by long-memory nucleotide sequences, this analysis is particularly relevant for the application to real cases. We generated long-memory nucleotide sequences by using either the RY rule or the SW rule and with different values of the Hurst exponent H. As discussed in Sec. II, in order to generate a RY long-memory genome we simulated a binary long-memory process with values $x_i=\pm 1$. Then, for each $x_i=+1$ we associated either A or G, each with probability 1/2; and for each $x_i=-1$, we associated either C or U, each with probability 1/2. Note that with this generation algorithm the simulated genomes have equal nucleotide frequencies, i.e., $p_a=p_c=p_g=p_t=1/4$. We then searched in the simulated genome for perfect repeats with a given stem length ℓ and loop length m and we compare the observed frequencies with the one expected by an i.i.d. genome. First, we find that also for long-memory sequences the occurrence of inverted or mirror repeats is essentially independent on the value of the loop length m. As for the Markovian case we find a small dependence for very small values of m. The behavior as a function of the stem length ℓ is very different from the i.i.d. case. In Fig. 6 we plot the quantity $P_{\mathrm{LM}}(\ell,m)/P_{i.i.d.}(\ell,m)$ as a function of ℓ, where $P_{\mathrm{LM}}(\ell,m)$ is the observed probability of inverted repeats in the long-memory sequence. The left-hand panel shows the SW (or hydrogen bond energy) rule and the right-hand panel shows the RY (or purine-pyrimidine) rule. In the RY case for $\ell \lesssim 5$ there is a decrease of the number of inverted repeats with respect to the i.i.d. case whereas for $\ell \gtrsim 5$ the number of observed inverted repeats is larger than the number expected in the i.i.d. case. However the value of the ratio $P_{\mathrm{LM}}(\ell,m)/P_{i.i.d.}(\ell,m)$ is never very large. For the SW rule a different behavior is observed. In the left-hand panel of Fig. 6 the y axis is in a logarithmic scale and the ratio $P_{\mathrm{LM}}(\ell,m)/P_{i.i.d.}(\ell,m)$ has a clear exponential dependence on ℓ. Very large values of the ratio are observed showing that using the i.i.d. formula for long-memory sequence can lead to a severe underestimation of the expected repeats. The difference observed between the two rules can be easily explained by recalling that an inverted repeat is formed when many bonds can be formed between complementary bases. Since in the SW rule the presence of, say, a C is strongly correlated with the presence of a G nearby, it is intuitive to understand why many more inverted repeats are observed in a SW than in a RY long-memory genome with the same Hurst exponent.

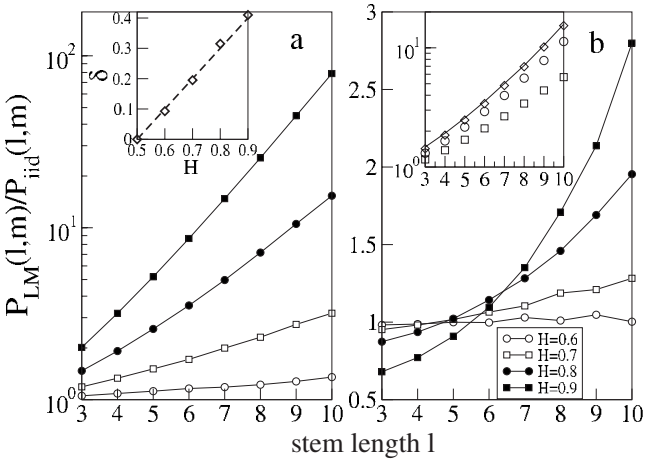

FIG. 6. Plots of the ratio $P_{\mathrm{LM}}(\ell,m)/P_{i.i.d.}(\ell,m)$ between the probability of observing an inverted or a mirror repeat with stem length ℓ and loop length $m>5$ in a long-memory and in an i.i.d. genome as a function of the stem length ℓ and of the Hurst exponent H. Data of panel (a) are generated according to the SW (or hydrogen bond energy) rule and the inset of panel (a) shows the fitted δ (see text) as a function of H. The dashed line is the function $\delta=H-1/2$. Data of panel (b) refer to inverted repeats and are generated according to the RY (or purine-pyrimidine) rule. The inset of panel (b) shows the ratio $P_{\mathrm{LM}}(\ell,m)/P_{i.i.d.}(\ell,m)$ for mirror repeat in a long-memory sequence with $H=0.8$ and generated according to the SW rule. The different symbols refer to different values of the CG content, specifically CG=50% (continuous line), CG=60% (diamonds), CG=70% (circles), and CG=80% (squares). For each value of H we simulated an artificial genome of length 10^8 bp.

Since it is difficult to develop a theory for the number of repeats in a long-memory genome, we try to get some intuition by considering the simplified model for Markovian genomes presented in Sec. V A. We remind that Eq. (21) predicts that the ratio $P(\ell)/P_{i.i.d.}(\ell)$ depends exponentially from ℓ according to $\exp[\ell \ln(1+4\delta^2)]$ where δ quantifies the "distance" of the model from the i.i.d. case. We fitted the curves in the left-hand panel of Fig. 6 with an exponential function and we estimated the corresponding value of δ as a function of H. The inset of the left-hand panel of Fig. 6 shows that to a good approximation $\delta=H-1/2$. This allows us to conjecture that the number of inverted repeats in SW long-memory sequences is

$$\begin{aligned} P_{\mathrm{LM}}(\ell,m) &= P_{i.i.d.}(\ell,m)\exp\{\ell \ln[1+4(H-1/2)^2]\} \\ &\simeq q^{\ell} \exp\{\ell \ln[1+4(H-1/2)^2]\}. \end{aligned} \tag{24}$$

For mirror repeats we find that long-memory sequences generated according to either SW or the RY rule show a behavior essentially indistinguishable from the one shown in the left-hand panel of Fig. 6. The reason is that both rules significantly increase the probability that two equal symbols are found at a short distance. As a consequence Eq. (24) holds also for mirror repeats according to either SW or RY rule. We stress again that this formula holds for sequences with approximately equal nucleotide frequencies. In many cases of interest the CG content is different from 50%. Since we

are not able to perform analytical calculations for this case we perform numerical simulations. The method we use to generate biased long-memory sequences is discussed in Sec. III. The inset of the right-hand panel of Fig. 6 shows the ratio $P_{\mathrm{LM}}(\ell,m)/P_{i.i.d.}(\ell,m)$ for mirror repeats in a long-memory genome with $H=0.8$ and generated according to the SW rule. Different symbols refer to different values of the CG content. We show CG values larger than 50% but for CG$<$50% we obtain a similar behavior, i.e., for example, the curve corresponding to CG=70% is indistinguishable from the one with CG=30%. We observe that for fixed ℓ the ratio $P_{\mathrm{LM}}(\ell,m)/P_{i.i.d.}(\ell,m)$ diminishes when the CG content increases. It is worth noting that the exponential behavior of the ratio is observed for all values of the CG content. For inverted repeats we observe a behavior similar to the one shown in the inset of the right-hand panel of Fig. 6. We believe that part of the decrease of the ratio $P_{\mathrm{LM}}(\ell,m)/P_{i.i.d.}(\ell,m)$ may be due to the under specification of the correlation structure discussed at the end of Sec. III. Therefore, as mentioned above, Eq. (24) holds exactly only for sequences with CG=50%. Our numerical results show that when 40%$<$CG$<$60% the error made in using Eq. (24) is smaller than 3%. It is important to stress that the results we have obtained in this section are clearly a first step toward a full understanding of the relation between long-memory sequences and occurrence of repeats. In fact the considered long-memory model does not reproduce the full correlation structure because it is based on an artificial binary correspondence (see discussion at the end of Sec. III).

In conclusion, differently from the Markov case, the exponential behavior of $P(\ell)/P_{i.i.d.}(\ell)$ expected from the simplified model is observable in long-memory sequences also for small values of ℓ. This is very important because it means that when the sequence is long memory (as in many noncoding sequences) the expected number of repeats can be significantly larger than the number expected in an i.i.d. sequence. The discrepancy between i.i.d. and long-memory models increases very quickly with $H-1/2$. Many regions of real genomes can have very large values of H. For example, parts of the human chromosome 22 have an estimated Hurst exponent $H=0.88$ [32]. In these cases a careful modeling of the nucleotide sequence is very important in estimating the expected number of repeats.

VIII. CONCLUSIONS

In conclusion we have developed many analytical and numerical results for the expected number of inverted and mirror repeats with different features (stem length, loop length, presence of mismatches or gap) under the assumption that the investigated sequence can be modeled with different types of sequence models. In general the computation of the number of repeats in model sequences is a complicated problem due to combinatorial difficulties, nonindependence of different occurrences (as in the case of gaps), and difficulties related to the sequence model (as for higher order Markov process and long-memory sequences). To the best of our knowledge, this is the most comprehensive study of the occurrence of inverted and mirror repeats in model sequences. However, the task of obtaining a full understanding of the relation between nucleotide correlation properties and occurrence of repeats is far from being achieved. Our results have been obtained by using the simplest model of long-memory nucleotide sequence, specifically the binary model. In this sense the results of Sec. VII are a first step toward the understanding of the relation between long-memory sequences and repeats.

A careful estimation of the expected number of repeats in a model sequence is crucial when the investigation of a real sequence displays the presence of a high number of repeats. Is this high number expected under some realistic hypothesis of the sequence model? Without a clear answer to this question it is very difficult to assess if the number of repeats observed in the real sequence has a potential biological role because the repeats are over-represented. The set of results we have obtained in this paper could usefully complement the repeat search algorithms to give a measure of the significance of the number detected occurrences.

ACKNOWLEDGMENTS

The authors wish to thank Rosario Mantegna and Salvatore Micciché for useful discussions. The authors acknowledge financial support from the NEST-DYSONET 12911 EU project.

APPENDIX A

In this appendix we derive Eq. (16) for the number of repeats with stem length ℓ and one gap.

There are $\ell-1$ possible positions for the gap in one arm. Let us call A_i, $(i=1,\ldots,\ell-1)$ the set of structures in which the gap has the ith position (see Fig. 2 for the case $\ell=3$). This ensemble of sets has the property that for any set of indices $i_1<i_2<\cdots<i_k$ it is

$$P(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k})=P(A_{i_1}\cap A_{i_k}). \tag{A1}$$

In fact if the sequence under consideration can form a structure with the gap both in the i_1 and the i_k position, then it can form the structure with the gap in any intermediate position.

We state the following theorem.

Theorem. Given an ensemble of sets $A_1,A_2,\ldots,A_N$ satisfying the property (A1), it holds

$$P(A_1\cup A_2\cup\ldots\cup A_N)=\sum_{i=1}^{N}P(A_i)-\sum_{i=1}^{N-1}P(A_i\cap A_{i+1}). \tag{A2}$$

In order to prove this theorem we need a lemma.

Lemma. Under the above hypothesis (A1), it is

$$P[\cup_{i=1}^{n}(A_i\cap A_{n+1})]=P(A_n\cap A_{n+1}). \tag{A3}$$

In fact

$$\begin{aligned} P[\cup_{i=1}^{n}(A_i \cap A_{n+1})] &= P\{[\cup_{i=1}^{n-1}(A_i \cap A_{n+1})] \cup (A_n \cap A_{n+1})\} \\ &= P\{[\cup_{i=1}^{n-1}(A_i \cap A_{n+1})] + P(A_n \cap A_{n+1})\} - P\{[\cup_{i=1}^{n-1}(A_i \cap A_{n+1})] \cap (A_n \cap A_{n+1})\} \\ &= P\{[\cup_{i=1}^{n-1}(A_i \cap A_{n+1})] + P(A_n \cap A_{n+1})\} - P\{\cup_{i=1}^{n-1}[(A_i \cap A_{n+1}) \cap (A_n \cap A_{n+1})]\}, \end{aligned} \tag{A4}$$

where we have used the inclusion-exclusion principle. By using 2 times the property (A1) we can rewrite

$$\begin{aligned} &P[\cup_{i=1}^{n-1}(A_i \cap A_{n+1})] + P(A_n \cap A_{n+1}) - P\{\cup_{i=1}^{n-1}[(A_i \cap A_{i+1} \cap \ldots \cap A_n \cap A_{n+1}) \cap (A_n \cap A_{n+1})]\} \\ &\quad = P[\cup_{i=1}^{n-1}(A_i \cap A_{n+1})] + P(A_n \cap A_{n+1}) - P[\cup_{i=1}^{n-1}(A_i \cap A_{i+1} \cap \ldots \cap A_n \cap A_{n+1})] \\ &\quad = P[\cup_{i=1}^{n-1}(A_i \cap A_{n+1})] + P(A_n \cap A_{n+1}) - P[\cup_{i=1}^{n}(A_i \cap A_{n+1})] \\ &\quad = P(A_n \cap A_{n+1}), \end{aligned} \tag{A5}$$

i.e., our thesis.

We can now prove the theorem. We prove it by induction. The theorem holds for $N=2$, because in this case Eq. (A2) is equivalent to the inclusion-exclusion principle. We assume that Eq. (A2) holds for N and we prove that it holds for $N+1$. In fact,

$$\begin{aligned} P(\cup_{i=1}^{N+1}A_i) &= P(\cup_{i=1}^{N}A_i \cup A_{N+1}) \\ &= P(\cup_{i=1}^{N}A_i) + P(A_{N+1}) - P[(\cup_{i=1}^{N}A_i) \cap A_{N+1}] \\ &= P(\cup_{i=1}^{N}A_i) + P(A_{N+1}) - P[\cup_{i=1}^{N}(A_i \cap A_{N+1})] \\ &= P(\cup_{i=1}^{N}A_i) + P(A_{N+1}) - P(A_N \cap A_{N+1}) \\ &= \sum_{i=1}^{N} P(A_i) - \sum_{i=1}^{N-1} P(A_i \cap A_{i+1}) + P(A_{N+1}) \\ &\quad - P(A_N \cap A_{N+1}) \\ &= \sum_{i=1}^{N+1} P(A_i) - \sum_{i=1}^{N} P(A_i \cap A_{i+1}), \end{aligned} \tag{A6}$$

i.e., our thesis. For the benefit of the reader we note that in the second equivalence we use the inclusion-exclusion principle, in the fourth we use the lemma, and in the fifth we use the induction hypothesis, i.e., that the thesis holds for N.

In the case of repeats considered in the paper it is $N=\ell-1$ and $P(A_i)=(1-q)^{\alpha}q^{\ell}$. Moreover for any i it is $P(A_i \cap A_{i+1})=(1-q)^{\alpha}q^{\ell-1}\tilde{q}$. From these values and the theorem [i.e., Eq. (A2)], Eq. (16) holds.

APPENDIX B

In this section we derive the expression (17) for the expected number of perfect inverted and mirror repeats in a Markovian genome.

Let us indicate the left-hand part of the stem with $n_1 n_2 \cdots n_\ell$, and consequently, the right-hand part of the stem will be $\bar{n}_\ell \cdots \bar{n}_2 \bar{n}_1$, where the bar indicates matching accordingly to the type of investigated repeats. We shall also indicate with $m_1 \cdots m_m$ the loop and with x_1 (x_2) the base before (after) the repeat. The repeat can be symbolically expressed as $x_1 n_1 n_2 \cdots n_\ell m_1 \cdots m_m \bar{n}_\ell \cdots \bar{n}_2 \bar{n}_1 x_2$. The probability for such a structure is

$$\begin{aligned} &p(x_1)p(n_1|x_1)p(n_2|n_1)\cdots p(m_1|n_\ell)p(m_2|m_1)\cdots p(\bar{n}_\ell|m_m)\cdots \\ &\quad p(\bar{n}_1|\bar{n}_2)p(x_2|\bar{n}_1). \end{aligned} \tag{B1}$$

Since we are not interested in the specific bases in x_1 and x_2 we can sum the probability in Eq. (B1) in x_1 and x_2 requiring that they are not complementary (remember that we are looking for *maximal* repeats). The expression becomes

$$\begin{aligned} &p(n_2|n_1)\cdots p(m_1|n_\ell)p(m_2|m_1)\cdots p(\bar{n}_\ell|m_m)\cdots p(\bar{n}_1|\bar{n}_2) \\ &\quad \times \sum_{x_1 \neq \bar{x}_2} p(x_1)p(n_1|x_1)p(x_2|\bar{n}_1). \end{aligned} \tag{B2}$$

The sum term in Eq. (B2) becomes

$$\begin{aligned} &\sum_{x_1 \neq \bar{x}_2} p(x_1)p(n_1|x_1)p(x_2|\bar{n}_1) \\ &\quad = p(n_1) - \sum_{x=1}^{4} p(x)p(n_1|x)p(\bar{x}|\bar{n}_1), \end{aligned} \tag{B3}$$

where we have used the property $\Sigma_{x=1}^{4}p(x|y)=1$.

In expression (B1) we need to sum over the possible loops, i.e., in the variables $m_1,\ldots,m_m$, by using the constraint $m_1 \neq \bar{m}_m$. We sum first over the internal bases of the loop $m_2,\ldots,m_{m-1}$ obtaining

$$\begin{aligned} &p(m_1|n_\ell)p(\bar{n}_\ell|m_m) \\ &\quad \times \sum_{m_2,\ldots m_{m-1}} p(m_2|m_1)p(m_3|m_2)\cdots p(m_m|m_{m-1}) \\ &\quad = p(m_1|n_\ell)p_{m-1}(m_m|m_1)p(\bar{n}_\ell|m_m), \end{aligned} \tag{B4}$$

where $p_k(b|a)$ is the k-step transition probability, i.e., the probability of having the symbol b conditioned to the fact that the k step before the symbol was a. For Markov chain the k-step transition probability matrix is easily obtained as the kth power of the one-step transition probability matrix. By obtaining Eq. (B4) we have used the Chapman-

Kolmogorov equation, that in its simpler form is $\Sigma_{z=1}^{4} p(y|z)p(z|x)=p_2(y|x)$.

Last we need to sum the expression (B4) over the variables m_1 and m_m by imposing that they are not complementary. By using again the Chapman-Kolmogorov equation we obtain

$$\sum_{m_1 \neq \bar{m}_m} p(m_1|n_\ell)p_{m-1}(m_m|m_1)p(\bar{n}_\ell|m_m) = p_{m+1}(\bar{n}_\ell|n_\ell) - \sum_{y=1}^{4} p(\bar{n}_\ell|y)p_{m-1}(y|\bar{y})p(\bar{y}|n_\ell). \quad \text{(B5)}$$

By setting all the terms together we finally obtain

$$\left(p(n_1) - \sum_{x=1}^{4} p(x)p(n_1|x)p(\bar{x}|\bar{n}_1)\right)p(n_2|n_1)\cdots p(n_\ell|n_{\ell-1}) \times [p_{m+1}(\bar{n}_\ell|n_\ell) - \sum_{y=1}^{4} p(\bar{n}_\ell|y)p_{m-1}(y|\bar{y})p(\bar{y}|n_\ell)] \times p(\bar{n}_{\ell-1}|\bar{n}_\ell)\cdots p(\bar{n}_1|\bar{n}_2) \quad \text{(B6)}$$

that can be simplified by noting that $p(n_1)p(n_2|n_1)\cdots p(n_\ell|n_{\ell-1})=p(n_1n_2\cdots n_\ell)$ is the probability of the ℓ word of the left-hand part of the stem. Likewise, $p(\bar{n}_{\ell-1}|\bar{n}_\ell)\cdots p(\bar{n}_1|\bar{n}_2)=p(\bar{n}_\ell\cdots\bar{n}_2\bar{n}_1)/p(\bar{n}_\ell)$ is proportional to the probability of the ℓ word of the right-hand part of the stem. Hence, the probability of a repeat with a specified sequence in the stem is

$$p(n_1n_2\cdots n_\ell)p(\bar{n}_\ell\cdots\bar{n}_2\bar{n}_1)\frac{\left(p(n_1) - \sum_{x=1}^{4} p(x)p(n_1|x\right)p(\bar{x}|\bar{n}_1)\left(p_{m+1}(\bar{n}_\ell|n_\ell) - \sum_{y=1}^{4} p(\bar{n}_\ell|y)p_{m-1}(y|\bar{y})p(\bar{y}|n_\ell)\right)}{p(n_1)p(\bar{n}_\ell)}. \quad \text{(B7)}$$

On the other hand, it is easy to see that the corresponding expression for an i.i.d. sequence is

$$P_{i.i.d.} = p(n_1n_2\cdots n_\ell)p(\bar{n}_\ell\cdots\bar{n}_2\bar{n}_1)\left(1 - \sum_{x=1}^{4} p(x)p(\bar{x})\right)^2. \quad \text{(B8)}$$

It is direct to show that Eq. (B7) reduces to Eq. (B8) when all the transition probabilities satisfy $p(x|y)=p(x)$, i.e., the process has no memory and becomes i.i.d.

In order to obtain the number of repeats of stem length ℓ and loop length m one needs to sum Eq. (B7) over the 4^ℓ possible ℓ words composing the left-hand part of the stem, i.e.,

$$P_{\text{Markov}}(\ell,m) = \sum_{n_1,n_2,\ldots,n_\ell=1}^{4} p(n_1n_2\cdots n_\ell)p(\bar{n}_\ell\cdots\bar{n}_2,\bar{n}_1) \times \frac{\left(p(n_1) - \sum_{x=1}^{4} p(n_1|x)p(\bar{x}|\bar{n}_1)\right)\left(p_{m+1}(\bar{n}_\ell|n_\ell) - \sum_{y=1}^{4} p(\bar{n}_\ell|y)p_{m-1}(y|\bar{y})p(\bar{y}|n_\ell)\right)}{p(n_1)p(\bar{n}_\ell)}, \quad \text{(B9)}$$

which is the result of Eq. (17).

[1] S. R. Eddy, Nat. Rev. Genet. **2**, 919 (2001).
[2] R. R. Sinden, *DNA Structure and Function* (Academic, San Diego, 1994).
[3] Y. D. Carafa, E. Brody, and C. Thermes, J. Mol. Biol. **216**, 835 (1990).
[4] E. A. Lesnik, R. Sampath, H. B. Levene, T. J. Henderson, J. A. McNeil, and D. J. Ecker, Nucleic Acids Res. **29**, 3583 (2001).
[5] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschi, Science **294**, 853 (2001).
[6] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel, Science **294**, 858 (2001).
[7] R. C. Lee and V. Ambros, Science **294**, 862 (2001).
[8] A. J. Hamilton and D. C. Baulcombe, Science **286**, 950 (1999).
[9] G. Hutvágner *et al.*, Science **293**, 834 (2001).
[10] G. P. Schroth and P. Shing Ho, Nucleic Acids Res. **23**, 1977 (1995).
[11] R. Cox and S. M. Mirkin, Proc. Natl. Acad. Sci. U.S.A. **94**, 5237 (1997).
[12] F. Lillo, S. Basile, and R. N. Mantegna, Bioinformatics **18**, 971 (2002).
[13] M. Spanò, F. Lillo, S. Micciche, and R. N. Mantegna, Fluct.

Noise Lett. **5**, L193 (2005).
[14] S. M. Mirkin and M. D. Frank-Kamenetskii, Annu. Rev. Biophys. Biomol. Struct. **23**, 541 (1994).
[15] P. Rice, I. Longden, and A. Bleasby, Trends Genet. **16**, 276 (2000).
[16] P. E. Warburton, J. Giordano, F. Cheung, Y. Gelfand, and G. Benson, Genome Res. **14**, 1861 (2004).
[17] M. Y. Leung *et al.*, J. Comput. Biol. **12**, 331 (2005).
[18] D. S. H. Chew, K. P. Choi, and M. Y. Leung, Nucleic Acids Res. **33**, e134 (2005).
[19] G. Reinert, S. Schbath, and M. S. Waterman, J. Comput. Biol. **7**, 1 (2000).
[20] C.-K. Peng *et al.*, Nature (London) **356**, 168 (1992).
[21] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[22] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[23] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994).
[24] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **51**, 5084 (1995).
[25] G. Bernardi *et al.*, Science **228**, 953 (1985).
[26] J. Beran, *Statistics for Long-Memory Processes* (Chapman and Hall, New York, 1994).
[27] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J. F. Muzy, and A. Arneodo, Phys. Rev. Lett. **86**, 2471 (2001).
[28] D. Holste, I. Grosse, and H. Herzel, Phys. Rev. E **64**, 041917 (2001).
[29] W. Li and D. Holste, Phys. Rev. E **71**, 041910 (2005).
[30] P. Carpena, P. Bernaola-Galvan, A. V. Coronado, M. Hackenberg, and J. L. Oliver, Phys. Rev. E **75**, 032903 (2007).
[31] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993).
[32] P. Bernaola-Galván, P. Carpena, R. Román-Roldán, and J. L. Oliver, Gene **300**, 105 (2002).