DNA viewed as an out-of-equilibrium structure

A. Provata,[1,2,*] C. Nicolis,[3,†] and G. Nicolis[2,‡]

[1] *Institute of Nanoscience and Nanotechnology, National Center for Scientific Research "Demokritos", 15310 Athens, Greece*
[2] *Interdisciplinary Center for Nonlinear Phenomena and Complex Systems, Université Libre de Bruxelles,*
Campus Plaine, CP. 231, 1050 Bruxelles, Belgium
[3] *Institut Royal Météorologique de Belgique, 3 Avenue Circulaire, 1180 Bruxelles, Belgium*

(Received 14 September 2013; published 5 May 2014)

The complexity of the primary structure of human DNA is explored using methods from nonequilibrium statistical mechanics, dynamical systems theory, and information theory. A collection of statistical analyses is performed on the DNA data and the results are compared with sequences derived from different stochastic processes. The use of χ^2 tests shows that DNA can not be described as a low order Markov chain of order up to $r = 6$. Although detailed balance seems to hold at the level of a binary alphabet, it fails when all four base pairs are considered, suggesting spatial asymmetry and irreversibility. Furthermore, the block entropy does not increase linearly with the block size, reflecting the long-range nature of the correlations in the human genomic sequences. To probe locally the spatial structure of the chain, we study the exit distances from a specific symbol, the distribution of recurrence distances, and the Hurst exponent, all of which show power law tails and long-range characteristics. These results suggest that human DNA can be viewed as a nonequilibrium structure maintained in its state through interactions with a constantly changing environment. Based solely on the exit distance distribution accounting for the nonequilibrium statistics and using the Monte Carlo rejection sampling method, we construct a model DNA sequence. This method allows us to keep both long- and short-range statistical characteristics of the native DNA data. The model sequence presents the same characteristic exponents as the natural DNA but fails to capture spatial correlations and point-to-point details.

DOI: 10.1103/PhysRevE.89.052105 PACS number(s): 02.50.Ey, 89.75.Fb, 87.18.Wd, 05.70.Ln

I. INTRODUCTION

The DNA molecule is one of the most complex systems encountered in nature. By its intricate, aperiodic structure it constitutes an information source for the synthesis of the different entities and for the occurrence of the multitude of delicately balanced processes within living cells. Yet, the connection between global DNA structure and its various functions remains to a large extent elusive, particularly in view of the coexistence of coding and noncoding regions and the realization of the important role of noncoding sequences in higher organisms [1–3].

One view of the DNA molecule frequently adopted in the literature is that of two nested nonoverlapping symbolic sequences, the coding and noncoding regions, each of which is expressed in terms of the four-symbol (-letter) alphabet corresponding to the four bases A, C, G, and T. Alternative expressions are provided by coarse-grained, two-letter alphabets. A widely used coarse graining of this kind is the AG-CT one. It is based on the chemical resemblance of the two purine (A, G) and pyrimidine (C, T) units, suggesting that each couple might originate from some primitive "ancestor" unit. A different coarse graining is provided by the AT-CG couple, based on the observation that (A, T) are grouped together to form a weak H-bond group, whereas (C, G) form a strong H-bond group. Its merit is to account for the compositional patchiness of CG content observed in human and more generally vertebrate DNA sequences [4]. The observed complexity of these nested sequences has been shaped during evolutionary time based on functional needs. Processes such as single nucleotide mutations, insertion and deletion of segments, multiple repetitions of elements acting simultaneously over different length and time scales have shaped the complexity of current day genomes producing intriguing statistical properties [3–9]. In this latter setting, early investigations have shown that the succession of bases along coding regions in higher organisms presents short-range correlations, whereas noncoding regions exhibit long-range correlations [10–12]. For these organisms, the coding segment length distribution has an exponentially falling tail, whereas the noncoding segment one falls off as a power law [13,14].

In this work, the structure of DNA, viewed as a symbolic sequence, is analyzed from the standpoint of nonequilibrium statistical mechanics, dynamical systems theory, and information theory. A first question raised concerns spatial asymmetry along the sequence, its signatures, and its role in information processing. A second question pertains to the identification and analysis of global indicators of the underlying complexity, beyond the linear correlations usually considered in the literature.

The above questions will be discussed using both coarse-grained two-letter and full four-letter representations of the genome. To arrive at a quantitative formulation, we view a DNA chain as the realization of a stationary stochastic process, i.e., a process where the joint probability distributions of the sequences generated remain invariant when shifted along the chain. As a corollary, the probabilities p_i of the individual states (symbols) attain rapidly limiting values as the sample size is increased. Here, the role of time step in the traditional setting of stochastic theory is played by a shift in space.

[*] aprovata@chem.demokritos.gr
[†] cnicolis@oma.be
[‡] gnicolis@ulb.ac.be

The data along with the results of a preliminary statistical analysis are compiled in Sec. II. Section III is devoted to a Markov chain analysis, leading to the conclusion that the data can not be fitted by a low order Markov chain, in accord with previous studies. The analysis here goes up to order 6. In Sec. IV, probability fluxes are evaluated and shown to be significantly different from zero in the four-letter alphabet, reflecting the breakdown of (generalized) detailed balance type conditions. The major conclusion drawn from this analysis is the presence of a systematic spatial asymmetry along the symbolic sequence defined by the DNA chain. In Sec. V, this analysis is complemented by the evaluation of a series of entropy and informationlike quantities, leading to interesting characterizations of the dynamical complexity as one advances along the original chain and along its reverse and of the information transfer between different parts of the chain. Exit distance and recurrence distance distributions, two global complexity indicators of special significance, are computed from the data and analyzed in Sec. VI. The existence of long tails in the distributions and of long-range correlations in the associated lengths is established in accord with previous results on long-range correlations in DNA sequences and is confirmed further by the evaluation of Hurst exponents. Building on the information provided by the exit distance distributions, a construction algorithm of a “model DNA” possessing the same statistical properties as the natural one, free of extra assumptions, is outlined in Sec. VII. Different criteria for comparing model and natural DNAs are also developed. The main conclusions are summarized in Sec. VIII.

II. DATA AND STATISTICAL ANALYSIS

For the needs of our analysis, we have employed the genomic data from two large human chromosomes (10 and 14) and two of the smaller ones (20 and 22). In the sequel, we frequently use as working data set a long contig in chromosome 20 of the *Homo sapiens* genome. This genomic contig is the locus N1_011387 (primary assembly) and contains 26 259 569 base pairs (bps), while the entire chromosome 20 contains $\sim 63 \times 10^6$ bps. This represents more than one third of chromosome 20 in a single sequence. This contig is a DNA entity long enough to ensure good statistics, when addressing both the short- and long-range spatial properties. Moreover, it is representative of the entire DNA molecule since it contains both coding and noncoding sequences and other functional elements in similar densities as for all other human chromosomes. In particular, the nucleotide frequencies for the contig are $p_A = 0.289\,341$, $p_C = 0.208\,691$, $p_G = 0.209\,448$, and $p_T = 0.292\,519$. Occasionally, unknown bps denoted by N, which still resist in today’s sequencing techniques, are found in genomes. The N percentage is very small and does not contribute significantly to the statistics. We can then choose either to eliminate all N’s or to replace them randomly with one of the other four {A,C,G,T}. For both choices, the presented results are indistinguishable, up to insignificant statistical errors. Very similar nucleotide frequencies are found in the other human chromosomes. The empirical frequencies of the four base pairs are not constant throughout the genome but vary locally on windows of constant length shifted along the sequences, depending on evolutionary factors and on the presence (or absence) of functional units. For example, the presence of the CpG dinucleotide is associated with the presence of isochores, DNA regions with high density of gene-coding regions. This feature seems to be captured by the AT-CG alphabet, where the frequencies $p_{AT} = 0.581\,861$ and $p_{CG} = 0.418\,139$ are found to differ substantially. In contrast, coarse graining the alphabet at the AG-CT level, the frequencies become very close: $p_{AG} = 0.498\,789$ and $p_{CT} = 0.501\,211$. Thus, information on possible presence of isochores faints. Alternatively, by refining the alphabet, for example by considering explicitly the frequencies of doublets, triplets, etc., information on finer and finer scales emerge, which can not be adherent from superpositions of previous levels of observation. This is one of the main elements which leads one to characterize these molecules as complex since different levels of complexity appear when varying the scale of observation.

For conciseness, we denote from now on (A, G) and (C, T) as states 1 and 2, respectively, in the two-letter purine-pyrimidine alphabet [or (A, T) and (C, G) as states 1 and 2, respectively, in the alternative two-letter alphabet] and (A, C, G, T) as states 1, 2, 3, 4, respectively, in the four-letter alphabet. To obtain the conditional probabilites, we compute first the probabilities $\mathcal{P}(i,j)$ to find the doublets ij (in this precise order) in the sequence and the single letter probabilities $\mathcal{P}(i)$. In Table I, the two-letter and four-letter conditional probabilities

$$w_{ji} = W(j|i) = \frac{\mathcal{P}(i,j)}{\mathcal{P}(i)} \tag{1}$$

obtained by counting the frequencies of adjacent bps i and j (symbol j following symbol i) averaged over the entire sequence are provided. Whereas W in the AG-CT case is nearly symmetric, it is markedly asymmetric in the AT-CG and the four-letter cases. In particular, w_{32}, the probability of encountering G after C, is noticeably smaller than the others. This difference is well known in the biology literature and is attributed to the specific regulatory function of the CpG complex in the human genome, being an essential structural element of the promoters. In spite of such differences, all w_{ij}’s keep statistically significant values (nonzero values), suggesting that no configuration of dinucleotides is excluded. In the language of the theory of stochastic processes, this is a signature of the property of ergodicity, inasmuch as time increments are here replaced by spatial shifts along the structure. Higher order probabilities are obtained in a similar way (available upon request).

III. MARKOV CHAIN ANALYSIS

In the preceding section, we have drawn inferences about probabilities of various orders from a long, unbroken data set. These data are viewed as defining the states of an underlying system at points n along the sequence, the succession of which is supposed to be governed by a set of probability laws. A natural question that comes then to mind relates to the type of stochastic process defined by these laws. As a reference, and to set the stage for what will follow later on, we briefly present in this section strong evidence that the data can in no way be expressed for all practical purposes as a low order Markov

TABLE I. Conditional probabilities w_{ij} as obtained from the DNA data.

		AG-CT		
Two-letter alphabet:	$w_{11} = 0.559\,964$	$w_{12} = 0.437\,910$	$w_{21} = 0.440\,036$	$w_{22} = 0.562\,090$
		AT-CG		
Two-letter alphabet:	$w_{11} = 0.559\,836$	$w_{12} = 0.612\,510$	$w_{21} = 0.440\,164$	$w_{22} = 0.387\,490$
Four-letter alphabet:	$w_{11} = 0.324\,143$	$w_{12} = 0.353\,3467$	$w_{13} = 0.286\,9938$	$w_{14} = 0.210\,9355$
	$w_{21} = 0.173\,548$	$w_{22} = 0.259\,233$	$w_{23} = 0.210\,515$	$w_{24} = 0.206\,090$
	$w_{31} = 0.245\,801$	$w_{32} = 4.575\,338 \times 10^{-2}$	$w_{33} = 0.259\,181$	$w_{34} = 0.254\,663$
	$w_{41} = 0.256\,508$	$w_{42} = 0.341\,667$	$w_{43} = 0.243\,310$	$w_{44} = 0.328\,311$

chain. This point has been addressed in the past by several authors [15–19]. The reason it is taken up again here is first that we dispose of data sets that are more extended and more reliable; second, that the type of test that we propose to apply has a firm theoretical foundation going back to the pioneering work of Billingsley [20]; and third, that it is applied to different alphabets.

A stochastic process $\{i_n\}$, in the form of a sequence of size n, is a Markov chain of order r if the conditional probability

$$W\,(i_n|i_1,i_2,\ldots,i_{n-1}) = \frac{\mathcal{P}\,(i_1,i_2,\ldots,i_n)}{\mathcal{P}\,(i_1,i_2,\ldots,i_{n-1})} \tag{2}$$

is independent of i_m for $m < n - r$. As stated in Sec. II, it is understood throughout that all these probabilities are considered as characterizing a stationary process. The simplest setting is that of a first order Markov chain $s = 1$. Suppose that the conditional probability matrix $W(j|i) = w_{ji}$ has been evaluated from some model. We denote by p_i the frequency of symbol i within the sequence of size n, while with p_{ij} we denote the frequency of occurrence of the doublet ij in the same sequence. Estimating the singlet and doublet frequencies np_i and np_{ij} from the data by different independent counts leads then one to test the legitimacy of the model as a first order Markov chain on the basis of the smallness of the differences $p_{ij} - p_i w_{ji}$. A fundamental result in this context is that the random vector (matrix) V_{ij},

$$V_{ij} = (np_{ij} - np_j w_{ij})/(np_j)^{1/2},$$

converges to the normal distribution with covariance matrix determined by w_{ij}. Keeping in mind the independence of the different samples, it follows [20] that the sum $\mathcal{V}$,

$$\mathcal{V} = \sum_{ij} \frac{(np_{ij} - np_j w_{ij})^2}{(np_j w_{ij})}, \tag{3}$$

obeys asymptotically, in the limit of large n, to the χ^2 distribution. This opens the way to testing the order of the Markov chain within a certain confidence interval by χ^2 type tests. These results can be extended rather straightforward to higher order Markov chains.

In many cases, including the problem addressed in this work, one disposes of no reliable model for estimating, *a priori*, the conditional probabilities w_{ij}. The question thus arises as to whether χ^2 type tests for the order of the Markov chain can still be conducted on the sole basis of the data. As suggested in Refs. [21–23], the answer is in the affirmative provided that the following χ^2 tests are used for the hypothesis that the chain is of the order r:

$$\chi^2 = \sum_{i_1,i_2\ldots i_s} \frac{\left[p_{i_1,i_2\ldots i_s} - p_{i_1,i_2\ldots i_{s-1}} W(i_s|i_{s-r}\ldots i_{s-1})\right]^2}{p_{i_1,i_2\ldots i_{s-1}} W(i_s|i_{s-r}\ldots i_{s-1})}, \tag{4}$$

where the W's are estimated from the data as ratios of frequencies of $i_1,i_2\ldots i_{s-1}$ and $i_1,i_2\ldots i_s$, s being the maximum order considered.

To apply this test to our data, we need to prescribe a confidence interval, which we have chosen to be 5%, and compare the corresponding χ^2 value as given in the tables to the value (4) obtained from the data. This requires specifying each time the number of degrees of freedom $\mathcal{F}$, which is related to s,r and the number of states N by

$$\begin{aligned} \mathcal{F} &= \text{number of degrees of freedom} \\ &= N^s - N^{s-1} - (N^r - N^{r-1}). \end{aligned} \tag{5}$$

The order of the process is then estimated to be the smallest value of r which produces a nonsignificant test statistics [15–19].

Tables II and III summarize the results from our data obtained using the χ^2 test for the purine-pyrimidine (AG-CT) and for the four-letter alphabets, respectively. In all cases tested, the χ^2 values obtained by applying (4) to the data turn out to be much larger than the confidence level ones for processes of order up to 6. In other words, the DNA data can not be fitted by a low order Markov chain. Similar conclusions hold for the AT-CG coarse-grained alphabet. For comparison, by simulating a first order Markov chain having the same p_i's and w_{ij}'s as the data and by applying the test leads to a value of $0.182\,155 \times 10^6$ for the first row of Table II and subsequently

TABLE II. χ^2 test (4) for the DNA data in the purine-pyrimidine (AG-CT) alphabet.

Orders compared	$\mathcal{F}$	χ^2 value (4)	χ^2 value at 5% level
0 1	1	$0.391\,189 \times 10^6$	3.84
1 2	2	$0.936\,032 \times 10^5$	5.99
2 3	4	$0.840\,413 \times 10^4$	7.81
3 4	8	$0.244\,684 \times 10^5$	9.48
4 5	16	$0.341\,158 \times 10^5$	11.07

TABLE III. χ^2 test (4) for the DNA data in the four-letter alphabet.

Orders compared	$\mathcal{F}$	χ^2 value (3.3)	χ^2 value at 5% level
0 1	9	$0.143\,217 \times 10^7$	17
1 2	36	$0.361\,137 \times 10^6$	51
2 3	144	$0.165\,965 \times 10^6$	150
3 4	576	$0.227\,638 \times 10^6$	633
4 5	2304	$0.322\,366 \times 10^6$	2417

for the second row to a value $0.392\,362 \times 10^1$, smaller than the χ^2 value of 5.99 at the 5% level.

IV. SPATIAL ASYMMETRY AND PROBABILITY FLUXES

The failure of the Markov property established in the preceding section leads us to search for alternative ways to characterize DNA viewed as a symbolic sequence or, alternatively, as a text written in the four-letter alphabet provided by the four nucleotides or in the restricted AG-CT or AT-CG coarse-grained alphabets. Now, a common syndrome of all languages is irreversibility in the form of spatial asymmetry, i.e., reading a text written in the language from, say, left to right produces a different result from reading it from right to left. In principle, spatial asymmetries in DNA sequences may be expected due to the extensive presence of repetitive elements and to large scale patchiness [8,9]. On the other hand, the second Chargaff parity rule [24–26] states that in a single DNA strand, oligonucleotides are present in equal frequencies with their reverse complements. This rule, recently shown to hold for oligonucleotides of size up to 3 in most organisms (excluding mitochondrial DNA) in coding and noncoding sequences alike [26,27], points on the contrary to the existence of some underlying symmetry. In this section, we address the issue of irreversibility and asymmetry for the DNA from the standpoint of the theory of stochastic processes and nonequilibrium statistical mechanics on the basis of the data summarized in Sec. II.

At the microscopic level of description, irreversibility and asymmetry are associated with the breakdown of the property of detailed balance, i.e., that in a given system the probability of an event leading from an initial state i to a final state j is counteracted by the probability of the reverse event leading from state j to state i. Transposed from the time domain to the one of the DNA symbolic sequence as it unfolds in space, the simplest expression of this property amounts to the joint probability $p(i,n;j,n+1)$ of two states i and j in adjacent positions n and $n+1$ along the chain satisfying the space reversal relation:

$$p(i,n;j,n+1) = p(j,n;i,n+1) \tag{6a}$$

or using the definition of conditional probabilities

$$w_{ji}p_i = w_{ij}p_j. \tag{6b}$$

Here, i,j run from 1 to 4 in the case of the four-letter alphabet defined by the nucleotides and from 1 to 2 for the two-letter AG-CT alphabet. Alternatively, the probability flux $J_{ij}^{(2)}$,

$$J_{ij}^{(2)} = w_{ji}p_i - w_{ij}p_j, \tag{7}$$

vanishes if the detailed balance condition is satisfied. In a similar vein, higher order space reversal conditions involving more than two sites can be introduced, e.g.,

$$p(i,n;j,n+1;k,n+2) = p(k,n;j,n+1;i,n+2) \tag{8a}$$

or, equivalently,

$$w_{kji}p_{ij} - w_{ijk}p_{kj} = p_i w_{ji} w_{kji} - p_k w_{jk} w_{ijk} = 0 \tag{8b}$$

expressing the vanishing of the probability flux

$$J_{ijk}^{(3)} = w_{kji}p_{ij} - w_{ijk}p_{kj}. \tag{9}$$

Notice that for an alphabet of more than two letters, there is more than one probability flux and more than one detailed balance conditions. For instance, in the four-letter alphabet there are six fluxes $J_{ij}^{(2)}$. If the process were first order Markov, these fluxes would be related by the stationarity condition

$$p_j = \sum_i w_{ji}p_i$$

or, using the normalization property $\sum_i w_{ij} = 1$,

$$\sum_i (w_{ji}p_i - w_{ij}p_j) = \sum_i J_{ij}^{(2)} = 0. \tag{10}$$

There would then be only three independent fluxes, say $J_{12}^{(2)}$, $J_{13}^{(2)}$, and $J_{23}^{(2)}$. One may also define composite fluxes, e.g., the flux from state 1 to the pyrimidines (C or T)

$$J_{1,\mathrm{CT}} = (w_{21}p_1 - w_{12}p_2) + (w_{41}p_1 - w_{14}p_4) = J_{12}^{(2)} + J_{14}^{(2)} \tag{11}$$

and the AG-CT flux as

$$\begin{aligned} J_{\mathrm{AG,CT}} &= J_{12}^{(2)} + J_{14}^{(2)} + J_{32}^{(2)} + J_{34}^{(2)} \\ &= J_{1,\mathrm{CT}} + J_{3,\mathrm{CT}}. \end{aligned} \tag{12}$$

The latter would be strictly zero had the process been a first order Markov.

We now proceed to the evaluation of the probability fluxes from the data and to the testing of the detailed balance condition. Table IV summarizes the main result for fluxes $J^{(2)}$ and $J^{(3)}$ in the case of the two-letter alphabets. As can be seen, the fluxes are very small. Actually, they are indistinguishable from those obtained from a random sequence of the same length and with probabilities p_i fitted from the data (not shown). Detailed balance holds therefore true in this case or, to put it differently, there is no overall spatial

TABLE IV. Probability fluxes $J^{(2)}$ and $J^{(3)}$ in the two-letter alphabets.

$J^{(2)}$	AG-CT	$J^{(2)}$	AT-CG
12	2.9802×10^{-8}	12	0.000000
$J^{(3)}$	AG-CT	$J^{(3)}$	AT-CG
112	5.2154×10^{-8}	112	0.000000
212	3.7252×10^{-8}	212	-7.450580×10^{-9}

TABLE V. Probability fluxes $J^{(2)}$ and $J^{(3)}$ in the four-letter alphabet.

$J^{(2)}$		$J^{(3)}$	
12	-2.3525×10^{-2}	123	1.1347×10^{-2}
14	1.2515×10^{-2}	124	7.7443×10^{-3}
32	3.4543×10^{-2}	134	6.6033×10^{-4}
34	-2.3533×10^{-2}	213	-1.2467×10^{-2}
		214	-3.7790×10^{-3}
		234	1.2393×10^{-2}
		312	1.2358×10^{-2}
		314	-3.6969×10^{-3}
		412	7.8714×10^{-4}
		413	7.5745×10^{-3}

asymmetry and irreversibility. For the purine-pyrimidine (AG-CT) alphabet, this is compatible with the symmetry of the associated conditional probability matrix pointed out in Sec. II. Furthermore, in this alphabet the existence of patches of CpG-rich regions is masked. It comes, on the contrary, at first sight, as a surprise in the case of the AT-CG alphabet, for which the corresponding matrix is markedly asymmetric and the CpG patchiness is not smeared out. In actual fact, since in a two-state system there exists only one flux, detailed balance is bound to be satisfied as a corollary of the property of stationarity.

Table V summarizes the results for $J^{(2)}$ and $J^{(3)}$ in the case of the four-letter alphabet. The results are now definitely significant, much larger than those obtained from a random sequence. We conclude that detailed balance does not hold here, in other words, there is an overall irreversibility in the form of spatial asymmetry. This hitherto unnoticed property clarifies further the status of DNA strings as a natural language. Furthermore, as we see in the next section, it is directly related to the role of DNA as an information source.

In the theory of stochastic processes and nonequilibrium statistical mechanics, asymmetry and deviation from detailed balance are attributed to the presence of a global constraint driving the system out of the state of thermodynamic equilibrium. This may, at a first sight, sound in contradiction with the usual view of DNA as a stable molecule in thermodynamic equilibrium with its environment. But, the contradiction is only apparent inasmuch as living matter in general, and DNA in particular as we observe it today, is to be viewed as the outcome of a primordial nonequilibrium evolutionary process that was eventually stabilized in a "fossil" form as a result of the action of local short-ranged intermolecular interactions. Otherwise, the waiting time to see this event happen spontaneously would be exceedingly large owing to the combined effects of detailed balance and of the explosion of the number of possible combinations of the constituting subunits among which only a small subset would possess the desired biological functions [28–30]. In this view, nonequilibrium is manifested at the level of the four-letter alphabet, but does not show up explicitly at the level of the coarse-grained two-letter alphabets. More work is needed to elucidate further the origin of this feature. For instance, it would be interesting to investigate how coarse-grained states of dynamical systems defined by various phase space partitionings respond in a different way to a nonequilibrium constraint acting on the system as a whole.

V. ENTROPY ANALYSIS AND INFORMATION TRANSFER

In this section, we introduce and analyze a set of quantities aiming to characterize the complexity of the DNA symbolic sequence, while accounting for its central role as information source [31–33] as well as for the spatial asymmetry and irreversibility established in the preceding section. The simplest quantity in this family is the information (Shannon) entropy

$$S_I = -\sum_i p_i \ln p_i \tag{13}$$

describing the amount of choice exercised by the information source and the associated uncertainly of the message recipient. By its static character, this quantity does not provide insights on the overall structure of the sequence. To handle this aspect, we divide the sequence into blocks of symbols $i_1, i_2, \ldots, i_n$ of length n and extend Eq. (13), which defines essentially the entropy per symbol, to the entropy per block of symbols over a window of length n:

$$S_n = -\sum_{i_1,i_2,\ldots,i_n} P(i_1,i_2,\ldots,i_n) \ln P(i_1,i_2,\ldots,i_n), \tag{14}$$

where the sum runs over all nonoverlapping windows of size n.

Now, suppose that the source has sent a message in the form of a particular n sequence. What is the probability that the next symbol will be i_{n+1}? Clearly, we are dealing here with a conditional event. The entropy excess associated with the addition of an extra symbol to the right of the n block ("word") is then [29,31]

$$h_n = -\sum_{i_1,i_2,\ldots,i_n,i_{n+1}} P(i_1,i_2,\ldots,i_n) W(i_{n+1}|i_1,i_2,\ldots,i_n) \times \ln W(i_{n+1}|i_1,i_2,\ldots,i_n). \tag{15}$$

The first nontrivial value h of h_n describes the amount of information obtained when one moves along the chain one step ahead of the initial state i_1:

$$h = h_1 = -\sum_{i,j} p_i w_{ji} \ln w_{ji}. \tag{16}$$

Actually, h_n would be n independent and equal to h had the sequence been compatible with the Markov property, which, as shown in Sec. III, is not the case. In spite of this failure, (16) keeps its significance whatever the nature of the process and will be referred to, in the sequel, as the Kolmogorov-Sinai (KS) entropy. To capture the asymmetry property analyzed in Sec. IV, it is also useful to introduce the reverse process in which the order of the states visited is running backwards, and to define the associated KS entropy as

$$h^R = -\sum_{i,j} p_i w_{ji} \ln w_{ij}. \tag{17}$$

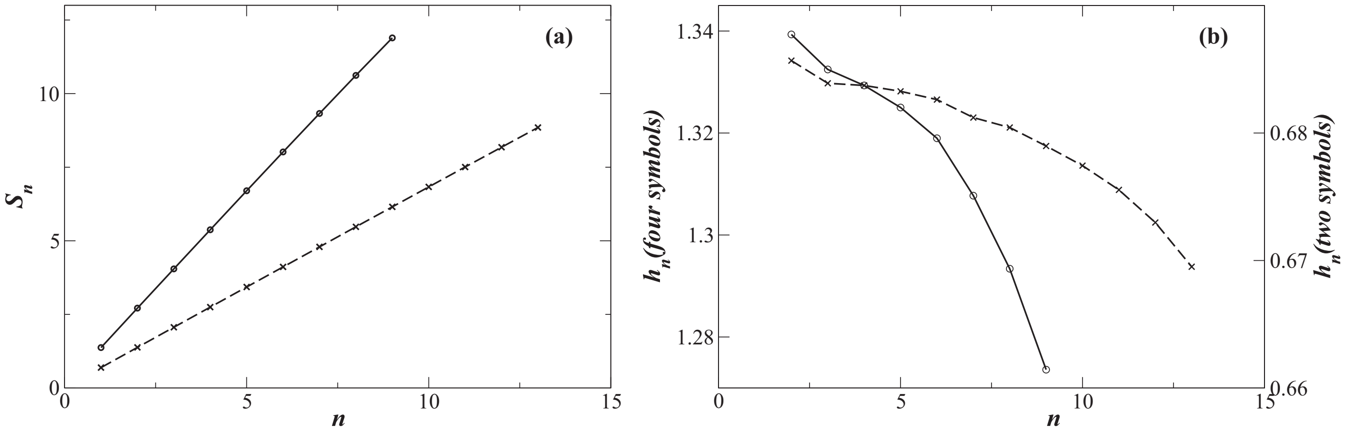

FIG. 1. (a) Information entropy S_n for blocks of size n for the DNA data in the four- and two-letter AG-CT alphabets. (b) Entropy excess h_n for blocks of size n for the DNA data in the four- and the two-letter AG-CT alphabets.

One can easily check that if the Markov property holds, h^R is larger than or equal to h. Indeed,

$$\sigma_I = h^R - h = \sum_{ij} p_i w_{ji} \ln \frac{w_{ji}}{w_{ij}} \tag{18}$$

or using the normalization and stationarity properties discussed in Sec. IV [29,34–36],

$$\sigma_I = h^R - h = \frac{1}{2} \sum_{ij} (w_{ji} p_i - w_{ij} p_j) \ln \frac{w_{ji} p_i}{w_{ij} p_j} \geqslant 0. \tag{19}$$

We can express this property by the statement that the direct sequence is more ordered than the reverse one as long as the probability flux $J_{ij}^{(2)}$ [Eq. (7)] does not vanish, i.e., as long as detailed balance does not hold. For this reason, we will refer to σ_I, which can be regarded as a distance from the regime of detailed balance, as the *information entropy production*. Conversely, in absence of the Markov property but knowing that $J_{ij}^{(2)}$ is different from zero, one may wonder whether h^R is still larger than h. As we see shortly, this is indeed the case for the DNA sequence in the four-letter alphabet.

Let now n be gradually increased. As stated earlier in a Markov process h would remain constant, entailing that S_n would increase linearly in n. Figure 1 depicts the dependence of S_n and h_n as defined from Eqs. (14) and (15) for the DNA data of Sec. II. As can be understood from the n dependence of the information excess h_n [Fig. 1(b)], the S_n versus n dependence is not strictly linear. Indeed, h_n varies (weakly but systematically) with n from a value 1.339 for $n = 1$ to 1.273 for $n = 8$, in the case of the four-letter alphabet (solid lines in Fig. 1). This is in accord with the conclusion drawn in Sec. IV on the non-Markovian character of the sequence and suggests the presence of long-range correlations (see also Sec. VI below).

Table VI summarizes the results of evaluation of h, h^R, and σ_I, using Eqs. (16)–(18), for the four-letter alphabet. For comparison, the corresponding values from a random sequence of the same length are also given. In this case, as expected h and h^R are both equal to the maximum entropy $h = h^R = \ln 4$ of the sequence and $\sigma_I = 0$.

The evaluation of the quantities in Table VI for the DNA sequence in the two-letter purine-pyrimidine (AG-CT) alphabet leads to the quite different conclusion that $h \sim h^R = 0.686$ and thus $\sigma_I \approx 0$, the corresponding h value for the random sequence being $h \sim \ln 2 = 0.693$. On the other hand, S_n and h_n still depend on n in a nontrivial way (see dashed lines in Fig. 1). For the alternative two-letter AT-CG grouping, the corresponding values are $h \sim h^R = 0.678$ and thus $\sigma_I = h^R - h \approx 0$.

On the basis of the above comparison between the two alphabets and between the DNA data and those associated to the random sequence, one is tempted to conclude that revealing the asymmetry of the DNA sequence in the four-letter alphabet, as established already in Sec. IV, has also some interesting signatures at the level of information processing: Information is being produced (in the sense $\sigma_I > 0$) as long as one advances along a preferred direction in sequence space, and this requires reading the "text" in a four-letter alphabet.

An alternative view of the DNA sequence in connection with both the presence of correlations and information processing is to consider two segments, typically of the same length: view the leftmost segment (say x) as the "source" and the second one (say y), as the "receiver" and evaluate the information transfer from x to y. We define this quantity, also referred to as *mutual information*, by [37]

$$I_{x \to y} = \sum_{i,j} P(x_i, y_j) \ln \frac{P(x_i, y_j)}{P(x_i) P(y_j)}. \tag{20}$$

We notice that

$$I_{x \to y} = S_I(y) + \sum_{i,j} P(x_i, y_j) \ln W(y_j | x_i).$$

Furthermore, if x and y are two adjacent sites of the sequence, the second term, which represents the conditional

TABLE VI. Kolmogorov entropy of the direct and reverse sequence and information entropy production as computed from the DNA data.

	h	h^R	σ_I
DNA data	1.339	1.416	0.077
Random sequence	1.373	1.373	4×10^{-7}

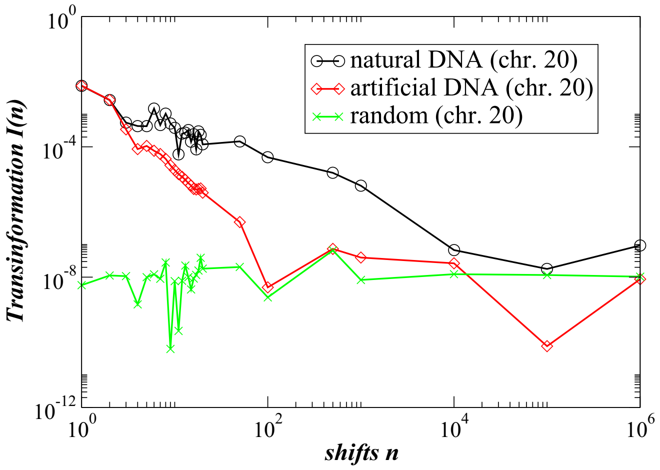

FIG. 2. (Color online) Information transfer $I(n)$ between a sequence and its shift by n symbols versus n (four-letter representation). Line with circles depicts the DNA sequence, line with crosses a random sequence, and line with diamonds a model-generated sequence (Sec. VI).

entropy of y given the state of x, reduces to the KS entropy [Eq. (16)]. In the following analysis, a sequence is compared with its shifts. For a specific shift of n sites, the information transfer represents then the capacity between adjacent symbols to interact down the sequence. The upper line in Fig. 2 depicts the dependence of $I_{x\rightarrow y}$ on n for two sequences of the same length: chromosome 20, working contig N1_011387 (in two-letter representation) and its shift by 1, 2, ... up to 10^5 symbols. The last excess n symbols in the comparison can either be reinjected at the beginning of the sequence, or they can be neglected without changing significantly the resulting I values. For comparison, the lower line (with crosses) stands for the results obtained from random sequences of the same length and bps frequencies as the working contig. The intermediate line (with diamonds) is associated with the model which will be discussed in Sec. VII. As can be seen from the figure, the information transfer as extracted from the DNA data remains higher with respect to the case of a random sequence for shifts up to 100, suggesting, once again, the presence of correlations and information transfer between successive bps up to the order of ∼ 100. At the level of functionality, this nontrivial information transfer may refer to cooperations between successive units related to the presence of codons in the coding regions and to the multiple presence of poly-A's, to frequent appearance of repetitive elements, to the regulatory elements, to the promoters and to other functional elements in the noncoding parts [38,39].

A different view of the presence of correlations in information transfer between two symbol sequences is provided by their Hamming distance, which determines the number of positions at which the corresponding symbols are different or counts the number of substitutions required to change one sequence into the other. The classical Hamming distance H_{1-2} between the two symbol sequences $S1$ and $S2$ of the same length L, as defined by Hamming in 1950 for error detection, is [40]

$$H_{1-2} = \frac{1}{L}\sum_{i=1}^{L} d_i, \quad \text{where}$$
$$d_i = \begin{cases} 0 & \text{if} \quad S1(i) = S2(i), \\ +1 & \text{if} \quad S1(i) \neq S2(i). \end{cases} \tag{21}$$

We shall also use a modified Hamming distance H'_{1-2} between $S1$ and $S2$, based on the particular nucleotide grouping. For example, for the purine-pyrimidine (AG-CT) grouping, the modified Hamming distance is defined as

$$H'_{1-2} = \frac{1}{L}\sum_{i=1}^{L} d_i, \quad \text{where}$$
$$d_i = \begin{cases} 0 & \text{if} \quad S1(i) = S2(i), \\ +0.5 & \text{if} \quad S1(i) \text{ and } S2(i) \text{ are both purines (A or G)}, \\ +0.5 & \text{if} \quad S1(i) \text{ and } S2(i) \text{ are both pyrimidines (C or T)}, \\ +1 & \text{otherwise}, \end{cases} \tag{22}$$

which considers the purine-purine variation as less important than the purine-pyrimidine one and penalizes by 0.5 if the two nucleotides belong to the same group and by 1 if they belong to a different group (similarly for the AT-CG reduction).

In the case of the original Hamming distance, the H distance between the contig sequence and a random one with the same symbol frequencies is $H(\text{contig-random}) = 0.74331$. Note that if the four symbol frequencies were equal, the value would be $\frac{12}{16} = 0.75$. The H value between two random sequences is of the same order $H(\text{random1-random2}) = 0.74329$. The H value between the contig sequence and its shifts shows, on the contrary, interesting correlations similar to the ones demonstrated by the information transfer [see Fig. 3(a)].

In the case of the modified Hamming distance, the H' distance between the contig sequence and a random one with the same symbol frequencies is $H'(\text{contig-random}) = 0.6216$. The H' value between two random sequences are of the same order $H'(\text{random1-random2}) = 0.6220$. Again, the H' value between the contig sequence and its shifts shows interesting correlations, similar to the ones demonstrated by the information transfer and the original H distance [see Fig. 3(b)].

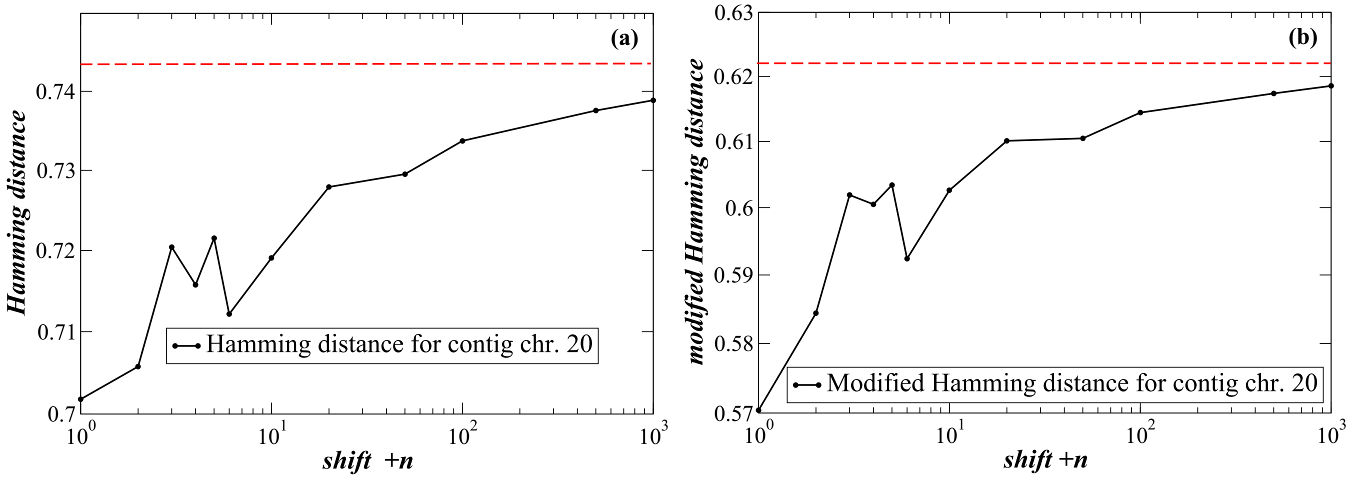

FIG. 3. (Color online) (a) The classical Hamming distance H_{1-2} between the contig sequence and its shifts. The red dashed line denotes the H value between two random sequences with the same symbol frequencies. (b) The modified Hamming distance H'_{1-2} between the contig sequence and its shifts. The red dashed line denotes the H' value between two random sequences with the same symbol frequencies. All sequences are represented in the four-letter alphabet.

Figures 3(a) and 3(b) demonstrate overall that there is a nontrivial information flow between each bps and its neighbors, while this information decreases as the bps become more and more distant on the chain.

VI. EXIT AND RECURRENCE DISTANCE DISTRIBUTIONS

So far, we have been concerned with global properties of the DNA sequences. In this section, we introduce a new set of quantities which allow probing features associated with the local structure. One example of special importance is the appearance of clusters in which a given symbol, or a given subsequence of symbols, is repeated for a certain number of steps, beyond which a transition to different symbols or subsequences is taking place.

To capture such features, we introduce the exit distance distribution, a concept analogous to the exit time distribution familiar from the theory of stochastic processes [41,42]. Distance distributions between single nucleotides and between oligonucleotides have been introduced and studied extensively in the past [43–45]. The exit distance distribution considered in this section provides a general tool to account for different kinds of transitions and for clusters of all orders. It also allows for useful comparisons between different alphabets and different starting states. In a two-letter alphabet, it will actually suffice to determine all the bulk statistical properties of the structure, i.e., the properties not related to spatial correlations. This feature is used in Sec. VII to create artificial DNA sequences sharing the statistical properties of the native ones. In the four-letter alphabet, it will not suffice by itself to fully characterize the structure but will provide, in conjunction with the conditional probability matrix, a very close approximation to its bulk statistical properties (again excluding properties related to spatial correlations).

To construct the exit distance distribution, we start with a certain state or symbol j, and we ask for the probability $q_{j,n}$ that an escape from it occurs at the nth step as one moves along the sequence [46]. If we denote by $\mathcal{P}(j,1)$ the probability to find the current symbol (numbered as 1 and set as the instantaneous origin of the coordinates) in the state j and by $\mathcal{P}(j,1:\overline{j},n)$, the probability to encounter a symbol different from j after n symbols, then $q_{j,n}$, is defined as

$$q_{j,n} = \mathcal{P}(j,1:\overline{j},n)/\mathcal{P}(j,1), \tag{23}$$

where $\overline{j}$ denotes the set of all allowed states with the exception of j.

Closely related to the above is the recurrence distance distribution in which starting again with a state j we ask what is the probability $F_{j,n}$ to encounter this state again n steps down the sequence, with the understanding that the sites between 1 and n are found in states $\overline{j}$, other than j:

$$F_{j,n} = \mathcal{P}(j,1;\overline{j},2;\ldots;j,n)/\mathcal{P}(j,1). \tag{24}$$

For a first order Markov process, both $q_{j,n}$ and $F_{j,n}$ can be evaluated explicitly on the sole basis of the conditional probability matrix $W = \{w_{ij}\}$. Specifically,

$$q_{j,n} = (w_{jj})^{n-1} - (w_{jj})^n \tag{25}$$

and $F_{j,n}$ is expressed in terms of its generating function $\tilde{F}_j(s)$ as

$$\tilde{F}_j(s) = [sW(I-sW)]^{-1}_{jj}\,, \tag{26}$$

where I is the unit matrix. As a corollary, both $q_{j,n}$ and $F_{j,n}$ are superpositions of exponentials in n and thus fall off exponentially for large n. Equations (25) and (26) can be extended rather straightforwardly to the case of a second order Markov process. The calculations are more involved, but the property of exponential decay for large n is again found to hold here.

We now proceed to the evaluation of these distributions from the DNA contig data in Sec. II, starting with $q_{j,n}$. For this purpose, the data are being read along the direct sequence.

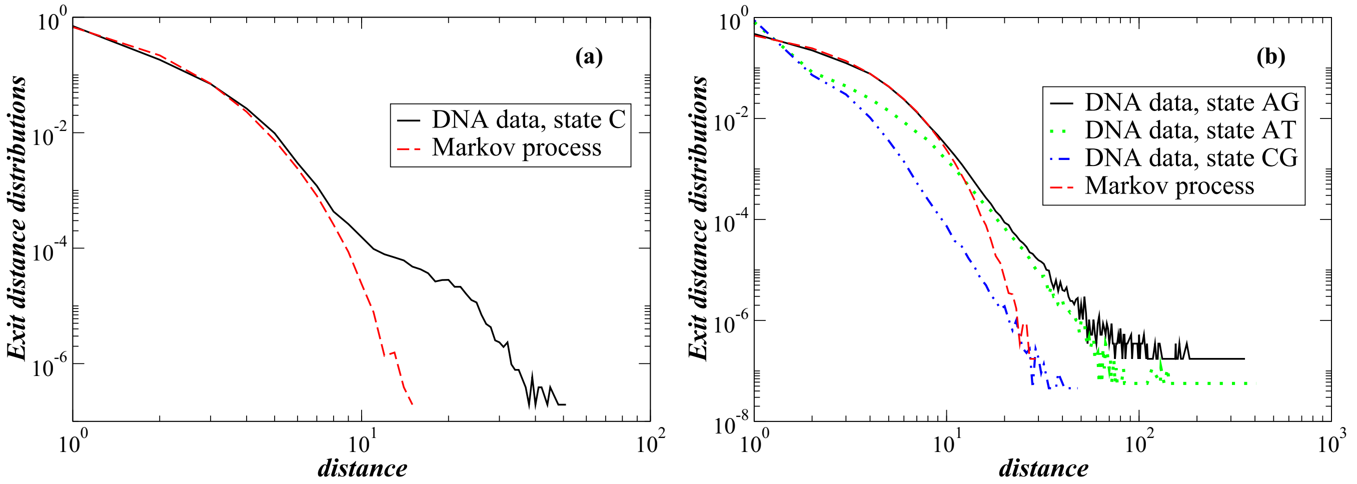

FIG. 4. (Color online) Exit distance distributions: (a) four-letter alphabet, C state; (b) two-letter alphabets, AG state (black solid line), AT state (dotted green line), and CG state (dashed-dotted blue line).

Whenever a state i is first spotted, the origin of coordinates is set on the corresponding site and the distance from the origin is recorded when a state different from i first appears along the sequence. Counting all the distances recorded in this way for each of the states, one arrives at the exit distance distribution. In Figs. 4(a) and 4(b), the distributions for state C and for purines (AG) in the case of the four- and the two-letter AG-CT alphabets, respectively, are depicted. In both cases, we observe the tendency for development of long tails (see also Refs. [47,48]). In particular in the case of the four-letter alphabet, the state C shows a linear region of low slope (~ -2) in the intermediate scales which is soon covered by finite size effects. Interestingly, the exit distributions from states A and T and from states C and G are indistinguishable. Furthermore, the distribution of A and T is longer ranged than the one of C and G (not shown), owing principally to the existence of poly(A) and poly(T) domains found in the human genome [38,39]. The tendency for development of long tails is better detected from comparison with the associated probability for a first order Markov process indicated in the figures by the dashed lines as obtained from a direct simulation of a Markov chain with conditional probabilities equal to those provided by the data.

Coming to the two-letter alphabets, as seen in Fig. 4(b) the distributions starting from the AG state in the AG-TC alphabet or from the AT state in the AT-CG alphabet are very close and are both long ranged, differing significantly from the associated probability for a first order Markov process. In contrast, for the CG state in the AT-CG alphabet, the range is noticeably shorter. This is due, presumably, to the low frequencies of the C and G states themselves and to the scarcity of the CpG combination.

An alternative manifestation of the log-log structure depicted in Figs. 4(a) and 4(b) is that the individual exit distances display long-range correlations as illustrated in Figs. 5(a) and 5(b) for the cases of four and two symbols [45]. If the exit distances are considered as an individual sequence $n_1^i, n_2^i, \ldots, n_R^i$, where n_k^i denotes the distance (in bps units) between the kth and the kth + 1 appearance of the symbol or state i in the original symbol sequence, then the exit distance correlation function $C(r)$ is defined as

$$C^i(r) = \frac{1}{R}\sum_{k=1}^{R}\left(n_k^i - \langle n^i\rangle\right)\left(n_{k+r}^i - \langle n^i\rangle\right), \tag{27}$$

where $\langle n^i\rangle$ denotes the average exit distance for the state i. From Fig. 5, power law decaying exponents close to $\frac{1}{3}$ and to $\frac{1}{2}$ for the four- and the two-symbol AG-CT cases are observed, respectively.

Table VII summarizes the means and variances of the exit distances for the four and for the AG-CT cases as compared to the corresponding quantities evaluated from the Markov chain simulation. We see that while the average values in the two cases are practically indistinguishable [and equal to the analytic results for a first order Markov chain $(1 - w_{ii})^{-1}$], the variances associated to the data are larger than the Markov ones. This reflects the delocalization of the DNA exit distance distribution in the state space, a property due among others to the presence of repeats along the sequence.

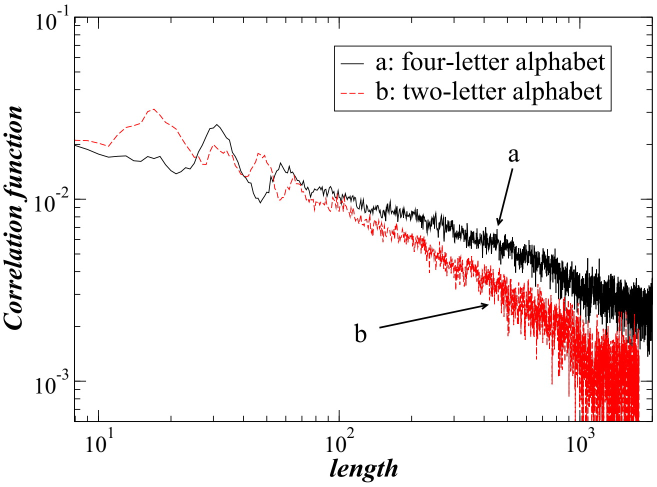

FIG. 5. (Color online) Correlation function: four-letter alphabet (curve a) and two-letter alphabet (curve b).

TABLE VII. Means and variances of exit distances for the DNA data and for a first order Markov chain.

		DNA Data		First order Markov	
		$\langle n\rangle$	$\langle\delta n^2\rangle$	$\langle n\rangle$	$\langle\delta n^2\rangle$
AG-CT alphabet:	1	2.272 5391	3.514 3790	2.271 9333	2.482 8568
	2	2.283 5724	3.726 2988	2.284 3783	2.524 2081
Four-letter alphabet:	1	1.479 6035	0.956 2388	1.479 1129	0.708 8170
	2	1.349 9513	0.459 9166	1.350 2789	0.473 1067
	3	1.349 8576	0.458 8480	1.349 8057	0.472 4476
	4	1.488 7851	0.983 2494	1.488 7587	0.727 2797

We turn next to the recurrence distribution $F_{j,n}$. We first observe that in the two-letter alphabet, the recurrence distribution of one of the two states is fully determined by the exit distance distribution of the other state. We are thus again in the presence of long-ranged distributions and long-range correlations of the individual recurrence distances (see Fig. 5 and the first two rows of Table VII). Coming to the four-letter alphabet, as for $q_{j,n}$, practically identical recurrence distributions for states A and T and for states C and G are observed (see Fig. 6). Both distributions are long ranged with the A and T falling off more slowly than for C and G for large n (and a crossover between the two distributions at $n \sim 5$). Furthermore, compared to the corresponding exit distance distributions, they are more delocalized as illustrated by Table VIII, to be compared with the last four rows of Table VII.

As was the case of Table VII, the averages $\langle n\rangle$ are very close to those obtained by simulating a first order Markov chain with a conditional probability matrix provided by the data, as well as with the well known analytic result $\langle n\rangle = 1/p_i$.

Closely related to recurrence is the concept of analogs, which finds its origin in the classification of atmospheric circulation patterns in meteorology. Translated in the language of the (coarse-grained) description of a symbolic sequence, the issue is to what extent there exist persistent patterns in different (distant) parts along the sequence, where symbols are found in a given prescribed order with an appreciable frequency [49].

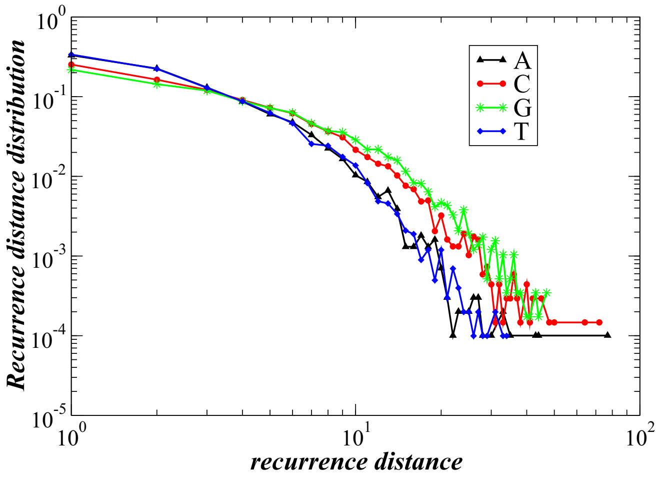

FIG. 6. (Color online) Recurrence distributions of the four symbols.

To address this question for the DNA symbolic sequence, we consider all pairs of n subsequences along the full sequence containing identical symbols in sites $1,\ldots$ up to m, and compute the "error" (in the sense of the Hamming distance, see Sec. V) as they start deviating from the $(m+1)$st site and onwards. The result for the two- and four-letter alphabets and for $n = 100$, $m = 8$ is depicted in Fig. 7(a). The dashed lines in this figure correspond to a random sequence. As expected, beyond $n = 8$ the symbols in the two members of the pair alternate indifferently between being identical (error 0) or being different (error 1), entailing that the error attains immediately a saturation value. The situation is very different for the DNA data, represented by the solid lines in Fig. 7(a). Here, a first stage of abrupt increase of the error is followed by a stage of very slow increase toward the saturation level, even though this level is not yet attained for n up to 100. This indicates a persistence trend or, alternatively, the presence of long-range correlations and is further confirmed by the plot of Fig. 7(b), suggesting a power law dependence of the error on n prior to the final decay to the saturation level with a power of the order of 0.5. This behavior can be viewed as the "spatial" analog of the error growth dynamics familiar from dynamical systems theory where, after an exponential stage [to be compared with the stage of fast growth in Fig. 7(a)], one observes a diffusive stage prior to the final stabilization to the saturation level.

As originally suggested in the early 1990s [10–12], within the set of quantities which probe the local structure of a sequence, the Hurst exponent H expresses the tendency of the future values of a sequence to persist or increase on average, or to fluctuate between small and large values [50]. In particular, for the range $0 \leqslant H < 0.5$, the sequence values tend to alternate, while for $0.5 < H \leqslant 1$, they tend to persist or increase on average. The value 0.5 is a border case, where the values are either completely uncorrelated or their correlations decay exponentially fast to zero.

TABLE VIII. Means and variances of recurrence distances for the DNA data and for a first order Markov chain.

	DNA Data		First order Markov	
	$\langle n\rangle$	$\langle\delta n^2\rangle$	$\langle n\rangle$	$\langle\delta n^2\rangle$
1	3.622 1180	12.571 706	3.627 2037	9.004 858
2	5.101 1710	29.414 993	5.105 8583	20.609 692
3	5.078 6543	29.119 154	5.079 6456	20.368 145
4	3.589 6053	12.082 341	3.592 3173	8.781 096

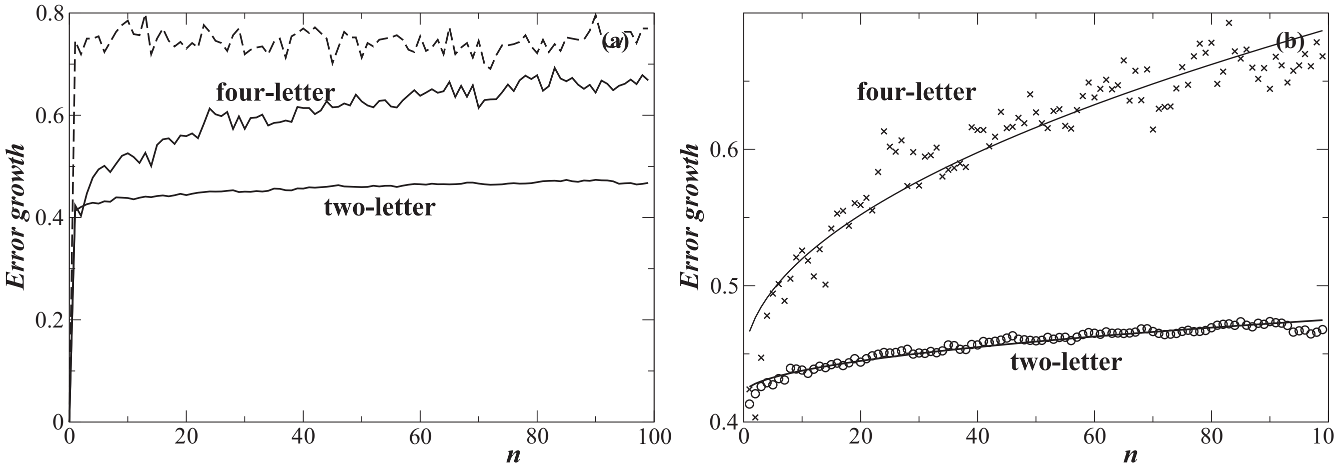

FIG. 7. (a) Error growth functions for the AG-CT alphabet and the four-letter alphabet; the dashed line represents a random sequence. (b) Same as (a) with solid lines depicting nonlinear fit.

To apply the concept of the Hurst exponent in DNA sequences (or any symbol sequence in general), we map the nucleotides (symbols) to numbers. For the two-letter alphabets we present here the calculation for the AG-CT grouping but similar calculations can be conducted for the AT-CG one. The mapping takes the form

$$\begin{aligned}(\text{A} \quad \text{or} \quad \text{G}) &\rightarrow 0,\\ (\text{C} \quad \text{or} \quad \text{T}) &\rightarrow 1.\end{aligned} \tag{28}$$

Thus, the symbol sequence turns into a corresponding numerical sequence, which carries all the information on the position of symbols. H is then directly calculated from the numerical series and is a significant measure which expresses the tendency of symbols to repeat themselves (persistence) or to alternate (antipersistence) down the sequence.

The calculation of the Hurst exponent is based on the computation of the range $R(n)$ between the maximum and the minimum cumulative values as one advances along the numerical sequence of size n, for various values of n. Cumulative values are essential in the H estimation because they keep track of the tendencies along the sequence. One then needs to rescale $R(n)$ by the standard deviation $S(n)$:

$$S(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x\rangle)^2} \tag{29}$$

in order to obtain the *rescaled range*. Once the rescaled range $R(n)/S(n)$ is calculated, it is averaged over many sequences (configurations) of the same length n:

$$E(n) = \left\langle \frac{R(n)}{S(n)} \right\rangle_{\text{confs}}. \tag{30}$$

The Hurst exponent is then defined as

$$E(n) = cn^H \tag{31}$$

and is computed from the slope of $E(n)$ versus n in a double logarithmic scale. When the sequence is characterized by fractality, with fractal dimension D, it can be shown that $H = 2 - D$, where $1 < D < 2$.

In Fig. 8, the rescaled ranges $E(n)$ are plotted as a function of n for the working contig N_011387 of chromosome 20 (solid line), the random sequence (dotted line), and the model DNA (stars) which will be discussed in the following section. The value calculated for the Hurst exponent is $H = 0.6145$ and is clearly distinct from that of the random sequence with the same letter frequency as the data. Calculations of the Hurst exponent in other human contigs give very similar H values. Values of $H > 0.5$ indicate persistence of the same symbols along the sequence, or to put it differently, clustering of similar nucleotides. This effect is a cause of correlations and can reflect the well known existence of poly(A) and poly(T) (in the complementary chain) motifs in the primary genomic DNA sequences that give rise to the corresponding poly(A) signals in mRNA [38]. Another source of the clustering of similar nucleotides is the Alu repeats [51,52] in the human genome which are also known to be associated with poly(A) sequences [39]. In addition, noncoding gene-poor (desert) regions are known to be rich in A and T inducing clustering of these symbols, while CpG-rich regions (isochores) are rich in genes and induce lower scale clustering [53–57].

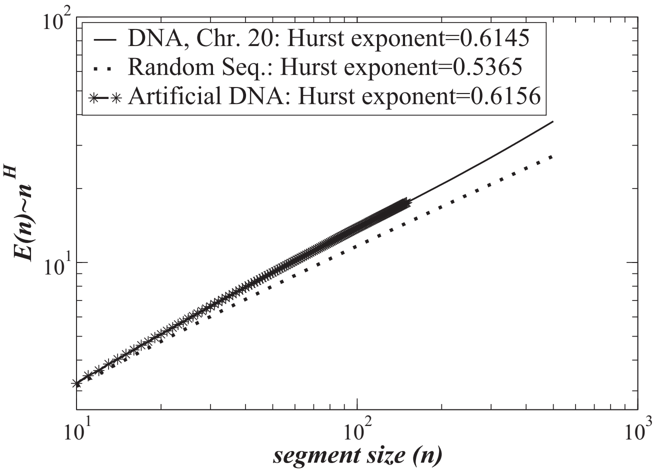

FIG. 8. The rescaled range $E(n)$ as a function of the sequence size n, for the calculation of the Hurst exponent H.

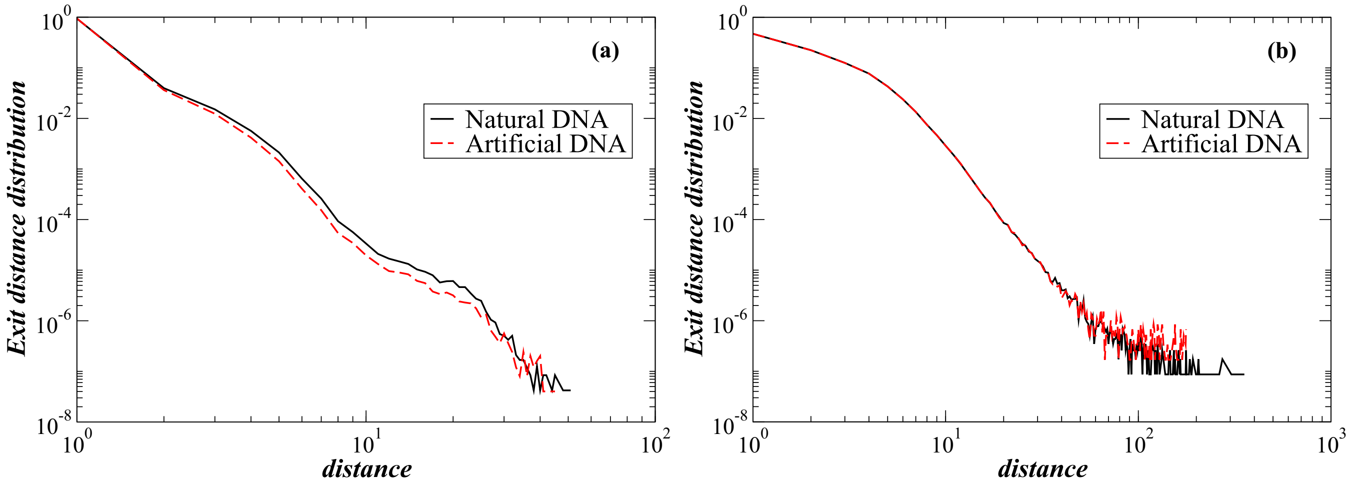

FIG. 9. (Color online) Exit distance distributions for natural (solid black lines) and artificial DNA (red dashed line): (a) four-letter alphabet, A state; (b) two-letter alphabet, AG state (purine).

VII. A MODEL DNA

DNA, a complex multicomponent structure which has evolved during billions of years in close contact with an ever-changing environment, can not be described or constructed based on a closed functional expression with a limited number of parameters. Statistical constructive methods or methods based on chaotic dynamics have been used since the early 1990s to create long nucleotide sequences with statistical properties mimicking those of specific DNA molecules [11,13,58–62]. All these attempts predict well some of the sequence properties but they fail in others and one needs to add an increasing number of parameters to probe into the structure's local details, even from the statistical point of view. In this section, we propose a "null" model of DNA based on a global statistical construction method. The method allows us to generate two- and four-letter sequences with statistical properties as close as possible to the ones of the original DNA data, on the basis of the exit distance distribution of the DNA sequences described in Sec. VI.

The construction method is known as "Monte Carlo rejection sampling," or simply rejection sampling, and dates back to von Neumann. Having calculated the exit distance distributions for the segments of all symbols in the natural DNA sequence, we use the rejection sampling method to create an equivalent model series. For simplicity, the method is described in the two-letter AG-CT alphabet and is easily extendable to the AT-CG and to the four-letter ones:

(1) Define first the initial symbol as an AG or CT, either randomly or as dictated by the contig sequence.

(2) Select an integer random number between $[1,N_{\rm AG}^{\rm max}]$ or $[1,N_{\rm CT}^{\rm max}]$ depending on whether a purine (AG) or a pyrimidine (CT) segment is to be created. ($N_{\rm AG}^{\rm max}$ and $N_{\rm CT}^{\rm max}$ are the maximum numbers of juxtaposed purines or pyrimidines which have been observed in the natural contig.) Call the selected number n.

(3) Choose a second random number $r \in [0,1]$ and compare it to the value of the exit distance distribution $q_{{\rm AG},n}$ or $q_{{\rm CT},n}$ depending on the current state on the chain.

(4) If $r \leqslant q_{{\rm AG},n}$ (or $r \leqslant q_{{\rm CT},n}$), then the sequence is extended by n units of purine (or pyrimidine).

(5) The algorithm returns to step 2 in order to make alternating additions of purine and pyrimidine clusters. More specifically, we always switch symbols when the algorithm passes from step 5 to step 2.

(6) The algorithm stops when the size of the artificially constructed sequence is equal to the size of the natural DNA contig.

By construction, the artificial sequences created with the rejection sampling method produce perfectly the exit distance distributions of the natural sequence. They possess the natural sequence's statistical properties, except those related to spatial arrangement of the clusters, as the bps clusters are placed in a random fashion. In particular, exit from a given state implies automatically entrance to the complementary state in the two-letter alphabet. The situation is different in the four-letter case. Here, one more assumption needs to be made regarding the alternation between the four symbols. Our procedure is based on the transition probabilities w_{ij} between the different letters as were presented in Table I and implies thus the assumption that higher order transition probabilities are not accounted for at this stage.

In Fig. 9, the exit distance distributions for the A symbols (four-letter alphabet) and the AG coarse-grained symbol (two-letter alphabet) are shown, both for the original and the artificial model-based DNA sequences; similar plots are obtained for the other symbols. As the construction was based on the reproduction of the exit distances, the native and model distributions are statistically identical by construction in the case of the two-letter alphabet [Fig. 9(b)], with small differences in the tails of the distribution attributed to the finite size of the sequences. The differences are nontrivial in the case of the four-letter alphabet [Fig. 9(a)], and this is attributed to the use of w_{ij}'s which account only for the pair correlations, while for the juxtaposition of segments of size n, higher order correlations (of range up to n) need to be taken into account.

In Sec. V, Fig. 2, the information transfer I between a sequence and its shifts was shown, both for the original and for the model-generated sequences. Both sequences show the same degree of information transfer in first and second neighbor positions. However, for more distant positions, the information transfer in the model-generated sequence undergoes an abrupt decay as compared with the natural DNA sequence where the information transfer persists for hundreds of units. This difference reflects the functional role of natural DNA sequence, as opposed to the statistical character of the artificial DNA. The nucleotides in a natural sequence need to control the information about neighboring positions since what dictates their functionality is their precise (not statistical) juxtaposition. In particular, information flow in decades of bps relates to the turn of the helix, while information flow in a few hundreds of bps is plausible since these are typical sizes for coding regions and for repetitive elements.

In a similar vein, the analog analysis shows that the error values as obtained from the model lie closer to the saturation level than those obtained from the natural DNA as depicted in Fig. 7. Interestingly, the Hamming and modified Hamming distances between the natural and the model sequences equal to 0.7444 and 0.6222, respectively, and are close to the values associated with distances between random sequences.

Finally, in Sec. VI, Fig. 8, the Hurst exponent H is depicted both for the natural (solid line) and for the model sequence (stars). By its nature, H is a nonlinear measure which takes into account size correlations of all orders and deals simultaneously with all segment sizes. As can be seen from the figure, the curves $E(n)$ for the natural and the model sequences are practically indistinguishable and the values of H are very close. This last result is very interesting and indicates that although the model fails to reproduce the quantities which characterize the DNA sequences pointwise (such as the Hamming distances and information transfer between the sequence and its nearest shifts), it succeeds in reproducing global nonlinear characteristics, such as the Hurst exponent and the details of the distributions.

VIII. CONCLUSIONS

In this study, the structure of global human chromosomal sequences has been analyzed using ideas and tools from nonlinear dynamics, information and complexity theories, and nonequilibrium statistical mechanics. Multiple analyses have been performed on the DNA data and compared with symbol sequences with two- and four-letter alphabets produced by different stochastic processes. In particular, we have shown that in the four-letter alphabet, DNA data exhibit spatial asymmetry and suggested on these grounds that the chromosomes can be viewed as out-of-equilibrium structures. We have established a connection between asymmetry and the processing of information along DNA sequences, using a series of entropylike quantities. We have introduced the exit and recurrence distance distributions, two further indicators of the complexity underlying the sequences, whose evaluation revealed a number of interesting features of their global structure, such as the generation of long-range correlations. Finally, we have designed an algorithm generating sequences that share the statistical properties of natural DNA, local as well as global, other than those related to spatial correlations, on the sole basis of the exit distance distribution. The results reported pertain mostly to human chromosome 20. Other chromosomes have been tested and shown to lead to similar conclusions.

It is worth stressing that while in the four-letter alphabet asymmetry coexists with long-range correlations, it can not be regarded as a prerequisite in the most general case. In fact, in the two-letter alphabets considered in this work, the sequences displayed reversibility (in the sense of detailed balance) for all practical purposes but were still exhibiting clear cut long-range correlations. Analogous situations are encountered in dynamical systems theory. For instance, autonomous Hamiltonian (and thus time-reversible) systems operating in the regime of weak chaos referred to as stochastic web give rise to anomalous diffusion indicative of the presence of long-range correlations in time [63,64].

The approach initiated in this work opens some interesting and worth-exploring perspectives. A first line of approach would be to apply the ideas of asymmetry, irreversibility, and information processing considered here in a global perspective to particular DNA building blocks such as coding DNA, noncoding DNA, repeats, etc. Another case to consider are higher eukaryotes, whose genomes share with human genome the existence of genes separated by long noncoding regions containing a high concentration of repeats. Similarly, in the spirit of comparative genomics, it would be interesting to apply the ideas developed here on organisms with intrinsically different genomic structure such as prokaryotes versus eukaryotes.

Finally, a quantitative comparison between the local and global statistical properties of the human genome derived in this work and those of the genome of higher mammals and especially of primates could lead to striking evolutionary insights.

ACKNOWLEDGMENTS

A.P. acknowledges financial support from the National Center for Scientific Research “Demokritos” for a Sabbatical visit to the Université Libre de Bruxelles and the European Science Foundation programme “Exploring the Physics of Small Devices” for the scientific exchange Grant No. EPSD-4308 with the same university. We also acknowledge an interesting discussion with Professor M. G. Velarde.

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell* (Garland Sciences, New York, 2007).

[2] I. Dunham, A. Kundaje, S. F. Aldred *et al.*, Nature (London) **489**, 57 (2012).

[3] A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. d’Aubenton-Carafa, and C. Thermes, Phys. Rep. **498**, 45 (2011).

[4] W. Deng and Y. H. Luan, Abstract Appl. Anal., Art. No. 926519 (2013).

[5] R. Roman-Roldan, P. Bernaola-Galvan, and J. L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).

[6] P. W. Messer and P. F. Arndt, Mol. Biol. Evol. **24**, 1190 (2007).

[7] P. Carpena, J. L. Oliver, M. Hackenberg, A. V. Coronado, G. Barturen, and P. Bernaola-Galvan, Phys. Rev. E **83**, 031908 (2011).

[8] P. Polak and P. F. Arndt, Genome Biol. Evol. **1**, 189 (2009).

[9] J. L. Oliver, P. Bernaola-Galvan, M. Hackenberg, and P. Carpena, BMC Evol. Biol. **8**, 107 (2008).

[10] W. Li and K. Kaneko, Nature (London) **360**, 635 (1992).

[11] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[12] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[13] Y. Almirantis and A. Provata, BioEssays **23**, 647 (2001).

[14] A. Provata and Th. Oikonomou, Phys. Rev. E **75**, 056102 (2007).

[15] M. Papapetrou and D. Kugiumtzis, Phys. A (Amsterdam) **392**, 1593 (2013).

[16] G. Simons, Y. C. Yao, and G. Morton, J. Stat. Planning Infer. **130**, 251 (2005).

[17] B. Ryabko and N. Usotskaya, in *Proceedings of the IEEE Region 8 International Conference on Computational Technologies in Electrical and Electronics Engineering, SIBIRCON 2008* (unpublished); N. Usotskaya and B. Ryabko, Comput. Stat. Data Anal. **53**, 1861 (2009); B. Ryabko and N. Usotskaya, Proc. IEEE Inf. Theory Workshop **2008**, 119 (2008).

[18] J. Besag and D. Mondal, Biometrics **69**, 488 (2013).

[19] M. Menendez, L. Pardo, M. C. Pardo, and K. Zografos, Methodol. Comput. Appl. Prob. **13**, 59 (2011).

[20] P. Billingsley, Ann. Math. Stat. **32**, 12 (1961).

[21] P. G. Hoel, Biometrika **41**, 430 (1954).

[22] W. Lowry and D. Guthrie, Mon. Weather Rev. **96**, 798 (1968).

[23] P. J. Avery and D. A. Henderson, J. R. Stat. Soc. C: Appl. Stat. **48**, 53 (1999).

[24] E. Chargaff, Experientia **6**, 201 (1950); R. Rudner, J. D. Karkas, and E. Chargaff, Proc. Natl. Acad. Sci. USA **60**, 921 (1968).

[25] S. G. Kong, W. L. Fan, H. D. Chen, Z. T. Hsu, N. Zhou, B. Zheng, and H. C. Lee, PLoS ONE **4**, e7553 (2009).

[26] A. Hart, S. Martínez, and F. Olmos, J. Stat. Phys. **146**, 408 (2012).

[27] G. Albrecht-Buehler, Proc. Natl. Acad. Sci. USA **103**, 17828 (2006).

[28] G. Nicolis, G. Subba Rao, J. Subba Rao, and C. Nicolis, in *Structure, Coherence and Chaos in Dynamical Systems*, edited by P. Christiansen and R. Parmentier (Manchester University Press, Manchester, 1989).

[29] G. Nicolis and C. Nicolis, *Foundations of Complex Systems*, 2nd ed. (World Scientific, Singapore, 2012).

[30] H. Frisch, Adv. Chem. Phys. **55**, 201 (1984).

[31] W. Ebeling and G. Nicolis, Europhys. Lett. **14**, 191 (1991); W. Ebeling, T. Poschel, and K.-F. Albrecht, Int. J. Bifurcation Chaos **5**, 51 (1995).

[32] J. S. Nicolis, *Chaos and Information Processing* (World Scientific, Singapore, 1991).

[33] R. Roman-Roldan, P. Bernaola-Galvan, and J. L. Oliver, Pattern Recognit. **29**, 1187 (1996).

[34] P. Gaspard, J. Stat. Phys. **117**, 599 (2004).

[35] J. L. Luo, C. Van den Broeck, and G. Nicolis, Z. Phys. B **56**, 165 (1984).

[36] D. Andrieux and P. Gaspard, Proc. Natl. Acad. Sci. USA **105**, 9516 (2008).

[37] J. S. Nicolis, *Dynamics of Hierarchical Systems* (Springer, Berlin, 1986).

[38] M. Kalkatawi, F. Rangkuti, M. Schramm, B. R. Jankovic, A. Kamau, R. Chowdhary, J. A. C. Archer, and V. B. Bajic, Bioinformatics **28**, 127 (2012).

[39] A. J. Lustig and T. D. Petes, J. Mol. Biol. **180**, 753 (1984).

[40] R. W. Hamming, Bell Syst. Tech. J. **29**, 147 (1950).

[41] W. Feller, *Introduction to Probability Theory and its Applications*, Vol. I (Wiley, New York, 1968).

[42] C. Gardiner, *Handbook of Stochastic Methods* (Springer, Berlin, 1983).

[43] P. Katsaloulis, T. Theoharis, and A. Provata, J. Theor. Biol. **258**, 18 (2009); P. Katsaloulis, T. Theoharis, W. M. Zheng, B. L. Hao, A. Bountis, Y. Almirantis, and A. Provata, Phys. A (Amsterdam) **366**, 308 (2006); P. Katsaloulis, T. Theoharis, and A. Provata, *ibid.* **316**, 380 (2002).

[44] V. Afreixo, C. A. C. Bastos, A. J. Pinho, S. P. Garcia, and P. J. S. G. Ferreira, Bioinformatics **25**, 3064 (2009); C. A. C. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. M. O. S. Rodrigues, and P. J. S. G. Ferreira, Adv. Intell. Soft Comput. **93**, 205 (2011).

[45] K. M. Frahm and D. L. Shepelyansky, Phys. Rev. E **85**, 016214 (2012).

[46] V. Balakrishnan, G. Nicolis, and C. Nicolis, J. Stat. Phys. **86**, 191 (1997).

[47] J. Masoliver, K. Lindenberg, and B. J. West, Phys. Rev. A **33**, 2177 (1986).

[48] E. Altmann, G. Cristadoro, and M. Esposti, Proc. Natl. Acad. Sci. USA **109**, 11582 (2012).

[49] C. Nicolis, J. Atmos. Sci. **55**, 465 (1998); A. Trevisan, *ibid.* **52**, 3577 (1995).

[50] J. Feder, *Fractals* (Plenum, New York, 1988).

[51] M. Hackenberg, P. Bernaola-Galvan, P. Carpena, and J. L. Oliver, J. Mol. Evol. **60**, 365 (2005).

[52] A. L. Price, E. Eskin, and P. A. Pevzner, Genome Res. **14**, 2245 (2004).

[53] W. Li, P. Bernaola-Galvan, P. Carpena, and J. L. Oliver, Computat. Biol. Chem. **27**, 5 (2003).

[54] J. L. Oliver, P. Carpena, R. Roman-Roldan, T. Mata-Balaguer, A. Mejas-Romero, M. Hackenberg, and P. Bernaola-Galvan, Gene **300**, 117 (2002).

[55] P. L. Luque-Escamilla, J. Martínez-Aroza, J. L. Oliver, J. F. Gómez-Lopera, and R. Román Roldán, Phys. Rev. E **71**, 061925 (2005).

[56] P. Bernaola-Galvan, J. L. Oliver, P. Carpena, O. Clay, and G. Bernardi, Gene **333**, 121 (2004).

[57] P. F. Arndt, T. Hwa, and D. A. Petrov, J. Mol. Evol. **60**, 748 (2005).

[58] J. M. Gutierrez, M. A. Rodriguez, and G. Abramson, Phys. A (Amsterdam) **300**, 271 (2001).

[59] H. Herzel and I. Grosse, Phys. A (Amsterdam) **216**, 518 (1995).

[60] A. Provata, Phys. A (Amsterdam) **264**, 570 (1999).

[61] P. Allegrini, M. Barbi, P. Grigolini, and B. J. West, Phys. Rev. E **52**, 5281 (1995).

[62] C. Beck and A. Provata, Europhys. Lett. **95**, 58002 (2011).

[63] R. S. MacKay, J. D. Meiss, and I. C. Percival, Phys. D (Amsterdam) **27**, 1 (1987).

[64] C. F. F. Karney, Phys. D (Amsterdam) **8**, 360 (1983).