Analysis of symbolic sequences using the Jensen–Shannon divergence

Article *in* Physical Review E · May 2002

DOI: 10.1103/PhysRevE.65.041905 · Source: PubMed

6 authors, including:

Ivo Grosse
Martin Luther University Halle-Wittenberg
284 PUBLICATIONS **12,428** CITATIONS
SEE PROFILE

Pedro Bernaola-Galvan
University of Malaga
87 PUBLICATIONS **4,463** CITATIONS
SEE PROFILE

Pedro Carpena
University of Malaga
116 PUBLICATIONS **5,958** CITATIONS
SEE PROFILE

Ramon Roman Roldan
University of Granada
41 PUBLICATIONS **1,904** CITATIONS
SEE PROFILE

Analysis of symbolic sequences using the Jensen-Shannon divergence

Ivo Grosse,[1,2] Pedro Bernaola-Galván,[2,3] Pedro Carpena,[2,3] Ramón Román-Roldán,[4] Jose Oliver,[5] and H. Eugene Stanley[2]

[1]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724*

[2]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*

[3]*Departamento de Física Aplicada II, ETSI de Telecomunicación, Universidad de Málaga, E-29071 Málaga, Spain*

[4]*Departamento de Física Aplicada, Universidad de Granada, E-18071 Granada, Spain*

[5]*Departamento de Genética e Instituto de Biotecnología, Universidad de Granada, E-18071 Granada, Spain*

(Received 22 December 2000; revised manuscript received 8 August 2001; published 25 March 2002)

We study statistical properties of the Jensen-Shannon divergence D, which quantifies the difference between probability distributions, and which has been widely applied to analyses of symbolic sequences. We present three interpretations of D in the framework of statistical physics, information theory, and mathematical statistics, and obtain approximations of the mean, the variance, and the probability distribution of D in random, uncorrelated sequences. We present a segmentation method based on D that is able to segment a nonstationary symbolic sequence into stationary subsequences, and apply this method to DNA sequences, which are known to be nonstationary on a wide range of different length scales.

DOI: 10.1103/PhysRevE.65.041905 PACS number(s): 87.15.Cc

I. INTRODUCTION

The statistical analysis of symbolic sequences is of central importance in various fields of science, such as symbolic dynamics [1,2], linguistics (following the pioneering works of Shannon [3]), or DNA sequence analysis [4–7]. One advantage of using information theoretical functionals for the analysis of symbolic sequences is that they do not require the symbolic sequence to be mapped to a numerical sequence, which is necessary in spectral or correlation analyses [8]. One of these functionals is the *Jensen-Shannon divergence D* [9–12], which quantifies the difference between two (or more) probability distributions, and which can be used to compare the symbol composition between different sequences.

There are three reasons why we choose D as a measure of divergence between probability distributions: (i) D is related to other information-theoretical functionals, such as the relative entropy or the Kullback divergence, and hence it shares their mathematical properties as well as their intuitive interpretability, (ii) D can be generalized to measure the distance between more than two distributions, and (iii) the compared distributions can be weighted, which allows us to take into account the different lengths of the subsequences from which the probability distributions are computed [13].

D has been used for measuring the distance between random graphs [10], for testing the goodness-of-fit of point estimations [12], in the analysis of DNA sequences [13,14], in the segmentation of textured images [15], and in the design of a statistical characterization of the mobility edge in disordered materials [16]. In addition, by making use of its ability to be generalized to an arbitrary number of probability distributions, D has been used to quantify the complex heterogeneity of DNA sequences [17–19] as well as to detect borders between coding and noncoding DNA [20].

Here we describe in detail some statistical properties of D as well as some theoretical background relevant for the above-mentioned applications. This paper is organized as follows: in Sec. II we introduce D and some of its mathematical properties. In Sec. III we provide three interpretations of D, one in the framework of statistical physics, one in the framework of information theory, and one in the framework of mathematical statistics. In Sec. IV we discuss some statistical properties of D, and we derive the mean, the variance, and the asymptotic probability distribution function of D. In Sec. V we apply the Jensen-Shannon divergence to the problem of segmenting a nonstationary sequence into stationary subsequences, and show that in this context the maximum value $D_{\max}$ of the Jensen-Shannon divergence D sampled along a sequence becomes a quantity of central importance. Hence, we study the probability distribution of $D_{\max}$ by means of Monte-Carlo simulations. In Sec. VI we present three examples of how D can be applied to the problem of segmenting nonstationary symbolic sequences (such as DNA sequences) into stationary subsequences, and Sec. VII concludes this paper.

II. THE JENSEN-SHANNON DIVERGENCE

Several measures have been proposed to quantify the difference (sometimes called *divergence*) between two (or more) probability distributions [9]. One of those measures is the Jensen-Shannon divergence, which is defined as follows: let $\mathbf{p}^{(1)}\equiv(p_1^{(1)},p_2^{(1)},\ldots,p_k^{(1)})$ and $\mathbf{p}^{(2)}\equiv(p_1^{(2)},p_2^{(2)},\ldots,p_k^{(2)})$ denote two probability distributions satisfying the usual constraints $\Sigma_{i=1}^{k}p_i^{(j)}=1$ and $0\leqslant p_i^{(j)}\leqslant 1$ for all $i=1,2,...,k$ and $j=1$, 2; and let $\pi^{(1)}$ and $\pi^{(2)}$ denote the *weights* of the distributions $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$, satisfying the constraints $\pi^{(1)}+\pi^{(2)}=1$ and $0\leqslant\pi^{(j)}\leqslant 1$. Then the Jensen-Shannon divergence D between the probability distributions $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ with weights $\pi^{(1)}$ and $\pi^{(2)}$ is defined by [11]

$$D[\mathbf{p}^{(1)},\mathbf{p}^{(2)}]\equiv H[\pi^{(1)}\mathbf{p}^{(1)}+\pi^{(2)}\mathbf{p}^{(2)}]-(\pi^{(1)}H[\mathbf{p}^{(1)}]+\pi^{(2)}H[\mathbf{p}^{(2)}]), \quad (1)$$

where

$$H[\mathbf{p}] = -\sum_{i=1}^{k} p_i \log_2 p_i \tag{2}$$

denotes the Shannon entropy of the probability distribution $\mathbf{p} \equiv (p_1, p_2, \ldots, p_k)$.

The Jensen-Shannon divergence D can be shown to be a special case of the *Jensen difference divergence* introduced by Burbea and Rao [21]. Also, D can be shown to be a special case of the φ *divergence* introduced by Csiszar [12,22]. Hence, the Jensen-Shannon divergence D shares all mathematical properties of both the Jensen difference divergence and the φ divergence. It is interesting to note that the Jensen-Shannon divergence is the only measure that simultaneously belongs to the family of Jensen difference divergences and the family of φ divergences [12], i.e., the intersection of the family of Jensen difference divergences and the family of φ divergences contains only a single measure, and that measure is the Jensen-Shannon divergence D.

In the following two paragraphs we list some mathematical properties of D that turn out to be important for its application as a divergence measure.

(1) By using the Jensen inequality [23] it is easy to see that

$$D[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}] \geqslant 0, \tag{3}$$

with $D\,[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}] = 0$ if and only if $\mathbf{p}^{(1)} = \mathbf{p}^{(2)}$.

(2) D is symmetric in its arguments $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$, i.e.,

$$D[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}] = D[\mathbf{p}^{(2)}, \mathbf{p}^{(1)}]. \tag{4}$$

(3) D is well defined even if $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ are not absolutely continuous, i.e., D is well-defined even if $p_i^{(1)}$ vanishes without vanishing $p_i^{(2)}$ or if $p_i^{(2)}$ vanishes without vanishing $p_i^{(1)}$.

D can be generalized to quantify the divergence between an arbitrary number of probability distributions. Let us consider m probability distributions $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}, \ldots,$ $\mathbf{p}^{(m)}$, and let us denote by $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(m)}$ the corresponding weights. We can define the Jensen-Shannon divergence between the m probability distributions $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$ with weights $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(m)}$ by

$$D[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}] = H\left[\sum_{j=1}^{m} \pi^{(j)} \mathbf{p}^{(j)}\right] - \sum_{j=1}^{m} \pi^{(j)} H[\mathbf{p}^{(j)}]. \tag{5}$$

It is interesting to note that the three mathematical properties mentioned above for the binary case can be generalized to the m-ary case as follows:

(1) The Jensen inequality [23] implies that

$$D[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}] \geqslant 0, \tag{6}$$

with $D\,[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}] = 0$ if and only if *all* probability distributions $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$ are identical, i.e., if and only if $\mathbf{p}^{(1)} = \mathbf{p}^{(2)} = \cdots = \mathbf{p}^{(m)}$.

(2) D is symmetric in its arguments $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$, i.e., D is invariant under any permutation of its arguments $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$.

(3) D is well defined even if the probability distributions $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$ are not absolutely continuous.

III. INTERPRETATIONS OF *D*

In the following three sections we will present three intuitive interpretations of the Jensen-Shannon divergence D.

A. Interpretation of *D* in the framework of statistical physics

In this section we show that D can be interpreted as the *intensive mixture entropy* in the following way: let us consider m vessels, each one containing a mixture of k ideal gases, let $\mathbf{f}^{(j)} \equiv (f_1^{(j)}, f_2^{(j)}, \ldots, f_k^{(j)})$ denote the vector of molar fractions of the k gases in the jth vessel for $j = 1,2,...,m$, and let $n^{(j)}$ denote the total number of molecules in the jth vessel. Then we know from the second law of thermodynamics that the sum of the Boltzmann entropies of the m separate vessels is smaller than (or equal to) the Boltzmann entropy of the joint vessel that we obtain after mixing the gases from all m vessels, and we can easily show that the difference of the sum of the entropies obtained before the ideal gases are mixed and the entropy obtained after the ideal gases are mixed is equal to

$$H_{\mathrm{mix}} = N k_B (\ln 2) H[\mathbf{f}] - \sum_{j=1}^{m} n^{(j)} k_B (\ln 2) H[\mathbf{f}^{(j)}], \tag{7}$$

where k_B denotes the Boltzmann constant, $N \equiv \sum_{j=1}^{m} n^{(j)}$ denotes the total number of ideal gas particles in all m vessels, and $\mathbf{f} \equiv \sum_{j=1}^{m} (n^{(j)}/N) \mathbf{f}^{(j)}$ denotes the vector of molar fractions of the k gases in the mixture containing the gas particles of all of the m vessels. H_{mix} is commonly called *mixing entropy*, and it is easy to see that

$$H_{\mathrm{mix}} = N k_B (\ln 2) D, \tag{8}$$

if the weights are chosen to be $\pi^{(j)} \equiv n^{(j)}/N$. Hence, D can be interpreted as the *intensive mixture entropy* measured in units of $k_B \ln 2$.

B. Interpretation of *D* in the framework of information theory

In this section we show that D can be interpreted as the mutual information in the following way: let us consider a sequence $\mathcal{S}$ of N symbols chosen from the alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$, and let us denote by p_i the probability of finding symbol a_i at an arbitrary but fixed position in sequence $\mathcal{S}$, for $i = 1,2,...,k$. Suppose that the sequence $\mathcal{S}$ is divided into m subsequences $\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(m)}$ of given lengths $n^{(1)}, n^{(2)}, \ldots, n^{(m)}$, and let us denote by $p_i^{(j)}$ the probability of finding symbol a_i at an arbitrary but fixed position in sequence $\mathcal{S}^{(j)}$, for $i = 1,2,...,k$ and $j = 1,2,...,m$.

In order to establish the connection between D and the mutual information defined in the framework of information theory, we define the random vector $(a,\ s)$, where the random variables $a \in \mathcal{A}$ and $s \in \{\boldsymbol{\mathcal{S}}^{(1),}\ \boldsymbol{\mathcal{S}}^{(2),\ldots,\boldsymbol{\mathcal{S}}^{(m)}}\}$ are generated as follows: draw a random position n with a uniform probability distribution along the sequence $\boldsymbol{\mathcal{S}}$, denote by a the symbol at position n, denote by s the subsequence that contains position n, and denote by p_{ij} the joint probability of $a = a_i$ and $s = \boldsymbol{\mathcal{S}}^{(j)}$ for $i = 1,2, \ldots ,k$ and $j = 1,2, \ldots m$. Then we obtain that the random variable a assumes the values $a_1, a_2, \ldots, a_k$ with probabilities $p_1, p_2, \ldots, p_k$, and the random variable s assumes the values $\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(m)}$ with probabilities $\pi^{(1)} \equiv n^{(1)}/N$, $\pi^{(2)} \equiv n^{(2)}/N \ldots$, $\pi^{(m)} \equiv n^{(m)}/N$, where the marginal possibilities p_i and $\pi^{(j)}$ are defined by

$$p_i \equiv \sum_{j=1}^{m} p_{ij} \quad \text{and} \quad \pi^{(j)} \equiv \sum_{i=1}^{k} p_{ij}$$

for $i = 1,2, \ldots ,k$ and $j = 1,2, \ldots ,m$.

Suppose that someone is drawing a symbol a from the entire sequence $\mathcal{S}$, not telling us from which subsequence s this symbol was drawn, and suppose it is our task to guess that subsequence $\mathcal{S}$ from which symbol a was drawn. One question answered by information theory is: "How much information I can we obtain from learning the identity of the symbol a about the identity of that subsequence s from which symbol a was drawn, provided we know the probability distribution $\{p_{ij}\}$?"

I is called the *mutual information in a about s* and defined by [3]

$$I \equiv \sum_{i=1}^{k} \sum_{j=1}^{m} p_{ij} \log_2 \frac{p_{ij}}{\pi^{(j)} p_i}. \tag{9}$$

Taking into account that $p_i^{(j)}$ denotes the conditional probability of finding symbol a_i at an arbitrary but fixed position in a given (fixed) sequence $\mathcal{S}^{(j)}$, it follows that $p_{ij} = \pi^{(j)} p_i^{(j)}$, and Eq. (9) can be rewritten as

$$I \equiv \sum_{i=1}^{k} \sum_{j=1}^{m} \pi^{(j)} p_i^{(j)} \log_2 \frac{p_i^{(j)}}{p_i}. \tag{10}$$

By rewriting Eq. (10) we obtain

$$I = \sum_{j=1}^{m} \pi^{(j)} \sum_{i=1}^{k} p_i^{(j)} \log_2 p_i^{(j)} - \sum_{i=1}^{k} \left(\sum_{j=1}^{m} \pi^{(j)} p_i^{(j)} \right) \log_2 p_i. \tag{11}$$

As $p_i = \Sigma_{j=1}^{m} \pi^{(j)} p_i^{(j)}$ defines the probability of finding symbol a_i in the whole sequence, we obtain

$$I = D[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}]. \tag{12}$$

Hence, D is identical to the mutual information in a about s, which quantifies the amount of information we obtain from learning the identity of the chosen symbol a about the identity of that subsequence s from which symbol a was chosen.

As I is symmetric in its arguments a and s, we may also consider the following game: suppose someone is drawing a symbol a from sequence $\mathcal{S}$, not telling us the identity of the drawn symbol a, but telling us the identity of that subsequence s from which symbol a was drawn. Suppose further that it is our task to guess the identity of the drawn symbol a. One question answered by information theory is: "How much information I can we obtain from learning the identity of the subsequence s about the identity of the drawn symbol a, provided we know the probability distribution $\{p_{ij}\}$." It can be mathematically proven that the mutual information in a about s is identical to the mutual information in s about a, and hence we can state that the Jensen-Shannon divergence D quantifies the amount of information we obtain from learning the identity of the subsequence s about the identity of the drawn symbol a.

If $\mathbf{p}^{(1)} = \mathbf{p}^{(2)} = \cdots = \mathbf{p}^{(m)}$, then it is clear that knowing the identity of the symbol a does not tell us anything about the identity of the subsequence s from which a was drawn, as the probability distributions of a are identical in all subsequences s. Likewise, it is clear that in this case knowing the subsequence s from which a was drawn does not tell us anything about the identity of a. Hence, it is intuitively clear that the mutual information in a about s (or the mutual information in s about a) is equal to zero, and hence it is also intuitively clear that in this case the Jensen-Shannon divergence D is equal to zero.

C. Interpretation of *D* in the framework of mathematical statistics

In this section we show that D can be interpreted as the *log-likelihood ratio* in the following way: consider the problem of estimating the probabilities $\mathbf{p} \equiv (p_1, p_2, \ldots, p_k)$ from a symbolic i.i.d. [24] sequence $\mathcal{S}$ of length N, in which at each position a symbol $a_i \in \mathcal{A} \equiv \{a_1, a_2, \ldots, a_k\}$ is randomly drawn with probability p_i. The maximum likelihood principle suggests to choose that probability vector $\mathbf{p}$ which maximizes the *likelihood*

$$L(\mathcal{S}|\mathbf{p}) \equiv \prod_{i=1}^{k} p_i^{F_i}, \tag{13}$$

where F_i denotes the number of occurrences of symbol a_i in sequence $\mathcal{S}$. As the logarithm is a strictly monotonic function, one may equally search for that $\mathbf{p}$ which maximizes $\ln L = \Sigma_{i=1}^{k} F_i \ln p_i$. It is easy to derive by using one Lagrange multiplier for the constraint $\Sigma_{i=1}^{k} p_i = 1$ that $p_i = F_i/N$ maximizes the *log-likelihood* $\ln L$. Hence, we obtain as maximum log-likelihood

$$\ln L_{\max} = N \sum_{i=1}^{k} f_i \ln f_i = -N(\ln 2) H[\mathbf{f}], \tag{14}$$

where $f_i \equiv F_i/N$ denotes the relative frequency of finding symbol a_i in sequence $\mathcal{S}$ of length N.

Now consider the slightly more complicated problem of a nonstationary sequence $\mathcal{S}$ of length N consisting of m stationary subsequences $\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(m)}$ with lengths $n^{(1)}, n^{(2)}, \ldots, n^{(m)}$, where the probability $p_i^{(j)}$ of generating symbol a_i in subsequence $\mathcal{S}^{(j)}$ may vary from subsequence to subsequence. The likelihood of obtaining the entire sequence $\mathcal{S}$ is equal to the product of the likelihoods of obtaining the m subsequences $\mathcal{S}^{(1)},\mathcal{S}^{(2)},\ldots,\mathcal{S}^{(m)}$. Hence, the maximum likelihood principle suggests to choose for each $j=1,2,\ldots,m$ that probability vector $\mathbf{p}^{(j)}\equiv(p_1^{(j)},p_2^{(j)},\ldots,p_k^{(j)})$ that maximizes the likelihood

$$L(\mathcal{S}^{(j)}|\mathbf{p}^{(j)})\equiv\prod_{i=1}^{k}(p_i^{(j)})^{F_i^{(j)}}, \tag{15}$$

where $F_i^{(j)}$ is the number of occurrences of symbol a_i in subsequence $\mathcal{S}^{(j)}$. It is again easy to derive by using m Lagrange multipliers for the m constraints $\Sigma_{i=1}^{k}p_i^{(j)}=1$ that $p_i^{(j)}=F_i^{(j)}/n^{(j)}$ maximizes the log-likelihood $\ln L^{(j)}$. Hence, we obtain as maximum log-likelihood

$$\ln L_{\max}^{(j)}=n^{(j)}\sum_i f_i^{(j)}\ln f_i^{(j)}=-n^{(j)}(\ln 2)H[\mathbf{f}^{(j)}], \tag{16}$$

where $f_i^{(j)}\equiv F_i^{(j)}/n^{(j)}$ denotes the relative frequency of finding symbol a_i in subsequence $\mathcal{S}^{(j)}$ of length $n^{(j)}$.

As problem one (with just one sequence) is a special case of problem two (of having m sequences), the sum of the maximum log-likelihoods $\Sigma_{j=1}^{m}\ln L_{\max}^{(j)}$ cannot be smaller than $\ln L_{\max}$, because in the "worst" case in which all of the m subsequences of problem two were *identical*, problem two would just reduce to problem one, giving the same log-likelihood as in problem one. Hence, the quantity

$$\Delta L\equiv\sum_{j=1}^{m}\ln L_{\max}^{(j)}-\ln L_{\max} \tag{17}$$

is non-negative, and ΔL is commonly called the log-likelihood ratio.

It is straightforward to see from Eqs. (14), (16), and (17) that

$$\Delta L=N(\ln 2)D. \tag{18}$$

Hence, in the framework of mathematical statistics ΔL can be interpreted as the increase of the log-likelihood when sequence $\mathcal{S}$, instead of being modeled as a sequence generated with a single probability vector $\mathbf{p}$, is modeled as a concatenation of m subsequences $\mathcal{S}^{(1)},\mathcal{S}^{(2)},\ldots,\mathcal{S}^{(m)}$ (in that order) generated from the probability vectors $\mathbf{p}^{(1)},\mathbf{p}^{(2)},\ldots,\mathbf{p}^{(m)}$.

The inequality $\Delta L\geqslant 0$ states that *any* partition of the original sequence into m subsequences increases the likelihood of the second model over the first model. In order to choose hypothesis two (m subsequences) in favor of hypothesis one (only one sequence), we require that ΔL be *significantly* greater than zero, and it is the goal of this paper to derive an approximation of the probability distribution function of ΔL.

Note that in all of the above interpretations of D the weights of the distributions $\pi^{(1)},\pi^{(2)},\ldots,\pi^{(m)}$ are proportional to the "sizes" $n^{(1)},n^{(2)},\ldots,n^{(m)}$ of the m elements considered: the number of particles of each of the m ideal vessels or the number of symbols in each of the m subsequences. It is interesting that this particular choice of weights arises in a natural way from all of the three interpretations presented above, and—as we will see later—this choice of weights endows the Jensen-Shannon divergence D with several statistical properties that make D particularly suitable for the analysis of symbolic sequences.

IV. STATISTICAL PROPERTIES OF *D*

Formally, D is a function of the probability distributions $\mathbf{p}^{(1)},\mathbf{p}^{(2)},\ldots,\mathbf{p}^{(m)}$, but in analyses of experimental data those probability distributions are not (directly) observable. However, when we study experimental symbolic sequences we can estimate those probability distributions $\mathbf{p}^{(j)}$ from the frequency distributions $\mathbf{f}^{(j)}\equiv(f_1^{(j)},f_2^{(j)},\ldots,f_k^{(j)})$, where $f_i^{(j)}$ denotes the relative frequency of symbol a_i in subsequence $\mathcal{S}^{(j)}$, for $i=1,2,\ldots,k$ and $j=1,2,\ldots,m$.

In all analyses of experimental data the Jensen-Shannon divergence D must be computed from those (observable) frequency distributions $\mathbf{f}^{(1)},\mathbf{f}^{(2)},\ldots,\mathbf{f}^{(m)}$ rather than from the (nonobservable) probability distributions $\mathbf{p}^{(1)},\mathbf{p}^{(2)},\ldots,\mathbf{p}^{(m)}$. As a consequence of replacing the probabilities $p_i^{(j)}$ by the corresponding relative frequencies $f_i^{(j)}$ in Eq. (1), the numerical values of D will fluctuate from data set to data set, even if those data sets can be assumed to be generated from the *same* probability distribution.

The fluctuation of $f_i^{(j)}$ from data set to data set may not only result in fluctuations of the numerical values of D, but also in a systematic shift (bias) of the numerical values of D computed from the observed data as compared to the numerical value of D computed from the unobservable probability distributions. In order to illustrate the presence of those fluctuations of D as well as its systematic shift (called *bias*), we perform the following control experiments:

We generate an ensemble of 2000 binary sequences ($k=2$) of $N=2500$ symbols each, obtained by joining $m=2$ subsequences as follows: we generate the left sequence of length $n=500$ by concatenating random, uncorrelated symbols drawn from the probability distribution $\mathbf{p}^{(1)}=(0.45,0.55)$, and the right sequence of length $N-n=2000$ symbols drawn from the probability distribution $\mathbf{p}^{(2)}=(0.55,0.45)$.

We move a cursor along the entire sequence, and we compute D between the subsequences at both sides of the cursor for all positions $n^{(1)}=1,2,\ldots,\ N-1$ and $n^{(2)}=N-1,\ N-2,\ldots,1$. In order to illustrate the effect of different choices of the weights $\pi^{(j)}$, we compute the Jensen-Shannon divergence in two different ways: (i) for the choice of equal weights $\pi^{(j)}=1/m$ for all subsequences $\mathcal{S}^{(j)}$, and (ii) for the *natural* choice of weights $\pi^{(j)}=n^{(j)}/N$. In the following we denote by $D_{1/m}$ the Jensen-Shannon divergence with the choice of equal weights (i), and we denote by D the Jensen-Shannon divergence with the natural choice of weights (ii).

An ideal estimator of D, which quantifies the difference between two probability distributions, should reach its maximum value exactly at that point which separates the subsequences generated by different probability distributions, i.e., it should reach its maximum value when $n^{(1)}=n=500$ and $n^{(2)}=N-n=2000$. Figure 1(a) shows $\langle D\rangle$ versus $n^{(1)}$ and $\langle D_{1/2}\rangle$ versus $n^{(1)}$, where the symbol $\langle\cdots\rangle$ denotes the en-

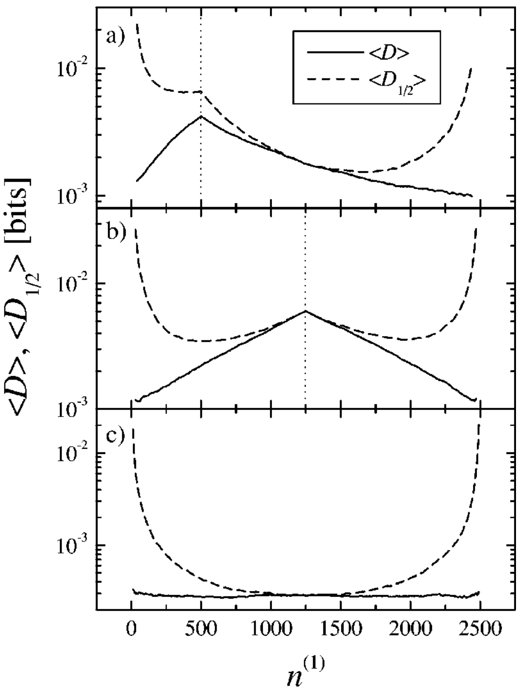

FIG. 1. Comparison of D and $D_{1/2}$. We generate an ensemble of 2000 binary sequences of length $N=2500$, obtained by joining two subsequences of lengths n and $N-n$, where the left subsequence of length n is generated from a probability distribution $(x,1-x)$ and the right subsequence of length $N-n$ is generated from a probability distribution $(y,1-y)$. We move a cursor along the entire sequence and we compute D and $D_{1/2}$ between the subsequences at both sides of the cursor. Finally we plot the ensemble averages $\langle D\rangle$ (solid line) and $\langle D_{1/2}\rangle$ (dashed line) as a function of the position of the cursor $n^{(1)}=1,2,...,N-1$. In (a) we choose $n=500$, $x=0.45$, and $y=0.55$, and find that D achieves its global maximum at $n^{(1)}\approx 500$ in the vicinity of the true fusion point of the two subsequences at $n=500$, whereas $D_{1/2}$ achieves its global maximum at the edges $n^{(1)}\rightarrow 0$ or $n^{(1)}\rightarrow 2500$ far away from the true fusion point of the two subsequences at $n=500$. This finding indicates that D might serve as an appropriate divergence measure to quantify the compositional differences between symbolic subsequences, whereas $D_{1/2}$ might not. In (b) we choose $n=1250$, $x=0.45$, and $y=0.55$, and find again that D achieves its global maximum at $n^{(1)}\approx 1250$ in the vicinity of the true fusion point of the two subsequences at $n=1250$, whereas $D_{1/2}$ achieves its global maximum at the edges $n^{(1)}\rightarrow 0$ or $n^{(1)}\rightarrow 2500$ far away from true fusion point of the two subsequences at $n=1250$, confirming the finding from (a) that D might serve as an appropriate divergence measure to quantify the compositional differences between symbolic subsequences, whereas $D_{1/2}$ might not. In (c) we choose $n=1250$ and $x=y=0.5$, and we find that D stays quite constant at a small value of approximately 2.9×10^{-4} bits, reflecting the fact that the analyzed sequences are stationary, whereas $D_{1/2}$ is clearly increasing as $n^{(1)}\rightarrow 0$ or $n^{(1)}\rightarrow 2500$, confirming the finding from (a) and (b) that D might serve as an appropriate divergence measure to quantify the compositional differences between symbolic subsequences, whereas $D_{1/2}$ might not. The effect that even in the case of i.i.d. sequences the expected value of D is greater than zero is referred to as finite-size effect, and we address this finite-size effect in Sec. IV.

semble average over all 2000 realizations.

Figure 1(a) shows that there are dramatic finite size effects when using $D_{1/2}$ (dashed line) instead of D (solid line). While D clearly achieves its global maximum at position $n^{(1)}\approx n=500$ [marked with a vertical dotted line in Fig. 1(a)], $D_{1/2}$ achieves its highest values at the beginning and the end of the horizontal axis, i.e., at very small and very large values of $n^{(1)}$.

We perform a second control experiment similar to the first experiment, in which we change the lengths of the two subsequences to $n=1250$ as well as $N-n=1250$, and in which we keep all other parameters the same as before. Figure 1(b) shows clearly that, again, D achieves its maximum at $n^{(1)}\approx n=1250$, while $D_{1/2}$ achieves its highest values at the beginning and the end of the horizontal axis, i.e., at very small and very large values of $n^{(1)}$.

These control experiments demonstrate two results: (i) the location of the maximum of D can separate regions of different composition and size in a symbolic sequence, and (ii) the estimation of $D_{1/2}$ and D from sequences of finite length is affected by finite size effects. In order to illustrate point (ii) directly, we perform a third control experiment in which we generate the two subsequences from *the same* probability distribution. In this case the experimentally obtained values of D that are nonzero are due only to statistical fluctuations.

Figure 1(c) shows $\langle D\rangle$ versus $n^{(1)}$ and $\langle D_{1/2}\rangle$ versus $n^{(1)}$ for an ensemble of 2000 stationary, binary sequences of length $N=2500$ in which each symbol is generated with probability 0.5. We find that, for all positions $n^{(1)}$, the values of D are approximately the same, whereas the values $D_{1/2}$ depend dramatically on $n^{(1)}$. Figure 1(c) also shows that $\langle D\rangle$ is not identical to zero, and we devote the following three sections to derivations of approximations of the mean, the variance, and the probability distribution function of D.

A. Mean of *D*

In this section we will derive an analytical approximation of the mean value of D when computed from an ensemble of finite i.i.d. sequences of length N.

It follows directly from the Jensen inequality that the expected value, $\langle H[\mathbf{f}]\rangle$, of the entropy computed from an ensemble of finite-length sequences cannot be greater than the theoretical value, $H[\mathbf{p}]$, of the entropy computed from the (unobservable) probabilities [25], i.e.,

$$\langle H|\mathbf{f}|\rangle \leqslant H[\mathbf{p}], \tag{19}$$

where $\langle\cdots\rangle$ denotes the expectation value over the ensemble of finite-length i.i.d. sequences generated by the probability distribution $\mathbf{p}$.

This mathematical statement is intuitively clear: due to the finite sample size, the relative frequency vector $\mathbf{f}$ fluctuates from sample to sample around the probability vector $\mathbf{p}$, and the majority of these fluctuations will make $\mathbf{f}$ less uniform than $\mathbf{p}$. Since the entropy $H[\mathbf{p}]$ quantifies the uniformity of the probability distribution $\mathbf{p}$, we expect that the majority of the values of $H[\mathbf{f}]$ computed from an ensemble of fluctuating frequency vectors $\mathbf{f}$ will be smaller than the value of $H[\mathbf{p}]$.

Up to first order the expected value of $H[\mathbf{f}]$ can be approximated by [26–30]

$$\langle H|\mathbf{f}|\rangle \simeq H[\mathbf{p}] - \frac{k-1}{2N\ln 2}, \tag{20}$$

where k is the number of components of the probability and frequency vectors $\mathbf{p}$ and $\mathbf{f}$, N is the sample size, and the symbol $\simeq$ indicates that we neglect terms of the order of $O(1/N^2)$. By applying Eq. (20) to each of the m subsequences we obtain

$$\langle H[\mathbf{f}^{(j)}]\rangle \simeq H[\mathbf{p}^{(j)}] - \frac{k-1}{2n^{(j)}\ln 2}, \tag{21}$$

for $j=1,2,...,m$, where the symbol $\simeq$ indicates that we neglect terms of the order of $O(1/(n^{(j)})^2)$. We will use approximations (20) and (21) to derive in the remainder of this section the expected value of the Jensen-Shannon divergence $D[\mathbf{f}^{(1)},\mathbf{f}^{(2)},\ldots,\mathbf{f}^{(m)}]$ computed from an ensemble of m i.i.d. sequences of total length N.

In order to avoid lengthy expressions, we define $D[\mathbf{F}] \equiv D[\mathbf{f}^{(1)},\mathbf{f}^{(2)},\ldots,\mathbf{f}^{(m)}]$ and $D[\mathbf{P}] \equiv D[\mathbf{p}^{(1)},\mathbf{p}^{(2)},\ldots,\mathbf{p}^{(m)}]$, and by substituting Eqs. (20) and (21) into Eq. (1) we obtain

$$\langle D[\mathbf{F}]\rangle \simeq D[\mathbf{P}] + \frac{k-1}{2N\ln 2}\left(\sum_{j=1}^{m} \pi^{(j)}\frac{N}{n^{(j)}} - 1\right). \tag{22}$$

This expression shows that, in general, the bias $\langle D[\mathbf{F}]\rangle - D[\mathbf{P}]$ depends on the lengths $n^{(j)}$ of the subsequences.

It is easy to see that one choice of weights that makes Eq. (22) independent of the subsequence lengths $n^{(j)}$ is

$$\pi^{(j)} \equiv n^{(j)}/N, \tag{23}$$

for $j=1,2,...,m$. This finding is interesting because this particular choice of weights turns out to be identical to the natural choice of weights in all of the three interpretations of D presented in Sec. III.

With this choice of weights, the expected value of the Jensen-Shannon divergence D becomes

$$\langle D[\mathbf{F}]\rangle \simeq D[\mathbf{p}] + \frac{k-1}{2N\ln 2}(m-1), \tag{24}$$

which is independent of the subsequence lengths $n^{(j)}$. Figure 2 illustrates the independence of the mean value of D of the subsequence lengths $n^{(j)}$, and it also shows that Eq. (24) is a reasonable approximation of the mean value of D.

Hence, expression (24) can be used as a reference to decide if a difference in composition between two sequences is larger than expected. Note that in Fig. 1(c) the average value of D fits the value predicted by Eq. (24), namely, $\langle D\rangle = 2.9 \times 10^{-4}$ bits. In addition, from Eq. (24) we see that the bias of the quantity ND is independent of the sequence length N, which allows us to compare Jensen-Shannon divergence values obtained from sequences of different sizes.

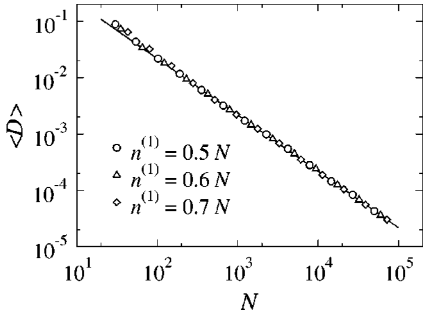

FIG. 2. Mean value of D as a function of the total sequence length N, ranging from $N=10$ to $N=10^5$, averaged over an ensemble of 2000 i.i.d. sequences generated from a four-letters alphabet ($k=4$), where each symbol occurs with probability 1/4. For each sequence length N we choose three different cutting points $n^{(1)}=0.5N$, $n^{(1)}=0.6N$, and $n^{(1)}=0.7N$, and we compute for each N and each $n^{(1)}$ and each of the 2000 i.i.d. sequences the Jensen-Shannon divergence D between the composition of the left subsequence of length $n^{(1)}$ and the composition of the right subsequence of length $n^{(2)}=N-n^{(1)}$. For each N and $n^{(1)}$ we compute the average of D over the ensemble of all 2000 i.i.d. sequences, and the figure shows the ensemble average $\langle D\rangle$ as a function of N and $n^{(1)}$. We find that $\log_{10}\langle D\rangle$ decays almost linearly as a function $\log_{10}N$, with a slope very close to -1, for each $n^{(1)}=0.5N$ (circles), $n^{(1)}=0.6N$ (triangles), and $n^{(1)}=0.7N$ (diamonds), and we also find that the approximation of $\langle D\rangle$ from Eq. (24) (solid line) agrees very well with the simulation results.

With the naive choice of weights $\pi^{(j)}=1/m$ we obtain for the expected value of the Jensen-Shannon divergence the approximation

$$\langle D_{1/m}[\mathbf{F}]\rangle \simeq D_{1/m}[\mathbf{P}] + \frac{k-1}{2N\ln 2}\left(\frac{N}{m}A - 1\right), \tag{25}$$

where $A \equiv \Sigma_{j=1}^{m} 1/n^{(j)}$ denotes the harmonic mean of the subsequence lengths $n^{(j)}$. Clearly $\langle D_{1/m}\rangle$ depends on the subsequence lengths $n^{(j)}$, and we see that $\langle D_{1/m}\rangle$ becomes minimal for $n^{(j)}=N/m$, while $\langle D_{1/m}\rangle$ diverges to infinity for $n^{(j)}\to 0$. This analytical approximation of the expected value of $D_{1/m}$ is consistent with the dramatic increase of the dashed line (corresponding to $\langle D_{1/m}\rangle$) close to the edges ($n^{(1)}\to 0$ or $n^{(2)}\to 0$) of the abscissa of Fig. 1.

There is another advantage of choosing the weights $\pi^{(j)}$ by Eq. (23). We will show in the following section that the choice of the weights $\pi^{(j)} \equiv n^{(j)}/N$ minimizes the quadratic deviation of the observed from the true Jensen-Shannon divergence. This advantage is more important than the advantage of having a bias that is independent of $n^{(j)}$, because the bias can be corrected analytically, in a first-order approximation, whereas the quadratic deviation of the observed from the true Jensen-Shannon divergence (i.e., the quadratic error) cannot be reduced. Hence, it is desirable to obtain an estimator of D that minimizes the quadratic deviation of the observed from the true Jensen-Shannon divergence (i.e., the quadratic error), and we will show in the following section that the choice of the weights $\pi^{(j)} \equiv n^{(j)}/N$ yields exactly that optimal estimator.

B. Variance of D

The variance of $D[\mathbf{F}]$ is given by

$$\sigma^2(D[\mathbf{F}])=\sigma^2\left(H[\mathbf{f}]-\sum_{j=1}^{m}\pi_j H[\mathbf{f}^{(j)}]\right)$$
$$=\sigma^2(H[\mathbf{f}])+\sum_{j=1}^{m}\pi_j^2\sigma^2(H[\mathbf{f}^{(j)}])$$
$$-2\sum_{j=1}^{m}\pi_j \mathrm{cov} H[\mathbf{f}],(H[\mathbf{f}^{(j)}])$$
$$+2\sum_{j=1}^{m}\sum_{l=j+1}^{m}\pi_j\pi_l \mathrm{cov}(H[\mathbf{f}^{(j)}],H[\mathbf{f}^{(l)}]). \tag{26}$$

As the set of vectors $\{\mathbf{f}^{(1)},\mathbf{f}^{(2)},\ldots,\mathbf{f}^{(m)}\}$ is product-multinomially distributed, we obtain that $H[\mathbf{f}^{(j)}]$ and $H[\mathbf{f}^{(l)}]$ are statistically independent for any $j\neq l$. Hence, the terms $\mathrm{cov}(H[\mathbf{f}^{(j)}],H[\mathbf{f}^{(l)}])$ are all equal to zero, and we need to consider only the terms $\sigma^2(H[\mathbf{f}])$, $\sigma^2(H[\mathbf{f}^{(j)}])$, and $\mathrm{cov}(H[\mathbf{f}],H[\mathbf{f}^{(j)}])$.

By Taylor-expanding $H[\mathbf{f}]$ about $\mathbf{p}$ we obtain a first-order approximation of the variance of $H[\mathbf{f}]$ [5,6,27,28],

$$\sigma^2(H[\mathbf{f}])\simeq\frac{1}{N}\sigma^2(\log_2\mathbf{p}), \tag{27}$$

where n_j denotes the length of subsequence $\mathcal{S}^{(j)}$, $\sigma^2(\log_2\mathbf{p}^{(j)})$ denotes the variance of the numbers $\log_2 p_i$ with respect to the probability distribution $\{p_i\}$, and the symbol $\simeq$ indicates that we neglect terms of the order of $O(1/N^2)$.

Likewise, we obtain a first-order approximation of the variance of $H[\mathbf{f}^{(j)}]$,

$$\sigma^2(H[\mathbf{f}^{(j)}])\simeq\frac{1}{n_j}\sigma^2(\log_2\mathbf{p}^{(j)}), \tag{28}$$

where N denotes the length of the whole sequence, $\sigma^2(\log_2\mathbf{p}^{(j)})$ denotes the variance of the numbers $\log_2 p_i^{(j)}$ with respect to the probability distribution $\{p_i^{(j)}\}$ for every $j=1,2,...,m$, and the symbol $\simeq$ indicates that we neglect terms of the order of $O(1/(n^{(j)})^2)$.

In the Appendix we derive a similar first-order approximation of the covariance terms, and under the null hypothesis that $\mathbf{p}^{(1)}=\mathbf{p}^{(2)}=\cdots=\mathbf{p}^{(m)}=\mathbf{p}$ we obtain

$$\mathrm{cov}(H[\mathbf{f}],H[\mathbf{f}^{(j)}])\simeq\frac{1}{N}\sigma^2(\log_2\mathbf{p}) \tag{29}$$

for all $j=1,2,...,m$, where $\sigma^2(\log_2\mathbf{p})$ denotes the variance of the numbers $\log_2 p_i$ with respect to the probability distribution $\{p_i\}$, and the symbol $\simeq$ indicates that we neglect terms of the order of $O(1/N^2)$. It is interesting to note that the first-order approximation of the covariance between $H[\mathbf{f}]$ and $H[\mathbf{f}^{(j)}]$ [Eq. (29)] is equal to the first-order approximation of the variance of $H[\mathbf{f}]$ [Eq. (27)].

By substituting the expressions from Eqs. (27), (28), and (29) into Eq. (26) we obtain for the variance of the Jensen-Shannon divergence with arbitrary weights $\pi^{(1)},\pi^{(2)},\ldots,\pi^{(m)}$,

$$\sigma^2(D)\simeq\left(\sum_{j=1}^{m}\pi^{(j)}\frac{\pi^{(j)}}{n^{(j)}}-\frac{1}{N}\right)\sigma^2(\log_2\mathbf{p}), \tag{30}$$

under the null hypothesis that $\mathbf{p}^{(1)}=\mathbf{p}^{(2)}=\cdots=\mathbf{p}^{(m)}=\mathbf{p}$, where the symbol $\simeq$ indicates that we neglect terms of the order of $O(1/N^2)$.

Let us now consider that choice of weights $\pi^{(j)}$ which minimizes the quadratic deviation of the observed from the true Jensen-Shannon divergence

$$\langle(D[\mathbf{F}]-D[\mathbf{P}])^2\rangle=\sigma^2(D)+(\langle D[\mathbf{E}]\rangle-D[\mathbf{P}])^2. \tag{31}$$

As the second term on the right hand side of Eq. (31) is of the order of $O(1/N^2)$, the minimization of the quadratic deviation of the observed from the true Jensen-Shannon divergence reduces to the minimization of the variance of the Jensen-Shannon divergence estimator.

By using one Lagrange multiplier for the normalization constraint $\Sigma_j\pi^{(j)}=1$ we obtain that the set of weights $\pi^{(j)}=n^{(j)}/N$ minimizes the variance of the Jensen-Shannon divergence D. This finding is intriguing, because this set of weights is (i) identical to the natural choice of weights in all of the three interpretations of D presented in Sec. III as well as (ii) identical to the special choice of weights that makes the bias of D independent of the subsequence lengths $n^{(j)}$ [Eq. (24)].

Furthermore, we find that for the special choice of weights $\pi^{(j)}\equiv n^{(j)}/N$ the variance of D vanishes in $O(1/N)$. This means that for the special choice of weights $\pi^{(j)}\equiv n^{(j)}/N$ the leading term of $\sigma^2(D)$ decreases with the sequence length N as $1/N^2$, whereas—in general—it decreases as $1/N$. It is clear that for the special choice of weights $\pi^{(j)}\equiv n^{(j)}/N$ the $O(1/N)$ term of $\sigma^2(D)$ becomes independent of both $n^{(j)}$ and $\mathbf{p}$, and it is interesting that for this special choice of weights the $O(1/N^2)$ term of $\sigma^2(D)$ also turns out to be independent of both $n^{(j)}$ and $\mathbf{p}$.

In contrast, we find that for the naive choice of weights $\pi^{(j)}\equiv 1/m$ the variance of $D_{1/m}$ neither vanishes in $O(1/N)$ nor does it become independent of the subsequence lengths $n^{(j)}$, and we obtain for the variance of the Jensen-Shannon divergence $D_{1/m}$,

$$\sigma^2(D)\simeq\frac{\sigma^2(\log_2 p)}{N}\left(\frac{N}{m^2}A-1\right), \tag{32}$$

where $A\equiv\Sigma_{j=1}^{m}1/n^{(j)}$ denotes the harmonic mean of the subsequence lengths $n^{(j)}$. Note that the expression inside the parentheses on the right-hand side of Eq. (32) is similar to the expression inside the parentheses on the right-hand side of Eq. (25). Hence, the variance of $D_{1/m}$ shows a singular behavior similar to that of the mean of $D_{1/m}$ when the length of at least one subsequence becomes very small.

C. Probability distribution of *D*

Expression (24) provides a good criterion to tell whether an experimentally observed Jensen-Shannon divergence D between m frequency distributions is greater than expected by chance, but it does not tell if D is *significantly* greater than expected by chance. In this section we will derive the probability distribution of D in order to quantify the statistical significance of experimentally observed values of D.

Given an observed value of $D=x$, we will calculate the probability of obtaining this value or a lower value by chance under the null hypothesis that all m sequences are generated from the same probability distribution. We call this probability the *significance threshold* of the given value x, and we denote it by

$$s(x) \equiv \text{Prob}\{D \leqslant x\}. \tag{33}$$

As $s(x)$ does not seem to admit an easy analytical expression, we will obtain an approximation by using the Taylor expansion

$$x \log_2 \frac{x}{a} = \frac{x-a}{\ln 2} + \frac{(x-a)^2}{a(2\ln 2)} + O((x-a)^3), \tag{34}$$

to approximate D in terms of quadratic functions as follows:

$$D \equiv \sum_{i=1}^{k} \sum_{j=1}^{m} p_i^{(j)} \pi^{(j)} \log_2 \frac{p_i^{(j)} \pi^{(j)}}{p_i \pi^{(j)}}$$

$$\simeq \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{p_i^{(j)} \pi^{(j)} - p_i \pi^{(j)}}{\ln 2} + \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(p_i^{(j)} \pi^{(j)} - p_i \pi^{(j)})^2}{p_i \pi^{(j)} (2\ln 2)} \tag{35}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(p_i^{(j)} \pi^{(j)} - p_i \pi^{(j)})^2}{p_i \pi^{(j)} (2\ln 2)}. \tag{36}$$

It is interesting to note that in this quadratic approximation of D there are no constant or linear terms because the first double sum of Eq. (35) vanishes exactly due to normalization of the probability distributions $p_i^{(j)}$, p_i, and $\pi^{(j)}$.

If we express the χ^2 statistic [31] in the same notation, we obtain

$$\chi^2 \equiv N \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(p_i^{(j)} \pi^{(j)} - p_i \pi^{(j)})^2}{p_i \pi^{(j)}} \simeq 2N(\ln 2) D. \tag{37}$$

The above χ^2 statistic is known to converge—for asymptotically large values of N—to the χ^2 distribution with $\nu=(k-1)(m-1)$ degrees of freedom [31]. Hence, also $2N(\ln 2)D$ converges—for asymptotically large values of N—to the χ^2 distribution with $\nu=(k-1)(m-1)$ degrees of freedom, i.e., we obtain for asymptotically large values of N the approximation

$$s(x) \simeq F_\nu[2N(\ln 2)x] \equiv \frac{\gamma[\nu/2, N(\ln 2)x]}{\Gamma(\nu/2)}, \tag{38}$$

where $\gamma(a,x)$ and $\Gamma(a)$ denote the *incomplete* and *complete* gamma functions, respectively [31,32].

The fact that D can be interpreted as mutual information agrees with Eq. (38), as it is known that, up to a multiplicative constant, the mutual information converges—for asymptotically large values of N—to the χ^2 probability distribution with $\nu=(k-1)(m-1)$ degrees of freedom [6].

V. STATISTICAL PROPERTIES OF $D_{\max}$

Expression (38) gives the significance threshold of a single value of D computed between two samples of fixed length. From the practical point of view this is equivalent to preselecting a fixed point that divides a sequence into two subsequences and asking for the probability that both subsequences have been generated from different probability distributions. But, in general, when facing an unknown sequence we do not have any *a priori* knowledge of the location of the possible cutting point.

The problem of finding the point where a nonstationary sequence can be most likely divided into two stationary subsequences has been widely studied in mathematics. There, the problem is known as the *change-point problem* [33–35], which consists of finding out (i) whether there exists a change point in the studied sequence, and (ii) at which position in the sequence the change point is located, provided it exists. Task (i) corresponds to determining whether the studied sequence is nonstationary, and task (ii) corresponds to determining the (most likely) location of the nonstationarity, provided it exists.

Since $2N(\ln 2)D$ can be interpreted as the log-likelihood ratio of the model with change point and the model without change point, the maximization of D along the sequence yields a natural way of determining the most likely location of the change point. Hence, we move a cursor along the entire sequence, compute D between the subsequences at both sides of the cursor for all positions, and choose that position as the optimal change point at which D reaches its maximum value $D_{\max}$.

In Sec. VI we describe a recursive segmentation algorithm that is based on this idea. The problem we will address in this section is to decide if the value $D_{\max}$ of the Jensen-Shannon divergence at the optimal change point is sufficiently large to partition the sequence at that point, or if the value $D_{\max}$ is sufficiently small to consider the entire sequence as stationary and not partition it at all. Hence, we will address in this section the problem of computing the statistical significance of experimentally observed values of $D_{\max}$.

Even if the studied sequence has been generated from a single probability distribution, we find $D_{\max}>0$ due to statistical fluctuations. Moreover, we find that $D_{\max}$ increases above any significance threshold s computed in Sec. IV as N increases. To decide if the obtained value $D_{\max}=x$ is statistically significant we need to compute the probability of obtaining this value or a lower value by chance in a random sequence, i.e., we need to compute

$$s_{\max}(x) = \mathrm{Prob}\{D_{\max} \leqslant x\}. \tag{39}$$

Obviously $s_{\max}(x) \neq s(x)$. In fact, if each value of D at each position of the cursor were independent of the others, we would obtain [36]

$$s_{\max}(x) = s(x)^N = \{F_{k-1}[2N(\ln 2)x]\}^N, \tag{40}$$

where N denotes the sequence length. Note that we are dealing with the comparison between only two distributions ($m = 2$), and hence the number of degrees of freedom is $\nu = k - 1$.

It is clear that the random variables D sampled at different positions of the same sequence are not statistically independent, because the value of D at a given position is almost identical to the value of D at the neighboring positions.

For binary ($k=2$) i.i.d. sequences Horvath [37] derives an analytic expression for $s_{\max}(x)$ in the limit of asymptotically large sequence lengths N, and Csorgo and Horvath [38] generalize that result to arbitrary k by deriving that the probability distribution function of $Z_N \equiv 2N(\ln 2)D_{\max}$ converges—for asymptotically large values of N—to

$$\mathrm{Prob}\{A_N Z_N \leqslant [B_N(\nu) + x]^2\} = \exp(-2e^{-x}), \tag{41}$$

where N denotes the sequence length, $\nu \equiv k-1$ denotes the number of degrees of freedom, A_N is defined by

$$A_N \equiv 2 \ln \ln N, \tag{42}$$

and $B_N(\nu)$ is defined by

$$B_N(\nu) \equiv 2 \ln \ln N + \frac{\nu}{2} \ln \ln \ln N - \ln \Gamma\left(\frac{\nu}{2}\right). \tag{43}$$

By converting Eq. (41) into our notation we obtain

$$s^{\infty}_{\max}(x) = \exp(-2e^{B_N(\nu) - \sqrt{A_N (2N \ln 2) x}}). \tag{44}$$

In the following paragraphs we test how accurately the asymptotic approximation $s^{\infty}_{\max}(x)$ agrees with the finite-size histogram $\hat{s}_{\max}(x)$ obtained by Monte-Carlo simulations of sequences of length N ranging from 10^2 to 10^8. For each sequence length $N = 10^2$, 10^4, 10^6, and 10^8, we generate an ensemble of 10^5 quaternary ($k=4$) i.i.d. sequences of length N, and for each sequence of each ensemble we move a cursor along the sequence and compute at each position $15 \leqslant n \leqslant N-15$ the Jensen-Shannon divergence D [39]. We define $D_{\max}$ as the maximum of all values of D computed from one sequence, and by collecting all values $D_{\max}$ of each ensemble of 10^5 random i.i.d. sequences of length N we obtain the histograms $\hat{s}_{\max}(x)$ for each N.

Figure 3(a) shows the histograms $\hat{s}_{\max}(x)$ for $k=4$ and $N = 10^2$, 10^4, 10^6, and 10^8 (symbols) together with the asymptotic approximations $s^{\infty}_{\max}(x)$ (solid lines). We find that the asymptotic approximations $s^{\infty}_{\max}(x)$ are not very accurate, and that even for sequence lengths as large as $N = 10^8$ there is still a significant deviation between $\hat{s}_{\max}(x)$ and $s^{\infty}_{\max}(x)$. Figure 3(a) also shows that the deviations between $\hat{s}_{\max}(x)$ and $s^{\infty}_{\max}(x)$ are particularly large in the right tail, where we desire both distributions agree particularly well.

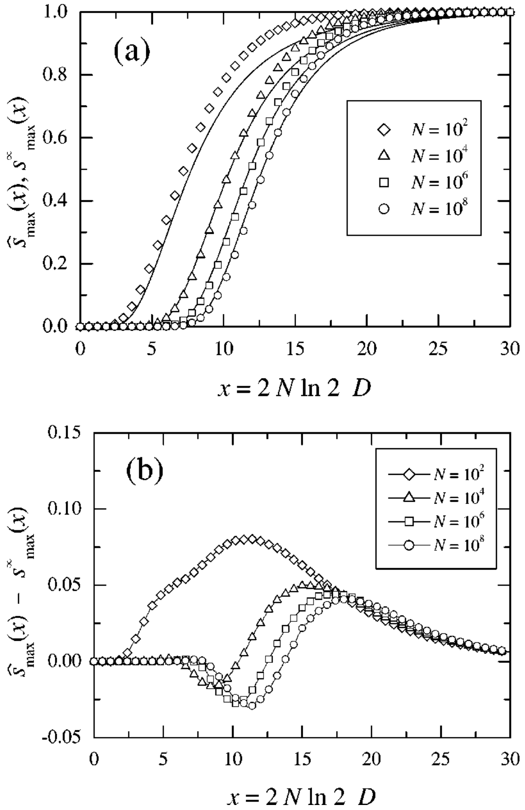

FIG. 3. Histograms $\hat{s}_{\max}(x)$ of $x = 2N(\ln 2)\; D_{\max}$ and their asymptotic approximations $s^{\infty}_{\max}(x)$ obtained from ensembles of 10^5 quaternary ($k=4$) i.i.d. sequences of length $N = 10^2$, 10^4, 10^6, and 10^8. (a) shows that the asymptotic approximations $s^{\infty}_{\max}(x)$ are not very accurate for finite-size sequences ranging in length N from 10^2 to 10^8, and that the largest deviations between $\hat{s}_{\max}(x)$ and $s^{\infty}_{\max}(x)$ occur in the right tails of the distributions. (b) shows a plot of the differences between the histograms $\hat{s}_{\max}(x)$ and their asymptotic approximations $s^{\infty}_{\max}(x)$ versus $x = 2N(\ln 2)\; D_{\max}$. We find that the accuracy of the approximations increases with increasing N, but that even for sequences of length $N = 10^8$ the deviations between $\hat{s}_{\max}(x)$ and $s^{\infty}_{\max}(x)$ are greater than 0.04.

Figure 3(b) illustrates the deviations between $\hat{s}_{\max}(x)$ and $s^{\infty}_{\max}(x)$ by plotting $\hat{s}_{\max}(x) - s^{\infty}_{\max}(x)$ versus $2N(\ln 2)x$. We find that the deviations between $\hat{s}_{\max}(x)$ and $s^{\infty}_{\max}(x)$ tend to become smaller as the sequence length N increases, but even for sequences of length $N = 10^8$ the deviations between $\hat{s}_{\max}(x)$ and $s^{\infty}_{\max}(x)$ are greater than 0.04.

As the asymptotic approximation $s^{\infty}_{\max}(x)$ is not very accurate for sequences ranging in length from $N = 10^2$ to 10^8, we recruit Monte-Carlo simulations to obtain numerical approximations of $\hat{s}_{\max}(x)$ as a function of the sequence length N and the alphabet size k. We find that the functional form of $\hat{s}_{\max}(x)$ seems to be very similar to the functional form stated in Eq. (40) if we replace the sequence length N by an *effective length* N_{eff}, and if we introduce a *scaling factor* $\beta < 1$, by which we multiply the argument of F_{k-1}.

Specifically, we find that the probability distribution of $D_{\max}$ may be approximated by

$$s_{\max}(x) \simeq [s(\beta x)]^{N_{\text{eff}}} = \{F_{k-1}[2N(\ln 2)\beta x]\}^{N_{\text{eff}}}. \quad (45)$$

N_{eff} can be understood as the effective number of independent cutting points, and the scaling factor β accomplishes that the variance of $D_{\max}$ is reduced due to correlations between the values of D computed at different positions of the same sequence.

Note that, in principle, both parameters N_{eff} and β depend on both N and k. To find an approximation of that dependence of N_{eff} and β on N and k, we perform the following simulations:

(1) We generate, for a given alphabet size k and a given sequence length N, an ensemble of 10^5 random i.i.d. sequences.

(2) For each sequence, we move a cursor along the sequence and compute at each position $15 \leqslant n \leqslant N-15$ the Jensen-Shannon divergence D [39], and we define $D_{\max}$ as the maximum of all values of D computed from one sequence.

(3) For each ensemble of 10^5 random i.i.d. sequences we obtain the histogram $\hat{s}_{\max}(x)$, and we fit the parameters N_{eff} and β of $s_{\max}(x)$ given by expression (45) to $\hat{s}_{\max}(x)$ by minimizing the Kolmogorov-Smirnov distance $|\hat{s}_{\max}(x) - s_{\max}(x)|$.

(4) We repeat the above procedure for different values of k and N.

Figure 4(a) shows the histograms $\hat{s}_{\max}(x)$ for $k=4$ and $N=10^2$, 10^4, 10^6, and 10^8 (symbols) together with the finite-size approximation $s_{\max}(x)$ obtained by the above procedure. We find by visual inspection of Fig. 4(a) and by extensive analysis of the Kolmogorov-Smirnov distances between $\hat{s}_{\max}(x)$ and $s_{\max}(x)$ for k varying from 2 to 12 and N varying from 10^2 to 10^8 that $s_{\max}(x)$ from Eq. (45) provides a good approximation of $\hat{s}_{\max}(x)$.

Figure 4(b) shows the deviations between $\hat{s}_{\max}(x)$ and $s_{\max}(x)$ by plotting $\hat{s}_{\max}(x) - s_{\max}(x)$ versus $2N(\ln 2)x$, and we find that the maximum deviation between $\hat{s}_{\max}(x)$ and $s_{\max}(x)$ stays below 0.02 for all of the cases we analyze, ranging from $k=2$ to $k=12$ and from $N=10^2$ to $N=10^8$. Moreover, we find that the maximum deviation between $\hat{s}_{\max}(x)$ and $s_{\max}(x)$ stays below 0.01 if we restrict the comparison of $\hat{s}_{\max}(x)$ and $s_{\max}(x)$ to the right tails of the distributions, where we want the approximations to be particularly accurate.

Next, we study how the parameters N_{eff} and β obtained by the fitting procedure described above depend on the alphabet size k and the sequence length N. Figure 5 shows N_{eff} and β versus N for varying values of k. First, we find that β is practically independent of N. Second, we find that for each k the effective number of cutting points N_{eff} admits a good linear fit as a function of $\ln N$, i.e.,

$$N_{\text{eff}} = a \ln N + b. \quad (46)$$

Both parameters a and b depend on the alphabet size k, and we present the least-squares values of a and b as a function of k in Table I.

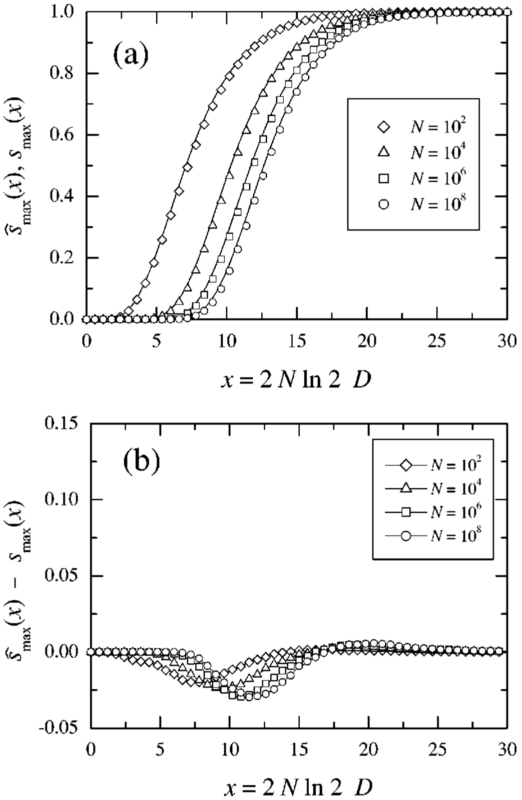

FIG. 4. Histograms $\hat{s}_{\max}(x)$ of $x = 2N(\ln 2)\,D_{\max}$ and their finite-size approximations $s_{\max}(x)$ obtained from ensembles of 10^5 quaternary ($k=4$) i.i.d. sequences of length $N=10^2$, 10^4, 10^6, and 10^8. (a) shows that the approximations $s_{\max}(x)$ are more accurate for sequences of length N ranging from 10^2 and 10^8 than the asymptotic approximations $s^{\infty}_{\max}(x)$ presented in Fig. 3, and that the largest deviations between $\hat{s}_{\max}(x)$ and $s^{\infty}_{\max}(x)$ do not occur in the right tails of the distributions, which we desire to approximate as accurately as possible. (b) shows a plot of the differences between the histograms $\hat{s}_{\max}(x)$ and their finite-size approximations $s_{\max}(x)$ versus $x = 2N(\ln 2)\ D_{\max}$. We find that the deviations between $\hat{s}_{\max}(x)$ and $s^{\infty}_{\max}(x)$ are smaller than 0.02. Moreover, we find that the deviations between $\hat{s}_{\max}(x)$ and $s^{\infty}_{\max}(x)$ are smaller than 0.01 if we restrict the comparison of $\hat{s}_{\max}(x)$ and $s_{\max}(x)$ to the tails of the distributions, which we desire to approximate as accurately as possible.

VI. APPLICATIONS OF *D*

In this section we illustrate how the results obtained in the previous sections may be used to develop an algorithm that can partition a nonstationary sequence into stationary subsequences. We describe this segmentation algorithm based on the Jensen-Shannon divergence D in detail, and we present three application examples of this recursive segmentation algorithm.

Many sequence analysis techniques rely on the stationarity of the analyzed sequence, i.e., they rely on the assumption that all portions of the sequence have at least the same composition. This *a priori* assumption is very often in conflict with experimental data, such as, for example, in case of DNA sequences [40]. The algorithm described here, which is an improved version of the algorithm presented in Refs. [13] and [18], allows us to decompose a nonstationary sequence

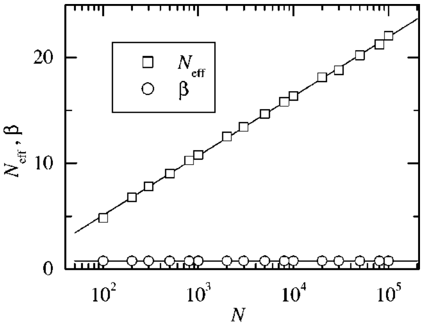

FIG. 5. Parameter values of $N_{\rm eff}$ (squares) and β (circles) as a function of the sequence length N, ranging from 200 to 10^5, for an alphabet size $k=4$. We find that β is almost independent of N, $\beta=0.80$, while $N_{\rm eff}$ admits a good linear fit to $\ln N$. The least-squares fit to $N_{\rm eff}=a \ln N+b$ yields $a=2.44$ and $b=-6.15$.

into stationary subsequences of homogeneous composition as follows:

First, we move along the sequence a cursor that divides at each position the sequence into two subsequences, and we compute D for each position of the cursor. We select that point at which D reaches its maximum value $D_{\rm max}$, and we compute its statistical significance $s_{\rm max}$. If this $s_{\rm max}$ exceeds a given threshold s_0, the sequence is cut at this point, and the procedure continues recursively for each of the two resulting subsequences. Otherwise, the sequence remains undivided. The process stops when none of the possible cutting points has a significance threshold exceeding s_0, and we say that the sequence is segmented at *significance threshold* s_0.

In the following three sections we present three examples that illustrate this recursive segmentation process.

A. Segmentation of a model sequence with known compositional domains

In order to test if the segmentation algorithm works, we generate a binary sequence of length 5×10^4 obtained by joining patches of different length and composition. We choose the sizes of the patches randomly from a power-law distribution in order to obtain a wide range of different sizes, and we choose the composition of the patches randomly from a truncated Gaussian distribution centered at 1/2.

TABLE I. Values of the parameters a, b, and β obtained by least-squares fitting of $s_{\rm max}(x)$ for three values of the alphabet size k.

k	a	b	β
2	2.96	−7.88	0.80
4	2.44	−6.15	0.80
12	2.32	−4.32	0.85

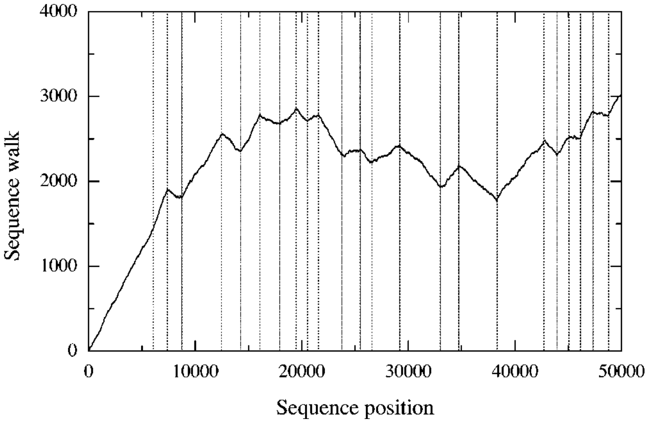

FIG. 6. Segmentation of a computer generated binary sequence of length 5×10^4 obtained by joining patches of different length and composition. The solid line represents the walk of the sequence (see text) and the vertical dotted lines represent the locations of the cuts obtained by the recursive segmentation procedure at significance threshold $s_0=95\%$. We find that the recursive segmentation procedure is indeed capable of partitioning the nonstationary input sequence into stationary subsequences at those points (vertical dotted lines) at which the local composition of the sequence changes, indicated by changes of the slope of the sequence walk (solid line).

To show graphically the variation in composition along this sequence, we plot in Fig. 6 the walk of the sequence. Given a binary sequence $\{y_i\}$, $i=1,...,N$, where y_i can assume the values $+1$ or -1, the walk of the sequence at position n is defined by [41]

$$w(n)=\sum_{i=1}^{n} y_i . \tag{47}$$

Regions with a positive slope in Fig. 6 correspond to an abundance of $+1$'s, and regions with a negative slope correspond to an abundance of -1's.

We apply the segmentation procedure presented above to this example sequence, and the vertical lines in Fig. 6 correspond to the cuts obtained by means of the segmentation procedure. Figure 6 shows clearly that the positions of the cuts coincide accurately with changes in the slope of $w(n)$. Moreover, regions without any cut do not seem to show a significant change of the slope of $w(n)$.

This observation allows us to conjecture that the subsequences obtained by the segmentation procedure are indeed homogeneous at the considered significance threshold. It is worth mentioning that the method does not rely on any initial assumption about the size distribution of the subsequences, and as we can verify by inspecting Fig. 6 the resulting subsequences have indeed a great variety of sizes.

B. Length distribution of compositionally stationary domains in prokaryotic and eukaryotic DNA

In this subsection we present one example in which we apply the recursive segmentation procedure to DNA sequences with the goal of studying the length distribution of compositionally stationary domains in prokaryotic and eukaryotic DNA. We segment at a significance threshold of $s_0=95\%$ the complete genome of the bacterium *Escherichia coli* [42] with a length of 4 639 221 base pairs (bp) as well as the human major histocompatibility complex (MHC) region of chromosome 6 [43] with a similar size of 3 673 777 bp. In both cases we use the natural four-letter alphabet $\mathcal{A}$

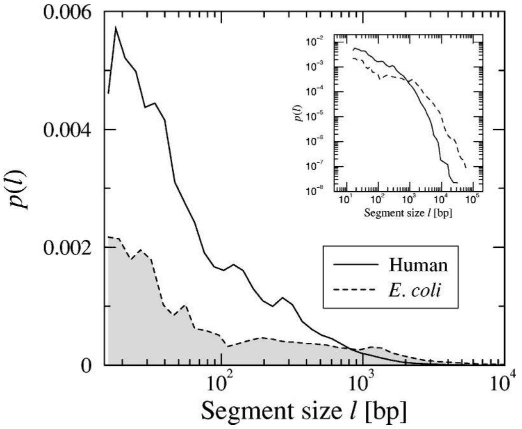

FIG. 7. Normalized distributions of segment sizes for the complete genome of the bacterium *E. coli* of length 4 639 221 bp and a contiguous human DNA sequence—the 3 673 777 bp long human MHC region of chromosome 6—of similar size. In both cases we use the natural four-letter alphabet and a significance threshold $s_0 = 95\%$. We find that the human MHC region is more heterogeneous than the *E. coli* genome, which is reflected by the longer tail (and the greater mean value) of the segment length distribution of the *E. coli* genome as compared to the human MHC region. The inset shows a double-logarithmic representation of the same distributions.

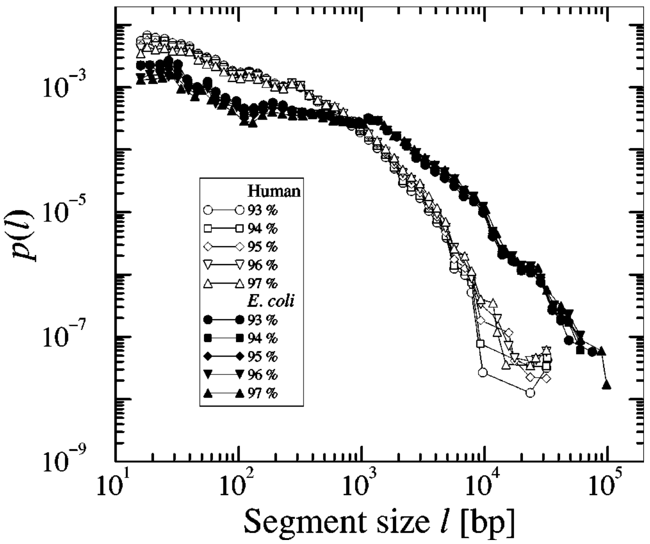

FIG. 8. Normalized distributions of segment sizes for several values of the significance threshold s_0, for the complete genome of the bacterium *E. coli* of length 4 639 221 bp and a contiguous human DNA sequence—the 3 673 777 bp long human MHC region of chromosome 6—of similar size. In all cases we use the natural four-letter alphabet, and we find that the length distributions are quite robust against changes of the significance threshold s_0.

$\equiv \{A,C,G,T\}$, where *A* stands for the base *adenine, C* stands for the base *cytosine, G* stands for the base *guanine*, and *T* stands for the base *thymine.*

We find that the recursive segmentation procedure partitions the human MHC region into 6169 segments with an average size of 595 bp, while it partitions the complete genome of the bacterium *E. coli* into 1534 segments with an average size of 3024 bp. This finding is consistent with the numbers of domains obtained by Li [44], who computes the significance threshold s_0 based on the Bayesian information criterion, and the finding that the number of domains obtained for the human MHC region is significantly greater than the number of domains obtained for the bacterium *E. coli* is consistent with reports on the presence of large compositional inhomogeneities in human DNA sequences [18,40,45].

Figure 7 shows the histogram of segment sizes for both the *E. coli* genome and the human MHC sequence. One noteworthy feature of these histograms is the high density of segments in the range below 30 bp. The high abundance of those short domains may be related to the presence of periodicities of about 10.5 bp in DNA sequences [46]. We find by inspection of the resulting segments in this small-size range that most of those short segments are made up of four types of stacks consisting of either a majority of A/T or a majority of AG/CT, respectively.

We find a weak signal indicating a second characteristic segment size in the range of 200–400 bp, which is again in agreement with previous studies [46,47]. The slower decay of the distribution of segment sizes found for the bacterium *E. coli* (inset of Fig. 7) indicates a larger abundance of long segments and seems to be a generic feature of the segment size distribution of most prokaryotes.

In order to check the robustness of the results against a change of the significance threshold s_0, we repeat the segmentation of these two sequences at different values of s_0. Figure 8 shows that the distributions are not identical, but that the main features of them, described above, remain unchanged.

C. Searching for borders between coding and noncoding DNA

In this section we describe a recently presented application of the recursive segmentation procedure to detect borders between coding and noncoding DNA [20].

One well-known statistical feature of coding regions is the nonuniform codon usage [48], which means that inside coding regions not all triplets of nucleotides (called codons) occur with the same probability. In particular, the probability p_i of finding nucleotide $a_i \in \{A,C,G,T\}$ varies from position to position [5,49,50]. This variation may originate from the restrictions imposed by the genetic code and also from some preferences in the synonymous codon usage, but irrespective of its origin, this variation is not present in noncoding DNA. Hence, this property can be used to distinguish coding from noncoding DNA, and in fact the first gene prediction programs [50] were based on the presence or absence of the positional variation of the nucleotide probabilities p_i.

In order to take into account this statistical property of coding DNA, we introduce the following 12-letter alphabet: define the phase of position n by $l \equiv n$ modulo 3. Hence, each of the nucleotides of the DNA sequences can be substituted by one of the following symbols from the alphabet $\mathcal{A}_{12} \equiv \{A_0, A_1, A_2, C_0, C_1, C_2, G_0, G_1, G_2 T_0, T_1, T_2,\}$, where, for example, T_2 denotes the nucleotide T with phase $l=2$. Using this alphabet we define the 12-letter frequency vector

$\mathbf{f}_{12} \equiv (f_{i,l})$, where $i \in \{A,C,G,T\}$, $l \in \{0,1,2\}$, and $f_{i,l}$ denotes the relative number of counts of nucleotide i in phase l.

Although coding and noncoding DNA may have the same or a similar composition when being described using the standard four-letter alphabet, the compositions given by $\mathbf{f}_{12}$ can be quite different. In noncoding DNA the probability of finding a given nucleotide is almost the same in all three phases, whereas in coding DNA this probability clearly varies from phase to phase. Even when comparing two coding regions whose starting positions are in different phases, the composition given by $\mathbf{f}_{12}$ is usually different. Hence, we propose the following modification of the segmentation procedure described above with the goal of detecting borders between coding and noncoding DNA.

Instead of computing D in terms of $\mathbf{f}_4$, we now compute D in terms of $\mathbf{f}_{12}$, and we hope that the resulting borders between stationary subsequences will be highly correlated to the borders between coding and noncoding DNA. The results obtained by segmenting complete prokaryotic genomes are fairly promising, taking into account that the segmentation procedure may be supplemented by additional biological information (see Ref. [20] for more details on the results).

A technical question related to the computation of the significance threshold for the 12-letter modification of the Jensen-Shannon segmentation procedure is worth mentioning: following Sec. V one could naively think that we should obtain $s_{\max}(x)$ from Eq. (45) with $k=12$, using Eq. (46) and the fitting parameters given in Table I. However, when using the frequency vector $\mathbf{f}_{12}$ we have to satisfy three constraints and not only one: $\Sigma_i f_{i,l} = 1/3$ for $l=0,1,2$, because the number of nucleotides in each phase is 1/3 of the total. Hence, the number of degrees of freedom is $\nu = k-3=9$, and for this case Eq. (45) reads

$$s_{\max}(x) = [s(\beta x)]^{N_{\text{eff}}} = \{F_9[\beta(2N \ln 2)x]\}^{N_{\text{eff}}}. \qquad (48)$$

By means of numerical simulations we obtain that $\beta_{3\times4}$ and N_{eff} are well fitted by $\beta_{3\times4}=0.84$ and $N_{\text{eff}} = a_{3\times4} \ln N + b_{3\times4}$, with $a_{3\times4}=2.34$ and $b_{3\times4}=-3.69$.

VII. CONCLUSIONS

One important task in analyses of experimental data is to partition a nonstationary sequence into stationary subsequences. This task is important because many statistical analysis techniques rely on the stationarity of the analyzed sequence, and the results of those analyses may be severely affected by nonstationarities of the analyzed data. Detecting nonstationarities in experimental data is nontrivial, and hence there is no standard solution to this problem. Many measures that can detect deviations from stationarity in one way or another have been proposed in the past, and one of the goals of this paper is to motivate the use of the Jensen-Shannon divergence as a measure of stationarity for symbolic sequences.

We propose to declare a sequence $\mathcal{S}$ stationary if we cannot find any point n at which $\mathcal{S}$ could be divided into two subsequences $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ with *significantly different* composition. In order to decide if the compositions of the two subsequences $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ are *different* we propose to compute the Jensen-Shannon divergence D between the two frequency vectors $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$ associated with $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$, and in order to decide if the maximum Jensen-Shannon divergence $D_{\max}$ is *significant* we propose to compute the probability that this (or a greater) value of $D_{\max}$ could have been obtained by chance.

One reason why we suggest the Jensen-Shannon divergence as a measure of stationarity is its easy interpretability in three different subfields of science. As we show in this paper, the Jensen-Shannon divergence can be interpreted as (i) the intensive mixture entropy in the framework of statistical physics, (ii) the mutual information in the framework of information theory, and (iii) the log-likelihood ratio in the framework of mathematical statistics.

In general, the weights $\pi^{(j)}$ enter the definition of the Jensen-Shannon divergence D as free parameters, which may be chosen in a problem-specific manner. It is interesting to note that all three interpretations of D suggest *one*, i.e., *the same*, natural choice of weights proportional to the sizes $n^{(j)}$ of the subsystems $\mathcal{S}^{(j)}$. Moreover, we find that this natural choice of weights makes the mean, the variance, and the probability distribution function of $2N(\ln 2)\, D$ independent of the subsystem sizes $n^{(j)}$, which is important for practical applications, where subsequences of different sizes must be compared.

We devote Sec. IV to the derivation of the mean, the variance, and the probability distribution function of $2N(\ln 2)\, D$, and we find that—for the natural choice of weights $\pi^{(j)} = n^{(j)}/N$—expressions (22) and (30) reduce to the classical results of the mean and the variance of the mixing entropy, mutual information, or log-likelihood ratio. We also show that for the naive choice of weights $\pi^{(j)} = 1/M$ the mean and the variance become singular as the length $n^{(j)}$ of at least one of the subsequences becomes very small. This singularity makes the naive choice of weights inappropriate for many practical applications, where subsequences with a wide range of different lengths $n^{(j)}$ are to be analyzed.

The natural choice of weights does not only make the mean, the variance, and the asymptotic probability distribution function of $2N(\ln 2)\, D$ independent of the subsequence lengths $n^{(j)}$, but also independent of the composition of the studied sequence. Moreover, we find that (i) the natural choice of weights minimizes the variance of D in a first-order approximation, and that (ii) with the natural choice of weights the variance of D decays as $1/N^2$ with the total sequence length N, whereas in general the variance of D decays as $1/N$. The combination of all of the above features are the reason why we prefer the natural choice of weights in our applications of the Jensen-Shannon divergence to analyses of symbolic sequences.

In order to declare a sequence stationary we require there be no point n at which the studied sequence could be partitioned into two subsequences of significantly different compositions. This requirement is the motivation for our goal of finding an approximation of the probability distribution function $s_{\max}(x) \equiv \text{Prob}\{D_{\max} \leqslant x\}$ for an ensemble of i.i.d. sequences of length N. If all of the D values computed along the same sequence were statistically independent, $s_{\max}(x)$

could be derived easily, but the nontrivial statistical dependences between the D values computed along the same sequence makes the derivation of $s_{\max}(x)$ hard.

Even in the limit of asymptotically large sequence lengths N, finding an approximation of $s_{\max}(x)$ is such a challenging problem that it could be attacked by mathematicians only in the last two decades. Pettitt, one of the pioneers in the field of change-point analysis, wrote in 1980 that "*the null distribution of the likelihood-ratio statistic is completely intractable*" [51], and it was only in 1989 when Horvath succeeded in deriving an asymptotic approximation $s^{\infty}_{\max}(x)$ of the probability distribution function of $D_{\max}$ for the special case of an ensemble of binary $(k=2)$ i.i.d. sequences [37].

One interesting feature of the asymptotic probability distribution function $s^{\infty}_{\max}(x)$ and its generalization [38] to the multinomial case is its scaling with $\ln\ln N$, which states that the expected value of $2N(\ln 2)\,D_{\max}$ diverges to infinity as $N\to\infty$, but that this divergence is extremely slow.

For practical applications the asymptotic scaling of $s^{\infty}_{\max}(x)$ is not as important as the accuracy of $s^{\infty}_{\max}(x)$ for finite N ranging from 10^2 to 10^8. The longest of the currently identified DNA sequences have a length of the order of 10^9 nucleotides, and the shortest identifiable DNA subsequences of homogeneous nucleotide composition have a length of the order of 10 nucleotides. Hence, we are interested in finding an approximation of $s_{\max}(x)$ that is accurate for lengths N ranging roughly from 10^2 to 10^8 nucleotides.

We find that the asymptotic approximation $s^{\infty}_{\max}(x)$ to the finite-size distribution $\hat{s}_{\max}(x)$ is not very accurate in that range of N, and so we recruit Monte-Carlo simulations to obtain a finite-size approximation $s_{\max}(x)$ that is more accurate than $s^{\infty}_{\max}(x)$ for N ranging from 10^2 to 10^8 and for k ranging from 2 to 12. In particular, we are interested in an approximation $s_{\max}(x)$ that is accurate in the right tail of the distribution, because this is the region where an accurate computation of the probability $s_{\max}(x)$ is needed in practical applications.

We find that $s_{\max}(x)$ may be well approximated by Eq. (45), which states that the probability distribution function of the maximum of all $N-1$ statistically *dependent* values of D computed along a sequence of length N is similar to the probability distribution function of the maximum of N_{eff} statistically *independent* random variables βD, where N_{eff} denotes the *effective* sequence length, and where β is a scaling factor that we introduce to account for the decrease of the variance of $D_{\max}$ due to correlations between the values of D computed at different positions of the same sequence. The finding that $s_{\max}(x)$ given by Eq. (45) yields an accurate approximation for N ranging from 10^2 to 10^8 and for k ranging from 2 to 12 is the central result of this paper.

When studying the dependence of β and N_{eff} on the sequence length N and the alphabet size k, we find that the scaling factor β is almost independent of both N and k, and that the effective sequence length N_{eff} admits a surprisingly accurate fit to $a\ln N+b$, where a and b are constants that depend only on the alphabet size k.

In the last section of this paper we introduce a recursive segmentation algorithm, which is an improved version of the algorithm proposed by Bernaola *et al.* [13], and which differs from the original algorithm by computing the probability of performing a segmentation step from the probability distribution function $s_{\max}(x)$ rather than from the probability distribution function $s(x)$. While the original algorithm tends to partition even a stationary sequence into domains of average size $1/(1-s_0)$, the recursive segmentation algorithm based on $s_{\max}(x)$ does not suffer from this artifact.

One question that has been raised in previous years is the question for the length distribution of compositionally homogeneous domains in DNA sequences of different organisms. Here we apply the recursive segmentation algorithm based on $s_{\max}(x)$ to the complete genome of the bacterium *E. coli* and the human MHC region on chromosome 6. Both DNA sequences have a similar length of approximately 4×10^6 nucleotides, and we find in Figs. 7 and 8 that the recursive segmentation algorithm based on $s_{\max}(x)$ yields in both cases compositionally homogeneous domains with a wide range of domain sizes. When comparing the two resulting segment size distributions to each other, we find that the human MHC region consists of more and shorter compositionally homogeneous domains than the *E. coli* genome, which is in agreement with previous findings on the complex organization of eukaryotic genomes.

In a second application example we study if the recursive segmentation algorithm could possibly be used to detect borders between coding and noncoding DNA sequences, and we find that—by choosing an appropriate representation of DNA sequences by 12 rather than four letters, encoding not only the identity of each nucleotide but also its position in the reading frame—the recursive segmentation algorithm based on $s_{\max}(x)$ can detect borders between coding and noncoding DNA sequences more accurately than conventional sliding-window techniques [20].

There is a whole plethora of problems in DNA sequence analysis that could be attacked by the recursive segmentation process, such as the identification of CpG islands or isochores, the determination of origins and termini of replication, or the detection of complex repeats or regulatory elements [52]. As the results presented in this paper are not restricted to quaternary sequences, they might possibly be useful in a wide variety of applications involving the problem of partitioning a nonstationary symbolic sequence into its stationary subsequences.

ACKNOWLEDGMENTS

We thank Sergey Buldyrev, Klaus Herman, Hanspeter Herzel, Dirk Holste, Piotr Kokoszka, Paul Krapivski, Wentian Li, Sidney Redner, Andrew Rukhin, Armin Schmitt, Edward Trifonov, Anja Victor, and Michael Zhang for valuable discussions, and the CSHL Association, the Spanish Government (Grant No. BIO99-0651-CO2-01), NIH, and NSF for financial support.

APPENDIX: APPROXIMATION OF THE ENTROPY COVARIANCE

In this appendix we derive a first-order approximation of the covariance between the entropy $H[\mathbf{f}]$ sampled from the entire sequence $\mathcal{S}$ of length N and the entropy $H[\mathbf{f}^{(j)}]$ sampled from subsequence $\mathcal{S}^{(j)}$ of length $n^{(j)}$ under the null hypothesis that $\mathcal{S}$ is an i.i.d. sequence.

We start with a Taylor expansion of $H[\mathbf{f}]$ about the vector $\mathbf{p}$, and by using the definitions

$$\hat{H} \equiv H[\mathbf{f}],$$

$$H \equiv H[\mathbf{p}],$$

$$\Delta H \equiv \hat{H} - H,$$

$$\Delta f_i \equiv f_i - p_i,$$

we obtain

$$\Delta H \simeq -\sum_{i=1}^{k} \Delta f_i \log_2 p_i - \sum_{i=1}^{k} \frac{(\Delta f_i)^2}{2 p_i \ln 2}, \qquad (A1)$$

where the symbol $\simeq$ indicates that we neglect terms of the order of $O((\Delta f_i)^3)$.

Analogously, we Taylor-expand $H[\mathbf{f}^{(j)}]$ about the vector $\mathbf{p}^{(j)}$, and by using the definitions

$$\hat{H}^{(j)} \equiv H[\mathbf{f}^{(j)}],$$

$$H^{(j)} \equiv H[\mathbf{p}^{(j)}],$$

$$\Delta H^{(j)} \equiv \hat{H}^{(j)} - H^{(j)},$$

$$\Delta f_i^{(j)} \equiv f_i^{(j)} - p_i^{(j)},$$

we obtain

$$\Delta H^{(j)} \simeq -\sum_{i=1}^{k} \Delta f_i^{(j)} \log_2 p_i^{(j)} - \sum_{i=1}^{k} \frac{(\Delta f_i^{(j)})^2}{2 p_i^{(j)} \ln 2}, \qquad (A2)$$

where the symbol $\simeq$ indicates that we neglect terms of the order of $O((\Delta f_i^{(j)})^3)$.

We next express the covariance $\mathrm{cov}(H[\mathbf{f}],H[\mathbf{f}^{(j)}])$ in terms of ΔH and $\Delta H^{(j)}$, and by using the above definitions we obtain

$$\begin{aligned}\mathrm{cov}(\hat{H},\hat{H}^{(j)}) &\equiv \langle (\hat{H}-H)(\hat{H}^{(j)}-H^{(j)})\rangle \\ &= \langle \Delta H \Delta H^{(j)}\rangle - \langle \Delta H\rangle\langle \Delta H^{(j)}\rangle. \qquad (A3)\end{aligned}$$

Since the product $\langle \Delta H\rangle\langle \Delta H^{(j)}\rangle$ is of the order of $O(1/N^2)$, we can neglect it in a first-order approximation of $\mathrm{cov}(\hat{H},\hat{H}^{(j)})$, and by plugging the Taylor expansions of Eqs. (A1) and (A2) into Eq. (A3) we obtain

$$\mathrm{cov}(\hat{H},\hat{H}^{(j)}) \simeq \sum_{g,i=1}^{k} \langle \Delta f_g \Delta f_i^{(j)}\rangle \log_2 p_g \log_2 p_i^{(j)}, \qquad (A4)$$

where the symbol $\simeq$ indicates that we neglect terms of the order of $O(1/N^2)$.

The derivation of $\langle \Delta f_g \Delta f_i^{(j)}\rangle$ is straightforward, because we can use the equalities

$$f_g = \sum_{h=1}^{m} \frac{n^{(h)}}{N} f_g^{(h)} \quad \text{and} \quad p_g = \sum_{h=1}^{m} \frac{n^{(h)}}{N} p_g^{(h)} \qquad (A5)$$

to obtain

$$\langle \Delta f_g \Delta f_i^{(j)}\rangle = \sum_{h=1}^{m} \frac{n^{(k)}}{N} \langle \Delta f_g^{(h)} \Delta f_i^{(j)}\rangle, \qquad (A6)$$

and we can work out the terms $\langle \Delta f_g^{(h)} \Delta f_i^{(j)}\rangle$ by completely elementary methods.

The product-multinomial sampling of the frequency vectors $\mathbf{f}^{(j)}$ implies that the drawing of symbol $a_g \in \mathcal{A}$ from subsequence $\mathcal{S}^{(h)}$ and the drawing of symbol $a_i \in \mathcal{A}$ from subsequence $\mathcal{S}^{(j)}$ are statistically independent, which in turn implies

$$\langle \Delta f_g^{(h)} \Delta f_i^{(j)}\rangle = \langle \Delta f_g^{(h)}\rangle\langle \Delta f_i^{(j)}\rangle = 0, \qquad (A7)$$

for all g, $i = 1,2,...,k$ and h, $j = 1,2,...,m$ with $h \neq j$. In case of $h = j$ we find

$$\langle \Delta f_g^{(j)} \Delta f_i^{(j)}\rangle = \frac{p_g^{(j)}(\delta_{gi} - p_i^{(j)})}{n^{(j)}}, \qquad (A8)$$

where δ_{gi} denotes Kronecker's delta, which is equal to 1 if $g = i$ and equal to 0 otherwise.

By plugging Eqs. (A7) and (A8) into Eq. (A6) we obtain

$$\langle \Delta f_g \Delta f_i^{(j)}\rangle = \frac{p_g^{(j)}(\delta_{gi} - p_i^{(j)})}{N}. \qquad (A9)$$

Under the null hypothesis that $\mathbf{p}^{(h)} = \mathbf{p}^{(j)}$ for all h, $j = 1,2,...,m$, Eq. (A9) simplifies to

$$\langle \Delta f_g \Delta f_i^{(j)}\rangle = \frac{p_g(\delta_{gi} - p_i)}{N}, \qquad (A10)$$

and by plugging Eq. (A10) into Eq. (A4) we obtain

$$\mathrm{cov}(\hat{H},\hat{H}^{(j)}) \simeq \frac{1}{N}\sigma^2(\log_2 p), \qquad (A11)$$

where the symbol $\simeq$ indicates that we neglect terms of the order of $O(1/N^2)$, and where $\sigma^2(\log_2 p)$ denotes the variance of the numbers $\log_2 p_i$ with respect to the probability distribution $\{p_i\}$.

[1] B. L. Hao, *Elementary Symbolic Dynamics and Chaos in Dissipative Systems* (World Scientific, Singapore, 1989).

[2] B. L. Hao, Physica D **51**, 161 (1991).

[3] C. E. Shannon and W. W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, IL, 1949).

[4] W. Li, J. Stat. Phys. **60**, 823 (1990).

[5] H. Herzel and I. Grosse, Physica A **216**, 518 (1995).

[6] H. Herzel and I. Grosse, Phys. Rev. E **55**, 800 (1997).

[7] V. E. Ramensky *et al.*, J. Comput. Biol. **7**, 1 (2000).

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

[9] A. A. Borovkov, *Mathematical Statistics* (Mir, Moscow, 1984).

[10] A. K. C. Wong and M. You, IEEE Trans. Pattern Anal. Mach. Intell. **7**, 599 (1985).

[11] J. Lin, IEEE Trans. Inf. Theory **37**, 145 (1991).

[12] M. L. Menéndez, J. A. Pardo, L. Pardo, and M. C. Pardo, J. Franklin Inst. **334B**, 307 (1997).

[13] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).

[14] W. Li, G. Stolovitzky, P. Bernaola-Galván, and J. L. Oliver, Genome Res. **8**, 916 (1998); W. Li, Phys. Rev. Lett. **86**, 5815 (2001).

[15] V. Barranco-López, P. Luque-Escamilla, J. Martínez-Aroza, and R. Román-Roldán, Electron. Lett. **31**, 867 (1995).

[16] P. Carpena and P. Bernaola-Galván, Phys. Rev. B **60**, 201 (1999).

[17] W. Li, Complexity **3**, 33 (1997).

[18] R. Roman Roldán, P. Bernaola-Galván, and J. L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).

[19] P. Bernaola-Galván, J. L. Oliver, and R. R. Román-Roldán, Phys. Rev. Lett. **83**, 3336 (1999).

[20] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).

[21] J. Burbea and C. R. Rao, IEEE Trans. Inf. Theory **28**, 489 (1982).

[22] I. Csiszár, Stud. Sci. Math. Hung. **2**, 299 (1967).

[23] T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[24] By a symbolic i.i.d. sequence we mean a sequence of independent and identically distributed (i.i.d.) symbols.

[25] M. Mansuripur, *Introduction to Information Theory* (Prentice-Hall, Englewood Cliffs, NJ, 1987).

[26] G. A. Miller, *Information Theory in Psychology*, edited by H. Quaster (Free Press, Glencoe, 1955).

[27] G. P. Basharin, Theor. Probab. Appl. **4**, 333 (1959).

[28] B. Harris, Topics Inf. Theory (Keszhtely) **16**, 323 (1975).

[29] H. Herzel, A. O. Schmitt, and W. Ebeling, Chaos, Solitons Fractals **4**, 97 (1994).

[30] M. S. Roulston, Physica D **125**, 285 (1999).

[31] W. H. Press, S. A. Teukolsky, W. T. Vettering, and B. P. Flannery, *Numerical Recipes in C* (Cambridge University Press, Cambridge, England, 1994).

[32] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1970).

[33] B. James and K. Ling, Biometrika **74**, 71 (1987).

[34] M. Pollak and D. Siegmund, Annu. Stat. Suppl. Soc. Secur Bull. **19**, 394 (1991).

[35] B. Witcher, P. Guttorp, and D. B. Percival, J. Stat. Comput. Simul. **68**, 65 (2000).

[36] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1971).

[37] L. Horvath, J. Multivariate Anal. **31**, 148 (1989).

[38] M. Csorgo and L. Horvath, *Limit Theorems in Change Point Analysis* (Wiley, New York, 1997).

[39] We choose an ‘‘excluded volume’’ of size 15 around the end points of the sequence, because the frequency vector **f** cannot be reliably computed if the length of the subsequence drops below approximately 15.

[40] C.-K. Peng *et al.*, Phys. Rev. E **49**, 1685 (1994).

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

[42] F. R. Blattner *et al.*, Science **277**, 1453 (1997).

[43] Sequence retrieved from http://www.sanger.ac.uk/HGP/Chr6/MHC.shtml.

[44] W. Li, in *Proceedings of the Fifth Annual International Conference on Computational Biology* (ACM, New York, 2001), p. 204.

[45] G. Bernardi *et al.*, Science **228**, 953 (1985).

[46] E. N. Trifonov, Physica A **249**, 511 (1998).

[47] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, and H. E. Stanley, Biophys. J. **72**, 866 (1997).

[48] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier, Nucleic Acids Res. **9**, R43 (1981).

[49] J. C. W. Shepherd, Proc. Natl. Acad. Sci. U.S.A. **78**, 1596 (1981).

[50] R. Staden and A. D. McLachlan, Nucleic Acids Res. **10**, 141 (1982); J. W. Fickett, *ibid.* **10**, 5303 (1982).

[51] A. N. Pettitt, Biometrika **67**, 79 (1980).

[52] W. Li, P. Bernaola-Galvan, F. Haghighi, and I. Grosse, Comput. Chem. (Oxford) (to be published).