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 Stanley2
1Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724
2Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215
3Departamento de Física Aplicada II, ETSI de Telecomunicación, Universidad de Málaga, E-29071 Málaga, Spain
4Departamento de Física Aplicada, Universidad de Granada, E-18071 Granada, Spain
5Departamento 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 $D_{\text{max}}$ of the Jensen-Shannon divergence $D$ sampled along a sequence becomes a quantity of central importance. Hence, we study the probability distribution of $D_{\text{max}}$ 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 $p^{(1)}=(p_1^{(1)}, p_2^{(1)}, \ldots, p_k^{(1)})$ and $p^{(2)}=(p_1^{(2)}, p_2^{(2)}, \ldots, p_k^{(2)})$ denote two probability distributions satisfying the usual constraints $\sum_{i=1}^{k}p_i^{(1)}=1$ and $0 \leq p_i^{(1)} \leq 1$ for all $i=1,2,\ldots,k$ and $j=1,2$; and let $\pi^{(1)}$ and $\pi^{(2)}$ denote the weights of the distributions $p^{(1)}$ and $p^{(2)}$, satisfying the constraints $\pi^{(1)}+\pi^{(2)}=1$ and $0 \leq \pi^{(1)}, \pi^{(2)} \leq 1$. Then the Jensen-Shannon divergence $D$ between the probability distributions $p^{(1)}$ and $p^{(2)}$ with weights $\pi^{(1)}$ and $\pi^{(2)}$ is defined by [11]

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

where

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

PHYSICAL REVIEW E, VOLUME 65, 041905

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 $D_{\text{max}}$ of the Jensen-Shannon divergence $D$ sampled along a sequence becomes a quantity of central importance. Hence, we study the probability distribution of $D_{\text{max}}$ 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 $p^{(1)}=(p_1^{(1)}, p_2^{(1)} \ldots, p_k^{(1)})$ and $p^{(2)}=(p_1^{(2)}, p_2^{(2)} \ldots, p_k^{(2)})$ denote two probability distributions satisfying the usual constraints $\sum_{i=1}^{k}p_i^{(1)}=1$ and $0 \leq p_i^{(1)} \leq 1$ for all $i=1,2,\ldots,k$ and $j=1,2$; and let $\pi^{(1)}$ and $\pi^{(2)}$ denote the weights of the distributions $p^{(1)}$ and $p^{(2)}$, satisfying the constraints $\pi^{(1)}+\pi^{(2)}=1$ and $0 \leq \pi^{(1)}, \pi^{(2)} \leq 1$. Then the Jensen-Shannon divergence $D$ between the probability distributions $p^{(1)}$ and $p^{(2)}$ with weights $\pi^{(1)}$ and $\pi^{(2)}$ is defined by [11]

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

where

DOI: 10.1103/PhysRevE.65.041905 PACS number(s): 87.15.Cc
denotes the Shannon entropy of the probability distribution \( p = (p_1, p_2, \ldots, p_k) \).

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 \( \varphi \) divergence introduced by Csiszar [12,22]. Hence, the Jensen-Shannon divergence \( D \) shares all mathematical properties of both the Jensen difference divergence and the \( \varphi \) 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 \( \varphi \) divergences. It is interesting to note that the three mathematical properties mentioned above for the binary case can be generalized to an arbitrary number of probability distributions. Let us consider \( m \) probability distributions \( p^{(1)}, p^{(2)}, \ldots, p^{(m)} \), and let us denote by \( \pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(m)} \) the corresponding weights.

The Jensen-Shannon divergence between the \( m \) probability distributions \( p^{(1)}, p^{(2)}, \ldots, p^{(m)} \) with weights \( \pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(m)} \) is given by

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

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

2. \( D \) is symmetric in its arguments \( p^{(1)}, p^{(2)}, \ldots, p^{(m)} \), i.e.,
\[
D[p^{(1)}, p^{(2)}] = D[p^{(2)}, p^{(1)}].
\]

3. \( D \) is well defined even if the probability distributions \( p^{(1)}, p^{(2)}, \ldots, p^{(m)} \) 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 \( f^{(j)} = (f_{1}^{(j)}, f_{2}^{(j)}, \ldots, f_{k}^{(j)}) \) denote the vector of molar fractions of the \( k \) gases in the \( j \)th vessel for \( j = 1, 2, \ldots, m \), and let \( n^{(i)} \) denote the total number of molecules in the \( j \)th 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

\[
H_{\text{mix}} = N k_B \ln(2) - \sum_{j=1}^{m} n^{(j)} k_B \ln(2) - H[f^{(j)}].
\]

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

\[
H_{\text{mix}} = N k_B \ln(2) D,
\]

if the weights are chosen to be \( \pi^{(j)} = n^{(j)}/N \). 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 \( i = 1, 2, \ldots, k \). 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 \( p_{ij} \) the probability of finding symbol \( a_i \) at an arbitrary but fixed position in sequence \( \mathcal{S}^{(j)} \), for \( i = 1, 2, \ldots, k \) and \( j = 1, 2, \ldots, m \).

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 ran-
dom variables \(a \in A\) and \(s \in \{S^{(1)}, S^{(2)}, \ldots, S^{(m)}\}\) are generated as follows: draw a random position \(n\) with a uniform probability distribution along the sequence \(S\), 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 \(s = S^{(j)}\) 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 \(S^{(1)}, S^{(2)}, \ldots, S^{(m)}\) with probabilities \(\pi^{(1)}=n^{(1)}/N, \pi^{(2)}=n^{(2)}/N, \ldots, \pi^{(m)}=n^{(m)}/N\), where the marginal possibilities \(p_i\) and \(\pi^{(j)}\) are defined by
\[
p_i = \sum_{j=1}^{m} p_{ij} \quad \text{and} \quad \pi^{(j)} = \sum_{i=1}^{k} p_{ij}
\]
for \(i = 1, 2, \ldots, k\) and \(j = 1, 2, \ldots, m\).

Suppose that someone is drawing a symbol \(a\) from the entire sequence \(S\), not telling us from which subsequence \(s\) this symbol was drawn, and suppose it is our task to guess the identity of that subsequence \(s\) from which symbol \(a\) was drawn, 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 \(p^{(1)}=p^{(2)}=\cdots=p^{(m)}\), 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 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 \(p_i=(p_1, p_2, \ldots, p_k)\) from a symbolic i.i.d. [24] sequence \(S\) of length \(N\), in which at each position a symbol \(a_i \in A=\{a_1, a_2, \ldots, a_k\}\) is randomly drawn with probability \(p_i\). The maximum likelihood principle suggests to choose that probability vector \(p\) which maximizes the likelihood
\[
L(S|p) = \prod_{i=1}^{k} p_i^{f_i},
\]
where \(f_i\) denotes the number of occurrences of symbol \(a_i\) in sequence \(S\). As the logarithm is a strictly monotonic function, one may equally search for that \(p\) which maximizes \(\ln L = \sum_{i=1}^{k} f_i \ln p_i\). It is easy to derive by using one Lagrange multiplier for the constraint \(\sum_{i=1}^{k} p_i = 1\) that \(p_i = F_i/N\) maximizes the log-likelihood \(\ln L\). Hence, we obtain as maximum log-likelihood
\[
\ln L_{\text{max}} = N \sum_{i=1}^{k} f_i \ln f_i = -N(\ln 2)H[f],
\]
where \(f_i = F_i/N\) denotes the relative frequency of finding symbol \(a_i\) in sequence \(S\) of length \(N\).

Now consider the slightly more complicated problem of a nonstationary sequence \(S\) of length \(N\) consisting of \(m\) stationary subsequences \(S^{(1)}, S^{(2)}, \ldots, S^{(m)}\) with lengths \(n^{(1)}, n^{(2)}, \ldots, n^{(m)}\), where the probability \(p_i^{(j)}\) of generating symbol \(a_i\) in subsequence \(S^{(j)}\) may vary from subsequence to subsequence. The likelihood of obtaining the entire sequence \(S\) is equal to the product of the likelihoods of obtain-
ing the m subsequences $S^{(1)},...,S^{(m)}$. Hence, the maximum likelihood principle suggests to choose for each $j = 1,2,...,m$ that probability vector $p^{(j)} = (p_{1}^{(j)}, p_{2}^{(j)}, ..., p_{k}^{(j)})$ that maximizes the likelihood
\[
L(S^{(j)}|p^{(j)}) = \prod_{i=1}^{k} (p_{i}^{(j)})^{f_{i}^{(j)}},
\]
where $F_{i}^{(j)}$ is the number of occurrences of symbol $a_{i}$ in subsequence $S^{(j)}$. It is again easy to derive by using m Lagrange multipliers for the m constraints $\sum_{i=1}^{k} p_{i}^{(j)} = 1$ that $p_{i}^{(j)} = F_{i}^{(j)}/n^{(j)}$ maximizes the log-likelihood $\ln L^{(j)}$. Hence, we obtain as maximum log-likelihood
\[
\ln L_{\text{max}}^{(j)} = n^{(j)} \sum_{i} f_{i}^{(j)} \ln f_{i}^{(j)} = -n^{(j)}(\ln 2)H[f^{(j)}],
\]
where $f_{i}^{(j)} = F_{i}^{(j)}/n^{(j)}$ denotes the relative frequency of finding symbol $a_{i}$ in subsequence $S^{(j)}$ 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 $\sum_{j=1}^{m} \ln L_{\text{max}}^{(j)}$ cannot be smaller than $\ln L_{\text{max}}$, 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
\[
\Delta L = \sum_{j=1}^{m} \ln L_{\text{max}}^{(j)} - \ln L_{\text{max}}
\]
is non-negative, and $\Delta L$ is commonly called the log-likelihood ratio.

It is straightforward to see from Eqs. (14), (16), and (17) that
\[
\Delta L = N(\ln 2)D.
\]

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

The inequality $\Delta L \geq 0$ 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 $\Delta L$ be significantly greater than zero, and it is the goal of this paper to derive an approximation of the probability distribution function of $\Delta L$.

Note that in all of the above interpretations of $D$ the weights of the distributions $\pi^{(1)}, \pi^{(2)}, ..., \pi^{(m)}$ are proportional to the “sizes” $n^{(1)}, n^{(2)}, ..., n^{(m)}$ 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 $p^{(1)}, p^{(2)}, ..., p^{(m)}$, 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 $p^{(j)}$ from the frequency distributions $f^{(j)} = (f_{1}^{(j)}, f_{2}^{(j)}, ..., f_{k}^{(j)})$, where $f_{i}^{(j)}$ denotes the relative frequency of symbol $a_{i}$ in subsequence $S^{(j)}$, for $i = 1,2,...,k$ and $j = 1,2,...,m$.

In all analyses of experimental data the Jensen-Shannon divergence $D$ must be computed from those (observable) frequency distributions $f^{(1)}, f^{(2)}, ..., f^{(m)}$ rather than from the (nonobservable) probability distributions $p^{(1)}, p^{(2)}, ..., p^{(m)}$.

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 $p^{(1)} = (0.45,0.55)$, and the right sequence of length $N-n = 2000$ symbols drawn from the probability distribution $p^{(2)} = (0.55,0.45)$.

We move a cursor along the entire sequence, and we compute $D$ between the subsequences at both sides of the cursor for all positions $n^{(1)} = 1,2,...,N-1$ and $n^{(2)} = N-n-1, N-n-2,...,1$. 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 $S^{(j)}$, and (ii) for the natural choice of weights $\pi^{(j)} = n^{(j)}/N$. In the following we denote by $D_{\text{m}}$ 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 $\langle D_{\text{m}} \rangle$ versus $n^{(1)}$, where the symbol $\langle \cdot \cdot \cdot \rangle$ denotes the en-
compositional differences between symbolic subsequences, whereas might serve as an appropriate divergence measure to quantify the sequences at subsequences at \( n \) and find again that \( D \) is stationary, whereas \( n = 3 \).\( D \sim 2000 \) binary sequences of length \( N \) or \( n \) subsequences of lengths \( 2 \) and \( 5 \) bits, reflecting the fact that the analyzed sequences are stationary, whereas \( D_{1/2} \) is clearly increasing as \( n^{(1)} \to 0 \) and \( n^{(1)} \to 2500 \), 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.

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 (\( D \)) 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, \( \langle H[f] \rangle \), of the entropy computed from an ensemble of finite-length sequences cannot be greater than the theoretical value, \( H[p] \), of the entropy computed from the (unobservable) probabilities [25], i.e.,

\[
\langle H[f] \rangle \leq H[p],
\]

where \( \langle \cdot \cdot \cdot \rangle \) denotes the expectation value over the ensemble of finite-length i.i.d. sequences generated by the probability distribution \( p \).

This mathematical statement is intuitively clear: due to the finite sample size, the relative frequency vector \( f \) fluctuates from sample to sample around the probability vector \( p \), and the majority of these fluctuations will make \( f \) less uniform than \( p \). Since the entropy \( H[p] \) quantifies the uniformity of the probability distribution \( p \), we expect that the majority of the values of \( H[f] \) computed from an ensemble of fluctuating frequency vectors \( f \) will be smaller than the value of \( H[p] \).
Up to first order the expected value of $H[f]$ can be approximated by [26–30]

$$
\langle H[f] \rangle \approx H[p] - \frac{k - 1}{2N \ln 2},
$$

(20)

where $k$ is the number of components of the probability and frequency vectors $p$ and $f$, $N$ is the sample size, and the symbol $\approx$ 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

$$
\langle H[f^{(j)}] \rangle \approx H[p^{(j)}] - \frac{k - 1}{2N^{(j)} \ln 2}.
$$

(21)

for $j = 1, 2, \ldots, m$, where the symbol $\approx$ 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 $D[f^{(1)}, f^{(2)}, \ldots, f^{(m)}]$ computed from an ensemble of $m$ i.i.d. sequences of total length $N$.

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

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

(22)

This expression shows that, in general, the bias $\langle D[F] \rangle - D[P]$ 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

$$
\pi^{(j)} = n^{(j)}/N,
$$

(23)

for $j = 1, 2, \ldots, m$. 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

$$
\langle D[F] \rangle \approx D[P] + \frac{k - 1}{2N \ln 2} (m - 1),
$$

(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, $\langle D \rangle = 2.9 \times 10^{-8}$ 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.
B. Variance of $D[F]$

The variance of $D[F]$ is given by

$$
\sigma^2(D[F]) = \sigma^2(H[f] - \sum_{j=1}^{m} \pi_j H[f^{(j)}])
$$

$$
= \sigma^2(H[f]) + \sum_{j=1}^{m} \pi_j^2 \sigma^2(H[f^{(j)}])
- 2 \sum_{j=1}^{m} \pi_j \text{cov}(H[f], H[f^{(j)}])
+ 2 \sum_{j=1}^{m} \sum_{l=j+1}^{m} \pi_j \pi_l \text{cov}(H[f^{(j)}], H[f^{(l)}]).
$$

(26)

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

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

$$
\sigma^2(H[f]) \approx \frac{1}{N} \sigma^2(\log_2 p),
$$

(27)

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

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

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

(28)

where $N$ denotes the length of the whole sequence, $\sigma^2(\log_2 p^{(j)})$ denotes the variance of the numbers $\log_2 p^{(j)}_i$ with respect to the probability distribution $\{p^{(j)}_i\}$ for every $j=1,2,\ldots,m$, and the symbol $\approx$ 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 $p^{(1)}=p^{(2)}=\cdots=p^{(m)}=p$ we obtain

$$
\text{cov}(H[f], H[f^{(j)}]) \approx \frac{1}{N} \sigma^2(\log_2 p)
$$

(29)

for all $j=1,2,\ldots,m$, where $\sigma^2(\log_2 p)$ denotes the variance of the numbers $\log_2 p_i$ with respect to the probability distribution $\{p_i\}$, and the symbol $\approx$ 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 $H[f]$ and $H[f^{(j)}]$ [Eq. (29)] is equal to the first-order approximation of the covariance of $H[f]$ [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 $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(m)}$,

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

(30)

under the null hypothesis that $p^{(1)}=p^{(2)}=\cdots=p^{(m)}=p$, where the symbol $\approx$ 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

$$
\langle (D[F] - D[P])^2 \rangle = \sigma^2(D) + \langle (D[E] - D[P])^2 \rangle.
$$

(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 $\sum_j \pi^{(j)} = 1$ we obtain that the set of weights $\pi^{(j)} \approx n^{(j)}/N$ 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 $\pi^{(j)} \approx n^{(j)}/N$ the variance of $D$ vanishes in $O(1/N)$. This means that for the special choice of weights $\pi^{(j)} \approx n^{(j)}/N$ the leading term of $\sigma^2(D)$ decreases with the sequence length $N$ as $1/N^2$, whereas—generally—it decreases as $1/N$. It is clear that for the special choice of weights $\pi^{(j)} \approx n^{(j)}/N$ the $O(1/N)$ term of $\sigma^2(D)$ becomes independent of both $n^{(j)}$ and $p$, 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 $p$.

In contrast, we find that for the naive choice of weights $\pi^{(j)} \approx 1/m$ 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 $D_{1/m},$

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

(32)

where $A = \sum_{j=1}^{m} 1/n^{(j)}$ 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 $D_{1m}$ 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 the significance threshold of the given value $x$, and we denote it by

$$s(x) = \text{Prob}\{D \leq x\}. \quad (33)$$

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

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

to approximate $D$ in terms of quadratic functions as follows:

$$D \approx \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{p_i^{(j)} \pi_j^{(j)} \log_2 \frac{p_i^{(j)} \pi_j^{(j)}}{p_j^{(i)} \pi_j^{(j)}}}{\ln 2} + \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(p_i^{(j)} \pi_j^{(j)} - p_j^{(i)} \pi_j^{(j)})^2}{p_j^{(i)} \pi_j^{(j)} (2 \ln 2) / \pi_j^{(j)}}. \quad (35)$$

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 $p_i^{(j)}, \pi_j^{(j)}$.

If we express the $\chi^2$ statistic [31] in the same notation, we obtain

$$\chi^2 = N \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(p_i^{(j)} \pi_j^{(j)} - p_j^{(i)} \pi_j^{(j)})^2}{p_j^{(i)} \pi_j^{(j)}} = 2 N (\ln 2) D. \quad (37)$$

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

$$s(x) = \frac{\gamma \left( \nu/2, N (\ln 2) x \right)}{\Gamma(\nu/2)}, \quad (38)$$

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 $\chi^2$ probability distribution with $\nu=(k-1)(m-1)$ degrees of freedom [6].

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

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 $2 N (\ln 2) D$ 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 $D_{\text{max}}$.

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 $D_{\text{max}}$ of the Jensen-Shannon divergence at the optimal change point is sufficiently large to partition the sequence at that point, or if the value $D_{\text{max}}$ 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 $D_{\text{max}}$.

Even if the studied sequence has been generated from a single probability distribution, we find $D_{\text{max}}>0$ due to statistical fluctuations. Moreover, we find that $D_{\text{max}}$ increases above any significance threshold $s$ computed in Sec. IV as $N$ increases. To decide if the obtained value $D_{\text{max}}=x$ 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
\[ s_{\text{max}}(x) = \text{Prob}[D_{\text{max}} \leq x]. \]  
(39)

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

\[ s_{\text{max}}(x) = s(x)^N = \{F_{k-1}[2N(\ln 2)x]\}^N, \]  
(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 \( s_{\text{max}}(x) \) 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 \( Z_N = 2N(\ln 2)D_{\text{max}} \) converges—for asymptotically large values of \( N \)—to

\[ \text{Prob}[A_N Z_N \leq (B_N(\nu) + x)^2] = \exp(-2e^{-x}), \]  
(41)

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

\[ A_N = 2 \ln \ln N, \]  
(42)

and \( B_N(\nu) \) is defined by

\[ B_N(\nu) = 2 \ln \ln N + \frac{\nu}{2} \ln \ln N - \ln \Gamma \left( \frac{\nu}{2} \right). \]  
(43)

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

\[ s_{\text{max}}^\infty(x) = \exp(-2e^{-B_N(\nu)}) \]  
(44)

\[ s_{\text{max}}^\infty(x) = \exp(-2e^{-B_N(\nu)}) \text{ for each } N. \]

In the following paragraphs we test how accurately the asymptotic approximation \( s_{\text{max}}^\infty(x) \) agrees with the finite-size histogram \( s_{\text{max}}(x) \) 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 \( 15 \leq n \leq N - 15 \) the Jensen-Shannon divergence \( D \) [39]. We define \( D_{\text{max}} \) as the maximum of all values of \( D \) computed from one sequence, and by collecting all values \( D_{\text{max}} \) of each ensemble of \( 10^5 \) random i.i.d. sequences of length \( N \) we obtain the histograms \( s_{\text{max}}(x) \) for each \( N \).

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

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

As the asymptotic approximation \( s_{\text{max}}^\infty(x) \) 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 \( s_{\text{max}}(x) \) as a function of the sequence length \( N \) and the alphabet size \( k \). We find that the functional form of \( s_{\text{max}}(x) \) seems to be very similar to the functional form stated in Eq. (40) if we replace the sequence length \( N \) by an effective length \( N_{\text{eff}} \), 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 \( D_{\text{max}} \) may be approximated by
\[ s_{\text{max}}(x) \approx \left[ s(\beta x) \right]^{N_{\text{eff}}} = \left\{ F_{k-1}[2N(\ln 2)\beta x] \right\}^{N_{\text{eff}}}. \]  

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

Note that, in principle, both parameters \( N_{\text{eff}} \) and \( \beta \) depend on both \( N \) and \( k \). To find an approximation of that dependence of \( N_{\text{eff}} \) and \( \beta \) 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 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 \( \hat{s}_{\text{max}}(x) \), and we fit the parameters \( N_{\text{eff}} \) and \( \beta \) of \( s_{\text{max}}(x) \) given by expression (45) to \( \hat{s}_{\text{max}}(x) \) by minimizing the Kolmogorov-Smirnov distance \( |\hat{s}_{\text{max}}(x) - s_{\text{max}}(x)| \).
4. We repeat the above procedure for different values of \( k \) and \( N \).

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

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

\[ N_{\text{eff}} = a \ln N + b. \]  

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.
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 $D_{\text{max}}$, and we compute its statistical significance $s_{\text{max}}$. If this $s_{\text{max}}$ 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 $5 \times 10^4$ 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.

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]

$$w(n) = \sum_{i=1}^{n} y_i.$$  \hfill (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 can be 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 $A$

---

TABLE 1. Values of the parameters $a$, $b$, and $\beta$ obtained by least-squares fitting of $s_{\text{max}}(x)$ for three values of the alphabet size $k$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$a$</th>
<th>$b$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.96</td>
<td>-7.88</td>
<td>0.80</td>
</tr>
<tr>
<td>4</td>
<td>2.44</td>
<td>-6.15</td>
<td>0.80</td>
</tr>
<tr>
<td>12</td>
<td>2.32</td>
<td>-4.32</td>
<td>0.85</td>
</tr>
</tbody>
</table>
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.

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 \( l = n \) modulo 3. Hence, each of the nucleotides of the DNA sequences can be substituted by one of the following symbols from the alphabet \( A_{12} = \{A_0,A_1,A_2,C_0,C_1,C_2,G_0,G_1,G_2,T_0,T_1,T_2\} \), where, for example, \( T_2 \) denotes the nucleotide \( T \) with phase \( l = 2 \). Using this alphabet we define the 12-letter frequency vector挽

\[挽\]
$f_{12}(i, l)$, where $i \in \{A, C, G, T\}$, $l \in \{0, 1, 2\}$, and $f_{12}$ 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 $f_{12}$ 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 $f_{12}$ 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 $f_i$, we now compute $D$ in terms of $f_{12}$, 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 highly correlated to the variance vectors $f^{(1)}$ and $f^{(2)}$ associated with $S^{(1)}$ and $S^{(2)}$, 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 $S^{(j)}$. Moreover, we find that this natural choice of weights makes the mean, the variance, and the probability distribution function of $2N/(2N)$ of independent of the system 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 $2N/(2N)$, 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 $2N/(2N)$ 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 $s_{\max}(x) = \text{Prob} \{D_{\max} \leq \alpha \}$ 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 \(s_{\text{max}}(x)\) hard.

Even in the limit of asymptotically large sequence lengths \(N\), finding an approximation of \(s_{\text{max}}(x)\) 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 \(s_{\text{max}}^\infty(x)\) of the probability distribution function of \(D_{\text{max}}\) 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 \(s_{\text{max}}^\infty(x)\) and its generalization [38] to the multinomial case is its scaling with \(\ln \ln N\), which states that the expected value of \(2N(\ln 2) D_{\text{max}}\) diverges to infinity as \(N \rightarrow \infty\), but that this divergence is extremely slow.

For practical applications the asymptotic scaling of \(s_{\text{max}}^\infty(x)\) is not as important as the accuracy of \(s_{\text{max}}^\infty(x)\) for finite \(N\) ranging from \(10^3\) to \(10^6\). The longest of the currently identified DNA sequences have a length of the order of \(10^6\) 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 \(s_{\text{max}}(x)\) that is accurate for lengths \(N\) ranging roughly from \(10^2\) to \(10^6\) nucleotides.

We find that the asymptotic approximation \(s_{\text{max}}^\infty(x)\) to the finite-size distribution \(s_{\text{max}}(x)\) is not very accurate in that range of \(N\), and so we recruit Monte-Carlo simulations to obtain a finite-size approximation \(s_{\text{max}}(x)\) that is more accurate than \(s_{\text{max}}^\infty(x)\) for \(N\) ranging from \(10^2\) to \(10^6\) and for \(k\) ranging from 2 to 12. In particular, we are interested in an approximation \(s_{\text{max}}(x)\) that is accurate in the right tail of the distribution, because this is the region where an accurate computation of the probability \(s_{\text{max}}(x)\) is needed in practical applications.

We find that \(s_{\text{max}}(x)\) 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 \(\mathcal{N}_{\text{eff}}\) statistically independent random variables \(\beta D\), where \(\mathcal{N}_{\text{eff}}\) denotes the effective sequence length, and where \(\beta\) is a scaling factor that we introduce to account for the decrease of the variance of \(D_{\text{max}}\) due to correlations between the values of \(D\) computed at different positions of the same sequence. The finding that \(s_{\text{max}}(x)\) given by Eq. (45) yields an accurate approximation for \(N\) ranging from \(10^2\) to \(10^6\) and for \(k\) ranging from 2 to 12 is the central result of this paper.

When studying the dependence of \(\beta\) and \(\mathcal{N}_{\text{eff}}\) on the sequence length \(N\) and the alphabet size \(k\), we find that the scaling factor \(\beta\) is almost independent of both \(N\) and \(k\), and that the effective sequence length \(\mathcal{N}_{\text{eff}}\) 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 \(s_{\text{max}}(x)\) 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 \(s_{\text{max}}(x)\) 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 \(s_{\text{max}}(x)\) 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 \(4 \times 10^6\) nucleotides, and we find in Figs. 7 and 8 that the recursive segmentation algorithm based on \(s_{\text{max}}(x)\) 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 \(s_{\text{max}}(x)\) 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 Koskosza, Paul Krapivsky, Wenzhian 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 \(H[f]\) sampled from the
entire sequence $S$ of length $N$ and the entropy $H[f^{(j)}]$ sampled from subsequence $S^{(j)}$ of length $n^{(j)}$ under the null hypothesis that $S$ is an i.i.d. sequence.

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

$$
\hat{H} = H[f],
$$
$$
H = H[p],
$$
$$
\Delta H = \hat{H} - H,
$$
$$
\Delta f_i = f_i - p_i,
$$
we obtain

$$
\Delta H \approx - \sum_{i=1}^{k} \Delta f_i \log_2 p_i - \sum_{i=1}^{k} \frac{(\Delta f_i)^2}{2p_i \ln 2},
$$
where the symbol $\approx$ indicates that we neglect terms of the order of $O((\Delta f_i)^3)$.

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

$$
\hat{H}^{(j)} = H[f^{(j)}],
$$
$$
H^{(j)} = H[p^{(j)}],
$$
$$
\Delta H^{(j)} = \hat{H}^{(j)} - H^{(j)},
$$
$$
\Delta f_i^{(j)} = f_i^{(j)} - p_i^{(j)},
$$
we obtain

$$
\Delta H^{(j)} \approx - \sum_{i=1}^{k} \Delta f_i^{(j)} \log_2 p_i^{(j)} - \sum_{i=1}^{k} \frac{(\Delta f_i^{(j)})^2}{2p_i^{(j)} \ln 2},
$$
where the symbol $\approx$ indicates that we neglect terms of the order of $O((\Delta f_i^{(j)})^3)$.

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

$$
\text{cov}(\hat{H},\hat{H}^{(j)}) = \langle (\hat{H} - H)(\hat{H}^{(j)} - H^{(j)}) \rangle
= \langle \Delta H \Delta H^{(j)} \rangle - \langle \Delta H \rangle \langle \Delta H^{(j)} \rangle.
$$

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 $\text{cov}(\hat{H},\hat{H}^{(j)})$, and by plugging the Taylor expansions of Eqs. (A1) and (A2) into Eq. (A3) we obtain

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

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

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

$$
f_g = \sum_{h=1}^{m} \frac{n_g^{(h)}}{N} f_g^{(h)} \quad \text{and} \quad p_g = \sum_{h=1}^{m} \frac{n_g^{(h)}}{N} p_g^{(h)}, \quad (A5)
$$
to obtain

$$
\langle \Delta f_g \Delta f_i^{(j)} \rangle = \sum_{h=1}^{m} \frac{n_g^{(h)}}{N} \langle \Delta f_g^{(h)} \Delta f_i^{(j)} \rangle, \quad (A6)
$$
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 $f^{(j)}$ implies that the drawing of symbol $a_g \in A$ from subsequence $S^{(h)}$ and the drawing of symbol $a_j \in A$ from subsequence $S^{(j)}$ are statistically independent, which in turn implies

$$
\langle \Delta f_g^{(h)} \Delta f_i^{(j)} \rangle = \langle \Delta f_g^{(h)} \rangle \langle \Delta f_i^{(j)} \rangle = 0, \quad (A7)
$$
for all $g, i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, m$ with $h \neq j$. In case of $h = j$ we find

$$
\langle \Delta f_g^{(h)} \Delta f_i^{(j)} \rangle = \frac{p_g^{(j)} (\delta_{gj} - p_i^{(j)})}{n_i^{(j)}}, \quad (A8)
$$
where $\delta_{gj}$ 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

$$
\langle \Delta f_g \Delta f_i^{(j)} \rangle = \frac{p_g^{(j)} (\delta_{gj} - p_i^{(j)})}{N}, \quad (A9)
$$
Under the null hypothesis that $p^{(h)} = p^{(j)}$ for all $h, j = 1, 2, \ldots, m$, Eq. (A9) simplifies to

$$
\langle \Delta f_g \Delta f_i^{(j)} \rangle = \frac{p_g (\delta_{gj} - p_i)}{N}, \quad (A10)
$$
and by plugging Eq. (A10) into Eq. (A4) we obtain

$$
\text{cov}(\hat{H},\hat{H}^{(j)}) = \frac{1}{N} \sigma^2 (\log_2 p), \quad (A11)
$$
where the symbol $\approx$ indicates that we neglect terms of order $O(1/N^2)$, 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\}.
[24] By a symbolic i.i.d. sequence we mean a sequence of independent and identically distributed (i.i.d.) symbols.
[28] B. Harris, Topics Inf. Theory (Keszthely) 16, 323 (1975).
[39] We choose an “excluded volume” of size 15 around the end points of the sequence, because the frequency vector cannot be reliably computed if the length of the subsequence drops below approximately 15.