Entropy and long-range memory in random symbolic additive Markov chains

S. S. Melnik and O. V. Usatenko

A. Ya. Usikov Institute for Radiophysics and Electronics Ukrainian Academy of Science, 12 Proskura Street, 61805 Kharkov, Ukraine

(Received 14 December 2015; revised manuscript received 29 April 2016; published 29 June 2016)

The goal of this paper is to develop an estimate for the entropy of random symbolic sequences with elements belonging to a finite alphabet. As a plausible model, we use the high-order additive stationary ergodic Markov chain with long-range memory. Supposing that the correlations between random elements of the chain are weak, we express the conditional entropy of the sequence by means of the symbolic pair correlation function. We also examine an algorithm for estimating the conditional entropy of *finite* symbolic sequences. We show that the entropy contains two contributions, i.e., the correlation and the fluctuation. The obtained analytical results are used for numerical evaluation of the entropy of written English texts and DNA nucleotide sequences. The developed theory opens the way for constructing a more consistent and sophisticated approach to describe the systems with strong short-range and weak long-range memory.

DOI: 10.1103/PhysRevE.93.062144

I. INTRODUCTION

Our world is complex, chaotic, and correlated. The most peculiar manifestations of this concept are human and animal communication, written texts of natural languages, DNA and protein sequences, data flows in computer networks, stock indexes, solar activity, weather, etc. For this reason, systems with long-range interactions (and/or sequences with long-range memory) and natural sequences with nontrivial information content have been the focus of a large number of studies in different fields of science for the past several decades. The unflagging interest in the systems with correlated fluctuations is also explained by the specific properties they demonstrate and their prospective applications as a creative tool for designing the devices and appliances with random components in their structure (different wave filters, diffraction gratings, artificial materials, antennas, converters, delay lines, etc. [1]).

Random sequences with a *finite number of states* exist as natural sequences (DNA or natural language texts) or arise as a result of coarse-grained mapping of the evolution of the chaotic dynamical system into a string of symbols [2,3]. Such random sequences are the subject of study of the algorithmic (Kolmogorov-Solomonoff-Chaitin) complexity, artificial intellect, information theory, compressibility of digital data, statistical inference problem, and computability, and have many application aspects mentioned above.

There are many methods for describing complex dynamical systems and random sequences connected with them: fractal dimensions, multipoint probability distribution functions, correlation functions, and many others. One of the most convenient characteristics serving the purpose of studying complex dynamics is entropy [4,5]. Being a measure of the information content and redundancy in a sequence of data, it is a powerful and popular tool in the examination of complexity phenomena. Among fields of science where the notion of entropy is of major significance, data compression [6], natural language processing [7], and artificial intelligence [8] are the most important. Recent advances in different fields of science have hinted at a deep connection between intelligence and entropy. The basic idea of compression is to exploit redundancy in data, expressed in terms of correlations, and transform this redundancy in a compression algorithm. The notion of entropy is also fundamental in the communication field. Some compression schemes are based on entropy (entropic coding), others are more likely based on algorithmic complexity (such as GZIP, based on the Lempel-Ziv complexity).

A standard method of understanding and describing statistical properties of a given random sequence of data requires the estimation of the joint probability function of words occurring for sufficiently large length L of words. For limited size sequences, reliable estimations can be achieved only for very small L because the number m^L (where m is the finite-alphabet length) of different words of the length L has to be much less than the total number $M - L$ of words in the whole sequence of the length M,

$$m^L \ll M - L \simeq M. \tag{1}$$

This is the crucial point because usually the correlation lengths R_c of natural sequences of interest are of the same order as the length of sequence. Inequality (1) cannot be fulfilled. The lengths of representative words that could correctly estimate the probability of words occurring are 4–5 for a real natural text of the length 10^6 (written on an alphabet containing 27–30 letters and symbols) or of the order of 20 for a coarse-grained text represented through a binary sequence. So, long-range memory that can exist in the sequences cannot be taken into account in such theories.

Here we present a complementary approach, which takes into account just the long-range memory. Specifically, we use an additive form of the conditional probability function. This function takes into account the weak long-range memory, which can be expressed in terms of the pair correlation function of symbols and can be found by numerical analysis of the sequence nearly at the same distances as the total length of the sequence.

We use the method developed earlier [9] for constructing the conditional probability function presented by means of a pair correlator, which makes it possible to calculate analytically the entropy of the sequence. It should be stressed that we suppose that the correlations are weak, but not short. Which kind of memory, i.e., long or short range, is more important depends on the intrinsic correlation properties of the sequence under study.

The scope of the paper is as follows. First, supposing that the correlations between symbols in the sequence are weak, we represent the conditional entropy in terms of the conditional probability function of the Markov chain and express the entropy as the sum of squares of the pair correlators. Then we discuss some properties of the results obtained. Next, a fluctuation contribution to the entropy due to finiteness of random chains is examined. The application of the developed theory to literary texts and DNA sequences of nucleotides is considered. In conclusion, some remarks on directions in which the research can progress are presented.

This work is a generalization of our previous paper [10] devoted to the binary random sequences, which we highly recommend to the reader before reading this paper.

II. ENTROPY OF THE ADDITIVE SYMBOLIC MARKOV CHAINS

Consider a semi-infinite random stationary ergodic sequence,

$$\mathbb{A} = a_0, a_1, a_2, \ldots, \tag{2}$$

of symbols (letters) a_i taken from the finite alphabet,

$$A = \{\alpha^1, \alpha^2, \ldots, \alpha^m\}, \quad a_i \in A, \quad i \in \mathbb{N}_+ = \{0, 1, 2, \ldots\}. \tag{3}$$

We use the notation a_i to indicate a position of the symbol a in the chain and the notation α^k to stress the value of the symbol $a \in A$.

We suppose that the symbolic sequence $\mathbb{A}$ is the *high-order Markov chain* [11–15]. Such sequences are also referred to as the multi- or the N-step [16–18] Markov chains, or categorical Markov chains [19]. The sequence $\mathbb{A}$ is the N-step Markov chain if it possesses the following property: the probability of symbol a_i to have a certain value $\alpha^k \in A$ under the condition that *all* previous symbols given depend only on N previous symbols,

$$P(a_i = \alpha^k | \ldots, a_{i-2}, a_{i-1}) = P(a_i = \alpha^k | a_{i-N}, \ldots, a_{i-2}, a_{i-1}). \tag{4}$$

Sometimes the number N is also referred to as the *order* or the *memory length* of the Markov chain. Note that definition (4) is valid for $i \geqslant N$; for $i < N$, we should use the well-known conditions of compatibility for the conditional probability functions (CPFs) of lower order [20].

The Markov chain with CPF of a general form, given by Eq. (4), is not convenient (compliant) to solve concrete problems. For this reason, we introduce a simplification for the CPF. Specifically, we suppose that the symbolic Markov chain under consideration is *additive*, i.e., its conditional probability is a linear function of random variables a_k, $k = i - N, \ldots, i - 1$,

$$P\left(a_i = \alpha \middle| a_{i-N}^{i-1}\right) = p_\alpha + \sum_{r=1}^{N} \sum_{\beta \in A} F_{\alpha\beta}(r)[\delta(a_{i-r}, \beta) - p_\beta], \tag{5}$$

where p_α is the relative number of symbols α in the chain, or their probabilities of occurring,

$$p_\alpha = \overline{\delta(a_i, \alpha)}. \tag{6}$$

Here, $\delta(\cdot)$ is the Kronecker delta symbol, playing the role of the characteristic function of the random variable a_i and converting symbols to numbers. Hereafter, we use the more concise notation a_{i-N}^{i-1} for N-word $a_{i-N}, \ldots, a_{i-1}$, and we often drop the superscript k from α^k to simplify the notations. It is evident that the memory function should satisfy some inequality of the type $\sum_{r=1}^{N} \sum_{\beta \in A} |F_{\alpha\beta}(r)| < \text{const}$ to provide the strict inequality (22), presented below, for arbitrary word a_{i-N}^{i}.

The additivity means that the previous symbols a_{i-N}^{i-1} exert an independent effect on the probability of the symbol $a_i = \alpha$ occurring. The first term in the right-hand side of Eq. (5) is responsible for the correct reproduction of statistical properties of uncorrelated sequences; the second one takes into account, and produces under generation, correlations among symbols of the random sequence. The conditional probability function in form (5) can correctly reproduce the binary (pair, two-point) correlations in the chain. Higher-order correlators and all correlation properties of higher orders are no longer independent. We cannot control them and correctly reproduce by means of the memory function $F_{\alpha\beta}(r)$ because the latter is completely determined by the pair correlation function; see Eq. (19) below.

The additive Markov chains are, in some sense, analogous to the chains described by autoregressive models [11,21]. In Appendix A, some suggestions on the form of Eq. (5) and its properties are presented.

To estimate the conditional entropy of stationary sequence $\mathbb{A}$ of symbols a_i, one could use the Shannon definition [4] for entropy per block of length L,

$$H_L = -\sum_{a_1, \ldots, a_L \in A} P\left(a_1^L\right) \log_2 P\left(a_1^L\right). \tag{7}$$

Here, $P(a_1^L) = P(a_1, \ldots, a_L)$ is the probability to find L-word a_1^L in the sequence. The conditional entropy, or the entropy per symbol, is given by

$$h_L = H_{L+1} - H_L. \tag{8}$$

This quantity specifies the degree of uncertainty of the $(L+1)$-th symbol occurring and measures the average information per symbol if the correlations of $(L+1)$-th symbol with preceding L symbols are taken into account. The conditional entropy h_L can be represented in terms of the conditional probability function $P(a_{L+1}|a_1^L)$,

$$h_L = \sum_{a_1, \ldots, a_L \in A} P\left(a_1^L\right) h\left(a_{L+1} \middle| a_1^L\right) = \overline{h\left(a_{L+1} \middle| a_1^L\right)}, \tag{9}$$

where $h(a_{L+1}|a_1^L)$ is the amount of information contained in the $(L+1)$-th symbol of the sequence conditioned on L previous symbols,

$$h\left(a_{L+1} \middle| a_1^L\right) = -\sum_{a_{L+1} \in A} P\left(a_{L+1} \middle| a_1^L\right) \log_2 P\left(a_{L+1} \middle| a_1^L\right). \tag{10}$$

The entropy rate (or Shannon entropy) is the conditional entropy at the asymptotic limit, $h = \lim_{L \to \infty} h_L$. This quantity measures the average information per symbol if *all* correlations, in the statistical sense, are taken into account; cf. with [22], Eq. (3).

Due to the supposed ergodicity of stationary sequence $\mathbb{A}$, the average value of any function $f(a_{r_1},a_{r_1+r_2},\ldots,a_{r_1+\cdots+r_s})$ of s arguments defined on the set A of symbols is the statistical (arithmetic, Cesaro's) average over the chain,

$$\overline{f}\,(a_{r_1},\ldots,a_{r_1+\cdots+r_s}) = \lim_{M\to\infty}\frac{1}{M}\sum_{i=0}^{M-1} f(a_{i+r_1},\ldots,a_{i+r_1+\cdots+r_s}). \tag{11}$$

Stationarity together with decay of correlations, $C_{\alpha,\beta}(r\to\infty)=0$ [see definition (13) below], leads, according to the Slutsky sufficient conditions [23], to mean ergodicity. This latter property is very useful in numerical calculations since the averaging can be done over the length of the sequence and the ensemble averaging can be avoided. Therefore, in our numerical as well as analytical calculations, we always apply averaging over the length of the sequence as implied in Eq. (9).

If the sequence, the statistical properties of which we would like to analyze, is given, then the conditional probability function of the Nth order can be found by a standard method (written below for subscript $i=N+1$),

$$P(a_{N+1}=\alpha^k|a_1,\ldots,a_N)=\frac{P(a_1,\ldots,a_N,\alpha^k)}{P(a_1,\ldots,a_N)}, \tag{12}$$

where $P(a_1,\ldots,a_N,\alpha^k)$ and $P(a_1,\ldots,a_N)$ are the probabilities of the $(N+1)$ subsequence $a_1,\ldots,a_N,\alpha^k$ and N subsequence $a_1,\ldots,a_N$ occurring, respectively.

There is a rather simple relation between the memory function $F_{\alpha\beta}(r)$ and the pair *symbolic* correlation function of the additive Markov chain. The two-point symbolic correlation function is defined as

$$C_{\alpha\beta}(r)=\overline{[\delta(a_i,\alpha)-p_\alpha][\delta(a_{i+r},\beta)-p_\beta]},\quad \alpha,\beta\in A. \tag{13}$$

This definition can also be rewritten in the form

$$\begin{aligned} C_{\alpha\beta}(r) &= \overline{[\delta(a_i,\alpha)\delta(a_{i+r},\beta)]}-p_\alpha p_\beta \\ &= P(a_i=\alpha,a_{i+r}=\beta)-p_\alpha p_\beta. \end{aligned} \tag{14}$$

The stationarity of sequence and condition of marginalization,

$$p_\alpha=\sum_{\beta\in A}P(a_i=\alpha,a_{i+r}=\beta),$$

imply that the function $C_{\alpha\beta}(r)$ possesses the following symmetric properties:

$$\begin{aligned} C_{\alpha\beta}(r) &= C_{\beta\alpha}(-r), \\ \sum_{\alpha\in A}C_{\alpha\beta}(r) &= \sum_{\beta\in A}C_{\alpha\beta}(r)=0. \end{aligned} \tag{15}$$

Let us suppose that there exists a one-to-one correspondence $a_i\leftrightarrow\varepsilon_i$ between the letters of symbolic sequence $\mathbb{A}$ and the numbers of the numeric sequence. Then, the ordinary "numeric" correlation function

$$C_\varepsilon(r)=\overline{(\varepsilon_i-\bar{\varepsilon})(\varepsilon_{i+r}-\bar{\varepsilon})} \tag{16}$$

of the sequence of ε_i can be expressed by means of symbolic correlator

$$C_\varepsilon(r)=\sum_{\alpha,\beta\in A}\varepsilon^\alpha\varepsilon^\beta C_{\alpha\beta}(r). \tag{17}$$

Here, ε^α is the numeric value of the random variable ε corresponding to the symbol α.

There were suggested two methods for finding $F_{\alpha\beta}(r)$ of a sequence with a known pair correlation function. The first one [9] is based on the minimization of the "distance,"

$$\text{Dist}=\overline{\left[\delta(a_i,\alpha)-P\left(a_i=\alpha\middle|a_{i-N}^{i-1}\right)\right]^2}, \tag{18}$$

between the conditional probability function, containing the sought-for memory function, and the given sequence $\mathbb{A}$ of symbols with a known correlation function. For any values of $\alpha,\beta\in A$ and $r\geqslant 1$, the minimization equation with respect to $F_{\alpha\beta}(r)$ yields the relationship between the correlation and memory functions (see Appendix B),

$$C_{\alpha\beta}(r)=\sum_{r'=1}^{N}\sum_{\gamma\in A}C_{\alpha\gamma}(r-r')F_{\beta\gamma}(r'). \tag{19}$$

The second method for deriving Eq. (19) given in Appendix B is a completely probabilistic straightforward calculation analogous to that used in [17].

Equation (19), despite its simplicity, can be analytically solved only in some particular cases: for one- or two-step chains, the Markov chain with a stepwise memory function, and so on. To avoid the various difficulties in solving it, we suppose that correlations in the sequence are weak (in amplitude, but not in length). We can obtain an approximate solution for the memory function in the form of the series (see Appendix C)

$$F_{\alpha\beta}(r)=\frac{C_{\beta\alpha}(r)}{p_\beta}-\frac{1}{p_\beta}\sum_{r'\neq r}\sum_{\gamma\in A}\frac{1}{p_\gamma}C_{\beta\gamma}(r-r')C_{\gamma\alpha}(r')+\cdots, \tag{20}$$

if we suppose the all components of the correlation function with $r\neq 0$ are small with respect to $C_{\alpha\beta}(0)$.

Equation (5) for the conditional probability function in the first approximation with respect to the small parameters $|C_{\alpha\beta}(r)|\ll|C_{\alpha\beta}(0)|$, $r\neq 0$, after neglecting the second term in Eq. (20), takes the form

$$P\left(a_i=\alpha\middle|a_{i-N}^{i-1}\right)\simeq p_\alpha+\sum_{r=1}^{N}\sum_{\beta\in A}\frac{C_{\beta\alpha}(r)}{p_\beta}[\delta(a_{i-r},\beta)-p_\beta]. \tag{21}$$

This formula provides a tool for constructing weak correlated sequences with a given pair correlation function [9]. Note that the i independence of the function $P(a_i=\alpha|a_{i-N}^{i-1})$ provides homogeneity and stationarity of the sequence under consideration; and the finiteness of N together with the strict inequalities

$$0<P\left(a_{i+N}=\alpha\middle|a_i^{i+N-1}\right)<1,\quad i\in\mathbb{N}_+=\{0,1,2,\ldots\}, \tag{22}$$

provides, according to the Markov theorem (see, e.g., Ref. [20]), ergodicity of the sequence.

The conditional probability $P(a_i=\alpha|a_{i-L}^{i-1})$ for a word of length $L<N$ can be obtained in the first approximation in the weak correlation parameter $\Delta_\alpha(a_{i-L}^{i-1})$ from Eqs. (5) and (21) by means of a routine probabilistic reasoning presented in

Appendix D,

$$P\left(a_i=\alpha|a_{i-L}^{i-1}\right)=p_\alpha+\Delta_\alpha\left(a_{i-L}^{i-1}\right),$$

$$\Delta_\alpha\left(a_{i-L}^{i-1}\right)=\sum_{r=1}^{L}\sum_{\beta\in A}\frac{C_{\beta\alpha}(r)}{p_\beta}[\delta(a_{i-r},\beta)-p_\beta]. \quad (23)$$

Taking into account the weakness of correlations,

$$\left|\Delta_\alpha\left(a_{i-L}^{i-1}\right)\right|\ll p_\alpha, \quad (24)$$

we expand Eq. (10) in Taylor series up to the second order in $\Delta_\alpha(a_1^L)$, $h(a_{L+1}|a_1^L)=h_0+(\partial h/\partial p_\alpha)\Delta_\alpha(a_1^L)+(1/2)(\partial^2 h/\partial p_\alpha^2)\Delta_\alpha^2(a_1^L)$, where the derivatives are taken at the point $P(a_i=\alpha|a_{i-L}^{i-1})=p_\alpha$ and h_0 is the entropy of the uncorrelated sequence,

$$h_0=-\sum_{\alpha\in A}p_\alpha\log_2(p_\alpha). \quad (25)$$

Then, the conditional entropy of the sequence in line with $\overline{\Delta_\alpha(a_1^L)}=0$ takes the form

$$h_L=\begin{cases}h_{L<N}=h_0-\dfrac{1}{2\ln 2}\displaystyle\sum_{\alpha\in A}\frac{\overline{\Delta_\alpha^2\left(a_1^L\right)}}{p_\alpha}\\ h_{L>N}=h_{L=N}.\end{cases} \quad (26)$$

If the length of block exceeds the memory length, $L>N$, the conditional probability $P(a_i=\alpha|a_{i-L}^{i-1})$ depends only on N previous symbols; see Eq. (4). Then, it is easy to show from (9) that the conditional entropy remains constant at $L\geqslant N$. Thus, the second line in Eq. (26) is consistent with the first line because, in the first approximation, in the weak correlations the parameter $\Delta_\alpha(a_{i-L}^{i-1})$ is constant at $L>N$ while the correlation function vanishes. The final expression, i.e., the main analytical result of the paper, for the conditional entropy of an infinite stationary ergodic weakly correlated random sequence of symbols is

$$h_L=h_0-\frac{1}{2\ln 2}\sum_{r=1}^{L}\sum_{\alpha,\beta\in A}\frac{C_{\alpha\beta}^2(r)}{p_\alpha p_\beta}. \quad (27)$$

In order to obtain this equation, we used Eq. (23) and replaced the term $C_{\alpha\beta}(r'-r)$ with $C_{\alpha\beta}(0)\delta(r,r')$ when calculating the summation.

III. DISCUSSION

It follows from Eq. (27) that the additional correction to the entropy h_0 of the uncorrelated sequence is negative. This is the anticipated result—the correlations decrease the entropy. The conclusion is not sensitive to the sign of correlations: persistent correlations, $C>0$, describing an "attraction" of the symbols of the same kind, and antipersistent correlations, $C<0$, corresponding to a "repulsion" between the same symbols, provide the corrections of the same negative sign. If the correlation function is constant at $1\leqslant r\leqslant N$, the entropy is a linear decreasing function of the argument L up to the point $r=N$.

Equation (27) takes a more simple form for a binary, $m=2$, chain of symbols, which can also be considered as a numeric chain of random variables a_i with the alphabet of symbols or numbers $A=\{0;1\}$. Let $p_1=\bar{a}$, $p_0=1-\bar{a}$. In order to calculate h_L, we should calculate four symbolic correlation functions:

$$\begin{aligned}C_{11}(r)&=\overline{\delta(a_i,1)\delta(a_{i+r},1)}-\bar{a}^2,\\ C_{00}(r)&=\overline{\delta(a_i,0)\delta(a_{i+r},0)}-(1-\bar{a})^2,\\ C_{01}(r)&=\overline{\delta(a_i,0)\delta(a_{i+r},1)}-(1-\bar{a})\bar{a},\\ C_{10}(r)&=\overline{\delta(a_i,1)\delta(a_{i+r},0)}-\bar{a}(1-\bar{a}).\end{aligned} \quad (28)$$

Taking into account that $\delta(a_i,1)=a_i$, $\delta(a_i,0)=1-a_i$, we obtain

$$\begin{aligned}C_{11}(r)&=C_{00}(r)=C(r),\\ C_{01}(r)&=C_{10}(r)=-C(r).\end{aligned} \quad (29)$$

Here, $C(r)$ is the ordinary numeric correlator

$$C(r)=\overline{(a_i-\bar{a})(a_{i+r}-\bar{a})}. \quad (30)$$

After simple algebra, we get

$$h_L=h_0-\frac{1}{2\ln 2}\sum_{r=1}^{L}K^2(r), \quad (31)$$

where $K(r)$ is the normalized pair correlation function of the binary sequence $K(r)=C(r)/C(0)$, the result obtained in Ref. [10]. A similar result containing only one term $K^2(L)$ for the mutual information,

$$M(L)=-\sum_{a_1,a_{L+1}\in A}P(a_1,a_{L+1})\log_2\frac{P(a_1,a_{L+1})}{p_{a_1}p_{a_{L+1}}},$$

of the binary chain was obtained earlier in Ref. [24].

IV. FINITE RANDOM SEQUENCES

The relative numbers p_α of symbols in the chain, correlation functions, and other statistical characteristics of random sequences are deterministic quantities only in the limit of their infinite lengths. It is a direct consequence of the law of large numbers. If the sequence length M is finite, the set of numbers a_1^M can no longer be considered as an ergodic sequence. In order to restore its status, we have to introduce the *ensemble* of finite sequences, $\{a_1^M\}_s$, $s\in\mathbb{N}=0,1,2,\ldots$. Yet, we would like to retain the right to examine *finite* sequences by using a single finite chain. So, for a finite chain, we should replace definition (13) of the correlation function with the following one:

$$\begin{aligned}C_{M,\alpha\beta}(r)&=\frac{1}{M-r}\sum_{i=0}^{M-r-1}[\delta(a_i,\alpha)-p_{M,\alpha}]\\ &\quad\times[\delta(a_{i+r},\beta)-p_{M,\beta}],\\ p_{M,\alpha}&=\frac{1}{M}\sum_{i=0}^{M-1}\delta(a_i,\alpha),\end{aligned} \quad (32)$$

which coincides with Eq. (13) in the limit $M\to\infty$. Now the correlation functions and the single-site probabilities $p_{M,\alpha}$ are random quantities, which depend on the particular realization of the sequence a_1^M. Fluctuations of these random quantities can contribute to the entropy of finite random chains even if the correlations in the random sequence are absent. It is well

known that the order of relative fluctuations of additive random quantity [as, e.g., the correlation function Eq. (32)] is $1/\sqrt{M}$.

Below we give a more rigorous justification of this explanation and show its applicability to our case. Let us present the correlation function $C_M(r)$ as the sum of two components,

$$C_{M,\alpha\beta}(r) = C_{\alpha\beta}(r) + C_{f,\alpha\beta}(r), \quad r \geqslant 1, \tag{33}$$

where the first summand $C_{\alpha\beta}(r) = \lim_{M\to\infty} C_{M,\alpha\beta}(r)$ is the correlation function determined by Eq. (32) (in the limit $M \to \infty$) obtained by averaging over the sequence with respect to index i, enumerating the elements a_i of sequence $\mathbb{A}$; and the second one, $C_{f,\alpha\beta}(r)$, is a fluctuation-dependent contribution. Function $C_{\alpha\beta}(r)$ can also be presented as the ensemble average $C_{\alpha\beta}(r) = \langle C_{M,\alpha\beta}(r)\rangle$ due to the ergodicity of the (infinite) sequence.

Now we can find a relationship between variances of $C_{M,\alpha\beta}(r)$ and $C_{f,\alpha\beta}(r)$. Taking into account Eq. (33) and the properties $\langle C_{f,\alpha\beta}(r)\rangle = 0$ at $r \neq 0$ and $C_{\alpha\beta}(r) = \langle C_{M,\alpha\beta}(r)\rangle$, we have

$$\left\langle C^2_{M,\alpha\beta}(r)\right\rangle = C^2_{\alpha\beta}(r) + \left\langle C^2_{f,\alpha\beta}(r)\right\rangle, \quad r \geqslant 1. \tag{34}$$

The correlation function $C_{\alpha\beta}(r)$ vanishes when r exceeds the correlation length R_c, $r \gg R_c$. This makes it possible to find the asymptotical value of $C^2_{f,\alpha\beta}(r)$,

$$\begin{aligned}\left\langle C^2_{f,\alpha\beta}(r)\right\rangle_{|r\gg R_c} &\cong \left\langle C^2_{M,\alpha\beta}(r)\right\rangle \\ &= \frac{1}{(M-r)^2}\Bigg\langle \sum_{i,j=0}^{M-r-1} [\delta(a_i,\alpha) - p_{M,\alpha}] \\ &\quad\times[\delta(a_{i+r},\beta) - p_{M,\beta}] \\ &\quad\times[\delta(a_j,\alpha) - p_{M,\alpha}][\delta(a_{j+r},\beta) - p_{M,\beta}]\Bigg\rangle.\end{aligned} \tag{35}$$

Neglecting the correlations between elements a_i and taking into account that the terms with $i = j$ give the main contribution to the result,

$$\begin{aligned}&\Bigg\langle \sum_{i,j=0}^{M-r-1} [\delta(a_i,\alpha) - p_{M,\alpha}][\delta(a_{i+r},\beta) - p_{M,\beta}] \\ &\qquad\times[\delta(a_j,\alpha) - p_{M,\alpha}][\delta(a_{j+r},\beta) - p_{M,\beta}]\Bigg\rangle \\ &\cong \sum_{i=0}^{M-r-1} \langle[\delta(a_i,\alpha) - p_{M,\alpha}]^2\rangle\langle[\delta(a_{i+r},\beta) - p_{M,\beta}]^2\rangle \\ &= (M-r)C_{f,\alpha\alpha}(0)C_{f,\beta\beta}(0),\end{aligned} \tag{36}$$

we obtain, after neglecting r in the term $M - r$, the averaged fluctuation-dependent contribution to the squared correlation function,

$$\begin{aligned}\left\langle C^2_{f,\alpha\beta}(r)\right\rangle &\simeq \frac{1}{M} C_{f,\alpha\alpha}(0)C_{f,\beta\beta}(0), \\ C_{M,\alpha\beta}(0) &= p_{M,\alpha}\delta(\alpha,\beta) - p_{M,\alpha}p_{M,\beta}.\end{aligned} \tag{37}$$

Note that Eq. (37) is obtained by means of averaging over the ensemble of chains. This is the shortest way to get the desired result. At the same time, for numerical simulations, we have only used the averaging over the chain as is seen from Eq. (32), where the summation over sites i of the chain plays the role of averaging.

Note also that the different symbols a_i in Eq. (36) are correlated. It is possible to show by direct evaluation of $C^2_{f,\alpha\beta}(r)$ with CPF (21) that the contribution of their correlations to $\langle C^2_{f,\alpha\beta}(r)\rangle$ is of the order of $\Delta/M^2 \ll 1/M$.

Equation (27), containing $C_{\alpha\beta}(r)$, is only valid for the infinite chain. In reality, we always work with sequences of finite length and can calculate $C_{M,\alpha\beta}(r)$, which contains the fluctuating part. To improve result (27), we have to subtract the fluctuating part of entropy, proportional to $\sum_{r=1}^{L}\langle C^2_{f,\alpha\beta}(r)\rangle$, from Eq. (27). Thus, Eqs. (34) and (37) yield the conditional entropy of the *finite* weakly correlated (approximately ergodic, $R_c \ll M$) random sequences,

$$h_L = h_0 - \frac{1}{2\ln 2}\left[\sum_{r=1}^{L}\sum_{\alpha,\beta\in A}\frac{C^2_{M,\alpha\beta}(r)}{p_{M,\alpha}p_{M,\beta}} - (m-1)^2\frac{L}{M}\right]. \tag{38}$$

This formula is the estimation of the conditional entropy of the additive Markov chain with the bias correction. It is clear that in the limit $M \to \infty$, this function transforms into Eq. (27). The last term in the right-hand side of Eq. (38) (the bias) describes the linearly decreasing fluctuation correction of the entropy.

For the binary chain, $m = 2$, we get the result obtained earlier in [10]. See also [24] where the bias correction was calculated for the mutual information of the binary random sequence.

Ordinarily, the bias correction for additive statistical characteristics of random finite continuous states data comes from the Edgeworth expansion [25]. In the present work, with the data being discrete states, such an approach is obviously not possible without profound modifications.

The squared correlation function $C^2_{M,\alpha\beta}(r)$ is normally a decreasing function of r, whereas the function $C^2_{f,\alpha\beta}(r)$ is nearly constant [see Eq. (37) for $r \ll M$]. Hence, the terms $\sum_{r=1}^{L}\sum_{\alpha,\beta\in A} C^2_{M,\alpha\beta}(r)/p_{M,\alpha}p_{M,\beta}$ and $(m-1)^2L/M$, being concave and linear functions, respectively, describe the competitive contributions to the entropy. It is not possible to analyze all particular cases of their relationship. Therefore, we indicate here the most interesting ones, keeping in mind monotonically decreasing correlation functions. An example of such a function is $C(r) = a/r^b$, $a > 0$, $b > 0$.

If the correlations are extremely small and compared with the inverse length M of the sequence, $\sum_{\alpha,\beta\in A} C^2_{M,\alpha\beta}(1)/p_{M,\alpha}p_{M,\beta} \sim 1/M$, the fluctuating part of the entropy exceeds the correlation part for almost all values of $L > 1$.

When the correlations are stronger, $\sum_{\alpha,\beta\in A} C^2_{M,\alpha\beta}(1)/p_{M,\alpha}p_{M,\beta} > 1/M$, there is at least one point where the contribution of the fluctuation and correlation parts of the entropy are equal. For monotonically decreasing function $\sum_{\alpha,\beta\in A} C^2_{M,\alpha\beta}(r)/p_{M,\alpha}p_{M,\beta}$, there is only one such point. Comparing the functions in square brackets in Eq. (38), we find that they are equal at some $L = R_s$, which hereafter will be referred to as a stationarity

length. If $L \ll R_s$, the fluctuations of the correlation function are negligibly small with respect to its magnitude, and hence for these L-words the finite sequence may be considered as the quasistationary one. At $L \sim R_s$, the fluctuations are of the same order as the genuine correlation function contribution, $\sum_{\alpha,\beta\in A} C^2_{M,\alpha\beta}(r)/p_{M,\alpha}p_{M,\beta}$. Here we have to take into account the fluctuation correction due to the finiteness of the random chain. At $L > R_s$, the fluctuation contribution exceeds the correlation one and Eq. (38) loses its meaning.

The other important parameter of the random sequence is the memory length N. If the length N is less than R_s, we have no difficulties to calculate the entropy of the finite sequence, which can be considered as quasistationary. If the memory length exceeds the stationarity length, $R_s \lesssim N$, we should take into account the fluctuation correction to the entropy.

V. APPLICATIONS TO NATURAL AND DNA TEXTS

The purpose of this section is to illustrate the applicability of the developed theory to some concrete sequences naturally arising in biology and linguistics.

In order to evaluate the conditional entropy of literature works, we calculate the probabilities $p_{M,\alpha}$ of each letter occurring in the simplified text and symbolic correlation functions $C_{M,\alpha\beta}(r)$. The simplification (some sort of coarse graining) consists of replacing all of the uppercase letters with the lowercase ones and neglecting all punctuation marks except blanks. Hence, we use the alphabet of 27 letters. The result for calculating the conditional entropy with the use of Eq. (38) is shown in Fig. 1. The entropy per one letter $h(0)$ (not shown in the picture) is 4 ± 0.1. It is evident that the difference between the one-letter entropy, in the case of the letters equipartition $\log_2 27 \approx 4.75$, and 4 ± 0.1 is due to the nonequipartition distribution of letters in the texts.

As we mentioned, the correlation length can be determined as the length where the entropy takes on a constant value. At first glance, the value of R_c is of the order of 9–11. But after this point, we observe a nearly linear small decrease of entropy extended over 2–3 decades. Probably, this phenomenon could be explained by small power-low correlation, observed and discussed in Ref. [17].

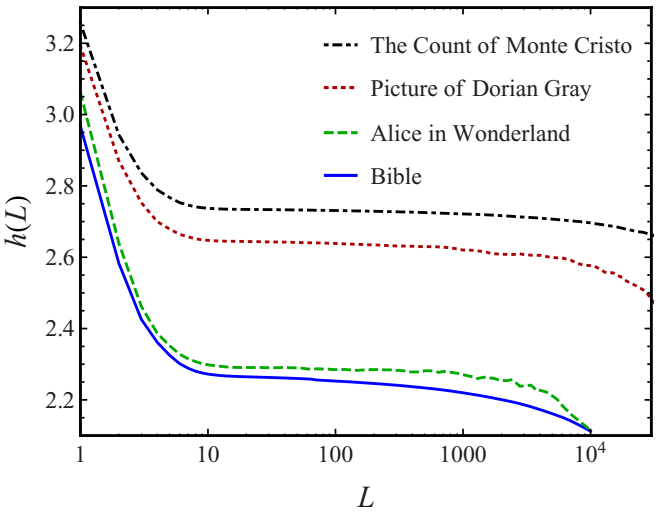

FIG. 1. The conditional entropy of the literature works (indicated in the legend near the curves) vs the length of words in the L-axis log scale. The curves correspond to the direct evaluations of Eq. (27) with fluctuation correction.

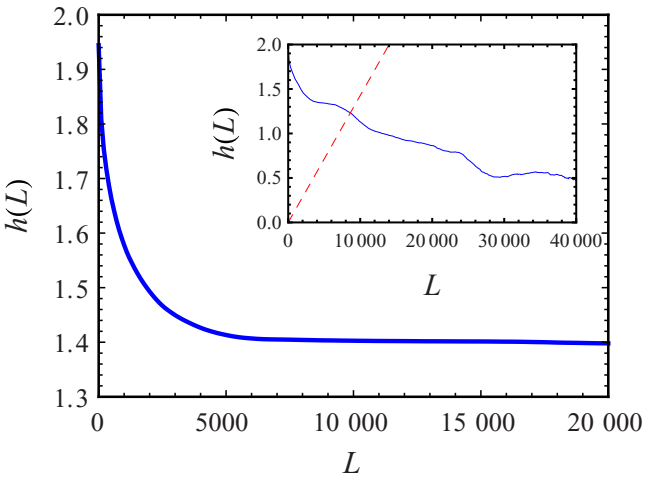

FIG. 2. The conditional entropy of *Homo sapiens* chromosome Y, locus NW 001842422 [26], of length $M \simeq 3.9 \times 10^6$ vs length L with the fluctuation correction. The curve is constructed using Eq. (27). The inset demonstrates the conditional entropy of *Homo sapiens* chromosome Y, locus NW 001842451, of length $M \simeq 4.5 \times 10^4$. The straight dashed line is fluctuation correction $9L/2\ln 2\, M$ due to finiteness of chain.

Application of the developed theory to nucleotide sequences of DNA molecules is shown in Fig. 2. In order to evaluate the entropy of the *Homo sapiens* chromosome Y, locus NW 001842422 [26], we calculate the probabilities $p_{M,\alpha}$ of each nucleotide occurring in the sequence and 9 different symbolic correlation functions $C_{M,\alpha\beta}(r)$.

It is clearly seen that the entropy in the interval $7 \times 10^3 < L < 2 \times 10^4$ takes on the constant value, $h_L \simeq 1.41$. It means that for $L > 7 \times 10^3$, all binary correlations, in the statistical sense, are taken into account. In other words, the correlation length of the *Homo sapiens* chromosome Y is of the order of 10^4. This length R_c is much greater than correlation length $R_c \approx 10$ observed for natural written texts.

In the inset of Fig. 2, the conditional entropy of *Homo sapiens* chromosome Y, locus NW 001842451, is shown. Here we cannot see a constant asymptotical region, which would be evidence for the existence of stationarity and finiteness of the correlation length. We suppose that the locus is not well described by our theory at long distances due to the relatively short length of sequence. The dashed line in the figure is the fluctuation correction of the conditional entropy. This correction should be small with respect to the correlation contribution in the region of reliability of the result. Thus, only for $L < 10^3$, the result can be considered as plausible.

It is interesting to compare our results with those obtained by estimation of block entropy (7) where the probabilities of words occurring are calculated with the standard likelihood estimate

$$P\left(a_1^L\right) = \frac{n\left(a_1^L\right)}{M - L + 1}. \tag{39}$$

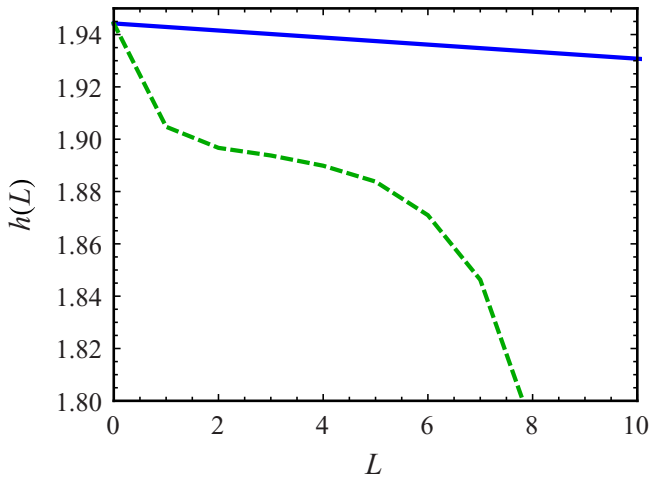

FIG. 3. Comparison of conditional entropies calculated by estimation of block occurring, Eq. (39) (bottom curve), and the result of Eq. (38) (top curve) for the *Homo sapiens* chromosome Y, locus NW 001842422.

Here, $n(a_1^L)$ is the number of occurrences of the word a_1^L in the sequence of the length M. In our paper [10], it was shown that there is a good agreement between the two approaches for the coarse-grained (binary) DNA sequence of R3 chromosome of *Drosophila melanogaster* of length $M \simeq 2.7 \times 10^7$ for $L \lesssim 5$–6 units. For the four-valued sequence (composed of adenine, guanine, cytosine, and thymine), we cannot draw a similar conclusion studying the conditional entropy of the *Homo sapiens* chromosome Y, locus NW 001842422, shown in Fig. 3. It is clear that at small L, strong short-range correlations or the exact statistics of the short words are more important than that which we took into account—the simple pair correlations.

It is difficult to come to an unambiguous conclusion as to which factor, i.e., the finiteness of the chain and violation of Eq. (1) or the strength of correlations, is more important for the discrepancy between the two theories and between the two studied sequences. Probably this is due to the triplet structure of the DNA molecules and strong correlations at short distances, $L \simeq 2 \div 3$.

VI. CONCLUSION AND PERSPECTIVES

(i) The main result of the paper, i.e., the conditional entropy of the stationary ergodic weakly correlated random sequence $\mathbb{A}$ with elements belonging to the finite alphabet, is given by Eq. (27). The other important point of the work is the calculation of the fluctuation contribution to the entropy due to the finiteness of random chains, which is the last term in Eq. (38).

(ii) In order to obtain Eq. (27), we used an assumption that the random sequence of symbols is the high-order Markov chain. Nevertheless, the final result contains only the correlation function and does not contain the conditional probability function of the Markov chain. This allows us to suppose that result (27) and the region of its applicability is wider than the assumptions under which it is obtained.

(iii) To obtain Eq. (27), we supposed that the correlations in the random chain are weak. It is not a very severe restriction. Many examples of such systems described by means of the pair correlator are given in Ref. [1]. The randomly chosen example of DNA sequences and the literary texts support this conclusion. The strongly correlated systems, opposed to weakly correlated chains, are nearly deterministic. For their description, we need a completely different approach. Their study is beyond the scope of this paper.

(iv) Equation (27) can be considered as an expansion of the entropy in series with respect to the small parameter Δ, where the entropy h_0 of the noncorrelated sequence is the zero approximation. Alternatively, for the zero approximation, we can use the exactly solvable model of the N-step Markov chain with the conditional probability function of words occurring taken in the form of the stepwise function [18]. Another way to choose the zero approximation can be based on the CPF obtained from probability of the block occurring, Eq. (7). Consequently, the developed theory opens the way to construct a more consistent and sophisticated approach describing the systems with strong short-range and weak long-range memory.

(v) Our consideration can be generalized to the Markov chain with the infinite memory length N. In this case, we should impose the condition of the decreasing rate of the correlation function and the conditional probability function at $N \to \infty$.

ACKNOWLEDGMENT

We are grateful for the helpful and fruitful discussions with G. M. Pritula, S. S. Apostolov, and Z. A. Maizelis.

APPENDIX A

The conditional probability function of the *binary additive* Markov chain of random variables $a_i \in \{0,1\}$, i.e., the probability of symbol a_i to have a value 1 under the condition that N previous symbols a_{i-N}^{i-1} are given, is of the following form [9,16]:

$$P\left(a_i = 1 \middle| a_{i-N}^{i-1}\right) = \bar{a} + \sum_{r=1}^{N} F(r)(a_{i-r} - \bar{a}). \tag{A1}$$

Analogously for $P(0|.)$,

$$P\left(a_i = 0 \middle| a_{i-N}^{i-1}\right) = 1 - P\left(1 \middle| a_{i-N}^{i-1}\right) = 1 - \bar{a} - \sum_{r=1}^{N} F(r)(a_{i-r} - \bar{a}). \tag{A2}$$

These two expressions are not symmetric with respect to the change $0 \leftrightarrows 1$ of generated symbol a_i. Let us show that Eqs. (A1) and (A2) can be presented in the symmetric form

$$P\left(a_i = \alpha | a_{i-N}^{i-1}\right) = p_\alpha + \sum_{r=1}^{N} \sum_{\beta \in \{0,1\}} F_{\alpha\beta}(r)[\delta(a_{i-r}, \beta) - p_\beta]. \tag{A3}$$

Taking into account the definitions $p_1 = \bar{a}$, $p_0 = 1 - \bar{a}$, using the evident equalities $\delta(a_{i-r}, 0) = 1 - a_{i-r}$, $\delta(a_{i-r}, 1) = a_{i-r}$, and putting $F_{11}(r) - F_{10}(r) = F_{00}(r) - F_{01}(r) = F(r)$, we easily obtain Eqs. (A1) and (A2). We should replace $\alpha, \beta \in \{0,1\}$ in Eq. (A3) by $\alpha, \beta \in A$ to obtain Eq. (5).

Note that there is no one-to-one correspondence between the memory function $F_{\alpha\beta}(r)$ and the conditional probability function $P(a_i = \alpha|a_{i-N}^{i-1})$. Indeed, it is easy to see that in view of Eqs. (5) and (6), the renormalized memory function $F'_{\alpha\beta}(r) = F_{\alpha\beta}(r) + \varphi_\alpha(r)$ provides the same conditional probability as $F_{\alpha\beta}(r)$.

Note that Eq. (5) can be considered as an approximate model expression simplifying the general form of the conditional probability function. As a matter of fact, the conditional probability (12) of the symbolic sequence of random variables $a_i \in \mathcal{A}$ can be represented exactly as a *finite* polynomial series containing N Kroneker delta symbols: a specific decomposed form of the CPF, which expresses some "independence" of the random variables a and spatial coordinates i,

$$P(.|.) = P(a_i = \alpha|a_{i-N},\ldots,a_{i-2},a_{i-1}) = p_\alpha + \sum_{\beta_1\ldots\beta_N\in\mathcal{A}}\sum_{r_1\ldots r_N} F_{\alpha;\beta_1\ldots\beta_N}(r_1,\ldots,r_N) \times \prod_{s=1}^{N}\delta(a_{i-r_s},\beta_s). \quad \text{(A4)}$$

Here, the arguments $r_1,\ldots,r_N$ of the function $F_{\alpha;\beta_1\ldots\beta_N}(r_1,\ldots,r_N)$, supposed to be ordered $r_1 \leqslant r_2 \leqslant ,\ldots, \leqslant r_{N-1} \leqslant r_N$, indicate the distances between the final "generated" symbol $a_i = \alpha$ and symbols $a_{i-1} = \beta_1,\ldots,a_{i-N} = \beta_N$. It is clear that there is one-to-one correspondence between $P(a_i = \alpha|a_{i-N}^{i-1})$ and the function $F_{\alpha;\beta_1\ldots\beta_N}(r_1,\ldots,r_N)$, which is referred to as the *generalized* memory function.

Hosseinia *et al.* [19] proved rigorously that the conditional probability can be written as a linear combination of the monomials of past process responses for the Markov chain. Earlier this idea was presented in Besag's paper [27].

APPENDIX B

The method for finding the memory function $F_{\alpha\beta}(r)$ of a sequence with a known pair correlation function is based on the minimization of the "distance" between the conditional probability function, containing the sought-for memory function, and the given sequence $\mathbb{A}$ of symbols with a known correlation function,

$$\text{Dist} = \overline{\left[\delta(a_i,\alpha) - P\left(a_i = \alpha\middle|a_{i-N}^{i-1}\right)\right]^2}, \quad \text{(B1)}$$

where the conditional probability $P(a_i = \alpha|a_{i-N}^{i-1})$ is defined by Eq. (5).

Let us express the distance in terms of the correlation functions (13). From Eqs. (5) and (B1), one obtains

$$\text{Dist} = \overline{[\delta(a_i,\alpha) - p_\alpha]^2} - 2\sum_{r=1}^{N}\sum_{\beta\in A} F_{\alpha\beta}(r)\overline{[\delta(a_i,\alpha) - p_\alpha][\delta(a_{i-r},\beta) - p_\beta]} + \sum_{r,r'=1}^{N}\sum_{\beta,\gamma\in A} F_{\alpha\beta}(r)F_{\alpha\gamma}(r') \times\overline{[\delta(a_{i-r},\beta) - p_\beta][\delta(a_{i-r'},\gamma) - p_\gamma]},$$

or, replacing the averages by corresponding correlation functions,

$$\text{Dist} = C_{\alpha\alpha}(0) - 2\sum_{r=1}^{N}\sum_{\beta\in A} F_{\alpha\beta}(r)C_{\beta\alpha}(r) + \sum_{r,r'=1}^{N}\sum_{\beta,\gamma\in A} F_{\alpha\beta}(r)F_{\alpha\gamma}(r')C_{\gamma\beta}(r'-r).$$

The minimization equation

$$\frac{\partial\text{Dist}}{\partial F_{\alpha\beta}(r)} = -2C_{\beta\alpha}(r) + 2\sum_{r'=1}^{N}\sum_{\gamma\in A} F_{\alpha\gamma}(r')C_{\gamma\beta}(r'-r) = 0 \quad \text{(B2)}$$

yields the relationship (19) between the correlation and memory functions.

Another way to derive Eq. (19) is a completely probabilistic straightforward calculation. Let us rewrite the correlation function (14) in terms of conditional probability $f_{\beta\alpha}(r) \equiv P(a_i = \alpha|a_{i-r} = \beta)$,

$$C_{\beta\alpha}(r) = P(a_i = \alpha,a_{i-r} = \beta) - p_\alpha p_\beta = p_\beta[f_{\beta\alpha}(r) - p_\alpha]. \quad \text{(B3)}$$

Obviously, the conditional probability can be expressed via the summation over all the variants of previous N-word $W = a_{i-N+1}^{i-1}$,

$$f_{\beta\alpha}(r) = \sum_W P(a_i = \alpha|W)P(W|a_{i-r} = \beta). \quad \text{(B4)}$$

Taking into account the additive form of the first multiplier (5), one can reverse the order of summations,

$$f_{\beta\alpha}(r) = \sum_W\left\{p_\alpha + \sum_{r'=1}^{N}\sum_{\gamma\in A} F_{\alpha\gamma}(r')[\delta(a_{i-r'},\gamma) - p_\gamma]\right\} \times P(W|a_{i-r} = \beta) = p_\alpha + \sum_{r'=1}^{N}\sum_{\gamma\in A} F_{\alpha\gamma}(r')\sum_W[\delta(a_{i-r'},\gamma) - p_\gamma] \times P(W|a_{i-r} = \beta).$$

Noting that $a_{i-r'}$ is one of the symbols of the word W, we conclude that the sum $\sum_W \delta(a_{i-r'},\gamma)P(W|a_{i-r} = \beta)$ is the conditional probability $P(a_{i-r'} = \gamma|a_{i-r} = \beta) = f_{\beta\gamma}(r-r')$.

Thus we obtain an equation for the values of f,

$$f_{\beta\alpha}(r) - p_\alpha = \sum_{r'=1}^{N}\sum_{\gamma\in A} F_{\alpha\gamma}(r')[f_{\beta\gamma}(r-r') - p_\gamma]. \quad \text{(B5)}$$

Multiplying it by p_β and expressing the $f_{\beta\alpha}(r)$ via $C_{\beta\alpha}(r)$ using (B3), we derive the desired relation (19).

APPENDIX C

Using definition (13) of the correlation function and its property,

$$C_{\alpha\beta}(0) = p_\alpha\delta_{\alpha\beta} - p_\alpha p_\beta, \quad \text{(C1)}$$

it is convenient to separate the term with $r' = r$ in Eq. (19),

$$C_{\alpha\beta}(r) = \sum_{\gamma\in A} C_{\alpha\gamma}(0) F_{\beta\gamma}(r) + \sum_{r'\neq r} \sum_{\gamma\in A} C_{\alpha\gamma}(r-r') F_{\beta\gamma}(r'). \tag{C2}$$

After using the symmetric properties of matrices $C_{\alpha\gamma}(0)$ and $F_{\beta\gamma}(r)$, we simplify the first term of the previous equation,

$$C_{\alpha\beta}(r) = p_\alpha F_{\beta\alpha}(r) + \sum_{r'\neq r} \sum_{\gamma\in A} C_{\alpha\gamma}(r-r') F_{\beta\gamma}(r'), \tag{C3}$$

and obtain the recurrent relation for the memory function,

$$F_{\alpha\beta}(r) = \frac{C_{\beta\alpha}(r)}{p_\beta} - \frac{1}{p_\beta} \sum_{r'\neq r} \sum_{\gamma\in A} C_{\beta\gamma}(r-r') F_{\alpha\gamma}(r'). \tag{C4}$$

In the case of weak correlations, the second term in the right-hand side of the equation is much smaller then the first one; then the first approximation for the memory function is

$$F_{\alpha\beta}(r) = \frac{C_{\beta\alpha}(r)}{p_\beta}. \tag{C5}$$

Substituting this result into the recurrent equation (C4), we obtain the second approximation for the $F_{\alpha\beta}(r)$,

$$F_{\alpha\beta}(r) = \frac{C_{\beta\alpha}(r)}{p_\beta} - \frac{1}{p_\beta} \sum_{r'\neq r} \sum_{\gamma\in A} \frac{1}{p_\gamma} C_{\beta\gamma}(r-r') C_{\gamma\alpha}(r'). \tag{C6}$$

APPENDIX D

Here we prove Eq. (23) using Eqs. (5) and (21) as a starting point. It follows from definition (12) of the conditional probability function,

$$P(a_i = a|W) = \frac{P(W,a)}{P(W)}, \quad W = a_{i-N+1}^{i-1}. \tag{D1}$$

Adding symbol $a_{i-N} = b$ to the string (W,a), we have

$$P(a_i = a|W) = \frac{\sum_{b\in A} P(b,W,a)}{P(W)}. \tag{D2}$$

Replacing here the probabilities $P(b,W,a)$ by the CPF $P(a_i = a|b,W)$ from the equation similar to that of Eq. (D1),

$$P(a_i = a|b,W) = \frac{P(b,W,a)}{P(b,W)}, \tag{D3}$$

we obtain, after some algebraic manipulations,

$$P(a_i = a|W) = p_a + \sum_{r=1}^{N-1} \sum_{b\in A} F_{ab}(r)[\delta(a_{i-r},b) - p_b] + \frac{1}{P(W)} \times \sum_{c\in A} F_{ac}(N) \sum_{b\in A} P(b,W)[\delta(b,c) - p_c]. \tag{D4}$$

The third term containing summation over b is of the form

$$P(c,W)(1-p_c) - P[\mathtt{not}(c),W]p_c, \tag{D5}$$

where the symbol $\mathtt{not}(c)$ stands for a complementary event to c. It is intuitively clear that in the zero approximation in Δ (i.e., for uncorrelated sequence), this term equals zero. In the next approximation, this term is of the order of Δ. These two statements can be verified by using the condition of compatibility for the Chapman-Kolmogorov equation (see, for example, Ref. [28]),

$$P\left(a_{i-N+1}^{i}\right) = \sum_{a_{i-N}\in A} P\left(a_{i-N}^{i-1}\right) P_N\left(a_i \middle| a_{i-N}^{i-1}\right). \tag{D6}$$

Hence, we have to neglect the third term in the right-hand side of Eq. (D4) because it is of the second order in Δ. So, Eq. (23) is proven for $L = N-1$. By induction, the equation can be written for arbitrary L.

[1] F. M. Izrailev, A. A. Krokhin, and N. M. Makarov, Phys. Rep. **512**, 125 (2012).
[2] P. Ehrenfest and T. Ehrenfest, *Encyklopädie der Mathematischen Wissenschaften* (Springer, Berlin, 1911), p. 742, Bd. II.
[3] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding* (Cambridge University Press, Cambridge, 1995).
[4] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, IL, 1949).
[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[6] D. Salomon, *A Concise Introduction to Data Compression* (Springer, Berlin, 2008).
[7] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval* (Cambridge University Press, Cambridge, 2008).
[8] A. D. Wissner-Gross and C. E. Freer, Phys. Rev. Lett. **110**, 168702 (2013).
[9] S. S. Melnyk, O. V. Usatenko, and V. A. Yampol'skii, Physica A **361**, 405 (2006).
[10] S. S. Melnik and O. V. Usatenko, Phys. Rev. E **90**, 052106 (2014).
[11] A. Raftery, J. R. Stat. Soc. B **47**, 528 (1985).
[12] W. K. Ching, E. S. Fung, and M. K. Ng, Naval Res. Logist. **51**, 557 (2004).
[13] W. K. Li and M. C. O. Kwok, Commun. Stat. Simul. Comput. **19**, 363 (1990).
[14] J. A. Cocho *et al.* Comput. Biol. Chem. **53**, 15 (2014).
[15] M. Seifert, A. Gohr, M. Strickert, and I. Grosse, PLoS Comput. Biol. **8**, e1002286 (2012).
[16] O. V. Usatenko, S. S. Apostolov, Z. A. Mayzelis, and S. S. Melnik, *Random Finite-Valued Dynamical Systems: Additive Markov Chain Approach* (Cambridge Scientific, Cambridge, 2010).
[17] S. S. Melnyk, O. V. Usatenko, V. A. Yampol'skii, and V. A. Golick, Phys. Rev. E **72**, 026140 (2005).
[18] O. V. Usatenko and V. A. Yampol'skii, Phys. Rev. Lett. **90**, 110601 (2003).
[19] R. Hosseinia, N. Leb, and J. Zideka, J. Stat. Theory Prac. **5**, 261 (2011).
[20] A. N. Shiryaev, *Probability* (Springer, New York, 1996).
[21] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, EURASIP J. Appl. Signal Process. **1**, 13 (2004).

[22] P. Grassberger, arXiv:physics/0207023.
[23] See, e.g., A. M. Yaglom, *Correlation Theory of Stationary and Related Random Functions* (Springer-Verlag, New York, 1987).
[24] W. Li, J. Stat. Phys. **60**, 823 (1990).
[25] O. J. J. Michel and P. Flandrin, Signal Proc. **53**, 133 (1996).
[26] ftp://ftp.ncbi.nih.gov/genomes/ (unpublished).
[27] J. Besag, J. R. Stat. Soc. B **36**, 192 (1974).
[28] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*, Springer Series in Synergetics, Vol. 13 (Springer-Verlag, Berlin, 1985).