PHYSICAL REVIEW E **98**, 042144 (2018)

Symbolic high-order Markov chains: Entropy and compressibility

S. S. Melnik and O. V. Usatenko
A. Ya. Usikov Institute for Radiophysics and Electronics Ukrainian Academy of Science, 12 Proskura Street, 61805 Kharkiv, Ukraine

(Received 28 January 2018; revised manuscript received 26 April 2018; published 29 October 2018)

Using the bilinear Markov chain approach, we study statistical properties of natural random symbolic sequences with complex correlation properties. In the limit of weak correlations, we present analytically the entropy of sequence by means of correlation functions of the second and third orders. We illustrate the applicability of the developed theory to some sequences naturally arising in biology and chaotic dynamics. We evaluate numerically the entropy of DNA nucleotide sequences and sequences obtained by dichotomization of a logistic map. Using the connection between the compressibility of random sequence and the algorithmic entropy, we compare the levels of entropies, obtained by means of a combination of analytical and numerical methods, with the algorithmic entropy calculated by the standard files archivers. We show that our method gives a much lower level of entropy as compared to the best archivers. Numerical simulations show also that an account of third-order correlation functions significantly decreases the entropy calculated in the framework of the additive Markov chain approach.

DOI: 10.1103/PhysRevE.98.042144

I. INTRODUCTION

Random sequences with a *finite number of state space* 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 [1,2]. The items of sequence can be phonemes, syllables, words, or DNA's base pairs according to the application.

Such random sequences are the subject of study of the algorithmic (Kolmogorov-Solomonoff-Chaitin) complexity, information theory, computability, statistical inference problem and have many application aspects. Among them are data compression [3] and the natural language processing [4], which is the important branch of the artificial intelligence.

One of the ways to get a correct insight into the nature of correlations in sequences with nontrivial information content consists in an ability to construct a correlated sequence of symbols having the same statistical characteristics as the initial system under study. There exist many algorithms for generating long-range correlated sequences—the high-order Markov chains are ones among the most important. Such random chains, the method of their generation and all their statistical properties are completely determined by the *conditional probability distribution function* (CPDF).

The main purpose of our work is to elaborate a reliable tool for reconstructing the CPDF of random *finite*-length sequences considering them as the high-order Markov chains with *finite alphabet*. The quality of different methods of CPDF's reconstruction is verified by studying the entropy and compressibility of random chains, numerically constructed with these CPDFs. The principle of such verification consists in the close connection between the entropy and compressibility explained below by Eq. (13)—the better CPDF is reconstructed, the lower entropy of the random sequence can be achieved, the higher level of its compression may be reached. The typical examples of sequences with correlated disorder—DNAs and literature texts—can be found in Refs. [5,6].

The classical method of the CPDF construction is based on calculation of the frequencies of word occurrence. The Nth-order CPDF, i.e., the probability of definite symbol $a_{N+1} = \alpha$ occurring after known N subsequence $a_1, \dots, a_N = a_1^N$ can be found by using the well-known standard likelihood method,

$$P\left(a_{N+1} = \alpha | a_1^N\right) = \frac{P\left(a_1^N, \alpha\right)}{P\left(a_1^N\right)}, \tag{1}$$

where $P(a_1^N)$ and $P(a_1^N, \alpha)$ are the *joint distribution functions* of the N subsequence a_1^N and $(N+1)$ subsequence (a_1^N, α) occurring, respectively; hereafter we use the concise notation a_{i-N}^{i-1} for N word $a_{i-N}, \dots, a_{i-1}$. In our context, the N *word* means a subsequence of N symbols.

For the long-range correlated systems another method is more appropriate. In this framework we study complexes of two and three symbols, transforming the frequencies of their occurrence into correlation functions. After that the entropy of sequence is presented analytically by means of correlation functions of the second and third orders. An applicability of each method for the sequence of finite length M is defined by the possible length of words or by the distances between letters in the complexes, which allows one the calculation of the CPDF with a sufficient precision.

Obviously, the average number of some word a_1^L occurring in a whole sequence of given length M and fixed dimension m of the alphabet exponentially decreases with word length L. Suppose that the sequence is weakly correlated (probabilities of all words of fixed length are almost equal). Then, the length $N_{\max}$ of word, that occurs on average one time, is determined from the evident equality $M \simeq m^L$, which gives $N_{\max} \simeq \ln M / \ln m$.

The correlations decrease the number of *typical* words that can be found in the sequence, and this phenomenon increases the length $N_{\max}$. In the presence of correlations, the frequencies of words occurrence differ from one another and, more specifically, the most common typical words contribute to entropy, and "discriminated" words drop out of the count. This conclusion follows from the famous result of the information theory, known as the Shennon-McMillan-Breiman theorem [7]. Above, in our qualitative reasoning, we have neglected this phenomenon of correlations, although in the below following consideration this effect is decisive and will be taken into account.

The words of length $L \ll N_{\max} \simeq \ln M/\ln m$ are well represented in the sequence of length M, so that one can use the statistical approach to these words and calculate the probabilities of their occurrence in the chain. By contrast, the statistics of longer words, $L \gtrsim N_{\max}$, is not well defined and the finite length sequences cannot be considered as a good probabilistic object for such words.

So, the statistical properties of the studied sequence can only be reconstructed up to the length of order $L \ll N_{\max}$. For the correlation length $R_c < N_{\max}$ the sequence should be considered as quasiergodic because the words of length R_c provide the statistically meaningful information for reconstructing the conditional probability distribution function of the sequence.

In the case $N_{\max} < R_c$ the statistically important information in the interval $N_{\max} < L < R_c$ is inaccessible in the framework of likelihood estimation method, Eq. (1), and can only be constructed with the proposed method of the bilinear Markov chains. Below in Figs. 2 and 3 we will see the realization of these two regimes.

A method that allows us to use the information on the symbols spaced by a distance $N_{\max} \lesssim r \ll \min(R_c, r_{\max})$, not only in the narrower region with $r \ll N_{\max}$, is connected with the high-order additive and bilinear Markov chains—constructions proposed in Ref. [8]. We highly recommend this paper to the reader before reading this article.

II. ANALYTICAL APPROACH

A. Symbolic Markov chains

Consider a semi-infinite random stationary ergodic sequence $\mathbb{S}$ of symbols (letters, characters) a_i, $\mathbb{S} = a_0, a_1, a_2, \dots,$ taken from the finite alphabet $\mathcal{A} = \{\alpha^1, \alpha^2, \dots, \alpha^m\}$, $a_i \in \mathcal{A}$, $i \in \mathbb{N}_+ = \{0, 1, 2 \dots\}$. We use the notation a_i to indicate a position i of the symbol a in the chain and the unified notation α^k to stress the value of the symbol $a \in \mathcal{A}$.

We suppose that the symbolic sequence $\mathbb{S}$ is the *high-order Markov chain* [6,9–11]. The Markov chain is of the N order if it possesses the following property: the probability of symbol a_i to have a certain value $\alpha^k \in \mathcal{A}$ under the condition that *all* previous symbols are fixed depends only on N previous symbols,

$$P\big(a_i = \alpha \big| a_0^{i-1}\big) = P\big(a_i = \alpha \big| a_{i-N}^{i-1}\big), \qquad i \geqslant N. \tag{2}$$

Below we represent the conditional probability function as the sum of multilinear monomials a_i of different orders (from zero up to the second order). The usefulness of such representation can be expressed in the two following sections.

B. Additive CPDF

Taking into account all difficulties to use the CPDF of general form Eq. (2), we should search some simplified models to reproduce properties of the real CPDF. In particular, at the first stage, we can suppose that the symbolic Markov chain under consideration is *additive*, i.e., its conditional probability is an additive function of random variables a_k, $k = i - N, \dots, i - 1$,

$$P^{(1)}\big(a_i = \alpha \big| a_{i-N}^{i-1}\big) = p_\alpha + \sum_{r=1}^{N} \sum_{\beta \in \mathcal{A}} F_{\alpha\beta}(r)[\delta(a_{i-r}, \beta) - p_\beta], \tag{3}$$

where $\delta(\cdots)$ is Kronecker delta, $p_\alpha = \overline{\delta(a_i, \alpha)}$ is the relative number of symbols α in the chain, or their probabilities of occurring, and $F_{\alpha\beta}(r)$ is proportionality factors, referred to below as *memory functions*.

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. (3) is responsible for correct reproduction of statistical properties of uncorrelated sequences (the so-called discrete white noise); the second one takes into account and reproduces binary correlations among symbols of the random sequence under generation. For any values of $\alpha, \beta \in \mathcal{A}$ and $r \geqslant 1$ the relationship between the correlation and memory functions is of the form (see Ref. [12])

$$C_{\alpha\beta}(r) = \sum_{r'=1}^{N} \sum_{\gamma \in \mathcal{A}} C_{\alpha\gamma}(r - r') F_{\beta\gamma}(r'), \tag{4}$$

where the *symbolic* two-point correlation function is defined as

$$C_{\alpha\beta}(r) = \overline{[\delta(a_i, \alpha) - p_\alpha][\delta(a_{i+r}, \beta) - p_\beta]}, \qquad \alpha, \beta \in \mathcal{A}. \tag{5}$$

An approximate solution of Eq. (4) for the memory function in the limit of weak correlations is

$$F_{\alpha\beta}(r) = \frac{C_{\beta\alpha}(r)}{p_\beta}. \tag{6}$$

C. Bilinear CPDF

The right-hand side of Eq. (3) contains two first terms of asymptotic expansion of the exact form, Eq. (2) (see details in Refs. [8,13]). The next term $Q^{(2)}$ is

$$\begin{aligned} Q^{(2)}\big(a_i = \alpha \big| a_{i-N}^{i-1}\big) = \sum_{\beta\gamma \in \mathcal{A}} \sum_{1 \leqslant r_1 < r_2 \leqslant N} & F_{\alpha;\beta\gamma}(r_1, r_2) \\ & \times \{[\delta(a_{i-r_1}, \beta) - p_\beta][\delta(a_{i-r_2}, \gamma) - p_\gamma] \\ & - C_{\gamma\beta}(r_2 - r_1)\}. \end{aligned} \tag{7}$$

The conditional probability function that contains linear term $P_{\text{add}} = P^{(1)}(\cdot|\cdot)$ and bilinear function $Q^{(2)}$ defines the *bilinear Markov chain* [8].

D. Entropy of the Markov chains

To estimate the conditional entropy of stationary sequence $\mathbb{S}$ of symbols a_i one could use the Shannon definition for *the* *block entropies* H_L of block length L,

$$H_L = -\sum_{a_1\cdots a_L\in\mathcal{A}} P\left(a_1^L\right)\log_2 P\left(a_1^L\right). \tag{8}$$

The *conditional* entropy, or the entropy per symbol, is given by

$$h_L = H_{L+1} - H_L = \overline{h\left(a_{L+1}\middle|a_1^L\right)}. \tag{9}$$

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 the $(L+1)$th symbol with preceding L symbols are taken into account. The conditional entropy h_L can be presented in terms of the conditional probability distribution function

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

E. Multilinear entropy presentation

The conditional entropy of a stationary ergodic weakly correlated random sequence can be approximately expressed [12] in terms of symbolic two-point correlation functions. The result of its analytical evaluation in the additive approximation is

$$h_L = h_0 - \frac{1}{2\ln 2}\sum_{r=1}^{L}\sum_{\alpha,\beta\in\mathcal{A}}\frac{C_{\alpha\beta}^2(r)}{p_\alpha p_\beta}. \tag{11}$$

The correction due to the third-order correlation obtained by the same method with using the bilinear CPDF is of the form [13]

$$\Delta h_L^{\mathrm{bil}} = -\frac{1}{2\ln 2}\sum_{r_1<r_2}^{L}\sum_{\alpha,\beta,\gamma\in\mathcal{A}}\frac{C_{\beta\gamma\alpha}^2(r_2,r_1)}{p_\alpha p_\beta p_\gamma}. \tag{12}$$

The correlation functions and other statistical characteristics of the Markov chain are deterministic quantities only in the limit of infinite lengths of the random sequence. For this reason, under calculations we eliminate fluctuations of random quantities that can contribute to the entropy of finite random chains, as was done in [12].

To obtain Eqs. (11) and (12), we have supposed that correlations in the random chain are weak. This is not a very severe restriction. Many examples of such systems, described in the framework of the weak correlation approximation, are given in Ref. [14]. The randomly chosen example of DNA sequences supports this conclusion. The strongly correlated systems, which are 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 article.

F. Prediction and arithmetic coding

According to the information theory, there is a straight connection between the information entropy, the ability of symbols prediction, and the compressibility of data arrays [15,16].

Having a message composed of symbols over some finite alphabet and knowing the probability of appearance of each symbol we seek to represent the message using the smallest possible number of bits. For this purpose, the most convenient seems the entropy encoders, namely, the *adaptive arithmetic coding* [17].

Arithmetic coding is a data compression technique that encodes the source binary string by creating another *binary code string*, which is a number between 0 and 1. The method starts with the interval [0, 1), and narrows it with each read symbol of sequence, taking into account its probability. The algorithm is designed such that a high-probability symbol narrows the interval less than a low-probability symbol (with the result that higher-probability symbols require less information to describe).

The length of the resulting binary chain is determined by the size of the final segment. The degree of compression is maximal in the case when the probabilities of the symbols occurring are predicted most accurately. The efficiency of the coding method can be expressed by means of its compression coefficient k defined in information theory [18],

$$k = \frac{\text{no. of bits after compression}}{\text{no. of bits before compression}} = \frac{h}{\log_2 m}. \tag{13}$$

This simple formula reflects a concept that a maximally compressed sequence is totally random, does not contain any correlations and its length, *no. of bits after compression*, is the algorithmic Kolmogorov complexity of the initial sequence.

It should be stressed that the model described in our article, as well as standard archivers, leads to the lossless compression: The calculation of each next character means a narrowing of the working segment, by choosing the left or right part of it. Therefore, as a result of processing, we obtain the final segment, whose coordinates are uniquely determined by the entire sequence. By setting the position of a segment with sufficient accuracy, we obtain the lossless encoding. For effectiveness of the arithmetic compression, it is necessary to know statistics of the words (or the two- and three-letter complexes in our case) of the sequence. An inaccurate evaluation of the statistics of the compressed sequence leads to the worst compression, which remains nevertheless a lossless one.

Let us illustrate the applicability of the developed theory to some random weakly correlated sequences naturally arising in biology and chaotic dynamics.

III. NUMERICAL SIMULATION

A. Entropy of a DNA sequence in the additive CPDF approach

In Fig. 1 the conditional entropies per symbol versus the length L for the DNA sequence of *Drosophila melanogaster*, NC 004354.1 taken from the NCBI base [19], are demonstrated. The sequence of nucleotides is transformed into a binary file, a sequence of bits, by coding each nucleotide with two bits: $A \Rightarrow 00$, $C \Rightarrow 01$, $G \Rightarrow 10$, $T \Rightarrow 11$ and then converted every eight bits into one byte.

The solid line corresponds to the entropy estimation based on the approximate analytical formula, Eq. (11), with numerically estimated symbolic correlation functions, Eq. (5). The compressibility level of the correlated sequence indicated in Fig. 1 by the dashed straight horizontal line is calculated by applying to the sequence the standard archiver 7Zip. The maximum compression is provided by the PPMd algorithm with the dictionary size equal to 1024 MB and the word size

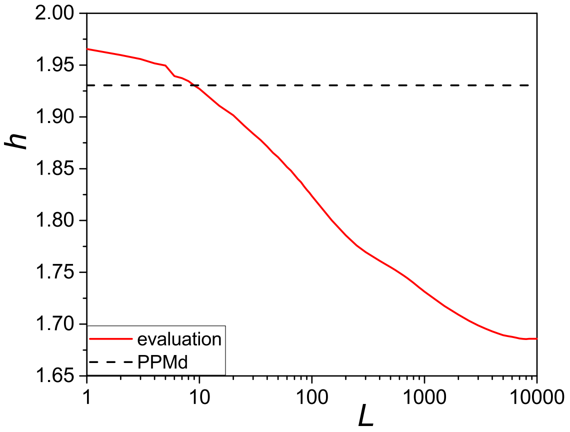

FIG. 1. The conditional entropies h per symbol vs the length L for R3 chromosome DNA from the *Drosophila melanogaster* of length $M \approx 2.7 \times 10^7$. The solid line is the entropy calculated in the weak pair correlations approximation, Eq. (11). The dashed horizontal line is the average compressibility of sequence by the standard file archiver, expressed in the entropy unities, Eq. (13).

equal to 32. We see a much lower level of entropy (the solid line in Fig. 1) with respect to algorithmic entropy obtained by the standard archivers.

As mentioned above, Eqs. (11) and (12) are valid only in the case of weak correlations. This condition can be repressed in terms of correlation functions, $|C_{\alpha\beta}(r)| \ll C_{\alpha\beta}(0)$, or in terms of CPDF, $|P(a_i = \alpha|a_{i-N}^{i-1}) - p_\alpha| \ll 1$. In practice, the verification of these conditions for each pair (α, β) of random variables can be quite cumbersome and complicated. Therefore, one can roughly estimate their fulfillment by comparing the contribution to the entropy caused by correlations, $h_0 - h_L$, with the uncorrelated entropy h_0. For the example of the DNA shown in Fig. 1, this condition is approximately $0.3 \ll 2$, which is good enough for the criterion of weak correlations.

B. Bilinear approximation

Now, let us study the effect of the third-order correlations on the CPDFs. We show that the third-order correlation functions can significantly lower the entropy calculated in the framework of the additive Markov chain approach and improve the compressibility of random sequences. In each of the cases considered below the results are obtained by taking into account all terms in Eqs. (11) and (12) with numerically calculated correlation functions (of the second and third order) for the given sequences. The first example is the chaotic deterministic system with relatively weak and short-range correlations and the second one is the same DNA sequence as that studied above.

1. Mapping

In Fig. 2 we present the plot of the conditional entropy per symbol versus the length L for the quasirandom sequence obtained by a coarse-graining of logistic map, $x_{n+1} = \mu x_n(1 - x_n)$. Dichotomization of continuous variables by the coarse-graining consists in replacement (after generation) of the numbers x_n by 0 or 1, if $x_n \in (0, 1/2)$ or $x_n \in [1/2, 1)$, respectively. The third-order correlations are taken into account at $L < 30$ because for the larger L their contributions disappear. It seems plausible to accept that the value of the block entropy per letter $h_{\rm bl}$ at $L \gtrsim 14$ becomes incorrect. At the same time the result for the entropy $h_{\rm bil}$ (not shown in the figure) remains valid until $L \leqslant 10^3$.

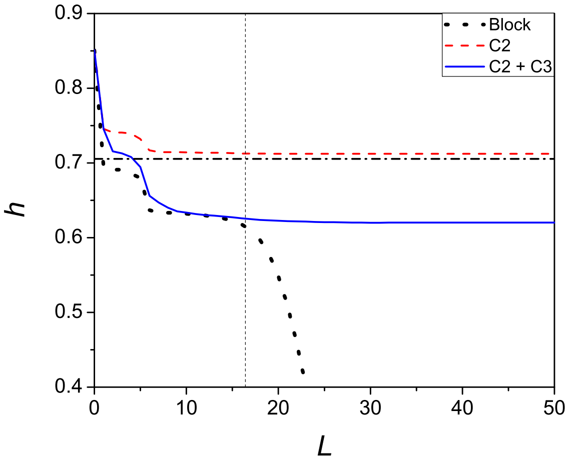

FIG. 2. The conditional entropies for the dichotomized logistic map: $x_{n+1} = \mu x_n(1 - x_n)$. The length of generated sequence is $M = 10^5$, the parameter $\mu = 3.80$, and the Lyapunov exponent $\Lambda = 0.43$. The bottom dotted curve presents the block entropy $h_{\rm bl}$, Eqs. (8)–(10), obtained by the estimation of the block probability occurring, Eq. (1); the top dashed curve is the entropy Eq. (11); the solid curve, $h_{\rm bil}$, is the result obtained by taking into account both terms Eqs. (11) and (12) with $r \leqslant 60$ for the pair correlation function and $r1 < r2 \leqslant 30$ for the third-order correlator. The vertical line at $L \approx 16$ indicates the length of words $N_{\max}$. The straight horizontal dash-dotted line indicates the average compression of the sequence by the standard 7Zip file archiver.

The compressibility level of the correlated sequence indicated in Fig. 2 by the straight horizontal dash-dotted line is calculated in the same way as for DNA, by applying the standard archiver 7Zip to the sequence. In this case, the optimal compression is achieved with the algorithm BZip2 (the dictionary size is 900 KB and the block size is 64 KB).

The proximity of the upper dashed line, where the pair correlations are only taken into account, and the straight horizontal dash-dotted line indicates that the standard archivers take into account the pair correlations only. A satisfactory coincidence of the bottom dotted curve and the solid line in the interval $9 < L < 15$ says that the proposed bilinear method correctly takes into account short-range statistics, unlike the standard compressor.

We can conclude that the entropy level for the chaotic data is higher than in the previous example of the DNA sequence. This is due to the higher level of correlations and larger value of the correlation length R_c in the DNA sequence as compared with the logistic map sequence (for the chosen value of the parameter μ).

2. Bilinear CPDF for DNA sequences

The importance of taking into account the third-order correlation functions is shown in Fig. 3, where we present the plot of the conditional entropy per symbol versus the length L

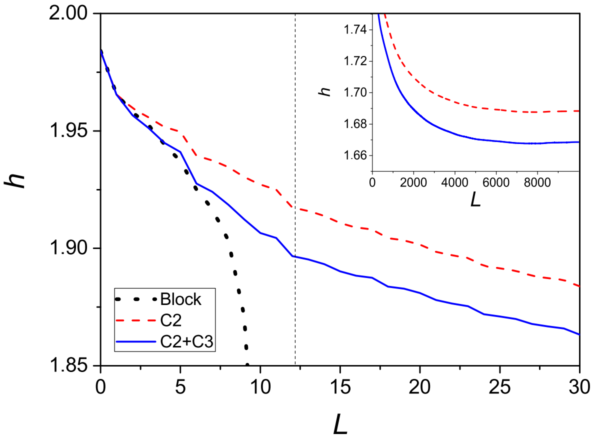

FIG. 3. The conditional entropies h per symbol vs the length L for R3 chromosome DNA from the *Drosophila melanogaster* nucleotide sequence translated into a binary file. Description of all curves coincides with that of Fig. 2. The pair and third-order correlation functions are taken into account in regions $r \leqslant 10^4$ and $r_1 < r_2 \leqslant 10$, respectively. The vertical line at $L \approx 12$ indicates the length of words $N_{\max}$. In the inset, the same curves are shown in the large scale.

for the DNA sequence of the R3 chromosome of *Drosophila melanogaster*. Under calculations we took into account the terms $C_{\beta\gamma\alpha}(r_2, r_1)$ at $r_1 < r_2 \leqslant 10$. For $r > 10$ the third-order correlations give a negligibly small contribution.

In distinction from Fig. 2 there is a common region of the entropy coincidence at $L < 6$ only. We see that the applicability of the likelihood method is lost long before (at $L \sim 10$, the lower curve) all pair and third-order correlations are taken into account (at $L \sim 7 \times 10^3$, in the inset).

Figure 3 shows that the gain in the level of compressibility, which we obtain by the method considered here (the combination of additive and bilinear parts of CPDF) with respect to the standard file archivers is

$$\frac{\Delta h_{\rm bil}}{\Delta h_{\rm stand}} = \frac{2 - 1.67}{2 - 1.93} \approx 4.7. \qquad (14)$$

IV. CONCLUSION

We have elaborated the accurate estimation for the CPDF of random symbolic sequences with complex correlation properties. The failure of the standard method, Eq. (1), even for moderate distances L is demonstrated. We have shown that for DNA nucleotide sequences a much lower level of entropy (and, probably, a higher level of compression) can be obtained by the method presented here than by the standard file archivers, even in the case of using the pair correlations only. The entropy of sequence was expressed by means of the correlation functions of the second and third orders, in the limit of weak correlations (11) and (12). Using this connection the entropy of the DNA nucleotide sequences and sequences obtained by dichotomization of the logistic map had been evaluated.

Using the connection between the compressibility of random sequence and the algorithmic entropy, the evaluated levels of entropies had been compared with the algorithmic entropy calculated by the standard files archivers. Numerical simulations show that the third-order correlations can significantly lower the entropy calculated in the framework of the additive Markov chain approach.

The result Eq. (14) allows us to hope that the method proposed in this work can be used for the creation of data compressors with the properties superior to the currently known archivers.

Our study was limited by the additive and bilinear Markov chain models. Nevertheless, there are no principal restrictions to use the higher-order memory and correlation functions.

---

[1] P. Ehrenfest and T. Ehrenfest, *Encyklopädie der Mathematischen Wissenschaften* (Springer, Berlin, 1911), p. 742, Bd. II.
[2] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding* (Cambridge University Press, Cambridge, 1995).
[3] D. Salomon, *A Concise Introduction to Data Compression*, 4th ed. (Springer, Berlin, 2008).
[4] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval* (Cambridge University Press, Cambridge, 2008).
[5] W. Ebeling and G. Nicolis, Europhys. Lett. **14**, 191 (1991).
[6] O. V. Usatenko and V. A. Yampol'skii, Phys. Rev. Lett. **90**, 110601 (2003).
[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (New York, Wiley, 2006).
[8] S. S. Melnik and O. V. Usatenko, Phys. Rev. E **96**, 012158 (2017).
[9] P. C. Shields, *The Ergodic Theory of Discrete Sample Paths*, Graduate Studies in Mathematics Vol. 13 (American Mathematical Society, Providence, 1996).
[10] O. V. Usatenko, V. A. Yampol'skii, K. E. Kechedzhy, and S. S. Mel'nyk, Phys. Rev. E **68**, 061107 (2003).
[11] 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).
[12] S. S. Melnik and O. V. Usatenko, Phys. Rev. E **93**, 062144 (2016).
[13] S. S. Melnik and O. V. Usatenko, Physica A **506**, 208 (2018).
[14] F. M. Izrailev, A. A. Krokhin, and N. M. Makarov, Phys. Rep. **512**, 125 (2012).
[15] A. N. Kolmogorov, Int. J. Comput. Math. **2**, 157 (1968).
[16] C. E. Shannon and W. Weaver, Bell Syst. Tech. J. **27**, 379 (1948).
[17] I. H. Witten, R. M. Neal, and J. G. Cleary, Commun. ACM **30**, 520 (1987).
[18] A. I. Khinchin, *Mathematical Foundations of Information Theory* (Springer, Berlin, 1957).
[19] See https://www.ncbi.nlm.nih.gov/nuccore/NC004354.1/.