OPEN ACCESS

GAVIN PUBLISHERS

Journal of Biostatistics & Biometrics

Short Communication

Melnik SS and Usatenko OV. J Biostat Biom: JBSB 104.
DOI: 10.29011/JBSB-104. 000004

DNA Sequences Entropy and Long-Rang Correlated Disorder Beyond the Pair Correlations

Sergey Melnik, Oleg Usatenko*

Department of Theoretical Physics, A. Ya. Usikov Institute for Radiophysics and Electronics Ukrainian National Academy of Science, Ukraine

***Corresponding author:** Oleg Usatenko, Department of Theoretical Physics, A. Ya. Usikov Institute for Radiophysics and Electronics Ukrainian National Academy of Science, Ukraine. Tel: +7988532380; Email: olegusatenko@mail.ru

Citation: Melnik SS, Usatenko OV (2018) DNA Sequences Entropy and Long-Rang Correlated Disorder Beyond the Pair Correlations. J Biostat Biom: JBSB 104. DOI: 10.29011/JBSB-104. 000004

Received Date: 02 May, 2018; **Accepted Date:** 28 May, 2018; **Published Date:** 01 June, 2018

Communication

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 the third orders. We evaluate numerically the entropy of some DNA nucleotide sequences. Numerical simulations show that the third-order correlations can significantly lower the entropy calculated in the framework of the additive Markov chain approach.

PACS numbers: 05.40.-a, 87.10+e, 07.05. Mh

One of the ways to get a correct insight into the nature of correlations in random sequences with nontrivialinformation content consists in an ability to construct a correlated sequence of symbols possessing 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) (known also as the transition probability function).

The main purpose of our work is to elaborate a reliable tool for constructing the CPDF for random sequences considering them as the high-order Markov chains with finite alphabet, supposing that the reference quasi-random chain of finite length is given. The quality of different methods of CPDF's reconstruction is verified by studying the levels of entropy of numerically constructed random chains with different CPDFs. The idea of such verification consists in the close connection between the entropy and the completeness of the statistical information the better CPDF is reconstructed, the lower entropy of the random sequence will be reached.

There are two methods of the CPDF construction. The first one is based on calculation the frequencies of word occurring [1,2], the second one takes into account the frequencies of the letters complexes appearance [3,4]. In the framework of the second method, we study complexes of two and three symbols, transforming the frequencies of their occurrence into correlation functions. After that we present analytically the entropy of sequence by means of correlation functions of the second and the third orders.

The N^{th} order CPDF can be found by using the well-known standard likelihood method via the N-symbols joint distribution functions. The average number of some word $a_1^N \equiv a_1,\ldots, a_N$ occurring in whole sequence expo-nentially decreases with word length N. For given length M of weakly correlated sequence with fixed dimension m of the alphabet, the length N_{max} of word, statistics for which can be calculated with suffcient accuracy can be evaluated as $N_{max} \approx \ln M/\ln m$. A method that allows us to use the information on the symbols spaced by a distance r > Nmax, not only in narrower region with r $\ll$ Nmax, is connected with the high-order additive and bilinear Markov chains constructions proposed in Ref [5].

We consider a semi-in finite random stationary ergodic sequence $\mathbb{S}$ of symbols a_i, $\mathbb{S}= a_0; a_1; a_{2,\ldots,}$ taken from the finite alphabet $\mathcal{A}= \{ a^1,a^2,\ldots,a^m\}$, $a_i \in \mathcal{A}$ $i \in \mathbb{N}_+ = \{0,1,2..\}$,

We use the notation a_i to indicate a position i of the symbol i 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 [6-8]. The conditional entropy, or the entropy per symbol, is given by

$$h_L = H_{L+1} - H_L = \overline{h(a_{L+1}|a_1^L)}, \qquad (1)$$

where HL is the Shannon block entropies of block length L. This quantity specifies the degree of uncertainty of $(L + 1)^{th}$ 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.

$$h(a_{L+1}|a_1^L) = -\sum_{a_{L+1}\in\mathcal{A}} P(a_{L+1}|a_1^L)\log_2 P(a_{L+1}|a_1^L). \quad (2)$$

The conditional entropy h_L can be presented in terms of the conditional probability function The conditional entropy of a stationary ergodic weakly correlated random sequence can be approximately ex-pressed [3] in terms of symbolic two-point correlation functions. The result of its analytical evaluation in the additive approximation is

$$h_L = h_0 - \frac{1}{2\ln 2}\sum_{r=1}^{L}\sum_{\alpha,\beta\in\mathcal{A}}\frac{C^2_{\alpha\beta}(r)}{p_\alpha p_\beta}. \qquad (3)$$

The correction due to the third-order correlation ob-tained by the same method with using the bilinear CPDF is of the form [4],

$$\Delta h_L^{bilin} = -\frac{1}{2\ln 2}\sum_{r_1<r_2}^{L}\sum_{\alpha,\beta,\gamma\in\mathcal{A}}\frac{C^2_{\beta\gamma\alpha}(r_2,r_1)}{p_\alpha p_\beta p_\gamma}. \qquad (4)$$

Let us illustrate applicability of the developed theory. In Figure 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 [9] are demonstrated. The sequence of nucleotides are trans-formed into a binary file, a sequence of bits, by coding each nucleotide with two bits: A ⇒00; C ⇒01; G⇒

10; T⇒11 and then converted every eight bits into one byte. The upper dashed line corresponds to the entropy estimation based on the approximate analytical formula, Equation (3), with numerically estimated symbolic correlation functions.

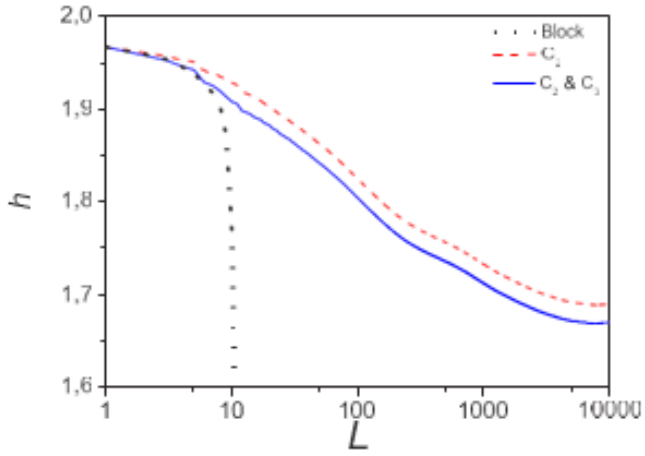

Figure 1: The conditional entropies h per symbol vs length L for R3 chromosome DNA from the *Drosophila melanogaster* nucleotide sequence translated into a binary file of length M ≈ $2{:}7\times10^7$. The low dotted line presents the entropy calculated in the framework of the method of likelihood estimation. The upper dashed line is the entropy calculated in the weak pair correlations approximation, Equation (3). The pair and third order correlation functions, solid line, are taken into account in regions r ≤ 104 and r1 < r2 ≤ 10, respectively.

The next result, solid line, is obtained by taking into account all terms in Equations (3,4) with numerically calculated correlation functions of the given sequence. The low dotted line in Figure 1 presenting the entropy is calculated in the framework of the method of likelihood estimation. Here, the conditional probability distribution function of N^{th} order is

$$P(a_{N+1} = \alpha^k|a_1^N) = \frac{P(a_1^N,\alpha^k)}{P(a_1^N)}, \qquad (5)$$

where p (a_1^L; a^k) and p (a_1^L) are the probabilities of the (N +1) -subsequence (a_1^L; a^k) and N-subsequence a_1^N occurring, respectively.

The distinction of the upper dashed line, where the pair correlations are only taken into account, and the solid line indicates that the bilinear component improves suffciently the accounting of the statistics. A satisfactory coincidence of the bottom dotted curve (entropy estimation via standard likelihood method) and the solid line in the interval $9 < L < 10$ says that the proposed bilinear model correctly takes into account short-range statistics, unlike the additive component itself. Under calculations we took into account the terms $C_{\alpha\beta\gamma}$ (r_2, r_1) at $r_1 < r_2 \leq 10$. For $r > 10$ the third-order correlations give a negligibly small contribution.

We see that applicability of likelihood method is lost long before (at L~10, the lower curve) than all pair and third order correlations are taken into account (at $L \sim 7 \times 10^3$). So, we have elaborated the accurate estimation for the CPDF of random symbolic sequences with complex correlation properties. The failure of the standard method, even for moderate distances L is demonstrated. We have shown that for DNA nucleotide sequences a much lower level of entropy can be obtained by the presented here method even in the case of using the pair correlations only. 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 allows us to hope that the method proposed in this work can be used for creation of data compressors with properties superior to the currently known archivers.

References

1. Ebeling W, Nicolis G (1991) Entropy of Symbolic Sequences: The Role of Correlations. Europhys Lett 14: 191.
2. Schurmann T, Grassberger (1996) Entropy estimation of symbol sequences. P Chaos 6: 414-427.
3. Melnik SS, Usatenko OV (2016) Entropy and long-range memory in random symbolic additive Markov chains. Phys Rev E 93: 062144.
4. Melnik SS, Usatenko OV (2017) Entropy of random symbolic high-order bilinear Markov chains. arXiv.org > cond-mat.stat-mech > 1709.06642; to be published in Physica A.
5. Melnik SS, Usatenko OV (2017) Decomposition of conditional probability for high-order symbolic Markov chains. Phys Rev E 96: 012158.
6. Usatenko OV, Yampol'skii VA (2003) Binary N-Step Markov Chains and Long-Range Correlated Systems. Phys Rev Lett 90: 110601.
7. P. C. Shields (1996) The ergodic theory of discrete sample paths (Graduate studies in mathematics, 13).
8. Usatenko OV, Apostolov SS, Mayzelis ZA, Melnik SS (2010) Random Finite-Valued Dynamical Systems: Additive Markov Chain Approach. Cambridge Scientific Publisher 171.
9. https://www.ncbi.nlm.nih.gov/nuccore/NC 004354.1/.