# New Structure in Genomes Manifests in Triplet Distribution Alongside a Sequence

[†]Michael Sadovsky [⋆] and [‡]Eugene Mirkes

[†]Institute of computational modelling of SB RAS and
[‡]Krasnoyarsk Institute of Railway Engineers
http://icm.krasn.ru

**Abstract.** A discovery of various structures in nucleotide sequences still challenges researchers. Here we present the structure determined by the distribution of triplets in chromosomes, where the distribution is defined to the nearest neighbour. We found a wonderful periodicity in triplet distribution to be observed alongside a nucleotide sequences of genomes of some higher mammalians (donkey, rat, chimpanzee, man). Besides, a variety of other structure patterns differing both from a periodical, and random ones have been detected. These patterns may not be explained within a framework of HMM of any relevant order (from 2 to 6).

**Keywords:** Order; Periodicity; Correlation; Decay; Long-Range Correlation

## 1 Introduction

A retrieval of various patterns and, in general, an order search in completely sequenced genomes is a great deal of up-to-date biophysics and bioinformatics. The correlations observed within these latter reflect some biological features of primary structures [1–4]. In particular, the sequence inhomogeneity manifesting in the difference of frequency dictionaries of non-overlapped coherent triplets counted for three different starting positions indicates the presence of protein coding regions in a genome; more exactly, non-coding regions are invariant against the frame shift of the triplet pattern, while the coding ones lack these invariance [6–8].

A complexity of patterns observed in a genetic sequence may vary significantly. The complexity itself is a matter of interests of mathematicians, biologists and biophysicists [9–16]. Screening a genome with respect to a complexity of different fragments of that latter, one my find various biologically important peculiarities in a nucleotide sequence. Here we present a new approach to figure out some patterns in the mutual distribution of triplets observed alongside a sequence. Probably, they are the shortest fragments in DNA sequences to be taken into consideration.

[⋆] This work was in part supported by Integration grant No. 21 from SB RAS.

This paper presents some preliminary results towards the observations on the distribution of triplets alongside a sequence, where that former has been found with respect to the nearest neighbour. To figure out the origin and role of the patterns, we compared the patterns observed both for real sequence, and for three surrogate sequences (of the same length) generated with Markov process; the processes of the order 2, 3 and 6 have been used to generate a surrogate sequence. All Markov processes have been generated on the basis of the relevant frequency dictionaries (bearing the strings of the length $q = 3$, 4 and 7, respectively) developed over the original nucleotide sequence under consideration.

The comparison shows significant difference between the patterns observed over a real sequence, and surrogate ones. Surprisingly, rather simple and apparent idea yet was not explored, and a number of various and intriguing structures have been found.

## 2 Materials and Methods

Consider a coherent symbol sequence $\mathfrak{T}$ of the length $N$ from four-letter alphabet $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$; the length is the number of symbols. Any other symbols (if any) occurred in a real genetic entity were omitted, and the sequence has been concatenated.

Triplet is a continuous string of three symbols: $\omega = \nu_1\nu_2\nu_3$. To observe the distribution pattern of triplets $\omega_1$ and $\omega_2$, we have counted the following function $f_{\langle\omega_1,\omega_2\rangle}(r)$:

$$F_{\langle\omega_1,\omega_2\rangle}(r) = \text{number } n_{\langle\omega_1,\omega_2\rangle}, \quad \rho\,(\omega_1,\omega_2) = r \tag{1}$$

that is the number of copies of the couple of triplets $\langle\omega_1,\omega_2\rangle$ located at the closest distance $r$ from each other. It other words, for each embedment of the triplet $\omega_1$, the nearest embedment of the triplet $\omega_2$ has been found, and the distance between them was determined. The *nearest neighbourhood* means that there is no triplet $\omega_2$ somewhere in between, for a couple under consideration.

Function (1) is not the density distribution of the couples of given triplets over the distance $r$. To do that, one must normalize it to develop a distribution function:

$$f_{\langle\omega_1,\omega_2\rangle}(r)\,, \quad \text{so that} \quad \sum_r f_{\langle\omega_1,\omega_2\rangle}(r) = 1\,. \tag{2}$$

If two triplets are located next each other with no gap between them, then $r = 0$; if a nucleotide $\nu$ is between the triplets, then $r = 1$, etc. If two triplets $\omega_1 = \nu_1\nu_2\nu^*$ and $\omega_2 = \nu^*\nu_2'\nu_3'$ are overlapping over a nucleotide $\nu^*$, then $r = -1$, etc.

An observed function (2) should be compared to a theoretical one. Yet, there are no special issues to figure out the theoretical distribution function, but to suppose that is must decrease. The decrease of the function is a trivial follow-up of the finiteness of an original sequence; a shape of the function at shorter distances is less obvious. Nevertheless, a decrease of the function is still the most expected pattern for that latter. Indeed, longer distances between two given

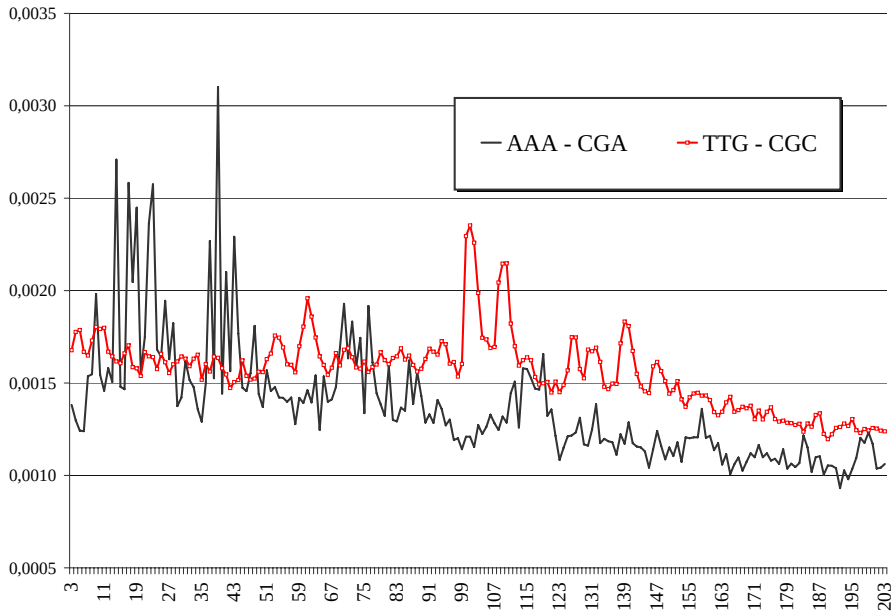


**Fig. 1.** Typical distribution function (2) pattern observed for AAA ↔ CGA (black) and TTG ↔ CGC (red) triplet couples, at the $22^{\text{nd}}$ chromosome of *Pan troglodytes* genome (accession number BA000046 in EMBL–bank).

triplets are less probable, than shorter ones; this may follow from combinatorial constraints, for example.

Counting the functions (1) and (2), one has to keep in mind the longest possible distance to be observed between two neighbouring given triplets. These functions are defined on the set of integers $\Omega = \{-2, -1, 0, 1, 2, \ldots, r_{\max}\}$, where $r_{\max}$ is the longest distance between two immediate (or closest) neighbours, for the given couple of triplets $\omega_1$ and $\omega_2$. Practically, one has to keep within the range of lengths defined arbitrary. Thus, we have chosen the value $r^* = 5000$; it means that any longer embedments of a given triplet couple were omitted. Meanwhile, it does not deteriorate the results.

### 2.1 Data source

The nucleotide sequences have been retrieved from EMBL–bank. We selected the sequences with the portion of any extra symbols (besides A, C, G, or T) less that 0.01. We have examined the sequences belonging to the genomes of organisms of various taxa ranging from bacteria to man; 461 sequences have been studied, totally.

Typical length of mammalian genome was $4 \times 10^7$ to $2 \times 10^8$ nucleotides; the shortest bacterial genome taken into analysis exceeds $10^6$ nucleotides, to avoid a finite sampling effects.

## 3 Results

We have examined the pattern of the function (2) for a number of chromosomes. To begin with, one faces the problem of a couple choice. Indeed, there exists $64 \times 64 = 4096$ triplet couples, for a genetic entity. Obviously, there is no way to show the function (2) for all the couples. Thus, one should figure out some order on the set of couples.

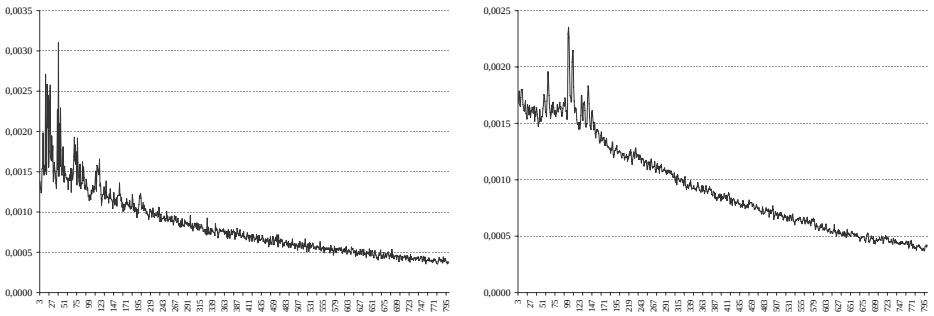


**Fig. 2.** Same functions as in Fig. 1 shown in smaller scale (up to $r = 800$; the exponential decrease of the function (2) becomes apparent). Left pattern shows the TTG ↔ CGC couple, and the right one shows AAA ↔ CGA couple.

First of all, one may concentrate on the study of the behaviour of the function (2) for some specific couples. Previously, a set of 32 couples has been identified [5]. This set consists of the couples making so called *complementary palindromes*, i. e. the couples of triplets read equally in opposite directions, with respect to the Chargaff's substitution rule. The couples CCC ↔ GGG, ATC ↔ GAT, TAA ↔ TTA, AAG ↔ CTT, ATG ↔ CAT, CAA ↔ TTG, GCA ↔ TGC, etc. are the examples of such palindromes. Such palindromic triplets are known for (rather good) proximity of the frequency of each one in a couple; in other words,

$$f_{\mathsf{TAA}} \approx f_{\mathsf{TTA}}, \quad f_{\mathsf{AAG}} \approx f_{\mathsf{CTT}}, \quad f_{\mathsf{ATG}} \approx f_{\mathsf{CAT}}, \quad f_{\mathsf{ACG}} \approx f_{\mathsf{CGT}},$$

and so on. Thus, one may expect to face the pattern of the distribution function (2) to be quite smooth and regular. Further, we shall see that some palindromes exhibit extremely complex and unusual behaviour of the function (2).

A nucleotide sequence yields 4096 couples of triplets, as the function (2); such abundance makes a problem of the analysis and visualization of the distributions.

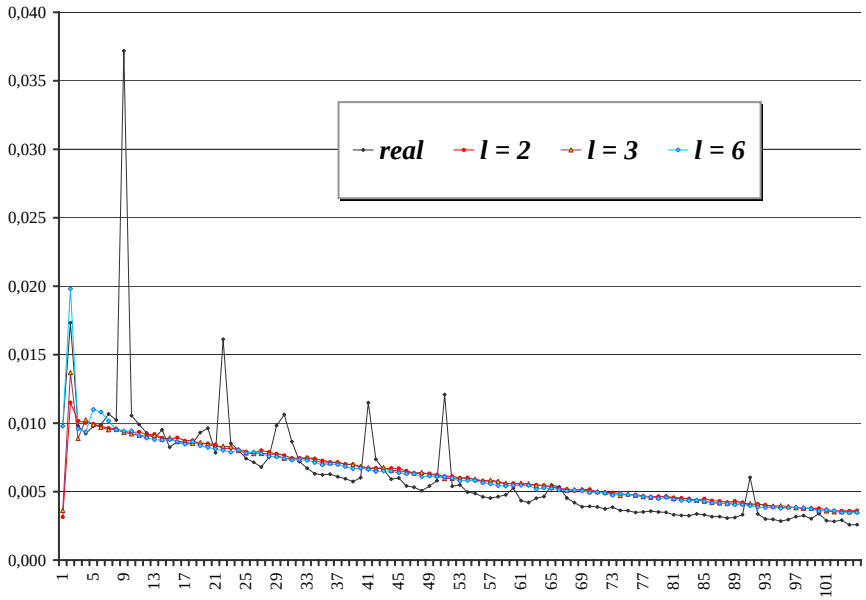


**Fig. 3.** Distribution function (2) pattern observed for CCC ↔ GGG couple of triplets. The charts are shown for the same *P.troglodytes* chromosome and three surrogate sequences generated with Markov random process of order 2, 3 and 6. The process order $l$ is shown in legend.

It is very natural that the function (2) decreases, as $r$ grows up. Thus, one may classify the couples on the basis of the decay character. Apparently, the function yields an exponential decay:

$$f_{\langle\omega_1,\omega_2\rangle}(r) \sim A \cdot \exp\{-\lambda \cdot r\}, \tag{3}$$

where the factor $\lambda$ might be specific for some couples (within the given species, of course).

Consider now Fig. 3 showing the head of the distribution functions (2) for the palindromic couple CCC ↔ GGG observed in the same genetic entity. The distribution for this couple exhibits strong and evident periodicity, with approximate period in 13 nucleotides. There are some other couples exhibiting similar behaviour of the distribution function (2). It should be said that there is no regularity in the patterns observed for the same couple of triplets, for various species, even for significantly close organisms (say, some bacterial or yeast strains).

In general, the complexity of the patter of the distribution function (2) increases with the clade level of a genome bearer. Simply speaking, bacterial genomes exhibit very smooth (quite often a single exponent trended) decrease of the distribution function (2). Quite similar picture is observed for yeast genomes, and some other fungi and protozoan genomes. Vertebrate organisms have usually more complicated pattern of the distribution function.

We have plotted the distribution function (2) determined for each couple of triplets within a given sequence with exponential decay function (3), through least square techniques. Parameters $\lambda$ and exponent factor $A$ were chosen to fit the function (2) best of all. Rather poor fitness was observed, resulted from two issued: the former is a structuredness observed for $0 \leq r \leq 100$, for many couples and sequences, and the latter is a very low (in comparison to the exponential pattern) decay of the distribution function (2) observed for long distances (for $r \leq 5000$). To kill the effects of a serious inhomogeneity observed for $0 \leq r \leq 100$, and the slow decay of the tail of the distribution function (2), three versions of the least square fitting have been implemented:

1) the distribution function (2) was used "as is";
2) a truncated distribution function (2) was used: we fitted the exponent for $100 \leq r \leq r_{\max}$, and
3) a "shifted" distribution function (2) was used: each value of the function (1) has been added with one, before the implementation of the function (2).

Thus, three couples of the parameters $A$ and $\lambda$ were derived for each couple, for a given sequence.

A comprehensive and consequent study of the distribution of the couples over the $\lambda$ value has been carried out, also. This distribution of exponent decay factors observed within the same DNA sequence was always found to be a bell-shaped; here we mean the distribution over the set of 4096 couples. Obviously, this is not the normal distribution, since it is determined for positive $r$, only. We did not studied the distribution pattern of the couples of the exponent decay factor $\lambda$ in detail, while it seems to be rather stable, in shape. The stability of the shape is accompanied with serious instability in the ordering of specific couples, in this distribution: the couples with the greatest $\lambda$ values are not numerous, for all the analyzed sequences, while they seem to be species specific. The taxonomy specificity becomes stronger, if a two-dimensional distribution is studied, in $\Re^2 = \{A\} \times \{\lambda\}$ space.

## 4 Discussion

Fig. 1 shows the distribution of the distances to the nearest neighbour observed for two triplet couples: AAA $\leftrightarrow$ CGA (black) and TTG $\leftrightarrow$ CGC (red) triplet couples, at the $22^{\text{nd}}$ chromosome of *Pan troglodytes* genome. Horizontal axis represents the distance $r$, while the vertical one represents the frequency of the nearest embedment found at this distance. Obviously, the peaks observed at $r = 101$ and $r = 109$ (see red line), and at $r = 14$ and $r = 39$ (see black line) proves the occurrence of a structure of a kind of a conspired periodicity, or a regularity of this type.

Fig. 2 shows the distribution function (2) for the same triplets, while observed for a wider range of the distances between the couples taken into consideration. Here the exponential pattern of the decrease is evident, unlike in Fig. 1. We have examined more than four hundred sequences of various taxa ranging from bacteria to higher primates. Such exponential decrease is absolutely universal,

for the studied sequences. Obviously, it results from a finiteness of a sequence, as well as from rather complex structure of these latter.

The exponential decrease of the function (2) forces a researcher to extract this trend from the distribution function, thus purifying the periodical, or quasi-periodical patterns in the genomes. Yet, this idea is quite hard for implementation. The point is that the trend actually is not exponential; at least, it is not purely exponential. First, some strings of the length $q = 4$ and $q = 5$ that could be combined from two (overlapping) triplets are extremely overrepresented, in comparison to the hexamers obtained due to concatenation of these two triplets; the difference in the frequency of the strings of the length $q = 4$ and $q = 5$ can exceed the frequency of the relevant hexamer more than $3 \times 10^3$ times.

Second, any frequency dictionary of the thickness $q$ (i. e. bearing the strings of the length $q$) unambiguously generates the relevant Markov process, of the order $q - 1$. We have generated the Markov process transition matrices, for a number of DNA sequences, and calculated the distribution function (2) theoretically. Indeed, changing the elements in the matrix that transfer into the given triplet $\omega_2$ for zero and raising such truncated matrix to the $l^{\text{th}}$ power ($l = 2, 3, 4, \ldots$), we calculated the distribution function. Surprisingly, the calculated function being plotted in the logarithmic coordinates, exhibited a broken line (of two segments) thus proving the non-uniform pattern of the decrease shape. Whether the theoretically determined distribution function (2) is a superposition of two exponents with different decay factors, or not should be studied separately. Less is evident for the real distribution functions (2) observed for real DNA sequences.

Third, the tail of the distribution function (2) observed over real DNA sequences differs drastically from the theoretical estimations. Indeed, a development of the function (2) due to the power calculation of the truncated transition matrix yields positive figures of the probability to meet a couple at the distance $l$ is positive, but $p_{\geq 100} \sim 10^{-8}$, while the real figures for the functions (2) are much greater: $f_{\geq 100} \sim A \cdot 10^{-4}$, where $1 < A < 10$. It means that real distribution functions (2) is definitely very far from any Markov approximation; here some further studies resembling the Boltzmann equation theory must be implemented.

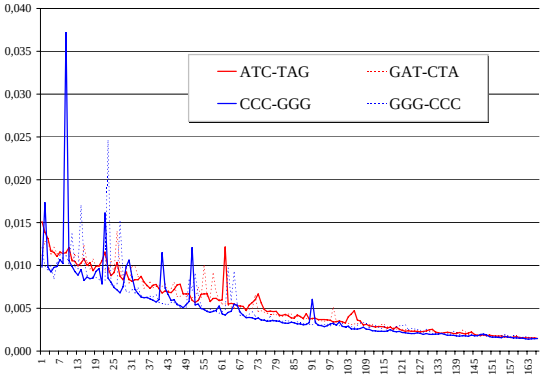


**Fig. 4:** Asymmetry in the triplet distribution (the same DNA sequence)

All these three points deteriorate the proper estimation of an exponential trend decrease, through the least square technique, as well as some other techniques of the approximation.

Let now list the basic properties of the observed patterns in the distribution of triplets alongside a nucleotide sequence:

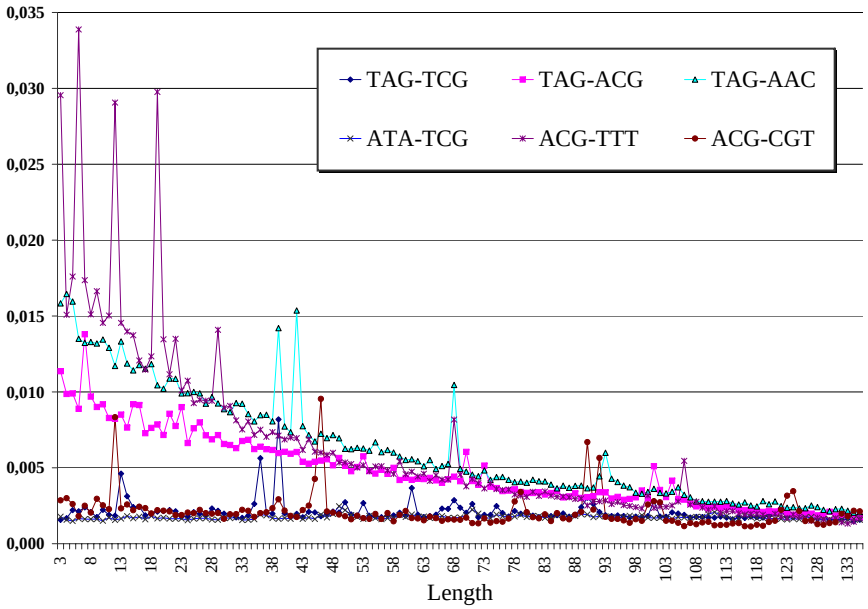


**Fig. 5.** The set of distribution functions (2) determined closest environment of the given triplet (see text for details).

1) the distribution of triplets alongside a genome reveals a new structure with yet unknown function or meaning;
2) this structure is extremely flexible: same triplets couples exhibit drastically different behaviour for different genomes;
3) the distribution exhibits a hidden periodicity;
4) the long-scale structure (tail of the distribution function) differs drastically from any Markovian approximation;
5) the distribution function is asymmetric one (see Fig. 4). This asymmetry means that distribution of a given couple, and the inverted couple differ.

Very few is known towards the origin and sense of the structure described above. The data and results shown above prove unambiguously that that is not a Markov property standing behind the patters, in real sequences. Another hypothesis is the long repeats. This hypothesis is disproved with a simple observation (see Fig. 5); we determined the charts of the distribution function (2) for six related couples: considering the couple $\mathsf{TAG} \leftrightarrow \mathsf{TCG}$ as the reference one, we traced the similar functions for the couples $\mathsf{TAG} \leftrightarrow \mathsf{ACG}$ and $\mathsf{TAG} \leftrightarrow \mathsf{AAG}$ that are the triplets overlapping over two and one nucleotide, respectively. Similarly, the distribution functions were determined for the couples $\mathsf{ATA} \leftrightarrow \mathsf{TCG}$, $\mathsf{ACG} \leftrightarrow \mathsf{TTT}$ and $\mathsf{ACG} \leftrightarrow \mathsf{CGT}$. A long-repeat theory of the origin of the structure described above must follow in the very high proximity of all these charts, while they are not. Moreover, it is clearly seen that the overall abundance of the

couples shown in Fig. 5 differs so much, that no way for a long-repeat mechanism may be found.

Let's outline in brief some further activities to be carried out to clarify the sense and function of the structure present here. First, a detailed study of the exponential trend should be carried out, whatever one understands for that former, under the constraints discussed above. A distribution of the couples over the exponent decay factor values may bring a lot for understanding the fine mechanisms of the structure.

Second, a Fourier transformation analysis should be implemented, to figure out the periodicity and other structural elements in the patterns of triplet distributions. Third, a relation between the patterns and taxonomy of the genome bearers should also be studied, in more detail.

## References


1. W.-H.Li, *Molecular Evolution* (Sineauer Associates, Sunderland, MA, 1997).
2. *Computational Models in Molecular Biology*, edited by S.L. Sazberg, D.B. Searls, and S. Kasif (Elsevier, Amsterdam, 1998).
3. J.K.Percus, *Mathematics of Genome Analysis* (Columbia University Press, Cambridge, 2002).
4. A.K.Konopka, All we need is truth. // *Comput.Biol.Chem.* (2004) **28**(1), 1–2.
5. M.G.Sadovsky, Yu.A.Putinzeva, N.A.Zajtzeva System Biology on Mitochondrion Genomes. // *Proc. of the* $3^{rd}$ *Int. Conf. on Bioinformatics, Biocomputational Systems and Biotechnologies*, 61–66.
6. A.Yu.Zinovjev, A.N.Gorban, T.G.Popova, Seven clusters in genomic triplet distribution. *In Siliko Biology* (2003) **3**, 471–482.
7. A.N.Gorban, A.Yu.Zinovjev, T.G.Popova, Self-organizing approach for automated gene identification. // *Open Systems & Information Dyn.* 2003 **10**, 321–333.
8. A.Carbone, A.Zinovyev, F.Kepes, Codon Adaptation Index as a measure of dominating codon bias. // *Bioinformatics.* (2003) **19**, 2005–2015.
9. S.Havlin, S.V.Buldyrev, A.L.Goldberger, R.N.Mantegna, C.K.Peng, M.Simons, H.E.Stanley, Statistical and linguistic features of DNA sequences. // *Fractals* (1995) **3**, 269–284.
10. L.Allison, L.Stern, T.Edgoose, T.I.Dix, Sequence complexity for biological sequence analysis. // *Comput.Chem.* (2000) **24**(1), 43–55.
11. V.N.Babenko, P.S.Kosarev, O.V.Vishnevsky, V.G.Levitsky, V.V.Basin, A.S.Frolov, Investigating extended regulatory regions of genomic DNA sequences. // *Bioinformatics* (1999) **15**, 644–653.
12. V.D.Gusev, L.A.Nemytikova, N.A.Chuzhanova, On the complexity measures of genetic sequences. // *Bioinformatics* (1999) **15**, 994–999.
13. E.Pizzi, C.Frontali, Low-Complexity Regions in Plasmodium falciparum Proteins. // *Genome Res.* (2001) **11**, 218–229.
14. H.E.Stanley, S.V.Buldyrev, A.L.Goldberger, S.Havlin, C.K.Peng, M.Simons, Scaling features of noncoding DNA. // *Physica* **A** (1999) **273**, 1–18.
15. P.Romero, Z.Obradovic, X.Li, E.C.Garner, C.J.Brown, A.K.Dunker, Sequence complexity of disordered protein. // *Proteins.* (2001) **42**(1), 38–48.
16. M.A.Jiménez-Montaño, W.Ebeling, Th.Pohl, P.E.Rapp, Entropy and complexity of finite sequences as fluctuating quantities // *BioSystems* (2002) **64**, 23–32.