A CORRELATION INVESTIGATION OF BACTERIAL DNA CODING SEQUENCES

OANA ZAINEA[*]*, V.V. MORARIU*

Department of Molecular and Biomolecular Physics, National Institute for R&D of Isotopic and Molecular Technology, 400293, Cluj-Napoca, Romania
*E-mail: oanaz@itim-cj.ro

Abstract. The coding sequences (CDS) length series in the genome of *E. coli* and *B. subtilis* were considered. The organization of the series was investigated by detrended fluctuation analysis (DFA), which gives information about the correlation characteristics. The CDS length series show a low level of correlation, which indicates close to randomness or almost lack of organization at the genome level. However, this is an apparent result. Correlation characteristics should remain constant if various segments of the series are analysed, and the correlation characteristic is uniform throughout the genome series. We have segmented the genome series into four quarters and performed DFA on each segment. The results showed a non-uniform correlation characteristic throughout the genome, ranging from high correlation or anti-correlation to almost randomness. High correlation is present in the second quarter of *B. subtilis* and in the first quarter of *E. coli* genome. This suggests that similar length genes are located preferentially in these segments.

Key words: bacterial DNA, coding sequences lengths, detrended fluctuation analysis, short-range correlation.

INTRODUCTION

Recently, it has been stated that "we know a lot about the bacterial DNA replication, recombination, repair and other aspects of cell biology, but still little about the organization of bacterial chromosome" [4, 7]. The traditional view was that bacterial chromosome represents a random, uncorrelated series of composing elements. Work published during the last few years has challenged such a view by revealing a non-uniform, "patching" structure of the chromosome [1, 2]. This involves large and small scale organizational patterns in many bacterial genomes. In fact, DNA organization at base pairs or nucleotides level has been a subject of investigation for a long time. Long-range correlation of the non-coding sequences has been reported, while no such correlation was found in the DNA coding regions [3].

Received December 2007;
in final form February 2008.

ROMANIAN J. BIOPHYS., Vol. 18, No. 1, P. 19–28, BUCHAREST, 2008

As a general rule, bacterial DNA is organized in a single circular chromosome or multiple circular chromosomes. Sometimes, it is organized in linear chromosomes and linear plasmids. An important characteristic of the bacterial DNA is its simple structure – compared to eukaryotic DNA – mainly composed of coding sequences (CDS), without introns. The non-coding sequences represent a small part of the genome. Contrary to the classical result of DNA coding regions in bacterial genomes as random sequences, it has been reported that DNA fluctuations of single bases are modulated by log-periodic variations [6].

While the above cited work was concerned with the order of bases in the DNA sequence, this work is focused on the length of coding sequences. A bacterial chromosome may contain thousands of coding sequences, which means they represent series containing that number of terms. The characteristics of such series are equally well-suitable for analysis by statistical physics methods [3]. One may ask if the length of CDS is randomly distributed along the genome or if there are any restrictions or rules which could cause some kind of genome organization at this level. This can be compared with the language analysis, in which the first stage is to look into the use of letters, than words and finally the combination of words (syntax). However, there are other reasons for such an investigation. Specialized literature indicates that gene location is not random, and genes contributing to a certain function are located close to each other [10]. Furthermore, a significant correlation between chromosome gene distribution and cell physical architecture was reported [5]. As a result it can be suspected that some kind of organization may exist in the genome at the level of CDS length.

A simple way to approach the CDS length series organization is to make a correlation investigation of the series. We chose detrended fluctuation analysis (DFA). This method was originally developed to investigate the long-range correlation in non-stationary series, as the method removes the trends in the series [9]. This is an important issue, because trends in the series of data introduce strong correlation and alter the correlation information on the fluctuation itself. In a recent work we illustrated the use of DFA to investigate any kind of correlation including short-range (or short memory) characteristics [8]. These results are based on autoregressive processes which are typical short-range correlated.

Our initial work was concerned with the correlation properties of the whole CDS length series corresponding to selected bacterial genomes. We found these results rather disappointing as the correlation exponents indicated a weak correlation of the data, even close to randomness. This was also noticed previously by other authors [11]. However, our preliminary investigation included an important result: the DFA method showed a possible short-range correlation (see the results below), which meant that the CDS length series should be regarded as series with non-uniform properties. A uniform correlation characteristic of the series would result

in the same correlation characteristic for any part of the series. A non-uniform organization of the series would give different correlation characteristics for different segments of the series. In this work we examine the (non-)uniform organization of the CDS lengths series for two bacterial genomes by segmenting the series into four quarters. Each quarter was analysed by DFA. We found that the correlation characteristics for each quarter were different, pointing to a certain genome region where the correlation was stronger than the rest. Each species had particular correlation characteristics of the CDS length series.

DATA AND METHODS

Data containing the genome details of *Escherichia coli* strain K12 MG1655 and *Bacillus subtilis* strain 168 were selected from the European Bioinformatics Institute (EBI) database (http://www.ebi.ac.uk/). A program written in Matlab version 7.0.1 was used to extract the start and end position of the coding sequences (this program is available on request from authors). The length of CDS was then calculated by the difference between these two positions. The length series were assembled by considering that the length of the first CDS was emitted at time $t = 1$, the second at time $t = 2$, etc.

The length series were subject to DFA analysis. Briefly, the DFA procedure consists of several steps. First, the initial series *u(i)* ($i = 1 \ldots N_{\max}$) are integrated:

$$y(i) = \sum_{i=1}^{j} \left[u(i) - \langle u \rangle \right] \tag{1}$$

where

$$\langle u \rangle = \frac{1}{N_{\max}} \sum_{j=1}^{N_{\max}} u(i) \tag{2}$$

The series are divided into boxes of equal size *n* and the local trend for each segment is calculated by least-squares fit of the data. If the fit of the trend is a first order polynomial, the method is known as DFA-1. The integrated time series *y(i)* is detrended by subtracting the local trend $y_{\text{fit}}(i)$ and the detrended fluctuation function is calculated as:

$$Y(i) = y(i) - y_{\text{fit}}(i) \tag{3}$$

Then, the root mean square fluctuation for a given box size *n* is calculated:

$$F(n) = \sqrt{\frac{1}{N_{\max}} \sum_{i=1}^{N_{\max}} \left[Y(i) \right]^2} \tag{4}$$

If the data are long-range power-law correlated, *F(n)* increases for large values of *n* as a power-law, $F(n) = n^{\alpha}$, where α is the correlation exponent. This implies that on a double log plot of *F(n) vs. n* there is a single straight line with a slope equal to the correlation exponent α. The exponent values are $\alpha = 0.5$ for random (uncorrelated) series, $\alpha = 1$ for $1/f$ series and $\alpha = 1.5$ for Brownian noise. Anti-correlation is associated to α values smaller than 0.5. It should be noted that the DFA plot is linear over the whole range of box sizes *n* if a long-range correlation is operative. Non-linearity of the DFA plot points to a short-range correlation and, therefore, to a non-power law description of the DFA plot [8]. The linear fit of the DFA plot over a limited range of *n* defines a short-range correlation and its slope is a local correlation exponent. This is the case for our results shown in the next section.

The segmentation procedure of the CDS length series was limited to four quarters in order to keep the segments reasonably long (≈ 1000 terms). This was necessary for obtaining meaningful results for the correlation exponents, as the DFA method requires series with lengths of at least two orders of magnitude.

RESULTS AND DISCUSSIONS

The length series for the coding sequences (CDS) in the genome of *B. subtilis* 168 is shown in Fig. 1. There are 4106 CDS lengths in total. A few long CDS are dispersed throughout the genome. The length series for *E. coli* K12 MG1655 genome consisting of 4341 CDS is illustrated in Fig. 2. Generally, *E. coli* CDS length series are shorter than those of *B. subtilis*.

The DFA plot for the entire genome of *B. subtilis* and *E. coli* (equivalent to the chromosome content for these bacterial species) is illustrated in Fig. 3. The two plots can be fitted by a straight line, with the slopes $\alpha = 0.644 \pm 0.007$ and $\alpha = 0.581 \pm 0.002$, respectively. However, a closer examination reveals that the plots are non-linear. This is more obvious for *B. subtilis* and only a slight non-linearity is apparent for *E. coli.* Such characteristics may suggest non-uniform organization of genomes. Hence, the segmentation procedure mentioned below is needed.

The DFA plot for the quarter segments of the genome in *E. coli* K12 MG1655 is shown in Fig. 4. The DFA plot is characterized by two linear domains for each quarter segment. These findings suggest that even the quarter segments have a non-uniform correlation, because the correlation at short distances, not larger than one order of magnitude, is different compared to large distances (the linear domains have different slopes). The DFA plots for the quarter series of *B. subtilis* genome are also non-linear (Fig. 5); therefore, the situation is qualitatively similar.

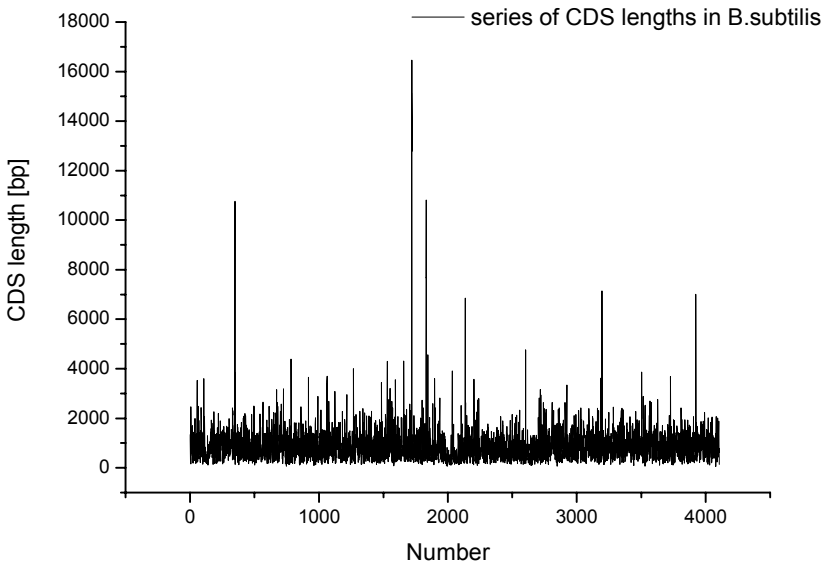

Fig. 1. Series of coding sequences (CDS) lengths for the genome of *B. subtilis* strain 168. There are 4106 coding sequences lengths, as indicated by the number on the *x*-axis. The unit for CDS length is base pairs (bp).

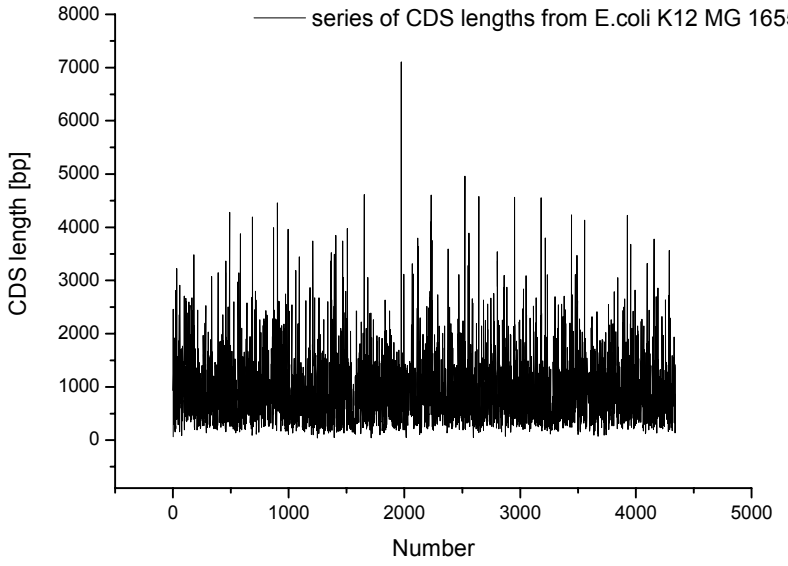

Fig. 2. Series of CDS lengths for the genome of *E. coli*, strain K12 MG 1655.

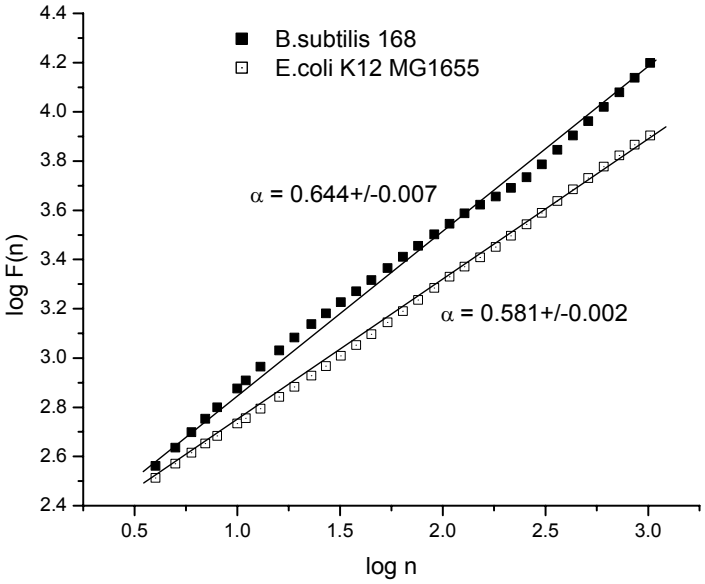

Fig. 3. Detrended fluctuation analysis of the length series for the coding sequences of *Bacillus subtilis* and *Escherichia coli*, strain K12 MG 1655.

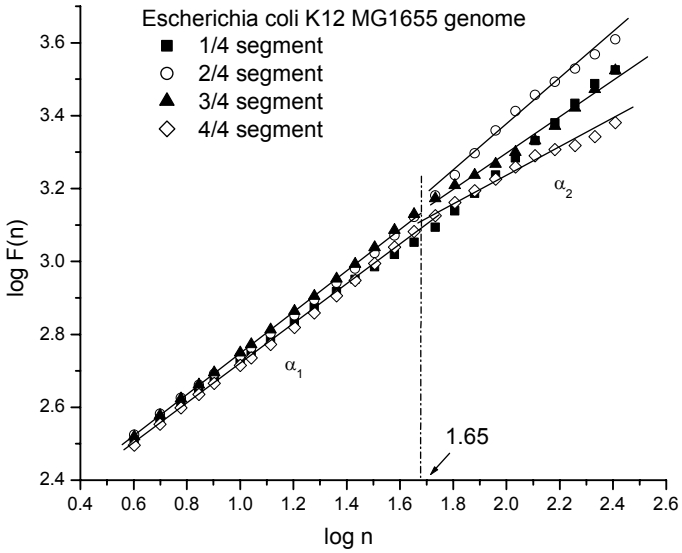

Fig. 4. Detrended fluctuation analysis plot: fluctuation function F(*n*) vs. box size *n* for the CDS length series of *E. coli*, strain K12 MG 1655. Note the non-linear characteristic of the plot: the slope of each straight line defines a correlation exponent corresponding to the box size interval. There are two obvious linear domains for each quarter segment, which indicate the existence of a short-range correlation. Their intersection is indicated by the vertical dotted line.

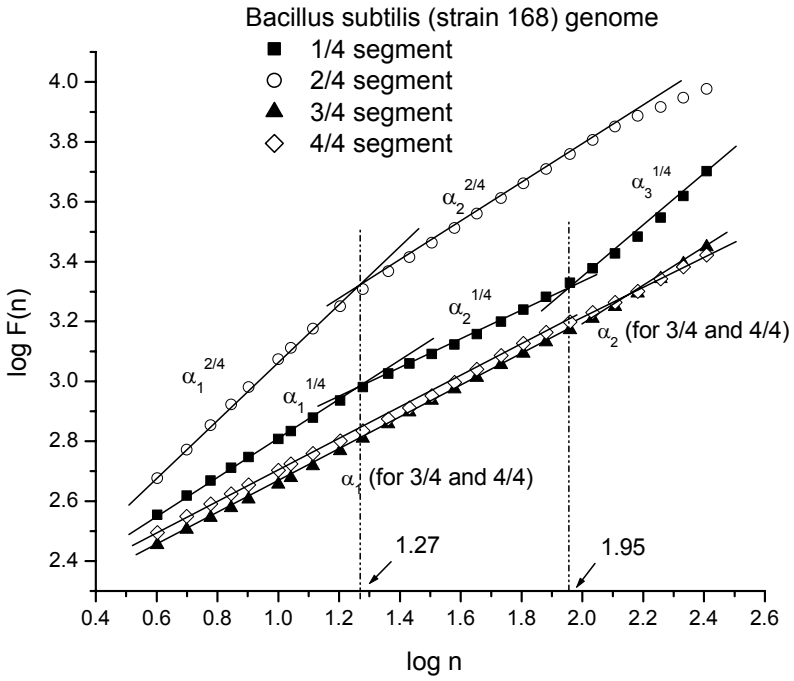

Fig. 5. Detrended fluctuation analysis plot: fluctuation function F(n) vs. box size n for the CDS length series of *B. subtilis* 168. There are two linear domains for all quarter segments, except the first one, 1/4, which has three domains. Their intersection is pointed by the two vertical dotted lines.

The DFA plots for the quarter segments of *B. subtilis* (Fig. 5) also revealed non-linear features, which were more evident for the first two quarter segments. A comparison among the DFA plots of the two species reveals significant qualitative differences. For example, the plot for the first quarter of *B. subtilis* shows three linear domains behaviour. Also, the first segment has an unusual high value for the correlation exponent α_1.

The α_1 and α_2 values corresponding to the four segments are further illustrated in Figs. 6 and 7. These representations allow a better comparison of the data. *E. coli* quarter segments show a rather low correlation, close to randomness, so the α_1 correlation exponent has values slightly above 0.5 (Fig. 5). On the other hand, α_2 shows a great variability among the quarter segments. The highest order or correlation is seen in the first quarter segment. This decreases steadily until the fourth quarter segment reaches a significant anti-correlation value $\alpha_2 \approx 0.35$. In other words, longer CDS are more probably followed by shorter CDS. The picture is different for *B. subtilis* (Fig. 6). The outstanding feature is the high value of α_1 correlation for the second quarter segment.

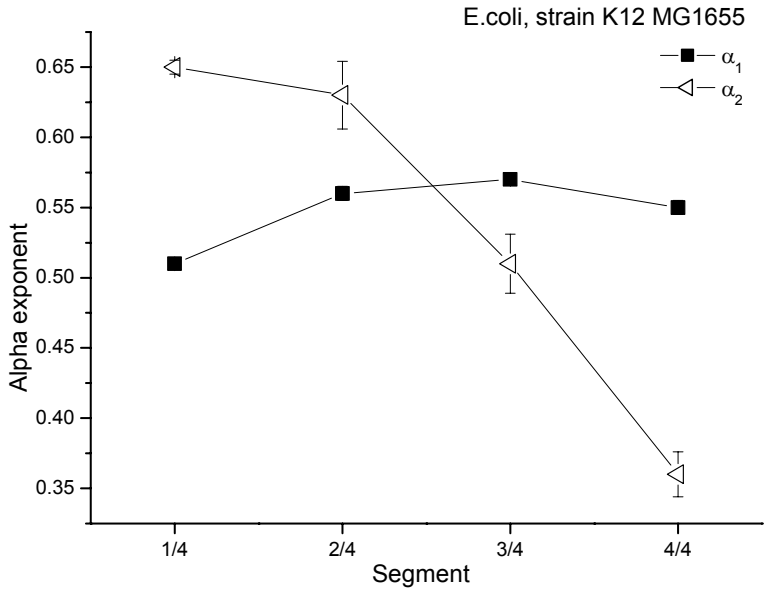

Fig. 6. Correlation exponents α_1 and α_2 for quarter segments of the coding sequence length series of *E. coli*, strain K12 MG 1655.

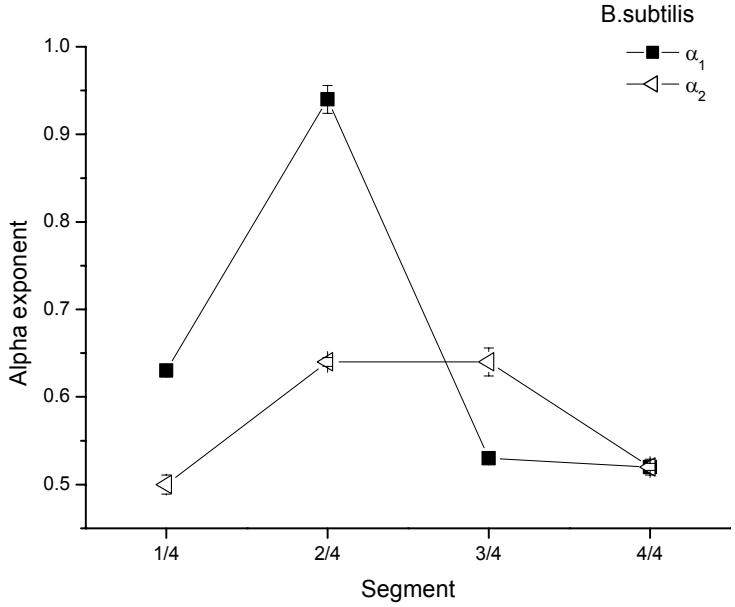

Fig. 7. Correlation exponents α_1 and α_2 for quarter segments of the coding sequence length series of *B. subtilis*.

The peculiar feature of the two bacterial species is that the correlation characteristics are reversed: the α_1 values for *E. coli* correspond to the α_2 values for *B. subtilis*.

Another characteristic of the DFA plots is the crossover location (Figs. 4–5). This is at log n = 1.67 for *E. coli*, and at log n = 1.27 and log n = 1.95 for *B. subtilis.* The corresponding values for n are 47, 19 and 90. This means, for example, that the first correlation exponent α_1 extends over the first 47 terms (that is the CDS lengths), while α_2, from this value onwards. It can be seen that for *B. subtilis* the first correlation domain is smaller than the one for *E. coli* (19 compared to 47). Therefore, differences between the correlation characteristics of the two bacterial species refer both to the magnitude and the correlation domain.

We looked for the biological significance of these findings. The results point to a strong correlation in the second quarter of *B. subtilis* genome and in the first quarter of *E. coli*. Such a local strong correlation may be the result of similar CDS lengths located in close proximity. In order to explain such a close proximity, we hypothesize that gene duplication might be the source of this correlation. An important percentage of duplicate genes (17–44%) may comprise the genome [12]. On the other hand, it is known that functional related genes are located close to each other in the chromosome [10]. While duplicated genes may increase the correlations of CDS length series, they are known to fulfill different functions.

CONCLUSIONS

Both *B. subtilis* and *E. coli* are characterized by a non-uniform organization of CDS lengths. This implies a short-range correlation, which is here described by two short-range correlation exponents. The non-uniform organization of the length series is revealed by the great variability of the correlation exponent in the four quarter segments of each genome. This suggests that genes with a similar length have a preferential location in these segments. There are notable differences between the two bacteria species, regarding the correlation exponents and the correlation domain. This might be helpful in comparative investigation of various bacterial genomes.

Acknowledgements. This work was funded by Romanian Authority for Scientific Research, Grant no. CEEX-05-D11-52/07.10.2005.

REFERENCES

1. ALLEN, T.A., N.D. PRICE, A.R. JOYCE, B.Ø. PALSSON, Long range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organisation, *PLoS Computational Biology*, 2006, **2(1)**, 0013–0021
2. AUDIT, B., C.A. OUZOUNIS, From genes to genomes: universal scale-invariant properties of microbial chromosome organisation, *J. Mol. Biol.*, 2003, **332,** 617–633.

3. BULDYREV, S.V., N.V. DOKHOLYAN, A.L. GOLDBERGER, S. HAVLIN, C.-K. PENG, H.E. STANLEY, G.M. VISWANATHAN, Analysis of DNA sequences using methods of statistical physics, *Physica A,* 1998, **249,** 430–438.
4. CARPENTIER, A-S., B. TORRÉSANI, A. GROSSMANN, A. HÉNAUT, Decoding the nucleoid organization of *Bacillus subtilis* and *Escherichia coli* through gene expression data, *BMC Genomics*, 2005, **6**, 84.
5. DANCHIN, A., P. GUERDOUX-JAMET, I. MOSZER, P. NITSCHKE, Mapping the bacterial cell architecture into the chromosome, *Philos. Trans. R. Soc. Lond. B*, 2000, **355**, 179–190.
6. JOSÉ, M.V., T. GOVEZENSKY, J.R. BOBADILLA, Statistical properties of DNA sequences revisited: the role of inverse bilateral symmetry in bacterial chromosomes, *Physica A,* 2005, **351**, 477–498.
7. LOVETT, S.T., A.M. SEGALL, New views of the bacterial chromosomes, *EMBO Rep.*, 2004, **5**, 860–864.
8. MORARIU, V.V., L. BUIMAGA-IARINCA, C. VAMOS, S.M. SOLTUZ, Detrended fluctuation analysis of autoregressive processes, *Fluct. Noise Lett.*, 2007, **7**, L249–L255.
9. PENG, C-K, S. HAVLIN, H.E. STANLEY, A.L. GOLDBERG, Quantification of scaling exponents and crossover phenomena in non-stationary heartbeat time series, *Chaos,* 1995, **5**, 82–87.
10. TAMAMES, J., G. CASARI, C.A. OUZOUNIS, A. VALENCIA, Conserved clusters of functionally related genes in bacterial genomes, *J. Mol. Evol*., 1997, **44**, 66–73.
11. YU, Z-G., V.V. ANH, B. WANG, Correlation property of length sequences based on global structure of the complete genome, *Phys. Rev. E.,* 2001, **63**, 011903.
12. ZHANG, J., Evolution by gene duplication: an update, *Trends in Ecology and Evolution,* 2003, **18,** 292–298.