U.P.B. Sci. Bull., Series A, Vol. 74, Iss. 1, 2012 ISSN 1223-7027

STRUCTURE FUNCTIONS AND MULTIFRACTAL DETRENDED FLUCTUATION ANALYSIS APPLIED TO THE CODING SEQUENCES: CASE STUDY - ESCHERICHIA COLI

Cristina STAN[1], Teofil MINEA[2], Teodora-Maria CRISTESCU[3], Luiza BUIMAGA-IARINCA[4], Constantin P. CRISTESCU[5]

În această lucrare prezentăm analiza multifractală a secvențelor de codare ale seriilor genomice şi aplicăm metodele studiate pe Escherichia Coli. Programele de calcul au fost realizate prin implementarea în Mathematica a algoritmilor pentru funcțiile de structură şi analiza multifractală bazată pe eliminarea tendințelor.

In this paper we present the multifractal analysis for the genomic coding sequences and apply the method to Escherichia Coli. The computer programs were implemented in Mathematica for two specific algorithms: structure functions (SF) and multifractal detrended fluctuation analysis (MF-DFA).

Keywords: structure functions, MF-DFA, Hurst exponent, genomic sequences

1. Introduction

A wide range of experimental signals (time-series or data sequences) from physics, biology, medicine, econophysics, etc. can be well modeled by multifractal processes [1-5]. The essence of multifractal analysis is to identify fractal dimensions of self-similar structures with varying regularities and to produce the distribution of indices of singularity, which constitutes the multifractal spectrum.

The multifractal formalism has been implemented using different algorithms, such as rescaled range analysis [6], wavelets analysis [7], detrended fluctuation analysis [8], fluctuation measurement by structure functions or singular measures [9], etc.

In this work we apply the methods based on structure functions and MF-DFA to the study of the multifractality of genome coding sequences, particularly for Escherichia Coli (EColi).

[1] Assoc. Prof., Department of Physics, Faculty of Applied Science, University POLITEHNICA of Bucharest (PUB), Romania, e-mail:cstan@physics.pub.ro

[2] Master Student, Faculty of Applied Science, UPB, e-mail: teofil_minea@yahoo.com

[3] Master Student, Faculty of Electronics and Telecommunications, UPB; mon1yka@yahoo.com

[4] PhD, Nat. Inst. of Res. and Dev. Isotop. Molec. Techn., Cluj-Napoca; iarinca@itim-cj.ro;

[5] Prof., Department of Physics, Faculty of Applied Science, UPB; cpcris@physics.pub.ro

2. Theoretical considerations

Let us consider an arbitrary signal $f(t_i)$, $(i=1,2,\cdots,N)$. If it represents a Wiener process (fractional Brownian motion), its variance is proportional to the time interval for which it is computed, power to the Hurst (*H*) exponent: i.e. $\sqrt{\left\langle (\Delta f)^2 \right\rangle} \propto (\Delta t)^H$ where H is ranging between 0 and 1. Hurst exponent describes the degree of the predictability of any signal and the long memory properties involved in the signal. The value of 0.5 is characteristic for the Brownian motion which is entirely non correlated (no memory). A value of $0 < H < 0.5$ indicates an anti-persistent signal (e.g. a decrease will more probably be followed by an increase), and a value of $0.5 < H < 1$ indicates a persistent signal (e.g. an increase will more probably be followed by another increase).

The first technique followed by us to evaluate the Hurst exponents uses the scaling properties of the structure functions (SF). The procedure is applicable to nonstationary data sequences with stationary gradients.

Structure function of order q>0 is defined as [10]:

$$S_q(\tau) = \left\langle \left(\left| f(t_i+\tau) - f(t_i) \right| \right)^q \right\rangle \tag{1}$$

$$S_q(\tau) = C_q \tau^{\zeta(q)} = C_q \tau^{qH(q)}. \tag{2}$$

Here C_q can depend slightly on q comparing with any power of τ. The log-log plot of $S_q(\tau)$ versus τ is a line with the slope $\zeta(q)$. For multifractal signals, $\zeta(q)$ versus q has a nonlinear dependence and H is not constant as in the case of monofractals and is a function of q:

$$H(q) = \frac{\zeta(q)}{q}. \tag{3}$$

The main Hurst exponent is computed for $q=1$.

The concept of multifractality refers to the fact that different sections of the series (different zones of the fractal object) are characterized by different values of the fractal dimension. The multifractal spectrum $D(h)$ can be computed using the Legendre transform of the structure function exponents $\zeta(q)$:

$$D_q(h) = \min_q\left(qh - \zeta(q) + 1\right). \tag{4}$$

The condition of minimum $\mathrm{d}D_q(h)/\mathrm{d}q = 0$ implies:

$$\mathrm{d}\zeta(q)/\mathrm{d}q = h. \tag{5}$$

Using relation (5), the multifractal spectrum can be determined by:

$$D(h) = qh - \zeta(q) + 1. \tag{6}$$

In this way, the structure functions analysis allows the computation of both the Hurst exponent and the order-dependence of the fractal dimension.

The second technique used by us to evaluate the Hurst exponents is MF-DFA analysis. In the first step we determine the "profile" as follows:

$$Y(j)=\sum_{i=1}^{j}(f_i-\langle f\rangle);\quad j=1,2,...,N \tag{7}$$

where <*f*> is the mean. Then, the signal is divided in $2N_s$ nonoverlaping segments (N_s=Int(N/s) of length *s,* obtained from the start to the end and reverse. The next stage deals with the detrended procedure using the best polynomial fit of the signal (y_υ) on each segment υ. This procedure implies the computation of the variances [11]:

$$F^2(s,\upsilon)=\frac{1}{s}\sum_{j=1}^{s}\left[Y\left((\upsilon-1)s+j\right)-y_\upsilon(j)\right]^2,\quad \upsilon=1,2,...N_s \tag{8}$$

$$F^2(s,\upsilon)=\frac{1}{s}\sum_{j=1}^{s}\left[Y\left(N-(\upsilon-N_s)s+j\right)-y_\upsilon(j)\right]^2,\quad \upsilon=N_s+1,...,2N \tag{9}$$

The fluctuation function of order 2 is defined as the square root of the relation (8) and (9). For the general case of a *q* order fluctuation function (*q* positive or negative, nonzero value) the formula can be modified as:

$$F_q(s)=\left\{\frac{1}{2N_s}\sum_{\upsilon=1}^{2N_s}\left[F^2(s,\upsilon)\right]^{q/2}\right\}^{1/q}. \tag{10}$$

For a self-similar signal the dependence of the fluctuation function on the "window" length *s* is expected to be exponential as $F_q(s)\sim s^{h(q)}$. The main Hurst exponent is computed for *q*=2. The log-log plot of the fluctuation exponent versus *s* is a line with the slope *h*(*q*) called generalized Hurst exponents.

3. Case study: coding sequences of Escherichia Coli

The local properties of the DNA sequence prove to be more informative than the global one in distinguishing coding and non-coding sequences. Recent work reported the important significance of the length and distribution of proteins (which is similar to the coding sequences) [12-15]. This is due to the fact that there is a profound relationship between protein length distributions and the mechanism of protein length evolution. As a result, the protein length distribution represents a comprehensive record of the evolutionary history of a species.

Our data were taken from the National Centre for Biotechnology Information (NCBI) website [16] and manipulated as in a signal consisting of sequences of coding length [2]. Figure 1 shows the coding sequences length (CDS) versus location of the Ecoli. The requirement of stationarity is fulfilled by the genomic data as proved by Fig.2, where the structure function is plotted as log-log graph versus the delay.

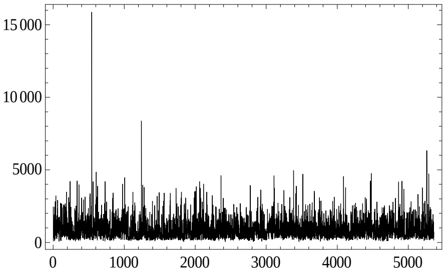

Fig.1. The CDS data sequences

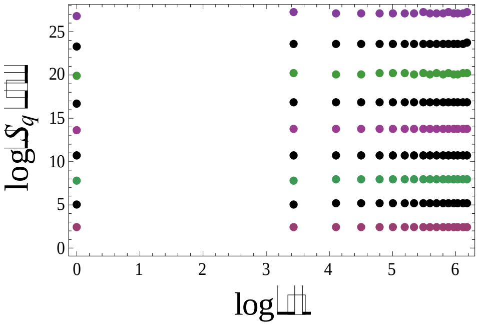

Fig.2. The log-log plot of the SF for the original data: *q* from 0.6 to 3 with step of 0.3 (bottom to top)

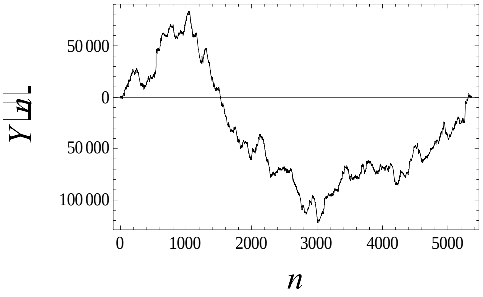

Fig.3. The CDS profile

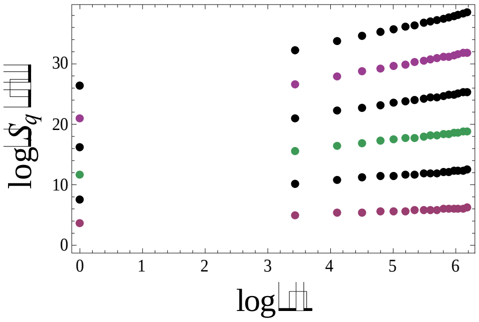

Fig.4. The log-log plot of SF for the profile: *q* from 0.5 to 3 with step of 0.5 (bottom to top)

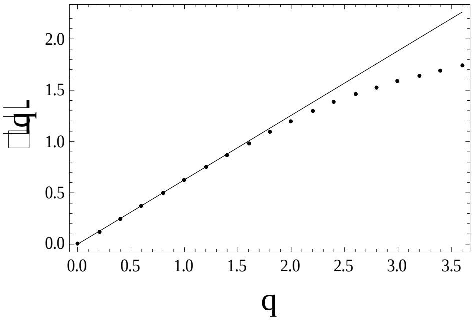

Fig.5. The structure function exponent versus *q* for the profile

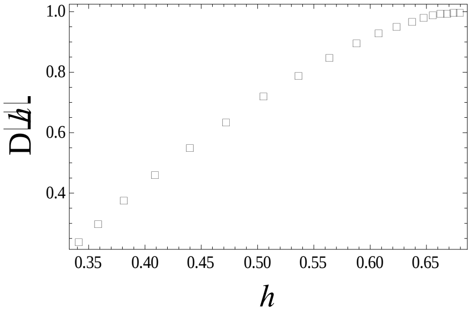

Fig.6. The fractal dimension versus the generalized Hurst exponent (SF)

The stationarity is reflected by the fact that the slopes of the curves are practically zero for all values of *q*. Consequently, the genomic data can be considered as the gradient of another process obtained by integration of the original data [10].

The main Hurst exponent will be obtained by applying the SF algorithm to the integrated series, usually known as the profile (Fig.3).

The dependence of the SF for the profile versus the delay for the specified values of *q* is plotted in Fig.4.

As we can see from Fig.5, the computed structure function exponent versus *q* is a nonlinear function showing the multifractal characteristics of the data. From $\zeta(q)$ we compute the main value of *H* using (3) for *q*=1.

The fractal dimension versus the generalized Hurst exponent is plotted in Fig.6.

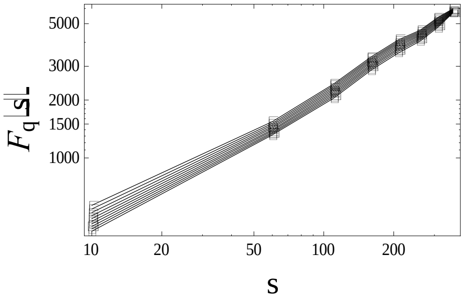

Fig.7. The fluctuation function in log-log scale for CDS profile: *q* from 0.6 to 3 with step of 0.3 (bottom to top) - (MF-DFA)

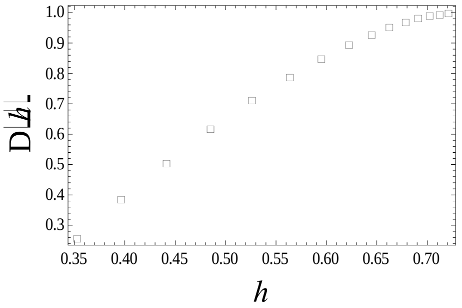

Fig.8. The fractal dimension versus the generalized Hurst exponent (MF-DFA)

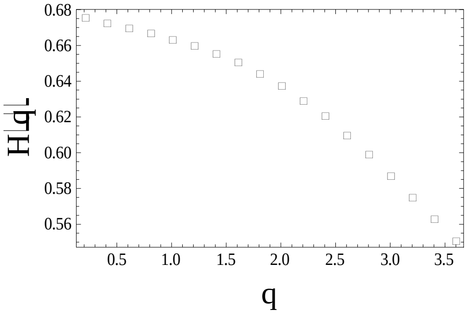

Fig.9. Generalized Hurst exponents versus *q* (SF)

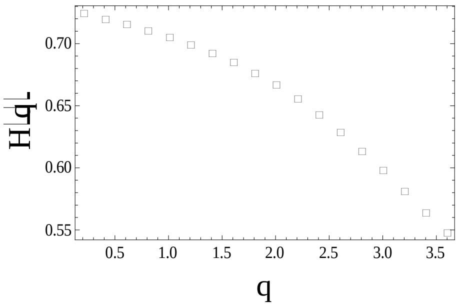

Fig.10. Generalized Hurst exponents versus *q* (MF-DFA)

For the MF-DFA analysis we illustrate the results of the fluctuation functions versus *s* for the specified values of *q* in Fig.7 and the multifractal spectrum computed with in Fig.8. The lines in Fig.7 are drawn for eyes guiding.

The computation of fractal dimension spectrum is detailed in [10].

Figures 9 and 10 present the dependence of the generalized Hurst exponents on *q* as computed from the SF and the MF-DFA algorithms, respectively. We observe the very good consistency of the results. The same conclusion can be obtained from the comparison of the fractal dimension spectrum from SF algorithm shown in Fig.6 and from MF-DFA algorithm shown in Fig.8.

The values of the main Hurst exponents computed using the two algorithms are: 0.663 with SF and 0.667 with MF-DFA. The remarkable coincidence of the two values from our analysis and the similar values reported using the wavelet method [17] confirms the correctness of our results.

6. Conclusions

Using the implementation in Mathematica of an improved SF algorithm and MF-DFA we demonstrate consistency of the results obtained for the EColi genomic sequence with results obtained by considerable more elaborate methods, such as the wavelet analysis. The good agreement encourages us to consider that the SF algorithm can be successfully applied to any genomic sequence and at the same time to benefit for the low computation cost of this algorithm.

REFERENCECES

[1] *J. D. Murray*, Mathematical Biology, Springer-Verlag Berlin, Heidelberg 2002.

[2] *O. Zainea, V. Morariu*, Fluctuation and Noise Letters, 7 (4), 2007, L501-L506.

[3] *L. Seuront, F. Schmitt, Y. Lagadeuc, D. Schertzer, S. Lovejoy,* J. Plankt. Res. 21, 1999, 877-22.

[4] *E. I. Scarlat*, *C. Stan, C. P. Cristescu,* Physica A, 379 (1), 2007, 188-198.

[5] *C. P. Cristescu, C. Stan, E. I. Scarlat,* UPB Sci. Bull., Series A, 69 (3), 2007, 37-45.

[6] J. Beran, Statistics for long-memory processes, Chapman & Hall, 1994.

[7] *J. F. Muzy, E. Bacry, A. Arneodo*, *Phys. Rev. Lett.* 67, pp. 3515-3518, 1991.

[8] *Peng C.-K, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberg*, Phys. Rev E 49, 1994, 1685-1689.

[9] *A. Davis, A. Marshak, W. Wiscombe, R. Cahalan*, J. Geograph. Res., 99,1994, 8055-8072.

[10] *C. X. Yu, M. Gilmore, W. A. Peebles, and T. L. Rhodes*, Phys. Plasmas 10, 2003, 2772 -83.

[11] *J. W. Kantelhardt, , S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, H. E. Stanley,* Physica A, 2002, 87-114.

[12] *C. Stan, T. Minea, T. M. Cristescu, L. Buimaga-Iarinca, V. Morariu and C. P. Cristescu,* Proc. GSP2011, 2-nd Intern. Workshop Genomic Sign. Processing, Bucharest, Jun. 2011, p. 87-90

[13] *D. J. Li, S. Zhang*, arXiv:0806.0205v1 [q-bio.GN]

[14] *D. J. Li, S. Zhang*, Modern Physics Letters B, 23, 2009, 3563-3580.

[15] *J. Song, A. Ware, S.-L. Liu*, BMC Genomics, 4 2003, 17-25.

[16] http://www.ncbi.nih.gov

[17] *A. Arneodo, B. Audit, N. Decoster, J.-F. Muzy & C. Vaillant,* In *The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes* (A. Bunde, J. Kropp & H. J. Schellnhuber, eds.), pp. 26-102. Springer Verlag, Berlin, 2002.