THE STATISTICAL PHYSICS OF MICROBIAL GENOMES: PART I. ORGANISATION OF CODING SEQUENCES IN THE CHROMOSOME OF *ESCHERICHIA COLI*

V.V. MORARIU[*], *OANA ZAINEA*[*], *A. BENDE*[*], *O. POPESCU*[**]

*Department of Molecular and Biomolecular Physics, National Institute for R&D of Isotopic and Molecular Technology, P.O.Box 700, Cluj-Napoca 5, 400293, Romania, E-mail: vvm@L40.itim-cj.ro
**Faculty of Biology, "Babeş-Bolyai" University, Cluj-Napoca, Romania

Abstract. This paper shows that the correlation characteristics of coding length sequences for several strains of *E. coli* have a non-random organization. Although modest, this correlation at the level of the chromosome is statistically significant. The methods of analysis included spectral analysis, detrended fluctuation analysis, Hurst exponent and correlation dimension. The methods were checked up against series of data with well-known characteristics. It is shown that the equivalent literature data are in serious error. Our analysis suggests that the organization of coding length sequences seems to be non-uniform, therefore, the evaluation as a simple long-range correlated system is only an approximation. The coding sequences are definitely not dynamic systems.

Key words: bacteria, chromosome, coding sequences, *Escherichia coli*, long-range correlation.

INTRODUCTION

Recent investigations revealed that there is no characteristic size pertinent to the description of chromosomes. Gene positions and gene orientations for many bacteria share a common scale invariant property known as long-range correlation [2]. This was regarded as an operon-like organization at all scales and implies that a complete scale range extending over more than three orders of magnitudes of chromosome segment lengths is necessary to properly describe the genome organization of prokaryotes [2]. More recently it has been shown that the large-scale organization of 135 bacterial and 16 archaeal organisms is significantly non-random [1]. These new results restarted the interest for the statistical physics approach to DNA investigation while the older investigations revealed random uncorrelated organization of nucleotides (or close to randomness) in coding sequences [3]. These earlier results proved to be rather disappointing especially for

Received April 2006;
in final form May 2006.

biologists who did not find of much use either the randomness of coding sequences or the non-randomness of non-coding sequences.

Our work is focused on the series of data composed of the lengths of coding sequences. The question is similar to previous papers in the field: is there any evidence for order in such series? Our work starts from a time series model based on the global structure of the complete genome [6]. This model is based on counting first the length of the coding and non-coding sequences respectively in terms of nucleotide contents. As a result, three kinds of integer sequences can be obtained: a) A coding length sequence represented by the successive gene lengths; b) A non-coding length sequence and, c) A whole length sequence represented by coding length sequence and non-coding length sequence replaced by their negative numbers respectively. Each of these three kinds of series is regarded as time series which can be subsequently subjected to various types of linear and nonlinear analysis in order to establish the degree of order or randomness. In the original paper the Hurst exponent and the correlation dimension were determined for a number of 21 bacteria and archaea [6]. The series consisting of the length of coding sequences was reported to indicate a significant long-range correlation or anti-correlation and a low correlation dimension depending on the species. These results are quite intriguing as the coding sequences based on the DNA walk model were basically found to be random as mentioned above [3]. Further the use of correlation dimension for a system which does not seem to be a dynamical system is here regarded as inappropriate. Consequently, both parameters seem quite unusual to indicate a significant degree of order in the series. Further, the power spectra of the series also appeared to be quite unusual by showing several negative peaks [5].

We decided to check up all these results by using established methods of analysis. The methods were first verified on series with well-known properties. We here report results which are thought to be the correct ones. They essentially show that the coding length sequences have a low yet a distinct long-range correlation. These results are totally at variance with the results published in the literature [5, 6].

MATERIALS AND METHODS

The genome of *Escherichia coli*, of the following strains: O157:H7 EDL933 (5,528445 base pairs); CFT073 (5,231428 bp); K12MG1655 (4,639675 bp); OH157:H7 (5,498450 bp) were chosen for investigation. The genome structures were downloaded from EMBL–EBI Data Bank. A program was written to extract the length of the coding sequences from the original files. This resulted for example in 5347 lengths of coding sequences for O157:H7 EDL933 strain, therefore a series consisting of 5347 terms. The length of coding sequences series of data were subject to the following analysis: Fast Fourier Transform (FFT), Detrended Fluctuation Analysis (DFA), Hurst Exponent Analysis (HEA), and Correlation Dimension Analysis (CDA). All these methods are well known in statistical physics, fractals and chaos. However, considering that different results were obtained compared to literature data we specify the source of the programs.

FFT was a facility in the ORIGIN program. The double log plot of the spectrum was fitted by a line and its slope was exponent β of the power law describing the spectrum ($P = 1/f^{\beta}$). Beta exponent varies between 0 for random series and almost 2 for Brownian noise. DFA program was written by us and it was checked out on characteristic series with well-known properties. DFA was in fact a DFA-1 program where the trend was approximated by a first degree polynomial. The result of DFA analysis is the so-called α exponent. Its value is 0.5 for random series and 1.5 for Brownian noise. In case of stationary series $\alpha = \beta = 1$ for $1/f$ noise. The relationship between the two exponents is $\beta = 2\alpha - 1$ but only for stationary series. The advantage of using DFA is the removal of trends or non-stationary contribution to correlation. HEA and CDA calculation was provided by the Sprott and Rowlands Chaos Data Analyzer, the professional version program. It is important to know that the methods were developed for stationary data as practically all the methods of linear and nonlinear methods of analysis. We checked the programs by using series of normal random numbers generated by the computer. Pink noise ($1/f$ noise), Brownian noise, and Henon attractor series of data were provided by Sprott and Rowlands program. The result of the check up is included in the table below.

Table 1

Check up of the programs on characteristic series of data

Method of analysis[*]	Parameters	Normal random numbers	Pink noise ($1/f$ noise)	Brownian noise	Henon attractor
FFT	β exponent	0.029 ± 0.028	0.994 ± 0.029	1.79 ± 0.048	Not defined
DFA	α exponent	0.492 ± 0.003	0.949 ± 0.003	1.486 ± 0.008	Not defined
HEA	Hurst exponent	–0.029	0.1597	0.526	–0.03.15
CDA	Correlation dimension	No dimension	4.582±0.116	≈0.45	1.239±0.08

[*] See Materials and Methods

This check up serves to check if all methods used in this work give reliable results. So, the normal random numbers are correctly characterized by a close to zero β exponent a close to 0.5 α exponent and a close to zero Hurst exponent. Such a series of data have no correlation dimension. Further, pink noise is characterized by beta and alpha exponents close to 1 and an anti-persistent Hurst exponent. Its correlation dimension is between 4 and 5. According to Sprott and Rowlands such values indicate randomness. However, we know from other sources [4] that such a low dimensional correlation dimension is an artifact which appears in $1/f$ noise and generally in fractional Brownian noise ($1/f^{\beta}$ where $0 < \beta < 2$). So, we rather use this test for further proving that the series is of $1/f$ or $1/f^{\beta}$ type. The Brownian noise is theoretically expected to have $\beta = 2$, but in reality all data sources experimented by us produced a lower value β ≈ 1.8. On the other hand, α is practically 1.5 as expected for Brownian noise. The Hurst exponent is slightly persistent as it is higher than the border line at 0.5. Henon attractor is a true dynamical system so it has a distinct and clear low correlation dimension while Hurst exponent is close to zero which indicates randomness of the series (deterministic randomness).

RESULTS AND DISCUSSION

An example of series of coding lengths is illustrated in Fig. 1. There are a few peaks which indicated longer coding sequences. An important feature of the plot is the apparent stationary characteristic, i.e. there are no visible trends in the plot. Although this does not represent a proof, the application of the methods of analysis appears to be acceptable.

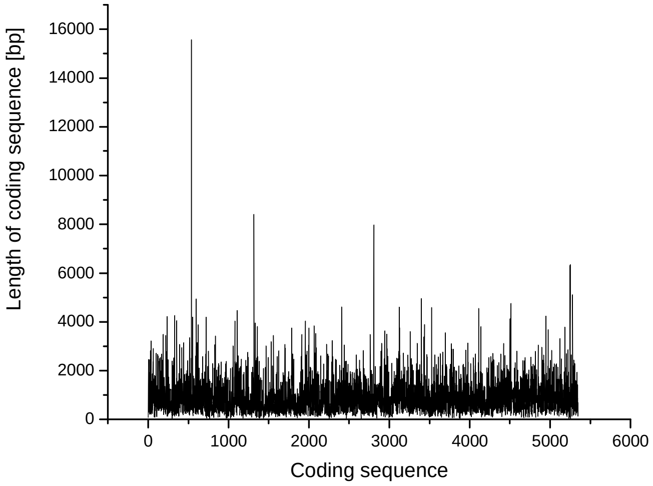

Fig. 1. Series of coding length sequences in the genome of *E. coli*, strain 0157 H7 EDL933.

The spectrum of this series is illustrated in Fig .2. The general look of the plot reminds about a typical 1/*f* spectrum. However, it is quite easy to observe that it has a biphasic character as two distinct slopes are evident at lower and higher frequencies respectively. The higher slope extends over at least two orders of magnitudes therefore one can say that a long-range correlation is present among the lower frequencies events.

DFA plot also shows that certain non-homogeneity of the correlation is present in the data as the plot is not a simple straight line. The straight fitting line has a slope α = 0.499 ± 0.008. This value which, in fact, is practically 0.5, indicates randomness. While such a feature might be regarded as an overall characteristic, the nonlinearity of the plot (Fig. 3) suggests that a certain part of the series deviates from randomness just as the spectral characteristic does.

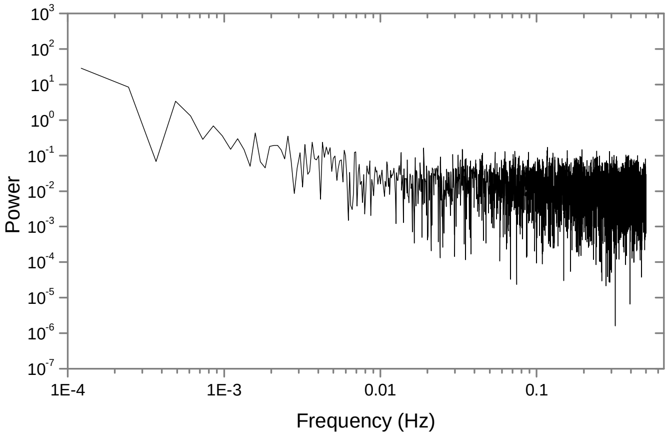

Fig. 2. Spectrum of coding length sequence in the genome of *E. coli* strain 0157 H7 EDL933.

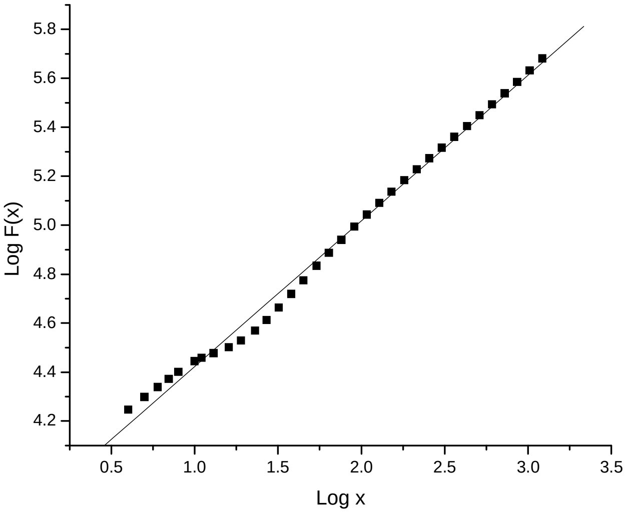

Fig. 3. Detrended fluctuation analysis of coding length sequence in the genome of *E. coli*, strain 0157 H7 EDL933.

The results for all four *E. coli* strains and for the unspecified strain reported in the literature [5, 6] are included in Table 2. The spectral characteristics for the other strains are similar except for strain K12MG1655 where a single beta exponent is evident. Its value is distinct to a slope zero which shows a residual correlation of the data. The other strains present a significant long-range correlation in the range of low frequencies. The values of beta exponent vary between 1.18 ÷ 1.39 which points out to an unusual strong long-range correlation. This kind of behavior was not mentioned before in the literature.

Table 2

Correlation characteristics of the length of coding sequences in *E. coli* strains

E. coli strain	Correlation dimension	Hurst exponent	DFA alpha exponent	Spectral beta exponent*	
O157:H7EDL933	4.3	0.0172	0.499±0.008	1.389 ± 0.138	0.157 ± 0.025
CFT073	4.282	0.0178	0.557±0.005	1.178 ± 0.168	0.177 ± 0.025
K12MG1655	4.43	0.0132	0.542±0.002	0.157±0.03	
OH157:H7	4.126	0.0154	0.561±0.003	1.392 ± 0.17	0.194 ± 0.025
Unspecified	3.098 [6]	0.5985 [6]	0.620 [5]	0.055 [5]	

*First column: low frequency range; second column: high frequency range.

The higher frequency range is characterized by a residual long-range correlation quite similar among the different strains. Although significant, from the statistical point of view, this is a modest value for a long-range correlation and it is doubtful to have a particular significance.

The literature [5] however presents a totally different spectrum compared to ours. In fact many of the spectra in the same reference for different microbial species show peculiar spectral shapes [5]. We noticed such abnormal spectra when the coding sequences were wrongly extracted from the original downloaded data. The series of data contained huge outliers which were not real, but simple computational faults. The FFT of these series resulted in spectra with one or more large negative peaks. Next, the authors of the paper of concern ignored the presence of these peaks in the spectra and considered only the "clean" lower frequency part of the spectra [5]. The figure in Table 2 is the result of such a calculation ($\beta = 0.055$). This value is however meaningless. It additionally can be criticized for not including the standard error. Therefore, β values reported by this paper are in serious doubt and offer a totally wrong picture on the correlation of the coding sequences.

The exponent α reported in Table 2 for *E. coli* strain also suggests a modest long-range correlation as the values are greater than 0.5. They are statistically significant. This is in agreement with the spectral data of β exponent. However, this analysis makes confident that the spectral analysis is not influenced by non stationary contributions as in the case of β exponent.

The Hurst exponent has close to zero values for all strains except the literature data where $H = 0.599$. This result adds to the list of disagreements with the present work.

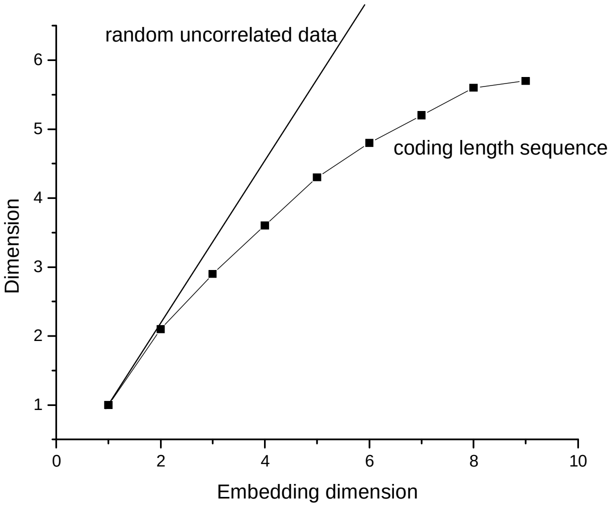

Fig. 4. Correlation dimension analysis of coding length sequence in the genome of *E. coli*, strain 0157 H7 EDL933.

Finally we discuss the correlation dimension. Again, a significant difference can be noticed between the present work and literature. Reference [6] does not report the correlation dimension versus embedding dimension in order to assess the plateau value of the correlation dimension. The authors only mention in their work that the estimate of the correlation dimension is taken when $D_2(m)$ does not change with m increasing. This is formally correct (Their notation is equivalent to our plot in Fig. 4. It should be reminded that it is common in the literature to use the term *correlation dimension* in an ambiguous way. A correlation dimension is calculated for a given value of embedding dimension as shown in Fig. 4. If the plot shows a clear plateau then the corresponding correlation dimension represents the attractor dimension. However, the attractor dimension is mistakenly coined as the "correlation dimension"). Further we know from Sprott and Rowlands Chaos Analyzer that such a plot does not always show a clear plateau (Fig. 4). For example all plots corresponding to our values for the correlation dimension in Table 2 represent such a case. In fact the values in Table 2 represent the choice of the Sprott and Rowlands Chaos Analyzer itself according to an unknown rule while no clear plateau exists in reality. Also, according to Sprott and Rowlands, a value of the correlation dimension higher than 5 indicates randomness. Pink noise and fractional Brownian noise are characterized by such correlation dimensions but they have no real significance as for the attractor of a dynamical system [4]. Therefore, the calculation of a "correlation dimension" for our coding sequences is not appropriate as the system is not a dynamical one. It only tells in an unspecified way that the system has an important randomness character yet it is not purely random. A true random series of data is, on the other hand, characterized by a straight line with a slope equal to one in Figure 4. Therefore, Fig. 4 seems only to vaguely tell that the series has a residual correlation. However, it should be stressed that the system under investigation is not a dynamical system as there is no clear plateau and therefore there is no low dimensional attractor present. The alternative left is that our series under investigation represent a stochastic system.

We suspected that our analysis is only a simplified approximation of the correlation characteristics of the series of data. More precisely, this means that the series of data were regarded as series with uniform correlation properties along the whole length of the chromosome. However, both the spectra and DFA plots do not show uniform linearity on the double logarithmic plots. This may suggest that the organization of coding length sequences is non-uniform, therefore, the evaluation as a simple long-range correlated system is only an approximation. We checked this by “chopping” the chromosome in smaller segments and evaluated the long range correlation characteristics of these segments. Indeed, we found that the correlation characteristics varied along the chromosome, hence the chromosome is a composite structure.

Finally, the overall conclusion of this work is that there is a weak long-range correlation in the coding length sequences of *E. coli*. We calculated what we believe to be the correct values which characterize this property of the sequences. The important aspect is that our analysis suggests that coding sequences are organized in a non-uniform manner. They do not represent simple long-range correlated systems with a uniform correlation along the whole chromosome. Further, the coding sequences definitely do no represent dynamic systems. It is likely they have a stochastic nature.

Acknowledgements. This work was supported by the Romanian Research Agency by a PNCD *CERES* grant.

REFERENCES

1. ALLEN, T.E., N.D. PRICE, A.R. JOYCE, B.O. PALSSON, Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization PLoS, *Computational Biology*, 2006, **2**, 1–9.
2. AUDIT, B., C.A. OUZOUNIS, From genes to genomes: universal scale-invariant properties of microbial chromosome organization, *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. THEILER, J., Some comments on the correlation dimension of 1/fα noise, *Phys. Lett. A.*, 1991, **155**, 480–493.
5. YU, Z.G., V. ANH, B. WANG, Correlation property of length sequences based on global structure of the complete genome, *Phys .Rev. E*, 2000, **63**, 011903.
6. YU, Z.G., V. ANH, Time series model based on global structure of complete genome, *Chaos, Solitons and Fractals*, 2001, **12**, 1827–1834.