COMPARATIVE ANALYSIS OF *BACILLUS SUBTILIS* AND *ESCHERICHIA COLI SAKAI* GENOMES. THE HIGH ORDER AUTOREGRESSIVE METHOD

LUIZA BUIMAGA-IARINCA, SILVIA NEAMTU, MIHAELA MIC, I. TURCU

Department of Molecular and Bio-molecular Physics, National Institute of R&D for Isotopic and Molecular Technology, 65–103, Donath St, P.O.Box 700 RO-400293 Cluj-Napoca, Romania, iarinca@itim-cj.ro

Abstract. Numerical series describing the length of coding sequence (CDS) of different bacterial genomes have been analyzed in a systematic manner. The differences and also the similarities shown by two investigated bacteria have been found and discussed. The statistical tool used in our study was the high order autoregressive model (AR*p*) which proves to be suitable to describe, in a simple way, the succession of length in the bacterial genome. The main result of the paper is that the succession of coding sequences in the bacterial genome is short-range correlated.

Key words: bacterial genome, coding sequences, autoregressive model, short-range correlation.

INTRODUCTION

The bacterial chromosome has a relatively simple structure given by a succession of coding and non-coding sequences. Each sequence consists of a specific chain of nucleobases. The coding sequences (CDS) correspond to genes and dominate the total content of nucleobases in the bacterial DNA. Recently published papers reveal the significance of the length distribution of the encoded proteins [5, 6, 7] signaling the possibility of similar correlations among the coding sequences.

A bacterial chromosome may contain thousands of coding sequences. The corresponding series has the same number of terms, each term being given by the length of corresponding CDS. The correlation properties of such series are suitable for analysis by statistical physics methods [2]. The main issue of such investigations is to find out if the length of coding sequence is randomly distributed along the genome or there are a number of rules which cause some kind of organization of the genome at this descriptional level.

Received: June 2010;
in final form June 2010.

ROMANIAN J. BIOPHYS., Vol. 20, No. 3, P. 203–211, BUCHAREST, 2010

An implicit hypothesis of earlier studies is that the distribution of CDS in the genome is random, *i.e*. the succession of the CDS lengths consists of uncorrelated data [9]. However, recent investigations suggest that some order exists in the CDS length series [11, 13] stressing the potential relevance of this property.

The main goal of this paper is to analyze in a systematic manner series consisting of CDS for different bacterial genomes and to evidence both the differences between the investigated species and also the similarities shown by various cell strains of the same species. The statistical tool used in our investigation is the high order autoregressive model (AR*p*). We are interested also in checking out the validity of autoregressive modeling procedure used for extracting the correlation properties. Specifically, special attention will be paid to the optimization of the AR*p* model order, suitable to describe, in a simple way, the succession of CDS length in the bacterial genome.

MATERIALS AND METHODS

BACTERIAL GENOME

The data considered for analysis consist of numerical series which correspond to the lengths of coding sequence from the genome of two bacteria: *Bacillus subtilis* and *Escherichia coli* respectively. The choice was motivated on the simple ground that they are two microbial organisms frequently used as examples.

The raw data have been obtained from the web site of European Molecular Biology Laboratory (EMBL) [14]. The original sequence file was compiled with Matlab R2008a (The MathWorks, Inc., Natick, MA, U.S.A.) in order to read out the start and end position of the CDS. The length of CDS was extracted by calculating the difference between the start and the end position.

AUTOREGRESSIVE MODEL

A discrete stochastic process $\{X_n,\ n = 0, \pm1, \pm2\ldots\}$ is called autoregressive process of order p, denoted ARp, if $\{X_n\}$ is stationary [10] and for any n:

$$X_n - \phi_1 X_{n-1} - \ldots - \phi_p X_{n-p} = Z_n \tag{1}$$

where $\{Z_n\}$ is a Gaussian white noise with zero mean and variance σ^2 .The real parameters ϕ_i, $i = 1,..,\ p$, can be interpreted as a measure of the influence of term i on the term i+1.

The properties of ARp processes have been studied in detail becoming the basis of the linear stochastic theory of time series [1, 3, 10].

Eq. 1 has a unique solution if the polynomial

$$\Phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p \tag{2}$$

has no roots z with $|z| = 1$. If additionally $\Phi(z) \neq 1$ for all $|z| > 1$, then the process is causal, i.e. the random variable X_n can be expressed only in terms of noise values Z_k with $k \leq n$.

The biological significance of the coefficients ϕ_i was emphasized in a previous work [8]; they describe the strength of the interaction after i steps in a discrete stochastic process.

For example, a value of $\phi = 0.9$ for AR1 (the first order autoregressive model) tells that the interaction between subsequent terms is quite strong compared to a $\phi = 0.2$ case. We can assimilate a strong *interaction* to a strong *resemblance.* On the contrary, a significant difference in the length of neighboring CDS would mean a weak *resemblance* among the terms of series. Therefore, in the case of our biological problem the meaning of parameter ϕ is the *resemblance* of the CDS lengths.

If the investigated process is autoregressive of first order, it gives only the *resemblance* between successive terms. *Resemblance* may also occur after more than one step, for example between the term n and the term $n–i$ from the series. Higher order autoregressive models must be applied when interaction occurs at various distances and intensities among several terms located in the neighborhood of X_n.

A plausible explanation for the validity of the AR model is the operon [4] structure of the bacterial genomes. It is known that functionally related genes are often organized into operons. If functionally related genes have similar sizes, then we would also expect that CDS length would show local correlations. The existence of operons in which the CDs have comparable lengths can be easily noticed by AR*p* analysis.

WORKING STEPS

The working steps are summarised as follows:

i) Data containing the succession of CDS lengths for *Bacillus subtilis* and *Escherichia coli Sakai* are first subjected to the detrending procedure as described in [12]. A simple and efficient detrending method consists in approximate the non-stationary features with a high order polynomial. In this previous work it was shown that in the genome analysis the trend is well approximated by a 10^{th} degree polynomial. This step is mandatory in the autoregressive analysis because the AR*p* models work properly only on stationary series.

ii)Series are subjected to discrete Fourier transform (FFT) to obtain the power spectrum. The obtained periodograms give us preliminary information about the intrinsic correlation which exists in the genome.

iii) As such processes can easily be confused and/or approximated by power-laws, the most important step in disclosing the nature of fluctuations is to average out their spectra.

iv) The averaged periodograms are fitted with several high order autoregressive models ARp. In the present paper we report the results obtained by using 5 ARp models with p = 1, 2, 3, 5 and 9. By increasing the order of the ARp model the fit becomes more accurate. From the practical point of view, it is recommendable to limit the value of p to the smallest integer which assures a relative error smaller than 0.2 for the returned values of the fit parameters ϕ_i.

RESULTS

The series containing the CDS length for the genome of *Bacillus subtilis* and *Escherichia coli Sakai* are shown in Figure 1. There are 4106 coding terms in the genome of *Bacillus subtilis* and 5365 in the genome of *Escherichia coli Sakai*. A few long CDS are seen dispersed throughout the two genomes. Analyzing the two genomes one observes that, most often, the *Bacillus subtilis* CDS are shorter as compared to *Escherichia coli Sakai*.

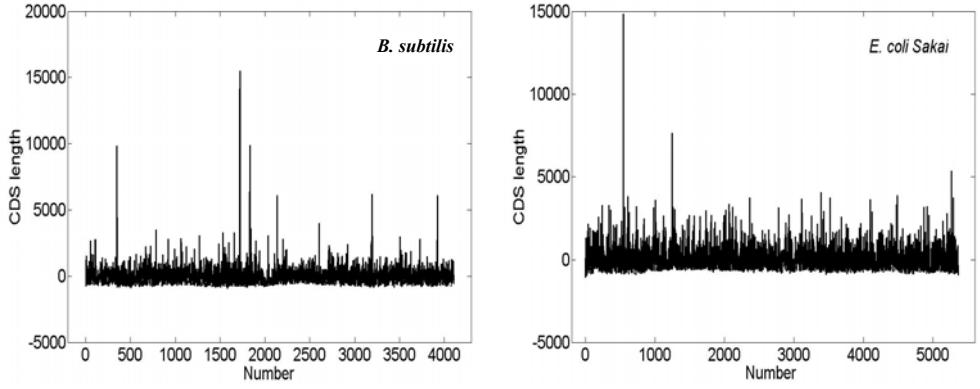


Fig. 1. Series of CDS lengths for the genome of *Bacillus subtilis* (*B. subtilis*) and *Escherichia coli Sakai* (*E. coli Sakai*). There are 4106 coding sequences lengths for *Bacillus subtilis* genome and 6341 in *Escherichia coli Sakai* genome as indicated by the number on the x-axis. The unit for CDS length is base pairs (bp). By number we mean index of a given gene in the genome structure.

The periodograms obtained by applying the discrete Fourier transform are relatively noisy and the information is embedded in the noise (Fig. 2).

Averaging the periodograms by using a mobile 21 terms window the relevant structure is evidenced for both genomes (Fig. 3). Even if the high frequency domain of both spectra is quite noisy, the fitted curves follow the main feature better than for the low frequency range.

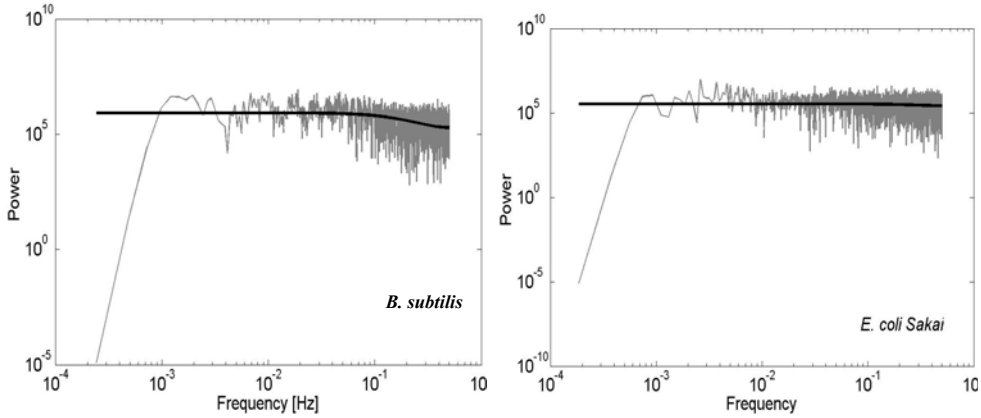
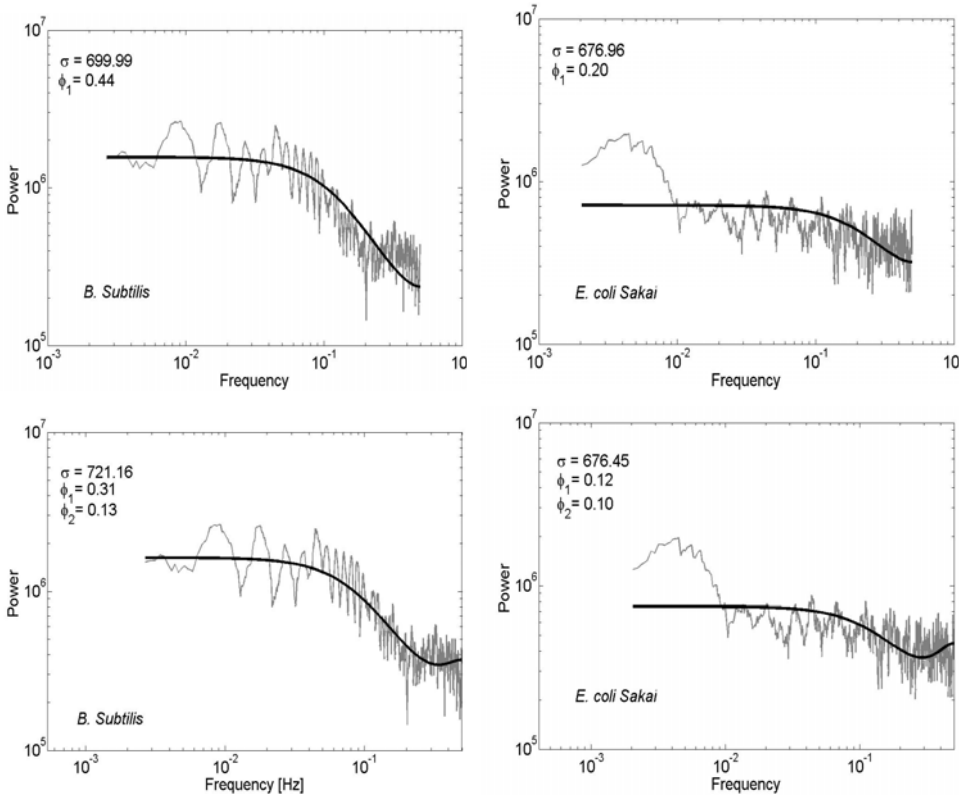


Fig. 2. FFT power spectrum represented in log-log coordinates for *Bacillus subtilis* and *Escherichia coli Sakai* respectively. The unaveraged periodograms are extremely noisy and at first glance they are well fitted by a linear function.



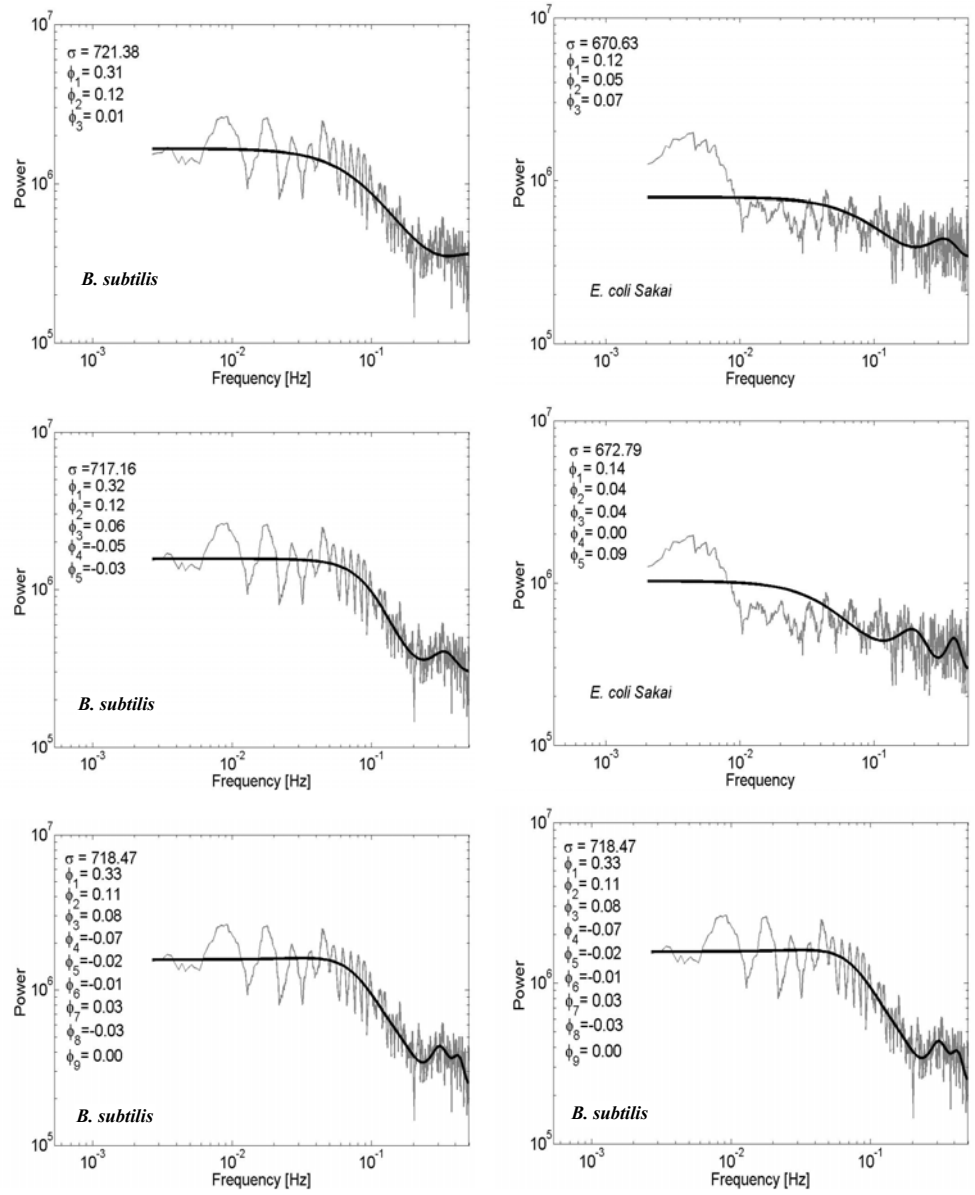


Fig. 3. The power spectra are fitted with several AR*p* models (grey line). Both power spectra are averaged by 21 terms in order to disclose the spectrum shape and are fitted relatively well with AR*p* models having $p \geq 3$.

The genomes analysis was performed using different autoregressive models. The parameters ϕ_i obtained from this analysis are shown in Table 1.

Table 1

The interaction parameters ϕ_i obtained by fitting the series of CDS length with several autoregressive models. The parameters ϕ_i describe quantitatively the resemblance between CDS lengths situated at distance *i* in the genome

	Bacillus subtilis	***E. coli Sakai***
AR1	ϕ_1= 0.440	ϕ_1= 0.200
AR2	ϕ_1= 0.310 ϕ_2= 0.125	ϕ_1= 0.160 ϕ_2= 0.123
AR3	ϕ_1= 0.310 ϕ_2= 0.120 ϕ_3= 0.010	ϕ_1= 0.120 ϕ_2= 0.050 ϕ_3= 0.045
AR5	ϕ_1= 0.310 ϕ_2= 0.120 ϕ_3= 0.060 ϕ_4=–0.050 ϕ_5=–0.030	ϕ_1= 0.140 ϕ_2= 0.046 ϕ_3= 0.044 ϕ_4= 0.006 ϕ_5= 0.010
AR9	ϕ_1= 0.330 ϕ_2= 0.110 ϕ_3= 0.083 ϕ_4=–0.066 ϕ_5=–0.018 ϕ_6=–0.014 ϕ_7= 0.029 ϕ_8=–0.033 ϕ_9= 0.004	ϕ_1= 0.140 ϕ_2= 0.037 ϕ_3= 0.050 ϕ_4=–0.001 ϕ_5= 0.053 ϕ_6= 0.010 ϕ_7= 0.047 ϕ_8= 0.000 ϕ_9= 0.074

DISCUSSION AND CONCLUSIONS

We can remark several similarities and also several significant differences between the two genomes.

a. Similarities:

i) The simplest autoregressive model AR1 describes only the mean overall decrease of the periodograms in the high frequency domain.

ii) The interaction coefficient returned by AR1 is positive for both species revealing the nonrandom succession of CDS length.

iii) The periodogram is described in a more detailed manner if one uses higher order AR*p*, the fitting procedure offering richer information about the investigated genomes.

iv) The values of the first 3 interaction parameters $\phi_1 - \phi_3$ are positive and decrease with the increase of the neighborhood order *i*. The coefficients corresponding to higher orders are in general small their significance being questionable. A value closer to zero means almost no resemblance between CDS lengths. The resemblance increases with the interaction factor value increment.

b. Differences:

v) The interaction parameter ϕ_1 returned by AR1 is more than double for *Bacillus subtilis* as compared to *Escherichia coli Sakai*.

vi) In the case of *Bacillus subtilis*, if the successive coefficients ϕ_i returned by different AR*p* models are summed together one obtains approximately the value of ϕ_1 from the AR1 model. A similar algorithm applied to *Escherichia coli Sakai* does not validate this feature.

Summarizing the results of the paper one can say that the succession of coding sequences from the bacterial chromosome proved to be short-range correlated. The high order autoregressive AR*p* models are suitable to describe the bacterial genomes by fitting relatively well the periodograms corresponding to CDS length series. The AR3 model appears to be accurate enough to describe the main features of the bacterial genomes, the additional information brought by higher order models consisting in relatively irrelevant details.

Comparing the two bacteria, we have found that the used theoretical models are able to reveal significant differences related to the local, short-range organization of two genomes. Accordingly, we are confident that the autoregressive model is a powerful theoretical tool able to reveal subtle correlation properties in the genome.

Acknowledgements. The work was supported by the National University Research Council (CNCSIS), PN-II-ID-PCE-2007-1, Project IDEI 32/2007. The authors are indebted to Vasile V. Morariu and Călin Vamos for fruitful discussions.

REFERENCES

1. BROCKWEL, P.J., R.A. DAVIES, *Time Series: Theory and Methods*, 2nd ed., New York: Springer, 1991.
2. 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.
3. HAMILTON, J.D., *Time Series Analysis*, Princeton University Press, 1994.
4. JACOB, F., D. PERRIN, C. SANCHEZ, J. MONOD, Operon: a group of genes with the expression coordinated by an operator, *C.R. Hebd. Seances Acad. Sci*., 1960, **250**, 1727–1729.
5. LI, D.J., ZHANG, S., The C-value enigma and timing of the Cambrian exposion, *arXiv Preprint Archive*, http://arxiv org/abs/0806.0108, 2008.
6. LI, D.J., S. ZHANG, Prediction of genomic properties and classification of life by protein length distributions, *arXiv Preprint Archive*, http://arxiv org/abs/ 0806.0205, 2008.
7. LI, D.J., S. ZHANG, Classification of life by the mechanism of genome size evolution, *arXiv Preprint Archive*, http://arXiv org/abs/0811.3164, 2008.
8. MORARIU, V., L. BUIMAGA-IARINCA, Autoregressive modeling of coding sequence lengths in bacterial genome, *Fluct. And Noise Letters*, 2010, **9(1)**, 47–59.

9. PENG, C.-K., S.V. BULDYREV, S. HAVLIN, M. SIMONS, H.E. STANLEY, A.L. GOLDBERGER, Mosaic organization of DNA nucleotides, *Phys. Rev. E,* 1994, **49**, 1685–1689.
10. SHIRYAEV, A.N., *Probability*, 2nd ed., Springer, 1996, pp. 405, 409.
11. VAMOS, C., S.M. SOLTUZ, M. CRACIUN, Order 1 autoregressive process of finite length, *Anal. Numer. Theor*., 2007, **2**, http://lanl.arxiv.org/abs/0709.2963.
12. VAMOS, C., Automatic algorithm for monotone trend removal, *Phys. Rev. E*, 2007, **75**, http://link.aps.org/doi/10.1103/PhysRevE.75.036705.
13. ZAINEA, O., V.V. MORARIU, The length of coding sequences in a bacterial genome: evidence for short-range correlation, *Fluct. Noise Lett*., 2007, **7**, 501–508.
14. http://www.ebi.ac.uk/Databases/ accessed on 01.06.2010.