Fluctuation and Noise Letters
Vol. 0, No. 0 (0000) 000–000

AUTOREGRESSIVE MODELING OF CODING SEQUENCE LENGTHS IN BACTERIAL GENOME

VASILE V. MORARIU [1, 2], LUIZA BUIMAGA-IARINCA [1]

[1]National Institute of Research and Development for Isotopic and Molecular Technologies, Department of Molecular and Biomolecular Physics, 400293, Cluj-Napoca, Romania, vvm@itim-cj.ro

[2] Romanian Academy of Scientists, 54, Splaiul Independentei, Sector 5, 050094 Bucharest, Romania

Received (received date)
Revised (revised date)
Accepted (accepted date)

Previous investigation of the length of coding sequence lengths (CDS) in the bacterial circular chromosome revealed short range correlation in the series of these data. We have further analyzed the averaged periodograms of these series and we found that the organization of CDS can be well described by first order autoregressive processes. This involves interaction between the neighboring terms. The autoregressive analysis may have great potential in modeling various physical and biological processes like light emission of galaxies, protein organization, cell flickering, cognitive processes and perhaps others.

Keywords: Bacterial genome; coding sequence lengths; autoregressive modeling .

1. Introduction

Bacterial chromosome has a relatively simple structure consisting of a succession of coding (CDS) and non-coding sequences. Each sequence consists of a specific series of types and number of bases. The coding sequences represent genes and they dominate the total content of bases in the bacterial chromosome. Recent work revealed the important significance of the length and distribution of proteins (which is similar to the coding sequences) [1-3]. The earlier studies assumed that the organization of the genome is random i.e. the succession of the CDS lengths consist of uncorrelated data. However later investigations suggest that some order exists in the CDS length series [4-7].

The Detrended Fluctuation Analysis (DFA) of CDS length series suggests that bacterial genomes are short range correlated [7]. However when the same series are subjected to Fourier Transform they produce a spectrum which appears to be linear in double log plots. This apparent scale invariance means that the series are long range correlated data which contradicts DFA results. One aim of this study was to solve this apparent contradiction. The solution of the problem lays mainly in Mandelbrot's general suggestion that a spectrum should be averaged out first in order to uncover its meaning [8-9]. Indeed the averaging procedure of the spectrum revealed its nonlinear characteristic. Further we report here that such spectra are described by autoregressive models which represent short range correlation.

2. Bacterial genome data and strategy of analysis

Each species of bacteria or archaea is characterized by a number of *n* coding sequences each having a specific length l_k where $k=1...n$. The length is expressed in number of bases. Each DNA molecule is known to consist of two strands *plus* and *minus* or *leading* and *lagging strands.* The following series of CDS lengths are considered in our analysis: A) The full series consisting of the natural sequence of lengths as they succeed in the genome, B) The series of CDS lengths in the strands *plus* and *minus* respectively. Therefore the subject of our analysis are the series *l(+/-), l(+)* and *l(-).* The data were extracted from the web site of European Molecular Biology Laboratory (EMBL) by using simple programs. An example of series for a bacteria species is illustrated in fig.1.

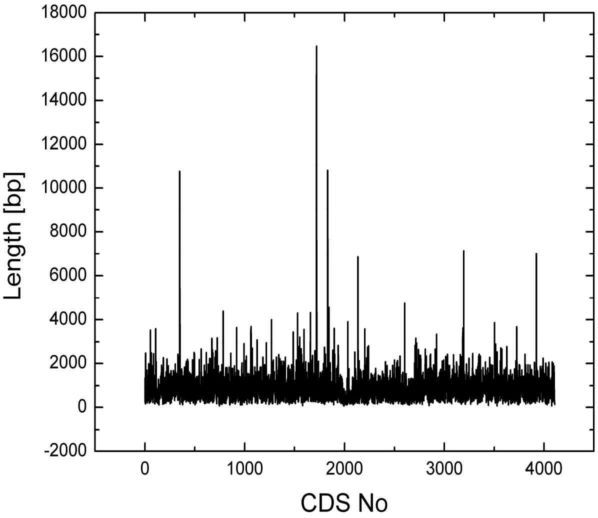

Figure 1 Coding sequence length series for *Bacillus* subtilis. The series refer to the full genome series denoted as l(+/-).

These series were analyzed by DFA and by an autoregressive model. The analysis resulted in two parameters: alpha – the correlation exponent and phi – the interaction factor respectively (see below).

It was considered of interest to explore comparatively the genome organization of some bacteria which are often used in explorative studies. We have also considered several other species of bacteria in order to see if differences exist among them. Some of bacteria exist as different cell strains so that various cell strains of the same species were analyzed. Although archaea forms a distinct domain they have a similar structure compared to the domain of bacteria so we included in our comparative analysis an archaea species. The full list of analyzed species is as follows: *Archaeoglobus fulgidus, Bacillus subtilis, Bacillus cereus ATCC14, Bacillus halonduras, Escherichia coli O157H7 EDL933, E.coli*

O157:H7 str. Sakai, E.coli str.K12 substr.MG1655 W3110, E.coli UTI 89, E.coli str. K-12 W311, E.coli APEC O1, Haemophilus influenzae 86, H. influenzae ATC, H. influenzae PittEE, H.influenzae PittGG, Helicobacter pylori 26695, Helicobacter pylori HPAG1, Helicobacter pylori J99.

First we present the result of the DFA analysis which proves that the series are short range correlated. Then we calculate the periodograms of the series and analyze them by an autoregressive model. Finally we compare the outcome of these types of analyses. While DFA simply shows that there is a correlation order in the bacterial genomes, the use of the autoregressive model is to show how the order is achieved in the genomes. This reveals that the length of CDS terms interact and the strength of their interaction. The succession of operations can be summarized as following: *i*) Detrend the series of data by subtracting various degrees of polynomial fits; *ii*) Discrete Fourier transform of the series; *iii*) Periodogram averaging using 1-21 terms; *iv*) Fit the spectrum to an AR(1) model. The resulting parameters are the interaction factor φ and the dispersion σ. Their values depend on the degree of the polynomial fit used for the detrending procedure, and on the number of spectral terms used for averaging procedure. If the data can be described by an AR(1) model, choose the final values of φ and σ, by analyzing the plot of φ and σ against the polynomial degree and the number of averaging terms. This procedure was applied to *l(+-), l(+)* and *l(-)* series

3. Methods of analysis

3.1. *Detrended fluctuation analysis of bacterial genomes*

Our initial method of investigation was the well known DFA. The DFA method was originally developed to investigate long-range correlation in non-stationary series [9]. First DFA integrates the series which is further divided into *n* boxes of equal length. A least square line is fit to the data in each box *n* which represents the trend in that box. The integrated time series *y(k)* is detrended by subtracting the local trend $y_n(k)$ in each box. Then the root-mean square of the resulting series is calculated as a fluctuation function:

$$F(n) = \sqrt{\frac{1}{N}\sum_{1}^{N}[y(k) - y_n(k)]^2}$$

Here *N* is the number of terms. *F(n)* typically increases with box size *n* and a linear relationship on a double log graph indicates the presence of scaling:

$$F(n) \propto n^{\alpha}$$

The value of α exponent is 0.5 for uncorrelated data, then values ranging between 0.5 < α < 1 indicates a long-range correlation while values between 0<α<0.5 represents anti-correlation. The key step in a long-range correlation decision is that DFA plot should be linear over the whole range of box sizes covering the series. Our earlier investigation was prompted by the observation that the DFA plot of CDS length series was generally non-linear over the whole chromosome series [7]. Strictly speaking a DFA plot should be linear if long-range correlation would extend over the whole chromosome of over, at least, two orders of magnitude of the box size *n*. Nonlinearity of the DFA plot can be a simple and direct evidence for short-range correlation as recently shown by a DFA investigation of autoregressive process [10]. An example of a DFA plot for a *l(+)* series is illustrated in figure 2. The plot is clearly nonlinear.

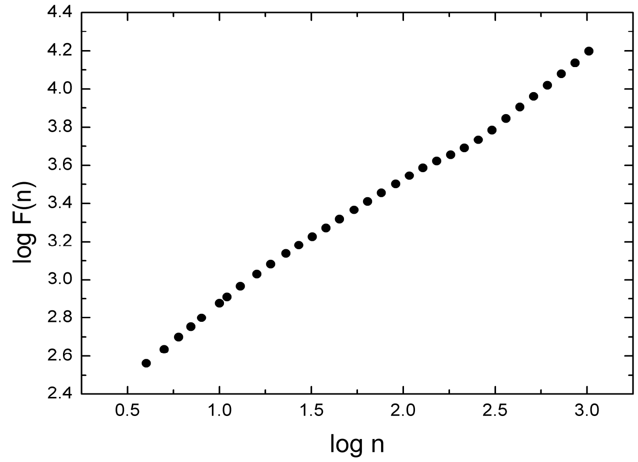

Figure 2 Detrended fluctuation analysis for the plus strand of Bacillus subtilis CDS length series.

A further simple way to check for short-range organization of the series is to apply a segmentation procedure of the CDS series and analyze the correlation in these segmented series by DFA. The correlation exponents should be unchanged if long-range correlation was operative. Short range correlation would result in different correlation exponents for different segments [10].

3.2. *Autoregressive modeling of the correlation spectra*

There is a key problem which recommends the analysis of the data by correlation spectroscopy. A simple Fourier transform results in noisy spectrum which can be fitted by a straight line. This is equivalent to scale invariance or a fractal like structure. However averaging the spectrum (as strongly recommended by Mandelbrot) results in non linear spectrum which contradicts the fractal like interpretation. This is illustrated in figure 4 for a bacteria species. Such a non linear spectrum is characteristic for autoregressive processes which can be easily described in an analytical way as shown below.

A discrete stochastic process $\{X_n, n=0, \pm1, \pm2,\ldots\}$ is called autoregressive process of order *p*, denoted AR(*p*), if $\{X_n\}$ is stationary and for any *n*:

$$X_n - \varphi_1 X_{n-1} - \ldots - \varphi_p X_{n-p} = Z_n \tag{1}$$

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 a stochastic process term on the next term after *i* time steps. The properties of AR(*p*) processes have been studied in detail and they are the basis of the linear stochastic theory of time series [11] and [12]. Equation 1 has a unique solution if the polynomial $\Phi(z)=1-\varphi_1 z-\ldots-\varphi_p z^p$ has no roots *z* with $|z| =1$. If in addition $\Phi(z) \neq 1$ 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 at previous moments and at the same moment.

The spectral density of an AR(p) process is:

$$f(\upsilon)=\frac{\sigma^2}{2\pi}\frac{1}{\left|\Phi(e^{-2\pi i\upsilon)}\right|^2},-0.5<\upsilon\leq 0.5\,, \tag{2}$$

where v is the frequency. For an AR(1) process, the spectral density in equation 2 becomes:

$$f(\upsilon)=\frac{\sigma^2}{2\pi}\frac{1}{1+\varphi^2-2\varphi\cos 2\pi\upsilon},-0.5<\upsilon\leq 0.5\,, \tag{3}$$

where φ is the only parameter φ_i in this case. The above mentioned formulas are valid for ideal stochastic processes of finite length.

The time series found in practice have a finite length and usually they are considered realizations of a finite sample of an AR(1) process of infinite length. Therefore, the changes of the equations 2 and 3 have to be analyzed for a sample with finite length $\{X_n, n=1, 2, ..., N\}$ extracted from an infinite process $\{X_n, n=0, \pm1, \pm2, ...\}$. A detailed analysis of the power spectrum of the AR(1) process and the influence of the finite length is contained in [13]. In this paper some of the main conclusions are discussed.

The sample estimator of the spectral density is the periodogram:

$$I_N(\upsilon)=\left|A_N(\upsilon)\right|^2\ , \tag{4}$$

where $A_N\,(v)$ is the discrete Fourier transform of the sample:

$$A_N(\upsilon)=\frac{1}{\sqrt{N}}\sum\nolimits_{n-1}^{N}X_n e^{2\pi i n\upsilon}\ , \tag{5}$$

Since the sample contains a finite number of components, there are only N independent values of $A_N(v)$ and $I_N(v)$. Usually, these values are computed for the Fourier frequencies $v_j = j/N$, where j is a integer satisfying the condition $-0.5 < v_j \leq 0.5$. The periodogram of an AR(p) process is an unbiased estimator of the spectral density:

$$\lim_{N\to\infty}\left\langle I_N(\upsilon_j)\right\rangle=2\pi f(\upsilon)\,, \tag{6}$$

where $(v_j -0.5N) < v \leq (v_j+0.5N)$ [13]. Hence, increasing the sample length N, while the time step t is kept constant, the average periodogram becomes a better approximation of the spectral density (equation 6). However, a single periodogram is not a consistent estimator, because it does not converge in probability to the spectral density, i.e. the standard deviation of $I_N\,(v_j)$ does not converge to zero, and two distinct values of the periodogram are uncorrelated, no matter how close the frequencies are when they are computed.

Usually, the spectral density and the periodogram are plotted on a log-log scale. The logarithmic coordinates strongly distort the shape of the graphic because by applying the logarithm, any neighborhood of the origin is transformed into an infinite length interval and the value of $f(0)$ cannot be plotted. For a sample with N terms, the first value of the spectral density is obtained for the minimum frequency v_{min} =1/N. Figure 2a shows the spectral density in equation 3 for N = 1024, σ =1 and different values of the parameter φ. One can observe that the AR(1) processes for higher φ values can be easily approximated by a fractal-like behavior. For φ =0.90 and especially for φ =0.99, a significant part of the power spectrum is almost linear with a slope equal to −2, which corresponds to β =2. A significant part of the spectrum could be regarded as linear for smaller value of φ (for example φ=0.5 in figure 3).

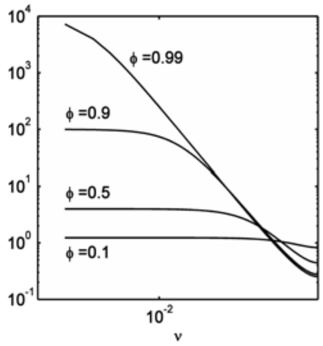

Figure 3 The spectral density of an AR(1) process for N = 1024, σ =1 and different values of the interaction factor among close terms φ.

For small frequencies, the AR(1) spectral density is strongly stretched in log-log coordinates such that a plateau appears (figure 2a) with a value given by:

$$f(0) = \frac{\sigma^2}{2\pi(1-\varphi)^2} \tag{8}$$

From equation 3 it follows that the plateau corresponds to the small values of v, when the variable term at the denominator can be neglected in comparison with the constant term. Using the quadratic approximation of the cosine function, the condition that the graph of the AR(1) power spectrum has a plateau $v < (1-\varphi)/2\pi\sqrt{\varphi}$ is obtained. If φ tends to 1, the plateau appears at smaller values of the frequency. Therefore, if N is large enough, the periodogram of an AR(1) sample has a plateau at small frequencies (if N is large, then $v_{min} \rightarrow 0$).

A time series $\{x_n, n=1, 2, ..., N\}$ as a realization of a sample $\{X_n, n=1, 2, ..., N\}$ from an AR(1) process is considered. Applying the discrete Fourier transform (equation 4) to the time series $\{x_n\}$, and then computing the periodogram (equation 5), values randomly distributed around the spectral density (equation 3) of the AR(1) are obtained. Since the periodogram is not a consistent estimator, by increasing the length N of the sample, the periodogram fluctuations around the theoretical spectral density are not reduced. Consistent estimation of the spectral density may be obtained using averaging of the periodogram on intervals with length of magnitude order of $\sqrt{N}$ [13]. Choosing the optimum weight function is a difficult task, because, if the periodogram is smoothed too much, then the bias with respect to the theoretical spectrum can become large. From various weight functions [16] the simplest one is used, i.e. the averaging with equal weights on symmetric intervals containing M Fourier frequencies, with $M = 1, 3, 5, ..., 21$. Then, the averaged periodogram contains $N - M + 1$ values, because for the first and last $(M - 1)/2$ values of the periodogram, the symmetric averaging cannot be performed.

Let us consider that an AR(1) model for an averaged periodogram is to be found, i.e. to find the values of the parameters φ and σ. The minimum of the quadratic norm of the difference between the averaged periodogram and the theoretical spectral density of the AR(1) model has to be determined. The sample standard deviation of the time series and $\varphi = 0$ are used as initial values for the optimization algorithm.All calculations were performed by using MATLAB [15].

Many series of data have non-stationary characteristics, so the application of Fourier transform to the data results in misleading spectra. A common procedure to avoid this complication is to use detrended fluctuation analysis[9]. This results in a correlation exponent free of the correlation introduced by the trend. However, in our case it is essential to obtain the corresponding spectrum, as the shape of the spectrum gives the relevant information (if either a power-law or a non-power law is operative). Consequently, an important preliminary step is to remove non-stationary characteristics in the series. We performed detrending by subtracting a polynomial fit from the original series. The problem is to determine the right polynomial fit. We performed 1 to 20 degree polynomial fits and generally found that a polynomial degree around 10 gives the most reliable result for φ and σ. The values of φ and σ also depended on the averaging procedure of the spectra so that optimizing the values of φ and σ involved optimizing both the detrending and the averaging procedures. The above mentioned polynomial fitting was chosen as its accuracy is comparable to the moving average method and to an automatic method for the estimation of a monotone trend. The same work also showed that polynomial fitting for a $1/f$ noise proved to have the best performance [14].

4. Results

Examples of non-averaged and averaged periodogram for the CDS length series *for Bacillus subtilis* species are illustrated in figure 4. The non-averaged periodogram (figure 4a) can be well fitted by a straight line which can mistakenly interpreted as a long range correlation. However the real shape of the periodogram is revealed after averaging. (figure 4 b). The averaged periodogram is well fitted by an AR(1) process.

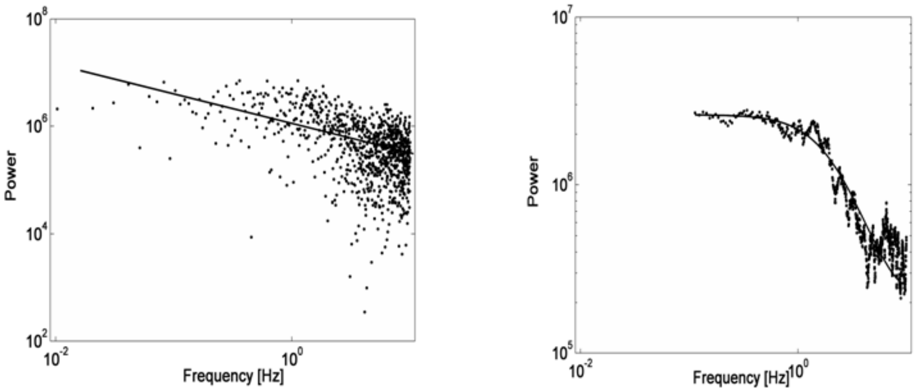

Figure 4 Periodogram for the length series of coding sequences of the plus (or leading strand) of *Bacillus subtilis*. a) Non-averaged periodogram; b) Averaged periodogram and fitting with a first order autoregressive process.

The corresponding AR(1) parameters φ as well as the correlation exponents resulting from DFA analysis are included in table 1 and 2 for a range o bacteria an archaea species. The parameter φ may vary between $0<\varphi<1$. In the case of bacteria and archaea presented in table 1 this parameter varies between 0.52 for *Bacillus subtilis* and almost 0 for some strands of *Haemophilus influenzae* and *Helicobacter pilori*. This means that the strenght of interaction between the the length of CDS terms may vary widely from a medium value down to almost no interaction. Further this interaction is sensitive to the *leading* or *lagging* strands and also to the species strain. On the other hand the correlation exponent (table 2) generally has a rather low value indicating a weak correlation among the terms of the series.

Higher order autoregressive processes where also tested. This involved additional interaction between terms at higher distances in the series. However it was concluded that the AR(1) modeling of the series proved to be quite satisfactory.

Table 1 The interaction strength φ of the first order autoregressive modeling of CDS lengths periodogram.

Species	+ strand	- strand	both strands
Archaeglobus Fulgidus	$\varphi = 0.01$	$\varphi = 0.16$	$\varphi = 0.14$
Bacilus subtilis	$\varphi = 0.52$	$\varphi = 0.43$	$\varphi = 0.44$
B Cereus ATCC14	$\varphi = 0.19$	$\varphi = 0.05$	$\varphi = 0.13$
B Halonduras	$\varphi = 0.11$	$\varphi = 0.10$	$\varphi = 0.16$
E. Coli O157H7	$\varphi = 0.15$	$\varphi = 0.21$	$\varphi = 0.24$
E. Coli Sakai	$\varphi = 0.21$	$\varphi = 0.26$	$\varphi = 0.20$
E. Coli K12	$\varphi = 0.12$	$\varphi = 0.20$	$\varphi = 0.18$
E. Coli UTI	$\varphi = 0.14$	$\varphi = 0.21$	$\varphi = 0.24$
E Coli W311	$\varphi = 0.17$	$\varphi = 0.14$	$\varphi = 0.18$
E Coli APEC	$\varphi = 0.18$	$\varphi = 0.15$	$\varphi = 0.19$
H Influ 86	$\varphi = 0.07$	$\varphi = 0.13$	$\varphi = 0.13$
H Influ ATC	$\varphi = 0.11$	$\varphi = 0.07$	$\varphi = 0.12$
H Influ Pitt EE	$\varphi = 0.11$	$\varphi = 0.05$	$\varphi = 0.11$
H Influ. Pitt GG	$\varphi = 0.13$	$\varphi = 0.00$	$\varphi = 0.10$
H Pylori 266	$\varphi = 0.05$	$\varphi = 0.00$	$\varphi = 0.02$
H Pylori HP	$\varphi = 0.05$	$\varphi = 0.02$	$\varphi = 0.03$
H Pylori J99	$\varphi = 0.04$	$\varphi = 0.03$	$\varphi = 0.04$

Table 2 The correlation exponent for different species of bacteria and archaea resulting from the detrended fluctuation analysis

Species	+ strand	- strand	both strands
Archaeglobus Fulgidus	$\alpha= 0.50$	$\alpha= 0.60$	$\alpha=0.57$
Bacilus subtilis	$\alpha= 0.84$	$\alpha= 0.67$	$\alpha= 0.79$
B Cereus ATCC14	$\alpha= 0.60$	$\alpha= 0.50$	$\alpha= 0.55$
B Halonduras	$\alpha= 0.56$	$\alpha= 0.55$	$\alpha= 0.56$
E. Coli O157H7	$\alpha= 0.56$	$\alpha= 0.60$	$\alpha= 0.60$
E. Coli Sakai	$\alpha= 0.56$	$\alpha= 0.62$	$\alpha= 0.56$
E. Coli K12	$\alpha= 0.55$	$\alpha= 0.57$	$\alpha= 0.58$
E. Coli UTI	$\alpha= 0.56$	$\alpha= 0.57$	$\alpha= 0.60$
E Coli W311	$\alpha= 0.59$	$\alpha= 0.54$	$\alpha= 0.57$
E Coli APEC	$\alpha= 0.56$	$\alpha= 0.55$	$\alpha= 0.58$
H Influ 86	$\alpha= 0.53$	$\alpha= 0.58$	$\alpha= 0.55$
H Influ ATC	$\alpha= 0.53$	$\alpha= 0.55$	$\alpha= 0.52$
H Influ Pitt EE	$\alpha= 0.55$	$\alpha= 0.55$	$\alpha= 0.54$
H Influ. Pitt GG	$\alpha= 0.62$	$\alpha= 0.50$	$\alpha= 0.54$
H Pylori 266	$\alpha= 0.50$	$\alpha= 0.50$	$\alpha= 0.51$
H Pylori HP	$\alpha= 0.54$	$\alpha= 0.49$	$\alpha= 0.51$
H Pylori J99	$\alpha= 0.53$	$\alpha= 0.52$	$\alpha= 0.54$

It is obvious that a relationship should exist among the correlation exponents α and the interaction factor φ for different species. This is illustrated in figure 5 for the three kind of series.It can be seen that the relationship is linear for *l(+)* and *l(-)* series while for the *l(+-)* series is non linear (figure 5). Non averaged periodograms give very similar results (not shown). This might suggests that the non-stationary contribution to the series are not important for the AR (1) modeling of the series.

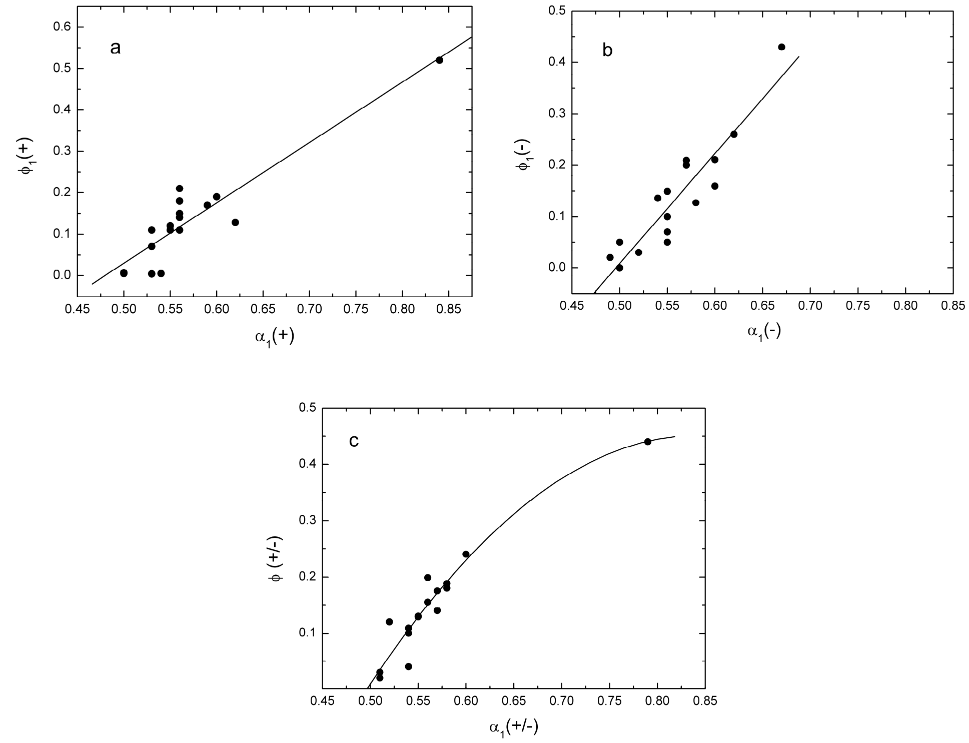

Figure 5 The relationship between the correlation exponent *alpha 1* (resulting from DFA analysis) and the strength of interaction *phi 1* (first order autoregressive modeling). a) the *strand* + data, b) *strand* - and c) *strand +/-* of *Bacillus subtilis.*

Parameter φ of the autoregressive fitting seems to be more sensitive compared to the correlation exponent. The impression at this stage is that first order autoregressive fitting of the CDS series is the best way to describe the order in these series. We further attempted to describe the order in the series of very different kind of phenomena which presented similar features of the spectrum. Such examples refer to length of protein series, biological cell flickering or cognitive phenomena. [9]. Therefore autoregressive processes proved to be a useful model for describing the order in these phenomena.

5. The biological significance of the AR model

A key parameter in the autoregressive model is the *interaction* factor φ. This characterizes the strength of interaction of a stochastic process term on the next term after *i* time steps. For example a value of φ=0.9 for a first order AR model tells that the interaction between succeeding terms is quite strong compared to say a φ=0.2 case. Suppose the terms of the series are x_i and x_{i+1}. If an interaction factor of φ=0.9 applies than the terms of the series will be x_i and $(x_i+x_{i+1})0.9$ as compared to x_i and $(x_i+x_{i+1})0.2$. A stronger interaction makes the second term to be higher than for a weaker interaction. Now we transpose the general physical description of the autoregressive model to our biological problem of coding sequences length. The terms of the series are in this case the length of the coding sequences. Suppose the terms of the series have similar or quite close values, or in other words the *resemblance* is strong. We can assimilate a strong *resemblance* to a strong *interaction.* At contrary, a significant difference in the length of 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. As the model is a first order autoregressive process, this means that it is a matter of *resemblance* between successive terms. *Resemblance* may also occur after more than one step, for example between the first and the third term of the series. Higher order autoregressive models may apply when interaction occurs between more that two terms at various distances. *Resemblance* may vary among different genomes as shown by the results in table 1.

A plausible explanation for the validity of the AR model is the operon 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 with similar CDS lengths can be easily noticed. However there is a diversity of operons as far as their content is concerned. Examples of various operons are listed in table 3.

Table 3 Examples of operons of *Eschericia coli* and the length of their enzymes

Operon Enzymes	**Length [bp]**	**Operon Enzymes**	**Length [bp]**	**Operon Enzymes**	**Length [bp]**
cys		ton		leu	
P	1017	E	1563	A	1572
U	834	D	1596	B	1092
W	876	C	1359	C	1401
A	1098	B	1194	D	606
M	912	A	807		

The *cis* operon shows a strong *resemblance* for the U and W enzymes and A and M enzymes respectively while P and U have a milder *resemblance*. There is a *strong resemblance* for the E and D enzymes of the *ton* operon while *resemblance* decreases down the column. The enzymes of *leu* operons have a weak *resemblance*. The overall *resemblance* for all successive terms is given by the value of *φ* in table 1. The striking case in table 1 is *Bacillus subtilis* which has the highest value of φ. This means that more similar CDS lengths exist in this genome compared to the rest of genomes. On the other hand there are clear differences between the plus and minus strands of the same genome. This must have a strong biological relevance as it points out to the well known differential distribution of genes in the two strands [16] For example the first four examples of E.coli in table 1 reveals that the minus strand is characterized by higher values of φ. That is *resemblance* of genes is higher in the minus strand compared to plus strand and this rule applies for all these four cases of *E.coli*. Higher values of φ could involve a more uniform and presumably more functionally related genes. This applies for the last two cases of E.coli in table 1, as well as for *Bacillus* species in the same table. At contrary lower values of φ could be associated to a larger diversity of genes. The frequency of leading strand genes is ~75 % in *B. subtilis* [17] but only ~55 % in *E. coli* [18]. A first systematic survey of gene strand bias showed that genomes could have from 55 to 80 % of genes in the leading strand, although systematically more than 85 % of ribosomal proteins were coded in the leading strand of these genomes [16]]. In fact, 78 % of the genes of Firmicutes (including mycoplasmas) are in the leading strand, compared with 58 % for the other genomes [19] This might be tentatively related to the generally higher values of φ for the + strand. It remains to establish the validity of this correlation for other genomes as well.

Despite the relationship between correlation and autoregressive parameters (fig 5) it is evident that the autoregressive parameter is more sensitive than the correlation parameter. The range of values for the φ parameter lay between 0.1 and 0.2 which can be regarded as low values. Other biological processes may show much higher values (such as 0.8-0.9 for flickering of red blood cells [9, 20].

We may conclude that the autoregressive modeling could be useful for comparative genomics.

Acknowledgements

The work was supported by the Romanian Agency for Scientific Research, PN-II-ID-PCE-2007-1, project IDEI 32/3007. The authors are indebted to Calin Vamos for computation work.

References

[1] D. J. Li, S. Zhang, *The C –value enigma and timing of the Cambrian exposion*, *arXiv Preprint Archive* [on line], http://arxiv org/abs/0806.0108 (2008)

[2] D. J. Li, S. Zhang, *Prediction of genomic properties and classification of life by protein length distributions*, *arXiv Preprint Archive* [on line], http://arxiv org/abs/ 0806.0205 (2008)

[3] D. J. Li, S. Zhang, *Classification of life by the mechanism of genome size evolution*, *arXiv Preprint Archive* [on line], http://arXiv org/abs/0811.3164 (2008)

[4] Y. Zhuang, F. Ma, J. Li-Ling, X. Xu, Y. Li, *Comparative analysis of amino acid usage and protein length distribution between alternatively and non-alternatively spliced genes across six eukarotic genomes*, *Mol.Biol. Evol.* **20 (12)** (2003) 1978-1985

[5] A. L. Berman, E. Kolker, E. N. Trifonov, *Underlying order in protein sequence organization*, *Proc. Natl. Acad. Sci. USA* **91** (1994) 4044-4047

[6] B. Rost, *Did evolution leap to create the protein universe?, Curr.Opin.Stru,Biol* **12** (2002) 409-416

[7] O. Zainea, V. V. Morariu, The *length of coding sequences in a bacterial genome: evidence for short-range correlation*, *Fluct,Noise Lett*. **7** (2007) 501-508

[8] B. B. Mandelbrot, *Multifractals and 1/f noise*, Springer, New York (1998)

[9] V. V. Morariu, C. Vamos, A. Pop, S. M. Soltuz, L. Buimaga-Iarinca, O. Zainea, *Autoregressive description of biological phenomena*, *arXiv Preprint Archive* [on line], http://arxiv org/abs/0808.1021 (2008)

[10]O. Zainea, V. V. Morariu, A *correlation investigation of bacterial DNA coding sequences*, *Romanian J. Biophys.* **18** (2008) 19-28

[11]P. J. Brockwel, R. A. Davies, *Time series: theory and methods*, 2nd *edn*. Springer New York (1991)

[12]J. D. Hamilton, *Time series analysis*, Princeton University Press (1994)

[13]C. Vamos, S. M. Soltuz, M. Craciun, *Order 1 autoregressive process of finite length. arXiv Preprint Archive* [on line], http://arxiv org/abs/0709.2963 (2007)

[14]C. Vamos, *Automatic algorithm for monotone trend removal*, *Phys. Rev. E.* **75** (2007) 036705

[15] MATLAB R2008a © 1994-2009 The MathWorks, Inc.

[16]E. P. C. Rocha, *The replication-related organization of bacterial genomes, Microbiology* **150** (2004), 1609-1627

[17] F. Kunst, N. Ogasawara, I. Moszer & 148 other authors. *The complete genome sequence of the Gram-positive bacterium Bacillus subtilis, Nature* (1997) **390,** 249-256

[18] F. R. Blattner, G. Plunkett, I. C. A.Bloch & 14 other authors, *The complete genome sequence of Escherichia coli K-12. Science* (1997)**2 77**, 1453–1461

[19]E. P. C Rocha, A. Danchin *Competition for scarce resources might bias bacterial genome composition, Trends Genet* (2002) **18**, 291–294

[20]L. Buimaga-Iarinca, V. Morariu, *Short range correlation of the erythrocyte membrane fluctuations*, *JPCS*, in press.