Dedicated to Prof P.T. Frangopol on the occasion of his 75th anniversary

Distribution and Correlation of the Coding Sequence Lengths in Bacterial Genomes

VASILE V. MORARIU*
National Institute for Research and Development of Isotopic and Molecular Technology,Department of Molecular and Biomolecular Physics, Donath Str., 400293, Cluj-Napoca, Romania

The length of coding sequence (CDS) series in bacterial genomes were regarded as a fluctuating system and characterized by the methods of statistical physics. The distribution and the correlation properties of CDS for 47 genomes were investigated. The distribution was found to be approximated by an exponential function while the correlation analysis revealed short range correlations.

Key words: length of coding sequences, short range correlation

The microbial genomes consist mainly of coding sequences (CDS) while non-coding sequences represent a minor proportion of the genome. Consequently the composition of coding genes, comprise most of the genome (~90%) and the bacterial genome size shows a strong positive relationship with gene number. The length of a CDS ranges between several hundreds and several thousands of base pairs. A typical species of bacteria, like *Escherichia coli* contains around $5 \cdot 10^6$ base pairs which are organized in several thousands of genes. The series of CDS lengths can be regarded as a space fluctuating series and therefore are suitable for the statistical analysis. It should be mentioned that while the correlation properties of DNA at the level of bases have been intensively studied for more than a decade, the series of coding sequence (CDS) lengths received very little attention. Although correlation characterization of CDS length series has been previously attempted on bacterial genomes, a full statistical characterization has not been reported to date [1-3]. While the older papers on the subject suggested a weak long–range correlation, the more recent papers, on contrary, brought evidence for short-range correlation and generally a non-uniform organization [3-4]. A systematic investigation of the CDS length distribution remained an open issue.

The aim of this paper was to establish the distribution and correlation characteristics of CDS length series in the genome of selected species and strains of bacteria.

Materials and Methods

A total number of 47 genomes of bacteria were investigated (table 1). The species were selected such as to cover all main divisions of bacteria as well as some of the main models used in literature (*E. coli* and *B. subtilis*). In case of *E. coli* and *Staphilococus aureus* various strains were included in order to compare the variability of the statistical properties of strains.

The extraction of the data, from the EMBL-EBI data base, was done with a program written in MATLAB. The length of coding sequences, expressed as number of base pairs, was calculated as the difference between the start and the end position of CDS in the genome. The mean value of the CDS length is also available in the proteome section of EMBL-EBI data base.

The distribution of the CDS lengths was analyzed by probability density of lengths. A probability density investigation presents the data as frequency counts *versus* bin centre while the width of the bin is either automatically or manually selected. The procedure is performed by the ORIGIN program. We found that our CDS length series however obey an exponential law. A continuous random variable X is said to have an exponential distribution if it has probability density function:

$$f_X(x,\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{for } x \le 0 \end{cases} \quad (1)$$

where $\lambda > 0$ is called the rate of the distribution. It can be seen that as λ gets larger, the process to happen tends to happen more quickly, hence we think of λ as a rate.

The correlation properties of the series were investigated by detrended fluctuation analysis (DFA). The DFA method was originally developed to investigate long-range correlation in non-stationary series [5]. First DFA integrates the series which is further divided into boxes of equal length, *n*. 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:

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

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:

$$F(n) \propto n^{\alpha} \quad (3)$$

The outcome of DFA analysis is the correlation exponent α. A single correlation exponent over at least two orders of magnitude of box sizes *n* describes long range correlation properties, while multiple correlation exponents generally describe short-range properties of the fluctuating system [6]. Their values range between $0.5<\alpha<1.5$ and $\alpha<0.5$ for anti correlated cases. When $\alpha=0.5$ the terms of the series are uncorrelated. 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. However the DFA plot of CDS length series was generally non-linear and it was approximated by two

* email: vvm@itim-cj.ro; Tel.: +40 264 584037

straight lines described by the slopes α_1 and α_2. The DFA results were checked against the shuffled series which proved the non random organization of the length series.

The results for the distribution and correlation analysis are included in table 1.

The Distribution Characteristics

As the frequency count of CDS length refers to intervals of values, many different but closely similar lengths are "melted" in a single frequency. We have investigated different types of distribution and the closest we found to match the data was an exponential like distribution. This was true for some bacterial species (fig. 1a) while for others such a distribution was only a first degree approximation (fig. 1b). We believe the reason for deviation from an exponential distribution is the fact that the CDS length series are non uniform on a long-range scale.

It can be seen that the rate constant of the distribution is double for the case shown in figure 1a compared to figure 1b. This higher rate constant seems to be associated to a precise exponential distribution. An exponential distribution occurs naturally when describing the lengths of the inter-arrival times in a homogeneous Poisson process. Exponential variables can be used to model situations where certain events occur with a constant probability per unit distance. On the other hand deviations from the exponential distributions, such as in figure 1b, clearly shows that change of CDS lengths occurs with a variable probability. Consequently the distribution cannot be regarded as uniform in such cases. When and why

Table 1
ANALYSIS OF CODING SEQUENCE LENGTHS: CORRELATION EXPONENT α_1 AND α_2, AND THE RATE RATE CONSTANT α OF THE EXPONENTIAL DISTRIBUTION FOR DIFFERENT BACTERIA SPECIES

No.	Bacteria	Alpha 1	Alpha 2	λ
1	Acidovorax avenae subsp. citrulli AAC00-1	0.564±0.001	0.564±0.001	0.051±0.001
2	Anabaena variabilis ATCC 29413	0.538±0.002	0.689±0.006	0.086±0.002
3	Azorhizobium caulinodans ORS 571	0.514±.0.003	0.659±0.007	0.100±0.005
4	Bradyrhizobium japonicum USDA 110	0.527±0.001	0.971±0.013	0.100±0.005
5	Bacteroides thetaiotaomicron VPI-5482	0.611±0.006	0.676±0.003	0.055±0.001
6	Bacillus subtilis	0.757±0.008	0.653±0.014	0.108±0.005
7	Bacillus pumilus SAFR-032	0.644±0.004	0.706±0.011	0.112±0.005
8	Bacillus halodurans C-125	0.561±0.002	0.710±0.012	0.070±0.001
9	Bacillus thuringiensis serovar konkukian str. 97-27	0.511±0.001	0.753±0.013	0.130±0.004
10	Clostridium beijerinckii NCIMB 8052	0.525±0.002	0.641±0.004	0.121±0.004
11	Delftia acidovorans SPH-1	0.555±0.004	0.633±0.002	0.099±0.003
12	Escherichia coli O157:H7 Sakai	0.558±0.002	0.757±0.009	0.117±0.002
13	Escherichia coli APEC O1	0.557±0.001	0.669±0.014	0.064±0.001
14	Escherichia coli O157:H7 EDL933	0.553±0.003	0.738±0.005	0.115±0.003
15	Escherichia coli CFT073	0.551±0.003	0.720±0.010	0.059±0.001
16	Escherichia coli K12	0.541±0.026	0.599±0.004	0.061±0.001
17	Enterococcus faecalis V583	0.542±0.002	0.599±0.005	0.111±0.003
18	Flavobacterium johnsoniae UW101	0.585±0.006	0.636±0.004	0.073±0.003
19	Frankia sp. EAN1pec	0.589±0.002	0.520±0.006	0.083±0.003
20	Hahella chejuensis KCTC 2396	0.573±0.001	0.543±0.005	0.076±0.004
21	Haemophilus influenzae (strain 86-028NP)	0.525±0.002	0.712±0.005	0.063±0.002
22	Haemophilus influenzae (strain ATCC 51907 / KW20 / Rd)	0.538±0.005	0.507±0.004	0.064±0.001
23	Haemophilus influenzae Pitt EE	0.562±0.008	0.649±0.003	0.264±0.017
24	Helicobacter pylori	0.504±0.004	0.553±0.003	0.123±0.004
25	Lactobacillus plantarum strain WCFS1	0.516±0.002	0.639±0.008	0.127±0.004
26	Lactobacillus casei ATCC 334	0.538±0.001	0.58±00.004	0.066±0.002
27	Mycoplasma penetrans HF-2 DNA	0.661±0.002	0.416±0.016	0.087±0.002
28	Mycoplasma pneumoniae M129	0.542±0.003	0.601±0.006	0.100±0.003
29	Mycobacterium smegmatis str. MC2 155	0.523±0.005	0.502±0.002	0.114±0.004

30	Microcystis aeruginosa NIES-843	0.568±0.004	0.643±0.005	0.094±0.004
31	Nocardia farcinica IFM 10152 DNA	0.560±0.003	0.471±0.008	0.158±0.016
32	Pseudomonas entomophila str. L48	0.572±0.003	0.504±0.001	0.151±0.013
33	Photorhabdus luminescens subsp. laumondii TTO1	0.584±0.002	0.632±0.007	0.112±0.011
34	Rhodococcus sp. RHA1	0.551±0.001	0.464±0.004	0.145±0.015
35	Sulfolobus solfataricus	0.583±0.003	0.481±0.003	0.077±0.002
36	Sorangium cellulosum 'So ce 56'	0.683±0.002	0.556±0.005	0.053±0.003
37	Staphylococcus aureus subsp. aureus JH1	0.524±0.001	0.639±0.007	0.264±0.017
38	Staphylococcus aureus subsp. aureus JH9	0.528±0.001	0.634±0.008	0.263±0.017
39	Staphylococcus aureus subsp. aureus Mu3	0.572±0.003	0.577±0.008	0.248±0.018
40	Staphylococcus aureus subsp. aureus Mu50	0.582±0.004	0.536±0.003	0.248±0.019
41	Staphylococcus aureus subsp. aureus MW2	0.529±0.001	0.595±0.010	0.249±0.018
42	Staphylococcus aureus subsp. aureus N315	0.578±0.004	0.577±0.005	0.261±0.017
43	Staphylococcus aureus subsp. aureus NCTC 8325	0.526±0.001	0.567±0.005	0.199±0.013
44	Staphylococcus aureus subsp. aureus str. Newman	0.569±0.003	0.589±0.004	0.257±0.017
45	Staphylococcus aureus subsp. aureus USA300 FPR3757	0.525±.0021	0.596±0.007	0.265±0.017
46	Staphylococcus aureus subsp. aureus USA300 TCH1516	0.528±0.001	0.638±0.013	0.267±0.018
47	Xanthobacter autotrophicus Py2	0.580±0.003	0.5600±.003	0.059±0.001

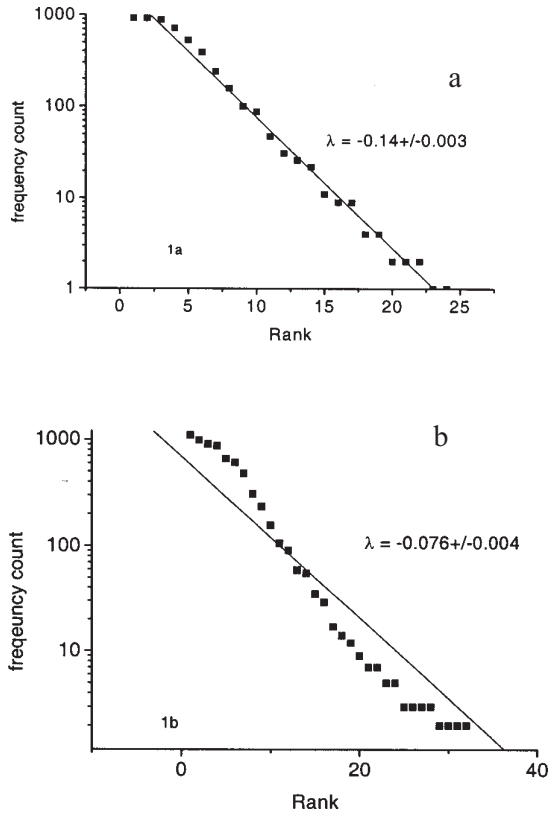

Fig. 1a. Distribution of coding sequence lengths for Baccillus thuringiensis serovar konkukian strain 97-27. The rate constant λ of the distribution is inserted in the figure 1b. The distribution for Hahella chejuensis KCTC 2396

deviations occur for some species is an interesting question which needs further investigation.

The correlation characteristics

The DFA plot for a microbial species and the corresponding shuffled series are illustrated in figure 2. It can be seen that at shorter distances (below about one order of magnitude) the correlation exponent α_1 has a value close to 0.5 yet slightly higher which indicate a residual correlation.

At higher distances among the terms of the series correlation increases to a significantly higher value α_2. The shuffled series remain close to the uncorrelated value of 0.5 as expected for random series.

It should be further mentioned that the kind of DFA plots illustrated in figure 2 was not seen for all genomes. Some of the species presented DFA plots with a downwards bend instead of upwards as in figure 2. However none of the investigated microbes showed a linear plot over the whole range of boxes *n i.e.* a long range correlation (or fractal)..

Although the short distance correlation characteristic as reflected in the value of α_1 is close to the random 0.5 value, its value seems significantly different from 0.5. This is further illustrated in figure 3 where α_2 is plotted against α_1 for different microbial genomes. This correlation "phase space" like picture shows the variability of correlation for different species of bacteria. *Bacillus subtilis* and *Mycoplasma penetrans* has extreme values for α_1 and α_2 respectively. On the other hand the distribution characteristics of these bacteria are placed into the middle of figure 3. Obviously the distribution and correlation characteristics are independent. Putting together the distribution and correlation data we can notice both similarity and difference among the organization of genome.

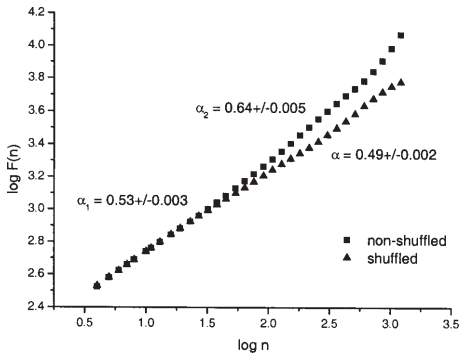

Fig. 2. Detrended fluctuation analysis of the coding sequences length series of Clostridium beijerincki. The lower plot is for the same series subjected to shuffling. The correlation exponents are included in the figure. The shuffled data are described by a single corrletion time α

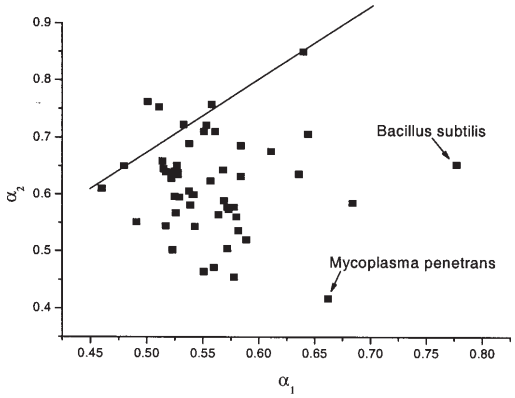

Fig. 3. The short distance α_1 and long distance α_2 correlation exponents for microbial CDS length series. The straight line is intended to guide the eye. Some of the species are also identified in the figure or the place of a species can be located by using the data of table 1

In terms of correlation a non uniform organization of the series is equivalent to short-range correlation. This is understood as a correlation $\alpha \neq 0.5$ extending over less than about one order of magnitude of box sizes *n* [8]. In other words CDS length values appear to be slightly or stronger correlated at short distances or at longer distances. Here "long" distances should not be confounded with "long range" correlation. We remind again that "long range" correlation means correlation over any scale in the series (also known as fractal) while correlation at "longer distances" remains within the short-range correlation class. To be more explicitly short-range correlation at "short distances" means for example that correlation between the first and second, first and third, etc., first and 9th term are higher than 0.5. Also short-range correlation may hold at longer distances, for example between the first term and the 10^{th} term, between the first and the 11^{th} term, the first and the 12^{th} term and so on. This example is clearly a short-range correlation as it extends over a limited range of data, lower than an order of magnitude.

Conclusions

The overall statistical characterization of the CDS length series show that distribution is approximated by exponential distribution. The correlation of data in the series is weak at distances shorter than an order of magnitude and more pronounced (either stronger or anti correlated) at distances longer than an order of magnitude. The reason is that the CDS length series seems to have a non uniform structure. It is suggested that the non-uniform organization of the coding sequences in the microbial genomes could be modeled by short-range memory systems.

Acknowledgements: This work was supported by a grant from the Romanian Authority for Scientific Research.

References

1. YU, Z.-G., ANH, V, Chaos Solitons and Fractals. 12, 2001, p.1827
2. YU, Z.-G., ANH, V. WANG, B., Phys. Rev.E. 63, 2001, 011903
3. ZAINEA, O., MORARIU, V.V., Fluct.Noise Lett. 7, 2007, p.L501
4. ZAINEA, O., MORARIU, V.V., Romanian J. Biophys. 18, 2008, p.19
5. PENG, C.-K., S.V. BULDYREV, S.V., HAVLIN, S., SIMONS, M., H.E. STANLEY, H.E., GOLDBERGER, A.L., Phys. Rev. E. 49, 1994, p.1685
6. MORARIU, V.V., BUIMAGA-IARINCA, L., VAMO[a], C., [a]OLTUZ, [a].M., Fluct.Noise Lett.7, 2007, p.L249

Manuscript received: 20.08.2008