Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2012, Article ID 287486, 12 pages
doi:10.1155/2012/287486

*Research Article*

# Sequence Complexity of Chromosome 3 in *Caenorhabditis elegans*

**Gaetano Pierro**

*System Biology, PhD School, University of Salerno, Via Ponte Don Melillo, 84084 Fisciano, Italy*

Correspondence should be addressed to Gaetano Pierro, gaetanopierro@hotmail.it

Received 29 February 2012; Revised 16 May 2012; Accepted 7 June 2012

Academic Editor: Ramana Davuluri



The nucleotide sequences complexity in chromosome 3 of *Caenorhabditis elegans* (*C. elegans*) is studied. The complexity of these sequences is compared with some random sequences. Moreover, by using some parameters related to complexity such as fractal dimension and frequency, indicator matrix is given a first classification of sequences of *C. elegans*. In particular, the sequences with highest and lowest fractal value are singled out. It is shown that the intrinsic nature of the low fractal dimension sequences has many common features with the random sequences.

## 1. Introduction

The *Caenorhabditis elegans* (*C. elegans*) is a 1 mm length transparent nematode. Thanks to its simple organic structure, it was taken as a model for research into genetic field. Early studies on *C. elegans* began in 1962 with some works on cell lineage and apoptosis [1, 2]. There are 2 distinct sexual types of the *C. elegans*, the hermaphrodite and the male. The second one is very rarely represented in nature (being approximately only the 0.05% of the population). We have 959 cells in the hermaphroditic species and 1031 cells for the male. The sexual difference at the chromosomal level provides: XX chromosomes for hermafrodite and X0 for the male. The sexual reproduction of *C. elegans* is realized by 2 distinct pathways: mating or, in case of the hermaphrodite, by a self-fertilization. The life cycle of *C. elegans* consists of 4 larval stages (from L1 to L4); however, if there exists some hard environment conditions, such as lacking of food, the *C. elegans* remains in the L3 larval stage, until the conditions improve.

The complete sequencing of *C. elegans* genome was completed in 2002. The *C. elegans* has 5 chromosomes autosomes plus the sex chromosome X. Totally, it is made up of nearly 100 million base pairs and 19000 genes [3–5]. Study on fractal analysis of multigenome of *C. elegans* has shown that chromosome 3 is the one with multifractal characteristics higher than the others, the less multifractal appears to be the chromosome sexual X [6]. For the first time, in this work, we have analyzed the different types of sequences belonging to the genome of *C. elegans*, focusing our investigation on those that show fractal characteristics. Thus, chromosome 3 of *C. elegans* has been carefully studied because its unsymmetrical and inhomogeneous statistical characteristics. Through the analysis of this chromosome we can investigate what are the features that make it more "complex" from a biostatistical point of view and in particular with the use of statistical parameters such as the complexity, the fractal dimension, the matrix correlation, and the nucleotide frequency. The concept of fractality in biology is further clarified.

On the chromosome 3 of *C. elegans,* 2780 genes have been identified. In this paper, almost all nucleotide sequences that are located on chromosome 3 of *C. elegans* were analyzed and compared with random sequences. In particular, it will be shown that the nucleotide sequences with a low fractal value have common features with random sequence with low fractal dimension. Moreover, the highest fractal dimension corresponds to sequence close to random sequence with high fractal value, and in particular, it is shown a high frequency of cytosine.

From mathematical point of view, a fractal is a geometric object, characterized by the self-similarity; that is, it repeats

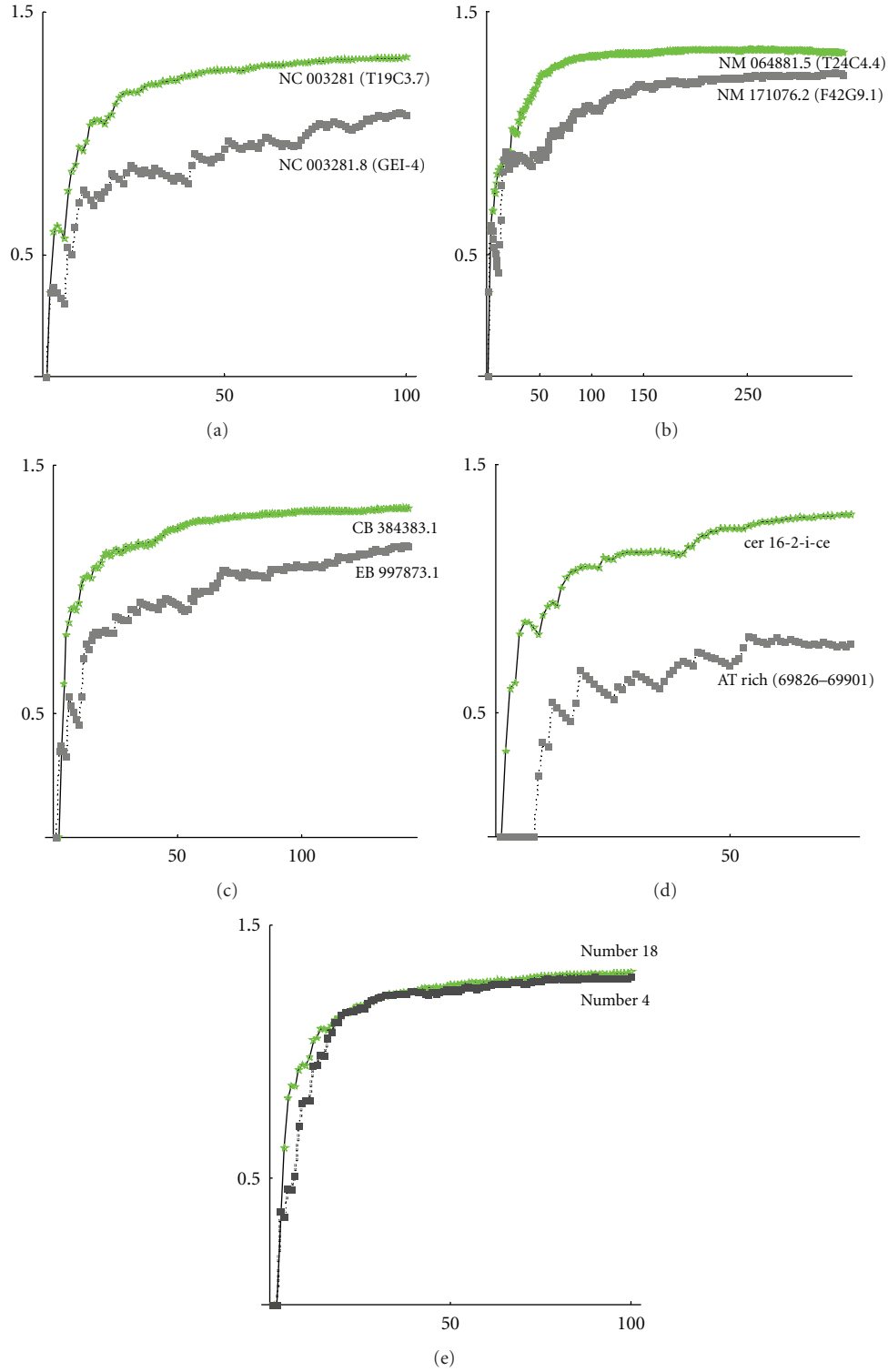


Figure 1: Curves of min-max complexity: (a) whole gene, (b) noncoding, (c) coding, (d) repeats, and (e) random sequences.

its structure cyclically in the same way at different scales. A more rigorous definition of a fractal is based on four properties: self-similarity, fine structure, irregularities, and noninteger dimension [7]. The fractal dimension is a parameter to compute the degree of complexity or disorder by measuring the unsmoothness of the object. This value enables to measure the amount of information contained in the sequence, the higher value corresponds to a higher information content. Generally, this value ranges between 1 and 2, so that the higher value corresponds to the higher

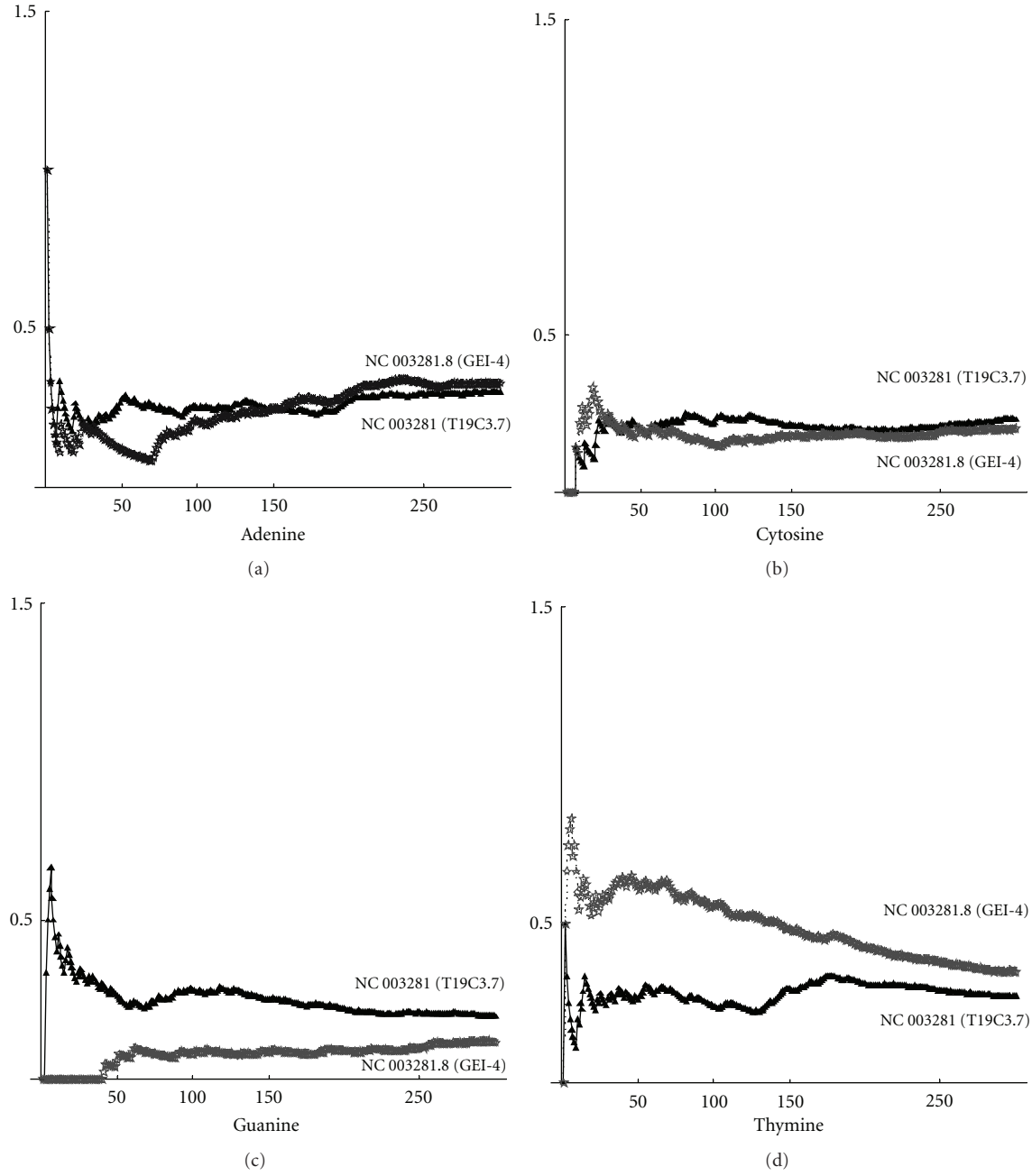


Figure 2: Max-min frequency curves for the whole sequence.

complexity. Fractality has been observed and measured in pathology and cancer models [8, 9], the study of branching blood vessels, or the irregularity of the contours of tumor cells [10, 11], the analysis of complete genomes [12], the correlation analysis of protein sequences [13] tissue pathology [14], in exons, introns [15], and nuclei [16], and it is involved in blood cancer [17, 18].

## 2. Materials and Methods

In the chromosome 3 of *C. elegans,* there have been singled out 2780 genes [19]. Some of them are very short, less than about 50 nucleotides, thus being useless for any statistical analysis, and some of them are still under investigation, so that some nucleotides are not yet properly identified. For this reason, there have been selected only some sequences with significant length, the shortest being about 100 nucleotides. In particular, we investigated 100 genes (whole sequence), 85 repeats sequences, 71 noncoding sequences (introns), and 100 coding sequences (exons lacks of UTR). In order to make a comparison with random sequences, 100 random sequences of 100 nucleotides have been generated. In this work, all sequences were downloaded from the National Center for Biotechnology Information

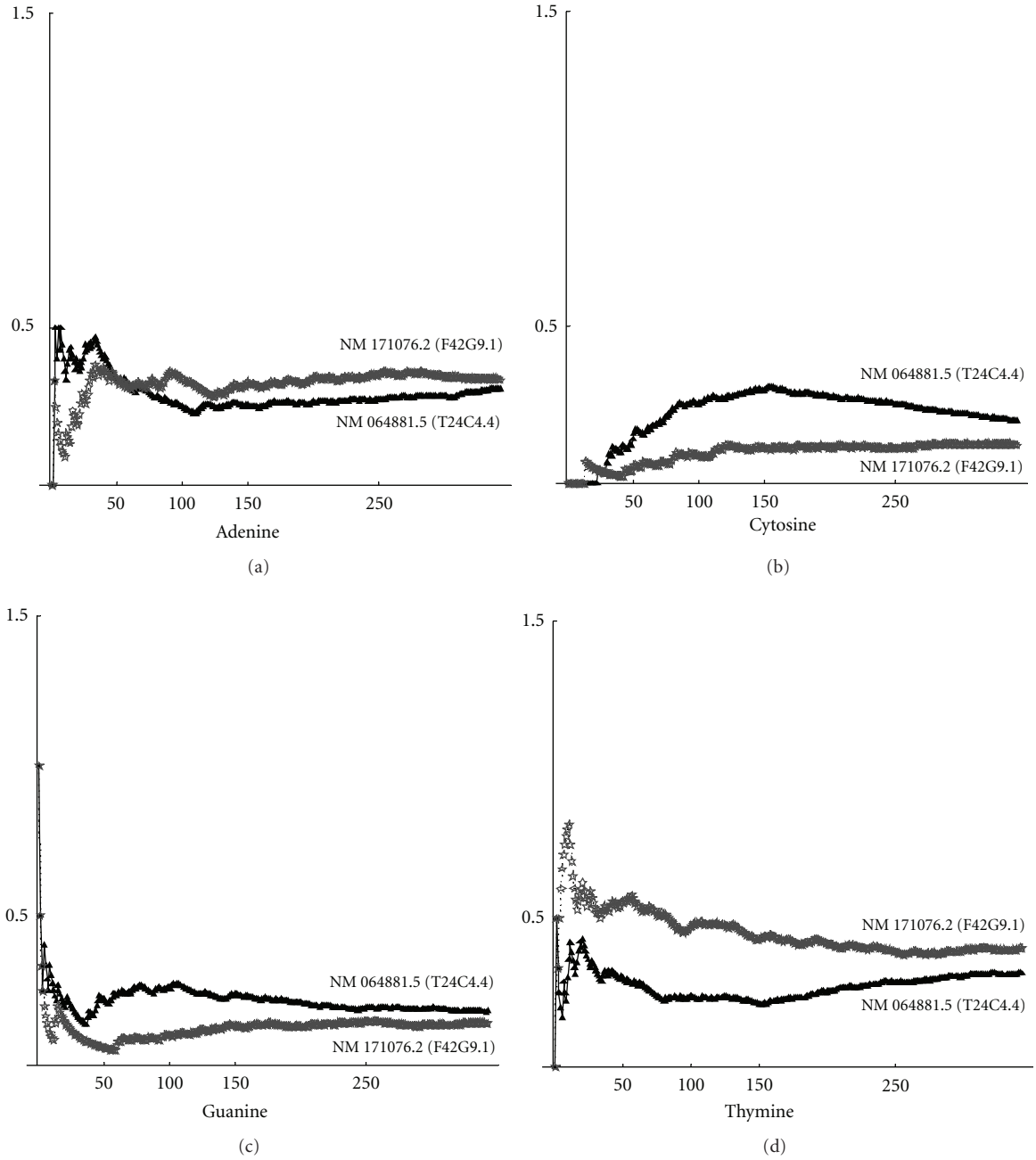


Figure 3: Max-min frequency curves for noncoding sequences.

[19]. A simple formula to estimate the fractal dimension has been given in [20, 21] and based on the correlation matrix, as follows. The fractal dimension is defined as the average of the number $p(n)$ of 1 in the randomly taken $n \times n$ minors of the $N \times N$ correlation matrix $u_{hk}$ (see also [20–24]).

In particular, let

$$\aleph_4 = \{A, C, G, T\} \tag{1}$$

be the finite set (alphabet) of nucleotides and $x \in \aleph_4$ any member of the 4 symbols alphabet.

A DNA sequence is the finite symbolic sequence $\mathfrak{D}(N) = ¥ \times \aleph_4$ so that

$$\mathfrak{D}(N) = \{x_h\}_{h=1,\ldots,N}, \quad N < \infty \tag{2}$$

being

$$x_h = (h, x) = x(h), \quad (h = 1, 2, \ldots, N;\ x \in \aleph_4) \tag{3}$$

the acid nucleic $x$ at the position $h$.

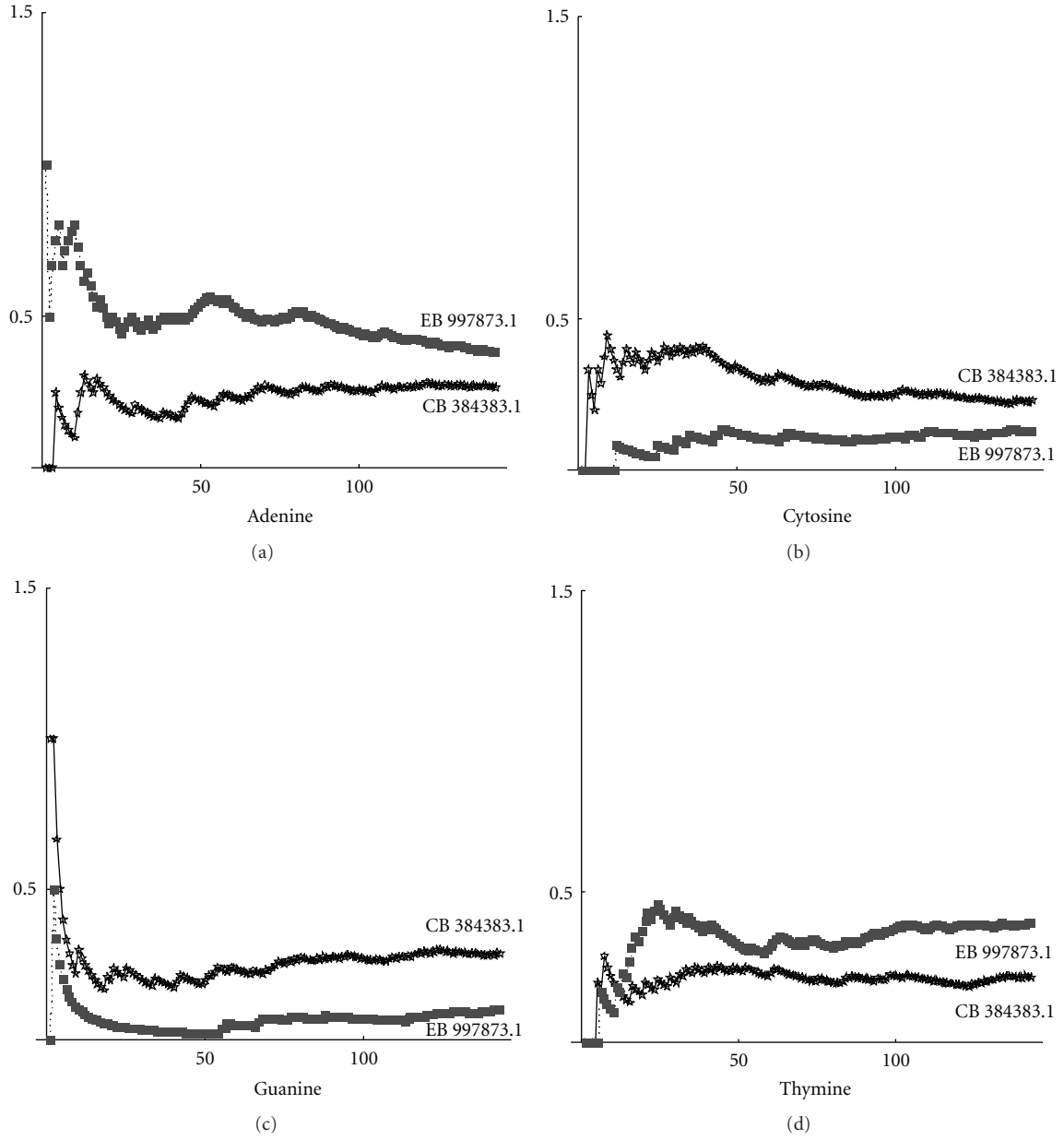


Figure 4: Max-min frequency curves for coding sequence.

Let $\mathfrak{D}_1(N)$, $\mathfrak{D}_2(N)$ be two DNA sequences, the indicator function [20, 22–26] is the map

$$u : \mathfrak{D}_1(N) \times \mathfrak{D}_2(N) \longrightarrow \{0, 1\} \tag{4}$$

such that the correlation matrix

$$u_{hk} = u(x_h, x_k) = \begin{cases} 1, & \text{if } x_h = x_k, \\ 0, & \text{if } x_h \neq x_k, \end{cases} \tag{5}$$
$$(x_h \in \mathfrak{D}_1(N),\ x_k \in \mathfrak{D}_2(N))$$

is a matrix of 0's and 1's showing the existence of correlation. When $\mathfrak{D}_1(N) \equiv \mathfrak{D}_2(N)$, the indicator function shows the existence of autocorrelation on the same sequence.

The probability distribution of nucleotides can be defined by the frequency

$$p_X(n) = \frac{1}{n}\sum_{i=1}^{n} u_{Xi}, \quad (X \in \aleph_4,\ x_i \in \mathfrak{D}(N);\ 1 \leq n \leq N) \tag{6}$$

that the acid nucleic $X$ can be found at the position $n$. This value can be approximated by the frequency count (on the indicator matrix) of the nucleotide distribution before $n$ [20, 21, 23, 24]

$$D = \frac{1}{2}\frac{1}{N}\sum_{n=2}^{N}\frac{\log p(n)}{\log n}. \tag{7}$$

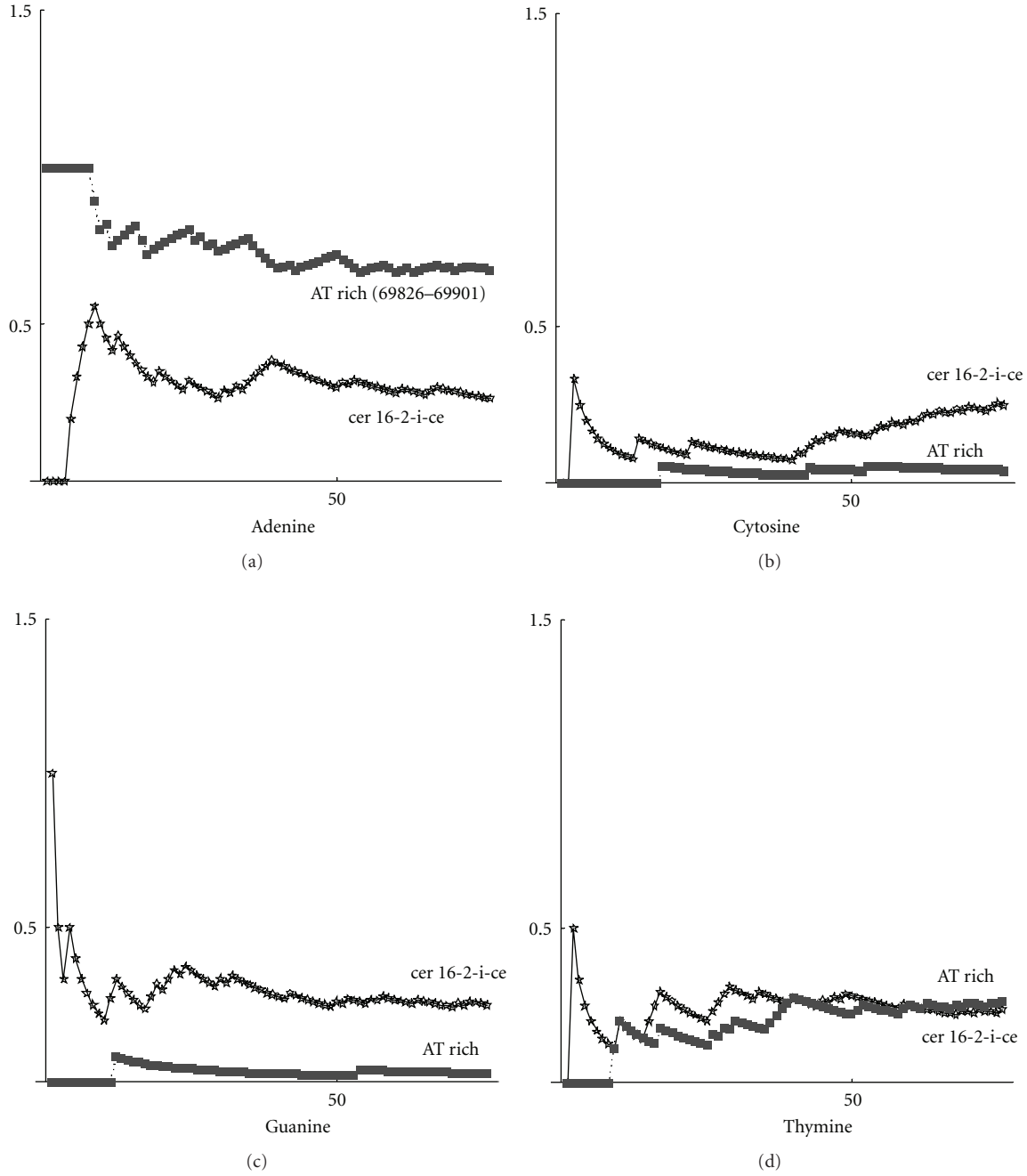


Figure 5: Max-min frequency curves for repeats sequence.

Table 1: Max value of fractal dimension of sequences.

| Type of sequence | Max value of fractal dimension | Tag of genomic sequence |
|---|---|---|
| Whole sequence of gene | 1.29850 | NC 003281 (T19C3.7) |
| Noncoding | 1.29808 | NM 064881.5 (T24C4.4) |
| Coding | 1.30639 | CB 384383.1 |
| Repeats | 1.31280 | CER 16-2-i-CE |
| Random sequence | 1.28452 | Number 18 |

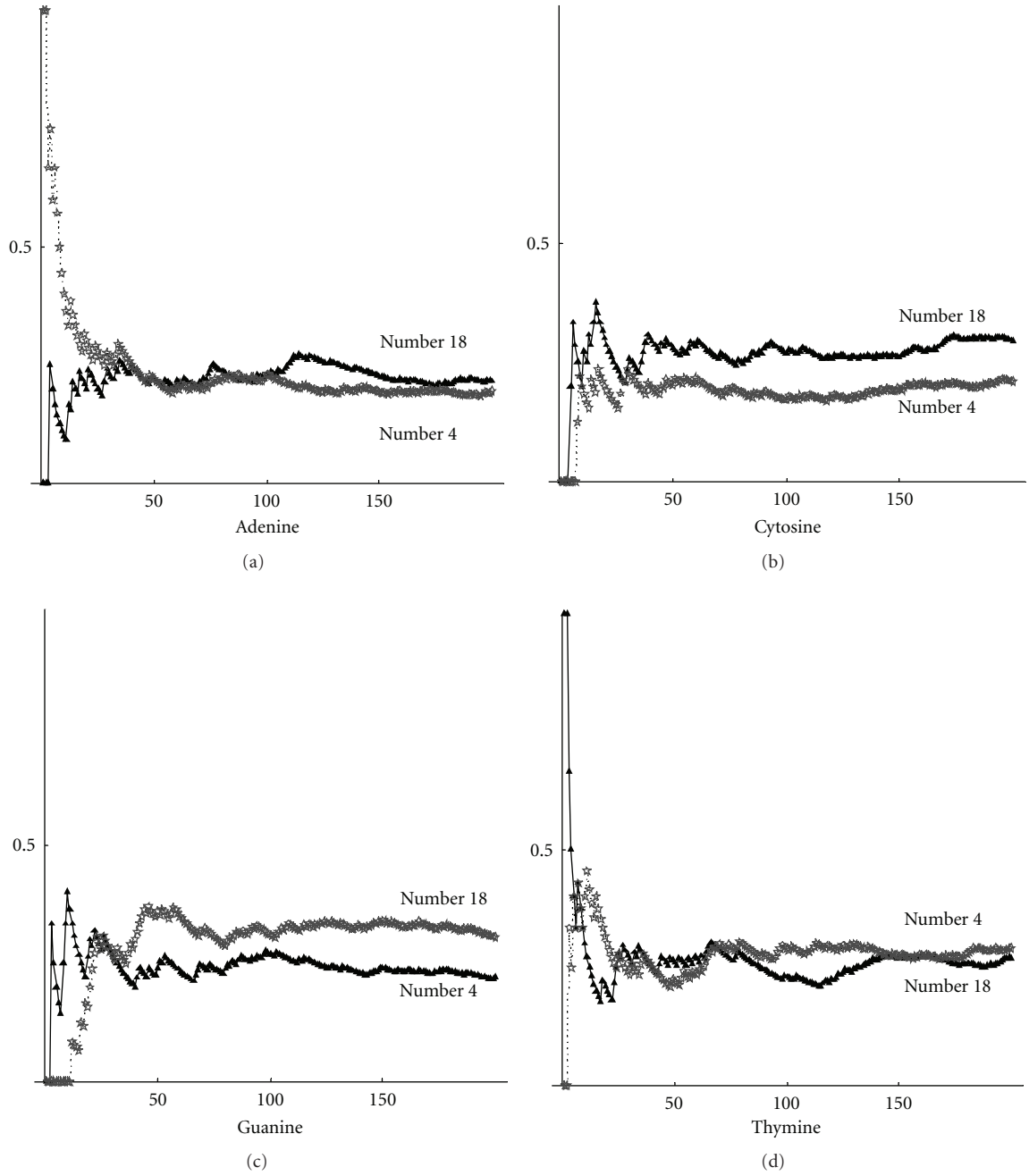


Figure 6: Max-min frequency curves for random sequences.

Table 2: Min value of fractal dimension of sequences.

| Type of sequence | Min value of fractal dimension | Tag of genomic sequence |
|---|---|---|
| Whole sequence of gene | 1.27016 | NC 003281.8 gei-4 |
| Noncoding | 1.27494 | NM 171076.2 (F42G9.1) |
| Coding | 1.27846 | EB 997873.1 |
| Repeats | 1.24155 | AT rich (69826–69901) |
| Random sequence | 1.28201 | Number 4 |

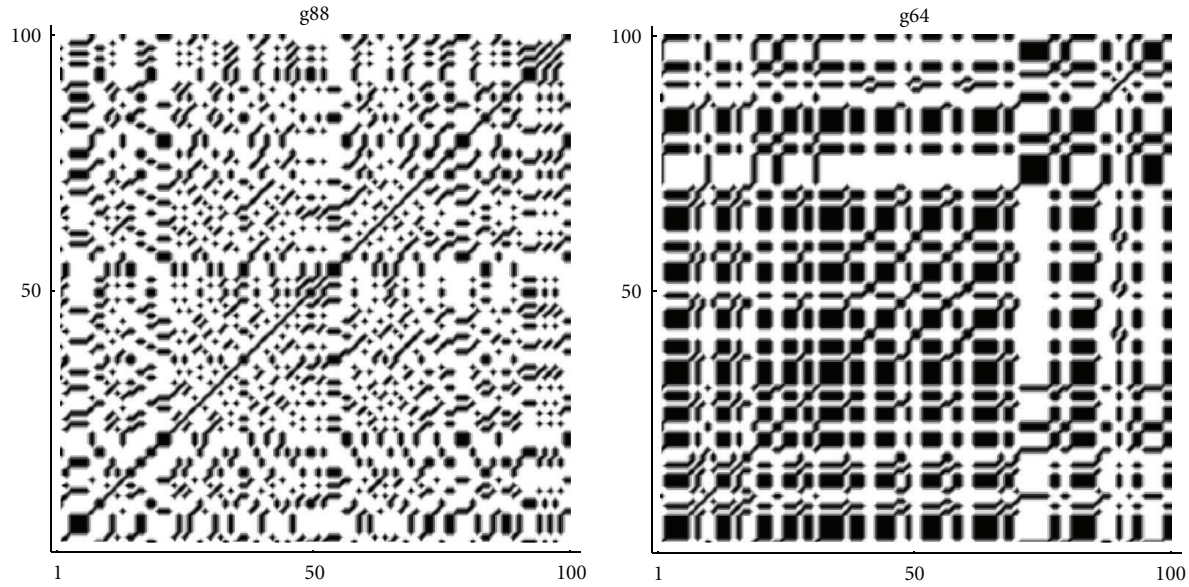


Figure 7: Autocorrelation plots on the whole sequence gene corresponding to max and min values of fractal dimension in (a) and (b), respectively.

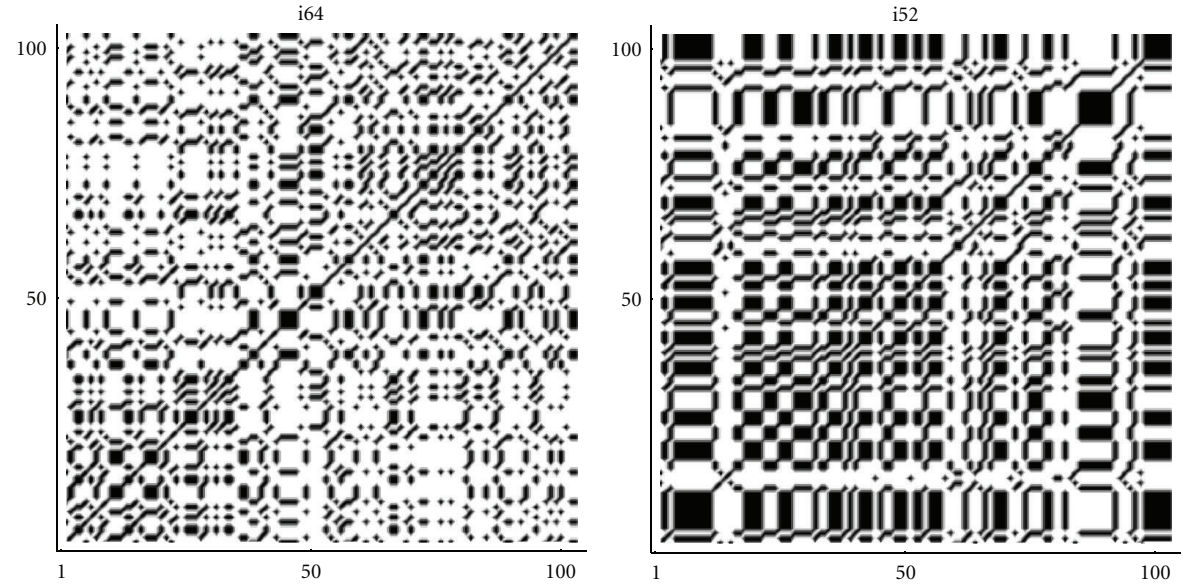


Figure 8: Autocorrelation plots on the noncoding sequences corresponding to max and min values of fractal dimension in (a) and (b), respectively.

In order to have a measure of complexity, for an $n$-length sequence, we use the following definition [20–24]:

$$K = \log\left(\frac{n!}{a_n!c_n!g_n!t_n!}\right)^{1/n} \tag{8}$$

with

$$a_n = \sum_{h=1,\ldots,n} u(A, x_h), \qquad c_n = \sum_{h=1,\ldots,n} u(C, x_h),$$
$$g_n = \sum_{h=1,\ldots,n} u(G, x_h), \qquad t_n = \sum_{h=1,\ldots,n} u(T, x_h). \tag{9}$$

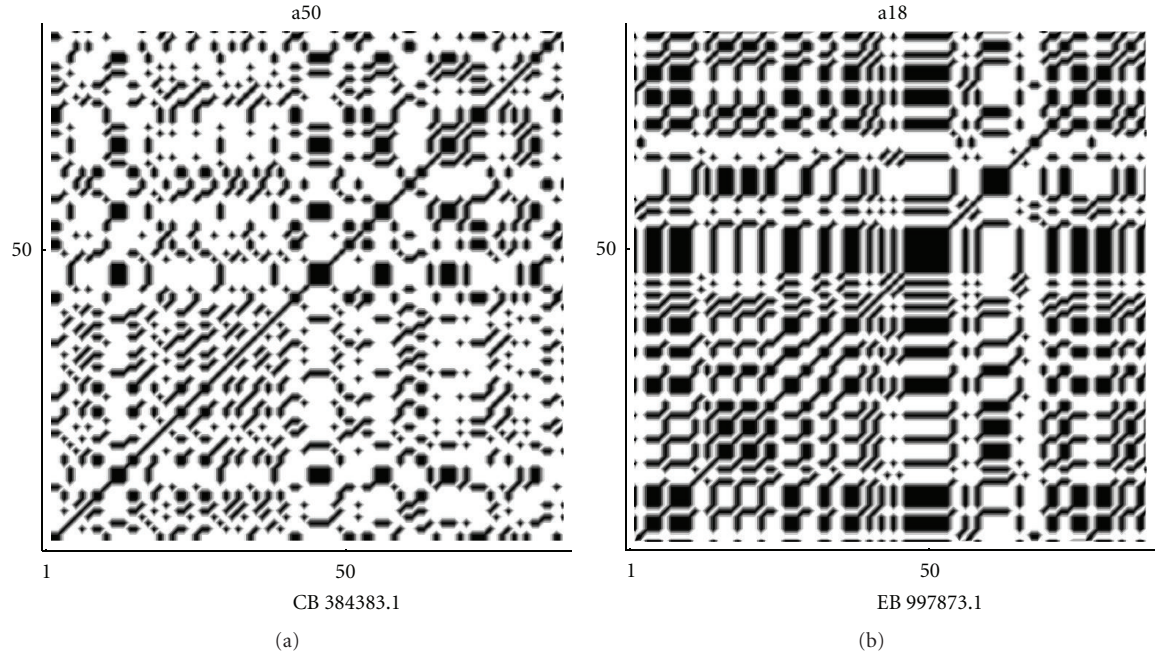


Figure 9: Autocorrelation plots on the coding sequences corresponding to max and min values of fractal dimension in (a) and (b), respectively.

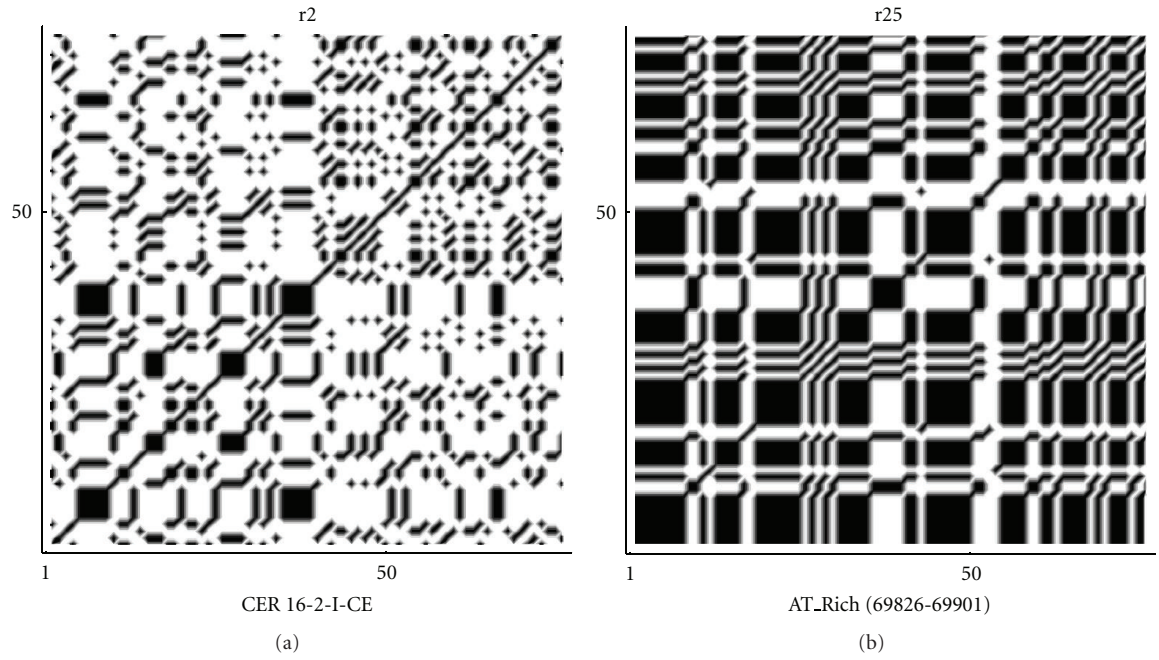


Figure 10: Autocorrelation plots on the repeats sequences corresponding to max and min values of fractal dimension in (a) and (b) respectively.

## 3. Results

By using formula (7), for each sequence of nucleotides, the corresponding fractal dimension has been computed, and obtained results are shown in Tables 1 and 2. In particular, the sequences with max/min values of fractal dimension among the whole sequences, coding/noncoding sequences, repeat sequences, random sequences have been singled out.

From these computations, we can see that the repeats sequence AT rich (69826–69901) has the lowest fractal value 1.24155. This could be explained because we have a large number of only 2 nucleotides, so that the sequence is simple

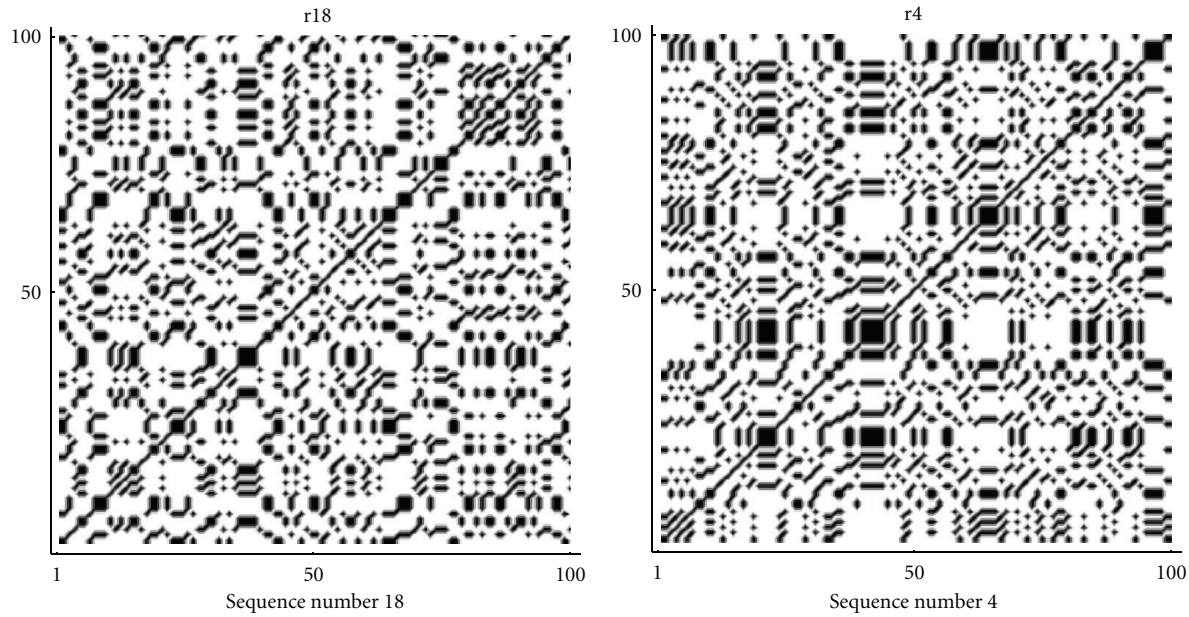


Figure 11: Autocorrelation plots on the random sequences corresponding to max and min values of fractal dimension in (a) and (b), respectively.

in the sense that there is a low variability and it shows a low complexity. Analogously, the sequence with the highest value of fractality is still a repeats sequence CER 16-2-i-CE with a fractal dimension 1.31280. Although there are some fluctuations, due to the fact that random generation, by a computer, is indeed a pseudorandom generation, the values of fractal dimension for random sequences are localized around 1.28, which appears to be the intermediate value between the maximum and minimum values obtained for all sequences examined. Further information about the heterogeneity of data is given by the complexity parameter (8). In Figure 1, the complexity curves corresponding to the sequences for maximum and minimum values of the fractal are plotted.

We investigated the complexity of the nucleotide sequences. In all cases, we obtained that the curve of higher complexity corresponds to the sequence with the highest fractal dimension. Thus, we can draw the conclusion that complexity and fractal dimension are equivalent parameters for studying the complexity. These results depend on the distribution of nucleotides. By using the definition (6), we can compute the frequency distribution on a sequence. Below are shown the frequencies for each nucleotide (adenine, cytosine, guanine, thymine). In particular, in Figure 2, the max-min curves for frequencies on the whole gene sequence are plotted. It can be seen that, in this case, adenine and cytosine tend to have the same value, while thymine and guanine maintain a significant distance between the max and min curves. Max-min frequency curves for noncoding sequences are shown in Figure 3. By taking into account the values of fractal dimensions, as given in Tables 1 and 2, we can observe that the higher frequency of cytosine corresponds to the higher fractal dimension. Thymine, instead, is more present in sequences with low fractal dimension. In Figure 4, the curves for max-min frequency of coding sequences are drawn. It can be seen, also in this case, that adenine and thymine are more present in the sequence with lower fractal dimension. As before, cytosine is more present in sequences with higher fractal dimension. Repeats and random sequences are given in Figures 5 and 6, respectively. In the first case for adenine, we have more frequencies rate for the low fractal sequence, while for cytosine we have more frequencies rate for the high fractal sequence. For random sequences, we have that the cytosine is more frequent in the sequence that has the highest value of fractal.

By the frequency analysis and the results of Tables 1 and 2 on the fractal dimension we can see that there is a correspondence between the frequencies of nucleotides and the fractal dimension. So that, sequences that show a lower fractal dimension have always a higher frequency for the adenine and thymine (in most cases), while the cytosine is more frequent in high fractal sequences. Almost the same results are true also for random sequences, especially for the thymine and cytosine. According to (5), the indicator map of the $N$-length sequence can be easily represented by the $N \times N$ sparse matrix of binary values $\{0, 1\}$ and this matrix can be visualized by the following (autocorrelation) dot-plots [20, 22] of Figures 7, 8, 9, 10, and 11. Figure 11(a) shows the sequences (of Table 1) with max value of fractal dimension, while in Figure 11(a), there are the sequences of Table 2 with min value of fractal dimension. We can see that also in these plots the distribution of nucleotides gives rise to some typical patterns.

All sequences with low fractal dimension (Figure 11(b)) turn out to have an important presence of nucleotide

correlation, this feature is less present in the sequences with higher fractal dimension, where we expect to have a more complex structure of the sequence.

## 4. Discussion

In this work, by means of statistical parameters such as indicator matrix, complexity, frequency, and fractal dimension, the different types of sequences (repeats, coding, noncoding, whole gene, random) of chromosome 3 (the one with the highest fractality) of the *C. elegans* have been analyzed. Our attempt was to give a statistical classification of these sequences and to understand the complexity of the sequences as a function of the nucleotides' distribution. By using (7) the values of the fractal dimension for all sequences are obtained. In detail, it was observed that the repeats sequences (which do not code for proteins) have a higher variability of values, since they assume the minimum and maximum on all sequences in the *C. elegans*. This leads us to analyze the role and the functional meaning of the repeats within the sequences of genes. Thereafter, we have verified the equivalence, with respect to the complexity, between the fractal dimension and complexity, since the sequences with highest fractality appear to have also a greater degree of complexity. Through the frequency distribution of nucleotide, it was noticed that the adenine is more present in sequences having a lower fractal dimension and, in particular, for the one being in absolute the lowest fractal (AT RICH). This result seems to be dependent on the fact that the sequence is made up of only 2 nucleotides, that is, adenine and thymine. Cytosine, instead, appears to be the most frequent nucleotide in the sequence with the highest fractal value and in particular for the sequence CER 16-2-i-CE. These results lead us to conjecture that there is a correlation between fractal dimension and the frequency of nucleotides such as adenine and cytosine. The information contents of a sequence of nucleotides depend on the different distribution of nucleotides, so that two sequences having the same nucleotides which are distributed according to two different permutations might have two different complexities (fractal dimension). In future work, this aspect of the different organization within the sequence will be further analyzed. Moreover, these results must be confirmed in other organisms which are evolutionarily distant from each other to better investigate the findings so far. At the moment, the obtained results were compared with some random sequences, which have a nucleotide random distribution, and in that case, we have obtained a significant correspondence with the complexity of the nucleotide sequences.

## References


[1] S. Brenner, "The genetics of Caenorhabditis elegans," *Genetics*, vol. 77, no. 1, pp. 71–94, 1974.

[2] C. Kenyon, "The nematode Caenorhabditis elegans," *Science*, vol. 240, no. 4858, pp. 1448–1453, 1988.

[3] J. Hodgkin, H. R. Horvitz, B. R. Jasny, and J. Kimble, "C. elegans: sequence to biology," *Science*, vol. 282, no. 5396, p. 2011, 1998.

[4] A. F. Bird and J. Bird, *The Structure of Nematodes*, Academic Press, San Diego, Calif, USA, 1991.

[5] D. L. Riddle, T. Blumenthal, R. J. Meyer, and J. R. Priess, *C. elegans II*, Cold Spring Harbor Laboratory Press, New York, NY, USA, 1997.

[6] P. E. Velez, L. E. Garreta, E. Martinez et al., "The Caenorhabditis elegans genome: a multifractal analysis," *Genetics and Molecular Research*, vol. 9, no. 2, pp. 949–965, 2010.

[7] B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman & Co, San Francisco, Calif, USA, 1982.

[8] S. S. Cross, "Fractals in pathology," *Journal of Pathology*, vol. 182, no. 1, pp. 1–8, 1997.

[9] J. W. Baish and R. K. Jain, "Fractals and cancer," *Cancer Research*, vol. 60, no. 14, pp. 3683–3688, 2000.

[10] S. S. Cross and D. W. K. Cotton, "The fractal dimension may be a useful morphometric discriminant in histopathology," *Journal of Pathology*, vol. 166, no. 4, pp. 409–411, 1992.

[11] A. L. Goldberger and B. J. West, "Fractals in physiology and medicine," *Yale Journal of Biology and Medicine*, vol. 60, no. 5, pp. 421–435, 1987.

[12] Z. G. Yu, V. Anh, and K. S. Lau, "Measure representation and multifractal analysis of complete genomes," *Physical Review E*, vol. 64, no. 3, Article ID 031903, pp. 319031–319039, 2001.

[13] Z. G. Yu, V. Anh, and K. S. Lau, "Multifractal and correlation analyses of protein sequences from complete genomes," *Physical Review E*, vol. 68, no. 2, Article ID 021913, pp. 021913-1–021913-10, 2003.

[14] G. A. Losa and T. F. Nonnenmacher, "Self-similarity and fractal irregularity in pathologic tissues," *Modern Pathology*, vol. 9, no. 3, pp. 174–182, 1996.

[15] Y. Xiao, R. Chen, R. Shen, J. Sun, and J. Xu, "Fractal dimension, of exon and intron sequences," *Journal of Theoretical Biology*, vol. 175, no. 1, pp. 23–26, 1995.

[16] J. G. McNally and D. Mazza, "Fractal geometry in the nucleus," *The EMBO journal*, vol. 29, no. 1, pp. 2–3, 2010.

[17] R. L. Adam, R. C. Silva, F. G. Pereira, N. J. Leite, I. Lorand-Metze, and K. Metze, "The fractal dimension of nuclear chromatin as a prognostic factor in acute precursor B lymphoblastic leukemia," *Cellular Oncology*, vol. 28, no. 1-2, pp. 55–59, 2006.

[18] D. P. Ferro, M. A. Falconi, R. L. Adam et al., "Fractal characteristics of May-Grünwald-Giemsa stained chromatin are independent prognostic factors for survival in multiple myeloma," *PLoS ONE*, vol. 6, no. 6, Article ID e20706, 2011.

[19] National Center for Biotechnology Information, http//www.ncbi.nlm.nih.gov/genbank/.

[20] C. Cattani, "Fractals and hidden symmetries in DNA?" *Mathematical Problems in Engineering*, vol. 2010, Article ID 507056, 31 pages, 2010.

[21] C. Cattani and G. Pierro, "Complexity on acute myeloid leukemia mRNA transcript variant," *Mathematical Problems in Engineering*, vol. 2011, Article ID 379873, 16 pages, 2011.

[22] C. Cattani, "Wavelet algorithms for DNA analysis," in *Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications*, M. Elloumi and A. Y. Zomaya, Eds., Wiley Series in Bioinformatics, chapter 35, pp. 799–842, John Wiley & Sons, New York, NY, USA, 2010.

[23] C. Cattani, "On the existence of wavelet symmetries in archaea DNA," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 673934, 21 pages, 2012.

[24] C. Cattani, "Complexity and simmetries in DNA sequences," in *Handbook of Biological Discovery, (Wiley Series in Bioinformatics)*, M. Elloumi and A. Y. Zomaya, Eds., Chapter 5, pp. 700–742, John Wiley & Sons, New York, NY, USA, 2012.

[25] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Physical Review Letters*, vol. 68, no. 25, pp. 3805–3808, 1992.
[26] R. F. Voss, "Long-range fractal correlations in DNA introns and exons," *Fractals*, vol. 2, no. 1, pp. 1–6, 1992.