SDI Paper Template Version 1.6 Date 11.10.2012

Fractal Analysis of DNA by Nonlinear Genome Signal Processing for Exon and Intron Separation

Ali Karmi, Ali Najafi, Peyman Gifani, and Sahand Khakabimamaghani*

Molecular Biology Research Center, Baqiyatallah University of Medical Sciences,Tehran, Iran

ABSTRACT

Aims: To provide a new reasonable measure for distinguishing between coding and non-coding regions of DNA sequences based on its fractal nature and self-similarity.
Study design: After conducting background studies on the fractal structure of DNA sequences, the application of Detrended Fluctuation Analysis for identifying coding and non-coding regions in those sequences was investigated. Finally, the propositions were tested on a standard dataset of 195 genes.
Place and Duration of Study: Sample: We use a common data set, "HMR 195", which has been used in conventional tools, between December 2012 and July 2013.
Methodology: The Fractal Scaling Exponent (FSE) of the numerical signal, produced by converting a DNA string to a numerical sequence via a number mapping algorithm, was calculated for exons and introns of 195 genes. This calculation was repeated twice: once for computing the optimal values of FSE, and once for non-optimal FSEs. Analysis of Variance (ANOVA) was used for investigating the significance of difference between the average FSE of exons versus that of introns in both optimal and non-optimal cases.
Results: ANOVA indicated a significant gap between the optimal mean FSE of exons (0.65) and introns (0.72). The difference, although smaller, was significant for non-optimal values as well.
Conclusion: Throughout this study, the FSE is proved to be a reliable measure for distinguishing between coding and non-coding regions of DNA gene sequences based on our experiments. Accordingly, this metric can be used for predicting exons/introns when embedded within current tools such as TestCode. However, its contribution to the predictive accuracy of current methods requires more investigation in the future works.

Keywords: DNA sequence; fractal scaling exponent; exon; intron

1. INTRODUCTION

The biological information is embedded in the DNA sequence which stores, produces and controls processes for growth and maintenance of living systems. This information organized in a structure of a nanowire, is encrypted in sequences of four bases, i.e. A, T, G, C. Each segment of every organism represents a processor to execute a particular biological process. Traditional approaches based on reductionism are hardly ever able to characterize more than a tiny subset of the full range of behaviors. In the past years, many well-known molecular biologists have pointed out the vital need for theoretical and computational tools to show the spatial and temporal organization of macromolecules interaction implicit in the way to create a living cell as a whole [1].

* Tel.: 00982182482548; fax: 00982188039883.
E-mail address: najafi74@ibb.ut.ac.ir, najafi74@yahoo.com

Gene identification in prokaryotes is easier, because the coding regions are small continuous strings of DNA. However, in higher eukaryotic organisms, genes are often split into a number of coding sequences (exons) separated by non-coding intervening sequences (introns). In gene identification, we can use intrinsic information derived from the query sequence itself, in addition to extrinsic information achieved by comparing the query sequence with other known sequences in public databases. Examples of intrinsic information are promoters, splice sites, and CpG islands. This information can also be derived according to the fact that coding region sequences in the DNA exhibit specific statistical properties.

Locating genes on DNA which has not been analyzed for potential coding regions involves using statistical detection methods by using probability models to predict where in a DNA sequence a gene is located. The nucleic acid sequence probabilities can be determined through analysis of known coding regions and can be categorized into measures that depend on coding DNA and measures that are independent of coding DNA. Model dependent statistics capture the specific features of coding DNA whereas model independent statistics capture the global features of coding DNA. Since model independent methods don't need a sample of coding DNA they can be used in the absence of previous knowledge of the under consideration species. Therefore, there are two approaches namely knowledge-based methods and *ab-initio* techniques. The knowledge-based methods suffer from some disadvantages. The methods such as hidden Markov models or Artificial Neural Network which uses training based system are organism/dataset-specific and the accuracy of this method is affected when the information of newly sequenced genomes or available organisms is limited [2].

The development of high-throughput data-collection techniques for sequencing DNA such as next generation sequencing brings vast nucleic acid sequence data rapidly. By sequencing the entire human genome and the genomes of several other species, a need for the rapid identification of genes on long stretches of sequenced DNA has been emerged. Although the conventional gene detection techniques, such as cDNA hybridization, are effective in locating transcribed genes, but these methods are based on reductionism approach, time-consuming and costly. Due to the present size and increasing rate of new raw data, we need systemic and integrative ways of thinking about information organization of genomes to check quickly for similarities and differences among them and to explore the interactions among genotypes, phenotypes, and the environment. By the advent of Bioinformatics the need for new computational tools to analyze and interpret the large amount of nucleotide sequences available in databases has been recently highlighted [3].

One of the whole genome structural features is the long-range correlation or scale-invariant property of DNA. This phenomenon implies that the occurrence of a nucleotide in a specific position depends on the previous nucleotides and also the occurrence of a small segment of nucleotides depends on large scale segments. Such long-range correlation is directly related to power-law and fractal structure of the DNA sequence. There is self-similarity among different scales of sequences which means that its fragments can be rescaled to resemble the original sequence itself.

In this paper we investigated this global feature of DNA sequence by calculating the Fractal Scaling Exponent (FSE) of the numerical signal which is produced by converting a DNA string to a numerical sequence by number mapping algorithm. By this approach, we have a numerical signal for a DNA string which could be analyzed by different signal processing algorithms. Based on fractal structure of DNA sequence, in this paper, we implemented Detrended Fluctuation Analysis to calculate the FSE of this signal. This measure, that captures another aspect of difference between coding and noncoding DNA sequences, can be used in existing *ab initio* prediction methods.

2. MATERIAL AND METHODS

Providing robust computing solutions for DNA sequence analysis is a challenging issue in Bioinformatics. Most of the Bioinformatics tools are currently searching for patterns or correlations existing in the DNA sequence based on codons, amino acids, and proteins using a variety of sophisticated computational techniques, including neural network algorithms, dynamic programming, decision trees, stochastic reasoning, and hidden Markov chain.

The application of chaos and fractal theory considering intrinsic patterns, correlations, and self-similarity measurement, is going to be highlighted in several areas of science. Correlation properties in DNA sequences have been studied in [4] through fractal landscape or DNA walk models, in which genometric method projected the sequence of position for each nucleotide on a two dimensional plot representation of the DNA landscape or walk, providing a landscape comparable among different genomes [5].

2.1 Scaling Exponents and Detrended Fluctuation Analysis

The scaling exponent is important in the characterization of long-range correlations in finite-length sequences. Peng and co-workers have developed an algorithm to estimate scaling exponents with local-detrending to remove the non-stationary components, known as Detrended Fluctuation Analysis (DFA) [4]. This technique can be applied to a self-similar process through simple integration. A DNA signal of length N, $x(s)$, which has been generated from DNA sequence by numerical mapping system, can be integrated to generate a self-similar series $y(k)$ by Equation (1).

$$y(j) = \sum_{i=1}^{j} \left[x(i) - \langle x \rangle \right] \tag{1}$$

Where the average value of x, denoted by $\langle x \rangle$, is given by

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x(i) \tag{2}$$

After partitioning the entire range of $y(k)$ into boxes of equal size n, we fit this integrated signal by using a polynomial function, $y_n(k)$, which is representing the local trend within each box. After removing the trend in the root-mean-square fluctuation, $F(n)$, is given by Equation (3):

$$F(n) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} \left[y(k) - y_n(k) \right]^2} \tag{3}$$

The above computation is repeated for boxes with different sizes (scales) to provide a relationship between $F(n)$ and the size n. A power-law relation between $F(n)$ and the size of

the box, indicates the presence of scaling $F(n) \sim n\alpha$. $F(n)$ is the average fluctuation and usually increases linearly with n.

This gives the scaling relation:

$$F(n) \approx n^{\alpha} \tag{4}$$

The scaling exponent (self-similarity parameter), α, should therefore be able to completely describe the significant correlation properties of DNA signal. Since the Equation (4) represents the scale transformation independent of n, then the exponent, α, provides a succinct measure of the dynamics across a range of n. For ideal mathematical fractals, such behavior persists without a limit in n. However, for real systems, its range is always finite, and may in addition be interrupted by dynamical mechanisms which introduce characteristic scales into the data. In our investigation this limitation occurred and we had to select an optimal range of box sizes. On the other hand, the scaling exponent estimation on the non-optimum region could also lead to meaningful results.

2-2- Converting Sequences to Signal

As we discussed before, to investigate the fractal property of DNA sequence by DFA, we should first convert DNA string to a sequence of numerical values via number mapping algorithm. A genome is simply a string of four nucleotide bases A, T, G, and C, and the mapping system is also compromised a number system of the base four. The system has four digits 0, 1, 2, and 3 assigned to the four bases according to their molecular weights. Smaller digits are assigned to higher molecular weights; that is G = 0, A = 1, T = 2, and C = 3.

By the fact that double strand of DNA are complementary to each other and according that in DNA structure the observed complementary pairing is GC and AT and when we add the values of the GC (0+3=3) and AT (1+2=3) a constant value of three is obtained, the signal generated by the DNA remains the same to its reverse, complementary, and reverse complementary sequence. Thus, in comparison to conventional gene analysis algorithms there is no need to take the sequence and run the algorithm and then take the reverse complement of the sequence and run the algorithm again in our approach and we have analyzed the DNA sequence only once. In order to convert the DNA string to a unique number string, a window of size three nucleotides is slid on the sequence to eliminate any ORF (Open Reading Frame) related bias. The codons are transformed to numbers using the formula $F(X_n, Y_n, Z_n) = 4*4*X_n+4*Y_n+Z_n$ for window n containing bases X, Y, and Z. For example, for sequence CTGTCA, the first triplet CTG is converted into a numerical value after obtaining the numerical value (3, 1, 2) through mapping system described above and then using the formula F(C,T,G) = 4*4*3+4*2+0 = 56; then the windows slide to the next triples, which are TGT, GTC, TCA and respectively.

2-3- Dataset

We use a common data set, “HMR 195” [6], which has been used in conventional tools. Below is a part of detailed description of this database exactly as stated in its website (http://blogs.ubc.ca/sanja/hmr195-dataset/):

DNA sequences were extracted from GenBank release 111.0 (April 1999). The basic requirements in sequence selection were:

- the sequence was entered in GenBank after August, 1997

- the source organism is *Homo sapiens*, *Mus musculus* or *Rattus norvegicus*
- only genomic sequences that contain exactly one gene were considered
- mRNA sequences and sequences containing pseudo genes or alternatively spliced genes were excluded.

Sequences collected according to those principles were further filtered to meet following requirement:

- all annotated coding sequences started with the ATG initiation codon and ended with one of the stop codons: TAA. TAG, TGA.
- all exons had dinucleotide AG at their acceptor site and dinucleotide GT at their donor site.
- sequences that did not contain any nucleotides in their 5' or 3' UTR were discarded.
- the sequences whose coding region contains in-frame stop codon were discarded.

HMR195 has the following characteristics:

- the ratio of Human: Mouse: Rat sequences is 103:82:10
- the mean length of the sequences in the set is 7,096 bp
- the number of single-exon genes is 43 and the number of multi-exon genes is 152.
- the average number of exons per gene is 4.86.
- the mean exon length is 208 bp, the mean intron length is 678 bp and the mean coding length of a gene is 1,015 bp (~330 amino acids).

3. RESULTS AND DISCUSSION

This section presents the results of our approach to analyzing the FSE of genome signal on mentioned dataset. As discussed before the first step is converting DNA sequence to genome signal by numerical mapping system presented in the previous section. The analysis was carried out after separating the intron and exon from the dataset. Any signal processing method can now be used to determine the variation or extract the biological feature form generated signals of introns and exons.

As we discussed before based on fractal structure of DNA sequence, we implemented DFA to calculate the FSE of this signal. The details of this algorithm have been mentioned in the previous section. The scaling exponent can be approximated as the slope of $\log(F(n))$ against $\log(n)$. The parameter α, called the scaling exponent or correlation exponent, represents the correlation properties of the signal. If $\alpha = 0.5$, there is no correlation and the signal is an uncorrelated signal (white noise); if $\alpha < 0.5$, the signal is anti-correlated; if $\alpha > 0.5$, there are positive correlations in the signal. Therefore, finding scaling exponents in the range 0.5 to 1.0 would indicate long-range power-law correlations of the kind which are ubiquitous in nature.

Although the algorithm is simple but the process of optimal feature selection on large amounts of data, is not a straightforward problem. In this section, we discuss the optimal FSE estimation by selecting the best box sizes which lead to meaningful exon and intron signal separation.

As illustrated in Fig. 1, when we use conventional DFA methods, in which all box sizes for slope estimation are used and a non-optimal scaling region has been defined, the diagram starts from small box size (n), for which the data is fully matched by detrending step and therefore the output value, $F(n)$, is so small. On the extreme of the diagram, when n becomes large, there is some box sizes in which detrending is saturated and no meaningful increase in fluctuation happens with an increase in the box size. Thus, the slope estimation must be performed on a limited region, where most of the discrimination takes place.

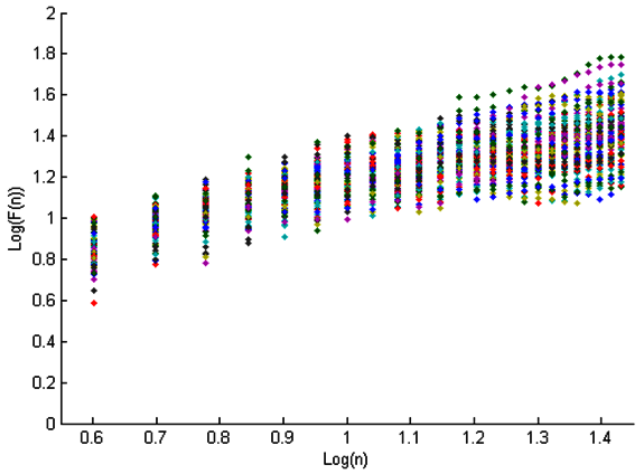

Fig. 1. The result of DFA for exon and intron signals. The scaling exponent can be approximated as the slope of log(*F*(*n*)) against log(*n*).

Fig. 2 and 3 show the effect of optimal box size in the slope estimation of log-log plot in Fig. 1. The FSE for both exon and intron is mostly greater than 0.5 which implies that there is a positive correlation in the genome signal. By adaptively checking different margins, we observed that it could be possible to choose the minimum best boxes for which the fluctuation rate statistically has the best meaningful relation to separate exon and intron instead of many boxes in the original DFA algorithm.
Fig. 2, shows the FSE for non-optimal box sizes derived from the conventional DFA algorithm in the slope estimation of log-log plot in Fig. 1. Fig. 3, shows the FSE for optimal box sizes in slope estimation of log-log plot.

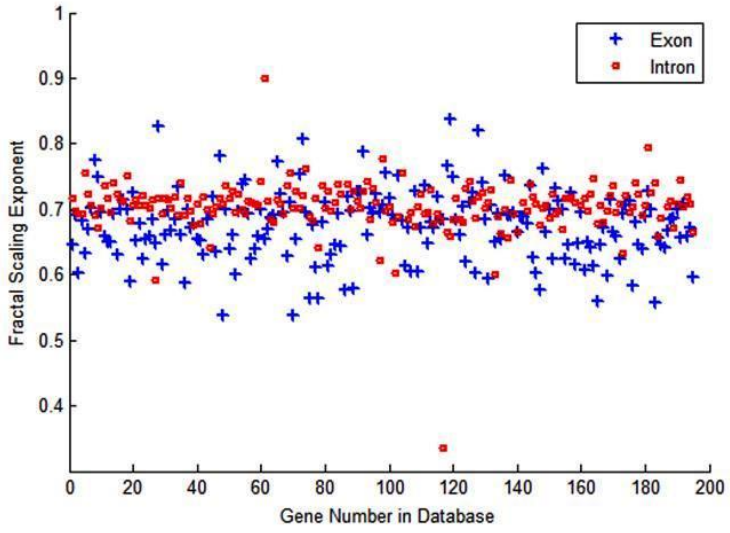

Fig. 2. FSE for non-optimal box sizes (Conventional DFA) in slope estimation of log-log plot in Fig. 1. FSE for both exon and intron is mostly greater than 0.5 which implies that there is a positive correlation in the genome signal.

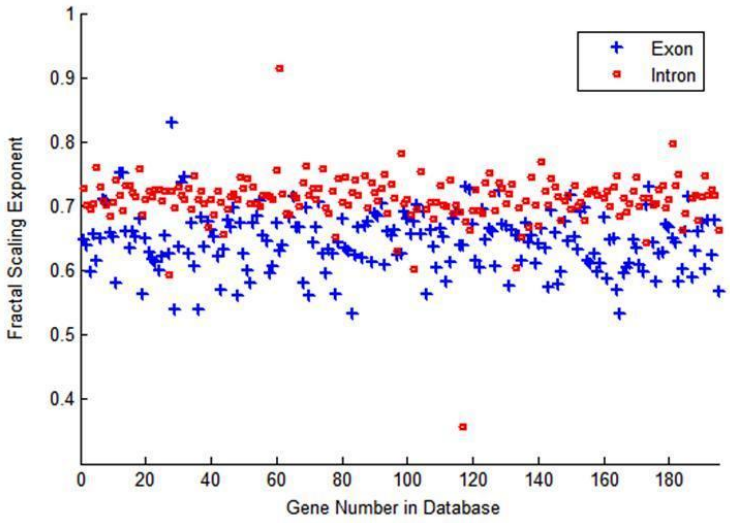

Fig. 3. FSE for optimal box sizes in slope estimation of log-log plot in Fig. 1. FSE for both exon and intron is mostly greater than 0.5 which implies that there is a positive correlation in the genome signal.

To determine whether differences in these two sets are statistically significant we use ANOVA (ANalysis Of VAriance). Since knowing the average and standard deviation is not sufficient to determine significance, the ANOVA can be used to see if a data set is significantly different from another. The results of this test revealed that FSE significantly separates the exon and intron groups and it can be seen that optimal box sizes differentiate these two groups more clearly regarding the gap between the mean FSEs of introns and exons. Table 1 summarized the result of this test for two different box size ranges.

Table 1: FSE separates the exon and intron groups and it can be seen that optimal box sizes differentiate these two groups more significantly.

DFA	Mean FSE (STD) Exon	Mean FSE (STD) Intron	Significance (*P*-Value)
Non-optimal FSE	0.6809 (0.07551)	0.7202 (0.0391)	$P < 0.01$
Optimal FSE Box sizes (4-28)	0.6446 (0.04179)	0.7231 (0.0938)	$P < 0.01$

Fractal analysis have been used to disclose long-range correlations in DNA sequences [7-9]. It has revealed complex patterns in natural objects [10-12]. For example, the genome fragments have been classified according to their fractal properties and a prokaryotic phylogenetic tree based on fractal analysis has been proposed in [13]. One of fractal analysis methods to study long-range correlations in genomes is the DFA [7, 14]. DFA is a scaling analysis method that provides a simple quantitative parameter (scaling exponent, α) to represent the correlation properties of sequence and the characteristic length scale of repetitive patterns.

The advantages of DFA over other methods are that it permits the detection of long-range correlations embedded in the apparent non-stationary signal produced by mapping of a sequence of the alphabet to numerical values. Conventional methods such as spectral analysis or root mean square fluctuation can be applied only to stationary signal. This method also avoids the counterfeit detection of long-range correlations that are artifacts of

non-stationarity in sequences and differentiates local patchiness, such as excess of one type of nucleotide in a specific region. DFA can be used for local heterogeneouse nucleotide content as well as for the entire sequence.

Moreover, DFA captures the fractal nature of DNA sequences which is not considered in traditional bias-based measures like Fickett's statistic [15]. While compositional bias is an important indicator of coding regions of DNA sequences, but it is not as specific to these regions as self-similarity is. Accordingly, FSE could provide a more reasonable metric for identifying informative regions in DNA sequences.

Methods such as Markov models have restrictions in dealing with dependencies at different scales, although they are more suitable for short-range nucleotide correlation analysis. Fast Fourier transform (FFT) method is also affected at high-frequency analysis of short-range correlations related to codon structure, whereas the signal is distorted by artifacts of the method especially at low frequencies. A meaningful relation between the self-similarity property of DNA sequence and evolution has been reported in [16] which suggested a link between long-range correlations and higher order structure of the DNA molecule [17]. Scale-independent correlations offer the best tradeoff between efficient information transfer and robustness to errors on all scales [16], whereas the information theory suggests that one can package the largest amount of information into characters of constant length when a sequence is self-similar [18].

4. CONCLUSION

Based on long-range correlation or scale-invariant property of DNA as one of the whole genome structural features, in this paper we use the self-similarity and fractal property of the numerical signal generated from a DNA sequence. FSE as a global feature has been extracted from the signal which belongs to exon and intron signals using DFA. The results imply that the FSE for both exon and intron of 195 genes in the dataset are mostly above 0.5 indicating the presence of long-range correlations and fractal nature of genome signal. More importantly the FSE of coding sequences (exon) was significantly lower than sequences that were primary noncoding (intron). The FSE of exon segments represents more variation compare to FSE of intron segments as illustrated in Fig. 2 and 4. It means that non-coding unit of a sequence has a simpler information structure in relation to coding segments (exon) which shows more complex information entropy.

Accordingly, like Fickett's statistic, FSE can be exploited in the existing exon/intron prediction algorithms like TestCode (http://rothlab.ucdavis.edu/genhelp/testcode.html). This could be the topic of next research complementing the present one. Moreover, the presented method can be used for other datasets with different target questions. Especially tracing the fractal property of the genome in evolution and among diverse species can be invaluable topics for future researches. Another issue in separating coding and non-coding regions not addressed in this research is the analysis of difference between 3' and 5' UTRs, as other non-coding segments, with other segments of the gene. This, also, can be another potential area of research.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

AUTHORS' CONTRIBUTIONS

Authors may use the following wordings for this section: " 'Payman Gifani' designed the study, performed the statistical analysis, wrote the protocol, and wrote the first draft of the manuscript. 'Ali Najafi' and 'Ali Karami' managed the analyses of the study. 'Sahand

Khakabimamaghani' managed the literature search. All authors read and approved the final manuscript."

REFERENCES

1. Tyson, J.J., K. Chen, and B. Novak, *Network dynamics and cell physiology.* Nat Rev Mol Cell Biol, 2001. **2**(12): p. 908-16.
2. Singhal, P., et al., *Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations.* Biophysical journal, 2008. **94**(11): p. 4173-4183.
3. Biran, A. and E. Meshorer, *Concise review: chromatin and genome organization in reprogramming.* Stem Cells, 2012. **30**(9): p. 1793-9.
4. Buldyrev, S.V., et al., *Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis.* Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics, 1995. **51**(5): p. 5084-91.
5. Ashida, H., K. Asai, and M. Hamada, *Shape-based alignment of genomic landscapes in multi-scale resolution.* Nucleic Acids Res, 2012. **40**(14): p. 6435-48.
6. Rogic, S., *Evaluating and improving the accuracy of computational gene-finding on mammalian DNA sequences*, 2000, The University of British Columbia.
7. Peng, C.K., et al., *Statistical properties of DNA sequences.* Physica A, 1995. **221**: p. 180-92.
8. Voss, R.F., *Evolution of long-range fractal correlations and 1/f noise in DNA base sequences.* Phys Rev Lett, 1992. **68**(25): p. 3805-3808.
9. Havlin, R.N.M., C.-K. Peng, and M. Simons, *Scale Invariant Features of Coding and Noncoding DNA Sequences.* Fractal Geometry in Biological Systems: An Analytical Approach, 1996: p. 15.
10. Berthelsen, C.L., J.A. Glazier, and M.H. Skolnick, *Global fractal dimension of human DNA sequences treated as pseudorandom walks.* Physical Review A, 1992. **45**(12): p. 8902.
11. Zu-Guo, Y., et al., *Fractals in DNA sequence analysis.* Chinese Physics, 2002. **11**(12): p. 1313.
12. de Sousa Vieira, M., *Statistics of DNA sequences: A low-frequency analysis.* Physical Review E, 1999. **60**(5): p. 5932.
13. Yu, Z.-G., et al., *The genomic tree of living organisms based on a fractal model.* Physics Letters A, 2003. **317**(3): p. 293-302.
14. Peng, C.-K., et al., *Mosaic organization of DNA nucleotides.* Physical Review E, 1994. **49**(2): p. 1685.
15. Fickett, J.W., *Recognition of protein coding regions in DNA sequences.* Nucleic Acids Res, 1982. **10**(17): p. 5303-18.
16. Voss, R.F., *Evolution of long-range fractal correlations and 1/fα noise in DNA base sequences.* Phys. Rev. Lett, 1992. **68**: p. 3805-3808.
17. Nyeo, S.-L., I.-C. Yang, and C.-H. Wu, *Spectral classification of archaeal and bacterial genomes.* Journal of Biological Systems, 2002. **10**(03): p. 233-241.
18. Nagai, N., et al., *Evolution of the periodicity and the self-similarity in DNA sequence: a Fourier transform analysis.* The Japanese journal of physiology, 2001. **51**(2): p. 159-168.

٣٥٦

٣٥٧