Vol 11 No 12, December 2002
1009-1963/2002/11(12)/1313-06

Chinese Physics

© 2002 Chin. Phys. Soc.
and IOP Publishing Ltd

Fractals in DNA sequence analysis*

Yu Zu-Guo(喻祖国)[a)b)], Vo Anh[a)],
Gong Zhi-Min(龚志民)[b)], and Long Shun-Chao(龙顺潮)[b)]

[a)] *Centre in Statistical Science and Industrial Mathematics, Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia*

[b)] *Department of Mathematics, Xiangtan University, Xiangtan 411105, China*

(Received 26 May 2002; revised manuscript received 28 July 2002)

Fractal methods have been successfully used to study many problems in physics, mathematics, engineering, finance, and even in biology. There has been an increasing interest in unravelling the mysteries of DNA; for example, how can we distinguish coding and noncoding sequences, and the problems of classification and evolution relationship of organisms are key problems in bioinformatics. Although much research has been carried out by taking into consideration the long-range correlations in DNA sequences, and the global fractal dimension has been used in these works by other people, the models and methods are somewhat rough and the results are not satisfactory. In recent years, our group has introduced a time series model (statistical point of view) and a visual representation (geometrical point of view) to DNA sequence analysis. We have also used fractal dimension, correlation dimension, the Hurst exponent and the dimension spectrum (multifractal analysis) to discuss problems in this field. In this paper, we introduce these fractal models and methods and the results of DNA sequence analysis.

Keywords: time series model, visual representation, multifractal analysis, DNA sequences
PACC: 8170, 4752

1. Introduction

Fractal methods have been successfully used to study many problems in physics, mathematics, engineering, finance, and even in biology. In the past decade or so, there has been an increasing interest in unravelling the mysteries of DNA. The heredity information of most organisms is encoded in a universal way in long chains of nucleic acids formed by four different nucleotides, i.e. adenine (A), cytosine (C), guanine (G) and thymine (T). The DNA sequence identifies a given species, distinguishing it from all other species, even those with the same nucleotide composition. The nucleotide sequence is called the primary structure, since it is the most basic level of a structure and also defines the higher level of the structure (secondary, tertiary, etc). The most basic step in the study of DNA is to determine its nucleotide sequence. After DNA is sequenced, we need to find different functional regions in it. How to gain more bioinformation from these DNA sequences is a challenging problem. The nucleotide sequences stored in GenBank have exceeded hundreds of millions of bases and they increase by ten times every five years. It has become important to improve on new theoretical methods to conduct DNA sequences analysis. Many biologists, physicists, mathematicians and computer specialists are attracted to this interesting research field. There are some important problems in DNA sequence analysis:

(i) the problem of how to distinguish the functional regions in the DNA sequence, in particular, to distinguish coding sequences from noncoding sequences is helpful to find new genes;

(ii) the search for patterns in these sequences;

(iii) problems related to the classification and evolution of organisms.[14]

A significant contribution in these many studies is the investigation of the long-range correlation in DNA sequences. One-dimensional (1D) DNA walk was first proposed by Peng *et al.*[20] They define the walk steps

*Project supported partially by the Youth Foundation of National Natural Science Foundation of China (Grant No 10101022) and Postdoctoral Research Support Grant No 9900658 of Queensland University of Technology, Australia to Yu Zu-Guo and the Foundation of Province Educational Committee of Hunan (Grant No 99803) to Long Shun-Chao.

†Permanent corresponding address of Zu-Guo Yu. Email: yuzg@hotmail.com, z.yu@qut.edu.au

"up" $u(i) = +1$ when a pyrimidine (C or T) occurs at position i along the DNA chain, while walk steps "down" $u(i) = -1$ if a purine (A or G) occurs at position i. They wish to answer the question of whether such a walk displays only short-range correlations or only long-range correlations (scale-free "fractal" phenomena). Using the exponent α from the fluctuation analysis, they found that

(i) for coding sequences, the exponent is close to 0.5 corresponding to the case when the walk is random or displays a local correlation;

(ii) for noncoding sequence, $\alpha \simeq 0.67 > 0.5$ explicitly, which corresponds to the case the walk displays a long-range correlation.

But their result has been criticized by other researchers because they did not distinguish A from G in purine and C from T in pyrimidine in their DNA walk model. Then in Refs.[3] and [4], they gave some explanations and improvements. By undertaking a more detailed analysis, Chatzidimitriou *et al*[5] concluded that both coding and noncoding sequences exhibit a long-range correlation. A subsequent work by Prabhu and Claverie[21] also substantially corroborates these results.

Luo *et al*[15,16] defined two- or three-dimensional DNA walk models which can distinguish C from T in pyrimidine, and A from G in purine. Through the average fractal dimension they found that the presence of base correlation has been found even in coding sequences and the average fractal dimension of exon is larger than that of intron.

Berthelsen *et al*[2] considered the global fractal dimension of human DNA sequences treated as pseudo-random walk. They considered the fractal dimension of this type of representation of long DNA sequences. Their conclusions are:

(i) the global fractal dimensions of coding sequences are different from those of noncoding sequences;

(ii) the DNA sequences they used are not random sequences.

Provata and Almirantis[22] regarded the coding parts of the DNA sequences as clusters of connected sites of a random Cantor-like set, with the noncoding part as the empty regions of the same set. Using the fractal dimension, they found that the coding/noncoding partition in the DNA sequences of lower organisms is homogeneous-like, while in the higher eucariotes the partition is fractal. But their method seems too rough to distinguish bacteria because the fractal dimensions of bacteria they give are all the same.

In recent years, our group has introduced a time series model (statistical point of view) and a visual representation (geometrical point of view) to DNA sequence analysis. We have also used fractal dimension, correlation dimension, the Hurst exponent and the dimension spectrum (multifractal analysis) to discuss the problems in this field. In this paper, we introduce these fractal models and methods and the results on DNA sequence analysis.

2. A time series model of DNA sequences

For a DNA sequence, we define a map f as follows:

$$f: \begin{array}{l} A \longrightarrow -2 \\ C \longrightarrow -1 \\ G \longrightarrow 1 \\ T \longrightarrow 2. \end{array} \tag{1}$$

Then we obtain a data sequence $\{x_k : k = 1, 2, \cdots, N\}$, where $x_k \in \{-2, -1, 1, 2\}$. We formally view this sequence as a time series. We expect it to reveal more information than the 1D DNA walk.[20]

2.1. Correlation dimension

The notion of correlation dimension, introduced by Grassberger and Procaccia,[7] is very suitable for experimental situations when only a single time series is available. It is now being used widely in many branches of physical science. Given a sequence of data from a computer or laboratory experiment

$$x_1,\ x_2,\ x_3, \cdots,\ x_N$$

where N is a large enough number. These numbers are usually sampled at an equal time interval $\Delta\tau$. We embed the time series into $\boldsymbol{R}^m$, choose a time delay $\tau = p\Delta\tau$, then obtain

$$y_i = (x_i, x_{i+p}, x_{i+2p}, \cdots, x_{i+(m-1)p}),$$

$$i = 1, 2, \cdots, N_m$$

where

$$N_m = N - (m-1)p.$$

In this way we obtain N_m vectors of embedding space $\boldsymbol{R}^m$.

For any y_i, y_j, we define the distance as

$$r_{ij} = \mathrm{d}(y_i, y_j) = \sum_{l=0}^{m-1} |x_{i+lp} - x_{j+lp}|.$$

If the distance is less than a given number r, we say that these two vectors are correlated. The correlation integral is defined as

$$C_m(r) = \frac{1}{N_m^2} \sum_{i,j=1}^{N_m} H_1(r - r_{ij}),$$

where H_1 is the Heaviside function

$$H_1(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0. \end{cases}$$

For a proper choice of m and a value of r which is not too large, Grassberger and Procaccia[7] have shown that the correlation integral $C_m(r)$ behaves like

$$C_m(r) \propto r^{D_2(m)}.$$

Thus, we can define correlation dimension as

$$D_2 = \lim_{m \longrightarrow \infty} D_2(m) = \lim_{m \longrightarrow \infty} \lim_{r \longrightarrow 0} \frac{\ln C_m(r)}{\ln r}.$$

2.2. Hurst exponent

To study the time series, Hurst[12] invented a statistical method—the rescaled range analysis (R/S analysis)—and later on Mandelbrot[17] and Feder [6] transplanted R/S analysis into fractal analysis. For any time series $x = \{x_k\}_{k=1}^N$ and any $2 \leq n \leq N$, we can define

$$< x >_n = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

$$X(i,n) = \sum_{u=1}^{i} [x_u - < x >_n],$$

$$R(n) = \max_{1 \leq i \leq n} X(i,n) - \min_{1 \leq i \leq n} X(i,n),$$

$$S(n) = \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - < x >_n)^2 \right]^{1/2}.$$

Hurst found that

$$R(n)/S(n) \sim \left(\frac{n}{2} \right)^H.$$

H is called the Hurst exponent.

As n changes from 2 to N, we obtain $N-1$ points in $\ln(n)$ versus $\ln(R(n)/S(n))$ plane. Then we can calculate the Hurst exponent H of the DNA sequence s using the least-squares linear fit.

In order to distinguish different function regions, with the map f given in (1), Yu and Chen[28] calculated the Hurst exponent of many functional DNA sequences and then averaged from these according to the function and species.

Conclusions: $H_{\text{exon}} > H_{\text{intron}}$ holds for the virus, but $H_{\text{intron}} > H_{\text{exon}}$ for the species which have the shape of cell except for drosophila. The latter means that exons are more complex than introns. This coincides with the results of Shen *et al.*[23] And the Hurst exponent of a DNA sequence is generally larger than $\frac{1}{2}$. This means that, when we use the fractional Brownian motion model to describe DNA sequences, we can say that it is a persistent system. In particular, we can see that H_{exon} are different from 1/2 explicitly. This indicates that the coding region of DNA is far from random. This is different from the result of Peng *et al*,[20] but coincides with the results of Luo *et al.*[15,16]

In order to discuss the classification and evolution of bacteria, Yu and Wang[30] used the average of the correlation dimensions of all coding sequences in the complete genome, and gave a rough classification of bacteria.

Viewed from the level of structure, the complete genome of an organism is made up of coding and noncoding segments. Here the length of a coding/noncoding segment means the number of its bases. Based on the lengths of coding/noncoding segments in the complete genome, we can obtain two types of integer sequences in the following ways:

(i) We order all lengths of coding segments according to the order of the coding segments in the complete genome. This integer sequence is called coding length sequence.

(ii) We order all lengths of the noncoding segments according to the order of the noncoding segments in the complete genome. This integer sequence is called the noncoding length sequence.

We have considered the above defined length sequences of the complete genomes in Ref.[24], and then viewed the length sequence as a time series. Using the correlation dimension, we have obtained some results better than that of Provata and Almirantis.[22]

The correlation property of the length sequence from DNA sequences also discussed by Yu *et al.*[27]

Conclusion: Long-range correlations must exist in most of these length sequences. The correlations have a rich variety of behaviours including the presence of anti-correlations. Furthermore, these correlations are all linear.

If we view the length sequences as a time series, through the multifractal analysis which will be introduced in section 3.2, we have:[26]

Conclusion: It is seen that there is a clear difference between the coding/noncoding length sequences of all organisms considered and a completely random sequence. The complexity of noncoding length sequences is higher than that of coding length sequences for bacteria. Almost all multifractal spectra for coding length sequences are flat, so their multifractality is small whereas almost all multifractal spectra for noncoding length sequences are multifractal-like.

3. Visual representations of DNA sequences

3.1. Two-dimensional portrait of DNA

Because there are only a few graphical illustration methods of sequence patterns (such as "Chaos game representation" proposed by Jeffrey,[13]) Hao *et al*[9] have proposed the following representation of DNA sequences.

We call any string made from K letters from the set $\{A, C, G, T\}$ a K-string. For a given K, there are 4^K different K-strings in total. In order to count the number of each type of K-string in a given DNA sequence, 4^K counters are needed. These counters may be arranged in a $2^K \times 2^K$ square. For a given K, we call this $2^K \times 2^K$ square a K-frame. Then we can put all K strings appearing in a complete genome into a K-frame and use the discrete colour code to represent the number of times of a K-string appears in the complete genome. A complete genome represented in a K-frame is called a portrait of the organism. Hao *et al*[9] found some fractal-like phenomena (self-similarity) which are introduced by the avoiding and under-represented strings.

For the alphabet $V = \{A, C, G, T\}$, and an avoiding string w_0, we represent all the K-strings which do not include w_0 as a factor into the K-frame, then we let K tend to infinity. Thus, we can obtain a fractal set F.

Problem: How to calculate the box dimension and Hausdorff dimension of F, and whether the box dimension equals the Hausdorff dimension?

We denote a_K as the number of K-strings in V that does not include w_0 as a factor, then we define the generating function of a_K as

$$f(t) = \sum_{K=0}^{\infty} a_K t^K,$$

where t is a complex variable. We can use Golden–Jackson cluster method[19] to obtain the expression of $f(t)$. Then Yu *et al*[10,29] obtain

Theorem 1 The Box dimension of F is given by

$$\dim_{\mathrm{B}}(F) = -\frac{\log |t_0|}{\log 2},$$

where t_0 is the minimal module zero point of $1/f(t)$, and $f(t)$ is the generating function of a_K.

With the help of the result of Mauldin and Williams,[18] we have

Theorem 2

$$\dim_H(F) = \dim_B(F) = \alpha$$
$$\text{and} \quad 0 < \mathcal{H}^{\alpha}(F) < \infty,$$

where $\dim_H(F)$ is the Hausdorff dimension of F.

3.2. 1D measure representation

In the two-dimensional portrait representation, it is difficult to know the accurate value of the frequencies of the substrings from the portrait representation. In order to improve it, Yu and Anh[25] used subintervals in one-dimensional space to represent substrings, thus we can directly obtain an accurate histogram of the substrings in the complete genome. We then view the histogram as a measure, which we call the measure representation of the complete genome. The rigorous definition is in the following.

We divide the interval $[0, 1)$ into 4^K disjoint subintervals, and we use each subinterval to represent a counter. Letting $s = s_1 \cdots s_K, s_i \in \{A, C, G, T\}, i = 1, \cdots, K$, be a substring with length K, we define

$$x_l(s) = \sum_{i=1}^{K} \frac{x_i}{4^i},$$

where

$$x_i = \begin{cases} 0, & \text{if } s_i = A, \\ 1, & \text{if } s_i = C, \\ 2, & \text{if } s_i = G, \\ 3, & \text{if } s_i = T, \end{cases}$$

and

$$x_r(s) = x_l(s) + \frac{1}{4^K}.$$

We then use the subinterval $[x_l(s), x_r(s))$ to represent the substring s. Let $N_K(s)$ be the number of times that substring s with length K appears in the complete genome. If the number of bases in the complete genome is L, we define

$$F_K(s) = N_K(s)/(L - K + 1)$$

as the frequency of substring s. It follows that $\sum_{\{s\}} F_K(s) = 1$. Now we can define a measure μ_K on $[0, 1)$ by $\mathrm{d}\mu_K(x) = Y_K(x)\mathrm{d}x$, where

$$Y_K(x) = F_K(s) \times 4^K, \quad \text{for} \quad x \in [x_l(s), x_r(s)).$$

We call μ_K the measure representation of the organism corresponding to the given K.

The most common operative numerical implementations of multifractal analysis are the so-called fixed-size box-counting algorithms.[8,11] In the one-dimensional case, for a given measure μ with a support $E \subset \boldsymbol{R}$, we consider the partition sum

$$Z_\epsilon(q) = \sum_{\mu(B)\neq 0} [\mu(B)]^q,$$

$q \in \boldsymbol{R}$, where the sum runs over all different non-empty boxes B of a given side ϵ in a grid covering of the support E, i.e.,

$$B = [k\epsilon, (k+1)\epsilon).$$

The exponent $\tau(q)$ is defined as

$$\tau(q) = \lim_{\epsilon\to 0} \frac{\log Z_\epsilon(q)}{\log \epsilon},$$

and the generalized fractal dimensions of the measure are defined as

$$D_q = \tau(q)/(q-1), \quad \text{for } q \neq 1,$$

and

$$D_q = \lim_{\epsilon\to 0} \frac{\sum_{\mu(B)\neq 0} \mu(B)\log\mu(B)}{\log\epsilon}, \quad \text{for } q = 1.$$

The generalized fractal dimensions are numerically estimated through a linear regression of

$$\frac{1}{q-1}\log Z_\epsilon(q)$$

against $\log\epsilon$ for $q \neq 1$, and similarly through a linear regression of

$$\sum_{\mu(B)\neq 0} \mu(B)\log\mu(B)$$

against $\log\epsilon$ for $q = 1$.

We calculated the D_q spectrum of all complete genomes available in GenBank.

Conclusion: When these measure representations, considered as time series, they exhibit a strong long-range correlation. For substrings with length $K = 8$, the D_q spectra of all organisms studied are multifractal-like and sufficiently smooth. With decreasing value of K, the multifractality lessens.

After we have the dimension spectra (D_q curves) of the measure representation of an organism, a question we may ask is

Question: What is the theoretical expression of this D_q curve?

Let $\varepsilon(t)$ be a positive stationary stochastic process on a bounded interval of $\boldsymbol{R}$, assumed to be the unit interval $[0, 1]$ for convenience, with $E\varepsilon(t) = 1$. For $j = 0, 1, 2, 3, \ldots$ we define

$$X_j(t) = \varepsilon_{r_j}(t) = \frac{1}{2^{-j+1}} \int_{t-2^{-j}}^{t+2^{-j}} \varepsilon(s)\,\mathrm{d}s. \qquad (2)$$

Then

$$K(q) = \lim_{J\to\infty} \frac{\ln E(X_J^q)}{-\ln 2^{-J+1}}, \qquad (3)$$

is usually used as another multifractal characterization in the multiplicative cascade generated by an infinite divisible distribution. Anh *et al.*[1] proposed a multiplicative cascade model to solve the above question. In that model, we have obtained the theoretical expression of the K_q curve

$$K(q) = \begin{cases} \kappa\left(q - \frac{(q\sigma+1)^{1-\alpha}-1}{(\sigma+1)^{1-\alpha}-1}\right), & \alpha \neq 1, \\ \kappa\left(q - \frac{\ln(q\sigma+1)}{\ln(\sigma+1)}\right), & \alpha = 1, \end{cases} \qquad (4)$$

where κ, A, α, σ are positive constants obtained from the infinitely divisible distribution. We can estimate their values from the numerical value of K_q calculated using Eq.(3). We found the expression of K_q fit the numerical value of K_q of the measure representation of the complete genome very well.[1] We have also found that the relationship $K(q) = -D(q)(q-1) + q - 1$ holds for the measure representation of the complete genome.[1] Therefore, from Eqs.(4) we can obtain the theoretical expression of the D_q curve.

Acknowledgments

One of the authors, Yu Zu-guo, would like to express his thanks to Prof. Hao Bai-lin of Institute of Theoretical Physics of Chinese Academy of Science for introducing him to this field, and giving useful discussions and continuous encouragement. Thanks are also due to Dr Cheng Guo-yi and Dr Wang Bin for a useful discussion.

References

[1] Anh V, Lau K S and Yu Z G 2001 *J. Phys. A: Math. Gen.* **34** 7127
[2] Berthelsen C L, Glazier J A and Skolnick M H 1992 *Phys. Rev.* A **45** 8902
[3] Buldyrev S V, Dokholyan N V, Goldberger A L, Havlin S, Peng C K, Stanley H E and Visvanathan G M 1998 *Physica* A **249** 430
[4] Buldyrev S V, Goldgerger A I, Havlin S, Peng C K and Stanley H E 1994 *Fractals in Science*, ed A. Bunde and S. Havlin (Berlin: Springer-Verlag) pp 49–87
[5] Chatzidimitriou-Dreismann C A and Larhammar D 1993 *Nature* **361** 212
[6] Feder J 1988 *Fractals* (New York: Plenum)
[7] Grassberger P and Procaccia I 1983 *Phys. Rev. Lett.* **50** 346
[8] Halsy T, Jensen M, Kadanoff L, Procaccia I and Schraiman B 1986 *Phys. Rev.* A **33** 1141
[9] Hao B L, Lee H C and Zhang S Y 2000 *Chaos, Solitons and Fractals* **11** 825
[10] Hao B L, Xie H M, Yu Z G and Chen G Y 2000 *Ann. Comb.* **4** 247
[11] Hentschel HGE and Procaccia I 1983 *Physica* D **8** 435
[12] Hurst H E 1951 *Trans. Am. Soc. Civil Eng.* **116** 770
[13] Jeffrey H J 1990 *Nucletic Acids Res.* **18** 2163
[14] Li W H and Graur D 1991 *Fundamentals of Molecular Evolution* (Sunderland, MA: Sinauer Associates)
[15] Luo L F, Lee W J, Jia L J, Ji F M and Tsai L 1998 *Phys. Rev.* E **58** 861
[16] Luo L F and Tsai L 1988 *Chin. Phys. Lett.* **5** 421
[17] Mandelbrot B B 1982 *The Fractal Geometry of Nature* (Oxford: Freeman)
[18] Mauldin R D and Williams S C 1988 *Tran. Am. Math. Soc.* **309** 811
[19] Noonan J and Zeilberger D, The Goulden-Jackson cluster method: extensions, applications and implementations, downloadable from http://www.math.temple.edu/zeilberg
[20] Peng C K, Buldyrev S, Goldberg A L, Havlin S, Sciortino F, Simons M and Stanley H E 1992 Long-range correlations in nucleotide sequences. *Nature* **356** 168
[21] Prabhu V V and Claverie J M 1992 *Nature* **359** 782
[22] Provata A and Almirantis Y 2000 *Fractals* **8** 15
[23] Shen R Q, Chen R S, Lin L J, Sun J, Xiao Y and Xu J 1993 *Chin. Sci. Bull.* **38** 1995 (in Chinese)
[24] Yu Z G and Anh V V 2001 *Chaos Solitons Fractals* **12** 1827
[25] Yu Z G, Anh V V and Lau K S 2001 *Phys. Rev.* E **64** 31903
[26] Yu Z G, Anh V V and Lau K S 2001 *Physica* A **301** 351
[27] Yu Z G, Anh V V and Wang B 2001 *Phys. Rev.* E **63** 11903
[28] Yu Z G and Chen G Y 2000 *Commun. Theor. Phys.* **33** 673
[29] Yu Z G, Hao B L, Xie H M and Chen G Y 2000 *Chaos Solitons Fractals* **11** 2215
[30] Yu Z G and Wang B 2001 *Chaos Solitons Fractals* **12** 519