Multifractal characterization of complete genomes

2001 J. Phys. A: Math. Gen. 34 7127

(http://iopscience.iop.org/0305-4470/34/36/301)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **34** (2001) 7127–7139 PII: S0305-4470(01)20618-0

Multifractal characterization of complete genomes

Vo Anh[1,4], **Ka-Sing Lau**[2] **and Zu-Guo Yu**[1,3]

[1] Centre in Statistical Science and Industrial Mathematics, Queensland University of Technology, GPO Box 2434, Brisbane, Q4001, Australia
[2] Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong, People's Republic of China
[3] Department of Mathematics, Xiangtan University, Hunan 411105, People's Republic of China

E-mail: v.anh@qut.edu.au, kslau@math.cuhk.edu.hk, yuzg@hotmail.com and z.yu@qut.edu.au

Received 3 January 2001, in final form 18 June 2001
Published 31 August 2001
Online at stacks.iop.org/JPhysA/34/7127

Abstract

This paper develops a theory for characterization of DNA sequences based on their measure representation. The measures are shown to be random cascades generated by an infinitely divisible distribution. This probability distribution is uniquely determined by the exponent function in the multifractal theory of random cascades. Curve fitting to a large number of complete genomes of bacteria indicates that the gamma density function provides an excellent fit to the exponent function, and hence to the probability distribution of the complete genomes.

PACS numbers: 05.45.Df, 05.10.Gg, 87.14.Gg

1. Introduction

DNA sequences are of fundamental importance in understanding living organisms, since all information on their hereditary evolution is contained in these macromolecules. One of the challenges of DNA sequence analysis is to determine the patterns of these sequences. It is useful to distinguish coding from noncoding sequences. Problems related to the classification and evolution of organisms using DNA sequences are also important.

Fractal analysis has proved useful in revealing complex patterns in natural objects. Berthelsen *et al* [2] considered the global fractal dimension of human DNA sequences treated as pseudorandom walks. Vieira [3] carried out a low-frequency analysis of the complete DNA of 13 microbial genomes and showed that their fractal behaviour does not always prevail through the entire chain and the autocorrelation functions have a rich variety of behaviours including

[4] http://www.maths.qut.edu.au/cissaim/anh.html

the presence of anti-persistence. Provata and Almirantis [14] proposed a fractal Cantor pattern of DNA. They mapped coding segments to filled regions and noncoding segments to empty regions of a random Cantor set and then calculated the fractal dimension of this set. They found that the coding/noncoding partition in DNA sequences of lower organisms is homogeneous-like, while in the higher eucariotes the partition is fractal. Yu and Anh [16] proposed a time series model based on the global structure of the complete genome and found that one can get more information from this model than that of the fractal Cantor pattern. Some results on the classification and evolution relationship of bacteria were found in [16]. The correlation property of length sequences was discussed in [17].

Although statistical analysis performed directly on DNA sequences has yielded some success, there has been some indication that this method is not powerful enough to amplify the difference between a DNA sequence and a random sequence as well as to distinguish DNA sequences themselves in more details. One needs more powerful global and visual methods. For this purpose, Hao *et al* [7] proposed a visualization method based on coarse-graining and counting of the frequency of appearance and strings of a given length. They called it the *portrait* of an organism. They found that there exist some fractal patterns in the portraits which are induced by avoiding and under-represented strings. The fractal dimension of the limit set of portraits was discussed in [8, 18]. There are other graphical methods of sequence patterns, such as the chaos game representation (see [5, 10]).

In the portrait representation, Hao *et al* [7] used squares to represent substrings and discrete colour grades to represent the frequencies of the substrings in the complete genome. It is difficult to know the accurate value of the frequencies of the substrings from the portrait representation. And they did not discuss the classification and evolution problem. In order to improve it, Yu *et al* [15] used subintervals in one-dimensional space to represent substrings to obtain an accurate histogram of the substrings in the complete genome. The histogram, viewed as a probability measure and was called the *measure representation* of the complete genome, gives a precise compression of the genome. Multifractal analysis was then proposed in Yu *et al* [15] to treat the classification and evolution problem based on the measure representation of different organisms.

In this paper, we go one step further and provide a characterization of the DNA sequences based on their measure representation. This is given in the form of the probability density function of the measure. We first show that the given measure is in fact a multiplicative cascade generated by an infinitely divisible distribution. This probability distribution is uniquely determined by the exponent $K(q)$, $q \geqslant 0$, in the multifractal analysis of the cascade. This theory will be detailed in the next section. We then apply the theory on a large number of typical genomes. It will be seen that the gamma density function provides an excellent fit to the $K(q)$ curve of each genome. This characterization therefore provides a needed tool to study the evolution of organisms.

2. Measure representation of complete genome

We first outline the method of Yu *et al* [15] in deriving the measure representation of a DNA sequence. Such a sequence is formed by four different nucleotides, namely adenine (a), cytosine (c), guanine (g) and thymine (t). We call any string made of K letters from the set {g, c, a, t} a K-string. For a given K there are in total 4^K different K-strings. In order to count the number of each kind of K-strings in a given DNA sequence, 4^K counters are needed. We divide the interval [0, 1[into 4^K disjoint subintervals, and use each subinterval to represent a

counter. Letting $s = s_1 \dots s_K$, $s_i \in \{\mathrm{a}, \mathrm{c}, \mathrm{g}, \mathrm{t}\}$, $i = 1, \dots, K$, be a substring with length K, we define

$$x_l(s) = \sum_{i=1}^{K} \frac{x_i}{4^i} \tag{2.1}$$

where

$$x_i = \begin{cases} 0 & \text{if} \quad s_i = \mathrm{a} \\ 1 & \text{if} \quad s_i = \mathrm{c} \\ 2 & \text{if} \quad s_i = \mathrm{g} \\ 3 & \text{if} \quad s_i = \mathrm{t} \end{cases} \tag{2.2}$$

and

$$x_r(s) = x_l(s) + \frac{1}{4^K}. \tag{2.3}$$

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

$$F(s) = \frac{N(s)}{(L - K + 1)} \tag{2.4}$$

to be the frequency of substring s. It follows that $\sum_{\{s\}} F(s) = 1$. Now we can define a measure μ_K on $[0, 1)$ by

$$\mu_K(\mathrm{d}x) = Y_K(x)\,\mathrm{d}x$$

where

$$Y_K(x) = 4^K F_K(s) \qquad x \in [x_l(s), x_r(s)). \tag{2.5}$$

We then have $\mu_K([0, 1)) = 1$ and $\mu_K([x_l(s), x_r(s))) = F_K(s)$. We call $\mu_K(x)$ the *measure representation* of an organism. As an example, the measure representation of *M. genitalium* for $K = 3, \dots, 8$ is given in figure 1. Self-similarity is apparent in the measures.

More than 33 bacterial complete genomes are now available in public databases. There are six Archaebacteria (*Archaeoglobus fulgidus*, *Pyrococcus abyssi*, *Methanococcus jannaschii*, *Pyrococcus horikoshii*, *Aeropyrum pernix* and *Methanobacterium thermoautotrophicum)*; five Gram-positive Eubacteria (*Mycobacterium tuberculosis*, *Mycoplasma pneumoniae*, *Mycoplasma genitalium*, *Ureaplasma urealyticum*, and *Bacillus subtilis)*. The others are Gram-negative Eubacteria, which consist of two Hyperthermophilic bacteria (*Aquifex aeolicus* and *Thermotoga maritima)*; five Chlamydia (*Chlamydia trachomatisserovar*, *Chlamydia muridarum*, *Chlamydia pneumoniae* and *Chlamydia pneumoniae AR39)*; two Spirochaete (*Borrelia burgdorferi* and *Treponema pallidum)*; one Cyanobacterium (*Synechocystis sp. PCC6803)*; and 13 Proteobacteria. The 13 Proteobacteria are divided into four subdivisions, which are the alpha subdivision (*Rhizobium sp. NGR234* and *Rickettsia prowazekii)*; gamma subdivision (*Escherichia coli*, *Haemophilus influenzae*, *Xylella fastidiosa*, *Vibrio cholerae*, *Pseudomonas aeruginosa* and *Buchnera sp. APS)*; beta subdivision (*Neisseria meningitidis MC58* and *Neisseria meningitidis Z2491)*; epsilon subdivision (*Helicobacter pylori J99*, *Helicobacter pylori 26695* and *Campylobacter jejuni)*. The complete sequences of some chromosomes of higher organisms are also currently available. We selected the sequences of chromosome 15 of *Saccharomyces cerevisiae*, chromosome 3 of *Plasmodium falciparum*, chromosome 1 of *Caenorhabditis elegans*, chromosome 2 of *Arabidopsis thaliana* and chromosome 22 of *Homo sapiens*.

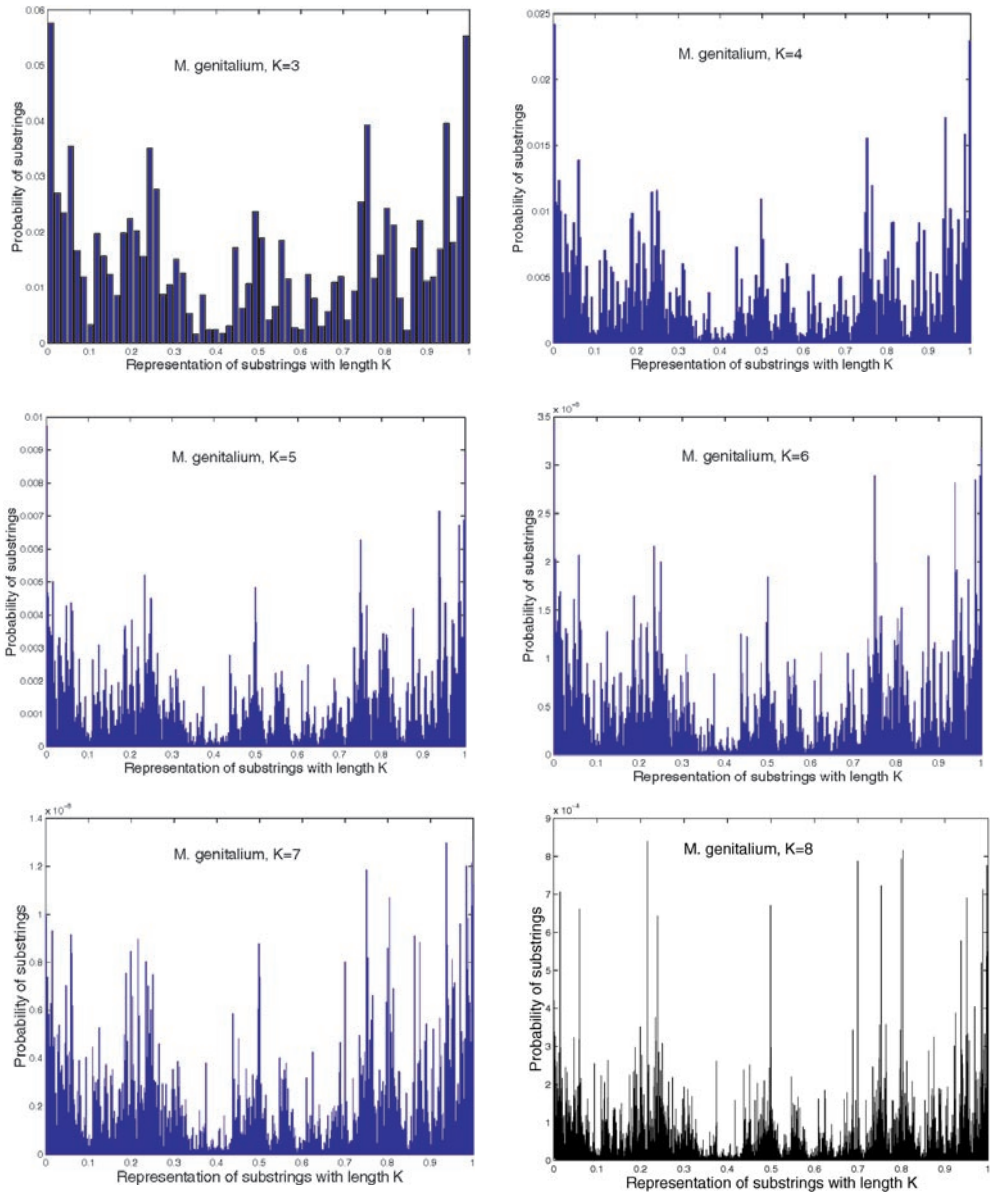

Figure 1. Histograms of substrings with different lengths.

In our previous work [15], we calculated the numerical dimension spectra D_q (defined in next section) for all above organisms and for different K. For small K, there are only a few different K-strings, so there is not enough information for any clear-cut result. We find that the D_q curves are very close to one another for $K = 6, 7, 8$ for each organism. Hence it would be appropriate to take $K = 8$ if we want to use the D_q curves to discuss the classification and evolution problem. It is still needed to know what is the analytical expression of the dimension spectra. The main aim of this paper is to establish a theoretical model to give such an analytical expression.

3. Multifractal models

Let $\varepsilon(t)$ be a positive stationary stochastic process on a bounded interval of $\mathbb{R}$, assumed to be the unit interval [0, 1] for convenience, with $E\varepsilon(t) = 1$. The smoothing of $\varepsilon(t)$ at scale $r > 0$ is defined as

$$\varepsilon_r(t) = \frac{1}{r}\int_{t-r/2}^{t+r/2} \varepsilon(s)\,\mathrm{d}s. \tag{3.1}$$

For $0 < r < u < v$, we consider the processes

$$X_{r,v}(t) = \frac{\varepsilon_r(t)}{\varepsilon_v(t)} \qquad t \in [0, 1].$$

Following Novikov [13], we assume the following scale invariance conditions:

(i) The random variables $X_{r,u}$ and $X_{u,v}$ are independent.
(ii) The probability distribution of each random variable $X_{u,v}$ depends only on the ratio u/v of the corresponding scales.

These conditions imply the power-law form for the moments of the processes $X_{u,v}$ if they exist. In fact, we may write

$$E(X_{u,v}(t))^q = g_q\left(\frac{u}{v}\right) \qquad q \geqslant 0 \tag{3.2}$$

from condition (ii) for some function g which also depends on q. From the identity

$$X_{r,v}(t) = X_{r,u}(t)X_{u,v}(t)$$

and condition (i) we get

$$g_q\left(\frac{r}{v}\right) = g_q\left(\frac{r}{u}\right) g_q\left(\frac{u}{v}\right). \tag{3.3}$$

Since u is arbitrary, we then have

$$g_q\left(\frac{r}{v}\right) = \left(\frac{r}{v}\right)^{-K(q)} \tag{3.4}$$

for some function $K(q)$ with $K(0) = 0$. It follows that

$$K(q) = \frac{\ln E(X_{r,v}(t))^q}{\ln(v/r)}.$$

Writing Y for $X_{r,v}$ we obtain

$$K'(q) = \frac{1}{\ln(v/r)}\frac{E(Y^q \ln Y)}{E(Y^q)}$$

$$K''(q) = \frac{1}{\ln(v/r)}\frac{(EY^q)E(Y^q(\ln Y)^2) - (E(Y^q \ln Y))^2}{(EY^q)^2}.$$

Since

$$\begin{aligned}(E(Y^q \ln Y))^2 &= (E(Y^{q/2}Y^{q/2} \ln Y))^2 \\ &\leqslant (EY^q)E(Y^q(\ln Y)^2)\end{aligned} \tag{3.5}$$

by Schwarz's inequality and $v/r > 1$, we get $K''(q) \geqslant 0$, that is, $K(q)$ is a convex function. It is noted that equality holds in (3.5) only if $K(q)$ is a linear function of q; other than this, $K(q)$ is a strictly convex function.

For $0 < q < 1$, we assume that $K(q) < 0$, which reflects the fact that, in this range, taking a qth-power necessarily reduces the singularity of $X_{u,v}$. Also, we assume that the probability

density function of $X_{u,v}$ is skewed in the positive direction. This yields that $K(q) > 0$ for $q > 1$. These assumptions, in conjunction with the strict convexity of $K(q)$, suggest the assumption that

$$K(1) = 0. \tag{3.6}$$

This implies that

$$EX_{u,v} = 1 \qquad \text{for arbitrary} \quad 0 < u < v. \tag{3.7}$$

In this paper, we will consider smoothing at discrete scales $r_j = 2^{-j+1}$, $j = 0, 1, 2, 3, \ldots$. Then the smoothed process at scale r_j is

$$X_j(t) = \varepsilon_{r_j}(t) = \frac{1}{2^{-j+1}} \int_{t-2^{-j}}^{t+2^{-j}} \varepsilon(s)\,\mathrm{d}s. \tag{3.8}$$

Under the condition $E\varepsilon(t) = 1$, it is reasonable to assume that

$$X_0(t) = 1 \qquad t \in [0, 1]. \tag{3.9}$$

Then, at generation J,

$$\begin{aligned} X_J(t) &= X_0(t)\frac{X_1(t)}{X_0(t)}\frac{X_2(t)}{X_1(t)} \cdots \frac{X_J(t)}{X_{J-1}(t)} \\ &= \frac{X_1(t)}{X_0(t)}\frac{X_2(t)}{X_1(t)} \cdots \frac{X_J(t)}{X_{J-1}(t)}. \end{aligned} \tag{3.10}$$

Under the scale invariance conditions (i) and (ii), the random variables X_j/X_{j-1} of (3.10) are independent and have the same probability distribution. Let W denote a generic member of this family. Note that $EW = 1$ from (3.7). Then (3.10) can be rewritten as

$$\begin{aligned} X_J(t) &= X_{J-1}(t)\frac{X_J(t)}{X_{J-1}(t)} \\ &= W_1(t)W_2(t)\cdots W_J(t) \qquad t \in [0, 1]. \end{aligned} \tag{3.11}$$

In other words, $X_J(t)$ is a multiplicative cascade process (see [6, 9]). Denote by μ_J the sequence of random measures defined by the density $X_J(t)$, that is,

$$\mu_J(\mathrm{d}t) = X_J(t)\,\mathrm{d}t \qquad J = 1, 2, 3, \ldots.$$

It can be checked that μ_J a.s. has a weak* limit μ_∞ since for each bounded continuous function f on $[0, 1]$, the sequence $\int_{[0,1]} f\,\mathrm{d}\mu_J$ is an L_1 -bounded martingale (see [9, 11, 12]). We denote the density corresponding to μ_∞ by $X_\infty(t)$. Then it is seen from (3.8) that

$$X_\infty(t) = \varepsilon(t) \qquad t \in [0, 1]. \tag{3.12}$$

Summarizing, we have established that

The positive stationary process $\varepsilon(t)$ is the limit of a multiplicative cascade with generator W.

We next want to characterize this random cascade. We first note that, for $j = 1, 2, 3, \ldots$,

$$\frac{X_j}{X_{j-1}} = 2\frac{\int_{t-2^{-j}}^{t+2^{-j}} \varepsilon(s)\,\mathrm{d}s}{\int_{t-2^{-(j-1)}}^{t+2^{-(j-1)}} \varepsilon(s)\,\mathrm{d}s} \leqslant 2 \tag{3.13}$$

from the positivity of $\varepsilon(t)$. Thus,

$$E\left(\frac{X_j}{X_{j-1}}\right)^q \leqslant 2^q.$$

This inequality together with (3.4) imply

$$K(q) \leqslant q \qquad q \geqslant 0. \tag{3.14}$$

We then have

$$\sum_{q=0}^{\infty} \left(E\left(\frac{X_j}{X_{j-1}} \right)^{2q} \right)^{-\frac{1}{2q}} = \sum_{q=0}^{\infty} \left(\frac{1}{2} \right)^{\frac{K(2q)}{2q}} = \infty.$$

In other words, the Carleman condition is satisfied (see [4], p 224). As a result, we get

The probability density function f_W of the generator W is uniquely determined by the set $\{K(q),\ q = 0, 1, 2, \ldots\}$.

It is seen that, if the function $K(q)$ has analytic continuation into the complex plane, then the characteristic function of $\ln W$ has the form

$$\psi(x) = E(\mathrm{e}^{\mathrm{i}x \ln W}) = (\tfrac{1}{2})^{-K(\mathrm{i}x)}. \tag{3.15}$$

Define $\psi_n(x) = (1/2^{1/n})^{-K(\mathrm{i}x)}$ for an arbitrary integer n. Then ψ_n is the characteristic function of the probability distribution corresponding to smoothing with scales $(2^{1/n})^{-j+1}$. Also, it holds that

$$\psi(x) = (\psi_n(x))^n.$$

Thus $\psi(x)$ is infinitely divisible (see [4], p 532); in other words,

$$\ln W \textit{ has an infinitely divisible distribution.} \tag{3.16}$$

It is noted from (3.13) that $-\ln \frac{W}{2} \geqslant 0$. The most general form for the characteristic function $\varphi(x)$ of positive random variables is given by

$$\varphi(x) = \exp\left\{ \int_0^{\infty} \frac{1 - \mathrm{e}^{\mathrm{i}xs}}{s} P(\mathrm{d}s) + \mathrm{i}ax \right\} \tag{3.17}$$

where $a \geqslant 0$ and P is a measure on the open interval $(0, \infty)$ such that $\int_0^{\infty} (1+s)^{-1} P(\mathrm{d}s) < \infty$ (see [4], p 539). On the other hand, it follows from (3.2) and (3.4) that the characteristic function of $-\ln \frac{W}{2}$ is given by

$$\begin{aligned} E(\mathrm{e}^{-\mathrm{i}x \ln \frac{W}{2}}) &= 2^{\mathrm{i}x} E(W)^{-\mathrm{i}x} \\ &= 2^{\mathrm{i}x} (\tfrac{1}{2})^{-K(-\mathrm{i}x)}. \end{aligned} \tag{3.18}$$

Using $q = -\mathrm{i}x$ and equating (3.17) with (3.18) then yields

$$K(q) = \left(1 - \frac{a}{\ln 2}\right) q - \int_0^{\infty} \frac{1 - \mathrm{e}^{-qs}}{s} \frac{P(\mathrm{d}s)}{\ln 2}. \tag{3.19}$$

As constrained by (3.6), the following condition must be satisfied by the measure $P(\mathrm{d}s)$:

$$\int_0^{\infty} \frac{1 - \mathrm{e}^{-s}}{s} \frac{P(\mathrm{d}s)}{\ln 2} = 1 - \frac{a}{\ln 2} \leqslant 1. \tag{3.20}$$

Equations (3.19) and (3.20) provide the most general form for the $K(q)$ curve of the positive random process $\{\varepsilon(t),\ 0 \leqslant t \leqslant 1\}$.

In practice, fitting this $K(q)$ curve to data requires a proper choice of the measure $P(\mathrm{d}s)$. Novikov [13] suggests the use of the gamma density function, namely,

$$f(x) = Ax^{\alpha - 1} \exp(-x/\sigma) \tag{3.21}$$

where $P(\mathrm{d}x) = f(x)\,\mathrm{d}x$ and A, α, σ are positive constants. From (3.19) and (3.21) we get

$$K(q) = \begin{cases} \kappa\left(q - \dfrac{(q\sigma+1)^{1-\alpha} - 1}{(\sigma+1)^{1-\alpha} - 1}\right) & \alpha \neq 1 \\ \kappa\left(q - \dfrac{\ln(q\sigma+1)}{\ln(\sigma+1)}\right) & \alpha = 1 \end{cases} \tag{3.22}$$

where $\kappa = 1 - a/\ln 2$, and from (3.20) we have

$$A = \frac{\kappa \ln 2}{\sigma^{\alpha-1}\Gamma(\alpha-1)}(1 - (\sigma+1)^{1-\alpha})^{-1}.$$

The form (3.22) will be used for data fitting in this paper. It is seen from (3.2) and (3.4) that the data for the $K(q)$ curve is provided by

$$K(q) = \lim_{J\to\infty} \frac{\ln E(X_J^q)}{-\ln 2^{-J+1}} \tag{3.23}$$

where it should be noted from (3.12) that $X_\infty(t) = \varepsilon(t)$, the given positive random process.

Since each smoothed process X_J may possess long-range dependence (see [1]), the ergodic theorem may not hold for these processes. As a result, the computation of $E(X_J^q)$ as sample averages may not be sufficiently accurate. There is an alternative form of the ergodic theorem developed by Holley and Waymire [9] for random cascades which we now summarize.

For random cascades with density $\varepsilon(t)$, limit measure μ_∞, branching number b and generator W, define

$$M_J(q) = {\sum_k}'(\mu_\infty(\Delta_k^J))^q \tag{3.24}$$

$$\tau(q) = \lim_{J\to\infty} \frac{\ln M_J(q)}{J \ln b} \tag{3.25}$$

$$D_q = \tau(q)/(q-1) \tag{3.26}$$

$$\chi_b(q) = \log_b E(W^q) - (q-1) \tag{3.27}$$

where the prime in (3.24) indicates a sum over those subintervals Δ_k^J of generation J which meet the support of μ_∞.

Theorem 1 [9]. *Assume that $W > a$ for some $a > 0$ and $W < b$ with probability 1, and that $E(W^{2q})/(EW^q)^2 < b$. Then, with probability 1,*

$$\tau(q) = -\chi_b(q). \tag{3.28}$$

In our case as developed above, $b = 2$, and (3.13) gives $W \leqslant 2$. In fact the scale $r_j = 2^{-j+1}$ used in (3.8) is arbitrary; it can be b^{-j+1} and the inequality $W \leqslant b$ still holds by definition of the smoothing and the positivity of $\varepsilon(t)$. In our development,

$$\begin{aligned} -K(q) &= \lim_{J\to\infty} \frac{\ln E(X_J^q)}{\ln 2^{-J+1}} \\ &= \lim_{J\to\infty} \frac{J \ln E(W^q)}{(J-1)\ln 2^{-1}} \text{ using (3.11)} \\ &= -\frac{\ln E(W^q)}{\ln 2}. \end{aligned}$$

Consequently,

$$K(q) = -\tau(q) + q - 1. \tag{3.29}$$

The above formula then provides a way to compute $K(q)$ via (3.25) and (3.29) using sums of qth powers of the limit measure instead of (3.23) using expectations. In fact, the ergodic theorem now takes the following form:

$$\lim_{J\to\infty}\frac{\ln E(X_J^q)}{(J-1)\ln 2}=\lim_{J\to\infty}\frac{\ln\sum_k'(\mu_\infty(\Delta_k^J))^q}{J\ln 2}+q-1.$$

4. Data fitting and discussion

For $K=8$, we first calculated $K(q)$ of the measure representation of all the above organisms directly from the definition of $K(q)$ (3.23). Figure 2 shows how to calculate this $K(q)$ curve. We give the $K(q)$ curves of *E. coli, S. cerevisiae Chr15, C. elegans Chr1, A. thaliana Chr2,* and *Homo sapiens Chr22* in figure 3. From figure 3, it is seen that the grade of the organism is lower when the K_q curve is flatter. Hence the evolution relationship of these organisms is apparent. We denote by $K_d(q)$ the value of $K(q)$ computed from the data using its definition (3.23) and define

$$\text{error}=\sum_{j=1}^{J}\left|\kappa\left(q_j-\frac{(q_j\sigma+1)^{1-\alpha}-1}{(\sigma+1)^{1-\alpha}-1}\right)-K_d(q_j)\right|^2.$$

Then the values of κ, σ and α can be estimated through minimizing error. In this minimization, we assume

$$0\leqslant\kappa,\sigma,\alpha\leqslant 20.$$

After obtaining the value of κ,σ and α, we then get the $K(q)$ curve from (3.22). The data fitting based on the form (3.22) was performed on all the organisms and shown in table 1 (from top to

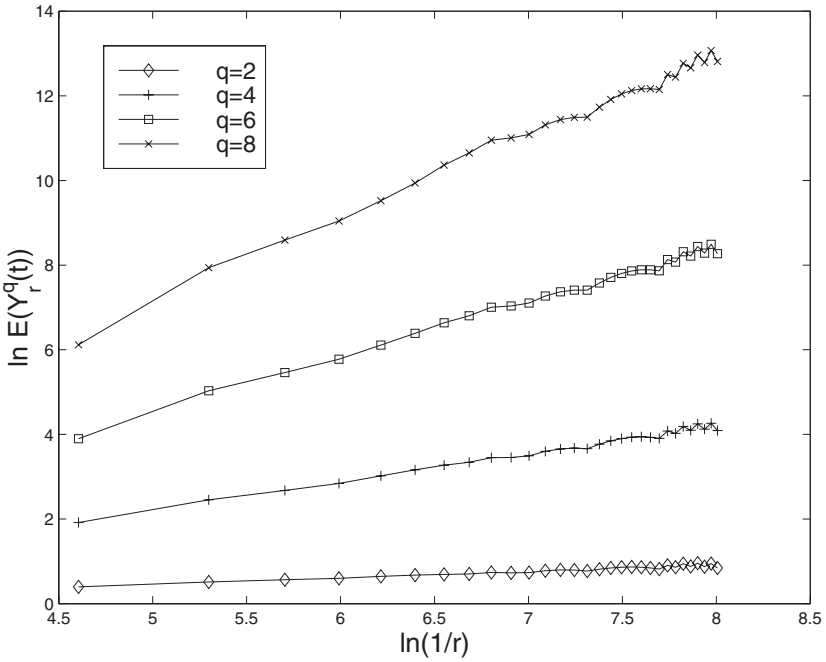

Figure 2. An example to show how to obtain the value of $K(q)$ directly using its definition.

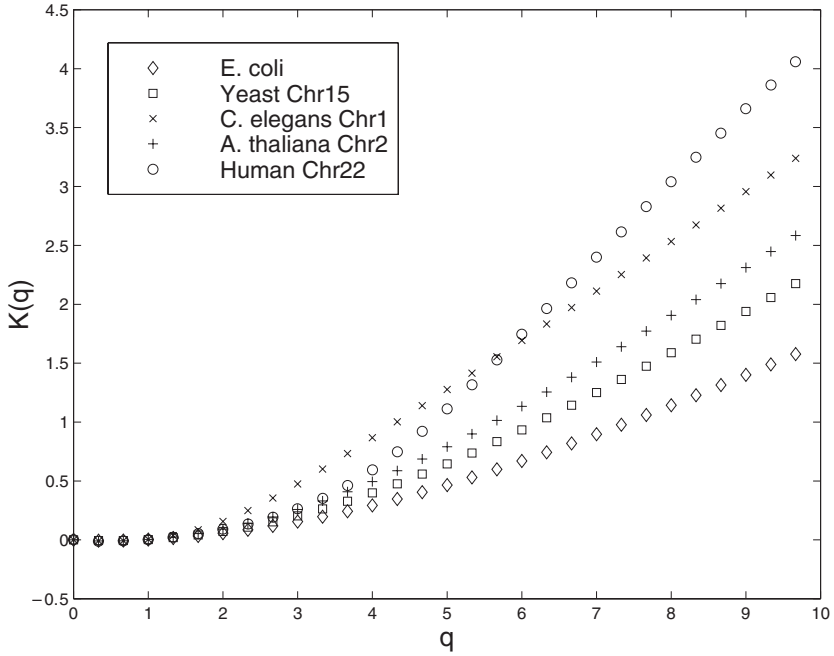

Figure 3. The values of $K(q)$ of chromosome 22 of *Homo sapiens*, chromosome 2 of *A. thaliana*, chromosome 1 of *C. elegans*, chromosome 15 of *S. cerevisiae* and *E. coli*.

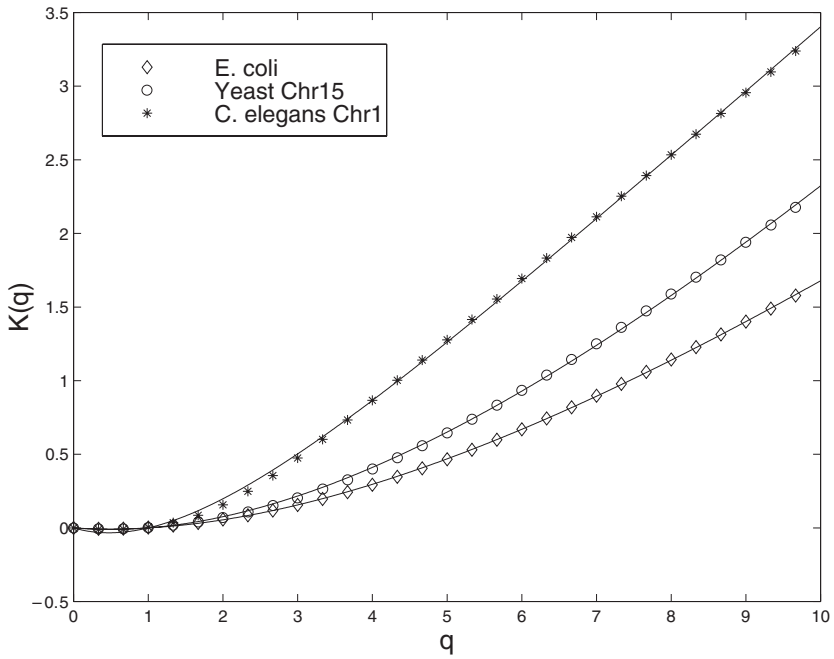

Figure 4. The data fitting of *E. coli*, chromosome 15 of *S. cerevisiae* and chromosome 1 of *C. elegans* based on the gamma model. The symbolled curves represent $K_d(q)$ computed from data, while the continuous curves represent $K(q)$ computed from formula (3.22).

Table 1. The values of κ, σ, α, error of all the organisms selected.

Species	κ	σ	α	error
Aquifex aeolicus	0.210 967	0.034 741	20.000 000	1.149 058E−03
Haemophilus influenzae	0.250 405	0.026 628	20.000 000	1.718 141E−04
Synechocystis sp. PCC6803	0.252 695	0.023 009	14.895 300	2.551 734E−04
Mycoplasma pneumoniae	0.260 598	0.028 227	14.468 067	1.367 545E−04
Chlamydia pneumoniae AR39	0.261 441	0.015 080	20.000 000	1.109 025E−02
Rhizobium sp. NGR234	0.269 307	0.141 332	1.974 406	1.037 725E−05
Chlamydia muridarum	0.282 757	0.021 999	20.000 000	5.528 718E−03
Chlamydia trachomatis	0.285 242	0.016 422	20.000 000	3.569 117E−03
Neisseria meningitidis MC58	0.287 688	0.021 525	20.000 000	3.869 811E−04
Helicobacter pylori 26695	0.296 316	0.042 743	20.000 000	3.999 003E−03
Helicobacter pylori J99	0.300 842	0.039 837	20.000 000	4.532 450E−03
Methanococcus jannaschii	0.305 624	0.034 413	19.737 016	9.356 220E−05
Rickettsia prowazekii	0.312 790	0.036 216	19.484 758	1.681 558E−04
Neisseria meningitidis Z2491	0.316 484	0.021 405	20.000 000	4.444 530E−04
Bacillus subtilis	0.325 036	0.015 238	20.000 000	5.327 829E−03
Aeropyrum pernix	0.325 043	0.024 461	20.000 000	1.056 628E−02
Mycoplasma genitalium	0.326 433	0.033 756	20.000 000	1.517 762E−03
Campylobacter jejuni	0.342 793	0.044 513	20.000 000	1.316 877E−03
M. tuberculosis	0.345 510	0.020 729	19.509 203	4.187 475E−04
Borrelia burgdorferi	0.350 140	0.045 101	20.000 000	2.282 837E−03
Thermotoga maritima	0.364 864	0.017 640	20.000 000	1.094 542E−03
Treponema pallidum	0.365 539	0.011 555	20.000 000	7.890 963E−03
Ureaplasma urealyticum	0.371 367	0.067 125	12.859 609	2.250 143E−04
Escherichia coli	0.386 280	0.024 556	6.404 487	2.418 786E−04
M. thermoautotrophicum	0.388 544	0.015 769	13.884 240	1.474 283E−03
Pseudomonas aeruginosa	0.412 200	0.753 456	0.918 436	4.798 280E−05
Caenorhabditis elegans Chr1	0.440 354	0.030 755	20.000 000	1.087 368E−02
Chlamydia pneumoniae AR39	0.484 163	0.018 637	20.000 000	2.701 796E−03
Archaeoglobus fulgidus	0.487 055	0.016 984	11.046 987	1.435 727E−03
S. cerevisiae Chr15	0.511 099	0.014 271	11.487 615	2.237 813E−03
Pyrococcus abyssi	0.513 144	0.016 623	7.295 978	7.294 311E−04
Buchnera sp. APS	0.536 577	0.031 866	20.000 000	4.064 171E−03
Arabidopsis thaliana Chr2	0.546 252	0.014 951	13.096 780	2.629 544E−03
Pyrococcus abyssi	0.562 316	0.015 389	11.328 229	1.574 777E−03
Vibrio cholerae	0.604 051	0.028 218	3.209 793	3.147 899E−04
Plasmodium falciparum Chr3	0.769 704	0.049 365	20.000 000	4.257 000E−02
Xylella fastidiosa	1.014 092	0.010 085	7.503 579	1.194 219E−02
Homo sapiens Chr22	1.290 643	0.008 267	12.966 19	1.900 450E−01

bottom, in the increasing order of the value of κ). It is found that the form (3.22) gives a perfect fit to the data for all bacteria. As an example, we give the data fitting of *E. coli, S. cerevisiae Chr15* and *C. elegans Chr1* in figure 4. But for higher organisms, for example, *Homo sapiens chromosome 22*, the fitting is not as good. Note that we only selected one chromosome for each higher organism. If all chromosomes for each higher organism are considered, the data fitting for K_q will be better. The fit for human chromosome is the worst in table 1. Since the length of human chromosome 22 is not larger than those of the complete genomes of all bacteria, there does not seem to be any relationship between the quality of fit and the length of the complete genome.

The parameter κ provides a tool to classify bacteria. From table 1, one can see *Helicobacter pylori 26695* and *Helicobacter pylori J99* group together, and three Chlamydia almost group

together. But this parameter κ alone is not sufficient, it must be combined with other tools to classify bacteria.

We also calculated the values of $\tau(q)$ using its definition (3.25). We found the values of $K_d(q)$ coincide with those obtained from (3.29). Hence we indeed can use (3.29) to calculate $K(q)$. Formula (3.22) gives an analytical expression for the quantity K_q. An analytical expressions for $\tau(q)$ can therefore be obtained from (3.29) and D_q from (3.26).

5. Conclusions

The idea of our measure representation is similar to the portrait method proposed by Hao *et al* [7]. It provides a simple yet powerful visualization method to amplify the difference between a DNA sequence and a random sequence as well as to distinguish DNA sequences themselves in more details. From our measure representation we can exactly know the frequencies of all the K-string appearing in the complete genome. But the representations alone are not sufficient to discuss the classification and evolution problem. Hence we need further tools.

In our previous work [15], when the measure representations of organisms were viewed as time series, it was found that they are far from being random time series, and in fact exhibit strong long-range dependence. Multifractal analysis of the complete genomes was performed in relation to the problem of classification and evolution of organisms. In this paper, we established a theoretical model of the probability distribution of the complete genomes. This probability distribution, particularly the resulting $K(q)$ curve, provides a precise tool for their characterization. Numerical results confirm the accuracy of the method of this paper.

For a completely random sequence based on the alphabet {a, c, g, t}, we have $D_q = 1$, $\tau(q) = q - 1$, $K(q) = 0$ for all q. From the $K(q)$ curves, it is seen that all complete genomes selected are far from being a completely random sequence.

Acknowledgments

The authors would like to express their gratitude to the referees for good comments and suggestions to improve this paper. This research was partially supported by QUT Postdoctoral Research Grant 9900658 to Zu-Guo Yu, and the RGC Earmarked Grant CUHK 4215/99P.

References

[1] Anh V V, Heyde C C and Tieng Q 1999 Stochastic models for fractal processes *J. Stat. Plan. Inference* **80** 123–35
[2] Berthelsen C L, Glazier J A and Raghavachari S 1994 *Phys. Rev.* E **49** 1860
[3] de Sousa Vieira M 1999 Statistics of DNA sequences: a low-frequency analysis *Phys. Rev.* E **60** 5932–7
[4] Feller W 1971 *An Introduction to Probability Theory and its Applications* vol 2 (New York: Wiley)
[5] Goldman N 1993 Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences *Nucl. Acids Res.* **21** 2487–91
[6] Gupta V K and Waymire E C 1993 A statistical analysis of mesoscale rainfall as a random cascade *J. Appl. Meteorol.* **32** 251–67
[7] Hao B-L, Lee H-C and Zhang S-Y 2000 Fractals related to long DNA sequences and complete genomes *Chaos Solitons Fractals* **11** 825–36
[8] Hao B-L, Xie H-M, Yu Z-G and Chen G-Y 2000 Avoided strings in bacterial complete genomes and a related combinatorial problem *Ann. Combin.* **4** 247–55
[9] Holley R and Waymire E C 1992 Multifractal dimensions and scaling exponents for strongly bounded random cascades *Ann. Appl. Probab.* **2** 819–45
[10] Jeffrey H J 1990 Chaos game representation of gene structure *Nucl. Acids Res.* **18** 2163–70

[11] Kahane J-P and Peyrière J 1976 Sur certaines martingales de benoit mandelbrot *Adv. Math.* **22** 131–45
[12] Mandelbrot B B 1974 Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier *J. Fluid Mech.* **62** 331–58
[13] Novikov E A 1994 Infinitely divisible distributions in turbulence *Phys. Rev.* E **50**
[14] Provata A and Almirantis Y 2000 Fractal cantor patterns in the sequence structure of DNA *Fractals* **8** 15–27
[15] Yu Z-G, Anh V and Lau K-S 2001 Measure representation and multifractal analysis of complete genomes *Phys. Rev.* E at press
[16] Yu Z-G and Anh V 2001 Time series model based on global structure of complete genome *Chaos Solitons Fractals* **12** 1827–34
[17] Yu Z-G, Anh V V and Wang B 2001 Correlation property of length sequences based on global structure of complete genome *Phys. Rev.* E **63** 11903
[18] Yu Z-G, Hao B-L, Xie H-M and Chen G-Y 2000 Dimension of fractals related to language defined by tagged strings in complete genome *Chaos Solitons Fractals* **11** 2215–22