Measure representation and multifractal analysis of complete genomes

Zu-Guo Yu,[1,2,*] Vo Anh,[1] and Ka-Sing Lau[3]
[1]*Centre in Statistical Science and Industrial Mathematics, Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia*
[2]*Department of Mathematics, Xiangtan University, Hunan 411105, People's Republic of China*[†]
[3]*Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong*
(Received 31 October 2000; revised manuscript received 1 May 2001; published 24 August 2001)

This paper introduces the notion of measure representation of DNA sequences. Spectral analysis and multifractal analysis are then performed on the measure representations of a large number of complete genomes. The main aim of this paper is to discuss the multifractal property of the measure representation and the classification of bacteria. From the measure representations and the values of the D_q spectra and related C_q curves, it is concluded that these complete genomes are not random sequences. In fact, spectral analyses performed indicate that these measure representations, considered as time series, exhibit strong long-range correlation. Here the long-range correlation is for the *K*-strings with dictionary ordering, and it is different from the base pair correlations introduced by other people. For substrings with length $K=8$, the D_q spectra of all organisms studied are multifractal-like and sufficiently smooth for the C_q curves to be meaningful. With the decreasing value of *K*, the multifractality lessens. The C_q curves of all bacteria resemble a classical phase transition at a critical point. But the ''analogous'' phase transitions of chromosomes of nonbacteria organisms are different. Apart from chromosome 1 of *C. elegans*, they exhibit the shape of double-peaked specific heat function. A classification of genomes of bacteria by assigning to each sequence a point in two-dimensional space (D_{-1},D_1) and in three-dimensional space (D_{-1},D_1,D_{-2}) was given. Bacteria that are close phylogenetically are almost close in the spaces (D_{-1},D_1) and (D_{-1},D_1,D_{-2}).

DOI: 10.1103/PhysRevE.64.031903 PACS number(s): 87.14.Gg, 87.10.+e, 47.53.+n

I. INTRODUCTION

DNA sequences are of fundamental importance in understanding living organisms, since all information of the hereditary and species evolution is contained in these macromolecules. The DNA sequence is formed by four different nucleotides, namely adenine (a), cytosine (c), guanine (g), and thymine (t). A large number of these DNA sequences are widely available in recent times. One of the challenges of DNA sequence analysis is to determine the patterns in these sequences. It is useful to distinguish coding from noncoding sequences. Problems related to the classification and evolution of organisms are also important. A significant contribution in these studies is to investigate the long-range correlation in DNA sequences [1–16]. Li and co-workers [1] found that the spectral density of a DNA sequence containing mostly introns shows $1/f^{\beta}$ behavior, which indicates the presence of long-range correlation when $0<\beta<1$. The correlation properties of coding and noncoding DNA sequences were first studied by Peng *et al.* [2] in their fractal landscape or DNA walk model. The DNA walk [2] was defined as that the walker steps ''up'' if a pyrimidine (c or t) occurs at position i along the DNA chain, while the walker steps ''down'' if a purine (a or g) occurs at position i. Peng *et al.* [2] discovered that there exists long-range correlation in noncoding DNA sequences while the coding sequences correspond to a regular random walk. By undertaking a more detailed analysis, Chatzidimitriou-Dreismann and Larhammer [5] concluded that both coding and noncoding sequences exhibit long-range correlation. A subsequent work by Prabhu and Claverie [6] also substantially corroborates these results. If one considers more details by distinguishing c from t in pyrimidine, and a from g in purine (such as two- or three-dimensional DNA walk models [15] and maps given by Yu and Chen [16]), then the presence of base correlation has been found even in coding sequences. On the other hand, Buldyrev *et al.* [12] showed that long-range correlation appears mainly in noncoding DNA using all the DNA sequences available. Based on equal-symbol correlation, Voss [8] showed a power law behavior for the sequences studied regardless of the proportion of intron contents. These studies add to the controversy about the possible presence of correlation in the entire DNA or only in the noncoding DNA. From a different angle, fractal analysis is a relatively new analytical technique that has proven useful in revealing complex patterns in natural objects. Berthelsen *et al.* [17] considered the global fractal dimensions of human DNA sequences treated as pseudorandom walks.

In the above studies, the authors only considered short or long DNA segments. Since the first complete genome of the free-living bacterium *Mycoplasma genitalium* was sequenced in 1995 [18], an ever-growing number of complete genomes has been deposited in public databases. The availability of complete genomes induces the possibility to establish some global properties of these sequences. Vieira [19] carried out a low-frequency analysis of the complete DNA of 13 microbial genomes and showed that their fractal behavior does not always prevail through the entire chain and the autocorrelation functions have a rich variety of behaviors including the pres-

*Corresponding author. Email address: yuzg@hotmail.com or z.yu@qut.edu.au
[†]Permanent corresponding address for Zu-Guo Yu.

ence of antipersistence. Yu and Wang [20] proposed a time series model of coding sequences in complete genomes. For fuller details on the number, size, and ordering of genes along the chromosome, one can refer to Part 5 of Lewin [21]. One may ignore the composition of the four kinds of bases in coding and noncoding segments and only consider the global structure of the complete genomes or long DNA sequences. Provata and Almirantis [22] 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 and/or noncoding partition in DNA sequences of lower organisms is homogeneouslike, while in the higher eucariotes the partition is fractal. This result does not seem refined enough to distinguish bacteria because the fractal dimensions of bacteria given by them [22] are all the same. The classification and evolution relationship of bacteria is one of the most important problems in DNA research. Yu and Anh [23] proposed a time series model based on the global structure of the complete genome and considered three kinds of length sequences. After calculating the correlation dimensions and Hurst exponents, it was found that one can get more information from this model than that of fractal Cantor pattern. Some results on the classification and evolution relationship of bacteria were found [23]. The correlation property of these length sequences has been discussed [24].

Although a 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 [25]. One needs more powerful global and visual methods. For this purpose, Hao *et al.* [25] proposed a visualization method based on counting and coarse-graining the frequency of appearance of substrings with 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 also discussed [26,27]. There are other graphical methods of sequence patterns, such as chaos game representation [28,29].

In the portrait representation, Hao *et al.* [25] used squares to represent substrings and discrete color 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. In order to improve it, in this paper we use subintervals in one-dimensional space to represent substrings and then 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. When the measure representation is viewed as a time series, a spectral analysis can be carried out.

Global calculations neglect the fact that DNA sequences are highly inhomogeneous. Multifractal analysis is a useful way to characterize the spatial inhomogeneity of both theoretical and experimental fractal patterns [30]. Multifractal analysis was initially proposed to treat turbulence data. In recent years it has been applied successfully in many different fields, including time series analysis [31,32] and financial modeling (see Anh *et al.* [33]). For DNA sequences, application of the multifractal technique seems rare (we have found only Berthelsen *et al.* [34]). In this paper, we pay more attention to this application. The quantities pertained to spectral and multifractal analyses of measures are described in Sec. III. Application of the methodology is undertaken in Sec. IV on a number of representative chromosomes. A discussion of the empirical results and some conclusions are drawn in Sec. V, where we also address the use of the multifractal technology in the classification problem of bacteria.

II. MEASURE REPRESENTATION

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-string 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 $s=s_1\cdots s_K, s_i\in\{a,c,g,t\}, i=1,\ldots,K$, be a substring with length K, we define

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

where

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

and

$$x_r(s)=x_l(s)+\frac{1}{4^K}. \tag{3}$$

We then use the subinterval $[x_l(s),x_r(s)[$ to represent 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

$$F_K(s)=N_K(s)/(L-K+1) \tag{4}$$

to be the frequency of substring s. It follows that $\Sigma_{\{s\}}F_K(s)=1$. Now we can define a measure μ_K on $[0,1[$ by $d\mu_K(x)=Y(x)dx$, where

$$Y_K(x)=4^K F_K(s), \quad \text{when} \quad x\in[x_l(s),x_r(s)[. \tag{5}$$

It is easy to see $\int_0^1 d\mu_K(x)=1$ and $\mu_K([x_l(s),x_r(s)[)=F_K(s)$. We call μ_K the *measure representation* of the organism corresponding to the given K. As an example, the histogram of substrings in the genome of *M. genitalium* for $K=3,\ldots,8$ are given in Fig. 1. Self-similarity is apparent in

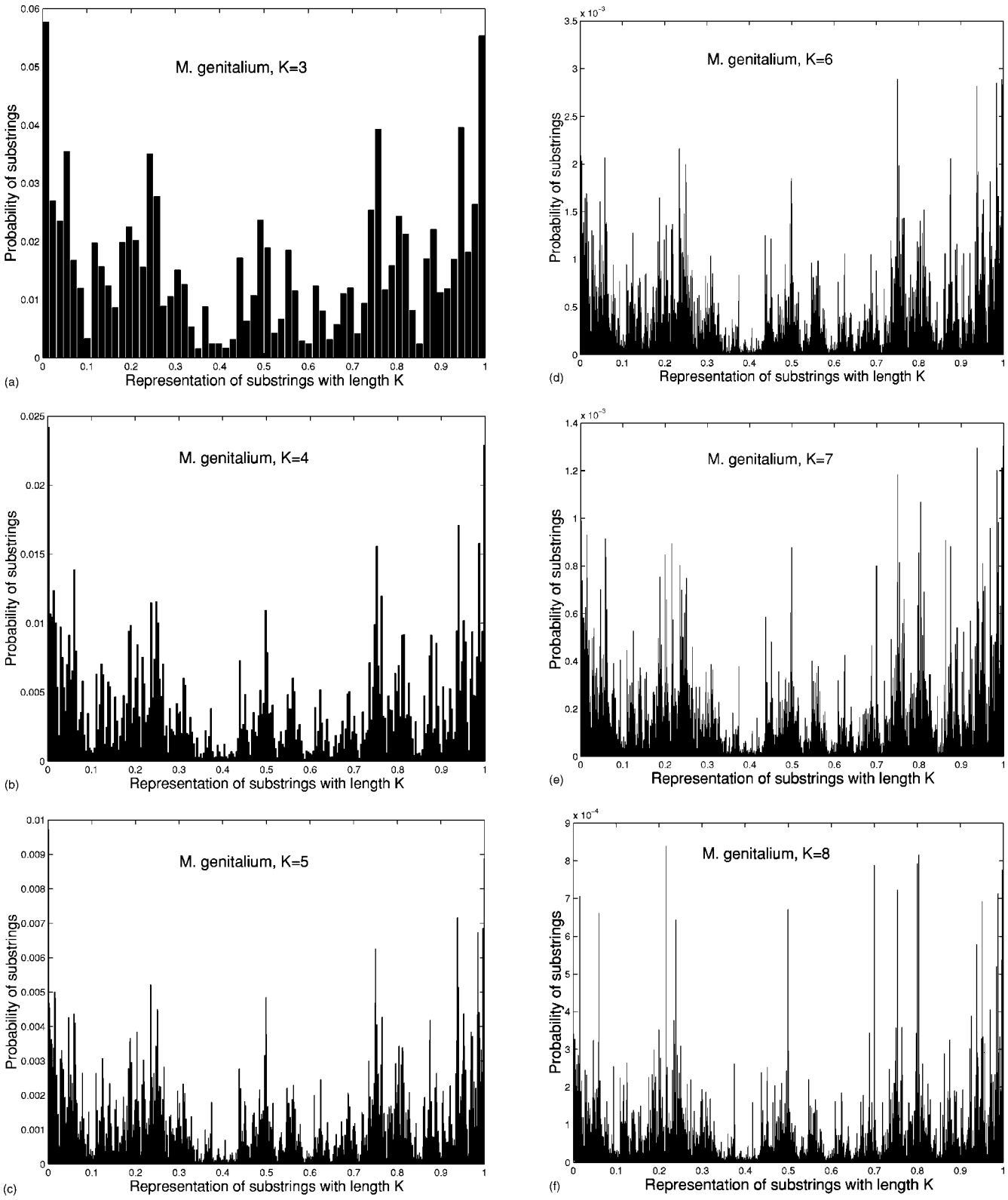

FIG. 1. Histograms of substrings with different lengths.

the measure. For simplicity of notation, the index K is dropped in $F_K(s)$, etc., from now on, where its meaning is clear.

Remark. The ordering of a,c,g,t in Eq. (2) will give the natural dictionary ordering of K-strings in the one-dimensional space. A different ordering of K-strings would change the nature of the correlations. But in our case, a different ordering of a,c,g,t in Eq. (2) gives almost the same D_q curve (therefore, the same with the C_q curve) which will be defined in the next section when the absolute value of q is

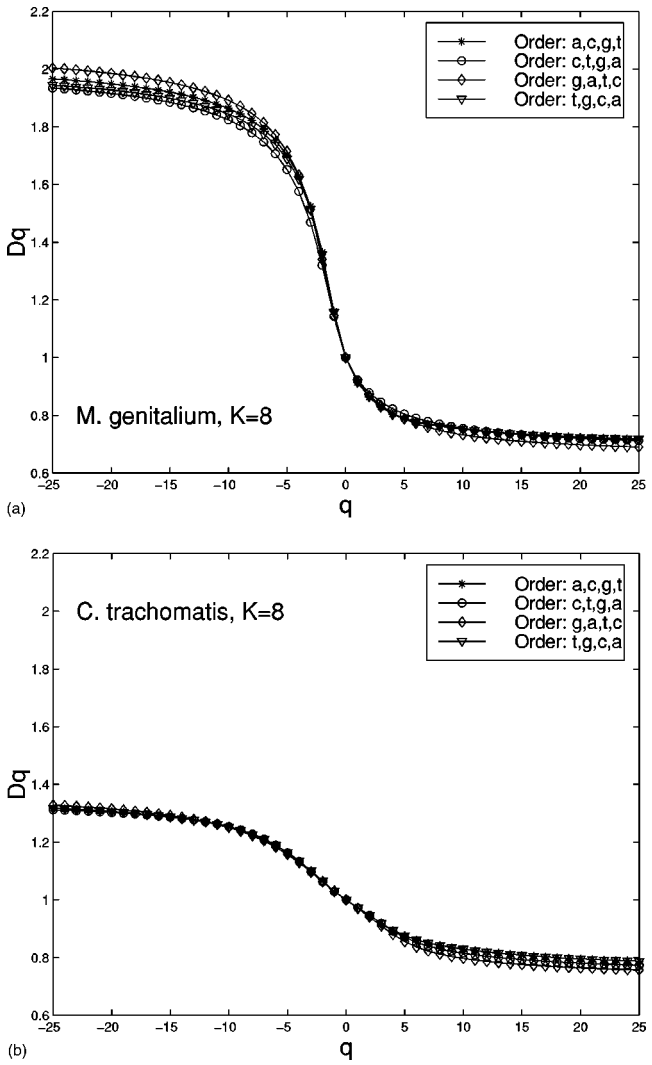

FIG. 2. The dimension spectra of measure representations given by different ordering of a,c,g,t in Eq. (2).

relatively small. We give Fig. 2 to support this point of view. Hence a different ordering of a,c,g,t in Eq. (2) will not change our result. When we want to compare different bacteria using the measure representation, once the ordering of a,c,g,t in Eq. (2) is given, it is fixed for all bacteria.

III. SPECTRAL AND MULTIFRACTAL ANALYSES

We can order all the $F(s)$ according to the increasing order of $x_l(s)$. We then obtain a sequence of real numbers consisting of 4^K elements that we denote as $F(t),t=1,\ldots,4^K$. Viewing the sequence $\{F(t)\}_{t=1}^{4^K}$ as a time series, the spectral analysis can then be undertaken on the sequence.

We first consider the discrete Fourier transform [35] of the time series $F(t),t=1,\ldots,4^K$, defined by

$$\hat{F}(f)=N^{-(1/2)}\sum_{t=0}^{N-1}F(t+1)e^{-2\pi ift}. \tag{6}$$

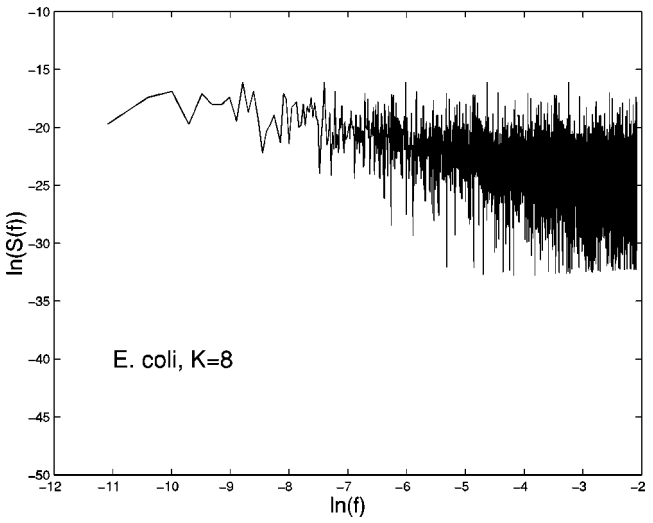

FIG. 3. The logarithmic power spectrum of the measure of E. coli corresponding to $K=8$. The estimated value of β is 0.598 691 2.

Then

$$S(f)=|\hat{F}(f)|^2 \tag{7}$$

is the *power spectrum of* $F(t)$. In recent studies, it has been found [36] that many natural phenomena lead to the power spectrum of the form $1/f^{\beta}$. This kind of dependence was named $1/f$ noise, in contrast to white noise $S(f)=\text{const}$, i.e., $\beta=0$. Let the frequency f take k values $f_k=k/N,k=1,\ldots,N/8$. From the $\ln[S(f)]$ vs $\ln(f)$ graph we can infer the value of β using the above low-frequency range. For example, we give the logarithmic power spectrum of the measure of *E. coli* with $K=8$ in Fig. 3.

The most common operative numerical implementations of multifractal analysis are the so-called *fixed-size box-counting algorithms* [37]. In the one-dimensional case, for a given measure μ with support $E\subset\mathbf{R}$, we consider the *partition sum*

$$Z_{\epsilon}(q)=\sum_{\mu(B)\neq 0}[\mu(B)]^q, \tag{8}$$

$q\in\mathbf{R}$, where the sum runs over all different nonempty boxes B of a given side ϵ in a grid covering of the support E, that is,

$$B=[k\epsilon,(k+1)\epsilon[. \tag{9}$$

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

$$\tau(q)=\lim_{\epsilon\to 0}\frac{\ln Z_{\epsilon}(q)}{\ln\epsilon} \tag{10}$$

and the generalized fractal dimensions of the measure are defined as

$$D_q=\tau(q)/(q-1)\quad\text{for}\quad q\neq 1 \tag{11}$$

and

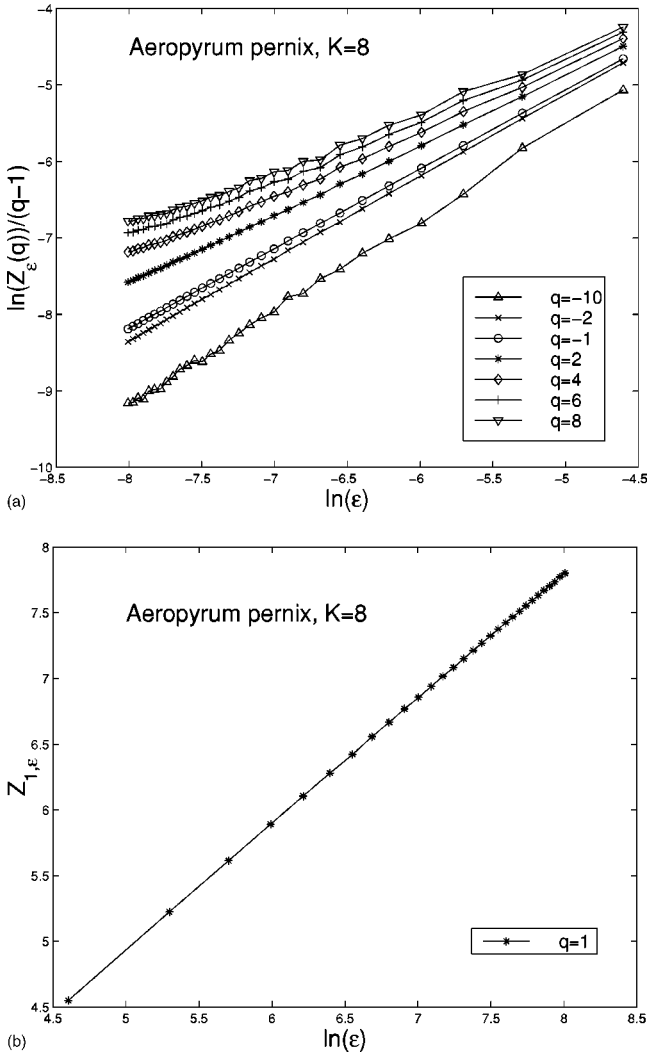

FIG. 4. The linear slopes in the D_q spectra.

$$D_q = \lim_{\epsilon \to 0} \frac{Z_{1,\epsilon}}{\ln \epsilon} \quad \text{for} \quad q=1, \tag{12}$$

where $Z_{1,\epsilon}=\Sigma_{\mu(B)\neq 0}\mu(B)\ln\mu(B)$. The generalized fractal dimensions are numerically estimated through a linear regression of

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

against $\ln\epsilon$ for $q\neq 1$, and similarly through a linear regression of $Z_{1,\epsilon}$ against $\log\epsilon$ for $q=1$. For example, we show how to obtain the D_q spectrum using the slope of the linear regression in Fig. 4. D_1 is called *information dimension* and D_2 is called *correlation dimension*. The D_q of the positive values of q give relevance to the regions where the measure is large, i.e., to the K-strings with high probability. The D_q of the negative values of q deal with the structure and the properties of the most rarefied regions of the measure.

Some sets of physical interest have a nonanalytic dependence of D_q on q. Moreover, this phenomenon has a direct analogy to the phenomenon of phase transitions in condensed-matter physics [38]. The existence and type of phase transitions might turn out to be a worthwhile characterization of universality classes for the structures [39]. The concept of phase transition in multifractal spectra was introduced in the study of logistic maps, Julia sets, and other simple systems. Evidence of a phase transition was found in the multifractal spectrum of diffusion-limited aggregation [40]. By following the thermodynamic formulation of multifractal measures, Canessa [32] derived an expression for the ''analogous'' specific heat as

$$C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q)-\tau(q+1)-\tau(q-1). \tag{13}$$

He showed that the form of C_q resembles a classical phase transition at a critical point for financial time series. In the next section we discuss the property of C_q for our measure representations of organisms.

IV. DATA AND RESULTS

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*; four 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 as follows. 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 nonbacteria organisms are also currently available. In order to discuss the classification problem of bacteria, we also 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*.

We obtained the dimension spectra and ''analogous'' specific heat of the measure representations of the above organisms and used them to discuss the classification problem. We calculated the dimension spectra and analogous specific heat of chromosome 22 of *Homo sapiens* for $K=1,\ldots,8$, and found that the D_q and C_q curves of $K=6,7,8$ are very close to one another (see Figs. 5 and 6). Hence it seems appropri-

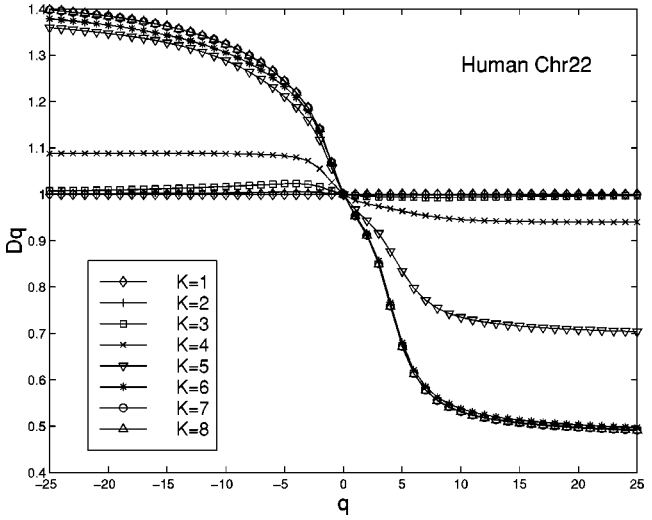

FIG. 5. Dimension spectra of measures of substrings with different lengths K in chromosome 22 of *Homo sapiens*.

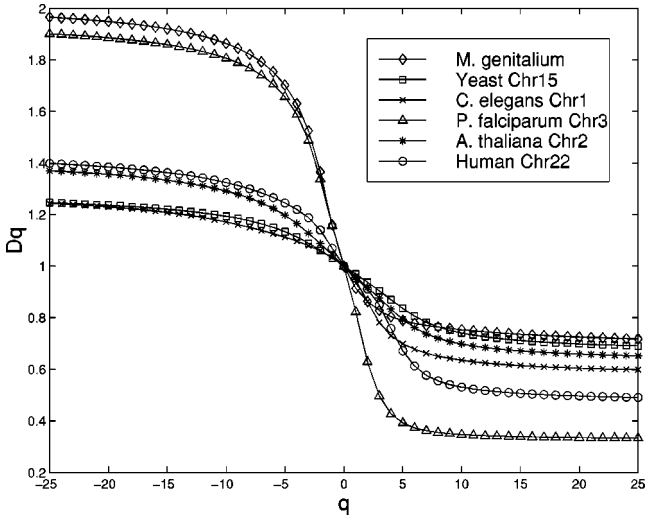

FIG. 7. Dimension spectra of chromosome 22 of *Homo sapiens*, chromosome 2 of *A. thaliana*, chromosome 3 of *P. falciparum*, chromosome 1 of *C. elegans*, and chromosome 15 of *S. cerevisiae* and *M. genitalium*.

ate to use the measure corresponding to $K=8$. For $K=8$, we calculated the dimension spectra, analogous specific heat and the exponent β of the measure representations of all the above organisms. As an illustration, we plot the D_q curves of *M. genitalium*, chromosome 15 of *Saccharomyces cerevisiae*, chromosome 3 of *Plasmodium falciparum*, chromosome 2 of *Arabidopsis thaliana*, and chromosome 22 of *Homo sapiens* in Fig. 7; and the C_q curves of these organisms in Fig. 8. Because all D_q are equal to 1 for the complete random sequence, from these plots it is apparent that the D_q and C_q curves are nonlinear and significantly different from those of the completely random sequence. From Fig. 7, we can claim that the curves representative of the organisms are clearly distinct from the curve representing a random sequence. From the plot of D_q, the dimension spectra of organisms exhibit a multifractal-like form. From Fig. 4, we can see the linear fits of $q=-2,-1,1,2$ are perfect and better than that of other values of q, Hence we suggest to use D_{-2},D_{-1},D_1,D_2 in the comparison of different bacteria. We give the numerical results for D_{-2},D_{-1},D_1,D_2 in Table I (from top to bottom, in the increasing order of the value of D_{-1}).

If only a few bacteria are considered at a time, we can use the D_q curve to distinguish them. This strategy is clearly not efficient when a large number of organisms are to be distinguished. For this purpose, we suggest using D_{-1},D_1, and D_{-2}, in conjunction with two-dimensional points (D_{-1},D_1) or three-dimensional points (D_{-1},D_1,D_{-2}). We give the

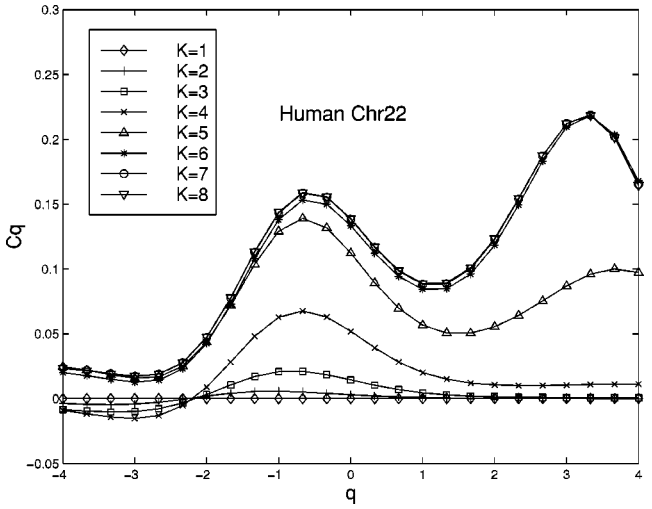

FIG. 6. Analogous specific heat of measures of substrings with different lengths K in chromosome 22 of *Homo sapiens*.

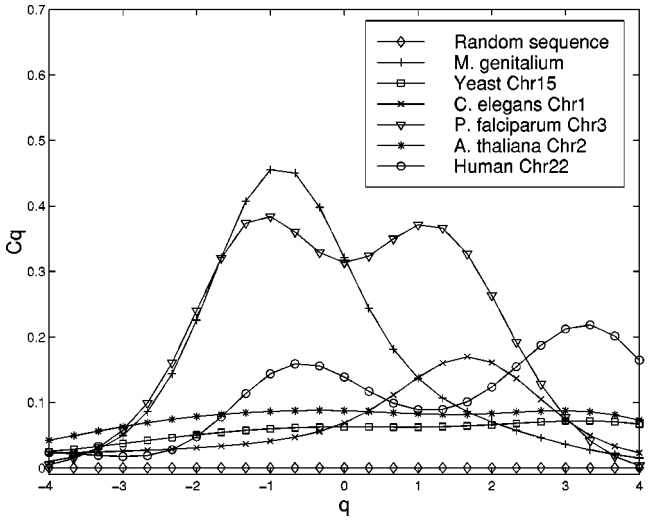

FIG. 8. Analogous specific heat of chromosome 22 of *Homo sapiens*, chromosome 2 of *A. thaliana*, chromosome 3 of *P. falciparum*, chromosome 1 of *C. elegans*, chromosome 15 of *S. cerevisiae* and *M. genitalium*, and a complete random sequence.

TABLE I. The values of D_{-1}, D_1, D_{-2}, and D_2 of all bacteria selected.

Species	Category	D_{-1}	D_1	D_{-2}	D_2
Xylella fastidiosa	Proteobacteria	1.023 935	0.973 450 5	1.046 237	0.943 400 7
Treponema pallidum	Spirochaete	1.024 096	0.974 452 9	1.048 537	0.945 687 9
Vibrio cholerae	Proteobacteria	1.027 849	0.975 419 3	1.060 974	0.952 940 2
Bacillus subtilis	Gram-positive Eubacteria	1.031 173	0.969 183 1	1.062 364	0.939 298 6
Chlamydia trachomatis	Chlamydia	1.031 900	0.970 572 3	1.067 158	0.942 124 1
Chlamydia pneumoniae	Chlamydia	1.034 190	0.969 118 9	1.075 935	0.939 613 8
Rhizobium sp. NGR234	Proteobacteria	1.034 821	0.968 923 3	1.068 532	0.943 014 1
Chlamydia muridarum	Chlamydia	1.036 608	0.964 696 0	1.075 166	0.929 364 0
Chlamydia pneumoniae AR39	Chlamydia	1.037 127	0.959 307 4	1.078 164	0.910 617 1
Pyrococcus abyssi	Archaebacteria	1.038 142	0.968 308 1	1.091 387	0.939 338 4
Aeropyrum pernix	Archaebacteria	1.040 248	0.953 563 0	1.074 807	0.903 315 9
Synechocystis sp. PCC6803	Cyanobacteria	1.045 674	0.965 713 7	1.127 265	0.936 414 1
Mycoplasma pneumoniae	Gram-positive Eubacteria	1.046 260	0.958 464 9	1.092 869	0.925 010 6
Archaeoglobus fulgidus	Archaebacteria	1.047 071	0.963 125 2	1.130 371	0.927 948 0
Escherichia coli	Proteobacteria	1.047 849	0.971 164 5	1.174 754	0.947 431 7
M. thermoautotrophicum	Archaebacteria	1.048 569	0.962 648 0	1.116 451	0.930 676 0
Thermotoga maritima	Hyperthermophilic bacteria	1.053 824	0.954 563 7	1.145 209	0.910 159 6
Aquifex aeolicus	Hyperthermophilic bacteria	1.055 210	0.954 089 3	1.134 702	0.914 536 1
Pyrococcus horikoshii	Archaebacteria	1.056 144	0.958 792 4	1.139 402	0.923 767 4
Neisseria meningitidis MC58	Proteobacteria	1.058 779	0.952 268 1	1.132 902	0.913 238 3
Neisseria meningitidis Z2491	Proteobacteria	1.058 805	0.949 750 3	1.133 201	0.906 516 7
M. tuberculosis	Gram-positive Eubacteria	1.061 496	0.941 034 1	1.115 466	0.892 054 0
Haemophilus influenzae	Proteobacteria	1.062 565	0.951 123 1	1.147 970	0.912 226 0
Buchnera sp. APS	Proteobacteria	1.085 581	0.895 585 1	1.152 650	0.790 422 1
Rickettsia prowazekii	Proteobacteria	1.088 237	0.919 265 5	1.173 883	0.856 704 4
Pseudomonas aeruginosa	Proteobacteria	1.109 776	0.915 498 0	1.187 378	0.862 232 1
Borrelia burgdorferi	Spirochaete	1.111 380	0.903 053 9	1.261 299	0.829 832 3
Campylobacter jejuni	Proteobacteria	1.123 096	0.905 343 7	1.279 505	0.834 979 3
Ureaplasma urealyticum	Gram-positive bacteria	1.124 616	0.884 348 1	1.260 287	0.806 591 6
Helicobacter pylori J99	Proteobacteria	1.128 590	0.929 961 4	1.390 791	0.875 844 3
Helicobacter pylori 26695	Proteobacteria	1.149 943	0.927 606 2	1.460 757	0.871 944 5
Mycoplasma genitalium	Gram-positive Eubacteria	1.160 435	0.914 271 8	1.365 716	0.863 178 9
Methanococcus jannaschii	Archaebacteria	1.165 208	0.911 373 1	1.349 664	0.862 822 6

distribution of two-dimensional points (D_{-1},D_1) and three-dimensional points (D_{-1},D_1,D_{-2}) of bacteria in Fig. 9.

V. DISCUSSION AND CONCLUSIONS

The idea of our measure representation is similar to the portrait method proposed by Hao *et al.* [25]. 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. If a DNA sequence is random, then our measure representation yields a uniform measure $(D_q=1,C_q=0)$.

From the measure representation and the values of D_q and C_q, it is seen that there exists a clear difference between the DNA sequences of all organisms considered here and the completely random sequence. Hence we can conclude that complete genomes are not random sequences.

We obtained the values of the exponent β of our measure representations ($\beta=0.393\,003$ for *V. cholerae*, $\beta=0.311\,623$ for *A. pernix*, $\beta=0.240\,601$ for *X. fastidiosa*, $\beta=0.381\,293$ for *T. pallidum*, $\beta=0.334\,057$ for *C. pneumoniae AR39*, and β is larger than 0.4 for all other bacteria selected). These values are far from 0. Hence when we view our measure representations of organisms as time series, they are far from being random time series, and in fact exhibit strong long-range correlation. Here the long-range correlation is for the *K*-strings with the dictionary ordering, and it is different from the base pair correlations introduced by other people.

Although the existence of the archaebacterial urkingdom has been accepted by many biologists, the classification of bacteria is still a matter of controversy [41]. The evolutionary relationship of the three primary kingdoms, namely archeabacteria, eubacteria, and eukaryote, is another crucial problem that remains unresolved [41].

When K is large $(K\geqslant 6)$, our measure representation contains rich information on the complete genomes. From Figs. 5 and 6 we find the curves of D_q and C_q are very close to

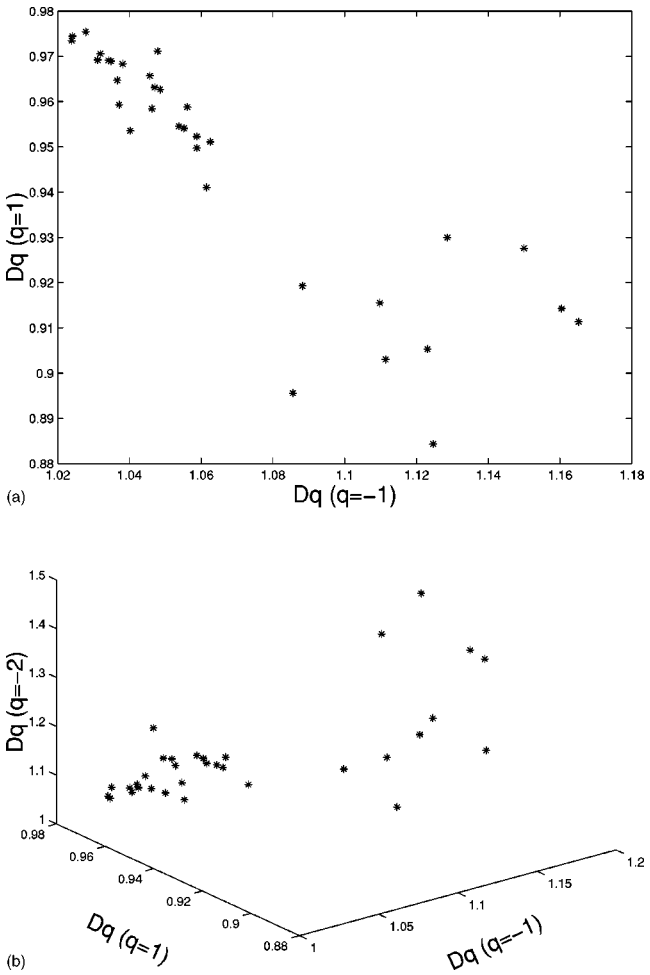

FIG. 9. Distributions of two-dimensional points (D_{-1},D_1) and three-dimensional points (D_{-1},D_1,D_{-2}) of the bacteria selected.

one another for $K=6,7,8$. Hence, for the classification problem, it would be appropriate to take $K=8$. We calculated the β, D_q, and C_q values of all organisms selected in this paper for $K=8$. We found that the D_q spectra of all organisms are multifractal-like and sufficiently smooth so that the C_q curves can be meaningfully estimated. From Fig. 5, with the decreasing of K, the multifractality becomes less severe. With $K=8$, we found that the C_q curves of all other bacteria resemble a classical phase transition at a critical point similar to that of *M. genitalium* shown in Fig. 8. But the analogous phase transitions of nonbacteria organisms are different. Apart from chromosome 1 of *C. elegans*, they exhibit the shape of a double-peaked specific heat function which is known to appear in the Hubbard model within the weak-to-strong coupling regime [42].

It is seen that the D_q curve is not clear enough to distinguish many bacteria themselves. In order to solve this problem we use two-dimensional points (D_{-1},D_1) and three-dimensional points (D_{-1},D_1,D_{-2}). From Fig. 9 it is clear that bacteria roughly gather into two classes (as shown in Table I). Using the distance among the points, one can obtain a classification of bacteria.

From Table I we can see all Archaebacteria belong to the same class except *M. jannaschii*. And four Chlamydia almost gather together. It is surprising that the closest pairs of bacteria, *Helicobacter pylori J99* and *Helicobacter pylori 26695* and *Neisseria meningitidis MC58* and *Neisseria meningitidis Z2491*, group with each other. Two hyperthermophilic bacteria group with each other and are linked with the Archaebacteria. It has previously been shown that *Aquifex* has a close relationship with Archaebacteria from the gene comparison of an enzyme needed for the synthesis of the amino acid trytophan [43] and using the length sequence of a complete genome [23]. In general, bacteria that are close phylogenetically are almost close in the spaces (D_{-1},D_1) and (D_{-1},D_1,D_{-2}).

ACKNOWLEDGMENTS

One of the authors, Zu-Guo Yu, would like to express his gratitude to Professor Bai-lin Hao of Institute of Theoretical Physics of the Chinese Academy of Science for introducing him into this field and also for his continuous encouragement. He also wants to thank Dr. Enrique Canessa of ICTP for pointing out the importance of the quantity C_q and for his useful comments, and Dr. Guo-Yi Chen of ITP for useful suggestions on the measure representation. This research was partially supported by QUT Grant No. 9900658 and the HKRGC Earmarked Grant CUHK 4215/99P.

[1] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992); W. Li, T. Marr, and K. Kaneko, Physica D **75**, 392 (1994).
[2] C.K. Peng, S. Buldyrev, A.L. Goldberg, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, Nature (London) **356**, 168 (1992).
[3] J. Maddox, Nature (London) **358**, 103 (1992).
[4] S. Nee, Nature (London) **357**, 450 (1992).
[5] C.A. Chatzidimitriou-Dreismann and D. Larhammar, Nature (London) **361**, 212 (1993).
[6] V.V. Prabhu and J.M. Claverie, Nature (London) **359**, 782 (1992).
[7] S. Karlin and V. Brendel, Science **259**, 677 (1993).
[8] (a) R. Voss, Phys. Rev. Lett. **68**, 3805 (1992); (b) Fractals **2**, 1 (1994).
[9] H.E. Stanley, S.V. Buldyrev, A.L. Goldberg, Z.D. Goldberg, S. Havlin, R.N. Mantegna, S.M. Ossadnik, C.K. Peng, and M. Simons, Physica A **205**, 214 (1994).
[10] H. Herzel, W. Ebeling, and A.O. Schmitt, Phys. Rev. E **50**, 5061 (1994).
[11] P. Allegrini, M. Barbi, P. Grigolini, and B.J. West, Phys. Rev. E **52**, 5281 (1995).
[12] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E **51**, 5084 (1995).
[13] A. Arneodo, E. Bacry, P.V. Graves, and J.F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).

[14] A.K. Mohanty and A.V.S.S. Narayana Rao, Phys. Rev. Lett. **84**, 1832 (2000).
[15] Liaofu Luo, Weijiang Lee, Lijun Jia, Fengmin Ji, and Lu Tsai, Phys. Rev. E **58**, 861 (1998).
[16] Zu-Guo Yu and Guo-Yi Chen, Commun. Theor. Phys. **33**, 673 (2000).
[17] C.L. Berthelsen, J.A. Glazier, and M.H. Skolnick, Phys. Rev. A **45**, 8902 (1992).
[18] C.M. Fraser *et al.*, Science **270**, 397 (1995).
[19] Maria de Sousa Vieira, Phys. Rev. E **60**, 5932 (1999).
[20] Zu-Guo Yu and Bin Wang, Chaos, Solitons Fractals **12**, 519 (2001).
[21] B. Lewin, *Genes VI* (Oxford University Press, Oxford, 1997).
[22] A. Provata and Y. Almirantis, Fractals **8**, 15 (2000).
[23] Zu-Guo Yu and Vo Anh, Chaos, Solitons Fractals **12**, 1827 (2001).
[24] Zu-Guo Yu, V.V. Anh, and Bin Wang, Phys. Rev. E **63**, 011903 (2001).
[25] Bai-lin Hao, Hoong-Chien Lee, and Shu-yu Zhang, Chaos, Solitons Fractals **11**, 825 (2000).
[26] Zu-Guo Yu, Bai-lin Hao, Hui-min Xie, and Guo-Yi Chen, Chaos, Solitons Fractals **11**, 2215 (2000).
[27] Bai-Lin Hao, Hui-Ming Xie, Zu-Guo Yu, and Guo-Yi Chen, Physica A **288**, 10 (2001).
[28] H.J. Jeffrey, Nucleic Acids Res. **18**, 2163 (1990).
[29] N. Goldman, Nucleic Acids Res. **21**, 2487 (1993).
[30] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).
[31] R. Pastor-Satorras, Phys. Rev. E **56**, 5284 (1997).
[32] E. Canessa, J. Phys. A **33**, 3637 (2000).
[33] V.V. Anh, Q.M. Tieng, and Y.K. Tse, International Transaction in Operations Research **7**, 349 (2000).
[34] C.L. Berthelsen, J.A. Glazier, and S. Raghavachari, Phys. Rev. E **49**, 1860 (1994).
[35] R. H. Shumway, *Applied Statistical Time Series Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1988).
[36] F. N. H. Robinson, *Noise and Fluctuations* (Clarendon Press, Oxford, 1974).
[37] T. Halsey, M. Jensen, L. Kadanoff, I. Procaccia, and B. Schraiman, Phys. Rev. A **33**, 1141 (1986).
[38] D. Katzen and I. Procaccia, Phys. Rev. Lett. **58**, 1169 (1987).
[39] T. Bohr and M. Jensen, Phys. Rev. A **36**, 4904 (1987).
[40] J. Lee and H.E. Stanley, Phys. Rev. Lett. **61**, 2945 (1988).
[41] N. Iwabe *et al.*, Proc. Natl. Acad. Sci. U.S.A. **86**, 9355 (1989).
[42] D. Vollhardt, Phys. Rev. Lett. **78**, 1307 (1997).
[43] E. Pennisi, Science **286**, 672 (1998).