ELSEVIER

Physica A 301 (2001) 351–361

PHYSICA A

www.elsevier.com/locate/physa

Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome

Zu-Guo Yu[a,b,*,1], Vo Anh[b], Ka-Sing Lau[c]

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

[b] *Department of Mathematics, Xiangtan University, Hunan 411105, People's Republic of China*

[c] *Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong*

Received 7 May 2001

Abstract

The coding and noncoding length sequences constructed from a complete genome are characterised by multifractal analysis. The dimension spectrum D_q and its derivative, the 'analogous' specific heat C_q, are calculated for the coding and noncoding length sequences of bacteria, where q is the moment order of the partition sum of the sequences. From the shape of the D_q and C_q curves, it is seen that there exists 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 D_q curves for coding length sequences are flat, so their multifractality is small whereas almost all D_q curves for noncoding length sequences are multifractal-like. It is seen that the 'analogous' specific heats of noncoding length sequences of bacteria have a rich variety of behaviour which is much more complex than that of coding length sequences. We propose to characterise the bacteria according to the types of the C_q curves of their noncoding length sequences. This new type of classification allows a better understanding of the relationship among bacteria at the global gene level instead of nucleotide sequence level.

PACS: 87.10+e; 47.53+n

Keywords: Coding/noncoding segments; Length sequence; Complete genome; Multifractal analysis; 'analogous' specific heat

* Corresponding author. School of Mathematical Science, Queensland University of Technology, Garden Point Campus, GPO Box 2434, Brisbane, Qld 4001, Australia. Tel.: +61-7-38645194; fax: +61-7-38642310. *E-mail address:* yuzg@hotmail.com (Z.-G. Yu).

[1] Permanent address: Department of Mathematics, Xiangtan University, Hunan 411105, People's Republic of China.

0378-4371/01/$ - see front matter
PII: S0378-4371(01)00391-0

1. Introduction

The rapidly accumulating complete genome sequences of bacteria and archaea provide a new type of information resource for understanding gene functions and evolution [1].

One can study the DNA sequences in detail by considering the order in which four kinds of nucleotides of DNA are assembled, namely adenine (a), cytosine (c), guanine (g), and thymine (t).

There has been considerable interest in the finding of long-range correlation (LRC) in DNA sequences at this level. Li et al. [2,3] found that the spectral density of a DNA sequence containing mostly introns shows $1/f^{\beta}$ behaviour, which indicates the presence of LRC. The correlation properties of coding and noncoding DNA sequences were also studied by Peng et al. [4] in their fractal landscape or DNA walk model. The DNA walk defined in Ref. [4] is 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. [4] discovered that there exists LRC in noncoding DNA sequences while the coding sequences correspond to a regular random walk. By doing a more detailed analysis, Chatzidimitriou-Dreismann and Larhammar [5] concluded that both coding and noncoding sequences exhibit LRC. A subsequent work by Prabhu and Claverie [6] also substantially corroborated 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 model [7] and maps given in Ref. [8]), then the presence of base correlation can be found even in coding sequences. In view of the controversy about the presence of correlation in all DNA or only in noncoding DNA, Buldyrev et al. [9] showed that the LRC appears mainly in noncoding DNA using all the DNA sequences available. Alternatively, Voss [10,11], based on equal-symbol correlation, showed a power-law behaviour for the sequences studied regardless of the percent of intron contents. Investigations based on different models seem to suggest different results, as they all look into only a certain aspect of the entire DNA sequence [12].

The avoided and under-represented strings in some bacterial complete genomes have been discussed [13–15]. A time series model of CDS in complete genome has been proposed [16]. Vieira [17] performed a low-frequency analysis of the complete DNA of 13 microbial genomes and found that their fractal behaviour does not always prevail through the entire chain and their autocorrelation functions have a rich variety of behaviours including the presence of anti-persistence.

For the importance of the numbers, sizes and ordering of genes along the chromosome, one can refer to Part 5 of Lewin [18]. Here, one may ignore the composition of the four kinds of bases in coding and noncoding segments and only consider the rough structure of the complete genome or long DNA sequences. Provata and Almirantis [19] proposed a fractal Cantor pattern of DNA. They map coding segments to filled regions and noncoding segments to empty regions of random Cantor set and then calculate the fractal dimension of the random fractal set. They found that the coding/noncoding partition in DNA sequences of lower organisms is homogeneous-like, while in the higher

eucaryotes the partition is fractal. This result seems too rough to distinguish bacteria because the fractal dimensions of bacteria they gave out are all the same.

Viewing 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, one can get two kinds of integer sequences by the following ways:

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

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

Yu and Anh [20] proposed a time series model for the length sequences of DNA. 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 [19]. The quantification of these correlations could give an insight into the role of the ordering of genes on the chromosome. Through detrended fluctuation analysis (DFA) [21] and spectral analysis, the LRC was found in these length sequences [22].

The correlation dimension and Hurst exponent are parameters of global analysis. Global calculations neglect the fact that length sequences from a complete genome are highly inhomogeneous. Thus multifractal analysis is a useful way to characterise the spatial inhomogeneity of both theoretical and experimental fractal patterns [23]. It was initially proposed to treat turbulence data. In recent years, it has been applied successfully in many different fields including time series analysis [24] and financial modelling [25,26]. For DNA sequences, application of the multifractal technique seems rare (we have found only Berthelsen et al. [27]). Recently, Yu et al. [28] considered the multifractal property of the measure representation of a complete genome. In this paper, we pay more attention to the multifractal characterisation of the coding and noncoding length sequences.

Some sets of physical interest have a nonanalytic dependence of dimension spectrum D_q on the q-moments of the partition sum of the sequences. Moreover, multifractality has a direct analogy to the phenomenon of phase transition in condensed-matter physics [29]. The existence and type of phase transitions might turn out to be a worthwhile characterisation of universality classes for the structures [30]. The concept of phase transition in multifractal spectra was introduced in the study of logistic maps, Julia sets and other simple systems. Evidence of phase transition was found in the multifractal spectrum of diffusion-limited aggregation [31]. By following the thermodynamic formulation of multifractal measures, where q represents an analogous temperature, Canessa [25] applied a standard expression for the 'analogous' specific heat and showed that its form resembles a classical phase transition at a critical point for financial time series.

In this paper, we calculate the 'analogous' specific heat of coding and noncoding length sequences. Our motivation to apply Canessa's framework to characterise stochastic sequences is to see whether there is a similar type of phase transition in the coding

and noncoding length sequences as in other time series. We show that based on the shape of the C_q curves and associated type of phase transitions, one can discuss the classification of bacteria. This new type of classification allows to better understand the relationship among bacteria at the global gene level instead of nucleotide sequence level.

2. Multifractal analysis

Let T_t, $t=1,2,\ldots,N$, be the length sequence of coding or noncoding segments in the complete genome of an organism. First we define

$$F_t = T_t \Big/ \left(\sum_{j=1}^{N} T_j \right) \tag{1}$$

to be the frequency of T_t. It follows that $\sum_t F_t = 1$. Now we can define a measure μ on $[0,1[$ by $\mathrm{d}\mu(x) = Y(x)\,\mathrm{d}x$, where

$$Y(x) = N \times F_t \quad \text{when } x \in \left[\frac{t-1}{N}, \frac{t}{N} \right[\,. \tag{2}$$

It is easy to see that $\int_0^1 \mathrm{d}\mu(x) = 1$ and $\mu([(t-1)/N, t/N[) = F_t$.

The most common numerical implementations of multifractal analysis are the so-called *fixed-size box-counting algorithms* [32]. In the one-dimensional case, for a given measure μ with support $E \subset \mathbf{R}$, we consider the *partition sum*

$$Z_\varepsilon(q) = \sum_{\mu(B)\neq 0} [\mu(B)]^q \,, \tag{3}$$

$q \in \mathbf{R}$, 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,

$$B = [k\varepsilon, (k+1)\varepsilon[\,. \tag{4}$$

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

$$\tau(q) = \lim_{\varepsilon\to 0} \frac{\log Z_\varepsilon(q)}{\log \varepsilon} \tag{5}$$

and the generalized fractal dimensions of the measure are defined as

$$D_q = \tau(q)/(q-1) \quad \text{for } q \neq 1 \tag{6}$$

and

$$D_q = \lim_{\varepsilon\to 0} \frac{Z_{1,\varepsilon}}{\log \varepsilon} \quad \text{for } q = 1 \,, \tag{7}$$

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

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

against $\log \varepsilon$ for $q \neq 1$, and similarly through a linear regression of $Z_{1,\varepsilon}$ against $\log \varepsilon$ for $q = 1$. D_1 is called the *information dimension* and D_2 the *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 coding or noncoding segments which are relatively long. The D_q of the negative values of q deal with the structure and the properties of the most rarefied regions of the measure, i.e., to the segments which are relatively short.

By following the thermodynamic formulation of multifractal measures, Canessa [25] derived an expression for the 'analogous' specific heat as

$$C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1)\,. \tag{8}$$

He showed that the form of C_q resembles a classical phase transition at a critical point for financial time series. In the following we calculate the 'analogous' specific heat of coding and noncoding length sequences for the first time. The types of phase transitions are helpful to discuss the classification of bacteria.

3. Data and results

More than 31 bacterial complete genomes are now available in public databases. There are five Archaebacteria: *Archaeoglobus fulgidus* (aful), *Pyrococcus abyssi* (pabyssi), *Methanococcus jannaschii* (mjan), *Aeropyrum pernix* (aero) and *Methanobacterium thermoautotrophicum* (mthe); five Gram-positive Eubacteria: *Mycobacterium tuberculosis* (mtub), *Mycoplasma pneumoniae* (mpneu), *Mycoplasma genitalium* (mgen), *Ureaplasma urealyticum* (uure), and *Bacillus subtilis* (bsub). The others are Gram-negative Eubacteria, which consist of two Hyperthermophilic bacteria: *Aquifex aeolicus* (aquae) and *Thermotoga maritima* (tmar); three Chlamydia: *Chlamydia trachomatisserovar* (ctra), *Chlamydia muridarum* (ctraM), and *Chlamydia pneumoniae* (cpneu); two Spirochaete: *Borrelia burgdorferi* (bbur) and *Treponema pallidum* (tpal); one Cyanobacterium: *Synechocystis* sp. *PCC6803* (synecho); and 13 Proteobacteria. The 13 Proteobacteria are divided into four subdivisions, which are alpha subdivision: *Rhizobium* sp. *NGR234* (pNGR234) and *Rickettsia prowazekii* (rpxx); gamma subdivision: *Escherichia coli* (ecoli), *Haemophilus influenzae* (hinf), *Xylella fastidiosa* (xfas), *Vibrio cholerae* (vcho1), *Pseudomonas aeruginosa* (paer) and *Buchnera* sp. *APS* (buch); beta subdivision: *Neisseria meningitidis MC58* (nmen) and *Neisseria meningitidis Z2491* (nmenA); epsilon subdivision: *Helicobacter pylori J99* (hpyl99), *Helicobacter pylori 26695* (hpyl) and *Campylobacter jejuni* (cjej).

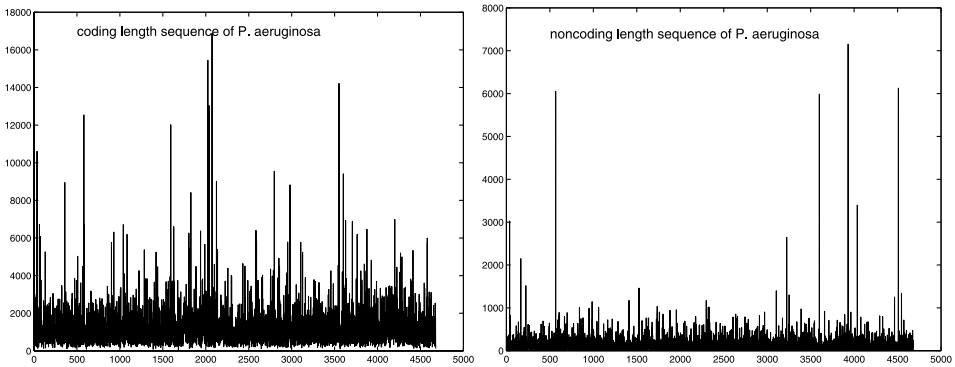

Fig. 1. The coding and noncoding length sequences of *Pseudomonas aeruginosa*.

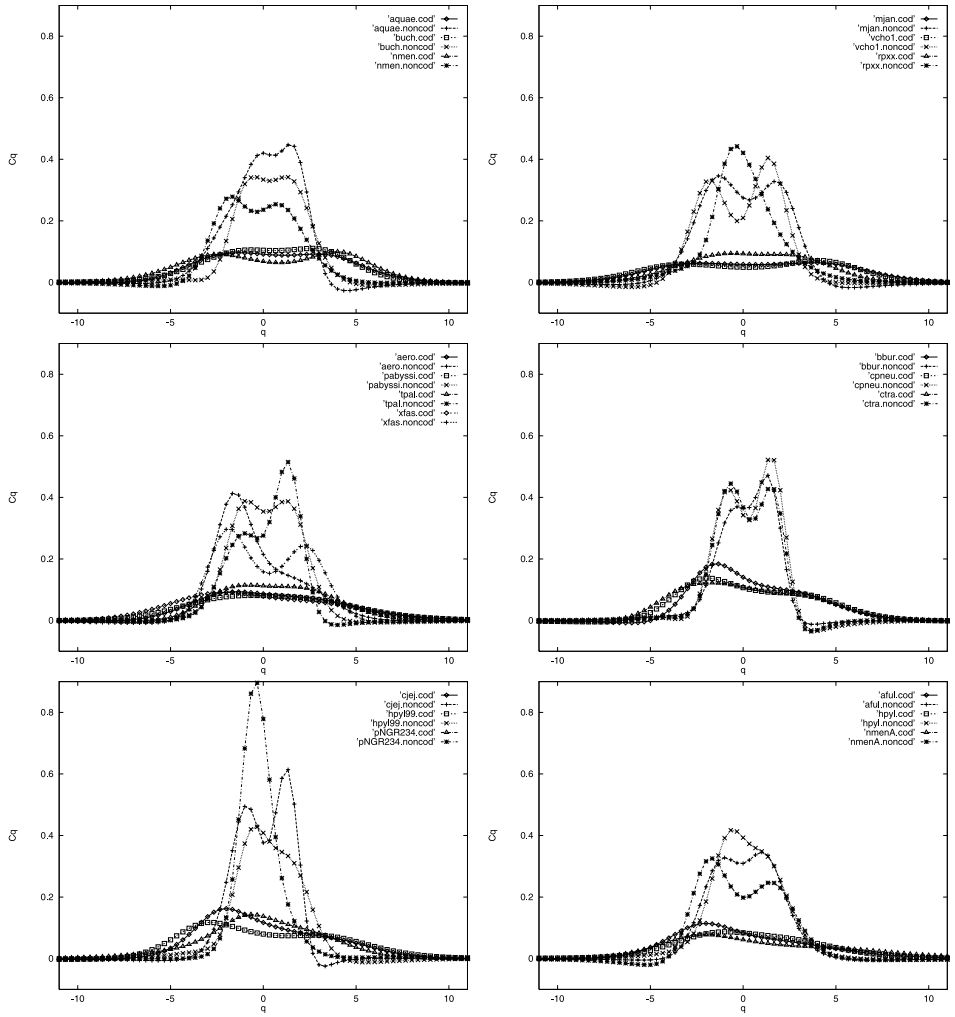

Fig. 2. C_q curves of coding and noncoding length sequences of 19 bacteria.

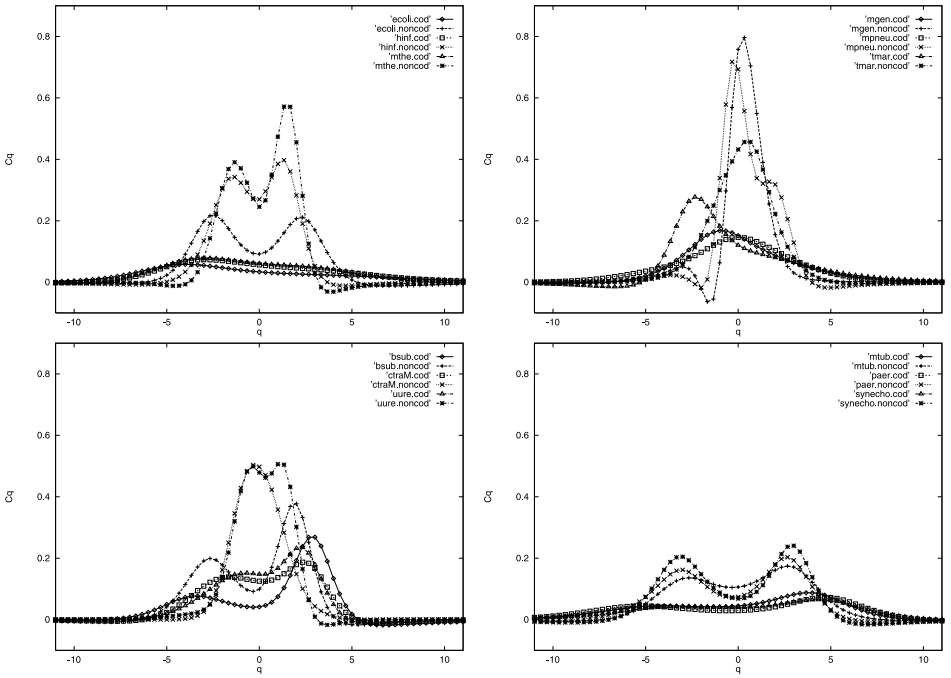

Fig. 3. C_q curves of coding and noncoding length sequences of another 12 bacteria.

First we counted out the length of coding and noncoding segments in the complete genomes of the above bacteria and obtained the coding and noncoding length sequences of these organisms. For example, we give the coding and noncoding length sequences of *Pseudomonas aeruginosa* (paer) in Fig. 1.

Then we calculated the dimension spectra D_q and 'analogous' specific heat C_q of the coding and noncoding length sequences of all the above bacteria according to the methods given in Section 2. In order to show the difference between coding and noncoding length sequences, we give the C_q curves of length sequences of all the above bacteria as Fig. 2 (for 19 bacteria) and Fig. 3 (for another 12 bacteria).

The hill behaviour of the dimension spectrum D_q for $q < 0$ is a well-known fact when using the box-counting method [24,25]. In Figs. 4 and 5, we present D_q of the coding or noncoding length sequences of all bacteria selected within the range $q \geqslant 0$.

4. Discussion and conclusions

If a length sequence is completely random, then our measure definition yields a uniform measure ($D_q = 1,\ C_q = 0$).

From the curves of D_q and C_q, it is seen that there exists a clear difference between the coding/noncoding length sequences of all organisms considered here and

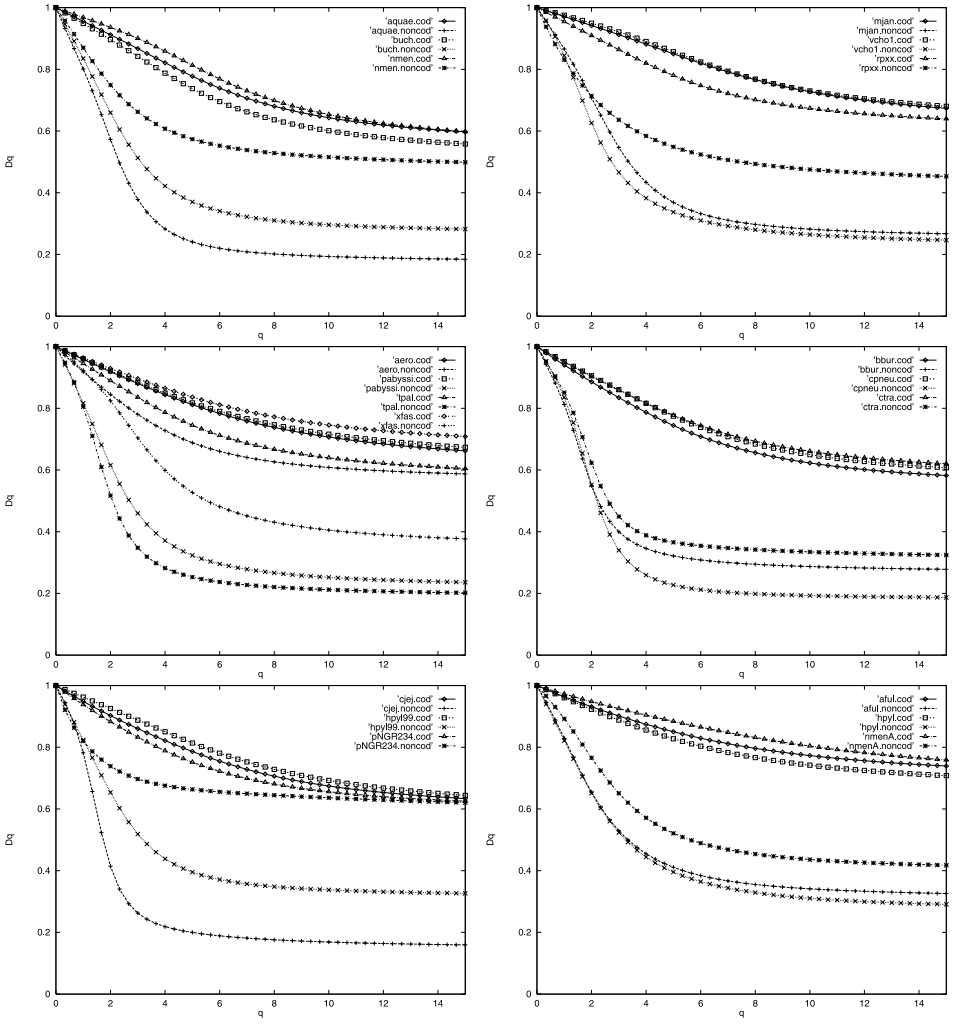

Fig. 4. D_q curves of coding and noncoding length sequences of 19 bacteria.

the completely random sequence. Hence we can conclude that complete genomes are not random sequences. But the D_q values of coding length sequences are closer to 1 than that of noncoding length sequences. In other words, noncoding length sequences are further away from a complete random sequence than coding length sequences. The property of the length sequences is the same as that of the DNA sequences [4].

We also found that for each bacterium selected, the D_q values for $q > 0$ of a noncoding length sequence are smaller than those of a coding length sequence, but for $q < 0$, the situation is reversed. It is well known that the dimension is a measure for complexity. Here the complexity of noncoding length sequences is higher than that of coding length sequences for bacteria.

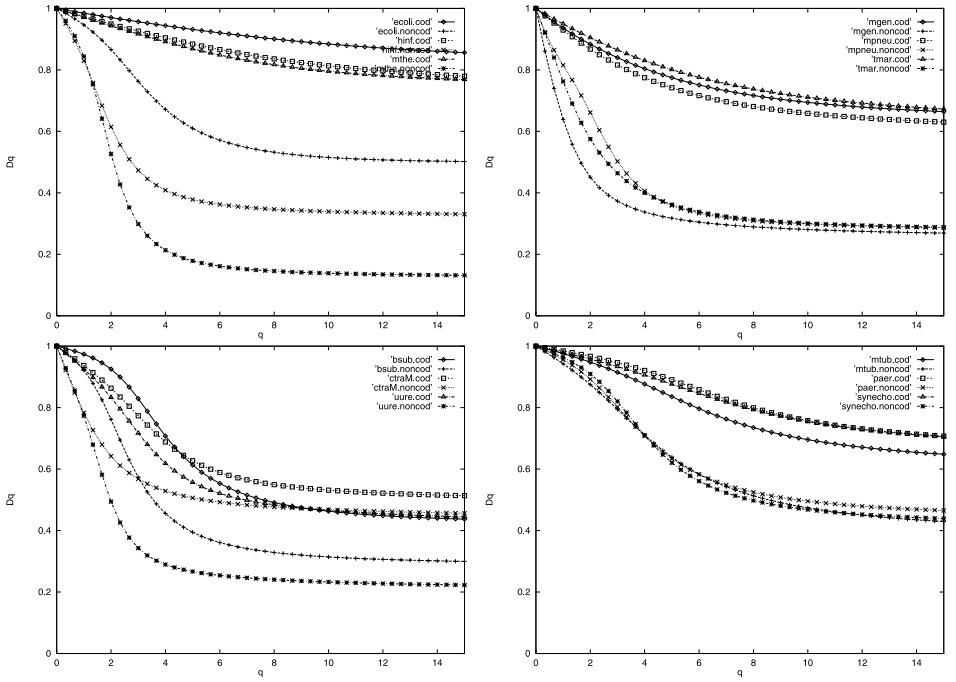

Fig. 5. D_q curves of coding and noncoding length sequences of another 12 bacteria.

From Figs. 4 and 5, almost all D_q curves for coding length sequences are flat, so their multifractality is not pronounced. On the other hand, almost all D_q curves for noncoding length sequences are multifractal-like.

In our previous paper [28], we counted out all substrings with fixed length appearing in the complete genome and gave a measure representation of the complete genome. We found that the shape of the C_q curves of all bacteria we selected are single-peaked. Hence this type of phase transition of the measure representation is not useful for classification of bacteria. On the other hand, from Figs. 2 and 3, one can see that the 'analogous' specific heats of noncoding length sequences of bacteria have a rich variety of behaviours which is much more complex than that of coding length sequences. Some have only one main single peak. In this class, some C_q curves display a shoulder to the right of the main peak, some display a shoulder to the left of the main peak, and some have no shoulder, which resembles a classical (first-order) phase transition at a critical point. In another class, the C_q curves display a balance double peak. So this provides a useful tool for classification of bacteria according to the types of 'analogous' specific heats of the noncoding length sequences. The relevant finding here is that noncoding length sequences display higher C_q peak heights and clear double-peaked structures than coding length sequences. This reveals different types of long-range correlations between the two classes of sequences. This new type of classification allows a better understanding of the relationship among bacteria at the global gene level instead of the

nucleotide sequence level. It can be useful to distinguish between sequence curves as given in the example of Fig. 1.

To conclude, multifractal analysis provides a simple yet powerful method to amplify the difference between a DNA length sequence and a random sequence. In particular, the multifractal characterisation given by the 'analogous' specific heat allows to distinguish DNA length sequences in more detail.

Acknowledgements

One of the authors, Zu-Guo Yu, would like to express his thanks to Prof. Bai-lin Hao of Institute of Theoretical Physics of Chinese Academy of Science for introducing him into this field and continuous encouragement. The authors also thank Dr. Enrique Canessa for many good suggestions and comments to improve this paper. The research was partially supported by QUT's Postdoctoral Research Support Grant No. 9900658 and the RGC Earmarked Grant CUHK 4215/99P.

References

[1] M.S. Gelfand, E.V. Koonin, A.A. Mironov, Prediction of transcription regulatory sites in Archaea by a comparative genomic approach, Nucleic Acids Res. 28 (3) (2000) 695–705.
[2] W. Li, K. Kaneko, Europhys. Lett. 17 (1992) 655.
[3] W. Li, T. Marr, K. Kaneko, Physica D 75 (1994) 392.
[4] C.K. Peng, S. Buldyrev, A.L. Goldberg, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature 356 (1992) 168.
[5] C.A. Chatzidimitriou-Dreismann, D. Larhammar, Nature 361 (1993) 212.
[6] V.V. Prabhu. J.M. Claverie, Nature 359 (1992) 782.
[7] Liaofu Luo, Weijiang Lee, Lijun Jia, Fengmin Ji, Lu Tsai, Phys. Rev. E 58 (1) (1998) 861–871.
[8] Zu-Guo Yu, Guo-Yi Chen, Rescaled range and transition matrix analysis of DNA sequences, Commun. Theor. Phys. 33 (4) (2000) 673–678.
[9] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.K. Peng, M, Simons, H.E. Stanley, Phys. Rev. E 51 (5) (1995) 5084–5091.
[10] R. Voss, Phys. Rev. Lett. 68 (1992) 3805.
[11] R. Voss, Fractals 2 (1994) 1.
[12] A.K. Mohanty, A.V.S.S. Narayana Rao, Phys. Rev. Lett. 84 (8) (2000) 1832–1835.
[13] Zu-Guo Yu, Bai-lin Hao, Hui-min Xie, Guo-Yi Chen, Dimension of fractals related to language defined by tagged strings in complete genome, Chaos, Solitons Fractals 11 (14) (2000) 2215–2222.
[14] Bai-lin Hao, Hoong-Chien Lee, Shu-yu Zhang, Fractals related to long DNA sequences and complete genomes, Chaos Solitons Fractals 11 (6) (2000) 825–836.
[15] Bailin Hao, Huimin Xie, Zuguo Yu, Guoyi Chen, Factorisable language: from dynamics to complete genomes, Physica A 228 (2000) 10–20.
[16] Zu-Guo Yu, Bin Wang, A time series model of CDS sequences on complete genome, Chaos Solitons Fractals 12 (3) (2001) 519.
[17] Maria de Sousa Vieira, Statistics of DNA sequences: a low-frequency analysis, Phys. Rev. E 60 (5) (1999) 5932–5937.
[18] B. Lewin, Genes VI, Oxford University Press, Oxford, 1997.
[19] A. Provata, Y. Almirantis, Fractal Cantor patterns in the sequence structure of DNA, Fractals 8 (1) (2000) 15–27.
[20] Zu-Guo Yu, Vo Anh, Time series model based on global structure of complete genome, Chaos Soliton Fractals 12 (10) (2001) 1827–1834.

[21] A.L. Goldberger, C.K. Peng, J. Hausdorff, J. Mietus, S. Havlin, H.E. Stanley, Fractals and the heart, in: P.M. Iannaccone, M. Khokha (Eds.), Fractal Geometry in Biological Systems, CRC Press, Boca Raton, FL, 1996, pp. 249–266.
[22] Zu-Guo Yu, V.V. Anh, Bin Wang, Correlation property of length sequences based on global structure of complete genome, Phys. Rev. E 63 (2001) 001903.
[23] P. Grassberger, I. Procaccia, Phys. Rev. lett. 50 (1983) 346.
[24] R. Pastor-Satorras, Phys. Rev. E 56 (5) (1997) 5284.
[25] E. Canessa, J. Phys. A: Math. Gen. 33 (2000) 3637.
[26] V.V. Anh, Q.M. Tieng, Y.K. Tse, Cointegration of stochastic multifractals with application to foreign exchange rates, Int. Trans. Oper. Res. 7 (2000) 349–363.
[27] C.L. Berthelsen, J.A. Glazier, S. Raghavachari, Phys. Rev. E 49 (3) (1994) 1860.
[28] Zu-Guo Yu, Vo Anh, Ka-Sing Lau, Measure representation and multifractal analysis of complete genome, Phys. Rev. E 64 (2001) 31 903.
[29] D. Katzen, I. Procaccia, Phys. Rev. Lett. 58 (1987) 1169.
[30] T. Bohr, M. Jensen, Phys. Rev. A 36 (1987) 4904.
[31] J. Lee, H.E. Stanley, Phys. Rev. Lett. 61 (1988) 2945.
[32] T. Halsy, M. Jensen, L. Kadanoff, I. Procaccia, B. Schraiman, Phys. Rev. A 33 (1986) 1141.