Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Theoretical Biology 226 (2004) 341–348

Journal of Theoretical Biology

www.elsevier.com/locate/jtbi

Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses

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

[a] Program in Statistics and Operations Research, Queensland University of Technology, G.P.O. Box 2434, Brisbane QLD 4001, Australia
[b] Department of Mathematics, Xiangtan University, Hunan 411105, China
[c] Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong, China

Received 28 April 2003; accepted 16 September 2003

Abstract

Similar to the chaos game representation (CGR) of DNA sequences proposed by Jeffrey (Nucleic Acid Res. 18 (1990) 2163), a new CGR of protein sequences based on the detailed HP model is proposed. Multifractal and correlation analyses of the measures based on the CGR of protein sequences from complete genomes are performed. The D_q spectra of all organisms studied are multifractal-like and sufficiently smooth for the C_q curves to be meaningful. The C_q curves of bacteria resemble a classical phase transition at a critical point. The correlation distance of the difference between the measure based on the CGR of protein sequences and its fractal background is also proposed to construct a more precise phylogenetic tree of bacteria.

Keywords: Chaos game representation of protein sequences; Multifractal analysis; Correlation analysis; Phylogenetic tree

1. Introduction

Modeling the three-dimensional structure of proteins is a complex physical, chemical and mathematical problem of prime importance in molecular biology, medicine, and pharmacology (Chothia, 1992; Shih et al., 2000). It is believed that the dynamical folding process and stable structure, or native conformation, of a protein are determined by its primary structure, namely its amino acid sequence (Shih et al., 2002). Twenty different kinds of amino acids are found in proteins. The prediction of the high level structures (secondary and space structures) from the amino acid sequence is a challenging problem in science. A well-known model of protein sequences is the HP model (Dill, 1985; Chan and Dill, 1989). In this model, 20 kinds of amino acids are divided into two types, hydrophobic (H) (or non-polar) and polar (P) (or hydrophilic). In recent years the HP model has been extensively studied (e.g. Shih et al., 2000; Li et al., 1996; Wang and Yu, 2000). From studying this model on lattices, Li et al. (1996) found that there are a small number of structures with exceptionally high designability which a large number of protein sequences possess as their ground states. These highly designable structures are found to have protein-like secondary structures (Shih et al., 2000; Li et al., 1996; Micheletti et al., 1998). But the HP model may be simplistic and lacks sufficient information on the heterogeneity and complexity of the natural set of residues (Wang and Wang, 2000). According to Brown (1998, p. 109), one can divide the polar class into three subclasses in the HP model: positive polar, uncharged polar and negative polar. As a result, 20 different kinds of amino acids can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar. This model, which we call the *detailed HP model* (Yu et al., 2002), provides more information than the HP model.

Since the sequencing of the first complete genome of the free-living bacterium *Mycoplasma genitalium* (Fraser et al., 1995), more and more complete genomes have been deposited in public databases such as Genbank at ftp://ncbi.nlm.nih.gov/genbank/genomes/. The complete genomes provide essential information for understanding gene functions and evolution. Retrieval of biological

*Corresponding author. School of Mathematical Sciences, Queensland University of Technology, G.P.O. Box 2434, Brisbane QLD 4001, Australia. Tel.: +61-7-38645194; fax: +61-7-38642310.

E-mail address: yuzg@hotmail.com, z.yu@qut.edu.au (Z.-G. Yu).

0022-5193/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2003.09.009

information from complete genomes and finding the appropriate proteins or coding/non-coding regions of a complete genome for a specific biological problem are some of the challenges for researchers in the bioinformatical field. The determination of patterns in DNA and protein sequences is also useful for many important biological problems such as identifying new genes and discussing the phylogenetic relationships among organisms.

Although statistical analysis performed directly on DNA sequences has yielded some success, there has been an 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 among themselves (Hao et al., 2000a). One needs more useful global and visual methods. Jeffrey (1990) proposed the chaos game representation (CGR) for DNA sequences, and found fractal patterns in these representations. The correlation properties of coding and noncoding DNA sequences were studied by Peng et al. (1992) in their fractal landscape or DNA walk model. Hao et al. (2000a) 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 also found 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 (Yu et al., 2000; Hao et al., 2000b). Yu et al. (2001) introduced a representation of a DNA sequence by a probability measure of K strings derived from the sequence.

Multifractal analysis is a useful way to characterize the spatial heterogeneity of both theoretical and experimental fractal patterns (Hentschel and Procaccia, 1983). A multifractal analysis based on the chaos game representation of DNA sequences was given in Gutierrez et al. (1998, 2001). Based on the measure representation of DNA sequences and the techniques of multifractal analysis, Anh et al. (2002) discussed the problem of recognition of an organism from fragments of its complete genome.

The CGR of DNA sequences has been extended to the representation of protein sequences (amino acid sequences) and protein structures (Fiser et al., 1994; Basu et al., 1997). In this paper, we propose a CGR of protein sequences based on the detailed HP model and then perform multifractal analysis on this new representation.

Works have been done to study the phylogenetic relationships based on correlation analysis of K strings of complete genomes (e.g. Yu and Jiang, 2001) and protein sequences from complete genomes (e.g. Qi et al., 2002; Li et al., 2001). Qi et al. (2002) pointed out that a phylogenetic tree based on protein sequences from complete genomes is more precise than a tree based on complete genomes, and subtracting random background from the probabilities of K strings of protein sequences can improve the phylogenetic tree from the biological point of view. For a given organism, we obtain a measure for each protein sequence using our CGR based on the detailed HP model. Using the observed frequency of each amino acid (i.e. each of the 20 letters in the alphabet), we also generate, through the chaos game algorithm (Barnsley, 1988), a simulation of the original protein sequence, with the same length. Then, through the CGR of the simulated protein sequence, we obtain another measure. In this paper, we propose to use the correlation distance based on the difference of these two measures to discuss the phylogenetic relationships of bacteria.

2. Chaos game representation of protein sequences based on the detailed HP model

The protein sequence is formed by 20 different kinds of amino acids, namely Alanine (A), Arginine (R), Asparagine (N), Aspartic acid (D), Cysteine (C), Glutamic acid (E), Glutamine (Q), Glycine (G), Histidine (H), Isoleucine (I), Leucine (L), Lysine (K), Methionine (M), Phenylalanine (F), Proline (P), Serine (S), Threonine (T), Tryptophan (W), Tyrosine (Y) and Valine (V) (Brown, 1998, p. 109). In the detailed HP model, they can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar. The eight residues A, I, L, M, F, P, W and V designate the non-polar class; the two residues D and E designate the negative polar class; the seven residues N, C, Q, G, S, T and Y designate the uncharged polar class; and the remaining three residues R, H and K designate the positive polar class.

For a given protein sequence $s = s_1 \cdots s_l$ with length l, where s_i is one of the 20 kinds of amino acids for $i = 1, \ldots, l$, we define

$$a_i = \begin{cases} 0 & \text{if } s_i \text{ is non-polar}, \\ 1 & \text{if } s_i \text{ is negative polar}, \\ 2 & \text{if } s_i \text{ is uncharged polar}, \\ 3 & \text{if } s_i \text{ is positive polar}. \end{cases} \tag{1}$$

We then obtain a sequence $X(s) = a_1 \cdots a_l$, where a_i is a letter of the alphabet $\{0, 1, 2, 3\}$. We next define the CGR for a sequence $X(s)$, similar to that of DNA sequences (Jeffrey, 1990), in a square $[0, 1] \times [0, 1]$, where the four vertices correspond to the four letters 0, 1, 2 and 3: the first point of the plot is placed half way between the center of the square and the vertex corresponding to the first letter of the sequence $X(s)$; the i-th point of the plot is then placed half way between the $(i-1)$-th point and the vertex corresponding to the i-th letter. We then call the obtained plot the *CGR* of the protein sequence s based on the detailed HP model.

Each coding sequence in the complete genome of an organism can be translated into a protein sequence using the genetic code (Brown, 1998, p. 122). We next link all translated protein sequences from a complete genome to a long protein sequence according to the order of the coding sequence in the complete genome. As a result, we obtain a linked protein sequence for each organism. In this paper, we only consider this kind of linked protein sequences for the organisms and view them as symbolic sequences. Then the CGR defined above of the linked protein sequence of an organism is called the CGR of the organism. For example, the CGR of *Buchnera* sp. APS is given in Fig. 1. A fractal pattern is apparent in this CGR. Considering the points in a CGR of an organism, we can define a measure μ by $\mu(B) = \sharp(B)/N_l$, where $\sharp(B)$ is the number of points lying in the subset B of the CGR and N_l is the length of the sequence. We can divide the square $[0,1] \times [0,1]$ into meshes of size 64×64, 128×128, 512×512 or 1024×1024. This results in a measure for each mesh. The measure μ based on a 64×64 mesh of *Buchnera* sp. APS is given in Fig. 2 as an example. We then can obtain a 64×64, 128×128, 512×512 or 1024×1024 matrix $\mathscr{A}$, where each element is the measure value on the corresponding mesh. We call $\mathscr{A}$ the *measure matrix* of the organism.

If s' is one of the 20 letters, we denote by $P(s')$ the frequency of the letter s' in the linked protein sequence. A new symbolic sequence $s = s_1, s_2 \cdots s_N$ is next generated, where $s_i = s' \in \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ with probability $P(s')$, $i = 1, \ldots, N$ (N being the length of the linked protein sequence from the complete genome of an organism). The CGR and the corresponding measure matrix $\mathscr{A}^f$ (with the same size as $\mathscr{A}$) can then be obtained for this generated sequence. $\mathscr{A}^f$ is called the *fractal background* of $\mathscr{A}$. We then define a new matrix $\mathscr{A}^d$ as

$$\mathscr{A}^d = \mathscr{A} - \mathscr{A}^f. \tag{2}$$

The matrix $\mathscr{A}^d$ will be used for correlation analysis later in this paper.

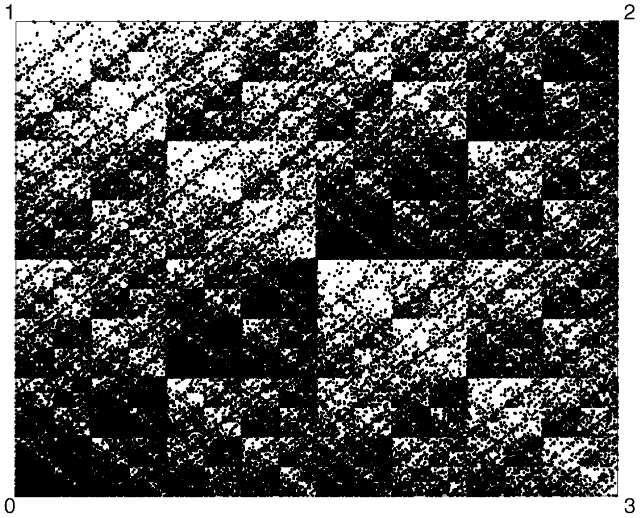

Fig. 1. Chaos game representation of *Buchnera* sp. APS (the linked protein sequence has 185827 amino acids).

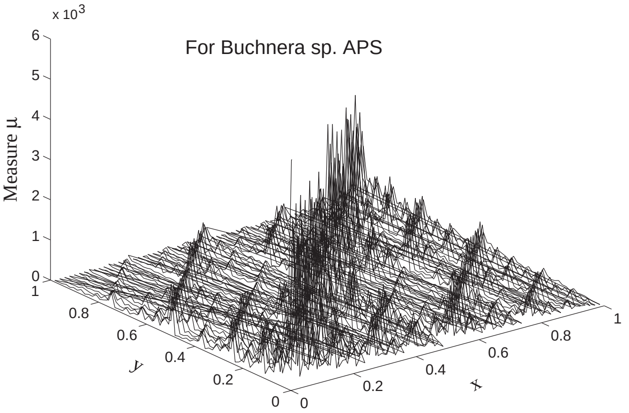

Fig. 2. The measure μ based on 64×64 meshes of *Buchnera* sp. APS.

As noted by Qi et al. (2002), we need to subtract the random background from the sequence $X(s)$ in order to get a good evolutionary tree. Qi et al. (2002) used a Markov model to do this. Here, we use the frequencies of the 20 kinds of amino acids appearing in the linked protein sequence. By the nature of its generation, this probability measure behaves as a multiplicative cascade and displays long memory. Hence, subtracting out the fractal background as described above has the effect of reducing long memory in the measure representation.

3. Multifractal analysis and correlation analysis

The multifractal spectrum of a measure M can be defined using the box-counting method (Halsy et al., 1986) as

$$D_q^{bc} = \lim_{\varepsilon \to 0} \frac{\ln(\sum_i (M_i/M_0)^q)}{\ln(\varepsilon)} \frac{1}{q-1}, \tag{3}$$

where ε is the ratio of the grid size to the linear size of the fractal, M_i the number of points fall in the i-th grid cell, M_0 the total number of points in the fractal.

The above definition is easier to understand, but not so good for the computation of the multifractal spectrum on real data. Tél et al. (1989) introduced a *sandbox* method which is defined by

$$D_q^{sb}(R/L) = \frac{\ln \langle [M(R)/M_0]^{q-1} \rangle}{\ln(R/L)} \frac{1}{q-1}. \tag{4}$$

It is derived from the box-counting method, but has better convergence. The basic idea is that one can randomly choose a point on the fractal, make a sandbox (a region with radius R) around it, then count the number of points of the fractal that fall in this sandbox of radius R, represented as $M(R)$ in the above definition. L is the linear size of the fractal, and q, M_0 have the same meaning as in the definition of D_q^{bc}. The brackets $\langle \bullet \rangle$ mean to take a statistical average over (many) randomly chosen centers of the sandboxes. Because of its dependence on statistical averaging, though the multifractal dimension is defined as $D_q = \lim_{R\to 0} D_q^{sb}(R/L)$, it is better to perform a linear fit on the logarithms of sampled data $\ln(\langle [M(R)]^{q-1}] \rangle)$ vs. $(q-1)\ln(R/L)$ and take its slope as the multifractal dimension in a practical use of the sandbox method. The idea can be illustrated by rewriting Eq. (4) as

$$\ln(\langle [M(R)]^{q-1} \rangle) = D_q^{sb}(R/L) \times (q-1)\ln(R/L) + (q-1)\ln(M_0). \tag{5}$$

First, we choose R in an appropriate range $[R_{\min}, R_{\max}]$. For each chosen R, we compute the statistical average of $[M(R)]^{q-1}$ over many radius-R sandboxes randomly distributed on the fractal, $\langle [M(R)]^{q-1} \rangle$, then plot the data on the $\ln(\langle [M(R)]^{q-1} \rangle)$ vs. $(q-1)\ln(R/L)$ plane. We next perform a linear fit on them and calculate the slope as an approximation of the multifractal dimension D_q. D_1 is called the *information dimension* and D_2 the *correlation dimension of the measure*. The D_q values for positive values of q are associated with the regions where the points are dense. The D_q values for negative values of q are associated with the structure and properties of the most rarefied regions. In addition to the multifractal dimension D_q, there is another exponent $\tau(q)$. One can calculate $\tau(q)$ from D_q by $\tau(q) = (q-1)\mathrm{D}_q$.

Some sets of physical interest have a non-analytic dependence of D_q on q. Moreover, this phenomenon has a direct analogy to the phenomenon of phase transitions in condensed-matter physics (Katzen and Procaccia, 1987). The existence and type of phase transitions might turn out to be a worthwhile characterization of universality classes for structures (Bohr and Jensen, 1987). 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 (Lee and Stanley, 1998). By following the thermodynamic formulation of multifractal measures, Canessa (2000) derived an expression for the "analogous" specific heat as

$$C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1). \tag{6}$$

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 the measure μ defined in Section 2.

For matrices $\mathscr{A}^d = (a_{ij})_{n\times n}$ and $\mathscr{B}^d = (b_{ij})_{n\times n}$ (defined in Section 2) of two different organisms, let

$$\langle \mathscr{A}^d \rangle = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}, \quad \langle \mathscr{B}^d \rangle = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij},$$

$$\delta(\mathscr{A}^d) = \sqrt{\frac{1}{n^2}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n}(a_{ij} - \langle \mathscr{A}^d \rangle)^2},$$

$$\delta(\mathscr{B}^d) = \sqrt{\frac{1}{n^2}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n}(b_{ij} - \langle \mathscr{B}^d \rangle)^2}.$$

Then we can view a_{ij} and b_{ij} as the sample values of random variables X_1 and X_2, respectively. Hence the covariance of X_1 and X_2 is

$$\mathrm{Cov}(\mathscr{A}^d, \mathscr{B}^d) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij} - \langle \mathscr{A}^d \rangle)(b_{ij} - \langle \mathscr{B}^d \rangle).$$

As a result, we obtain the correlation coefficient between X_1 and X_2 as

$$\rho(\mathscr{A}^d, \mathscr{B}^d) = \frac{\mathrm{Cov}(\mathscr{A}^d, \mathscr{B}^d)}{\delta(\mathscr{A}^d)\delta(\mathscr{B}^d)}. \tag{7}$$

We have $-1 \leqslant \rho(\mathscr{A}^d, \mathscr{B}^d) \leqslant 1$. If it is equal to zero, X_1 and X_2 are uncorrelated. We next define the *correlation distance* between these two organisms by

$$\mathrm{Dist}(\mathscr{A}^d, \mathscr{B}^d) = \frac{1 - \rho(\mathscr{A}^d, \mathscr{B}^d)}{2}. \tag{8}$$

4. Data and results

Currently, there are more than 50 complete genomes of Archaea and Eubacteria available in public databases (for example in Genbank at the web site ftp://ncbi.nlm.nih.gov/genbank/genomes/). These include eight **Archae Euryarchaeota**: *Archaeoglobus fulgidus* DSM4304 (Aful), *Pyrococcus abyssi* (Paby), *Pyrococcus horikoshii* OT3 (Phor), *Methanococcus jannaschii* DSM2661 (Mjan), *Halobacterium* sp. NRC-1 (Hbsp), *Thermoplasma acidophilum* (Taci), *Thermoplasma volcanium* GSS1 (Tvol), and *Methanobacterium thermoautotrophicum* deltaH (Mthe); two **Archae Crenarchaeota**: *Aeropyrum pernix* (Aero) and *Sulfolobus solfataricus* (Ssol); three **Gram-positive Eubacteria (high G + C)**: *Mycobacterium tuberculosis* H37Rv (MtubH), *Mycobacterium tuberculosis* CDC1551 (MtubC) and *Mycobacterium leprae* TN (Mlep); 12 **Gram-positive Eubacteria (low G + C)**: *Mycoplasma pneumoniae* M129 (Mpne), *Mycoplasma genitalium* G37 (Mgen), *Mycoplasma pulmonis* (Mpul), *Ureaplasma urealyticum* (serovar 3)(Uure), *Bacillus subtilis* 168 (Bsub), *Bacillus halodurans* C-125 (Bhal), *Lactococcus lactis* IL 1403 (Llac), *Streptococcus pyogenes* M1 (Spyo), *Streptococcus pneumoniae* (Spne), *Staphylococcus aureus* N315

(SaurN), *Staphylococcus aureus* Mu50 (SaurM), and *Clostridium acetobutylicum* ATCC824 (CaceA). The others are **Gram-negative Eubacteria**, which consist of two **hyperthermophilic bacteria**: *Aquifex aeolicus* (Aqua) VF5 and *Thermotoga maritima* MSB8 (Tmar); four **Chlamydia**: *Chlamydia trachomatis* (serovar D) (Ctra), *Chlamydia pneumoniae* CWL029 (Cpne), *Chlamydia pneumoniae* AR39 (CpneA) and *Chlamydia pneumoniae* J138 (CpneJ); two **Cyanobacterium**: *Synechocystis* sp. PCC6803 (Syne) and *Nostoc sp*. PCC6803 (Nost); two **Spirochaete**: *Borrelia burgdorferi* B31 (Bbur) and *Treponema pallidum* Nichols (Tpal); and 16 **Proteobacteria**. The 16 Proteobacteria are divided into four subdivisions, which are **alpha subdivision**: *Mesorhizobium loti* MAFF303099 (Mlot), *Sinorhizobium meliloti* (smel), *Caulobacter crescentus* (Ccre) and *Rickettsia prowazekii* Madrid (Rpro); **beta subdivision**: *Neisseria meningitidis* MC58 (NmenM) and *Neisseria meningitidis* Z2491 (NmenZ); **gamma subdivision**: *Escherichia coli* K-12 MG1655 (EcolK), *Escherichia coli* O157:H7 EDL933 (EcolO), *Haemophilus influenzae* Rd (Hinf), *Xylella fastidiosa* 9a5c (Xfas), *Pseudomonas aeruginosa* PA01 (Paer), *Pasteurella multocida* PM70 (Pmul) and *Buchnera* sp. APS (Buch); and **epsilon subdivision**: *Helicobacter pylori* J99 (HpylJ), *Helicobacter pylori* 26695 (Hpyl) and *Campylobacter jejuni* (Cjej).

We downloaded the long protein sequences from the complete genomes of the above bacteria and calculated the dimension spectra and "analogous" specific heat of the measure μ from their CGRs. As an illustration, we plot the D_q curves of the measure μ in Fig. 3 and the C_q curves of the measure μ in Fig. 4. Because all the D_q are equal to 2 if the CGR of a protein sequence is completely random, it is apparent from the plots that the D_q and C_q curves are nonlinear and significantly different from those of completely random sequences. Hence the CGRs of linked protein sequences from complete genomes are not completely random sequences.

From the plot of D_q, the dimension spectra of the measure μ exhibit a multifractal-like form.

If only a few organisms 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 found it more informative to use C_1 and C_2 in conjunction with the two-dimensional vectors (C_1, C_2). The distribution of the vectors (C_1, C_2) also shows some patterns useful for classification. We show the result for the measure μ in Fig. 5.

But the above result using (C_1, C_2) is still not precise enough to yield a satisfactory phylogenetic relationship

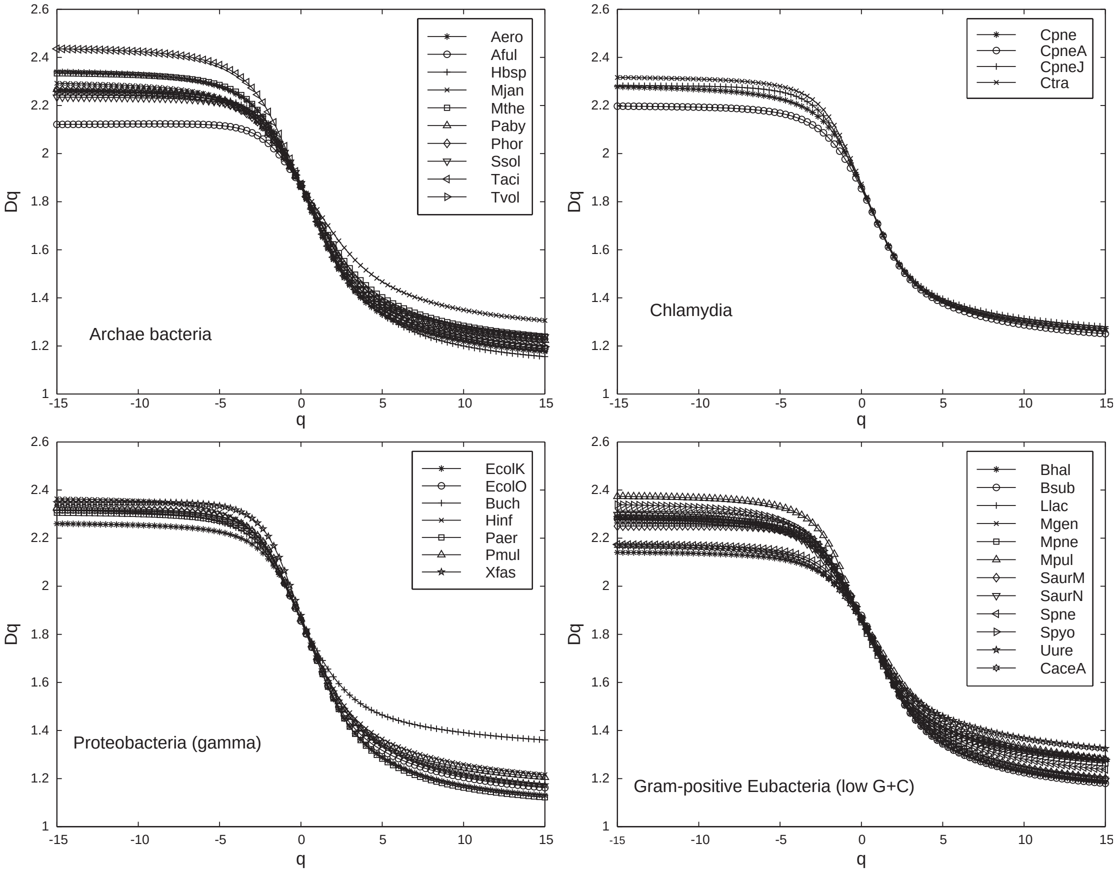

Fig. 3. Dimension spectra of measure μ from the CGRs of protein sequences of some organisms.

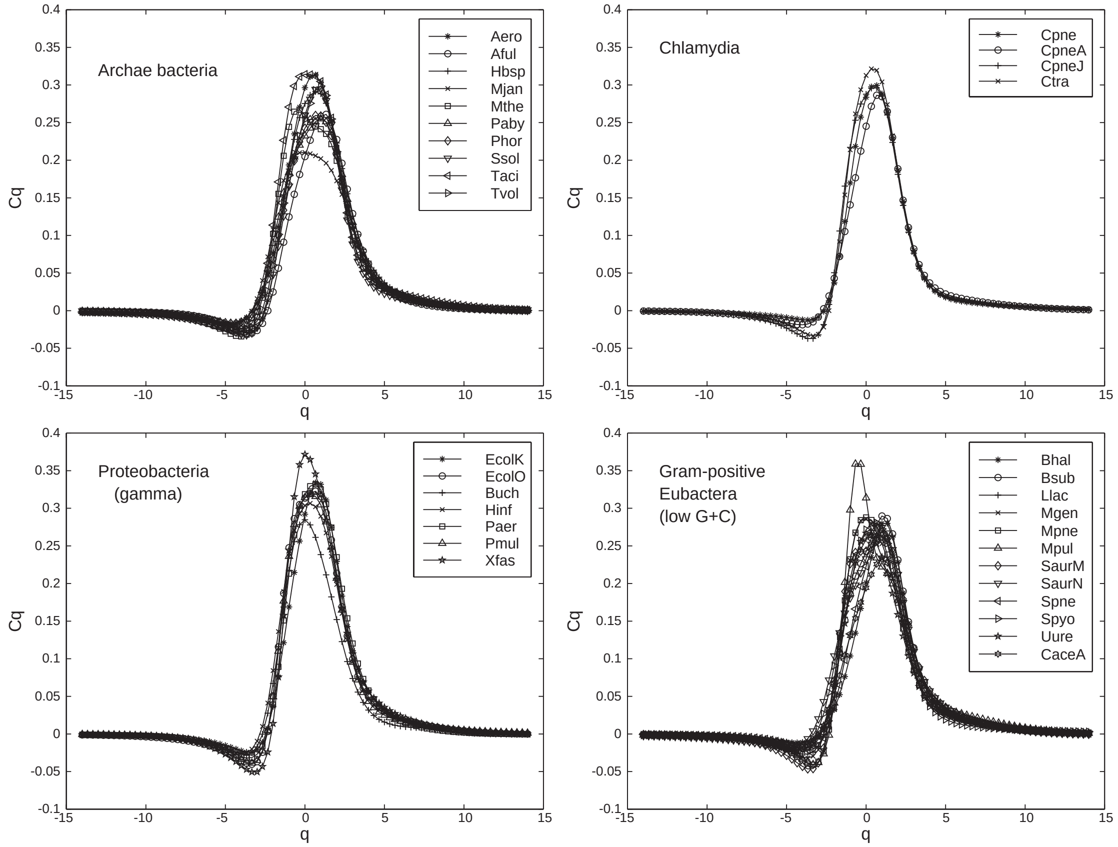

Fig. 4. “Analogous” specific heat of measure μ from CGRs of protein sequences of some organisms.

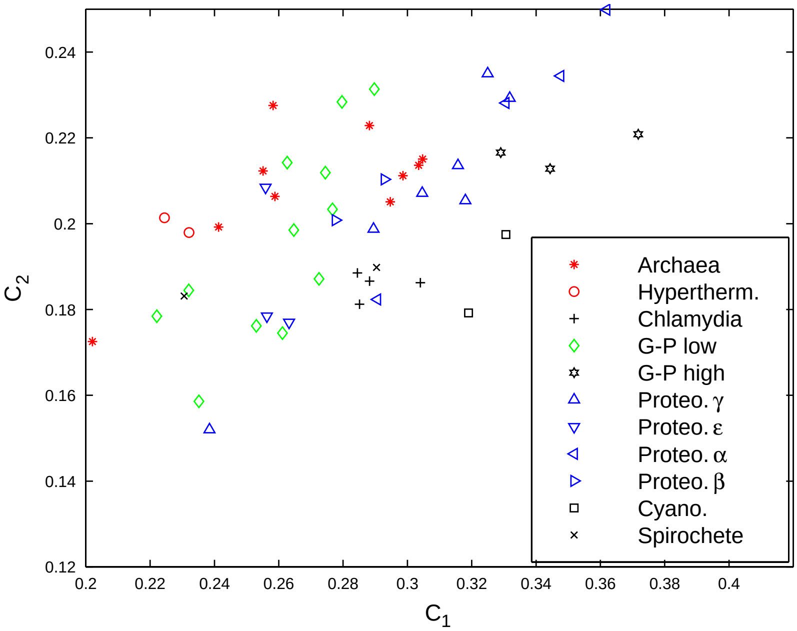

Fig. 5. Distribution of two-dimensional points (C_1, C_2) of organisms selected.

for the organisms selected. For this purpose, we used the distance matrices from the correlation analysis to construct the phylogenetic tree with the help of the neighbor-joining program in the PHYLIP package of Felsenstein (The Phylip software, http://evolution.genetics.washington.edu/phylip.html). We found that the phylogenetic tree based on the correlation distance becomes more precise with the increasing size of the

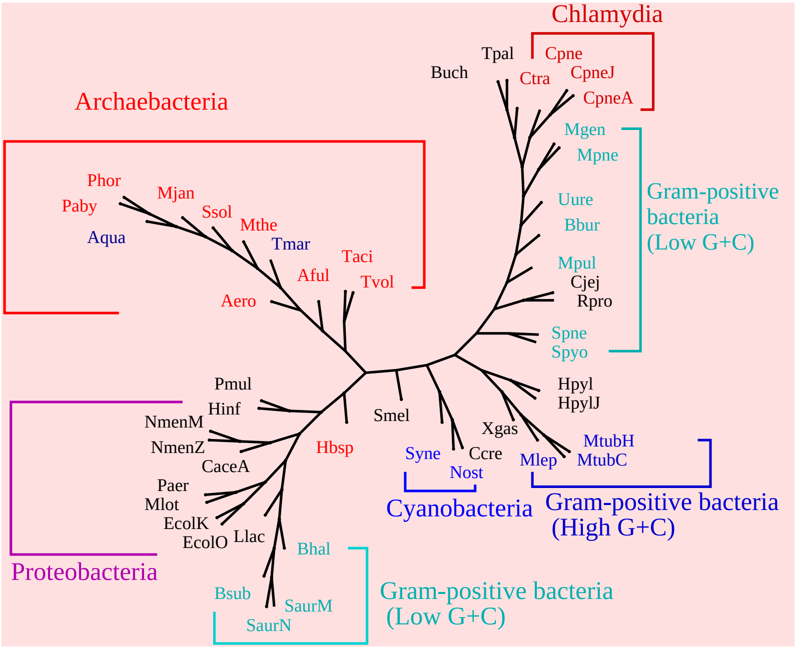

Fig. 6. The neighbor-joining phylogenetic tree based on the correlation distance using matrix $\mathscr{A}^d$ with size 1024 × 1024.

matrix $\mathscr{A}^d$ (we tried the sizes 64 × 64, 128 × 128, 512 × 512 and 1024 × 1024). The phylogenetic tree using matrix $\mathscr{A}^d$ with size 1024 × 1024 is given in Fig. 6.

5. Discussion and conclusions

The frequent errors exist in identified proteins in genomes. Because of the size of the genomes, these errors may not have large effect overall on the results of our study.

The chaos game representation based on the detailed HP model of protein sequences provides a simple yet powerful visualization method to distinguish protein sequences themselves in more details. From the D_q and C_q curves, it is concluded that the point sequences in the CGR of all organisms considered here are not completely random. It should be noted that we cannot conclude that the protein sequences are not completely random since we used the detailed HP model. We also found that the C_q curves of all studied bacteria resemble a classical phase transition at a critical point as shown in Fig. 4.

Although the existence of the archaebacterial kingdom has been accepted by many biologists, the classification of bacteria is still a matter of controversy (Iwabe et al., 1989). The evolutionary relationship of the three primary kingdoms, namely archeabacteria, eubacteria and eukaryote, is another crucial problem that remains unresolved (Iwabe et al., 1989).

Fig. 5 shows some patterns which are helpful for the classification problem. In fact, the points corresponding to organisms from the same category are not far from each other. But the multifractal analysis is still not sufficient to give a satisfactory phylogenetic relationship for the organisms selected. The correlation distance using the matrix $\mathscr{A}^d$ with the fractal background removed from the original CGR gives a more satisfactory phylogenetic tree. Fig. 6 shows all Archaebacteria except *Halobacterium* sp. NRC-1 (Hbsp) staying in a separate branch with the Eubacteria. The bacteria in the Chlamydia, Cyanobacteria and Gram-positive (high G + C) groups gather together, respectively. So at the general global level of complete genomes, our result supports the genetic annealing model for the universal ancestor (Woese, 1998). Furthermore, the two hyperthermophilic bacteria *Aquifex aeolicus* (Aqua) VF5 and *Thermotoga maritima* MSB8 (Tmar) stay in the Archaebacteria branch, we noticed that these two bacteria, like most Archaebacteria, are hyperthermophilic. It has previously been shown that Aquifex has close relationship with Archaebacteria from a gene comparison of an enzyme needed for the synthesis of the amino acid trytophan (Pennisi, 1998).

Qi et al. (2002) pointed out that the subtraction of random background is an essential step in their method. Our results show that subtraction of the fractal background is also an essential step in our correlation method. The correlation analysis is more precise than the multifractal analysis for the phylogenetic problem. Similar to the method in (Qi et al., 2002), lateral gene transfer (Lawrence and Ochman, 1998) might not affect our results since the correlation method also does not depend on the choice of one or another gene.

The phase transition-like phenomenon in the C_q curves can indicate the complexity of organisms. In our previous research work (Yu et al., 2001, 2003), we found that the C_q curves of bacteria resemble a classical phase transition. But the C_q curves of eukaryotes studied exhibit the shape of double-peaked phase transition. From the phase transition theory, we can say eukaryotes are more complex than bacteria. It coincide the conclusion in the biological theory.

Acknowledgements

One of the authors, Zu-Guo Yu, expresses his gratitude to Prof. Bai-lin Hao and Dr. Ji Qi of Institute of Theoretical Physics of the Chinese Academy of Science for useful discussion on the phylogenetic problem. This work was supported by Grant No. 10101022 of the Youth Foundation of National Natural Science Foundation of China, QUT Postdoctoral Research Support Grant No. 9900658, the Australian Research Council Grant A10024117 and the HKRGC Earmark Grant CUHK 4215/99P.

References

Anh, V.V., Lau, K.S., Yu, Z.G., 2002. Recognition of an organism from fragments of its complete genome. Phys. Rev. E 66, 031910.

Barnsley, M.F., 1988. Fractal Everywhere. Springer, Berlin, New York.

Basu, S., Pan, A., Dutta, C., Das, J., 1997. Chaos game representation of proteins. J. Mol. Graphics Model. 15, 279–289.

Bohr, T., Jensen, M., 1987. Order parameter, symmetry breaking, and phase transitions in the description of multifractal sets. Phys. Rev. A 36, 4904–4915.

Brown, T.A., 1998. Genetics, 3rd Edition. Chapman & Hall, London.

Canessa, E., 2000. Multifractality in time series. J. Phys. A: Math. Nucl. Gen. 33, 3637–3651.

Chan, H.S., Dill, K.A., 1989. Compact polymers. Macromolecules 22, 4559–4573.

Chothia, C., 1992. One thousand families for the molecular biologist. Nature (London) 357, 543–544.

Dill, K.A., 1985. Theory for the folding and stability of globular proteins. Biochemistry 24, 1501–1509.

Fiser, A., Tusnady, G.E., Simon, I., 1994. Chaos game representation of protein structures. J. Mol. Graphics 12, 302–304.

Fraser, C.M., et al., 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270, 397–404.

Gutierrez, J.M., Iglesias, A., Rodriguez, M.A., 1998. Analyzing the multifractal structure of DNA nucleotide sequences. In: Barbi, M., Chillemi, S. (Eds.), Chaos and Noise in Biology and Medicine. World Scientific, Singapore, pp. 315–319.

Gutierrez, J.M., Rodriguez, M.A., Abramson, G., 2001. Multifractal analysis of DNA sequences using novel chaos-game representation. Physica A 300, 271–284.

Halsy, T., Jensen, M., Kadanoff, L., Procaccia, I., Schraiman, B., 1986. Fractal measures and their singularities: the characterization of strange set. Phys. Rev. A 33, 1141–1151.

Hao, B.L., Lee, H.C., Zhang, S.Y., 2000a. Fractals related to long DNA sequences and complete genomes. Chaos, Solitons Fractals 11 (6), 825–836.

Hao, B.L., Xie, H.M., Yu, Z.G., Chen, G.Y., 2000b. Avoided strings in bacterial complete genomes and a related combinatorial problem. Ann. Combin. 4, 247–255.

Hentschel, H.G.E., Procaccia, I., 1983. The infinite number of generalized dimensions of fractals and stranger attractors. Physica D 8, 435–444.

Iwabe, N., et al., 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA 86, 9355–9359.

Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18, 2163–2170.

Katzen, D., Procaccia, I., 1987. Phase transitions in the thermodynamic formalism of multifractals. Phys. Rev. Lett. 58, 1169–1172.

Lawrence, J.G., Ochman, H., 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA 95, 9413–9417.

Lee, J., Stanley, H.E., 1998. Phase transition in the multifractal spectrum of diffusion-limited aggregation. Phys. Rev. Lett. 61, 2945–2948.

Li, H., Helling, R., Tang, C., Wingreen, N.S., 1996. Emergence of preferred structures in a simple model of protein folding. Science 273, 666–669.

Li, M., et al., 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17, 149–154.

Micheletti, C., Banavar, J.R., Maritan, A., Seno, F., 1998. Steric constraints in model proteins. Phys. Rev. Lett. 80, 5683–5686.

Peng, C.K., Buldyrev, S., Goldberg, A.L., Havlin, S., Sciortino, F., Simons, M., Stanley, H.E., 1992. Long-range correlations in nucleotide sequences. Nature 356, 168.

Pennisi, E., 1998. Genome data shake tree of life. Science 280, 672–674.

Qi, J., Wang, B., Hao, B.L., 2002. Prokaryote phylogeny based on complete genomes–tree construction without sequence alignment. J. Mol. Evol. (in press).

Shih, C.T., Su, Z.Y., Gwan, J.F., Hao, B.L., Hsieh, C.H., Lee, H.C., 2000. The HP model, designability, and alpha-helices in protein structures. Phys. Rev. Lett. 84 (2), 386–389.

Shih, C.T., Su, Z.Y., Gwan, J.F., Hao, B.L., Hsieh, C.H., Lee, H.C., 2002. Geometric and statistical properties of the mean-field HP model, the large–small model and real protein sequences. Phys. Rev. E 65, 041923.

Tél, T., Fülöp, Á., Vicsek, T., 1989. Determination of fractal dimensions for geometrical multifractals. Physica A 159, 155–166.

Wang, J., Wang, W., 2000. Modeling study on the validity of a possibly simplified representation of proteins. Phys. Rev. E 61, 6981–6986.

Wang, B., Yu, Z.G., 2000. One way to characterize the compact structures of lattice protein model. J. Chem. Phys. 112, 6084–6088.

Woese, C.R., 1998. The universal ancestor. Proc. Natl. Acad. Sci. USA 95, 6854–6859.

Yu, Z.G., Jiang, P., 2001. Distance, correlation and mutual information among portraits of organisms based on complete genome. Phys. Lett. A 286, 34–46.

Yu, Z.G., Hao, B.L., Xie, H.M., Chen, G.Y., 2000. Dimension of fractals related to language defined by tagged strings in complete genome. Chaos, Solitons Fractals 11 (14), 2215–2222.

Yu, Z.G., Anh, V.V., Lau, K.S., 2001. Measure representation and multifractal analysis of complete genomes. Phys. Rev. E 64, 031903.

Yu, Z.G., Anh, V.V., Lau, K.S., 2002. Fractal analysis of measure representation of large proteins based on the detailed HP model, submitted for publication.

Yu, Z.G., Anh, V.V., Lau, K.S., 2003. Multifractal and correlation analyses of protein sequences from complete genomes. Phys. Rev. E 68, 021913.