Phylogenetic tree of prokaryotes based on complete genomes using fractal and correlation analyses

Zu-Guo Yu[1,2] **and Vo Anh**[1]

[1]Program in Statistics and Operations Research, Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia.
[2]Department of Mathematics, Xiangtan University, Hunan 411105, China.
Email: z.yu@qut.edu.au (Zu-Guo Yu), v.anh@qut.edu.au (Vo Anh)

Abstract

We develop a fast algorithm for deriving species phylogeny based on the measure representations of DNA sequences and protein sequences proposed in our previous papers (Yu *et al.*, *Phys. Rev. E* **64**, 031903 (2003); *Phys. Rev. E* **68**, 021913 (2003)). Due to the way they are constructed, these two measures will be treated as a random multiplicative cascades. Such multiplicative cascades commonly have built-in multifractal structures. In this paper we propose to use an iterated function system (IFS) model to simulate the multifractal structures. After removing the mulifractal structures from the original measures, the two kinds of obtained sequences will become stationary time series suitable for cross-correlation analysis. We then can define two kinds of correlation distances between two organisms using these obtained sequences respectively. Using a large data set of prokaryote genomes, we produce two species trees that are largely in agreement with previously published trees using different methods. These trees also agree well with currently accepted phylogenetic theory.

Keywords: Measure representation of DNA and protein sequences, Iterated function system (IFS), Phylogenetic tree.

1 Introduction

Since the sequencing of the first complete genome of the free-living bacterium *Mycoplasma genitalium* in 1995 (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/. Complete genomes provide essential information for understanding gene functions and evolution. To be able to determine the patterns of DNA and protein sequences is very useful for studying many important biological problems such as identifying new genes and establishing the phylogenetic relationship among organisms.

A DNA sequence is formed by four different nucleotides, namely, adenine (a), cytosine (c), guanine (g) and thymine (t). A protein sequence is formed by twenty 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,Page 109). The protein sequences from complete genomes are translated from their coding sequences (DNA) through the genetic code (Brown 1998, Page 122).

A useful result is the establishment of long memory in DNA sequences (Li et al.1994; Peng et al. 1992; Chatzidimitriou-Dreismann and Larhammar 1993; Prabhu and Claverie 1992). Li *et al.* (1994) found that the spectral density of a DNA sequence containing mostly introns shows $1/f^{\beta}$ behaviour, which indicates the presence of long-range correlation when $0 < \beta < 1$. The correlation properties of coding and noncoding DNA sequences were also studied by Peng *et al.* (1992) in their fractal landscape or DNA walk model. Peng *et al.* (1992) 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 Larhammar (1993) concluded that both coding and noncoding sequences exhibit long-range correlation. A subsequent work by Prabhu and Claverie (1992) also corroborated these results. From a different angle, fractal analysis is a relatively new analytical technique that has proved useful in revealing complex patterns in natural phenomena. Berthelsen *et al.* (1992) considered the global fractal dimension of human DNA sequences treated as pseudorandom walks. Vieira (1999) 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 the presence of anti-persistence.

Although statistical analyses performed directly on DNA sequences have 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 (Hao et al. 2000). One needs more powerful global and visual methods. For this purpose, Hao *et al.* (2000) proposed a visualisation 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 (Yu et al. 2000; Hao et al. 2001). There are other graphical methods of sequence patterns, such as the chaos game representation (Jeffrey 1990; Goldman 1993).

Yu *et al.* (2001) introduced a representation of a DNA sequence by a probability measure of K-strings derived from the sequence. This probability measure is in fact the histogram of the events formed by all the K-strings in a dictionary ordering. Anh *et al.* (2001)

took a further step in providing a theory to characterise the multifractality of the probability measures of complete genomes. Based on the measure representation of DNA sequence and the technique of multifractal analysis (Yu et al. 2001; Grassberger and Procaccia 1983), Anh *et al.* (2002) discussed the problem of recognition of an organism from fragments of its complete genome. The significant self-similarity between genome fragments was also found (Anh *et al.* 2002) if the fragment is not too short.

There have been a number of recent attempts to develop methodologies that do not require sequence alignment for deriving species phylogeny based on overall similarities of complete genome data (Stuart et al. 2002, Li 2001). Works have been done to study the phylogenetic relationship based on correlation analyses of the K-strings of complete genomes (Yu and Jiang 2001; Yu et al. 2003a) and protein sequences from complete genomes (Qi et al. 2003; Chu et al. 2003). Qi *et al.* (2003) pointed out that a phylogenetic tree based on the protein sequences from complete genomes is more precise than a tree based on the complete genomes (DNA) themselves, and removing the random background from the probabilities of K-strings of protein sequences can improve a phylogenetic tree from the biological point of view.

Recently Yu *et al.* (2003) introduced the notion of measure representation of protein sequences similar to that for DNA sequences introduced in (Yu et al 2001) and their multifractal analysis. Due to the way the measure representations of DNA (Yu et al. 2001) and protein sequences (Yu et al 2003) are constructed, the measures can be treated as random multiplicative cascades. Such multiplicative cascades commonly have built-in multifractal structures. In this paper we propose to use an iterated function system (IFS) model to simulate the multifractal structures. After removing the mulifractal structure from the original measures, two kinds of obtained sequences will become stationary time series suitable for cross-correlation analysis. So we can define two kinds of correlation distances between two organisms using these obtained sequences respectively. Then two phylogenetic trees are constructed based on their correlation analyses.

2 Measure representations of DNA and protein sequences

We first outline the method of Yu *et al.* (2001,2003) in deriving the measure representation of DNA and protein sequences. First we link all coding sequences from a complete genome to form a long DNA sequence according to the order of the coding sequences in the complete genome. Secondly each coding sequence in the complete genome of an organism is translated into a protein sequence using the genetic code (Brown 1998,Page 122). We then link all translated protein sequences from a complete genome to form a long protein sequence according to the order of the coding sequences in the complete genome. In this way, we obtain a linked coding sequence and a linked protein sequence for each organism. In this paper we only consider these kinds of linked coding and protein sequences and view them as symbolic sequences.

We call any string made of K letters from the alphabet $\{g, c, a, t\}$ or $\{A,\ C,\ D,\ E,\ F,\ G,\ H,\ I,\ K,\ L,\ M,\ N,\ P,\ Q,\ R,\ S,\ T,\ V,\ W,\ Y\}$ which corresponds to four kinds of nuleotides or twenty kinds of amino acids a K-string. For a given K there are in total 4^K or 20^K different K-strings in DNA or protein sequences respectively. In order to count the number of each kind of K-strings in a given protein sequence, 4^K or 20^K counters are needed. We divide the interval $[0, 1[$ into 4^K (for a DNA sequence) or 20^K (for a protein sequence) disjoint subintervals, and use each subinterval to represent a counter.

In the DNA sequence case, letting $s = s_1 \cdots s_K$, $s_i \in \{a, c, g, t\}, i = 1, \cdots, K$, be a substring with length K, we define

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

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

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

We then use the subinterval $[x_l^c(s), x_r^c(s))$ to represent substring s. Let $N(s)$ be the times the substring s appears in the linked coding sequence and $N_K(total)$ the total number of times that all substrings with length K appear in the linked coding sequence (we use an open reading frame and slide one position each time to count the times; $N_K(s)$ may be zero). We define

$$F_K^c(s) = N_K(s)/N_K(total) \tag{4}$$

to be the frequency of substring s. It follows that $\sum_{\{s\}} F^c(s) = 1$. We now define a measure μ_K^c on $[0, 1)$ by $\mu_K^c(x) = Y_K(x)\, dx$,where

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

We then have $\mu_K^c([0, 1[) = 1$ and $\mu_K^c([x_l^c(s), x_r^c(s)[) = F_K^c(s)$. We call $\mu_K^c(x)$ the *measure representation* of the linked coding sequence of the organism corresponding to the given K. As an example, a fragment of the histogram of substrings in the linked coding sequence of *Buchnera sp. APS* for $K = 11$ is given in the top figure of Figure 1.

For protein sequences, letting $s = s_1 \cdots s_K$, $s_i \in \{A,\ C,\ D,\ E,\ F,\ G,\ H,\ I,\ K,\ L,\ M,\ N,\ P,\ Q,\ R,\ S,\ T,\ V,\ W,\ Y\}$, $i = 1, \cdots, K$, be a substring with length K, we define

$$x_l^p(s) = \sum_{i=1}^{K} \frac{x_i}{20^i}, \tag{6}$$

where x_i is one of the integer values from 0 to 19 corresponding to $s_i = A,\ C,\ D,\ E,\ F,\ G,\ H,\ I,\ K,\ L,\ M,\ N,\ P,\ Q,\ R,\ S,\ T,\ V,\ W,\ Y$ respectively, and

$$x_r^p(s) = x_l^p(s) + \frac{1}{20^K}. \tag{7}$$

We then use the subinterval $[x_l^p(s), x_r^p(s)[$ to represent substring s. Let $N_K(s)$ be the number of times that substring s with length K appears in the linked protein sequence and $N_K(total)$ the total number of times that all substrings with length K appear in the linked protein sequence (we use an open reading frame and slide one position each time to count the times; $N_K(s)$ may be zero). We define

$$F_K^p(s) = N_K(s)/N_K(total) \tag{8}$$

to be the frequency of substring s. It follows that $\sum_{\{s\}} F_K^p(s) = 1$. We define a measure μ_K^p on $[0, 1[$ by $d\mu_K^p(x) = Y_K(x)dx$, where

$$Y_K(x) = 20^K F_K^p(s), \quad \text{when} \quad x \in [x_l^p(s), x_r^p(s)[. \tag{9}$$

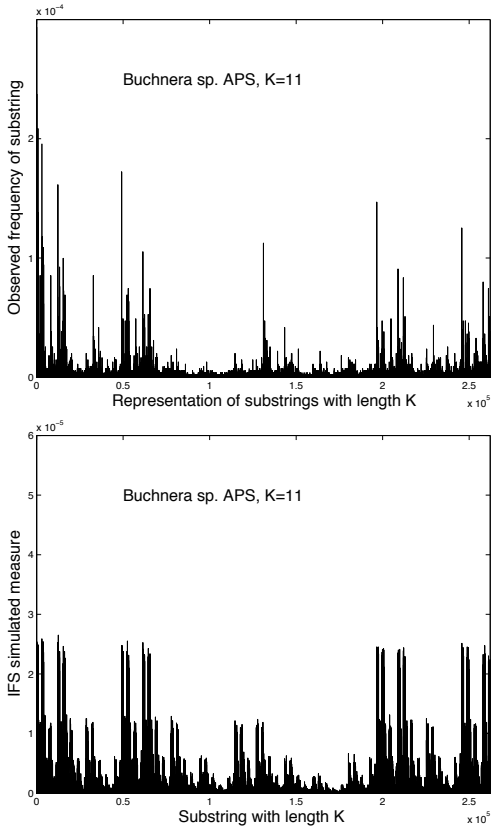

Figure 1: A segment of measure representation of the linked coding sequence for *Buchnera sp. APS* in the top figure and the IFS simulation for the same set of K-strings in the bottom figure.

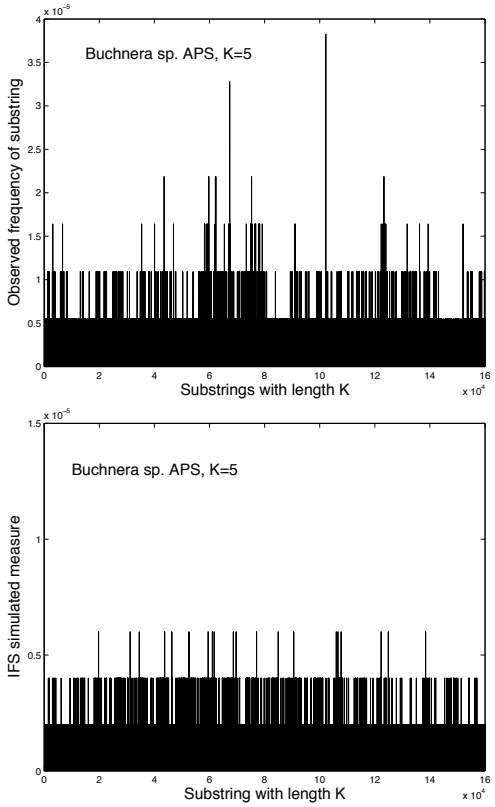

Figure 2: A segment of measure representation of the linked protein sequence for *Buchnera sp. APS* in the top figure and the IFS simulation for the same set of K-strings in the bottom figure.

It is easy to see that $\int_0^1 d\mu_K^p(x) = 1$ and $\mu_K^p([x_l^p(s), x_r^p(s)[) = F_K^p(s)$. We call μ_K^p the *measure representation* of the linked protein sequence of the organism corresponding to the given K. As an example, a fragment of the histogram of substrings in the linked protein sequence of *Buchnera sp. APS* for $K = 5$ is given in the top figure of Figure 2.

For simplicity of notation, the index K is dropped in $F_K^c(s)$, etc., from now on, where its meaning is clear. We can order all the $F^c(s)$ or $F^p(s)$ according to the increasing order of $x_l^c(s)$ or $x_l^p(s)$. We then obtain a sequence of real numbers consisting of 4^K or 20^K elements which we denote as $F^c(t), t = 1, \cdots, 4^K$ or $F^p(t), t = 1, \cdots, 20^K$.

3 Iterated function systems

In order to simulate the measure representations of the linked coding and protein sequences, we propose the *iterated function system* (IFS) model. IFS is the name given by Barnsley and Demko (1985) originally to a system of contractive maps $w = \{w_1, w_2, \cdots, w_N\}$. Let E_0 be a compact set in a compact metric space, $E_{\sigma_1\sigma_2\cdots\sigma_n} = w_{\sigma_1} \circ w_{\sigma_2} \circ \cdots \circ w_{\sigma_n}(E_0)$ and

$$E_n = \cup_{\sigma_1,\cdots,\sigma_n \in \{1,2,\cdots,N\}} E_{\sigma_1\sigma_2\cdots\sigma_n}.$$

Then $E = \cap_{n=1}^{\infty} E_n$ is called the *attractor* of the IFS. The attractor is usually a fractal and the IFS is a relatively general model to generate many well-known fractal sets such as the Cantor set and the Koch curve. Given a set of probabilities $p_i > 0$, $\sum_{i=1}^N p_i = 1$, pick an $x_0 \in E$ and define the iteration sequence

$$x_{n+1} = w_{\sigma_n}(x_n), \qquad n = 0, 1, 2, 3, \cdots, \qquad (10)$$

where the indices σ_n are chosen randomly and independently from the set $\{1, 2, \cdots, N\}$ with probabilities $P(\sigma_n = i) = p_i$. Then every orbit $\{x_n\}$ is dense in the attractor E (Barnsley and Demko 1985;Vrscay 1991). For n large enough, we can view the orbit $\{x_0, x_1, \cdots, x_n\}$ as an approximation of E. This process is called *chaos game*.

Let μ be the invariant measure on the attractor E of an IFS, χ_B the characteristic function for the Borel subset $B \subset E$, then from the ergodic theorem for IFS (Barnsley and Demko 1985),

$$\mu(B) = \lim_{n\to\infty} [\frac{1}{n+1} \sum_{k=0}^{n} \chi_B(x_k)].$$

In other words, $\mu(B)$ is the relative visitation frequency of B during the chaos game. A histogram approximation of the invariant measure may then be obtained by counting the number of visits made to each pixel on the computer screen.

3.1 Moment method to estimate the parameters of the IFS model

The coefficients in the contractive maps and the probabilities in the IFS model are the parameters to be estimated for a real measure which we want to simulate. Vrscay (1991) detailed a moment method to perform this task. If μ is the invariant measure and E the attractor of IFS in $\mathbf{R}$, the moments of μ are

$$g_i = \int_E x^i d\mu, \qquad g_0 = \int_E d\mu = 1. \qquad (11)$$

If $w_i(x) = c_i x + d_i,\ i = 1, \cdots, N$, then the following well-known recursion relations hold for the IFS model:

$$[1 - \sum_{i=1}^{N} p_i c_i^n] g_n = \sum_{j=1}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) g_{n-j} (\sum_{i=1}^{N} p_i c_i^{n-j} d_i^j). \tag{12}$$

Thus, setting $g_0 = 1$, the moments $g_n,\ n \geq 1$, may be computed recursively from a knowledge of $g_0, \cdots, g_{n-1}$ (Vrscay 1991).

If we denote by G_k the moments obtained directly from the real measure using (11), and g_k the formal expression of moments obtained from (12) for IFS model, then through solving the optimal problem

$$\min_{c_i, d_i, p_i} \sum_{k=1}^{n} (g_k - G_k)^2, \qquad \text{for some chosen } n, \tag{13}$$

we can obtain the estimated values of the parameters in the IFS model.

From the measure representation of a linked coding or protein sequence, we see that it is natural to choose $N = 4$ and

$$w_i(x) = x/4 + (i-1)/4, \quad i = 1, 2, 3, 4$$

or $N = 20$ and

$$w_i(x) = x/20 + (i-1)/20, \quad i = 1, 2, \cdots, 20$$

respectively in the IFS. For a given measure representation of a linked coding or protein sequence, The probabilities $p_i, i = 1, 2, 3, 4$ or $p_i, i = 1, 2, \cdots, 20$ in the IFS model are the parameters to be estimated. Once we have obtained the estimated parameters in the IFS model, we can generate a measure μ^{cf} (for linked coding sequence) or μ^{pf} (for linked protein sequence) through the formula obtained from the ergodic theorem. As an example, a fragment of IFS simulated measure for the linked coding sequence of *Buchnera sp. APS* for $K = 11$ is given in the bottom figure of Figure 1 and a fragment of IFS simulated measure for the linked protein sequence of *Buchnera sp. APS* for $K = 5$ is given in the bottom figure of Figure 2.

4 Correlation distance

Due to the way they are constructed, the measures μ^c and μ^p of the linked coding and protein sequences can be treated as random multiplicative cascades. Such multiplicative cascades commonly have built-in multifractal structures. We propose to use the IFS measure to characterize a multifractal structure. From the point of view of (Qi et al. 2003), we need to subtract the random background from the sequence $\{F^p(s)\}$ in order to get a more satisfactory evolutionary tree. Qi *et al.* (2003) used a Markov model to do this. In this paper we propose to remove the IFS simulated multifractal strucure from the measure representation by defining

$$F^{cr}(s) = \begin{cases} \frac{F^c(s)}{\mu^{cf}([x_l^c(s), x_r^c(s)[)}, & \text{if } \mu^{cf}([x_l^c(s), x_r^c(s)[) \neq 0, \\ 1, & \text{otherwise,} \end{cases} \tag{14}$$

$$F^{pr}(s) = \begin{cases} \frac{F^p(s)}{\mu^{pf}([x_l^p(s), x_r^p(s)[)}, & \text{if } \mu^{pf}([x_l^p(s), x_r^p(s)[) \neq 0, \\ 1, & \text{otherwise.} \end{cases} \tag{15}$$

For all 4^K or 20^K different K-strings, we can also order the $F^{cr}(s)$ or $F^{pr}(s)$ sequence according to the dictionary order of s.

For two random variables $\mathbf{X}$ and $\mathbf{Y}$ with samples $X(1), X(2), \cdots, X(N_1)$ and $Y(1), Y(2), \cdots, Y(N_1)$ respectively, let

$$< \mathbf{X} >= \frac{1}{N_1} \sum_{i=1}^{N_1} X(i), \quad < \mathbf{Y} >= \frac{1}{N_1} \sum_{i=1}^{N_1} Y(i),$$

$$\delta(\mathbf{X}) = \sqrt{\frac{1}{N_1} \sum_{i=1}^{N_1} (X(i) - < \mathbf{X} >)^2},$$

$$\delta(\mathbf{Y}) = \sqrt{\frac{1}{N_1} \sum_{i=1}^{N_1} (Y(i) - < \mathbf{Y} >)^2}.$$

Then the sample covariance of $\mathbf{X}$ and $\mathbf{Y}$ is

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \frac{1}{N_1} \sum_{i=1}^{N_1} (X(i) - < \mathbf{X} >)(Y(i) - < \mathbf{Y} >). \tag{16}$$

The sample correlation coefficient between $\mathbf{X}$ and $\mathbf{Y}$ is therefore given by

$$\rho(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{Cov}(\mathbf{X}, \mathbf{Y})}{\delta(\mathbf{X}) \delta(\mathbf{Y})}. \tag{17}$$

We have $-1 \leq \rho(\mathbf{X}, \mathbf{Y}) \leq 1$. If it is equal to zero, the random variables $\mathbf{X}$ and $\mathbf{Y}$ are considered uncorrelated.

Then for two organisms $\mathbf{X}$ and $\mathbf{Y}$ and a given K, we use $\rho^c(\mathbf{X}, \mathbf{Y})$ to denote the correlation coefficient between their $F^{cr}(s)$ sequences and $\rho^p(\mathbf{X}, \mathbf{Y})$ to denote the correlation coefficient between their $F^{pr}(s)$ sequences. We next define the *correlation distances* by

$$\mathrm{Dist}^c(\mathbf{X}, \mathbf{Y}) = \frac{1 - \rho^c(\mathbf{X}, \mathbf{Y})}{2} \tag{18}$$

and

$$\mathrm{Dist}^p(\mathbf{X}, \mathbf{Y}) = \frac{1 - \rho^p(\mathbf{X}, \mathbf{Y})}{2}. \tag{19}$$

5 Data and results

Currently there are more than 50 complete genomes of Archaea and Eubacteria available in public databases, for example Genbank at 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); twelve **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)

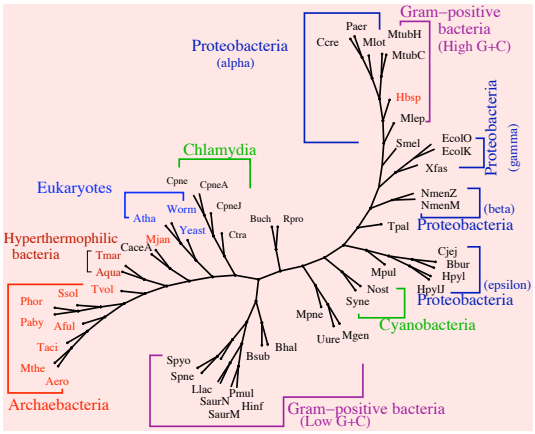

Figure 3: The neighbor-joining phylogenetic tree based on the correlation distance using the linked coding sequences and $\{F^{cr}(s)\}$ with $K = 11$.

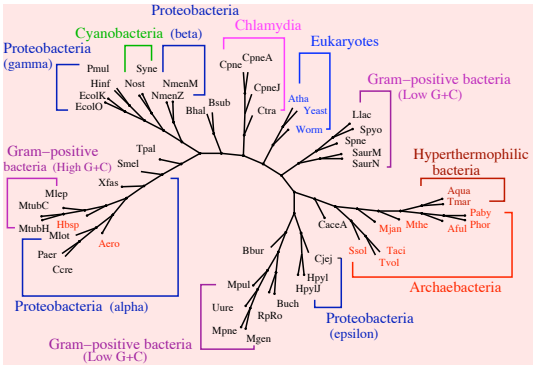

Figure 4: The neighbor-joining phylogenetic tree based on the correlation distance using the linked protein sequences and $\{F^{pr}(s)\}$ with $K = 5$.

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 sixteen **Proteobacteria**. The sixteen 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). Besides these prokaryotic genomes, the genomes of three eukaryotes: the yeast *Saccharomyces cerevisiae* (yeast), the nematode *Caenorhabdites elegans* (chromosome I-V, X) (Worm), and the flowering plant *Arabidopsis thaliana* (Atha), were also included in our analysis.

We downloaded the coding and protein sequences from the complete genomes of the above organisms. The numerical results showed that it is appropriate to use the measures of linked protein sequences with $K = 5$ (see Qi et al. (2003) and the measures of linked coding sequences with $K = 11$. The case $K = 6$ (for linked protein sequences) and $K = 12$ (for linked coding sequences) are worth trying but beyond our computing power for the time being. When we use the linked coding sequences with $K = 11$, we calculate the the correlation distances based on $F^{cr}(s)$ sequences of all the above organisms. When we use the linked protein sequences with $K = 5$, we calculate the the correlation distances based on $F^{pr}(s)$ sequences of all the above organisms.

We use the distance matrices from the correlation analysis to construct the phylogenetic tree with the help of neighbor-joining program in the PHYLIP package of J. Felsenstein (his web site in the references). We show the phylogenetic tree using $F^{cr}(s)$ sequences with $K = 11$ in Figure 3 and the phylogenetic tree using $F^{pr}(s)$ sequences with $K = 5$ in Figure 4 .

6 Discussion and conclusions

Although the existence of the archaebacterial urkingdom 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).

The correlation distance based on $\{F^{cr}(s)\}$ or $\{F^{pr}(s)\}$ after removing the multifractal structure from the original information gives a satisfactory phylogenetic tree. Figure 3 shows that all Archaebacteria except *Halobacterium* sp. NRC-1 (Hbsp) stays in a separate branch with the Eubacteria and Eukaryotes. Figure 4 shows that all Archaebacteria except *Halobacterium* sp. NRC-1 (Hbsp) and *Aeropyrum pernix* (Aero) stay in a separate branch with the Eubacteria and Eukaryotes. In the two trees obtained, the three Eukaryotes also group in one branch and almost all other bacteria in different traditional categories stay in the right branch. At a general global level of complete genomes, our result supports the genetic annealing model for the universal ancestor (Woese 1998). The two hyperthermophilic bacteria: *Aquifex aeolicus* (Aqua) VF5 and *Thermotoga maritima* MSB8 (Tmar) gather together and stay in the Archaebacteria branch in the two trees. We notice that these two bacteria, like most Archaebacteria, are hyperthermophilic. It has previously been shown that Aquifex has close relationship with Archaebacteria from the gene comparison of an enzyme needed for the synthesis of the amino acid trytophan (Pennisi 1998).

It has been pointed out (Qi et al. 2003) that the subtraction of random background is an essential step. Our results show that removing the multifractal structure is also an essential step in our correlation method. In Yu et al. (2003a), we proposed to use the recurrent IFS model (Vrscay 1991) to simulate the measure representation of complete genome and define the phylogenetic distance based on the parameters from the recurrent IFS model. The method of Yu et al. (2003a) does not include the step of removing multifractal structure, so we obtained a tree in which archaebacteria, eubacteria and eukaryotes intermingle with one another. Although the result from the correlation method of Qi et al. (2003) is slightly better than the result from our correlation method (*Halobacterium* sp. NRC-1 (Hbsp) and *Aeropyrum pernix* (Aero) stay with other Archaebacteria in their phylogetic tree), our algorithm seems simpler, faster and more efficient in using computer space. The reason is that Qi *et al.* (2003) used a Markov model to

subtract the random background. Hence their algorithm needs to retain all information of K-, $(K-1)$- and $(K-2)$-strings. When K is large, considerable computer space is needed to store this information. On the other hand, our method only requires the information of K-strings and the 4 or 20 parameters estimated from the IFS model. Similar to the method in Qi et al (2003), lateral gene transfer (Jeffrey et al. 1998) might not affect our results since the correlation method does not depend on the selection of a specific gene.

Acknowledgment

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 discussions on the phylogenetic problem. This work was supported by of the Youth Foundation of National Natural Science Foundation of China No. 10101022, QUT Postdoctoral Research Support Grant No. 9900658, the Australian Research Council grant A10024117.

References

[1] Anh V. V., Lau K. S. and Yu Z. G., (2001) Multifractal characterisation of complete genomes, *J. Phys. A: Math. Gene.* **34**, 7127-7139.

[2] Anh V. V., Lau K. S. and Yu Z. G., (2002) Recognition of an organism from fragments of its complete genome, *Phys. Rev. E* **66**, 031910.

[3] Barnsley M.F. and Demko S.,(1985) Iterated function systems and the global construction of Fractals, *Proc. Roy. Soc. London A* **399**, 243-275.

[4] Berthelsen C. L., Glazier J. A. and Skolnick M. H., (1992) Global fractal dimension of human DNA sequences treated as pseudorandom walks, *Phys. Rev. A* **45(12)**, 8902-8913.

[5] Brown T. A., (1998), *Genetics* (3rd Edition), CHAPMAN & Hall, London.

[6] Chatzidimitriou-Dreismann C.A. and Larhammar D., (1993) Long-range correlation in DNA, *Nature (London)* **361**, 212-213.

[7] Chu K.H., Qi J., Yu Z.G. and Anh V., (2003) Origin and phylogeny of chloroplasts: A simple correlation analysis of complete genomes, *Mol. Biol. Evol.*,(Accepted for publication).

[8] Felsenstein J., The Phylip software, http: //evolution.genetics.washington.edu/phylip.html

[9] Fraser C. M. *et al.*, (1995) The minimal gene complement of Mycoplasma genitalium, *Science*, **270**, 397-404.

[10] Goldman N., (1993) Nucleotide,dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences, *Nucleic Acids Research* **21(10)**, 2487-2491.

[11] Grassberger P. and Procaccia I.,(1983) Characterization of strange attractors, *Phys. Rev. lett.* **50**, 346-349.

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

[13] Hao B.L.,Xie H.M., Yu Z.G. and Chen G.Y., (2001) Factorizable language: from dynamics to bacterial complete genomes, *Physica A* **288**, 10-20.

[14] 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.

[15] Jeffrey H. J., (1990) Chaos game representation of gene structure, *Nucleic Acids Research* **18(8)**, 2163-2170.

[16] Jeffrey G., Lawrence J.G., Ochman H., (1998) Molecular archaeology of the *E. coli* genome, *Proc. Natl. Acad. Sci. USA* **95**, 9413-9417.

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

[18] Li W., Marr T., and Kaneko K.,(1994) Understanding long-range correlations in DNA sequences, *Physica D* **75**, 392-416.

[19] Peng C.K., Buldyrev S., Goldberg A.L., Havlin S., Sciortino F., Simons M., and Stanley H.E.,(1992) Long-range correlations in nucleotide sequences, *Nature* **356**,168-170.

[20] Prabhu V.V. and Claverie J. M., (1992) Correlations in intronless DNA, *Nature* **359**, 182-182.

[21] Qi J., Wang B. and Hao B.L., (2003) Prokaryote phylogeny based on complete genomes–tree construction without sequence alignment, *J. Mol. Evol.* (Accepted for publication).

[22] Pennisi E., (1998) Genome Data shake tree of life, *Science* **286**, 672-674.

[23] Stuart G.W., Moffett K. and Baker S., (2002) Integrated gene species phylogenies from unaligned whole genome protein sequences, *Bioinformatics* **18(1)**, 100-108.

[24] Vieira Maria de Sousa ,(1999) Statistics of DNA sequences: A low frequency analysis, *Phys. Rev. E* **60(5)**, 5932-5937.

[25] Vrscay E. R., (1991) Iterated function systems: theory, applications and the inverse problem, in *Fractal Geometry and analysis*, Eds, J. Belair, (NATO ASI series, Kluwer Academic Publishers).

[26] Woese C.R., (1998) The universal ansestor, *Proc. Natl. Acad. Sci. USA* **95**, 6854-6859.

[27] 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 and Fractals* **11(14)**, 2215-2222.

[28] Yu Z. G., Anh V. V. and Lau K. S., (2001) Measure representation and multifractal analysis of complete genome, *Phys. Rev. E* **64**, 031903.

[29] Yu Z. G., Anh V. V. and Lau K. S., (2003), Multifractal and correlation analysis of protein sequences from complete genome,*Phys. Rev. E* **68**, 021913.

[30] Yu Z.G., Anh V., Lau K.S. and Chu K.H., (2003) The genomic tree of living organisms based on a fractal model. *Phys. Lett. A*,**317**, 293-302.

[31] Yu Z.G. and Jiang P., (2001) Distance, Correlation and Mutual information among portraits of organisms based on complete genomes, *Phys. Lett. A* **286**, 34-46.