The genomic tree of living organisms based on a fractal model *

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

[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.

[3]Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong, China.

[4]Department of Biology, Chinese University of Hong Kong, Shatin, Hong Kong, China.

Abstract

Accumulation of complete genome sequences of living organisms creates new possibilities to discuss the phylogenetic relationships at the genomic level. In the present study, a fractal model is proposed to simulate a kind of visual representation of complete genome. The estimated parameters in the fractal model is used to define the genetic distance between two organisms. Because we take into account all genome content including both coding and non-coding regions, the phylogenetic tree from such an analysis leads to alternate classification of genomes that is called a genomic tree. This method of phylogenetic analysis does not require sequence alignment of homologous genes and relies instead on our fractal analysis, so it can avoid artefacts associated with sequence alignment. The similarity in related organisms based on the fractal model of the complete genome is global. Our result from such an analysis of more than 50 genomes indicates that lateral gene transfer must have been very common in the early history of life and thus constitutes a major source of variations in a substantial proportion of prokaryotic genome.

Key words: Measure representation, IFS (RIFS) model, complete genome, genomic tree.

1 Introduction

Although foreshadowed by earlier suggestions, so far the realisation by Chatton [1] was the most important advance made in our understanding of the living world as a whole. The classification of Chatton is that there are two major groups of organisms, the prokaryotes (bacteria) and the eukaryotes (organisms with nucleated cells). This classification was confirmed and made more widely known by Stanier and van Niel [2], and it is now universally accepted by biologists.

The classification of prokaryotes was chaotic before the works of Woese and colleagues [3-5]. Based on comparison of the small subunit ribosomal RNA (rRNA), Woese and colleagues proposed that there are two fundamentally different groups of bacteria, namely eubacteria and archaebacteria.

*The research was partially supported by QUT's Postdoctoral Research Support Grant No. 9900658 and the Youth Foundation of Chinese National Natural Science Foundation (No. 10101022), the RGC Earmarked Grant CUHK 4215/99P.

†Corresponding author. Email: yuzg@hotmail.com,z.yu@qut.edu.au

With eukaryotes, they constitute four kingdoms of life, namely Protoctista, Plantae, Fungi and Animalia. Phylogenetic trees based on the small and large rRNAs [6], and duplicated genes [7] support the monophyly of the archaebacteria originally proposed by Woese and colleagues. Although the existence of the archaebacterial domain is accepted by many biologists, its phylogenetic status is still a matter of controversy: Lake and colleagues [8, 9] argued that archaebacteria are paraphyletic; sulfobacteria (eocytes) are more closely related to eukaryotes than to other archaebacteria, whereas halobacteria are more closely related to eubacteria than to other archaebacteria. However this argument has been criticized on different grounds [10, 11].

Many genes (particularly those encoding metabolic enzymes) give different phylogenies of the same organisms or even fail to support the three-domain classification of living organisms [10,12-15]. A recurrent question concerns the controversial proximity of archaea to either eukarya or eubacteria [16]. Archaebacteria appear to be close to eukarya when the protein synthesis machinery (transcription and translation) is considered, but close to bacteria if metabolic genes are compared [17]. And the evidence presented by Mayr [11] shows clearly that the archaebacteria are so much more similar to the eubacteria than to eukaryotes that their removal from the prokaryotes is not justified.

Evolutionary inference based on DNA sequences traditionally compares homologous segments of a single gene in different organisms. These comparisons are generally reliable indicators of phylogenetic relationships, but are limited in being based on point mutations only [18]. The availability of complete genome sequences allows the reconstruction of organismal phylogeny, taking into account the genome content. Many previous attempts to analyse the macrostructure of genomes for phylogenetic reconstruction have been based on a number of well-known techniques such as DNA hybridization studies and restriction enzyme fragment analyses [19]. Sankoff *et al.* [18] proposed a measure of gene order rearrangement based on the minimal set of chromosomal inversions, transpositions, insertions, and deletions necessary to convert the order in one genome to that of the other and discussed the phylogenetic inference of eukaryotes based on this measure. Only the genes appear in all genomes of organisms considered can be used in Sankoff 's method. A more integrative view of genome evolution is feasible with the shared gene trees proposed recently by Snel *et al.* [20]. The genome-based phylogenetic analyses using protein-encoding genes [21] and the gene content and overall similarity [22] have been reported and the concept *genomic tree* have also been proposed.

In most of genome-based phylogenetic analyses till now, the authors only used the coding region of the complete genome. The similarity based on the coding region may not representative of the global similarity. The non-coding region of a genome also provides important information for phylogenetic reconstruction and should be taken into account when we compare complete genomes. Many measures of similarity used in the literature mostly relied on the sequence alignment. There doesn't seem to be any good standard for defining the score in sequence alignment. The another way to discuss the evolutionary problem is looking at the frequencies of strings [23-25] or the lengths of the functional regions in DNA sequences [26, 27].

In the present study, we attempt to elucidate the classification of genomes of living organisms based on a simple fractal model of complete genomes. The fractal method has been successfully used to study many problems in Physics, Mathematics, Engineering, and Biology (e.g. [26-30]) in the past two decades or so. In our model, we use mainly the primary structure (namely the nucleotide sequence) of the complete genome, including both the coding and non-coding regions. The genetic distance defined between two organisms is based only on the parameters derived from the fractal model, so that we can avoid artefacts associated with sequence alignment. The similarity based on the fractal model of the complete genome is global.

2 Measure representation of complete genomes

It is hard to get a global outline of the complete genome from its nucleotide sequence. It is helpful to develop some visual methods to the complete genome. From the visual representation, one can have the auto-visual outline of the complete genome. For this purpose, Peng *et al.* [28] proposed their well-known DNA walk model. Chaos game representation of DNA sequence was also proposed [31]. Hao *et al.* [32] proposed the 2-dimensional portrait representation for complete genomes. Using Hao's portrait representation, Yu and Jiang [24] proposed a simple way to construct the phylogenetic tree of bacteria (just 14 complete genomes used). Inspired by the idea of Hao's portrait representation [32], Yu *et al.* [30] proposed one-dimensional measure representation for the complete genome. In all the above visual representations of complete genomes, self-similarity was found. In other words, there exist fractal patterns in the visual representations of complete genomes. In the following we introduce the measure representation proposed by Yu *et al.* [30].

We call any string made of K letters from the set $\{g, c, a, t\}$ a K-string. For a given K there are in total 4^K different K-strings. In order to count the number of each kind of K-strings in a given DNA sequence 4^K counters are needed. We divide the interval $[0, 1[$ into 4^K disjoint subintervals, and use each subinterval to represent a counter. Letting $s = s_1 \cdots s_K, s_i \in \{a, c, g, t\}, i = 1, \cdots, K$, be a substring with length K, we define

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

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

$$x_{right}(s) = x_{left}(s) + \frac{1}{4^K}. \tag{3}$$

We then use the subinterval $[x_{left}(s), x_{right}(s)[$ 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 total number of K-strings appeared in the complete genome is denoted as $N_K(total)$, we define

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

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

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

It is easy to see $\int_0^1 d\mu_K(x) = 1$ and $\mu_K([x_{left}(s), x_{right}(s)[) = F_K(s)$. We call μ_K the *measure representation* of the organism corresponding to the given K. As an example, the histogram of substrings in the complete genome of *Buchnera sp. APS* for $K = 8$ is given in the left figure of Figurre 1. Self-similarity is apparent in 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: For organisms with more than one chromosome, we count out the occurrence of all K-strings in all chromosomes. The ordering of a, c, g, t in (2) will give the natural dictionary ordering of K-strings in the one-dimensional space. When we want to compare different organisms using the measure representation, once the ordering of a, c, g, t in (2) is chosen, it is fixed for all organisms considered. And later on in this paper we will show that a different ordering of a, c, g, t will induce a similar genomic tree.

3 (Recurent) iterated function systems model

In order to simulate the measure representation of the complete genome, Anh *et al.* [33] proposed the *iterated function systems* (IFS) model and the recurrent IFS model. IFS is the name given by Barnsley and Demko [34] 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

$$x_{n+1} = w_{\sigma_n}(x_n), \qquad n = 0, 1, 2, 3, \cdots, \tag{6}$$

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 [34, 35]. 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.*

Given a system of contractive maps $w = \{w_1, w_2, \cdots, w_N\}$ on a compact metric space E^*, we associate with these maps a matrix of probabilities $\mathbf{P} = (p_{ij})$ which is row stochastic, i.e. $\sum_j p_{ij} = 1,\ i = 1, 2, \cdots, N$. Consider a random chaos game sequence generated by

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

where x_0 is any starting point. The fundamental difference between this process and the usual chaos game Eq. (6) is that the indices σ_n are not chosen independently, but rather with a probability that depends on the previous index σ_{n-1}:

$$P(\sigma_{n+1} = i) = p_{\sigma_n, i}$$

Then $(E^*, w, \mathbf{P})$ is called a *recurrent IFS* (RIFS).

Let μ be the invariant measure on the attractor E of an IFS or RIFS, χ_B the characteristic function for the Borel subset $B \subset E$, then from the ergodic theorem for IFS or RIFS [34],

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

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.

4 Moment method to estimate the parameters of the IFS (RIFS) model

The coefficients in the contractive maps and the probabilities in the IFS or RIFS model are the parameters to be estimated for a real measure which we want to simulate. Vrscay [35] introduced a moment method to perform this task. If μ is the invariant measure and E the attractor of IFS or RIFS in $\mathbf{R}$, the moments of μ are

$$g_i = \int_E x^i d\mu, \qquad g_0 = \int_E d\mu = 1. \tag{7}$$

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:

$$[1 - \sum_{i=1}^{N} p_i c_i^n] g_n = \sum_{j=1}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) g_{n-j} (\sum_{i=1}^{N} p_i c_i^{n-j} d_i^j). \tag{8}$$

Thus, setting $g_0 = 1$, the moments $g_n,\ n \geq 1$, may be computed recursively from a knowledge of $g_0, \cdots, g_{n-1}$ [35].

For the RIFS model, we have

$$g_n = \sum_{j=1}^{N} g_n^{(j)}, \tag{9}$$

where $g_n^{(j)},\ j = 1, \cdots, N$, are given by the solution of the following system of linear equations:

$$\sum_{j=1}^{N} (p_{ji} c_i^n - \delta_{ij}) g_n^{(j)} = - \sum_{k=0}^{n-1} \left(\begin{array}{c} n \\ k \end{array} \right) [\sum_{j=1}^{N} c_i^k d_i^{n-k} p_{ji} g_k^{(j)}],\ i = 1, \cdots, N,\ n \geq ?. \tag{10}$$

For $n = 0$, we set $g_0^{(i)} = m_i$, where m_i are given by the solution of the linear equations

$$\sum_{j=1}^{N} p_{ji} m_j = m_i, \quad i = 1, 2, \cdots, N, \quad \text{and } g_0 = \sum_{i=1}^{N} m_i = 1. \tag{11}$$

If we denote by G_k the moments obtained directly from the real measure using (7), and g_k the formal expression of moments obtained from (8) for IFS model and from (9-11) for RIFS model, then through solving the optimal problem

$$\min_{c_i, d_i, p_i \text{ or } p_{ij}} \sum_{k=1}^{n} (g_k - G_k)^2, \qquad \text{for some chosen } n, \tag{12}$$

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

5 Definition of distance betweeen two organisms

From the measure representation of a complete genome, we see that it is natural to choose $N = 4$ and

$$w_1(x) = x/4,\ w_2(x) = x/4 + 1/4,\ w_3(x) = x/4 + 1/2,\ w_4(x) = x/4 + 3/4$$

in the IFS or RIFS model. For a given measure representation of a complete genome, we obtain the estimated values of the probabilities p_1, p_2, p_3, p_4 in IFS model or the matrix of probabilities $\mathbf{P} = (p_{ij})$ by solving the optimisation problem (12). For example, when $K = 8$, the estimated values of the matrix of probabilities of *Buchnera sp. APS* is

$$\begin{pmatrix} 0.423483 & 0.207054 & 0.099711 & 0.269753 \\ 0.354290 & 0.187515 & 0.129088 & 0.329107 \\ 0.299749 & 0.167843 & 0.148956 & 0.383452 \\ 0.290126 & 0.100192 & 0.179554 & 0.430129 \end{pmatrix}.$$

Based on the estimated values of probabilities, we can use the chaos game to generate a histogram approximation of the invariant measure of IFS or RIFS which we can compare with the real measure representation of the complete genome. For example, the histogram approximation of the generated measure of *Buchnera sp. APS* using the RIFS model is shown in the right figure of Figurre 1. It is seen that the left and right figures in Figurre 1 are quite similar. In order to clarify how close the simulation measure is to the original measure representation, we convert the measure to its walk representation. If $t_j,\ j = 1, 2, \cdots, 4^K$ is the histogram of a measure and t_{ave} is its average, then we define $T_j = \sum_{k=1}^{j}(t_k - t_{ave}),\ j = 1, 2, \cdots, 4^K$. So we can plot the two walks of the real measure representation and the measure generated by chaos game of IFS or RIFS model in one figure. In Figure 2, we show the walk representations of the measures in Figure 1. From Figure 2, one can see that the difference between the two walk representations is very small. We simulated the measure representations of the complete genomes of many organisms using the IFS and RIFS models. We found that RIFS is a good model to simulate the measure representation of complete genome of organisms. From above, once the matrix of probabilities is determined, the RIFS model is obtained. Hence the matrix of probabilities obtained from the RIFS model can be used to represent the measure of the complete genome of an organism. Different organisms can be compared using their matrix of probabilities obtained from the RIFS model. If $\mathbf{P} = (p_{ij}),\ \mathbf{P'} = (p'_{ij}),\ i, j = 1, 2, 3, 4$, are the matrices of probabilities of two different organisms obtained from the RIFS model for a fixed K, we propose to define the distance between the two organisms as

$$Dist = \sqrt{\sum_{i,j=1}^{4} (p_{ij} - p'_{ij})^2}. \tag{13}$$

The genetic distance defined between two organisms is based only on the parameters derived from the fractal model, so that we can avoid artefacts associated with sequence alignment. The similarity based on the fractal model of the complete genome is global.

6 Data and genomic trees

Till now more than 50 complete genomes of Archaea and Eubacteria are available in public databases (for example in Genbank at web site ftp://ncbi.nlm.nih.gov/genbank/genomes/ or in KEGG at web site http://www.genome.ad.jp/kegg/java/org_list.html). And the traditional classification of these organism is also provided. There are eight **Archae Euryarchaeota**: *Archaeoglobus fulgidus* DSM4304, *Pyrococcus abyssi*, *Pyrococcus horikoshii* OT3, *Methanococcus jannaschii* DSM2661, *Halobacterium* sp. NRC-1, *Thermoplasma acidophilum*, *Thermoplasma volcanium* GSS1, and *Methanobacterium thermoautotrophicum* deltaH; two **Archae Crenarchaeota**: *Aeropyrum pernix* and *Sulfolobus solfataricus*; three **Gram-positive Eubacteria (high G+C)**: *Mycobacterium tuberculosis* H37Rv (lab strain), *Mycobacterium tuberculosis* CDC1551 and *Mycobacterium leprae* TN; twelve **Gram-positive Eubacteria (low G+C)**: *Mycoplasma pneumoniae* M129, *Mycoplasma genitalium* G37, *Mycoplasma pulmonis*, *Ureaplasma urealyticum* (serovar 3), *Bacillus subtilis* 168, *Bacillus halodurans* C-125, *Lactococcus lactis* IL 1403, *Streptococcus pyogenes* M1, *Streptococcus pneumoniae*, *Staphylococcus aureus* N315, *Staphylococcus aureus* Mu50, and *Clostridium acetobutylicum* ATCC824. The others are **Gram-negative Eubacteria**, which consist of two **hyperthermophilic bacteria**: *Aquifex aeolicus* VF5 and *Thermotoga maritima* MSB8; five **Chlamydia**: *Chlamydia trachomatis* (serovar D), *Chlamydia trachomatis MoPn*, *Chlamydia pneumoniae* CWL029, *Chlamydia pneumoniae* AR39 and *Chlamydia pneumoniae* J138; one **Cyanobacterium**: *Synechocystis* sp. PCC6803; two **Spirochaete**: *Borrelia burgdorferi* B31 and *Treponema pallidum* Nichols; and sixteen **Proteobacteria**. The sixteen Proteobacteria are divided into four subdivisions, which are **alpha subdivision**: *Mesorhizobium loti* MAFF303099, *Sinorhizobium meliloti*, *Caulobacter crescentus* and *Rickettsia prowazekii* Madrid; **beta subdivision**: *Neisseria meningitidis* MC58 and *Neisseria meningitidis* Z2491; **gamma subdivision**: *Escherichia coli* K-12 MG1655, *Escherichia coli* O157:H7 EDL933, *Haemophilus influenzae* Rd, *Xylella fastidiosa* 9a5c, *Pseudomonas aeruginosa* PA01, *Pasteurella multocida* PM70 and *Buchnera* sp. APS; and **epsilon subdivision**: *Helicobacter pylori* J99, *Helicobacter pylori* 26695 and *Campylobacter jejuni.* Besides these prokaryotic genomes, the genomes of three eukaryotes: the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabdites elegans* (chromosome I-V, X), and the flowering plant *Arabidopsis thaliana*, were also included in our analysis.

Based on the evolutionary distance matrix obtained by using Eq. 13, a genomic tree was inferred by the neighbouring-joining method [36] using *MEGA* (version 2.1) [37].

7 Discussion and conclusions

A question one may ask on our method is whether the value of K will affect our final result of the genomic tree. In our RIFS model, for the measure representation when K is small, there are only a few possible K-strings, so this would make no statistical sense. We constructed the genomic trees of the selected organisms from $K = 6$ to $K = 9$, and found the topologies of these trees are the same. Because of the limitation on processing power of our computer, we are not able to construct the genomic tree for K larger than 9. We present the genomic tree for $K = 8$ (i.e. only consider the substrings with length 8) in Figure 3.

Another question one may ask on our method is whether the ordering of $\{a, c, g, t\}$ in Eq. (2) will affect the final result of the genomic tree. Although the values of the matrix of probabilities of an organism in the RIFS model will change if the ordering of $\{a, c, g, t\}$ in Eq. (2) changes, we believe that the phylogenetic trees based on different orderings of $\{a, c, g, t\}$ in Eq. (2) will be similar. In Figure 4, we present the genomic tree based on the order g, a, t, c in Eq. (2), the topology of this tree is very similar to that in Figure 3 which is based on the order a, c, g, t in Eq. (2). Some discrepancies were noted however. Noticeably the two hyperthermophilic bacteria, *Aquifex aeolicus* and *Thermotoga maritima*, are closely related in Figure 3, but more distantly related in Figure 4.

In Ref. [22] , Tekaia *et al.* pointed out “There are pitfalls of traditional methods such as variable changes in sequence and reliability of sequence alignments.”. Because our fractal method just pick out the scaling property of the complete genome and does not require sequence alignment of homologous genes, the genomic tree present here does not suffer from such problems. And our methodology is not intended to substitute for evolutionary inference of the traditional method but, rather to provide a classification of genomes (include both coding and non-coding regions) using the global similarity on the genome level. The fractal method can characterize the sequence similarity and the phylogenetic relationship is based on some kinds of similarity in the genome, so we can expect our distance based on fractal method carry a phylogenetic signal.

In spite of the success of microbial taxonomy based on DNA sequences of genes coding for rRNA and other biomolecules, the evolutionary relationships between major groups of prokaryotes are still obscure because phylogenetic analyses of single gene sequences often fail to resolve these deep branches due to mutational saturation. Further, artefacts-related sequence misalignment and the different evolutionary rates between lineages of the gene in question could produce misleading topology in phylogenetic analyses [21]. Another complication is incidences of gene transfer between species, i.e. lateral, or horizontal, gene transfer, which means trees based on individual genes do not represent organismal phylogeny. It was our intention that the genomic tree (Figure 3) based on the complete genome which includes both coding and non-coding sequences can reflect the evolutionary histories of living organisms despite lateral gene transfer, gene duplication and gene loss such as the tree resulting from the method proposed by Fitz-Gibbon and House [21]. And in the defining distance, we only used the parameters which reflect the fractal scaling property of the complete genomes, and no sequence alignment is necessary. So our method can avoid the mistakes induced by misalignment.

The aim of including complete genomic data from different strains of the same species (including *Mycobacterium tuberculosis*, *Staphylococcus aureus*, *Chlamydia trachomatis*, *Chlamydia pneumoniae*, *Escherichia coli*, *Neisseria meningitidis* and *Helocobacter pylori*) is to test whether our method is reasonable because each of these species should group together at the most detailed level from any point of view. In our genomic tree different strains of the same species do cluster together, or at least very closed related. This aspect agrees with that of the previous work by Qi *et al.* [25].

In our genomic tree, the two species of *Clamydia* are closely related, suggesting the similarity of the genome organization within this genus. However, in other cases, species of the same genus (including *Pyrococcus, Thermoplasma, Mycobacterium, Bacillus, Mycoplasma*, and *Streptococcus*) do not constitute a grouping. Moreover, members of well-defined taxonomic groups within archaea (Euryarachaeota and Crenarchaeota) or eubacteria (such as Gram-positive bacteria, Spirochaete and Proteobacteria) do not cluster together. The only exception is a number of species in euryarachaeta (*Archaeoglobus fulgidus, Methanobacterium thermoautotrophicum, Pyrococcus abyssi* and *Thermoplasma acidophilum* in Figure 3) are closely related; yet other members of this group are scattily distributed in the tree. The Gram-positive bacteria with high GC content are found in a cluster distinct from their counterparts with low GC content but the evolutionary significance of this result is unknown, as in each case, species in other groups intermingle with one another. As a whole, the archaebacteria mix together with the eubacteria in our genomic tree. With regard to the eukaryotes, although *Arabidopsis thaliana* and *Caenorhabdites elegans* are closely related, the yeast *Saccharomyces cerevisiae* is more closely related to some of the eubacteria than to the other two eukaryotes, possibly reflecting the presence of prokaryotic genes in this eukaryote. So at the most general global level of complete genome, our result tends to support the genetic annealing model for the universal ancestor [38] and the scenario of a reticulate tree in the early history of life as presented by Doolittle [39]. This scenario represents frequent incidences of lateral gene transfer during prokaryote evolution, which would account for the observation that many genes give different phylogenies from the same organisms. Lateral gene transfer between eubacteria and archaea would explain the fact that some eubacteria possess genes of archaeal origin, and vice versa. For instance, the bacterial pathogen *Borrelia burgdorferi* bears an archaeal-type lysyl-tRNA synthetase [40]. On the other hand, the archaeon *Archaeoglobus filgidis* has many genes (such as the gene coding for 3-hydroxy-3-methylglutaryl Coenzyme A reductase) apparently of bacterial origin [17]. This enzyme subsequently identified in other members of Archaeoglobales as well as in Thermoplasmatales possibly results from an initial lateral gene transfer event from eubacteria to archaea, followed by another event between archaea [41]. The extent of lateral gene transfer can be massive: it was estimated that 18open reading frames of E. coli resulted from 234 lateral transfer incidences since its divergence from the Salmonella lineage [42]. Analysis of complete genomes based on a smaller number of organisms than those analysed in the present study has also suggested that extensive, and continuous, gene transfer occurred early in the evolution of prokaryotes [43, 44]. The genomic tree from our analysis based on more than 50 genomes supports this point of view. Our result indicates that lateral gene transfer must have been very common in the early history of life and thus constitutes a major source of variations in a substantial proportion of the prokaryotic genome. Therefore our analysis on the complete genome would not give an organismal phylogeny consistent with gene trees based on single molecules. In this regard, Doolittle [39] believes that a universal organismal tree cannot be resolved through molecular phylogenetics. As Woese [38] argues, there is no conventional organismal phylogenetic tree in the early history of life.

In most of genome-based phylogenetic analysis such as Refs.[20-22,25], the authors only used the coding regions or the translated protein sequences from the complete genome. Because of the limitation and the increasing of genome data, the previous works (such as Ref. [25]) usually compare their trees with the traditional classification obtained by Woese *et al.* [3-5]. That is very reasonable they can obtain the tree whose overall topology strongly resemble the SSU rRNA-based evolutionary trees [3-5] because all biological phenomena are expressed by proteins (so as the coding regions). Qi *et al.* [25] found that the tree resembles the traditional evolutionary tree if you use the translated protein sequences from the complete genome and subtract the random background from the original compositional vector. Using all the available completely sequenced genomes through whole proteome comparsons, the proximity between the Archaea and the Eubacteria was observed in Ref.[22]. Furthermore, in this paper, we consider contents of both coding and non-coding regions, Archaea , Eubacteria and Eukaryotes are mixing together. Hence the contents of non-coding regions will affect the topology of the genomic tree very much. Hence we can see the coding regions play more important role under the traditional sense of phylogenetic relationship.

Acknowledgments

One of the author Z.G. Yu would like to express his thanks to Prof. Bai-Lin Hao of Institute of Theoretical Physics of Chinese Academy of Science for continuous encouragement and some useful comments on this paper. The authors also thank Chi-Pang Li of Chinese University of Hong Kong for helps in generating the genomic trees.

References

[1] E. Chatton, *Titres et travaux scientifiques* (Sette,Sottano, Italy, 1937).

[2] R.Y. Stanier, C.B. van Niel, *Arch. Microbiol* **42**, 17 (1962).

[3] C.R. Woese, G.E. Fox, *Proc. Natl. Acad. Sci. USA* **74**, 5088 (1977).

[4] C.R. Woese, *Microbiol. Rev.* **51**, 221 (1987).

[5] C.R. Woese, O. Kandler, M.L. Wheelis, *Proc. Natl. Acad. Sci. USA* **87**, 4576 (1990).

[6] M. Gouy, W.-H. Li, *Nature (London)* **339**, 145 (1989).

[7] N. Iwabe, K.-I. Kuma, M. Hasegawa, S. Osawa, T. Miyata, *Proc. Natl. Acad. Sci. USA* **86**, 9355(1989).

[8] J.A. Lake, *Nature (London)* **331**, 184 (1988).

[9] J.A. Lake, E. Henderson, M. Oakes, M.W. Clark, *Proc. Natl. Acad. Sci. USA* **81**, 3796 (1984).

[10] R.S. Gupta, *Microbiol. Mol. Biol. Rev.* **62**, 1435 (1998).

[11] E. Mayr, *Proc. Natl. Acad. Sci. USA* **95**, 9720 (1998).

[12] T. Cavalier-Smith, *Nature* **339**, 100 (1989).

[13] P. Forterre, *et al.*, *Biosystems* **28**, 15 (1992).

[14] J.R. Brown, W.F. Doolittle, *Microbiol. Mol. Biol. Rev.* **61**, 456 (1997).

[15] R.F. Doolittle, *Nature* **392**, 339 (1998).

[16] H. Brinkmann, H. Philippe, *Mol. Biol. Evol.* **16**, 429 (1999).

[17] W.F. Doolittle, J.M. Logsdon, *Curr. Biol.* **8**, R209 (1998).

[18] D. Sankoff, et al, *Proc. Natl. Acad. Sci. USA* **89**, 6575 (1992).

[19] W.-H. Li, *Molecular evolution* (Sinauer, Sunderland, MA, 1997).

[20] B. Snel, P. Bork, M.A. Huynen, *Nat. Genet.* **21**, 108 (1999).

[21] S. Fitz-Gibbon, C.H. House, *Nucleic Acids Res.* **27**, 4218 (1999).

[22] F. Tekaia, A. Lazcano, B. Dujon, *Genome Research* **9**, 550 (1999).

[23] B. Wang, *J. Mol. Evol.* **53**, 244 (2001).

[24] Z.G. Yu,. P. Jiang, *Phys. Lett. A* **286**, 34 (2001).

[25] J. Qi, B. Wang, B.L. Hao, Prokaryote pgylogeny based on complete genomes–tree construction without sequence alignment, *J. Mol. Evol.* (revised version).

[26] Z.G. Yu and V. Anh, *Chaos, Solitons and Fractals* **12**, 1827(2001).

[27] Z.G. Yu, V. Anh and B. Wang, *Phys. Rev. E* **63**, 011903(2001).

[28] C.K. Peng, *et al.* , *Nature* **356**, 168 (1992).

[29] Z.G. Yu, B.L. Hao, H.M. Xie and G.Y. Chen, *Chaos, Solitons and Fractals* **11**, 2215 (2000).

[30] Z.G. Yu, V. Anh and K. S. Lau, *Phys. Rev. E* **64**, 031903 (2001).

[31] H.J. Jeffrey, *Nucleic Acids Res.* **18**, 2163 (1990).

[32] B.L. Hao, H.C. Lee, S.Y. Zhang, *Chaos, Solitons and Fractals* **11**, 825 (2000).

[33] V.V. Anh, K.S. Lau and Z.G. Yu, *Phys. Rev. E* **66**, 031910 (2002).

[34] M.F. Barnsley and S. Demko, *Proc. Roy. Soc. London A* **399**, 243 (1985).

[35] E. R. Vrscay, in *Fractal Geometry and analysis*, Eds, J. Belair, (NATO ASI series, Kluwer Academic Publishers, 1991).

[36] N. Saitou, M. Nei, *Mol. Biol. Evol.* **10**, 471 (1987).

[37] S. Kumar, K. Tamura, I.B. Jakobsen, M. Nei, *Bioinformatics* **17**, 1244 (2001).

[38] C.R. Woese, *Proc. Natl. Acad. Sci. USA* **95**, 6854 (1998).

[39] W.F. Doolittle, *Science* **284**, 2124 (1999).

[40] M. Ibba, J.L. Bono, P.A. Rosa, D. Sll, *Proc. Natl. Acad. Sci. USA* **94**, 14383 (1997).

[41] Y. Boucher, H. Huber, S. L'Haridon, K.O. Stetter, W.F. Doolittle, *Mol. Biol. Evol.* **18**, 1378 (2001).

[42] J.G. Lawrence, H. Ochman, *Proc. Natl. Acad. Sci. USA* **95**, 9413 (1998).

[43] M.C. Rivera, R. Jain, J.E. Moore, L.A. Lake, *Proc. Natl. Acad. Sci. USA* **95**, 6239 (1998).

[44] R. Jain, M.C. Rivera, J.A. Lake, *Proc. Natl. Acad. Sci. USA* **96**, 3801 (1999).

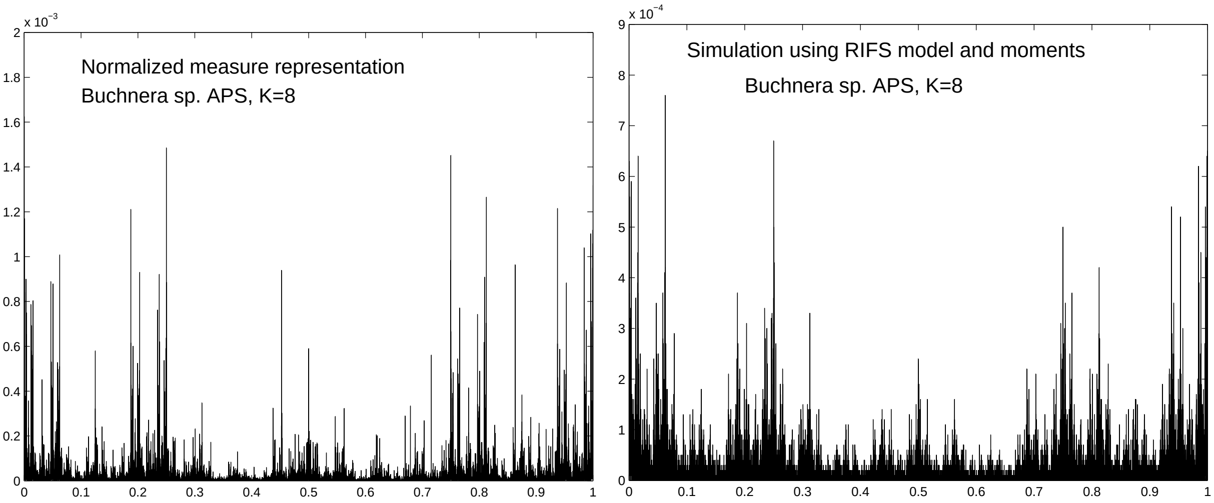

Figure 1: The measure representation (left) and the RIFS simulation (right) of the complete genome of *Buchnera sp.* APS when K=8.

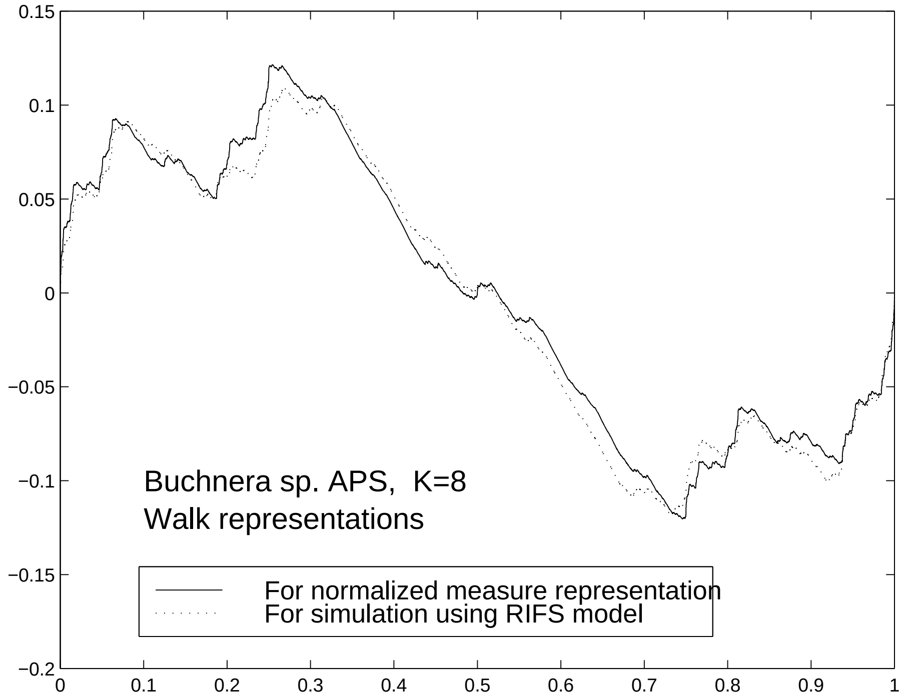

Figure 2: The walk representations of measures in Figure 1.

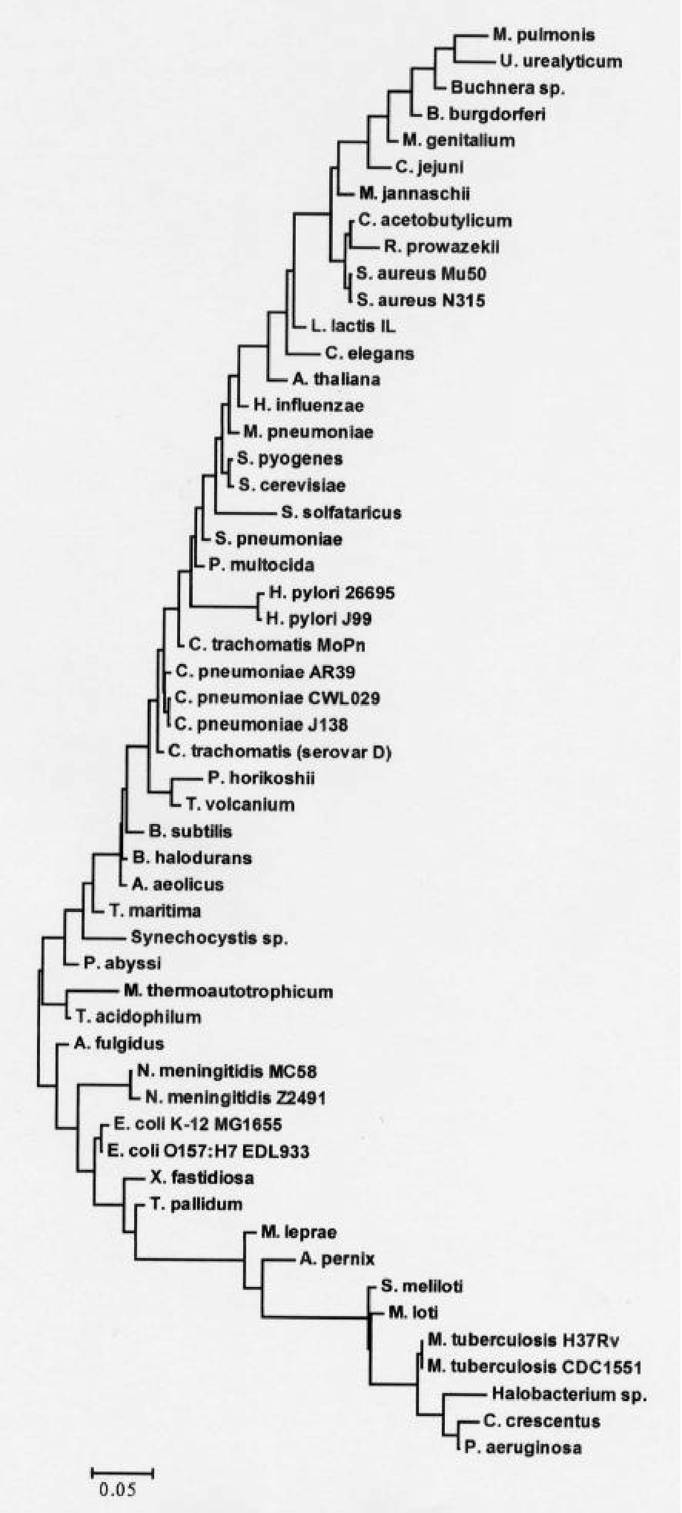

Figure 3: The genomic tree of living organisms using the neighbour-joining method and the distance based on the parameters in the RIFS model, with the order $\{a, c, g, t\}$ in Eq. (2).

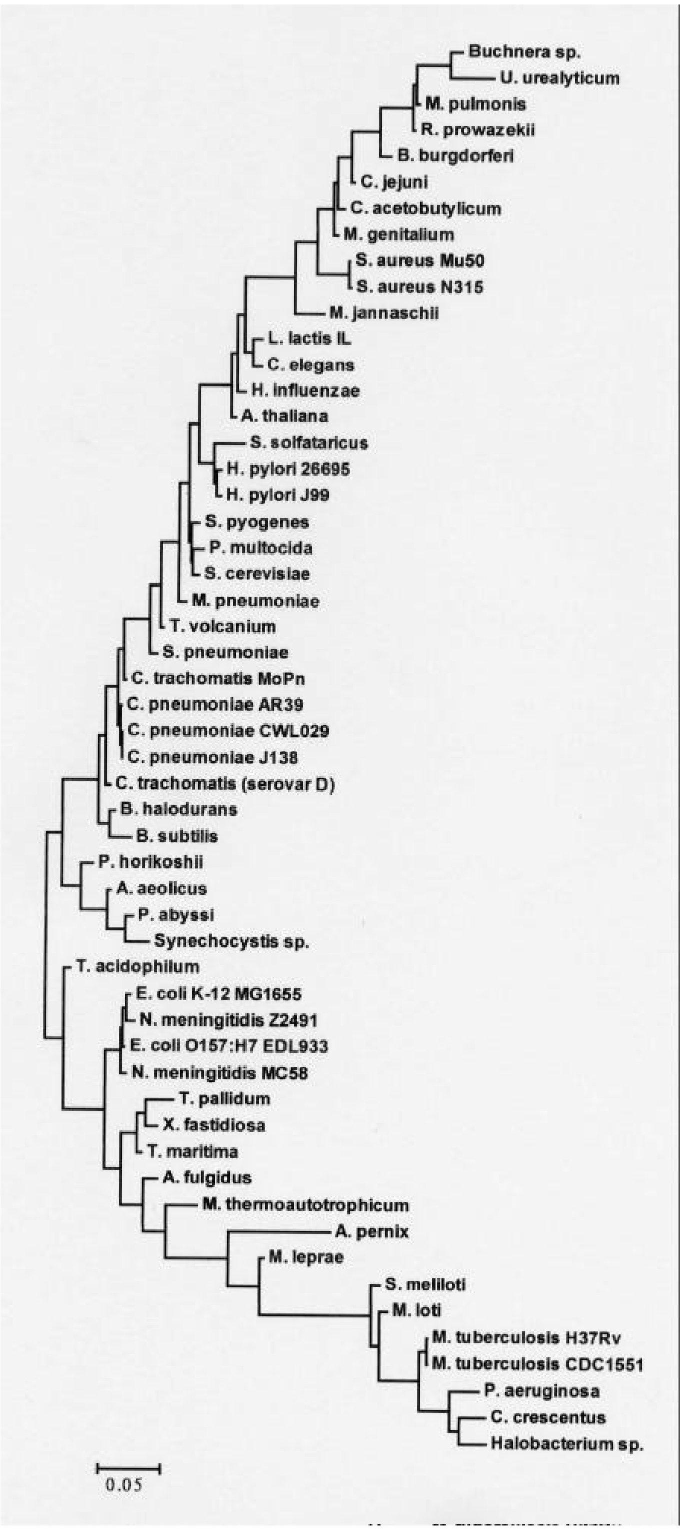

Figure 4: The genomic tree of living organisms using the neighbour-joining method and the distance based on the parameters in the RIFS model, with the order $\{g, a, t, c\}$ in Eq. (2).