Phylogenetic tree based on complete genomes using fractal and correlation analyses without sequence alignment

Z.G. Yu
Program in Statistics and Operations Research, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia.
School of Mathematics and Computing Science, Xiangtan University, Hunan 411105, China.

and

V.V. Anh
Program in Statistics and Operations Research, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia.

ABSTRACT

The complete genomes of living organisms have provided much information on their phylogenetic relationships. Similarly, the complete genomes of chloroplasts have helped resolve the evolution of this organelle in photosynthetic eukaryotes. In this review, we describe two algorithms to construct phylogenetic trees based on the theories of fractals and dynamic language using complete genomes. These algorithms were developed by our research group in the past few years. Our distance-based phylogenetic tree of 109 prokaryotes and eukaryotes agrees with the biologists’ “tree of life” based on the 16S-like rRNA genes in a majority of basic branchings and most lower taxa. Our phylogenetic analysis also shows that the chloroplast genomes are separated into two major clades corresponding to chlorophytes s.l. and rhodophytes s.l. The interrelationships among the chloroplasts are largely in agreement with the current understanding on chloroplast evolution.

Keywords: phylogeny; genome; fractal analysis; correlation analysis.

1. INTRODUCTION

Since the sequencing of the first complete genome of the free-living bacterium Mycoplasma genitalium in 1995 [1], 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.

In our understanding of the classification of the living world as a whole, the most important advance was made by Chatton [2], whose classification is that there are two major groups of organisms, the prokaryotes (bacteria) and the eukaryotes (organisms with nucleated cells). Then the universal tree of life based on the 16S-like rRNA genes given by Woese and colleagues [3, 4] led to the proposal of three primary domains (Eukarya, Bacteria, and Archaea). Although the archaeabacterial domain is accepted by biologists, its phylogenetic status is still a matter of controversy [5, 6]. Analyses of some 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 [5, 7, 8].

It is generally accepted that genome sequences are excellent tools for studying evolution [9]. In building the tree of life, analysis of whole genomes has begun to supplement, and in some cases to improve upon, studies previously done with one or few genes [9]. The availability of complete genomes allows the reconstruction of organismal phylogeny, taking into account the genome content, for example, based on the rearrangement of gene order [10], the presence or absence of protein-coding gene families [11], gene content and overall similarity [12], and occurrence of folds and orthologs [13]. All these approaches depend on alignment of homologous sequences, and it is apparent that much information (such as gene rearrangement and insertions/deletions) in these data sets is lost after sequence alignment, in addition to the intrinsic problems of alignment algorithms [14–16]. 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 the complete genomes (e.g., [14–23]).

By overcoming the problem of noise and bias in the protein sequences through the use of appropriate models, whole-genome trees have now largely converged to the rRNA-sequence tree [24]. Qi et al. [17] have developed a simple correlation analysis of complete genome sequences based on compositional vectors without the need of sequence alignment. The compositional vectors calculated from the frequency of amino acid strings are converted to distance values for all taxa, and the phylogenetic relationships are inferred from the distance matrix using conventional tree-building methods. An analysis based on this method using 109 organisms (prokaryotes and eukaryotes) yields a tree separating the three domains of life, Archaea, Eubacteria and Eukarya, with the relationships among the taxa correlating with those based on traditional analyses [17]. A correlation analysis based on a different transformation of compositional vectors was also reported by Stuart et al. [15] who demonstrated the applicability of the method in revealing phylogeny using vertebrate mitochondrial genomes.
Chloroplast DNA is a primary source of molecular variations for phylogenetic analysis of photosynthetic eukaryotes. During the past decade the availability of complete chloroplast genome sequences has provided a wealth of information to elucidate the phylogeny of photosynthetic eukaryotes at deeper levels of evolution. There have been many phylogenetic analyses based on comparison of sequences of multiple protein-coding genes in chloroplast genomes (e.g., [25-31]). The approach proposed by Qi et al. [17] has also been adopted to analyze the complete chloroplast genomes [32] and found to reveal a phylogeny of this organelle that is largely consistent with the phylogeny of the photosynthetic eukaryotes based on traditional analyses, thus demonstrating the value of this methodology in analyzing genomes of a smaller size.

In the approach proposed by Qi et al. [17], a key step is to subtract the noise background in the composition vectors of the protein sequences from complete genomes through a Markov model. In the past few years, we proposed two alternative methods to model the noise background in the composition vector. One method [21] is based on the iterated function system (IFS) model [19, 20, 33] in fractal geometry; the other method is based on the relationship between a word and its two sub-words in the theory of symbolic dynamics [23]. Here we review and apply these two methods to construct phylogenetic trees of 109 prokaryotes and eukaryotes. The results are as good as those previously reported in Qi et al. [17] and Chu et al. [32].

2. METHODS

The phylogenetic signal in the protein sequences is often obscured by noise and bias [24]. There is always some randomness in the composition of protein sequences, revealed by their statistical properties at single amino acid or oligopeptide level (see Weiss et al. [34] for a discussion on this point). In order to highlight the selective diversification of sequence composition, we subtract the random background (noise and bias) from protein sequences.

Method 1: Measure Representation of Protein Sequences and IFS Simulation

Yu et al. [19] proposed the measure representation of protein sequences. 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), Threone (T), Tryptophan (W), Tyrosine (Y) and Valine (V) [35, p109]. Each coding sequence in the complete genome of an organism is translated into a protein sequence using the genetic code [35, p122].

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 protein sequence for each organism. Here we only consider these kinds of linked protein sequences and view them as symbolic sequences.

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

Letting s=s_1s_2...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=1,2,...,K, be a substring with length K, we define

\[ x_j(s) = \sum_{i=1}^{K} \frac{x_i}{20^K}, \]

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

\[ x_j(s) = x_j(s) + \frac{1}{20^K}. \]

We then use the subinterval \( \{x_j(s), x_j(s)\} \) 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 (N_K(s) may be zero). Denoting the total number of K-strings appearing in the linked protein sequence as \( N_K(total) \), we define

\[ F_K(s) = N_K(s) / (N_K(total)) \]

to be the frequency of substring s. It follows that

\[ \sum_{s} F_K(s) = 1. \]

Now we can define a measure \( \mu_K \) on [0,1] by

\[ d\mu_K(x) = F_K(x)dx, \]

where

\[ Y_K(x) = 20^k F_K(s) \text{ when } x \in [x_j(s), x_j(s)], \]

We call \( \mu_K \) the measure representation of the organism corresponding to the given K.

We can order all the \( F(s) \) according to the increasing order of \( x_j(s) \). According to the IFS model described in Yu et al. [19], we can get the IFS simulation of all \( F(s) \). We denote this IFS simulation as \( F^{\text{IFS}}(s) \). In this method, we view \( F^{\text{IFS}}(s) \) of the 20^K kinds of K-strings as the noise background.

Method 2: Dynamical Language Model

Let \( N = 20^K \). We use a window of length K and slide it through each protein sequence in a genome by shifting one position at a time to determine the frequencies of each of the \( N \) kinds of strings. A protein sequence is excluded if its length is shorter than K. The observed frequency \( p(s_1s_2...s_K) \) of a K-string \( s_1s_2...s_K \) is defined as

\[ p(s_1s_2...s_K) = n(s_1s_2...s_K) / (L-K+1), \]
where \( n(s_1s_2...s_K) \) is the number of times that
\( s_1s_2...s_K \) appears in this sequence. Denoting by \( m \)
the number of protein sequences from each complete genome, the
observed frequency of a \( K \)-string \( s_1s_2...s_K \) is defined as
\[
\frac{(\sum_{j=1}^{m} n_j(s_1s_2...s_K))}{(\sum_{j=1}^{m} (L_j - K + 1))}.
\]
Here \( n_j(s_1s_2...s_K) \) means the number of times that
\( s_1s_2...s_K \) appears in the \( j \)th protein sequence and \( L_j \)
the length of the \( j \)th protein sequence in this complete genome.

In this method, we consider an idea from the theory of dynamical language that a \( K \)-string \( s_1s_2...s_K \) is possibly
constructed by adding a letter \( s_K \) to the end of the \((K-1)\)-string
\( s_1s_2...s_{K-1} \) or a letter \( s_1 \) to the beginning of the
\((K-1)\)-string \( s_2...s_{K-1}...s_K \). Suppose that we have performed
direct counting for all strings of length \((K-1)\) and the 20
kinds of letters, the expected frequency of appearance of \( K \)-
strings is predicted by
\[
q(s_1s_2...s_K) = \frac{p(s_1s_2...s_{K-1})p(s_K) + p(s_1)p(s_2...s_K)}{2},
\]
where \( q \) denotes the predicted frequency, and \( p(s_1) \) and
\( p(s_K) \) are frequencies of amino acids \( s_1 \) and \( s_K \) appearing
in this genome. (In [17, 31], the authors use Markov model to
characterize the predictor, in which the information of the
\((K-1)\)-strings and \((K-2)\)-strings.) is needed. In this
method we view \( q(s_1s_2...s_K) \) of the 20\(^K\) kinds of \( K \)-
strings as the noise background.

**Subtraction of the noise background and the correlation distance**

We then subtract the noise background before performing a
cross-correlation analysis (similar to removing a time-varying
mean in time series before computing the cross-correlation of
two time series).

We calculate a new measure \( X \) of the shaping role of selective
evolution as
\[
X(s_1s_2...s_K) = \begin{cases} 
P(s_1s_2...s_K) & \text{if } F(\cdot s_K) = 1, \text{ if } F(\cdot s_K) \neq 0 \\ 
0 & \text{if } F(\cdot s_K) = 0 \end{cases}
\]
in Method 1 [21] and

\[
X(s_1s_2...s_K) = \begin{cases} 
p(s_1s_2...s_K) - 1, & \text{if } q(s_1s_2...s_K) \neq 0 \\ 
0, & \text{if } q(s_1s_2...s_K) = 0 \end{cases}
\]
in Method 2 [23]. The transformation
\[
X(s) = F(s) / F^{\mu} (s) - 1
\]
or
\[
X(s) = p(s) / q(s) - 1
\]
has the desired effect of subtraction of random background
(noise and bias) from \( F \) or \( p \) and renders it a stationary time
series suitable for subsequent cross-correlation analysis.

For all possible \( K \)-strings \( s_1s_2...s_K \), we use
\( X(s_1s_2...s_K) \) as components to form a composition vector
for a genome. To further simplify the notation, we use \( X_i \) for
the \( i \)-th component corresponding to the string vector \( i, i = 1,..., N \) (the \( N \) strings are arranged in a fixed alphabetical order). Hence we construct a composition vector
\( X = (X_1, X_2,..., X_N) \) for genome \( X \), and likewise
\( Y = (Y_1, Y_2,..., Y_N) \) for genome \( Y \).

If we view the \( N \) components in vectors \( X \) and \( Y \) as samples of
two zero-mean random variables respectively, the sample
correlation \( C(X,Y) \) between any two genomes \( X \) and \( Y \)
is defined in the usual way in probability theory as
\[
C(X,Y) = \frac{\sum_{i=1}^{N} X_i \times Y_i}{(\sum_{i=1}^{N} X_i^2 \times \sum_{i=1}^{N} Y_i^2) \frac{1}{2}}.
\]
The distance \( D(X,Y) \) between the two genomes is then
defined by the equation
\[
D(X,Y) = (1 - C(X,Y)) / 2.
\]
A distance matrix for all the genomes under study is then
generated for construction of phylogenetic trees.

**Genome Data Sets and Tree Construction**

We retrieve the complete genomes from NCBI database

To test Method 1, in [21] we selected 51 bacteria genomes and
3 euukaryotes genomes. These include eight Archaea 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 deltalH (Mith); two Archaea Crenarchaeota: Aeropyrum pernix (Aper) and Sulfolobus solfataricus (Ssol); three Gram-positive Eubacteria (high G+C): Mycobacterium tuberculosis H37Rv (MtbB1), Mycobacterium tuberculosis CDC1551 (MtbC) and Mycobacterium leprae TN (Mlep); twelve Gram-positive Eubacteria (low G+C): Mycoplasma pneumoniae M129 (Mpmn), Mycoplasma genitalium G37 (Mgen), Mycoplasma pneumoniae (Mpunl), Ureaplasma urealyticum (serovar 3) (Uure), Bacillus subtilis 168 (Bsub), Bacillus halodurans C-125 (Bhal), Lactococcus lactis IL 1403 (Llac), Streptococcus pyogenes M1 (Spyo), Streptococcus pneumoiae (Spne), Staphylococcus aureus 315 (SaurN), Staphylococcus aureus Mu50 (SaurM), and Clostridium acetobutyllicum ATCC8284 (CaceA). The others are Gram-negative Eubacteria, which consist of two hyperthermophile bacteria: Aquifex aeolicus (Aqua) VF5 and Thermotoga maritima MS88 (Tmar); four Chlamydia: Chlamydia trachomatis (serovar D) (Ctra), Chlamydia pneumoniae (CpneA) and Chlamydia pneumoniae J138 (CpneJ); two Cyanobacterias: Synechocystis sp. PCC6803 (Syne) and Nostoc sp. PCC6803 (Nost); two Spirochaetes: Borrelia burgdorferi B31 (Bbur) and Treponema pallidum Nichols (Tpal); and sixteen Firmicutes: Staphylococcus aureus sp. IL 1403 (Llac), and Clostridium acetobutylicum ATCC8284 (CaceA). To test Method 2, in [23] we used two data sets: **Data set 1** (used in [17]), We selected 109 organisms for prokaryote phylogenetic analysis. These include four Archaea Crenarchaeota: Aeropyrum pernix (Aper), Sulfolobus solfataricus (Ssol), Sulfolobus tokodaii (Salto) and Pyrobasilum aerophilum (Pyrae); twelve Archaea Euryarchaeota: Archaeoglobus fulgidus (Archf), Halobacterium sp. NRC-1 (Halsp), Methanocarcina acetivorans str. C 2A (Meta), Methanococcus jannaschii (Metja), Methanopyrus kandleri AV19 (Meta), Methanosarcina mazei Goel (Metma), Methanobacterium thermoautotrophicum (Metth), Chlorobacterium abyssus (Pyrab), Pyrococcus furiosus (Pyru), Pyrococcus horikoshii (Pyro), Thermoplasma acidophilum (Theac) and Thermoplasma volcanium (Thevo); two Hyperthermophilic bacteria: Aquifex aeolicus (Aquea) and Thermotoga maritima (Thema); one Deinococcus-Thermus: Deinococcus radiodurans R1 (Deir); three Cyanobacteria: Cyanobacterium Nostoc sp. PCC7120 (Anasp), Cyanobacterium Synecystis PCC6803 (Syene) and Thermosynechococcus elongatus BP-1 (Theel); one Green sulphur bacteria: Chlorobium tepidum TLS (Chlte); nine Proteobacteria alpha subdivision: Agrobacterium tumefaciens C58 (Agtr5), Agrobacterium tumefaciens C58 U267 (Agtr5U267), Bacillus megaterium (Brum), Bacillus subtilis 1330 (Bsub), Caulobacter crescentus (Cauc), Mesorhizobium loti (Rhilo), Sinorhizobium meliloti 1021 (Rhime), Rickettsia conorii (Riccon) and Rickettsia prowazekii (Ripro); three Proteobacteria beta subdivision: Neisseria meningitidis MC8 (NeimeN) and Neisseria meningitidis Z2491 (NeimeZ) and Ralstonia solanacearum (Rals); twenty two Proteobacteria gamma subdivision: Buchnera sp. AP5 (Bucai), Buchnera aphidicola Sg (Bucap), Escherichia coli CTI073 (EcotlC), Escherichia coli O157:H7 EDL933 (EcoliE), Escherichia coli K-12 (EcoliK), Escherichia coli O157:H7 (EcoliO), Haemophilus influenzae Rd (Haein), Pasteurella multocida PM70 (Pasmu), Pseudomonas aeruginosa PA01 (Paeae), Pseudomonas putida KT2440 (Pepsu), Salmonella typhi (Salti), Salmonella typhimurium LT2 (Salty), Shewanella oneidensis MR-1 (Sheon), Shigella flexneri 2a str. 301 (Shifl), Vibrio cholerae (Vibch), Vibrio vulniﬁcus CMCP6 (Vibvu), Wigglesworthia sibiricus (Wigsib), Xanthomonas campestris ATCC 33913 (Xanca), Xylella fastidiosa (Xyfla), Yersinia pestis strain C902 (YerpC) and Yersinia pestis KIM (YerpE); three Proteobacteria epsilon subdivision: Campylobacter jejuni (Camje), Helicobacter pylori J99 (Heljy) and Helicobacter pylori 26695 (Helpy); twenty seven Firmicutes: Bacillus anthracis A2012 (Bacan), Bacillus halodurans (Bachd), Bacillus subtilis (Basu), Clostridium acetobutylicum ATCC824 (Cloab), Clostridium perfringens (Clope), Lactococcus lactis sp. IL 1403 (Laca), Listeria monocytogenes EGD-e (Listi), Listeria innocua (Lisun), Mycoplasma genitalium (Mycge), Mycoplasma penetrans (Mycpe), Oceanobacillus iheyensis (Ocei), Mycoplasma pneumoniae (Mycop), Mycoplasma pulmonis UAB CTIP (Mycp), Staphylococcus aureus N315 (Staun), Staphylococcus aureus Mu50 (StaunM), Staphylococcus epidermidis ATCC 12228 (Staup), Streptococcus galaccae NEM316 (StrangN), Streptococcus gallaciae 2603V/R (StragV), Streptococcus mutans UA159 (Strmu), Streptococcus pyogenes R6 (StrprR), Streptococcus pyogenes TIGR4 (StrprT), Streptococcus pyogenes MGA58232 (Strpy8), Streptococcus pyogenes MGA58315 (Strpyg), Strepococcus pyogenes SF370 (StrpyS), Thermococcus thermocatenarius tengsenga (Thet) and Ureaplasma urealyticum (Ueru); seven Actinobacteria: Bacillus subtilis (Bacsu) and Caulobacter crescentus NCC2705 (Biflo), Campylobacter jejuni (Cjje), Besila bedfordiae strains Mr1-2661 (Cito1-2661), Clostridium perfringens (Clope), Lactobacillus casei sp. IL 1403 (Laca), Listeria monocytogenes EGD-e (Listi), Listeria innocua (Lisun), Mycoplasma genitalium (Mycge), Mycoplasma penetrans (Mycpe), Oceanobacillus iheyensis (Ocei), Mycoplasma pneumoniae (Mycop), Mycoplasma pulmonis UAB CTIP (Mycp), Staphylococcus aureus N315 (Staun), Staphylococcus aureus Mu50 (StaunM), Staphylococcus epidermidis ATCC 12228 (Staup), Streptococcus galaccae NEM316 (StrangN), Streptococcus gallaciae 2603V/R (StragV), Streptococcus mutans UA159 (Strmu), Streptococcus pyogenes R6 (StrprR), Streptococcus pyogenes TIGR4 (StrprT), Streptococcus pyogenes MGA58232 (Strpy8), Streptococcus pyogenes MGA58315 (Strpyg), Strepococcus pyogenes SF370 (StrpyS), Thermococcus thermocatenarius tengsenga (Thet) and Ureaplasma urealyticum (Ueru); seven Actinobacteria: Bacillus subtilis (Bacsu) and Caulobacter crescentus NCC2705 (Biflo), Corynebacterium efficiens YS-314 (Coref), Corynebacterium glutamicum (Corg), Mycobacterium leprae TN (Mycel), Mycobacterium tuberculosis CDC1551 (MyctuC), Mycobacterium tuberculosis H37Rv (MyctuH) and Streptomyces coelicolor A3(2) (Strco); five Chlamydia: Chlamydia muridarum (Chlmu), Chlamydia pneumoniae AR39 (ChlnPa), Chlamydia pneumoniae CWL029 (ChlpC), Chlamydia pneumoniae J138 (Chlpj), and Chlamydia trachomatis (Chlre); three Spirochaetes: Borrelia burgdorferi (Borbu), Leptospira interrogans serovar lasi str. 5601 (Lepisa), and Treponema pallidium (Trepa); and one Fusobacteria: Fusobacterium nucleatum ATCC 25586 (Fusnu). We also included in the analysis six eukaryotes: Saccharomyces cerevisiae (yeast), Caenorhabditis elegans (worm), Arabidopsis thaliana (Ath), Encephalitozoon cuniculi (Enccu), Plasmodium falciparum (Plafa) and Schizosaccharomyces pombe (Schpo). **Data set 2** (used in [32]), We selected the following genomes of Chloroplast, Archaea, Eubacteria and Eukaryotes for
chloroplast phylogenetic analysis. These include twenty one chloroplast genomes (Cyanophora paradoxa, Cyanidium caldarium, Porphyra purpurea, Guillardia theta, Odontella sinensis, Euglena gracilis, Chlorella vulgaris, Nephroselmis olivacea, Mesoestigma viride, Chaetosphaeridium globosum, Marchantia polymorpha, Psilotum nudum, Pinus thunbergii, Oenothera elata, Lotus japonicus, Spinacia oleracea, Mesostigma viride, Chaetosphaeridium globosum, Marchantia polymorpha, Psilotum nudum, Pinus thunbergii, Oenothera elata, Lotus japonicus, Spinacia oleracea, Nicotiana tabacum, Arabidopsis thaliana, Oryza sativa, Triticum aestivum and Zea mays), two archaea genomes (Archaeoglobus fulgidus and Sulfolobus solfataricus), eight eubacteria genomes (Helicobacter pylori, Neisseria meningitides, Rickettsia prowazekii, Borrelia burgdorferi, Chlamydophila pneumoniae, Mycobacterium leprae, Nostoc sp. and Synechocystis sp.) and three eukaryotes genomes (Saccharomyces cerevisiae, Arabidopsis thaliana and Caenorhabditis elegans).

The words in the brackets are the abbreviations of the names of these organisms used in our phylogenetic trees (Figures 1 and 2).

Qi et al. [36] pointed out that the Fitch-Margoliash method [37] is not feasible when the number of species is as large as 100 or more and an algorithm such as maximum likelihood is not based on the distance matrix alone. So we construct all trees using the neighbour-joining (NJ) method [38] in the PHYLIP package [39].

3. Results and Discussion

Although the existence of the archaebacterial uringdom has been accepted by many biologists, the classification of bacteria is still a matter of controversy [40]. The evolutionary relationship of the three primary kingdoms, namely archaebacteria, eubacteria and eukaryote, is another crucial problem that remains unresolved [40].

It has been pointed out [17] 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 [20], we proposed to use the recurrent IFS model [41] 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. [20] does not include the step of removing multifractal structure, and yielded a tree in which archaebacteria, eubacteria and eukaryotes intermingle with one another.

In both methods presented here, K must be larger than 3. We can only calculate the distance matrices and construct the trees for K from 3 to 6 because of the limitation on the computing capability of our PCs and supercomputers. We find that the topology of the trees converges with K increasing from 3 to 6 and it becomes stable for K ≥ 5. We show the phylogenetic tree using X(s) sequences through Method 1 with K=5 in Fig. 1. For Method 2, we present the results based on K= 6 in Figures 2 and 3.

The correlation distance based on Method 1 after removing the multifractal structure (IFS simulation) from the original information gives a satisfactory phylogenetic tree. Fig. 1 shows that all Archaeabacteria except Halobacterium sp. NRC-1(Hbsp) and Aeropyrum pernix (Aero) stay in a separate branch.
with the Eubacteria and Eukaryotes. 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 [42]. The two hyperthermophilic bacteria: *Aquifex aeolicus* (Aqua) VF5 and *Thermotoga maritima* MSB8 (Tmar) gather together and stay in the Archaeabacteria branch in the tree. We notice that these two bacteria, like most Archaeabacteria, are hyperthermophilic. In the phylogenetic analyses based on a few genes, the tendency of the two hyperthermophilic bacteria, *Aquae* and *Thema*, to get into Archaea, has intensified the debate on whether there has been wide-spread lateral or horizontal gene transfers among species [43–45]. Eisen and Fraser [9] claimed that analyses of complete genomes suggest that lateral gene transfer has been rare over the course of evolution and it has not distorted the structure of the tree. Our results using Method 1 based on the complete genome ([Fig. 1](#)) do not seem to support the views of Eisen and Fraser [9]. Hence more works are required for this problem.

**Fig. 2** shows the $K=6$ tree based on the NJ analysis for the selected 109 organisms using Method 2. The selected Archaea group together as a domain (except *Pyrobaculum aerophilum*). The six eukaryotes also cluster together as a domain, and all Eubacteria fall into another domain. So the division of life into three main domains Eubacteria, Archaeabacteria and Eukarya is a clean and prominent feature. At the interspecific level, it is clear that Archaea is divided into two groups of Euryarchaeota and Crenarchaeota. Different prokaryotes in the same group (Firmicutes, Actinobacteria, Cyanobacteria, Chlamydia, Hyperthermophilic bacteria) all cluster together. Proteobacteria (except epsilon division) cluster together. In Proteobacteria, prokaryotes from alpha and epsilon divisions group with those from the same division. It is clear that the branch of Firmicutes is divided into sub-branches Bacillales, Lactobacillales, Clostridia and Mollicutes. Our phylogenetic tree of organisms supports the 16S-like rRNA tree of life in its broad division into three domains and the grouping of the various prokaryotes. So after subtracting the noise and bias from the protein sequences as described in our method, the whole-genome tree converges to the rRNA-sequence tree as asserted in Charlebois et al. [24].

In our tree ([Fig. 2](#)) the two hyperthermophilic bacteria group together and stay in the domain of eubacteria. This result is the same as in Qi et al. [17] and also supports the point of view in Eisen and Fraser [9]. We gave more comparison between Method 2 and the Markov model proposed by Qi et al. [17] in our recent work [23].

**Fig. 3** shows the $K=6$ tree based on NJ analysis for the
chloroplasts (data set 2) using Method 2. All the chloroplast genomes form a clade branched in Eubacteria domain and share a most recent common ancestor with cyanobacteria, which agrees with the widely accepted endosymbiotic theory that chloroplasts arose from cyanobacteria-like ancestor [46-48]. Apparently, despite massive gene transfer from the endosymbiont to the nucleus of the host cell [28, 29, 45], our analysis is able to identify cyanobacteria as the most closely related prokaryotes of chloroplasts. The chloroplasts are separated into two major clades, one of which corresponds to the green plants sensu lato, or chlorophytes s.l. [49], which include all taxa with a chlorophyte chloroplast, both primary and secondary endosymbioses in origin, and the other comprising the glaucophyte Cyanophora and members of rhodophytes s.l., which refers to rhodophytes (or red algae, Cyanidium and Porphyra in the tree) and their secondary symbiotic derivatives (the heterokont Odontella and the cryotphyte Guillardia). The close relationship between Cyanophora and rhodophytes s.l. (Cyanophora is mixed into rhodophytes s.l.) agrees with some of the previous analyses [26,50], although most recent studies suggest that the glaucophyte represents the earliest branch in chloroplast evolution with the green plants s.l. and rhodophytes s.l. as sister taxa [25, 28, 29, 51]. In chlorophyte s.l., the green algae (i.e., Chlorella, Mesostigma, and Nephroselmis) and Euglena are basal in position and the seed plants cluster together as a derived group, although the relationships among the other taxa (i.e., Marchantia, Psilotum, and Chaetosphaeridium) are somewhat different from our traditional understanding, probably due to limited taxon sampling in these primitive green plants.

Fig. 3 Phylogeny of chloroplast genomes based on Method 2 in the case K=6 [23].

To sum up, our simple correlation analysis on the complete chloroplast genomes has yielded a tree that is in good agreement with our current knowledge on the phylogenetic relationships of different groups of photosynthetic eukaryotes in general (see [48, 49, 52] for reviews). The only difference between the trees obtained by the present method and the one in Chu et al [32] is the placement of Pinus in the clade of Chlorophyte s.l. (for K=5 and 6).

Our approach circumvents the ambiguity in the selection of genes from complete genomes for phylogenetic reconstruction, and is also faster than the traditional approaches of phylogenetic analyses, particularly when dealing with a large number of genomes. Moreover, since multiple sequence alignment is not used, the intrinsic problems associated with this complex procedure can be avoided.
4. 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, and Dr. Li-Qian Zhou of Xiangtan University for useful discussions on the phylogenetic problem. Financial support was partially provided by the Chinese National Natural Science Foundation (grant no. 05B007) (Z.-G. Yu), and the Youth foundation of Educational Department of Hunan province in China (grant no. 05B007) (Z.-G. Yu), the Australian Research Council (grant no. DP0559807) (V. Anh).

5. REFERENCES


