Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on complete genome without sequence alignment

Z. G. Yu[1,2*], L. Q. Zhou[1], V. V. Anh[2], K. H. Chu[3], J. Qi[4], S. C. Long[1] and J. Q. Deng[1]

[1]School of Mathematics and Computing Science, Xiangtan University, Hunan 411105, China.

[2]Program in Statistics and Operations Research, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia.

[3]Department of Biology, Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China.

[4]Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100080, China.

Keywords: phylogeny; prokaryotes; chloroplast; genome; composition vector.

Abstract

The complete genomes of living organisms have provided much information on their phylogenetic relationships. Similarly, the complete genomes of the chloroplast have helped to resolve the evolution of this organelle in photosynthetic eukaryotes. In this paper we use these sequences to test a compositional approach without sequence alignment for phylogenetic analysis of complete genomes based on correlation analysis. All protein sequences from 54 complete prokaryote and eukaryote genomes were analyzed first. Then all protein sequences from 21 complete chloroplast genomes were analyzed. Our distance-based phylogenic tree of the 54 prokaryotes and eukaryotes agrees with the biologists' "tree of life" based on 16S rRNA comparison in a predominantly majority of basic branchings and most lower taxa. Our phylogenetic analysis also shows that the chloroplast genomes are separated to two major clades corresponding to chlorophytes (green plants) *s.l.* and rhodophytes (red algae) *s.l.* The interrelationships among the chloroplasts are largely in agreement with the current understanding on chloroplast evolution. Thus this study establishes the value of our simple compositional approach of phylogenetic analysis in elucidating the evolutionary relationships among genomes.

* Corresponding author, post address: School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia. Tel: 061-7-38645194; Fax: 061-7-38642310; e-mail: z.yu@qut.edu.au or yuzg@xtu.edu.cn

Introduction

In our understanding of the classification of the living world as a whole, the most important advance was made by Chatton (1937), 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 (Woese 1987; Woese et al. 1990) led to the proposal of three primary domains (Eukarya, Bacteria, and Archaea). Although the archaebacterial domain is accepted by biologists, its phylogenetic status is still a matter of controversy (Gupta 1998; Mayr 1998). Analysis of some genes, particularly those encoding metabolic enzymes, gives different phylogenies of the same organisms or even fail to support the three-domain classification of living organisms (Brown and Doolittle 1997; Doolittle 1998; Gupta 1998).

It is generally accepted that genome sequences are excellent tools for studying evolution (Eisen and Fraser 2003). 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 a few genes (Eisen and Fraser 2003). 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 (Sankoff et al. 1992), the presence and absence of protein-coding gene families (Fitz-Gibbon and House 1999), gene content and overall similarity (Tekaia et al. 1999), and occurrence of folds and orthologs (Lin and Gerstein 2000). All these above 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, let alone the intrinsic problems of alignment algorithms (Li et al. 2001; Stuart et al. 2002). 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., Li et al. 2001; Yu and Jiang 2001; Yu et al 2003a, 2003b, 2004; Edwards et al. 2002; Stuart et al. 2002).

By overcoming the problem of noise and bias in the protein sequences through the use of better models, whole-genome trees have now largely converged to the rRNA-sequence tree (Charlebois et al. 2003). Qi et al. (2004) have developed a simple correlation analysis of complete genome sequences based on compositional vectors without the need of sequence alignment. The compositional vectors calculated based on 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 (Qi et al. 2004). A correlation analysis based on a different transformation of compositional vectors was also reported by Stuart et al. (2002) 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 the deep levels of evolution. There have been many phylogenetic analyses based on comparison of sequences of multiple protein-coding genes in chloroplast genomes (e.g., Martin et al. 1998, 2002; Turmel et al. 1999, 2002; Adachi et al. 2000; Lemieux et al. 2000; De Las Rivas et al. 2002). The approach proposed by Qi et al. (2004) has also been adopted to analyze the complete chloroplast genomes (Chu et al., 2004) 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. (2004), a key step is to substract the noise background in the composition vector of the protein sequences from complete genomes through a Markov model. In this study, we intend to improve the approach proposed by Qi et al. (2004). We propose to model the noise background in the composition vector through the relationship between word and its two sub-words in theory of symbolic dynamics. This approach is easier to understand and faster than the approach by Qi et al. (2004). We apply our new approach to the phylogenetic analyses of 54 organisms (mainly prokaryotes), and the same chloroplast genomes used in our previous paper (Chu et al. 2004). The results are as good as those previously reported in Qi et al. (2004) and Chu et al. (2004).

Materials and Methods

Genome Data Sets

We retrieve the complete genomes from NCBI database (ftp://ncbi.nlm.nih.gov /genbank/genomes/).

We selected 54 organisms for prokaryote phylogenetic analysis. These include ten **Archaea**: *Archaeoglobus fulgidus* (Aful), *Pyrococcus abyssi* (Paby), *Pyrococcus horikoshii* (Phor), *Methanococcus jannaschii* (Mjan), *Halobacterium* sp. NRC-1 (Hbsp), *Thermoplasma acidophilum* (Taci), *Thermoplasma volcanium* (Tvol), *Methanobacterium thermoautotrophicum* (Mthe), *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* (Mlep); twelve **Gram-positive Eubacteria (low G+C)**: *Mycoplasma pneumoniae* (Mpne), *Mycoplasma genitalium* (Mgen), *Mycoplasma pulmonis* (Mpul), *Ureaplasma urealyticum* (Uure), *Bacillus subtilis* (Bsub), *Bacillus halodurans* (Bhal), *Lactococcus lactis* (Llac), *Streptococcus pyogenes* (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) and *Thermotoga maritima* (Tmar); four **Chlamydia**: *Chlamydia trachomatis* (Ctra), *Chlamydia pneumoniae CWL029* (Cpne), *Chlamydia pneumoniae AR39* (CpneA) and *Chlamydia pneumoniae J138* (CpneJ); **two Cyanobacteria**: *Synechocystis sp. PCC6803* (Syne) and *Nostoc* sp. *PCC6803* (Nost); two **Spirochaetes**: *Borrelia burgdorferi* (Bbur) and *Treponema pallidum* (Tpal); and sixteen **Proteobacteria**. The sixteen Proteobacteria are divided into four subdivisions, which are **alpha subdivision**: *Mesorhizobium loti* (Mlot), *Sinorhizobium meliloti* (smel), *Caulobacter crescentus* (Ccre) and *Rickettsia prowazekii* (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* (Hinf), *Xylella fastidiosa* (Xfas), *Pseudomonas aeruginosa* (Paer), *Pasteurella multocida* (Pmul) and *Buchnera sp. APS* (Buch); and **epsilon subdivision**: *Helicobacter pylori J99* (HpylJ), *Helicobacter pylori 26695* (Hpyl) and *Campylobacter jejuni* (Cjej). We also included in the analysis three **eukaryotes**: the yeast

Saccharomyces cerevisiae (yeast), the nematode *Caenorhabdites elegans* (chromosome I-V, X) (Worm), and the flowering plant *Arabidopsis thaliana* (Atha). The words in the brackets are the abbreviations of the name of these organisms used in our phylogenetic tree.

Complete sequences of 21 chloroplast genomes (*Cyanophora paradoxa, Cyanidium caldarium, Porphyra purpurea, Guillardia theta, Odontella sinensis, Euglena gracilis, Chlorella vulgaris, Nephroselmis olivacea, Mesostigma viride, Chaetosphaeridium globosum, Marchantia polymorpha, Psilotum nudum, Pinus thunbergii, Oenothera elata, Lotus japonicus, Spinacia oleracea, Nicotiana tabacum, Arabidopsis thaliana, Oryza sativa, Triticum aestivu* and *Zea mays*) are selected for our phylogenetic analysis on chloroplast, using the cyanobacterium *Synechocystis* as the outgroup taxon as in many previous analyses (e.g., Turemel et al., 1999; Martin et al., 2002).

Composition Vectors and Distance Matrix

We base our analysis on all protein sequences including hypothetical reading frames from each genome, regarding sequences of the 20 amino acids as symbolic sequences. In such a sequence of length L, there are a total of $N = 20^K$ possible types of strings of length K. We use a window of length K and slide it through the sequences by shifting one position at a time to determine the frequencies of each of the N kinds of strings in each genome. A protein sequence is excluded if its length is shorter than K. The observed frequency $p(\alpha_1\alpha_2...\alpha_K)$ of a K-string $\alpha_1\alpha_2...\alpha_K$ is defined as $p(\alpha_1\alpha_2...\alpha_K) = n(\alpha_1\alpha_2...\alpha_K)/(L-K+1)$, where $n(\alpha_1\alpha_2...\alpha_K)$ is the number of times that $\alpha_1\alpha_2...\alpha_K$ appears in this sequence. Denoting by m the number of protein sequences from each complete genome, the observed frequency of a K-string $\alpha_1\alpha_2...\alpha_K$ is defined as $(\sum_{j=1}^{m} n_j(\alpha_1\alpha_2...\alpha_K))/(\sum_{j=1}^{m}(L_j-K+1))$; here $n_j(\alpha_1\alpha_2...\alpha_K)$ means the number of times that $\alpha_1\alpha_2...\alpha_K$ appears in the jth protein sequence and L_j the length of the jth protein sequence in this complete genome.

The phylogenetic signal in the protein sequences is too often obscured by noise and bias (Charlebois et al. 2003). There is always some randomness in the composition of protein sequences, revealed by statistical properties of protein sequences at single amino acid or oligopeptide level (see Weiss et al. 2000 for a recent discussion on this point). In order to

highlight the selective diversification of sequence composition, we subtract the random background (noise and bias) from the simple counting results. In the present study, we consider an idea from the theory of dynamical language that a K-string $\alpha_1\alpha_2...\alpha_K$ is possibly constructed by adding a letter α_K to the end of $(K-1)$-string $\alpha_1\alpha_2...\alpha_{K-1}$ or a letter α_1 to the beginning of $(K-1)$-string $\alpha_2\alpha_3...\alpha_K$. 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

$$q(\alpha_1\alpha_2..\alpha_K)=\frac{p(\alpha_1\alpha_2..\alpha_{K-1})p(\alpha_K)+p(\alpha_1)p(\alpha_2\alpha_3..\alpha_K)}{2}$$

where *q* denotes the predicted frequency. [In the previous papers of our group (Qi et al. 2004; Chu et al. 2004), we use Markov model to characterize the predictor, in which we need to know the information of the $(K-1)$-strings and $(K-2)$-strings.] We then subtract the above random background (noise and bias) 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 then calculate a new measure *X* of the shaping role of selective evolution as

$$X(\alpha_1\alpha_2...\alpha_K)=\begin{cases} p(\alpha_1\alpha_2...\alpha_K)/q(\alpha_1\alpha_2...\alpha_K), & if \quad q(\alpha_1\alpha_2...\alpha_K)\neq 0 \\ 1, \quad if \quad q(\alpha_1\alpha_2...\alpha_K)=0. \end{cases}$$

The transformation $X = p/q$ has the desired effect of subtraction of random background (noise and bias) in *p* and rendering it a stationary time series suitable for subsequent cross-correlation analysis.

For all possible K-strings $\alpha_1\alpha_2...\alpha_K$, we use $X(\alpha_1\alpha_2...\alpha_K)$ 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 type i, i = 1,…, *N* (the *N* strings are arranged in a fixed order as the 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.

The vector *p* that we described is identical to the peptide frequency vector used by Stuart et al. (2002). We have pointed out in our previous paper (Chu et al. 2004) that their method of structure removal is entirely different. Starting from the vector *p*, these authors used Singular Value Decomposition (SVD) and then Dimension Reduction on their constructed matrix. The correlation distance is then used to construct the tree. In the method used in Qi et al. (2004) and Chu et al. (2004), we subtract random background through a Markov model for *q* and *X*. The SVD step is much more complicated than the method proposed by Qi et al. (2004) in both theoretical and practical considerations. In the present study, we subtract the random background through a dynamic language formula. We only need the information for all strings of length $(K-1)$ and the 20 kinds of letters instead of that for all strings of length $(K-1)$ and $(K-2)$ which is needed in the Markov model. In particular, our new method is much faster than the one used in Qi et al. (2004) and Chu et al. (2004) when K is relatively large.

Tree Construction

Different distance methods, including Fitch-Margoliash (Fitch and Margoliash 1967), neighbour-joining (Saitou and Nei 1987) and minimum evolution (Saitou and Imanishi 1989), are used to construct the phylogenetic trees. A previous study on prokaryotes shows that the topology of the trees stabilized for $K \geq 5$ (Qi et al. 2004). In the present study, we used $K = 4$ or 5 and the topologies of the resulting trees are similar. Here we present the results based on $K= 5$. The distance matrix generated from this analysis can be provideded via email z.yu@qut.edu.au.

Results and Discussion

The topologies of the trees generated by distance methods including Fitch-Margoliash (FM), neighbour-joining (NJ) and minimum evolution (ME) are identical or very similar.

Fig. 1 shows the tree based on the NJ analysis for the selected 54 organisms. The ten Archaea group together as a domain. The three eukaryotes also cluster together as a domain. And all Eubacteria fall into another domain. So the division of life into three main domains Eubacteria, Archaebacteria and Eukarya is a clean and prominent feature. No mixing among members of different domains appears in the tree. At the interspecific level, different prokaryotes in the same group (Spirochete, Cyanobacteria, Chlamydia, Gram-positive bacteria (High G+C), Gram-positive bacteria (Low G+C), Hyperthermophilic bacteria and Proteobacteria) all group together. 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 in the protein sequences as described in our method, the whole-genome tree converges to the rRNA-sequence tree as asserted in Charlebois et al. (2003).

In the phylogenetic analyses based on a few genes, the tendency of the two hyperthermophilic bacteria, *Aquae* and *Thema*, to get into Archaea, have intensified the debate on whether there has been wide-spread lateral or horizontal gene transfers among species (Doolittle 1999; Ragan 2001; Martin and Herrmann 1998). It is a consensus now that one should not equate a tree inferred from a single or a few genes to the organismic tree of life (Qi et al. 2004; Pennisi 1999). Analysis 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 (Eisen and Fraser 2003). In our tree (Fig. 1) the two hyperthermophlic bacteria group together and stay in the domain of eubacteria. This result is same as in Qi et al. (2004) and also supports the point of view in Eisen and Fraser (2003).

Fig. 2 shows the trees based on NJ analysis for the chloroplasts. The chloroplasts are separated into two major clades, one of which corresponds to the green plants *sensu lato*, or chlorophytes *s.l.* (Palmer and Delwiche 1998), 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 *Guillarida*). The close relationship between *Cyanophora* and rhodophytes *s.l.* agrees with some of the previous analyses (Stirewalt et al. 1995; De Las Rivas et al. 2002), 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 (Martin et al. 1998, 2002; Adachi et al. 2000; Moreira et al. 2000). 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. 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 Palmer and Delwiche 1998, McFadden 2001a,b for reviews).

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 analysis, particularly when dealing with a large number of genomes. Moreover, since multiple sequence alignment is not necessary, the intrinsic problems associated with this complex procedure can be avoided. In contrast to a recent similar analysis on mitochondrial genomes based on compositional vector (Stuart et al. 2002), our approach does not require prior information on gene families in the genome and is also simpler in the method used for subtraction of random background from the data (see Materials and Methods). The present method improves on the method used in the previous papers of our group (Qi et al. 2004; Chu et al. 2004). In the present method, we only need the information for all strings of length $(K-1)$ and the 20 kinds of letters instead of that for all strings of length $(K-1)$ and $(K-2)$ which is needed in the Markov model (Qi et al. 2004; Chu et al. 2004). In particular, our new method is much faster than the one used in Qi et al. (2004) when K is relatively large. We have shown that this approach is applicable for analyzing the prokaryotes as well as the much smaller genomes of chloroplasts. We believe that the present approach is an important step towards the analysis of the wealth of information provided by genome projects.

Acknowledgments

One of the author Zu-Guo Yu would like to express his thanks to Prof. B.-L. Hao ITP of Chinese Academy of Science for discussion. Financial support was provided by Youth Foundation of the Chinese National Natural Science Foundation (grant no. 10101022), and Postdoctoral Research Support Grant (no. 9900658) of Queensland University of Technology (Z.-G. Yu), an AoE Fund of The Chinese University of Hong Kong (K.H. Chu).

LITERATURE CITED

Adachi, J., P. J. Waddell, W. Martin, and M. Hasegawa. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. **50**:348-358.

Brown, J.R., Doolittle,W.F., (1997). Archaea and the prokaryote-to-eukaryote transition, *Microbiol. Mol. Biol. Rev.* **61:** 456-502.

Charlebois R.L., R.G. Beiko and M. A. Ragan, (2003). Branching out. *Nature*, **421**:217-217.

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

Chu K.H., J. Qi, Z.G. Yu and V.V. Anh, (2004). Origin and Phylogeny of Chloroplasts revealed by a simple correlation analysis of complete genome. Mol. Biol. Evol., **21**:200-206.

De Las Rivas, J., J. J. Lozano, and A. R. Ortiz, (2002). Comparative analysis of chloroplast genomes: Functional annotation, genome-based phylogeny, and deduced evolutionary patterns. *Genome Res.*, **12**:567-583.

Doolittle, R.F., (1998). Microbial genomes opened up. *Nature,* **392:** 339-342.

Doolittle, R.F., (1999), Phylogenetic classification and the universal tree. *Science,* **284:**2124-2128.

Edwards, S. V., B. Fertil, A. Giron, and P. J. Deschavanne, (2002). A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst. Biol*., **51**:599-613.

Eisen, J.A. and C.M. Fraser, (2003). Phylogenomics: intersection of evolution and genomics. *Science*, **300**:1706-1707.

Fitch, W. M., and E. Margoliash, (1967). Construction of phylogenetic trees. *Science* 1**55**:279-284.

Fitz-Gibbon, S. T., and C. H. House, (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.,* **27**:4218-4222.

Gray, M. W., (1992). The endosymbiont hypothesis revisited. *Int. Rev. Cytol.*, **141**:233-357.

Gray, M. W., (1999). Evolution of organellar genomes. *Curr. Opin. Genet. Dev.*, **9**:678-687.

Gupta, R.S., (1998), Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among Archaebacteria, Eubacteria, and Eukaryotes. *Microbiol. Mol. Biol. Rev.*, **62:** 1435- 1491.

Lemieux, C., C. Otis, and M. Turmel, (2000). Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature **403**:649-652.

Li, M., J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, **17**:149-154.

Lin, J., and M. Gerstein, (2000). Whole-genome trees based on the occurrence of folds and orthologs, implications for comparing genomes at different levels. *Genome Res.*, **10**:808-818.

Martin, W., and R. G. Herrmann, (1998). Gene transfer from organelles to the nucleus: How much, what happens, and why? *Plant Physiol*., **118**:9-17.

Martin, W., B. Stoebe, V. Goremykin, S. Hansmann, M. Hasegawa, and K. V. Kowallik, (1998). Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*, **393**:162-165.

Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny, (2002). Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. U.S.A.*, **99**:12246-12251.

Mayr, E., (1998), Two empires or three, *Proc. Natl. Acad. Sci. U.S.A.*, **95:** 9720-9723.

McFadden, G. I., (2001a). Primary and secondary endosymbiosis and the origin of plastids. *J. Phycol*., **37**:951-959.

McFadden, G. I., (2001b). Chloroplast origin and integration. *Plant Physiol.*, **125**:50-53

Moreira, D., H. LE Guyader, and H. Ppilippe, (2000). The origin of red algae and the evolution of chloroplasts. *Nature,* **405**:69-72.

Palmer, J. D., and C. F. Delwiche, (1998). The origin and evolution of plastids and their genomes. In *Molecular Systematics of Plants II DNA Sequencing* (eds. Soltis, D.E., Soltis, P.S. and Doyle, J.J.), pp. 345-409. Kluwer, London.

Pennisi, E., (1999). Is it the time to uproot the tree of life? *Science* **284**:1305-1308.

Qi, J., B. Wang, and B. Hao, (2004). Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol*., **58**:1-11.

Ragan M.A., (2001). Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Gen. Dev*., **11**:620-626.

Saitou, N., and T. Imanishi, (1989). Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Mol. Biol. Evol.*, **6**:514-525.

Saitou, N., and M. Nei, (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**:406-425.

Sankoff, D., G. Leaduc, N. Antoine, B. Paquin, B. F. Lang, and R. Cedergren, (1992). Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. U.S.A.*, **89**:6575-6579.

Stirewalt, V. L., C. B. Michalowski, W. Loffelhardt, H. J. Bohnert, and D. A. Bryant, (1995). Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*. *Plant Mol. Biol. Rep.*, **13**:327-332.

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

Tekaia, F., A. Lazcano, and B. Dujon, (1999). The genomic tree as revealed from whole proteome

comparisons. *Genome Res*., **9**:550-557.

Turmel, M., C. Otis, and C. Lemieux, (1999). The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: Insights into the architecture of ancestral chloroplast genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **96**:10248-10253.

Turmel, M., C. Otis, and C. Lemieux, (2002). The chloroplast and mitochondrial genome sequences of the charophyte C*haetosphaeridium globosum:* Insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. *Proc. Natl. Acad. Sci. U.S.A.*, **99**:11275-11280.

Weiss, O., M. A. Jimenez, and H. Herzel, (2000). Information content of protein sequences. *J. Theor. Biol*., **206**:379-386.

Woese, C.R., (1987), Bacterial evolution, *Microbiol. Rev*., **51:** 221-271.

Woese, C.R., Kandler, O. & Wheelis, M.L., (1990), Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya, *Proc. Natl. Acad. Sci. USA,* **87:** 4576-4579.

Yu, Z.G., and P. Jiang, (2001). Distance, correlation and mutual information among portraits of organisms based on complete genomes. *Phys. Lett. A*, **286**:34-46.

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

Yu, Z.G., V. Anh and K. S. Lau, (2004). Chaos game representation, and multifractal and correlation analysis of protein sequences from complete genome based on detailed HP model. *J. Theor. Biol*. **226**:341-348

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

Figures

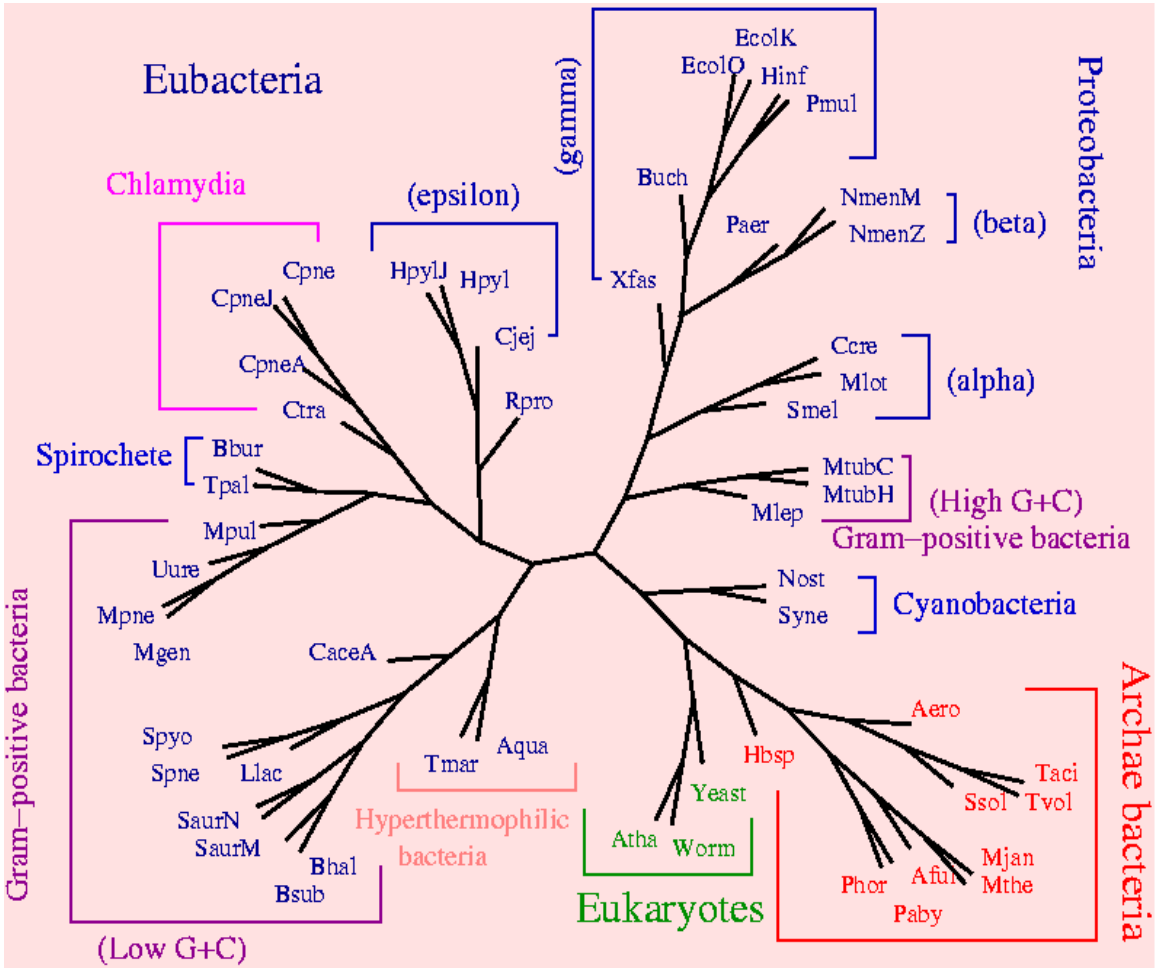

Fig. 1 Phylogeny of 54 organisms (prokaryotes and eukaryotes) selected on our new compositional approach..

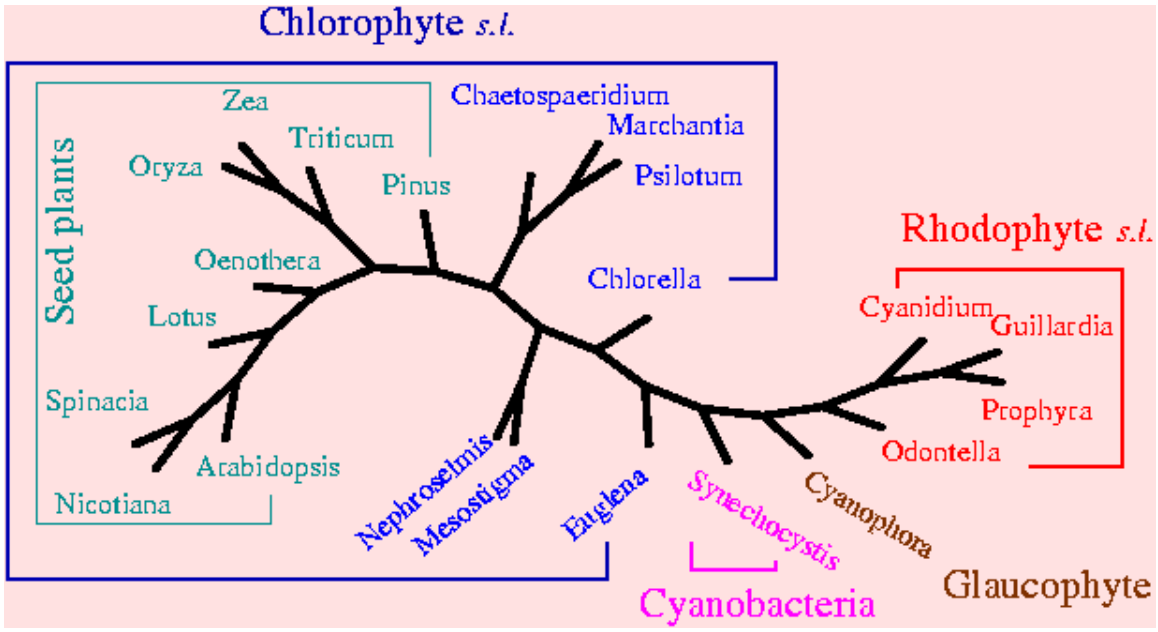

Fig. 2 Phylogeny of chloroplast genomes based on our new compositional approach.