

Classification of Non-Animals and Invertebrates Based on Amino Acid Composition of Complete Mitochondrial Genomes

Kenji Sorimachi^{1,2} & Teiji Okayasu³

¹ Life Science Research Center, Higashi-Kaizawa, Takasaki, Gunma, Japan

² Educational Support Center, Dokkyo Medical University, Mibu, Tochigi, Japan

³ Center for Medical Informatics, Dokkyo Medical University, Mibu, Tochigi, Japan

Correspondence: Kenji Sorimachi, Life Science Research Center, Higashi-Kaizawa, Takasaki, Gunma 370-0041, Japan. Tel: 81-27-352-2955. Email: kenjis@jcom.home.ne.jp

Received: September 11, 2013 Accepted: October 9, 2013 Online Published: October 21, 2013

doi:10.5539/ijb.v6n1p1

URL: <http://dx.doi.org/10.5539/ijb.v6n1p1>

Abstract

Amino acid compositions were predicted from data sets of 47 non-animal and 58 invertebrate animal complete mitochondrial genomes, which were chosen alphabetically based on scientific names without considering biological characteristics. Using Ward's clustering method with amino acid composition or nucleotide content as traits, non-animals were classified into Plantae, Chromalveolata, and Fungi, and invertebrates were classified into Animalia and primitive groups, Amoebozoa, Excavata, Protista, and Choanozoa. A combined sample set of primitive eukaryotes was also examined by cluster analysis using amino acid composition and nucleotide content. Some Amoebozoa comprised a single cluster, whereas other Amoebozoa were grouped with other organisms (Excavata, Prosista, Chromaleolata, Fungi and Plantae), indicating their close relationships. Choanozoa (choanoflagellates; *Monosiga brevicollis*), considered the closest living relatives of animals, were found to be instead closely related to Fungi (*Smittium culisetae*, *Pleurotus ostreatus*, and *Epidermophyton floccosum*) and Excavata (*Malawimonas jakobiformis*). Our results demonstrate that amino acid composition and nucleotide content are useful indices for characterizing non-animal and invertebrate complete mitochondrial genomes.

Keywords: Ward's clustering analysis, evolution, phylogenetic tree, complete mitochondrial genome, amino acid composition, nucleotide content, plant, animal

1. Introduction

Methodology for analyzing nucleotide gene sequences was first developed in 1975 (Sanger & Coulson, 1975; Maxam & Gilbert, 1977). A comprehensive analysis of *Haemophilus influenzae* was carried out in 1995 (Fleischmann et al., 1995), with a draft of the complete human genome obtained in 2001 (Lander et al., 2001; Venter et al., 2001). Because nucleotide mutations are associated with biological evolution, nucleotide and amino acid sequence data have been used to construct an enormous number of phylogenetic trees (Dayhoff, Park, & McLaughlin, 1977; Sogin, Elwood, & Gunderson, 1986; Doolittle & Brown, 1994; Maizels & Weiner, 1994; DePouplana, Turner, Steer, & Schimmel, 1998; Woese & Fox, 1977; Weisburg, Brns, Pelletier, & Lane, 1991) that have helped us to understand biological evolution. Because nucleotide and amino acid substitution rates differ among genes, however, universal phylogenetic trees accurately modeling true phylogenies cannot be reconstructed based on current knowledge levels. For instance, different analytical methods, such as Ward's clustering (Ward, 1963) and neighbor-joining (Saitou & Nei, 1987), have yielded different phylogenetic trees using the same data set, with analysis of different traits yielding different results (Sorimachi & Okayasu, 2013). Although we cannot presently construct phylogenetic trees that are universally representative of actual phylogenies, the scientific validity of phylogenetic trees cannot be denied.

Phylogenetic analyses have primarily utilized nucleotide and amino acid sequence data, with nucleotide content and amino acid composition rarely used to investigate biological phenomena such as evolution. Studies based on nucleotide or amino acid sequences are applicable to genes or regions of relatively small length, but not to entire genomes consisting of huge numbers of nucleotides and many genes. Nevertheless, simple comparison of sequence differences between genes, both within and among species, is of course still useful.

Sueoka (1961) was the first to analyze bacterial cellular amino acid composition. More recently, our laboratory

has independently analyzed cellular amino acid composition of bacteria, archaea, and eukaryotes (Sorimachi, 1999). Graphical representations and diagrammatic approaches to the study of complicated biological systems can provide intuitive pictures and useful insights (Chou, 1990; Qi X. Q., Won, & Qi Z. H., 2007). With the aid of certain graphical representations, simple patterns representing complicated organisms can be easily discerned from huge genomic data sets. For example, when radar charts are used to visualize cellular amino acid compositions, their star-shaped patterns are similar among various organisms, with any differences appearing to reflect biological evolution (Sorimachi, 1999). In addition, amino acid compositions deduced from complete genomes resemble those obtained from amino acid analyses of cell lysates (Sorimachi et al., 2001).

Intra-species nucleotide content was first analyzed by Chargaff, who reported that $G = C$, $A = T$, and $(G + A) = (C + T)$. This rule has been dubbed Chargaff's first parity rule (Chargaff, 1950), and is understandable based on the double-stranded structure of DNA (Watson & Crick, 1953). This rule is also applicable to each single strand of nuclear DNA from individual species, a case that has been termed Chargaff's second parity rule (Runder, Karkas, & Chargaff, 1968). Because these rules are based on values normalized to 1 (i.e., $G + C + A + T = 1$), nucleotide contents are expressed by their ratios. The second parity rule is more difficult to understand, because it is difficult to imagine how G and C or T and A pairs are formed in a single DNA strand. This puzzle has recently been solved mathematically using similarity of the forward and reverse strands and homogeneity of the DNA strand over genome structure (Sorimachi, 2009). Although Chargaff's parity rules were originally formulated as intra-species phenomena, they can be expanded to encompass inter-species relationships using data from a large number of complete genomes (Mitchell & Brigde, 2006). These results indicate that nucleotide content, similar to amino acid composition (Sorimachi & Okayasu, 2003), can be used to characterize whole genomes (Sorimachi & Okayasu, 2004a).

We have recently demonstrated the existence of natural selection in vertebrate evolution using phylogenetic trees derived from amino acid composition or nucleotide content of complete mitochondrial genomes (Sorimachi & Okayasu, 2013). Vertebrate mitochondrial DNA contains 13 genes, whereas gene number varies in plant and invertebrate mitochondrial DNA. In the present study, we investigated evolution in other eukaryotes using phylogenetic trees constructed from amino acid composition or nucleotide content data.

2. Materials and Methods

Mitochondrial genome data were obtained from the National Center for Biotechnology Information (NCBI, <http://www.ncbi.nlm.nih.gov/sites>). In an earlier study, organisms were chosen according to the alphabetical order of their scientific names without considering their characteristics (Sorimachi & Okayasu, 2008). Nucleotide contents of complete mitochondrial genomes were calculated from their complete corresponding single-strand DNA (Sorimachi & Okayasu, 2008) and normalized to 1 ($G + C + T + A = 1$). Predicted amino acid compositions were estimated for mitochondrial genome coding regions. Ratios of individual amino acids to total amino acids were calculated as percentages of total amino acids, and the presentation order of each amino acid on radar charts was determined based on their HPLC elution orders (Sorimachi, 1999). In our previous studies (Sorimachi & Okayasu, 2013; Sorimachi, Okayasu, Ohhira, Masawa, & Fukasawa, 2013), phylogenetic trees obtained from Ward's clustering method (Ward, 1963) using amino acid compositions or nucleotide contents predicted from complete vertebrate mitochondrial genomes slightly differed from those obtained from neighbor-joining method (Saitou & Nei, 1987) using 16S rRNA sequences. However, their basic results were consistent between two different methods. Therefore, Ward's clustering method was carried out in the present study. Classifications based on Ward's clustering method (Ward, 1963) were conducted using multivariate software developed by ESMI (Tokyo, Japan).

3. Results

3.1 Amino Acid Compositions Encoded in Mitochondrial DNA

In a previous study of vertebrate mitochondrial DNA, encoded amino acid compositions were similar among species. Contents (percentages) of some specific amino acids differed significantly among species, however, allowing classification of vertebrates into terrestrial and aquatic species (Sorimachi & Okayasu, 2013). To expand the scope of our investigation using amino acid compositions, we examined a wider range of eukaryote species, i.e., liverworts (*Marchantia polymorpha*), fungi (*Monoblepharella* and *Epidermophyton floccosum*), magnoliopsida (*Arabidopsis thaliana*), algae (*Cyanidioschyzon merolae*), and protists (*Phytophthora sojae*) (Figure 1). As shown in Figure 1, amino acid compositions differed among the various species examined. In every species, both Leu and Ile contents were high compared with other amino acids. Leu content was higher than that of Ile in some species, with the opposite true in other cases. In vertebrate mitochondrial DNA, Leu contents are usually higher than those of both Ile and Phe, giving rise in the case of *Monoblepharella* to a

“pennon shape” in the radar plot (Sorimachi & Okayasu, 2013). In addition, Asp and Ser contents were high; these high Asp and Lys contents, coupled with low Glu levels, were the cause of a characteristic “V-shape” pattern observed in representations of amino acid composition. These results indicate that non-animal mitochondrial DNA can be classified based on amino acid composition.

3.2 Cluster Analysis of Non-Animal Mitochondria

As new knowledge accumulated after Linnaeus formulated his well-known historical classification of organisms, five-kingdom (Whittaker, 1969), three-domain (Woese, Kandler, & Wheelis, 1990), and six-kingdom (Cavalier-Smith, 1998) theories were subsequently proposed. In the present study, we have primarily followed the six-kingdom classification of Cavalier-Smith (1998). Cluster analysis of non-animal mitochondria was carried out based on Ward’s method using contents of 20 amino acids as traits. As shown in Figure 2, two major clusters were obtained. The first major cluster in the tree consists of three sub-clusters: (1) Chromalveolata (*Phytophthora sojae*, *Phytophthora ramorum*, *Saprolegnia ferax*, *Phytophthora infestans*, and *Rhodomonas salina*); (2) Fungi (*Smittium culisetae*, *Pleurotus ostreatus*, and *Epidermophyton floccosum*), and Plantae (*Cyanidioschyzon merolae* and *Chondrus crispus*); (3) Fungi (*Podospira anserina*, *Gibberella zeae*, and *Moniliophthora perniciosa*) and Chromalveolata (*Rhodomonas salina* and *Cafeteria roenbergensis*).

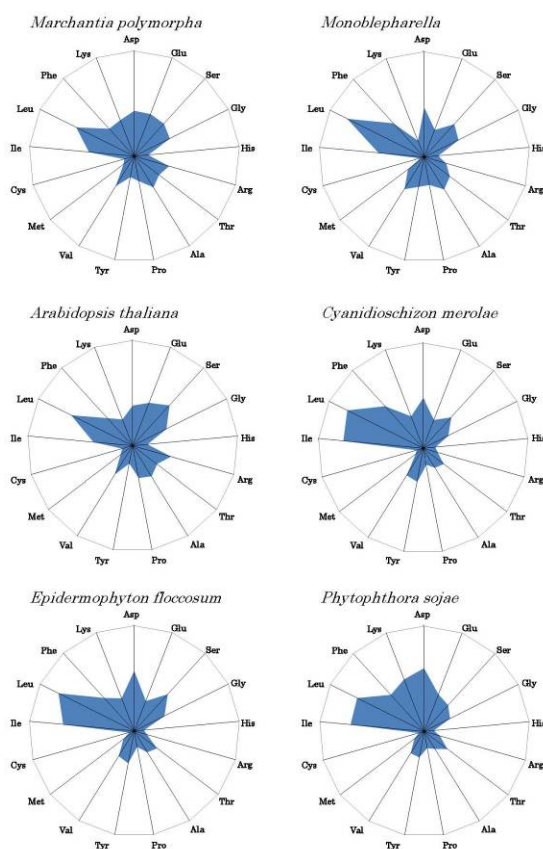


Figure 1. Radar charts of amino acid composition predicted from complete mitochondrial genomes. Values represent percentages of total amino acids

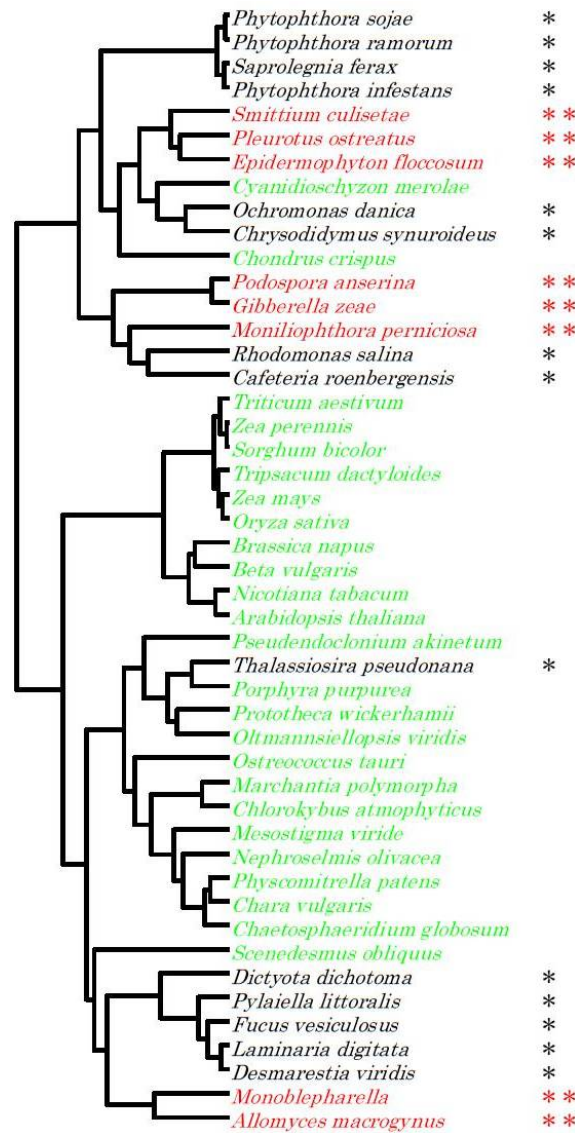


Figure 2. Phylogenetic tree based on Ward's clustering method using amino acid composition. Black (*), red (**), and green (null) characters represent Chromalveolata, Fungi, and Plantae, respectively

The second major cluster in Figure 2 consists of three sub-clusters. One sub-cluster comprises land plants (angiosperms or Magnoliophyta), which are classified into monocots or Liliopsida (*Triticum aestivum*, *Zea perennis*, *Sorghum bicolor*, *Tripsacum dactyloides*, *Zea mays*, and *Oryza sativa*) and into Magnoliopsida (*Brassica napus*, *Beta vulgaris*, *Nicotiana tabacum*, and *Arabidopsis thaliana*). Members of Magnoliopsida are completely separated from Liliopsida. The second sub-cluster consists of Plantae (*Pseudendozonium akinetum*, *Thalassiosira pseudonana*, *Porphyra purpurea*, *Oltmannsiellopsis viridis*, *Ostreococcus tauri*, *Chlorokybus atmophyticus*, *Mesostigma viride*, *Nephroselmis olivacea*, *Chara vulgaris*, *Chaetosphaeridium globosum*, *Marchantia polymorpha*, and *Physcomitrella patens*). The third sub-cluster contains representatives of Plantae (*Scenedesmus obliquus*), Chromalveolata (*Dictyota dichotoma*, *Pylaiella littoralis*, *Fucus vesiculosus*, *Laminaria digitata*, and *Desmarestia viridis*), and Fungi (*Monoblepharella* and *Allomyces macrogynus*) – all of which are aquatic organisms except for moss.

3.3 Comparison of Amino Acid Compositions in Major Clusters

The sub-cluster in Figure 2 consisting of Protista (*Phytophthora sojae*) and Chromalveolata (*Phytophthora ramorum*, *Saprolegnia ferax*, and *Phytophthora infestans*) is characterized by higher Ile content

relative to Leu content, with radar plots displaying a “mountain-shape” because of high Phe, Lys, and Asp content and lower Glu, Ser, and Gly content (Figure 3). The sub-cluster comprising Fungi (*Smittium culisetæ*, *Pleurotus ostreatus*, and *Epidermophyton floccosum*), Plantae (*Cyanidioschyzon merolæ* and *Chondrus crispus*), and Chromalveolata (*Ochromonas danica* and *Chrysodidymus synuroideus*) is characterized by higher Leu content than Ile content and by the formation of a “V-shape” reflecting Asp, Glu, and Ser content. In the third sub-cluster, Fungi (*Podospora anserina*, *Gibberella zeae*, and *Moniliophthora perniciosa*) have lower Phe and Lys content than do Chromalveolata (*Rhodomonas salina* and *Cafeteria roenbergensis*).

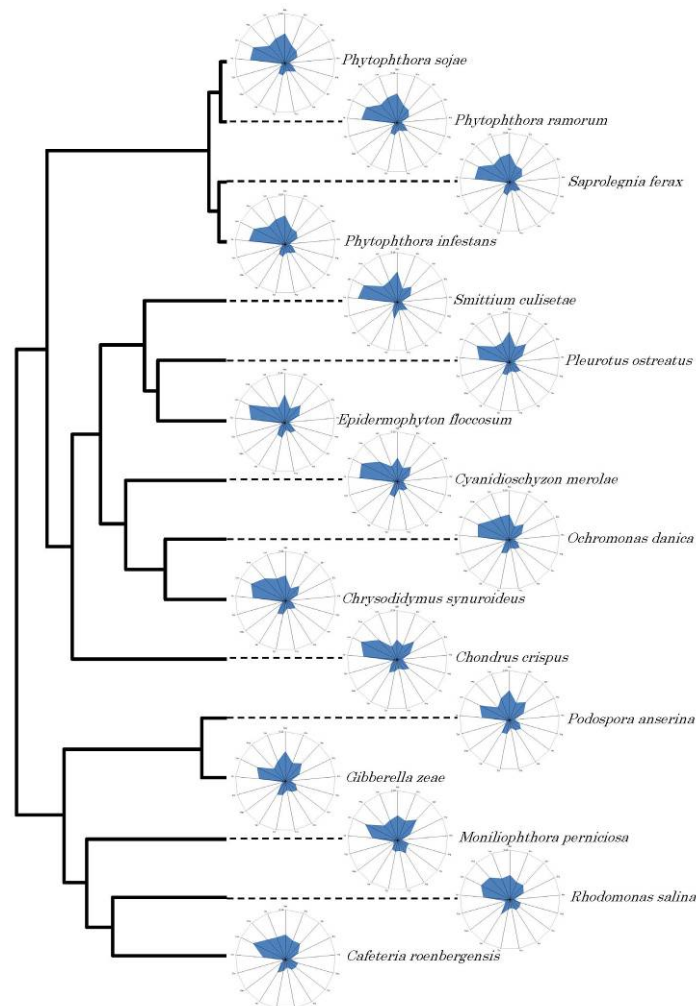


Figure 3. The first major cluster from Figure 2, a phylogenetic tree based on Ward's clustering using amino acid composition

The other major cluster of Figure 2 is made up of two sub-clusters. In the first sub-cluster, consisting of land plants divided into Liliopsida (*Triticum aestivum*, *Zea perennis*, *Sorghum bicolor*, *Tripsacum dactyloides*, *Zea mays*, and *Oryza sativa*) and Magnoliopsida (*Brassica napus*, *Beta vulgaris*, *Nicotiana tabacum*, and *Arabidopsis thaliana*) (Figure 4), Leu content is higher than Ile content, and Phe content is higher than Lys content.

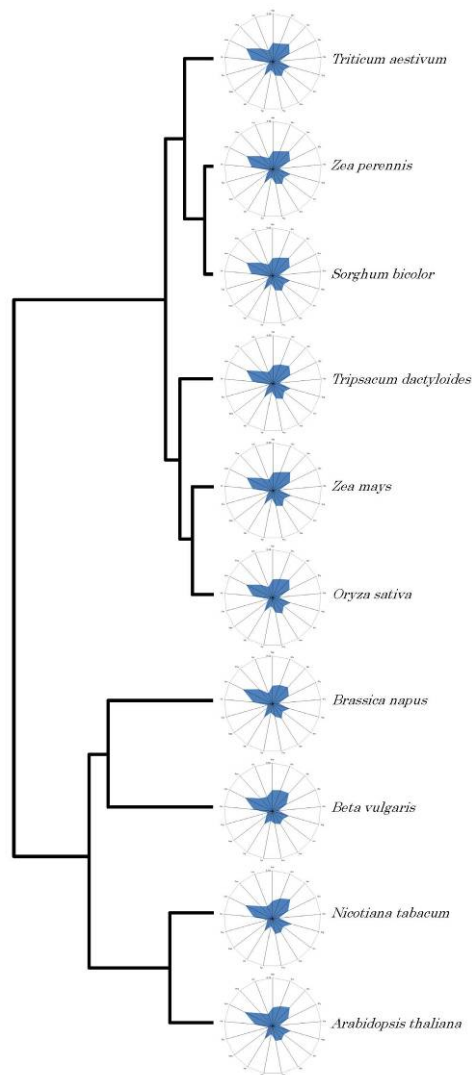


Figure 4. The second major cluster from Figure 2, a phylogenetic tree based on Ward's clustering using amino acid composition

The other sub-cluster, which includes Plantae (*Pseudendozonium akinetum*, *Porphyra purpurea*, *Prototheca wickerhamii*, *Oltmannsiellopsis viridis*, *Ostreococcus tauri*, *Marchantia polymorpha*, *Chlorokybus atmophyticus*, *Mesostigma viride*, *Nephroselmis olivacea*, *Physcomitrella patens*, *Chara vulgaris*, and *Chaetosphaeridium globosum*) and a representative of Chromalveolata (*Thalassiosira pseudonana*), is characterized by higher Leu content than Ile content (Figure 5). In Fungi (*Monoblepharella* and *Allomyces macrogynus*), also included in this sub-cluster, Glu content is lower than that of both Asp and Ser, giving rise to a “V-shape” relationship among Asp, Glu, and Ser contents on the plot. Conversely, the “V-shape” is absent, with a weak convex shape noted in some cases, in Plantae and Chromalveolata of this sub-cluster (*Dictyota dichotoma*, *Pylaiella littoralis*, *Fucus vesiculosus*, *Laminaria digitata*, and *Desmarestia viridis*).

3.4 Cluster Analysis Based on Nucleotide Content

Using nucleotide content calculated from complete mitochondrial genomes, Ward's clustering method yielded results similar to those based on amino acid composition (Figure 6). Three major clusters were obtained: two clusters consisting of Chromalveolata, Fungi, and Plantae, and a clearly separated cluster comprising land plants and Fungi. This is consistent with results obtained using amino acid composition (Figure 2).

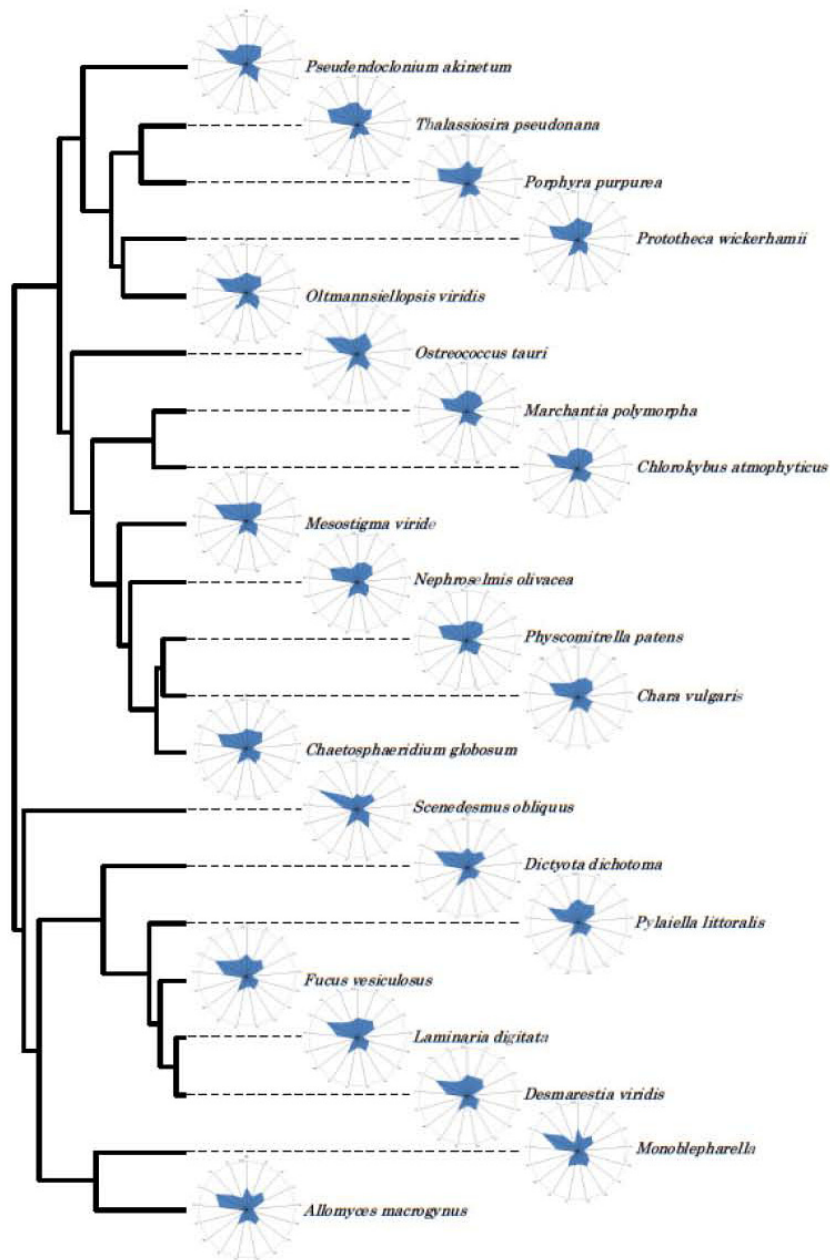


Figure 5. The third major cluster from Figure 2, a phylogenetic tree based on Ward's clustering using amino acid composition

3.5 Cluster Analysis Based on Nucleotide Content

Using nucleotide content calculated from complete mitochondrial genomes, Ward's clustering method yielded results similar to those based on amino acid composition (Figure 6). Three major clusters were obtained: two clusters consisting of Chromalveolata, Fungi, and Plantae, and a clearly separated cluster comprising land plants and Fungi. This is consistent with results obtained using amino acid composition (Figure 2).

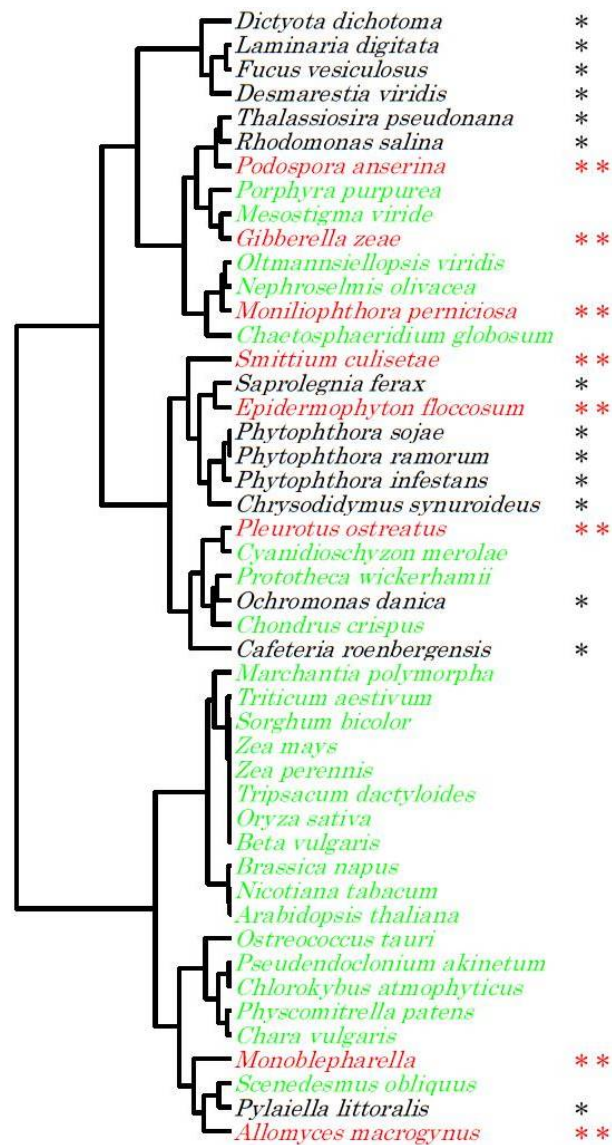


Figure 6. Phylogenetic tree based on Ward's clustering using nucleotide content. Black (*), red (**), and green (null) characters indicate Chromalveolata, Fungi, and Plantae, respectively

3.6 Amino Acid Compositions of Invertebrates

In our previous studies (Sorimachi & Okayasu, 2008), 58 invertebrates were chosen according to the alphabetical order of their scientific names without considering their biological characteristics. Amino acid compositions of seven species are represented in Figure 7. As shown in this figure, Protista (*Tetrahymena pyriformis*), Choanozoa (*Monosiga brevicollis*), Excavata (*Malawimonas jakobiformis* and *Reclinomonas americana*), and Amoebozoa (*Physarum polycephalum*, *Polysphondylium pallidum*, and *Dictyostelium discoideum*) are characterized by comparatively high Leu and Ile content. Ile content is significantly higher than Leu content in *Malawimonas jakobiformis* and *Physarum polycephalum*. Asp content is much higher than both Lys and Glu content, forming a “steep mountain-shape” in *Tetrahymena pyriformis*, *Monosiga brevicollis*, and *Malawimonas jakobiformis* and a “gentle slope mountain-shape” in *Physarum polycephalum*, *Reclinomonas americana*, and *Polysphondylium pallidum*.

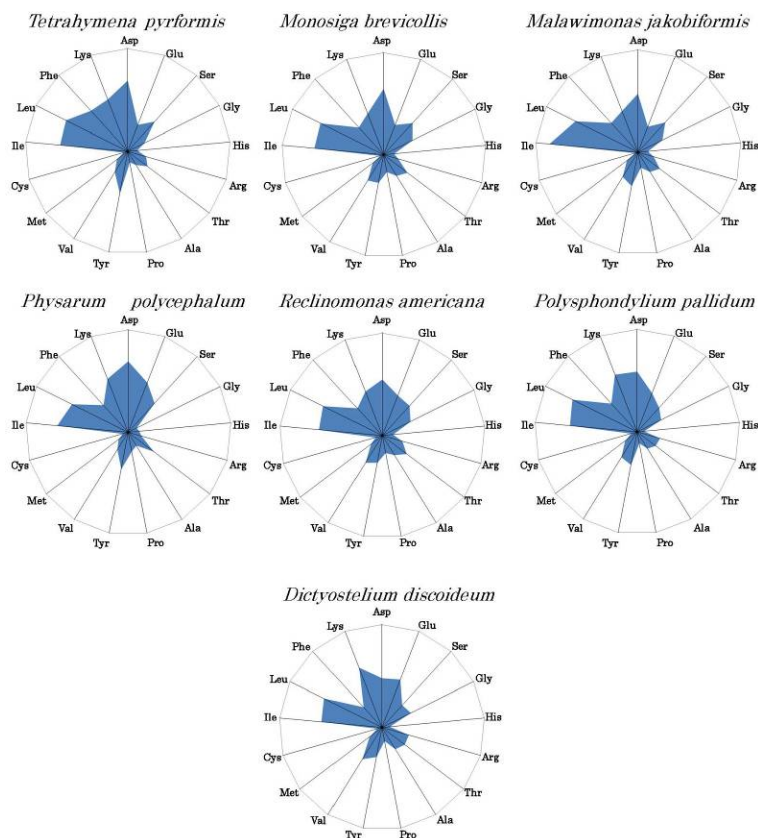


Figure 7. Radar charts of amino acid composition predicted from complete mitochondrial genomes. Values correspond to percentages of total amino acids

3.7 Cluster Analysis of Invertebrate Mitochondria

In the present study, the above seven species plus the remaining 51 invertebrates were categorized based on amino acid composition using Ward's clustering method (Figure 8 and Supplemental Figure 1). The set of 58 invertebrates was classified into two groups: kingdom Animalia, divided into clusters I-IV, and a group consisting of Amoebozoa (*Dictyostelium citrinum*, *Dictyostelium discoideum*, *Polysphondylium pallidum*, and *Physarum polycephalum*), Excavata (*Reclinomonas americana* and *Malawimonas jakobiformis*), Choanozoa (*Monosiga brevicollis*), and Protista (*Tetrahymena pyriformis*).

Within Animalia, Cluster I consists of Platyhelminthes (*Paragonimus westermani*, *Schistosoma mansoni*, *Echinococcus granulosus*), Nematoda (*Thaumamermis cosgrovei*, *Strelkovimermis spiculatus*, *Romanomermis culicivorax*, and *Caenorhabditis elegans*), and Arthropoda (*Varroa destructor*, *Bombyx mori*, and *Adoxophyes honmai*). Cluster II comprises Placozoa (*Trichoplax adhaerens*), Cnidaria (*Porites porites*, *Pavona clavus*, *Briareum asbestinum*), Porifera (*Geodia neptuni*), Mollusca (*Venerupis philippinarum*, *Lampsilis ornata*, and *Haliotis rubra*), Arthropoda (*Tigriopus japonicus*, *Tigriopus californicus*, *Lepeophtheirus salmonis*, *Vargula hilgendorfii*, *Ligia oceanica*, and *Daphnia pulex*) and Chordata (*Branchiostoma belcheri*). Cluster III is made up of Protozoa (*Plasmodium vivax*), Mollusca (*Sepioteuthis lessoniana*, *Octopus vulgaris*, *Todarodes pacificus*, and *Dosidicus gigas*) and Arthropoda (*Tricholepidion gertschi*, *Reticulitermes hageni*, *Reticulitermes flavipes*, *Hutchinsoniella macracantha*, *Pagurus longicarpus*, *Eriocheir sinensis*, *Drosophila simulans*, *Drosophila melanogaster*, *Bactrocera oleae*, *Bactrocera dorsalis*, and *Anopheles gambiae*). Cluster IV includes Xenoturbellida (*Xenoturbella bocki*), Hemichordata (*Saccoglossus kowalevskii* and *Balanoglossus carnosus*), and Echinodermata (*Luidia quinalia*, *Asterias amurensis*, *Acanthaster planci*, and *Acanthaster brevispinus*).

Members of kingdom Animalia were completely separated from other kingdoms based on Ward's clustering (Figure 8 and Supplemental Figure 1), and amino acid compositions of Animalia differed accordingly from those of other groups consisting of Amoebozoa, Excavata, Choanozoa and Protista. The latter group is characterized

by Ile content higher than or similar to Leu content, and by Asp content higher than that of Lys and Glu, with the exception of *Dictyostelium* (Figure 7). In animal mitochondria, high Leu, Ile, and Phe contents are characteristic. The highest observed amino acid percentages were those for Leu, giving rise to a “pen-point shape” for the representation of the content of these three amino acids. High Ser and Gly contents were also noted, resulting in a “knife-edge shape” in the plots (Supplemental Figure 1). Both characteristic shapes resemble a flying bird.

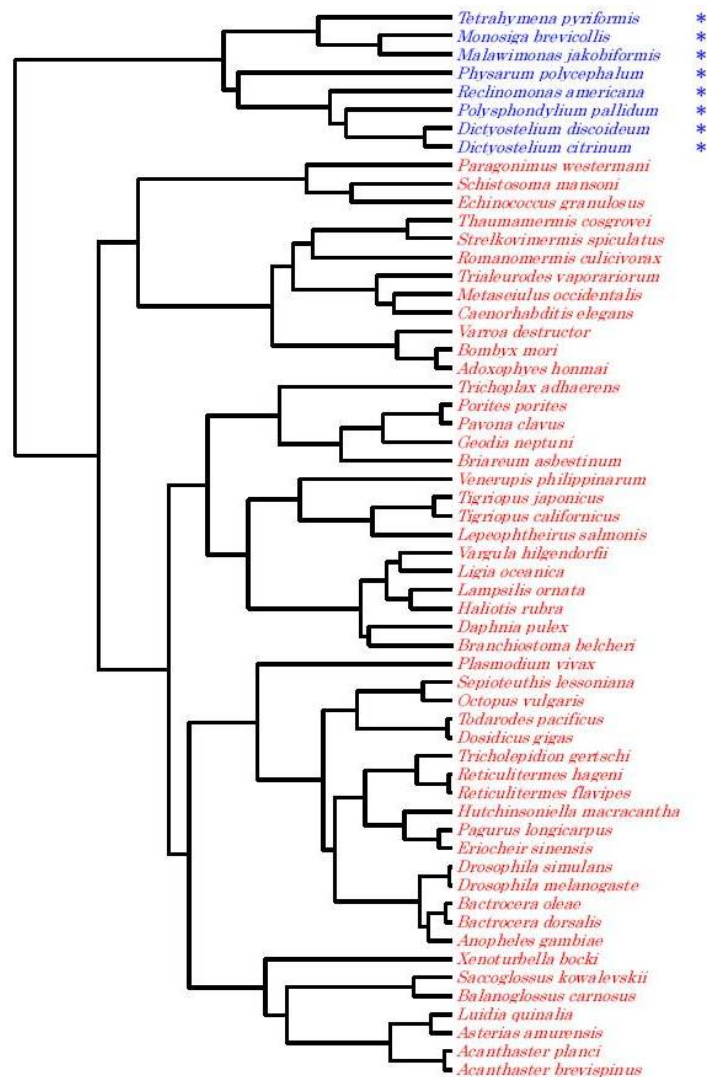


Figure 8. Phylogenetic tree based on Ward's clustering using amino acid composition. Blue characters (*) correspond to Amoebozoa, Excavata, and Choanozoa, and red (null) characters indicate Animalia

3.8 Cluster Analysis of Primitive Organisms

In the trees generated by cluster analysis, organisms belonging to one of the major clusters comprising Chromalveolata, Fungi, and Plantae are clearly separated from other plants (Figures 2 and 6), and the cluster consisting of Protista, Choanozoa, Excavata, and Amoebozoa is also separated from Animalia (Figure 8 and Supplemental Figure 1). Because these organisms seem to be primitive in Plantae or Animalia, Ward's clustering based on amino acid composition was carried out with a combined sample set (Figure 9). Amoebozoa (*Polysphondylium pallidum*, *Dictyostelium discoideum*, and *Dictyostelium citrinum*) formed a single cluster, while other Amoebozoa (*Physarum polycephalum*) and Excavata (*Reclinomonas americana*) fell into a cluster that included Fungi. Protista (*Tetrahymena pyriformis*), Choanozoa (*Monosiga brevicollis*), and Excavata (*Malawimonas jakobiformis*) were placed into another cluster consisting of Plantae (*Cyanidioschyzon merolae*) and Fungi (*Smittium culisetae*, *Pleurotus ostreatus*, and *Epidermophyton floccosum*). When a sample set

including many other plants was analyzed, similar results were obtained (Supplemental Figure 2).

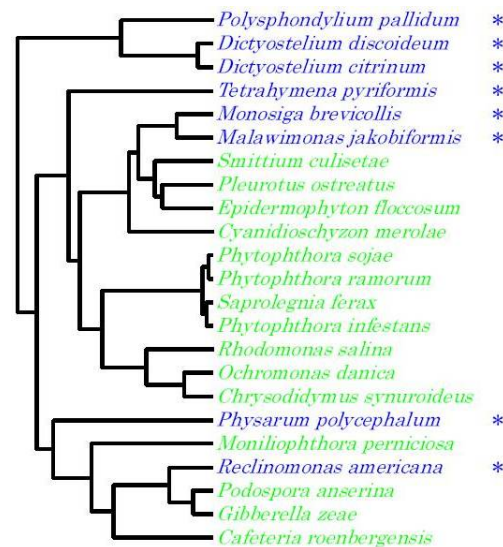


Figure 9. Phylogenetic tree based on Ward's clustering analysis using amino acid composition. Blue (*) characters designate Amoebozoa, Excavata, and Choanozoa, and green (null) characters correspond to Chromalveolata, Fungi, and Plantae

4. Discussion

In our previous study related to vertebrate evolution, vertebrates were clearly differentiated into two groups – terrestrial and aquatic – based on a data set chosen according to the alphabetical order of species names without considering species characteristics (Sorimachi & Okayasu, 2013). Species selection criteria are important for classification, because sampling affects classification results. We indeed found that when we varied which samples were chosen, differences were observed in the tree topologies recovered from the analyses (unpublished data).

Using amino acid composition as the examined trait in cluster analysis generated a better classification than did nucleotide content (Figures 2 and 6). The number of traits associated with amino acid composition (20) was greater than the number representing nucleotide content (4), resulting in good differentiation among organisms. Although amino acid composition provided better classification results than nucleotide content in the vertebrate evolutionary analysis, a significant separation between terrestrial and aquatic vertebrates was also obtained in the cluster analysis using nucleotide content (Sorimachi & Okayasu, 2013). In general, an increased number of traits reduce the probability of coincidental similarity in cluster analyses, resulting in better species classification. In contrast, an increase in the number of samples increases the probability of coincidental similarity, worsening species classification.

Terrestrial and aquatic vertebrates were completely separated in vertebrate phylogenetic trees based on amino acid composition or nucleotide content, with the exception of hagfish (*Eptatretus burgeri*), which fell into the terrestrial group (Sorimachi & Okayasu, 2013). This anomalous placement seems to be a consequence of its primitive characteristics (Janvier, 2010). Amino acid compositions predicted from complete vertebrate mitochondrial genomes were very similar to one another, although some amino acid contents differed significantly between the two groups and can thus be used to characterize them. Clear separation between terrestrial and aquatic vertebrates was thus obtained (Sorimachi & Okayasu, 2013). In contrast, species such as algae, mosses, and fungi are evolutionarily highly diverged, as can be seen by their large amino acid composition differences (Figures 1, 3-5). This means that evolutionary divergence in Plantae and Fungi may proceed in several different directions, resulting in multiple characteristic changes in amino acid composition. Consequently, the probability of coincidental similarity may increase, such that classification results obtained for Plantae, Chromalveolata, and Fungi were worse than for vertebrates. The latter have diverged in just two directions – terrestrial and aquatic – providing a good separation.

Within kingdom Plantae, members of Angiospermae were completely separated from liverworts (*Marchantiophyta* and *Bryophyta*) among land plants (Embryophyta), while land plants were separated from algae (Figure 2). Land plants and algae have thus evolved independently under natural selection in terrestrial and aquatic spheres, respectively, as observed in vertebrate evolution (Sorimachi & Okayasu, 2013). In addition, Magnoliopsida was completely separated from Liliopsida in Angiospermae, although amino acid compositions were similar between both groups (Figure 4). Although terrestrial and aquatic vertebrates were clearly separated in our previous study (Sorimachi & Okayasu, 2013), amino acid compositions were substantially similar between the two groups and among various vertebrates (Sorimachi & Okayasu, 2013). These clear separations may be due to significant changes in amino acid composition between the two groups or among many organisms. Characteristic differences in amino acid composition are indeed observed between terrestrial and aquatic vertebrates (Sorimachi & Okayasu, 2013). Similarly, bacteria are classified into two groups, “S-type” represented by *Staphylococcus aureus* and “E-type” represented by *Escherichia coli*, based on differences in amino acid compositions (Sorimachi & Okayasu, 2004b; Okayasu & Sorimachi, 2009). This phenomenon has been confirmed by other bioinformatics analyses (Qi X. Q., Won, & Qi Z. H., 2007).

Molecular phylogenetic trees are typically constructed from nucleotide or amino acid sequences of gene(s), such as cytochrome C (Dayhoff, Park, & McLaughlin, 1977), 18S rRNA (Anguinaldo et al., 1997), 16S rRNA (Sorimachi & Okayasu, 2013; Sorimachi et al., 2013; Weiseburg, Brns, Pelletier, & Lane, 1991; Puslednik & Serb, 2008; Poulakakis, Pakaki, Mylonas, & Lymberakis, 2008), 12S rRNA (Puslednik & Serb, 2008; Poulakakis, Pakaki, Mylonas, & Lymberakis, 2008), and tRNA (Maizels & Weiner, 1994; DePouplana, Turner, Steer, & Schimmel, 1998). Vertebrate mitochondrial genomes are of similar size among different species, and comprise 13 coding genes. This size appears to be too long, however, to use the entire nucleotide sequence for phylogenetic construction. In addition, plant mitochondrial genomes vary widely in size (2.5×10^4 – 7.0×10^5 bp). Based on these facts, methods based on nucleotide or amino acid sequences do not therefore seem to be applicable for phylogenetic tree construction. However, because the ratios of individual amino acids to total amino acids or nucleotides to total nucleotides are independent of genome size and species differences, these indices, which represent whole genome characteristics, are useful tools for genomic research. Consistent with this expectation, reasonable results based on amino acid composition and nucleotide content were obtained in this investigation and in our previous study (Sorimachi & Okayasu, 2013).

Choanoflagellates (e.g., *Monosiga brevicollis*), which have a unique cellular morphology consisting of a single flagellum surrounded by a “collar” of microvilli, are unicellular aquatic flagellates. This cellular morphology is very similar to that of the collared cells (choanocytes) of sponges. On this basis, colonial choanoflagellates are thought to be the closest living relatives of animals. Recent molecular phylogenetic studies have investigated the relationship between Choanozoa and Metazoa (King et al., 2008), but it is difficult to assert that the evolutionary process leading from a highly differentiated unicellular organism to an ancient multi-cellular organism with organized tissues has been elucidated because of the large difference between two organisms (Lavrov, Forget, Kelly, & Lang, 2005). Based on cluster analysis using amino acid composition in our current study, *Monosiga brevicollis* was found to be closely related to Excavata (*Malawimonas jakobiformis*), Fungi (*Smittium culisetae*, *Pleurotus ostreatus*, and *Epidermophyton floccosum*), and Plantae (*Cyanidioschyzon merolae*) (Figure 9) rather than to animals including sponges (Figure 8). Differing results between our study and those of others are due to differences in algorithms and analyzed traits, as phylogenetic trees are not universally representative (unpublished data).

In the present study, primitive organisms, i.e., water-mold (*Phytophthora sojae*, *Phytophthora ramorum*, *Phytophthora infestans*, and *Saprolegnia ferax*), were separated from both Plantae and Fungi (Figure 2), while other primitive organisms, i.e., Protozoa (Amoebozoa, Choanozoa, and Excavata: *Tetrahymena pyriformis*, *Monosiga brevicollis*, *Malawimonas jakobiformis*, *Physarum polycephalum*, *Reclinomonas americana*, *Polysphondylium pallidum*, *Dictyostelium discoideum*, and *Dictyostelium citrinum*), were separated from Animalia (Figure 8 and Supplemental Figure 1). Using a combined sample set of both primitive organisms, only a subset of Amoebozoa (*Polysphondylium pallidum*, *Dictyostelium discoideum*, and *Dictyostelium citrinum*) clustered together, whereas other Amoebozoa (*Physarum polycephalum*), Excavata (*Malawimonas jakobiformis* and *Reclinomonas americana*) and Prosista (*Tetrahymena pyriformis*) were grouped into the same clusters with different kingdoms (Chromalveolata, Fungi and Plantae) (Figure 9 and Supplemental Figure 2). The two primitive organismal groups could not evidently be separated into clusters independent of one another. These results thus suggest that the primitive organisms examined diverged from a common origin. The hypothesis that these primitive lineages may have a common origin is supported by genomic analyses using nucleotide content relationships (Sorimachi, 2010).

5. Conclusions

The ratios of amino acids to the total amino acids or of nucleotides to total nucleotides predicted from complete mitochondrial genomes consisting of huge number of nucleotides can characterize a whole organism. As these values are independent of species and genome size, these indexes are very useful for genome research, as well as single gene research. Indeed, Ward's clustering method using amino acid compositions or nucleotide contents predicted from complete mitochondrial genomes provided consistent phylogenetic trees.

Acknowledgements

KS was employed at Dokkyo Medical University until March 31st, 2011.

References

- Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., & Garey, J. R. (1997). Evidence for a clade of nematodes, arthropods, and other moulting animals. *Nature*, *387*, 489-493. <http://dx.doi.org/10.1038/387489a0>
- Cavalier-Smith, T. (1998). A revised six-kingdom system of life. *Biol. Rev. Camb. Philos. Soc.*, *73*, 203-266. <http://dx.doi.org/10.1017/S0006323198005167>
- Chargaff, E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, *6*, 201-209. <http://dx.doi.org/10.1007/BF02173653>
- Chou, K-C. (1990). Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady-state systems. *Biophys. Chem.*, *35*, 1-24. [http://dx.doi.org/10.1016/0301-4622\(90\)80056-D](http://dx.doi.org/10.1016/0301-4622(90)80056-D)
- Dayhoff, M. O., Park, C. M., & McLaughlin, P. J. (1977). Building a phylogenetic trees: cytochrome C. In M. O. Dayhoff (Ed.), *Atlas of protein sequence and structure* (Vol. 5, pp. 7-16). National Biomedical foundation, Washington D. C..
- DePouplana, L., Turner, R. J., Steer, B. A., & Schimmel, P. (1998). Genetic code origins: tRNAs older than their synthetases? *Proc. Natl. Acad. Sci. USA*, *95*, 11295-11300. <http://dx.doi.org/10.1073/pnas.95.19.11295>
- Doolittle, W. F., & Brown, J. R. (1994). Tempo, mode, the progenote, and the universal root. *Proc. Natl. Acad. Sci. USA*, *91*, 6721-6728. <http://dx.doi.org/10.1073/pnas.91.15.6721>
- Janvier, P. (2010). Micro RNAs revive old views about jawless vertebrate divergence and evolution. *Proc. Natl. Acad. Sci. USA*, *107*, 19137-19138. <http://dx.doi.org/10.1073/pnas.1014583107>
- King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J ... Rokhsar D. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature*, *451*, 783-788. <http://dx.doi.org/10.1038/nature06617>
- Lander, E. S., Linton, M. L., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., ... Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860-921. <http://dx.doi.org/10.1038/35057062>
- Lavrov, D. V., Forget, L., Kelly, M., & Lang, B. F. (2005). Mitochondrial genomes of two demosponges provide insights into an early stage of animal evolution. *Mol. Biol. Evol.*, *22*, 1231-1239. <http://dx.doi.org/10.1093/molbev/msi108>
- Maizels, N., & Weiner, A. M. (1994). Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc. Natl. Acad. Sci. USA*, *91*, 6729-6734. <http://dx.doi.org/10.1073/pnas.91.15.6729>
- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA*, *74*, 560-564. <http://dx.doi.org/10.1073/pnas.74.2.560>
- Mitchell, D., & Bridge, R. (2006). A test of Chargaff's second rule. *Biochem. Biophys. Res. Commun.*, *340*, 90-94. <http://dx.doi.org/10.1016/j.bbrc.2005.11.160>
- Okayasu, T., & Sorimachi, K. (2009). Organisms can essentially be classified according to two codon patterns. *Amino Acids*, *36*, 261-271. <http://dx.doi.org/10.1007/s00726-008-0059-0>
- Poulakakis, N., Pakaki, V., Mylonas, M., & Lymberakis, P. (2008). Molecular phylogenetics of the Pectinidae (Mollusca: Bivalvia) and effect of increased taxon sampling and outgroup selection on tree topology. *Mol. Phylogenet. Evol.*, *47*, 396-402. <http://dx.doi.org/10.1016/j.ympev.2007.10.014>
- Puslednik, L., & Serb, J. M. (2008). Molecular phylogenetics of the ectinidae (Mollusca: Bivalvia) and effect of increased taxon sampling and outgroup selection on tree topology. *Mol. Phylogenet. Evol.*, *48*, 1178-1188.

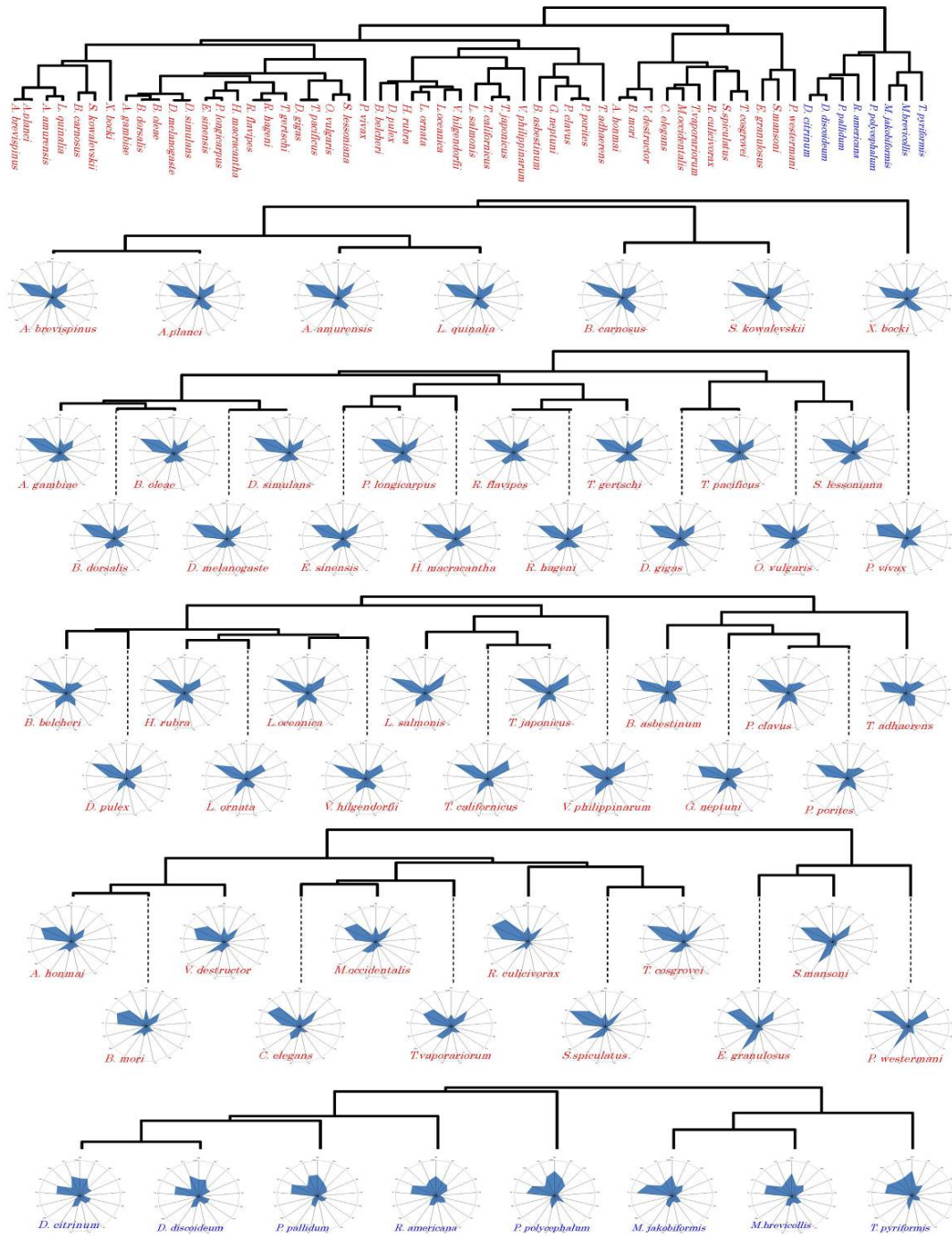
<http://dx.doi.org/10.1016/j.ympcv.2008.05.006>

- Qi, X. Q., Won, J., & Qi, Z. H. (2007). New 3D graphical representation of DNA sequence based on dual nucleotides. *J. Theoret. Biol.*, *249*, 681-690. <http://dx.doi.org/10.1016/j.jtbi.2007.08.025>
- Rundner, R., Karkas, J. D., & Chargaff, E. (1968). Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. *Proc. Natl. Acad. Sci. USA*, *60*, 921-922. <http://dx.doi.org/10.1073/pnas.60.3.921>
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, *4*, 406-425.
- Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, *94*, 441-448. [http://dx.doi.org/10.1016/0022-2836\(75\)90213-2](http://dx.doi.org/10.1016/0022-2836(75)90213-2)
- Sogin, M. L., Elwood, H. J., & Gunderson, J. H. (1986). Evolutionary diversity of eukaryotic small subunit rRNA genes. *Proc. Natl. Acad. Sci. USA*, *83*, 1383-1387. <http://dx.doi.org/10.1073/pnas.83.5.1383>
- Sorimachi, K. (1999). Evolutionary changes reflected by the cellular amino acid composition. *Amino Acids*, *17*, 207-226. <http://dx.doi.org/10.1007/BF01361883>
- Sorimachi, K. (2009). A proposed solution to the historic puzzle of Chargaff's second parity rule. *Open Genom. J.*, *2*, 12-14. <http://dx.doi.org/10.2174/1875693X00902010012>
- Sorimachi, K. (2010). Genomic data provides simple evidence for a single origin of life. *Natural Science*, *2*, 519-525. <http://dx.doi.org/10.4236/ns.2010.25065>
- Sorimachi, K., & Okayasu, T. (2003). Gene assembly consisting of small units with similar amino acid composition in the *Saccharomyces cerevisiae* genome. *Mycoscience*, *44*, 415-417. <http://dx.doi.org/10.1007/S10267-003-0131-2>
- Sorimachi, K., & Okayasu, T. (2004a). An evaluation of evolutionary theories based on genomic structures in *Saccharomyces cerevisiae* and *Encephalitozoon cuniculi*. *Mycoscience*, *45*, 345-350. <http://dx.doi.org/10.1007/S10267-004-0192-X>
- Sorimachi, K., & Okayasu, T. (2004b). Classification of eubacteria based on their complete genome: where does Maycoplasmataceae belong? *Proc. Roy. Soc., B* *271*, 127-130. <http://dx.doi.org/10.1098/rsbl.2003.0141>
- Sorimachi, K., & Okayasu, T. (2008). Universal rules governing genome evolution expressed by linear formulas. *Open Genom. J.*, *1*, 33-43. <http://dx.doi.org/10.2174/1875693X00801010033>
- Sorimachi, K., & Okayasu, T. (2013). Phylogenetic tree construction based on amino acid composition and nucleotide content of complete vertebrate mitochondrial genomes. *IOSR J. Pham.*, *3*, 51-60.
- Sorimachi, K., Itoh, T., Kawarabayasi, Y., Okayasu, T., Akimoto, K., & Niwa, A. (2001). Conservation of the basic pattern of cellular amino acid composition of archaeobacteria during biological evolution and the putative amino acid composition of primitive life forms. *Amino Acids*, *21*, 393-399. <http://dx.doi.org/10.1007/s007260170004>
- Sorimachi, K., Okayasu, T., Ohhira, S., Masawa, N., & Fukasawa, I. (2013). Natural selection in vertebrate evolution under genomic and biosphere biases based on amino acid composition: primitive vertebrate hagfish (*Eptatretus burgeri*). *Natural Science*, *5*, 221-227. <http://dx.doi.org/10.4236/ns.2013.52033>
- Sueoka, N. (1961). Correlation between base composition of deoxyribonucleic acid and amino acid composition in proteins. *Proc. Natl. Acad. Sci. USA*, *47*, 1141-1149. <http://dx.doi.org/10.1073/pnas.47.8.1141>
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. ... Zhu, X. (2001). The sequence of the human genome. *Science*, *291*, 1304-1351. <http://dx.doi.org/10.1126/science.1058040>
- Ward, J. H. (1963). Hierarchic grouping to optimize an objective function. *J. Amer. Statistic Assoc.*, *58*, 236-244. <http://dx.doi.org/10.1080/01621459.1963.10500845>
- Watson, J. D., & Crick, F. H. (1953). Genetical implications of the structure of deoxyribonucleic acid. *Nature*, *171*, 964-967. <http://dx.doi.org/10.1038/171964b0>
- Weisburg, W. G., Brns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.*, *137*, 697-703. <http://dx.doi.org/10.1080/01621459.1963.10500845>
- Whittaker, R. H. (1969). New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. *Science*, *163*, 150-160. <http://dx.doi.org/10.1126/science.163.3863.150>

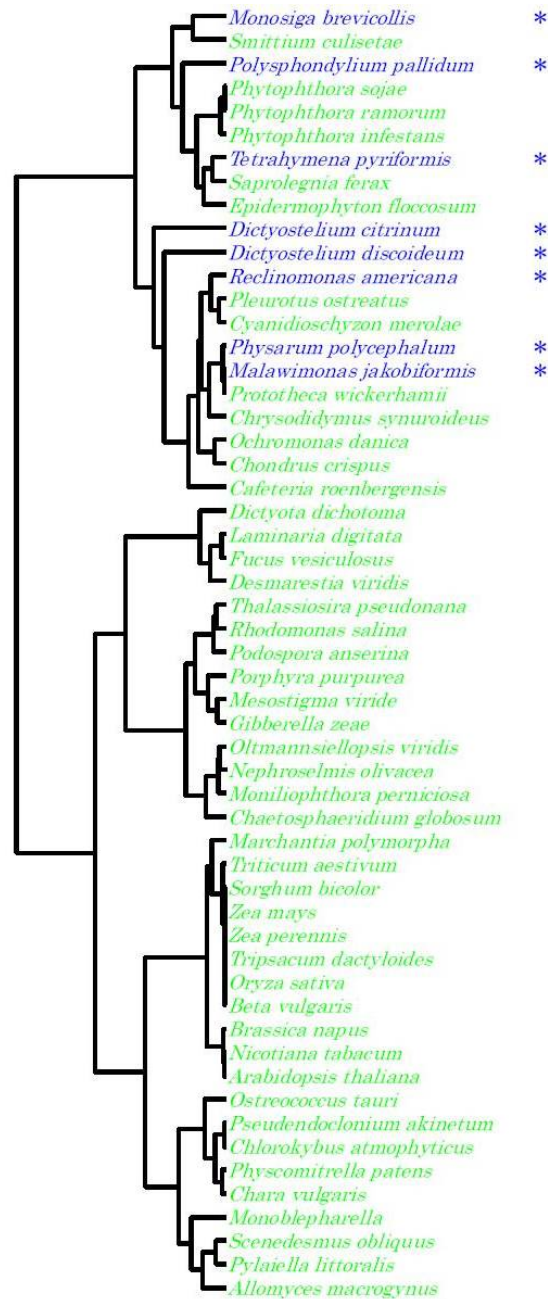
Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms 2. *Proc. Natl. Acad. Sci. USA*, 74, 5088-5090. <http://dx.doi.org/10.1073/pnas.74.11.5088>

Woese, C. R., Kandler, O., & Wheelis, M. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA*, 87, 4576-4579. <http://dx.doi.org/10.1073/pnas.87.12.4576>

Supplemental Figures



Supplemental Figure 1. Phylogenetic tree based on Ward’s clustering using amino acid composition, and radar charts of amino acid composition predicted from complete mitochondrial genomes. Blue characters (the right cluster in the highest phylogenetic tree and the lowest phylogenetic tree) correspond to Amoebozoa, Excavata, and Choanozoa, and red characters (the other clusters and phylogenetic trees) indicate Animalia



Supplemental Figure 2. Phylogenetic tree based on Ward's clustering using amino acid composition. Blue (*) characters correspond to Amoebozoa, Excavata, and Choanozoa, and green (null) characters indicate Non-Animalia

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).