IOSR Journal Of Pharmacy
(e)-ISSN: 2250-3013, (p)-ISSN: 2319-4219
Www.Iosrphr.Org Volume 3, Issue 6 (July 2013), Pp 51-60

Phylogenetic tree construction based on amino acid composition and nucleotide content of complete vertebrate mitochondrial genomes

Kenji Sorimachi[1], Teiji Okayasu[2]
1 Life Science Research Center, Higashi-Kaizawa, Takasaki, Gunma 370-0041, Japan
2 Centor for Medical Informatics, Dokkyo Medical University, Mibu, Tochigi 321-0293, Japan

ABSTRACT: *To evaluate the appropriateness of phylogenetic trees in biological evolution, we expanded a pre-existing baseline data set of randomly selected organisms by incorporating a collection of intentionally chosen organisms. Using two different clustering algorithms—Ward's method and neighbor-joining—we constructed phylogenetic trees based on nucleotide sequences as well as amino acid composition and nucleotide content of complete mitochondrial genomes and 16S rRNA and NADH dehydrogenase subunit 5 genes. We compared classifications derived from cluster analyses with data based on mathematical calculations of complete mitochondrial genomes. Analyses of predicted amino acid composition from complete mitochondrial genomes and of 16S rRNA sequences clearly differentiated terrestrial and aquatic vertebrates. Although no truly representative phylogenetic tree exists, phylogenetic trees provide scientifically appropriate information about biological evolution.*

KEYWORDS : *Amino acid composition, complete mitochondrial genome, evolution, nucleotide content, phylogenetic tree*

I. INTRODUCTION

The concept of evolution by natural selection was established by Charles Darwin and Alfred Wallace 150 years ago. This theory was formulated from observations of specific differences and similarities in the phenotypes of organisms living on geographically isolated islands. Based on this theory, Charles Darwin presumably developed the phylogenetic tree model to represent biological evolution. The theory of biological evolution has been further confirmed by paleontology [1], using phenotypic changes in fossils, and by molecular biology [2], using genotypic changes (nucleotides or amino acids) in living organisms.

Studies of biological evolution have generally focused on nucleotide or amino acid sequences of certain genes. Most phylogenetic tree constructions have been carried out using nucleotide or amino acid sequences [3-10], with amino acid composition or nucleotide content rarely used for this purpose. Sueoka was the first to analyze cellular amino acid composition in bacteria [11]; more recently, our laboratory has independently analyzed the cellular amino acid composition of bacteria, archaea, and eukaryotes [12]. When radar charts are used to express cellular amino acid compositions, their patterns—a "star-shape"—are similar among various organisms, with their differences seeming to reflect biological evolution [12]. Amino acid and nucleotide content predicted from complete genomes have been used to classify bacteria, archaea, and eukaryotes [13] into two categories: "GC-rich" and "AT-rich". By analyzing amino acid content, we recently classified vertebrates into two groups—terrestrial and aquatic—based on natural selection [10]. In the same study, an identical classification was obtained through analysis of 16S rRNA sequences.

Because organisms were randomly chosen in our previous study [10], it was difficult to evaluate the accuracy of the resulting classifications. In that study, vertebrates were completely classified into two groups—terrestrial and aquatic vertebrates—based on cluster analyses using as characters nucleotide content or predicted amino acid composition from complete mitochondrial genomes. When organisms belonging to the same group are analyzed, their resulting positions in certain cluster(s) can be used to evaluate the appropriateness of the data for use in classification. To better elucidate the effect of natural selection in biological evolution, we have added several species to our previously examined sample set [10].

Using multivariate analysis, samples can be classified mathematically into various clusters regardless of sample characteristic relationships. Identical results are not always obtained from the same organisms, however, even when the same traits are analyzed using the same clustering methods. Furthermore, addition or removal of samples yields different phylogenetic patterns. Such results indicate that there is no truly representative phylogenetic tree in biological evolution. In the study reported here, an intentionally designed

organismal group was analyzed and the resulting classifications used to evaluate results obtained with different methods and from different targets. In our evaluation, we weighed the stability and appropriateness of phylogenetic trees against several factors affecting phylogenetic tree reconstruction.

II. 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 [14]. In the present study, we added the following groups and organisms to the previous data set: primates (human: *Homo sapiens*; chimpanzees: *Pan paniscus* and *Pan troglodytes*; gorilla: *Gorilla gorilla*; monkey: *Macaca mulatta*), other mammals (goat: *Capra hircus*; boar: *Sus scrofa*; deer: *Cervus nippon yakushimae*; rhinoceros: *Diceros bicornis*), Oceanian vertebrates (opossum: *Metachirus nudicaudatus*; wallaby: *Macropus robustus*; platypus: *Ornithorhynchus anatinus*), cetaceans (whale: *Balaenoptera musculus*; dolphins: *Sousa chinensis*; porpoises: *Phocoena phocoena*), reptiles (*Crocodylus niloticus*; *Gavialis gangeticus*; *Alligator mississippiensis*), chondrichthyes (*Plesiobatis daviesi*; *Chiloscyllium punctatum*; *Heterodontus francisci*; *Amblyraja radiata*), and other fish (eel: *Anguilla marmorata*; seahorse: *Hippocampus kuda*; lamprey: *Lampetra fluviatilis*).

Nucleotide contents of coding and non-coding regions of mitochondrial genomes were compared with the content of their complete corresponding single-strand DNA [14] and normalized to 1 (G + C + T + A = 1). Predicted amino acid compositions of mitochondrial genome coding regions were estimated. Calculations were performed using Microsoft Excel (version 2003). Classifications based on Ward's clustering method [15] were conducted using multivariate software developed by ESUMI (Tokyo, Japan). Multiple sequence alignment of 16S rRNA, a well-characterized gene region [8,9], was performed using ClustalW. Phylogenetic trees were constructed from the aligned 16S rRNA sequence data according to the neighbor-joining method [16] using DINASIS Axon (Hitachi, Tokyo, Japan).

III. RESULTS

3.1 Complete mitochondrial genome

In our previous study [10], cluster analyses using amino acid composition were used to classify vertebrates into two groups: terrestrial and aquatic vertebrates. In that study, however, sampled vertebrates were chosen randomly, without any selection criteria. In this study, we intentionally incorporated additional vertebrate species belonging to some of the previously analyzed groups to evaluate the resulting phylogenetic trees.

Phylogenetic trees were first constructed by Ward's method [15] using amino acid composition as the analyzed character. The resulting phylogenetic tree (Fig. 1) consists of three major clusters, i.e., aquatic, mammalian, and other vertebrates (including reptiles, amphibians, birds, and chondrichthyes). Among mammalian vertebrates, primates (*Homo sapiens, Pan paniscus, Pan troglodytes, Gorilla gorilla, and Macaca mulatta*) fall into the same cluster; other animals (*Capra hircus, Sus scrofa, Cervus nippon yakushimae,* and *Diceros bicornis*) and Oceanian vertebrates are also found within the mammalian cluster. Unexpectedly, however, turtle (*Chelonia mydas*) is placed within the primate cluster. Within the mammalian cluster, cetaceans (*Balaenoptera musculus, Sousa chinensis,* and *Phocoena phocoena*) form a small cluster that is closely related to hippopotamus. This placement is consistent with results of other studies [17,18], and supports the evolution of cetaceans from terrestrial vertebrates. Primitive vertebrates are thought to have been derived from aquatic animals. The hagfish (*E. burgeri*) has in fact been thought to be a primitive vertebrate [19], and its position in the phylogenetic tree is different from that of other aquatic vertebrates (Fig. 1). Of the three Oceanian vertebrates, platypus (*Ornithorhynchus anatinus*) is incorporated into the small whale cluster, whereas the opossum (*Metachirus nudicaudatus*) and wallaby (*Macropus robustus*) group together with mouse (*Mus musculus*) and rat (*Rattus norvegicus*). All of these animals belong to the large cluster of mammals.

Newly added fish (*Anguilla marmorata* and *Hippocampus kuda*) are placed among aquatic vertebrates. Lamprey (*Lampetra fluviatilis*) and a chondrichthyes clade (*Plesiobatis daviesi*, *Chiloscyllium punctatum*, *Heterodontus francisci*, and *Amblyraja radiata*), however, are found in the major cluster comprising reptiles, amphibians, and birds. Within this major cluster, the reptiles (*Crocodylus niloticus*, *Gavialis gangeticus*, and *Alligator mississippiensis*) also form a small cluster along with *Taeniopygia guttata* (zebra finch).

When nucleotide contents of coding regions of the 13 mitochondrial genes were analyzed using Ward's method, two major clusters were formed (Fig. 2). Rather than clustering together, related species are widely separated in the tree. Similar results were observed upon analysis of nucleotide content estimated from both non-coding regions as well as complete mitochondrial genomes (data not shown).

As shown in Fig. 3, no major clusters were evident when amino acid composition estimated from complete mitochondrial genomes was analyzed using neighbor-joining [16]. Chondrichthyes are more closely associated with other aquatic vertebrates, whereas hagfish (*E. burgeri*) and lamprey (*Lampetra fluviatilis*) are extremely distant from them. Newly added cetaceans, reptiles, chondrichthyes, and Oceanian vertebrates form independent clusters. Although mammals form a single group, many of the remaining reptiles are scattered throughout the tree.

3.2 16S rRNA

In our previous study [10], we investigated natural selection using phylogenetic trees constructed from 16S rRNA sequences [8,9] of randomly chosen vertebrates. For the current study, we intentionally added several additional species groups, as described in Materials and Methods.

In the neighbor-joining tree generated from 16S rRNA nucleotide sequences (Fig. 4), terrestrial and aquatic vertebrates are clearly separated from each other, with the exception of two species, *Xenopus laevis* and *Lyciasalamandra atifi*. Species belonging to a particular group cluster together. These results are basically consistent with those obtained from analyses based on amino acid composition (Fig. 1). Chondrichthyes are placed within aquatic vertebrates, as are hagfish (*E. burgeri*) and lamprey (*Lampetra fluviatilis*). Oceanian animals (*Metachirus nudicaudatus, Macropus robustus*, and *Ornithorhynchus anatinus*) form a small cluster, and cetaceans are closely related to hippopotamus (*Hexaprotodon liberiensis*). Reptiles and birds are clustered together.

When Ward's method was used to construct a tree from nucleotide content of 16S rRNA sequences, terrestrial and aquatic vertebrates were not separated into two groups (Fig. 5). In addition, species that should cluster together were widely separated.

IV. FIGURES

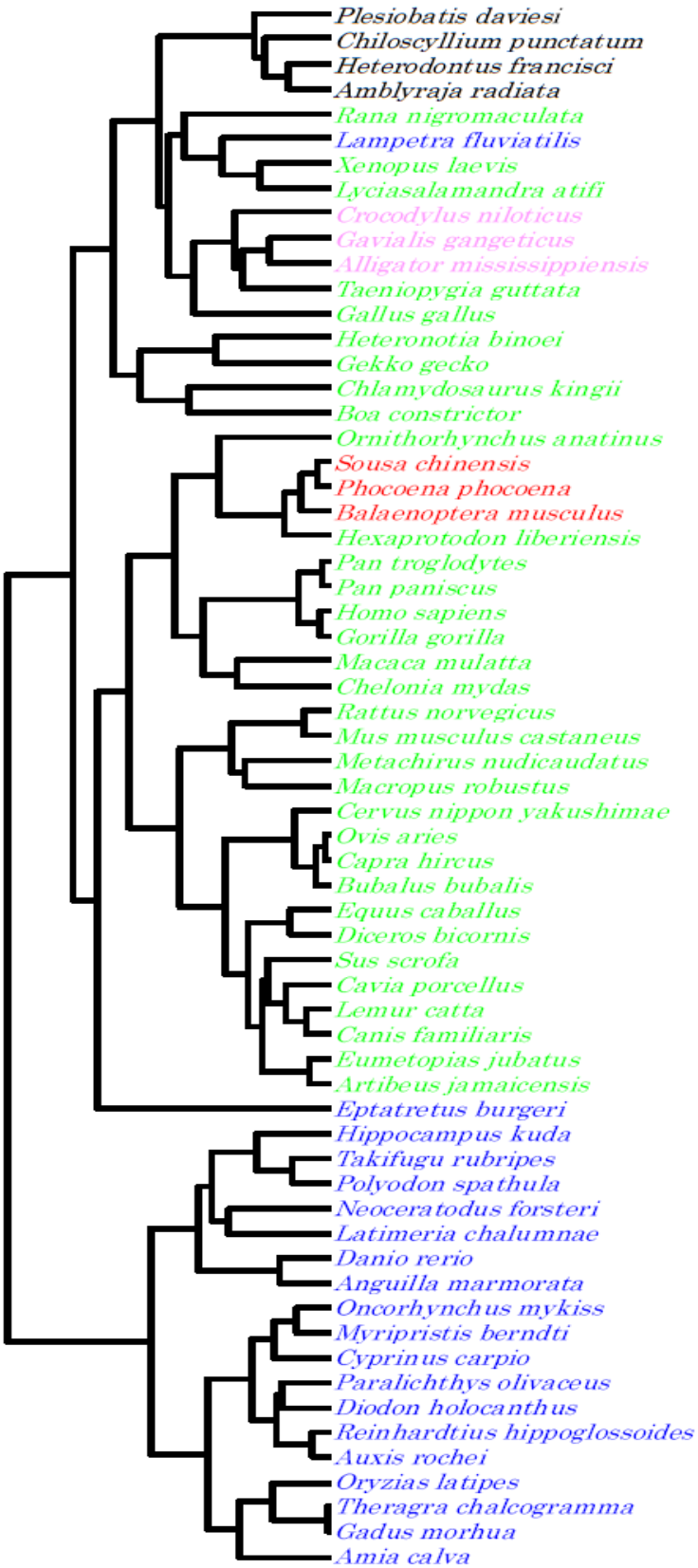

Figure 1. Phylogenetic tree generated using Ward's cluster analysis method [15] from predicted amino acid composition of complete mitochondrial genomes. Green and blue characters represent terrestrial and aquatic vertebrates, respectively. Newly added samples are indicated as follows: chondrichthyes (black), reptiles (pink), and cetaceans (red).

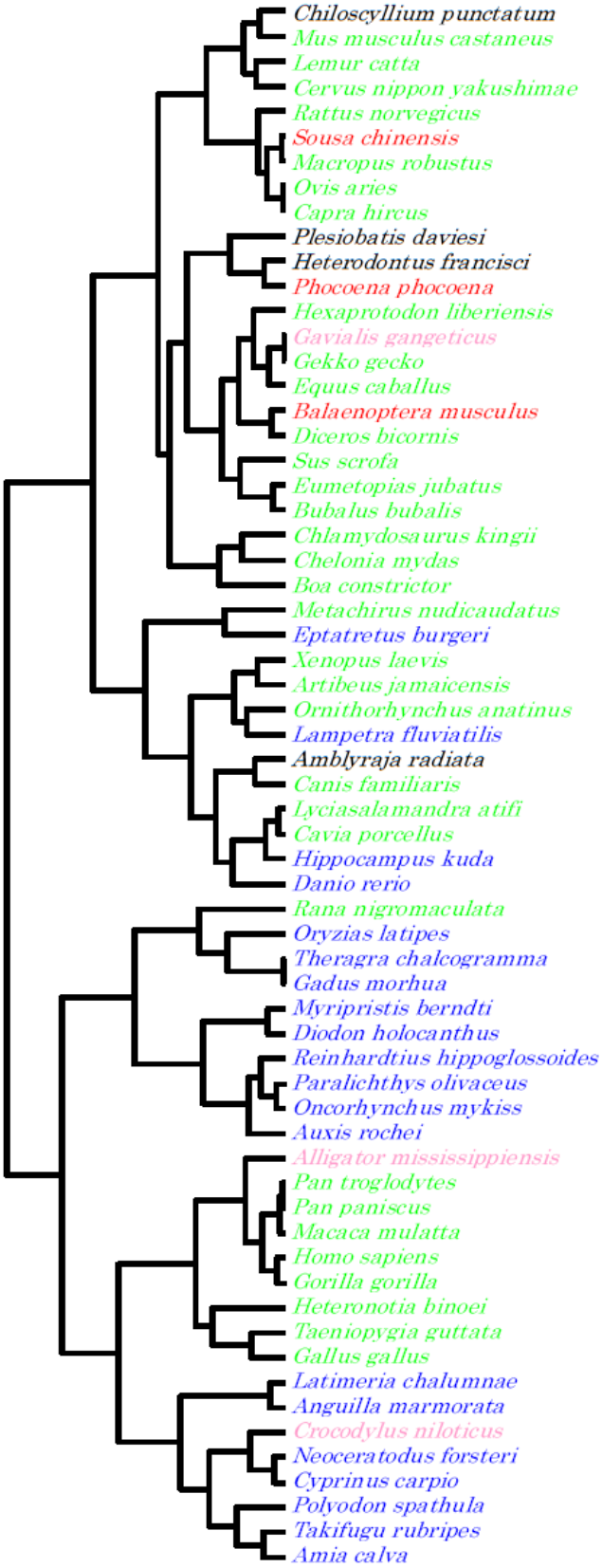

Figure 2. Phylogenetic tree generated using Ward's cluster analysis method [15] from nucleotide content of complete mitochondrial genomes. Green and blue characters represent terrestrial and aquatic vertebrates, respectively. Newly added samples are indicated as follows: chondrichthyes (black), reptiles (pink), and cetaceans (red).

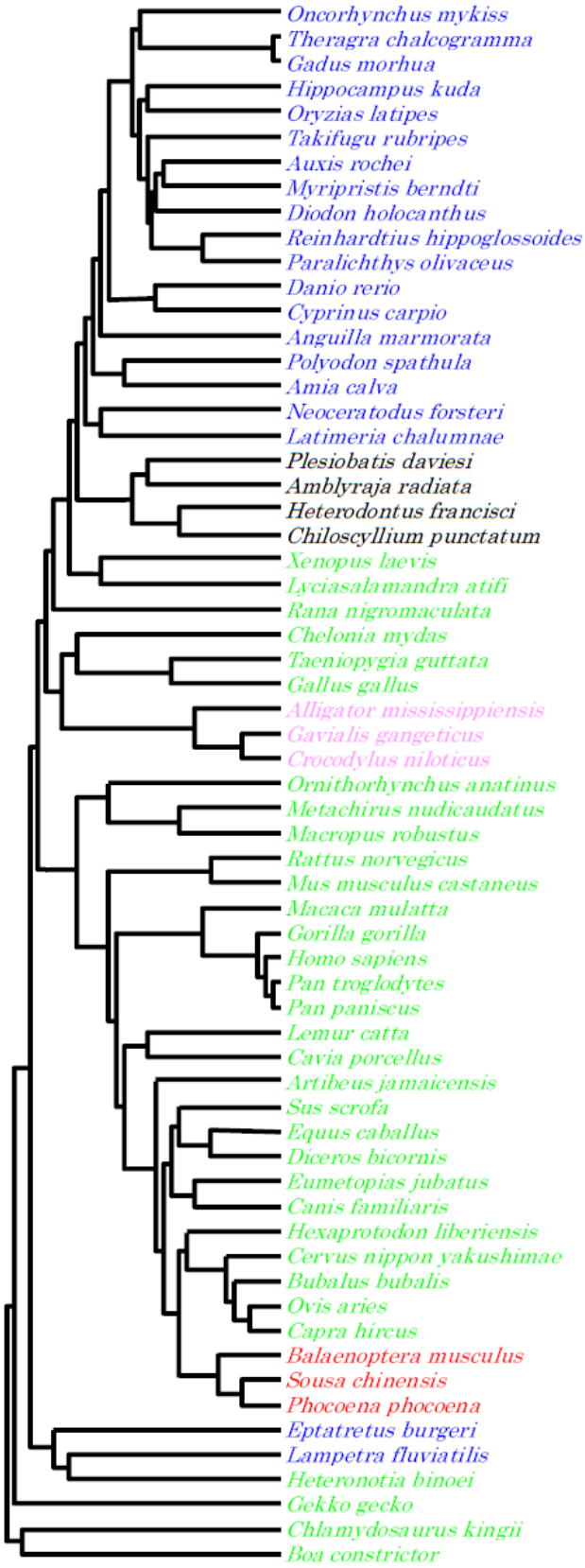

Figure 3. Phylogenetic tree generated using the neighbor-joining method [16] from amino acid composition of complete mitochondrial genomes. Green and blue characters represent terrestrial and aquatic vertebrates, respectively. Newly added samples are indicated as follows: chondrichthyes (black), reptiles (pink), and cetaceans (red).

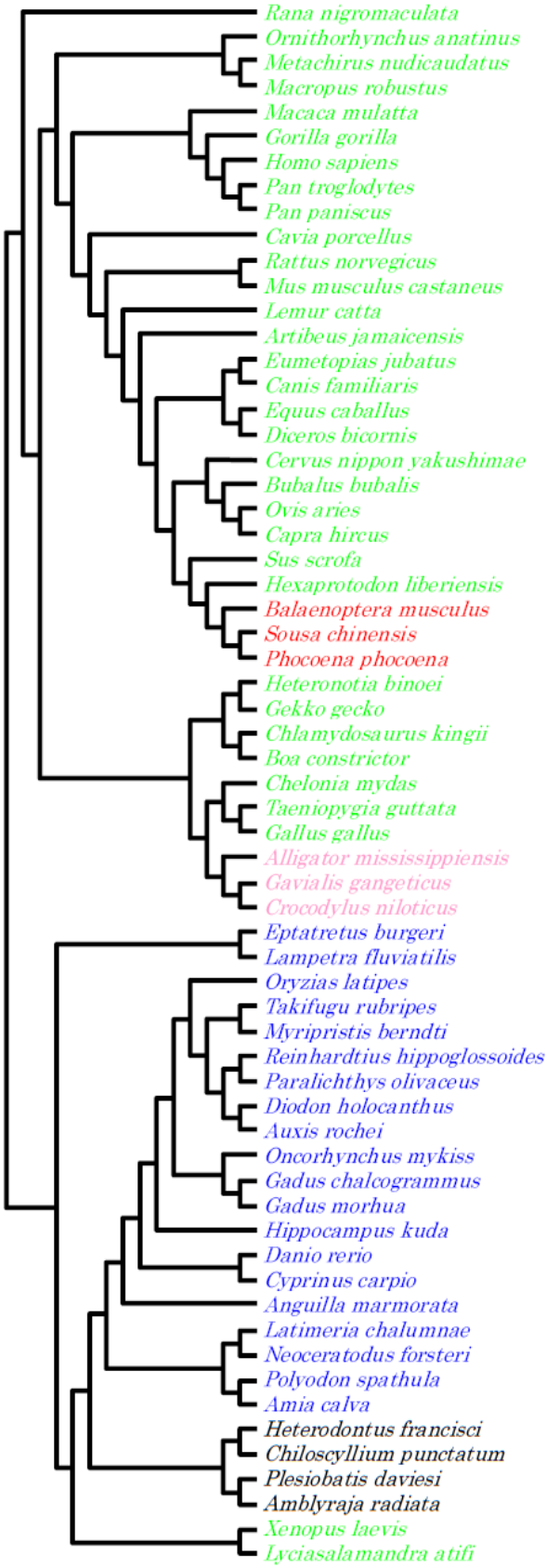

Figure 4. Phylogenetic tree generated using the neighbor-joining method [16] from 16S rRNA nucleotide sequences. Green and blue characters represent terrestrial and aquatic vertebrates, respectively. Newly added samples are indicated as follows: chondrichthyes (black), reptiles (pink), and cetaceans (red).

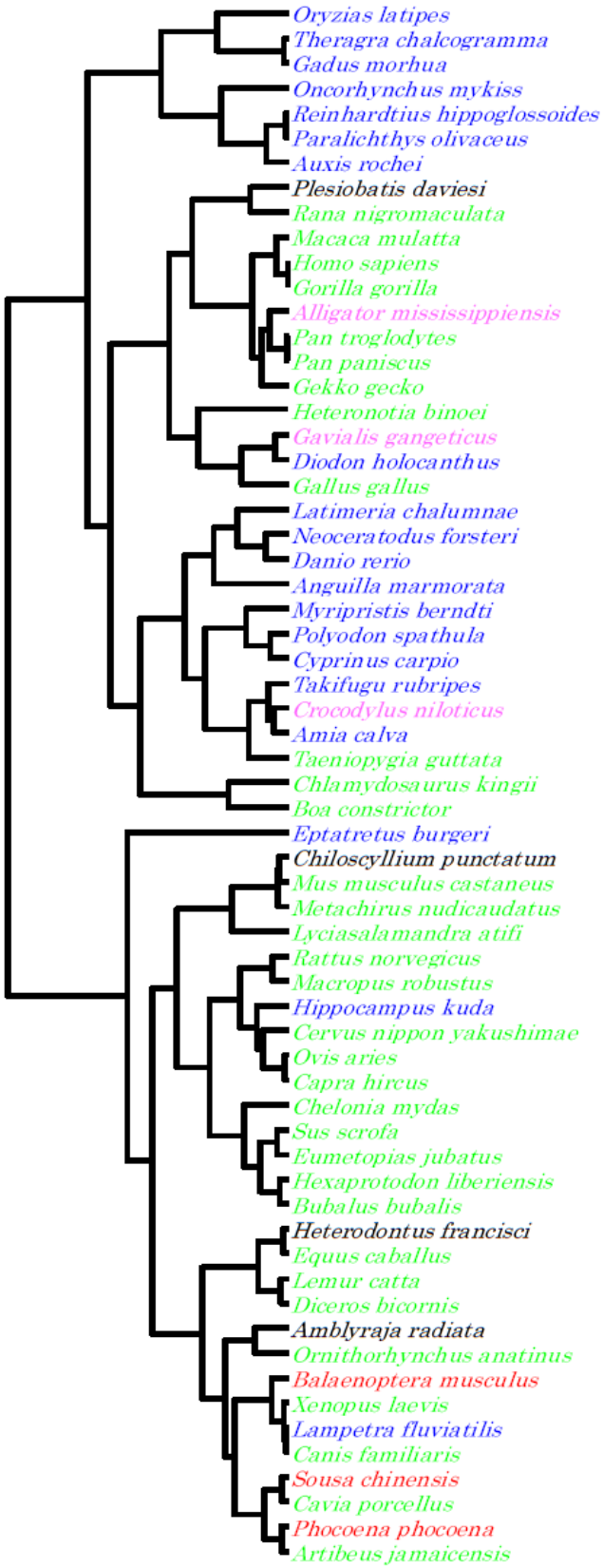

Figure 5. Phylogenetic tree generated by cluster analysis using neighbor-joining [16] from nucleotide content of 16S rRNA sequences. Green and blue characters represent terrestrial and aquatic vertebrates, respectively. Newly added samples are indicated as follows: chondrichthyes (black), reptiles (pink), and cetaceans (red).

V. DISCUSSION

Phylogenetic trees are a simple way to visualize biological evolution, an abstract phenomenon. As a consequence, many phylogenetic trees have been constructed since they were first popularized by Charles Darwin. Phylogenetic trees are not universally representative, however, because they are constructed mathematically based on comparative similarity among samples. Mathematical calculations sometimes lead to coincidental similarity when genetic relationships are absent among independent or even related samples. For example, in our analysis, turtle (*Chelonia mydas*) was unexpectedly placed in the primate cluster. This is a consequence of the algorithms used, not a scientific error. When using phylogenetic trees based on mathematical analyses to explain biological evolution, a scientist must therefore pay careful attention to this point, and evaluate whether the result is based on coincidence or real phenomena. In phylogenetic trees in our study, for example, hagfish (*E. burgeri*) was distant from aquatic vertebrates; in this case, however, it was concluded that this placement was in fact because of hagfish characteristics [19].

Phylogenetic trees have generally been constructed from amino acid or nucleotide sequences of target gene(s) [3-10]. The first attempts to classify bacteria, archaea, and eukaryotes from complete nuclear genome data, however, were based on predicted amino acid composition and nucleotide content [13]. Although all organisms were eventually classified into two groups—"GC-rich" and "AT-rich", it was difficult to evaluate phylogenetic trees derived from amino acid composition and nucleotide content because the initially analyzed organisms were selected randomly [13]. Using Ward's clustering method with amino acid composition or nucleotide content from complete mitochondrial genome data, we were recently able to classify vertebrates into aquatic and terrestrial groups [10]. In the present study, however, we obtained better classifications using amino acid composition as a character than we did using nucleotide content (Figs. 1 and 2). Similar results were actually obtained in the previous study [10]. These differing levels of success may be due to the different number of character states analyzed: 20 amino acids vs. 4 nucleotides. Because comparisons of amino acid or nucleotide gene sequences are based on a large number of characters, phylogenetic tree constructions using these data generally return reasonable results. Amino acid or nucleotide sequences are not applicable to the construction of phylogenetic trees or cluster analyses of whole genomes consisting of huge numbers of various genes; however, ratios of amino acids to total amino acids or nucleotides to total nucleotides calculated from whole genomes can be applied to these purposes, because these ratios are independent of genome size and species differences [as reviewed in 20]. Scientifically reasonable organismal classifications based on complete genomes have in fact been obtained in this manner [10,13]. When carrying out analyses based on amino acid composition, classifications using complete mitochondrial genome data were better than those derived solely from the NADH dehydrogenase subunit 5 gene (unpublished data). The larger genome target yielded better classifications because coincidental similarity was reduced. With respect to the number of samples used in cluster analyses, 39 species were included in the previous study compared with 63 in this investigation. Increased sampling may increase statistical support for meaningless similarity, resulting in unreasonable classifications. Increases in the number of characters analyzed and in the size of target yield good results in cluster analyses or phylogenetic trees, because these increases reduce the probability of coincidental similarity. Conversely, increases in the number of samples yield opposite results because of increased statistical support for coincidental similarity. When cluster analyses based on neighbor-joining were carried out using amino acid composition, terrestrial and aquatic vertebrates were clearly differentiated (Fig. 3). Two distinct major clusters were not observed, however. Consequently, the two different clustering algorithms yielded different phylogenetic trees, even though they both generated significant results.

As described above, many phylogenetic trees have been constructed based on single genes. In our previous study [10], in fact, we used 16S rRNA sequences and neighbor-joining to successfully classify vertebrates into terrestrial and aquatic groups, and obtained similar results in the current study after intentional addition of many other organisms (Fig. 4). When we analyzed nucleotide content of 16S rRNA gene sequences, however, clear differentiation of terrestrial and aquatic vertebrates was not observed in the phylogenetic tree, even when the neighbor-joining method was used (Fig. 5). Similarly, neighbor-joining using nucleotide sequences of NADH dehydrogenase subunit 5 gave better separation of terrestrial and aquatic vertebrates than when nucleotide content was analyzed (unpublished data). These results may be due to the small number of characters available using nucleotide content.

VI. CONCLUSION

Based on the results of this study, phylogenetic tree topology is influenced by the method, characters, and data used, even when the same organisms are examined (Figs. 1–5). Furthermore, removal or addition of samples sometimes causes basic changes to phylogenetic trees because of the inherent nature of the mathematical algorithms employed. Because it is impossible to construct a universally representative phylogenetic tree using our present knowledge, one must remember that phylogenetic trees are changeable based on various conditions.

VII. ACKNOWLEDGEMENTS

We would like to thank Professor Masaru Kojima for financial support from his research grant.

REFERENCES

[1]. A. Cobbett, M. Wilkinson, and M. Wills, Fossils impact as hard as living taxa in parsimony analyses of morphology. *Systems Biol.* 17, 2007, 753-766.

[2]. E. Zuckerkandl, and L.B. Pauling, Molecular disease, evolution, and genetic heterogeneity" in Kasha M and Pullman B (Eds.). *Horizons in Biochemistry*, (Academic Press, New York, 1962) 189-225.

[3]. M.O. Dayhoff, C.M. Park, and P.J. McLaughlin, Building a phylogenetic trees: cytochrome C. In: M.O. Dayhoff (Ed.) *Atlas of protein sequence and structure*, 5, (National Biomedical Foundation, Washington, D.C. 1977), 7–16.

[4]. M.L. Sogin, H.J. Elwood, and J.H. Gunderson, Evolutionary diversity of eukaryotic small subunit rRNA genes. *Proc.Natl. Acad. Sci.USA* 83, 1986, 1383–1387.

[5]. W.F. Doolittle, and J.R. Brown, Tempo, mode, the progenote, and the universal root. *Proc.Natl. Acad. Sci.USA* 91, 1994, 6721–6728.

[6]. N. Maizels, and A.M. Weiner, Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc.Natl. Acad. Sci.USA* 91, 1994, 6729–6734.

[7]. L. DePouplana, R.J. Turner, B.A. Steer, and P. Schimmel, Genetic code origins: tRNAs older than their synthetases? *Proc.Natl. Acad. Sci.USA* 95, 1998, 11295–11300.

[8]. C.R. Woese, and G.E. Fox, Phylogenetic structure of the prokaryotic domain: the primary kingdoms 2. *Proc. Natl. Acad. Sci. USA* 74, 1977, 5088-5090.

[9]. W.G. Weisburg, S.M. Barns, D.A. Pelletier, and D.J. Lane, 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* 173, 1991, 697-703.

[10]. K. Sorimachi, T. Okayasu, S. Ohhira, I. Fukasawa, and N. Masawa, Natural selection in vertebrate evolution under genomic and biosphere biases based on amino acid content: primitive vertebrate hagfish (Eptatretsu burgeri). *Natural Science* 5, 2013, 221-227.

[11]. N. Sueoka, Correlation between base composition of deoxyribonucleic acid and amino acid composition in proteins. *Proc. Natl. Acad. Sci. USA* 47, 1961, 1141-1149.

[12]. K. Sorimachi, Evolutionary changes reflected by the cellular amino acid composition. *Amino Acids* 17, 1999, 207-226.

[13]. T. Okayasu, and K. Sorimachi, Organisms can essentially be classified according to two codon patterns. *Amino Acids* 36, 2009, 261-271.

[14]. K. Sorimachi, and T. Okayasu, Universal rules governing genome evolution expressed by linear formulas. *Open Genom. J.* 1, 2008, 33-43.

[15]. J.H. Ward, Hierarchic grouping to optimize an objective function. *J. Amer. Statistic. Assoc.* 58, 1963, 236-244.

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

[17]. J. Gatesy, C. Hayashi, M.A. Cronin, and P. Arctander, Evidence from milk casein genes that cetaceans are close relatives of hipoppotamid artiodactyls. *Mol. Biol. Evol.* 13, 1996, 954-963.

[18]. B.M. Ursing, and U. Arnason, Analyses of mitochondrial genomes strongly support a hipoppotumus-whale clade. *Proc. Biol. Sci.* 265, 1998, 2251-2255.

[19]. P. Janvier, Micro RNAs revive old views about jawless vertebrate divergence and evolution. *Proc. Natl. Acad. Sci. USA* 107, 2010, 19137-19138.

[20] K. Sorimachi, Evolution from primitive life to *Homo sapiens* based on visible genome structures: the amino acid world. *Natural Science* 1, 2009, 107-119.