Phylogenetic Position of Cetaceans Relative to Artiodactyls: Reanalysis of Mitochondrial and Nuclear Sequences

Masami Hasegawa and Jun Adachi
The Institute of Statistical Mathematics, Tokyo, Japan

By a maximum likelihood analysis of mitochondrial DNA sequences, we examine Graur and Higgins’ hypothesis of the Ruminantia/Cetacea clade with Suiformes as an outgroup. Graur and Higgins analyzed these sequences by the neighbor-joining and parsimony methods, as well as by the maximum likelihood method under the assumption that the substitution rate is the same for all sites. The Ruminantia/Suiformes clade assumed by the traditional taxonomy was rejected strongly by this analysis and the Ruminantia/Cetacea clade was supported. Adoption of a more realistic model distinguishing among rates at different codon positions in the maximum likelihood analysis of the same data, however, grossly reduces the significance level of the Graur–Higgins hypothesis. Thus, although the Ruminantia/Suiformes grouping is indeed least likely from Graur and Higgins’ data set of mitochondrial DNA, this traditional tree cannot be rejected with statistical significance under the new analysis, and more data are needed to settle the issue. In the same way, we examine Irwin and Árnason’s suggestion of the *Hippopotamus*/Cetacea clade by using cytochrome *b* and hemoglobins *α* and *β*, and it turns out that their suggestion is also fragile. This analysis demonstrates the importance of selecting an appropriate model among the alternatives in the maximum likelihood analysis and of using many different genes from many relevant species in order to make reliable phylogenetic inferences.

Introduction

Cetacea, consisting of whales, dolphins, and porpoises, is among the most specialized of mammalian orders, and its evolutionary origin has been a subject of great interest to many researchers. The close relationship of the order Cetacea to the Artiodactyla seems to have been well established by morphological (e.g., Novacek 1992) as well as by molecular studies (e.g., Irwin, Kocher, and Wilson 1991; Milinkovitch, Orti, and Mayer 1993). In traditional taxonomy the artiodactyls are assumed to form a monophyletic clade. Recently, however, Graur and Higgins (1994) challenged this view. From phylogenetic analyses of mitochondrial DNA (mtDNA) and nuclear-encoded proteins, they suggested that cetaceans are more closely related to one subgroup of artiodactyls, the Ruminantia (i.e., cow), than either ruminants or cetaceans are to another artiodactyl subgroup, the Suiformes (i.e., pig).

The strongest support for their hypothesis comes from the mtDNA data. They analyzed mtDNA sequences from cow (Ruminantia), pig (Suiformes), and finback whale (Cetacea) with mouse as an outgroup by neighbor-joining and parsimony methods, as well as by the maximum likelihood (ML) method under the assumption that the substitution rate is the same for all sites. In this analysis, the Ruminantia/Suiformes grouping assumed by traditional taxonomy is strongly rejected and the Ruminantia/Cetacea grouping is favored. As for the ML analysis, the log-likelihoods of the Ruminantia/Suiformes grouping and of the Suiformes/Cetacea grouping are lower by 62.1 ± 24.4 and 39.8 ± 26.8 (±1 SE), respectively, than that of the Ruminantia/Cetacea grouping. Thus, Graur and Higgins seem to have convincingly shown that the traditional tree of artiodactyl monophyly is not true. However, because in their ML analysis they assume rate homogeneity of nucleotide substitution among different codon positions, and because this issue is critical to the study of mammalian evolution, the problem deserves reexamination by the ML method using a more realistic substitution model.

In another analysis, Irwin and Árnason (1994) suggested possible paraphyly of the order Artiodactyla relative to Cetacea in a different way. From parsimony and neighbor-joining analyses of the cytochrome *b* gene, they suggested that the cetaceans might be more closely related to *Hippopotamus* (traditionally considered a suiform) than either of them are to ruminants and other suiforms. Therefore, we will also reexamine their suggestion by the ML analyses of the cytochrome *b* gene, as well as of other protein sequences such as hemoglobins *α* and *β*. Because Graur and Higgins’ analysis did not include hippopotamus, their hypothesis is not necessarily incompatible with that of Irwin and Árnason.

Materials and Methods

Sequence Data

We analyze the mtDNA data used by Graur and Higgins (1994); these were kindly provided by Dr. Higgins. These data encompass the genes for cytochrome *b* (cyt-b; 1,137 bp), nicotinamide adenine dinucleotide dcytochrome oxidase subunit III (ATP6–COIII; 115 bp), the 12s ribosomal RNA (864 bp), and a composite of four tRNAs (Leu, Val, Phe, and Ile; 262 bp), from four species: *Balaenoptera physalus* (finback whale), *Bos taurus* (cow), *Sus scrofa* (pig), and *Mus musculus* (mouse).

We also analyze the mtDNA sequences of 13 species, including 9 additional eutherian species of which complete mtDNA sequences are available: i.e., *Balaenoptera musculus* (blue whale, EMBL/GenBank/DDBJ database accession number: X72204), *Phoca vitulina*...
(harbor seal, X63726), Halichoerus grypus (grey seal, X72004), Rattus norvegicus (rat, X14848), Homo sapiens (human, D38112), Pan troglodytes (chimpanzee, D38113), Pan paniscus (bonobo, D38116), Gorilla gorilla (gorilla, D38114), and Pongo pygmaeus (orangutan, D38115). Throughout our analysis, we assume the relationships of (Rodentia, (Primates, (Carnivora, in-group)) (Cao et al. 1994; Cao, Adachi, and Hasegawa 1994) and of (((H. sapiens, (P. troglodytes, P. paniscus)), G. gorilla), P. pygmaeus) in Primates (Horai et al. 1995). Although Equus caballus (horse, X79547) data are also available, we did not include it in the analysis, because the Cetacea/Artiodactyla clade is firmly established but the branching order among Carnivora, Perissodactyla, and the Cetacea/Artiodactyla clade cannot be resolved (Adachi and Hasegawa 1996c). We use the alignment of proteins by Adachi and Hasegawa (1996c). In analyzing the 12S ribosomal RNA gene, we use the alignment by Cao, Adachi, and Hasegawa (1994).

The relationships among Cetacea, Ruminantia, Suiformes (excluding hippopotamus), and Hippopotamus were examined by using cyt-b data from 29 species, which included 5 cetaceans, 6 ruminants, 5 suiforms, 1 hippopotamus, and 12 outgroup mammals (species names and database accession numbers are given in table 2). We included as many species as we could under the condition that the phylogenetic relationships within the respective groups have been well established. Relationships of (((B. physalus, B. musculus), C. marginata), (S. attenuata, S. longiroshis)) in Cetacea, (((B. taurus, B. bubalis), (D. domo, O. hemiminus)), (T. napu, T. javanicus)) in Ruminantia, (((S. scrofa, S. barbatus), P. africanus), B. babyrussa), T. tajacu) in Suiformes, and (((P. vitulina, H. grypus), E. jubatus), U. maritimus), P. leo) in Carnivora are well established by local phylogenetic analyses (data not shown) and are therefore fixed in the further analyses. Because the phylogenetic position of sperm whales within Cetacea is in dispute (Milinkovitch, Orti, and Meyer 1993; Árnason and Gullberg 1994; Adachi and Hasegawa 1995) and because inclusion of the sperm whales will raise another problem which is not directly relevant to the issue of the present study, the sperm whales are not included in the main analysis.

In addition, the relationships among Cetacea, Ruminantia, Suiformes, and Hippopotamus are examined using hemoglobin α and β sequences from 17 species (species names and database accession numbers are given in table 3). The relationships of (((A. melanoleuca, U. maritimus), P. vitulina), (F. catus, C. crocuta)) in Carnivora and of (((H. sapiens, M. mulatta), (A. geoffroyi, S. fuscicolis))) in Primates are well established from the present data, and are fixed in the analyses. The relationship of (Rodentia, (Primates, (Carnivora, in-group)) is again assumed.

Methods for Phylogenetic Analyses

The NucML program in our MOLPHY package (Adachi and Hasegawa 1996a) is used in analyzing the DNA sequence data. We assume Hasegawa, Kishino, and Yano’s (1985) model for nucleotide substitution, in which a nucleotide i is replaced by another nucleotide j in an infinitesimally short time interval, dt, with a probability of

\[ P_{ij}(dt) = \begin{cases} \alpha \pi_i dt & \text{for transition } (T \leftrightarrow C, A \leftrightarrow G) \\ \beta \pi_j dt & \text{for transversion } (T, C \leftrightarrow A, G) \end{cases} \]

(1)

where \( \pi_i \) is the frequency of nucleotide j, and \( \alpha \) and \( \beta \) are parameters that determine transition and transversion rates, respectively. The \( \alpha/\beta \) ratio is determined so as to maximize the likelihood for respective trees.

On the other hand, the DNAML program (Felsenstein 1993) used by Graur and Higgins assumes

\[ P_{ij}(dt) = \frac{(k/\Pi_j + 1) \mu \pi_j dt}{\mu \pi_j dt} \]

(2)

for transversion

where \( \Pi_j = \pi_T + \pi_C \) if j is T or C, and \( \pi_A \) or \( \pi_G \) if j is A or G (Hasegawa and Kishino 1989). When \( \pi_T + \pi_C = \pi_A + \pi_G \), this model coincides with the model of equation (1), and these two models are generally close with each other.

The ProtML program in MOLPHY is also applied to the amino acid sequences of proteins. Using 22 completely sequenced mitochondrial genomes of vertebrates, Adachi and Hasegawa (1996c) estimated a transition probability matrix of the general reversible Markov (mtREV) model of amino acid substitution for mtDNA-encoded proteins. This matrix is considered to approximate closely the pattern of substitution of mitochondrion-encoded genes as Graur and Higgins did. For the amino acid frequencies; mtREV model). In analyzing hemoglobins α and β, we used the Jones, Taylor, and Thornton (1992) model with the amino acid frequencies of the data used as equilibrium frequencies; mtREV-F model). This model was developed mainly for approximating the evolution of nuclear-encoded proteins.

A bootstrap probabilistics (BP) of a particular tree of being the highest likelihood tree among alternatives is estimated by the RELL (resampling of estimated log-likelihood) method (Kishino, Miyata, and Hasegawa 1990), which is a good approximation to the computationally intensive bootstrap method (Felsenstein 1985) in estimating BP (Hasegawa and Kishino 1994). SEs of log-likelihood differences between trees are estimated by Kishino and Hasegawa’s (1989) formula. When rates at different codon positions are distinguished, SE is estimated for each codon position, and then variances for the three codon positions are summed up.

Results and Discussion

Analyses of mtDNA Sequences Used by Graur and Higgins

In our analysis we first assumed that the evolutionary rate was identical among different positions and therefore among different codon positions for protein-encoding genes as Graur and Higgins did. For the tan-
This shows that the two models represented by equations (1) and (2) are essentially the same. The summation of the analyses of five genes (α/β ratios are optimized for each gene but site homogeneity within a gene is assumed) gives a result nearly identical to that of the tandemly combined sequences; that is, the cow/pig grouping has a lower log-likelihood than the cow/cetacean grouping by 41.1 ± 15.1 (0.02% BP; table 1). Among the individual genes, analysis of cyt-b most strongly rejects the cow/pig grouping with lower log-likelihood by 31.8 + 12.4 (0.03% BP). However, it may be that the assumption of equal rates among different codon positions is unrealistic, and we can test for the superiority of models assuming among-site heterogeneity in rate by comparing the Akaike Information Criterion defined by AIC = −2 log(likelihood) + 2 × (number of parameters). A model that minimizes AIC is considered to be the most appropriate model (Akaike 1973, 1974). Therefore, we an-

### Table 1: NucML and ProtML Analyses of the mtDNA Sequence Data Used by Graur and Higgins (1994)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Length</th>
<th>Δl  (BP)</th>
<th>Cow–Pig</th>
<th>Cow–Cetacean</th>
<th>Pig–Cetacean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NucML analysis: single-rate model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyt-b 1st</td>
<td>379</td>
<td>−11.8 ± 6.6 (0.0024)</td>
<td>(−7.4 ± 8.3 (0.1929)</td>
<td>(−1,033.9) (0.8047)</td>
<td>3.8</td>
</tr>
<tr>
<td>2nd</td>
<td>379</td>
<td>−2.0 ± 7.5 (0.4005)</td>
<td>(−681.2 (0.5944)</td>
<td>6.7 ± 5.2 (0.0051)</td>
<td>3.7</td>
</tr>
<tr>
<td>3rd</td>
<td>379</td>
<td>−4.1 ± 3.1 (0.0140)</td>
<td>(−1,480.5 (0.8532)</td>
<td>−3.8 ± 3.5 (0.1328)</td>
<td>8.7</td>
</tr>
<tr>
<td>Total</td>
<td>1,137</td>
<td>−10.5 ± 10.4 (0.0432)</td>
<td>(−3,203.0 (0.5914)</td>
<td>−3.1 ± 6.3 (0.3654)</td>
<td></td>
</tr>
<tr>
<td>ATP6-COII 1st</td>
<td>37</td>
<td>−7.89 (0.6916)</td>
<td>(−0.8 ± 1.4 (0.1650)</td>
<td>−0.8 ± 1.4 (0.1434)</td>
<td>4.8</td>
</tr>
<tr>
<td>3rd</td>
<td>37</td>
<td>−2.6 ± 2.2 (0.0089)</td>
<td>(−140.3 (0.8846)</td>
<td>−2.6 ± 2.2 (0.1065)</td>
<td>5.6</td>
</tr>
<tr>
<td>Total</td>
<td>74</td>
<td>−1.8 ± 2.2 (0.2246)</td>
<td>(−220.0 (0.7427)</td>
<td>2.6 ± 2.6 (0.0327)</td>
<td></td>
</tr>
<tr>
<td>ND1 1st</td>
<td>318</td>
<td>−0.53 ± 2.7 (0.7871)</td>
<td>−7.3 ± 4.9 (0.0126)</td>
<td>−4.8 ± 6.0 (0.2003)</td>
<td>3.0</td>
</tr>
<tr>
<td>2nd</td>
<td>318</td>
<td>−1.2 ± 2.7 (0.2092)</td>
<td>(−587.6 (0.5462)</td>
<td>−1.2 ± 2.8 (0.2446)</td>
<td>2.6</td>
</tr>
<tr>
<td>3rd</td>
<td>318</td>
<td>−0.7 ± 2.6 (0.22/4)</td>
<td>(−1,278.3 (0.490)</td>
<td>−0.2 ± 2.9 (0.3636)</td>
<td>9.5</td>
</tr>
<tr>
<td>Total</td>
<td>954</td>
<td>−2.82 ± 21.0 (0.6741)</td>
<td>(−3.4 ± 4.9 (0.1115)</td>
<td>−4.2 ± 7.2 (0.2144)</td>
<td></td>
</tr>
<tr>
<td>Total for protein genes</td>
<td>2,165</td>
<td>−6.9 ± 11.3 (0.1709)</td>
<td>(−6,249.4 (0.5509)</td>
<td>−4.5 ± 9.9 (0.2782)</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>3,334</td>
<td>−16.2 ± 14.2 (0.0503)</td>
<td>(−9,427.5 (0.6268)</td>
<td>−6.1 ± 10.9 (0.3229)</td>
<td></td>
</tr>
<tr>
<td>ProtML analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyt-b</td>
<td>379</td>
<td>6.5 ± 4.9* (0.0281)</td>
<td>(−1,662.0 (0.7607)</td>
<td>−4.4 ± 5.9 (0.2112)</td>
<td></td>
</tr>
<tr>
<td>ATP6-COII</td>
<td>37</td>
<td>−1.1 ± 1.7 (0.1629)</td>
<td>(125.0 (0.6773)</td>
<td>−1.1 ± 1.7 (0.1598)</td>
<td></td>
</tr>
<tr>
<td>ND1</td>
<td>318</td>
<td>(−1,481.8) (0.4656)</td>
<td>−2.2 ± 4.2 (0.1524)</td>
<td>−0.6 ± 5.0 (0.3820)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>734</td>
<td>−3.4 ± 5.2 (0.1323)</td>
<td>(−3,270.9 (0.6284)</td>
<td>−3.9 ± 7.9 (0.2393)</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>734</td>
<td>−3.7 ± 5.9 (0.1463)</td>
<td>(−3,301.6 (0.570)</td>
<td>−1.7 ± 6.7 (0.3267)</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** The log-likelihood of the ML tree is given in ( ), and the differences in log-likelihood of alternative trees from that of the ML tree (Δl) are shown with their SE following ±, which were estimated by Kishino and Hasegawa's (1989) formula. The bootstrap probabilities (BP) given in parentheses were estimated by the RELL method (Kishino, Miyata, and Hasegawa 1990) with 10^4 replications. The α/β ratio was optimized for each tree topology, and the ratio for the ML tree is shown in the right most column.
alyzed the three codon positions in the protein-encoding genes separately and then summed up the log-likelihoods for the three different categories of positions. The \( \alpha/\beta \) ratio was optimized for each codon position. Although the cow/cetacean grouping is still the ML tree for cyt-b, it turned out that the difference of log-likelihood is only 10.5 \( \pm \) 10.4 and cannot be rejected at the 5% level of significance (table 1). The analyses separating and combining codon positions of the cyt-b data gave an AIC value of 6,460.0 and 7,278.2, respectively, for the cow/cetacean grouping, and the separate analysis turned out to be much better than the combined analysis that assumed the site homogeneity in approximating the data.

Because of the small number of substitutions, ML estimates cannot be obtained for the second codon position of ATP6–COI; therefore, these positions were not used in a separate analysis. As for the total for protein-encoding genes, the log-likelihood of the cow/pig grouping is lower than that of the cow/cetacean grouping only by 6.9 \( \pm \) 11.3 (17% BP), which is in sharp contrast to 31.8 \( \pm \) 12.4 (0.03% BP) estimated by assuming equal rate among the different codon positions. The total of the NucML analyses of the five sets of mtDNA data including 12S rRNA and tRNAs gives lower log-likelihood for the cow/pig grouping than that of the cow/cetacean grouping by 16.2 \( \pm \) 14.2 (5% BP; table 1). Thus, although Graur and Higgins’ hypothesis is still preferred to the traditional tree of artiodactyl monophyly (trees 6, 7, and 8) has 5% BP, the cow/cetacean grouping. The combined analysis of proteins separately and then summed up the log-likelihoods over the three proteins (table 1). The ProtML analysis is therefore consistent with the NucML analysis, which distinguishes between different codon positions, the support is not as great as they thought.

Next, we applied the ProtML program to the amino acid sequences of the mtDNA-encoded proteins. Although the cow/cetacean grouping is again preferred to the traditional tree with the cow/pig grouping, the log-likelihood of the cow/pig grouping is lower only by 5.4 \( \pm \) 5.2 (13% BP) than that of the ML tree for the sum of log-likelihoods over the three proteins (table 1). The ProtML analysis is therefore consistent with the NucML analysis, which distinguishes between different codon positions.

Analyses Including Additional Species

Graur and Higgins (1994) used only mouse and/or seal as an outgroup, and only cow and pig as representatives of Ruminantia and Suiformes. However, phylogenetic analyses based on a small number of relevant species are often unstable (Adachi and Hasegawa 1995). Philippe and Douzery (1994) and Adachi and Hasegawa (1996b) demonstrated that Graur and Higgins’ hypothesis is not necessarily supported when different species are sampled as representatives of Ruminantia, Suiformes, Cetacea, and outgroup mammals. When other mammalian species, from which complete mtDNA sequences are available (one cetacean and eight outgroup species), are additionally included in the ProtML analysis, the log-likelihood of the cow/pig grouping becomes lower by 8.8 \( \pm \) 6.1 (1% BP) than that of the cow/cetacean grouping. The combined analysis of proteins also gives a similar result; i.e., the log-likelihood of the cow/pig grouping is lower by 8.1 \( \pm \) 6.0 (2% BP) than that of the cow/cetacean grouping. On the other hand, the summed NucML analyses support the pig/cetacean grouping over the cow/cetacean grouping either by the separate or combined analyses of different codon positions. The sum of log-likelihoods over all protein genes assuming different rates among different codon positions gives BP of as high as 30% for the cow/pig grouping (with lower log-likelihood by 2.0 \( \pm \) 9.9 than the pig/cetacea grouping), and using all of the protein and RNA genes gives a BP of 13% for this traditional hypothesis (with lower log-likelihood by 11.3 \( \pm \) 14.0). Therefore, although the Ruminantia/Cetacea grouping is best supported from Graur and Higgins’ data of mtDNA from the four species, more data are obviously needed to determine which one of the phylogenetic hypothesis is the truth.

Phylogenetic Place of Hippopotamus

Table 2 gives the results of the NucML analyses of the cyt-b gene sequences from 29 species for the relationships among Cetacea, Ruminantia, Suiformes, and Hippopotamus. Consistent with Irwin and Árnason’s (1994) findings, the Cetacea/Hippopotamus grouping excluding Ruminantia and Suiformes as outgroups (trees 1, 2, and 3) is supported with 85% BP (total of BPs of trees 1, 2, and 3), whereas the traditional tree of the artiodactyl monophyly (trees 6, 7, and 8) has 5% BP. Tree 2 is the ML tree, and log-likelihood of tree 6 of the artiodactyl monophyly is lower by 15.6 \( \pm \) 12.4. Inclusion of sperm whale Physeter macrocephalus (accession number: X75589), which cluster with the baleen whales consistently with Milinkovitch, Ortiz, and Meyer’s (1993) hypothesis, does not give any significant change in the argument; that is, tree 2 is again the ML tree, and log-likelihood of tree 6 is lower by 17.0 \( \pm \) 15.8. The ProtML analysis of cyt-b (table 3) is consistent with the NucML analysis and gives 80% BP for the Cetacea/Hippopotamus grouping and 2% BP for the artiodactyl monophyly. Tree 2 is again the ML tree, and log-likelihood of tree 6 is lower by 14.2 \( \pm \) 10.7. However, the ProtML analyses of hemoglobins \( \alpha \) and \( \beta \) (fig. 1, table 3) do not confirm the suggestion of the cyt-b data, and the summed log-likelihood of hemoglobins \( \alpha \) and \( \beta \) instead supports artiodactyl monophyly (61% BP) over the Cetacea/Hippopotamus grouping (19% BP). Tree 3 is the ML tree with 31% BP for the total of cyt-b and hemoglobins, and this tree is consistent both with Graur and Higgins’ and with Irwin and Árnason’s hypotheses. Although the total of cyt-b and hemoglobins still marginally supports the Cetacea/Hippopotamus grouping with 60% BP (total of trees 1, 2, and 3 in table 3), the artiodactyl monophyly hypothesis should not be dismissed, because its BP is as high as 28% (total of trees 6, 7, and 8).

From the NucML and ProtML analyses of cyt-b from all 29 species, the total BPs of trees that support Graur and Higgins’ hypothesis (trees 3, 10, 12, 13, and 14) are only 6 and 3%, respectively. Summation of the analyses of cyt-b and hemoglobin supports their hypothesis with only 39% BP. Thus, more data are obvi-
Table 3
ProtML Analyses of Cytochrome b and Hemoglobins α and β

<table>
<thead>
<tr>
<th>TREE</th>
<th>Δι (BP)</th>
<th>Δι (BP)</th>
<th>Δι (BP)</th>
<th>Δι (BP)</th>
<th>TOTAL (BP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-2.1 ± 5.9 (0.2893)</td>
<td>-7.9 ± 8.4 (0.0680)</td>
<td>-7.0 ± 7.4 (0.0157)</td>
<td>-0.2 ± 5.9 (0.0293)</td>
<td>-12.2 ± 12.7 (0.1525)</td>
</tr>
<tr>
<td>2. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-4.2 ± 5.9 (0.3066)</td>
<td>-10.8 ± 9.2 (0.0064)</td>
<td>-11.4 ± 10.2 (0.0009)</td>
<td>-11.4 ± 10.2 (0.0009)</td>
<td>-12.8 ± 11.8 (0.1316)</td>
</tr>
<tr>
<td>3. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-5.2 ± 5.9 (0.0656)</td>
<td>-7.9 ± 10.5 (0.1004)</td>
<td>-14.7 ± 7.7 (0.0000)</td>
<td>-13.7 ± 13.5 (0.0068)</td>
<td>-17.4 ± 13.5 (0.0068)</td>
</tr>
<tr>
<td>4. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-8.8 ± 9.1 (0.1126)</td>
<td>-14.7 ± 7.7 (0.0000)</td>
<td>-9.2 ± 6.3 (0.0000)</td>
<td>-9.2 ± 6.3 (0.0000)</td>
<td>-10.6 ± 13.8 (0.0000)</td>
</tr>
<tr>
<td>5. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-15.1 ± 10.3 (0.0016)</td>
<td>-14.7 ± 7.7 (0.0000)</td>
<td>-14.7 ± 7.7 (0.0000)</td>
<td>-14.7 ± 7.7 (0.0000)</td>
<td>-20.6 ± 13.8 (0.0000)</td>
</tr>
<tr>
<td>6. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-14.2 ± 5.9 (0.0170)</td>
<td>-14.2 ± 5.9 (0.0170)</td>
<td>-14.2 ± 5.9 (0.0170)</td>
<td>-14.2 ± 5.9 (0.0170)</td>
<td>-27.3 ± 13.8 (0.0097)</td>
</tr>
<tr>
<td>7. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-18.3 ± 11.3 (0.0034)</td>
<td>-12.2 ± 3.1 (0.2048)</td>
<td>-12.2 ± 3.1 (0.2048)</td>
<td>-12.2 ± 3.1 (0.2048)</td>
<td>-30.5 ± 13.8 (0.0046)</td>
</tr>
<tr>
<td>8. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-19.0 ± 10.1 (0.0020)</td>
<td>-20.0 ± 2.8 (0.0568)</td>
<td>-20.0 ± 2.8 (0.0568)</td>
<td>-20.0 ± 2.8 (0.0568)</td>
<td>-40.0 ± 13.8 (0.0032)</td>
</tr>
<tr>
<td>9. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-15.1 ± 11.2 (0.0128)</td>
<td>-10.5 ± 7.0 (0.0101)</td>
<td>-10.5 ± 7.0 (0.0101)</td>
<td>-10.5 ± 7.0 (0.0101)</td>
<td>-25.6 ± 13.8 (0.0007)</td>
</tr>
<tr>
<td>10. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-17.3 ± 11.3 (0.0053)</td>
<td>-10.7 ± 8.1 (0.0180)</td>
<td>-10.7 ± 8.1 (0.0180)</td>
<td>-10.7 ± 8.1 (0.0180)</td>
<td>-28.0 ± 13.8 (0.0007)</td>
</tr>
<tr>
<td>11. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-14.6 ± 11.5 (0.0136)</td>
<td>-14.4 ± 7.7 (0.0000)</td>
<td>-9.2 ± 6.3 (0.0000)</td>
<td>-9.2 ± 6.3 (0.0000)</td>
<td>-23.8 ± 14.9 (0.0005)</td>
</tr>
<tr>
<td>12. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-19.0 ± 11.0 (0.0005)</td>
<td>-8.0 ± 8.8 (0.0540)</td>
<td>-2.6 ± 8.8 (0.1313)</td>
<td>-2.6 ± 8.8 (0.1313)</td>
<td>-21.6 ± 15.6 (0.0005)</td>
</tr>
<tr>
<td>13. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-17.8 ± 11.4 (0.0032)</td>
<td>-9.3 ± 10.1 (0.0074)</td>
<td>-2.5 ± 7.9 (0.0781)</td>
<td>-2.5 ± 7.9 (0.0781)</td>
<td>-24.3 ± 17.2 (0.0019)</td>
</tr>
<tr>
<td>14. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-16.6 ± 11.7 (0.0194)</td>
<td>-12.6 ± 8.5 (0.0015)</td>
<td>-7.4 ± 5.7 (0.0023)</td>
<td>-7.4 ± 5.7 (0.0023)</td>
<td>-24.0 ± 15.6 (0.0005)</td>
</tr>
<tr>
<td>15. (<strong>(Cet,Hip),(Rum,Sui))</strong></td>
<td>-12.5 ± 7.6 (0.0007)</td>
<td>-12.4 ± 7.1 (0.0007)</td>
<td>-12.4 ± 7.1 (0.0007)</td>
<td>-12.4 ± 7.1 (0.0007)</td>
<td>-24.9 ± 12.7 (0.0038)</td>
</tr>
</tbody>
</table>

Note.—Cytochrome b data are from the same source as those listed in the table 2. Species names and Swiss-Prot database accession numbers for hemoglobin α/β used in the analysis are as follows: ruminants (Alces alces, European moose, P01970/P02073; Capra hircus, goat, P01970/P02077), Hippopotamus amphibius (P19015/P194016). S. scrofa (P01965/P02067), cetaceans (Balaeopectra acutirostrata, minke whale, P11897/P118984; Tursiops truncatus, Atlantic bottlenosed dolphin, P18978/P18990), carnivores (Ailuropoda melanoleuca, giant panda, P18970/P18983; Ursus maritimus, polar bear, P07423P07422; Phoca vitulina, harbor seal, P09090/P09099; Felis catus, cat, P07405/P07412; Crocuta crocuta, spotted hyena, P18973/P18986), primates (Homo sapiens, P01922/P02025; Macaca mulatta, rhesus macaque, P01925/P02026; Ateles geoffroii, black-handed spider monkey, P01927/P02034; Saguinus fuscicollis, brown-headed tamarin, P01929/P02039), and rodents (Mus musculus, P01942/P02088; Rattus norvegicus, P01946/P02091).
The optimum cx/p ratios for the first and third positions thereafter decreases under the model of equation (1) when \( \alpha > \beta \) (Hasegawa, Kishino, and Yano 1985). These results suggest that combining different codon positions that differ in the evolutionary rate and \( \alpha/\beta \) ratio and then assigning a single value to each of these can produce a seemingly strange estimate of the \( \alpha/\beta \) ratio, which is not intermediate to those of the components.

As an example, let us combine the first and third codon position data of cyt-b gene from the 13 species. The optimum \( \alpha/\beta \) ratios for the first and third positions are 4.7 and 53.0, respectively. The optimum ratio for the combined data turns out to be 4.4, which is lower than both of the ratios for the first and third positions.

Given \( \alpha \) and \( \beta \), the expected numbers of transitions and transversion-type differences between two sequences diverged \( t \) time ago under the model of equation (1), \( S(t) \) and \( V(t) \), are

\[
S(t) = 2n\left(\pi_Y \pi_C + \pi_A \pi_G\right) + \left(\pi_T \pi_T \pi_T \pi_T + \pi_A \pi_C \pi_Y \pi_T \right) \exp(-2\beta t)
- \left(\pi_T \pi_T \pi_T \pi_T \right) \exp(-2\alpha \pi_T \pi_T \pi_T \pi_T) - \left(\pi_A \pi_A \pi_T \pi_T \right) \exp(-2\beta \pi_T \pi_T \pi_T \pi_T) \right)
\]

\[
V(t) = 2n\pi_Y \pi_R \left[1 - \exp(-2\beta t)\right]
\]

where \( \pi_Y = \pi_T + \pi_C, \pi_R = \pi_A + \pi_G, \) and \( n \) is the number of sites (Hasegawa, Kishino, and Yano 1985).

Figure 2a gives \( S/n \) vs. \( V/n \) plots for the first and third codon positions of cyt-b gene from the 13 species. The solid and dotted curves in the figure show the relationships between \( S(t)/n \) and \( V(t)/n \) expected under the assumption that the theoretical curves in figure 2a hold for the component sites, and this curve fits the combined data of distant as well as close pairs. For moderately large \( t \), \( S(t)/n \) becomes smaller than both of those of the two categories of sites for a given \( V(t)/n \). However, by using a model with a single set of parameters \( \alpha \) and \( \beta \), such a fit, retaining a fit to closely related pairs, cannot be attained. Fitting the model to the close pairs compromises the fitting of the model to distant pairs and vice versa. When distant pairs predominate in the combined data set, the optimum fit with a single set of parameters \( \alpha \) and \( \beta \) gives a spuriously low \( \alpha/\beta \) ratio, one that is lower than both of the individual components. This is exactly the case for the combined data as represented in figure 2b, and the theoretical curve with a single set of parameters is shown by a dotted curve, which clearly does not fit the data, particularly for the closely related pairs.

On the other hand, when close pairs predominate, the optimum fit would give an \( \alpha/\beta \) value that falls between the two values of the components. The optimum \( \alpha/\beta \) ratios for the first, second, and third codon positions of protein-encoding genes in Horai et al.'s (1995) complete mtDNA sequences from primates (three humans, chimpanzee, bonobo, gorilla, and orangutan) are, 11.9, 9.3, and 37.8, respectively, whereas that for the combined data is an intermediate value of 12.7 (data not shown). Even in such a case, a single analysis of the combined sequences cannot be justified, because fitting of a homogeneous model to the data is again much worse than conducting an analysis that separates rates (AICs for the combined and separate analyses are 58,215.4 and 53,200.2, respectively).

To the extent that the difference in the rate among different sites is continuous, the gamma distribution is a useful approximation (Yang 1993). However, when the rate and \( \alpha/\beta \) ratio are distinctly different such as between the third codon positions and the other codon positions in protein-encoding genes, we should analyze them separately in the ML analyses.

Highly variable sites have low probability of occurrence and contribute more to the likelihood than conservative sites do. When the difference in rate among sites is not taken into account (Yang, Goldman, and Friday 1994). Furthermore, because the \( \alpha/\beta \) ratio is grossly underestimated when third positions are combined with first/second positions, transitions in the third positions contribute more to the likelihood than they should despite the fact that they are mostly saturated. A merit of the ML method is that because it is based on an explicit model, it can incorporate the complexity of the real underlying process, and the parameters of the model for the underlying process can be estimated from the data under analysis. We should make more use of this advantage.

**Conclusion**

The possible paraphyly of the order Artiodactyla is a highly interesting problem in mammalian phylogenet-
Fig. 2.—Transition ($S/n$) vs. transversion ($V/n$) differences for cyt-b gene sequences from the 13 species. (A) Squares and triangles represent observed differences for the first and third codon positions, respectively. The solid and dotted curves are theoretical ones for the $\alpha/\beta$ ratios of 4.7 and 53.0 with the nucleotide frequencies of the first and third codon positions as equilibrium frequencies. (B) X's represent observed differences for the combined data of the first and third codon positions. The solid curve is expected for the combined data under the assumption that each component sites evolve under the parameter values used in figure 2a (the ratio of $\beta$ for the third positions to that for the first positions is assumed to be 10). The dotted curve is for a single set of parameters $\alpha$ and $\beta$ ($\alpha/\beta = 4.4$).
ies, but none of the proposed hypotheses is convincingly supported by the existing sequence data when analyzed carefully by the ML method. In this work we demonstrate the importance of using an appropriate model of nucleotide or amino acid substitution in the ML analysis.

Even though our rate-heterogeneous model for nucleotide substitution in different codon positions is surely an improvement over models that assume equal rates among codon positions, it is merely an approximation of the real process and cannot be complete. Therefore, we must continue to improve models used in ML analyses, as well as accumulate more data in order to obtain a reliable phylogenetic conclusion on the origin of Cetacea.

Acknowledgments

We thank D. Higgins for providing us with the aligned sequence data and S. Edwards for helpful comments on an earlier version of the manuscript. Thanks are also due to two anonymous reviewers for helpful comments to improve the manuscript. This work was supported by grants from the Ministry of Education, Science, and Culture of Japan.

LITERATURE CITED


