Efficiencies of Different Genes and Different Tree-building Methods in Recovering a Known Vertebrate Phylogeny

Claudia A. M. Russo, Naoko Takezaki, and Masatoshi Nei
Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University

The relative efficiencies of different protein-coding genes of the mitochondrial genome and different tree-building methods in recovering a known vertebrate phylogeny (two whale species, cow, rat, mouse, opossum, chicken, frog, and three bony fish species) was evaluated. The tree-building methods examined were the neighbor joining (NJ), minimum evolution (ME), maximum parsimony (MP), and maximum likelihood (ML), and both nucleotide sequences and deduced amino acid sequences were analyzed. Generally speaking, amino acid sequences were better than nucleotide sequences in obtaining the true tree (topology) or trees close to the true tree. However, when only first and second codon positions data were used, nucleotide sequences produced reasonably good trees. Among the 13 genes examined, Nd5 produced the true tree in all tree-building methods or algorithms for both amino acid and nucleotide sequence data. Genes Cytb and Nd4 also produced the correct tree in most tree-building algorithms when amino acid sequence data were used. By contrast, Co2, Nd1, and Nd41 showed a poor performance. In general, large genes produced better results, and when the entire set of genes was used, all tree-building methods generated the true tree. In each tree-building method, several distance measures or algorithms were used, but all these distance measures or algorithms produced essentially the same results. The ME method, in which many different topologies are examined, was no better than the NJ method, which generates a single final tree. Similarly, an ML method, in which many topologies are examined, was no better than the ML star decomposition algorithm that generates a single final tree. In ML, the best substitution model chosen by using the Akaike information criterion produced no better results than simpler substitution models. These results question the utility of the currently used optimization principles in phylogenetic construction. Relatively simple methods such as the NJ and ML star decomposition algorithms seem to produce as good results as those obtained by more sophisticated methods. The efficiencies of the NJ, ME, MP, and ML methods in obtaining the correct tree were nearly the same when amino acid sequence data were used. The most important factor in constructing reliable phylogenetic trees seems to be the number of amino acids or nucleotides used.

Introduction

It is well known that the phylogenetic trees reconstructed from different genes for the same set of organisms are often different (e.g., Goodman et al. 1982; Hedges 1994). This is true even with mitochondrial DNA (mtDNA), where all genes are inherited together without recombination and there is no confusion of orthologous and paralogous genes (e.g., Cao, Adachi, and Hasegawa 1994; Cao et al. 1994; Simon et al. 1994; Honeycutt et al. 1995). The differences between phylogenetic trees reconstructed may be caused by sampling error of nucleotides or codons, different patterns of nucleotide or amino acid substitutions, etc., but in most cases it is difficult to know which of the reconstructed trees is the correct one because the true tree is unknown. It is possible that some genes are more suitable for reconstructing a phylogenetic tree than others, but it is usually difficult to know which gene is the best.

This problem can be solved if the true phylogeny of the organisms is known. Hillis, Huelsenback, and Cunningham (1994) studied the accuracy of a reconstructed tree by producing an artificially generated phylogeny in phages, inducing mutation by chemical mutagens. While this experiment is interesting, it is desirable to examine the accuracy of the trees reconstructed by using existing organisms. Actually, there are organisms of which the phylogeny is firmly established by fossil records and morphological characters. For example, no one would dispute the phylogenetic relationships of humans, chimpanzees, macaques, marsupials, birds, and bony fishes. If we use these groups of organisms, it is possible to determine how reliable a particular gene is for obtaining the correct tree.

One of the purposes of this paper is to study this problem by using 13 protein-coding genes in the mitochondrial genome for a group of vertebrate species, of which the phylogeny is known. Because the number of codons and the extent of sequence divergence differ considerably among the genes, it is possible to examine their effects on the phylogenetic tree reconstructed. In this paper we are also interested in studying the effi-

---

1Present address: Departamento de Genética, Universidade Federal do Rio de Janeiro, Brazil.

Key words: mtDNA genes, neighbor joining, minimum evolution, maximum parsimony, maximum likelihood, vertebrate species, molecular phylogeny.

Address for correspondence and reprints: Masatoshi Nei, Institute of Molecular Evolutionary Genetics and Department of Biology, 328 Mueller Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802. Email: NXM2@PSU.EDU.

© 1996 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038
The known tree used in this paper. The phylogenetic relationships of the 11 vertebrate species are based on morphological characters and fossil records. The branch lengths are least-squares estimates of the neighbor-joining tree with Poisson-correction distance.

**Materials and Methods**

**Organisms and Sequence Data**

There are now many complete mtDNA sequences from diverse groups of organisms. For our study, we have chosen sequences from 11 vertebrate organisms for which the evolutionary relationships are established and the complete sequence is available. The organisms used (and the source of the sequences with the GenBank accession numbers) are two whale species (Balaenoptera physalus—X61145, Arason, Gullberg, and Widegren 1991; Balaenoptera musculus—X72204, Arason and Gullberg 1993), cow (Bos taurus—V00654 and J01394, Anderson et al. 1982), mouse (Mus musculus—V0071, Bibb et al. 1981), rat (Rattus norvegicus—X14848, Gadaleta et al. 1989), opossum (Didelphis virginiana—Z2957, Janke et al. 1994), chicken (Gallus gallus—X52392, Desjardins and Morais 1990), African clawed frog (Xenopus laevis—X02890, M10217, X01600, and X01601, Roe et al. 1985), rainbow trout (Oncorhynchus mykiss—L29771, R. Zardoya, J. M. Bautista, and A. Garrido-Pertierra, unpublished data), loach (Clostostoma lacustre—M91245, Tseng et al. 1992), and carp (Cyprinus carpio—X61010, Chang, Huang, and Lo 1994).

In the analysis of DNA sequences, we used the nucleotide sequences of each gene determined by the alignment of amino acid sequences. Therefore, the number of nucleotides for each gene is three times the number of codons. In this study the nucleotide sequences for rRNAs, tRNAs, control region, and intergenic regions were not used.

**Phylogenetic Analysis**

To reconstruct phylogenetic trees from the 13 genes and the entire set of genes, we used four commonly used

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Some Statistical Properties of 13 Mitochondrial Genes for Amino Acid Sequence Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>Atp6</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>Number of codons</td>
<td>219</td>
</tr>
<tr>
<td>Minimum p distance (%)</td>
<td>5</td>
</tr>
<tr>
<td>Maximum p distance (%)</td>
<td>50</td>
</tr>
<tr>
<td>Gamma parameter</td>
<td>2.1</td>
</tr>
<tr>
<td>Parsimony sites*</td>
<td>119</td>
</tr>
<tr>
<td>RCI (%)†</td>
<td>71</td>
</tr>
</tbody>
</table>

* Phylogenetically informative sites for parsimony analysis.
† RCI stands for the average rescaled consistency index for the reconstructed unweighted parsimony tree(s).
‡ There are small proportions of overlapping codons between Atp6 and Atp8 and between Nd4 and Nd41, but they were treated as though they were independent codons.
Amino Acid Sequences

The NJ and ME methods are so-called distance methods and require pairwise distances. For amino acid sequence data, we used three different distance measures: proportion of amino acid differences (p distance), Poisson-correction distance, and gamma distance (see MEGA; Kumar, Tamura, and Nei 1993). The gamma distance takes into account the heterogeneity of evolutionary rates among different amino acid sites, and the extent of heterogeneity is measured by the gamma parameter a. The a value was estimated by using the computer program GAMMA (S. Kumar, personal communication), which is based on parsimony analysis (Kocher and Wilson 1991). GAMMA requires an input tree, and we used the known tree in all cases. NJ trees were constructed by using MEGA, whereas ME trees were constructed by the program METREE (Rzhetsky and Nei 1994). ME trees were constructed only for p distance and Poisson-correction distance, because METREE does not compute gamma distance. In the search for ME trees, we examined only the NJ tree or trees whose topological distance (d_T; see below) from the NJ tree (not the true tree) was 2 or 4, since the ME and NJ trees are known to be very similar to each other (Rzhetsky and Nei 1992). In the present case the number of trees with d_T = 2 is 16, but the number of trees with d_T = 4 varies with the topology of the NJ tree (Rzhetsky and Nei 1992). In the case of the true tree in figure 1 the number is 160.

Maximum-parsimony trees were constructed by using the default option of the branch and bound search of the software PAUP (Swofford 1993). Both weighted and unweighted MP trees were produced. Unweighted parsimony generated a single most parsimonious tree for all genes except for Atp6, Atp8, and Col, where 3, 3, and 2 equally parsimonious trees were produced, respectively. When two or more parsimonious trees were obtained, we constructed a strict consensus tree. Weighted parsimony was performed by using the rescaled consistency index (Farris 1989) as a weighting factor for each parsimony site. This index varies from 0 (high degree of homoplasy) to 1 (no homoplasy). (Homoplasy is equivalent to parallel and backward mutations.) In this study we applied this weighting procedure only for one cycle. Weighted parsimony always produced a single most parsimonious tree.

To produce ML trees, we used Adachi and Hasegawa’s (1994) program ProtML. Because the ML method for protein data requires a large amount of computer time, we first used the star decomposition (SD) algorithm, which generates one final tree, as in the case of the NJ method. This tree (SD tree) may not be the real ML tree, so we also used the specific-tree algorithm, which examines any set of specified trees. We examined all trees whose topological distance from the SD tree was equal to 2 or 4 and chose the tree showing the highest likelihood as the ML tree. This algorithm is similar to that of finding ME trees (Rzhetsky and Nei 1992).

We used four different ProtML substitution models for constructing trees, i.e., the Poisson, Dayhoff, JTT, and JTT-f models. The Poisson model uses a Poisson model of amino acid substitution, whereas the Dayhoff and JTT algorithms use the empirical amino acid substitution models based on the data compiled by Dayhoff, Schwartz, and Orcutt (1978) and Jones, Taylor, and Thornton (1992), respectively. The JTT-f model uses Jones, Taylor, and Thornton’s substitution model under the assumption that the relative frequencies of the 20 different amino acids in a sequence are identical with the average observed frequencies for all sequences and remain the same for the entire evolutionary process. Actually, we also used the Poisson and Dayhoff models with the observed amino acid frequencies, but this modification hardly affected the topologies of the trees obtained. Therefore, we shall not consider these models in this paper.

Nucleotide Sequences

To compare the utility of DNA sequences for phylogenetic construction with that of amino acid sequences, we constructed phylogenetic trees for all genes by using either all three codon positions of the sequences or first and second codon positions only. In both sets of data we again used the NJ, ME, MP, and ML methods. The distance measures used for NJ were the p, Jukes–Cantor, Kimura-2-parameter, and gamma distances (see Nei 1991). The values of gamma parameter a were also computed by using the program GAMMA. In the case of the ME method we used only the p, Jukes–Cantor, and Kimura-2-parameter distances. The NJ and ME trees were constructed by MEGA and METREE, respectively. In the case of MP trees, we again used PAUP for both unweighted and weighted parsimony. Weighting was done by using the rescaled consistency index. The ML trees were constructed by using the NucML program by Adachi and Hasegawa (1994). We again used the SD algorithm and the specific-tree algorithm for trees with d_T = 2 and 4. The substitution models used were the Poisson, proportional, and HKY (Hasegawa, Kishino, and Yano 1985) models (see Adachi and Hasegawa 1994).
Table 2
Pairwise p Distances (p × 100) for Atp6 Amino Acid Sequences from 11 Vertebrate Organisms

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Whale-p</td>
<td>5</td>
<td>23</td>
<td>29</td>
<td>26</td>
<td>30</td>
<td>45</td>
<td>48</td>
<td>49</td>
<td>47</td>
<td>48</td>
</tr>
<tr>
<td>2 Whale-m</td>
<td>22</td>
<td>27</td>
<td>25</td>
<td>30</td>
<td>46</td>
<td>47</td>
<td>47</td>
<td>46</td>
<td>47</td>
<td>49</td>
</tr>
<tr>
<td>3 Cow</td>
<td>23</td>
<td>21</td>
<td>24</td>
<td>28</td>
<td>44</td>
<td>45</td>
<td>47</td>
<td>47</td>
<td>49</td>
<td>48</td>
</tr>
<tr>
<td>4 Rat</td>
<td>24</td>
<td>42</td>
<td>43</td>
<td>44</td>
<td>49</td>
<td>49</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>46</td>
</tr>
<tr>
<td>5 Mouse</td>
<td>27</td>
<td>47</td>
<td>49</td>
<td>49</td>
<td>48</td>
<td>46</td>
<td>47</td>
<td>47</td>
<td>49</td>
<td>48</td>
</tr>
<tr>
<td>6 Opossum</td>
<td>23</td>
<td>21</td>
<td>24</td>
<td>31</td>
<td>27</td>
<td>27</td>
<td>13</td>
<td>27</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>7 Chicken</td>
<td>5</td>
<td>28</td>
<td>48</td>
<td>50</td>
<td>49</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>8 Xenopus</td>
<td>31</td>
<td>31</td>
<td>27</td>
<td>27</td>
<td>20</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>9 Trout</td>
<td>47</td>
<td>46</td>
<td>47</td>
<td>20</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>10 Loach</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>11 Carp</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Note:—All insertions/deletions were removed from the entire data set, and the distances were computed by using the remaining 219 amino acid sites. p, Balaenoptera physalus; m, Balaenoptera musculus. Note that mammalian species show similar distances from chicken, Xenopus, and fishes. This makes it difficult to obtain the correct tree.

Accuracy of the Topology Obtained
The accuracy of the tree topology obtained was measured by the topological distance of the tree obtained from the true tree. This distance (dT) is based on the work by Robinson and Foulds (1981) and Penny and Hendy (1985) and is given by the following formula (Rzhetsky and Nei 1992).

\[ d_T = 2[\min(q_r,q_t) - r] + |q_r - q_t|, \]

where \( q_r \) and \( q_t \) are the total numbers of ways of sequence partitions (equal to the number of interior branches) for the tree reconstructed from a given data set and for the true tree, respectively, and \( r \) is the number of partitions (interior branches) that are identical for the two trees. For bifurcating trees, \( d_T \) is equal to twice the number of sequence partitions for which the two trees compared are different (incorrect interior branches of the reconstructed tree). Thus, \( d_T = 0 \) means that the tree obtained is the same as the true tree, and as \( d_T \) increases, the deviation from the true tree increases. In the case of MP trees we also computed the average rescaled consistency index (RCI) for all parsimony sites to examine the reliability of the tree obtained.

To compare the efficiencies of different tree-building methods, we used the sum of \( d_T \) for all genes and the number of genes generating the correct topology (\( n_C \)) among the 13 genes examined.

Results
Statistical Properties of Different Genes
Table 1 shows various statistical properties of the 13 mitochondrial genes used in this study. The number of codons or amino acids encoded varies from 52 (Atp8) to 582 (Nd5), and thus we can examine the effect of gene size on the accuracy of the tree reconstructed. The minimum and maximum p distances indicate that some genes (e.g., Co1 and Co3) are highly conserved whereas others (e.g., Atp8 and Nd6) are quite divergent. Of course, the distance value varies with sequence pair, and table 2 shows the magnitude of variation in p among different pairs of sequences for the gene Atp6 as an example. The gamma parameter also varies from gene to gene, but the extent of heterogeneity of substitution rate (\( 1/\alpha \)) is not so great as in the case of the control region of mtDNA sequences, where \( \alpha = 0.15 \) has been obtained (Kocher and Wilson 1991). The number of amino acid sites that are informative for parsimony analysis (parsimony sites) varies from 44 to 278, suggesting that some genes are much more useful for parsimony analysis than others. The average rescaled consistency index, which is supposed to be negatively correlated to the extent of parallel or backward mutations, varies from gene to gene, but this index seems to have no correlation with the maximum p distance.

Reconstruction of Phylogenetic Trees
Amino Acid Sequences
Table 3 shows the topological distances (\( d_T \)) of reconstructed trees from the true tree for different genes and different tree-building methods. Gene Nd5 produced the correct tree in all tree-building methods and algorithms used, and genes Cytb and Nd4 also produced the correct tree in all the methods except in a few ML algorithms. Co3 produced the correct tree in all parsimony and ML algorithms and in the NJ and ME methods with p distance. By contrast, genes Co2, Nd1, and Nd4l generated incorrect trees in all tree-building methods. The topologies produced by different tree-building algorithms were often the same or very similar, though there were several exceptions (Atp8, Co2, Nd3, Nd4l, and Nd6). These results indicate that when a tree-building
method produces a given topology other methods or algorithms also tend to produce the same topology, whether it is correct or not. This finding is similar to that observed in Saitou and Imanishi's (1989) computer simulation with respect to DNA sequences.

If the differences in \(d_T\) among different tree-building methods are primarily caused by sampling error of nucleotides, one would expect that large genes tend to produce the correct tree more often than small genes. There is certainly such a tendency, but some large genes (e.g., Col) sometimes produced incorrect trees. Therefore, there must be other factors that affect the accuracy of the tree reconstructed. One such factor could be the pattern of amino acid substitution. If this pattern varies from gene to gene, different genes may produce different topologies. We therefore examined this pattern by estimating the transition matrix of amino acids for each gene. In this study we used Yang's (1995) computer program PAML to estimate all the elements of the general reversible amino acid substitution model. However, all genes produced very similar transition matrices. Therefore, it was difficult to explain the differences in \(d_T\) by different amino acid substitution patterns. In the case of Col, however, the small extent of sequence divergence (table 1) is probably responsible for its relatively poor performance.

Among different tree-building methods, NJ tends to show small \(d_T\)'s, whereas ML tends to show large \(d_T\)'s. The other methods show intermediate \(d_T\) values. Table 3 also includes the number of genes that produced the correct topology \((n_c)\). According to this criterion, the ML star decomposition with the Poisson model is best in topology construction and is followed by NJ with Poisson-correction distance and gamma distance. However, because the number of genes examined is small, the differences in \(n_c\) are not statistically significant. Clearly, we need more genes to compare different tree-building methods, including nuclear genes.

In both NJ and ME methods, we used two or three different distance measures, but they had little effect on the topology of the tree reconstructed. One might expect that a distance measure that is proportional to the number of amino acid substitutions performs better in phylogenetic reconstruction than other distances and thus either Poisson-correction distance or gamma distance is better than \(p\) distance. However, for the reconstruction of phylogenetic trees the variance as well as the linearity of a distance measure with the number of substitutions plays an important role (Nei, Tajima, and Tateno 1983; Goldstein and Pollock 1994; Tajima and Takezaki 1994). Therefore, a simple measure such as \(p\) distance or Poisson-correction distance often shows a better perfor-

Table 3
Topological Distances (\(d_T\)) of Reconstructed Trees from the True Tree

<table>
<thead>
<tr>
<th>Genes</th>
<th>Atp6</th>
<th>Atp8</th>
<th>Col</th>
<th>Co2</th>
<th>Co3</th>
<th>Cytb</th>
<th>Nd1</th>
<th>Nd2</th>
<th>Nd3</th>
<th>Nd4</th>
<th>Nd41</th>
<th>Nd5</th>
<th>Nd6</th>
<th>All</th>
<th>Sum</th>
<th>(n_c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NJ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(p)</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gamma</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ME</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(p)</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unweighted</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>21</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unweighted-b</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>na</td>
<td>71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>20</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted-b</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>na</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML star decom.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dayhoff</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JTT</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JTT-f</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML specific-trees</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td>0</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dayhoff</td>
<td>?</td>
<td>4</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JTT</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JTT-f</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>25</td>
<td>31</td>
<td>21</td>
<td>99</td>
<td>6</td>
<td>4</td>
<td>45</td>
<td>26</td>
<td>48</td>
<td>8</td>
<td>66</td>
<td>0</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—Unweighted-b stands for unweighted bootstrap consensus trees and weighted-b weighted bootstrap consensus trees. na: Because of a large amount of computer time required, the tree was not constructed. \(n_c\): Numbers of genes that produced the true tree. The \(d_T\) values in boldface letters indicate that the ML value for the tree obtained is lower than that of the true tree.
parsimony methods. The JTT-f raises a question about the efficiency of weighted parsimony with the RCI. It is also interesting to note that in the cases of genes Col and Nd3 with Poisson-correction distance the NJ method identified the correct topology, whereas the ME did not. This indicates that the criterion of minimum evolution is to choosing an incorrect topology rather than to improving or keeping the NJ tree. This may have happened by chance (Rzhetsky and Nei 1992), but it seems that an optimization process such as the ME criterion does not always work well with real data (see Discussion).

We have used four different algorithms to produce MP trees, but in most of the genes used the four algorithms produced essentially the same topology. This raises a question about the efficiency of weighted parsimony with the RCI. It is also interesting to note that there is virtually no correlation between RCI and dT in parsimony methods. The dT value is generally small when the number of parsimony sites is large, but the correlation between the two quantities is not very high.

The four different models for the ML star decomposition algorithm produced the same tree for six genes but different trees for the others. Kishino and Hasegawa (1989, 1990) suggested that when different models produce different ML trees, the most likely tree should be chosen by using the Akaike information criterion (AIC), which is defined as $-2 \times$ (estimated log likelihood) $+ 2 \times$ (number of free parameters) (Akaike 1974). The tree with the smallest AIC value is supposed to be the best. The number of free parameters in AIC is the same for the Poisson, Dayhoff, and JTT models, but the number for the JTT-f is greater than that for the others by 19, because the amino acid frequencies are estimated from data.

Table 4 shows the AIC values for a few genes for the four different substitution models. It is clear that AIC is smallest for the JTT-f model and largest for the Poisson model. Actually, this was the case for all genes. However, table 3 shows that JTT-f is not efficient in obtaining the correct tree and the Poisson model is slightly better. Therefore, AIC does not seem to be a good criterion to choose the correct tree at least in the present case. However, this result might have been obtained because we used the SD algorithm, which does not examine a large number of trees. Indeed, when we computed the likelihood for the true tree, the value was higher than that of the SD tree in eight cases ($d_T$ values in boldface letters in table 3). Therefore, the inefficiency of the ML criterion observed in table 3 may be due to the inefficiency of the SD algorithm. However, when we examined all trees that are different from the SD tree by $d_T = 2$ or 4, the $d_T$ value of the ML tree did not always decrease. Rather it increased in some genes. The $n_c$ value also remained nearly the same. Interestingly, many of the ML trees obtained were not the true tree, but all of them had a likelihood value higher than that for the true tree. These results suggest that the ML criterion may not always be very efficient in obtaining the correct topology for protein sequences.

The reliability of a phylogenetic tree reconstructed is expected to increase as the number of codons used increases. To examine whether this expectation is true or not, we constructed a tree based on the entire set of genes using all statistical methods except the bootstrap consensus trees by parsimony. The bootstrap consensus trees required too much computer time to be completed in a reasonable time. As is seen from table 3, the trees obtained showed the correct topology regardless of the method used. When the bootstrap test (Felsenstein 1985) was applicable, all interior branches had a bootstrap value of 100%, and the interior branch test (Rzhetsky and Nei 1992) gave a confidence probability of 99.9% or higher for all interior branches. This clearly supports the idea that the number of codons (or characters) used is very important for constructing a reliable phylogenetic tree.

**Nucleotide Sequences**

As mentioned earlier, synonymous substitutions between mammalian and nonmammalian sequences are almost certainly saturated or near saturation. Therefore, nucleotide sequences are expected to be subject to a large extent of noise. However, a number of authors (e.g., Cummings, Otto, and Wakeley 1995) have used all three codon positions of nucleotide sequences for constructing trees for distantly related organisms. Some authors have used only first and second codon positions, because these positions are less affected by synonymous substitutions than third codon positions (Cao, Adachi, and Hasegawa 1994; Cao et al. 1994). It is therefore interesting to examine the efficiencies of these approaches in obtaining the correct topology.

Table 5 shows the $d_T$ values of the trees obtained for each gene for the cases of all three codon positions data and first two codon positions data (the latter in parentheses). When all codon positions are used, many
genes show a higher $d_T$ value than that for amino acid sequences. This suggests that nucleotide sequences are less appropriate for constructing trees for distantly related organisms. However, $Nd5$ again shows $d_T = 0$ for all tree-building methods and algorithms. $Nd4$ and $Nd6$ also show $d_T = 0$ or 2 for all algorithms. Actually, $Nd6$ performs slightly better for nucleotide sequences than for amino acid sequences, but this is exceptional. The $d_T$ values for first and second codon positions data are generally slightly smaller than those for all three positions data as expected, but they are still generally higher than the values for amino acid sequence data. Some genes such as $Atp6$ and $Nd3$ tend to show a smaller $d_T$ value for this case than for the case of amino acid sequences.

Table 5 shows that when all three codon positions are used the MP method shows high $d_T$ and low $n_C$ values, whereas the ML star decomposition algorithm tends to have low $d_T$ and relatively high $n_C$ values. In the present case the difference between $n_C = 1$ and $n_C = 7$ is significant at the 5% level if we use Fisher's exact test. Therefore, the ML star decomposition algorithm is significantly better than the MP method in obtaining the correct tree. It seems that MP is affected by substitution noise more strongly than the other methods. However, when only first and second codon positions are used, all tree-building methods show similar $d_T$ and $n_C$ values, and thus all methods seem to be equally efficient.

As in the case of amino acid sequences, the efficiencies of NJ and ME in obtaining the correct tree does not improve by using distance measures that are supposed to reflect the number of nucleotide substitutions better than $p$ distance. Similarly, in ML a more sophisticated model does not necessarily give the correct tree more often than the others. In the present set of data the HKY model showed the lowest AIC value among the three models for all genes (data not shown). The HKY model has some tendency to generate a better tree particularly in the case of the SD algorithm. However, the merit of this model is small and could be due to chance.

The ME method examined about 180 trees in addition to the NJ tree. However, this algorithm has hardly affected the final tree chosen. In other words, the NJ and ME trees were the same in most cases. In the case of ML, the specific-tree algorithm examined a large number of trees including the true tree in most cases, but the performance of the algorithm in terms of $d_T$ and $n_C$ was also no better than that of the SD algorithm. Actually, the former algorithm chose an incorrect tree slightly more often than the latter, and the incorrect tree chosen always had a higher likelihood value than the true tree.

As in the case of amino acid sequence data, when all 13 genes were used, the correct tree was obtained in
all cases examined. However, the bootstrap value was about 70% for one branch of the NJ trees when all three codon-positions data were used, though the CP value was 99.9% or higher. (No bootstrap tests were done for the ME, MP, and ML methods.)

Discussion
Comparison of Genes

As mentioned earlier, one of the purposes of this paper is to find mitochondrial genes that are suitable for constructing phylogenetic trees for distantly related organisms. Our results have shown that amino acid sequences are generally more informative than nucleotide sequences for constructing reliable trees and that Nd5 is the most appropriate gene for this purpose. The next best genes are Nd4 and Cytb. Actually, Nd5 seems to be most appropriate even for nucleotide sequences. By contrast, genes Co2, Ndi, and Nd4l seem to be least appropriate. This conclusion is consistent with the results obtained by Cao et al. (1994), though their purpose was not to find appropriate genes. These authors conducted ML analysis of protein sequences to study the phylogenetic relationships of cow, fin whale, harbor seal, human, mouse, and rat. Genes Cytb, Nd4, and Nd5 produced the same tree as the genome tree obtained by using all genes, whereas Co2, Ndi, and Nd4l did not. This result did not change when blue whale and gray seal were added (Cao, Adachi, and Hasegawa 1994). However, note that in this case the true tree is not well established (Novacek 1992), and thus this conclusion is dependent on the assumption that the genome tree is identical with the true tree.

In the past Co1, Co2, and Cytb have been used extensively for phylogenetic analyses (e.g., Adkins and Honeycutt 1994; Cantatore et al. 1994), but Nd4 and Nd5 have not. Both Nd4 and Nd5 are large genes, and their protein sequence divergences seem to be appropriate for constructing trees for distantly related organisms of higher vertebrates. We therefore recommend that these genes should be used more often.

Comparison of Models and Algorithms

We have seen that in the cases of the NJ and ME methods Poisson-correction and gamma distances, which are supposed to reflect the number of amino acid substitutions better than p distance, are no better than the latter distance in producing the correct topology. This seems to be strange. Using computer simulation, however, Saitou and Nei (1987), Sourdip and Krimbas (1987), Saitou and Imanishi (1989), and Nei (1991) have shown that in the case of DNA sequence data p distance produces the correct tree slightly more often than the Jukes–Cantor distance, unless the rate of nucleotide substitution varies considerably with evolutionary lineage and sequence divergence is high. This is partly because p distance has a smaller variance relative to the mean than corrected distances. Similar results were obtained by Schöniger and von Haeseler (1993) and Tajima and Takezaki (1994).

Figure 1 shows that the sequences from the three fish species evolved considerably slower than the others. These three species formed one cluster outside the other sequences, and all mammalian sequences evolved nearly at the same rate. Furthermore, if we impose an artificial root in the middle of the branch between mammals and nonmammals, the tree looks like a linear tree with a molecular clock. This may have helped p distance to generate trees as good as those obtained by Poisson-correction or gamma distance.

A somewhat similar statement can be made about the weighted and unweighted parsimony methods used. For the present data set, both methods showed essentially the same efficiency of obtaining the correct topology, though weighted parsimony is supposed to be better than unweighted parsimony. These results are somewhat different from those of Hillis, Huelsenbeck, and Cunningham (1994), Tateno, Takezaki, and Nei (1994), Huelsenbeck (1995), and Nei, Takezaki, and Sitnikova (1995), who have shown that in the case of nucleotide sequence data weighted parsimony is often more efficient than unweighted parsimony in obtaining correct topologies. One possible reason for this difference is that in our data set the extent of sequence divergence was not very large (tables 1 and 2) and in this case weighting does not give much advantage.

In the case of likelihood methods we already indicated that the improvement of the substitution model in terms of the AIC value does not necessarily increase the chance of obtaining a better topology. Similar results have been obtained by Cao, Adachi, and Hasegawa (1994), Cao et al. (1994), and Yang, Goldman, and Friday (1994). These results are somewhat unexpected, because the differences in AIC among different substitution models are substantial. Using computer simulation, however, Gaut and Lewis (1995) showed that unless the extent of sequence divergence is very large the probability of obtaining the correct topology for a less complicated substitution model is essentially the same as that for a more complicated one even if sequence data are generated by the latter model. Yang (1996) also showed that a simple model (Jukes–Cantor model) may give a higher probability of obtaining the correct topology than a sophisticated model (substitution rate varying among different sites) (depending on the model tree used) even when sequence data are generated by the latter model. These studies were done with nucleotide sequence data, but the same conclusion is expected to hold with amino acid sequence data as well. However,
conducting a computer simulation with the Huelsenbeck (1995) type trees of two very long branches and two very short branches, Hasegawa and Fujiwara (1993) produced examples in which the Dayhoff model recovers the correct tree more often than the Poisson and proportional models when amino acid sequences are generated by the Dayhoff model. Therefore, if the rate of substitution varies extensively with evolutionary lineage, the above conclusion may not apply.

So far we emphasized the effectiveness of simple distance measures or simple substitution models in obtaining correct topologies. However, for estimating branch lengths or evolutionary times between sequences, unbiased distances (unbiased estimators of substitutions) or correct substitution models generally give better results (Tateno, Takezaki, and Nei 1994). That is, a distance measure or substitution model that is appropriate for estimating branch lengths is not necessarily good for topology inference. In this case one can use different distance measures or substitution models for topology construction and branch length estimation (Nei, Tajima, and Tateno 1983; Nei and Takezaki 1994). For a data set similar to the present one, however, it is still advisable to use corrected distances or appropriate substitution models, because they would give more reliable branch length estimates though they may not improve topology construction.

Optimization Principle in Phylogenetic Analysis

In the present set of data, ME was no better than NJ in obtaining the correct topology. This is somewhat counterintuitive, because Rzhetsky and Nei (1993) have shown that the true topology has the smallest expected value of the sum of branch lengths when unbiased distances are used and this has been the theoretical basis of the ME method. In practice, however, distance estimates are subject to sampling error, and for this reason the ME method is not necessarily better than the NJ method.

The criterion of maximum likelihood also does not always choose the correct topology. In fact, the ML specific-tree algorithm used here showed a tendency to be inferior to the ML star decomposition algorithm. This is partly due to sampling error, because the difference in likelihood value between the ML tree and the true tree was not statistically significant in most cases when the topologies of the two trees were different and the differences were tested by Kishino and Hasegawa's (1989) method. However, there seem to be some other factors that make the topology estimation by likelihood complicated. At the present time, the theoretical basis of likelihood method of topology estimation remains unclear (Nei 1987, p. 324-325; Yang 1994, 1996; Yang, Goldman, and Friday 1995), and there seems to be no mathematical proof that the correct topology gives the highest expected likelihood value among all possible topologies, as will be discussed below.

Comparison of Different Tree-building Methods

It is a difficult and tricky problem to compare the relative efficiencies of different tree-building methods in obtaining correct topologies, because the theoretical basis of each method is not well established. Recently Edwards (1995) stated that in the study of evolution, which is an "after-trial" evaluation, the ML method is known to be the best. However, it should be noted that the topology estimation in phylogenetic analysis is not the same as the estimation of parameters in the classical theory of the ML method, because the maximization of likelihood is conducted separately for different topologies (different probability spaces) (Nei 1987, p. 323-325; Yang 1994, 1996). Indeed, it is not difficult to construct examples in which the ML method is inferior to the MP and NJ methods (Yang 1996; N. Takezaki and M. Nei, unpublished). Figure 2 shows one such example. Of course, a series of computer simulations have shown that ML is generally slightly better than other methods in many different situations (Saitou and Im-
Amino Acid vs. Nucleotide Sequences

Examining the efficiencies of protein-coding genes in estimating the topology of the genome tree for mtDNA, Cummings, Otto, and Wakeley (1995) concluded that the topology of the tree produced by a gene is a poor indicator of the topology of the genome tree. This conclusion is different from ours, and there are two reasons for this. First, it is still unclear whether their genome tree represents the true phylogenetic tree of the organisms used (whale, cow, seal, human, mouse, rat, frog, carp, and loach) (Novacek 1992), though recent molecular data tend to support it. If the genome tree is not the correct one, their conclusion is not very meaningful. Second, they used all three codon positions of nucleotide sequences and did not attempt to enhance the efficiency of any of the tree-building methods used (NJ, MP, and ML). Probably for these reasons, individual genes did not produce the genome tree (supposed to be the true tree) as often as in our data analysis. If they had used first and second codon positions only instead of the three positions, they might have obtained better results.

Data Set Used

However, it should be mentioned that the results obtained in this paper are dependent on the data set used. As mentioned earlier, the unrooted version of the tree in figure 1 roughly satisfies the condition of a linear tree with a molecular clock, and probably for this reason we obtained the true tree when we used large genes or the entire set of genes. Actually, if several groups of fast-evolving sequences and slowly evolving sequences are mixed in the true tree, construction of the correct topology usually becomes more difficult, though theoretically all the methods used here are supposed to take care of varying substitution rates. For example, if lamprey and sea urchin sequences, which have evolved much faster than fish and Xenopus sequences, are included in our data set, we obtain an incorrect tree for all tree-building methods even if we use the entire set of genes. In this case the cluster of lamprey and sea urchin is attached to the interior branch connecting mammals and nonmammals and this branching pattern is statistically significant (data not shown). At the present time we do not know the real reason for this, but it suggests that we must be very careful about the tree obtained when many fast-evolving and slowly evolving sequences are mixed. We are now investigating this problem from a theoretical point of view.

Acknowledgments

We thank Sudhir Kumar for his help in the alignment of sequences and for the computation of gamma parameters and Ziheng Yang, Andrey Rzhetsky, and Walter Fitch for their comments on an earlier version of this paper. We are also grateful to Tanya Sitnikova for developing a computer program to identify all the trees that are different from a given tree by $d_T = 2$ and 4. This paper is part of the Ph.D. thesis of Claudia A. M. Russo at the Departamento de Genética da Universidade Federal do Rio de Janeiro. Claudia A. M. Russo was sponsored by CNPq (National Research Council) from the Education Ministry, Brazil. This study was supported by research grants from the National Institutes of Health and National Science Foundation to M.N.

Literature Cited


KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis, version 1.01. Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, Penn.


TAKASHI GOJOBORI, reviewing editor

Accepted December 7, 1995