Tie Trees Generated by Distance Methods of Phylogenetic Reconstruction

Naoko Takezaki
National Institute of Genetics, Shizuoka, Japan; and Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University

In examining genetic data in recent publications, Backeljau et al. showed cases in which two or more different trees (tie trees) were constructed from a single data set for the neighbor-joining (NJ) method and the unweighted pair group method with arithmetic mean (UPGMA). However, it is still unclear how often and under what conditions tie trees are generated. Therefore, I examined these problems by computer simulation. Examination of cases in which tie trees occur shows that tie trees can appear when no substitutions occur along some interior branch(es) on a tree. However, even when some substitutions occur along interior branches, tie trees can appear by chance if parallel or backward substitutions occur at some sites. The simulation results showed that tie trees occur relatively frequently for sequences with low divergence levels or with small numbers of sites. For such data, UPGMA sometimes produced tie trees quite frequently, whereas tie trees for the NJ method were generally rare. In the simulation, bootstrap values for clusters (tie clusters) that differed among tie trees were mostly low (<60%). With a small probability, relatively high bootstrap values (at most 70%–80%) appeared for tie clusters. The bias of the bootstrap values caused by an input order of sequences can be avoided if one of the different paths in the cycles of making an NJ or UPGMA tree is chosen at random in each bootstrap replication.

Introduction

Backeljau et al. (1996) showed that there are cases in which distance-tree-making methods produce two or more tree topologies (tie trees) from the same set of gene frequency or DNA sequence data. These authors suggested that if tie trees are not thoroughly examined, erroneous conclusions may be made with respect to evolutionary relationships of populations or genes. They also indicated that if tie trees are not treated properly in testing the reliability of branching patterns by the bootstrap test (Felsenstein 1985), confidence levels of certain clusters may be inflated, depending on the input order of genes or populations for a computer program. Although Backeljau et al. (1996) examined published data sets and showed examples in which tie trees appeared, it is still not clear why and how often tie trees occur. Also, ways to treat tie trees that appear during the bootstrap test should be investigated.

To examine these problems, a computer simulation was conducted in this study. The probabilities that tie trees appear were computed for the neighbor-joining (NJ) method (Saitou and Nei 1987) and the unweighted pair group method with arithmetic mean (UPGMA; Sneath and Sokal 1973), which are the methods used by Backeljau et al. (1996). Furthermore, I computed the bootstrap confidence levels for tie trees that appeared in the simulation to examine how high the bootstrap value can be for clusters that differ among tie trees. Also, the difference in the bootstrap values was examined with and without giving treatment for tie trees during a bootstrap test. Two methods for treating tie trees during the bootstrap test were examined.

Conditions Under Which Tie Trees Occur

In order to examine why tie trees are generated, let us consider the case of four operational taxonomic units (OTUs) for the NJ method. There are three different topologies for unrooted trees of four OTUs (fig. 1). The NJ tree is a tree which has the smallest sum of branch lengths among the three topologies. Therefore, the conditions under which tree A is an NJ tree are given as $S_A < S_B$ and $S_A < S_C$, where $S_A$, $S_B$, and $S_C$ are the sums of branch lengths for trees A, B, and C, respectively. These conditions can be shown to be equivalent to

\[ D_A < D_B \]

and

\[ D_A < D_C, \]

(1)

respectively, where $D_A = d_{12} + d_{34}$, $D_B = d_{13} + d_{24}$, $D_C = d_{14} + d_{23}$, and $d_{ij}$ is the distance between OTUs $i$ and $j$ (Saitou and Nei 1987). Tree A becomes a tie with tree B or C if $D_A = D_B$ and $D_A < D_C$, or $D_A < D_B$ and $D_A = D_C$, respectively.

When there are no substitutions along an interior branch, the data give no resolution for the branching pattern of the sequences. In this case, even if the model tree (a tree with branch lengths of expected numbers of substitutions per site) is bifurcating, the realized tree (a tree with branch lengths of actual numbers of substitutions; Kumar 1996) is multifurcating. For example, if no substitutions occur along the interior branch on tree A (fig. 1) and there are no multiple hits at any sites, $D_A = D_B = D_C$. Therefore, all of the three topologies (fig. 1A–C) become tie trees. However, even when substitutions occur along the interior branch, tie trees can occur by chance. Figure 2A shows such an example, in which tie trees occur for four nucleotide sequences. There is one substitution along the interior branch and two substitutions along the exterior branches. If the actual numbers of substitutions are used as distance measures for this case (fig. 2A), $D_A = 2$, $D_B = 4$, and $D_C = 4$. Because relations (1) hold, tree A is an NJ tree. If no mul-

Key words: tie trees, phylogenetic analysis, distance methods, neighbor-joining method, UPGMA.

Address for correspondence and reprints: Naoko Takezaki, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411, Japan. E-mail: ntakezak@lab.nig.ac.jp.

© 1998 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038
mple substitutions occur at any site, the $D$ values obtained by nucleotide differences between sequences are always the same as those obtained by the actual number of substitutions. However, if multiple substitutions occur at some sites, nucleotide differences may give different $D$, values and tie trees may appear in some cases. Let us assume that the substitutions in figure 2A actually occurred at two sites in figure 2B and C. In this case, if nucleotide differences are used as distance measures, trees A and B become tie trees. Let us denote the number of nucleotide differences between sequences $i$ and $j$ at a site by $n_{ij}$. $D_A$, $D_B$, and $D_C$ are the sums of $k_A = n_{12} + n_{34}$, $k_B = n_{13} + n_{24}$, and $k_C = n_{14} + n_{23}$ for all sites, respectively. At the site in figure 2B, one substitution occurred along the interior branch, and $k_A = 0$, $k_B = 2$, and $k_C = 2$. If only the nucleotide differences at this site are considered, relations (1) hold, and tree A is the NJ tree. At the site in figure 2C, however, parallel substitutions occurred along exterior branches leading to sequences 2 and 4. At this site, $k_A = 2$, $k_B = 0$, and $k_C = 2$. By adding the nucleotide differences at this site to the $D$ values, $D_A = 2$, $D_B = 2$, and $D_C = 4$. Thus, trees A and B become tie trees.

At the site in figure 2D, similar to the parallel substitutions in figure 2C, substitutions occurred along two of the exterior branches in different clusters. However, unlike parallel substitutions that result in the same nucleotide, they resulted in nucleotides that were different from each other. At this site, $k_A = 2$, $k_B = 1$, and $k_C = 2$. Therefore, if there are two sites as in figure 2D instead of the site in figure 2C, trees A and B also become tie trees because $D_A = 4$, $D_B = 4$, and $D_C = 6$.

Thus, tie trees can appear not only due to an interior branch of zero substitutions but also due to a combination of different site patterns by chance when multiple hits occur at some sites. The details of how the different patterns of substitutions at sites affect the occurrence of tie trees for the NJ method are given for four sequences in the appendix. As seen in the above example, the site patterns produced by a substitution along an interior branch give a smaller $k$ value for a correct tree than for incorrect trees (table 1). Also, substitutions along exterior branches within the same cluster (fig. 2E) give a smaller $k$ value for a correct tree than for incorrect trees. Note that the same site pattern (fig. 2E) can occur by a substitution along the interior branch and a substitution along an exterior branch (fig. 2F). In contrast, the site patterns produced by substitutions along the exterior branches in different clusters (figs. 2C and D) give a larger $k$ value for a correct tree than for incorrect trees.

**Fig. 1.**—Three topologies for unrooted trees of four OTUs.

**Fig. 2.**—Example of the case in which tie trees occur when there is a substitution along an interior branch (A), and examples of substitutions that produce different nucleotide patterns at a site (B–F). Numbers along branches indicate the number of substitutions. The divergence of sequences is assumed to be relatively low; therefore, the minimum number of substitutions that generate the site pattern are considered.

Tie trees tend to occur when no substitutions occur along an interior branch(es). However, tie trees are also likely to occur when few substitutions occur along an interior branch. At the sites where a substitution occurred along an interior branch, the $k$ value for a correct tree is smaller than those for incorrect trees. Therefore, tie trees can be produced by few sites of multiple hits where the $k$ value for a correct tree is larger than those for incorrect trees. Thus, when there is a short interior branch(es), tie trees are expected to occur relatively of-

**Table 1**

The Effects of Different Nucleotide Site Patterns on the Values of $D_A$, $D_B$, and $D_C$ when Nucleotide Differences Are Used as Distance Measures

<table>
<thead>
<tr>
<th>Site Pattern</th>
<th>$n_{12} + n_{34}$</th>
<th>$n_{13} + n_{24}$</th>
<th>$n_{14} + n_{23}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. XXXX</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2. XXXY</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3. XXYX</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4. XYXX</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5. YYYY</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6. XYYY</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7. XYYX</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8. XYYY</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>9. XXXZ</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10. XXZX</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>11. YXXZ</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>12. YYXZ</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>13. YYZX</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>14. XYYY</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>15. XYYZ</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: $n_{ij}$ is the number of nucleotide differences between sequences $i$ and $j$ at a site. $D_A$, $D_B$, and $D_C$ are the sums of $k_A = n_{12} + n_{34}$, $k_B = n_{13} + n_{24}$, and $k_C = n_{14} + n_{23}$, for all sites, respectively. X, Y, Z, and W represent any of the nucleotides A, T, G, and C. Each site pattern shows nucleotides of sequences 1, 2, 3, and 4 in order.
ten due to zero or few substitutions on the interior branch. Note that even when there are no substitutions along an interior branch(es), tie trees may not be generated if multiple hits occur at some sites. Therefore, when there is a short interior branch, tie trees tend to appear if the exterior branches are short and multiple hits do not occur.

When nucleotide difference was used as distance measure, distance values were discrete numbers. Therefore, it is relatively easy for $D$ values to have the same value. Let us assume, say, $D_A = D_B = 10$. $D_A = N_{12} + N_{34}$ and $D_B = N_{13} + N_{24}$, where $N_{ij}$ is a total number of nucleotide differences between sequences $i$ and $j$. There are many different combinations of values for $N_{12}$, $N_{34}$, $N_{13}$, and $N_{24}$ to have this equality. However, in computing corrected distances, logarithms are taken. For Jukes and Cantor’s (1969) distance, $D_A = -\frac{3}{4}\{\log[1 - (4N_{12}/(3N))] + \log[1 - (4N_{34}/(3N))]\}$ and $D_B = -\frac{3}{4}\{\log[1 - (4N_{13}/(3N))] + \log[1 - (4N_{24}/(3N))]\}$, where $N$ is the total number of sites. In this case, values of $D_A$ and $D_B$ are different unless $N_{12} = N_{13}$ and $N_{34} = N_{24}$, or $N_{12} = N_{24}$ and $N_{34} = N_{13}$. Therefore, tie trees for corrected distances are expected to occur less frequently than those for nucleotide differences. Note that the conditions for nucleotide differences under which an NJ tree or a UPGMA tree is chosen and those under which tie trees appear are the same for proportion of nucleotide differences ($p$ distance).

In the case of the UPGMA, the pair of OTUs that have the smallest distance are combined to make a cluster. If the root of the UPGMA tree is not considered, tree $A$ is the UPGMA tree if $d_{12}$ or $d_{34}$ is the smallest among all the distance values. Therefore, tree $A$ becomes a tie with tree $B$ or $C$ if $d_{12}$ or $d_{34}$ has the same value as any other distance. If the root of the tree is considered, tie trees may also occur in the cycle of clustering in which the root is generated. Thus, there are many more different conditions under which tie trees occur for the UPGMA than there are for the NJ method. Therefore, tie trees for the UPGMA seem to appear more often than do those for the NJ method. We can expect that tie trees for the UPGMA can appear relatively frequently when there is a short interior branch, similar to the case for the NJ method. Also, tie trees for the UPGMA can appear by chance when there are multiple hits at some sites, although the conditions for tie UPGMA trees to occur are different from those for tie NJ trees.

Effect of Rounding

In actual construction of an NJ or a UPGMA tree, tie trees may also appear when a small number of digits are used for distance values even when there are actually no tie trees. For example, when an NJ tree is constructed from the distance values in table 2, no tie trees appear ($D_A = 1.048$, $D_B = 1.058$, and $D_C = 1.072$). However, if we construct an NJ tree after rounding the third digit after the decimal point of the distance values, trees $A$ and $B$ become tie trees ($D_A = 1.05$, $D_B = 1.05$, and $D_C = 1.07$).

### Table 2

<table>
<thead>
<tr>
<th>OTU</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>...</td>
<td>0.53</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
<td>0.526</td>
<td>0.55</td>
<td>0.53</td>
</tr>
<tr>
<td>3</td>
<td>...</td>
<td>0.524</td>
<td>0.554</td>
<td>0.52</td>
</tr>
<tr>
<td>4</td>
<td>...</td>
<td>0.518</td>
<td>0.534</td>
<td>0.522</td>
</tr>
</tbody>
</table>

**Note**: The numbers in the lower triangular matrix are the original distance values, for which no tie trees occur. The numbers in the upper triangular matrix are the distance values after the third digit was rounded.

In their analysis of tie trees, Backeljau et al. (1996) constructed NJ and UPGMA trees with distance matrices from published papers. In those distance matrices, two or three digits are usually shown. Therefore, in Backeljau et al.’s analysis, tie trees might be generated more often than when an NJ or a UPGMA tree is constructed directly from gene frequency or sequence data. For example, Backeljau et al. (1996) found tie NJ trees in one case (data by Randi et al. 1994). Two digits were given in the distance matrix of Randi et al. (1994). However, if distance values are computed directly from the numbers of transitions and transversions shown in their table, tie trees do not appear.

### Method of Computer Simulation

The method for computer simulation was similar to that of Sourdis and Nei (1988). Figure 3A–D shows the model trees for generating eight nucleotide sequences. Figure 3A and $B$ is for the cases in which all the lineages evolve with a constant rate of substitution and are the same model trees as those in Sourdis and Nei’s figure 2A and $C$. Figure 3C and $D$ is for the cases in which the substitution rate varies with lineages. Kimura’s (1980) two-parameter model was used for nucleotide substitution with a transition–transversion rate ratio ($\alpha/\beta$) of two. Two levels of sequence divergence were considered. The expected numbers of substitutions per site from the root to the tip of the model trees ($U$) were 0.5 and 0.05 for the constant-rate cases (fig. 3A and $B$). The expected number of substitutions per site for the shortest interior branch ($a$) of the model trees for the constant-rate cases was one eighth of $U$ for the symmetric tree (fig. 3A) and one tenth of $U$ for the asymmetric tree (fig. 3B). The same $a$ value was used for the varying-rate cases (fig. 3C and $D$).

In each replication, sequences of 300, 600, or 1,200 nucleotide sites were generated, and an NJ tree and a UPGMA tree were constructed. The distance measures used were the proportion of nucleotide differences ($p$ distance), Jukes and Cantor’s (1969) (JC) distance, and Kimura’s (1980) two-parameter (K2P) distance.

When an NJ or a UPGMA tree is constructed, a pair of OTUs is combined to make a cluster in each cycle of tree construction. When two or more pairs of OTUs (tie OTU pairs) give the smallest sum of branch lengths for the NJ method or the smallest distance for the UPGMA, tree topologies generated by combining all tie OTU pairs are examined to see whether or not dif-
Fig. 3.—Model trees for computer simulation. The expected numbers of nucleotide substitutions per site from the root of the model trees to the tip (U) were assumed to be 0.5 and 0.05 for the constant-rate cases. A, Symmetric tree with constant rate. B, Asymmetric tree with constant rate. U = 8a. C, Symmetric tree with varying rate. D, Asymmetric tree with varying rate. The a value in the constant rate cases was used for the varying-rate cases.

Different tree topologies (tie trees) and the correct tree topology appear. Note that in the case where tie OTU pairs appear in some cycles, tie trees do not always appear, because the same tree topology may be generated by combining different OTU pairs. When tree topologies were compared, the root of the UPGMA tree was removed because the NJ method produces an unrooted tree.

After 20,000 replications, the percentages of replications in which tie trees appeared (Pt) and those in which the correct tree topology was generated (Pc) were computed. Replications in which tie trees appeared were included for computation of P_C if one of the tie trees gave a correct tree topology.

In the simulation, the double-precision floating-point type was used for computation of the distance measures and the sum of branch lengths. Rounding errors seem to occur after 12 to 15 digits for the double-precision type. Note that rounding errors for the regular floating-point type occur after about six digits. To avoid the effect of rounding error, the sums of branch lengths or the distances were regarded as equal if the difference (tie tolerance [TT]) was smaller than or equal to 10^{-10}. Pt values may change for different TTs. Therefore, in the preliminary study, we used TT = 10^{-5} and 10^{-15}. Pt values were slightly higher or lower than those for TT = 10^{-10}, but the results were essentially the same.

Backeljau et al. (1996) found a case in which MEGA (Kumar, Tamura, and Nei 1993) generated a UPGMA tree that was different from the UPGMA trees generated by other computer programs (data by Wingfield et al. 1994). They found that the UPGMA tree generated by MEGA was included in the tie trees that appeared for TT = 0.005. They suspected that MEGA generated a different UPGMA tree because the computer programs had different TT values or rounding errors. Actually, the early version of MEGA (version 1.0) had a bug in constructing UPGMA trees, and this seems to be the reason why a different UPGMA tree was generated by MEGA. (This bug was removed in version 1.01, which is widely used.) Backeljau et al. (1996) also reported that different UPGMA trees were produced by three computer packages (SAS, MEGA, and NTSYS) for Plötner, Becker, and Plötner’s (1993) data. In this case, however, the reason is not clear. There were no tie trees when TT was changed to 0.005 (data not shown).

Bootstrap Test for Tie Trees

In order to see whether or not the reliability of clusters differing among tie trees (tie clusters) are assessed properly, the bootstrap test was conducted for tie trees that appeared in the above simulation. One thousand bootstrap replications were conducted in the cases with p distance, because tie trees appeared relatively frequently in these cases. In each bootstrap replication, nucleotide sites were resampled at random with replacement, and an NJ or a UPGMA tree was constructed. If a clustering pattern in the tie trees coincided with one of clusters in a tree constructed in a bootstrap replication, the bootstrap value for the cluster was increased by one.

Tie trees can also appear in bootstrap replications. Without proper treatment for the tie trees, bootstrap confidence values may be inflated or deflated, depending on the input order of sequences (Backeljau et al. 1996). This happens because the computer software chooses one of the tie OTU pairs in a cycle, depending on the order in which they appear in computation, and this order of tie OTU pairs is determined by the input order of sequences. For example, as we examined before, if no substitutions occur along the interior branch of the four-sequence tree in fig. 1A and there are no multiple hits at any sites, the three tree topologies (fig. 1A–C) are tie. In this case, the three topologies are always tie
in bootstrap replications. However, if the computer program chooses an OTU pair with the earliest input order among the tie OTU pairs, sequences 1 and 2 always make a cluster. In such a case, tree A will be supported by 100% bootstrap probability (BP).

Two methods for treating this problem were examined in the simulation. In the first method, all tie trees were searched in each bootstrap replication. If n tie trees appeared in a bootstrap replication and one of them had the same cluster as in the original tree, 1/n was added to the bootstrap value of the cluster (MEGA manual, p. 46). In the second method, when tie OTU pairs appeared in cycles of tree making, a tree was constructed by choosing one of the tie OTU pairs at random in order to avoid the effect of an input order of sequences. Since the former method searches all tie trees in each bootstrap replication, this method will give a more accurate estimate of the bootstrap confidence level than the latter. However, if many tie trees appear in bootstrap replications, the computation will become intensive, and it may become difficult to conduct a sufficiently large number of bootstrap replications. By contrast, the latter always constructs a single tree in a bootstrap replication. Therefore, it will not be difficult to conduct a large number of bootstrap replications. In the simulation, the bootstrap test was conducted by both methods. The bootstrap values were also computed without treatment for tie trees during the bootstrap. The bootstrap test was not conducted for the asymmetric tree with 300 sites in the low-divergence case because it became computation-intensive.

In the PHYLIP package (Felsenstein 1995), the input order of sequences can be randomized to search for tie trees. This should have the same effect as the latter method described above. This randomization of sequence order can be done for construction of an original NJ or UPGMA tree in the PHYLIP package. However, the randomization cannot be conducted in bootstrap replications by the NJ method or the UPGMA, contrary to what Backeljau et al. (1996) indicated. Note that sequence order can be randomized during bootstrap for the maximum-parsimony (MP) method and for the maximum-likelihood method as well as for the distance method by FITCH and KITSCH in the PHYLIP package.

### Results

The Probabilities \((P_T \text{ and } P_C)\) that Tie Trees and Correct Tree Topologies Appear

Table 3 shows \(P_T \text{ and } P_C\) values for the NJ method. First, let us look at the \(P_T\) values. When the divergence level is high \((U = 0.5)\), \(P_T\) values are quite low for both the constant- and varying-rate cases. Tree ties are never observed for JC and K2P distances. For \(p\) distance, tie trees appear in about 1%–2% of replications with 300 sites. However, as the number of nucleotide sites increases, the \(P_T\) decreases. When the divergence is low \((U = 0.05)\), tie trees appear more frequently than for \(U = 0.5\). This seems to have happened because an interior branch(es) of no substitutions or few substitutions is likely to occur in the low-divergence case. With 300 sites, \(P_T\) values for \(p\) distance are generally about 10%, whereas \(P_T\) values for JC and K2P distances are usually low (<0.4%). \(P_T\) values are exceptionally high in the varying-rate case for the asymmetric tree. With 300 sites, \(P_T\) for \(p\) distance is 25.8% and those for JC and K2P distances are about 4% and 3%, respectively. As mentioned before, even when there is an interior branch of no substitutions, tie trees may not appear if multiple hits occur at some sites. However, in the asymmetric tree with varying rate (fig. 2D), many exterior branches...
Table 4
Percentages of Replications in Which Correct Tree Topologies Were Obtained ($P_C$) and Those in Which Tie Trees Appeared ($P_T$) for the UPGMA

<table>
<thead>
<tr>
<th>NO. OF SITES</th>
<th>SYMMETRIC</th>
<th></th>
<th></th>
<th></th>
<th>ASYMMETRIC</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_C$</td>
<td>$P_T$</td>
<td>$P_C$</td>
<td>$P_T$</td>
<td>$P_C$</td>
<td>$P_T$</td>
<td>$P_C$</td>
<td>$P_T$</td>
</tr>
<tr>
<td>High divergence, constant rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300 . . . 43.89</td>
<td>9.38</td>
<td>43.77</td>
<td>5.32</td>
<td>41.85</td>
<td>0.35</td>
<td>44.63</td>
<td>6.21</td>
<td>42.08</td>
</tr>
<tr>
<td>600 . . . 77.97</td>
<td>3.89</td>
<td>77.97</td>
<td>1.64</td>
<td>77.51</td>
<td>0.23</td>
<td>75.24</td>
<td>2.54</td>
<td>73.39</td>
</tr>
<tr>
<td>1,200 . . . 98.14</td>
<td>0.00</td>
<td>98.14</td>
<td>0.00</td>
<td>98.14</td>
<td>0.00</td>
<td>97.68</td>
<td>0.00</td>
<td>97.22</td>
</tr>
<tr>
<td>High divergence, varying rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300 . . . 27.43</td>
<td>1.41</td>
<td>28.45</td>
<td>0.57</td>
<td>28.71</td>
<td>0.00</td>
<td>0.00</td>
<td>6.15</td>
<td>0.00</td>
</tr>
<tr>
<td>600 . . . 13.47</td>
<td>0.46</td>
<td>15.52</td>
<td>0.00</td>
<td>15.53</td>
<td>0.00</td>
<td>0.00</td>
<td>2.04</td>
<td>0.00</td>
</tr>
<tr>
<td>1,200 . . . 4.62</td>
<td>0.46</td>
<td>5.08</td>
<td>0.00</td>
<td>5.08</td>
<td>0.00</td>
<td>0.00</td>
<td>1.22</td>
<td>0.00</td>
</tr>
<tr>
<td>Low divergence, constant rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300 . . . 12.70</td>
<td>35.34</td>
<td>12.58</td>
<td>23.73</td>
<td>10.27</td>
<td>4.79</td>
<td>8.35</td>
<td>33.85</td>
<td>7.66</td>
</tr>
<tr>
<td>600 . . . 28.92</td>
<td>12.12</td>
<td>28.69</td>
<td>12.12</td>
<td>25.83</td>
<td>0.69</td>
<td>17.91</td>
<td>19.71</td>
<td>13.08</td>
</tr>
<tr>
<td>1,200 . . . 66.67</td>
<td>8.33</td>
<td>66.67</td>
<td>4.17</td>
<td>63.90</td>
<td>0.00</td>
<td>52.91</td>
<td>6.74</td>
<td>52.03</td>
</tr>
<tr>
<td>Low divergence, varying rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300 . . . 30.75</td>
<td>24.39</td>
<td>30.12</td>
<td>14.76</td>
<td>29.09</td>
<td>3.83</td>
<td>0.00</td>
<td>41.45</td>
<td>0.00</td>
</tr>
<tr>
<td>600 . . . 38.72</td>
<td>12.64</td>
<td>37.56</td>
<td>7.79</td>
<td>36.62</td>
<td>1.15</td>
<td>0.00</td>
<td>21.07</td>
<td>0.00</td>
</tr>
<tr>
<td>1,200 . . . 32.88</td>
<td>1.61</td>
<td>32.42</td>
<td>1.61</td>
<td>32.20</td>
<td>0.00</td>
<td>0.00</td>
<td>8.55</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note: $\alpha = 10^{-10}$.

are short, and multiple hits tend not to occur. Therefore, in this case, tie trees can occur quite frequently due to a zero interior branch.

Now let us look at the $P_C$ values. In the constant-rate cases, $p$ distance has higher $P_C$ values than JC and K2P distances. The difference between $P_C$ values for $p$ distance and for JC or K2P distance is larger for $U = 0.5$ than for $U = 0.05$. This indicates that $p$ distance is more efficient in obtaining a correct tree topology than are corrected distances such as JC or K2P when the substitution rate is constant and the divergence level is high. However, when the substitution rate varies extensively, it is known that performance of corrected distances becomes better than that of $p$ distance (Nei 1991; Schöniger and von Haeseler 1993; Tajima and Takezaki 1994; Rzhetsky and Sitnikova 1996). In the varying-rate case, $P_C$ values for $p$ distance are much lower (<11%) than for JC or K2P distance for the asymmetric tree with the high divergence level, and $P_C$ values decrease as the number of sites increases (table 3). However, even in the varying-rate cases $P_C$ values for $p$ distance are still higher than those for JC or K2P distance for the low divergence level and for the symmetric tree with the high divergence level. This indicates that $p$ distance may be useful for obtaining a correct tree topology even when the rate of evolution varies to some extent. However, the condition under which $p$ distance is more efficient than corrected distance in the varying-rate case is still unclear, and $p$ distance may give a positively misleading result in certain cases. Therefore, it is not safe to construct a phylogenetic tree with $p$ distance when the rate varies with lineages (Rzhetsky and Sitnikova 1996).

Table 4 shows $P_T$ and $P_C$ values for the UPGMA. As mentioned in the previous section, the root of a UPGMA tree was removed when the tree topology was compared. If the root of a UPGMA tree was considered, $P_C$ values were slightly smaller than those for an unrooted UPGMA tree, and $P_T$ values were larger (data not shown). $P_T$ values for the UPGMA are much higher than those for the NJ method in table 3. Even for $U = 0.5$, tie trees appear for $p$ distance in 5%–10% of replications with 300 sites. $P_T$ values for JC distance are about half those for $p$ distance or lower. However, even in this case, $P_T$ values for the K2P distance are very low (<0.5%). For $U = 0.05$, $P_T$ values are much higher than those for $U = 0.5$, as in the case for the NJ method. With 300 sites, $P_T$ values for $p$ distance are 35%–40%, and even for the K2P distance, $P_T$ values are about 5% or higher. This indicates that tie trees may appear quite often for the UPGMA for sequences with a low divergence level.

It has been shown that the UPGMA is inefficient in obtaining a correct tree when the assumption of rate constancy is violated (Chakraborty 1977; Saitou and Nei 1987; Takezaki and Nei 1996). In table 4, $P_C$ values for the UPGMA are indeed low in the varying-rate cases. Particularly for the asymmetric tree, $P_C$ values are always zero. In the constant-rate cases, $P_C$ values for the UPGMA are quite low when the divergence level is low. However, when the divergence is high, the $P_C$ values for the UPGMA are slightly higher than those for the NJ method for the asymmetric tree.

The Number of Tie Trees

Figure 4 shows the frequency distribution of the number of tie trees. The cases shown in the figure are the low-divergence cases for $p$ distance with 300 sites. The frequency distribution is not shown for the other cases, because $P_T$ values were generally very low and the numbers of tie trees were mostly two or three. As shown in figure 4, there are two tie trees in most of the
cases in which tie trees occur. Generally, the shape of the distribution is similar for the NJ method and for the UPGMA. However, tie trees occur more frequently for the UPGMA than they do for the NJ method, and the number of tie trees for the UPGMA is larger than that of those for the NJ method (fig. 4A–C). However, for the asymmetric tree with varying rate (fig. 4D), the number of tie trees for the NJ method tends to be large. This was the case in which $P_T$ values were unusually high for the NJ method (table 3). In this case, three tie trees occur almost as frequently as two tie trees, and a very large number of tie trees appear occasionally.

Bootstrap Test for Tie Trees

In order to see how much BPs are inflated if no treatment is given for tie trees during bootstrap replications, the BPs obtained by giving no treatment for tie trees during bootstrap were compared to those obtained by searching all tie trees in each bootstrap replication. Figure 5 shows the case in which the largest difference was observed for the BPs by the two methods. In this case, the length of the interior branch leading to the cluster of sequences 1 and 2 is zero in the original tree. Therefore, the branching patterns $((1, 2), 3)$, $((1, 3), 2)$ and $(1, (2, 3))$ are tie (fig. 5A–C). When no treatment was given for tie trees during bootstrap, the branching pattern $((1, 2), 3)$ was supported by 76% and $((1, 3), 2)$ and $(1, (2, 3))$ by 15% and 7% BP, respectively. If the computer program always chooses the tie OTU pair with the earliest input order, $((1, 2), 3)$ should have 100% BP. However, sometimes $((1, 3), 2)$ and $(1, (2, 3))$ were chosen. This may be due to the instability of the floating-point values in the computer program. When all tie trees are searched in bootstrap replications, all the three branching patterns have 33% BP, as we expect for this case.
The case in which the largest difference was observed in the bootstrap probabilities (BPs) obtained by giving no treatment for tie trees and those obtained by searching all tie trees in bootstrap replications. The different BPs given by the two methods are shown in boldface. The numbers before the slashes are BPs obtained by giving no treatment for tie trees during bootstrap, and those after the slashes are BPs obtained by searching all tie trees. There are three tie trees (A, B, and C) from the original data. Tie trees differ by the branching patterns of sequences 1, 2, and 3. The branching patterns are ((1, 2), 3), ((1, 3), 2), and (1, (2, 3)) in trees A, B, and C, respectively. The different BPs were given by the two methods for the clusters (tie clusters) that differ among the tie trees. When no treatment was given for tie trees during bootstrap, the computer program tended to choose the OTU pair with the earliest input order. BPs for the tie clusters (1, 2), (1, 3), and (2, 3) were 76%, 15%, and 7%, respectively. However, when all tie trees were searched during bootstrap, BPs for all three tie clusters were 33%. This case occurred for the asymmetric tree with varying rate, 600 sites, and low divergence.

In other cases in which large differences in BPs by the two methods were observed (>30%), there was also an interior branch of zero length in the original tree. However, such a large difference in BPs occurred rarely for the two methods. In most of the cases, the differences in the BPs were relatively small (<5%). The discrepancy in the BPs for the two methods became even smaller for the high divergence level or the larger number of sites. Therefore, it seems that the BPs are greatly inflated by giving no treatment for tie trees when there is an interior branch(es) of length zero or nearly zero. However, if tie trees occur by chance due to multiple hits, the discrepancy of BPs by the two methods seems to be very small.

BPs computed by searching all tie trees in each bootstrap replication were also compared with those obtained by randomly choosing one of the tie OTU pairs in cycles of tree making. The differences in BPs for the two methods were generally very small (at most 5%) even when many tie trees occurred. Because the latter method is less computation-intensive, BPs can be estimated by this method in practice.

Next, in order to see whether or not the accuracy of clusters (tie clusters) that differ among tie trees is assessed properly by the bootstrap test, the BPs for tie clusters were examined. Figure 6 shows the frequency distributions of the BPs for tie clusters. BPs were com-

![Figure 5](image-url)

![Figure 6](image-url)
The cases in which the highest bootstrap probabilities (BPs) for clusters (tie clusters) that differ among tie trees were observed. The BPs for tie clusters are shown in boldface. Asterisks indicate the highest BPs for tie clusters. A and B, The case for the NJ method. C and D, The case for the UPGMA. The highest BPs for tie clusters were 69% for the NJ method (tree A), and 76% for the UPGMA (tree C). The case for the NJ method is for the asymmetric tree with varying rate and 600 nucleotide sites. The case for the UPGMA is for the symmetric tree with constant rate and 300 sites. Both cases are for the high divergence level.

Computation of tie trees was carried out by searching all tie trees in bootstrap replications. The figure shows the constant-rate cases for the symmetric tree with 300 sites. However, note that the shape of the distribution was similar in the other cases. In the low-divergence case (fig. 6A), the distributions are shifted to the left side compared with those in the high-divergence case (fig. 6B). This seems to have happened because a larger number of tie trees tend to appear in one replication in the low-divergence case than in the high-divergence case. Thus, when there are many tie trees, each tie cluster tends to be supported by a small BP. In the low-divergence case, the peak of the distribution is 15%–20% for the UPGMA, whereas that for the NJ method is 35%–40%. This also seems to be because the number of tie trees in one replication is larger for the UPGMA than for the NJ method (fig. 4). When the number of sites increased, the distribution became narrower and converged to about 50% (data not shown).

As shown in figure 6, most of the BPs for tie clusters are 60% or less. However, relatively high BPs appeared with a small probability even when all tie trees were searched during bootstrap. Figure 7 shows the cases in which the highest BPs are observed for tie clusters. The highest BPs for tie clusters were 69% for the NJ method (fig. 7A) and 76% for the UPGMA (fig. 7C). Interestingly, these are both high-divergence cases and there are no zero interior branches. Therefore, the tie trees occurred in these cases due to multiple hits.

In any event, this result indicates that the BPs are mostly low (<60%) for tie clusters, reflecting the presence of other tie trees, and that high BPs (e.g., >95%; Berry and Gascuel 1996; Efron, Halloran, and Holmes 1996) are unlikely to occur for tie clusters. Therefore, the reliability of tie clusters appears to be properly determined by the bootstrap test.

Discussion

This study has shown that tie trees for distance methods tend to appear for closely related sequences with a small number of sites, whereas tie trees rarely occur for sequences with a high divergence level or with a large number of sites. With the NJ method, tie trees tend to be rare. In the low-divergence case, tie trees appeared with a small probability for p distance with a small number of sites. However, tie trees for the NJ method rarely occurred for corrected distances. Quite high $P_T$ values were observed for the NJ method only when there were many short branches on the model tree. In the high-divergence case, tie trees for the NJ method appeared with a small probability only for $p$ distance. Tie trees for the NJ method did not appear for the corrected distances. In contrast, tie trees for the UPGMA occurred quite frequently in the low-divergence case. Even in the high-divergence case, tie trees for the UPGMA appeared relatively frequently if the number of sites was small.

The MP method (Fitch 1971) is commonly used to construct phylogenetic trees. It is known to often generate equally parsimonious trees (tie MP trees) (Sourdis
and Nei 1988). Therefore, \( P_T \) values for the MP method were compared for the UPGMA and for the NJ method (unpublished data). The same simulation scheme was used for this comparison. \( P_T \) values for the MP method were generally higher than those for the UPGMA and the NJ method. Particularly with a small number of sites, \( P_T \) values for the MP method were much higher than those for the NJ method and the UPGMA. This seems to have happened because a relatively small number of parsimony-informative sites (Fitch 1977) are available when the number of sites is small (Sourdis and Nei 1988). However, \( P_T \) values for the UPGMA became larger than those for the MP method for the symmetric tree with a larger number of sites. For the symmetric tree, tie trees for the UPGMA tend to occur often in the low-divergence case, because the same distance value can easily appear for different pairs of sequences.

In this study, the probability that tie trees occur was examined for a relatively small number of sequences. If the number of sequences becomes large, the probability that tie trees will appear may increase. It is known that there are a large number of equally parsimonious trees for human mitochondrial DNA sequences (Hedges et al. 1992). Therefore, to have some idea of how tie trees appear for a large number of sequences, tie trees were searched for the 133 human mitochondrial sequences of Redd et al. (1995). In the case of the UPGMA, the search for tie trees did not finish because there are so many tie OTU pairs in cycles of tree making and the computation became intensive. Therefore, it is not clear how many tie trees occur and how tie trees are different from one another for the UPGMA. For the NJ method, there were 945 tie trees for \( p \) distance and 105 tie trees for JC and K2P distances. However, the difference in the tie trees among the tie trees occurred only for the branching pattern of the six sequences at the very tip of the tree (data not shown). These six sequences are almost identical (one or two nucleotide differences), such that the branching among them is close to multifurcation. Indeed, the bootstrap values for the branching pattern for the six sequences were virtually zero. Thus, there were many tie trees for the NJ method. However, the topologies of the tie trees were very similar to one another and the difference in the topologies of tie trees has virtually no effect on the resolution of the branching pattern of the NJ method. This is quite different from the case of tie trees for the MP method. When a consensus tree was constructed for equally parsimonious trees for the human mitochondrial sequences, there was only a poor resolution of the branching pattern (Hedges et al. 1992).

As this study has shown, tie trees for distance methods can appear under some conditions. However, if a large number of sites are used or the divergence of sequences is high, tie trees rarely appear. Further, tie trees for corrected distances tend to be rare, although tie trees for \( p \) distance can occur relatively frequently in some cases. In inferring phylogenetic trees, it may be useful to examine tie trees. However, even when tie trees occur, the reliability of the tree topologies seems to be estimated properly by the bootstrap test. Therefore, testing the reliability of the branching pattern appears to be important in actual analyses.

**Acknowledgments**

This study was motivated by a discussion with Dr. Masatoshi Nei. It was supported by research grants from the National Institutes of Health and the National Science Foundation to Dr. Masatoshi Nei and by a Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Science and Culture to Dr. Takashi Gojobori.

**APPENDIX**

Let us examine how different patterns of substitutions at sites affect the occurrence of tie trees when nucleotide differences between sequences are used as distance measures for the NJ tree. With respect to the nucleotide difference, there are 15 patterns at a site for four nucleotide sequences (table 1). Table 1 shows \( k_A = n_{12} + n_{34} \), \( k_B = n_{13} + n_{24} \), and \( k_C = n_{14} + n_{23} \) for all the site patterns. In site patterns 2–8, three nucleotides are the same and one nucleotide is different from the others. These site patterns can appear as a result of a substitution along an exterior branch. For such sites, \( k_A = k_B = k_C = 1 \), so these sites give no resolution to the branching pattern of the four sequences. Also, if all the nucleotides at a site are different from each other, as in pattern 15, \( k_A = k_B = k_C = 2 \). As stated before, \( D_A \), \( D_B \), and \( D_C \) are sums of \( k_A \), \( k_B \), and \( k_C \) for all sites. Therefore, in order to compare the values of \( D_A \), \( D_B \), and \( D_C \), it is necessary to take into account only site patterns 6–14. Relations (1) are equivalent to

\[
2m_7 + m_{10} + m_{13} < 2m_6 + m_9 + m_{14}
\]

and

\[
2m_8 + m_{11} + m_{12} < 2m_6 + m_9 + m_{14}, \quad (A1)
\]

where \( m_i \) is the number of sites that have pattern \( i \). Therefore, a tie tree for tree A (fig. 1) occurs whenever an equality occurs in either of the above relations.

**LITERATURE CITED**


Naruya Saitou, reviewing editor

Accepted February 26, 1998