When is it Safe to Use an Oversimplified Substitution Model in Tree-Making?

Andrey Rzhetsky and Tatyana Sitnikova
Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University

The choice of an "optimal" mathematical model for computing evolutionary distances from real sequences is not currently supported by easy-to-use software applicable to large data sets, and an investigator frequently selects one of the simplest models available. Here we study properties of the observed proportion of differences (p-distance) between sequences as an estimator of evolutionary distance for tree-making. We show that p-distances allow for consistent tree-making with any of the popular methods working with evolutionary distances if evolution of sequences obeys a "molecular clock" (more precisely, if it follows a stationary time-reversible Markov model of nucleotide substitution). Next, we show that p-distances seem to be efficient in recovering the correct tree topology under a "molecular clock," but produce "statistically supported" wrong trees when substitution rates vary among evolutionary lineages. Finally, we outline a practical approach for selecting an "optimal" model of nucleotide substitution in a real data analysis, and obtain a crude estimate of a "prior" distribution of the expected tree branch lengths under the Jukes-Cantor model. We conclude that the use of a model that is obviously oversimplified is inadvisable unless it is justified by a preliminary analysis of the real sequences.

Introduction

An essential part of the modern evolutionary studies involves the inference of phylogenetic relationships among species through comparison of their genes. To allow for this inference, the majority of the tree-making methods require that one of a few existing models of nucleotide substitution be chosen for describing evolution of genes. The problem of selecting the mathematical model for estimating evolutionary distances or tree-making with the maximum-likelihood method is far from trivial (see Goldman 1993a, 1993b; Yang, Goldman, and Friday 1994; Rzhetsky and Nei 1995) because the multiparameter mathematical models that fit the real data well are not always more successful in recovering the correct tree topology than the models with fewer parameters (see Saitou and Nei 1987; Soursis and Krimbas 1987; Nei 1991; Zharkikh and Li 1992, 1993; Huls 1993; Tateno, Takezaki, and Nei 1994; Gauth and Lewis 1995; Yang 1996).

Since the choice of an "optimal" mathematical model for computing evolutionary distances from real sequences is not currently supported by easy-to-use software applicable to large data sets, evolutionists often use for tree-making the simplest substitution model available, neglecting potential pitfalls of this approach. In particular, the observed proportion of differences (p-distance) between a pair of genes is occasionally used as an estimate of the evolutionary distance between corresponding species (e.g., see Gouy and Li 1989; Poinar, Cano, and Poinar 1993; Vingron and von Haeseler 1994).

In this paper we analyze behavior of the distance-based tree-making methods when the distances are estimated under an oversimplified model. Specifically, we (1) identify the conditions of consistency of tree-making with p-distances, (2) study the efficacy of tree-making from "proper" and "improper" distance estimates with a computer simulation, (3) attempt to give an intuitive explanation of the observed regularities, and (4) outline practical steps to simplify selection of substitution models in the analysis of the actual data.

Tree-Making with p-Distances is Consistent Under a "Molecular Clock"

The good news for those interested in using p-distances is that any consistent distance-based method combined with p-distances provides a consistent estimate of the correct tree topology if the sequences under analysis are generated under any stationary time-reversible Markov model of nucleotide substitution. (Time-reversibility is an essential part of our argument in support for this statement. Readers not interested in the technical details of this part are advised to skip the rest of this section.)

To begin with, we give a verbal definition of a stationary time-reversible Markov model. Most of the models used in evolutionary analyses describe a gene as a collection of independent sites each being an instance of the same random process, a Markov chain. The Markov property of the process implies that the probability, \( P_j(s, s + t) \), of observing a nucleotide \( j \) substituted with a nucleotide \( k \) within time interval \([s, s + t]\) is independent of the history of the site before time \( s \). Furthermore, the Markov process is called stationary if the probability \( P_j(s, s + t) \) does not depend on \( s \),

\[
P_j(s, s + t) = P_j(t),
\]

and the vector of probabilities, \( \pi = \{ \pi_A, \pi_T, \pi_C, \pi_G \} \), of observing the process in each of four possible states, A, T, C and G, does not change with time. When interpreted in evolutionary terms, a stationary Markov process of gene evolution incorporates a "molecular clock." In phylogenetic literature (e.g., see Tavaré 1986), the Markov models are commonly defined through the
"instantaneous substitution rate" matrix, \( Q \), connected with the matrix of substitution probabilities, \( P(t) = [P_{ij}(t)] \), by relation

\[
P(t) = \exp(Qt).
\]

A Markov process is called time-reversible if matrix \( HQ \) is symmetrical, where \( H = \text{diag} \{ \pi_i \} \) is a square matrix with elements of vector \( \pi = [\pi_A, \pi_C, \pi_G, \pi_T] \) on the main diagonal; for rigorous definition of the time-reversible processes see Tavaré (1986) and Keilson (1979). Virtually all models of nucleotide substitution that are presently used in data analysis, e.g., those by Jukes and Cantor (1969), Kimura (1980), and Hasegawa, Kishino, and Yano (1985), belong to the class of time-reversible models.

In our proof we make use of a theorem which was independently discovered by several mathematicians (Zaretskii, Buneman, Patrinos and Hakimi, and Dobson, see Barthélemy and Guénoche 1991, p. 52); here this theorem is referred to as "the ZBPHD theorem."

The ZBPHD-Theorem (Zaretskii, Buneman, Patrinos and Hakimi, Dobson)

Let \( d \) be a metric on a set \( S \) (\( S \) is a set of present-day sequences) which satisfies the "four-point condition" (explained below). Then (1) there is a tree \( T \) (topology and strictly positive branch lengths) which contains the members of \( S \) among its pending vertices and which induces \( d \), i.e., for any two sequences, \( i \) and \( j \), the length of the shortest path connecting them in \( T \) is equal to \( d_{ij} \), and (2) the tree \( T \) and the set of branch lengths are defined in a unique way and can be recovered in polynomial time.

Applied to \( p \)-distances, the ZBPHD theorem states that if (1) the expected values of \( p \)-distances, \( E(p_{ij}) \)'s, have metric properties and (2) the expected \( p \)-distances satisfy the four-point condition, the complete set of the expected \( p \)-distances forms an additive distance matrix (i.e., there is a unique weighted tree such that each entry of the matrix, \( d_{ij} \), corresponds to the length of the shortest path between pending vertices \( i \) and \( j \) in the tree). Starting with an additive distance matrix, every consistent tree making method, by definition of consistency, recovers the same tree. To prove that this unique tree is indeed the correct one, we need to show that the expected \( p \)-distances define the correct tree topology for each quartet of sequences. To prove our assertion, we need to show two things: (1) that the \( p \)-distance is a metric and (2) that the four-point condition,

\[
E(p_{ij} + p_{ij}) \leq E(p_{ik} + p_{jk}) = E(p_{ik} + p_{jk}).
\]

is satisfied whenever the unrooted four-sequence tree \( (i,j,k,l) \) is correct. (It suffices to study only two rooted four-species trees, shown in the figure 1A and B, because all other rooted topologies for these sequences are obtainable by relabeling sequences \( i, j, k, \) and \( l \).

It is easy to check that the \( p \)-distance is indeed a metric (see also Saitou and Nei 1987), because for any three sequences \( i, j, \) and \( k \) we have (1) \( p_{ij} \geq 0 \), (2) \( p_{ij} = 0 \) if \( i = j \), and (3) \( p_{ij} + p_{jk} \leq p_{ik} \).

It takes a little longer to show that the expected values of \( p \)-distances satisfy the four-point condition. Using the properties of time-reversibility and stationarity of the Markov model generating the present-day sequences, we can express the expected \( p \)-distances between sequences \( i, j, k, \) and \( l \) related by the tree shown in figure 1A by

\[
E(p_{ij}) = 1 - \text{tr}(\text{IIe}^2Q'),
\]

\[
E(p_{ij}) = 1 - \text{tr}(\text{IIe}^2Q''),
\]

\[
E(p_{ij}) = 1 - \text{tr}(\text{IIe}^2Q'''),
\]

\[
E(p_{ij}) = 1 - \text{tr}(\text{IIe}^2Q''''),
\]

Here \( \text{tr} \) stands for the sum of entries on the main diagonal of a square matrix \( A \) ("trace"), and \( \text{II} = \text{diag} \{ \pi \} \). Similarly, for tree \( B \) in figure 1 we obtain

\[
E(p_{ij}) = 1 - \text{tr}(\text{IIe}^2Q'),
\]

\[
E(p_{ij}) = 1 - \text{tr}(\text{IIe}^2Q''),
\]

\[
E(p_{ij}) = 1 - \text{tr}(\text{IIe}^2Q'''),
\]

\[
E(p_{ij}) = 1 - \text{tr}(\text{IIe}^2Q''''),
\]

Clearly, the four-point condition is satisfied for both trees if

\[
\text{tr}(\text{IIe}^2Q') \geq \text{tr}(\text{IIe}^2Q'') \geq \text{tr}(\text{IIe}^2Q''')
\]

for \( t_1 \approx t_2 \approx t_3 \),

\[
\text{tr}(\text{IIe}^2Q') \geq \sum_{i=1}^{4} \pi_i^2 + \sum_{i=1}^{4} \sum_{j=1}^{4} (w_{ij} - \lambda_i)(\text{cos}^2(t_3)).
\]

where \( \lambda_i \)’s and \( w_{ij} \)’s are real numbers such that \( \lambda_i < 0 \) (\( u = 1, 2, 3, 4 \) and \( u \)’s are not all zero. From examining equation (7) it is obvious that \( \text{tr}(\text{IIe}^2Q') \) is strictly negative for finite \( t \), and \( \text{tr}(\text{IIe}^2Q') \) is a monotone decreasing function of \( t \). Therefore, the four-point condition (eq. 3) follows from the validity of inequality (eq. 6), and this concludes our proof.

The above proof can be easily extended to more general cases. For example, it can be easily modified to
cover the models of protein evolution (for this, matrix \( Q \) and vector \( \pi \) should be defined as \( 20 \times 20 \) matrix and a \( 1 \times 20 \) vector, respectively; the argument remains exactly the same), the models allowing for rate variation across sequence sites, and mixtures of time-reversible models.

**Tree-Making with p-Distances is Dangerous if the “Molecular Clock” is Violated**

The bad news for those using \( p \)-distances is that one can easily get a misleading tree supported by high “bootstrap values” (see Felsenstein 1985) when substitution rates in the true tree do vary among evolutionary lineages.

Being curious about the shape of the “consistency zone” associated with \( p \)-distances, we tried to visualize it (assuming the Jukes-Cantor model of nucleotide substitution) for an unrooted four-sequence tree by “slicing” the five-dimensional parameter space defined by the expected branch lengths of the tree \( b_1, b_2, b_3, b_4, \) and \( b_5 \) into 27 two-dimensional projections (see fig. 2). Each of the large squares in figure 2 corresponds to a “slice” of the five-dimensional space with fixed values of \( b_3, b_4, \) and \( b_5 \) (see fig. 2), while the values of \( b_1, b_2, \) and \( b_7 \) run from 0.0375 to 0.7125 with increment 0.075 within each of the squares. Each small square, therefore, corresponds to a fixed value of vector \( (b_1, b_2, b_3, b_4, b_5) \) and is shown in black or in white, depending on whether or not, respectively, the corresponding set of \( p \)-distances specifies a wrong tree topology.

Although figure 2 indicates that the tree-making with \( p \)-distances is more often consistent than not, the use of \( p \)-distances as estimates of evolutionary distances is obviously a risky business.

**Efficiency of Tree-Making with \( p \)-distances and with “Unbiased” Distances**

During the last 10 years, a number of authors (Saitou and Nei 1986, 1987; Sourdis and Krimbas 1987; Sourdis and Nei 1988; Zharkikh and Li 1992, 1993; Schöniger and von Haeseler 1993; Tajima and Takezaki 1994) have remarked that under a “molecular clock” tree-making with \( p \)-distances is slightly more efficient than tree-making with “proper” distances corresponding to the true model. That is, several distance-based methods were studied in computer simulation and were found to recover the model tree topology slightly more often (a few percent) when “proper” distances (i.e., those corresponding to the mathematical model used to generate sequences) were substituted with \( p \)-distances. The extent of generality, utility, and practical implications of this observation are still unclear because the computer simulation was done for only a few arbitrarily chosen model trees, and a clear-cut procedure for selecting distance measure in practice was not suggested.

Our computer simulation confirmed that the distance-based methods tend to work better with \( p \)-distances than with “proper” distances under a “molecular clock.” Unlike the earlier studies (see references in the preceding paragraph), limited to a set of arbitrarily chosen model trees, we used a scheme of computer simulation for four-sequence trees (similar to those by Huelsenbeck and Hillis 1993) aiming at sampling the complete parameter space in a regular way (see figs. 3 and 4 for details). That is, we simulated evolution of four nucleotide sequences under the Jukes-Cantor model and a “molecular clock” (see model trees in fig. 3A and B), and estimated the probabilities, \( P_p \) and \( P_a \), of recovering the correct tree topology with \( p \)-distances and with “proper” estimates under the Jukes-Cantor model, respectively. Instead of using the correction formula by Jukes and Cantor (1969), we used Tajima’s (1993) “almost unbiased” estimator to avoid “infinite” distance estimates obtained when large observed \( p \)-distances (>\( 3/4 \)) are substituted into the Jukes-Cantor formula. As shown in figure 4, the estimated ratio \( P_d/P_p \) was always equal to or smaller than 1, indicating that \( p \)-distances allow for recovering the correct tree more frequently than “proper” distances (note that occasionally \( P_p \) was twice as large as \( P_d \); see fig. 4).

However, in the absence of a “molecular clock” tree-making with \( p \)-distances was positively misleading within the “inconsistency” zone, and it did not show any apparent advantages over tree-making with “unbiased” distances in the parameter area where both methods were consistent (see figs. 3C and 5). Although the ratio \( P_d/P_p \) is close to 1 within the “consistency zone” (the area below the dotted line in the upper left corner of the density plot in fig. 5), it rapidly increases to infinity in the vicinity of area where tree-making with \( p \)-distances is inconsistent (the upper left corner of the density plot in fig. 5). Figures 4 and 5 clearly indicate that the same phenomenon of “attraction of the long branches” (Penny, Foulds, and Hendy 1982) stands behind both the improved performance of tree-making with \( p \)-distances under a “molecular clock” and the pathological behavior of tree-making with \( p \)-distances in the absence of a “molecular clock.”

The conclusion that we have drawn from our computer simulations is that it is positively dangerous to use \( p \)-distances in the data analyses unless there is solid evidence that the sequences under analysis evolve in a “clock-like” fashion. In the following sections we shall attempt to give an intuitive explanation of the better performance of \( p \)-distances under a “molecular clock” and outline an approach for obtaining the prior distribution for tree-making.

**Analytic and Intuitive Explanation of Good Performance of \( p \)-Distances Under a “Molecular Clock”**

The correct four-sequence topology is selected by the majority of distance-based tree-making methods (including the neighbor joining method; Saitou and Nei 1987) if the following inequalities hold:
FIG. 2. Shape of the "consistency zone" associated with p-distances for a four-sequence unrooted tree. The white squares indicate the regions in the parameter space where the neighbor-joining method (Saitou and Nei 1987) is consistent in combination with p-distances; the black squares indicate the regions where tree-making from p-distances is inconsistent. The five-dimensional parameter space is defined by all possible values of branch lengths of the tree shown on the top of the figure; the parameter space is "sliced" into 27 two-dimensional projections (large squares). To understand the plot first consider the upper nine "slices" of the parameter space. All these "slices" correspond to a set of trees with the same (short) interior branch. The upper left corner corresponds to a symmetric tree with all five branches short and equal; the lower right corner represents a symmetric tree with long exterior branches and a short interior branch. The lower left and upper right corners of the nine-square group are almost completely black because they correspond to extremely asymmetric trees with short interior branches and one long exterior branch at each side (see schemes in the figure); the nature of inconsistency of the tree-making method in this case is usually explained by "attraction of the long branches" (Penny, Foulds, and Hendy 1992). It can be seen that the tree-making method works well for symmetric or almost symmetric trees but has difficulties with recovering asymmetric trees. The second and the third nine-slice groups are analogous to the first nine-slice group except that the interior branch becomes longer for the second group and is very long for trees in the third nine-slice group. As a result, performance of the tree-making method considerably improves for the second group; the method becomes "one hundred percent" consistent when the interior branch is very long—even extremely asymmetrical trees can be correctly recovered from p-distances when the sequences under analysis are sufficiently long. The values of the branch lengths of the model tree corresponding to each of the small squares are indicated on the axes in the upper left corner (0.0375, 0.375, and 0.75 for branches $h_i$, $h_2$, and $h_5$, and 0.0375 + 0.075 $\times k_i$ for branches $b_1$ and $b_2$, $k_i = 0, 1, 2, \ldots, 9$ for branches $b_1$ and $b_2$); all values are expressed in terms of the expected proportions of differences. The sequences were assumed to follow the Jukes-Cantor model of nucleotide substitution.

Alternatively, one can use the following approximate formula:

$$P_k = \frac{1}{\sqrt{\pi} X \Sigma X^T} \int_{-\infty}^{0} \int_{-\infty}^{0} \exp \left[ -\frac{1}{2} (X - X^T)(y - X^T) \right] / \sqrt{\pi} dy.$$  

where $k$ stands for either "p" or "d." Equation (9) is
FIG. 3.—Three model trees for four sequences used in computer simulation. Tree A is compatible with a “molecular clock” for any a, b, and c, whereas trees B and C are consistent with a “molecular clock” only when \( |a - c| \leq b \) and \( a = b \), respectively. Note that the expected branch lengths of the model trees are measured here in terms of the expected proportion of differences rather than in terms of the number of nucleotide substitutions per site.

The point that we are trying to make in this article is that it is dangerous to use iron arrows (p-distances) instead of bronze arrows (“proper” distances) unless prior information about the exact position of the magnet (a prior distribution of the model parameters) is available. In the following section we shall discuss an approach for estimating the prior distribution from the real data.

The Prior Distribution of Model Parameters and Tree Topologies

One can substantially improve one’s chances of recovering the correct tree and reduce computational expenses by using “prior” information about the relationships among taxa and the likely values of phylogenetic parameters. It is rather intuitive that different tree topologies and different combinations of parameter values may have very different likelihoods of being encountered in real phylogenetic analysis and therefore the vast majority of possible tree topologies and parameter values should not be considered in the real data analysis. For example, a professional systematist considers in his/her study only a tiny portion of all possible classifications of living creatures, because the vast majority of the possible dendrograms are totally incompatible with the total morphological and molecular evidence accumulated to date. Similarly, we know a sufficient number of sequences of genes and proteins to make at least approximate statements about the likelihoods of encountering various combinations of parameter values under any specified model.

Here we estimate the prior distribution of phylogenetic parameter values, starting with formulation of assumptions about the underlying model. Let \( T \) be a tree topology and \( \Theta \) be a set of values of the parameters, such as branch lengths of the tree and substitution model parameters. Let us consider an imaginary situation where the model tree is a random variable itself. (We define the model tree as union of a tree topology, tree branch lengths, and a set of parameters corresponding to the chosen model of nucleotide substitution. The model tree is used to generate a set of present-day sequences which are in turn used to reconstruct the phylogenetic tree. Note that tree topology is defined as an unweighted graph, i.e., it does not include branch lengths.) In our “quick-and-dirty” estimation we assume that (1) the tree topology, \( T \), and a set of parameter values, \( \Theta \), under a model of nucleotide substitution, are
randomly sampled from a multivariate prior distribution, $f(T, \Theta)$. (2) Distributions of $T$ and $\Theta$ are independent. (3) All possible tree topologies are equally likely, so that distribution of $T$ can be ignored in this estimation, (4) the actual genes evolve according to the Jukes-Cantor model, and (5) the expected exterior and interior branch lengths are independently sampled from two independent univariate distributions. Under these assumptions, the prior distribution of four-sequence unrooted trees is computed as follows:

$$f_T(\Theta) = f(b_1, b_2, b_3, b_4, b_5) = [f(b_1)]^4 \times f(b_5), (11)$$

where $b_1, b_2, b_3, b_4$, and $b_5$ are the expected branch lengths of the four-sequence tree shown in fig. 1C, and $f(b_1)$ and $f(b_5)$ are the prior distributions of the expected branch lengths.
lengths of the exterior and the interior branches, respectively.

Our estimate of the prior distribution of the "expected branch lengths," obtained under the above simple model, indicated that the vast majority of trees encountered in real data analysis is concentrated within an area that is very small compared to the complete parameter space (see fig. 7A, B, and C). This is because our estimates of $f(b_1)$ and $f(b_2)$ (see fig. 7A and B, respectively) are obviously quite different from uniform distributions, and this nonuniformity is preserved in the estimate of prior distribution (eq. 11). (We find it difficult to plot the estimated five-dimensional distribution of four-sequence trees, and decided to show here only an estimate of the prior distribution [see fig. 7C] of trees isomorphic to model tree C in fig. 3; the distribution was computed with eq. 11 and renormalized to sum to 1.)

An estimate of the prior distribution can be particularly useful for comparison of alternative tree-making methods, models of nucleotide substitution, and estimators of evolutionary distances. To give an example of application of our estimate for such purposes, we estimated the "unconditional" probabilities (table 1) of recovering the correct tree topology with $p$- and $d$-distances in the case of a four-sequence data set. The "unconditional" probabilities were computed by approximating the following expression.

$$\sum \text{prob}(\tilde{T}(x) = T | T, \Theta) f_T(\Theta) \text{prob}(T),$$

where $\text{prob}(\tilde{T}(x) = T | T, \Theta)$ is either probability $P_p$ or probability $P_d$ computed for specified $T$ and $\Theta$, and $\text{prob}(T)$ is a probability of encountering topology $T$. Summation is made over all possible values of $T$ and $\Theta$. The table shows that the differences in unconditional probabilities are rather trivial both in the absence of a "molecular clock" (tree C and "general" tree) and in its presence (trees A and B), and that the probability of encountering a data set such that tree-making with $p$-distances is inconsistent is close to 8%.

**Discussion**

Computer simulation and actual data analyses indicate that the rule of thumb in the real data analysis should be to avoid oversimplified models of nucleotide or amino acid substitution as much as possible when sequences under analysis are not very close, say, the maximum $p$-distance exceeds 0.1 (e.g., see Discussion in Jin and Nei 1990). Hopefully, in the near future an arbitrary choice of the substitution model in each particular case will be replaced with rigorous mathematical algorithms implemented in easy-to-use computer programs.

Distance-based tree-making with $p$-distances has much in common with tree-making with the maximum-parsimony method. Both methods have ill-defined "consistency zones" (Felsenstein 1978), and both seem to work well in certain regions of the parameter space. It would be interesting to estimate the probability of encountering an actual data set belonging to the "inconsistency zone" of the maximum-parsimony method; estimates of this kind might eventually help to resolve the notorious debate about the practical utility of the maximum-parsimony method.
Fig. 7.—Estimates of the “prior distributions” of the branch lengths. A. Frequency distribution of the exterior branch lengths. B. Frequency distribution of the interior branch lengths. These two distributions were estimated by reanalysis of phylogenetic trees published in Molecular Biology and Evolution and The Journal of Molecular Evolution in 1994 and the beginning of 1995 (Adell and Dopazo 1994; Clark, Maddison, and Kidwell 1994; Dopazo 1994; Goldman and Yang 1994; Jeffs, Holmes, and Ashburner 1994; Juan, Papeete, and Quintana 1994; Lopez et al. 1994; Pawlowski et al. 1994; Slade, Moritz, and Heidman 1994; Sullivan et al. 1994; Yang 1994a, 1994b; Escalante, Barrio, and Ayala 1995; Frye and Hedges 1995; Lee, Ota, and Vaquier 1995; Porter et al. 1995; Russo, Takezaki, and Nei 1995; Rzhetsky, Kumar, and Nei 1995; Smith, Smith, and Spratt 1995; Yang and Roberts 1995). Only those trees that had their branch length estimates corrected for multiple hits were considered in the analysis. To obtain distributions A and B we first decomposed each published tree into a set of four-sequence trees. Then, we recorded separately the lengths of the interior and the exterior branches of these four-sequence trees. Frequency distributions A and B were computed using 1,642 arc 882 data points, respectively. For the sake of convenience of presentation, the branch lengths were converted to the probabilities of observing different nucleotides at the ends of each branch under Jukes and Cantor’s model of nucleotide substitution. C. Probability of encountering a model tree with particular combination of the expected branch lengths, given that the tree topology and the ratios of the expected branch lengths are as shown in figure 3C. This conditional distribution was estimated using (1) frequency distributions A and B shown in this figure, and (2) the assumption that the expected branch lengths of the model tree are drawn randomly and independently from prior distributions of the expected interior and exterior branch lengths. The small area encapsulated by a solid line contains approximately 95% of the distribution. The branch lengths are measured in terms of p-distance, i.e., they were transformed into proportions of difference using the inverse of Jukes and Cantor’s (1969) formula.

The problem of choosing between “oversimplified” and “complex” mathematical models is obviously not unique for the distance-based methods (see Gaut and Lewis [1995] and Yang [1996] for discussion of this problem in context of the maximum-likelihood estimation). The maximum-likelihood framework allows for direct comparison of two rival evolutionary models in terms of goodness-of-fit to data, although the current implementations of this technique are very computationally intensive (see Goldman 1993a, 1993b; Yang, Goldman, and Friday 1994). We believe that resolution of the model-selection problem for “everyday” tree-making could be found through estimation of the prior distribution of phylogenetic parameters. That is, one can think about creating a database recording the “optimal” models for data sets representing particular taxonomic groups or particular protein families. The “database” solution might be more practical than choosing the “optimal” mathematical model in each instance of tree-making.

We also anticipate the importance of building a database of likelihoods (or “prior probabilities”) of alternative tree topologies. The optimal algorithm for com-
Table 1
(A) Probability of Recovering the Correct Tree Topology by a Distance-Based Method in Combination with p- and d-Distances and (B) the Portion of the Prior Distribution, Contained in the “Inconsistency Zone” Associated with Tree-Making with p-Distance

<table>
<thead>
<tr>
<th>MODEL TREE</th>
<th>d-distance</th>
<th>p-distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>n = 150</td>
<td>n = 500</td>
</tr>
<tr>
<td></td>
<td>0.864</td>
<td>0.956</td>
</tr>
<tr>
<td>B</td>
<td>0.933</td>
<td>0.981</td>
</tr>
<tr>
<td>C</td>
<td>0.942</td>
<td>0.991</td>
</tr>
<tr>
<td>“General”</td>
<td>0.884</td>
<td>0.952</td>
</tr>
</tbody>
</table>

Note.—The probability values given for model trees A, B, and C (fig. 3) are conditional, i.e., they were normalized by the prior probabilities of encountering each kind of model tree in data analysis.

computing such likelihoods still needs to be developed, and alternative approaches may range from ascribing to alternative trees “intuitive likelihoods” suggested by systematics experts, to computing the frequencies of occurrence of alternative tree topologies in tree-making from molecular data. Such database approaches can drastically change the everyday practice, reliability, and efficiency of tree-making as we know it today.

Acknowledgments

We are grateful to Masatoshi Nei for pointing out the problem and to Simon Tavaré for reference to J. Keilson’s book. Special thanks are due to Masatoshi Nei, Ziheng Yang, Spencer Muse, Andrey Zharkikh, and anonymous reviewers who provided helpful comments on earlier versions of this paper. This study was supported by grants from the National Science Foundation and National Institute of Health to Masatoshi Nei.

APPENDIX
Computing the Probabilities of Obtaining the Correct Four-Species Tree from the p-Distances and Jukes-Cantor Distances

Consider four present-day sequences, 1, 2, 3, and 4, of equal length n and without gaps, generated under the Jukes-Cantor model according to tree C in figure 1. Denote by \( k = [k_1, k_2, \ldots, k_{15}] \) the observed frequencies of the four-nucleotide patterns belonging to the following 15 categories

\[
\begin{align*}
\text{Category} & \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12 \quad 13 \quad 14 \quad 15 \\
\text{Sequence 1} & \quad x \quad x \quad x \quad y \quad x \quad x \quad x \quad x \quad x \quad x \quad y \quad y \quad y \quad x \\
\text{Sequence 2} & \quad x \quad x \quad x \quad x \quad y \quad y \quad x \quad y \quad y \quad x \quad x \quad z \quad y \\
\text{Sequence 3} & \quad x \quad x \quad x \quad x \quad y \quad y \quad y \quad x \quad x \quad z \quad z \quad x \quad z \\
\text{Sequence 4} & \quad x \quad y \quad x \quad x \quad y \quad x \quad y \quad z \quad x \quad z \quad x \quad z \quad x \quad x \quad h
\end{align*}
\]

(A1)

where \( x, y, z, \) and \( h \) are all different nucleotides. Clearly, vector \( k \) follows a multinomial distribution (e.g., see Saitou and Nei, 1986).

\[
\text{prob}\{k: \pi, n\} = \frac{n!}{\prod_{i=1}^{15} k_i!} \prod_{i=1}^{15} \pi_i^{k_i}, \quad (A2)
\]

where \( \Sigma_{i=1}^{15} k_i = n, \Sigma_{i=1}^{15} \pi_i = 1 \), and the entries of vector \( \pi \) can be easily computed. The observed counts of pairwise differences between four sequences can be expressed in terms of \( k_i \)'s in the following way.

\[
\mathbf{n} = \{n_1, n_2, n_3, n_4, n_5, n_6\} = \mathbf{M} \mathbf{k}, \quad (A3)
\]

where \( t \) denotes transpose of a vector and matrix \( \mathbf{M} \) is as follows:

\[
\mathbf{M} = \begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix} \quad (A4)
\]

The vector of proportions of differences between the four sequences is defined by

\[
\mathbf{p} = \{p_{12}, p_{13}, p_{14}, p_{23}, p_{24}, p_{34}\}' = \mathbf{n}/n, \quad (A5)
\]

where \( p_{ij} \) is the observed proportion of differences between the sequences \( i \) and \( j \). Therefore, the distribution of vector \( \mathbf{p} \) is as follows:

\[
\text{prob}\{\mathbf{p} = \mathbf{n}/n\} = \sum_{k_{15}=0}^{M_{15}} \sum_{k_{14}=0}^{M_{14}} \ldots \sum_{k_{9}=0}^{M_{9}} \left( \prod_{i=1}^{15} a_i! \right)^{15} \prod_{j=8}^{15} \prod_{j=8}^{15} \pi_j^{k_j}, \quad (A6)
\]

The values of \( a_i \)'s, and \( M_i \)'s are defined as follows:

\[
\begin{align*}
a_7 &= (-n_2 + n_3 + n_4 - n_5 + 2k_8 \\
- k_{10} + k_{11} + k_{12} - k_{13})/2, \\
a_6 &= (-n_1 + n_3 + n_4 - n_6 + 2k_8 \\
- k_9 + k_{11} + k_{12} - k_{14})/2, \\
a_5 &= (n_1 + n_2 - n_4 - 2k_8 - k_{11} \\
- 2k_{12} - k_{13} - k_{14} - k_{15})/2, \\
a_4 &= (n_1 - n_3 + n_5 - 2k_8 - k_{10} \\
- 2k_{11} - k_{12} - k_{14} - k_{15})/2, \\
a_3 &= (n_2 - n_3 + n_6 - 2k_8 - k_9 \\
- 2k_{11} - k_{12} - k_{13} - k_{15})/2, \\
a_2 &= (-n_4 + n_5 + n_6 - 2k_8 - k_9 \\
- k_{10} - k_{11} - 2k_{12} - k_{15})/2, \\
a_1 &= n - \sum_{i=2}^{15} a_i - \sum_{j=8}^{15} k_j,
\end{align*}
\]
Therefore, the probability of obtaining a particular vector of distance estimates we note that there is a one-to-one correspondence between vectors \( p \) and \( d \), where \( d \) is a vector of distance estimates obtained by substituting each entry of vector \( p \) into the Jukes-Cantor formula. Therefore, the probability of obtaining a particular vector \( d \) of the Jukes-Cantor distance estimates can be computed by equation (A6) where vector \( p \) is chosen appropriately. Since the explicit distributions of vectors \( p \) and \( d \) are known, it is easy to calculate the probability of recovering the correct tree topology. For example, to compute the probability \( P_p \) one needs to sum expression (A6) over all admissible values of \( p \) such that the following inequalities are satisfied.

\[
\begin{align*}
P_{12} - P_{13} - P_{24} + P_{34} &< 0, \\
P_{12} - P_{14} - P_{23} + P_{34} &< 0.
\end{align*}
\]  

To derive the joint distribution of the Jukes-Cantor distance estimates we note that there is a one-to-one correspondence between vectors \( p \) and \( d \), where \( d \) is a vector of distance estimates obtained by substituting each entry of vector \( p \) into the Jukes-Cantor formula. Therefore, the probability of obtaining a particular vector \( d \) of the Jukes-Cantor distance estimates can be computed by equation (A6) where vector \( p \) is chosen appropriately. Since the explicit distributions of vectors \( p \) and \( d \) are known, it is easy to calculate the probability of recovering the correct tree topology. For example, to compute the probability \( P_p \) one needs to sum expression (A6) over all admissible values of \( p \) such that the following inequalities are satisfied.

\[
\begin{align*}
P_{12} - P_{13} - P_{24} + P_{34} &< 0, \\
P_{12} - P_{14} - P_{23} + P_{34} &< 0.
\end{align*}
\]  

**LITERATURE CITED**


Takashi Gojobori, reviewing editor

Accepted August 7, 1996