Calculating Bootstrap Probabilities of Phylogeny Using Multilocus Sequence Data

Tae-Kun Seo

Professional Programme for Agricultural Bioinformatics, University of Tokyo, 1-1-1 Yayoi Bunkyo-Ku, Tokyo 113-8657, Japan

Phylogeny estimation is extremely crucial in the study of molecular evolution. The increase in the amount of available genomic data facilitates phylogeny estimation from multilocus sequence data. Although maximum likelihood and Bayesian methods are available for phylogeny reconstruction using multilocus sequence data, these methods require heavy computation, and their application is limited to the analysis of a moderate number of genes and taxa. Distance matrix methods present suitable alternatives for analyzing huge amounts of sequence data. However, the manner in which distance methods can be applied to multilocus sequence data remains unknown. Here, we suggest new procedures to estimate molecular phylogeny using multilocus sequence data and evaluate its significance in the framework of the distance method. We found that concatenation of the multilocus sequence data may result in incorrect phylogeny estimation with an extremely high bootstrap probability (BP), which is due to incorrect estimation of the distances and intentional ignorance of the intergene variations. Therefore, we suggest that the distance matrices for multilocus sequence data be estimated separately and these matrices be subsequently combined to reconstruct phylogeny instead of phylogeny reconstruction using concatenated sequence data. To calculate the BPs of the reconstructed phylogeny, we suggest that 2-stage bootstrap procedures be adopted; in this, genes are resampled followed by resampling of the sequence columns within the resampled genes. By resampling the genes during calculation of BPs, intergene variations are properly considered. Via simulation studies and empirical data analysis, we demonstrate that our 2-stage bootstrap procedures are more suitable than the conventional bootstrap procedure that is adopted after sequence concatenation.

Introduction

Reconstruction of phylogeny is one of the most important issues in the study of evolution. In the past when limited genomic information was available, molecular phylogeny was typically estimated using single locus data; however, due to the increase in the amount of available genomic data, it is currently being estimated using multilocus sequences (e.g., Adachi et al. 2000; Cao, Fujiwara, et al. 2000; Cao, Sorensen, et al. 2000; Nikaido et al. 2003; Nishihara et al. 2007). A simple approach to deal with multilocus sequence data is concatenation, that is, merging multilocus sequences and treating the merged set as a single locus. It has been noted that separate analysis is preferred over sequence concatenation (e.g., Adachi et al. 2000; Cao, Fujiwara, et al. 2000; Cao, Sorensen, et al. 2000; Nikaido et al. 2003; Nishihara et al. 2007). Because different genes may have been subject to different evolutionary processes, it would be appropriate to investigate their individual evolutionary histories. Evolutionary histories are typically summarized using parameters such as tree topology, nucleotide frequencies, branch lengths of phylogeny, etc. Sequence concatenation does not reflect the natural data partitioning; hence, the evolutionary parameters estimated from the concatenated sequences may not adequately represent the evolutionary implications of individual data partitions. Another problem concerning sequence concatenation occurs during the bootstrap procedure for measuring the uncertainty in the estimated phylogeny. The conventional bootstrap procedures (e.g., Felsenstein 1985; Kishino and Hasegawa 1989; Shimodaira and Hasegawa 1999) assume that the sequence columns are independently and identically distributed (i.i.d.). With this assumption, pseudodata are generated by resampling of the sequence columns from the concatenated sequences. However, it is inappropriate to assume an i.i.d. for all the concatenated sequence columns because different genes have been subject to different evolutionary processes.

Recently, Seo et al. (2005) suggested new procedures for determining and testing optimal tree topology using multilocus sequence data in the framework of the likelihood method. Instead of "combining (i.e., concatenating) sequence data," they suggested "combining the log-likelihood scores" obtained from the separate analysis. The approach of combining log-likelihood scores is conventional and has been adopted in many analyses (e.g., Adachi et al. 2000; Cao, Fujiwara, et al. 2000; Cao, Sorensen, et al. 2000; Nikaido et al. 2003; Nishihara et al. 2007). Calculation of phylogenetic uncertainty by conventional separate analysis involves 1) merging the pools of log-likelihood scores and applying the bootstrap procedure (Kishino and Hasegawa 1989; Shimodaira and Hasegawa 1999) or 2) stratifying the pools of log-likelihood scores and applying the bootstrap procedure within the strata (Yoder and Yang 2000). The latter approach is referred to as "stratified sampling." In contrast to the conventional bootstrap procedures after separate analysis, Seo et al. (2005) suggested 2-stage bootstrap procedures in which genes are resampled and the sequence columns are subsequently resampled from within the resampled genes. By resampling the genes, the intergene variations are appropriately considered while testing for the significance of tree topologies.

Although the bootstrap procedures for likelihood methods that were proposed by Seo et al. (2005) compare the competing tree topologies as a whole, they cannot quantify the significance of subclades within the phylogeny. Typically, the significance of subclades is represented by bootstrap probabilities (BPs) that are determined by parsimony and distance methods (Felsenstein 1985). However, the manner in which parsimony and distance methods can be applied to multilocus sequence data without sequence concatenation remains unknown. Bayesian approaches can be applied to multilocus sequence data without sequence concatenation (Rannala and Yang 1996; Ronquist
and Huelsenbeck 2003; Edwards et al. 2007). However, Bayesian methods generally require considerable computation, and their application is limited to a moderate number of taxa and sequences.

In our study, we suggest new distance methods for reconstructing the phylogeny and calculating the BPs using multilocus sequence data. By a simulation study, we demonstrate the advantage of separate analysis over sequence concatenation and the importance of considering intergene variations while measuring the BPs for the subclades. Further, we discuss the applicability of our procedures with the analysis of mammalian sequence data.

Methods

In this section, we begin by explaining the potential problem associated with sequence concatenation and the origin of the problem with reference to the framework of the distance method. Subsequently, we discuss 1) how to combine the individual results obtained using multilocus sequence data in order to determine optimal tree topology and 2) how to calculate BPs.

Potential Problem Associated with Sequence Concatenation

Recently, it was noted that heterotachy can result in incorrect tree estimation under the parsimony and likelihood framework (Kolaczkowski and Thornton 2004). When a single locus data set comprises multiple partitions having identical tree topologies but different branch lengths, the tree reconstructed using the entire single locus data set may be inconsistent with that of the individual partitions. The mechanism by which the problem of heterotachy occurs in single locus data set straightforwardly applies to concatenated sequence data. Due to the differences in the evolutionary processes, different genes possess different sets of branch lengths even when they have identical tree topologies. Here, we show that heterotachy is also a concern in the application of the distance method, and we investigate the origin of incorrect estimation of the phylogenetic tree.

Suppose, we measure the evolutionary distance between taxa A and B using sequence data from 2 loci, that is, loci 1 and 2. We denote the evolutionary distance between A and B at the loci 1 and 2 as \(d_{AB}^{(1)}\) and \(d_{AB}^{(2)}\), respectively. The evolutionary distance is measured in terms of the number of nucleotide substitutions per nucleotide site. To simplify the problem, we assume that both loci have the number of nucleotide substitutions per nucleotide site.

The evolutionary distance is measured in terms of the number of nucleotide substitutions per nucleotide site.

![Diagram of Bootstrap Probabilities of Phylogeny](https://academic.oup.com/mbe/article-abstract/25/5/960/1202339)

**FIG. 1.**—Underestimation by \(d_{AB}^{(\text{concat})}\); \(p_{AB}^{(i)}(i=1, 2)\) represents the proportion of the different sites between taxa A and B in the \(i\)th gene. The \(d_{AB}^{(1)}\) represents the evolutionary distance derived from \(p_{AB}^{(1)}\). Because of the concave relationship between \(p_{AB}^{(1)}\) and \(d_{AB}^{(1)}\), \(d_{AB}^{(\text{concat})}\) which is derived from \((p_{AB}^{(1)} + p_{AB}^{(2)})/2\) is always less than \(d_{AB}^{(\text{mean})} = (d_{AB}^{(1)} + d_{AB}^{(2)})/2\).

Statistically, we can assume that \(d_{AB}^{(1)}\) and \(d_{AB}^{(2)}\) are random samples from a certain unknown distribution. The mean of this distribution would be a good summary measure of the evolutionary distance between taxa A and B. The arithmetic mean of the 2 distances is the unbiased estimator for the mean of distance distribution:

\[
\frac{d_{AB}^{(\text{mean})}}{2} = \frac{d_{AB}^{(1)} + d_{AB}^{(2)}}{2}. \tag{2}
\]

On sequence concatenation of the 2 loci, the proportion of the different sites is \((p_{AB}^{(1)} + p_{AB}^{(2)})/2\). The JC evolutionary distance estimated from the concatenated sequences is as follows:

\[
d_{AB}^{(\text{concat})} = -\frac{3}{4} \log \left(1 - \frac{4}{3} \cdot \frac{p_{AB}^{(1)} + p_{AB}^{(2)}}{2}\right). \tag{3}
\]

Based on the concave relationship between \(p_{AB}^{(i)}\) and \(d_{AB}^{(i)}\) in equation (1), we can show that

\[
d_{AB}^{(\text{concat})} < d_{AB}^{(\text{mean})}.
\]

That is, \(d_{AB}^{(\text{concat})}\) is a “biased” estimator for the mean of distance distribution. The underestimation by \(d_{AB}^{(\text{concat})}\) is more severe when distantly related sequences are concatenated with closely related sequences. Figure 1 is the schematic representation of the underestimation by \(d_{AB}^{(\text{concat})}\).

Figure 2 shows an example in which the underestimation by the concatenation results in an incorrect tree estimation with perfect confidence. We compare the phylogenetic reconstruction of 4 taxa using the approaches specified in equations (2) and (3), followed by the Neighbor-Joining (NJ) algorithm (Saitou and Nei 1987). The 2 loci have identical and infinite sequence lengths. They evolve along
identical tree topologies but have different branch lengths. Because the sequence lengths are infinite, we can perfectly estimate the pairwise distances and reconstruct the true individual phylogenies of the 2 loci. In this case, the pairwise distances are consistent with the reconstructed tree for each gene. That is, the distance between the 2 taxa is equal to the sum of the lengths of all the branches connecting the 2 taxa. This consistency is maintained when we combine the 2 distances with equation (2), and the estimated tree topology is identical to each individual tree topology (refer to the right lower box of fig. 2). However, the consistency is not guaranteed on concatenation of the 2 loci, which may result in incorrect estimation of the phylogenetic tree (refer to the right upper box of fig. 2). Because the concatenated sequences are infinitely long, the resampled data are very similar to the original concatenated data in terms of the estimated distances, and we obtain 100% BP for the incorrect phylogeny.

Although the example in figure 2 may appear to be somewhat artificial due to the infinite sequence lengths, it clearly reveals the potential problem associated with sequence concatenation. It has been noted that the consistent phylogenetic signal of each gene, which is too weak to be detected individually, can be amplified and detected in the concatenated sequences (e.g., Queiroz and Gatesy 2006). This “hidden support” is occasionally taken into consideration during phylogeny reconstruction. However, the example in figure 2 reveals that the hidden support is not necessarily a real signal but may be an artifact. The clade (A, C) in the phylogenetic tree estimated from the concatenated sequences is not supported at all by the individual loci.

Determining the Optimal Tree by Separate Analysis

Because there is a potential problem associated with sequence concatenation, we suggest that sequence data be subjected to separate analysis and the results be combined for phylogeny reconstruction. For convenience of mathematical notation, we denote the pairwise distances in the vector form. Let the true distance vector of the $i$th gene be $d_i$. For given $m$ taxa, there exist $m(m-1)/2$ pairwise distances, and $d_i$ has $m(m-1)/2$ components. Based on the variability of the $d_i$s among genes, we assume that $d_i$ is a random variable whose mean and variance are denoted as follows:

$$E[d_i] = \Delta, \quad \text{Var}[d_i] = \Sigma.$$  \hspace{1cm} (4)

When the effects of the systematic biases—such as incomplete lineage sorting and horizontal gene transfer (Degnan and Rosenberg 2006)—are relatively minor, $\Delta$ would be consistent with the true species tree. Therefore, phylogeny reconstruction using $\Delta$ is quite reasonable (refer to Discussion).

For given $n$ genes, the unbiased estimators of $\Delta$ and $\Sigma$ are calculated as follows:

$$\hat{\Delta} : = \frac{1}{n} \sum_{i=1}^{n} d_i,$$

$$\hat{\Sigma} : = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{d}_i - \hat{\Delta}) (\hat{d}_i - \hat{\Delta})^T.$$

The vectors of $\hat{d}_i$s are not directly observed but estimated from the sequence data and are then denoted as $\tilde{d}_i$s. Thus,
we can consider the following more practical estimators of \( \Delta \) and \( \Sigma \):

\[
\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\delta}_i \\
\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{\delta}_i - \hat{\Delta})(\hat{\delta}_i - \hat{\Delta})^T.
\]

(5)

The phylogenetic tree can be estimated by applying the tree reconstruction algorithm \( T(\cdot) \) to \( \hat{\Delta} \):

\[
\hat{T} = T(\hat{\Delta}).
\]

(7)

Calculating BPs Using Multilocus Sequence Data

The variance of \( \hat{\Delta} \) can be calculated as follows:

\[
\text{Var}[\hat{\Delta}] = \frac{1}{n} \sum_{i=1}^{n} \text{Var}[\hat{\delta}_i] \approx \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[\hat{\delta}_i] \\
= \frac{1}{n} \hat{\Sigma} \approx \frac{1}{n^2} \Sigma
\]

(8)

where the first and the second approximations are valid when the sequence lengths and \( n \) are sufficiently large.

With regard to the nonnegative definite matrix \( \hat{\Sigma}/n \) in equation (8), we can determine a unique nonsingular matrix \( \hat{P} \) such that (Anderson 1984)

\[
\hat{P} \cdot \left( \frac{1}{n} \hat{\Sigma} \right) \cdot \hat{P}^T = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}
\]

where \( \hat{r} \) is the rank for \( \hat{\Sigma} \) and the estimate for the rank of \( \Sigma \). We note that \( \Sigma \) is not necessarily of full rank. With regard to the bifurcating phylogeny of \( m \) taxa, \( m(m-1)/2 \) elements of \( \delta_i \) are determined by the lengths of \( 2m-3 \) branches. When all the genes have identical tree topologies, the variability of \( m(m-1)/2 \) elements can be explained by the variability of \( 2m-3 \) branches, and the rank of \( \Sigma \) is \( 2m-3 \). In practice, the rank of \( \Sigma \) may differ from \( 2m-3 \) because each gene may not be perfectly consistent with the tree structure.

The purpose of obtaining \( \hat{P} \) from \( \hat{\Sigma}/n \) is to normalize \( \hat{\Delta} \); that is, when \( n \) and the sequence lengths are large:

\[
\text{E}[\hat{P}(\hat{\Delta} - \Delta)] \approx 0 \\
\text{Var}[\hat{P}(\hat{\Delta} - \Delta)] = \hat{P} \cdot \text{Var}[\hat{\Delta}] \cdot \hat{P}^T \approx \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}
\]

We adopt 2-stage bootstrap procedures (Seo et al. 2005) for quantifying the uncertainty of distance measure. In the first stage, we resample genes with replacement from the original data set, which will be denoted by the “\( \ast \)” sign. In the second stage, we resample the sequence columns with replacement within the resampled genes, which will be denoted by the “\( \ast \ast \)” sign. With regard to the resampled data set, we consider the following statistics, similar to those in equations (5) and (6):

\[
\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\delta}_i \\
\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{\delta}_i - \hat{\Delta})(\hat{\delta}_i - \hat{\Delta})^T.
\]

(6)

For large \( n \) and long sequence lengths, we can show that

\[
\text{E}[\hat{\Delta}^{\ast \ast \ast}] \approx \hat{\Delta} \\
\text{Var}[\hat{\Delta}^{\ast \ast \ast}] \approx \frac{1}{n} \hat{\Sigma}^{\ast \ast \ast}.
\]

With regard to the nonnegative definite matrix \( \hat{\Sigma}^{\ast \ast \ast}/n \), we can determine a unique nonsingular matrix \( \hat{P}^{\ast \ast \ast} \) (Anderson 1984) such that

\[
\hat{P}^{\ast \ast \ast} \cdot \left( \frac{1}{n} \hat{\Sigma}^{\ast \ast \ast} \right) \cdot \hat{P}^{\ast \ast \ast T} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},
\]

where \( \hat{r}^{\ast \ast \ast} \) is the rank for \( \hat{\Sigma}^{\ast \ast \ast} \). Here, \( \hat{P}^{\ast \ast \ast} \) is obtained in order to normalize \( \hat{\Delta}^{\ast \ast \ast} \). That is,

\[
\text{E}[\hat{P}^{\ast \ast \ast}(\hat{\Delta}^{\ast \ast \ast} - \hat{\Delta})] \approx 0 \\
\text{Var}[\hat{P}^{\ast \ast \ast}(\hat{\Delta}^{\ast \ast \ast} - \hat{\Delta})] = \hat{P}^{\ast \ast \ast} \cdot \text{Var}[\hat{\Delta}^{\ast \ast \ast}] \cdot \hat{P}^{\ast \ast \ast T} \approx \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.
\]

Following the 2 guidelines—centering and pivoting—for bootstrap hypothesis testing (Hall and Wilson 1991), we approximate the distribution of the normalized \( \hat{\Delta} \) by using the distribution of the normalized \( \hat{\Delta}^{\ast \ast \ast} \):

\[
\hat{P}(\hat{\Delta} - \Delta) \sim \hat{P}^{\ast \ast \ast}(\hat{\Delta}^{\ast \ast \ast} - \hat{\Delta}),
\]

(9)

where the “\( \sim \)” sign signifies that the distribution of the left term can be approximated with that of the right term.

In equation (9), \( \hat{\Delta} \) is an unknown parameter. If we adopt the Bayesian idea of Efron et al. (1996), we can consider the posterior probability of \( \Delta \). That is, with uniform prior of \( \Delta \), equation (9) leads to

\[
\Delta - \hat{\Delta} \sim \hat{P}^{-1}\hat{P}^{\ast \ast \ast}(\hat{\Delta}^{\ast \ast \ast} - \hat{\Delta}).
\]

(10)

With regard to the original data set, we calculate \( \hat{\Delta} \) and \( \hat{P} \) once. With regard to each resampled data set, we calculate \( \hat{\Delta} \) and \( \hat{P} \). Using iteration to generate the pseudodata set and calculate \( \hat{\Delta}^{\ast \ast \ast} \) and \( \hat{P}^{\ast \ast \ast} \), we obtain the posterior distribution of the unknown parameter \( \Delta \). The posterior probability of \( T(\Delta) \) can be used to calculate the BPs for the phylogenetic tree.

Simple Bootstrap Strategies

Equation (10) represents the most general bootstrap procedure in our study, hereafter referred to as B5. We consider 2 other bootstrap procedures—B3 and B4—that are
simpler than B5 and involve resampling of genes. The performance of the procedures B3–B5 will be compared with that of the conventional procedures—B1 and B2 (see Results and Discussion). The procedures B1–B4 are explained below. Whereas B1 is used with sequence concatenation, B2–B5 are used with separate analysis of equations (5) and (7).

1. B1: Multilocus sequences are concatenated for tree reconstruction, and the sequence columns are resampled with replacement from the concatenated data in order to calculate the BPs. The BPs are calculated in a manner similar to that used for the analysis of a single gene (Felsenstein 1985).

2. B2: To generate pseudodata, we resample the sequence columns within each gene without resampling the genes. That is, the BPs for the phylogenetic tree are calculated by iteration of \( T(\hat{A}) \). This “stratified sampling” has been used for likelihood methods (e.g., Yoder and Yang 2000). By comparing the results of B1 and B2, we can see the effects of separate analysis.

3. B3: Following the 2-stage bootstrap procedures (Seo et al. 2005), we resample the genes and sequence columns. However, we do not consider centering and pivoting of the bootstrap guidelines set forth by Hall and Wilson (1991). The BPs for the phylogenetic tree are calculated by iterating \( T(\hat{\Delta}^{**} \hat{P}^{**}) \). By comparing the results of B2 and B3, we can see the effects of the resampling genes.

4. B4: The Bayesian idea of Efron et al. (1996) is adopted to calculate the BPs. The centering scheme proposed by Hall and Wilson (1991) is considered; however, the pivoting scheme is ignored. Ignoring pivoting in equation (10) implies ignoring \( \hat{P}^{**} \) and \( \hat{P}^{**} \), leading to:

\[
\Delta - \hat{\Delta} - (\hat{\Delta}^{**} - \hat{\Delta}) = 2\hat{\Delta} - \hat{\Delta}^{**}.
\]

The BPs for the phylogenetic tree are calculated using the iteration of \( T(2\hat{\Delta} - \hat{\Delta}^{**}) \). By comparing the results of B3 and B4, we can see the effects of the centering scheme of bootstrap procedures. Further, by comparing the results of B4 and B5, we can see the effects of the pivoting scheme because centering and pivoting are both considered in B5.

Results
Simulation Studies

We performed simulation in order to compare the performance of the bootstrap procedures B1–B5. We used sequence concatenation and separate analysis as represented in equations (5) and (7) to reconstruct the phylogenies of randomly generated genes. When generating genes for the 4 taxa, we considered the variations in terms of the tree topology, branch lengths, and sequence lengths. We assumed that each gene evolved along either of the 2 tree topologies, namely, ((A,B),(C,D)) and ((A,C),(B,D)), which will hereafter be referred to as Trees 1 and 2, respectively. Each gene was determined to evolve along Tree 1 (or Tree 2) with the probability \( P \) (or \( 1 - P \)). The terminal and internal branch lengths of all genes were random samples obtained from the uniform distributions Unif[0, 0.2] and Unif[0, 0.02], respectively. The sequence length of each gene was a random sample obtained from the uniformly distributed integers between 501 and 3,500. In each gene, random DNA sequences of the 4 taxa were generated using the JC model (Jukes and Cantor 1969) after determination of the tree topology, branch lengths, and sequence lengths.

Once we generated \( n \) random genes, the pairwise distances for the concatenated sequences and separate genes were calculated using the JC model. For tree reconstruction, the NJ algorithm (Saitou and Nei 1987) was applied to the distance matrix of the concatenated sequences and to the average distance matrix of equation (5). To calculate the BPs, we generated 100 pseudosamples by following the procedures B1–B5. We calculated the BPs for 3 clades, that is, (A,B), (A,C), and (A,D), hereafter referred to as Nodes 1, 2, and 3, respectively.

For a given \( n \) and \( P \), the procedure for generating genes and calculating the BPs at Nodes 1–3 was repeated 1,000 times. Subsequently, we investigated the distribution of the BPs at the 3 Nodes. Node 3 is not supported by either Tree 1 or Tree 2. Nodes 1 and 2 compete with and are complimentary to each other, that is, a high BP at Node 1 implies a low BP at Node 2 and vice versa. Therefore, we will focus on the description of the results at Node 1, although we have shown the figures for all 3 Nodes. Using different values for \( n \) and \( P \), we consider the following 4 scenarios:

1. Scenario 1 (\( P = 0.6, n = 5 \)). In this Scenario, we consider a situation in which the number of multilocus genes is relatively small (\( n = 5 \)) and one of the 2 trees is favored. Because \( P > 0.5 \), Tree 1 is favored, and it is reasonable to expect high BPs at Node 1. When the procedures B1 and B2 are used, some data sets exhibit very high BPs (\( \geq 95\% \)) at Node 1; however, other data sets exhibit very low BPs (\( \leq 5\% \)) at Node 1 (fig. 3). A low BP at Node 1 is inconsistent with the fact that Tree 1 is favored (\( P = 0.6 \)). The difference between the calculated values of the BPs for different data sets is more exaggerated when the B1 procedure is used. Of the 1,000 iterations in the B1 procedure, 199 data sets display \( \leq 5\% \) BP at Node 1. Although extremely low BPs are observed with B3 and B4, their proportions out of 1,000 iterations are smaller than those of B1 and B2. In contrast to B1–B4, B5 displays neither extremely high nor extremely low BPs at Node 1 (see Discussion).

2. Scenario 2 (\( P = 0.6, n = 100 \)). As in Scenario 1, Tree 1 is favored (\( P = 0.6 \)); however, the number of multilocus genes is relatively large (\( n = 100 \)). In figure 4, the histograms of the BPs for B1–B5 at Node 1 are concentrated in the higher range, which is consistent with the fact that Tree 1 is favored (\( P = 0.6 \)). It is notable that some data sets (36 out of 1,000) for B1 exhibit very low BPs (\( \leq 5\% \)) at Node 1, despite the proportion being small. Because the intergene variations are taken into consideration, B3–B5 have a more scattered distribution than B1 and B2.

3. Scenario 3 (\( P = 0.5, n = 5 \)). The number of multilocus genes is relatively small (\( n = 5 \), and both trees are
equally favored ($P = 0.5$). When the procedures B1 and B2 are used, a majority of the data sets display extremely high or low BPs at Node 1 (fig. 5). However, as in the result of Scenario 1, the proportions of the extreme BPs for B3 and B4 are smaller than those of B1 and B2. When B5 is used, the distribution of BPs at Node 1 is similar to that observed in Scenario 1, and most of the data sets are concentrated in the moderate range of BPs.

4. Scenario 4 ($P = 0.5$, $n = 100$). As in Scenario 3, both trees are equally favored ($P = 0.5$). However, the number of multilocus genes is relatively large ($n = 100$). When the procedures B1 and B2 are used, a majority of the data sets display extremely high or low BPs at Node 1 (fig. 6). When the procedures B3–B5 are used, the distribution of BPs at Node 1 is scattered between 0 and 1 and is similar to uniform distribution (see Discussion).

Empirical Data Analysis

To compare the performance of the procedures B1–B5, we analyzed the sequence data published by Nishihara et al. (2007). They collected a sequence data set comprising 2,789 protein-coding data partitions from 10 mammalian species and investigated their tree topologies and significance. They considered 3 candidate trees—Trees 1, 2, and 3 (fig. 7)—representing different relationships among Boreotheria, Xenarthra, and Afrotheria. Using the maximum likelihood (ML) method and the nucleotide substitution model, they observed that the concatenation approach yields the monoclade of Xenarthra and Afrotheria (Tree 3 of fig. 7) with a 100% BP (see similar result by Hallstrom et al. [2007]). However, by separate analysis, Tree 1 was observed to be the ML tree. Although the BPs of Tree 1 were generally higher than those of Tree 3, the BPs of Trees 1 and 3 varied considerably according to the differences in the evolutionary models and the categorization of the sequence data. Therefore, they noted that Tree 3 cannot be rejected against Tree 1, and a greater amount of sequence data are needed to resolve the phylogenetic relationship.

The total length of 2,789 data partitions is 1,011,870 nt. This implies that on an average, 1 data partition has approximately 395 nt. We found that short sequences can create a problem in our distance method because distance estimates have large variances (see Discussion). Therefore, from among the 2,789 data partitions, we selected partitions whose lengths are moderately large. In this data set, there are 508 and 103 data partitions that are longer than 500 and 1,000 nt, respectively, hereafter referred to as data sets A and B, respectively. The total lengths of the data sets A and B are 422,937 and 146,958 nt, respectively. For each data partition within the data sets A and B, we calculated the
distances using the Tamura–Nei (TN; Tamura and Nei 1993) model with a gamma distribution of the rate heterogeneity among sites (Yang 1993, 1994b).

We adopted the value of 0.375 as the gamma rate heterogeneity parameter and calculated the TN distances using the formula of Nei and Kumar (2000, p. 44; for justification of the value of 0.375, see Discussion). Reconstruction of NJ trees with sequence concatenation of the data sets A and B yielded Tree 3 as shown in figure 7. Reconstruction of NJ trees with separate analysis using equations (5) and (7) for the data sets A and B yielded Tree 1. Table 1 shows the BPs for the data sets A and B obtained using the 5 bootstrap procedures at Nodes 1–9 of figure 7.

Discussion
New Distance Methods Using Multilocus Sequence Data

Here, we suggest new procedures for determining the optimal tree topology and calculating BPs using the distance method and multilocus sequence data. For calculating BPs using the B4 and B5 approaches, we adopt the Bayesian idea of Efron et al. (1996). Therefore, the BPs calculated using the procedures B4 and B5 can be regarded as the posterior probabilities for the subclades. The advantage of adopting the Bayesian idea is that interpretation of the BP is rendered easy and intuitive. The interpretation of a conventional BP is “repeatability” (Felsenstein 1985).

The pseudodata repeatedly generated around the original data represent information regarding the true data-generating mechanism for the original data. This interpretation is rather complicated and is not intuitive. By adopting the Bayesian idea, BP can be interpreted as the posterior probability that will enable the determination of the existence of a subclade.

The advantage of our procedures over the conventional Bayesian methods (e.g., Rannala and Yang 1996; Ronquist and Huelsenbeck 2003) is that they require considerably less computation and can deal with a larger sequence data set. The application of the conventional Bayesian methods is limited to a moderate number of genes and taxa because it requires heavy computation. Our procedures could be used as alternatives to the conventional Bayesian methods when large amounts of sequence data are analyzed.

The Importance of Considering Intergene Variations

Based on the results of figure 3, it is evident that the problem associated with concatenation (B1) can be very serious when the number of genes is small. When \( n \) is small, stochastic errors have a greater influence on the sampled genes, and all genes may be incidentally sampled from a tree topology that does not represent the true phylogenetic relationship. In such situations, B1 after sequence
concatenation and B2 of stratified sampling may yield high BPs for incorrect subclades. Considering the existence of intergene variations, more care should be exercised during confidently formulating a conclusion based on a small number of genes. Even when all the sampled genes support identical tree topologies, if there are considerable intergene variations in terms of the distance, we reasonably expect that there may be other genes that support different topologies but are incidentally not selected. This expectation prevents us from formulating a confident decision until we obtain sufficient information, that is, a large number of genes. As a result, the BPs calculated using B3–B5 are distributed mostly in the intermediate range when \( n \) is small (fig. 3) and are concentrated in the higher range at Node 1 when \( n \) is large (fig. 4). When there is little variation or consistency among genes, the procedures B3–B5 can yield confident results even with a small number of genes. In all Scenarios, Node 3 is supported by neither Tree 1 nor Tree 2, which is consistent among all genes. Therefore, B3–B5 yield low BPs at Node 3 even with a small number of genes (figs. 3 and 5).

The poor performance of B1 after concatenation and that of B2 with stratified sampling is very clear in Scenario 4 (fig. 6). The number of genes is large (\( n = 100 \)). Because \( P = 0.5 \), Trees 1 and 2 are equally probable. However, when B1 and B2 are used, a majority of the 1,000 iterations exhibit extremely high (\( \geq 95\% \)) or low (\( \leq 5\% \)) BPs at Node 1. Because intergene variations are considered, a more flat and widely scattered distribution is obtained using B3–B5. The distributions of BPs calculated using B3–B5 are similar to the uniform distribution between 0 and 1. This is against our intuitive expectation that BPs would be concentrated around 0.5 at Node 1. A similar problem was noted under the Bayesian framework (Suzuki et al. 2002; Lewis et al. 2005). When the multifurcating node is true, the posterior probabilities of a bifurcating node are uniformly distributed between 0 and 1 (Lewis et al. 2005). The situation of \( P = 0.5 \) in Scenario 4 is somewhat similar to the situation of the multifurcating node that Lewis et al. (2005) considered. When \( P = 0.5 \), the true \( \Delta \) is located at the boundary between the space of Trees 1 and 2. The distance estimates \( \hat{\Delta} \) are symmetrically distributed around the boundary. If \( \hat{\Delta} \) is incidentally located in the space of Tree 1, a majority of the resampled estimates \( \hat{\Delta} \)'s will be located in the space of Tree 1, resulting in a high BP at Node 1. If \( \hat{\Delta} \) is incidentally far away from the boundary, the BP at Node 1 can be extremely high. Similarly, if \( \hat{\Delta} \) is incidentally located in the space of Tree 2, the BP at Node 1 can be low, even extremely low. This intuitively explains why the distributions of the BPs at Node 1 calculated using B3–B5 converge to a uniform distribution between 0 and 1 instead of converging to a point mass at 0.5. If the \( \hat{\Delta} \)'s are located exactly at the boundary between the space of Trees 1 and 2, the \( \hat{\Delta} \)'s should be symmetrically distributed around the boundary, and the distributions of BPs should converge to a point mass at 0.5. However, this does not happen at all. However, large the amount of data, the \( \hat{\Delta} \)'s are symmetrically distributed around the boundary and are not located exactly at it.
In Scenario 3, the BPs at Node 1 calculated using B5 are more concentrated around 0.5, whereas the BPs at Node 1 calculated using B3–B4 are rather uniformly distributed between 0 and 1. The deviation of B5 from the uniform distribution can be explained on the basis of poor approximation for the variance. The approximation of equation (8) is valid when the sequence lengths and number of genes are sufficiently large. Otherwise, $\hat{P}$ and $\hat{P}_q$ are not reliable approximates; therefore, B3 and B4 should be preferred over B5. When the number of genes is large, the distribution of the BP calculated using B5 is similar to the distributions of the BPs calculated using B3 and B4 as shown in Scenarios 2 and 4. This implies that the effects of centering and pivoting are relatively minor when the data set is large. Therefore,

\[ \text{FIG. 6.—Scenario 4 (P = 0.5, n = 100): Refer to the legend of figure 3.} \]

\[ \text{FIG. 7.—The 3 major trees of 10 mammalian species. The Nodes 1–6 are common among the 3 trees. The Nodes 7, 8, and 9 are consistent with the Trees 1, 2, and 3 respectively. Group A, B, and C represent Boreotheria, Xenarthra, and Afrotheria, respectively.} \]
we mainly focused on the procedure using weighting scheme between these 2 extremes. In our study, it is considered. It is possible to create an intermediate sequence length is completely ignored and the other, in which 2 extremes: one, in which the information regarding sequence heterogeneity among sites during the calculation of rate heterogeneity among sites may be another possibility.

Therefore, it is appropriate to take into consideration the rate heterogeneity among sites during the calculation of the pairwise distances. It is very important to consider rate heterogeneity among sites during tree reconstruction and ignoring it has been noted to possibly result in the reconstruction of an incorrect phylogeny (e.g., Yang 1995).

The rate heterogeneity among sites is typically modeled with a gamma distribution whose mean and variance are 1 and $1/\lambda$, respectively (Yang 1993). In principle, it is not possible to estimate $\lambda$ by the distance method, and we applied the ML method to determine the appropriate $\lambda$ for calculating the pairwise distances. PAML (Yang 2007) was used to estimate $\lambda$ using the TN model (Tamura and Nei 1993) for the concatenated sequences of the 2,789 data partitions. The estimated $\lambda$s for Trees 1, 2, and 3 of figure 7 are 0.375, 0.375, and 0.374, respectively. Similarly, $\lambda$s for the concatenated sequences of the data sets A and B were estimated, and the estimated $\lambda$s for Trees 1, 2, and 3 of the data set A are 0.363, 0.363, and 0.362, respectively, and those of the data set B are 0.371, 0.372, and 0.370, respectively. The value of 0.375 that we adopted as an $\alpha$ parameter for calculating the pairwise distances is qualitatively similar to the $\lambda$s estimated from the concatenated data partitions. When $\alpha = 0.375$ is considered during the calculation of the pairwise distances, the rodent group is properly positioned within the Boreothere group. The NJ trees for 2,789 data partitions reconstructed using the TN model with sequence concatenation and with separate analysis were observed to be Trees 3 and 1, respectively. This result is consistent with that of Nishihara et al. (2007). They applied the ML method with the General Time Reversible (Yang 1994a) model, and the ML trees reconstructed by sequence concatenation and by separate analysis were observed to be Trees 3 and 1, respectively.

Variations among the Data Partitions

Because different data partitions have been subject to different evolutionary mechanisms, the optimal $\alpha$ for different data partitions will be different. However, we did not consider this variation in calculating the pairwise distances and applied $\alpha = 0.375$ to all data partitions. It is possible to estimate $\lambda$s for separate data partitions by the ML method and to apply the separate $\lambda$s for calculating the pairwise distances for separate data partitions. However, this approach

### Table 1

<table>
<thead>
<tr>
<th>Bootstrap Procedures</th>
<th>Partial Data Sets</th>
<th>Node 1</th>
<th>Node 2</th>
<th>Node 3</th>
<th>Node 4</th>
<th>Node 5</th>
<th>Node 6</th>
<th>Node 7</th>
<th>Node 8</th>
<th>Node 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>Dataset A</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>Dataset A</td>
<td>99.9</td>
<td>99.9</td>
<td>100.0</td>
<td>98.3</td>
<td>98.9</td>
<td>97.2</td>
<td>47.8</td>
<td>3.5</td>
<td>48.5</td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>99.4</td>
<td>100.0</td>
<td>100.0</td>
<td>80.0</td>
<td>5.7</td>
<td>14.3</td>
</tr>
<tr>
<td>B3</td>
<td>Dataset A</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>97.2</td>
<td>99.9</td>
<td>98.6</td>
<td>70.4</td>
<td>3.2</td>
<td>26.4</td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>93.9</td>
<td>100.0</td>
<td>99.9</td>
<td>71.1</td>
<td>13.0</td>
<td>15.9</td>
</tr>
<tr>
<td>B4</td>
<td>Dataset A</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>95.7</td>
<td>99.1</td>
<td>95.7</td>
<td>43.3</td>
<td>7.2</td>
<td>49.3</td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>90.6</td>
<td>100.0</td>
<td>98.8</td>
<td>61.4</td>
<td>13.7</td>
<td>24.9</td>
</tr>
<tr>
<td>B5</td>
<td>Dataset A</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>99.8</td>
<td>100.0</td>
<td>99.9</td>
<td>42.4</td>
<td>0.6</td>
<td>57.0</td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>94.2</td>
<td>100.0</td>
<td>100.0</td>
<td>64.0</td>
<td>13.2</td>
<td>22.8</td>
</tr>
</tbody>
</table>

Note.—Out of the 2,789 data partitions, we selected only those partitions whose lengths are larger than 500 and 1,000 nt, referred to as data sets A and B, respectively. The number of data partitions in data set A (data set B) is 508 (103).
would be impractical when a huge amount of data need to be analyzed, and the ML method also requires a lot of computational time. Our general recommendation is to apply the ML method to concatenated sequences to obtain the appropriate value for $\hat{x}$; this is relatively easy even when a huge amount of sequence data need to be analyzed. As shown in our results, the $\hat{z}$ estimates appear to be quite robust for different trees and different lengths of concatenated data.

We note that a substantial portion of the 2,789 data partitions comprises relatively short sequences. For example, there are 645 (1,991) data partitions whose lengths are smaller than 200 (400) nt. In our distance method, we found that short sequences could cause a problem because distance estimates have a large variance. In the analysis using the ML method conducted by Nishihara et al. (2007), they adopted the resampling estimated log-likelihood (RELL; Kishino and Hasegawa 1989) procedure. That is, instead of resampling the sequence data and recalculating the likelihood scores for the resampled data, they resampled the likelihood scores obtained from the original sequence data. When the sequences are short, the RELL procedure may underestimate the variance because it ignores the variance of the estimation conducted using resampled data. In our distance method, pairwise distances are recalculated for resampled sequence data sets. Because the distance estimates from short sequences are subject to large variations, BPs are also subject to large variations. Although Tree 1 represents the NJ tree reconstructed using the 2,789 data partitions and $\hat{\Delta}$ of equation (5), the BPs at Nodes 1–6 obtained by the 2-stage bootstrap procedures are relatively low (data not shown). Because of the large variance in the distance estimates, the distance matrices from the resampled data are distributed over the space of many trees, resulting in relatively low BPs for a specific clade. To reduce the problem caused by short sequences, we arbitrarily set the threshold value as 500 nt (for data set A) and 1,000 nt (for data set B) for selection of data partitions. With moderately long data partitions, the NJ trees were identical to those obtained from all 2,789 data partitions and the BPs at Nodes 1–6 were high (table 1).

Problem Associated with Sequence Concatenation

The phylogenetic relationship among Boreotheria, Xenarthra, and Afrotheria remains controversial, and Nishihara et al. (2007) noted that a greater amount of sequence data and a more appropriate taxon sampling are required to resolve the phylogenetic relationship among these 3 taxa. Nishihara et al. criticized the sequence concatenation approach because it presents a 100% BP for the clade of Xenarthra and Afrotheria (Node 9 of fig. 7). In our analysis, we obtained results that are consistent with those of Nishihara et al. When the number of data partitions is moderate (data set B), the BP at Node 9 calculated using B1 is not extreme; however, when the number of data partitions is large (data set A), it is 100.0%. This leads us to arrive at an incorrect conclusion, which is also demonstrated in our simulation study. For data sets A and B, the BPs at Nodes 7 and 9 calculated using B2–B5 were moderately high, which is consistent with the results of Nishihara et al.’s ML method. Whereas there is a drastic difference between B1 and B2–B5 (table 1), the differences among B2–B5 are relatively minor. However, this does not generally imply that B2 is as acceptable as B3–B5. It appears that the effect of separate analysis is dominant over the effect of the 2-stage bootstrap procedures in this example involving mammalian data. As shown in the simulation study, there is a potential problem associated with B2 because it ignores the intergene variations. Therefore, we emphasize that B3–B5 are generally preferable over B1 and B2.

Gene Tree versus Species Tree

In our procedures, we assume that $\Delta$ is consistent with the species tree. Therefore, the species tree can be estimated using $\hat{\Delta}$—the arithmetic mean of $\hat{\delta}$s. The gene trees reconstructed using $\hat{\delta}$s are distributed around the species tree, and we assume that systematic bias has a minor effect on the distribution of $\hat{\delta}$s.

Recently, it was noted that gene trees are not necessarily consistent with the species tree due to various evolutionary events, such as incomplete lineage sorting and horizontal gene transfer (Degnan and Rosenberg 2006). In such situations, the species tree cannot be estimated using a consensus gene tree. Our tree reconstruction method using equations (5) and (7) does not consider systematic bias caused by incomplete lineage sorting and horizontal gene transfer. However, it is notable that incomplete lineage sorting is generally not a major concern when the species are moderately distant (e.g., Degnan and Rosenberg 2006). Moreover, only a small portion of the genomic loci may be subject to horizontal gene transfer during the evolution of higher organisms. Tree estimation using equations (5) and (7) is suitable for situations in which the effect of systematic bias is minor, and the consensus gene tree is consistent with the species tree.

Concluding Remarks

Here, we introduce new procedures for estimating phylogeny and calculating BPs using multilocus sequence data in the context of the distance method. By theoretical investigation, simulation studies, and empirical data analysis, we demonstrated that there is a potential problem associated with the sequence concatenation and that intergene variations should be considered during the measurement of phylogenetic uncertainty. Although B5 is the most sophisticated procedure, its application appears to be limited to data sets comprising a large number of genes and long sequences. Therefore, B3 and B4 would be preferable in most circumstances. Because our distance methods require considerably less computation, they are good alternatives to the Bayesian and likelihood methods when huge amounts of sequence data need to be analyzed.

Acknowledgments

The author thanks Jeffrey L. Thorne, Asger Hobolth, Sang Chul Choi, Hirohisa Kishino, Marcy Uyenoyama, and
2 anonymous reviewers for their valuable comments and Hidenori Nishihara for providing data. The author was supported by Japan Society for the Promotion of Science (EYS B-18770216).

Literature Cited


Marcy Uyenoyama, Associate Editor

Accepted February 7, 2008