Two simple methods have been proposed by Kishino et al. for estimating bootstrap probabilities of alternative trees from a maximum-likelihood analysis without performing maximum-likelihood estimation for each resampled data set; i.e., a resampling estimated log-likelihood method and a multivariate normal distribution method. To examine the extent to which these two methods provide good approximations, bootstrap probabilities estimated by these methods from real data were compared with those estimated by repeated bootstrap resampling and maximum-likelihood estimation. It turned out that both of these simple methods are good approximations to the computationally intensive bootstrap method. This finding should motivate people in this field to use the maximum-likelihood method when inferring molecular evolutionary trees.

Introduction

A maximum-likelihood (ML) method for inferring molecular evolutionary trees from DNA sequence data was originally developed by Felsenstein (1981). Later, Kishino et al. (1990) and Adachi and Hasegawa (1992) developed an ML method for analyzing protein sequence data. These methods do not assume rate constancy among lineages and hence are robust against departures from constancy (Hasegawa et al. 1991; Hasegawa and Fujiwara 1993).

In a series of papers, we have developed a method to evaluate the confidence limits of the ML tree. Hasegawa and Kishino (1989) noted that, since the likelihood is dependent on a particular realization of a random variable, it is important to estimate the variance of the log of the likelihood ratio between alternative trees. They estimated the variance by using bootstrap resampling (Felsenstein 1985). Kishino and Hasegawa (1989) developed a simple method for estimating the variance by expressing it explicitly.

Rather than this sort of pairwise comparison between alternative trees, a bootstrap probability—i.e., a frequency of a particular tree being the highest-likelihood tree among alternatives during bootstrap resampling—might be more appealing to molecular evolutionists.

Key words: approximate bootstrap probability, maximum likelihood, molecular evolutionary tree.

Address for correspondence and reprints: Masami Hasegawa, The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan.

evolutionary tree. The ML method infers the evolutionary tree from these sequences.

The data can be represented as follows: species 1—\(X_{11}, X_{12}, X_{13}, \ldots, X_{1n}\); species 2—\(X_{21}, X_{22}, X_{23}, \ldots, X_{2n}\); \(\ldots\); species \(k—X_{k1}, X_{k2}, X_{k3}, \ldots, X_{kn}\), where \(X_{ij}\) is either of T, C, A, or G for DNA sequences and is either of the 20 amino acids for protein sequences. We write the whole data \((X_{ij})\) as \(X\) and the value of the \(h\)th site \((X_{1h}, X_{2h}, \ldots, X_{kh})^T\) (a superscript “T” denotes a transposed vector) as \(X_h\).

Under the assumption of independent identical distributions (i.i.d.) among sites, the likelihood is given by

\[
L(\theta | X) = \prod_{h=1}^{n} f(X_h | \theta),
\]

where \(f(x | \theta) = f(x_1, x_2, \ldots, x_k | \theta)\) is the probability that species 1 has the base or amino acid \(x_1\), species 2 has \(x_2\), \ldots, and species \(k\) has \(x_k\) at a homologous site. The vector \(\theta\) denotes the unknown parameters, such as the branching dates and the substitution rates, along the respective branches of the tree.

The likelihood of each tree is computed by altering \(\theta\) in an optimal fashion. Then, the fit of the tree to the data is compared with those of the alternative trees by the criterion of likelihood. The tree with the highest likelihood is chosen as the most likely candidate for the true tree.

The log-likelihoods of \(m\) alternative trees are represented by

\[
l_{(i)}(\hat{\theta}(i) | X) = \sum_{h=1}^{n} \log f(X_h | \hat{\theta}(i)),
\]

where each term of the right-hand side follows i.i.d. When \(\theta\) is replaced by the ML estimate, \(\hat{\theta}\), each term of the right-hand side of

\[
l_{(i)}(\hat{\theta}(i) | X) = \sum_{h=1}^{n} \log f(X_h | \hat{\theta}(i)),
\]

no longer follows i.i.d. However, when \(n\) is large, the distribution of eq. (3) coincides asymptotically with that of eq. (2). Therefore, the estimated log likelihoods

\[(l_{(1)}, l_{(2)}, \ldots, l_{(m)})\]

asymptotically follow an MND whose mean and variance-covariance can be estimated by

\[
l_{(i)}(\hat{\theta}(i) | X),
\]

and

\[
\frac{n}{n-1} \sum_{h=1}^{n} \left\{ \log f(X_h | \hat{\theta}(i)) - \frac{1}{n} \sum_{h'=1}^{n} \log f(X_h | \hat{\theta}(i)) \right\}
\]

\[
\times \left\{ \log f(X_h | \hat{\theta}(j)) - \frac{1}{n} \sum_{h'=1}^{n} \log f(X_h | \hat{\theta}(j)) \right\}
\]

(Kishino and Hasegawa 1989), respectively.

The MND method estimates a bootstrap probability that a particular tree is selected as the best tree from comparison among components of a random number that follows the MND presented above (Kishino et al. 1990).

The bootstrap probabilities can be estimated also by the RELL method, which resamples the estimated log likelihoods of sites as follows:

\[
l_{(i)}^{(B)} = \sum_{h=1}^{n} \log f(X_h^{(B)} | \hat{\theta}(i)),
\]

where \(X_h^{(B)}\) refers to a site resampled by bootstrap (Schork and Schork 1989; Kishino et al. 1990). Either method can give the bootstrap probabilities of candidate trees without performing ML estimation for each resampled datum.

Comparison of the RELL and MND Methods with the Standard Bootstrap Method

To examine the extent to which the RELL and MND methods provide good approximations, bootstrap probabilities estimated by these methods were compared with those estimated by the standard bootstrap method. The data used in this study are the third-codon positions of the 3' end of the NADH-dehydrogenase subunit 4 (ND4 gene) and of the 5' end of the ND5 gene from Hominoidea (human, chimpanzee, gorilla, and orangutan) (Brown et al. 1982). The data consist of 232 nucleotides and have been analyzed elsewhere (Hasegawa and Kishino 1989; Kishino and Hasegawa 1989).

Three tree topologies are possible among human, chimpanzee, and gorilla, when orangutan is taken as an outgroup to these three species (Hasegawa and Kishino 1989; Kishino and Hasegawa 1989): tree 1—(((human, chimp), gorilla), orangutan); tree 2—(((human, gorilla), chimp), orangutan); and tree 3—(((chimp, gorilla), human), orangutan). The log likelihood for each
tree topology was computed by DNAML of Felsenstein’s (1988) PHYLIP program package (version 3.2). The transition/transversion parameter (=26) was chosen so as to maximize the likelihood. Bootstrap resampling and ML estimation were repeated 5,000 times, and, for each resampling of sites, equation (4) was also applied to estimate bootstrap probabilities of the three alternative trees by the RELL method. Bootstrap probabilities were also estimated by the MND method.

Estimated log-likelihoods of trees 1, 2, and 3 ($l_1$, $l_2$, and $l_3$) are -661.52, -663.16, and -665.22, respectively, and tree 1 is the highest-likelihood tree. Figure 1 shows distributions of $l_2 - l_1$ and $l_3 - l_1$ estimated by the bootstrap and RELL methods. It is apparent that the simple method of RELL gives nearly the same result as the cumbersome method of bootstrap.

Table 1 represents bootstrap probabilities estimated by the bootstrap method as well as by the RELL and MND methods. It turns out that the RELL and MND methods give almost identical estimates of bootstrap probabilities and that these estimates approximately coincide with the result obtained by repeated bootstrap resampling and ML estimation. This result suggests that the RELL and MND methods are useful as good approximations for estimating the bootstrap probabilities.

### Discussion

The present study demonstrates that the simple RELL and MND methods give bootstrap probabilities that approximate well the result obtained by the standard bootstrap method. Once an ML analysis has been done, the CPU time for the RELL method is proportional to the length of the sequences, while that of the MND method is independent of it. Hence, for long sequences, the MND method is more rapid than the RELL I method. However, the CPU time for the RELL method, as well as for the MND method, is negligible compared with that in estimating the log-likelihood for each site. Because of the simplicity of the RELL method in its application, the method has been implemented in a PROTML program for an ML analysis of protein sequences (Adachi and Hasegawa 1992).

The RELL method would be useful not only in molecular phylogenetic analyses, but also in various fields of applied statistics in which bootstrap resampling and ML estimation must be repeated. This method is related to the parsimony test of Templeton (1983).

The ML method has a disadvantage in that it requires a large computational burden when compared with other widely used methods of molecular phylogenetics. However, this disadvantage might be compensated to some extent by the introduction of the simple methods, discussed in the present paper, for estimating the bootstrap probability of an ML tree.

### Acknowledgments

We are grateful to M. Fujiwara and J. Adachi for help with the computer. We also thank J. Felsenstein for information and J. Reeves and the two reviewers for helpful comments on an earlier version of the manuscript. This work was carried out under the Institute of Statistical Mathematics Cooperative Research Program (91-ISM·CRP-69 and 92-ISM·CRP-81) and was supported by grants from the Ministry of Education, Science, and Culture of Japan.
LITERATURE CITED


TAKASHI GOJOBORI, reviewing editor

Received October 6, 1992

Accepted August 12, 1993