Letter to the Editor

Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference

Hidetoshi Shimodaira and Masami Hasegawa

Institute of Statistical Mathematics, Tokyo, Japan

The maximum-likelihood method for inferring molecular phylogeny (Felsenstein 1981) is being widely used. The probabilistic model for generating the molecular sequences is specified by the substitution process and the tree topology. The parameters for the substitution process and the branch lengths are estimated by maximizing the likelihood, and then the tree topology is estimated by maximizing the maximized likelihood. To obtain the confidence limit of the topology, the test of Kishino and Hasegawa (1989), referred to as the KH test, is often used in practice. The same idea that is the basis for the KH test is also found in the statistical literature (Linhart 1988; Vuong 1989). The KH test was designed for comparing two topologies but is often used for comparing many topologies. This use of the KH test leads to overconfidence for a wrong tree, because the sampling error due to the selection of the topology is overlooked in it. In this note, we present a modification of the KH test to take into account a multiplicity of testings.

Let a index the topologies and $L_a$ be the maximum log-likelihood under the probabilistic model specified by the topology a. We have as candidates $M$ topologies, labeled 1, 2, ..., $M$. The KH test is a normal approximation test which can be used to compare $L_a$ with $L_b$ of another prespecified topology b. In fact, $(L_a - L_b)/S_{ab}$ is asymptotically distributed as normal with unit variance under certain conditions, where $S_{ab}$ is an estimate of the variance of $L_a - L_b$. However, the KH test is often used to compare $L_a$ with $L_b$ 5 max$(L_{a1}, ..., L_{aM})$, where $a$ is the maximum-likelihood topology. We have to consider the effect of selection of $a$ to derive the distribution of $L_a - L_b$.

An application of multiple-comparison techniques to statistical model selection is illustrated by Shimodaira (1993, 1998), and the procedure described below is regarded as a modified KH test which automatically corrects the selection bias.

Step 1. Calculate the test statistic $T_a = 5 \max(L_1, L_2, ..., L_M) - L_a$. The replicates $\tilde{L}_a$, $\tilde{L}_b$, ... are stored in an $M \times N$ array. For resampling $L_{ai}$, the RELM method and the normal approximation method of Kishino, Miyata, and Hasegawa (1990) are computationally useful.

Step 2. Generate $N$ bootstrap replicates of vector $(L_{a1}, ... , L_{aM})$. The replicates $\tilde{L}_{ai}$, $\tilde{L}_{aj}$ with $i \neq j$ are regarded as a modified KH test which automatically corrects the selection bias.

Step 3. Subtract the average of each row from the entries of the array. Now, we have the array of $\tilde{R}_{ai} = \tilde{L}_{ai} - 2N^{2/3}S_{ai1}^{-1}L_{ai}$. This step is called “centering,” and $\tilde{R}_{ai}$ is regarded as a replicate of $L_a$ generated under the least favorable configuration (f.f.c.), explained later.

Step 4. For each column ($i = 1, ..., N$) of the array, calculate $S_{ai} = 5 \max(\tilde{R}_{ai}/2 \tilde{R}_{ai}, ..., \tilde{R}_{ai}/2 \tilde{R}_{ai})$ and replace the entries with these values. $S_{ai}$ is a replicate of $T_a$.

Step 5. For each row ($a = 1, ..., M$) of the array, count the number of entries which exceeded $T_a$, then calculate the $p$-value, defined by $P_a$ 5 (number of $\{S_{ai} > T_a\})/N$. $P_a$ is denoted by $T$ and is regarded as a confidence limit of the topology.

The above procedure reduces to a resampled version of the KH test if $M = 2$. It also reduces to the KH test to compare the topologies with $a$ if we replace $\tilde{S}_{ai}$ with $\tilde{R}_{ai}/2\tilde{R}_{ai}$ in step 4. Our definition of $T_a$ takes into account the possibility that any of $b = 1, ..., M$ could be the maximum-likelihood topology.

Let $E(L_a)$ be the expected value of $L_a$ with respect to the true model specified by the true substitution process and the true topology. We are interested in finding the best topology $a^*$ among the candidates which maximizes the expected log-likelihood; $E(L_a)$ 5 $\max(E(L_{a1}), ..., E(L_{aM}))$. The coverage probability, denoted by $P_C$, is the probability of $a^*$ being included in $T$. The distinctive property of our $T$ is that

$$
P_C \geq 1 - 2 \frac{P}{a^*}$$

holds approximately when the sequence length and $N$ are sufficiently large. If there are several a’s for which $E(L_{a^*}) < E(L_{a_m})$, inequality (1) holds for each of them.

The confidence limit is derived from the following hypothesis test. Let $H_a$ be the null hypothesis that $E(L_a)$ 5 $E(L_{a^*})$, i.e., $a$ is the best topology. $H_a$ will be rejected if $T_a$ is large, and the distribution of $T_a$ under $H_a$ is needed for calculating the $p$-value as the upper probability. This test is treated in the above procedure, and $P_a$ is in fact the $p$-value of the test of $H_a$. $T$ consists of $a$ for which $H_a$ is not rejected, and thus $1 - 2P_C$ is equivalent to the probability of rejecting $H_a$. This probability is controlled by $P_{a^*}$, and (1) follows. Several remarks on our method are given below:

1. The distribution function of $T_a$ depends on the true model. To control the error probability in testing $H_a$, we consider all the possible values of $(E(L_{a1}), ..., E(L_{aM}))$ to find the maximum probability that $T_a$ is greater than a critical constant. Maximum is attained at the l.f.c., i.e., $E(L_{a1}) < 5 \cdots 5 E(L_{aM})$, which

Key words: confidence limit of tree topology, Kishino-Hasegawa test, mammalian phylogeny, multiple comparisons, multiplicity of testings, selection bias.

Address for correspondence and reprints: Hidetoshi Shimodaira, The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan. E-mail: shimo@ism.ac.jp.

© 1999 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038
4. The procedure is valid, and the coverage probability
3. Our method is different from another type of testing
2. It follows from the nonnegativity of the Kullback-

(1) holds even if we replace
approach because of the misspecification of the sub-
models (topologies) are often rejected in the latter
the correct model. Since the model is specified by both
(1962). The former tests which model is better than
of nonnested models, such as that described in Cox
among the candidates and that the substitution pro-
tution process is correctly specified. Therefore, $a^*$ is
is attributed to this misspecification for long
nating extremely unlikely topologies from the set of
candidates. All possible topologies are not necessarily-
to be included among the candidates when we are
interested in biological hypotheses represented by
their typical topologies.

6. Our procedure to test $H_2$ is conditioned on the shape of the joint distribution function of $L_a \geq L_b$, $b \leq 1$, ...
$M$ except for the means, and we did not consider the
effect of this approximation here. A detailed ana-
ysis is given in Shimodaira (1997) for the case in
which $M \leq 2$.

As an example of an application of our method,
mitochondrial protein sequences of 3,414 amino acids
for six mammal species were analyzed and the results
are shown in table 1. These data are a subset of the data
used in Waddell et al. (1999). The program AAML in
PAML (Yang 1997) is used to calculate the sitewise log-
likelihoods for each topology, and the RELL method
of Kishino, Miyata, and Hasegawa (1990) is used to resam-
ple $L_{a,i}$. The mtREV model (Adachi and Hasegawa
1996) is used for amino acid substitutions, and site het-
erogeneity is modeled by the discrete gamma distribu-
tion (Yang 1996). The clade (P, B) is significantly sup-
ported in a preliminary analysis, and thus only the 15
 bifurcating tree topologies containing this clade are con-
sidered here.

According to table 1, the bootstrap selection prob-
ability (BP) and the KH test suggest that the confidence
limit consists of the best three trees at $P^B = 0.1$. On the
other hand, the multiple-comparisons (MC) method sug-
gests that the confidence limit consists of the best seven
trees at the same level, and the multiple comparisons of
the standardized statistics (MS) method includes tree 10
also. The MC and MS methods appear to be conserva-
tive, and indeed they are so. The BP and KH lead to

---

Table 1  
Log-Likelihood Differences and $p$-Values for the 15 Bifurcating Topologies of the Mammal Data Set

<table>
<thead>
<tr>
<th></th>
<th>$L_a$</th>
<th>$L_b$</th>
<th>BP</th>
<th>KH</th>
<th>MC</th>
<th>MS</th>
<th>Topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.583</td>
<td>0.640</td>
<td>0.941</td>
<td>0.944</td>
<td>(((H, (P, B)), O), M, D)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.7</td>
<td>0.317</td>
<td>0.360</td>
<td>0.811</td>
<td>0.805</td>
<td>(((H, (P, B)), O), M, D)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7.4</td>
<td>0.038</td>
<td>0.121</td>
<td>0.577</td>
<td>0.422</td>
<td>((H, O), (P, B)), M, D)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>17.6</td>
<td>0.012</td>
<td>0.040</td>
<td>0.169</td>
<td>0.203</td>
<td>((H, (P, B)), O, M), D)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>18.9</td>
<td>0.030</td>
<td>0.066</td>
<td>0.139</td>
<td>0.296</td>
<td>((H, (P, B)), (O, M), D)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>20.1</td>
<td>0.006</td>
<td>0.050</td>
<td>0.109</td>
<td>0.100</td>
<td>((H, (P, B)), O), M, D)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>20.6</td>
<td>0.011</td>
<td>0.048</td>
<td>0.107</td>
<td>0.248</td>
<td>((H, (O, M)), (P, B), D)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>22.2</td>
<td>0.001</td>
<td>0.032</td>
<td>0.070</td>
<td>0.048</td>
<td>((H, M), (P, B), O), D)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>25.4</td>
<td>0.000</td>
<td>0.001</td>
<td>0.029</td>
<td>0.013</td>
<td>(((H, (P, B)), O), M, D)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>26.3</td>
<td>0.002</td>
<td>0.018</td>
<td>0.032</td>
<td>0.124</td>
<td>((H, (O, M)), (P, B), D)</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>28.9</td>
<td>0.000</td>
<td>0.008</td>
<td>0.017</td>
<td>0.069</td>
<td>((H, (O), M), (P, B), D)</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>31.6</td>
<td>0.000</td>
<td>0.003</td>
<td>0.006</td>
<td>0.032</td>
<td>((H, (M), (P, B)), O, D)</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>31.7</td>
<td>0.000</td>
<td>0.003</td>
<td>0.006</td>
<td>0.035</td>
<td>((H, (P, B), M), O, D)</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>34.7</td>
<td>0.000</td>
<td>0.001</td>
<td>0.002</td>
<td>0.012</td>
<td>((H, O), (P, B), M), D)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>36.2</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.007</td>
<td>((H, (P, B), M), O, D)</td>
<td></td>
</tr>
</tbody>
</table>

Note.—BP is the bootstrap selection probability of Felsenstein (1985) estimated by the RELL method (Kishino,
Miyata, and Hasegawa 1990). KH is the $p$-value of the KH test, MC is the $p$-value of the multiple-comparisons method
with $w_{1,i} \leq 1$, and MS is that with $w_{1,i} \leq 5$?z. See Remark 4 in text. The number of replicates is $N = 5 \times 10^4$. The labels
for the taxa are as follows: H 5 Homo sapiens (human), P 5 Pongo pygmaeus (orangutan), B 5 Bos taurus (cow), O 5
Oryctolagus cuniculus (rabbit), M 5 Mus musculus (mouse), and D 5 Didelphis virginiana (opossum).
smaller confidence limits and seem to be preferable, but they are not guaranteed to satisfy inequality (1).

Although both the MC and the MS methods satisfy (1) at least approximately, the confidence limit suggested by MC is smaller than that for MS in this example. It is not clear yet what kind of weight matrix produces smaller confidence limits in practical data analyses of phylogeny inference.

From a molecular phylogenetic analysis, Graur, Duret, and Gouy (1996) strongly suggested that Lagomorpha (rabbit) is closer to Primates than to Rodentia, contrary to the traditional view of Glires (Lagomorpha Rodentia) (e.g., Novacek 1992). However, Halanych (1996) criticized Graur, Duret, and Gouy’s analysis and demonstrated that the Primates-Lagomorpha grouping is not preferred if the complexity of the problem is taken into account. Our result is consistent with Halanych in that trees 4, 5, and 7 with the Lagomorpha/Rodentia clade cannot be dismissed. The BP and KH give smaller estimates of the confidence limits and can give overconfidence for a wrong tree in terms of (1). Overconfidence can be given also by using a wrong model for the substitution process (e.g., Hasegawa and Adachi 1996). Together with the improvement of the models for substitutions, the improvement of the method for estimating the confidence limit given in this paper should be important in not giving overconfidence to a wrong tree.

Acknowledgments

The authors thank Joe Felsenstein, H. Kishino, and the reviewers for helpful discussions and comments.

LITERATURE CITED


