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 of each $a = 1, \ldots, M$ with L_b of another prespecified topology b. In fact, $(L_b - L_a)/\hat{s}_{a\cdot b}$ is asymptotically distributed as normal with unit variance under certain conditions, where $\hat{s}^2_{a\cdot b}$ is an estimate of the variance of $L_b - L_a$. However, the KH test is often used to compare $L_{\hat{a}}$ with $L_{\hat{a}} = \max\{L_1, \ldots, L_M\}$, where $\hat{a}$ is the maximum-likelihood topology. We have to consider the effect of selection of $\hat{a}$ to derive the distribution of $L_{\hat{a}} - L_a$.

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 $T_a = \max\{L_1 - L_a, \ldots, L_M - L_a\} = L_{\hat{a}} - L_a$ for $a = 1, \ldots, M$.

Step 2. Generate N bootstrap replicates of vector $(L_1, \ldots, L_M)$. The replicates $\tilde{L}_{a\cdot i}$, $a = 1, \ldots, M$, $i = 1, \ldots, N$ are stored in an $M \times N$ array. For resampling $\tilde{L}_{a\cdot i}$, the RELL method and the normal approximation method of Kishino, Miyata, and Hasegawa (1990) are computationally useful.

Step 3. Subtract the average of each row from the entries of the array. Now, we have the array of $\tilde{R}_{a\cdot i} = \tilde{L}_{a\cdot i} - N^{-1}\sum_{j=1}^{N}\tilde{L}_{a\cdot j}$. This step is called "centering," and $\tilde{R}_{a\cdot i}$ is regarded as a replicate of L_a generated under the least favorable configuration (l.f.c.), explained later.

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

Step 5. For each row ($a = 1, \ldots, M$) of the array, count the number of entries which exceeded T_a, then calculate the p-value, defined by $P_a =$ (number of $\{\tilde{S}_{a\cdot i} \geq T_a\})/N$.

Step 6. Compare P_a with a prespecified significance level P^*. The set consisting of topologies with $P_a \geq P^*$ is denoted by $\mathcal{T}$ 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 $\hat{a}$ if we replace $\tilde{S}_{a\cdot i}$ with $\tilde{R}_{a\cdot i} - \tilde{R}_{a\cdot i}$ in step 4. Our definition of $\tilde{S}_{a\cdot i}$ takes into account the possibility that any of $b = 1, \ldots, M$ 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; $E(L_{a^*}) = \max\{E(L_1), \ldots, E(L_M)\}$. The coverage probability, denoted by P_C, is the probability of a^* being included in $\mathcal{T}$. The distinctive property of our $\mathcal{T}$ is that

$$P_C \geq 1 - P^* \qquad (1)$$

holds approximately when the sequence length and N are sufficiently large. If there are several a's for which $E(L_a) = E(L_{a^*})$, 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 $E(L_a) = E(L_{a^*})$, 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 is done in the above procedure, and P_a is in fact the p-value of the test of H_a. $\mathcal{T}$ consists of a for which H_a is not rejected, and thus $1 - P_C$ is equivalent to the probability of rejecting H_{a^*}. This probability is controlled by P^*, 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 $(E(L_1), \ldots, E(L_M))$ to find the maximum probability that T_a is greater than a critical constant. This maximum is attained at the l.f.c., i.e., $E(L_1) = \cdots = E(L_M)$, 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 106-8659, Japan. E-mail: shimo@ism.ac.jp.

Mol. Biol. Evol. 16(8):1114–1116. 1999

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

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

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 $w_{a \cdot b} = 1$, and MS is that with $w_{a \cdot b} = \hat{s}_{a \cdot b}^{-2}$. See Remark 4 in text. The number of replicates is $N = 10^4$. The labels for the taxa are as follows: H = *Homo sapiens* (human), P = *Phoca vitulina* (harbor seal), B = *Bos taurus* (cow), O = *Oryctolagus cuniculus* (rabbit), M = *Mus musculus* (mouse), and D = *Didelphis virginiana* (opossum).

is approximated by the centering in step 3. This l.f.c. usually does not correspond to a tree topology, but to a mixture of models specified by topologies by concatenating the sequences generated under several topologies. This may be an artifact, but it is a good representation of the misspecification of the substitution process. The uncertainty of topology selection is often attributed to this misspecification for long sequence length.

2. It follows from the nonnegativity of the Kullback-Leibler relative entropy that the correct topology maximizes the expected log-likelihood if the substitution process is correctly specified. Therefore, $\hat{a}^*$ is the correct topology provided that it is included among the candidates and that the substitution process is not terribly wrong.
3. Our method is different from another type of testing of nonnested models, such as that described in Cox (1962). The former tests which model is better than the other, while the object of the latter is to find the correct model. Since the model is specified by both the substitution process and the topology, all of the models (topologies) are often rejected in the latter approach because of the misspecification of the substitution process.
4. The procedure is valid, and the coverage probability (1) holds even if we replace T_a with $T_a = \max_{b \neq a} w_{a \cdot b}(L_b - L_a)$ in step 1 and $\tilde{S}_{a \cdot i}$ with $\tilde{S}_{a \cdot i} = \max_{b \neq a} w_{a \cdot b}(\tilde{R}_{b \cdot i} - \tilde{R}_{a \cdot i})$ in step 4, where $w_{a \cdot b}$ is a prespecified weight matrix. Shimodaira (1998) used $w_{a \cdot b} = \hat{s}_{a \cdot b}^{-2}$ to standardize $L_b - L_a$. In the KH test, the weight matrix is specified after the selection, since $w_{a \cdot b} = 0$ except for $b = \hat{a}$.
5. $P_C = 1 - P^*$ holds only at the l.f.c., and $P_C \geq 1 - P^*$ in general. This can lead to an unnecessarily large $\mathcal{T}$ if many topologies are compared simultaneously. We should make M as small as possible by eliminating 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_a is conditioned on the shape of the joint distribution function of $L_b - L_a$, $b = 1, \ldots, M$ except for the means, and we did not consider the effect of this approximation here. A detailed analysis is given in Shimodaira (1997) for the case in which $M = 2$.

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 resample $\tilde{L}_{a \cdot i}$. The mtREV model (Adachi and Hasegawa 1996) is used for amino acid substitutions, and site heterogeneity is modeled by the discrete gamma distribution (Yang 1996). The clade (P, B) is significantly supported in a preliminary analysis, and thus only the 15 bifurcating tree topologies containing this clade are considered here.

According to table 1, the bootstrap selection probability (BP) and the KH test suggest that the confidence limit consists of the best three trees at $P^* = 0.1$. On the other hand, the multiple-comparisons (MC) method suggests 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 conservative, and indeed they are so. The BP and KH lead to

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 1 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

ADACHI, J., and M. HASEGAWA. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

COX, D. R. 1962. Further results on tests of separate families of hypotheses. J. R. Stat. Soc. B **24**:406–424.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

GRAUR, D., L. DURET, and M. GOUY. 1996. Phylogenetic position of the order Lagomorpha (rabbits, hares and allies). Nature **379**:333–335.

HALANYCH, K. 1996. Testing hypotheses of chaetognath origins: long branches revealed by 18S ribosomal DNA. Syst. Biol. **45**:223–246.

HASEGAWA, M., and J. ADACHI. 1996. Phylogenetic position of cetaceans relative to artiodactyls: reanalysis of mitochondrial and nuclear sequences. Mol. Biol. Evol. **13**:710–717.

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. **29**:170–179.

KISHINO, H., T. MIYATA, and M. HASEGAWA. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **30**:151–160.

LINHART, H. 1988. A test whether two AIC's differ significantly. S. Afr. Stat. J. **22**:153–161.

NOVACEK, M. 1992. Mammalian phylogeny: shaking the tree. Nature **356**:121–125.

SHIMODAIRA, H. 1993. A model search technique based on confidence set and map of models. Proc. Inst. Stat. Math. **41**:131–147 [in Japanese].

———. 1997. Assessing the error probability of the model selection test. Ann. Inst. Stat. Math. **49**:395–410.

———. 1998. An application of multiple comparison techniques to model selection. Ann. Inst. Stat. Math. **50**:1–13.

VUONG, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica **57**:307–333.

WADDELL, P., Y. CAO, J. HAUF, and M. HASEGAWA. 1999. Using novel phylogenetic methods to evaluate mammalian mtDNA, including AA invariant sites-LogDet plus site stripping, to detect internal conflicts in the data, with special reference to the position of hedgehog, armadillo, and elephant. Syst. Biol. **48**:31–53.

YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. **11**:367–372.

———. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555–556.

NARUYA SAITOU reviewing editor

Accepted April 28, 1999