Ann. Inst. Statist. Math.
Vol. 52, No. 1, 43–56 (2000)

COMPARING THE LIKELIHOOD FUNCTIONS OF PHYLOGENETIC TREES

A. BAR-HEN[1]* AND H. KISHINO[2]

[1] *The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku,
Tokyo 106-8569, Japan*
[2] *Department of Social and International Studies, University of Tokyo,
3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan*

(Received March 29, 1996; revised August 10, 1998)

Abstract. DNA sequence data provide a good source of information on the evolutionary history of organisms. Among the proposed methods, the maximum likelihood methods require an explicit probabilistic model of nucleotide substitution that makes the assumption clear. However, procedures for testing hypotheses on topologies have not been well developed. We propose a revised version of the maximum likelihood estimator of a tree and derive some of its properties. Then we present tests to compare given trees and to derive the most likely candidates for the true topology, applying to maximum likelihoods the notion of contrast, as defined in the framework of the analysis of variance, and the procedures used in multiple comparison. Finally, an example is presented.

Key words and phrases: Maximum likelihood, multiple comparisons, phylogeny.

1. Introduction

Since the time of Darwin, biologists have attempted to determine evolutionary relationships among various species. These relationships can be illustrated in a phylogenetic tree that shows the course of evolution among a group of organisms. Originally, phylogenetic relationships were inferred from morphological data. Species were considered closely related if they were physically similar, or if analogous parts of the organisms functioned similarly. Methods based on morphological data tend to be limited to the study of groups of organisms that are somewhat closely related. It is difficult to compare the physical characteristics of a human being and a bacterium.

The advent of nucleic acid sequencing has revolutionized the field of phylogenetic inference. Nucleotid acid sequences (DNA and RNA) from any organisms can be aligned, providing a basis for the construction of phylogenies for organisms over a wide range. The problem of inferring phylogenetic relations among a group of organisms using nucleotide sequence data is one of continuing interest to researchers in the field of molecular evolution. There are a variety of approaches to the problem in current use, see Swofford and Olsen (1990) for a review.

If the occurrence of evolutionary change is modelled with stochastic process, statistical methods can be used to infer a phylogeny. The oldest statistical techniques are based on maximum likelihood. A review can be found in Felsenstein (1983).

* Now at Laboratoire de statistique et probabilites, Université de Lille I Bat. M2, 59655 Villeneuve d'Ascq Cedex, France.

Such models have been in use for some time now but interest in them heightened following the revelation by Felsenstein (1978) that the popular parsimonious criterion can cause serious biases when the rates of evolutionary change in the true phylogenetic tree differ greatly from one branch to another.

To apply this method, one regards a tree topology as a statistical model containing several unknown parameters, such as branch lengths and evolutionary rates. The maximum likelihood value of each possible topology is computed by altering the unknown parameters in an optimum fashion. Then the fit of the tree topologies to the data is compared by the criterion of maximum likelihood. The model with the highest likelihood is chosen as the most likely candidate for the true topology. The use of statistical models fitted by maximum likelihood is currently considered as the best method of inferring phylogenies (see, e.g. Felsenstein (1981), Barry and Hartigan (1987), and Navidi *et al.* (1993)).

Nevertheless, a crucial problem is the reliability of a proposed phylogenetic tree. By considering the possible trees as different hypotheses, this problem is closely connected to the problem of hypothesis testing. Since the statistical models expressing different tree topologies are not nested, the usual theory of likelihood ratio test does not apply directly. Cox (1961) studied testing separate families of hypothesis based on the distributions of the likelihood ratio statistics under the hypothesis. Partly because this procedure is a pairwise comparison procedure, and partly because there are lots of trees (hypothesis) to be considered, it is difficult to answer satisfactorily to the requirement such as testing the significance of a group of trees against others or testing if a set of organisms are in a family.

Recent work showed the validity of the interior-branch test (IBT) in the framework of distance methods (Sitnikova *et al.* (1995)). Distance methods estimate the branch lengths from the converted distances between sequence pairs by least square procedure. Therefore, once we can assume that the converted distances follow multivariate normal distribution, the estimated branch lengths also follow normal distributions. IBT tests the zero length of an interior-branch with the alternative hypothesis of positive length. In the framework of maximum likelihood procedure, however, since the likelihood can only be defined for non-negative branch lengths, the likelihood ratio test cannot be applied in an ordinary way to the one sided test.

Churchill *et al.* (1992) and Navidi *et al.* (1993) consider procedures likelihood ratio test and linear invariant, to evaluate the topologies. As a likelihood approach, it considers the ratio of the maximum log likelihood of a topology with wide classes of evolutionary processes and the full model multinomial distribution. Goldman (1993) considered the parametric bootstrap to get the distribution of the log likelihood ratio statistics under the hypothesis, because of the large degree of freedom.

Because of the large degree of freedom again, the power may not be high. If several topologies are not rejected in the hypothesis test, next step is to compare the topologies, which is the aim of this article. Navidi *et al.* (1993) also uses the above log likelihood ratio statistics to get the confidence region of the parameters. The procedure is valid only when the topology is correct. If we take account of the uncertainty with regard to topologies, the threshold value of the log likelihood ratio is not the one from the χ^2 distribution with the specified degree of freedom.

Felsenstein (1985) proposed nonparametric bootstrap procedure to set the reliability on the trees, calculating proportions of trees selected from analyses of pseudo samples. Kishino and Hasegawa (1989) approximated the set of log likelihood ratios by (multivariate) normal distribution, estimating the variances and covariances from the values of the sites. Although they are used by many researchers in molecular phylogeny, the statistical property of the procedure has not been fully studied, and is still under argument (Zharkikh and Li (1992*a*, 1992*b*), Hillis and Bull (1993)). Felsenstein and Kishino (1993) pointed out that the bootstrap proportions can be used as a conservative significance level in the frame work of hypothesis testing, or as elements composing confidence sets.

In this paper, we provide an inferential scheme to decide about the reliability of proposed trees, using the theory of multiple comparisons. We derive some of the properties of a revised version of the maximum likelihood and derive a test of homogeneity for the likelihood of the tree. Then we derive tests to compare given trees and to derive the most likely candidates for the true topology. Finally an example is presented.

2. Maximum likelihood of phylogenetic trees

Most frequently observed events of molecular evolution are nucleotide substitutions, and the process is modelled as a Markov process. Table 1 includes six models of nucleotide substitution that have often been used for phylogenetic inference. The simplest model is Jukes-Cantor's and the transition probability is expressed by

$$\begin{aligned} p_{ii}(t) &= \frac{1}{4} + \frac{3}{4}\exp\left(-\frac{4}{3}\lambda t\right), \\ p_{ij}(t) &= \frac{1}{4}\left(1 - \exp\left(-\frac{4}{3}\lambda t\right)\right) \qquad (i \neq j). \end{aligned} \tag{2.1}$$

Felsenstein's model takes account of unequal nucleotide composition, and the transition probability becomes

$$\begin{aligned} p_{ii}(t) &= \exp(-ut) + (1 - \exp(-ut))\,\pi_i, \\ p_{ij}(t) &= (1 - \exp(-ut))\,\pi_j \qquad (i \neq j), \end{aligned}$$

where π_k is the equilibrium probability of the state k. Kimura's two-parameter model notes the different rate between transition (substitution within purine or pyrimidine) and transversion (substitution between purine or pyrimidine). The transition probability is expressed by

$$\begin{aligned} p_{TT}(t) &= p_{CC}(t) = p_{AA}(t) = p_{GG}(t) \\ &= \frac{1}{4} + \frac{1}{4}\exp\left(-4\beta t\right) + \frac{1}{2}\exp\left(-2(\alpha+\beta)t\right), \\ p_{TC}(t) &= p_{CT}(t) = p_{AG}(t) = p_{GA}(t) \\ &= \frac{1}{4} + \frac{1}{4}\exp\left(-4\beta t\right) - \frac{1}{2}\exp\left(-2(\alpha+\beta)t\right), \\ p_{TA}(t) &= p_{TG}(t) = p_{CA}(t) = p_{CG}(t) = p_{AT}(t) = p_{AC}(t) = p_{GT}(t) = p_{GC}(t) \\ &= \frac{1}{4} - \frac{1}{4}\exp\left(-4\beta t\right). \end{aligned}$$

Hasegawa *et al.* model takes account of both unequal nucleotide composition and the difference between transition and transversion.

Usually, the sequence data does not have information on the direction of the evolution, and time reversible processes are considered. Only an unrooted tree is identifiable. To set the root of the tree, an outgroup is incorporated in the sequence.

Table 1. Matrices of the instantaneous rates of nucleotide substitution.

	Mutant			
Original	A	T	C	G
1. Jukes-Cantor model:				
A	$\cdots$	λ	λ	λ
T	λ	$\cdots$	λ	λ
C	λ	λ	$\cdots$	λ
G	λ	λ	λ	$\cdots$
2. Felsenstein model:				
A	$\cdots$	$\pi_T\lambda$	$\pi_C\lambda$	$\pi_G\lambda$
T	$\pi_A\lambda$	$\cdots$	$\pi_C\lambda$	$\pi_G\lambda$
C	$\pi_A\lambda$	$\pi_T\lambda$	$\cdots$	$\pi_G\lambda$
G	$\pi_A\lambda$	$\pi_T\lambda$	$\pi_C\lambda$	$\cdots$
3. Kimura model:				
A	$\cdots$	β	β	α
T	β	$\cdots$	α	β
C	β	α	$\cdots$	β
G	α	β	β	$\cdots$
4. Hasegawa *et al.* model:				
A	$\cdots$	$\pi_T\beta$	$\pi_C\beta$	$\pi_G\alpha$
T	$\pi_A\beta$	$\cdots$	$\pi_C\alpha$	$\pi_G\beta$
C	$\pi_A\beta$	$\pi_T\alpha$	$\cdots$	$\pi_G\beta$
G	$\pi_A\alpha$	$\pi_T\beta$	$\pi_C\beta$	$\cdots$
5. Tamura-Nei model:				
A	$\cdots$	$\pi_T\beta$	$\pi_C\beta$	$\pi_G\alpha_1$
T	$\pi_A\beta$	$\cdots$	$\pi_C\alpha_2$	$\pi_G\beta$
C	$\pi_A\beta$	$\pi_T\alpha_2$	$\cdots$	$\pi_G\beta$
G	$\pi_A\alpha_1$	$\pi_T\beta$	$\pi_C\beta$	$\cdots$
6. Rzhetsky-Nei model:				
A	$\cdots$	β_2	β_3	α_4
T	β_1	$\cdots$	α_3	β_4
C	β_1	α_2	$\cdots$	β_4
G	α_1	β_2	β_3	$\cdots$

Note: In all matrices the ellipses on the diagonal replace the entry required to ensure that row sums are zero. π_A, π_T, π_C, π_G are the nucleotide compositions in the equilibrium.

Let us consider s homologous nucleotide sequences that consist of n nucleotide sites. The data can be represented in the following tabular form:

$$\begin{array}{cccccc}
\text{Species}\,1 & X_{11} & \cdots & X_{1q} & \cdots & X_{1n} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{Species}\,p & X_{p1} & \cdots & X_{pq} & \cdots & X_{pn} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{Species}\,s & X_{s1} & \cdots & X_{sq} & \cdots & X_{sn}
\end{array}$$

where X_{pq} is T, C, A or G and denotes the state of the q-th site in the species p. Let $\boldsymbol{X} = (X_{pq})$ be the matrix of data and

$$\boldsymbol{X}_h = (X_{1h}, \ldots, X_{sh})'$$

be the value of the h-th site. The superscript $'$ denotes the transpose operator.

Assuming a model of substitution, the log-likelihood of a given tree i is:

$$l_i(\boldsymbol{\theta}_i \mid \boldsymbol{X}) = \sum_{h=1}^{n} \log\left(f_i(\boldsymbol{X}_h \mid \boldsymbol{\theta}_i)\right) \tag{2.2}$$

where $f_i(\boldsymbol{X}_h \mid \boldsymbol{\theta}_i)$ is the probability that the species j has x_{jh} at the homologous site h. Given the topology describing the branching order, it is expressed in terms of the transition probabilities. The vector $\boldsymbol{\theta}_i$ denotes the unknown parameters such as the branching lengths, the ratio of transition rate to transversion rate. As is seen from equations (2.1), the transition probabilities are expressed in terms of the expected number of substitution (the evolutionary time along lineages multiplied by the substitution rate), it is impossible to estimate the rate and the branching dates separately without supplementary information. We therefore adopt the expected numbers of substitutions as free parameters and refer them as branch lengths, particularly when we do not assume constant rate of evolution.

Figure 1 illustrates how the likelihood of a site is calculated. At this stage, there is two conditional likelihoods $L_{A1}(\boldsymbol{X}_{1h} \mid k_1)$, $L_{A2}(\boldsymbol{X}_{2h} \mid k_1)$ of the subtrees of the subsets $\boldsymbol{X}_{1h}$ and $\boldsymbol{X}_{2h}$ connected to the internal node A, given that the state at the node is k_1. Writing the branch length of the interior-branch connecting the nodes A and B by v, the extended conditional likelihood of $\boldsymbol{X}_{0h} = (\boldsymbol{X}'_{1h}, \boldsymbol{X}'_{2h})'$ given that the state at the node B is k_2 is given by

$$L_{B1}(\boldsymbol{X}_{0h} \mid k_2) = \sum_{k_1} \bar{p}_{k_1,k_2}(v) L_{A1}(\boldsymbol{X}_{1h} \mid k_1) L_{A2}(\boldsymbol{X}_{2h} \mid k_1), \tag{2.3}$$

where $\bar{p}_{k_1,k_2}(v)$ is the transition probability expressed as a function of branch length. Combining the subsets step by step, finally we get at some node Q the all of the three conditional likelihoods $L_{Q1}(\boldsymbol{X}'_{1h} \mid k)$, $L_{Q2}(\boldsymbol{X}'_{2h} \mid k)$, $L_{Q3}(\boldsymbol{X}'_{3h} \mid k)$ of the subtrees of the subsets $\boldsymbol{X}'_{1h}$, $\boldsymbol{X}'_{2h}$, and $\boldsymbol{X}'_{3h}$ divided by the node. The likelihood of the h-th site is obtained by

$$f_i(\boldsymbol{X}_h \mid \boldsymbol{\theta}_i) = \sum_{k} \pi_k L_{Q1}(\boldsymbol{X}'_{1h} \mid k) L_{Q2}(\boldsymbol{X}'_{2h} \mid k) L_{Q3}(\boldsymbol{X}'_{3h} \mid k). \tag{2.4}$$

The unknown parameters $\boldsymbol{\theta}_i$ of a given tree topology i is estimated by maximizing the log likelihood. One may remark that the method for updating likelihood down a tree is different from Felsenstein (1981), but only because the stopping places are different (at the bottom of a branch rather than at its top).

A mathematical model of nucleotide substitution with many parameters may fit various sets of sequence data, but the estimate of evolutionary distance has a larger variance compared with that of a simpler model with fewer parameters when both models are applicable to the same data set. On the other hand, errors in estimating trees are more often caused by models that are too simple than by models that are too complex.

If s species are under study, it is not difficult to see that there exists

$$t = \frac{(2s-5)!}{2^{s-3}(s-3)!}$$

distinct unrooted bifurcating trees. Therefore it is important to be able to compare the different topologies. Since the models are not nested, the classical likelihood ratio test does not apply (see e.g. Rao (1973)).

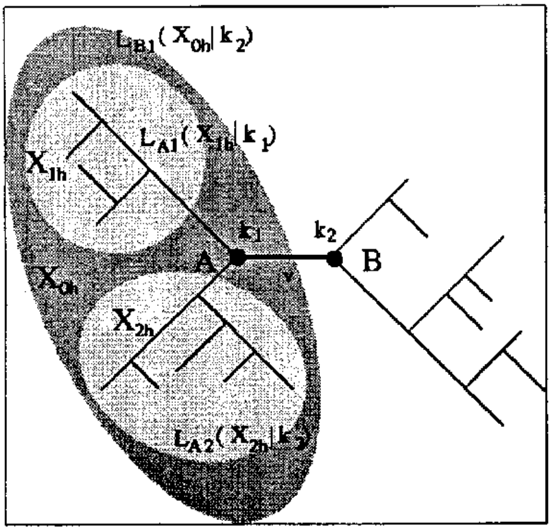

Fig. 1. Likelihood of the h-th site.

An alternative way to handle the problem is to use the Kullback-Leibler information between the true (but unknown) distribution $g(\cdot)$ and a candidate model $f_i(\cdot \mid \theta_i)$ defined as (Kullback (1959)):

$$I\left(g(\cdot); f_i(\cdot \mid \boldsymbol{\theta}_i)\right) = E_Z\left[\log \frac{g(Z)}{f_i(Z \mid \boldsymbol{\theta}_i)}\right]$$

where E_Z represents expectation with regard to Z. Since it is the difference between $E_Z\left[\log g(Z)\right]$ and $I'\left(g(\cdot); f_i(\cdot \mid \boldsymbol{\theta}_i)\right) = E_Z\left[\log f_i(Z \mid \boldsymbol{\theta}_i)\right]$, and the former does not depend on the parameter values or the type of models, we compare the latter for statistical model comparison and estimation of the parameters in the models.

If $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimator of $\boldsymbol{\theta}_i$, then $\hat{\boldsymbol{\theta}}_i$ is a consistent estimator of $\boldsymbol{\theta}_i^*$, the unique parameter that minimizes the Kullback-Leibler information quantity under weak condition on the identifiability (see e.g. White (1982)).

PROPOSITION 2.1. *Under the classical hypotheses for the expansion of the maximum likelihood, we have asymptotically*:

$$\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) = \frac{1}{n} l_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) \sim \mathcal{N}\left(\mu_i, \frac{\sigma_i^2}{n}\right) \tag{2.5}$$

with

$$\mu_i = I'\left(g(\cdot); f_i(\cdot \mid \boldsymbol{\theta}_i^*)\right),$$
$$\sigma_i^2 = \mathrm{Var}_Z\left[\log f_i(Z \mid \boldsymbol{\theta}_i^*)\right].$$

PROOF. A first order Taylor expansion of $l_i(\boldsymbol{\theta}_i^* \mid \boldsymbol{X})$ at the neighbourhood of $\hat{\boldsymbol{\theta}}_i$ gives:

$$l_i(\boldsymbol{\theta}_i^* \mid \boldsymbol{X}) = l_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \frac{1}{2} n(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*)' \boldsymbol{J}_i(\hat{\boldsymbol{\theta}}_i)(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*) + \text{Remainder} \tag{2.6}$$

where $\boldsymbol{J}_i(\hat{\boldsymbol{\theta}}_i)$ represents the Fisher information matrix and the Remainder tends to zero when the number of observations (i.e., the length of the sequences) tends to infinity.

Since $\sqrt{n}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*) = O_p(1)$, i.e. bounded, the second term of RHS of equation (2.6) is negligible compared with the first term.

Using equation (2.2), we have:

$$l_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) = \sum_{h=1}^{n} \log(f_i(\boldsymbol{X}_h \mid \hat{\boldsymbol{\theta}}_i)). \tag{2.7}$$

Since a classical hypothesis is the independence of sites (Felsenstein (1983)), the RHS of equation (2.7) is the sum of independent and identically distributed random variable. The result is obtained by applying the central limit theorem.

It has to be noted that different rates of evolution at different sites do not affect the assumption of independence if the rates are assigned independently to sites.

As noted by Kishino and Hasegawa (1989), the mean and the variance of the likelihood estimator tend to infinity as n tends to infinity. Therefore the use of $\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X})$, as defined in equation (2.5) is preferable. Various authors have studied its properties (see e.g. Steiger *et al.* (1985), White (1981, 1982)).

PROPOSITION 2.2.

$$\widehat{\mathrm{Var}}(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X})) = \frac{1}{n(n-1)} \sum_{h=1}^{n} \left\{ \log f_i(\boldsymbol{X}_h \mid \hat{\boldsymbol{\theta}}_i) - \frac{1}{n} \sum_{h=1}^{n} \log f_i(\boldsymbol{X}_h \mid \hat{\boldsymbol{\theta}}_i) \right\}^2$$

and

$$\widehat{\mathrm{Cov}}(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}), \tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X})) = \frac{1}{n(n-1)} \sum_{h=1}^{n} \left\{ \log f_i(\boldsymbol{X}_h \mid \hat{\boldsymbol{\theta}}_i) - \frac{1}{n} \sum_{h=1}^{n} \log f_i(\boldsymbol{X}_h \mid \hat{\boldsymbol{\theta}}_i) \right\} \times \left\{ \log f_i(\boldsymbol{X}_h \mid \hat{\boldsymbol{\theta}}_i) - \frac{1}{n} \sum_{h=1}^{n} \log f_i(\boldsymbol{X}_h \mid \hat{\boldsymbol{\theta}}_j) \right\}.$$

PROOF. Using the hypothesis that the sites are independently and identically distributed, the variance within sites can be estimated by the variance between site. The result is direct. Linhart (1988) and Vuong (1989) proved this result in a general setting.

One may note that $\hat{\boldsymbol{\theta}}_i$ cannot be the maximum likelihood estimator if $f_i(\boldsymbol{X}_h \mid \hat{\boldsymbol{\theta}}_i) = 0$.

Kishino *et al.* (1990) have also proposed a procedure to estimate the variance and the covariance of the likelihood estimates with the help of bootstrap techniques. It is direct to adapt it to the revised version of the maximum likelihood estimates.

3. Significance of information on phylogenetic relation

The first step of the analysis is to determine if there exists enough information to reconstruct the evolutionary tree. If it is not true, the tree will not have intermediate branch. In molecular biology, this hypothesis is commonly called the star hypothesis. Let us consider at first the four species cases. We have three possible unrooted trees. Let us call $\tilde{l}_1(\hat{\boldsymbol{\theta}}_1 \mid \boldsymbol{X})$, $\tilde{l}_2(\hat{\boldsymbol{\theta}}_2 \mid \boldsymbol{X})$ and $\tilde{l}_3(\hat{\boldsymbol{\theta}}_3 \mid \boldsymbol{X})$ the revised version of the likelihood associated with each tree.

Let us consider the vector

$$\boldsymbol{l} = \begin{pmatrix} \tilde{l}_1(\hat{\boldsymbol{\theta}}_1 \mid \boldsymbol{X}) - \tilde{l}_2(\hat{\boldsymbol{\theta}}_2 \mid \boldsymbol{X}) \\ \tilde{l}_1(\hat{\boldsymbol{\theta}}_1 \mid \boldsymbol{X}) - \tilde{l}_3(\hat{\boldsymbol{\theta}}_3 \mid \boldsymbol{X}) \\ \tilde{l}_2(\hat{\boldsymbol{\theta}}_2 \mid \boldsymbol{X}) - \tilde{l}_3(\hat{\boldsymbol{\theta}}_3 \mid \boldsymbol{X}) \end{pmatrix}. \tag{3.1}$$

Let us first derive a test to conclude if the data have enough information to discriminate between the topologies. We consider the hypothesis that the three topologies are equi-distant from the true topology, measured by the Kullback-Leibler information quantity.

PROPOSITION 3.1. *Under*

$$H_0 : I'(g(\cdot); f_1(\cdot \mid \boldsymbol{\theta}_1^*)) = I'(g(\cdot); f_2(\cdot \mid \boldsymbol{\theta}_2^*)) = I'(g(\cdot); f_3(\cdot \mid \boldsymbol{\theta}_3^*))$$

the test statistic

$$\boldsymbol{l}'\boldsymbol{l} = (\tilde{l}_1(\hat{\boldsymbol{\theta}}_1 \mid \boldsymbol{X}) - \tilde{l}_2(\hat{\boldsymbol{\theta}}_2 \mid \boldsymbol{X}))^2 + (\tilde{l}_1(\hat{\boldsymbol{\theta}}_1 \mid \boldsymbol{X}) - \tilde{l}_3(\hat{\boldsymbol{\theta}}_3 \mid \boldsymbol{X}))^2 + (\tilde{l}_2(\hat{\boldsymbol{\theta}}_2 \mid \boldsymbol{X}) - \tilde{l}_3(\hat{\boldsymbol{\theta}}_3 \mid \boldsymbol{X}))^2$$

is approximately distributed as:

$$\boldsymbol{l}'\boldsymbol{l} \approx a\chi_b^2$$

with

$$a = \frac{\operatorname{tr}(\boldsymbol{A}^2)}{\operatorname{tr}(\boldsymbol{A})},$$
$$b = \frac{(\operatorname{tr}(\boldsymbol{A}))^2}{\operatorname{tr}(\boldsymbol{A}^2)}$$

where $\boldsymbol{A}$ *is the variance-covariance of* $\boldsymbol{l}$.

PROOF. Since topologies are separate families of hypotheses, using the central limit theorem, asymptotically (Kishino and Hasegawa (1989)):

$$\boldsymbol{l} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{A})$$

with

$$\boldsymbol{\mu}' = E((\tilde{l}_1(\boldsymbol{\theta}_1^* \mid \boldsymbol{X}) - \tilde{l}_2(\boldsymbol{\theta}_2^* \mid \boldsymbol{X}), \tilde{l}_1(\boldsymbol{\theta}_1^* \mid \boldsymbol{X}) - \tilde{l}_3(\boldsymbol{\theta}_3^* \mid \boldsymbol{X}), \tilde{l}_2(\boldsymbol{\theta}_2^* \mid \boldsymbol{X}) - \tilde{l}_3(\boldsymbol{\theta}_3^* \mid \boldsymbol{X})).$$

Under $H_0 : E(\tilde{l}_1(\hat{\boldsymbol{\theta}}_1 \mid \boldsymbol{X}) = E(\tilde{l}_2(\hat{\boldsymbol{\theta}}_2 \mid \boldsymbol{X})) = E(\tilde{l}_3(\hat{\boldsymbol{\theta}}_3 \mid \boldsymbol{X})$ then $\boldsymbol{\mu} = \boldsymbol{0}$. Therefore, using classical theory of quadratic forms,

$$\boldsymbol{l}'\boldsymbol{l} \sim \sum_i \lambda_i u_i^2 \tag{3.2}$$

where λ_i are the eigenvalues of $\boldsymbol{A}$, the variance-covariance of $\hat{\boldsymbol{l}}$ and $u_i \sim \mathcal{N}(0,1)$ independent.

There exist many ways of computing the tail probabilities of (3.2), the first one is to use a mathematical package that can perform this kind of integral. An easier way is to approximate this weighted sum of central chi-square by a pondered chi-square. This approximation is known to give correct result even for the tail of the distribution (Grad and Solomon (1955)). If we may notice that $\sum_i \lambda_i = \operatorname{tr}(\boldsymbol{A})$ and $\sum_i \lambda_i^2 = \operatorname{tr}(\boldsymbol{A}^2)$, the proposition is obtained by equalizing the two first moments.

The degrees of freedom of the chi-square can be non integer. With most of the statistical packages, this is not a restriction. In other case it is always possible to use approximation as Wilson-Hilferty's one (Johnson and Kotz (1970)).

Let us go to the next step that is the problem of s $(s > 4)$ species. Therefore there exists t distinct unrooted trees. As a generalisation of equation (3.1), one may write:

$$\boldsymbol{l} = \begin{pmatrix} \tilde{l}_1(\hat{\boldsymbol{\theta}}_1 \mid \boldsymbol{X}) - \tilde{l}_2(\hat{\boldsymbol{\theta}}_2 \mid \boldsymbol{X}) \\ \tilde{l}_1(\hat{\boldsymbol{\theta}}_1 \mid \boldsymbol{X}) - \tilde{l}_3(\hat{\boldsymbol{\theta}}_3 \mid \boldsymbol{X}) \\ \vdots \\ \tilde{l}_{t-1}(\hat{\boldsymbol{\theta}}_{t-1} \mid \boldsymbol{X}) - \tilde{l}_t(\hat{\boldsymbol{\theta}}_t \mid \boldsymbol{X}) \end{pmatrix}.$$

$\boldsymbol{l}$ is a vector of dimension $\frac{t(t-1)}{2}$ and asymptotically

$$\boldsymbol{l} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{A}).$$

Under .

$$H_0 : I'(g(\cdot); f_1(\cdot \mid \boldsymbol{\theta}_1^*)) = I'(g(\cdot); f_2(\cdot \mid \boldsymbol{\theta}_2^*)) = \cdots = I'(g(\cdot); f_t(\cdot \mid \boldsymbol{\theta}_t^*))$$

$\boldsymbol{\mu}'\boldsymbol{\mu} = 0.$

Thus with the same arguments as before we also end up on a weighted chi-square and the degree of freedom and the weighted term are of the same form than in the Proposition 3.1.

4. Tests for more specific aims

If H_0 cannot be rejected it means that the data does not permit to discriminate between the topologies. In other word, it means that the data does not have enough information on the phylogenetic relations. One may expect that we do not end to this case and thus the hypothesis that all the likelihoods are equal has to be rejected. In this case, the practionner wants to be able to compare the different topologies. The aim of the two next sections is to answer to some of the most classical questions. In the first section, we derive the distribution of linear combinations of the likelihood estimates. It is especially useful when a specific comparison is wanted or if a specific regroupment is under consideration. In the second section, we derived a test to determine if a candidate tree can be the best tree.

It is well known in the context of analysis of variance that a global test can be significant even if the multiple comparisons procedures cannot permit to discriminate between the various hypothesis. Since the same problem can arise with the proposed method, it is better to conduct at first the test derived in Section 3.

4.1 *Linear combinations of the likelihood estimates*

In many applications, some inference had been made about the candidate for the true topologies among the taxa. Since different methods can lead to different results, it is important to be able to test predetermined hypothesis. This question is very close to the notion of contrasts in the framework of the analysis of variance (see e.g. Hochberg and Tamhane (1987)). One application is when a given tree has to be compared to alternative trees. Another case is when a particular association is under interest. In this case, the practionner wants to compare the trees with the given association to the tree without this association.

PROPOSITION 4.1.. *Under the hypothesis* H_{1a} : $\sum_i c_i\mu_i = 0(\sum_i c_i = 0)$ *the test statistic*

$$\frac{(\sum_i c_i\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}))^2}{\sum_i c_i^2 \operatorname{Var}(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X})) + \sum_{i\neq j} c_i c_j \operatorname{Cov}(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}), \tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X}))} \tag{4.1}$$

follows asymptotically a chi-square distribution with one degree of freedom.

PROOF.

$$\sum_i c_i\tilde{l}_i(\boldsymbol{\theta}_i \mid \boldsymbol{X}) = \frac{1}{k}\sum_{i>j}(c_i - c_j)(\tilde{l}_i(\boldsymbol{\theta}_i \mid \boldsymbol{X}) - \tilde{l}_j(\boldsymbol{\theta}_j \mid \boldsymbol{X})).$$

Since $\boldsymbol{l}$ follows a multivariate normal law, any linear combination of $\boldsymbol{l}$ follows a normal law.

From a practical point of view, the sequence is generally long enough to replace the variance and the covariance by their estimate and therefore, it is direct to obtain that the equation (4.1) follows a χ_1^2 under the hypothesis H_{1a}.

To preserve the level of the tests the contrasts have to be independent and it is easy to see that the comparison of t likelihoods permits $t-1$ independent contrasts.

4.2 *Candidate for the best tree*

In some applications, the candidate for the true tree cannot be inferred from previous studies. In this case the parameter of interest is (see Edwards and Hsu (1983); Hsu (1984); Hochberg and Tamhane (1987)):

$$E(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X})) - \max_{j\neq i} E(\tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X})).$$

These are nonlinear functions of $\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X})$'s and therefore the problem is slightly different from the contrasts.

PROPOSITION 4.2. *Let's consider the case of s species, a simultaneous confidence interval at the* $(1-\alpha)\%$ *level for* $I'(g(\cdot); f_i(\cdot \mid \theta_i^*)) - I'(g(\cdot); f_j(\cdot \mid \boldsymbol{\theta}_j^*))$ *(i fixed and* $j \in \{1,\ldots,t\}\backslash\{i\}$*) is given by:*

$$[\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X}) - d, \tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X}) + d] \qquad j \in \{1,\ldots,t\}\backslash\{i\} \tag{4.2}$$

where d is the critical value of a standardized normal distribution at the $(1-\frac{\alpha}{2(t-1)})\%$ *level, multiplied by* $u = \max_{j\neq i}\sqrt{\widehat{\operatorname{Var}}(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \tilde{l}_i(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X}))}$.

PROOF. At first, we note that:

$$\frac{(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X})) - (E(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X})))}{\sqrt{\widehat{\operatorname{Var}}(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X}))}} \sim \mathcal{N}(0,1).$$

Generally n, the length of the sequence, is large enough to ensure the validity of the normal approximation. The pairwise comparisons are not independent. To take this fact into account a correction factor has to be used. The easiest way of doing is to apply the (conservative) Bonferroni correction.

COROLLARY 4.1. *A simultaneous confidence interval at the* $(1-\alpha)\%$ *level for*

$$\left(E(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X})) - \max_{j\neq i} E(\tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X}))\right) \qquad (i \textit{ fixed and } j \in \{1,\ldots,t\}\backslash\{i\})$$

is given by:

$$\left[\left(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \max_{j\neq i}\tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X}) - d\right)^-, \left(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \max_{j\neq i}\tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X}) + d\right)^+\right] \tag{4.3}$$

with $x^- = \min(x,0)$, $x^+ = \max(x,0)$ *and d defined as in equation* (4.2).

It is a direct application of a result of Hochberg and Tamhane ((1987), p. 150).

5. Example

Horai *et al.* (1995) determined the entire mitochondrial DNA (mtDNA) sequences (16.5 Kb) from the following five species:

1. Human;
2. Common Chimpanzee (*Pan troglodytes*);
3. Pygmy Chimpanzee (*bonobo*; *Pan paniscus*);
4. Gorilla (*Gorilla gorilla*);
5. Orangutan (*Pongo pygmaeus*).

Codons are translated to amino acids of proteins. If the translation into an amino acid does not depend on the base at the third position of a particular codon, this codon is called four-fold degenerate. One may expect that the evolution of these sites is neutral (Kimura (1983)), and the substitution pattern in evolution reflects that of mutation. In this set of data, there is 1669 four-fold degenerate homologeous sites that are not ambiguous and not overlapping. We focus our attention on these sites.

Since 5 species are under study, it is possible to construct 15 unrooted bifurcating trees.

Table 2 shows, for each tree, $\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X})$ and $\sqrt{\widehat{\mathrm{Var}}(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}))}$ for $i = 1,\ldots,15$. Hasegawa *et al.* model of Table 1 was used. The trees are highly correlated. Therefore the variance of the difference is very small.

A test of homogeneity was performed for the 15 unrooted trees. Under $H_0 : E(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X})) = \cdots = E(\tilde{l}_{15}(\hat{\boldsymbol{\theta}}_{15} \mid \boldsymbol{X}))$

$$13.41 \sim \chi^2_{1.76}.$$

Therefore the null hypothesis is rejected even at the 0.1% level of confidence. From many previous studies, the generally accepted tree is Tree 1 (Hasegawa and Yano (1984), Sibley and Ahlquist (1984)) and the classical other candidates are Tree 2 (Templeton (1983), Martin (1985)) and Trees 3 (Ueda *et al.* (1985)).

From the Table 2, it is obvious that the hypothesis of equality of the revised likelihood of the Tree 2 and the Tree 3 will not be rejected, even at the 5% level. Under the

Table 2. Sample statistics of the likelihood of the 1669 sites for each of the possible unrooted trees.

No of the Tree	Topology	$\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X})$	$\sqrt{\widehat{\mathrm{Var}}(\tilde{l}_i(\hat{\boldsymbol{\theta}}_1 \mid \boldsymbol{X}) - \tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}))}$
Tree 1	(((Chimp, Bonob), Human), Goril, Orang)	−3.39341	0
Tree 2	((Chimp, Bonob), Goril, (Orang, Human))	−3.40198	0.003686
Tree 3	((Chimp, Bonob), (Goril, Human), Orang)	−3.40198	0.003684
Tree 4	(((Chimp, Human), Bonob), Goril, Orang)	−3.40851	0.005642
Tree 5	((Chimp, (Bonob, Human)), Goril, Orang)	−3.40851	0.005642
Tree 6	(Chimp, ((Goril, Human), Bonob), Orang)	−3.41342	0.005679
Tree 7	(Chimp, (Goril, Bonob), (Orang, Human))	−3.42025	0.007343
Tree 8	(Chimp, Goril, ((Orang, Human), Bonob))	−3.42049	0.007280
Tree 9	(Chimp, (Goril, Human), (Orang, Bonob))	−3.41372	0.005603
Tree 10	((Chimp, Human), (Goril, Bonob), Orang)	−3.42379	0.008717
Tree 11	(Chimp, ((Goril, Bonob), Human), Orang)	−3.42361	0.008768
Tree 12	(Chimp, (Goril, (Bonob, Human)), Orang)	−3.42385	0.008697
Tree 13	(Chimp, Goril, (Orang, (Bonob, Human)))	−3.42403	0.008653
Tree 14	(Chimp, Goril, ((Orang, Bonob), Human))	−3.42403	0.008653
Tree 15	((Chimp, Human), Goril, (Orang, Bonob))	−3.42409	0.008628

hypothesis that the pseudo-likelihood of the Tree 1 is equal to the mean of the revised likelihood of the Trees 2 and 3, we have:

$$5.41 \sim \chi_1^2.$$

Therefore, the Tree 1 *is significantly different from the Trees 2 and 3 at the 5% level.*

For the three first trees, the closest species are the two species of chimpanzees. Therefore, it is possible to test this particular association by making a contrast between trees 1, 2, 3 and the others. The coefficients for the three first trees are -4, -4, -4 and the coefficients for the 12 last threes are $-1, \ldots, -1$. After some computations we obtain:

$$9.15 \sim \chi_1^2$$

which is significantly different at the 5% level.

Even if there are many studies that tried to infer the true tree, we apply, by sake of curiosity, the techniques of the Subsection 4.2 to determine the trees that can be candidate for the best tree.

$$u = \max_{j \neq i} \sqrt{\widehat{\mathrm{Var}}(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \tilde{l}_j(\hat{\boldsymbol{\theta}}_j \mid \boldsymbol{X}))} = \sqrt{\widehat{\mathrm{Var}}(\tilde{l}_i(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{X}) - \tilde{l}_{11}(\hat{\boldsymbol{\theta}}_{11} \mid \boldsymbol{X}))} = 0.00877$$

and the critical value of a standardized normal law at the $(1 - \frac{0.05}{2*14})\%$ is 2.914.

Therefore, we have, at the 5% level of confidence,

$$d = 0.00877 \times 2.914 = 0.0256.$$

Not only Tree 1 is selected but also Tree 2 to Tree 9 are selected at the 5% level.

6. Conclusions

Many methods have been proposed to recover the structure of the evolutionary trees. One of the main argument for the maximum likelihood techniques is that the hypothesis have to be clearly stated even if the same hypothesis are often implied by the other method. Although a lot of attention has been paid to the hypothesis, little attention has been paid to the statistical properties of the maximum likelihood estimates. In this article we developed some useful procedure to test the pertinence of the reconstructed trees but for specific case, other tests can be derived such as the multiple comparison procedure or the Dunnett test (see e.g. Hochberg and Tamhane (1987)). We might consider the normalized statistics to increase the power. This is left for further study.

Since the amount of available data and the capacity of the computer is increasing very quickly, the need of this kind of procedure is very important to understand the evolutionary process of the species. In this article we focus our interest on branching order but the problem of the precision of the branch length is also an important question and work is currently be done on this subject.

Acknowledgements

We would like to thank Hidetoshi Shimodaira and Masami Hasegawa for helpful comments on an earlier version of this paper, Jun Adachi for providing the data of the example and a referee whose remarks have led to great improvements of the paper.

References

Barry, D. and Hartigan, J. A. (1987). Statistical analysis of hominoid molecular evolution, *Statist. Sci.*, **2**, 191–210.

Churchill, G. A., van Haeseler, A. and Navidi, W. C. (1992). Sample size for a phylogenetic inference, *Molecular Biology and Evolution*, **9**, 753–769.

Cox, D. R. (1961). Tests of separate families of hypotheses, *Proc. 4th Berkeley Symp. on Math. Statist. Prob.*, Vol. 1, 105–123, University of California Press, Berkeley.

Edwards, D. G. and Hsu, J. C. (1983). Multiple comparisons with the best treatment, *J. Amer. Statist. Assoc.*, **78**, 965–971.

Felsenstein, J. (1978). Cases in which parsimony or competibility methods will be positively misleading, *Systematic Zoology*, **27**, 401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach, *Journal of Molecular Evolution*, **17**, 368–376.

Felsenstein, J. (1983). Statistical inference of phylogenies, *J. Roy. Statist. Soc. Ser. A*, **146**, 246–272.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, **39**, 783–791.

Felsenstein, J. and Kishino, H. (1993). Is there something wrong with the bootstrap on phylogenies?, A reply to Hillis and Bull, *Systematic Biology*, **42**, 193–200.

Goldman, N. (1993). Statistical tests of models of DNA substitution, *Journal of Molecular Evolution*, **36**, 182–198.

Grad, A. and Solomon, H. (1955). Distribution of quadratic forms and some applications, *Ann. Math. Statist.*, **26**, 464–477.

Hasegawa, M. and Yano, T. (1984). Phylogeny and classification of Hominoidea as inferred from DNA sequence data, *Proceedings of the Japan Academy*, **B60**, 389–392.

Hillis, D. M. and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis, *Systematic Biology*, **42**, 182–192.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*, Wiley, New York.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. and Takahata, N. (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs, *Proc. Nat. Acad. Sci. U.S.A.*, **92**, 532–536.

Hsu, J. C. (1984). Constrained simultaneous confidence intervals for multiple comparisons with the best, *Ann. Statist.*, **12**, 1136–1144.

Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions*, Wiley, New York.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea, *Journal of Molecular Evolution*, **29**, 170–179.

Kishino, H., Miyata, T. and Hasagewa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts, *Journal of Molecular Evolution*, **31**, 151–160.

Kullback, S. (1959). *Information Theory and Statistics*, Dover, New York.

Linhart, H. (1988). A test whether two AIC's differ significantly, *South African Statist. J.*, **22**, 153–161.

Martin, L. (1985). Significance of enamel thickness in hominoid evolution, *Nature*, **314**, 260–263.

Navidi, W. C., Churchill, G. A. and von Haeseler, A. (1993). Phylogenetic inference: linear invariants and maximum likelihood, *Biometrics*, **49**, 543–555.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed., Wiley, New York.

Sibley, C. G. and Ahlquist, J. E. (1984). The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization, *Journal of Molecular Evolution*, **20**, 2–15.

Sitnikova, T., Rzhetsky, A. and Nei, M. (1995). Interior-branch and bootstrap tests of phylogenetic trees, *Molecular Biology and Evolution*, **12**, 319–333.

Steiger, J. H., Shapiro, A. and Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics, *Psychometrika*, **50**, 253–264.

Swofford, D. L. and Olsen, G. J. (1990). Phylogeny reconstruction, *Molecular Systematics* (eds. D. M. Hillis and C. Moritz), 411–501, Sinauer, Sunderland, Massachusetts.

Templeton, A. R. (1983). Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes, *Evolution*, **37**, 221–244.

Ueda, S., Takenaka, O. and Honjo, T. (1985). A truncated immunoglobulin ϵ pseudogene is found in gorilla and man but not in chimpanzee, *Proc. Nat. Acad. Sci. U.S.A.*, **82**, 3712–3715.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non nested hypotheses, *Econometrica*, **57**, 307–333.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models, *J. Amer. Statist. Assoc.*, **76**, 419–433.

White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 253–264.

Zharkikh, A. and Li, W. H. (1992*a*). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: four taxa with a molecular clock, *Molecular Biology and Evolution*, **9**, 1119–1147.

Zharkikh, A. and Li, W. H. (1992*b*). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: four taxa without molecular clock, *Journal of Molecular Evolution*, **35**, 356–366.