THE ROYAL SOCIETY

Received 8 June 2001
Accepted 14 September 2001
Published online 20 December 2001

Inferring confidence sets of possibly misspecified gene trees

Korbinian Strimmer* **and Andrew Rambaut**

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

The problem of inferring confidence sets of gene trees is discussed without assuming that the substitution model or the branching pattern of any of the investigated trees is correct. In this case, widely used methods to compare genealogies can give highly contradicting results. Here, three methods to infer confidence sets that are robust against model misspecification are compared, including a new approach based on estimating the confidence in a specific tree using expected-likelihood weights. The power of the investigated methods is studied by analysing HIV-1 and mtDNA sequence data as well as simulated sequences. Finally, guidelines for choosing an appropriate method to compare multiple gene trees are provided.

Keywords: model selection; model misspecification; confidence set; expected likelihood weight; bootstrap; variance reduction

1. INTRODUCTION

The assessment of competing evolutionary trees inferred from DNA sequences is a very important issue in molecular sequence analysis. Consequently, in recent years a number of statistical procedures to test gene trees and to construct corresponding confidence sets have been suggested, most of them based on computing the likelihood of trees. Interestingly, a recent review by Goldman *et al.* (2000) showed that the available methods essentially fall into two classes that can give highly contradicting evaluations of the confidence in the compared trees. This violates the intuitive notion that two different but equally valid approaches to analysing the same data should give the same answer.

Here, this problem is investigated and it is argued that this apparent difference is due to the potential misspecification of the investigated genealogies to which one class of tree comparison methods is susceptible whereas the other is not. Gene trees can be misspecified either because the tree topology or the employed model of substitution is incorrect. In addition to reviewing this question a simple method based on expected likelihood weights is proposed to robustly infer confidence sets of gene trees.

The rest of the paper is organized to provide an introduction to statistical methods and model comparison with special emphasis on model misspecification. Following this, methods to construct confidence sets of gene trees are described. Then the datasets reported in Shimodaira & Hasegawa (1999) and in Goldman *et al.* (2000) are reanalysed and biological reasons are discussed to determine why the investigated genealogies for these sequences might be misspecified. Using computer simulation the efficiency of the investigated methods for constructing confidence sets are studied. Finally, guidelines are presented for choosing an appropriate method for the comparison of gene trees.

*Author for correspondence (korbinian.strimmer@zoo.ox.ac.uk).

2. THEORY

(a) *Models*

A statistical model for a random variable X is provided by a probability distribution for all states x assumed by X. A dataset $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a vector of n independent realizations of X. Typically, the true model F with distribution $f(x)$ that gave rise to the observed data $\boldsymbol{x}$ is not known. Instead, to explain the data one usually considers a number of candidate models $M_1, M_2, \ldots, M_r$ with some proposed distributions $m_1(x), m_2(x), \ldots, m_r(x)$. A set of models form a composite or parameterized model $M(\boldsymbol{\theta})$ if their distributions have the form $m(x; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ represents the parameters. It is usually not known whether the true model F is included in the set of candidate models (if it is not, the model set is said to be misspecified).

(b) *Likelihood*

The likelihood framework provides a means of evaluating data as evidence for a given model (Birnbaum 1962; Edwards 1972). The likelihood L of a model M is defined by

$$L = L(M|\boldsymbol{x}) = \Pr(\boldsymbol{x}|M) = \prod_{i=1}^{n} m(x_n), \tag{2.1}$$

and provides a measure of fit between the model and the data. The 'law of likelihood' states that for any two models M_1 and M_2, model M_1 is better supported by the observed data $\boldsymbol{x}$ than model M_2 if $L_1 > L_2$, and the likelihood ratio L_1/L_2 measures the strength of evidence in favour of M_1 versus M_2 (Hacking 1965). Similarly, the evidence w_i of the data for a model M_i relative to the competing models $M_1, M_2, \ldots, M_r$ is given by the 'likelihood weight'

$$w_i = w(M_i|\boldsymbol{x}) = L_i \Big/ \sum_{j=1}^{r} L_j. \tag{2.2}$$

The model $M_{\max}$ with the highest likelihood, and hence with the largest likelihood weight $w_{\max}$, is called the maximum-likelihood (ML) model. Similarly, the parameter vector $\boldsymbol{\theta}_{\max}$ that selects the ML model in a parameterized model set $M(\boldsymbol{\theta})$ is the ML estimate of this vector.

(c) *Model comparison*

The likelihood ratio (LR) provides a natural test statistic for the comparison of the goodness-of-fit of two competing models. The sampling distribution of the LR statistic, usually under the null hypothesis of the less likely model, is easily obtained by Monte Carlo simulation (Cox 1961, 1962). If two fully nested model families are compared, minus twice the logarithm of the LR statistic has a limiting central χ^2-distribution under suitable regularity conditions (Wilks 1938). For two non-nested models the limiting distribution is Gaussian (Cox 1961, 1962; White 1982*b*). These tests implicitly assume that at least one of the models is correct. However, often it cannot be guaranteed that the true data-generating model is among the investigated candidate models. Consequently, in the case of misspecified models the above tests can be invalid (Foutz & Srivastana 1977; Kent 1982; White 1982*a*; Golden 1995). However, misspecification of a candidate model can be tested (White 1982*a*). Moreover, LR tests robust against model misspecification are constructed using the null hypothesis that the compared models are equally close to the (unknown) true model, rather than singling out a particular model as the null model (Vuong 1989). This also allows extension to compare multiple models (Shimodaira 1998).

(d) *Confidence sets*

The objective of a confidence set is to provide an interval estimate that gives a measure of precision for a point estimate (the ML model). An estimated confidence set can be defined as the smallest subset of the investigated models that contains the true model for a prespecified fraction C, of all possible datasets of size n, generated under the true model (e.g. Garthwaite *et al.* 1995). An alternative definition of a confidence set is based on model selection probabilities. In this perspective, a confidence set is the smallest subset of the candidate models that have together probability C to be selected as outcomes for a random dataset of length n drawn from the true distribution. These two interpretations of confidence sets are equivalent for correctly specified model sets but the latter is also applicable in situations where the true model is not included in the set of candidate models. It also generalizes to multidimensional problems and is implicit in Monte Carlo procedures for the construction of confidence sets (Buckland 1984). Note that in either definition the confidence set takes hypothetical data other than the observed $\boldsymbol{x}$ into account.

(e) *Inferring confidence sets*

Confidence sets are closely related to hypothesis tests: the acceptance region $1 - \alpha$ of a test H_0 (true model) versus H_1 is a confidence set with coverage $1 - \alpha$. This allows the inference of confidence sets of models using the LR tests described earlier.

However, confidence sets can also be constructed more directly using model selection probabilities (following the second definition of a confidence set). The selection probability c_i for an individual model M_i given a random data sample from the true model can be estimated by the expected relative evidence for that model, i.e. by $c_i = E_F(w_i)$, where the expectation is taken with respect to the true model F. If two models M_1 and M_2 have the same likelihood for all possible samples from the true distribution, then by definition they also have the model selection probability ($c_1 = c_2$). The expected likelihood weight c_i can be directly interpreted as confidence in a model. As it is additive (note that $\sum_{i=1}^{r} c_i = 1$) the confidence of a subset of the compared models $M_1, M_2, \ldots M_r$ is the sum of the confidence values of all models in that subset. The smallest confidence set with maximum probability for a prescribed level of confidence C is constructed by collecting the models with the largest confidence values in decreasing order, until the accumulated level of confidence meets the threshold C.

(f) *Computing confidence values*

To calculate the expected likelihood weights c_i the true model F for the data sample $\boldsymbol{x}$ needs to be known. As this is hardly ever the case, an approximation such as

$$c_i = E_F(w_i) \approx E_{\hat{F}}(w_i) \approx \frac{1}{B}\sum_{b=1}^{B} w(M_i|\boldsymbol{x}^{*b}) \qquad (2.3)$$

is useful, where $\boldsymbol{x}^{*b}$ is one of B bootstrap samples drawn with replacement from the data $\boldsymbol{x}$ and $\hat{F}$ is the nonparametric empirical distribution (Efron 1982; Efron & Tibshirani 1993). Under suitable regularity conditions the first approximation ($\hat{F} \rightarrow F$) is valid for large sample size n, whereas the second requires a large number B of bootstrap replicates. In this expression c_i can also be interpreted as the 'bagged' variance-reduced estimator for w_i (Breiman 1996). The variance reduction, with typically no significant increase in bias, is implicit in the bootstrap averaging (Efron & Tibshirani 1997). It is expected that more ambitious bootstrap estimators for c_i can also be fashioned (Efron 1987; DiCiccio & Efron 1996).

(g) *Bootstrap weights and coverage*

Related to the expected likelihoods weights are the weights $s_i = E_F(I_i)$, where $I_i = I(M_i|\boldsymbol{x})$ is an indicator function of the ML model, i.e. $I_i = 1$ if i is the best-fit model otherwise $I_i = 0$. In comparison, for each random data sample the indicator function gives evidence only to one model, whereas the likelihood weights give evidence to all models. Several bootstrap estimators for s_i exist (Felsenstein 1985; Efron *et al.* 1996). Most often, the estimated values are interpreted as p-values (Hillis & Bull 1993), which contrasts with the understanding of expected likelihood weights as a confidence distribution.

Here it is argued that the likelihood weights $c_i = E(w_i)$ are preferable as model selection probabilities over the weights $s_i = E(I_i)$ for four reasons. First, if prior information (e.g. in the form of a likelihood) is available it can be incorporated easily into the likelihood weight (Edwards 1972). For small sample size this information will be recovered in $E(w_i)$ but not in $E(I_i)$. Second, confidence sets inferred using $E(I_i)$ tend to undercover and usually need upward calibration, e.g. by employing the double bootstrap (Efron & Tibshirani 1993). By contrast, the distribution given by $E(w_i)$ is wider than that given by $E(I_i)$. Preliminary simulations indicate (data not shown) that as a result, confidence sets based on $E(w_i)$ rarely undercover. Third, the expected likelihood weight c_i has a further interpretation as a predictive model selection criterion emphasizing generalizability of a model in addition to

goodness-of-fit (Akaike 1974; Linhart & Zucchini 1986; Myung *et al.* 2000). Fourth, for large sample size, the likelihood weight w_i degenerates to the indicator function I_i for the best-fit model, and hence the weights s_i can be considered a special case of the expected likelihood weights w_i.

(h) *Gene trees*

Gene trees describe the evolutionary relationship of genes. The statistical dependencies between the gene sequences in an evolutionary tree are commonly explained using directed graphical models (Felsenstein 1981; Hendy *et al.* 1994; Strimmer & Moulton 2000). These probabilistic models consist of two distinct parts: a graph (the tree structure and branch lengths) and an associated stochastic process (the substitution model); both determine together a distribution for all possible site patterns observable in a column of a sequence alignment. A sequence dataset $\boldsymbol{x}$ of length n corresponds to n samples from this distribution.

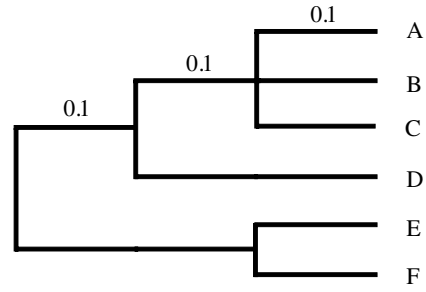
Statistical comparison of gene trees is widely applied. Goldman (1993) was the first to propose parametric bootstrapping using the LR statistic (Cox 1961, 1962). A variety of applications of LR tests for gene trees are reviewed in Huelsenbeck & Rannala (1997). Felsenstein (1985) used bootstrap proportions to assess the reliability of a tree. Kishino & Hasegawa (1989) devised an approach similar to that reported by Vuong (1989) to compare two gene trees. An extension of this method to multiple comparison of trees, following Shimodaira (1998), is described by Shimodaira & Hasegawa (1999). A further likelihood approach to multiple comparison of trees is given by Bar-Hen & Kishino (2000). Goldman *et al.* (2000) provided a recent technical overview of LR tests for gene trees and associated confidence sets; they also discuss the appropriate use of the test by Kishino & Hasegawa (1989).

3. APPLICATION

(a) *Mammalian protein sequences*

The first example to illustrate the inference of confidence sets of gene trees is taken from Shimodaira & Hasegawa (1999) who analysed mitochondrial protein sequences from six mammalian species (human, harbour seal, cow, rabbit, mouse, opossum). The alignment has length $n = 3414$ amino acids. For all 105 possible topologically different trees for the six sequences ML branch lengths were estimated using the mtREV+Γ amino acid substitution model (Adachi & Hasegawa 1996). Subsequently, two sets of candidate models were investigated. The first test set consisted of the 15 most likely gene trees as reported in Shimodaira & Hasegawa (1999); in the second set, all 105 genealogies were included as candidate models.

Shimodaira (2001) points out that in this example the gene trees are misspecified as they are all rejected in LR tests (Cox 1961, 1962). This does not necessarily imply that the data do not fit to a tree; it can also indicate that the substitution model is not adequate. Therefore, to infer 95% confidence sets, only procedures robust against model misspecification were employed. In particular, the KH method (Kishino & Hasegawa 1989), the SH method (Shimodaira & Hasegawa 1999) and the approach based on the expected likelihood weight as measure of confidence were used. The results are summarized in table 1.



T_1

Figure 1. Tree T_1 used for simulating data. Branch lengths denote expected nucleotide substitutions per site.

Two distinct patterns can be observed. First, the confidence set inferred using the expected likelihood weights is the smallest and contains four trees. The KH test produced a slightly larger confidence set with five trees, whereas the SH test gives the most conservative estimate (8–16 trees). Second, the SH test is sensitive to the inclusion of additional but unlikely gene trees. When the 15 best genealogies are investigated the SH confidence set contains only the gene trees 1–8, but when all 105 trees are compared it consists of the trees 1–15 and 17.

These confidence sets contrast with the result of a Cox-type LR test for the same data presented in Goldman *et al.* (2000), which strongly favours the ML gene tree as the only appropriate explanation for the data. However, this test may be misleading; in the presence of model misspecification, LR tests based on the assumption that one of the candidate models is the true data-generating model are not applicable (Foutz & Srivastana 1977; Kent 1982; White 1982*a*).

(b) *HIV nucleotide sequences*

The second dataset was taken from Goldman *et al.* (2000). It consists of six HIV-1 nucleotide sequences of length $n = 2000$ bp from the *gag* and *pol* genes. The sequences are from four subtypes (A, B, D, E) and are referred to as A1 (HIV-1 common name Q23), A2 (U455), B (BRU), D (NDK), E1 (90CF11697) and E2 (93TH057). For all 105 possible genealogies for these sequences, ML branch lengths were estimated under the REV+Γ nucleotide substitution model (e.g. Yang 1994).

These data are interesting as HIV-1 is known to be subject to frequent recombination (e.g. Robertson *et al.* 1995), and hence sites along the sequence alignment may well have different evolutionary histories. In other words, any tree-like model for the relationship of the genes is likely to be incorrect. In this case application of methods robust against model misspecification is advised. Table 2 shows the 95% confidence sets as inferred by the KH and SH methods as well as by our approach.

The KH test and the approach based on expected likelihood weights agree that the three most likely trees form a suitable confidence set. By contrast, the SH test is much more conservative and includes 12 more gene trees including one with a log-likelihood difference to the best tree as large as $\Delta l = 35.83$. In this example, the small confidence set has a straightforward interpretation. The three

Table 1. Confidence sets of mammalian gene trees.
(Trees are listed in the order of their likelihood. Abbreviations: Δl, log-likelihood difference to most likely tree; c, confidence value (expected likelihood weight); KH, p-value for the two-sided KH test; SH, p-value for SH test. The number of bootstrap replicates (c, SH) was $B = 1000$. Bold type indicates trees included in the corresponding 95% confidence set ($p > 0.1$ for the KH test; see Goldman *et al.* 2000).)

		number of candidate trees					
		the best 15 trees			all 105 trees		
tree	Δl	c	KH	SH	c	KH	SH
1	0.00	**0.5603**	**1.0000**	**1.0000**	**0.5460**	**1.0000**	**1.0000**
2	2.52	**0.3263**	**0.7395**	**0.8060**	**0.3295**	**0.7395**	**0.9300**
3	7.41	**0.0397**	**0.2391**	**0.5820**	**0.0507**	**0.2391**	**0.8390**
4	17.74	0.0110	0.0814	**0.1770**	0.0173	0.0814	**0.5730**
5	19.07	**0.0347**	**0.1311**	**0.1590**	**0.0275**	**0.1311**	**0.5470**
6	20.09	0.0073	**0.1021**	**0.1130**	0.0067	**0.1021**	**0.5320**
7	20.79	0.0165	0.0956	**0.1000**	0.0157	0.0956	**0.4700**
8	22.22	0.0026	0.0634	**0.0770**	0.0030	0.0634	**0.4630**
9	25.67	0.0000	0.0025	0.0290	0.0000	0.0025	**0.3750**
15	36.54	0.0000	0.0006	0.0020	0.0000	0.0006	**0.1650**
16	48.73				0.0000	0.0003	0.0480
17	49.39				0.0000	0.0002	**0.0500**
size of set		4	5	8	4	5	16

Table 2. Confidence sets of HIV-1 gene trees.
(For definition of abbreviations see table 1. The test set includes all 105 possible trees.)

tree	topology	Δl	c	KH	SH
1	(E1, E2,(A2,(A1,(D,B))))	0.00	**0.6780**	**1.0000**	**1.0000**
2	(E1, E2,(A1,(A2,(D,B))))	3.61	**0.1609**	**0.4289**	**0.8600**
3	(E1, E2,((A1,A2),(D,B)))	3.90	**0.1526**	**0.3840**	**0.8600**
4	(E1,(E2,(D, B)),(A1,A2))	19.67	0.0033	0.0268	**0.4000**
5	(E1,(E2,(A1, A2)),(D,B))	19.67	0.0032	0.0268	**0.4000**
15	(E1,((E2,A2),A1),(D,B))	35.83	0.0000	0.0007	**0.1390**
size of set			3	3	15

included trees indicate that the genealogical relationship between the sequences A1 and A2 relative to the two groups E1/E2 and D/B is unresolved.

Parametric bootstrap tests as described in Goldman *et al.* (2000) reject all trees but the ML tree. As before, it can be argued that this result may be biased, as a result of the misspecification of the investigated gene trees. In contrast to the mammalian sequences, where the substitution model is likely to be incorrect, in the case of the HIV-1 data, net-like rather than tree-like evolution may be the cause of the misspecification.

(c) *Efficiency of SH confidence set*

In the previous examples, the SH method gave the most conservative estimates of confidence sets and also appeared to be sensitive to the inclusion of unlikely models in the test set. To study further the statistical efficiency of the SH confidence set we simulated sequences along tree T_1 in figure 1. This tree relates six sequences A–F and contains a multifurcation next to sequences A, B and C. Using a Kimura (1980) substitution model with a transition–transversion ratio of 2, we generated datasets of various lengths ($n = 100$, 500, 1000, 2000, 3000, 4000 and 5000 bp). As candidate trees we considered all 105 possible binary trees. As the true multifurcating genealogy T_1 is not available in this set, we expect the confidence set to include the three trees necessary to resolve the polytomy in T_1.

Table 3 shows the sizes of the confidence sets inferred for these data using the KH and SH tests and the expected likelihood weight. All approaches essentially agree on a confidence set containing three trees. However, in contrast to the alternative methods the SH test requires an order of magnitude more data (4000 bp versus 500 bp) to restrict its confidence set to the best three trees. For very short sequences (100 bp) all methods are conservative, with the expected likelihood weight leading to the overall smallest confidence set (10 trees).

Table 3. Size of confidence sets in dependence of sequence length.
(For definition of abbreviations see table 1. Sequences were evolved along T_1 in figure 1. The tree test set includes all 105 possible genealogies. Confidence sets that contain three trees correspond to those trees necessary to resolve the polytomy in figure 1.)

sequence length	c	KH	SH
100	10	15	15
500	3	3	15
1000	3	3	15
2000	3	3	15
3000	3	3	15
4000	3	3	3
5000	3	3	3

4. DISCUSSION

(a) *Multiple comparison of gene trees*

Three methods for constructing confidence sets of gene trees were compared. The KH and SH confidence sets are based on LR tests; in addition, a simple approach using expected likelihood weights as a measure of confidence in a model was described. By construction, all these methods are robust against model misspecification.

The KH confidence set and the confidence set derived from expected likelihood weights are very similar, with the latter being slightly smaller. By contrast, the SH confidence set is much more conservative and requires a large sample size to eliminate unlikely models. This was observed in the two sequence examples and also in the simulated datasets.

The SH test is conservative as it aims at multiple comparison with the unknown best model (Hsu 1996). By contrast, the KH test is a pairwise method, and special provisions are necessary for comparison with the ML model (Goldman *et al.* 2000). The method based on expected-likelihood weights is the most direct approach. It provides a simple and intuitive method for multiple comparison of models and construction of corresponding confidence sets.

(b) *Misspecification of gene trees*

Goldman *et al.* (2000) show that there can be dramatic differences in the outcome of LR tests to compare gene trees. In this paper it is argued that the contradictory results are a result of model misspecification. In particular, if either the substitution process or the actual branching pattern is incorrect in the investigated gene trees, then LR tests based on the assumption that the true data-generating model is among the candidate models may be misleading. For the sequence examples studied here and in Goldman *et al.* (2000), the reasons for misspecification are likely to be insufficient complexity of the substitution model (mtDNA) or recombination (HIV-1).

Here, it is emphasized that methods for comparing gene trees are available that are robust against model misspecification. For example, both the KH and SH tests and the method based on expected-likelihood weights do not require correct specification of the candidate gene trees. These approaches are conservative and avoid overconfidence in the ML model. Hence, unless further precautions against model misspecification have been taken, they should be preferred over parametric bootstrap tests based on one particular gene tree.

This paper benefited greatly from discussions with Nick Goldman and Tim Massingham. Valuable comments from Carsten Wiuf, Robert Freckleton and two anonymous referees were also highly appreciated. This work was supported by an Emmy Noether research fellowship by the Deutsche Forschungsgemeinschaft (K.S.) and the Wellcome Trust (A.R.).

APPENDIX A: COMPUTER PROGRAMS AND IMPLEMENTATION

All described methods for the robust construction of confidence sets of gene trees (KH and SH methods, expected-likelihood weights) have been implemented in Java and are available in the software library PAL (Drummond & Strimmer 2001); see the PAL Web page at http://www.pal-project.org for further details.

Direct implementation of equation (2.2) for computing the likelihood weight is not possible in most standard programming languages (this would require accurate arithmetic for extremely small floating point numbers). This problem is circumvented by rewriting equation (2.2) as

$$w_i = \frac{e^{l_i - l_{\max}}}{\sum_{j=1}^{r} e^{l_j - l_{\max}}}, \qquad \text{(A 1)}$$

where $l_i = \log L_i$ and $l_{\max}$ is the log likelihood of the ML model.

REFERENCES

Adachi, J. & Hasegawa, M. 1996 Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42**, 459–468.

Akaike, H. 1974 A new look at the statistical model identification. *IEEE Trans. Automat. Cont.* **19**, 716–723.

Bar-Hen, A. & Kishino, H. 2000 Comparing the likelihood functions of phylogenetic trees. *Ann. Inst. Statist. Math.* **52**, 43–56.

Birnbaum, A. 1962 On the foundations of statistical inference. *J. Am. Statist. Assoc.* **57**, 269–326.

Breiman, L. 1996 Bagging predictors. *Machine Learning* **24**, 123–140.

Buckland, S. T. 1984 Monte Carlo confidence intervals. *Biometrics* **40**, 811–817.

Cox, D. R. 1961 Tests of separate families of hypotheses. In *Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, pp. 105–123. Berkeley, CA: University of California Press.

Cox, D. R. 1962 Further results on tests of separate families of hypotheses. *J. R. Statist. Soc. Ser.* B **24**, 406–424.

DiCiccio, T. J. & Efron, B. 1996 Bootstrap confidence intervals (with discussion). *Statist. Sci.* **11**, 189–228.

Drummond, A. & Strimmer, K. 2001 PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics* **17**, 662–663.

Edwards, A. W. F. 1972 *Likelihood.* Cambridge University Press.

Efron, B. 1982 *The jacknife, the bootstrap and other resampling plans*. Philadelphia, PA: SIAM.

Efron, B. 1987 Better bootstrap confidence intervals (with discussion). *J. Am. Statist. Assoc.* **82**, 171–200.

Efron, B., Halloran, E. & Holmes, S. 1996 Bootstrap confidence levels for phylogenetic trees. *Proc. Natl Acad. Sci. USA* **93**, 13 429–13 434.

Efron, B. & Tibshirani, R. J. 1993 *An introduction to the bootstrap*. London: Chapman and Hall.

Efron, B. & Tibshirani, R. J. 1997 Improvements on cross-validation: the 632+ bootstrap method. *J. Am. Statist. Assoc.* **92**, 548–560.

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

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

Foutz, R. V. & Srivastana, R. C. 1977 The performance of the likelihood ratio test when the model is incorrect. *Ann. Statistics* **5**, 1183–1194.

Garthwaite, P. H., Jolliffe, I. T. & Jones, B. 1995 *Statistical inference*. London: Prentice-Hall.

Golden, R. M. 1995 Making correct statistical inferences using a wrong probability model. *J. Math. Psychol.* **38**, 3–20.

Goldman, N. 1993 Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**, 182–198.

Goldman, N., Anderson, J. P. & Rodrigo, A. G. 2000 Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* **49**, 652–670.

Hacking, I. 1965 *Logic of statistical inference*. Cambridge University Press.

Hendy, M. D., Penny, D. & Steel, M. A. 1994 Discrete Fourier analysis for evolutionary trees. *Proc. Natl Acad. Sci. USA* **91**, 3339–3343.

Hillis, D. M. & Bull, J. J. 1993 An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**, 182–192.

Hsu, J. C. 1996 *Multiple comparisons—theory and methods*. London: Chapman and Hall.

Huelsenbeck, J. P. & Rannala, B. 1997 Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**, 227–232.

Kent, J. T. 1982 Robust properties of likelihood ratio tests. *Biometrika* **69**, 19–27.

Kimura, M. 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120.

Kishino, H. & Hasegawa, M. 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.

Linhart, H. & Zucchini, W. 1986 *Model selection*. New York: Wiley.

Myung, I. J., Balasubramanian, V. & Pitt, M. A. 2000 Counting probability distributions: differential geometry and model selection. *Proc. Natl Acad. Sci. USA* **97**, 11 170–11 175.

Robertson, D. L., Sharp, P. M., McCutchan, F. E. & Hahn, B. H. 1995 Recombination in HIV-1. *Nature* **374**, 124–126.

Shimodaira, H. 1998 An application of multiple comparison techniques to model selection. *Ann. Inst. Statist. Math.* **50**, 1–13.

Shimodaira, H. 2001 Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection. *Comm. Stat.* A **30**, 1751–1772.

Shimodaira, H. & Hasegawa, M. 1999 Multiple comparison of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116.

Strimmer, K. & Moulton, V. 2000 Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.* **17**, 875–881.

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

White, H. 1982*a* Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

White, H. 1982*b* Regularity conditions for Cox's test of non-nested hypotheses. *J. Econometrics* **19**, 301–318.

Wilks, S. S. 1938 The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**, 60–62.

Yang, Z. 1994 Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105–111.