Confidence Regions and Hypothesis Tests for Topologies Using Generalized Least Squares

Edward Susko

Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia

A confidence region for topologies is a data-dependent set of topologies that, with high probability, can be expected to contain the true topology. Because of the connection between confidence regions and hypothesis tests, implicitly or explicitly, the construction of confidence regions for topologies is a component of many phylogenetic studies. Existing methods for constructing confidence regions, however, often give conflicting results. The Shimodaira-Hasegawa test seems too conservative, including too many topologies, whereas the other commonly used method, the Swofford-Olsen-Waddell-Hillis test, tends to give confidence regions with too few topologies. Confidence regions are constructed here based on a generalized least squares test statistic. The methodology described is computationally inexpensive and broadly applicable to maximum likelihood distances. Assuming the model used to construct the distances is correct, the coverage probabilities are correct with large numbers of sites.

Introduction

The use of generalized least squares (GLS) for estimation and hypothesis tests about topologies was mentioned by Cavalli-Sforza and Edwards (1967) and is considered in more detail in Bulmer (1991). Given a set of distances d_{ij}, the GLS test statistic is of the form

$$\sum_{i<j,k<l} w_{ij,kl}(d_{ij} - \delta_{ij})(d_{kl} - \delta_{kl}), \qquad (1)$$

where the weights $w_{ij,kl}$ are entries of the inverse of the covariance matrix for the estimated distances. For a given topology, δ_{ij} is the sum of the branch lengths along the path from i to j; the branch lengths are chosen to minimize equation (1). Because the choice of δ_{ij} is topology dependent, the test statistic calculated in equation (1) will be dependent on the topology T of interest; call it the GLS test statistic, g_T. The reason that the GLS test statistic is more suitable for hypothesis testing and confidence region construction than, for instance, the unweighted least squares test statistic is that, under the null hypothesis the GLS test statistic has a known chi-square distribution. The null hypothesis here is that the given topology is the true topology. Implicitly the null hypothesis also assumes that the substitution model used in constructing distances is correct. Under this null hypothesis, a random g_T has a chi-square distribution with degrees of freedom $T(T-1)/2 - (2T-3)$, where T is the number of taxa. In contrast, the least squares value would have a complex distribution that requires knowledge of covariances between distances for the computation of P values.

For the GLS test statistic, a P value is calculated as the probability, under the chi-square distribution, of observing a value at least as large as g_T. This is a P value for a test of the null hypothesis that the given topology is the true topology. Equivalently, because of the more general duality between testing and confidence region construction, the GLS test statistic can be seen as providing a means for constructing confidence regions of topologies. A $(1-\alpha)\times 100\%$ confidence region for the true topology is a data dependent, and hence random, set of topologies that contains the true topology with probability $1-\alpha$. The confidence region based on the GLS test statistic is the set of all topologies with P values $\geq\alpha$.

As a brief illustration, consider the mammal data set previously considered in Shimodaira and Hasegawa (1999), Goldman, Anderson, and Rodrigo (2000), and Shimodaira (2002). There are six taxa in the data set, so that 105 topologies are possible. For each of these 105 topologies, a modification of the non-negative least squares routine of Lawson and Hanson (1974) was used to obtain the optimal δ_{ij} for equation (1) with weights $w_{ij,kl}$ calculated using the sample average approximation described below. With these δ_{ij}, the GLS test statistic was calculated for each of the topologies. The P values corresponding to these test statistics were calculated as the probabilities that a chi-squared random variable with 6 degrees of freedom is greater than the test statistic (generally, the degrees of freedom are $T(T-1)/2-(2T-3)$, where T is the number of taxa). The topologies were then ranked from largest P value to smallest P value. The topologies corresponding to P values greater than or equal to 0.05 give a 95% confidence region and are given in table 1. Note that some of the topologies have exactly the same P values. This occurred because some of the estimated branch lengths were 0, making these topologies equivalent.

The use of GLS for hypothesis tests of topologies was considered previously in Bulmer (1991). However, the methods presented there, in particular the formulas given for the covariances that give rise to the $w_{ij,kl}$, were specific to the distances described in Tajima and Nei (1984). The work presented here extends Bulmer (1991): The methods for the calculation of covariances presented are generally applicable to most maximum likelihood (ML) distances and the use of chi-square distributions for calculating P values is given a justification that will apply to most ML distances. Software derived from the PHYLIP package source code (Felsenstein 1993) that will sort a set of input trees according to their P values, as in the mammal example, is available for download at http://www.mathstat.dal.ca/~tsusko.

Although GLS could provide a framework for both estimation and testing of topologies, our primary interest is

Key words: generalized least squares, phylogeny, statistical tests.

E-mail: susko@mathstat.dal.ca.

Mol. Biol. Evol. 20(6):862–868. 2003
DOI: 10.1093/molbev/msg093
 ISSN: 0737-4038

in testing and the construction of confidence regions. Other methods that can be used to construct confidence regions for trees include the bootstrap selection probability (BP) of Felsenstein (1985), which is usually used to assign confidence levels for clades but can be used to assign confidence levels for trees. The approximately unbiased (AU) test of Shimodaira (2002) adjusts the BP for possible curvature. Hillis and Bull (1993) and Newton (1996) raise concerns about BP, however. A number of likelihood-based statistical tests have been developed, and the more commonly used ones include the Kishino-Hasegawa (KH) test (Kishino and Hasegawa 1989), the Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa 1999), and the Swofford-Olsen-Waddell-Hillis (SOWH) test; the SOWH test is a parametric bootstrap likelihood ratio test described in Swofford et al. (1996) and discussed in Goldman, Anderson, and Rodrigo (2000). Each of these tests has difficulties. The KH test is appropriate for comparisons of competing topologies but is inappropriate for the construction of confidence regions (cf. Goldman, Anderson, and Rodrigo 2000). The SH test seems to be too conservative in the sense that hypotheses are too infrequently rejected, whereas the SOWH test, in contrast, frequently rejects every tree except the ML tree (Goldman, Anderson, and Rodrigo 2000) and requires long computational times for the repeated ML fitting required. In contrast, the GLS method for testing and confidence region construction presented here does not require expensive calculation, is widely applicable, and has correct large-site coverage probabilities: as the number of sites gets large, a 95% confidence set of trees will contain the true tree with probability 0.95. As with most other testing procedures, however, the GLS method requires that the substitution model used in reconstructing the distances be correct. Poor substitution models can lead to distances that are not consistent with any trees or, worse, attribute high confidence to topologies having certain artifacts like long branch attraction.

Methods

Matrix Formulation of the GLS Test Statistic

To be clear about how the $w_{ij,kl}$ weights in equation (1) are obtained, it is necessary to give a matrix algebra expression for the test statistic. Let the taxa under consideration be labeled $1, \ldots, T$ and let $\mathbf{y}$ be the vector of distances $(d_{12}, \ldots, d_{1T}, d_{23}, \ldots, d_{T-1T})$. This vector will be estimated from the data and will not require a topology for estimation. In contrast, the corresponding vector $\boldsymbol{\delta}$ of distances consistent with a topology will depend on what topology is being considered. Fixing a particular topology, let the vector of branch lengths of that topology be $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{2T-3}]^T$. The distance for the kth pair, say taxa i and taxa j, is the sum of the branch lengths for branches along the path from i to j. This can be expressed as

$$\delta_{i,j} = \sum_l x_{kl}\alpha_l, \tag{2}$$

where x_{kl} is 1 if the lth branch is in the path from i to j and 0 otherwise. Let X be the matrix with k, l entry x_{kl}. Then equation (2) can be expressed in matrix notation as

$$\boldsymbol{\delta} = X\boldsymbol{\alpha}.$$

Note that the ordering of the branch lengths is not important but that a change of ordering would have to be accompanied by a corresponding change in the ordering of the columns of the X matrix. We will let V denote the covariance matrix for the estimated branch length vector $\mathbf{y}$: the i, j entry of V is the covariance between the distance for pair i and pair j, and the ith diagonal entry is the variance for the ith distance y_i. The GLS test statistic can then be expressed as

$$(\mathbf{y} - X\boldsymbol{\alpha})^T V^{-1} (\mathbf{y} - X\boldsymbol{\alpha})$$

This expression is the same as the one given in equation (1) with $w_{ij,kl}$ equal to the r, s entry of the V^{-1} matrix, where r is the index for the (i, j) pair and s is the index for the (k, l) pair.

In most applications of GLS outside of phylogenetic estimation, the vector $\boldsymbol{\alpha}$ is unrestricted. Here, as a vector of branch lengths, it is desirable to restrict the $\boldsymbol{\alpha}$ vector to be non-negative. In practice this can give estimates of branch lengths that differ from the usual GLS estimates. In theory, as long as the branch lengths for the true tree are non-negative, which we will assume throughout, with a large number of sites, the difference between the unrestricted GLS estimate and the restricted GLS estimate will be negligible for the true topology.

The GLS Test Statistic and Its Chi-Square Distribution

If the $\mathbf{y}$ vector is normally distributed, the GLS test statistic will have a χ^2 distribution with degrees of freedom equal to the difference of the length of the $\mathbf{y}$ vector and the length of the $\boldsymbol{\alpha}$ vector. In the present case this gives $T(T-1)/2 - (2T-3)$ degrees of freedom. However, this result requires that the $\mathbf{y}$ vector be at least approximately normally distributed. In the phylogenetic applications being considered here, the $\mathbf{y}$ vector is a vector of distances that we will assume are ML distances with respect to some substitution model.

THEOREM. *Assume that the vector* $\mathbf{y}$ *of distances is a set of ML distances. Then, with a large number, n, of sites, the distribution of* $\mathbf{y}$ *is approximately multivariate normal. Let* $p_j(x; d)$ *denote the probability of the data x at a site for the jth pair of taxa., calculated when the distance between them is d. Then the variance of the jth estimated distance is*

$$V_{jj} = Var(y_j) = \left(E\left\{ -\frac{\partial^2}{\partial d_j^2} \log[p_j(x; d_j)] \right\} \right)^{-1} \Big/ n. \tag{3}$$

The covariance between the jth distance and the kth distance is

$$V_{jk} = nV_{jj}V_{kk}E\left\{ \frac{\partial}{\partial d_j} \log[p_j(x; d_j)] \frac{\partial}{\partial d_k} \log[p_k(x; d_k)] \right\}. \tag{4}$$

PROOF. The result is very similar to general ML estimation results like that of Lehman (1983, Theorem 4.1,

p. 429). The main difference arises because, whereas ML estimation is usually done jointly for all of the unknown parameters of interest, ML distance estimation is done separately for each of the distances.

Let the *score function* for the jth pair be

$$u_j(d) = \sum_{i=1}^{n} \frac{\partial}{\partial d} \log[p_j(x_i; d)],$$

where there are n sites. Then general ML results give that

$$\sqrt{n}(y_j - d_j) = -\sqrt{n}V_{jj}u_j(d_j) + o_p(1), \qquad (5)$$

where d_j is the true unknown distance between the pair. Let $\mathbf{u}^*(\mathbf{d})$ be the vector with jth entry $-\sqrt{n}V_{jj}u_j(d_j)$; then equation (5) can be combined across all pairs in vector notation as

$$\sqrt{n}(\mathbf{y} - \mathbf{d}) = \mathbf{u}^*(\mathbf{d}) + o_p(1).$$

The quantity $\sqrt{n}\mathbf{u}^*(\mathbf{d})$ is a sum of independent random variables and so by the multivariate central limit theorem, $\mathbf{u}^*(\mathbf{d})$ has a multivariate normal distribution with covariance matrix nV where the entries of V are as given in equations (3) and (4). Since, up to higher order terms, $\sqrt{n}(\mathbf{y} - \mathbf{d})$ is equal to $\mathbf{u}^*(\mathbf{d})$, the result follows. ■

The implication of the large-site normality of $\mathbf{y}$ is that, under the null hypothesis that the true topology is the one that is being used to construct the GLS test statistic, the GLS test statistic has a chi-square distribution with $T(T-1)/2 - (2T-3)$ degrees of freedom. Because the entries of the covariance matrix V are not known, they need to be estimated. As long as they are estimated with statistically consistent methods of estimation, the chi-square limiting distribution will still be applicable. Because this is the distribution used to calculate the P values, it follows that a P value calculated for the true topology will be greater than α, and hence the true topology will be contained in the confidence region, $(1-\alpha) \times 100\%$ of the time. Thus, in theory, the coverage probabilities (the probability that the true topology is in the confidence region) will be correct with large numbers of sites. In practice, the coverage probabilities will differ from $1-\alpha$ for 3 reasons: (1) The number of sites is always finite. (2) The substitution model used in constructing distances is always an approximation to the true substitution process. (3) In theory, V^{-1} and all other quantities are calculated exactly. In practice they contain numerical errors.

Estimation of the Covariances Matrix V

A complication enters into construction of the GLS test statistic because the covariances V_{ij} depend on the unknown topological relationship between the 4 taxa involved in the two pairs. We present here two methods for estimating the covariance matrix entries.

Sample Average Method

The first method of estimation utilizes the forms of equations (3) and (4) and uses a sample average to approximate these expectations. The estimate of the variance is given as

$$\hat{V}_{jj} = \left\{ n^{-1} \sum_{i=1}^{n} -\frac{\partial^2}{\partial d^2} \log[p_j(x_i; d)] \right\}^{-1} \Big/ n,$$

and the estimate of the covariance is

$$\hat{V}_{jk} = \hat{V}_{jj}\hat{V}_{kk} \sum_{i=1}^{n} \frac{\partial}{\partial d_j} \log[p_j(x_i; d_j)] \frac{\partial}{\partial d_k} \log[p_k(x_i; d_k)].$$

Bootstrap Estimation

The second method of estimation is a nonparametric bootstrap method (Efron and Tibshirani 1993).

1. From the original data set of n sites, select n sites at random and *with replacement*, giving a new data set.
2. For the new data set calculate the ML estimates. Label the vector of estimated distances $\mathbf{y}^{(b)}$ to distinguish it from the original vector of estimated distances, $\mathbf{y}$.
3. Repeat steps 1 and 2 B times where B is fairly large. This will give a set of bootstrapped distance estimates $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(B)}$.
4. The estimates of the variances are then

$$\hat{V}_{jj} = B^{-1} \sum_{b=1}^{B} (y_j^{(b)} - y_j)^2, \qquad (6)$$

and the estimates of the covariances are

$$\hat{V}_{jk} = B^{-1} \sum_{b=1}^{B} (y_j^{(b)} - y_j)(y_k^{(b)} - y_k). \qquad (7)$$

The bootstrap estimate of V is easier to implement but requires much longer computational times. Both methods are large-site approximations to V, although they can be expected to give reasonable results with moderate numbers of sites if the number of taxa is not too large.

Implementation

The calculation of the GLS test statistic requires that branch lengths be calculated once for each topology considered. In contrast, the calculation of bootstrap support or the implementation of the SOWH test requires repeated fitting of branch lengths for bootstrap samples from the original data. Thus the GLS test will generally be relatively fast compared to these procedures, especially by comparison with cases when ML estimation is used. GLS estimation will, however, require more computation than ordinary or weighted least squares estimation. Because the calculation can be implemented by converting to an ordinary least squares problem, one can consider the additional cost more carefully.

The calculation of the covariance matrix V requires on the order of nT^4 operations (additions and multiplications). Because the matrix V is a covariance matrix, it is positive definite, and so a Cholesky decomposition $V = U^T U$ can be obtained for it, where U is an upper triangular matrix; calculation of U requires on the order

Table 1
The Topologies for the Mammal Data Set Having a *P* Value >0.00001

GLS Test Statistic	*P* Value	Topology
6.12	0.410	((((harbor seal, cow), human), rabbit), (mouse, opposum))
6.40	0.380	((((mouse, opposum), human), rabbit), (harbor seal, cow))
6.66	0.353	((human, rabbit), (harbor seal, cow), (mouse, opposum))
12.58	0.050	2 topologies equivalent to (((human, mouse, opposum), rabbit), (harbor seal, cow))
14.51	0.024	2 topologies equivalent to (((harbor seal, cow), human), (mouse, rabbit, opposum))
16.05	0.013	2 topologies equivalent to ((human, opposum), (harbor seal, cow), (rabbit, mouse))
16.14	0.013	6 topologies equivalent to ((human, mouse, rabbit, opposum), (harbor seal, cow))

of T^6 additional operations. The problem, originally stated as that of minimizing

$$(\mathbf{y} - X\boldsymbol{\alpha})^T V^{-1}(\mathbf{y} - X\boldsymbol{\alpha})$$

can then be restated as minimizing

$$(U^{-1}\mathbf{y} - U^{-1}X\boldsymbol{\alpha})^T(U^{-1}\mathbf{y} - U^{-1}X\boldsymbol{\alpha}). \quad (8)$$

Solving the triangular systems of equations

$$\mathbf{y}^* = U\mathbf{y} \qquad X^* = U\boldsymbol{X},$$

which requires on the order of T^5 operations, allows one to express equation (8) as

$$(\mathbf{y}^* - X^*\boldsymbol{\alpha})^T(\mathbf{y}^* - X^*\boldsymbol{\alpha}),$$

which is an ordinary least squares problem with $\mathbf{y}^*$ and $\boldsymbol{X}^*$ replacing $\mathbf{y}$ and $\boldsymbol{X}$. Note that the computation of V, U, and $\mathbf{y}^*$ needs to be done only once for a given data set; it does not matter how many topologies the GLS test statistic is calculated for.

In theory, the GLS test statistics can be calculated with or without the restriction that the branch lengths be non-negative. However, the simulation results of Kuhner and Felsenstein (1994) indicate that branch lengths and topologies are more accurately estimated with non-negativity restrictions using Fitch-Margoliash least squares; because of the similarity of the forms of estimation, it can be expected that those results would generalize to GLS estimation. Perhaps more important, while for ''good'' topologies branch lengths can expected to be non-negative or close to 0, for very poorly supported topologies this need not be the case, and thus negative branch lengths allow additional freedom in fitting estimated distances that could give rise to small GLS test statistics. The version of GLS used here imposes the non-negativity constraint using a version of the NNLS routine of Lawson and Hanson (1974), a globally convergent method that successively solves unrestricted least squares problems with successive subsets of branches set to 0 or, equivalently, ignored in the estimation.

Results

As examples we will consider two data sets, the amino acid mammal data set considered in Shimodaira and Hasegawa (1999), Goldman, Anderson, and Rodrigo (2000), and Shimodaira (2002) and the nucleotide data set considered in Goldman, Anderson, and Rodrigo (2000).

Mammal Data

The mammal data set consisted of 3,414 aligned amino acids from six mammals: human, harbor seal, cow, rabbit, mouse, and opossum. The PAM substitution model (Dayhoff, Schwartz, and Orcutt 1979) as implemented in PHYLIP 3.6a2 (Felsenstein 1993) was used, and a gamma rates-across-sites correction was used in calculating the distances. The α parameter for the gamma model was estimated using Tree-Puzzle version 4.02 (Strimmer and von Haeseler 1996) as 0.34. Because there are six taxa, there are 105 possible topologies. The GLS test statistics were obtained for all 105 topologies and the topologies were ranked from smallest GLS statistic to largest GLS test statistic. Again because there are six taxa, the *P* values for the individual tests for the topologies were calculated using a chi-square distribution with $6 \cdot 5/2 - (2 \cdot 6 - 3) = 6$ degrees of freedom. The topologies with *P* values ≥ 0.05 give a 95% confidence region for the true topology and are tabulated in table 1.

The GLS values gave very clear evidence for the grouping (harbor seal, cow). Each of the topologies that did not have this pair had a *P* value <0.00001. Note that this illustrates more generally how to test an alternative hypotheses with a less specific topological structure: the *P* value for the test is the maximum *P* value from GLS test statistics calculated for all of the topologies inconsistent with the topological structure. In the present example the alternative hypothesis is that the grouping (harbor seal, cow) is present in the true topology and the *P* value is calculated as 5×10^{-6}, the largest *P* value from the set of 95 six-taxa topologies without the grouping (harbor seal, cow).

In theory the GLS test statistics can be calculated with or without the restriction that the branch lengths be non-negative. We expect that the GLS statistic will be better able to detect very poor topologies with the non-negativity restriction imposed and have chosen to implement a variation of the NNLS routine of Lawson and Hanson (1974) in order to accomplish this. This is the reason that some of the topologies estimated have exactly the same *P* values. Some branch lengths for these topologies were estimated as 0, making them equivalent.

The PAM substitution model used here is different from the mtREV model (Adachi and Hasegawa 1996*a*) that was used in Shimodaira and Hasegawa (1999), Goldman, Anderson, and Rodrigo (2000), and Shimodaira (2002). Nevertheless the main conclusions and results are in general agreement. Differences arise only for inferences where there is no agreement between competing methods.

The grouping (harbor seal, cow) inferred here was an a priori feature of the topologies considered in Shimodaira and Hasegawa (1999), Goldman, Anderson, and Rodrigo (2000), and Shimodaira (2002). The rankings of the top

Table 2
The Topologies for the HIV Data Set Having a *P* value >0.00001

GLS Test Statistic	*P* Value	Topology
15.53	0.017	((((E1, E2), A2), A1), (D, B))
18.58	0.005	((E1, E2), (A1, A2), (D, B))

three topologies, done here on the basis of the GLS test statistic, are the same as the rankings of topologies from the log likelihoods. All three of these topologies are included in the GLS 95% confidence region for trees, and the same holds true for the AU test and SH test reported in Shimodaira and Hasegawa (1999) and Shimodaira (2002). The BP test, however, rejects the hypothesis that the third topology is the true topology, and a big difference arises between the other methods and the SOWH test for the second topology, which gave a P value of <0.001; the P value here is 0.380 and was greater than 0.30 for all of the tests reported in Shimodaira (2002).

Differences arise with the remaining topologies. The version of GLS used here, with branch lengths restricted to be non-negative, seems much more likely to assign zero branch lengths than ML estimation, and so some of the different topologies reported in SH are equivalent here. The topology (((human,mouse,opposum),rabbit), (harbor seal,cow)) gives a P value of 0.050 and is included marginally in a 95% confidence region. The remaining topologies would not be. In this case the GLS region seems to provide a trade-off between the SOWH test, which only includes the ML topology in a 95% confidence region, and the SH test, which as reported in Shimodaira (2002), includes the 15 topologies with a (harbor seal, cow) grouping as well as, marginally, an additional topology without that grouping.

HIV Data

The HIV data set consisted of a set of six homologous sequences, each containing 2,000 base pairs from the *gag* and *pol* genes for isolates of HIV-1 subtypes A, B, D, and E: A1 (Q23), A2 (U455), B (BRU), D (NDK), E1 (90CF11697), and E2 (93TH057). Maximum likelihood distances were calculated using an F84 model (Felsenstein 1984) and gamma rate correction. The transition/transversion ratio was estimated as 4.70; the α parameter for the gamma rate distribution was estimated as 0.23 in Tree-Puzzle. Only one topology is included in a 99% confidence region for the topologies, although one other is marginally excluded. Interestingly, the P value for the best GLS topology is only 0.016, which suggests that this topology can be rejected, leaving no suitable topology. This may be in indication of a lack of fit for the F84 model for this data. Generally, the possibility of a poor substitution model giving a small P value needs to be considered when applying the test.

The SOWH and SH tests were applied in Goldman, Anderson, and Rodrigo (2000) for the null hypothesis that the second topology listed in table 2 was the true topology. The SH test gave a P value of 0.002, smaller than the GLS P value 0.005 but resulting in the same conclusion. The SH test P value was 0.26; however, as reported in Buckley (2002), this P value was highly dependent on the substitution process used and was 0.077 for the HKY85 model, which is similar to the F84 model considered here.

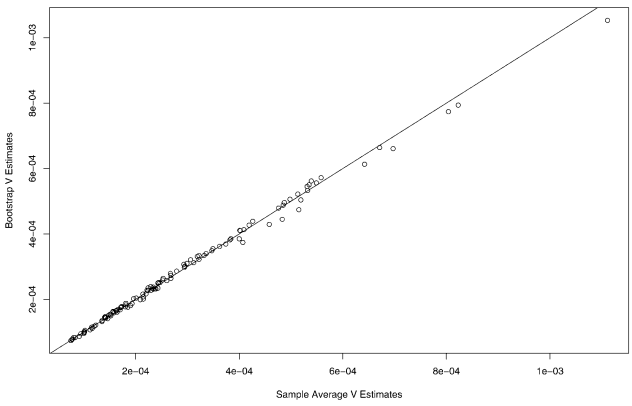

FIG. 1.—The bootstrap estimates with $B = 5000$ bootstrap replicates of the entries of the V matrix plotted against the corresponding estimates from the sample average method for the mammal data set.

Estimation of Covariance Matrices

The results for the HIV data set and for the mammal data set used the sample average method for the calculation of the covariance matrix V. To cross-check the estimation of the covariance matrix, we did the estimation using the bootstrap method. Plots of the estimated covariance matrices from the bootstrap method against the estimated covariance matrices from the sample average method are given in figures 1–3. The estimates are very similar. Interestingly, a comparison of figures 2 and 3 suggests that the sample average method may give better bootstrap estimates than a bootstrap with a smaller number of bootstrap replicates. The bootstrap estimates in figure 3, being based on $B = 5,000$ bootstrap replicates, should be considered better than the bootstrap estimates in figure 2, which are based on $B = 100$. The sample average estimates are more similar to the bootstrap estimates with $B = 5,000$ that with $B = 100$, so much similar that it is

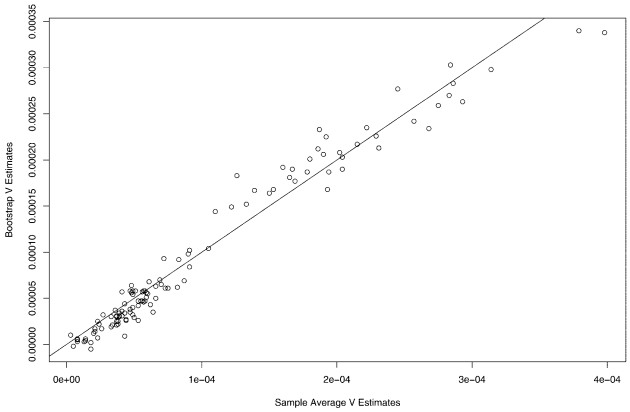

FIG. 2.—The bootstrap estimates with $B = 100$ bootstrap replicates of the entries of the V matrix plotted against the corresponding estimates from the sample average method for the HIV data set.

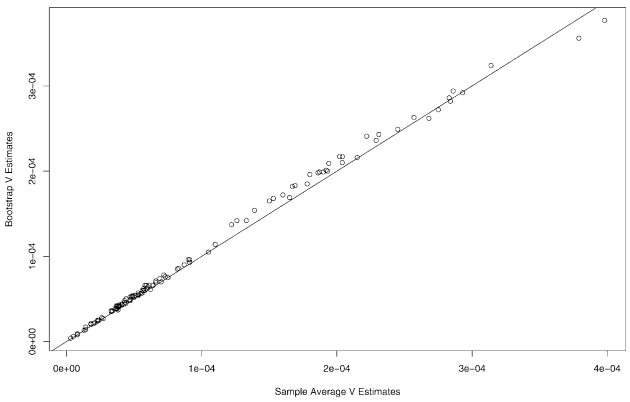

FIG. 3.—The bootstrap estimates with $B = 5000$ bootstrap replicates of the entries of the V matrix plotted against the corresponding estimates from the sample average method for the HIV data set.

clear that they provide better approximations to the bootstrap estimates with $B = 5,000$ than the estimates with $B = 100$.

Discussion

The GLS method for constructing confidence regions for topologies provides a reasonable way of addressing uncertainty in topological estimation. It is does not require long computational times and, assuming the correct substitution model is used in constructing the distances, is theoretically sure to have the correct coverage probability. With poor substitution models, it is possible that the GLS test statistics will reflect a lack of fit and be larger than expected, giving rise to smaller coverage probabilities.

The methods presented here are an extension of Bulmer (1991) that allow for general ML distances and provide more general methods for the calculations of the V matrix. One other difference is the restriction of branch lengths to be non-negative, which was not a feature of the Bulmer implementation. For estimation, this restriction is not expected to be very important; the true topology can be expected to have branch lengths that are estimated as almost all non-negative anyway. However, in constructing confidence regions it is possible that for poor topologies there will exist choices of $\boldsymbol{\alpha}$ vectors with negative entries that make no topological sense but give small GLS test statistics.

Two methods for the calculation of covariance matrices have been presented: a bootstrap method and a sample average method. The examples considered here, and others not reported here, suggest that the estimates from either of these methods will be very similar. Because the bootstrap can require much longer computational times, we have chosen to implement the sample average method in the available software.

The theory behind the methods presented here assumes that the number of sites is large. In practice it will usually be the case that the number of sites required will be more if there are a large number of taxa than if there are a small number of taxa. This is so because the number of variances and covariances requiring estimation in the V matrix increases with the number of taxa, so that the effects of errors in estimation of the weights in the original GLS test statistic given in equation (1) are aggregated over more terms. Careful analysis decisions may allow one to avoid difficulties associated with larger numbers of taxa without losing much information. One strategy is to use a full data set to estimate transition/transversion ratios, rate distributions, and any other auxilliary parameters in the model. Once estimates for these parameters have been obtained, the number of taxa in the data set can be reduced by taking out closely related taxa, leaving a few representative taxa to address the topological questions of interest. One must be cautious, however, in reducing the number of taxa. Several studies involving four taxon subsamples with a constant set of three taxa but differing choices for an outgroup taxon indicate that inferences about the placement of the three taxa is highly dependent on the choice of the outgroup taxon (Philippe and Douzery 1994; Adachi and Hasegawa 1995, 1996*b*).

In some cases, investigators will have a relatively small prespecified set of topologies that are of interest because of the results of previous analyses. These topologies can be tested for inclusion in, say, a 95% confidence region by checking whether P values are ≥ 0.05 or not. In other cases, the entire 95% confidence region is desired. In principle this requires calculation of the GLS test statistics for all possible topologies, which is feasible with a relatively small number of taxa but becomes infeasible with larger numbers of taxa. With larger numbers of taxa, tree search algorithms (stepwise addition, star decomposition, tree bisection, and reconnection; cf. Hillis, Moritz, and Mable (1996) pp. 478–485) might be implemented as they are for least squares estimation and ML estimation. Keeping all trees found in the search with P values ≥ 0.05 would give an approximate 95% confidence region. The advantage with this approach is that the number of trees for which GLS test statistics are calculated would be relatively small. However, because trees with large GLS values may be missed during a search, the region might miss trees that would have been in the region had an exhaustive search been done.

Acknowledgments

The author was supported by the Natural Sciences and Engineering Research Council of Canada. This work is part of a Genome Atlantic/Genome Canada Large-Scale Project.

Literature Cited

Adachi, J., and M. Hasegawa. 1995. Phylogeny of whales: dependence of the inference on species sampling. Mol. Biol. Evol. **12**:177–179.

———. 1996*a*. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

———. 1996*b*. Instability of quartet analyses of molecular sequence data bit the maximum likelihood method: the Cetace/Artidactyla relationships. Mol. Phyl. Evol. **6**:72–76.

Buckley, T. J. 2002. Model misspecification and probabilistic

tests of topology: evidence from empirical data sets. Syst. Biol. **51**:509–523.
Bulmer, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. **8**:868–883.
Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic analysis: models and estimation prodcedures. Am. J. Hum. Genet. **19**:233–257.
Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1979. A model of evolutionary change in proteins. Pp 345–352 *in* M. O. Dayhoff, ed. Atlas of Protein Sequence and Structure, Vol. 5. National Biomedical Research Foundation, Silver Spring, Md.
Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman & Hall, New York.
Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.
———. 1993. PHYLIP (phylogeny inference package). Version 3.6a2. Distributed by the author, Department of Genetics, University of Washington, Seattle.
Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. **49**:652–670.
Hillis, D., and J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. **42**:182–192.
Hillis, D. M., C. Moritz, and B. K. Mable. 1996. Molecular systematics, 2nd edition. Sinauer, Associates, Sunderland, Mass.
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.
Kuhner, M. K., and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.
Lawson, C. L., and R. J. Hanson. 1974. Solving least squares problems. Prentice-Hall, Englewood Cliffs, N.J.
Lehmann, E. L. 1983. Theory of point estimation. Wiley, New York.
Newton, M. A. 1996. Bootstrapping phylogenies: large deviations and dispersion effects. Biometrika **83**:315–328.
Philippe, H., and E. Douzery. 1994. The pitfalls of molecular phylogeny based on four species as illustrated by the Cetacea/Artiodactyla relationships. J. Mamm. Evol. **2**:133–152.
Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. **51**:492–508.
Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114–1116.
Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.
Swofford, D. L., G. J., Olsen, P. J. Waddel, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. Hilllis, C. Mortiz, and B. K. Mable, eds. Molecular systematics, 2nd edition, Sinauer Associates, Sunderland, Mass.
Tajima, F., and M. Nei. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**:269–285.

Manolo Gouy, Associate Editor

Accepted January 21, 2003