Model Misspecification and Probabilistic Tests of Topology: Evidence from Empirical Data Sets

THOMAS R. BUCKLEY

Department of Biology, Duke University, Durham, North Carolina, USA, and Landcare Research, Mt. Albert, Auckland, New Zealand (address for correspondence); E-mail: buckleyt@landcareresearch.co.nz

Abstract.—Probabilistic tests of topology offer a powerful means of evaluating competing phylogenetic hypotheses. The performance of the nonparametric Shimodaira–Hasegawa (SH) test, the parametric Swofford–Olsen–Waddell–Hillis (SOWH) test, and Bayesian posterior probabilities were explored for five data sets for which all the phylogenetic relationships are known with a very high degree of certainty. These results are consistent with previous simulation studies that have indicated a tendency for the SOWH test to be prone to generating Type 1 errors because of model misspecification coupled with branch length heterogeneity. These results also suggest that the SOWH test may accord overconfidence in the true topology when the null hypothesis is in fact correct. In contrast, the SH test was observed to be much more conservative, even under high substitution rates and branch length heterogeneity. For some of those data sets where the SOWH test proved misleading, the Bayesian posterior probabilities were also misleading. The results of all tests were strongly influenced by the exact substitution model assumptions. Simple models, especially those that assume rate homogeneity among sites, had a higher Type 1 error rate and were more likely to generate misleading posterior probabilities. For some of these data sets, the commonly used substitution models appear to be inadequate for estimating appropriate levels of uncertainty with the SOWH test and Bayesian methods. Reasons for the differences in statistical power between the two maximum likelihood tests are discussed and are contrasted with the Bayesian approach. [Bayesian statistics; Markov chain Monte Carlo; maximum likelihood; nucleotide substitution models; parametric bootstrapping; SH test; SOWH test; statistical tests.]

In molecular phylogenetics, a topology estimated from one character partition will often conflict with a topology estimated from other characters, or from the evolutionary relationships implied by taxonomy or geographic distributions. In such cases it is useful to construct some form of confidence limits on the optimal topology and determine whether this region of uncertainty includes alternative topologies of interest. Three probabilistic approaches to testing such hypotheses include parametric maximum likelihood (ML) tests, nonparametric ML tests, and Bayesian posterior probabilities. Parametric ML tests of topology have been described by Huelsenbeck and Bull (1996), Huelsenbeck et al. (1996), Swofford et al. (1996), and Goldman et al. (2000). These approaches use Monte Carlo simulation to generate the null distribution of the test statistic under a fully specified evolutionary model. Goldman et al. (2000) refer to the parametric ML test of topology as the Swofford–Olsen–Waddell–Hillis (SOWH) test, as briefly described by Swofford et al. (1996), and this usage is followed here.

The second ML approach is a nonparametric test described by Shimodaira and Hasegawa (1999; SH test), related to the more commonly used KH test (Kishino and Hasegawa, 1989). The KH test is applicable to the comparison of two a priori hypotheses only, whereas the SH test can be applied to a set of topologies that contains a mixture of a priori and a posteriori topologies. The SH test also corrects for the comparison of multiple topologies. The biological utility of the SH test is illustrated by Buckley et al. (2001).

The third approach is based on Bayesian statistics coupled with Markov-chain Monte Carlo (MCMC) methods (Rannala and Yang, 1996; Yang and Rannala, 1997; Larget and Simon, 1999; Mau et al., 1999; Li et al., 2000; Huelsenbeck and Ronquist, 2001). Bayesian methods are claimed to have several advantages over traditional, ML-based hypothesis testing: Some argue that posterior probabilities provide a more intuitive estimate of uncertainty than do P-values (Gelman et al., 1995; Lewis, 2001); moreover, the use of MCMC to sample from the posterior distribution provides a measure of uncertainty in a point estimate of topology without having repeatedly to reoptimize the likelihood over multiple bootstrap replicates. Finally, MCMC methods offer a convenient means for incorporating multiple sources of uncertainty into the calculation of probabilities.
(Huelsenbeck and Bollback, 2001), which is much more difficult to achieve in a frequen-
talist framework. However, three potential pitfall in quantifying prior knowledge, misspecification of the model, and failure of the MCMC algorithm to achieve convergence on the posterior distrib-
ution (reviewed by Gelman et al., 1995). Systematists therefore need to start accumulat-
ing information regarding both limitations and strengths of Bayesian methods before use of these methods becomes widespread.

This study addresses two issues central to probabilistic tests of topology. First, the
different types of tests can indicate quite different biological conclusions (Goldman et al., 2000; Shimodaira, 2001; Strimmer and Rambaut, 2001). In part, however, this ref-
lects the fundamental statistical differences between the likelihood and Bayesian ap-
proaches, and indeed between parametric and nonparametric ML tests. Second, all the
probabilistic tests examined here require a substitution model for the calculation of like-
lihoods. Not surprisingly, therefore, the as-
sumptions of these substitution models can have large effects on the inferred results of
the tests themselves (Sullivan and Swoford, 1997), although this effect has yet to be exam-
ined in a Bayesian framework.

Here, the statistical properties of ML and Bayesian tests of topology are investigated by using five data sets about which we can be confident of most of the phylogenetic rel-
ationships among taxa and another data set that has been analyzed by Goldman et al.
(2000). I have assumed a priori that all the substitution models implemented here are mis-
specified to some extent. This makes it possible, therefore, to examine the robust-
ness of various tests when their assumptions are violated, as is the case with real data.
Additionally, several of these data sets are characterized by branch length heterogene-
ity, which can pose a problem for any phy-
logenetic method. Although drawing gen-
eral conclusions from a small number of data sets can be problematic, the “known phy-
genotype” approach (e.g., Russo et al., 1996; Cunningham, 1997) can be viewed as com-
plementary to simulation studies of phyloge-
netic methods (e.g., Huelsenbeck and Hillis,
1993).

The results presented here agree with those of previous studies (e.g., Goldman et al.,
2000; Strimmer and Rambaut, 2001) in find-
ing that the SH test is much more conserva-
tive than the SOWH test, and that the SOWH test is much more susceptible to Type I errors when the null hypothesis is true (i.e., when the suboptimal topology is correct). Simi-
larly, I demonstrate several examples where
the Bayesian approach is misleading and
again attribute this failure to misspecification of the likelihood model.

**Materials and Methods**

**Data Sets**

**HIV gag and pol sequences.**—Goldman et al. (2000) analyzed six gag and pol sequences from four subtypes (A, E, B, and D) of the HIV virus. Because of disagreement in the lit-
erature as to whether each of these four sub-
types is monophyletic (e.g., Anderson et al.,
2000), the true topology for this data set is not known with certainty. However, these sequences have been analyzed by Goldman et al. (2000) and Strimmer and Rambaut (2001) and so are included here for compar-
isons. The two trees compared are shown in

**Figure 1a.**

**Primate mitochondrial DNA (mtDNA) sequences.**—The alignment of 888 base pairs (bp) of mtDNA sequences from nine primate species published by Hayasaka et al. (1988) and distributed with the PAML package (Yang, 1997) was examined. Yang
and Rannala (1997) and Larget and Simon
(1999) have also analyzed this data set by
Bayesian methods. The relationships among
these primate species are corroborated by a wide variety of molecular markers (e.g.,
Shoshani et al., 1996; Goodman et al., 1998;
Waddell et al., 1999b; Zietkiewicz et al.,
1999; Satta et al., 2000; Murphy et al., 2001; Page
and Goodman, 2001) and morphological
and fossil data (Goodman et al., 1998). I
compared the support for the well-supported
topology and also for an incorrect topology in
which the chimpanzee and gorilla are
presented as sister species (Fig. 1b).

**Vertebrate tRNA sequences.**—In a phyloge-
netic analysis of protein and nucleotide se-
quencies from whole mitochondrial genomes, several authors (Cao et al., 1998; Rasmussen and Arnason, 1999a,b; Waddell et al., 1999b) have obtained trees in which the lamprey se-
quency rooted the vertebrate tree such that
all extant fishes form a monophyletic group.
Traditional systematic studies and analyses
FIGURE 1. Assumed true and false topologies for the six data sets analyzed with branch lengths optimized under the GTR + Γ model. (a) HIV gag and pol, (b) primate mtDNA, (c) vertebrate tRNA, (d) bird mtDNA, (e) arthropod EF1α, and (f) sigmodontine rodent mtDNA. For each pair of topologies, the top tree is assumed to be true and the bottom tree is false, except for the HIV data, where the correct topology is not known with certainty.
of nuclear rRNA genes (Mallatt and Sullivan, 1998) have placed the root of the vertebrate tree between the cartilaginous fishes and the bony fishes + tetrapods. This mitochondrial data set is generally acknowledged as being problematic because the failure of various phylogenetic reconstruction methods to select the well-supported phylogeny (Cao et al., 1998). A subset of the tRNA sequences from Waddell et al. (1999b) was analyzed that included human, platypus, alligator, chicken, ostrich, trout, shark, and lamprey. The two topologies compared are shown in Figure 1c.

**Bird whole mtDNA genomes.**—The traditional consensus on bird phylogeny states that the earliest divergence among extant birds is between the Paleognathae, which includes the ratites (rhea, ostrich, kiwi, etc.), and the Neognathae, which includes all other living birds. This pattern is supported by phylogenetic analyses of morphological characters (Cracraft, 2001), immunological data (Prager and Wilson, 1980), and various gene sequences (Stapel et al., 1984; Groth and Barrowclough, 1999; García-Moreno and Mindell, 2000; van Tuinen et al., 2000). However, recent studies of sequence data from whole mitochondrial genomes have obtained trees in which the root of the avian tree occurs within the Passeriformes, or songbirds (Härliid and Arnason, 1999; Mindell et al., 1999; Waddell et al., 1999a). Here, whole mitochondrial genomes (14,043 bp) from five taxa were analyzed: alligator, rhea, falcon, duck, and Oscine songbird (Mindell et al., 1999). All substitution models favored the joining of the extremely long alligator branch to the Oscine songbird branch. The two alternative topologies are shown in Figure 1d.

**Arthropod elongation factor 1α (EF1α) sequences.**—Phylogenetic analyses of gene sequences (Friedrich and Tautz, 1995; Regier and Schultz, 1997, 2001; Schultz and Regier, 2000; Boore et al., 1998; García-Monchado et al., 1999; Kusche and Burmester, 2001) have all supported a grouping of hexapods and crustaceans (Pancrustacea) to the exclusion of chelicerates and myriapods. Nine EF1α sequences from Schultz and Regier (2000) were analyzed: those from *Periplaneta americana* (cockroach), *Pedetontus saltator* (Archaeognatha insect), *Tomocerus* sp. (collembo), *Nebalia hessleri* (Malacostracan crustacean), *Armadillidium vulgare* (Malacostracan crustacean), *Narceus americanus* (Myriapoda), *Scutigera coleoptrata* (Myriapoda), *Dysdera crocata* (spider), and *Dinothermobium pandorae* (spider). ML analyses of these species
support one of two topologies, depending on the substitution model. One topology corresponds to the well-supported phylogeny, whereas the other topology groups the two crustacean species with the collembo-lan (Fig. 1e).

Rodent mt12S rRNA sequences.—The phylogenetic relationships among a group of sigmodontine rodents from the genera *Peromyscus* and *Onychomys* are well established from analyses of both molecular and morphological data (reviewed by Sullivan et al., 1995). Sullivan et al. (1995), however, observed that data from the mitochondrial 12S rRNA gene supported an incorrect topology, regardless of the phylogenetic method used. Here I examined eight species from the genera *Peromyscus* and *Onychomys*, for which all the substitution models implemented led to the selection of an incorrect topology (Figure 1f).

Nucleotide Substitution Models

The models implemented here are those of Jukes and Cantor (1969; JC69), Kimura (1980; K80), Hasegawa et al. (1985; HKY85), and the General Time Reversible model (e.g., Yang, 1994a; GTR). Among-site variation models included equal rates, gamma-distributed rates (Yang, 1994b; GTR + Γ), invariable sites (e.g., Hasegawa et al., 1985; GTR + I), mixed invariable sites and gamma-distributed rates (Gu et al., 1995; GTR + I + Γ), and site-specific rates (e.g., Swofford et al., 1996; GTR + SSR).

The SOWH Test

The particular version of the SOWH test implemented here, following the notation of Goldman et al. (2000), is as follows: \( \delta = \text{difference in likelihood between the two trees (likelihood ratio test statistic)}, \) \( T_0 = \text{null topology and branch lengths}, \) \( T_{ML} = \text{ML topology and branch lengths}, \) \( \theta_0 = \text{fully specified substitution model optimized under the null hypothesis,} \) \( \delta^{(i)} = \text{likelihood ratio test statistic for replicate} i, \) and \( n = \text{number of simulated data sets}. \) The hypotheses are:

\[
H_0: \text{That } T_0 \text{ is the true tree} \\
H_A: \text{That another topology is the true tree}
\]

To test the significance of \( \delta \), use the following procedure. First, simulate \( n \) data sets using \( T_0 \) under the substitution model (\( \theta_0 \)). Estimate the most likely tree for each replicate data set, using \( \theta_0 \) as estimated from the original data. Calculate \( \delta^{(i)} \) for each replicate data set \( i \), and use these values to generate the null distribution.

To simulate nucleotide sequences, I used Seq-Gen 1.1-1.21 (Rambaut and Grassly, 1997), generating 1,000 replicates for all data sets except the bird mtDNA data set, for which 500 replicates were generated. Tree searching and likelihood calculations were done with PAUP* 4.0b4a-4.0b8 (Swofford, 1998). Branch and bound searches were used for the rodent 12S rRNA, HIV, and bird mtDNA sequences. For the vertebrate tRNA, arthropod EF1α, and primate mtDNA sequences I used a heuristic search strategy, obtaining start trees by way of stepwise addition (3–10 random addition replicates), followed by TBR branch swapping.

The SH Test

The particular version of the SH test implemented in PAUP* 4.0 (Swofford, 1998) is referred to as the MC test by Shimodaira and Hasegawa (1999). The hypotheses compared in the SH test take the following form:

\[
H_0: \text{That all topologies are equally good explanations of the data} \\
H_A: \text{That some or all of the topologies are not equally good explanations of the data}
\]

This implementation of the SH test corresponds to Test *posNPrncd*, with RELL approximation (Kishino et al., 1990) as described by Goldman et al. (2000). A weighted version of the SH test is also possible, as described by Shimodaira and Hasegawa (1999) and implemented by Buckley et al. (2001). I reoptimized substitution model parameters and branch lengths for each topology tested and performed 1,000 replicates in the RELL approximation. I restricted attention to only two topologies when performing the SH test. With most data sets, one will need to sample more topologies than this; however, restricting the number of candidate topologies to only two allows a more direct comparison with the SOWH test.

Bayesian Posterior Probabilities

The Bayesian analyses were performed by using the program MrBayes 1.0–2.01 (Huelsenbeck and Ronquist, 2001).
MrBayes uses a Metropolis-coupled MCMC (MCMCMC) algorithm to sample from the posterior distribution. This algorithm runs multiple “heated” chains and has the advantage of being able to move across valleys in the posterior distribution more readily than some other MCMC algorithms. Each Markov chain was started from a random tree and run for $2 \times 10^6$ to $5 \times 10^6$ cycles. Every 100th cycle of the chain was sampled to minimize the size of the output files and to ensure that these samples were independent. Four chains were run simultaneously with a “temperature” of 0.2. The initial 10% of cycles were discarded as burn-in. Because the number of cycles required for the MCMCMC algorithm to converge on the target posterior distribution (stationarity) is unknowable a priori, each analysis was repeated twice to confirm that convergence had been achieved (Gelman et al., 1995; Larget and Simon, 1999; Huelsenbeck and Ronquist, 2001). Vague or uninformative prior distributions were assumed for the rate matrix (0–100), gamma-shape parameter (0–10), proportion of invariant sites (0–1), SSR parameters (0–10), branch lengths (0–10), and topology. A Dirichlet distribution was assumed for the base frequency parameters. Posterior probabilities of individual topologies were estimated by importing the output of the MCMC simulations into the program Summarize from the Bambep package (Simon and Larget, 2000).

**RESULTS**

Comparing Different Tests of Topology

The results from the tests of topology are shown in Tables 1–6. For each data set, the likelihood of alternative topologies, the test statistics, $P$-values for the SOWH and SH tests, and the posterior probabilities are shown. Table 7 shows the sizes of the 0.95 posterior intervals (the set of topologies obtained by ranking the topologies in order of decreasing posterior probability and taking the set of topologies for which the posterior probabilities sum to 0.95).

The results from all data sets analyzed demonstrate that the three probabilistic tests differ strongly in their ability to discriminate among alternative topologies. For example, when the SOWH test was applied to the HIV data set, the null hypothesis was able to be rejected with all $P$-values $\leq 0.006$ (Table 1). Similarly, the posterior probabilities of the suboptimal topology under all models were very small (Table 1). On the other hand, with the SH test, the null hypothesis was able to be rejected only under the JC69 model for the HIV data set (Table 1), suggesting much greater uncertainty than was inferred by using the SOWH test and the Bayesian approach. When all data sets were considered, the $P$-values from the SH test were always greater than those calculated under the SOWH test for the same model. Additionally, Type 1 errors were much more common for the SOWH test when the null hypothesis was assumed to be true. In analyses of the rodent 12S, vertebrate tRNA, and bird mtDNA data sets, all SOWH test results are Type 1 errors.

Small differences in likelihood often led to significant results under the SOWH test. For example, under the GTR + I model for the primate mtDNA data set, a likelihood difference of only 1.786 was significant ($P = 0.004$; Table 2). Similarly, for the Arthropod EF1α data under the GTR + $\Gamma$ model, a likelihood difference of 0.508 was marginally significant suggesting much greater uncertainty than was inferred by using the SOWH test and the Bayesian approach. When all data sets were considered, the $P$-values from the SH test were always greater than those calculated under the SOWH test for the same model. Additionally, Type 1 errors were much more common for the SOWH test when the null hypothesis was assumed to be true. In analyses of the rodent 12S, vertebrate tRNA, and bird mtDNA data sets, all SOWH test results are Type 1 errors.

Small differences in likelihood often led to significant results under the SOWH test. For example, under the GTR + I model for the primate mtDNA data set, a likelihood difference of only 1.786 was significant ($P = 0.004$; Table 2). Similarly, for the Arthropod EF1α data under the GTR + $\Gamma$ model, a likelihood difference of 0.508 was marginally significant suggesting much greater uncertainty than was inferred by using the SOWH test and the Bayesian approach. When all data sets were considered, the $P$-values from the SH test were always greater than those calculated under the SOWH test for the same model. Additionally, Type 1 errors were much more common for the SOWH test when the null hypothesis was assumed to be true. In analyses of the rodent 12S, vertebrate tRNA, and bird mtDNA data sets, all SOWH test results are Type 1 errors.

**Table 1.** Results of the tests of topology from the HIV gag and pol sequences.

<table>
<thead>
<tr>
<th>Null tree</th>
<th>ML tree</th>
<th>$\delta$</th>
<th>$-\log$ likelihood</th>
<th>$P$-value</th>
<th>Posterior probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC69</td>
<td>5494.415</td>
<td>5464.935</td>
<td>29.48</td>
<td>$&lt;0.001^{**}$</td>
<td>0.032*</td>
</tr>
<tr>
<td>K80</td>
<td>5281.217</td>
<td>5260.130</td>
<td>21.087</td>
<td>$&lt;0.001^{**}$</td>
<td>0.057</td>
</tr>
<tr>
<td>HKY85</td>
<td>5168.530</td>
<td>5151.143</td>
<td>17.387</td>
<td>$&lt;0.001^{**}$</td>
<td>0.077</td>
</tr>
<tr>
<td>GTR</td>
<td>5150.876</td>
<td>5134.730</td>
<td>16.146</td>
<td>$&lt;0.001^{**}$</td>
<td>0.089</td>
</tr>
<tr>
<td>GTR + I</td>
<td>5072.499</td>
<td>5069.427</td>
<td>3.072</td>
<td>0.006*</td>
<td>0.222</td>
</tr>
<tr>
<td>GTR + $\Gamma$</td>
<td>5076.233</td>
<td>5070.722</td>
<td>5.512</td>
<td>$&lt;0.001^{**}$</td>
<td>0.138</td>
</tr>
<tr>
<td>GTR + I + $\Gamma$</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.137</td>
</tr>
<tr>
<td>GTR + SSR</td>
<td>5010.956</td>
<td>5001.516</td>
<td>9.440</td>
<td>$&lt;0.001^{**}$</td>
<td>0.185</td>
</tr>
</tbody>
</table>

The maximum likelihood values given in bold are optimal for the corresponding model. The difference in log likelihood between the two topologies is represented by $\delta$. The likelihood was identical under both the GTR + I and GTR + I + $\Gamma$ models for the HIV data. For the SOWH and SH tests, * and ** indicate $P$-values significant at 0.05 and 0.001, respectively.
TABLE 2. Results of the tests of topology from the primate mtDNA sequences.

<table>
<thead>
<tr>
<th></th>
<th>-log likelihood</th>
<th>SOWH test</th>
<th>SH test</th>
<th>Posterior probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True tree</td>
<td>False tree</td>
<td>δ</td>
<td>P-value</td>
</tr>
<tr>
<td>JC69</td>
<td>5571.693</td>
<td>5569.513</td>
<td>2.179</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>K80</td>
<td>5382.290</td>
<td>5384.288</td>
<td>1.997</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>HKY85</td>
<td>5232.712</td>
<td>5234.199</td>
<td>1.487</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>GTR</td>
<td>5189.523</td>
<td>5191.429</td>
<td>2.106</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>GTR + I</td>
<td>5053.381</td>
<td>5055.167</td>
<td>1.786</td>
<td>0.004*</td>
</tr>
<tr>
<td>GTR + Γ</td>
<td>5028.315</td>
<td>5033.817</td>
<td>5.502</td>
<td>0.002*</td>
</tr>
<tr>
<td>GTR + I + Γ</td>
<td>5028.001</td>
<td>5033.096</td>
<td>5.095</td>
<td>0.001*</td>
</tr>
</tbody>
</table>

TABLE 3. Results of the tests of topology from the vertebrate mitochondrial tRNA sequences.

<table>
<thead>
<tr>
<th></th>
<th>-log likelihood</th>
<th>SOWH test</th>
<th>SH test</th>
<th>Posterior probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True tree</td>
<td>False tree</td>
<td>δ</td>
<td>P-value</td>
</tr>
<tr>
<td>JC69</td>
<td>8136.354</td>
<td>8129.218</td>
<td>7.136</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>K80</td>
<td>7588.333</td>
<td>7580.915</td>
<td>7.417</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>HKY85</td>
<td>7509.906</td>
<td>7503.966</td>
<td>5.940</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>GTR</td>
<td>7462.817</td>
<td>7456.896</td>
<td>5.921</td>
<td>0.002*</td>
</tr>
<tr>
<td>GTR + I</td>
<td>7353.159</td>
<td>7350.388</td>
<td>2.771</td>
<td>0.018*</td>
</tr>
<tr>
<td>GTR + Γ</td>
<td>7343.037</td>
<td>7340.663</td>
<td>2.375</td>
<td>0.046*</td>
</tr>
<tr>
<td>GTR + I + Γ</td>
<td>7340.286</td>
<td>7338.062</td>
<td>2.224</td>
<td>0.023*</td>
</tr>
</tbody>
</table>

TABLE 4. Results of the tests of topology from the bird mtDNA sequences.

<table>
<thead>
<tr>
<th></th>
<th>-log likelihood</th>
<th>SOWH test</th>
<th>SH test</th>
<th>Posterior probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True tree</td>
<td>False tree</td>
<td>δ</td>
<td>P-value</td>
</tr>
<tr>
<td>JC69</td>
<td>58337.895</td>
<td>58253.948</td>
<td>83.947</td>
<td>&lt;0.002*</td>
</tr>
<tr>
<td>K80</td>
<td>57738.370</td>
<td>57664.811</td>
<td>73.559</td>
<td>&lt;0.002*</td>
</tr>
<tr>
<td>HKY85</td>
<td>56815.229</td>
<td>56743.293</td>
<td>71.937</td>
<td>&lt;0.002*</td>
</tr>
<tr>
<td>GTR</td>
<td>56238.067</td>
<td>56168.401</td>
<td>69.666</td>
<td>&lt;0.002*</td>
</tr>
<tr>
<td>GTR + I</td>
<td>55123.399</td>
<td>55106.408</td>
<td>16.990</td>
<td>&lt;0.002*</td>
</tr>
<tr>
<td>GTR + Γ</td>
<td>55019.122</td>
<td>55003.543</td>
<td>15.579</td>
<td>&lt;0.002*</td>
</tr>
<tr>
<td>GTR + I + Γ</td>
<td>55018.712</td>
<td>55003.410</td>
<td>15.302</td>
<td>&lt;0.002*</td>
</tr>
</tbody>
</table>

TABLE 5. Results of the tests of topology from the arthropod EF1α sequences.

<table>
<thead>
<tr>
<th></th>
<th>-log likelihood</th>
<th>SOWH test</th>
<th>SH test</th>
<th>Posterior probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True tree</td>
<td>False tree</td>
<td>δ</td>
<td>P-value</td>
</tr>
<tr>
<td>JC69</td>
<td>7481.409</td>
<td>7476.026</td>
<td>5.384</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>K80</td>
<td>7377.036</td>
<td>7373.207</td>
<td>3.829</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>HKY85</td>
<td>7357.046</td>
<td>7353.766</td>
<td>3.281</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>GTR</td>
<td>7297.599</td>
<td>7294.259</td>
<td>3.540</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>GTR + I</td>
<td>6865.350</td>
<td>6865.680</td>
<td>0.330</td>
<td>0.005</td>
</tr>
<tr>
<td>GTR + Γ</td>
<td>6856.155</td>
<td>6856.664</td>
<td>0.508</td>
<td>0.030*</td>
</tr>
<tr>
<td>GTR + I + Γ</td>
<td>6841.362</td>
<td>6841.946</td>
<td>0.584</td>
<td>0.038*</td>
</tr>
</tbody>
</table>
Table 6. Results of the tests of topology from the rodent 12S rRNA sequences.

<table>
<thead>
<tr>
<th>Model</th>
<th>-log likelihood</th>
<th>SOWH test</th>
<th>SH test</th>
<th>Posterior probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True tree</td>
<td>False tree</td>
<td>δ</td>
<td>P-value</td>
</tr>
<tr>
<td>JC69</td>
<td>1719.271</td>
<td>1704.357</td>
<td>14.914</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>K80</td>
<td>1693.839</td>
<td>1681.301</td>
<td>12.538</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>HKY85</td>
<td>1651.285</td>
<td>1639.569</td>
<td>11.716</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>GTR</td>
<td>1639.614</td>
<td>1628.644</td>
<td>10.970</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>GTR + I</td>
<td>1617.542</td>
<td>1612.843</td>
<td>4.699</td>
<td>0.018*</td>
</tr>
<tr>
<td>GTR + Γ</td>
<td>1617.349</td>
<td>1612.623</td>
<td>4.726</td>
<td>0.015*</td>
</tr>
<tr>
<td>GTR + I + Γ</td>
<td>1617.394</td>
<td>1612.623</td>
<td>4.771</td>
<td>0.018*</td>
</tr>
</tbody>
</table>

(P = 0.03; Table 5). Conversely, the lowest likelihood difference that was significant under the SH test was 29.48 for the HIV data under the JC69 model.

In cases where the SOWH test indicated strong rejection of a topology, the posterior probability of that topology also tended to be low. For example, the assumed true topologies for the rodent 12S rRNA, vertebrate tRNA, and bird mtDNA data sets had extremely low posterior probabilities, and these topologies were in general strongly rejected by the SOWH test. The exception to this pattern was the arthropod EF1α data, for which the posterior probabilities estimated by using the among-site rate variation models suggested considerable uncertainty in the data. On the other hand, the SOWH test indicated marginal rejection of the false topology, except for the GTR + I model.

The Effects of Substitution Model Assumptions

As expected, the relationship between substitution model complexity and the power of the tests was not straightforward (see also Buckley and Cunningham, 2002). For the primate mtDNA sequences I observed that the posterior probability of the assumed true tree increased from 0.116 to 0.970 as the model increased in complexity from JC69 to GTR + I + Γ (Table 2). Similarly, the 0.95 posterior intervals included only one topology under the GTR + I + Γ model but two topologies under the less parameter-rich submodels (Table 7). The P-values estimated under the SH test decreased from 0.405 (K80) to 0.138 (GTR + I + Γ) for the primate data set (Table 2). Results were similar for the arthropod EF1α sequences, the posterior probability of the assumed true tree increasing from 0.004 under JC69 to 0.409 under GTR + I + Γ (Table 5); however, the size of the 0.95 posterior intervals increased from one to seven topologies (Table 7). Other data sets showed different patterns. For example, the SH test P-values increased from 0.032 (JC69) to 0.137 (GTR + I + Γ) for the HIV sequences (Table 1), and the posterior intervals widened (Table 7), indicating a drop in resolution.

For some of the data sets, small shifts in model structure led to rather different biological conclusions. For the primate mtDNA data, the incorrect topology was optimal under the JC69 model, and the null hypothesis was easily rejected (P < 0.001). However, with the addition of a single parameter to the model (transition:transversion ratio), the assumed true topology became optimal,

Table 7. Number of distinct topologies in 0.95 posterior intervals for each data set.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>HIV gag and pol</th>
<th>Primate mtDNA</th>
<th>Vertebrate tRNA</th>
<th>Bird mtDNA</th>
<th>Arthropod EF1α</th>
<th>Rodent 12S rRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC69</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>K80</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>HKY85</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>GTR</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>GTR + I</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>GTR + Γ</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>GTR + I + Γ</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>GTR + SSR</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
and the false topology was strongly rejected ($P < 0.001$). Similarly, the posterior probabilities differed greatly between these two models. In contrast, the null hypothesis under the SH test was unable to be rejected, regardless of the substitution model (Table 2). The arthropod EF1α sequences exhibited a similar pattern. Under the GTR model, the false topology was optimal and the null hypothesis was easily rejected. With the addition of the gamma parameter (GTR + $\Gamma$), however, the assumed true topology became optimal and the false topology was rejected (Table 5).

When the suboptimal topology is correct, an increase in the conservativeness of the test is desirable. The $P$-values from the SH test for the bird mtDNA data were <0.05 under the equal rates models. However, when among-site rate variation parameters were included in the substitution model, the SH test became more conservative and the assumed true tree was included within the confidence limits of the optimal tree. The HIV sequences showed a similar pattern.

**Discussion**

**Caveats**

The interpretation of the results presented in this study is in part dependent on whether the topologies in Figure 1 are correct. Although many of these phylogenetic relationships have been challenged at some stage, the topologies presented here are supported by multiple sources of independent character data. However, the markers analyzed here could conceivably have evolved along different phylogenies to the species tree as a result of lineage sorting effects. For the bird, primate, and rodent mtDNA data sets this seems unlikely because further taxon sampling (the bird data set; Buckley and Cunningham, 2002) and longer mtDNA sequences (primate and rodent data sets; Sullivan et al., 1995; Waddell et al., 1999b) all lead to convergence on the topologies that are assumed to be correct. If further phylogenetic studies revealed that some of these topologies were in fact not correct, then inferences concerning the Type 1 error rate would be erroneous. However, the overall conclusions regarding the relative statistical power of the different tests would remain unchanged.

**Comparing the SOWH and SH Tests of Topology**

A direct comparison of the results of the SH test and those of the SOWH test is not straightforward, because the null hypotheses that these two tests are examining are not identical (Goldman et al., 2000). The SOWH tests evaluates the null hypothesis that a given topology is the true topology, whereas the SH test evaluates whether two or more topologies are equally good explanations of the data. Comparing the relative power of the two tests is useful because the overall biological goal of both tests is the same: Are the observed patterns of variation within a data set compatible with a specified topology? Below, I expand on the possible explanations given by Goldman et al. (2000) for the differences in statistical power between the two ML tests, contrast the results of the ML tests with the Bayesian approach, and discuss the role of substitution model assumptions in assessing phylogenetic hypotheses.

The results presented here are consistent with those of other authors who have observed large differences in statistical power between parametric and nonparametric ML tests of topology (Mallatt and Sullivan, 1999; Clark et al., 2000; Goldman et al., 2000; Fishbein et al., 2001; Strimmer and Rambaut, 2001). In every comparison between the two ML tests performed here, the $P$-values from the SOWH test were smaller than those generated under the SH test. In some cases, the magnitude of this difference was very large (e.g., the primate mtDNA sequences). Additionally, the Type 1 error rate was much greater for the SOWH test, even under the complex among-site rate variation models. The SH test was extremely conservative, and nonsignificant $P$-values were observed only in the HIV and bird data sets under the JC69 and equal rates models, respectively.

The SH test simultaneously compares a set of topologies, and confidence intervals are estimated, to account for the multiple comparisons. This explains the conservativeness of the SH test relative to other nonparametric ML tests (e.g., the KH test). Because the SOWH test compares only two topologies at a time, it is potentially less conservative, as noted by Goldman et al. (2000). However, for the tests used here, only pairs of topologies were examined and therefore this factor cannot explain the observed difference in
power between the two tests. Because the SH test requires that all reasonable topologies be made available for testing (Shimodaira and Hasegawa, 1999), and because the results of the test are in part dependent on the number of topologies examined (Goldman et al., 2000; Shimodaira, 2001), one must consider very carefully what topologies to include. As in any form of model selection (e.g., Burnham and Anderson, 1998) the number of candidate topologies should be minimized through the prudent application of prior knowledge. The SH test will be more useful when a manageable number of topologies can be specified. The weighted version of the SH test, implemented by Buckley et al. (2001), is less sensitive to changes in the size of the set of candidate topologies (Shimodaira, 2001).

Because of the computational burden of the SOWH test, a heuristic search for the optimal topology is required for virtually all data sets; there is therefore no guarantee that the most likely topology will be located for each replicate. When a tree search converges on a suboptimal topology for replicate \( i \), the corresponding value of \( \delta^{(i)} \) will be decreased. Over many replicates, this effect could potentially cause a tightening of the null distribution and make the SOWH test more liberal. The magnitude of this potential artifact on the power of the test for any given data set will be difficult to anticipate, but it can be minimized by using sophisticated search strategies. For all the data sets analyzed here, either branch and bound searches or thorough heuristic searches (TBR branch swapping from multiple start trees) were used, and therefore this effect cannot explain the large number of Type 1 errors observed. Goldman et al. (2000) also discuss the potential effects of fixing parameter values while searching for the ML topology for each replicate data set.

Parametric tests tend to be more powerful than nonparametric tests (Sokal and Rolf, 1995:168) because the former rely on a model derived from the data to construct the null distribution of the test statistic. The cost of this dependence is that the test may be biased if the model is inadequate in some manner (Lee, 1994; Strimmer and Rambaut, 2001). The basic reason for this dependence is as follows. Because the optimal topology for each replicate is being estimated with the exact substitution model that was used to generate the sequences (although not necessarily the same parameter estimates), the probability of recovering the generating topology will be high. The resulting likelihood ratio values will then be equal to or very close to zero, and consequently, the null distributions will be very tight. Similarly, each replicate will more easily converge on the generating topology when the data are simulated under a simple substitution model than when a more complex model is used. This happens because it is easier to recover a topology that has been generated under a simple substitution model relative to a more complex model, even when the model is known. Consequently, the SOWH test is expected to be very powerful because the same model is used both to generate the data and to construct the null distribution. Depending on whether the optimal topology is correct, this property of the SOWH test could potentially cause either a high Type 1 error rate or over-confidence in a topology.

This prediction is supported by the simulation study of Huelsenbeck et al. (1996), who observed that when their parametric test of monophyly was implemented with a misspecified model and the rates of change were varied across the tree, the test became too liberal. The magnitude of this bias increased when the total rate of change over the tree was high. Importantly, the model had to be violated by only one parameter for this bias to become noticeable. Other authors have also examined the effects of model misspecification on likelihood ratio tests in phylogenetics. Shimodaira (2001) performed a series of likelihood ratio tests of individual topologies against the “full model,” which corresponds to the total set of branches from all of the topologies. Shimodaira (2001) was able to reject all the topologies, which indicated either that the sequences had not evolved under a strictly bifurcating pattern (e.g., there has been recombination) or, more probably, that the test is overly sensitive to model misspecification. Gaut and Lewis (1995) also noted that the likelihood ratio test for zero length branches, based on \( \chi^2 \) statistics, had a high Type 1 error rate when the substitution model was incorrect. Zhang (1999) has also demonstrated that Monte Carlo–based likelihood ratio tests of nested substitution models can be too liberal when the model is misspecified.
Why does the SOWH test fail for some of the data sets analyzed here? Examining the relative branch lengths and the total rate of change over the tree, and searching for non-stationarity can give us clues. For example, the rodent 12S data set is characterized by branch length heterogeneity, a low rate of change over the tree (Fig. 1f) and inevitably, model misspecification. Huelsenbeck et al. (1996) examined a similar situation in their simulation study, and they too observed that the parametric test of monophyly tends to assign overconfidence to a topology under these conditions. Nonstationarity of base frequencies was detected in the vertebrate tRNA, arthropod EF1α, primate mtDNA, and bird mtDNA data sets (data not shown). This very common phenomenon, unaccounted for by the models implemented here, could potentially mislead the SOWH test.

The sensitivity of the SOWH test to model misspecification is probably the major reason for its greater Type 1 error rate relative to the SH test for the data presented here. These results support the recommendation of Huelsenbeck et al. (1996) that, to minimize Type 1 errors, parametric tests of topology should be implemented with highly complex substitution models. The results reported here indicate that the commonly used substitution models may not be complex enough for at least some applications of the SOWH test. Unfortunately, the dependence of the SOWH test on highly complex models will preclude its application to many real data sets. Results from the SOWH test based on the standard Markov models implemented here should be viewed with caution if there is any good evidence for model misspecification, such as a lack of independence between sites and shifts in base frequencies, rate matrices, or the distribution of among-site rate variation across the tree. The results of the SOWH test should also be interpreted cautiously when there are large differences in substitution rates among different lineages.

Assessing the power or bias of the SOWH test in other empirical studies is difficult because whether the null hypotheses being examined are correct is usually not known. I found 32 publications in the literature that contained a total of 58 individual ML parametric tests of topology. Of those 58 tests, only 7 had nonsignificant P-values (e.g., Huelsenbeck and Bull, 1996; Huelsenbeck et al., 1997; Sullivan and Swofford, 1997; Hrincevich et al., 2000). The studies that used parsimony to generate the null distribution were excluded (e.g., Ruedi et al., 1998; Emerson et al., 2000). The null hypotheses being tested in most of these ML studies are probably incorrect, making it doubtful that the application of the SOWH test has been seriously misleading in general. However, for many of the above studies in which the null hypothesis is incorrect, the uncertainty associated with the optimal, true tree may have been underestimated.

Given the lack of fit between the assumptions made by parsimony and the model used to generate the data, perhaps using the parsimony method to estimate the optimal topology for each replicate will be less discriminatory than the full likelihood approach. The only two studies that have compared the effects of using likelihood and parsimony to generate the null distribution for the SOWH test support this contention (Hillis et al., 1996; Sullivan et al., 2000).

Model Misspecification and Bayesian Posterior Probabilities

The results of the SOWH test and the Bayesian posterior probabilities were generally correlated. In cases where the SOWH test indicated strong rejection of the null hypothesis, the suboptimal topology tended to have a very low posterior probability. Intuitively, one might expect to observe greater uncertainty under the hierarchical Bayesian approach, relative to the SOWH test (with approximations), because MCMC accounts for uncertainty in the parameter estimates of the substitution model (see Huelsenbeck and Bollback, 2001) and, indeed, this prediction is consistent with the observations from the arthropod EF1α data set. In the case of the rodent 12S rRNA, vertebrate tRNA, and bird mtDNA data sets, the Bayesian approach was highly misleading because the posterior probabilities of the assumed true topologies were extremely low. Conversely, the Bayesian approach appeared to work well for the primate mtDNA and arthropod EF1α data sets, provided that among-site rate variation was modeled adequately. The possibilities of convergence problems of the MCMC algorithm, inappropriate priors, and model misspecification are discussed below.
If the MCMC algorithm has not achieved convergence, then any estimates of confidence in a topology will be misleading. Because all the estimated posterior probabilities were very similar across different simulations, it is most likely this potential pitfall has not affected the results. In all of the simulations, $2 \times 10^6$ to $5 \times 10^6$ cycles was apparently sufficient for obtaining stable posterior probabilities; however, this number may be inadequate for larger phylogenies and more complex models.

The most common criticism of Bayesian methods is that subjective prior distributions can bias the results (e.g., Royall, 1997). Because vague priors for the topology, branch lengths, and substitution model parameters were used here, the value of the likelihood function should dominate the shape of the posterior distribution (Gelman et al., 1995). Therefore, I deem it unlikely that any misleading results are attributable to inappropriate prior distributions. The sensitivity of Bayesian phylogenetic analyses to changes in the prior distributions of parameters from the substitution model has only been touched upon (Yang and Rannala, 1997).

Because problems with convergence and inappropriate priors can be excluded, the failure of the Bayesian approach for some of the data sets analyzed here can be attributed to misspecification of the likelihood model. The observation that the true tree is not optimal under the Bayesian approach for a given data set is not serious problem. Such observations are to be expected, given moderate model misspecification and finite sample size. Of greater concern is the absence of the true tree from a specified region of uncertainty (i.e., the 0.95 posterior intervals) around the optimal topology. Systematists should be more concerned with identifying the total set of trees that can be reasonably supported by the data, rather than focusing on point estimates of topology. Indeed, one of the greatest strengths of the Bayesian approach is its inherent ability to quantify the uncertainty associated with a parameter estimate, by using MCMC to sample from the posterior distribution.

Because the use of flat priors has the effect of making the posterior distribution largely dependent on the structure of the likelihood model, a seriously misspecified likelihood model can potentially yield strong and unreliable posterior inferences. Shimodaira (2001) argues that the Bayesian approach may be misleading under model misspecification and gives an empirical example in which approximate Bayesian posterior probabilities credit much greater confidence to an optimal topology than do nonparametric ML tests (see also Buckley et al., 2002). These findings and the data presented here are consistent with the recommendations of Larget and Simon (1999), and Huelsenbeck and Bollback (2001): Users of Bayesian phylogenetic methods need to be more concerned with maximizing the fit of the model to the data rather than minimizing the associated variance or computational inefficiency. For some of the data sets analyzed here, models that are widely considered to be “parameter-rich” (e.g., GTR + I + Γ) led to gross overestimates of support for incorrect topologies. For some data sets one may have to resort to greater complexity, such as models based around the codon (e.g., Yang et al., 2000) and mixed distribution models (e.g., Yang, 1996), or to average inferences over a set of plausible models (e.g., Hoeting et al., 1999). Both methods are expected to increase the uncertainty associated with a parameter estimate. Simulation studies to investigate the effects of substitution model misspecification on Bayesian posterior probabilities will be particularly informative.

Conclusions

The analyses presented here indicate that corrections for multiple comparisons and insufficient explorations of tree and parameter space can be discounted as adequate explanations for the failure of the SOWH and Bayesian approaches for some of these data sets. More probably the misleading results obtained are due to model misspecification. The power of the SOWH test results from the use of an explicit substitution model to construct the null distribution. However, when the assumptions of the model are violated, overconfidence in a topology can result. The effects of model misspecification are exacerbated when the trees contain a mixture of short and long branches (e.g., the bird mtDNA and rodent mtDNA data sets). Although these effects have been known for some time (Huelsenbeck et al., 1996), the results presented here and the large number of highly significant $P$-values observed in the
literature suggest that model misspecification may have had a larger effect on empirical studies than previously appreciated.

These results support the contention (Huelsenbeck et al., 1996; Larget and Simon, 1999; Huelsenbeck and Bollback, 2001) that users of the SOWH test and Bayesian posterior probabilities should be more concerned with the fit of the substitution model than with computational efficiency or minimizing the associated variance. As demonstrated in some cases, the more complex models commonly implemented in the literature (e.g., GTR + I + Γ) were simply inadequate. The tendency of the SOWH test to generate Type I errors in the examples presented here, and in simulation studies (Huelsenbeck et al., 1996), can in principle be alleviated by increasing the complexity of the substitution model. However, this will severely restrict the utility of the SOWH test because of the computational burden of repeated ML topology searches and the ever increasing size of molecular data sets. The SH test is certainly a safer approach to phylogenetic hypothesis testing, although it has several problems with implementation that require further study. The most significant of these is the high level of conservativeness and the difficulty in selecting the set of candidate trees to be tested.

Simulation studies are now required to clarify the relationships among the different probabilistic tests and their relative sensitivities to model misspecification. One possible approach would be to simulate data sets under parameter-rich models, such as codon models (Yang et al., 2000), mixed distribution models (Yang, 1996), or nonstationary models (Galtier and Guoy, 1998). Tree searches and tests of topology would then be performed under the more standard substitution models (i.e., GTR + I + Γ and submodels). Measurement of Type I and 2 error rates and the corresponding posterior probabilities would indicate which tests have acceptable statistical properties. Branch length heterogeneity, total tree length, and sequence length also are important factors to take into consideration.

ACKNOWLEDGMENTS

I thank Nick Goldman, Allen Rodrigo, Hidetoshi Shimodaia, and Jack Sullivan for helpful discussions and unpublished manuscripts. David Mindell and Ziheng Yang provided some of the data sets. Cliff Cunningham gave support and encouragement during the initial stages of this study. Bruce Rannala and three anonymous reviewers gave comments on the manuscript. This work was supported by funding from Duke University Comprehensive Cancer Center, and Department of Biology, and the Foundation for Research, Science and Technology, New Zealand (contract LR/CX0004).

REFERENCES


Galtier, N., and M. Guoy. 1998. Inferring pattern and process: Maximum likelihood implementation of

2002

BUCKLEY—LIKELIHOOD AND BAYESIAN TESTS OF TOPOLOGY

521


First submitted 9 November 2001; revision submitted 5 December 2001; final accepted 18 February 2002

Associate Editor: Bruce Rannala