Selection of evolutionary models for phylogenetic hypothesis testing using parametric methods

B. C. EMERSON, K. M. IBRAHIM* & G. M. HEWITT
School of Biological Sciences, University of East Anglia, Norwich, UK

Introduction

Felsenstein (1985) proposed the use of the general statistical procedure of repeated sampling with replacement from a data set (bootstrapping) to estimate confidence limits of the internal branches in a phylogenetic tree. Although there have been debates on the interpretation of bootstrap results (e.g. Felsenstein & Kishino, 1993; Hillis & Bull, 1993) this is still the commonest method for assessing confidence in phylogenies. However, this nonparametric bootstrap method does not provide the opportunity to test a priori hypotheses about the phylogeny of a group. The alternative, parametric bootstrap method requires the specification of a particular substitutional model, the parameters of which will be used to generate simulated, replicate DNA sequence data sets. It has been both suggested that, (a) the method appears robust to changes in the model of evolution, and alternatively that, (b) as realistic model of DNA substitution as possible should be used to avoid false rejection of a null hypothesis. Here we empirically evaluate the effect of suboptimal substitution models when testing hypotheses of monophyly with the parametric bootstrap using data sets of mitochondrial COI and COII sequences for Macaronesian Calathus beetles, and mitochondrial 16S rDNA and nuclear ITS2 sequences for European Timarcha beetles. Whether a particular hypothesis of monophyly is rejected or accepted appears to be highly dependent on whether the nucleotide substitution model being used is optimal. It appears that a parameter rich model is either equally or less likely to reject a hypothesis of monophyly where the optimal model is unknown. A comparison of the performance of the Kishino–Hasegawa (KH) test shows it is not as severely affected by the use of suboptimal models, and overall it appears to be a less conservative method with a higher rate of failure to reject null hypotheses.

Keywords:
- hypothesis
- Kishino–Hasegawa test
- parametric bootstrap
- phylogeny.

Abstract

Recent molecular studies have incorporated the parametric bootstrap method to test a priori hypotheses when the results of molecular based phylogenies are in conflict with these hypotheses. The parametric bootstrap requires the specification of a particular substitutional model, the parameters of which will be used to generate simulated, replicate DNA sequence data sets. It has been both suggested that, (a) the method appears robust to changes in the model of evolution, and alternatively that, (b) as realistic model of DNA substitution as possible should be used to avoid false rejection of a null hypothesis. Here we empirically evaluate the effect of suboptimal substitution models when testing hypotheses of monophyly with the parametric bootstrap using data sets of mitochondrial COI and COII sequences for Macaronesian Calathus beetles, and mitochondrial 16S rDNA and nuclear ITS2 sequences for European Timarcha beetles. Whether a particular hypothesis of monophyly is rejected or accepted appears to be highly dependent on whether the nucleotide substitution model being used is optimal. It appears that a parameter rich model is either equally or less likely to reject a hypothesis of monophyly where the optimal model is unknown. A comparison of the performance of the Kishino–Hasegawa (KH) test shows it is not as severely affected by the use of suboptimal models, and overall it appears to be a less conservative method with a higher rate of failure to reject null hypotheses.
parsimony based approach is an acceptable alternative with only a slightly lower level of discrimination (Hillis et al., 1996).

Many recent molecular studies have used the parametric bootstrap method for testing a priori hypotheses against their molecular based phylogenies (Mallat & Sullivan, 1998; Ruedi et al., 1998; Flook et al., 1999; Jackman et al., 1999; Oakley & Phillips, 1999). Other studies have utilized variations of the method to examine the possibility of long branch attraction (Felsenstein, 1978; Hendy & Penny, 1989; Huelsenbeck, 1997) causing erroneous topologies (Maddison et al., 1999; Pang et al., 1999), to estimate the sequence length required to resolve a particular phylogenetic tree (Halanych, 1998; Flook et al., 1999), to evaluate the efficiency of different methods of phylogenetic reconstruction (Bull et al., 1993; Hwang et al., 1998), and to test the adequacy of models of DNA sequence evolution (Goldman, 1993; Yang et al., 1994).

A critical first step of the parametric bootstrap is specifying a particular substitutional model, whose parameters will be estimated from the empirical data, and which will subsequently be used when generating the simulated, replicate DNA sequences. Hillis et al. (1996) stated that in limited studies the method appeared to be robust to changes in the model of evolution. However, Huelsenbeck et al. (1996a) suggest as realistic model of DNA substitution as possible should be used to reduce Type I error (false rejection of a null hypothesis) implying that parameter-rich models may perform better. Zhang (1999) has shown that if the substitution model is inadequate then the estimates of the substitution parameters will be biased. By using biased estimates of the parameters the simulated sequence data will deviate from the empirical data, such that the null distribution of tree score differences will be distorted.

In this study we evaluate empirically the effect of the choice of substitutional model when testing hypotheses of monophyly using DNA sequence data. We have chosen three gene regions (two mitochondrial and one nuclear) commonly used in phylogenetic analyses: cytochrome oxidase subunit I and II (COI and COII), the large mitochondrial ribosomal subunit gene (16S), and the nuclear ribosomal internal transcribed spacer 2 (ITS2). For COI and COII we have used published sequence data for Macaronesian Calathus beetles (Carabidae) (Emerson et al., 1999, 2000) and for 16S and ITS2 we have used published sequence data for European Timarcha beetles (Chrysomelidae) (Gomez-Zurita et al., 2000a, b). For each of these data sets we also examine the comparative performance of the nonparametric Kishino–Hasegawa (KH) test (Kishino & Hasegawa, 1989) which compares competing phylogenetic hypotheses under the maximum likelihood (ML) optimality criterion. This method uses likelihood ratio tests of the statistical significance of competing tree topologies. All three data sets offer clear a priori hypotheses of monophyly for testing against empirical DNA sequence data.

**Materials and methods**

The following steps summarize the methodology used. For each DNA sequence data set (Calathus COI and COII, Timarcha 16S and Timarcha ITS2) we: (1) defined an appropriate model of nucleotide substitution for the purpose of parameterizing a maximum likelihood phylogenetic analysis of the data. For each of the three data sets the results of the phylogenetic analysis are in conflict with null hypotheses of monophyly at specified nodes. (2) To test the significance of the conflict in each case, we have carried out a parametric bootstrap analysis to evaluate the probability of such conflict arising by chance. (3) As a comparison of performance KH tests of the null hypotheses against conflicting maximum likelihood phylogenetic results were also performed. (4) For both the parametric bootstrap analyses and KH tests the effect of suboptimal parameter estimates for the model of nucleotide substitution were also evaluated.

**DNA sequences**

For COI and COII we have analysed 71 sequences [924 base pairs (bp) of COI and 687 bp of COII] for 37 Calathus species, and an outgroup, Calathidius accuminatus (Emerson et al., 1999, 2000). In the case of 16S a subset of the sequences presented by Gomez-Zurita et al. (2000a) from 19 species of the genus Timarcha were analysed. Ambiguous sites and sites with insertions or deletions were removed to obtain 34 sequences of length 495. Similarly, for ITS2 we analysed a subset of the sequences presented by Gomez-Zurita et al. (2000b) representing 19 Timarcha species. After removing ambiguous sites and sites with insertions or deletions, 30 sequences of length 532 were obtained. For both the 16S and ITS2 data sets T. metallicus was used as an outgroup.

**Model selection, phylogenetic analysis, and null hypotheses**

For each of the three data sets the fit of sequence data to 56 models of base substitution, ranked in Table 1 roughly according to the complexity of their matrix of transition probability, was tested using Modeltest v.3.0 (Posada & Crandall, 1998). Modeltest uses the log likelihood ratio tests and the Akaike information criterion (AIC) (Akaike, 1973) to determine which of the models best describes the data. For each data set a maximum likelihood tree is constructed with the parameters defined as best describing the data set by Modeltest. We then explore null hypotheses of monophyly for each of these data sets in order to evaluate
The phylogeny in Fig. 1 is unresolved with regard to a number of the internal branches that are characterized by short length and lack of nonparametric bootstrap support. We consider five null hypotheses of monophyly with regard to island clades: (1) that with the exception of C. subfuscus which is clearly related to the continental C. ambiguus, the remaining Macaronesian Calathus are monophyletic (clades A, B, C, D); (2) that all Canary Island Calathus are monophyletic (clades B, C, D); (3) that excluding clade D the Macaronesian Calathus are monophyletic (clades A, B, C); that Madeira is either (4) monophyletic with the eastern Canary Islands (clades A, B) or (5) with the main Canary Island clade (clades A, C).

Timarcha 16S

For the Timarcha 16S sequence data the transition model of substitution (TIM), Rodríguez et al. (1990) fitted the data best with estimated substitution rates of: A–C: 1, A–G: 13.185, A–T: 6.077, C–G: 6.077, C–T: 45.935, G–T: 1. The proportion of invariant sites estimated to be 0.781, a gamma shape parameter of \( \hat{\Gamma} = 1.871 \), and, base frequencies: A: 0.404, C: 0.128, G: 0.089, T: 0.378. Figure 2a is a ML tree of mtDNA sequence data for the 34 Iberian Timarcha obtained using PAUP*. Within the phylogeny the T. goettingensis species complex, previously characterized as monophyletic based on cytogenetic studies (Petitpierre, 1970) is not monophyletic but includes the species T. hispanica, T. granadensis and Timarcha sp. Here we test the null hypothesis of monophyly for the T. goettingensis species complex.

Timarcha ITS2

For the Timarcha ITS2 sequence data the transversion model of substitution with equal base frequencies (TVMef), Rodríguez et al. (1990) fitted the data best with estimated substitution rates of: A–C: 1.404, A–G: 3.193, A–T: 3.787, C–G: 1.067, C–T: 3.193, G–T: 1. The proportion of invariant sites estimated to be 0.361, and a gamma shape parameter of \( \hat{\Gamma} = 0.832 \). Figure 2b is a ML tree of mtDNA sequence data for the 30 Iberian Timarcha obtained using PAUP*. Again the T. goettingensis species complex is not monophyletic but includes the species T. hispanica, T. granadensis and Timarcha sp. Similar to the 16S data set the null hypothesis we test is monophyly for the T. goettingensis species complex.

Parametric bootstrap

For a given hypothesis of monophyly, a constrained NJ tree (enforcing monophyly) was constructed from the sequence data in PAUP* using ML distances from the best fit model and the parameter estimates derived from the Modeltest analysis (Figs 3 and 4). Next, for each of the seven hypotheses, 500 replicate DNA sequence data sets were obtained using PAUP*.
were generated using Seq-Gen v.1.1 (Rambaut & Grassly, 1997) which simulates the evolution of DNA sequences along a defined phylogeny under a specified model of the substitution process. Again, the sequences were generated using the best-fit model of substitution and the parameter estimates derived from the empirical data. However, because of software limitations within Seq-Gen v.1.1 we were unable to generate sequences with a
defined proportion of invariant sites as in the best-fit model identified using Modeltest. Instead, parameters were used from a similar model but without a proportion of invariant sites defined. For each of the null hypotheses, each of the 500 replicate data sets were subject to heuristic parsimony searches first with and then without the specified constraint. The resulting distribution of differences between the step-length of constrained and
unconstrained trees was then compared with the tree length differences for the empirical constrained and nonconstrained trees. Alternatively we could have constructed ML trees and produced a distribution of likelihood differences but this was not feasible due to computational limitations.

This process was then repeated using the following less adequate substitution models for each of the three sequence datasets. For the *Calathus* COI and COII data: (1) GTR without $\Gamma$, (2) HKY85 (Hasegawa et al., 1985), (3) Kimura 2-parameter (Kimura, 1980) with $\Gamma$, (4) Kimura 2-parameter without $\Gamma$, (5) Jukes–Cantor (Jukes & Cantor, 1969). For the *Timarcha* 16S data: (1) TIM without $\Gamma$, (2) HKY85, (3) Kimura 2-parameter with $\Gamma$, (4) Kimura 2-parameter without $\Gamma$, (5) Jukes–Cantor. For the *Timarcha* ITS2 data: (1) TVM without $\Gamma$, (2) Kimura 3-parameter (Kimura, 1981), (3) Kimura 2-parameter with $\Gamma$, (4) Kimura 2-parameter without $\Gamma$, (5) Jukes–Cantor. Again, for each of these models, parameter estimates were derived from PHYLIP*, constrained NJ trees were constructed from the empirical data, sequence data was simulated under the same model, constrained and unconstrained parsimony trees were obtained from heuristic searches, and distributions of differences generated.

**Kishino–Hasegawa test**

This parametric test evaluates alternative tree topologies by testing if the difference between their log

---

**Fig. 3** Constrained trees constructed using neighbour joining with maximum likelihood distances under the GTR + I + G substitution model for the *Calathus* COI and COII sequence data. The first tree is the unconstrained tree and the other five represent null hypotheses about monophyly. Unconstrained and constrained trees were also generated under 5 other sub-optimal models (see text for details).
likelihood values is statistically significant. This can be carried out either by bootstrap re-sampling which requires a high amount of computation (Hasegawa & Kishino, 1989) or using explicit estimates of the variance of the difference between log likelihoods as proposed in Kishino & Hasegawa (1989). The latter were carried out as implemented in PAUP*. Neighbour joining trees (Saitou & Nei, 1987) were constructed using ML distances for unconstrained and constrained trees using parameter estimates for each of the aforementioned models and each of the five Calathus null hypotheses evaluated. The smaller 16S and ITS2 data sets for Timarcha meant it was computationally feasible to construct trees under the optimality criterion or maximum likelihood.

Results and discussion

Parametric bootstrap

For the Calathus COI and COII data the distribution of differences in step length between hypothesis-constrained and unconstrained trees from the data sets simulated under the best-fit model (GTR + \( \Gamma \)) resulted in the rejection of hypotheses 1, 2 and 3 (\( P \leq 0.01 \)) but hypotheses 4 and 5 could not be rejected at the 5% significance level (Fig. 5, Table 2). Thus, monophyly of Madeira with the eastern Canary Islands (clades A, B) or with the main Canary Island clade (A, C) cannot be excluded. For the Timarcha 16S data the hypothesis of monophyly for the T. goettingensis complex could not be rejected (\( P > 0.25 \)) but was rejected for the ITS2 data (\( P < 0.01 \)) (Fig. 5, Table 2).

Figure 6 shows the distribution of differences for each of the seven hypotheses when less than adequate substitution models were used both on the empirical data and for generating the replicate sequences using simulation. The significance of each of the null hypotheses under each of the models is summarized in Table 2. All null hypotheses are rejected with all the suboptimal models with the exception of the Kimura 2-parameter + \( \Gamma \) model. For this model of substitution the \( P \)-value for null hypothesis 5 (AC monophyly) is 0.14, and that for the null hypothesis of monophyly for the T. goettingensis complex using 16S sequences is 0.17.

The simplest of the models, the Jukes Cantor, for which base frequencies are assumed to be equal with only one substitution rate, produces the most left skewed distribution for all seven hypotheses (Fig. 6). The Kimura 2-parameter model is an improvement on the Jukes Cantor model by allowing one substitutional rate for transitions and one for transversions, but this has an apparently minor effect on the distribution of the step length differences. For the Calathus COI and COII data and the Timarcha 16S data, the HKY85 model further refines the substitutional model by allowing for unequal base frequencies, and both the TIM and GTR models improve on the HKY85 model by allowing for more substitutional rates (4 and 6, respectively). For the Timarcha ITS2 data the K3P model is an improvement upon the K2P model by incorporating an additional substitutional rate. The TVMeF model further refines this by defining a total of five substitutional rates. These improvements impact on the distribution of expected differences in that the right tail is longer and the left skew is reduced, but again only marginally.

Fig. 4 Constrained trees constructed using neighbour joining with maximum likelihood distances under the TIM + I + \( \Gamma \) substitution model for the Timarcha 16S sequence data and the TVMeF + I + \( \Gamma \) substitution model for the Timarcha ITS2 sequence data. For each of these data sets unconstrained and constrained trees were also generated under five other suboptimal models (see text for details).
The addition of an estimate of the gamma shape parameter stands out as being an important factor in generating higher and more frequent step differences between the hypothesis constrained and unconstrained trees. This translates to a higher probability of accepting the null hypotheses, which we shall phrase in terms of failure to reject because the hypotheses were initially framed relative to an empirical tree which, although indicative, does not support them.

The addition of a gamma shape parameter to the GTR model meant the Calathus null hypotheses 4 (AB) and 5 (BC) could not be rejected. Adding a gamma shape parameter to the TIM model resulted in the failure to reject the Timarcha 16S null hypothesis. Similarly, the addition of a gamma shape parameter to the suboptimal Kimura 2-parameter model results in the inability to reject Calathus null hypothesis 4 and the Timarcha 16S null hypothesis. All these null hypotheses are rejected when a gamma shape parameter is not included. Although the gamma shape parameter stands out as the singularly greatest factor contributing to the reduction of the left skew of the distribution of expected differences, it is also clear that additional parameters that apparently cause little effect individually can be important in combination with the gamma shape parameter.

Two conclusions can be drawn from the above observations. First, whether a particular hypothesis of monophyly is rejected or accepted appears to be highly dependent on whether the nucleotide substitution model being used is optimal. Second, as the available substitution models are essentially improvements over a basic model using additional parameters, it seems safe to conclude that a parameter rich model is either equally or
Table 2 Results of parametric bootstrap analyses for hypotheses of monophyly for Macaronesian Calathus beetles and for the Timarcha goettingensis species complex under different models of sequence evolution. COI and COII sequence data were used for the former, and 16S and ITS2 sequence data in the case of T. goettingensis. See Fig. 1 for the five hypotheses of monophyly represented by letter combinations. The P-values represent the proportion of the 500 simulated data sets which yielded a length difference equal to or greater than the empirical difference between the hypothesis constrained and unconstrained tree. See text for explanation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Hypothesis</th>
<th>T. goettingensis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABCD</td>
<td>BCD</td>
</tr>
<tr>
<td>Jukes Cantor</td>
<td>P &lt; 0.01</td>
<td>P &lt; 0.01</td>
</tr>
<tr>
<td>K2P</td>
<td>P &lt; 0.01</td>
<td>P &lt; 0.01</td>
</tr>
<tr>
<td>K2P + Γ</td>
<td>P &lt; 0.01</td>
<td>P &lt; 0.01</td>
</tr>
<tr>
<td>GTR</td>
<td>P &lt; 0.01</td>
<td>P &lt; 0.01</td>
</tr>
<tr>
<td>GTR + Γ</td>
<td>P &lt; 0.01</td>
<td>P &lt; 0.01</td>
</tr>
<tr>
<td>TIM</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TIM + Γ</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>K3P</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TVMef</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TVMef + Γ</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

less likely to reject a hypothesis of monophyly where the optimal model is unknown.

Kishino–Hasegawa test

In contrast to the parametric bootstrap, the KH tests did not reject hypothesis 3 (ABC) for the Calathus for the two substitution models with Γ correction, GTR + Γ and Kimura 2-parameter + Γ. Hypothesis 4 (AB), which was rejected by the bootstrap test in the case of all suboptimal models, could not be rejected under any of the suboptimal substitution models. Similarly, the hypothesis of monophyly for the T. goettingensis complex with the 16S sequence data which was rejected by the bootstrap test in the case of all models lacking the Γ parameter, could not be rejected by any of the suboptimal substitution models including those lacking Γ correction. For hypothesis 5 (AC), only the GTR + Γ and the Kimura 2-parameter + Γ models failed to reject the null hypothesis (P ≥ 0.17). This and the fact that all suboptimal models resulted in the rejection of hypotheses 1 (ABCD), 2 (BCD) and the hypothesis of T. goettingensis monophyly with the ITS2 sequence data are consistent with the results of the bootstrap test.

Given that the KH test (Table 3) produced three times as many acceptable model/hypotheses combinations compared with the parametric bootstrap method (Table 2), we conclude that its performance is not as severely affected as the latter by the use of suboptimal models. Furthermore, as all the model/hypothesis combinations rejected by the KH test have also been rejected by the parametric bootstrap but not the vice versa, we conclude that testing hypotheses of monophyly using the parametric bootstrap method is more conservative. By this we mean that the KH test fails to reject hypotheses of monophyly more often than does the parametric bootstrap. It has recently been pointed out by Goldman et al. (2000) that the KH test will give inflated P-values unless both topologies being tested are defined a priori. However, they observed that a corrected test (Shimodaira & Hasegawa, 1999) also tends to fail to reject a null hypothesis more often than a parametric test.

It has been demonstrated that for a given data set, suboptimal models can often produce the same best supported tree as optimal models (Yang et al., 1994). However, from our analyses it appears that the use of suboptimal models for hypothesis testing using parametric methods can lead to hypothesis rejection, when the use of an optimal model results in failure to reject the hypothesis. In particular the need to account for a gamma distribution of rates over sites appears to be a singularly important parameter to reduce this type I error. This is more apparent for the parametric bootstrap, but also true to some extent for the KH test. For our examples of the KH test, when there is failure to reject the null hypothesis with the optimal model, but with a comparatively low P-value (Table 3, hypotheses ABC and AC), suboptimal models not incorporating a Γ parameter lead to hypothesis rejection. However when there is failure to reject the null hypothesis with a comparatively high P-value [Table 3, hypotheses AB, and T. goettingensis monophyly (16S)] for the optimal model, suboptimal models also result in failure to reject. We have used a broad range of empirical data with a number of relevant hypotheses to address the issue of choice of evolutionary models for hypothesis testing using parametric methods. The general applicability of our conclusions will be enhanced by additional support from a numerical simulation study which we plan to undertake.
Fig. 6 Distribution of steplength differences from parametric bootstrap analysis of five hypotheses of monophyly for Calathus beetles from Macaronesia for a dataset of COI and COII sequence data, and hypotheses of monophyly for the Timarcha goettingensis species complex for data sets of 16S and ITS2 sequence data. Hypotheses are tested under six different substitutional models. $x$ axis = steplength differences between constrained and unconstrained maximum parsimony analyses for each simulated data set. $y$ axis = the frequency of simulated data sets giving a particular steplength difference.
Table 3 Results of the Kishino–Hasegawa test for hypotheses of monophyly for Macaronesian *Calathus* beetles and for the *Timarcha goettingensis* species complex under different models of sequence evolution. COI and COII sequence data were used for the former, and 16S and ITS2 sequence data in the case of *T. goettingensis*. See Fig. 1 for the five hypotheses of monophyly represented by letter combinations. The P-values represent the proportion of the 500 simulated data sets which yielded a length difference equal to or greater than the empirical difference between the hypothesis constrained and unconstrained tree. See text for explanation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Hypothesis</th>
<th>T. goettingensis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16S</td>
<td>ITS2</td>
</tr>
<tr>
<td>Jukes Cantor</td>
<td>0.01</td>
<td>0.50</td>
</tr>
<tr>
<td>K2P</td>
<td>0.01</td>
<td>0.84</td>
</tr>
<tr>
<td>K2P + G*</td>
<td>0.01</td>
<td>0.89</td>
</tr>
<tr>
<td>HKY85</td>
<td>0.01</td>
<td>0.70</td>
</tr>
<tr>
<td>GTR</td>
<td>0.01</td>
<td>0.78</td>
</tr>
<tr>
<td>GTR + G*</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>TIM</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TIM + G*</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>K3P</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TVMef</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TVMef + G*</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Acknowledgments

We would like to thank Jesús Gómez-Zurita for kindly providing us with 16S and ITS2 sequence data for *Timarcha*, and two anonymous referees for helpful comments on an earlier draft.

References


Received 11 April 2001; accepted 17 April 2001