Syst. Biol. 53(6):949–962, 2004
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490888868

Empirical Problems of the Hierarchical Likelihood Ratio Test for Model Selection

DIEGO POL

Division of Paleontology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024, USA; E-mail: dpol@amnh.org

Abstract.—Advocates of maximum likelihood (ML) approaches to phylogenetics commonly cite as one of their primary advantages the use of objective statistical criteria for model selection. Currently, a particular implementation of the likelihood ratio test (LRT) is the most commonly used model-selection criterion in phylogenetics. This approach requires the choice of a starting point and a parameter addition (or removal) sequence that can affect all ML inferences (i.e., topology, model, and all evolutionary parameters). Here, several alternative starting points and parameter sequences are tested in empirical data sets to assess their influence on model selection and optimal topology. In the studied data sets, varying model-selection protocols leads to selection of different models that, in some cases, lead to different ML trees. Given the sensitivity of the LRT, some possible solutions to model selection (within the hypothesis testing approach) are outlined, and alternative model-selection criteria are discussed. Some of the suggested alternatives seem to lack these problems, although their behavior and adequacy for phylogenetics needs to be further explored. [hierarchical likelihood ratio test; maximum likelihood; model selection.]

Maximum likelihood (ML) methods of statistical inference are characterized by the likelihood function, which is proportional to the probability of observing the data given a probabilistic model and its associated parameters (Edwards, 1992). The point estimates of this method are obtained selecting the parameter values of the model that maximize the probability of observing the data. A wide variety of methods have been proposed for inferring trees from nucleic acid sequences, but during the last decade a particular kind of ML approach became increasingly popular, the maximum relative likelihood (sensu Steel and Penny, 2000). This approach was first proposed by Felsenstein (1973, 1981) for the analysis of nucleotide sequences and requires the specification of a probabilistic model of evolution in order to calculate the likelihood score of a given tree (topology plus a set of branch lengths) based on the sequence data.

Numerous probabilistic models of sequence evolution have been proposed to date, differing in their complexity and in which parameters are used to describe the process of nucleotide substitution. Many of these have been implemented in software packages, allowing researchers to use most of them in empirical analyses (e.g., Swofford, 2002; Yang, 1997; Felsenstein, 2002; Olsen et al., 1994). The wide range of available alternative models forces the question of what model is appropriate for the problem at hand. This question is central because it affects the likelihood-based inferences of the evolutionary process and can be a decisive factor in the estimation of the optimal topology. The influence of an appropriate model in likelihood analyses has been thoroughly documented in analytical (e.g., Chang, 1996), simulation-based (e.g., Khuner and Felsenstein, 1994; Tateno et al., 1994; Gaut and Lewis, 1995), and empirical studies (e.g., Sullivan and Swofford, 1997; Kelsey et al., 1999; Buckley et al., 2001).

Several authors considered one of the key advantages of the relative ML approach the availability of objective statistical methods for model selection (e.g., Huelsenbeck and Crandall, 1997; Sullivan and Swofford, 1997; Swofford et al., 1996). Several model-selection methods have been proposed in phylogenetics, including statistical hypothesis testing approaches such as the likelihood ratio test (Felsenstein, 1973, 1981; Goldman, 1993; Yang et al., 1995; Huelsenbeck and Crandall, 1997), information-theoretic approaches such as the Akaike information criterion (Kishino and Hasegawa, 1989; Tamura, 1994; Muse, 1999), or Bayesian methods (Morozov et al., 2000; Bollback, 2002; Suchard et al., 2002; Minin et al., 2003; Huelsenbeck et al., 2004). However, as Posada and Crandall (2001) noted, until recently it was remarkably common to find published empirical studies that used given model without including any kind of justification. Furthermore, as they noted, the LRT has been used much more extensively than the other model selection methods.

In phylogenetics, the LRT consists of successively performing multiple pairwise tests of goodness of fit between two models (a simpler [null hypothesis] and a more complex model [alternative hypothesis]). These tests are usually implemented by starting with a simple model and progressively adding parameters to the null model (if this increases significantly the likelihood score). Thus, the different parameters are successively added until the current null model is not rejected (e.g., Cunningham et al., 1998). Alternatively, a similar procedure can be done starting with the most complex model and progressively removing parameters (if this does not decrease significantly the likelihood score). Different parameters are successively removed until the current null model (the simpler model) is rejected (e.g., Frati et al., 1997; Wilgenbusch and de Queiroz, 2000). Due to the structure in which the multiple tests are conducted, Posada and Crandall (1998, 2001) dubbed these iterative approaches, involving the addition or elimination of parameters, the hierarchical LRT (hLRT).

This protocol is based on a statistical hypothesis-testing framework widely applied in linear regressions (Miller, 1990) and aims to select a model with high predictive power, a critical property in experimental sciences. A model is selected based on the bias-variance trade-off (Cox and Snell, 1974; Draper and Smith, 1998), choosing models with as many parameters as needed to obtain an adequate fit and as few parameters as possible in order to keep the total error of prediction reasonably small (and avoid excessive computational costs). The hypothesis-testing approach for model selection has been criticized based on different arguments regarding the use of χ^2 critical distribution in some cases (e.g., Yang et al., 1995; Zhang, 1999; Sanderson and Kim, 2000; Posada and Buckley, 2004), an arbitrary alpha value (e.g., Burnham and Anderson, 1998), and its general adequacy for the phylogenetic problem (e.g., Grant and Kluge, 2003).

However, even if the validity and all assumptions of the hLRT are accepted, some problems may remain in its current implementation. As noted by Sanderson and Kim (2000), this approach suffers from pitfalls such as the need for arbitrary choice between sequential addition or removal of parameters as well as election of the order in which they are to be removed. These arbitrary choices may influence which model is selected by the hLRT (Cunningham et al., 1998; Zhang, 1999; Sanderson and Kim, 2000; Posada and Crandall, 2001; Posada and Buckley, in press). In such cases, these choices could influence any ML estimation (or any Bayesian or distance-based analysis), possibly affecting our understanding and characterization of the underlying evolutionary process and the estimation of the ML topology.

Posada and Crandall (2001) evaluated the performance of four different implementations of hLRT in simulated data sets, concluding that these may select different models under some circumstances, but regardless, their accuracy in selecting the simulation model was almost identical most of the time. Cunningham et al. (1998) also tested different hLRT but using experimental phylogenies, arriving at different selected models with different parameter addition sequences (but see Posada and Crandall (2001) for a critique of their approach). Although the degree to which different implementations of the hLRT affect the results of empirical data analyses has not been yet assessed, most empirical ML analyses that used a model selection criterion used the single hLRT implemented in the program Modeltest (Posada and Crandall, 1998). Here, 32 variants of hLRT are explored for 18 empirical data sets in order to test their influence on the selected models and their effect on the estimation of the optimal topology.

Materials and Methods

Data Sets

Eighteen data sets were gathered for this study. Some were provided by the authors and the rest were downloaded from two websites (Treebase: http://herbaria.harvard.edu/treebase, and Systematic Biology's homepage: http://systematicbiology.org). The size of most of these varies from 9 to 25 taxa and 0.277 and 2.072 kb (a list of the matrices analyzed is given in Table 1), although 5 are much larger (i.e., 133 to 287 taxa and 0.686 to 1.850 kb).

Parameter Estimation Protocol

ML scores (with specifications of the parameter values) were estimated under the 56 different models of character change implemented in Modeltest 3.06 (Posada and Crandall, 1998). Although 56 different Markov models seem to be a somewhat large quantity, it represents only a small fraction of possible models (Sanderson and Kim, 2000).

ML estimations were performed using PAUP* 4.0b10 (Swofford, 2002). Optimal branch lengths were calculated using a maximum of 20 smoothing passes implemented in the Raphson-Newton iteration procedure. Nucleotide-substitution parameters were heuristically estimated using a parsimony-based approximation as starting point. The Γ distribution for heterogeneous rates was estimated through the discrete-Γ approximation proposed by Yang (1994) using four rate categories represented by their mean. Among the analyzed datasets,

TABLE 1. Data sets analyzed in this work. ntax = number of taxa included in the data matrix; bp = total number of aligned nucleotides included in the data set (number of parsimony-informative sites given between parentheses).

Data set	ntax	bp	Gene	Reference
1	9	903 (202)	Cyt b	Wiens and Whollingsworth (2000)
2	13	901 (227)	ND4 (+ tRNAs)	Wiens and Whollingsworth (2000)
3	11	1045 (156)	Cyt b + tRNAthr	McCracken et al. (1999)
4	16	987 (127)	12s	Lundrigan et al. (2002)
5	11	1007 (331)	12s	Springer et al. (1999)
6	18	1134 (283)	12s + tRNAval + 16s	Stanhoppe et al. (1998)
7	16	278 (47)	β2-microglobulin	Lundrigan et al. (2002)
8	16	1145 (321)	Cyt b	Lundrigan et al. (2002)
9	10	543 (138)	COI	Canatella et al. (1998)
10	16	1949 (206)	18s + 28s	Hedges et al. (1990)
11	25	2001 (615)	Cyt b + COI + COII	Wayne et al. (1997)
12	19	1189 (242)	12s + tRNAval,phe + 16s	Burk et al. (1998)
13	21	2072 (946)	18s	Halanych (1996)
14	200	1850 (412)	18s	Soltis et al. (1997)
15	133	1512 (704)	18s	Giribet and Ribera (1998)
16	200	1120 (601)	Cyt b	Johnson (2001)
17	287	1134 (653)	Cyt b	Weksler (personal communication)
18	140	686 (209)	rDNA ITS + trnL intr	Wojciechowski et al. (1999)

there were not cases of extreme rate heterogeneity (as measured by an $\alpha < 0.001$). However, those data sets in which the α parameter of the Γ distribution was optimized as smaller than 0.2 (for any model) were analyzed using eight rate categories. The estimation of the 56 likelihood scores was conducted on the same tree (arbitrarily chosen from the set of MPTs of each data set). The maximum parsimony trees were found through a heuristic search of 10 replicates of a random stepwise addition plus a round of TBR branch swapping for the smaller datasets. The larger data sets were analyzed using more efficient tree search strategies (Goloboff, 1999) implemented in the software TNT (Goloboff et al., 2000).

Hierarchical Likelihood Ratio Test

The likelihood scores of the assumed topology for each data set were evaluated using different variations of the hierarchical LRTs. This method is a classic statistical hypothesis testing of relative goodness of fit based in the statistic δ and follows this form:

$$\delta = 2(\ln L_1 - \ln L_0)$$

where L_0 is the likelihood under the null hypothesis (simple model) and L_1 is the likelihood under the alternative hypothesis (complex model). When the simpler model is a special case of the complex model (i.e., taking the derivatives of the likelihood function with respect to the parameters, the simpler model is a linear subspace of the more complex model's differentials), the significance of the δ statistic is usually assessed assuming it to be asymptotically distributed as χ^2 with q degrees of freedom, where q is the difference in number of free parameters between the two models (Yang et al., 1995; Huelsenbeck and Crandall, 1997; Posada and Crandall, 2001). In cases where the null model has the parameter being tested fixed at a boundary of its parameter space, a mixed χ^2 distribution was used to assess significance (Ota et al., 2000; Goldman and Whelan, 2000; Posada and Crandall, 2001). In order to evaluate all the considered models, this test was implemented successively performing pairwise comparisons of nested models in a hierarchical way (i.e., hLRT as defined by Posada and Crandall, 1998, 2001).

In order to test the effects of different hLRT in empirical datasets, 32 different hLRTs were employed here. These consisted of 16 different orders in which parameters were tested, starting either from the simplest model (bottom-up approach adding parameters) or from the most complex model (top-down approach deleting parameters). The different orders of parameter addition (or removal) used in these variants of the hLRTs are shown in Figure 1 and listed in Table 2.

The 32 different hLRTs were implemented here by modifying the original source code of Modeltest 3.06 (Posada and Crandall, 1998) that uses a particular bottom-up hLRT. Modeltest is a free software published under the GNU license, the source code of which is available at no charge from the author's website

TABLE 2. Sequence in which the parameters were tested in each of the 32 alternative hLRT implemented here. bu = bottom-up sequence in which the addition of parameters are successively tested; td = top-down sequence in which the removal of parameters are successively tested; π = parameters of base frequencies; $ts\neq tv$ = transition/transversion rate ratio parameter; $ts_1\neq ts_2$ = independent rate parameters for each of the two transitions types; $tv_1 \neq tv_2$ = independent rate parameters for two different transversion types; $tv_1 \neq tv_2 \neq tv_3\neq tv_4$ = independent rate parameters for each of the four transversion types; Γ = rate heterogeneity parameter using gamma distribution; p_{inv} = invariants parameter.

hLRT	Parameter addition (or removal) sequence
bu1	$\pi \rightarrow ts\neq tv \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow \Gamma \rightarrow p_{inv}$
bu2	$\pi\ ts\neq tv \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow ts_1\neq ts_2 \rightarrow \Gamma \rightarrow P_{inv}$
bu3	$\pi\ ts\neq tv \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow p_{inv} \rightarrow \Gamma$
bu4	$\pi\ ts\neq tv \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow ts_1\neq ts_2 \rightarrow p_{inv} \rightarrow \Gamma$
bu5	$\Gamma \rightarrow p_{inv} \rightarrow \pi\ ts\neq tv \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4$
bu6	$\Gamma \rightarrow p_{inv} \rightarrow \pi\ ts\neq tv \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow ts_1\neq ts_2$
bu7	$p_{inv} \rightarrow \Gamma\pi\ ts\neq tv \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4$
bu8	$p_{inv} \rightarrow \Gamma\pi\ ts\neq tv \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow ts_1\neq ts_2$
bu9	$\pi\ \Gamma \rightarrow p_{inv} \rightarrow ts\neq tv \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4$
bu10	$\pi \rightarrow \Gamma \rightarrow p_{inv} \rightarrow ts\neq tv \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow ts_1\neq ts_2$
bu11	$\pi\ p_{inv} \rightarrow \Gamma\ ts\neq tv \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4$
bu12	$\pi\ p_{inv} \rightarrow \Gamma\ ts\neq tv \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow ts_1\neq ts_2$
bu13	$ts\neq tv \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2 \neq tv_3\neq tv_4 \rightarrow \pi\Gamma \rightarrow p_{inv}$
bu14	$ts\neq tv \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow ts_1\neq ts_2 \rightarrow \pi\Gamma \rightarrow P_{inv}$
bu15	$ts\neq tv \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow \pi\ p_{inv} \rightarrow \Gamma$
bu16	$ts\neq tv \rightarrow tv_1\neq tv_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow ts_1\neq ts_2 \rightarrow \pi\ p_{inv} \rightarrow \Gamma$
td1	$\pi \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts_1\neq ts_2 \rightarrow ts\neq tv \rightarrow \Gamma \rightarrow p_{inv}$
td2	$\pi \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts\neq tv \rightarrow \Gamma \rightarrow p_{inv}$
td3	$\pi \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts_1\neq ts_2 \rightarrow ts\neq tv \rightarrow p_{inv} \rightarrow \Gamma$
td4	$\pi \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts\neq tv \rightarrow p_{inv} \rightarrow \Gamma$
td5	$\Gamma \rightarrow p_{inv} \rightarrow \pi \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts_1\neq ts_2 \rightarrow ts\neq tv$
td6	$\Gamma \rightarrow p_{inv} \rightarrow \pi \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts\neq tv$
td7	$p_{inv} \rightarrow \Gamma \rightarrow \pi \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts_1\neq ts_2 \rightarrow ts\neq tv$
td8	$p_{inv} \rightarrow \Gamma \rightarrow \pi \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts\neq tv$
td9	$\pi \rightarrow \Gamma \rightarrow p_{inv} \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts_1\neq ts_2 \rightarrow ts\neq tv$
td10	$\pi \rightarrow \Gamma \rightarrow p_{inv} \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts\neq tv$
td11	$\pi \rightarrow p_{inv} \rightarrow \Gamma \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts_1\neq ts_2 \rightarrow ts\neq tv$
td12	$\pi \rightarrow p_{inv} \rightarrow \Gamma \rightarrow ts_1\neq ts_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts\neq tv$
td13	$tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts_1\neq ts_2 \rightarrow ts\neq tv \rightarrow \pi \rightarrow \Gamma \rightarrow p_{inv}$
td14	$ts_1\neq ts_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts\neq tv \rightarrow \pi \rightarrow \Gamma \rightarrow p_{inv}$
td15	$tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts_1\neq ts_2 \rightarrow ts\neq tv \rightarrow \pi \rightarrow p_{inv} \rightarrow \Gamma$
td16	$ts_1\neq ts_2 \rightarrow tv_1\neq tv_2\neq tv_3\neq tv_4 \rightarrow tv_1\neq tv_2 \rightarrow ts\neq tv \rightarrow \pi \rightarrow p_{inv} \rightarrow \Gamma$

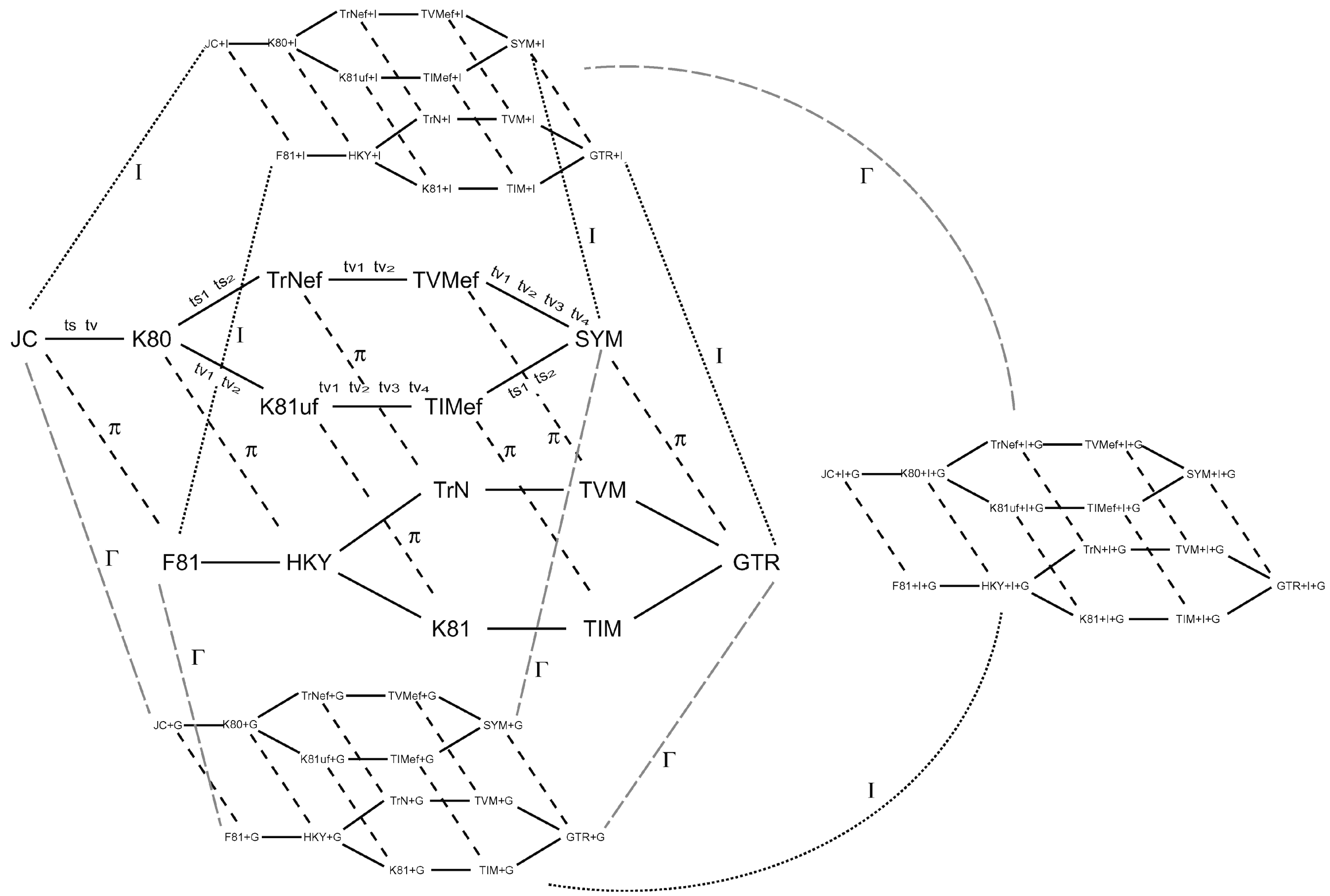

FIGURE 1. Relationship among the 56 models used in this study. These models are linked by lines representing the individual likelihood ratio tests used to choose between two given models in each step of the hLRT (i.e., testing the addition or removal of the parameters in which the two models differ). Solid black lines represent LRT of nucleotide-substitution parameters (the parameters tested are specified above each line one of the set of models). Dashed black lines represent LRT testing the presence of unequal base frequencies (π parameter). Dashed gray lines represent LRT testing the presence of rate heterogeneity through the gamma distribution (Γ). Dotted black lines represent LRT testing the inclusion of the invariants parameter (p_{inv}). The models of DNA substitution follow the notation of Posada and Crandall (1998): JC (Jukes and Cantor, 1969), F81 (Felsenstein, 1981), K80 (Kimura, 1980), HKY (Hasegawa et al., 1985), TrN (Tamura and Nei, 1993), TrNef (TrN equal base frequencies), K81 (Kimura, 1981), K81uf (K81 unequal base frequencies), TIM (Posada, 2003), TVM (Posada, 2003), GTR (Rodríguez et al., 1990), TIMef (TIM with equal base frequencies), TVMef (TVM with equal base frequencies), and SYM (Zharkikh, 1994). The addition of G and I to the model name represents the inclusion of rate heterogeneity using the gamma distribution and the proportion of invariable sites, respectively.

http://darwin.uvigo.es/software/modeltest.html). In addition to the hLRT, the original source code of Modeltest was also used to calculate the Akaike information criterion (AIC), an information-theoretic approach to model selection (as described in Posada and Crandall, 2001). The modified versions made for this study are similarly available at request from the author of this paper.

Effects on Topology Estimation

In order to test the influence of the different models selected by distinct hLRT, heuristic ML searches assuming the selected model (with their estimated parameter values) were performed for the smaller datasets. These analyses were conducted using the most dissimilar models selected for each of these datasets (in particular those models that differ markedly in their substitution parameters). The heuristic likelihood searches were conducted in PAUP* 4.0b10 (Swofford, 2002) and consisted of performing a round of TBR branch swapping on the most parsimonious tree used for selecting the optimal models. In cases where different selected models produced different ML trees, more exhaustive tree search strategies were conducted (i.e., 10 replicates of random addition sequences followed by TBR branch swapping). After the completion of these tree searches, additional rounds of model selection procedure were conducted using the same hLRT but using the ML tree (instead of the parsimony tree as in the original model selection). The newly selected models were then assumed during a subsequent heuristic likelihood tree search. This iterative procedure of model selection and tree estimation was performed until a stable solution was achieved (following the protocol proposed by Sullivan and Swofford, 1997).

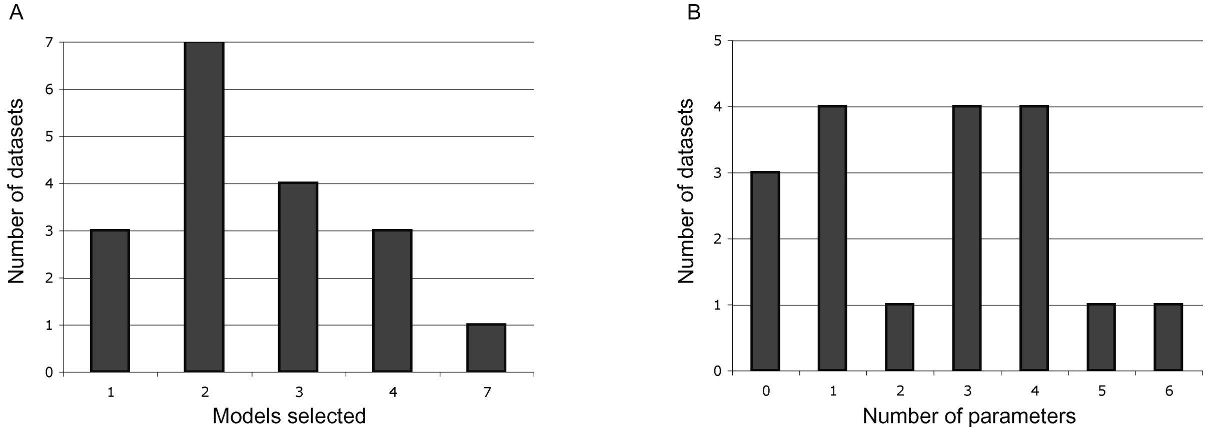

FIGURE 2. Influence of the different hLRTs in the selected models. (A) Frequency histogram of number of different models selected by different hLRTs per data set. (B) Frequency histogram of maximum difference in number of parameters between the selected models in each data set.

RESULTS

The analyses of the data sets considered here showed that, more often than not, different hLRTs selected different models. In 15 out of the 18 data sets analyzed, there were at least two different substitution models selected as optimal depending on the hLRT used to evaluate them. In some cases, only two models were alternatively selected as optimal, but in other datasets up to seven different models were considered as optimal by different hLRT (Fig. 2A). For some data sets (4 out of the 18), the different selected models were very similar, differing only in one parameter. However, in more than half of the matrices analyzed here (10 data sets), the selected models varied markedly in their complexity (differing in a maximum of three to six parameters; see Fig. 2B).

Bottom-Up versus Top-Down

The effect of these two approaches was compared when the order of parameters addition/removal was identical (e.g., bu1 versus td1). Different models were selected as optimal in 39% to 61% of the data sets, depending on the order in which parameters were tested (see Table 3 for a detailed summary). The top-down approach often selected a more parameter-rich model than the bottom-up approach (Table 3); however, for some data sets the bottom-up approach yielded more complex models for most parameter addition sequences (Fig. 3). The AIC was also employed to select the optimal model in each of these datasets (Table 4). This measure usually selected more complex models than most of the implementations of hLRT. In most of the other data sets (11 out of the 18), the AIC selected the most complex model of the set of models selected by the different hLRTs. Furthermore, in two of the analyzed data sets (data sets 7 and 13), the AIC selected more complex models than any of the models selected by the different hLRTs. Despite these trends, the outcome of AIC and hLRT is still data set dependent, because in three of the analyzed data sets (2, 3, and 16), some of the hLRTs selected a more complex model than the AIC. Additionally, in two data sets the AIC and the 32 different hLRTs selected the same model.

In sum, the AIC and the two LRT approaches usually selected different models and in most cases the top-down hLRT and the AIC yielded more complex models, although this seems to be data set dependent.

Order of Parameters

The multiple pairwise models tests can be done in different ways depending on the order in which parameters are being added or deleted. In the results obtained here, the bottom-up hLRT was much more affected by parameter addition order than the top-down approach. The 16 variants of the top-down approach selected a unique model in 16 data sets, although in data sets 6 and 7 two alternative models were selected in each case (depending

TABLE 3. Summary of comparisons between the bottom-up and top-down approach. dif = number of data sets in which an equivalent top-down and bottom-up approach (i.e., that use the same parameter-addition/removal sequence) yielded different models; bu = number of datasets in which a bottom-up approach; yielded more complex models than a top-down approach; td = number of data sets in which a bottom-up approach yielded more complex models than a bottom-up approach.

hLRT	dif	bu	td
1	7	3	3
2	7	3	3
3	9	6	3
4	9	6	3
5	9	1	8
6	8	1	7
7	10	3	7
8	9	3	6
9	8	0	8
10	7	0	7
11	10	3	7
12	9	3	6
13	10	4	5
14	10	4	5
15	11	6	5
16	11	6	5

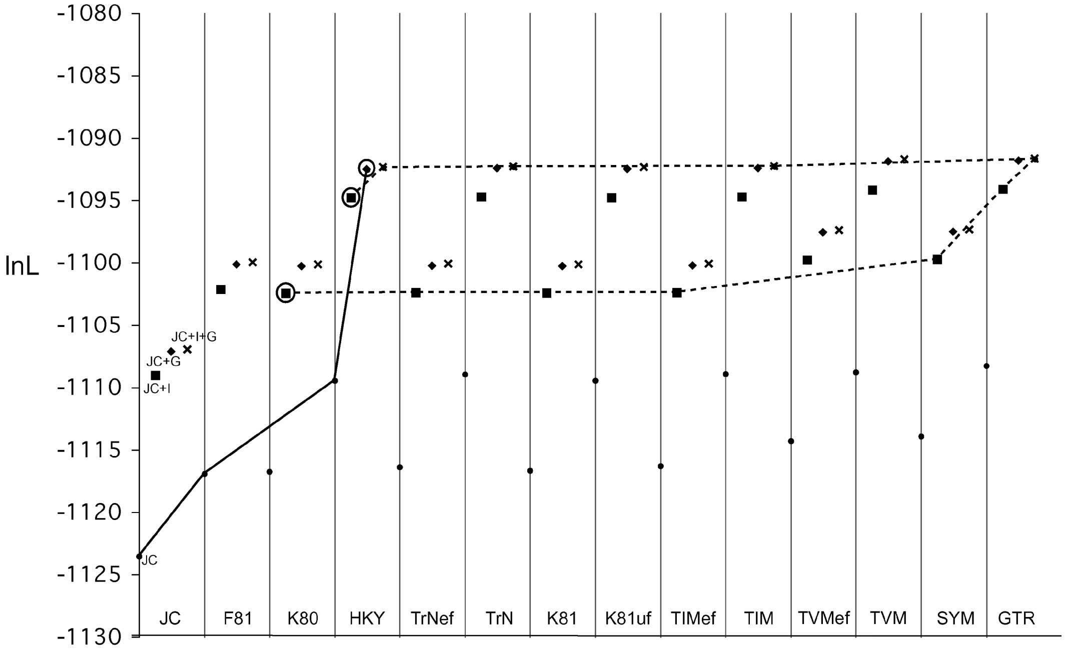

FIGURE 3. Paths of the alternative hLRTs for the data set 7 (Lundrigan et al., 2002). The y-axis represents the likelihood scores of each of the 56 different models on the maximum parsimony tree assumed during the initial round of model selection. The 56 models are ordered along the abscissa according to their substitution and base frequency parameters. Within each subdivision the models are ordered from left to right: no rate heterogeneity (●), with invariants parameter (■), with gamma distribution (♦), and with invariants plus gamma distribution (+). Solid lines represent the path of bottom-up hierarchical likelihood ratio test from the simplest model (JC) to the model selected as optimal. Dashed lines represent the path of top-down hLRT, from the most complex model (GTR+I+Γ) to the model selected as optimal. At the end of these paths, the selected models are circled. Each of the solid line-segments that form the path of a bottom-up hLRT represents an individual likelihood ratio test that accepted the alternative (more complex) model (i.e., the addition of the parameter being tested). Each of the dashed line-segments that form the path of a top-down hLRT represents an individual likelihood ratio test that rejected the alternative (more complex) model (i.e., accepted the removal of the parameter being tested). In this data set, bottom-up parameter-addition sequences selected either HKY+Γ (e.g., bu1 shown in figure) or HKY+I (e.g., bu3). Both models are more complex than the one selected by the top-down parameter-removal sequences td5 and td6 (dashed lines).

on the parameter removal sequence; see Fig. 3). In contrast, the 16 different bottom-up hLRTs showed a marked dependence on the order of parameter addition. This approach selected different models in 12 out of the 18 data sets and sometimes these models differed widely in their constituent parameters (Fig. 4).

In half of the data sets, differences among the results of the bottom-up hLRT involved the presence of rate heterogeneity parameters. Usually, the inclusion of any of these parameters greatly increases the likelihood score. However, in all the cases analyzed here, the gamma distribution parameter (α) increased the likelihood score more drastically than the inclusion of the invariants parameter (p_{inv}). When the presence of the invariants parameter is tested before the gamma distribution, the selected model usually have a lower score than an alternative model with the same number of parameters (e.g., selecting HKY+I instead of HKY+Γ; Fig. 5) or have more parameters than needed (e.g., selecting GTR+I+Γ when GTR+Γ is equally optimal; Fig. 6).

In seven of the analyzed data sets, the results of the alternative bottom-up hLRT resulted in selected models that differ in their substitution parameters. In these cases, the results of the significance tests of the substitution parameters were highly affected by the presence of other parameters in the pair of models being tested. This suggest the lack of independence between most of the parameters involved in these models, a dependence previously noted by several researchers (e.g., Sullivan et al., 1996, 1999). As expected, the presence of nucleotide frequency (π) and rate heterogeneity parameters (p_{inv} and α) often caused the selection of models with fewer substitution parameters than when substitution parameters are tested before the frequency and heterogeneity parameters (Fig. 7). However, in some cases, if nucleotide frequency parameters (π) are tested first, the selected model has more nucleotide-substitution parameters (Fig. 8).

Topology Estimation

The different models selected by the alternative hLRT can, theoretically, affect the estimation of the ML topology. In order to assess the degree to which this happens, heuristic ML searches were conducted for some data sets assuming the models selected by the hLRTs. Due to time constraints, this was done only for some of the smaller data sets (matrices 1, 2, 3, 5, 8, 9, 11, and 12), alternatively assuming the most dissimilar models selected by the different hLRTs. In six out of the eight data sets, the different models assumed selected the same ML topology in the

TABLE 4. Summary of the alternative models selected by different hLRTs and AIC for each dataset during the initial round of model selection. Examples of the hLRTs that found each of the selected models is given (bu$_n$ and td$_n$ represent cases in which every order of parameter-addition or parameter-removel resulted in the same model).

Data set	hLRT model	AIC model
1	TVM+Γ (bu$_1$)	GTR+I+Γ
	TVM+I+Γ (bu$_3$)	
	GTR+Γ (bu$_{13}$)	
	GTR+I+Γ (bu$_{16}$)	
	TVM+Γ (td$_n$)	
2	GTR+Γ (bu$_1$)	K81uf+I+Γ
	GTR+I+Γ (bu$_3$)	
	K81uf+I+Γ (bu$_{11}$)	
	K81uf+Γ (td$_n$)	
3	GTR+Γ (bu$_1$)	GTR+Γ
	GTR+I+Γ (bu$_3$)	
	TrN+I+Γ (bu$_7$)	
	TVM+Γ (bu$_{13}$)	
	TVM+I+Γ (bu$_{15}$)	
	TrN+Γ (td$_1$)	
	HKY+Γ (td$_2$)	
4	GTR+Γ (bu$_1$, td$_n$)	GTR+I+Γ
	GTR+I+Γ (bu$_3$)	
5	GTR+I+Γ (bu$_3$)	GTR+I+Γ
	TrN+Γ (bu$_5$)	
	TrN+I+Γ (bu$_9$)	
	GTR+Γ (td$_n$)	
6	HKY+Γ (bu$_1$)	TVM+I+Γ
	HKY+I+Γ (td$_5$)	
7	HKY+G (bu$_1$)	HKY+Γ
	HKY+I (bu$_3$)	
	K80+I (td$_5$)	
8	TrN+I+Γ (bu$_5$)	GTR+I+Γ
	GTR+I+Γ (bu$_{1,}$td$_n$)	
9	GTR+I+Γ (bu$_n$, td$_n$)	GTR+I+Γ
10	GTR+I+Γ (bu$_n$, td$_n$)	GTR+I+Γ
11	HKY+I+Γ (bu$_5$)	GTR+I+Γ
	GTR+I+Γ (bu$_{1,}$td$_n$)	
12	TrN+Γ (bu$_1$)	GTR+I+Γ
	TrN+I+Γ (bu$_3$)	
	GTR+I+Γ (td$_n$)	
13	TrN+I+Γ (bu$_n$, td$_n$)	TIM+I+Γ
14	TrN+I+Γ (bu$_n$)	GTR+I+Γ
	GTR+I+Γ (td$_n$)	
15	TrNef+I+Γ (bu$_1$)	GTR+I+Γ
	TrN+I+Γ (bu$_5$)	
	GTR+I+Γ (td$_n$)	
16	GTR+I+Γ (bu$_1$)	TVM+I+Γ
	TVM+I+Γ (td$_n$)	
17	TVM+I+Γ (bu$_5$)	GTR+I+Γ
	GTR+I+G (td$_n$)	
18	TrN+I+Γ (bu$_1$)	SYM+I+Γ
	TrNef+I+Γ (bu$_{15}$)	
	SYM+I+Γ (td$_n$)	

initial heuristic tree search. In data sets 5 and 11, however, the two models used here resulted in different ML topologies in the initial tree searches.

In data set 5, the bottom-up approach hLRT$_{bu3}$ selected the most complex model (GTR+I+Γ), whereas an alternative bottom-up approach, hLRT$_{bu5}$, selected a simpler model (TrN+Γ). ML searches under the two models (assuming the optimal parameter values) yielded different ML trees. Additional iterations of hLRT model selection (applying hLRT$_{bu3}$ and hLRT$_{bu5}$, assuming the respective ML topologies) were conducted applying iteratively the procedure proposed by Sullivan and Swofford (1997). This procedure consistently yielded the same dissimilar results for these two hLRTs, as in the initial model selection procedure (and consequently the same two alternative topologies shown in Fig. 9). In this case, if these12S rRNA mammal sequences are analyzed, the arbitrary choice between hLRT$_{bu3}$ and hLRT$_{bu5}$ results in choosing between these two hypotheses on the evolutionary history of this group (Fig. 9).

In dataset 11, hLRT$_{bu1}$ selected the most complex model (GTR+I+Γ), whereas another bottom-up approach, hLRT$_{bu5}$, selected a much simpler model (HKY+I+Γ) (Fig. 10). The initial tree searches resulted in different maximum likelihood trees for each of these two models (Fig. 11). As in data set 5, the use of the iterative procedure (Sullivan and Swofford, 1997) of hLRT model selection consistently yielded the same dissimilar results as the initial model selection procedure (and consequently the same two alternative topologies shown in Fig. 11).

It must be noted that in both cases the alternative topologies are similar to each other (differing in either one or two NNI) and similarly supported by the data. For instance, in data set 5, the likelihood scores of the two alternative topologies are notably similar (under GTR+I+Γ: −6439.10142 versus −6439.65909, and under TrN+Γ: −6469.89133 versus −6470.18839). The only clade supported under GTR+I+Γ but not under TrN+Γ (i.e., mouse/rat+rabbit) is only moderately supported by the data (Bayesian posterior probability under GTR+I+Γ is 0.753). Similarly, in data set 11, the likelihood scores are very similar under GTR+I+Γ (−13219.50525 versus −13219.44782) and under HKY+I+Γ (−13226.32427 versus −13226.61326). In this case, support for the alternative clades is markedly low (Bayesian posterior probability under GTR+I+Γ for the clade 1 is 0.639 and for clade 2 is 0.117; see Fig. 11). Thus, in these cases, topological differences between different models appear to involve only poorly supported clades. It could be argued, as suggested by one of the reviewers, that these topological differences are not a real problem because the data do not show a large degree of support to either of the two trees (or models). Clearly, in these cases, the uncertainties in model and topology can be incorporated in the results of the phylogenetic analysis. However, these uncertainties (and possibly others) may remain hidden when a single (or even a few) hLRT scheme is conducted.

DISCUSSION

As previously recognized (Sanderson and Kim, 2000), there is no justification for preferring any one of the multiple options of sequence of parameter addition (or removal). The results presented here show that, in most of the empirical data sets analyzed here, more than one model is selected as optimal by the different sequences of parameter addition of the alternative hLRT. Furthermore, in some cases, the alternative models can lead to different ML topologies. In some cases, a pair of alternative models selected by the different hLRTs could

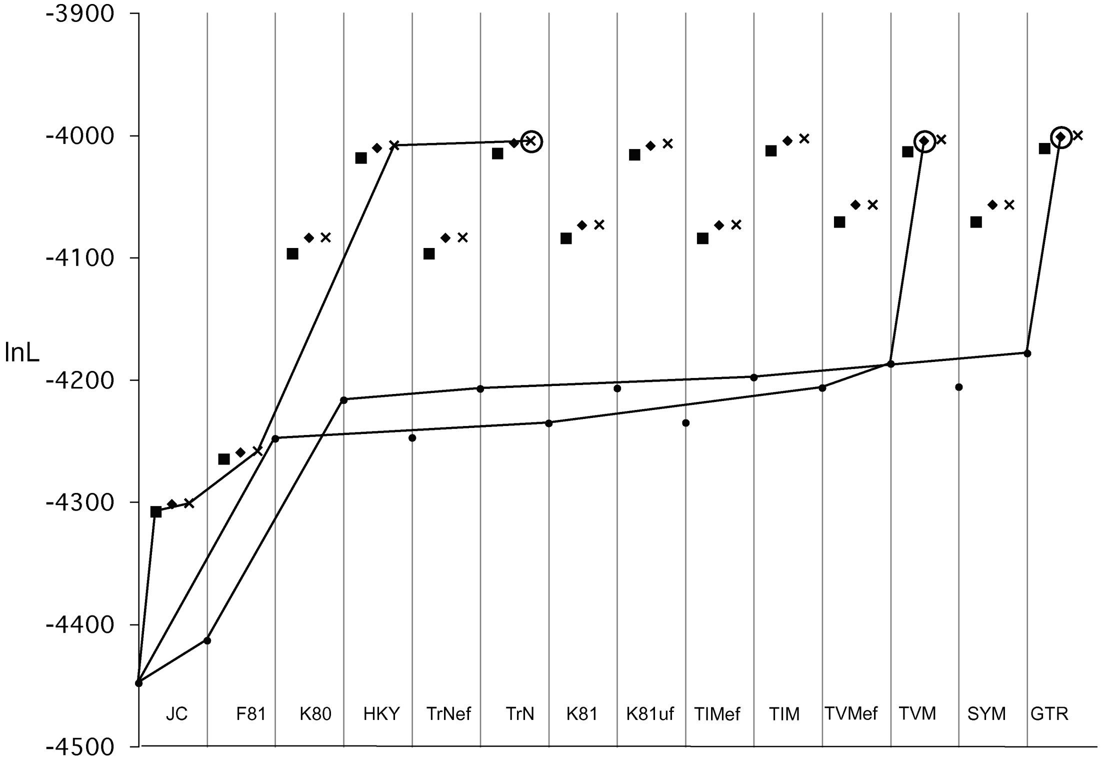

FIGURE 4. Paths of the alternative hLRTs for data set 3 (McCracken et al., 1999) in which three different bottom-up parameter-addition sequences selected three different models (TrN+I+Γ, TVM+Γ, and GTR+Γ) of dissimilar complexity (i.e., number of parameters). The graph follows the same convention as Figure 3.

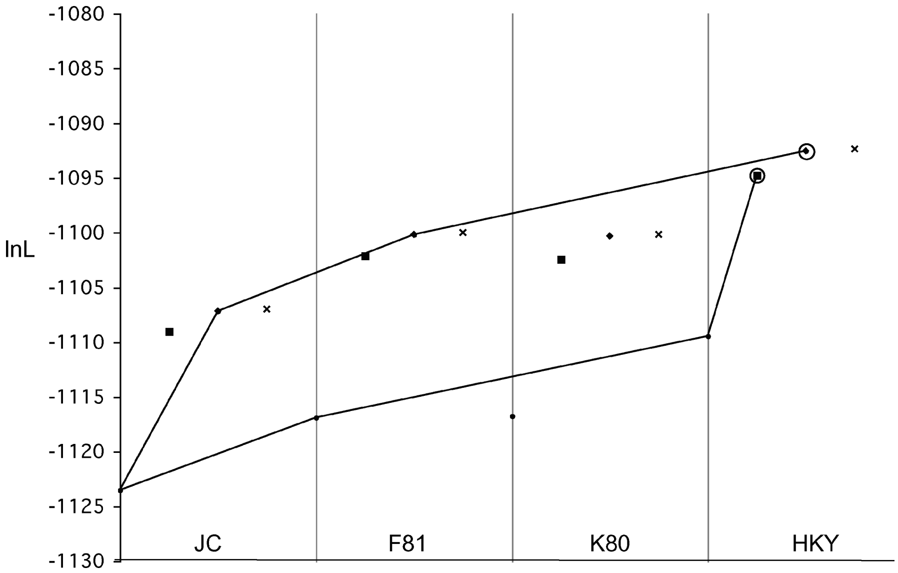

FIGURE 5. Paths of the alternative hLRTs for data set 7 (Lundrigan et al., 2002). Two bottom-up parameter-addition sequences are shown. In one of them the invariants parameter is tested before the gamma distribution. This hLRT selects a model (HKY+I) with lower score and equal number of parameters than the model selected when the gamma distribution is tested before the invariants (HKY+Γ). The graph follows the same convention as Figure 3.

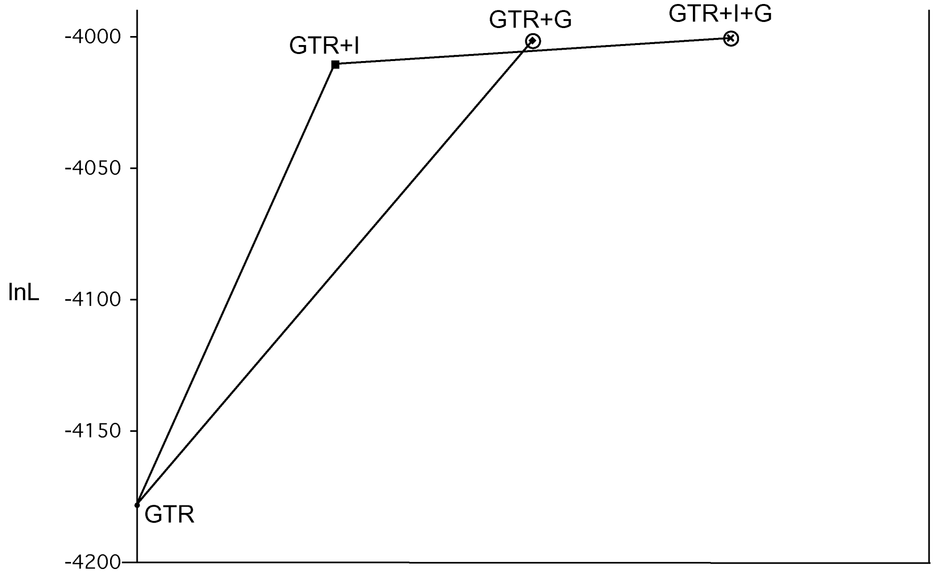

FIGURE 6. Paths of the alternative hLRTs for data set 3 (McCracken et al., 1999). Final steps of two bottom-up parameter-addition sequence are shown. In one of them the invariants parameter is tested before the gamma distribution. This hLRT selects a model (GTR+I+Γ) with nonsignificantly different score and higher number of parameters than the model selected when the gamma distribution is tested before the invariants (GTR+Γ). The graph follows the same convention as Figure 3.

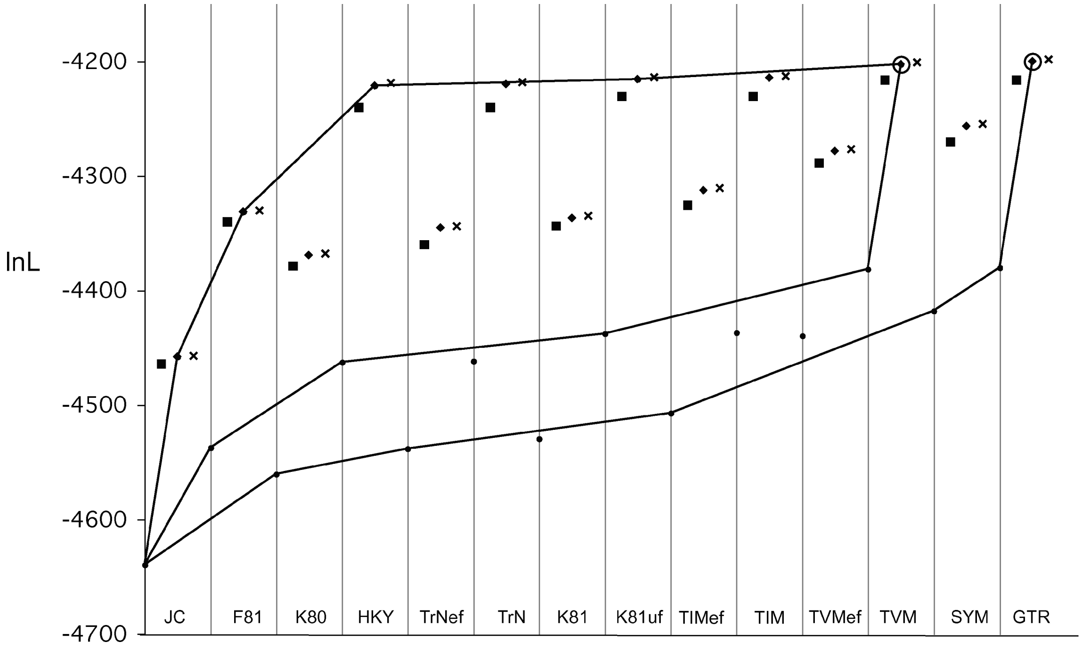

FIGURE 7. Paths of the alternative hLRTs for data set 1 (Wiens and Whollingsworth, 2000). Three bottom-up parameter-addition sequences are shown. In two of them, the rate heterogeneity or base frequency are tested before the substitution parameters. These two hLRTs select a simpler model (TVM+Γ) than the model selected when the substitution parameters are the first to be tested (GTR+Γ). The graph follows the same convention as Figure 3.

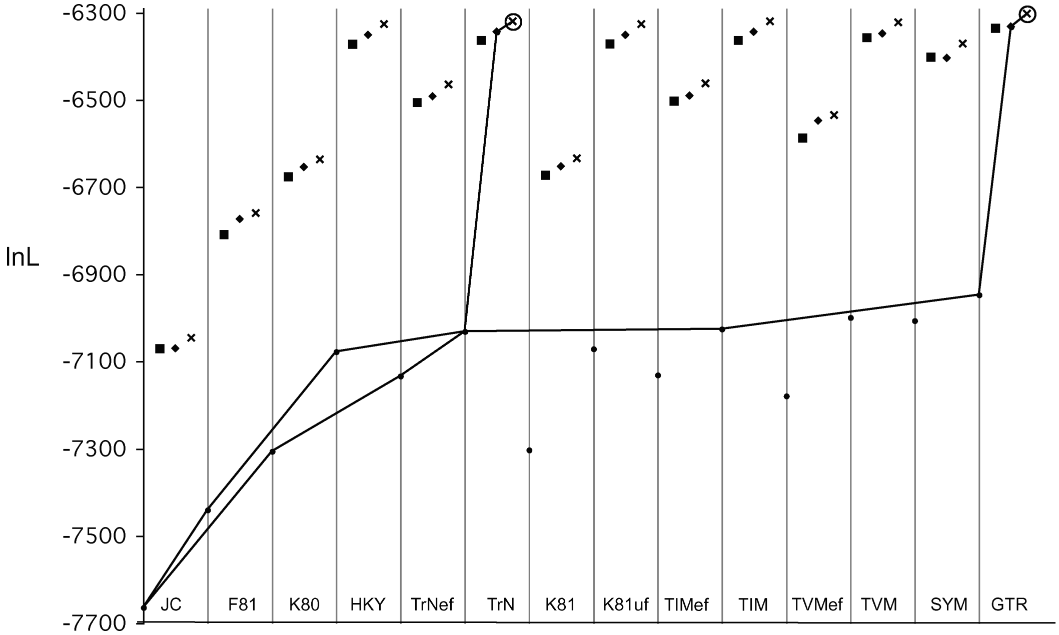

FIGURE 8. Paths of the alternative hLRTs for data set 8 (Lundrigan et al., 2002). Two bottom-up parameter-addition sequences are shown. In one of them, the base frequency is tested before the substitution parameters. This hLRT selects a more complex model (GTR+I+Γ) than the model selected when the substitution parameters are tested first (TrN+I+Γ). The graph follows the same convention as Figure 3.

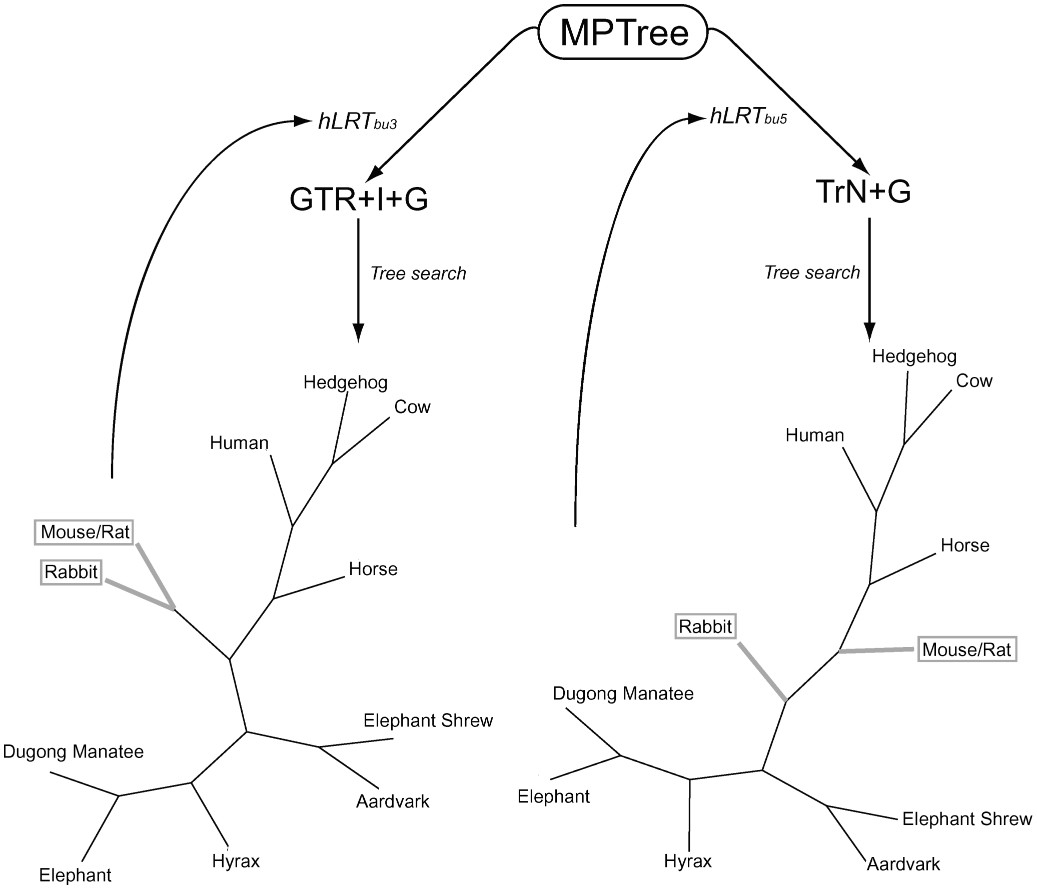

FIGURE 9. Scheme of the protocol of two alternative iterative sequences of hLRTs (a top-down and a bottom-up approach) and likelihood heuristic tree searches for data set 5 (12S rRNA of Springer et al., 1999) that results in two different models and topologies.

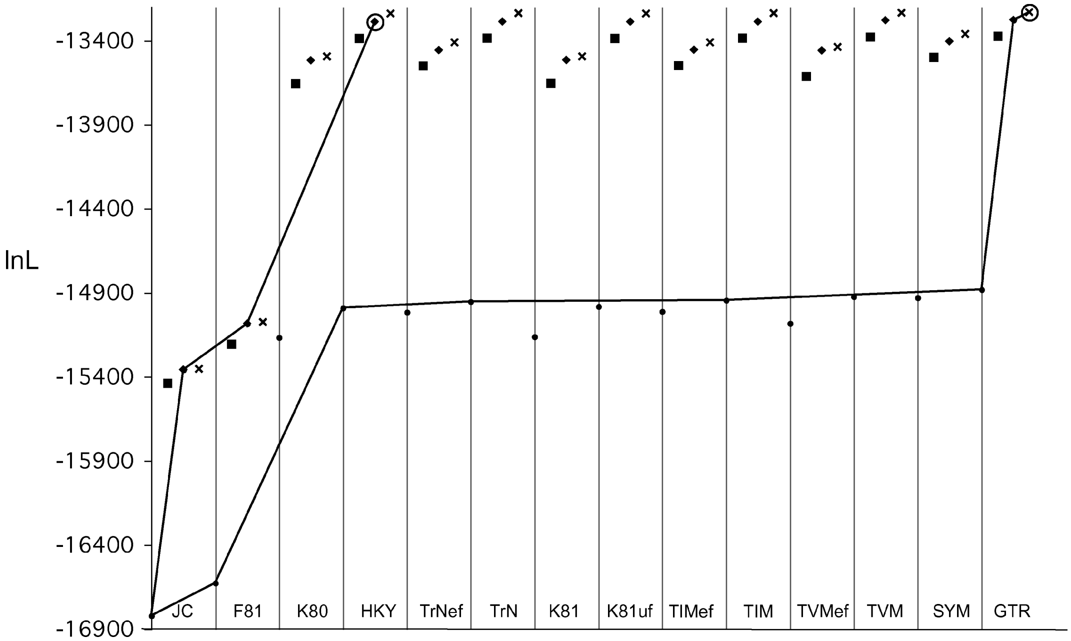

FIGURE 10. Paths of the alternative hLRTs for data set 11 (Wayne et al., 1997). The two different bottom-up parameter-addition sequences showed here select drastically different models that causes differences in their respective maximum likelihood topologies. The graph follows the same convention as Figure 3.

be compared through an additional LRT. However, these comparisons suffer from some limitations: (1) they can only be done in the case of nested models (or alternatively requiring the time-consuming Monte Carlo procedures); and (2) when more than two models are selected, multiple comparisons are needed (requiring the problematic choice of testing order and starting point).

As noted by Sanderson and Kim (2000), these problems are common to any sequential hypothesis-testing approach, as seen in the case of model selection for multiple linear regressions. In contrast to the recent concern (but see Minin et al., 2003) on this problem within phylogenetics, it has been extensively treated in the field of linear regressions since the availability of computers allowed the consideration of numerous alternative models and the dangers of mechanically using automatic and prefixed model selection procedures became evident (Miller, 1984, 1990; Draper and Smith, 1998). Possible improvements to current implementations of hLRTs include the commonly used forward selection or backward elimination methods (Hamaker, 1962; Abt, 1967; Mantel, 1970; Cox and Snell, 1974), in which the order of parameter addition is not predetermined but is decided considering which of the alternative models has a better fit (similar to the dynamic LRT of Posada and Crandall, 2001). Additionally, Efroymson (1960) proposed the stepwise procedure as a solution to this problem. Here, the LRTs are also conducted sequentially, but alternatively testing the addition of a parameter to the current model and the removal of a parameter from the current model until the algorithm converges. These and many other solutions could be implemented for model selection in phylogenetics and could eliminate the need of the above-mentioned arbitrary decisions in ML analyses. However, despite these possible solutions to these problems, they need to be further explored.

Alternatively, other model-selection criteria have been proposed that differ from the classic statistical testing approach to model selection, since it was noted to be a frequently abused and poorly understood method of applied statistics that lacks a firm theoretical base (Copas, 1984; Burnham and Anderson, 1998). Some options to the hypothesis-testing approach include information-theoretic approaches such as those based on the AIC (Kishino and Hasegawa, 1989; Tamura, 1994; Muse, 1999; Posada and Buckley, in press) or Bayesian methods (Morozov et al., 2000; Bollback, 2002; Suchard et al., 2002; Minin et al., 2003; Huelsenbeck et al., 2004). These avoid some of the problems noted here for hLRTs because they evaluate all the models simultaneously, avoiding all the problems related with the starting point and the sequence of parameter addition (or removal). Furthermore, they also avoid the need for critical distributions and arbitrary significance values, seen as pitfalls of the hypothesis-testing approach by some authors (e.g., Burnham and Anderson, 1998; Sanderson and Kim, 2000).

Some of these methods also allow incorporating uncertainty in model selection. This issue is of special interest because, in some cases, the data may not contain enough information to clearly discriminate among all possible models (with a precise estimate of all their parameters), being the likelihood scores of alternative models remarkably similar to each other. In these cases, the data sets will probably be prone to have sensitivity in the parameter-addition sequence and starting point during model

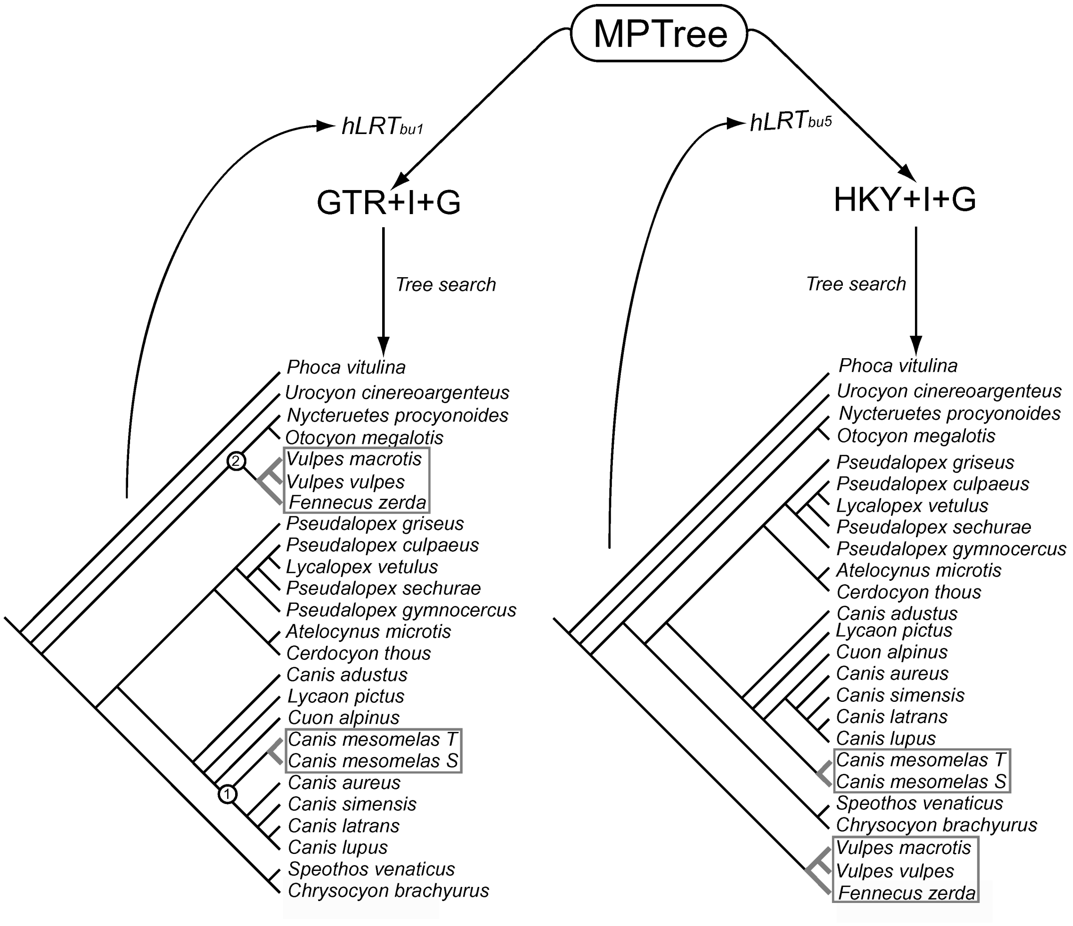

FIGURE 11. Scheme of the protocol of two alternative iterative sequences of bottom-up hLRTs and likelihood heuristic tree searches for data set 11 (Wayne et al., 1997) that results in two different models and topologies.

selection by hLRT. Cases such as data sets 5 and 11 would certainly be benefited from the inclusion of model uncertainty within the phylogenetic analysis. If these data sets cannot discriminate among the competing models selected by different hLRTs, the topological differences of their ML trees (of the alternative models) should be regarded as uncertainties in the ML analysis of the phylogenetic relationships of these groups.

Finally, it is important to note that the effects of the alternative approaches to hLRT (e.g., top-down versus bottom-up) and their final consequences in the likelihood analysis seem to be markedly data set dependent. This pattern (and the small number of data sets sampled here) precludes the formulation of general recommendations on which hLRT would be best for empirical analyses.

CONCLUSIONS

The existence of objective, statistical, model-selection procedures is commonly cited as an important advantage of ML approaches to phylogenetics (Swofford et al., 1996; Huelsenbeck and Crandall, 1997; Sullivan and Swofford, 1997), and the current implementation of hLRTs is clearly the most widely used model-selection criterion in phylogenetics.

Recently, the adequacy of the hypothesis testing approach for model selection in phylogenetics was criticized on other grounds (e.g., Burnham and Anderson, 1998; Sanderson and Kim, 2000; Grant and Kluge, 2003). The present study of a small sample of empirical datasets shows that in most of these cases, at least some of the hLRTs lead to different models selected as optimal. These models are usually similar, but in some cases they can differ widely in their complexity and therefore affect all inferences, including the estimation of the ML topology. Thus, arbitrary decisions on the starting point and the parameter addition (or removal) sequences can have important consequences for ML analyses.

These results suggest that caution should be used when applying current implementations of hLRT protocols for model selection. The use of iterative cycles of hLRT model selection followed by ML tree searches (Sullivan and Swofford, 1997) might ameliorate the impact of arbitrary choices between alternative hLRT schemes, although in the cases tested here still lead to different results (e.g., data sets 5 and 11). If the current hLRT approach were used, it would be advisable to investigate as thoroughly as possible the impact of the hLRT scheme choice on the phylogenetic relationships of the group under study. In particular, it seems that this problem might be more conspicuous in data sets that are not informative enough to select a single optimal model.

Possible solutions to the problems of current hLRT protocols include posterior comparisons among the selected models or other algorithms developed within the hypothesis-testing approach (e.g., Efroymson, 1960). Alternatively, hLRTs could be abandoned altogether in favor of other model selection criteria (e.g., AIC, BIC; see Posada and Buckley, in press), although their behavior in phylogenetics needs to be further explored.

Acknowledgments

I would like to thank W. Wheeler, P. Goloboff, D. Posada, N. Giannini, J. Faivovich, T. Grant, J. de Laet, J. Clarke, and M. Weksler for useful comments and suggestions on this subject, although they may disagree with what was written here. Additional thanks are extended to J. Sullivan, A. Driskell, J. Kim, and C. Simon for their careful review and editorial suggestions that markedly improved the quality of this manuscript.

References

Abt, K. 1967. On the identification of the significant independent variables in linear models. Metrika 12:1–15.

Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.

Buckley, T. R., C. Simon, and G. K. Chambers. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. 50:67–86.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: A practical information theoretic approach. Springer, New York.

Burk, A., M. Westerman, and M. Springer. 1998. The phylogenetic position of the musky rat-kangaroo and the evolution of bipedal hopping in kangaroos (Macropodidae:Diprotodontia). Syst. Biol. 47:457–474.

Cannatella, D. C., D. M. Hillis, P. T. Chippindale, L. Weight, A. S. Rand, and M. J. Ryan. 1998. Phylogeny of frogs of the *Physalaemus pustulosus* species group with an examination of data incongruence. Syst. Biol. 47:311–335.

Chang, J. T. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math. Biosci. 134:189–215.

Copas, J. B. 1984. Discussion of Dr. Miller's paper. J. R. Stat. Soc. A 147:410–412.

Cox, D. R., and E. J. Snell. 1974. The choice of variables in observational studies. Appl. Stat. 23:51–59.

Cunningham, C. W., H. Zhu, and D. M. Hillis. 1998. Best maximum-likelihood models for phylogenetic inference:Empirical tests with known phylogenies. Evolution 52:978–987.

Draper, N. R., and H. Smith. 1998. Applied regression analysis, 3rd edition. Wiley & Sons Inc., New York.

Edwards, A. W. F. 1992. Likelihood, 2nd edition. Johns Hopkins University Press, Baltimore.

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. Pages 67–76 *in* Phenetic and phylogenetic classiécation (J. McNeill, ed.). Systematics Association Publication, London.

Efroymson, M. A. 1960. Multiple regression analysis. Pages 191–203 *in* Mathematical methods for digital computers (A. Ralston and H. S. Wilf, eds.). Wiley, New York.

Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22:240–249.

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

Felsenstein, J. 2002. Phylip ver. 3.6a3. Software package distributed by the author. Seattle, Washington, USA.

Frati, F., C. Simon, J. Sullivan, and D. L. Swofford. 1997. Gene evolution and phylogeny of the mitochondrial cytochrome oxidase gene in collembola. J. Mol. Evol. 44:145–158.

Gaut, B. S., and P. O. Lewis. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12:152–162.

Giribet, G., and C. Ribera. 1998. The position of arthropods in the animal kingdom: A search for a reliable outgroup for internal arthropod phylogeny. Mol. Phyl. Evol. 9:481–488.

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

Goldman, N., and A. S. Whelan. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol. Biol. Evol. 17:975–978.

Goloboff, P. A. 1999. Analyzing large data sets in reasonable times: Solutions for composite optima. Cladistics 15:415–428.

Goloboff, P. A., J. S. Farris, and K. Nixon. 2000. TNT 1.0. Software and documentation distributed by the authors. San Miguel de Tucumán, Argentina. Available at www.zmuc.dk/public/phylogeny.

Grant, T., and A. G. Kluge. 2003. Data exploration in phylogenetic inference: Scientific, heuristic, or neither. Cladistics 19:379–418.

Halanych, K. M. 1996. Testing hypotheses of chaetognath origin: Long branches revealed by 18S ribosomal DNA. Syst. Biol. 45:223–246.

Hamaker, H. C. 1962. On multiple regression analysis. Stat. Neerlandica 16:31–56.

Hasegawa M., K. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Hedges, S. B., K. D. Moberg, and L. R. Maxson. 1990. Tetrapod phylogeny inferred from 18s and 28s ribosomal RNA sequences and a review of the evidence for amniote relationships. Mol. Biol. Evol. 7:607–633.

Huelsenbeck, J. P., and K. A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. 28:437–466.

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21:1123–1133.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. M. Munro, ed.). Academic Press, New York.

Kelsey, C. R., K. A. Crandall, and A. F. Voevodin. 1999. Different models, different trees: The geographic origin of PTLV-I. Mol. Phylogenet. Evol. 13:336–347.

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

Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA 78:454–458.

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 algorithmsunder equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

Lundrigan, B. L., S. A. Jansa, and P. K. Tucker. 2002. Phylogenetic relationships in the genus *Mus*, based on paternally, maternally, and biparentally inherited characters. Syst. Biol. 51:410–431.

Mantel, N. 1970. Why stepdown procedures in variable selection. Technometrics 12:621–625.

McCracken, K. G., J. Harshman, D. A. McClellan, and A. D. Afton. 1999. Data set incongruence and correlated character evolution:An example of functional convergence in the hind-limbs of stifftail diving ducks. Syst. Biol. 48:683–714.

Miller, A. J. 1984. Selection of subsets of regression variables. J. R. Stat. Soc. A 147:389–410.

Miller, A. J. 1990. Subset selection in regression. Chapman and Hall, New York.

Minin, V., Z. Abido, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674–683.

Morozov, P., T. Sitnikova, G. Churchill, F. J. Ayala, and A. Rzhetsky. 2000. A new method for characterizing replacement rate variation in molecular sequences: Application of the Fourier and Wavelet models to *Drosophila* and mamalian proteins. Genetics 154:381–395.

Muse, S. 1999. Modeling the molecular evolution of HIV sequences. Pages 122–152 *in* The evolution of HIV (K. A. Crandall, ed.). Johns Hopkins University Press, Baltimore.

Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek. 1994. fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. 10:41–48.

Ota, R., P. J. Waddell, M. Hasegawa, H. Shimodaira and H. Kishino. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol. Biol. Evol. 17:798–803.

Posada, D. 2003. Selecting a model of nucleotide substitution. Pages 6.5.1–6.5.14 *in* Current protocols in bioinformatics (A. D. Baxevanis et al., eds.). John Wiley & Sons, Inc., New York.

Posada, D., and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: Advantages of the AIC and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793–808.

Posada, D., and K. A. Crandall. 1998. MODELTEST: Testing the model of DNA substitution. Bioinformatics 14:817–818.

Posada, D., and K. A. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.

Rodríguez, F., J. F. Oliver, A. Marín, and J. R. Medina. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142:485–501.

Sanderson, M. J., and J. Kim. 2000. Parametric phylogenetics? Syst. Biol. 49:817–829.

Soltis, D. E., P. S. Soltis, D. L. Nickrent, L. A. Johnson, W. J. Hahn, S. B. Hoot, J. A. Sweere, R. K. Kuzoff, K. A. Kron, M. W. Chase, S. M. Swensen, E. A. Zimmer, S.-M. Chaw, L. J. Gillespie, W. J. Kress, and K. J. Sytsma. 1997. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. Ann. Miss. Bot. Gard. 84:1–49.

Springer, M. S., H. M. Amrine, A. Burk, and M. J. Stanhope. 1999. Additional support for Afrotheria and Paenungulata, the performance of mitochondrial versus nuclear genes, and the impact of data partitions with heterogeneous base composition. Syst. Biol. 48:65–75.

Stanhope, M. J., V. G. Waddell, O. Madsen, W. de Jong, S. B. Hedges, G. C. Cleven, D. Kao, and M. S. Springer. 1998. Molecular evidence for multiple origins of Insectivora and for a new order of endemic African insectivore mammals. Proc. Natl. Acad. Sci. USA 95:9967–9972.

Steel, M. A., and D. Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2002. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.

Sullivan, J., K. E. Holsinger, and C. Simon. 1996. The effect of topology on estimates of among-site rate variation. J. Mol. Evol. 42:308–312.

Sullivan, J., and D. L. Swofford. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mammal. Evol. 4:77–86.

Sullivan, J., D. L. Swofford, and G. J. P. Naylor. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum likelihood models. Mol. Biol. Evol. 16:1347–1356.

Swofford, D. L. 2002. PAUP* 4.0 vers. b10. Phylogenetic analysis using parsimony and other methods. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis. 1996. Phylogeny reconstruction. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C.Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

Tamura, K. 1994. Model selection in the estimation of the number of nucleotide substitutions. Mol. Biol. Evol. 11:154–157.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

Tateno, Y., N. Takezaki, and M. Nei. 1994. Relative efficiencies of the maximum likelihood, neighbor-joining, and maximum parsimony methods when the substitution rate varies with site. Mol. Biol. Evol. 11:261–277.

Wayne, R. K., E. Geffen, D. J. Girman, K. P. Koepfli, L. M. Lau, and C. R. Marshall. 1997. Molecular systematics of the Canidae. Syst. Biol. 46:622–653.

Wilgenbusch, J., and K. de Queiroz. 2000. Phylogenetic relationships among the phrynosomatid sand lizards inferred from mitochondrial DNA sequences generated by heterogeneous evolutionary processes. Syst. Biol. 49:592–612.

Wojciechowski, M. F., M. J. Sanderson, and J.-M. Hu. 1999. Evidence on the monophyly of Astragalus (Fabaceae) and its major subgroups based on nuclear ribosomal dna its and chloroplast dna trnl intron data. Syst. Bot. 24:409–437.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Yang, Z., N. Goldman, and A. Friday. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Syst. Biol. 44:384–399.

Zhang, J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. Mol. Biol. Evol. 16:868–875.

Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39:315–329.

First submitted 10 November 2003; reviews returned 3 April 2004;
final acceptance 13 August 2004
Associate Editor: Junhyong Kim