Theoretical Basis of Likelihood Methods in Molecular Phylogenetic Inference

Rhiju Das, Centre of Mathematics and Physical Sciences applied to Life science and EXperimental biology (CoMPLEX), University College London.

Supervisors: Z. Yang and J. Mallet, Galton Laboratory, Department of Biology University College London.

Dissertation submitted for completion of the degree of Master of Research in University College London.

4 September, 2000

This dissertation does not exceed 18,000 words in length, exclusive of tables, figure captions, appendices, and references.

Abstract

Phylogenetic inference for molecular data by the maximum-likelihood approach has been attacked from a theoretical point of view, because the likelihood functions take different forms for different trees, so that optimised likelihood values for different trees do not appear to be directly comparable (Nei, 1987). Here, a new "super-tree" perspective is introduced to refute these criticisms. A super-tree likelihood expression is constructed which is a function of all possible bipartition lengths; it reduces to the individual tree likelihood functions when bipartitions not in a given tree are set to zero. From this perspective, the problem of phylogeny inference is seen to be a classical statistical problem involving selection between composite hypotheses. In particular, the usual ML procedure is well-justified, and, moreover, the likelihood ratio between two trees does indeed indicate the posterior odds of the trees. This "literal" interpretation of the likelihood values is shown by simulation to provide a more intuitive indication of tree selection accuracy than the "integrated" likelihood posterior probabilities of Rannala and Yang (1996) and bootstrap supports. Thus, the likelihood framework for phylogenetic inference for molecular phylogenetic inference has a good theoretical basis – provided that an adequately realistic model of molecular mutation is used to fit the data. To test the adequacy of such molecular mutation models, a set of straightforward "consistency checks", based on likelihood ratio statistics, are also presented. Predicted distributions of these statistics are shown to agree with simulation. These consistency checks, as well as a likelihood-based tree selection procedure, have been applied to several data sets: mtDNA from five primates, α and β globin genes from five mammals, mtDNA and wingless genes from sixty *Heliconiini* butterflies, and mtDNA from forty mimicking races of *Heliconius melpomene* and *Heliconius erato* butterflies. These consistency checks, as well as the presence of internal contradictions, reject the commonly used HKY85+Γ model when applied to many of these data sets. It is concluded that while maximum likelihood is rigorous in principle (and preferable to theoretically unjustified methods like maximum parsimony), it should be considered a heuristic procedure for phylogenetic inference until the complex biological processes influencing molecular mutation are fully understood.

Theoretical Basis of Likelihood Methods in Molecular Phylogenetic Inference

Rhiju Das, Centre of Mathematics and Physical Sciences applied to Life science and EXperimental biology (CoMPLEX), University College London.

1. Introduction

The inference of phylogenies based on DNA or amino acid sequences of living species has been one of the most powerful techniques of modern genetics – and also one of the most controversial. Molecular data has been used to find man's place among the primates (Schwartz, 1984; Hasegawa, 1991), the relationship of mammals to birds and dinosaurs (Hedges *et al.*, 1990; Hedges, 1994), and the primordial branching pattern of the earliest living organisms (Olsen *et al.*, 1994b). Yet, for each problem, different researchers have published conflicting trees – sometimes based on the same data set!

So far, there has been no rigorous statistical framework to guide phylogeny inference. Instead, researchers depend mostly on heuristic methods, like maximum parsimony analysis (Farris *et al.*, 1970) or the minimum evolution distance method [and its cousins, least-squares distance fitting, and neighbour-joining; see (Cavalli-Sforza and Edwards, 1967), (Fitch and Margoliash, 1967), and (Rzhetsky and Nei, 1993)]. A more rigorous technique to have gained wide popularity is the maximum likelihood (ML) method, introduced in a computationally tractable form by Felsenstein (1973, 1981). For each possible tree hypothesis, the likelihood of a molecular data set can be computed, using a Markov process to model the changes among possible molecular states. The lengths of the tree branches and the parameters of the evolutionary model (like the transition / transversion bias for DNA) are then varied until this likelihood function is maximised, producing a single likelihood value, $L_{\text{tree}} = \exp(l_{\text{tree}})$, for each tree. Finally, the tree with the highest log-likelihood l_{tree} is picked as the phylogeny hypothesis best supported by the data. Often, the data is re-sampled with replacement, or "bootstrapped" (Felsenstein, 1985), to mimic statistical variation; if the ML method finds the same tree for, say, 95% of the bootstrap data replicates, the tree is considered well-supported.

With its likelihood / probability values and explicit description of the molecular evolutionary model, the ML method appears statistically rigorous, at first glance. However applications of the ML method to real data sets using simple evolutionary models of molecular mutations have often led to bizarre results. One example (among many) is the analysis by Zardoya *et al.* (1998) of the whole mitochondrial genome of the lungfish, the coelecanth, and several tetrapods to determine the closest living ancestor to the tetrapods. In their investigation, each mitochondrial gene yielded an ML tree topology (with, e.g., lungfish+tetrapods, or coelecanth+tetrapods, as sister groups) with a likelihood several orders of magnitude (often 10^5 or more) greater than alternative trees – but different proteins favoured contradicting tree topologies.

In addition to this inconsistency in practical applications, there appear to be some theoretical problems with the ML procedure for molecular phylogenetic inference. As pointed out by Nei (1987), the classical likelihood theory of parameter estimation does not seem to directly apply to ML tree inference. In particular, the likelihood function has a different form for different trees – and the tree is not a continuous parameter (Yang *et al.*, 1995). So it is not readily obvious that likelihoods for different trees can be properly compared to pick out the "best" tree. Without the apparent backing of classical likelihood theory, the ML method for phylogeny inference instead finds its support mostly from computer simulation studies and from the theoretical demonstration of its consistency (Yang, 1994b), i.e., its ability to pick out the correct tree in the limit of infinitely long molecular sequences. Several theoretical questions remain open: How can one be sure that the assumed evolutionary model is properly describing all relevant aspects of the data set, including purifying / positive selection and recombination? Is the ratio of two tree likelihoods $L_1 / L_2 = \exp(\Delta l_{12})$ an appropriate measure of the ratio of posterior

probabilities? Or should one take into account the variance $\sigma(\Delta l_{12})$ in the log-likelihood difference (Kishino and Hasegawa, 1989), looking at, say, $\exp[\Delta l_{12}/\sigma(\Delta l_{12})]$ (Jermiin *et al.*, 1997) – or should one trust the ratio of the trees' bootstrap supports?

The main objective of this report is to address these uncertainties lying at the heart of the ML method for molecular phylogenetic inference. In particular, it will be argued that ML is indeed statistically sound for accurate models of molecular mutation. To support this claim, a new "super-tree" likelihood function is introduced which is a function of all possible bipartitions[1] of the taxa. It is designed so that when all internal bipartitions except for the subset found in a given tree are constrained to zero, the "super-tree" likelihood reduces to the form of the usual likelihood function defined for that tree.[2] From the super-tree perspective, the problem of phylogeny inference is then seen to be a classical statistical problem involving composite hypotheses, refuting the theoretical doubts expressed in (Yang *et al.*, 1995) and elsewhere. In particular, a likelihood ratio between different tree topologies can indeed be interpreted literally as an estimate of the trees' posterior odds, "the ratio of the frequencies with which, in the long run, the two hypotheses will deliver the observed data" [section 3.4 of Edwards(1972)] – as long as an adequately realistic model of the molecular mutation process is used.

This last disclaimer is very important, however. The secondary objective of this report is to introduce some simple tests ("consistency checks") which can check if the evolutionary model assumed in the ML analysis is realistic enough to describe a given data set. In fact, it will be found that these tests quite often reject the most general evolutionary models implemented in the current generation of phylogeny inference programs. Therefore, while the ML method is, in theory, statistically well-founded, at present it is best considered a heuristic method in practice, since all the biological subtleties of molecular mutation are not yet completely understood.

This report is divided into seven main sections, including this Introduction; mathematical details are collected in the appendices. In the next section, a well-defined super-tree likelihood function is introduced explicitly for the simplest model of binary characters, illustrated with several four-taxon examples, and then extended to the evolutionary models most commonly used in the literature. The third section builds on the intuition obtained from the super-tree perspective to describe how ML can be used to choose among tree hypotheses; in particular, likelihood ratios between trees are shown, by simulation, to be a better indicator of the accuracy of phylogeny inference than bootstrap supports. The fourth section develops some straightforward likelihood ratio tests to check the adequacy of the evolutionary model in describing a data set; several novel predictions regarding the distributions of these likelihood ratio statistics are checked against simulations. The fifth section applies the ML procedure and consistency checks to several real data sets: a segment of mitochondrial DNA (mtDNA) from five primates; the α and β globin genes from five mammals; mtDNA and a nuclear gene for a sixty-taxon data set of South American passion-vine (*Heliconiini*) butterflies; and mtDNA from forty mimicking races of *Heliconius melpomene* and *Heliconius erato* butterflies. The properties of the ML method, in the light of these theoretical results, simulations, and analyses of real data, are discussed in terms of self-consistency and statistical evaluation, and compared with other heuristic techniques of phylogeny estimation in the sixth section. The seventh section concludes the report, comparing what has been achieved to what was proposed before starting the project.

---

[1] In this report, the terms "bipartition" and tree "branch" are used interchangeably. The former term will be used more often when discussing abstract extensions to usual tree structures.

[2] Strimmer and Moulton (2000) have recently published another method for generalising the likelihood function for a phylogeny to a more general "phylogenetic network". However their approach introduces several new, unspecified parameters and does not produce a unique likelihood function. Unlike the work presented here, the method of Strimmer and Moulton does not provide a theoretical basis for the usual likelihood methods, as will be discussed in Section 6.

2. The super-tree likelihood function

This section introduces a perspective where different tree topologies are seen to be special cases of a more general "super-tree" problem with a single likelihood function. The first three subsections, describing likelihood basics and the trivial two-taxon and three-taxon problems for the simplest binary model of molecular mutation, say nothing particularly new and are intended mainly to establish notation. Subsection 2.4 introduces a super-tree likelihood function for the non-trivial four-taxon case, and briefly discusses how this provides a theoretical justification of the usual ML procedure of phylogeny inference. The last two subsections sketch how the super-tree perspective can be extended to data sets with more taxa and with more general evolutionary models.

2.1. Likelihood basics.

First, the basic formalism is described. Given a data set of m taxa aligned molecular sequences with n sites each, a general likelihood function takes the binomial form:

$$L = \frac{n!}{n_0!\, n_1! \cdots n_N!} p_0^{n_0} p_1^{n_1} \cdots p_N^{n_N} \ , \qquad (1)$$

where N is the number of possible "site patterns", the n_0, n_1, ... n_N are the observed numbers of the site patterns in the data, and p_0, p_1, ... p_N, are the probabilities of each pattern under the given evolutionary model and tree topology. Explicitly, a site pattern is defined as a set of molecular characters that exist at a given site in the taxa. So, for example, n_0 might be the number of sites that are adenine for all the taxa in a DNA data set; n_1 might be the number of sites that have thymine in the first taxon, but adenine in the rest; etc. If there are c possible character states, there are thus $N = c^m$ possible site patterns. Note that there is a constraint on the observed and the predicted probabilities, that the frequencies sum to one: $\sum n_i/n = \sum p_i = 1$.

For a given data set, the n_i are constants, and the p_i are varied until the likelihood is maximised to best fit the $N-1$ degrees of freedom. If the evolutionary model has enough independent parameters (at least $N-1$ of them) one can hope to attain a "perfect" fit $p_i = n_i/n$, which is a global maximum.[3]

Since L is often a very small number, it is convenient to deal with logarithm of the likelihood

$$l = \log L = \sum_{i=1}^{N} n_i \log p_i \ , \qquad (2)$$

with the constant term $\log n! - \sum \log[n_i!]$ suppressed. A perfect fit gives the value $l_{\max} = \sum n_i \log[n_i/n]$.

The remainder of this section describes the super-tree likelihood for the simplest evolutionary model, for binary characters with equal frequencies (Neyman, 1971). This model might be appropriate for, e.g., DNA sequences where only pyrimidines and purines are distinguished. Extensions to more realistic models of DNA and amino acid evolution are given in the last two sub-sections.

---

[3] To see this, it may help the reader (especially if he/she comes from a physics background like the author) to note that

$$L \propto \prod_{i=1}^{N} (np_i)^{n_i} \frac{e^{-np_i}}{n_i!} \ .$$

The expression is simply the product of probabilities for an integer n_i to be picked in a Poisson distribution with expectation np_i, and clearly has its only local maximum with respect to independent variation of the predicted frequencies p_i at $n_i = np_i$. Z. Yang (priv. comm.) has pointed out, however, that some readers would consider it obvious that a perfect fit is a global maximum of the original binomial form of L, and that this rearrangement into Poisson form is a distraction. To each his own!

2.2. The simplest binary model, two-taxon case.

Writing the probabilities of nucleotides to be in either of the two states, 0 or 1, as a column vector $\mathbf{P} = [p_0, p_1]^T$, a simple binary evolutionary model is described by $d\mathbf{P}/dt = Q\,\mathbf{P}$, where the instantaneous rate matrix is

$$Q = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{3}$$

The solution at any time t is

$$\{p_{ij}\} = \exp(Qt) = \begin{bmatrix} p_{\text{same}} & p_{\text{diff}} \\ p_{\text{diff}} & p_{\text{same}} \end{bmatrix}, \tag{4}$$

where the probability that two nucleotides are different after separation time t is

$$p_{\text{diff}} = \frac{1}{2}[1 - e^{-2t}], \tag{5}$$

and $p_{\text{same}} = 1 - p_{\text{diff}}$. The equilibrium character frequencies are $\pi_0 = \pi_1 = 1/2$. Now a data set of two taxa yields four site pattern numbers n_{00}, n_{01}, n_{10}, and n_{11}. Since the variables must sum to n, there are three degrees of freedom. The log-likelihood function is

$$\begin{aligned} l &= n_{00} \log\left[\frac{p_{\text{same}}}{2}\right] + n_{01} \log\left[\frac{p_{\text{diff}}}{2}\right] + n_{10} \log\left[\frac{p_{\text{diff}}}{2}\right] + n_{11} \log\left[\frac{p_{\text{same}}}{2}\right] \\ &= n_{\text{same}} \log p_{\text{same}} + n_{\text{diff}} \log p_{\text{diff}} - n \log 2 \end{aligned} \tag{6}$$

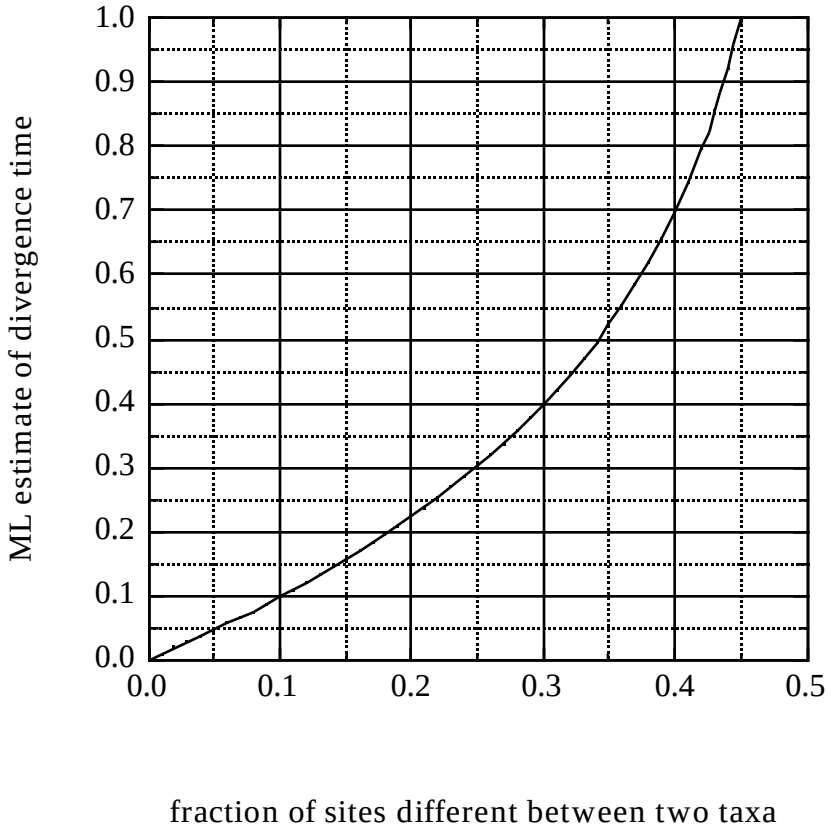


Figure 1. Maximum likelihood estimate of the divergence time between two taxa based on the simplest binary model – see equation (7).

So the assumption of equal character frequencies allows one to "collapse" together site patterns which are inverses of each other, $n_{\text{diff}} = n_{01} + n_{10}$ and $n_{\text{same}} = n_{00}+n_{11}$. Since one has the constraint $n_{\text{same}} + n_{\text{diff}} = n$, there is actually only one collapsed degree of freedom left. Thus, with a single parameter, the separation time t, one can reach the global maximum where

$$t^{(\max l)} = -\frac{1}{2}\log\left[1 - 2\,\boldsymbol{f}_{\text{diff}}\right] \tag{7}$$

with $\boldsymbol{f}_{\text{diff}} = n_{\text{diff}}/n$. See Figure 1 for a plot of this function. The equation above shows that, in this simple model, the ML estimate of the divergence time between two taxa is only dependent on $\boldsymbol{f}_{\text{diff}}$, the number of character states that are different between the two taxa.

Note that while equation (7) gives a perfect fit of the collapsed site pattern frequencies (one degree of freedom), it is not a perfect fit of the three original degrees of freedom. To accomplish that, one might consider a more general evolutionary model, with, say, fittable equilibrium frequency $\pi_0^{\text{eq}} \neq 1/2$ and fittable root frequency $\pi_0^{\text{root}} \neq 1/2$. Also, note that expression (7) is not well-defined if $\boldsymbol{f}_{\text{diff}} > 1/2$. If one finds such a data set, one must set $t^{(\max l)}$ at the boundary $t^{(\max l)} \to \infty$; or, more palatably perhaps, one might be able to obtain a non-boundary fit by using a more general evolutionary model.

2.3. Three-taxon case.

The three-taxon case is almost as straight-forward to solve as the two-taxon case. Suppose one has three taxa A, B, and C, for which there is the single possible connecting tree shown in Figure 2.[4] One needs to find the three optimum branch lengths t_A, t_B, and t_C.

For the binary model, there are eight site patterns (or 7 degrees of freedom), and as before, one can collapse them into $n_{\text{same}} = n_{000} + n_{111}$; $n_{A|BC} = n_{100} + n_{011}$; $n_{B|AC} = n_{010} + n_{101}$; and $n_{C|AB} = n_{001} + n_{110}$.

The likelihood function is then:

$$l = n_{\text{same}} \log p_{\text{same}} + n_{A|BC} \log p_{A|BC} + n_{B|AC} \log p_{B|AC} + n_{C|AB} \log p_{C|AB} - n \log 2 \tag{8}$$

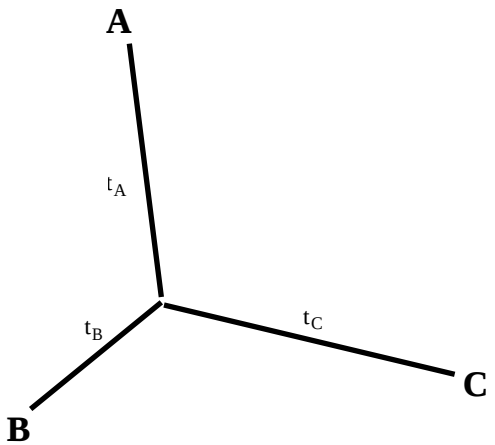


Figure 2. The only unrooted three-taxon tree.

[4] In this and the following two subsections, one does not consider rooted trees since, with a reversible evolutionary Markov model and without a hypothesis like the molecular clock, ML cannot in general find the root of a tree – the pulley principle of Felsenstein (1981). See (Yang, 2000) for the solution of the tree-taxon rooted case under a molecular clock assumption.

with the pattern probabilities:

$$
\begin{aligned}
p_{\text{same}} &= \frac{1}{4}\left[1 + e^{-2(t_A+t_B)} + e^{-2(t_A+t_C)} + e^{-2(t_B+t_C)}\right] \\
p_{A|BC} &= \frac{1}{4}\left[1 - e^{-2(t_A+t_B)} - e^{-2(t_A+t_C)} + e^{-2(t_B+t_C)}\right] \\
p_{B|AC} &= \frac{1}{4}\left[1 - e^{-2(t_A+t_B)} + e^{-2(t_A+t_C)} - e^{-2(t_B+t_C)}\right] \\
p_{C|AB} &= \frac{1}{4}\left[1 + e^{-2(t_A+t_B)} - e^{-2(t_A+t_C)} - e^{-2(t_B+t_C)}\right]
\end{aligned}
\qquad (9)
$$

The above expression can be derived using the standard sum over internal character states [see, e.g., (Felsenstein, 1981)],

$$
p_{\text{same}} = \sum_{i=\{0,1\}} \pi_i \sum_{j=\{0,1\}} p_{ij}(t_A) p_{ij}(t_B) p_{ij}(t_C), \text{ etc.,} \qquad (10)
$$

or using a sum over "pathsets" as described in Appendix A. Note that the expressions in (9) are sums over exponentials; this will be a repeating theme in following subsections.

There are three collapsed degrees of freedom, and three branch parameters, so one can again obtain a perfect fit of the collapsed site frequencies. Explicitly,

$$
t_A^{(\max l)} = \frac{1}{4}\left\{-\log\left[1 - 2\left(f_{A|BC} + f_{B|AC}\right)\right] - \log\left[1 - 2\left(f_{A|BC} + f_{C|AB}\right)\right] + \log\left[1 - 2\left(f_{B|AC} + f_{C|AB}\right)\right]\right\}, \qquad (11)
$$

where $f_{A|BC} = n_{A|BC}/n$, etc. Expressions for $t_B^{(\max l)}$ and $t_C^{(\max l)}$ can be obtained from (11) by symmetry.

As with the two-taxon case, some disclaimers apply. There may be a problem in obtaining the global maximum if $f_{A|BC}+f_{B|AC}$, $f_{A|BC}+f_{C|AB}$, or $f_{B|AC}+f_{C|AB}$ is greater than $1/2$, in which case the above expression is not well-defined. To find a non-boundary solution – or to obtain a perfect fit of all seven un-collapsed degrees of freedom – one might consider a more general evolutionary model, say with a root frequency π_0^{root} different from $1/2$ at a chosen root taxon, plus three different values of $\pi_0^{eq}(t_A)$, $\pi_0^{eq}(t_B)$, and $\pi_0^{eq}(t_C)$, to describe evolution to different equilibrium frequencies on each branch.

2.4. Four-taxon case.

Now, consider the four-taxon case. There are three possible trees (see Figure 3) and the non-trivial phylogenetic inference problem is to choose between them. There are eight (collapsed) site patterns that might show up in the data and they can be labelled n_{same}, $n_{A|BCD}$, $n_{B|ACD}$, $n_{C|ABD}$, $n_{D|ABC}$, $n_{AB|CD}$, $n_{AC|BD}$, and $n_{AD|BC}$. Explicitly, $n_{A|BCD}$ is the number of sites where taxon A has a different character than the other three taxa; the other site pattern numbers are similarly defined. Label site pattern frequencies as before: $f_{A|BCD} = n_{A|BCD}/n$, etc. In a loose sense, one might consider the observed frequencies of site patterns ($n_{A|BCD}$, $n_{B|ACD}$, $n_{C|ABD}$, $n_{D|ABC}$, $n_{AB|CD}$, $n_{AC|BD}$, $n_{AD|BC}$) as raw approximations to the branch lengths (t_A, t_B, t_C, t_D, t_I, t_{II}, t_{III}).

In the usual ML procedure, the likelihood function is different depending on the assumed tree topology. Suppose one could define a "super-tree" likelihood function, which is dependent on all possible internal bipartition lengths, with the following property: if t_{II} and t_{III} are constrained to zero, the super-tree likelihood function takes the form of the usual likelihood function for tree I. Similarly, setting $t_I = t_{III} = 0$ or $t_I = t_{II} = 0$ yields the likelihood functions for trees II or III, respectively. With such a super-tree likelihood function, the three tree hypotheses in Figure 3 correspond to particular parameter configurations (five-dimensional hyper-planes) within a seven-dimensional super-tree space with a single likelihood function.

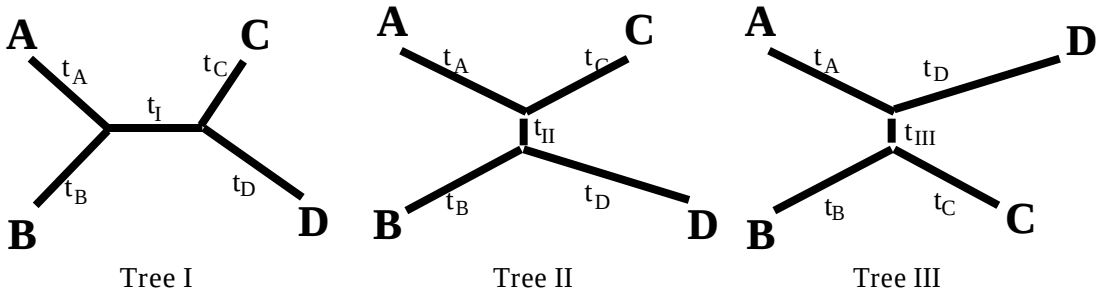


Figure 3. The three unrooted four-taxon tree topologies.

Before presenting the explicit form of a super-tree likelihood function, it is worth clarifying this perspective with a cartoon; see Figure 4. The contours of a putative super-tree likelihood function are plotted for a hypothetical data set; to make the diagram two-dimensional, t_{III} is assumed to vanish (i.e., tree III is ignored as a possible topology), and at each point in (t_I, t_{II}) space, the likelihood has been optimised with respect to external branch lengths (t_A, t_B, t_C, t_D). The problem is to decide between tree topologies I and II; biologically realistic hypotheses correspond only to points on the positive t_I and t_{II} semi-axes. The usual ML procedure maximises the likelihood along each of these semi-axes separately, producing parameter estimates marked "tree I" and "tree II" on Figure 4. The theoretical uncertainty of the usual ML procedure lies in the fact that the likelihood functions maximised for tree topologies I and II appear to be different, being functions of different sets of parameters. But having a picture like Figure 4 illustrates that the likelihood functions for different trees are indeed related – they are special cases of a single super-tree likelihood. Thus, likelihood ratios like L_I/L_{II} are indeed directly interpretable as posterior odds of the two trees, given the data and the assumed evolutionary model.

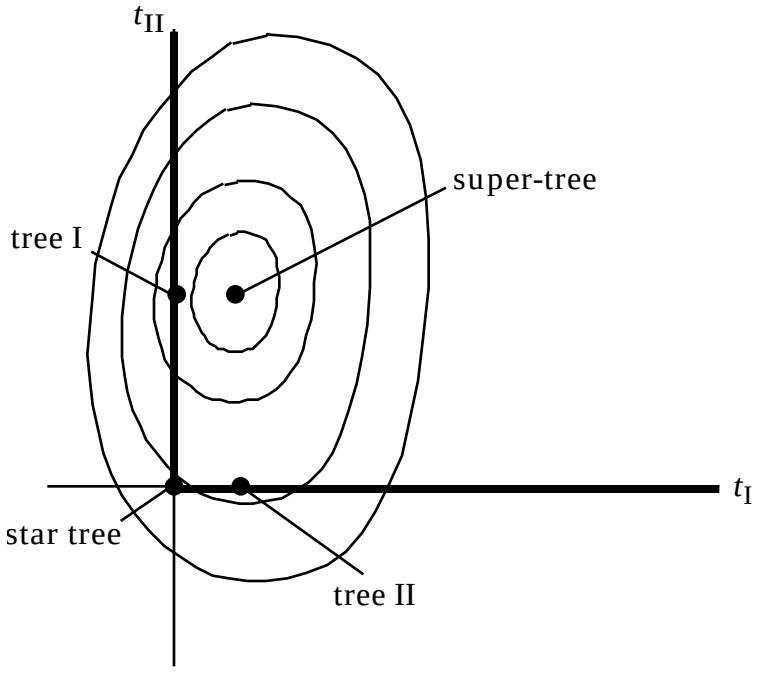


Figure 4. "Cartoon" of likelihood contours in the super-tree perspective. Points on the positive t_I semi-axis and positive t_{II} semi-axis (dark lines) correspond to biologically realistic parameter sets with tree topologies I and II (see Figure 3), respectively. The super-tree likelihood function at those points reduces to the usual ML functions for those tree topologies.

Further discussion of comparisons of L_{I} and L_{II} to choose the tree topology is given in Section 3. Another insight from the super-tree picture is that when the likelihood is maximised overall (t_{I}, t_{II}), giving the parameters marked "super-tree" in the Figure 4, this global super-tree likelihood maximum should not be too much better supported than a biologically realistic parameter set (i.e., $L_{\text{super-tree}}/L_{\mathrm{I}}$ or $L_{\text{super-tree}}/L_{\mathrm{II}}$ should be small). This observation is the basis of a set of "consistency checks" that test the adequacy of an assumed evolutionary model in describing the data, which is further discussed in Section 4.

The super-tree likelihood function must take the same form as in the previous sub-sections:

$$l = \sum_{i=\{\text{site patterns}\}} n_i \log p_i - n \log 2 \tag{12}$$

So the challenge is to come up with a general form of the predicted site pattern probabilities as a function of all possible bipartition lengths (t_{A}, t_{B}, t_{C}, t_{D}, t_{I}, t_{II}, t_{III}). In fact, this is fairly trivial. One starts by writing the site pattern probabilities for the individual tree topologies, which turn out to be linear combinations of a few exponentials [like equation (9)]; see Appendix A. The following form (given in matrix form, for brevity) for the generalised site pattern probabilities, designed to reduce properly to the expressions for the three separate tree topologies, is then more or less obvious:

$$\begin{bmatrix} p_{\text{same}} \\ p_{\mathrm{A|BCD}} \\ p_{\mathrm{B|ACD}} \\ p_{\mathrm{C|ABD}} \\ p_{\mathrm{D|ABC}} \\ p_{\mathrm{AB|CD}} \\ p_{\mathrm{AC|BD}} \\ p_{\mathrm{AD|BD}} \end{bmatrix} = \frac{1}{8} \begin{bmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & -1 & -1 & -1 & +1 & +1 & +1 & -1 \\ +1 & -1 & +1 & +1 & -1 & -1 & +1 & -1 \\ +1 & +1 & -1 & +1 & -1 & -1 & -1 & -1 \\ +1 & +1 & +1 & -1 & +1 & -1 & -1 & -1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \end{bmatrix} \begin{bmatrix} 1 \\ e^{-2(t_{\mathrm{A}}+t_{\mathrm{II}}+t_{\mathrm{III}}+t_{\mathrm{B}})} \\ e^{-2(t_{\mathrm{A}}+t_{\mathrm{I}}+t_{\mathrm{III}}+t_{\mathrm{C}})} \\ e^{-2(t_{\mathrm{A}}+t_{\mathrm{I}}+t_{\mathrm{II}}+t_{\mathrm{D}})} \\ e^{-2(t_{\mathrm{B}}+t_{\mathrm{I}}+t_{\mathrm{II}}+t_{\mathrm{C}})} \\ e^{-2(t_{\mathrm{B}}+t_{\mathrm{I}}+t_{\mathrm{III}}+t_{\mathrm{D}})} \\ e^{-2(t_{\mathrm{C}}+t_{\mathrm{II}}+t_{\mathrm{III}}+t_{\mathrm{D}})} \\ e^{-2(t_{\mathrm{A}}+t_{\mathrm{B}}+t_{\mathrm{C}}+t_{\mathrm{D}})} \end{bmatrix} \tag{13}$$

See Appendix A for a more detailed discussion of the derivation of the above result.[5]

Note that there are seven degrees of freedom, and seven independent parameters in (12) and (13). The global maximum likelihood $l_{\max}$ can thus be obtained when $n_i = np_i$ (barring parameters hitting a boundary; see previous sub-sections).

But do equations (12) and (13) define the *unique* extension of L into a function of t_{I}, t_{II}, and t_{III} to reduce to the usual tree likelihood functions when all bipartitions not in a given tree are set to zero? Clearly not – one can see that special terms can be added to (13) that disrupt the likelihood terrain everywhere except on the hyper-planes corresponding to the three tree hypotheses. For the simplest binary model, however, the form (13) for the site pattern probabilities is unique in having several desirable mathematical properties [including those which make is amenable to analysis by Hadamard conjugation (Steel *et al.*, 1998)]. In particular, the parameters $(t_{\mathrm{A}}, t_{\mathrm{B}}, t_{\mathrm{C}}, t_{\mathrm{D}}, t_{\mathrm{I}}, t_{\mathrm{II}}, t_{\mathrm{III}})^{(\max l)}$ display a nice additivity property. Explicitly, if one estimates the divergence time t_{AB} between two taxa A and B based only on the fraction of sites which are different in A and B – see equation (7).

[5] Interestingly, after this above result (and generalisations discussed below) was formulated, the author discovered that the mathematical expression had in fact been derived independently in the work on Hadamard conjugations of Hendy, Penny, Steel, Waddell, and collaborators, using a different mathematical approach (Steel *et al.*, 1998). However, those authors have focused on heuristic (parsimony and tree-fitting) analyses of Hadamard conjugated data patterns, outside the likelihood framework. They also have not been able to extend their Hadamard conjugation formulas to the more general evolutionary models with unequal character frequencies used in ML analyses. The disadvantages of these "spectral analysis" methods in comparison to the ML methods of this report, will be further discussed in Section 6.4.

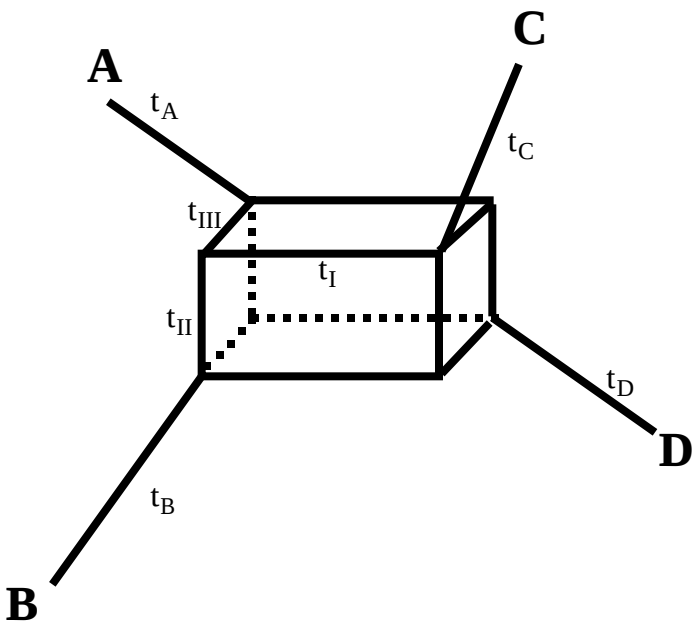


Figure 5. Geometrical interpretation of the four-taxon super-tree structure.

it is easy to show that:

$$t_{AB} = t_A^{(\max l)} + t_B^{(\max l)} + t_{II}^{(\max l)} + t_{III}^{(\max l)}, \qquad (14)$$

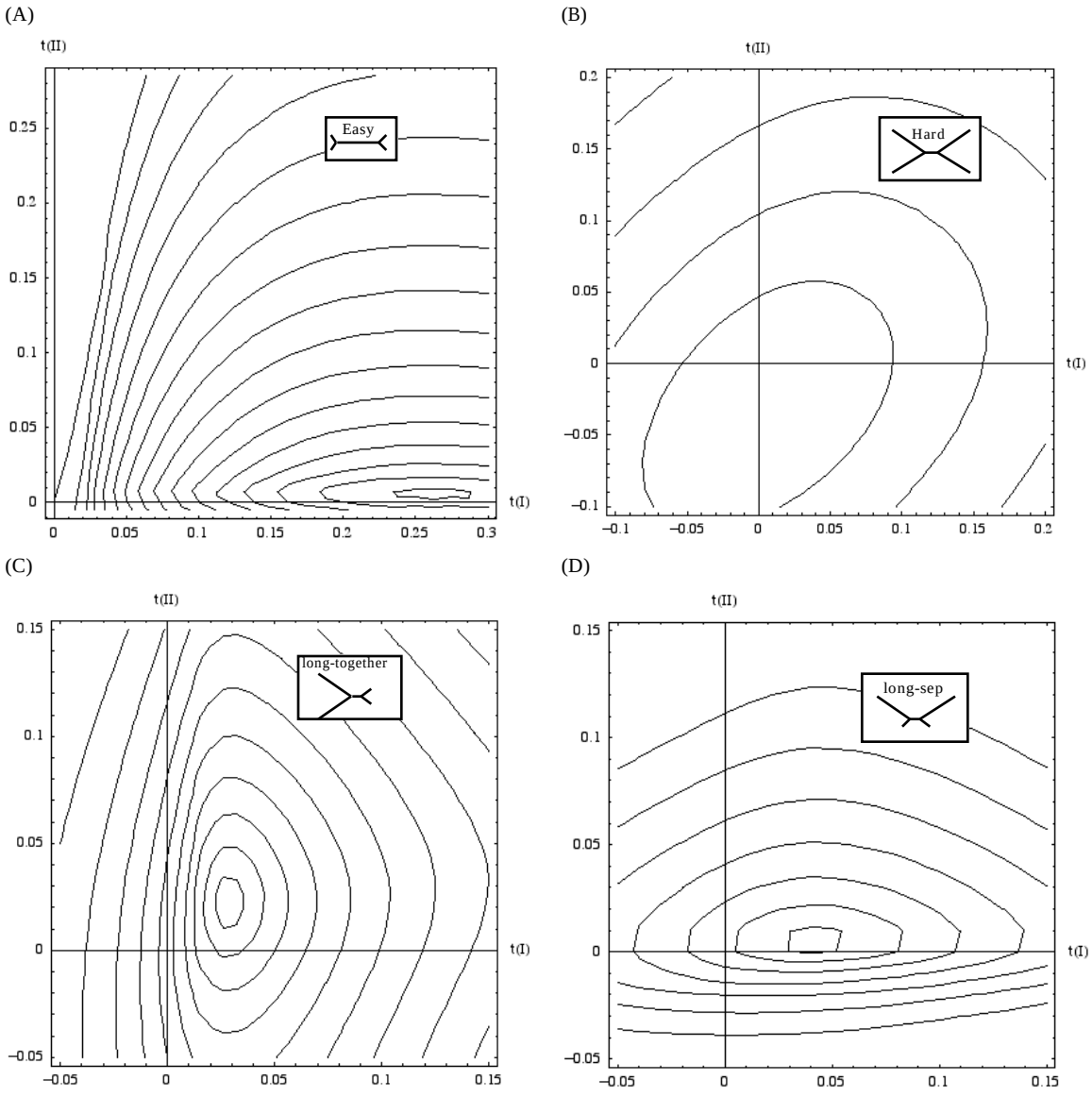
that is, the distance between the two taxa is simply the sum of the maximum likelihood lengths of all bipartitions which might separate the two taxa in a tree. Based on this additivity, one might attempt to visualise a geometrical structure corresponding to the general super-tree as in Figure 5, a parallelepiped of the internal bipartitions stuck with four external branches.

The reader may protest that the super-tree likelihood of (12) and (13) is essentially a mathematical artifice, since points with non-zero t_I and t_{II} do not have a clear biological interpretation. However, the main objective here is not to find such artificial solutions, but to show that the likelihoods of the biologically realistic hypotheses (the positive t_I and t_{II} semi-axes in Figure 4) are directly comparable, in that they are special cases of a single likelihood function; and (12) and (13) accomplish this generalisation.

Before discussing how to obtain super-tree likelihood functions for more taxa and for more complicated evolutionary models, it is worth getting a better intuitive understanding of the super-tree space for this simplest four-taxon problem. The super-tree likelihood for six sample data sets has been investigated. Plotted in Figures 6A-F are the likelihood contours in the t_I–t_{II} plane [to create these plots, t_{III} is constrained to zero; external branch lengths are varied by a simplex algorithm (Numerical Recipes, 1992) at each (t_I, t_{II}) to optimise the likelihood, as in Figure 6].

The first four data sets correspond to 100 character sequences simulated using *evolver* from the PAML package (Yang, 1997) based on the model trees of topology I. The first is an "easy" tree (long internal branch, short external branches); the second is a "hard" tree (vice versa); the third has two long branches separated by the internal branch; and the fourth has the two long branches together. The fifth data set is a mixture, where half the sites correspond to an easy tree with topology I, and half correspond to an easy tree with topology II – mimicking the effects of a recombination event. The sixth data set has the same simulation parameters as the hard tree (second data set), but three-quarters of the sites are forced to be invariant, so that the effects of site-rate variation can be investigated. See caption to Figure 6 for exact simulation parameters. In each figure the plotted log-likelihood contours correspond to $2(l - l_{max}) = 1, 4, 9, \ldots \chi^2$, and are therefore directly comparable between figures.

Figure 6. Contours of the super-tree likelihood function, for 100-binary-character data sets simulated with various combinations of long (t=0.25) and short (t=0.02) branches, with the tree topology I. Here, for each point t_I (the x-axis) and t_{II} (the y-axis), the other bipartition length t_{III} is constrained to zero, and the likelihood is optimised over the four external branch lengths by a simplex algorithm. (A) The easy tree (short,short)-long-(short,short). (B) The hard tree (long,long)-short-(long,long). (C) The tree (long,long)-short-(short,short). (D) The tree (long,short)-short-(long,short). (E) The data is an equal mixture of tree topologies I and II with the easy branch lengths. (F) Same as (B), but three quarters the sites are forced to be invariant. the plotted log-likelihood contours correspond to $2(l - l_{max}) = 1, 4, 9, \ldots j^2$; the outer-most likelihood contours are not plotted for some of the figures, due to irregularities in Mathematica's plotting algorithm.



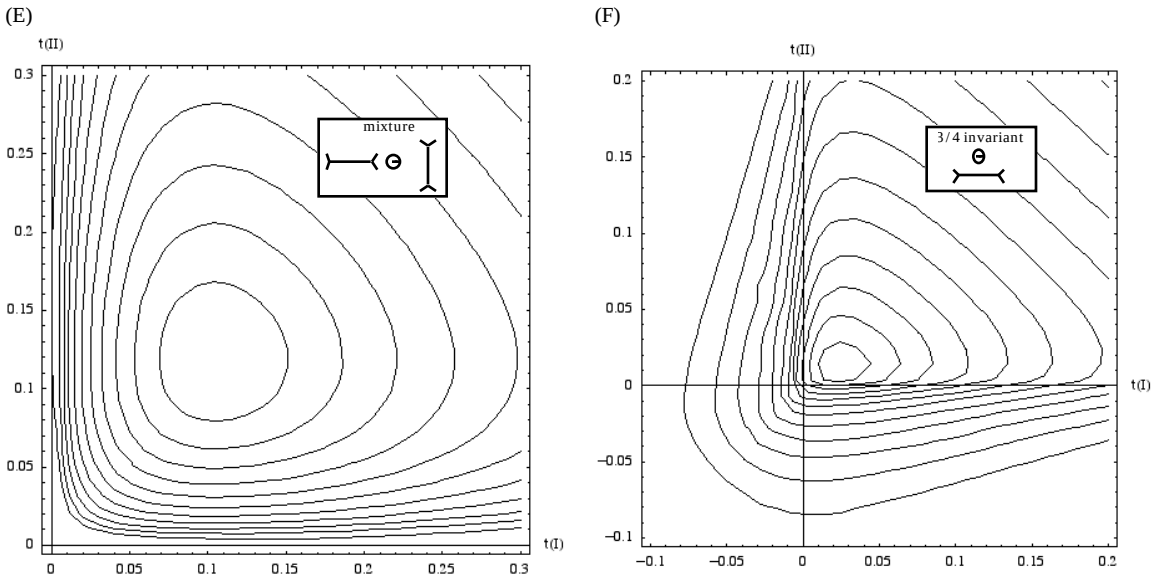


Figure 6, continued from previous page.

To begin, note that the super-tree likelihood function in each plot is quite regular, with only one local maximum, as expected. However, when constrained to the positive t_I and t_{II} semi-axes, there can be more than one maximum – hence the observation by many authors that the likelihood terrain for separate tree hypotheses under the usual ML procedure is highly irregular (see, e.g., Yang *et al.*, 1995).

Secondly, note that in Figures 6A–D, there are always points on the positive t_I semi-axis lying within the highest one or two likelihood contours, as would be expected – the data sets were simulated with trees of topology I. In Figure 6E, the fifth data set, neither the t_I nor the t_{II} axes cross the highest likelihood contours, betraying the non-"tree-like" nature of the data, which is actually a mixture of data sets simulated with two different trees. Figure 6F is also anomalous in that the t_I and t_{II} axes do not cross well within the two highest likelihood contours. This sixth data set has been simulated with three quarters of its sites being invariant; if one takes this site rate variation into account in the super-tree likelihood function [see Sections 2.6 and 4(a)], the contour spacing increases to look like Figure 6B.

Thirdly, compare the various tree-like data sets represented in Figures 6A-D. The contours in Figure 5B (a hard tree) are much more widely spaced than in Figure 6A (an easy tree), as expected for a "harder" tree. Figure 6C (long branches separate) represents a data set where the raw site pattern frequency $f_{AC|BD}$ is more than twice $f_{AB|CD}$, which, if naively taken to indicate $t_{II} > t_I$, appears to favour tree topology II. This is a data set where (uncorrected) maximum parsimony (MP) would fail and would choose the wrong tree II, succumbing to MP's infamous tendency to cluster long branches (Felsenstein, 1978). The likelihood function of Figure 6C, however, appropriately favours the t_I axis, corresponding to the correct tree topology I. The large down-correction that the relation (13) implicitly applies in going from observed $n_{AC|BD}$ to the parameter t_{II} leaves its imprint in the large likelihood variance in the t_{II} direction – compare Figure 6C to Figure 6D, simulated with a different tree (long branches together). In Figure 6D, the likelihood variance is largest in the t_I direction, corresponding to the large down-correction of $n_{AB|CD}$.

Finally, note that the plots in Figures 6B–E (the "hard" trees) show a positive correlation (especially strong for the tree in Figure 6B) between t_I and t_{II}.[6] A similar result

[6] The normalized covariance, given by

was found in (Waddell *et al.*, 1994). In fact, this is expected to be true for all "hard" cases of phylogeny inference, where the internal branch is relatively small. This can be intuitively understood as follows. If the external branches are long, compared to the internal branch, the observed site pattern frequency corresponding to the external branches ($f_{A|BCD}$, $f_{B|ACD}$, $f_{C|ABD}$, and $f_{D|ABC}$) are larger than those corresponding to internal branches ($f_{AB|CD}$, $f_{AC|BD}$, and $f_{AD|BC}$). Thus, the allowed variance [proportional to n_i by Poisson statistics; see footnote to Section 2.1] of the predicted site pattern numbers is larger for those corresponding to external bipartitions than those for internal bipartitions – that is, the external bipartition lengths are more free to change than the internal ones. Suppose an internal bipartition, say t_I, is forced to decrease from its maximum super-tree likelihood value. The most dramatic effect of such a change will be to reduce the predicted site pattern frequency $f_{AB|CD}$. Now, to optimise the likelihood, the external bipartition lengths (more free to move than t_{II} and t_{III}) will increase to help restore the predicted $f_{AB|CD}$; but in doing so, they will also increase the predicted $f_{AC|BD}$ and $f_{AD|BC}$. To counter-balance this, the t_{II} and t_{III} parameters will decrease (but less dramatically than the external bipartition lengths). As a result, there is a negative correlation of the internal bipartition length t_I with its surrounding branch lengths, and a positive correlation of t_I with "alternative" bipartition lengths t_{II} and t_{III}.

The intuition developed above, regarding regularity of the super-tree likelihood, overlap of high likelihood contours with the true tree hypothesis, and correlation between bipartition length parameters, will be used in Section 4 to produce statistical tests to apply to real data sets. With the simplest super-tree likelihood function (13) explicitly introduced, generalisations to more taxa and more complicated evolutionary models can now be discussed.

2.5. Generalisation to more taxa.

For a general problem of phylogeny inference with m taxa, there are more parameters and more degrees of freedom. There are $(2m–5) \times (2m–3) \times \ldots \times 3$ possible (unrooted) tree hypotheses [see, e.g., (Felsenstein, 1978)]. For the binary model which has been considered so far, there are $N = 2^m$ possible site patterns to fit, collapsed to 2^{m-1} patterns if one collects together inverses (e.g., $n_{ABF|CDEG} = n_{0011101} + n_{1100010}$ in a seven-taxon case), and therefore there are $2^{m-1}–1$ collapsed degrees of freedom, due to the constraint that the site pattern frequencies must sum to one. Also, the number of bipartitions is:

$$\frac{1}{2}\left[\binom{m}{2}+\binom{m}{3}+\ldots+\binom{m}{m-1}\right] = 2^{m-1}-1, \tag{15}$$

which is just enough parameters to describe the collapsed degrees of freedom.

A well-defined diagrammatic process for writing the super-tree likelihood function is given in Appendix A. The form is quite similar to (13), with $l = \sum n_i \log p_i - n \log 2$, and with each predicted site pattern probability p_i taking the form of the sum of several exponentials,

$$p_i = 2^{-(m-1)} \sum_{j=1}^{2^m} H_{ij} \exp(-\rho_j). \tag{16}$$

The exponents ρ_j correspond to "pathsets", defined as pairs, quartets, hextets, etc. of taxa. The element ρ_j for a given pathset is defined as the sum of bipartition lengths that would separate a single taxon of the pathset from the others. For example, for a five-taxon data set, a pathset corresponding to pair (AB) would correspond to

---

$$\frac{\partial^2 L}{\partial t_I \partial t_{II}} \Bigg/ \sqrt{\frac{\partial^2 L}{\partial^2 t_I}\frac{\partial^2 L}{\partial^2 t_{II}}},$$

has been numerically evaluated at the super-tree maximum to be –0.04, +0.4,+0.04,+0.04, 0.00, and +0.1 for the data sets shown Figures 6A–F, respectively. A positive value corresponds to positive correlation between t_I and t_{II}.

$$\rho_{(AB)} = t_{A|BCDE}+t_{AC|BDE}+t_{AD|BCE}+t_{AE|BCD}+t_{B|ACDE}+t_{BC|ADE}+t_{BD|ACE}+t_{BE|ACD} \quad (17)$$

and a pathset corresponding to quartet (ABCD) would correspond to

$$\rho_{(ABCD)} = t_{A|BCDE}+t_{B|ACDE}+t_{C|ABDE}+t_{E|ABCD}+t_{E|ABCD}+t_{BE|ACD}+t_{CE|ABD}+t_{DE|ABC}. \quad (18)$$

Note that the total number of pathsets is

$$\binom{m}{2}+\binom{m}{4}+\ldots = 2^{m-1}-1, \quad (19)$$

so the number of degrees of freedom are conserved. The components of the matrix $\boldsymbol{H} = \{H_{ij}\}$ are determined by assigning ±1 to each taxon depending on its character in a particular site pattern i, and multiplying these factors for each pathset j; see the derivation in Appendix A.

There is again a simple relation like (14) between sums of internal bipartition lengths ρ at the maximum super-tree likelihood point, and pair-wise distances t between taxa; indeed, $\rho_{(AB)}^{(max)} = t_{AB}$. The super-tree likelihood construction presented here is thus well-defined, being unique if one desires this property.

Whether the general m-taxon super-tree has a simple geometrical interpretation as in Figure 5 is not clear; if such a picture does exist, it would be a complicated multi-dimensional box with m appendages sticking out. It might make an interesting problem for a geometer / topologist to characterise this beast.

The general idea of the 4-taxon case carries through here. The super-tree likelihood is quite regular, with a single local maximum, obtained by solving the perfect fit $n_i = np_i$ (see Section 2.1). Individual tree hypotheses correspond to particular $(2m–3)$-dimensional hyper-planes cutting through the $2^{m–1}–1$ dimensional space. The length of a small internal bipartition is expected to exhibit a negative correlation with surrounding branches, and a positive correlation with alternative internal bipartitions.

In practice, fully characterising such a large-dimensional space for a given data set is difficult. The best one might hope to do is to maximise the likelihood for several viable tree hypotheses – the usual ML procedure – and to compare those values with each other and with the super-tree likelihood {given directly by the value $\sum n_i \log[n_i / n] – n \log 2$, where the collapsed site pattern frequencies are perfectly fitted}. In fact, these appear to provide sufficient information to select a given tree and to evaluate its statistical confidence; see Section 3.

2.5 Generalisation to more complicated evolutionary models

(a) General mutation matrix

Current analyses of molecular sequences with $c>2$ characters are more sophisticated than the simplest binary model discussed above:

- To model DNA changes, the 4×4 mutation matrix Q usually allows for the equilibrium frequencies π_A, π_G, π_C, and π_T to be different from 1/4, and for there to be a bias favouring transitions (A↔G, C↔T) over transversions (A↔C, G↔T, A↔T, C↔G), parameterised by κ [HKY85; (Hasegawa *et al.*, 1985)].
- To model amino acid changes, the 20×20 matrix Q is fitted to data tabulated from a wide range of proteins [see, e.g., (Jones *et al.*, 1992)].
- To model codon changes, a 61×61 matrix Q (there are $4^3–3$ codons, ignoring stop codons) can be approximately parameterised by unequal codon frequencies π_{AAA}, π_{AAU}, etc.; a transition / transversion ratio κ; and a non-synonymous / synonymous substitution ratio ω to mimic positive (ω > 1) or purifying (ω < 1) selection (Goldman and Yang, 1994).

Some analyses even allow the matrix Q to change from branch to branch. Can a super-tree likelihood function be found for these models? Indeed, it can be done, although it is not necessarily unique. Appendix A demonstrates that formulas for site pattern probabilities p_i (and thus a likelihood function) as functions of all possible bipartition lengths can be obtained; these functions reduce to the usual formulas for a given tree when bipartitions not in that tree are set to zero, as desired. For the simplest binary model, as well as for the Kimura 3-substitution-type (K-3ST) model for DNA with equal base frequencies (Kimura, 1980), diagrammatic procedures are given in Appendix A which produce unique super-tree likelihood functions of particularly simple forms.

There are some differences between the super-tree likelihood for the general model and the one for the simplest binary model. For the general model, the formula for each p_i is still a linear combination of exponentials of pathset sums ρ_j of bipartition lengths. Unlike before, the coefficients of the exponentials are no longer $\pm 1/2^{m-1}$. Also, unlike the binary model, for general unequal character frequencies, categories of site patterns cannot be collapsed together, as there is no longer the symmetry $0 \leftrightarrow 1$. There are thus a full $c^m - 1$ degrees of freedom to be fitted, with only $2^{m-1} - 1$ bipartition lengths (plus possibly a few parameters of Q) to fit them. As such, one cannot expect a perfect fit, where $n_i = np_i$, and the value of its likelihood at its overall maximum in super-tree space cannot be easily estimated without direct optimisation.

Despite the complexity of the super-tree formalism for more general models, the point is that it can be done in principle – therefore, likelihoods for different tree topologies, as determined by the usual ML procedure, can be considered special cases of a single likelihood function and can be directly compared.

(b) Rate variation among sites

Besides having a sophisticated rate matrix, ML analyses generally need to take into account site rate variation to adequately describe real data sets. The super-tree likelihood function can easily accommodate such extensions.

Consider the commonly used model where the evolutionary rates μ at different sites are assumed to be taken from a Gamma distribution, i.e.,

$$\frac{dP}{d\mu} \propto \mu^{\alpha^{-1}-1} e^{-\mu/\alpha} \qquad \text{(Gamma distribution)} \qquad (20)$$

Then the formulas for predicted site pattern frequencies p_i must be averaged over this distribution. It is trivial to show that, in fact, the only necessary modification to the super-tree predicted frequencies p_i is to replace all the exponentials $\exp(-\rho_i)$ in the formulas with $(1 - \alpha\,\rho_i)^{-1/\alpha}$. Other rate distributions, including uniform and Inverse Gaussian distributions, or discrete distributions with, e.g., a fraction of invariant sites ($f_{\text{invariant}}$), can also easily be implemented. See, e.g., (Waddell *et al.*, 1997).

Inclusion of site-rate-variation, however, should be considered a rather different procedure than changing the parameters of the rate matrix Q. For the binary model, varying the shape α of a Gamma distribution shifts the super-tree maximum likelihood parameters, but it does not affect the super-tree maximum likelihood value, $\sum n_i \log[n_i/n] - n\log 2$, since, at maximum likelihood, there will always be a perfect fit of the (collapsed) site pattern probabilities. This means that an ML value for α (or any other such site-rate-variation parameter, like the fraction of invariant sites) cannot be solved in the simplest binary model without invoking a particular tree hypothesis.

Note also that one must be careful about the number of parameters used to model the site-rate distribution. Suppose, for example, one investigates a four-taxon binary data set with the model discussed previously. If one models the site-rate-variation as a Gamma distribution with shape parameter α plus a fraction $f_{\text{invariant}}$ of invariant sites, there will be a problem. For each possible tree hypothesis, optimising the extra two parameters, α and $f_{\text{invariant}}$, counteracts the two constraints on internal bipartitions, e.g., $t_{\text{II}} = t_{\text{III}} = 0$, and the maximum super-tree likelihood, with perfect fit of (collapsed) site pattern probabilities, will be obtained for each tree topology. There will thus be no way to distinguish the ML values for each tree!

3. Comparison of tree likelihoods

How does the super-tree perspective clarify the problem of interest – inference of the correct phylogeny for *m* molecular sequences? In the super-tree likelihood framework, there is a continuous, well-defined likelihood function for any given set of $2^{m-1}-1$ bipartition lengths, e.g., (t_A, t_B, t_C, t_D, t_I, t_{II}, and t_{III}) in the four-taxon problem, and n_p evolutionary parameters like character frequencies and distribution of site rates, such that the predicted site pattern frequencies are greater than zero. The problem then is to choose between several composite hypotheses, corresponding to the regions of the parameter space where bipartition lengths not present in a given tree hypothesis are set to zero, e.g., ($t_I > 0$; $t_{II} = t_{III} = 0$) for tree topology I in the four-taxon case.

This section describes four main ways to carry out the statistical comparison of trees: (a) a "Bayesian" approach based on likelihood, multiplied by some prior, integrated over each tree hypothesis; (b) a more direct approach which compares likelihood values maximised over each tree hypothesis (the usual ML procedure) and which interprets likelihood ratios literally as posterior odds of trees; (c) a frequentist approach which selects a tree based on a set of statistics like the ML values for each tree hypothesis and then finds (by simulation) the probability of committing errors; and (d) bootstrapping, which can be interpreted as providing an approximation to any of the first three approaches, depending on one's prejudices. These approaches are described below in the super-tree likelihood framework, and then compared against computer simulations. In the end, it will be readily apparent that the second approach (b) is the most appropriate one for phylogenetic inference, as long as the assumed molecular evolutionary model is adequately realistic. The next section will describe a series of straightforward nested likelihood ratio tests which are useful in checking whether one is using a fully appropriate model when conducting ML phylogeny inference.

(a) Bayesian decision theory

In the Bayesian approach, the super-tree likelihood function, multiplied by a specified prior probability, is integrated over all the branch length parameters **t** of a given tree hypothesis. This gives the posterior probability value for that tree topology, by Bayes' Theorem:

$$P(\text{tree topology} \mid \text{data}) = \frac{P(\text{tree topology})P(\text{data} \mid \text{tree topology})}{P(\text{data})}, \qquad (21)$$

where P(tree topology) is a "prior" probability for a given tree topology (usually assumed to be the same for all tree topologies); P(data) is a normalisation factor that insures that the sum of the posterior probabilities of all tree topologies is unity; and

$$P(\text{tree topology} \mid \text{data}) = \int L(\mathbf{t}) f(\mathbf{t}) d\mathbf{t}, \qquad (22)$$

where the integration is over the $(2m-3)$-dimensional sub-space of the full parameter space defined by non-zero values of bipartition lengths which are in the given tree topology and by zero values for bipartitions not in the tree. The approach of picking the tree topology with maximum posterior probability was first proposed (although not from a super-tree perspective) by Rannala and Yang (1996); it is sometimes called maximum integrated likelihood (Steel and Penny, 2000). There is however a difficulty in this approach: one needs to define the prior probability function $f(\mathbf{t})$, encoding prior knowledge of putative branch lengths and evolutionary parameters for each tree topology. Note that the chosen prior must necessarily be "proper", i.e., integrate to unity over each tree hypothesis, since the likelihood function does not vanish for large branch lengths – see, e.g., the simple two-taxon case in Section 2.2. A uniform prior over all branch lengths is therefore not allowed. Rannala and Yang (1996) have used a proper prior for clock-like trees inspired by a model of speciation / extinction as a random birth / death process. Larget and Simon (1999) have assumed a different prior, uniform on all clock-like trees with total branch length sums less than a large constant.

The main problem with the Bayesian approach appears to be computational complexity. While the Markov Chain Monte Carlo algorithms of Yang and Rannala (1997) and Larget and Simon (1999) allow for the integration over a tree hypothesis to be carried out, they are restricted to being able to explore the full tree space only for small numbers of taxa (less than ten for the first algorithm, less than forty for the second method); also, the effect of site-rate distributions on the analyses has not yet been investigated.

(b) Direct likelihood comparison

Following the approach delineated by Edwards (1972) in Chapter 3 of his treatise *Likelihood*, comparing log-likelihood estimates maximised under each tree hypothesis – the usual ML procedure (Felsenstein, 1981), sometimes called maximum relative likelihood (Penny and Steel, 2000) – is expected to be valid, and much simpler, than the full Bayesian decision-theoretic analysis above. All tree hypotheses have the same "simplicity", i.e., number of fitting parameters, and the likelihoods of different trees are indeed comparable as is most clearly evident from the super-tree perspective, where the values are separate evaluations of a single super-tree likelihood function. Therefore, the ratios of the maximum likelihoods of two tree hypotheses provides an appropriate measure of their posterior odds.[7] Thus, one can define a likelihood-based support[8] value P^{l}(tree) for a given tree hypothesis as its (maximal) likelihood divided by the sum of the (maximal) likelihoods of all the considered tree hypotheses. Previously, the likelihood ratios have generally not been interpreted so literally [for an exception, see (Strimmer and von Haeseler, 1997)] due to the theoretical uncertainty regarding the apparent difference in the likelihood form for different tree topologies [see (Nei, 1987), (Yang *et al.*, 1995)]; the super-tree perspective of Section 2, however, mitigates this uncertainty.

The ML approach, which can be applied with a uniform prior, is expected to produce more conservative support estimates than the full Bayesian integration described above as $n \to \infty$. Explicitly, the ratio of the likelihood of the ML tree to the likelihood of any alternative tree is expected to be lower than the corresponding ratios of integrated posterior probabilities, since integration over the alternative tree parameters (which have ML internal branch lengths closer to the boundary at zero) will pick up less of the higher likelihood regions than the best tree. Indeed, analysis of a 895-bp mitochondrial DNA data set from six primates by Rannala and Yang (1996) show that the $\Delta l = \Delta[\log L]$ values from the usual ML analysis are consistently lower than the values $\Delta[\log P(\text{tree} \mid \text{data})]$ based on integrated posterior probabilities.

(c) Frequentist approach; critical regions.

A frequentist approach is commonly assumed in the literature of phylogeny simulation. Instead of simply assigning a "degree of belief" (like a single integrated posterior probability or likelihood value) to each tree as in the analyses above, the idea is to find a set of the statistics, and then a tree selection procedure based on comparison of these statistics that minimizes the possibility of making the wrong choice.

Unfortunately the phylogeny estimation problem requires a comparison between composite, non-nested hypotheses, and statistical theory offers little help in selecting a set of statistics or a selection criterion; see Chapter 23 of (Kendall and Stuart, 1961). Instead, biologists use a heuristic method; common statistics are the branch length sum S of each tree (in minimum-evolution distance methods), the total number of evolutionary changes N(steps) (in parsimony / cladistic analyses), or the values of the maximum log-likelihood l for each tree hypothesis. The tree with the smallest value of S, N(steps), or $-l$ is chosen as the best tree.[9]

---

[7] This is the approach in, e.g., example 3.7.1 of Edwards (1972). Note that if it is possible to reformulate a problem so that the likelihood ratio of composite hypotheses is independent of the unspecified, "nuisance" parameters, this should, of course, be done [section 6.3, Edwards(1972)]. However, it does not seem possible to remove explicit dependence on the internal branch lengths from likelihood ratios in the phylogeny inference problem.

[8] "Support" is used here with the meaning of a "degree of belief" (in the range of 0% to 100%). It is not meant in Edwards'(1972) sense of a log-likelihood difference.

[9] Among the statistics listed, maximum likelihood is almost certainly the best criterion in the frequentist approach to composite hypotheses. In particular, it turns out that for easier problems involving comparison of simple hypotheses, as well as for very special cases involving composite hypotheses, using a likelihood criterion (possibly with a slight

To statistically evaluate the selection procedure, one needs to know the power function, the probability of accepting the tree topology given that the true tree shares or does not share the same topology as the found tree – the "true positive" rate (probability of not making a Type I error) and the "false positive" rate (probabilities of making a Type II error), respectively. See, e.g., section 22.24 of (Kendall and Stuart, 1961). Simulation-based papers focusing on phylogeny reconstruction have been able to estimate the true positive rate, by simulating several replicates of a single known tree. However, it seems difficult to estimate a false positive rate in the general comparison between composite hypotheses. In particular, for phylogeny estimation, it is not known beforehand which alternative trees are expected to occur in Nature and therefore might produce a "background" signal. Therefore, the frequentist approach applied to the composite hypotheses of phylogenetic inference seems incomplete, as well as computationally burdensome.

As an example of the inappropriateness of the frequentist approach, consider the tree simulated in Figure 6D, with long branches together. It is well known that maximum parsimony (MP) is "better" (i.e., has larger true positive rate) than maximum relative likelihood at selecting this tree from data simulated with the tree [see, e.g., (Yang, 1994a)]. However, MP (but not necessarily ML) will also incorrectly select this tree if the data was simulated with an alternative tree with the long branches separate; this is the infamous bias of parsimony to collect together long branches, the Felsenstein zone (1978). Thus, in a loose sense, MP has higher false positive rate than ML. However, a biologist might claim that the alternative long-branches-separate tree does not arise in Nature due to the molecular clock constraining the shape of trees connecting four extant taxa, so that truly the false positive rate of MP is acceptably low for biologically relevant trees. But in that case, one could just as well include the molecular clock assumption in the ML analysis (by the appropriate constraint on branch lengths), and one expects then that both its true positive rate and false positive rate would improve (increase and decrease, respectively) over MP. Nevertheless, in either case, it appears that the rather important step of estimating a false positive rate is not very well-defined in the frequentist approach to statistically evaluating tree-selection methods.

(d) Bootstrapping

Bootstrapping describes the process of generating (pseudo)replicates of the given data set by resampling its sites, with replacement; see, e.g., (Efron and Tibshirani, 1993). Given a tree selection procedure like maximum parsimony or maximum likelihood, the bootstrap support value P^{boot}(tree) for a given tree is the frequency at which is selected by the procedure among bootstrap replicates (Felsenstein, 1985); it is very commonly used in the current literature to statistically evaluate tree selection procedures on real molecular data. Efron *et al.* (1996) have claimed that this support value is in fact a reasonable assessment of the Bayesian posterior probability of the tree, given a uniform prior and the assumption that the cut on the selection statistic would correctly separate the trees in the limit of no statistical noise. Their analysis, however, appears to rely on a picture of a continuous, convex parameter space divided into several contiguous regions corresponding to trees. From the super-tree perspective, one might tentatively construct such a picture by chopping up the super-tree parameter space so that, e.g., the region ($t_{\text{I}} > t_{\text{II}}$; $t_{\text{I}} > t_{\text{III}}$) corresponds to tree topology I in the four-taxon problem. However, this is a different, and arguably incorrect, formulation of the tree selection problem from the one described above, in terms of composite hypotheses.

An alternative interpretation of the bootstrap support value is that, in being taken from pseudo-replicates of the correct tree, it is an estimate of the true positive rate, the probability of not making a type I error. This appears to be the idea implemented by Kishino and Hasegawa (1989) with their relative estimation of log-likelihood (RELL) technique, which estimates the distribution of likelihood ratio between the two simple (fully specified) hypotheses by bootstrapping. This interpretation has essentially been discounted, however, by several simulation studies; see, e.g., (Zharkikh and Li, 1992) and (Hillis and Bull, 1993). In particular, when the true positive rate is high, the bootstrap

---

modification to correct bias) does indeed yield a uniformly most powerful (UMP) or uniformly most powerful unbiased (UMPU) test. See sections 22.10 and 24.24 in (Kendall and Stuart, 1961).

support tends to underestimate it; and when the true positive rate is low, the bootstrap support can give misleadingly high values for the correct or incorrect tree.

(e) Comparison of statistical approaches by simulation.

This sub-section describes tree selection for four-taxon and six-taxon simulated data sets (1000 replicates each) for the simplest binary model. There are four main simulated data sets, using the simplest binary model for character evolution with mutation rates uniform across all sites, using the four-taxon (100 bases and 500 bases) and six-taxon trees (500 bases and 5000 bases) in Figure 7. A thousand replicates were simulated using PAML's *evolver* and analysed using PAML's *baseml*. Figure 8 shows histograms of two possible measures of statistical confidence – the likelihood-based support value $P^{l}(\text{tree}) = L_{\text{tree}} / (L_{I}+L_{II}+L_{III})$ and the bootstrap support value $P^{boot}(\text{tree})$ – for the correct tree and for an alternative tree. Note that for the six-taxon case, only trees of topology I, II, and III in Figure 7 were assumed to be valid trees. The bootstrap supports have been computed using the RELL technique of Kishino and Hasegawa (1989), implemented in PAML's *rell* application. For the six-taxon data sets, where the simulated tree is clock-like, Bayesian posterior probabilities (integrating the likelihood over each tree topology) have also been calculated using PAML's *mcmctree* application.[10] These results are shown in Figure 9, and compared to $P^{l}(\text{tree})$ where the likelihoods have been recomputed with a molecular clock assumption.

Studying the histograms in Figure 8 yields several insights. On one hand, the behaviour of the likelihood-based support value $P^{l}(\text{tree})$ is intuitive. For data sets with low sequence lengths (and little phylogenetic information), the support is distributed around 30–40% for both the correct and wrong trees. For longer sequence lengths, a large fraction of data sets are phylogenetically informative. The likelihood-based support values for the correct and incorrect tree are then clustered near 90–100% and 0–10%, respectively, as expected.

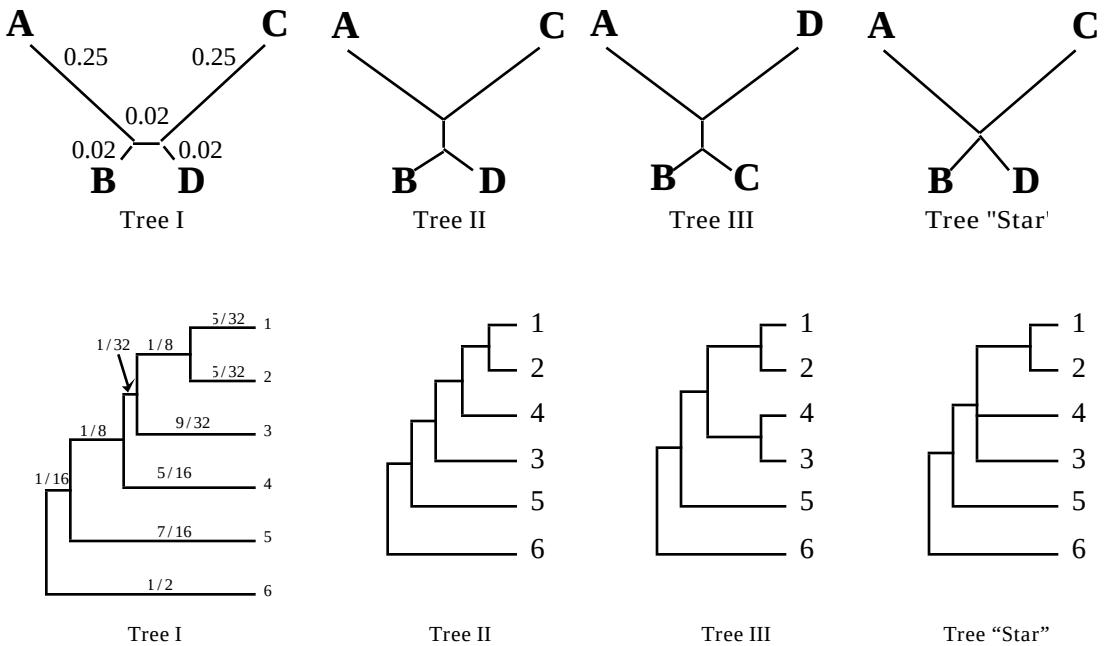
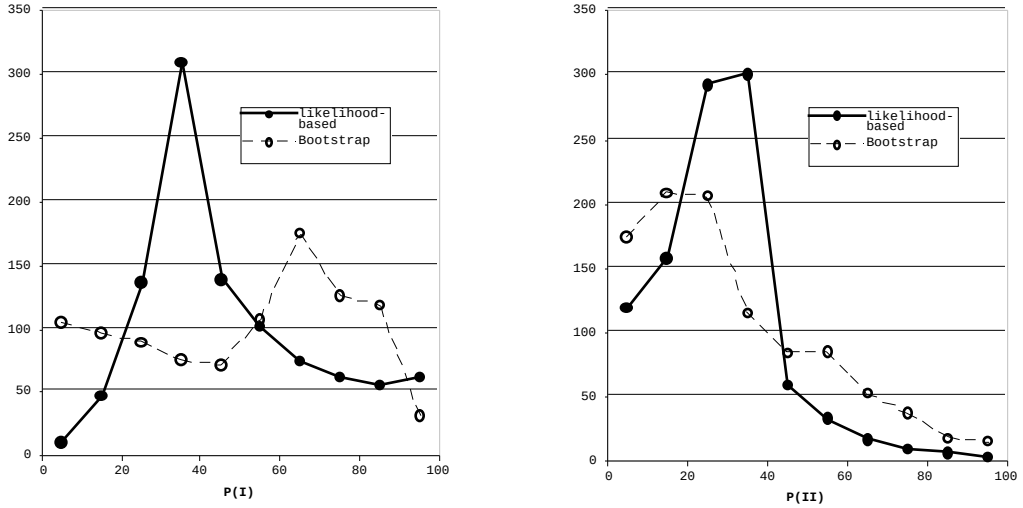
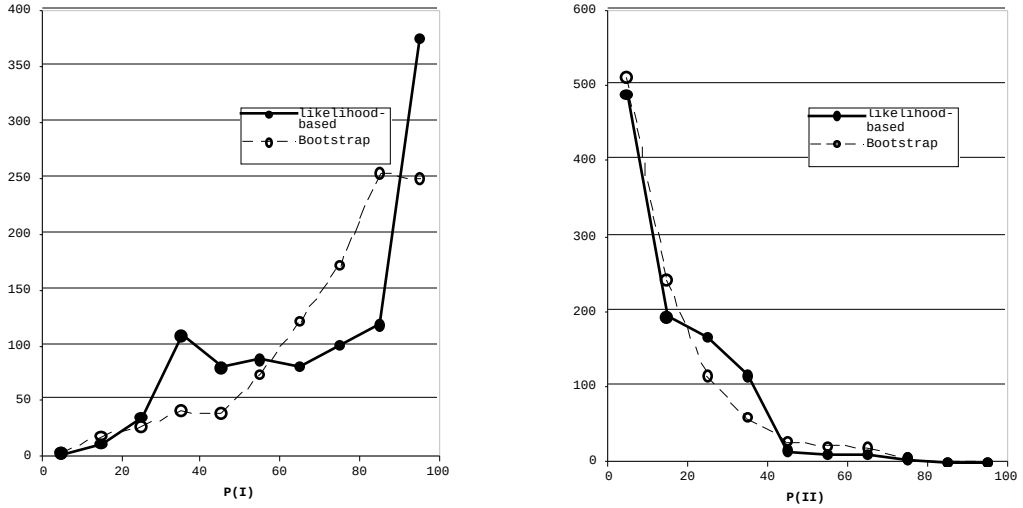


Figure 7. Four-taxon and six-taxon trees considered in the simulation studies. Tree topology I is the simulated (true) phylogeny in both sets.

[10] Input parameters to the mcmctree program were as follows: empirical Bayesian analysis option; iteration accuracy parameter $\delta_1 = 0.5$; average number of mutations per site from root to present $m = 0.5$ (the simulated value); birth rate $\lambda = 6.7$; death rate $\mu = 2.5$; species sampling fraction $\rho = 0.06$. The last three parameters are those used in Yang and Rannala (1997) to describe a primate data set; they describe a fairly flat prior distribution for divergence times relative to present. It has been checked that for a dozen replicates that lowering _ _, or changing _, _, or _ does not change the posterior probabilities by more than a few percent.

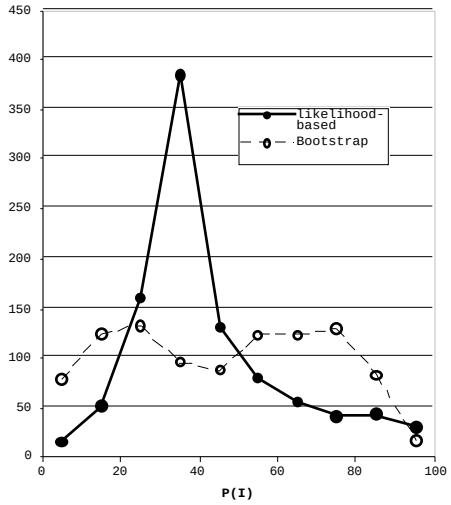


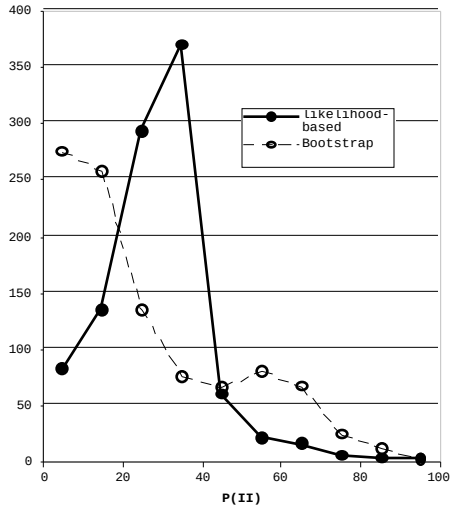
Four taxa, 100 characters



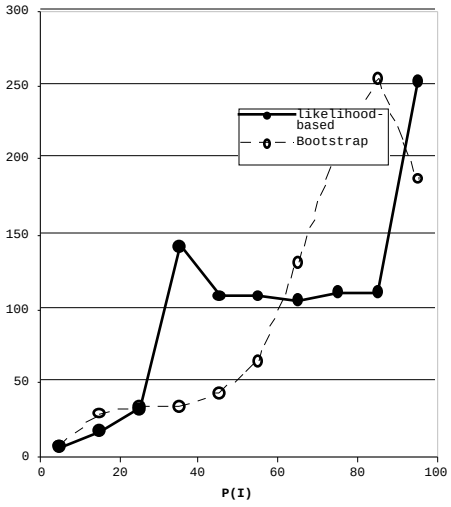
Four taxa, 500 characters

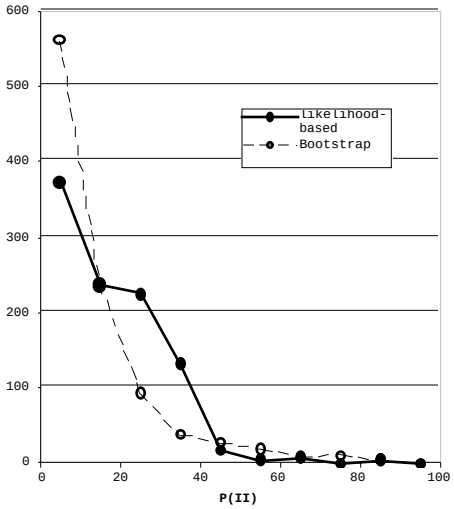
Figure 8 (continued on next page). Comparison of likelihood-based support values $P^{l}(\text{tree}) = L_{\text{tree}}/(L_{\text{I}}+L_{\text{II}}+L_{\text{III}})$ and bootstrap support values $P^{\text{boot}}(\text{tree})$ for the correct tree (I) and the wrong tree (II). Histograms are shown for the data sets simulated with the four-taxon and six-taxon trees of Figure 7, with different sequence lengths.





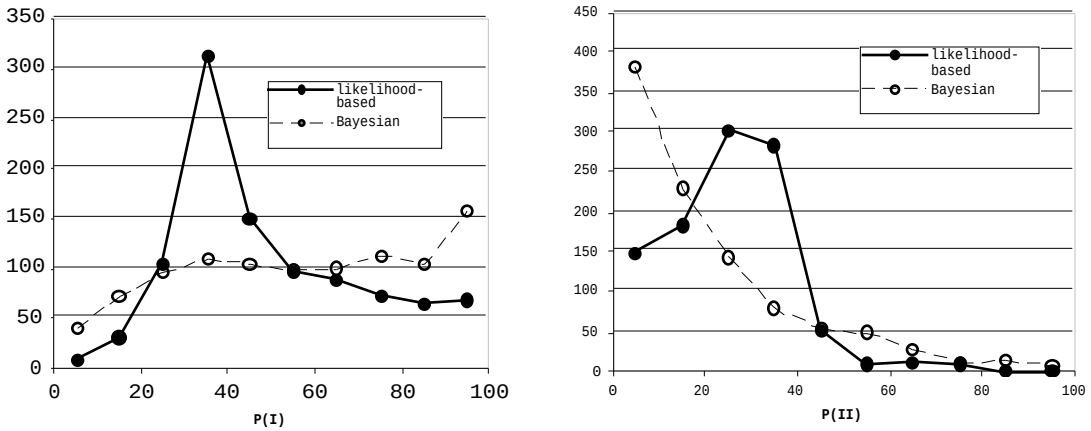
Six taxa, 500 characters



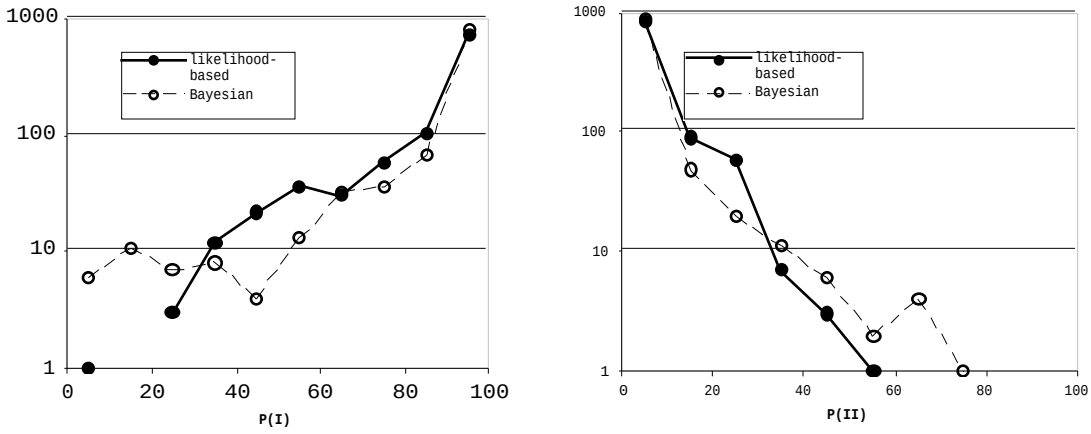


Six taxa, 5000 characters

Figure 8, continued from previous page.



Six taxa, 500 characters



Six taxa, 5000 characters

Figure 9. Comparison of likelihood-based support values $P^{l}(\text{tree}) = L_{\text{tree}}/(L_{\text{I}}+L_{\text{II}}+L_{\text{III}})$ and Bayesian integrated likelihood posterior probabilities for the correct tree (I) and the wrong tree (II); likelihoods are calculated with a molecular clock assumption. Histograms (1000 replicates) are shown for the simulated six-taxon data sets. Note the change to a log-scale, to better show the tails of the distributions, in the bottom plots.

number of chars.	MP	ML (no clock)	ML (clock)	MAP (clock)
500	54.0%	53.1%	69.2%	67.1%
5000	91.2%	87.1%	99.2%	96.8%

Table 1. Probability of accepting the correct tree by maximum parsimony (MP), maximum (relative) likelihood (ML) with and without a molecular clock assumption, and the maximum (integrated) likelihood MAP analyses (with clock assumption). The results are for the six-taxon tree in Figure 7, simulated with the simplest binary model (1000 replicates).

On the other hand, the bootstrap support value P^{boot}(tree) has undesirable properties. In the low-sequence-length data sets, it ranges almost uniformly from 0–100% for the correct tree and can be misleadingly high for the wrong tree. Furthermore, for higher statistics data sets, the bootstrap support value for the correct tree clusters near a value less than 100%.

The comparison for the six-taxon case of the Bayesian posterior probabilities of the MAP analysis of Yang and Rannala (1997) and the likelihood-based support values is also interesting. The right-hand column of Figure 9 shows that the MAP support value is more likely to be large (> 50%) for the wrong tree than the likelihood-based support value. Also, as predicted, the Bayesian posterior probabilities are generally higher than the P^{l}(tree) values, even for the wrong tree.

Finally, consider the "performance" of all the methods as measured by the probability of accepting the correct tree, summarised in Table 1 for the six-taxon simulations for maximum parsimony, for maximum likelihood (with and without molecular clock assumption), and for the MAP analysis (with molecular clock assumption). As discussed above, this value, used by e.g., (Yang, 1996a) and (Penny and Steel, 2000), is only half the story in assessment of phylogenetic methods; one would also like a measure of the false positive rate, the probability that one has not accepted the wrong tree. Nevertheless, concentrating on the true positive rate, Table 1 shows that maximum parsimony "out-performs" the usual ML analysis without a molecular clock assumption, as would be expected since the relevant internal branch (see tree diagram in Figure 7) is connected to long external branches to taxon 1 and taxon 2. However, when the molecular clock assumption is made, the performance of ML improves dramatically over MP. Somewhat surprisingly, the ML-clock analysis also outperforms the MAP analysis by a statistically significant margin (based on 1000 replicates). However, this may be a result of assuming an inefficient MAP prior, and needs to be investigated further.

Based on the above results, the simplest valid procedure for phylogeny inference appears to be the usual ML procedure, which is to find optimal likelihoods for each tree hypothesis, and to pick the one with the highest likelihood as the best. The ratio of each tree ML value to the sum of all tree likelihood values can then be interpreted as a useful and completely intuitive estimator of statistical support. Note, of course, that this procedure is only valid if the evolutionary model of the molecular sequences is correct – the next section described a set of consistency checks which should be applied to test the adequacy of an evolutionary model in describing a data set, before any final phylogeny inferences are made.

4. Proposed statistical tests to check model adequacy

The previous section has argued that the usual ML procedure is statistically sound, and that the likelihood ratios are indeed directly interpretable between different trees as posterior odds of the trees (unlike the bootstrap), if the assumed model of molecular mutation is correct. But real data sets are certainly very complex – recombination, purifying / positive selection (possibly acting differently on different taxa), and compensating mutations, among other biological effects, may conceivably undermine the simple molecular mutation models used in programs like PAML. This section proposes a set of "consistency checks" – likelihood ratio tests which are able, in some cases, to reject inadequate models.

Table 2 summarises the hierarchy of nested models whose likelihoods one can evaluate for a given bipartition t_{I} in a given data set with a given likelihood model. The value $l_{\max}$ is defined as $\sum n_i \log[n_i / n]$. The star tree is defined by constraining $t_{\mathrm{I}} = t_{\mathrm{II}} = t_{\mathrm{III}} = 0$. Four likelihood ratio tests will described, based on the following statistics:

(a) $(l_{\text{super-tree}} - l_{\mathrm{I}})$
(b) $(l_{\max} - l_{\text{super-tree}})$
(c) $(l_{\mathrm{II/III}} - l_{\text{star}})$
(d) $(l_{\max} - l_{\mathrm{I}})$.

Statistics (a) and (c) have been chosen as indicators of how close the higher likelihood contours pass near parameters corresponding to the tree I topology; see, e.g., the diagram in Figure 4. Statistics (b) and (d) indicate how much worse the likelihood gets if one uses a finite set of parameters to fit the data rather than an infinite set, which would provide a perfect fit ($l \to l_{\max}$). The two statistics (a) and (b) require computation of $l_{\text{super-tree}}$, which is trivial for the simplest binary model (where there is a perfect fit to the collapsed site pattern frequencies; see Section 2.5), but for general models is rather more complicated (and possibly not unique; see Appendix A), since it involves the optimisation of at least 2^{m-1} parameters. The two other statistics (c) and (d) give much the same information as the two involving $l_{\text{super-tree}}$, but are somewhat less sensitive.

Note that even more sensitive tests of unequal frequencies, site-rate-variation, etc. can be applied if one solves the maximum likelihood for a given tree under a more general model and sees the change $2\Delta l$ in going to the nested, less general model [see, e.g., (Yang, 1994a), (Goldman and Whelan, 2000)]. However, such tests do not give the researcher a way to check the overall adequacy of the final, most general model considered; the likelihood ratio tests described in this section do offer such a consistency check.

Likelihood value	Fitted parameters	Number of parameters (simplest binary model)	Number of parameters (general model)
$l_{\max}$	as many as possible (perfect fit)	$2^m - 1$	$c^m - 1$
$l_{\text{super-tree}}$	all bipartition lengths, plus evolutionary params.	$2^{m-1} - 1$	$2^{m-1} - 1 + n_{\text{param}}$
$l_{\mathrm{I/II/III}}$	bipartition lengths in the ML tree I (or nearest-neighbour alternatives II / III), plus evolutionary params.	$(2m - 3)$	$(2m - 3) + n_{\text{param}}$
l_{star}	bipartition lengths in a given tree with a forced multifurcation, plus evolutionary params.	$(2m - 4)$	$(2m - 4) + n_{\text{param}}$

Table 2. Summary of the models considered in these reports, listed in nested order, from most general to least general.

Data set (model)	$l_{super} - l_I$	$l_{super} - l_{II}$	$l_{super} - l_{III}$	$l_{max} - l_{super}$	$l_I - l_{star}$	$l_{II} - l_{star}$	$l_{III} - l_{star}$	$l_{max} - l_I$
4 taxa, 100 characters	1.3±1.0 (1.0±1.0)	1.9±1.5 (>1)	2.1±1.5 (>1)	4.1±1.9 (4.0±2.0)	0.82±1.1 (>0.25)	0.21±0.50 (<0.25)	0.10±0.30 (<0.25)	5.5±2.1 (4.5±2.2)
4 taxa, 500 characters	1.1±1.1 (1.0±1.0)	3.7±2.8 (>1)	3.8±2.8 (>1)	4.2±2.2 (4.0±2.0)	2.8±2.5 (>0.25)	0.16±0.41 (<0.25)	0.05±0.21 (<0.25)	5.3±2.4 (4.5±2.2)
6 taxa, 500 characters	11.3±3.4 (11.0±3.4)	11.7±3.5 (>11)	11.8±3.5 (>11)	17.5±4.5 (16.0±4.0)	0.58±0.93 (>0.25)	0.15±0.43 (<0.25)	0.16±0.40 (<0.25)	28.8±5.5 (29.5±5.4)
6 taxa, 5000 characters	10.8±3.2 (11.0±3.4)	12.7±3.7 (>11)	12.7±3.7 (>11)	16.0±4.0 (16.0±4.0)	2.0±1.9 (>0.25)	0.07±0.26 (<0.25)	0.08±0.29 (<0.25)	26.9±5.3 (29.5±5.4)
Unequal freq. (not taken into account)	2.9±2.3 (1.0±1.0)	5.0±3.2 (>1)	6.5±3.6 (>1)	145±13 (4.0±2.0)	3.7±2.8 (>0.25)	1.6±1.8 (<0.25)	0.13±0.38 (<0.25)	148±13 (4.5±2.2)
Unequal freq. (taken into account)	— (0.5±0.7)	— (>0.5)	— (>0.5)	— (3.5±1.9)	2.5±2.3 (>0.25)	0.09±0.24 (<0.25)	0.02±0.09 (<0.25)	4.8±2.3 (4.5±2.2)
Gamma rate var. (not taken into account)	6.2±3.5 (1.0±1.0)	7.6±4.3 (>1)	10.3±4.7 (>1)	4.3±2.1 (4.0±2.0)	4.7±3.5 (>0.25)	3.2±2.6 (<0.25)	0.57±1.00 (<0.25)	10.4±4.0 (5.0±2.2)
Gamma rate variation (taken into account)	0.6±0.8 (0.5±0.7)	2.2±2.0 (>0.5)	2.2±2.0 (>0.5)	4.3±2.1 (4.0±2.0)	1.7±1.8 (>0.25)	0.06±0.21 (<0.25)	0.07±0.25 (<0.25)	2.8±2.3 (4.5±2.2)

Table 3. Summary of mean±standard deviation (with expected values in parantheses if the assumed evolutionary model is correct) of ($l_{super\text{-}tree} - l_{I/II/III}$), ($l_{max} - l_{super\text{-}tree}$), ($l_{I/II/III} - l_{star}$), and ($l_{max} - l_I$) in the simulated binary data sets (1000 replicates each). The last four rows, which show tests of the presence of unequal character frequencies and site-rate-variation, are 4-taxa sets of 500 characters each. For the model with unequal character frequencies, the super-tree likelihood is not uniquely defined (see text) and difficult to evaluate without direct optimisation; so those entries are not evaluated.

The expected asymptotic distributions of these statistics is discussed below, and are checked against computer simulations. The four main simulated data sets have been described in the previous section. Histograms of likelihood ratio statistics from these runs are shown in Figures 10–12. In addition, one data set of 1000 replicates of 500 bases each, with the four-taxon tree, was simulated with unequal character frequencies ($\pi_0 = 0.8$, $\pi_1 = 0.2$); and one data set (also 1000 replicates, 500 bases) was simulated with site-rate-variation described by a Gamma distribution with $\alpha = 0.5$, but with equal character frequencies. Summary likelihood ratio statistics for all of the simulations are given in Table 3.

(a) The statistic ($l_{\text{super-tree}} - l_{\text{I}}$)

One of the main observations of Section 2.4 was that for simulated data sets based on a single tree topology, a point in super-tree space corresponding to the correct tree topology was always enclosed by one of the high-likelihood contours. More quantitatively, for large n, one expects the statistic $2(l_{\text{max}} - l_{\text{I}})$ to be distributed as χ^2_k, where k is the difference in the number of parameters of the full model and the nested model [chapter 24 in (Kendall and Stuart, 1961)]. Explicitly,

$$\frac{dP}{dl} \propto l^{k/2-1} e^{-l}, \qquad (\chi^2_k \text{ distribution}) \tag{23}$$

with the proportionality constant determined by normalising the distribution to have unit integral. In this case,

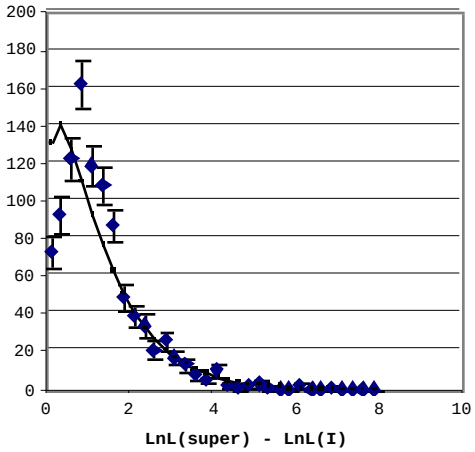
$$k = (\#\text{ super-tree parameters}) - (\#\text{ tree parameters}) = (2^{m-1} - 1) - (2m - 3),$$

which equals 2 (m = 4), 8 (m = 5), 22 (m = 6) etc., in the asymptotic limit $n \to \infty$. If, in a given data set with a given evolutionary model, the best tree hypothesis yields a $(l_{\text{super-tree}} - l_{\text{I}})$ larger than a critical value, as determined by a χ^2 test at some specified significance level, one can reject the assumed evolutionary model. Note that in using the likelihood l_{I} of the ML tree hypothesis rather than some guess of the "true tree" (which may be unknown), the test is checking a "best-case" scenario – if the ML tree is rejected by the $(l_{\text{super-tree}} - l_{\text{I}})$ test than so will any other tree.

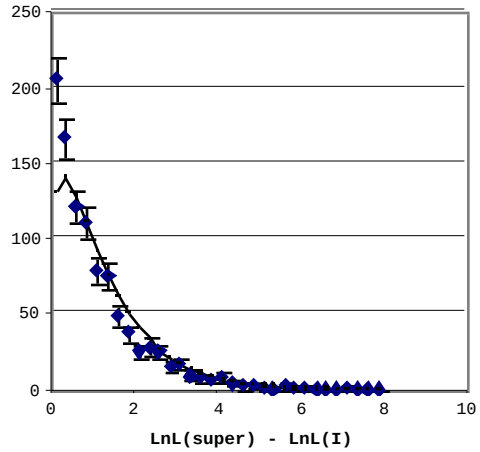
The histograms of Figure 10 show the actual $(l_{\text{super-tree}} - l_{\text{I}})$ distributions for four simulated data sets fitted to χ^2_k distributions. The four-taxon data sets fit[11] to $k \approx 2.6$ (100 bases) and $k = 2.1$ (500 bases), which are close to the expected limit $k = 2$. In particular, while the fits are not perfect for the lowest values of $(l_{\text{super-tree}} - l_{\text{I}})$, they are excellent for the right-hand tail of the distribution, which is the crucial part for carrying out confidence tests. The six-taxon data sets fit well through the whole range of $(l_{\text{super-tree}} - l_{\text{I}})$, with $k = 22.6$ (500 bases) and $k = 21.6$ (5000 bases), quite close to the expected $k = 22$.

To illustrate the power of the test, consider the results of a more complex data simulations with unequal character frequencies, and with site-rate-variation parameterised by a Γ distribution ($\alpha = 0.5$), summarised in Table 3. When the unequal frequencies are not taken into account [i.e., when one forces $\pi_0 = \pi_1 = 1/2$] in the former data set, the $(l_{\text{super-tree}} - l_{\text{I}})$ statistic is found to be 2.9±2.3 (*mean* ± *standard deviation*), with 1.0±1.0 being the expected value. And when, in the latter data set, the site-rate-variation is ignored and then taken into account, $(l_{\text{super-tree}} - l_{\text{I}})$ decreases dramatically from 6.2±3.5 (expected: 1.0±1.0) to 0.6±0.8 (expected: 0.5±0.7). Thus, in the majority of data sets, if one fails to take into account either unequal frequencies or site-rate-variation, checking $(l_{\text{super–tree}} - l_{\text{I}})$ will provide a necessary warning, especially in the latter case.

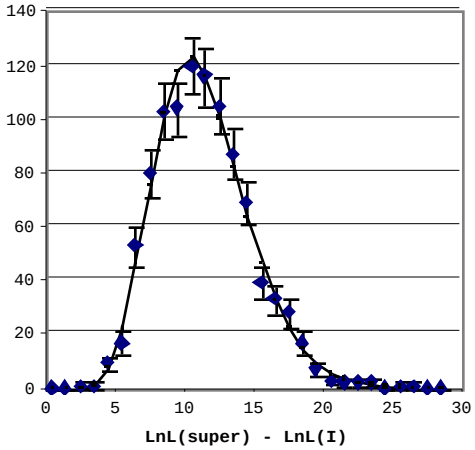
[11] Fits of the χ^2_k distribution to simulations here (and later) are accomplished by setting $k/2$ equal to the mean of the log-likelihood-difference found in the simulation.



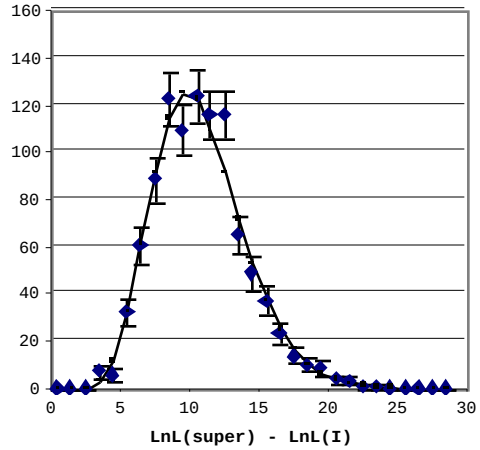
Four taxa, 100 characters



Four taxa, 500 characters



Six taxa, 500 characters



Six taxa, 5000 characters

Figure 10. Distributions of the $(l_{\text{super-tree}} - l_i)$ statistic compared to predicted distributions of the form χ^2_k for simulations (1000 replicates) of the simplest binary model for the four-taxon and six-taxon trees shown in Figure 7. Fitted values of the parameter *k* are given in the text.

(b) The statistic $(l_{max} - l_{super\text{-}tree})$

A second way to check the description of the evolutionary process is by looking at the statistic $(l_{max} - l_{super\text{-}tree})$.[12] In the limit $n \to \infty$, $2(l_{max} - l_{super\text{-}tree})$ is expected to follow a χ^2_k distribution with

$$k = (\#\text{ site patterns}) - (\#\text{ super-tree parameters}) = (c^m - 1) - (2^{m-1} - 1) = c^m - 2^{m-1}$$

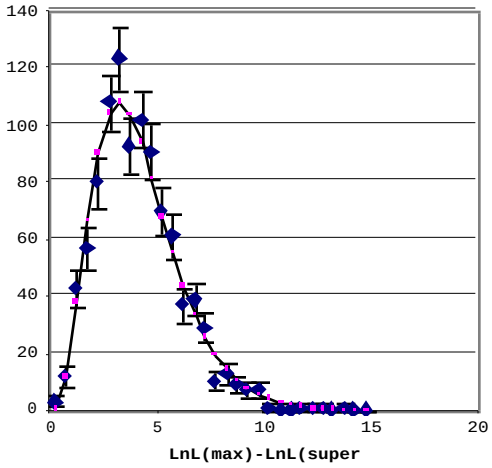
(where c is the number of possible character states). For finite n, not all site patterns show up in the data, and k is usually different from than the asymptotic limit but can still be estimated fairly easily; see the discussion of $(l_{max} - l_I)$ in (d) below.

The histograms of Figure 11 show the $(l_{max} - l_{super\text{-}tree})$ distributions for four simulated data sets fitted to χ^2_k distributions. The four taxa data sets fit well with $k = 8.2$ (100bases) and $k = 8.4$ (500 bases), with the expected asymptotic limit being $k = 8$. The six-taxon data sets fit well to $k = 34$ (500 bases) and $k = 32$ (5000 bases), with expected asymptotic limit being $k = 32$.

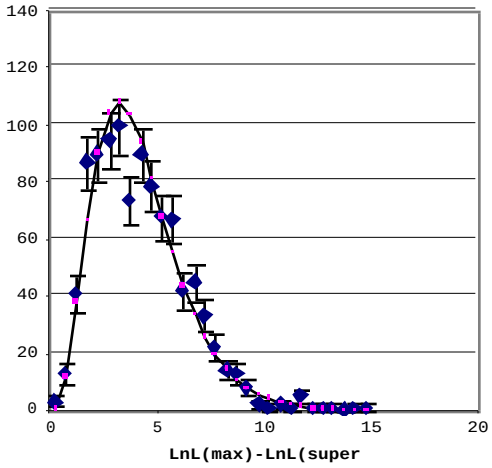
The $(l_{max} - l_{super\text{-}tree})$statistic, like $(l_{super\text{-}tree} - l_I)$ above, provides a particularly good consistency check on data sets which have unequal frequencies; see Table 3. For the simulated data set with $(\pi_0 = 0.8, \pi_1 = 0.2)$, not taking into account the unequal frequencies yields $(l_{max} - l_{super\text{-}tree}) = 145 \pm 13$, much larger than the expected value being 4.5 ± 2.2. Note that this statistic cannot provide a sensitive test for site-rate-variation, since $l_{super\text{-}tree}$ does not vary with, e.g., the shape parameter of an assumed Γ distribution of site rates. Indeed, for the simulated data set with site-rate-variation, $(l_{max} - l_{super\text{-}tree}) = 4.8 \pm 2.3$, with the expected value, 4.5 ± 2.2.

Thus the $(l_{max} - l_{super\text{-}tree})$ statistic is especially good at catching inconsistencies in unequal frequencies, and, presumably more generally, other parameters of the mutation matrix. The other $(l_{super\text{-}tree} - l_I)$ statistic described above should be used to check site-rate-variation.

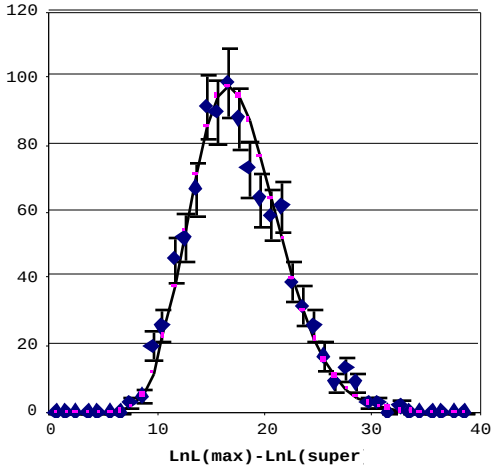
[12] The reader may wonder why this statistic doesn't vanish for the simplest binary model, where one can attain a "perfect fit". The answer is that the super-tree can perfectly fit the *collapsed* degrees of freedom – but in collapsing pairs of site patterns which are inverses of each other (under the switch $0 \leftrightarrow 1$), an assumption has been made that the inverses in each pair occur with equal probabilities. Testing $(l_{max} - l_{super\text{-}tree})$ checks this assumption.



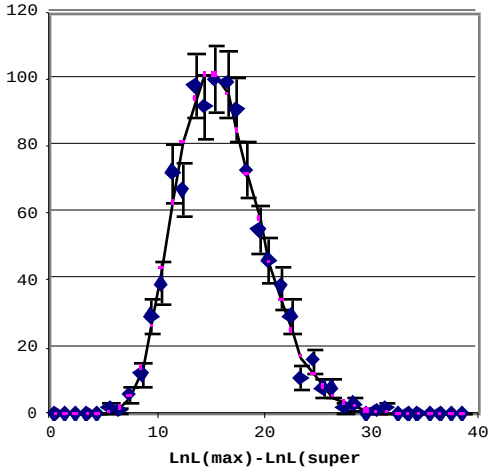
Four taxa, 100 characters



Four taxa, 500 characters



Six taxa, 500 characters



Six taxa, 5000 characters

Figure 11. Distributions of the $(l_{max}-l_{super\text{-}tree})$ statistic compared to predicted distributions of the form χ^2_k for simulations (1000 replicates) of the simplest binary model for the four-taxon and six-taxon trees shown in Figure 7. Fitted values of the parameter k are given in the text.

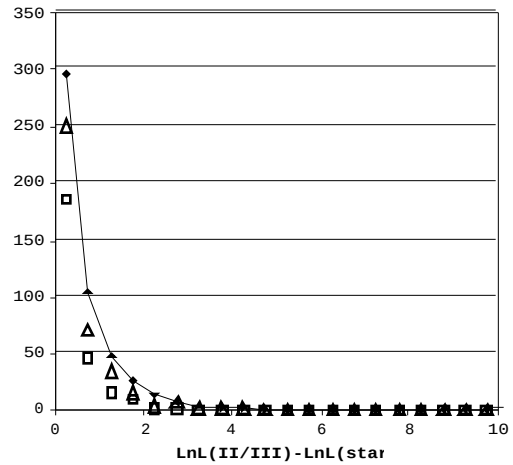
(c) The statistics $(l_{II} - l_{star})$ and $(l_{III} - l_{star})$

Consider $(l_{II} - l_{star})$ and $(l_{III} - l_{star})$. Loosely speaking, these ratios, like $(l_{super\text{-}tree} - l_I)$ statistic in (a) above, provide an indication of how close the t_I axis crosses near the higher super-tree likelihood contours (see Figure 4). The test has been mentioned in, e.g., (Yang *et al.*, 1994), but the distribution of the test statistics has not been predicted. The star tree model has one less parameter than the tree II (or tree III) models, and therefore one might naively expect $2(l_{II} - l_{star})$ to asymptotically follow a $(1/2)\chi^2_0+(1/2)\chi^2_1$ distribution. The reason for mixing in the χ^2_0 distribution, defined as a Dirac delta function at zero, is that the optimal t_{II} under tree II hypothesis (i.e., $t_I = t_{III} = 0$, $t_{II} > 0$) will be at the boundary, zero, half the time; see (Ota *et al.*, 2000).

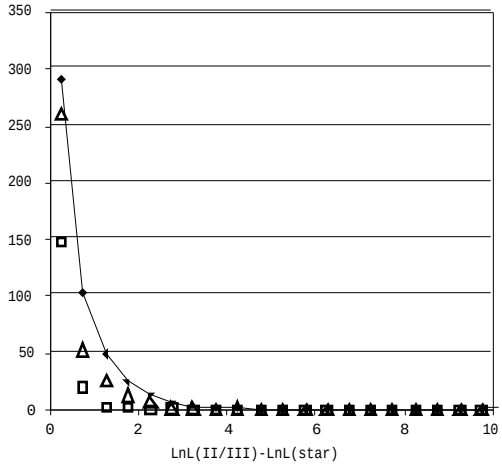
Actually, $(1/2)\chi^2_0+(1/2)\chi^2_1$ is not the correct asymptotic distribution. For most phylogenetic cases, a positive correlation between t_I and t_{II} is expected. See Section 2.4. Therefore, for large *n*, decreasing t_I to zero from its positive value at the overall maximum super-tree likelihood point, as is done to consider tree hypothesis II, will tend to also push t_{II} towards negative values. Thus, under tree hypothesis II, t_{II} will end up at the boundary zero more than half the time. Therefore, using the $(1/2)\chi^2_0+(1/2)\chi^2_1$ distribution is expected to be conservative (i.e., less powerful at rejecting the model under consideration than if one uses the true null distribution), and to get more conservative for larger sequence lengths *n* and for harder phylogenies, i.e., larger external branch length/internal branch length ratios.

These $(l_{II} - l_{star})$ and $(l_{III} - l_{star})$statistics are plotted for the four main simulated data sets in the histograms of Figure 12. Clearly, the $(1/2)\chi^2_0+(1/2)\chi^2_1$ distribution is over-conservative, especially for the data sets with larger *n*, as predicted.

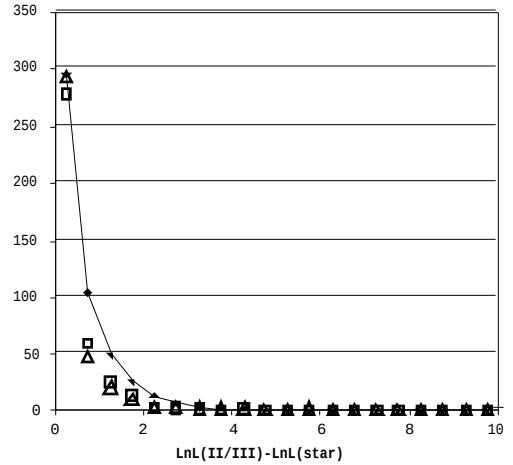
Like the $(l_{super\text{-}tree} - l_I)$ statistic, the $(l_{II} - l_{star})$ statistic should provide a sensitive test of the site-rate-variation (see also Figure 6F). In the simulated data set with site-rate-variation parameterised by a Γ distribution ($\alpha = 0.5$), the statistic has value $(l_{II} - l_{star}) = 3.2\pm2.6$, with expected value less than the 0.25 ± 0.43, when the site-rate-variation is not taken into account. See Table 3. Fitting α, however, yields the proper value $(l_{II} - l_{star}) = 0.06\pm0.21$, consistent with a distribution far more conservative than the predicted $(1/2)\chi^2_0+(1/2)\chi^2_1$ distribution. Also, for the simulated data set with unequal frequencies, taking into account $\pi_0 \neq 1/2$ decreases $(l_{II} - l_{star})$ from 1.6 ± 1.8 to 0.17 ± 0.50, with expected value less than 0.25 ± 0.43. Values for $(l_{III} - l_{star})$ are similar. So the test is moderately sensitive to parameters of the mutation matrix and very sensitive to site-rate-variation.



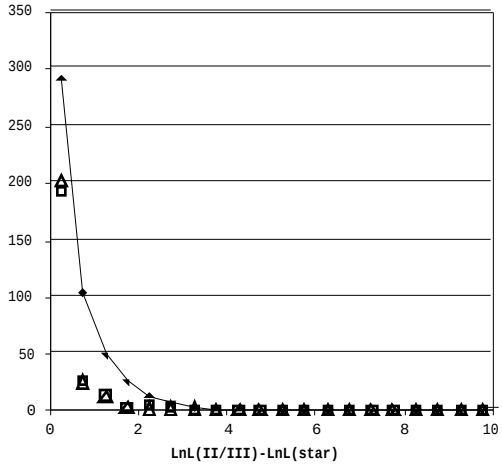
Four taxa, 100 characters



Four taxa, 500 characters



Six taxa, 500 characters



Six taxa, 5000 characters

Figure 12. Distributions of the $(l_{II} - l_{star})$ (squares) and $(l_{III} - l_{star})$ (squares) statistics compared to predicted $(1/2)\chi^2_0+(1/2)\chi^2_1$ distribution for simulations (1000 replicates) of the simplest binary model for the four-taxon and six-taxon trees shown in Figure 7. In these histograms, the bin at zero is not shown, as its value is an order-of-magnitude higher than for any other bins; for each of the four data sets, more than half of the 1000 replicates yielded $(l_{II} - l_{star})=0$ or $(l_{III} - l_{star})=0$.

(d) The statistic $(l_{max} - l_I)$

The $(l_{max} - l_I)$ provides a test that combines the $(l_{super\text{-}tree} - l_I)$ and $(l_{max} - l_{super\text{-}tree})$ discussed in (a) and (b) above; it is expected to be less sensitive than conducting tests (a) and (b) separately, but is more straightforward to carry out for general evolutionary models where $l_{super\text{-}tree}$ is difficult to find or not unique. In the asymptotic limit, $2(l_{max} - l_I)$ is expected to follow a χ^2_k distribution with

$$k = (\#\text{ site patterns}) - (\#\text{ tree parameters}) = (c^m - 1) - n_p,$$

where n_p is the number of optimised model parameters, including branch lengths, character frequencies, etc. This test statistic has been described before in e.g., (Navidi *et al.*, 1991), (Reeves, 1992), (Goldman, 1993), and (Yang, 1994b), but its predicted distribution depended on either assuming the infinite-sequence-length limit or conducting numerically intensive simulation.

For finite sequence length n, however, the distribution will not be this χ^2_k distribution, since all site patterns will generally not show up in the data – in fact, it is common to see data sets where the vast majority (sometimes >90%) of sites are "constant", the same for all taxa. Yang *et al.* (1995) have attempted a strategy of combining into larger categories those data points (site patterns) which have similar low probabilities, and then checking the data against prediction with a likelihood ratio or Pearson X^2 statistic. However, the method of Yang *et al.* (1995) seems to involve some bias in determining which site patterns to combine; so, in this report, the $(l_{max} - l_I)$ statistic (without such re-grouping of the data) is used.

To approximate the distribution of this statistic, note that the mean $(l_{max} - l_I)$ is expected to be rather large (exponentially growing with the number of taxa), and it is therefore valid to approximate its distribution as a normal distribution, by the Law of Large Numbers. The mean and variance of the normal distribution of $(l_{max} - l_I)$ can in fact be easily estimated by a procedure described in Appendix B, based on summing contributions from Poisson distributions for individual site patterns.

For the simulated data sets, the predicted $(l_{max} - l_I)$ distribution will simply be the convolution of the $(l_{max} - l_{super\text{-}tree})$ and $(l_{super\text{-}tree} - l_I)$ distributions, so the fit of the simulated statistic to the predicted asymptotic distribution, as well as the sensitivity to parameters of the mutation matrix, follows directly from the discussion above. See also the last column of Table 3.

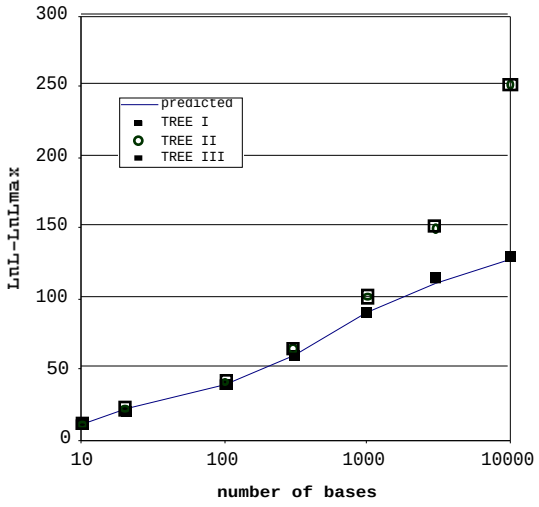


Figure 13. Comparison of predicted mean $(l_{max} - l_I)$ to actual $(l_{max} - l_{I/II/III})$ from 1000 simulated replicates, as a function of the number of bases in the sequence. Data is simulated with the four-taxon tree of Figure 7, and the JC69 model for nucleotides. See also Table 4.

n	Predicted ($l_{max} - l_I$)	Observed ($l_{max} - l_I$)	Observed ($l_{max} - l_{II}$)	Observed ($l_{max} - l_{III}$)
10	11.3±4.7	11.4±4.2	11.5±4.1	11.7±4.2
20	22.1±5.6	21.5±5.3	21.9±5.5	22.1±5.4
100	39.5±6.9	39.2±7.1	40.9±7.4	41.0±7.4
300	61.1±7.5	60.6±8.3	64.3±9.2	64.4±9.3
1000	89.7±8.5	89.6±8.7	102.1±9.6	102.2±9.6
3000	111.3±9.3	113.7±9.0	150.5±15.2	150.7±15.0
10000	127.1±10.3	129.6±9.7	251.4±21.1	251.4±21.1

Table 4. Predicted mean and variance of the statistic ($l_{max} - l_I$) for small number of characters n_{sites}, as determined by the one-time procedure of Appendix B, compared to observed values in 1000 replicates. A four-taxon tree (I in Figure 7) is used, with the JC69 four-nucleotide model.

It is useful to check whether the prediction of the mean and variance of ($l_{max} - l_I$)for finite *n* agrees with more complicated simulations with more than two character states, where all possible site patterns will not usually be present for moderate *n*. Figure 13 and Table 4 show the results of a simulation with four character states, using a JC69 model (Jukes and Cantor, 1969), with the four-taxon tree of Figure 7. The ($l_{max} - l_I$) found by the simple one-time procedure described in the appendix fits very well to that found by intensive multi-replicate simulation.

To summarise the results of this section, a set of consistency checks has been presented to test the adequacy of an evolutionary model which should be applied to molecular data before proceeding with phylogenetic inference by the ML procedure. Two tests, on the statistics ($l_{super\text{-}tree} - l_I$) and ($l_{max} - l_{super\text{-}tree}$), are based on the overall maximum super-tree likelihood value and have been shown to be especially sensitive to inadequacies in modelling site-rate-variation and mutation matrix parameters, respectively. Two more tests, on the statistics ($l_{II/III} - l_{star}$) and ($l_{max} - l_I$) do not depend on evaluating the super-tree likelihood and give good, but less sensitive, checks on the site-rate-variation and evolutionary parameters.

5. Real data sets

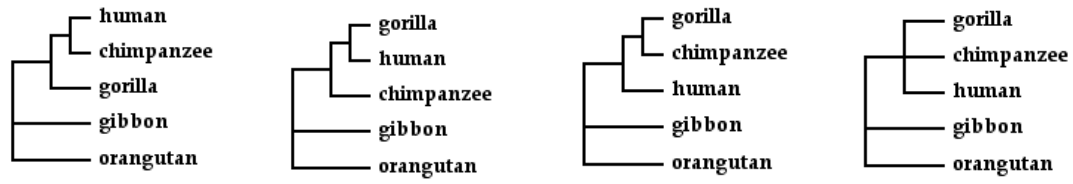
Before discussing the theoretical results introduced above, it is worth looking at some real data sets. Does ML analysis using common evolutionary models on real DNA data pass the "consistency checks" described in Section 4? Are the resulting likelihood-ratio-based supports reasonable? This section applies ML analysis to phylogeny inference in four data sets: a mitochondrial DNA (mtDNA) segment from five primates; α and β globin genes from five mammals; mtDNA and the *wingless* gene from almost sixty species (the genera *Heliconius* and *Eueides*, and outgroups) of passion-vine butterflies; and mtDNA from several mimicking races of *Heliconius erato* and *Heliconius melpomene* butterflies.

5.1. Mitochondrial DNA from five primates.

This data set is a "classic" in molecular phylogenetic studies, having been previously analysed by several authors, e.g., Yang *et al.* (1994b), and it is included with the PAML package. The data consists of an aligned 895-bp segment of DNA (including a tRNA gene and parts of two protein-coding genes) from human, chimpanzee, gorilla, orangutan, and gibbon (Brown *et al.*, 1982). The tRNA bases, and the first, second, and third codon positions are considered separate site categories, and are analysed as separate data sets. The three trees (plus a star tree) of Figure 14 are tested.

Before comparing likelihoods of the trees, the consistency checks described in Section 3.1 must be applied. Since estimating the value of $l_{\text{super-tree}}$ is difficult in this six-taxon case, only the ($l_{\max} - l_{\text{I}}$) and ($l_{\text{II/III}} - l_{\text{star}}$) tests have been performed. A succession of evolutionary models are tested, starting with the simplest JC69 model (equal base frequencies; no transition/transversion bias κ; no site-rate-variation), then F81 [allowing for unequal base frequencies, measured empirically; (Felsenstein, 1981)], then HKY85 (allowing for unequal base frequencies and a fitted $\kappa \neq 1$), and finally the more general HKY85+Γ and HKY85+3cat (modelling the site-rate distribution as a Gamma distribution with fitted shape parameter α, and as three discrete categories with fitted frequencies and rates, respectively).

To illustrate the tests, l_{I}, l_{II}, l_{III}, and l_{star} are shown in Table 5 for the third codon position bases, which might be expected to be a relatively "neutral" marker, since substitutions at the third codon positions are usually synonymous. Models JC69 and F81 are easily rejected (with 95% confidence) by the ($l_{\max} - l_{\text{I}}$) test, where the distribution of this statistic is approximated as a normal distribution with mean and variance as estimated in Appendix B; the HKY85 models (with and without site-rate-variation) cannot be rejected at 90% confidence by this test. Furthermore, the HKY85 and HKY85+3cat models are also rejected with 95% confidence by the ($l_{\text{II}} - l_{\text{star}}$) test, whose null distribution is (conservatively) taken to follow the $(1/2)\chi^2_0+(1/2)\chi^2_1$ distribution. Apparently site-rate-variation must be taken into account – and with an appropriately shaped distribution. The results for the other codon positions and the tRNA bases are similar to the results at the third codon position, except the ($l_{\text{II}} - l_{\text{star}}$) test is not able to reject any of the models for the slow-mutating second codon position bases.



Tree I Tree II Tree III Tree Star

Figure 14. Possible unrooted phylogenies considered in five-primate mitochondrial DNA set.

Model	No. of fitted evol. params.	Predicted l_{best} (±std. dev.)	l_I	l_{II}	l_{III}	l_{star}
JC69	0	–856.5±9.5	–1087.5** ††	–1093.2**	–1092.5** †	–1094.0*
F81	3	–855.0±9.5	–1026.6** ††	–1031.4**	–1029.4** ††	–1031.4*
HKY85	4	–854.5±9.5	–864.1††	–867.1††	–869.7*	–869.7*
HKY85+Γ	5	–854.0±9.5	–857.4††	–859.1	–860.5	–860.5
HKY85+3cat	9	–852.0±9.5	–858.4††	–858.4††	–858.7††	–863.0

(*) and (**) marks those values that are rejected by the ($l_{max} - l$) test (one-sided, Gaussian distribution) with 90% confidence and 95% confidence, respectively. The best tree is expected to pass this test.

(†) and (††) marks those values that are rejected by the ($l - l_{star}$) test [distribution of $2(l - l_{star})$ assumed to be $(1/2)\chi^2_0+(1/2)\chi^2_1$] with 90% confidence and 95% confidence, respectively. The alternatives to the best tree are expected to pass this test.

Table 5. Tests of various evolutionary models applied to third codon position nucleotides in five-primate mitochondrial DNA data set. Observed values of l_I, l_{II}, l_{III}, and l_{star}, compared to the predicted value for the best tree, based on the expected value of ($l_{max} - l_{best}$) estimated by the procedure in Appendix B, and the observed value l_{max} = –780.1. Results for other codon positions and tRNA nucleotides are similar.

Having performed these consistency checks of the evolutionary model, the relevant log-likelihood values under the HKY85+Γ model for the three codon positions and tRNA bases are given in Table 6. In addition the values are given for two analyses combining information from all four site categories. The first combination analysis SEP fits all parameters (base frequencies, κ, α, and branch lengths) separately for the four kinds of sites; and the second analysis, here called HOMOGENOUS, mimics the analysis of Yang *et al.* (1994b) by completely ignoring the codon/tRNA information about site categories, and by treating them all as one big, homogenous data set. See (Yang, 1996b). It is seen that none of the four site categories, when taken separately, disfavour tree I, and indeed the third codon position and tRNA bases both favour it with likelihood support values P^l(tree I) around 80%. There are thus no conflicting signals within the different rate categories – a reassuring observation.

Adding the separate likelihood values (the SEP analysis) then allows tree I to be well-supported with 97.3% probability. In the simpler HOMOGENOUS analysis, the likelihood values under HKY85+Γ both pass the ($l_{max} - l_I$) and ($l_{II/III} - l_{star}$) consistency checks, and the support for tree I becomes 96.6%. So barely any phylogenetic resolution is lost in ignoring site category information, in this case.

Data set	l_{max}	Predicted l_{best}	l_I	$l_{II}-l_I$	$l_{III}-l_I$	$l_{star}-l_I$	P^l(I)	P^l(II)	P^l(III)
First codon position	–571.2	–633.3±9.4	–635.0	+0.5	+1.0	+0.0	17%	32%	51%
Second codon position	–437.7	–458.7±5.7	–464.4	–0.3	–0.3	–0.3	40%	30%	40%
Third codon position	–780.1	–854.0±9.5	–857.4	–1.7	–3.1	–3.1	81%	15%	4%
tRNA	–424.4	463.9±7.8	–470.2	–2.5	–2.3	–2.3	85%	7%	8%
SEP	–2427.3	–2409.9±15.3	–2427.0	–4.0	-4.7	(–5.7)	97.3%	1.8%	0.9%
HOMOGENOUS	–2477.0	–2613.5±12.0	–2622.4	–4.0	–4.0	–4.0	96.6%	1.7%	1.7%

Table 6. Analyses of all codon positions in the HKY85+Γ model. See text for discussion of combined analyses. Note that the likelihood function for the HOMOGENOUS analysis is different from the others since it ignores information about site categories. Predicted values of l_{best} are based on the expected value of ($l_{max} - l_{best}$) estimated by the procedure in Appendix B, and the observed value l_{max}.

5.2. α/β-globin codons from six mammals.

A second data set analysed here is the combined α- and β- globin codon sequences from human, goat(α)/cow(β), rabbit, rat, and marsupial included with the PAML package. These globin genes are, of course, textbook examples in most fields of molecular biology, from structural genomics (Branden and Tooze, 1999) to biochemistry (Stryer, 1995). In phylogenetics, they have played an important role is the search for positive selection in genes (Yang *et al.*, 2000) and in controversies of the relation of birds to reptiles and mammals (Hedges, 1994).

The data set has been analysed in several ways: as one large DNA sets, with the HKY85+Γ model; at each codon position separately as three DNA data sets (HKY85+Γ); as an amino acid data set [with empirical transition matrices JTT, Dayhoff, mtREV24, and mtMAM included in the PAML package, modified to reflect the codon frequencies empirically determined from this data set]; and as a codon data set to be analysed with a model including various distributions for a non-synonymous/synonymous ratio parameter. There are fifteen (unrooted) bifurcating trees; Table 7 shows the results for the log-likelihood values of the eight trees of Figure 15, for each of these analyses. Apparently Tree III' is favoured by most of the analyses; but the true tree is known to be Tree I [see, e.g., (Janke *et al.*, 1997)]. However, none of the analyses pass the basic consistency checks, despite the generality of some of these models. That is, the best tree in any of these analyses is rejected by the ($l_{max} - l_I$) test, indicating that the models are not complex enough to adequately describe the data. See Section 5.2 for further discussion.

What would happen if one ignored the rejection of the model by the consistency checks and still analysed this data set using, e.g., the amino acids with the best-fitting JTT matrix? The marsupial branch would be more favoured to attach to the goat branch than to the correct rat branch by a factor of almost 100, based on likelihood-based support values, and by a factor of 5, based on RELL bootstrap support values. Thus, dangerously misleading phylogeny inferences can be made if one fails to test if the evolutionary model can be rejected by the ($l_{max} - l_I$) and ($l_{alternative} - l_{star}$) consistency checks.

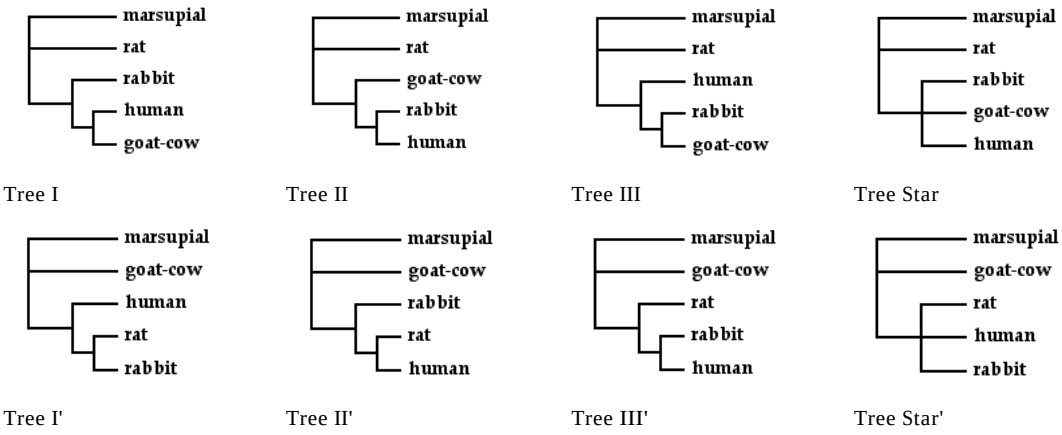


Figure 15. Eight (out of fifteen) possible unrooted phylogenies for which illustrative log-likelihood values are presented (see Table 4), for the five-mammal α/β-globin data set.

Data set	l_{max}	Pred. l_{best}	lI	lII	lIII	lstar	lI'	lII'	lIII'	lstar'
DNA:										
(no categorisation)	–2872.8	–3124.2±15.0	-3148.2**††	-3148.7**†	-3149.3**	-3150.7**	-3149.3**††	-3150.2**†	-3151.7**	-3151.8**
DNA:										
1st codon position	–790.8	–909.8±12.7	-934.3**††	-936.1**	-935.7**	-936.2**	-935.6**	-935.6**	-935.1**	-935.6**
2nd codon position	–685.8	–786.2±12.2	-810.7**	-810.9**	-808.6**	-810.9**	-807.5**††	-809.9**	-809.7**	-809.9**
3rd codon position	–1059.6	–1206.9±13.1	-1250.5**	-1248.6**††	-1250.5**	-1250.5**	-1252.0**	-1250.1**††	-1252.0**	-1252.0**
Amino acids:										
JTT	–1189.1	–1655.2±35.1	-1743.4**††	-1740.9**††	-1743.9**††	-1746.1**	-1733.0**††	-1738.2**††	-1739.9**††	-1742.5**
JTT+_	–1189.1	–1654.7±35.1	-1720.1**†	-1719.1**††	-1720.8**	-1721.7**	-1711.2**††	-1716.9**††	-1717.7**	-1718.2**
JTT+__ +empF	–1189.1	–1645.2±35.1	-1694.7**††	-1695.2**††	-1696.5**	-1697.5**	-1688.2**††	-1693.8**††	-1694.5**	-1695.5**
Dayhoff+__ +empF	–1189.1	–1645.2±35.1	-1682.1**††	-1681.6**††	-1683.0**	-1684.1**	-1676.1**††	-1680.8 **	-1682.0**	-1682.9**
mtREV24+__ +empF	–1189.1	–1645.2±35.1	-1714.1**†	-1713.8**††	-1714.6**	-1715.7**	-1706.3**††	-1711.6**†	-1711.8**	-1713.0**
mtMam+ Γ +empF	–1189.1	–1645.2±35.1	-1732.8**†	-1732.6**††	-1733.8**	-1734.6**	-1725.5**††	-1730.2**	-1730.8**	-1731.5**
Codons:										
4x3 base freqs (single ω)	–1493.9	–2694.4±52.0	-3055.0**††	-3051.5**††	-3053.7**††	-3060.5**	-3048.8**††	-3050.1**††	-3054.3**††	-3056.8**
61 empF (3cat for ω)	–1493.9	–2667.9±52.0	-2804.5**††	-2805.5**	-2806.4**	-2806.7**	-2800.7**††	-2805.5**	-2806.3**	-2806.3**

(*) and (**) marks those values that are rejected by the (lmax – l) test (one-sided, Gaussian distribution) with 90% confidence and 95% confidence,
respectively. The best tree is expected to pass this test.
(†) and (††) marks those values that are rejected by the (l – lstar) test [distribution of 2(l – lstar) assumed to be (1/2) _20+(1/2) _21] with 90% confidence and
95% confidence, respectively. The alternatives to the best tree are expected to pass this test.

Table 7. Tests of various evolutionary models applied to combined DNA, separate codon positions, encoded amino acids, and codons of five-mammal ___-globin data set. Observed values of lI, lII, lIII, and lstar, compared to the predicted value for the best tree, based on the expected value of (lmax – lbest) estimated by the procedure in Appendix B, and the observed values of l_{max}.

5.3. Mitochondrial DNA and *wingless* data for *Eueides* and *Heliconius* butterflies

The *Heliconius* butterflies and its close cousins are well-studied neotropical butterflies (*Lepidoptera: Nymphalidae*) which exhibit a wide-range of colourful wing patterns that advertise their unpalatability to predators. Geographic variability and inter-species mimicry have complicated phylogenetic analysis of these insects based on morphological and behavioural characteristics (Brower, 1997); see Figure 16. The *Heliconius* butterflies have played a central role in the theories of mimicry (Bates, 1862), the evolution of unpalatibility (Mallet and Gilbert, 1995), and the co-evolution of insects and plants [the cyanogen-containing passion vines, whose leaves the *Heliconius* caterpillars feed on, are poisonous to most other animals; see, e.g., (Murawski, 1983)]. The recent availability of short DNA sequences promises to answer many questions regarding the evolution of these insects, by clarifying their phylogeny.

This section investigates the relationship of *Eueides* and *Heliconius* using mitochondrial DNA and nuclear data (the gene *wingless*) collected by A. Brower for 59 butterfly species. Previous analysis of this data, using the heuristic maximum parsimony (MP) method, has led to confusing results. Brower (1994) originally claimed, based on analysis of the mtDNA sequences alone, that the *Heliconius* genus is paraphyletic to (inclusive of) *Eueides*, which is traditionally considered its sister genus (Brown, 1981). More recently, Brower and Egan (1997) retracted this controversial hypothesis, after combining mtDNA and *wingless* data to obtain a better resolved tree, again using maximum parsimony. Does an ML analysis clarify the position of *Eueides* with respect to *Heliconius*?

The data are 1105-bp and 378-bp sequences from the mitochondrial DNA (mtDNA) and the nuclear *wingless* (*wg*) gene, respectively; the mtDNA data contains the end of cytochrome oxidase I (COI) gene, the tRNA leu gene, and the start of the cytochrome oxidase II (COII) gene. For the likelihood analyses, sites which are ambiguous in some of the species are removed from the data; the resulting sequence lengths, as well as average nucleotide frequencies, fitted transition / transversion bias, and fitted site-rate-variation shape parameter α (for a Gamma distribution) obtained by PAML are given in Table 8.

The initial tree search on 59 species was carried out in the *fastDNAml* package, which is faster than PAML, but does not take into account site-rate variation. Figure 17 shows the ML tree obtained by *fastDNAml* (Olsen *et al.*, 1994a) assuming an HKY85 model with κ=85, using all of the mtDNA and *wingless* data.[13]

To evaluate the support for *Heliconius* being monophyletic with respect to *Eueides*, log-likelihood values (obtained separately for different codon positions in different genes, under the HKY85+Γ model) are given in Table 8 for this tree (called tree I) and the two trees obtained by forcing the *Eueides* clade to be the sister group of the *sara-sapho-erato* clade, and of the *melpomene-cydno-silvaniform-*"primitive" clade. These latter trees (called tree II and tree III respectively) are the ones obtained by nearest-neighbour interchange around the branch leading to the *Heliconius* clade. Collapsing this branch yields the "star" tree, for which log-likelihoods are also given in Table 8.

[13] The parameter settings for *fastDNAml* were the following: empirical frequencies (F); jumble (J) the taxon order for step-wise addition; transition / transversion bias (T) = 5; NNI rearrangements during taxon step-wise addition (G 1 1); outgroup (O) =Speyeria; and categorize (C) according to gene content. The relative rates of the ten different categories of sites (see Table 8) were taken to be (1, 0.4, 7, 0.4, 1, 0.4, 7; 0.7, 0.4, 4). A similar phylogeny (with some re-arrangement of long-branched groups like *demeter* and the outgroups, but the same high-level clades) is obtained with *fastDNAml* without categorization, by the MP analysis of Brower and Egan (1997), and also by the (heuristic) distance-based minimum evolution method implemented in PAUP* (Swofford, 1998) using a variety of distance estimators (an ML-based distance, the uncorrected P-distance, and the LogDet distance), so the topology appears to be quite robust to changes in the tree search method.

Figure 16. Mimicry rings amongst the Heliconius species of East Peru; figure from J. Mallet. Top row, with bright yellow hind-wing band, are *H. melpomene* and *H. erato*. Bottom ten species form a "rayed" mimicry ring dominating a separate geographical area [Left, *H. melpomene*, *H. elevatus*, *H. demeter*; centre, *Laparus doris*, *Neruda aoede*, *Eueides tales* and a pericopine moth; right, *H. erato*, *H. burneyi*, and *H. xanthocles*].

Site category	# bases	κ	α	rel. mut. rate	l_{max}	*pred.* l_I	l_I	$l_{II} - l_I$	$l_{III} - l_I$	$l_{star} - l_I$
COI 1st cod. pos.	41	1.9	0.2	1.0	–102.0	–267.7±34.0	–242.9	–2.4	–2.4	–2.4
COI 2nd cod. pos.	42	2.7	0.2	0.5	–77.5	–109.8±9.5	–109.8	–0.0	–0.0	–0.0
COI 3rd cod. pos.	41	33	0.6	9.5	–138.6	–702.2±51.4	–738.0	–2.0	–1.4	–2.0
tRNA leu	69	8.5	(0.5)	0.4	–146.5	–231.5±18.5	–237.5	–0.0	–0.0	–0.0
COII 1st cod. pos.	207	9.3	0.2	0.9	–553.2	–1186.0±104.7	–1071.4	–6.9	–6.9	–6.9
COII 2nd cod. pos.	207	2.5	0.1	0.3	–432.5	–669.2±57.0	–622.0	–0.0	–0.0	–0.0
COII 3rd cod. pos.	205	52	0.5	11	–978.0	–3967.6±119.0	–3992.4	–0.7	+3.3	–0.7
wg 1st cod. pos.	108	3.6	0.3	0.5	–273.9	–416.3±31.2	–451.7	–0.0	–0.0	–0.0
wg 2nd cod. pos.	109	1.8	(0.5)	0.3	–232.5	–344.9±19.4	–349.1	–0.0	–0.0	–0.0
wg 3rd cod. pos.	108	8.1	1.1	2.7	–468.0	–1481.0±85.1	–1551.4	–2.4	–2.4	–2.4

Table 8. Likelihood parameters for analysis of the position of the *Eueides* clade relative to the *Heliconius* butterflies; data set contains 59 species. For tree definitions, see Figure 15. Sequence lengths are given for data after ambiguous sites have been removed. The transition-transversion bias κ and the shape parameter α of the Gamma distribution for site rates have been fitted for each site category separately (with tree I) by PAML, except in the tRNA and *wingless* 2nd codon position data where the mutation rate was so low that α =0.5 was imposed to help the program converge. Mutation rates (relative to COI 1st codon position) were found by fitting tree I with all site categories simultaneously with the constraint that corresponding branch lengths for each site category were proportional The predicted values of l_I are based on the observed l_{max} for each site category and a predicted ($l_{max} - l_I$) obtained from simulations (ten replicates for each site category).

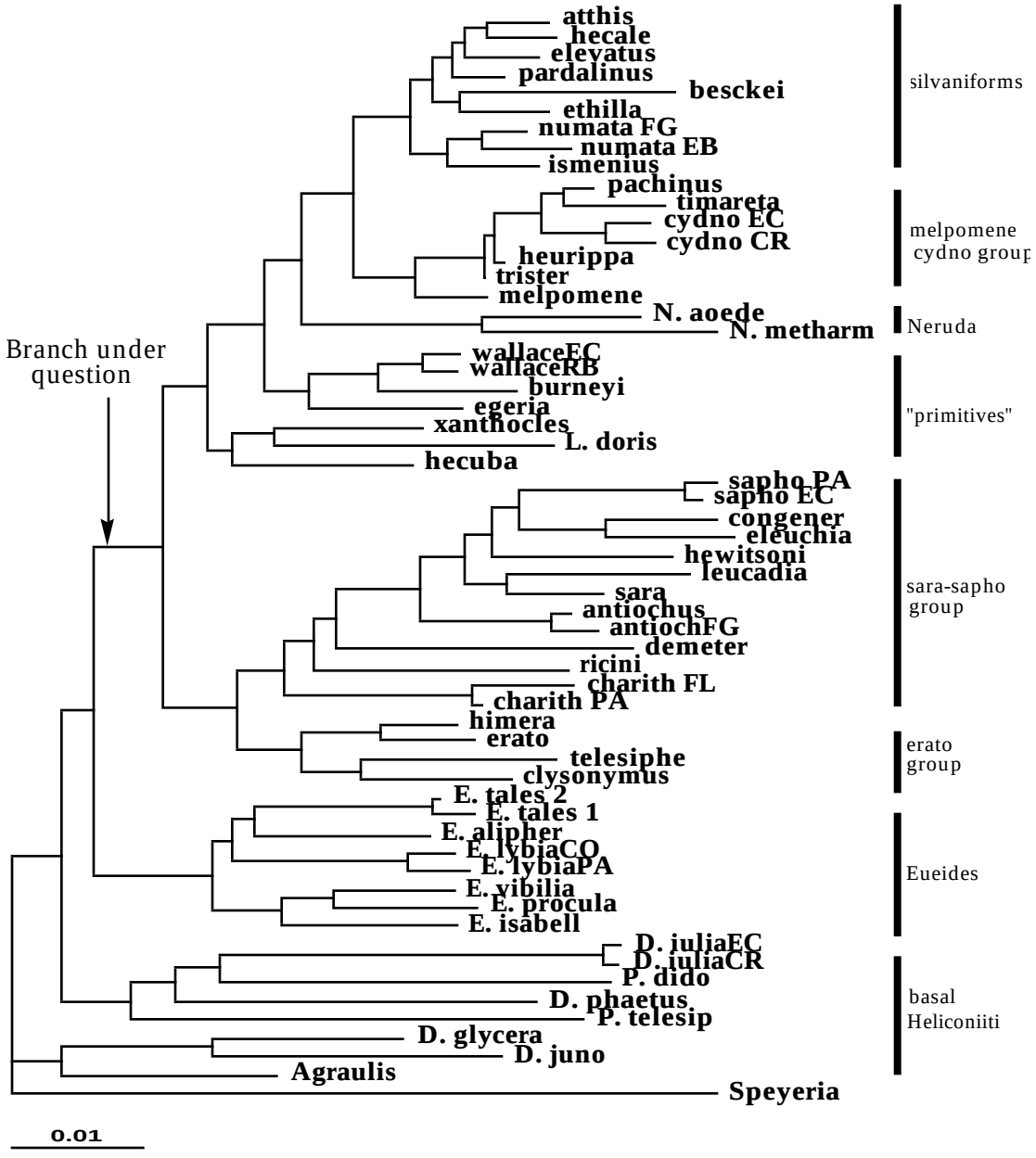


Figure 17. Tree (with ML branches) found by *fastDNAml* using combined mitochondrial DNA and *wingless* data. Clade / group labelling follows that of (Brower and Egan, 1997). The branch investigated in the text by ML analysis is the one leading to *Heliconius*, also marked on the figure. This tree is denoted as tree I; and the two trees obtained by forcing the *Eueides* clade to be the sister group of the *sara-sapho-erato* clade, and of the *melpomene-cydno-silvaniform-* "primitive" clade are denoted trees II and III, respectively.

For each codon position separately, the HKY85+Γ model cannot be rejected with 90% confidence by either the ($l_{max} - l_I$) or the ($l_{II/III} - l_{star}$) tests of Section 4. However, Table 8 does reveal a contradiction within the data: the COII third position data appears to strongly favour a different tree than the rest of the data set! Taking the likelihood values literally, the COII third codon positions prefer Tree II (where *Eueides* is the sister clade of *Heliconius melpomene-cydno-silvaniform*-"primitive" clade) over Tree I (where *Eueides* is outside *Heliconius*) with a likelihood ratio of 25. In contrast, using information from the first codon positions from the mtDNA data and the *wg* data, gives Tree I a likelihood higher by a factor of 10^4 and 10, respectively, over each of the alternative trees.

It is not difficult to guess a reason for this contradiction. At the third codon positions, DNA sequences from different species have rather unbalanced nucleotide compositions. The biological selection process that is responsible for the low G content is not well understood [see, e.g., (Rodríguez-Trelles *et al.*, 2000)], and may be much more complicated (involving, say, dramatic variation in selection pressure over time) than in the simple HKY85 mutation matrix. The HKY85+Γ model used above assumes, however, that the nucleotide composition has the same equilibrium value throughout the tree. The PAML package does have an option (*nhomo* = 4) to fit different nucleotide frequencies at different branches of the tree. However, the program did not converge properly when this option was used; different initial conditions and different values of the *smalldiff* parameter led to different optimisations.

Thus, the maximum likelihood analysis, using the HKY85+Γ model, while passing the consistency checks of Section 4, still appears to give contradicting results at different codon positions in the mtDNA data. There is thus still some uncertainty in applying ML to determine the relation of *Eueides* and *Heliconius* – although maximum likelihood does seem preferable to the maximum parsimony analyses of (Brower, 1994) and (Brower and Egan, 1997), which do not take into account site-rate-variation or the transition / transversion bias. A definitive analysis awaits a more sophisticated likelihood program which can smoothly take into account variation in nucleotide frequencies across lineages for the third codon position data. For now, however, one can say that the weight of the evidence is in favour of *Heliconius* being monophyletic with respect to *Eueides*. In particular, if the third codon position data are ignored, the mtDNA and the *wingless* data sets separately (and together) favour this topology with strong confidence.

5.4. Mitochondrial DNA data applied to race formation in *Heliconius erato* and *Heliconius melpomene*

Brower (1996) has also collected mtDNA data for mimicking races of *Heliconius erato* (53 specimens) and *Heliconius melpomene* (44 specimens). Based on a maximum parsimony analysis, he found an unexpected result: sympatric sub-species which exhibit remarkably similar wing-pattern phenotypes do not appear to have shared a common recent evolutionary history. Moreover, the parsimony mtDNA cladogram shows evidence for *H. melpomene* to be paraphyletic with respect to the species *H. cydno*. Are these phylogenetic results also supported by an ML analysis?

The mtDNA segment for the *H. melpomene* and *H. erato* data sets is the same as for the more general *Heliconiinae* of the previous section. For a description of the wing patterns and geographic location of the sub-species, see (Brower, 1996). For the tree search, the program *fastDNAml*[14] was used to find initial topologies for the *erato* and *melpomene* data sets; then, the nearest-neighbour tree search with PAML was carried out, taking into account site-rate variation and imposing a molecular clock[15]. Since the divergence times relevant to race formation are short, the slow-mutating second codon position and tRNA data are not phylogenetically informative and are not analysed here; the data at first and third codon positions (for COI and COII genes combined) are analysed separately. The assumed transition / transversion bias parameters, gamma site-rate-variation parameters, and final log-likelihoods (with and without the molecular clock assumption) are given in Table 9. The trees, with ML branch lengths, are shown in Figure 17.

Before studying these phylogenies, the consistency checks of Section 4 need to be applied to the data set to check the adequacy of the assumed evolutionary model; see Table 9. Interestingly, the assumed HKY85+Γ model is rejected with 95% confidence when applied with these trees to the third codon position data in the *erato* analysis, and to the first codon position data in the *melpomene* analysis. The problem in the former case is probably related to the very low G content in that data, as discussed in the previous section. For now, the likelihood ratios will still be interpreted as literal (but heuristic) measures of posterior odds; but when a better understanding of the mutation processes for unbalanced nucleotide composition is available, the *erato* data should be re-analysed. The problem in the latter case, the first codon position of the *melpomene* data, may have to do with selection effects (first codon position substitutions are more often non-synonymous than third codon positions). The problem is not so drastic, however – removing the first codon positions does not significantly change the phylogeny obtained or its statistical evaluation.

Statistical evaluation of the tree is accomplished in terms of likelihood ratios. In particular, for each branch in the ML tree (call its likelihood L_{I}), the likelihoods L_{II} and L_{III} are obtained for each of the two trees II and III obtained by nearest-neighbour interchange (NNI) around the given branch. An estimate of the stability of that branch is then given by $P_{\mathrm{NNI}} = L_{\mathrm{I}} / (L_{\mathrm{I}}+L_{\mathrm{II}}+L_{\mathrm{III}})$. So, for example, if this value is near 33% the branch is not well resolved, as nearest-neighbour interchanges produce trees with the same likelihood. When P_{NNI} is larger than 60% for a branch, it means that the branch is fairly well-supported – the likelihood of the ML tree is better by a factor of 3 than each of the alternative NNI trees. Figure 18 shows P_{NNI} values, based on likelihoods obtained with and without the molecular clock assumption, for such well-supported branches in the *erato* and *melpomene* trees.

Comparison of the trees in Figure 18 with the cladogram in (Brower, 1996) – with respect to placement of outgroups, to definition and arrangement of race clades, and to resolution within the clades – shows that this maximum likelihood analysis gives better resolution than the maximum parsimony analysis.

---

[14] Settings for *fastDNAml* are the same as in the previous section.

[15] The molecular clock was useful particularly in resolving the arrangement of the outgroups in each data set. A molecular clock has not been imposed in previous sections because the clock assumption can be rejected for the previous data sets, based on a likelihood ratio test [see, e.g., (Yang, 1996b)]. In the *erato* and *melpomene* data sets in this section, however, imposing the molecular clock constraint only increases the log-likelihood l for third codon positions by 22.9 and 18.5, respectively, with expected values being 24.5±5.0 and 21.0±4.6 – this data (the first codon positions give similar results) appears consistent with the molecular clock assumption.

H. erato data set

Site category	# bases	κ	α	rel. mut. rate	l_{max}	*pred.* l_{tree}	l_{tree}
COI+COII 1st cod. pos.	274	65	0.1	1	–498.9	–572.2±18.0	–570.7
COI+COII 3rd cod. pos.	272	65	0.5	10	–747.3	–1105.0±49.5	–1260.6

H. melpomene data set

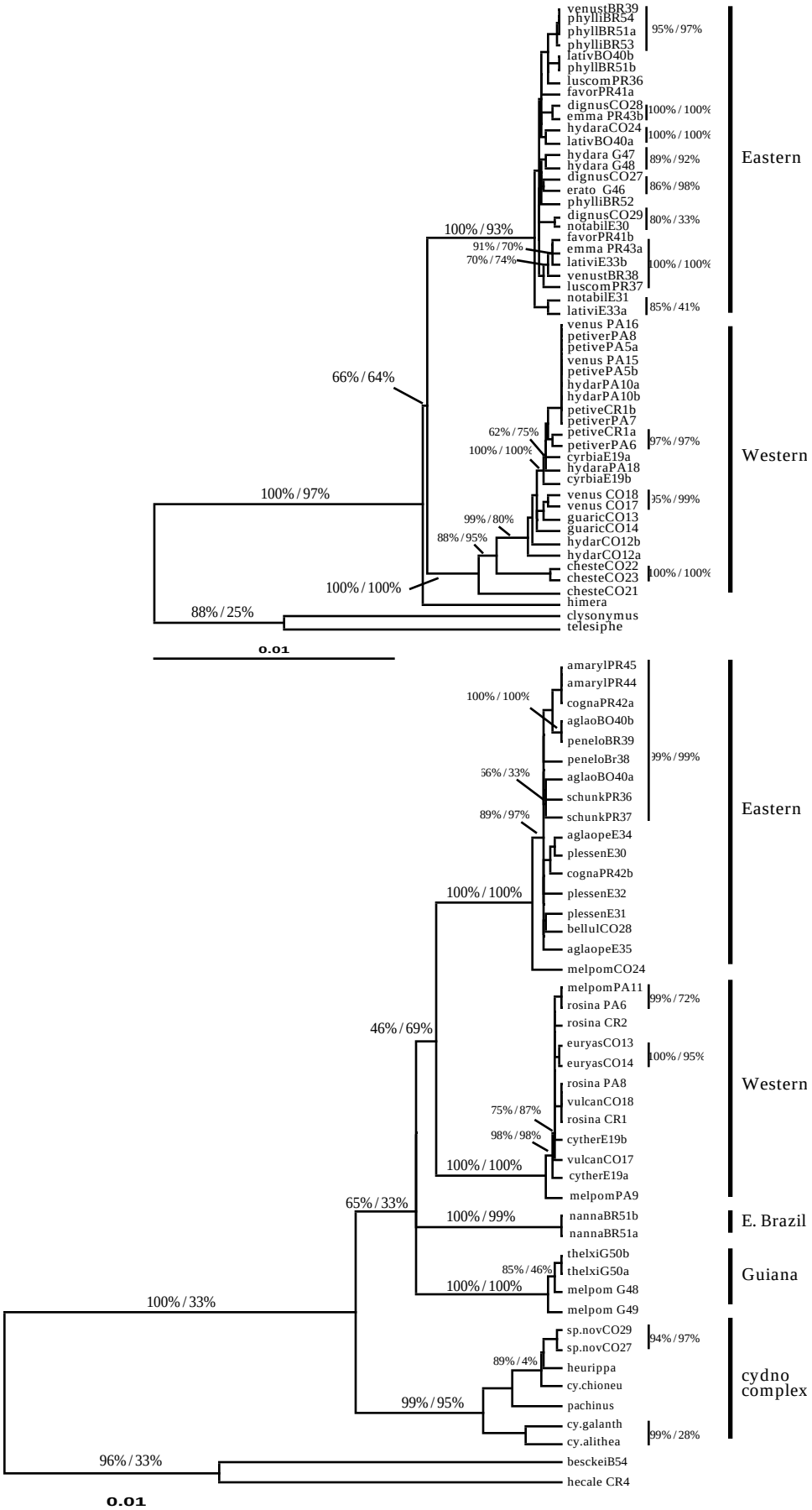
Site category	# bases	κ	α	rel. mut. rate	l_{max}	*pred.* l_{tree}	l_{tree}
COI+COII 1st cod. pos.	277	65	0.1	1	–469.6	–509.3±12.4	–538.9
COI+COII 3rd cod. pos.	276	65	0.5	10	–660.7	–938.6±39.2	–1002.3

Table 9. Likelihood analysis for *H. erato* (53 species) and *H. melpomene* (44 species) data sets. For tree definitions, see Figure 16. Sequence lengths are given for data after ambiguous sites have been removed. The assumed values of κ (transition / transversion bias) and α (shape parameter for Gamma distribution of site rates) are those fitted to the *melpomene* data set; in fact, doubling or halving these values for either *melpomene* or *erato* data set barely changes the fit ($\Delta l < 1$), because the divergence times are so low. The slow-mutating second codon position and tRNA data are not phylogenetically informative and are not analysed here; the data at first and third codon positions (for COI and COII genes combined) are analysed separately. The predicted values of l_i are based on the observed l_{max} for each site category and a predicted ($l_{max} - l_i$) obtained from simulations (ten replicates for each site category).

Firstly, the putative outgroups in the ML analysis are indeed put outside the *H. erato* and *H. melpomene* clades. In particular, the ML tree supports the traditional view of putting *H. himera* outside the *H. erato* races with fairly good confidence (P_{NNI} = 66% and 64%, with or without the molecular clock assumption, respectively); the position of *himera* was not resolved in the MP analysis. More strikingly, the molecular clock assumption allows the *H. cydno* complex to be placed outside *H. melpomene* (P_{NNI} = 65%), also the traditional assumption; the MP cladogram of Brower (1996) made *melpomene* paraphyletic to *cydno*.

Secondly, the memberships of the race clades are very well-defined. In *H. erato*, an eastern clade (containing butterflies from Bolivia, Brazil, Peru, Guiana, and parts of Ecuador and Colombia) and a western clade (containing butterflies from Panama, Costa Rica, and other parts of Ecuador and Colombia), are very well-resolved (P_{NNI} =100% and >90% for clade monophyly, with and without the molecular clock assumption, respectively). Brower (1996) found the same clades – geographically separated by the Andes – except that he was not able to determine the placement of the *H. erato chestertonii*. In contrast, the ML analysis places the *chestertonii* as the basal sub-species in the western *erato* clade with strong confidence (P_{NNI} >80% for relevant branches). In *H. melpomene*, the ML analysis produces the same clades (with strong confidence P_{NNI} =100%) as the MP analysis: an eastern and western clade, as well as separate clades for the two *nanna* sub-species in the sample from Southeast Brazil and for the four Guiana butterflies. The ML analysis, like the MP analysis, does not provide much information about the relative arrangement of these four *H. melpomene* clades; obtaining data from more specimens, especially from the rather under-sampled Amazonian basin area, would help to resolve this higher level of the tree.

Figure 18. (next page) Trees found by PAML (with ML branches under molecular clock assumption) for the *H. erato* (above) and *H. melpomene* (below) data set. Labels of *erato* and *melpomene* races are given as abbreviation of race name, country of origin [BO=Bolivia; BR=Brazil; CO=Colombia; CR=Costa Rica; E=Ecuador; G=French Guiana; PA=Panama], and locale number [see (Brower, 1996)]. Branches are statistically evaluated in terms of $P_{NNI} = L_I / (L_I + L_{II} + L_{III})$ (where II and III are trees obtained by nearest-neighbour interchange around a given branch); these support values are given on the trees (with / without molecular clock assumption) for well-supported branches. The taxon *H. sara* has been left out of the top tree for clarity; its position is external to all the other taxa (P_{NNI} = 100% / 33%).


venustBR39
phylliBR54
phyllBR51a
phylliBR53
lativBO40b
phyllBR51b
luscomPR36
favorPR41a
dignusCO28
emma PR43b
hydaraCO24
lativBO40a
hydara G47
hydara G48
dignusCO27
erato G46
phylliBR52
dignusCO29
notabilE30
favorPR41b
emma PR43a
lativiE33b
venustBR38
luscomPR37
notabilE31
lativiE33a
venus PA16
petiverPA8
petivePA5a
venus PA15
petivePA5b
hydarPA10a
hydarPA10b
petiveCR1b
petiverPA7
petiveCR1a
petiverPA6
cyrbiaE19a
hydaraPA18
cyrbiaE19b
venus CO18
venus CO17
guaricCO13
guaricCO14
hydarCO12b
hydarCO12a
chesteCO22
chesteCO23
chesteCO21
himera
clysonymus
telesiphe
95% / 97%
100% / 100%
100% / 100%
89% / 92%
86% / 98%
80% / 33%
100% / 100%
85% / 41%
97% / 97%
95% / 99%
100% / 100%
Eastern
Western
100% / 93%
91% / 70%
70% / 74%
66% / 64%
62% / 75%
100% / 100%
100% / 97%
99% / 80%
88% / 95%
100% / 100%
88% / 25%
0.01
amarylPR45
amarylPR44
cognaPR42a
aglaoBO40b
peneloBR39
peneloBr38
aglaoBO40a
schunkPR36
schunkPR37
aglaopeE34
plessenE30
cognaPR42b
plessenE32
plessenE31
bellulCO28
aglaopeE35
melpomCO24
melpomPA11
rosina PA6
rosina CR2
euryasCO13
euryasCO14
rosina PA8
vulcanCO18
rosina CR1
cytherE19b
vulcanCO17
cytherE19a
melpomPA9
nannaBR51b
nannaBR51a
thelxiG50b
thelxiG50a
melpom G48
melpom G49
sp.novCO29
sp.novCO27
heurippa
cy.chioneu
pachinus
cy.galanth
cy.alithea
besckeiB54
hecale CR4
99% / 99%
99% / 72%
100% / 95%
99% / 28%
94% / 97%
Eastern
Western
E. Brazil
Guiana
cydno complex
100% / 100%
56% / 33%
89% / 97%
100% / 100%
46% / 69%
75% / 87%
98% / 98%
100% / 100%
65% / 33%
100% / 99%
85% / 46%
100% / 100%
100% / 33%
89% / 4%
99% / 95%
96% / 33%
0.01

Finally, the radiation of races into eastern and western *erato*/*melpomene* clades has occurred rapidly and recently, so full resolution within each clade would require much longer segments of mtDNA to be sequenced for each sub-species; but a few phylogenetic statements relevant to intra-clade relationships can be made. On one hand, like the MP analysis of Brower (1996), the ML analysis is able to pair some of the *erato* taxa (e.g., an *emma* specimen with a *dignus* specimen in the *erato* eastern clade) and *melpomene* taxa (e.g., a *penelope* specimen with an *aglaope* specimen in the *melpomene* western clade). Less trivially, ML finds *H. melpomene melpomene* specimens to be basal to both the eastern and western *melpomene* clades, also like MP. On the other hand, the ML analysis provides a few intriguing insights into the internal topologies which have been missed by the MP analysis. In the *erato* western clade, the sub-species from Panama, Costa Rica, and Ecuador are grouped by the ML analysis into a strongly supported subclade (P_{NNI} =100%, with or without the molecular clock assumption) to the exclusion of the butterflies from Colombia. Also, in the *melpomene* western clade, the sub-species from Brazil, from Bolivia, and from most of Peru are grouped into a subclade.

This additional, unexpected geographic substructure within the race clades, combined with the new information on *chestertonii*, *cydno*, and *himera*, discussed above, brings up new evolutionary questions which were not brought up in the MP analysis of Brower (1996). For example, what historic event might have caused a subclade of the western *erato*s to spread away from Colombia into the peripheral regions in Ecuador and Panama? The divergence time for this subclade is remarkably similar to the divergence time for the western *melpomene* clade. Could the western *melpomene* radiation have somehow impelled the *erato* subclade radiation, or were both radiations catalysed by some external geographic event? Also, regarding wing pattern evolution, *H. erato hydara* is found in both eastern and western *erato* clades, but *H. erato chestertonii* – with an incredibly different wing pattern – is basal to the western clade. So was the primal wing pattern of *chestertonii* similar to that of *hydara*? Or is the mtDNA phylogeny, being solely a representation of matrilineal heritage, not completely relevant to the complex evolution of wing patterns, which requires an understanding of the nuclear genes and recombination? Answering such questions is well beyond the scope of this essay – but they demonstrate how a maximum likelihood tree search, statistically evaluated by comparing the likelihoods of nearest-neighbour trees, can squeeze out more information from a given data set than a maximum parsimony analysis.

6. Discussion

6.1.Theoretical basis for the ML method in molecular phylogenetic inference.

This report has presented an extension of the usual likelihood function for a tree describing a molecular data set. This super-tree likelihood is a function of the lengths of all possible taxon bipartitions. It is defined to have the desirable property that for a given tree, the super-tree likelihood function reduces to the usual likelihood function (Felsenstein, 1981) under that tree hypothesis when bipartitions not in the tree are set to zero. Thus, theoretical uncertainties that appear to arise in the usual ML method – where the likelihood function appears to take a different form for different tree topologies (Nei, 1987) – are not a problem from the super-tree perspective. In particular, the phylogeny estimation problem is seen to be one of the traditional form of choosing between composite hypotheses, given a likelihood function. The solution is to maximise the likelihood over each tree hypothesis: the usual ML method. It should be noted that for general models, the super-tree extension of the likelihood function is not unique or necessarily simple (as it is for the simplest binary model), but it is always definable – and the fact that it exists in principle provides the theoretical basis for the usual ML procedure.

The resulting likelihood value for each tree (normalised to the sum of likelihood values over all considered trees) gives an estimate of confidence P^{l}(tree) in the tree which behaves intuitively, as has been shown by numerical simulation. This estimate of support – unlike the bootstrap support value or, in some cases, the MAP value in the Bayesian analysis of Rannala and Yang (1996) – does not give a misleadingly high value in the cases where the wrong tree happens to be the ML tree.

Why then have most authors avoided directly interpreting likelihood ratios as giving quantitative estimates of posterior odds of different trees? The validity of the likelihood analysis depends on the accuracy of the assumed evolutionary model. If, for example variation of the mutation rate at different sites is not taken into account, the log-likelihood differences between different trees tends to be exaggerated (see, e.g., Table 5, for primate mtDNA data under the simplest JC69 model). The likelihood-based support value in such an analysis would therefore be skewed near 100% for the maximum likelihood tree, even if it is the wrong tree. It is no wonder then that the more conservative support measures, like $\exp[\Delta l_{12}/\sigma(\Delta l_{12})]$ [where $\sigma(\Delta l_{12})$ is the estimated variance of the log-likelihood difference; see (Jermiin *et al.*, 1997)] and the bootstrap support have been proposed as support indicators in such incomplete analyses. As this report has shown, the likelihood-based supports are more appropriate (and easier to compute) than these alternative, possibly misleading measures, as long as the assumed evolutionary model is indeed a correct description. To check the model's validity, some basic consistency checks have been presented in this report, two of which [comparison of $(l_{max} - l_{best})$ and $(l_{alternative} - l_{star})$ to a normal distribution estimated as in Appendix B, and to a $(1/2)\chi^2_0+(1/2)\chi^2_1$ distribution, respectively] require little extra computation.

6.2. What if the analysis does not pass a consistency check?

For rather general evolutionary models, none of the DNA-based, amino-acid-based, or codon-based analyses of the α/β globin data set of Section 4.2 pass both the $(l_{max} - l_{best})$ and the $(l_{alternative} - l_{star})$ consistency checks, even for the third codon position nucleotide data. The fact that the likelihood ratio tests are actually quite conservative tests [the statistics $(l_{max} - l_{super})$ and $(l_{super} - l_{best})$, which are more difficult to evaluate, provide more powerful checks; see Section 4.] means that the data is not described at all well by the evolutionary model – it would be dangerous to proceed with the phylogeny inference.

Why would the data fail the consistency check? Several possible problems can be identified, including use of wrong transition matrix, the presence of different selection pressures on different lineages, recombination, and correlations of ("compensating") site

mutations. In the first case, the fit can be improved with a more appropriate transition matrix. For example, one might use structural information to find different matrices for, e.g., trans-membrane helices and hydrophilic polar residues. See, e.g., (Liò *et al.*, 1998); (Liò and Goldman, 1999) for promising approaches.

In the second case, where selection pressures are different among different lineages, the fit might be somewhat improved by allowing non-synonymous substitutions to become more favourable along different lineages. The best bet, however, might be to guess which sites are under strongest selection pressure, and to either throw them out of the phylogenetic analysis, or to somehow carefully model the selection process. For example, the third codon positions of the α/β globin data set appear to have statistically distinguishable nucleotide frequencies in the different taxa; if one can understand the cause of this variance, it could be included in the likelihood model.

In the third case, recombination, the analysis will fail to resolve short internal branches corresponding to fast radiations of species. The data set would then be expected to fail the consistency check if it contains separate segments that support different trees (corresponding to the different ways genes have fixed after being shuffled around by proto-species during the radiation). A better analysis might then be to divide the data into physically separate genes, and to compare the ML trees from each gene.

Finally, most ML analyses assume that mutations at different sites are independently distributed (i.d.) of each other; but for protein-coding and RNA genes this is usually not the case, as mutations often occur in tandem to prevent distortion of the three-dimensional structure or sabotaging the chemistry of the relevant protein/RNA molecule. In such cases where several sites are under selection to change simultaneously, the i.d. assumption gives too much credence to several mutations which should really all be considered a single "event". To avoid this danger, one might hope to incorporate structural information or other such prior knowledge in designing the likelihood model that accounts for correlation of mutations at separate sites.

6.3. Comparison of super-tree function to "phylogenetic networks".

Strimmer and Moulton (2000) have also recently given a generalisation of the likelihood function for evolutionary trees to a more general "phylogenetic network" as a directed graphical model. Despite sharing some superficial similarity with the super-tree model presented here [for example, Figure 5c in their paper is identical to Figure 5 in this report], the approaches are quite distinct.

In the phylogenetic network approach, a graphical "network" model containing the full set, or a subset, of bipartitions is drawn. It is rooted at a prescribed node, and all edges in the tree are given a direction pointing away from the node. Then, the usual calculation of site pattern probabilities [see, e.g., eq. (10)] needs to be modified, since now each node may have more than one "parent". In particular, Strimmer and Moulton (2000) specify that the probability $P(x|y,z)$ of observing a state x given parents y and z separated by edges t_y and t_z, respectively, to be $p_y P_{yx}(t_y) + p_z P_{zx}(t_z)$, where p_y and p_z are pre-specified "prior" probabilities.

As a simple example, take the four taxon tree, but with the constraint $t_A = t_B = t_C = t_{III} = 0$, with the simplest binary model. The super-tree approach would give the probability of observing the same character at a given site in all four taxa as [see eq. (13)]:

$$p_{\text{same}}^{(\text{super-tree})} = \frac{1}{8}\left(1+e^{-2t_{\text{I}}}\right)\left(1+e^{-2t_{\text{II}}}\right)\left(1+e^{-2t_{\text{D}}}\right) \qquad (24)$$

In the approach of Strimmer and Moulton, if the root is specified at the internal node closest to with B, C, or D (see Figure 5), the site pattern probability for the phylogenetic network will be the same as the expression above. If the root is taken at node A, however, one obtains:

$$p_{\text{same}}^{(\text{network})} = p_{\text{same}}^{(\text{super-tree})} + \frac{1}{8}\left[p_{\text{recomb}}\left(1+e^{-2t_{\text{I}}}\right)\left(1-e^{-2t_{\text{II}}}\right)+\left(1-p_{\text{recomb}}\right)\left(1-e^{-2t_{\text{I}}}\right)\left(1+e^{-2t_{\text{II}}}\right)\right]\left(1-e^{-2t_{\text{D}}}\right), \qquad (25)$$

where p_{recomb} is some pre-specified prior probability of the sequence flowing from the root through I and then II. Strimmer and Moulton suggest interpreting p_{recomb} as a "recombination frequency". The differences between the phylogenetic network and super-tree approaches is now clear. Firstly, the phylogenetic network approach is not unique, since it requires the specification of a root, and a set of prior probabilities like p_{recomb}. In contrast, the super-tree likelihood is independent of rooting (for reversible models) and of the values of any new parameters.

Secondly, the question arises: how should one specify the root and p_{recomb} in the phylogenetic network model? Strimmer and Moulton (2000) suggest the choice of a root which produces a maximum likelihood; however the likelihood function in the different-root cases will be different, so the comparison of the different rooting hypotheses is questionable, for the very reason that the usual ML method of phylogeny estimation was questioned (Yang *et al.*, 1995). Even if the root can be specified (for example, by a known outgroup), determining p_{recomb} poses a harder problem. For the simplest binary model, there are exactly as many degrees of freedom (assuming equal character frequencies) as there are possible bipartitions. Therefore, there is no way to constrain p_{recomb} by the data in the simplest binary model! [Even for more general models, it is clear that attempting to fit all the extra p_{recomb} parameters (which is Strimmer and Moulton's approach in their sample data set) will dramatically increase the variance in the estimated branch length parameters.] On the other hand, in the super-tree approach, there are no extra parameters or necessary rootings.

Finally, the phylogenetic network and super-tree methods approach the limit of usual evolutionary tree hypotheses in different ways. On one hand, in the equation (25) above, both constraints $t_{\text{II}} = 0$ and $p_{\text{recomb}} = 1$ are required to obtain Tree topology I; the constraints $t_{\text{I}} = 0$ and $p_{\text{recomb}} = 0$ give tree topology II. Note that the likelihood functions for Tree I and Tree II have different parameter values for p_{recomb}; thus the phylogenetic network model provides no theoretical justification for comparing tree likelihoods. Statistical evaluation of Strimmer and Moulton's approach relies on simulation, or RELL bootstrapping (Kishino and Hasegawa, 1989). On the other hand, in the super-tree method, one simply sets all bipartition lengths not in a given tree to zero. So, in the example of (24), one sets $t_{\text{II}} = 0$ or $t_{\text{I}} = 0$ to get the Tree I or Tree II topology, respectively. Thus, the same form of the likelihood function with all the same evolutionary parameter values are being compared in the different trees, and the theoretical basis for likelihood methods in tree selection is clear. In particular likelihood ratios between different trees can be properly interpreted as posterior odds of the trees.

6.4. Comparison of ML to other phylogeny inference methods.

Having established that the likelihood is theoretically justified in its application to phylogeny inference, other existing methods seem less desirable. Here, the methods of parsimony, minimum evolution, and spectral analysis are described in terms of consistency and statistical evaluation.

Maximum parsimony is known to be inconsistent for trees with long branches separated by internal branches; and corrected versions of parsimony do not have a theoretical basis, aside from a vague invocation of Occam's razor. Statistical evaluation usually relies on bootstrapping, or Bremer/decay supports (Bremer, 1998) measured in integer units, but it is not clear how to generally interpret these values.

Distance-based methods may be more promising for phylogenetic inference based on non-molecular data. The theoretical basis for the "minimum evolution" procedure has been established for the problem of phylogeny inference based on a pair-wise distance matrix with statistically independent entries (as would be obtained from DNA hybridisation experiments). See (Rzhetsky and Nei, 1993). This distance-based theory, however, has not been extended to phylogeny inference based on comparison of known molecular sequences. Again, statistical estimation usually relies on the hard-to-interpret bootstrapping, or testing the positivity of internal branch lengths, which is known to be liberal (Sitnikova *et al.*, 1995).

Finally, the spectral analysis (Hadamard conjugation) method (Hendy *et al.*, 1994; Waddell *et al.*, 1997) has some similarity with the super-tree perspective in that it

produces length estimates for all possible internal bipartitions. However, the Hadamard transform used in that method does not extend to more general evolutionary models with, e.g., unequal character frequencies. Moreover, in spectral analysis, heuristic methods, like parsimony or a "closest-tree" fit, are applied to the bipartition length estimates to yield a tree; and also, statistical evaluation, except via a heuristic measure of "conflict", bootstrapping (Lento *et al.*, 1995), or extensive numerical simulation (Waddell *et al.*, 1994), appears difficult.

Thus, ML phylogeny inference appears to be the only existing method with a firm theoretical basis in addition to an easy-to-measure estimate of statistical confidence (the likelihood ratio) that conforms to intuition. But maximum likelihood does have its disadvantages. Firstly, for large numbers of taxa and for more general evolutionary models, ML requires heavy computational power. However, modern programs like *fastdnaml* (Olsen *et al.*, 1994a) take less than half an hour on modern workstations to estimate the overall maximum likelihood tree for data sets as large as 60 taxa (1000bp DNA sequences) via an uphill climb with an assumed HKY85 model, as shown in Section 5.4 for *Heliconius* race formation data. The guess can then be refined by, e.g., checking nearest-neighbour trees with more general evolutionary models implemented in PAML. Evaluation of the relevant star trees, and the maximum possible value of the likelihood is also quick; it is therefore straightforward to perform the consistency checks based on the $(l_{max} - l_{best})$ or the $(l_{alternative} - l_{star})$ statistics. Then comparison of likelihood ratios of the best tree with, say, the trees obtained by nearest neighbour interactions, allows for likelihood-based support values to be assigned to each internal branch.

Possibly the biggest problem with ML (and indeed with the other phylogenetic inference methods available) is our incomplete understanding of the basic biological processes involved in molecular mutation, as is evidenced by the rejection of several of the data sets in Section 5 by the likelihood ratio tests discussed in Section 4. Hopefully, as recombination, purifying/positive selection due to chemical and structural constraints on the coded tRNA/proteins, compensating mutations, etc., are better understood, these effects will be properly incorporated into likelihood-based programs. The resulting, ever more sophisticated analyses, in a likelihood framework, will surely illuminate the contradictions that currently abound in molecular phylogenetics.

Conclusions

Before summarising the main points of this report, it is informative to compare what was originally intended to be researched to what has actually been investigated for this M.Res. summer project. The original proposal (from 2 May, 2000) was to study the relation of "distance-based" methods as approximations to the likelihood-based methods of molecular phylogenetic inference. In fact, some progress was made in approximating a tree likelihood by a formula containing only the tree branch lengths (results not reported here). However, after reading about the conflicting results of maximum likelihood analyses applied to real data [see, e.g., (Zardoya *et al.*, 1998), and (Hedges *et al.*, 1990)] and the theoretical uncertainties of ML as posed by Nei (1987) and Yang *et al.* (1995), the author became concerned with his basic assumption – that the likelihood function was worth approximating!

The author's research has therefore been pushed to deeper questions: Is maximum likelihood theoretically justified for molecular phylogenetic inference? Can likelihood ratios between different trees be "literally" interpreted as posterior odds of the trees? Are evolutionary models assumed in the current generation of ML analysis programs adequate for describing the complexities of real molecular data? The following conclusions have been reached:

- A "super-tree" likelihood expression has been constructed – explicitly for the simplest binary model – which is a function of all possible bipartition lengths, and which reduces to the individual tree likelihood functions when bipartitions not in a given tree are set to zero. The construction can be generalised to more general models of molecular mutation. Thus, different tree hypotheses can be considered composite hypotheses residing in a single super-tree space, and described by a single super-tree likelihood function.
- Therefore, the likelihood ratios of different trees can indeed be interpreted as the posterior odds of the trees. As is shown by simulation, tree support values based on likelihood ratios provide an intuitive indicator of tree selection accuracy, while bootstrap supports [and possibly the "integrated" likelihood values of the MAP analysis of Rannala and Yang (1996)] can be misleading.
- A set of "consistency checks" have been presented to test how adequate an evolutionary model describes a given data set. Two of the tests are particularly straightforward, involving the log-likelihood differences between the maximum possible log-likelihood value and the ML tree, and between an alternative tree (obtained by nearest-neighbour interchange around a given branch in the ML tree) and a star tree.
- The consistency checks have been applied to four real data sets, and are shown to reject the HKY85+Gamma model (which takes into account unequal nucleotide frequencies, transition / transversion bias, and site-rate-variation modelled as a Gamma distribution) for several of these ML analyses, including the ones for third codon positions in α and β globin genes from five mammals and in mtDNA data from *Heliconius erato* butterflies.
- Considering the inadequacy of present models of molecular mutation in describing these real data sets, ML analysis should, at present, be considered a "heuristic" procedure for most molecular data. However, such (tentative) ML analyses can still illuminate more phylogenetic information in a given data set than, for example, maximum parsimony methods, as is exemplified by likelihood-based analyses in Section 5 of the position of the *Eueides* clade with respect to the passion-vine butterflies of genus *Heliconius*, and of the geographic structure of race formation in *Heliconius erato* and *Heliconius melpomene*.

Thus, the likelihood framework is theoretically justified for molecular phylogenetic inference. As the complex biological processes that affect molecular mutation – from special structural and chemical constraints on the coded protein / tRNA, to correlation between different mutations, to variation in selection pressure over time – are better understood, they will hopefully be included in (and analysed by) ever-improving likelihood programs, which will, in turn, make more reliable phylogenetic inferences.

Acknowledgements

The author wishes to thanks supervisors Z. Yang and J. Mallet for their help in initial literature search and in defining the objectives of this project, and also M. Anisimova for many helpful discussions. The author gratefully acknowledges the Marshall Aid Commemoration Commission for funding his study at UCL through a British Marshall Scholarship.

References

Bates HW (1862). Contributions to an insect fauna of the Amazon valley. Lepidoptera: Heliconidae. Trans Linn Soc Lond. 23:495-566.

Branden C, Tooze J (1999). Introduction to protein, 2nd ed. Garland Pub. New York.

Bremer K (1988). Branch support and tree stability. Cladistics. 10: 295-304.

Brower A (1994). Phylogeny of Heliconius butterflies inferred from mitochondrial DNA sequences (Lepidoptera: Nymphalidae), Molecular Phylogenetics and Evolution. 3(2):159-174.

Brower A (1996). Race formation and the evolution of mimicry in Heliconius butterflies: a phylogenetic hypothesis from mitochondrial DNA sequences. Evolution. 50:195-221.

Brower A (1997). The evolution of ecologically important characters in Heliconius butterflies (Lepidoptera: Nymphalidae): a cladistic review. Zool J Linn Soc 119:457-472.

Brower A, Egan, M. 1997. Cladistic analysis of Heliconius butterflies and relatives (Nymphalidae: Heliconiiti): a revised phylogenetic position for Eueides based on sequences for mmtDNA and a nuclear gene, Proc R Soc Lond B. 264:969-977.

Brown K S, Jr (1981). The biology of Heliconius and related genera. Ann Rev Entomology. 26:427-256.

Brown WM, Prager EM, Wang A, Wilson AC (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol. 18(4):225-39.

Cavalli-Sforza LL, Edwards AWF (1967). Phylogenetic analysis: models and estimation procedures. Evolution 32:550-570.

Edwards AWF (1972). Likelihood. Cambridge University Press, Cambridge, UK.

Efron B, Tibshirani R (1993). An introduction to the bootstrap. Chapman & Hall, New York.

Efron B, Halloran E, Holmes S (1996). Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci U S A. 93(23):13429-34.

Farris JS, Kluge AG, Eckardt MJ (1970). A numerical approach to phylogenetic systematics. Syst. Zool. 19:172-189.

Felsenstein J (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet. 25(5):471-92 .

Felsenstein J (1978). Cases in which parsimony and compatibility methods will be positively misleading. Syst Zool. 27: 401-410.

Felsenstein J (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17(6):368-76.

Felsenstein J (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Fitch WM, Margoliash E (1967). Construction of phylogenetic trees. Science 155: 279-284.

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

Goldman N, Whelan S (2000). Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol Biol Evol. 17(6):975-8.

Goldman N, Yang Z (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11(5):725-36.

Hasegawa M (1991). Molecular phylogeny and man's place in Homonoidea. J. Anthropol. Soc. Nippon 99: 49-61.

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

Hedges SB (1994). Molecular evidence for the origin of birds. Proc Natl Acad Sci U S A. 91(7):2621-4.

Hedges SB, Moberg KD, Maxson LR (1990). Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. Mol Biol Evol. 7(6):607-33.

Hendy MD, Penny D (1993). J. Classif. 10:5-24.

Hendy MD, Penny D, Steel MA (1994). A discrete Fourier analysis for evolutionary trees. Proc Natl Acad Sci U S A. 91(8):3339-43.

Hillis DM, Bull JJ (1993). An empirical test of bootstrapping as a method for assessing the confidence in phylogenetic analysis. Syst Biol. 42:182-192.

Janke A, Xu X, Arrnason U (1997). The complete mitochondrial genome of the wallaroo (Macropus robustus) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. Proc Natl Acad Sci. 94:1276-1281.

Jermiin LS, Olsen GJ, Mengersen KL, Easteal S (1997). Majority-rule consensus of phylogenetic trees obtained by maximum likelihood analysis. Mol Biol Evol. 14:1296-1302.

Jones DT, Taylor WR, Thornton JM (1992). The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8(3):275-82.

Jukes TH, Cantor CR (1969) Evolution of protein molecules, in: Munro HN (ed). Mammalian protein metabolism. Academic Press, New York.

Kendall DG, Stuart A (1961). The advanced theory of statistics. Vol 2: Inference and relationships. Griffin, London.

Kimura M (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16(2):111-20.

Kishino H, Hasegawa M (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol. 29(2):170-9.

Lake JA (1987). A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol. 4(2):167-91.

Larget B, Simon D (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol. 16:750-759.

Lento GM, Hickson RE, Chambers GK, Penny D (1995). Use of spectral analysis to test hypotheses on the origin of pinnipeds. Mol Biol Evol. 12(1):28-52.
Liò P, Goldman N (1999). Using protein structural information in evolutionary inference: transmembrane proteins. Mol Biol Evol. 16(13):1696-710.
Liò P, Goldman N, Thorne JL, Jones DT (1998). PASSML: combining evolutionary inference and protein secondary structure prediction. Bioinformatics. 14(8):726-33.
Mallet J, Gilbert LE (1995): Why are there so many mimicry rings? Correlations between habitat, behaviour and mimicry in Heliconius butterflies. Biol J Linn Soc. 55:159-180.
Murawski DA (1993). A taste for poison. National Geographic. 184 (6): 122-137.
Navidi WC, Churchill GA, von Haeseler A (1991). Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol Biol Evol. 8(1):128-43.
Nei M, Kumar S, Takahashi K (1998). The optimisation principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc Natl Acad Sci U S A. 95(21):12390-7.
Nei M (1987). Molecular evolutionary phylogenetics. Columbia Univ. Press, New York.
Neyman J (1971). Molecular studies of evolution: a source of novel statistical problems, in: Gupta, SS, Tackel, J (eds). Statistical decision theory and related topics. Academic Press, NY.
Numerical recipes in C: the art of scientific computing, 2nd ed. (1992). Eds. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Cambridge University Press, Cambridge, UK.
Olsen GJ, Matsuda H, Hagstrom R, Overbeek R (1994a). fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput Appl Biosci. 10(1):41-8.
Olsen GJ, Woese CR, Overbeek R (1994b). The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol. 176(1):1-6.
Rannala B, Yang Z (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol. 43(3):304-11.
Reeves JH (1992). Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J Mol Evol. 35:17–31.
Rodríguez-Trelles F, Tarrio R, Ayala FJ (2000). Fluctuating mutation bias and the evolution of base composition in Drosophila. J Mol Evol. 50(1):1-10.
Rzhetsky A, Nei M (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol Biol Evol. 10(5):1073-95.
Schwartz JH (1984). The evolutionary relationships of man and orang-utans. Nature 308: 501-505
Sitnikova T, Rzhetsky A, Nei M (1995). Interior-branch and bootstrap tests of phylogenetic trees. Mol Biol Evol. 12(2):319-33.
Steel M, Penny D (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol Biol Evol. 17(6):839-50.
Steel M, Hendy MD, Penny D (1998). Reconstructing phylogenies from nucleotide pattern probabilities: A survey and some new results. Discr Appl Math. 88: 367-396.
Strimmer K, Moulton V (2000). Likelihood analysis of phylogenetic networks using directed graphical models. Mol Biol Evol. 17(6):875-81.
Strimmer K, von Haeseler A (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci U S A. 94(13):6815-9.
Stryer, L (1995). Biochemistry, 4th ed. W.H. Freeman . New York.
Waddell PJ, Penny D, Moore T (1997). Hadamard conjugations and modeling sequence evolution with unequal rates across sites. Mol Phylogenet Evol. 8(1):33-50.
Waddell PJ, Penny D, Hendy MD, Arnold G (1994). The sampling distributions and covariance matrix of phylogenetic spectra. Mol Biol Evol. 11: 630-642.
Yang Z (1997). How often do wrong models produce better phylogenies? Mol Biol Evol. 14(1):105-8.
Yang Z (1994a). Estimating the pattern of nucleotide substitution. J Mol Evol. 39(1):105-11.
Yang Z (1994b). Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst Biol.43:349-342.
Yang Z (1996a). Phylogenetic analysis using parsimony and likelihood methods. J Mol Evol. 42(2):294-307.
Yang Z (1996b). Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol. 42(5):587-96.
Yang Z (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13(5):555-6.
Yang Z (2000). Complexity of the simplest phylogenetic estimation problem. Proc R Soc Lond B Biol Sci. 22; 267(1439):109-16.
Yang Z, Rannala B (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol. 14(7):717-24.
Yang Z, Goldman N, Friday A (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol. 11(2):316-24.
Yang Z, Goldman N, Friday A (1995). Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst Biol. 44(3):384-399.
Yang Z, Nielsen R, Goldman N, Pedersen A-MK (2000). Codon-substitution models for heterogenous selection pressure at amino acid sites. Genetics. 155:431-449.
Zardoya R, Cao Y, Hasegawa M, Meyer A (1998). Searching for the closest living relative(s) of tetrapods through evolutionary analyses of mitochondrial and nuclear data.
Mol Biol Evol. 15(5):506-17.
Zharkikh A, Li WH (1992). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: I. Four taxa with a molecular clock. Mol Biol Evol. 9: 1119-1147.

Appendix A

This appendix outlines the procedure for designing a super-tree likelihood function which reduces to the usual tree likelihood function when the lengths of bipartitions not in a given tree are set to zero.

The basic idea here will be to re-write the formulas for the predicted site pattern probabilities in a given *m*-taxon tree as the sum of several exponentials of "path-sets" (i.e., linear combinations of bipartition lengths). The two-taxon case is the simplest example. Given the $c \times c$ instantaneous transition matrix Q of an evolutionary model for c characters [so that the time evolution of the column vector of character probabilities **P** is given by $d\mathbf{P}/dt = Q\,\mathbf{P}$] the transition probability matrix for two taxa separated by time t can be found by diagonalizing Q. In particular, the c eigenvectors of Q, with eigenvalues $\lambda_0, \lambda_1, \ldots \lambda_{c-1}$, are arranged into the columns of a matrix $U = \{u_{ij}\}$. Also, denote $U^{-1}=\{v_{ij}\}$. Then it is easy to show that the probability matrix is

$$\left\{p_{ij}(t)\right\} = \exp(Qt) = U \begin{bmatrix} \exp(\lambda_0 t) & & \\ & \mathbf{O} & \\ & & \exp(\lambda_{c-1} t) \end{bmatrix} U^{-1} = \sum_m u_{im} \exp(\lambda_m t) v_{mj} , \qquad \text{(A1)}$$

Note the convention here: $p_{ij}(t)$ is the probability of the character state j becoming the character state i after time t. The eigenvectors ("channels") provide an alternative basis from the usual character states (which correspond to unit vectors); in the channel basis, the time dependence is particularly simple (exponential). Note that if non-diagonal entries of Q are positive (the usual case), and if probability is conserved, one is guaranteed to have the largest eigenvalue as $\lambda_0=0$, with the components of the corresponding eigenvector proportional to the equilibrium character frequencies.

Now, this process of decomposing formulas into "channels" can be extended to the general *m*-taxa case. Consider finding the probability of finding site pattern *xyzw* for a four-taxon tree like Tree I in Figure 3. The usual sum over internal states, arbitrarily picking taxon A as the "root" (the final answer is independent of the choice of root for reversible models; Felsenstein, 1981) is, with summation over indices suppressed:

$$\begin{aligned} p_{xyzw} &= \pi_x \sum_{i,j} p_{ix}(t_A) p_{yi}(t_B) p_{ji}(t_I) p_{zjx}(t_C) p_{wj}(t_D) \\ &= \pi_x \sum_{i,j} \sum_{m,n,k,r,s} \left[u_{im} \exp(\lambda_m t_A) v_{mx}\right]\left[u_{yn} \exp(\lambda_n t_B) v_{ni}\right]\left[u_{jk} \exp(\lambda_k t_I) v_{ki}\right]\left[u_{zr} \exp(\lambda_r t_C) v_{rj}\right]\left[u_{ws} \exp(\lambda_s t_D) v_{sj}\right] \\ &= \sum_{m,n,k,r,s} \exp\left[\lambda_m t_A + \lambda_n t_B + \lambda_k t_I + \lambda_r t_C + \lambda_s t_D\right]\left\{\pi_x v_{mx} u_{yn} u_{zr} u_{ws}\right\}\left\{\sum_i u_{im} v_{ni} v_{ki}\right\}\left\{\sum_j u_{jk} v_{rj} v_{sj}\right\} \end{aligned} \qquad \text{(A2)}$$

For a general *m*-taxon tree and site pattern, the formula is similar. There is a sum over exponentials corresponding to possible "channel"-assignments (or "path-sets") to the branches of the tree. The coefficients of these exponentials is determined by a set of terms like $\{\pi_x v_{mx} u_{yn} u_{zr} u_{ws}\}$ which give the transformation from the basis of usual character states to the basis of "channels", eigenvectors of Q. Finally, there are a set of "vertex" operators like $V_{mnk}=\sum u_{im} v_{ni} v_{ki}$ which encode factors to assign for each trisection of channels. For simple models, the vertex operator is zero for many types of channel intersections; the operator can then be concisely presented diagramatically as a set of "allowed" vertices (those producing non-zero factors) with corresponding operator values.

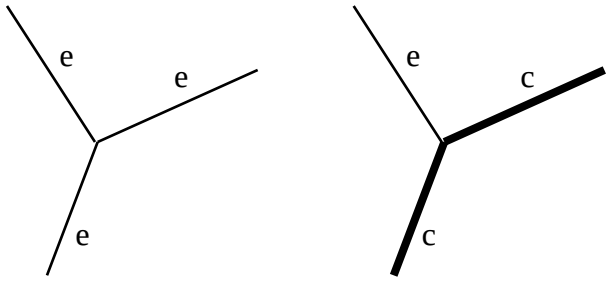


Figure A1. The vertex operator, in diagrammatic representation, for the simplest binary model. If the following channel intersections occur in the tree, a factor of 1 is given to the path-set; other intersections give factors of zero (i.e., are disallowed).

For the simplest binary model,

$$U = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}; \qquad \text{(A3)}$$

$U^{-1} = 2\,U$; and the vertex operator is diagrammatically presented in Figure A1. There are two channels, the "equilibrium" ($\lambda_e = 0$) and "changing" ($\lambda_c = -2$) channels. Based on these diagrammatic rules, it is easy to see that the only path-sets that will be included in a formula for a given site pattern are those which have *c*-channels threading through the tree in connected, non-intersecting lines; see the examples in Figure A2. For a given site pattern the coefficient of the exponential is determined by assigning a factor 1/2 to each taxon starting in an *e*-channel, and a factor ±1/2 (sign depending on the taxon's character for that site pattern) for each taxon starting in a *c*-channel. The factors are multiplied to yield a coefficient $\pm 1/2^m$. With these rules for finding the formulas for the site pattern frequencies, the log-likelihood function for a given tree is $\mathrm{l} = \sum n_i \log p_i$, summed over all site patterns.

How then can one define a super-tree likelihood function? For an *m*-taxon tree, the predicted site pattern probabilities take the form of a sum of several exponentials of sums of the bipartition lengths *t*. For the simplest binary model above, there are 2^{m-1} such exponential terms, each corresponding to an allowed path-set; and each path-set term is uniquely determined by which pair (or quartet, hextet, etc.) of external branches is given channel label *c*. For different trees, there will be the same path-sets contributing to a given site pattern probability; only the internal bipartition lengths going into the exponent of each path-set term will be different. To produce a formula for a super-tree site-pattern probability, one includes in the exponent of each path-set term any bipartition length that might show up for any tree. By construction, then the site-pattern probability reduces to the usual formula for a given tree when bipartitions not in that tree are set to zero. This is the procedure used to find eq. (13) for the four-taxon tree in Section 2.4.

A similar, well-defined procedure can be defined for finding the super-tree site pattern probabilities for a Kimura 3-state model, where the instantaneous transition matrix is:

$$Q = \begin{bmatrix} -1-\kappa_1-\kappa_2 & 1 & \kappa_1 & \kappa_2 \\ 1 & -1-\kappa_1-\kappa_2 & \kappa_2 & \kappa_1 \\ \kappa_1 & \kappa_2 & -1-\kappa_1-\kappa_2 & 1 \\ \kappa_2 & \kappa_1 & 1 & -1-\kappa_1-\kappa_2 \end{bmatrix} \qquad \text{(A4)}$$

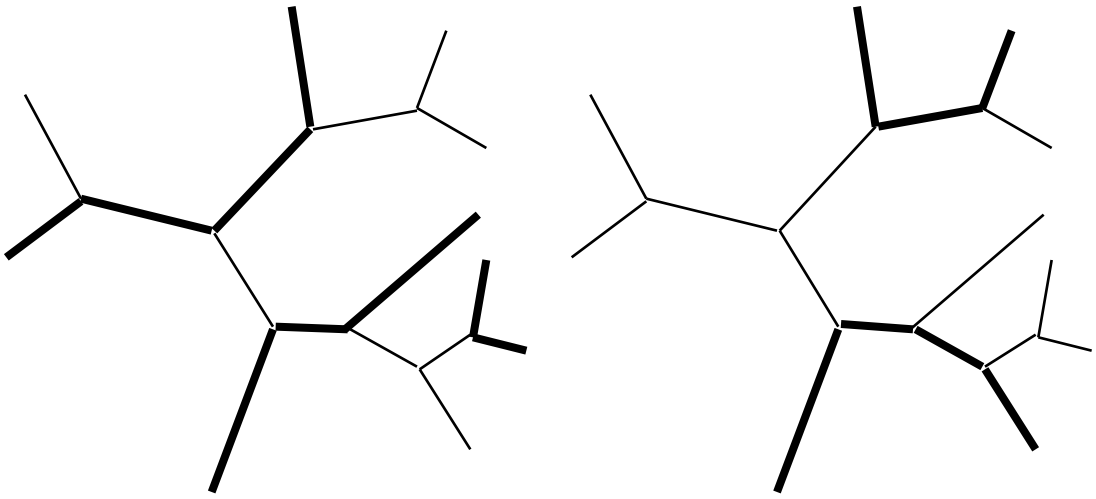


Figure A2. Two sample path-sets for a 10-taxon tree. For each path-set – a labelling of each branch with a channel (light lines are the "*e*" channel; dark lines are the "*c*" channel) – there is an exponential term in the site pattern probability, with the exponent proportional to the sum of the lengths of all *c*-branches.

The matrix of eigenvectors is then:

$$U = \frac{1}{4}\begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix} \qquad \text{(A5)}$$

and $U^{-1} = 4\,U$. The eigenvalues of the channels are $\lambda_e = 0$; $\lambda_{c1} = -2 - 2\kappa_2$; $\lambda_{c2} = -2\kappa_1 - 2\kappa_2$; and $\lambda_{c3} = -2 - 2\kappa_1$. The diagrammatic vertex rules are given in Figure A3. Again, each site pattern probability is a linear combination of path-set exponentials. It can be shown that for any given tree, each path-set can be uniquely indexed based on the channel-labeling of the external branches only. Thus the site pattern probability for the super-tree can again be determined by including in the exponents of each path-set term all bipartition lengths that might show up for any tree.

The above construction of the super-tree fails for more general models. For example, in the binary model with unequal frequencies, the instantaneous rate matrix is:

$$Q = \begin{bmatrix} -\pi_1 & +\pi_0 \\ +\pi_1 & -\pi_0 \end{bmatrix}, \qquad \text{(A6)}$$

and the matrix of eigenvectors, and its inverse can be written as:

$$U = \begin{bmatrix} \pi_0 & +1/2 \\ \pi_1 & -1/2 \end{bmatrix}; \qquad U = \begin{bmatrix} 1 & +2\pi_1 \\ 1 & -2\pi_0 \end{bmatrix};$$

with eigenvalues $\lambda_e = 0$ and $\lambda_c = -2$. The diagrammatic vertex rules are those given in Figure A4. However, here the path-set cannot be indexed solely based on the channel-labelling of the external branches; see Figure A5. Thus, it is somewhat arbitrary to decide which pathset in a given tree corresponds to a given pathset in a different tree. The procedure for defining the super-tree site pattern probability is still possible, but not unique.

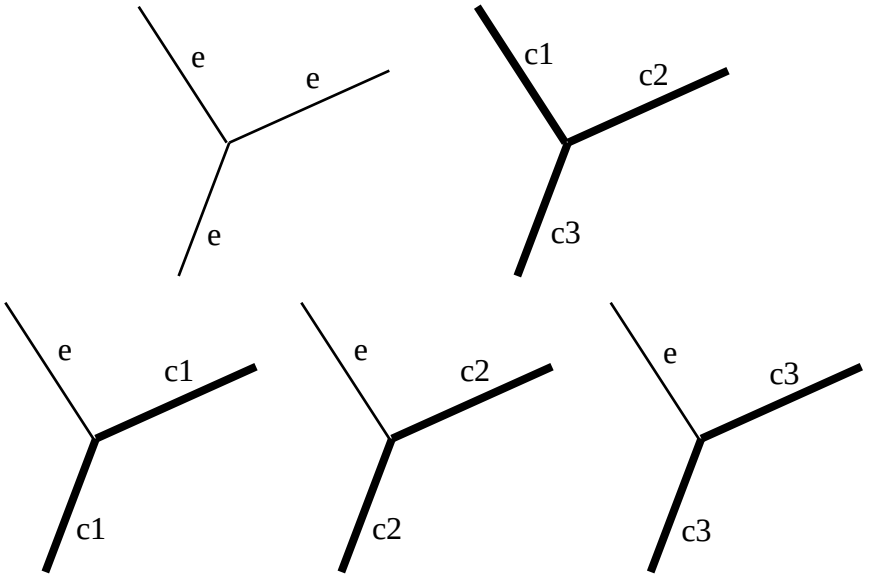


Figure A3. The vertex operator, in diagrammatic representation, for the Kimura 3-state model. If the following channel intersections occur in the tree, a factor of 1 is given to the path-set; other intersections give factors of zero (i.e., are disallowed).

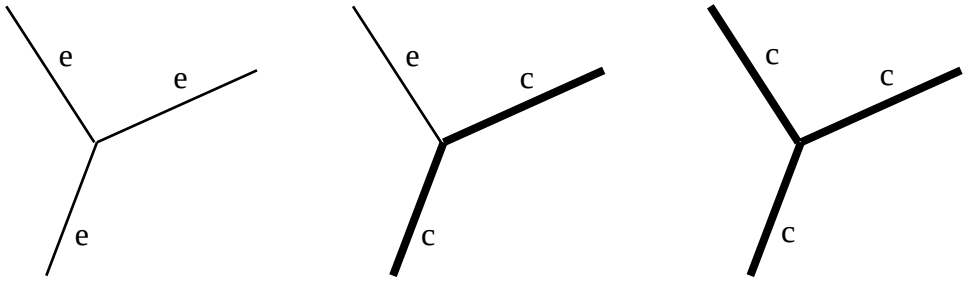


Figure A4. The vertex operator, in diagrammatic representation, for the binary model with unequal frequencies. The following channel intersections give non-zero values. In particular, the third vertex gives a factor proportional to $\pi_0 - \pi_1$, and is therefore zero only in the equal frequency case (see Figure A1).

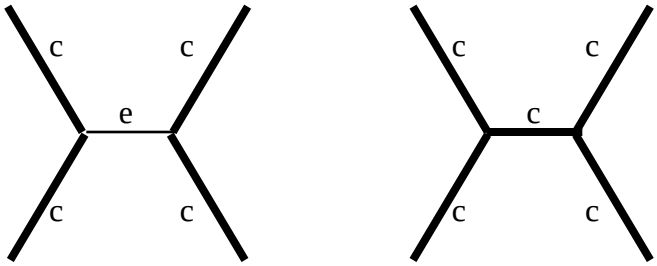


Figure A5. Two path-sets which contribute separate exponential terms in the formulas site pattern probabilities for the binary model with unequal frequencies, but which have the same channel labelling of external branches.

For models like HKY85 for DNA sequences the Dayhoff matrix for amino-acid sequences, the result is the same: the super-tree site pattern probabilities can be defined, but not necessarily uniquely.[16] Indeed, the super-tree probabilities given above for the simplest binary model and the Kimura three-state model are not unique, either. Those simple procedures can easily be modified by extra terms which will not perturb their limits for individual tree hypotheses. However, the procedures above are useful for obtaining well-defined, especially simple formulas, with intuitive properties like the additivity relation for pair-wise distances given in Section 2.4 for the simplest binary model. Defining such procedures for more general models, possibly based on geometrical pictures like Figure 3, is currently under investigation.

Appendix B

The mean and variance of the normal distribution of statistics like $(l_{max} - l_I)$ described in Section 3.1 are approximated as follows. For a finite number of sites n, the full $c^m - 1$ site patterns are not expected to be seen in the data set; indeed, many real data sets are dominated by sites which are the same in all or most taxa ["sparseness" of the data; see, e.g., (Goldman, 1993), (Yang *et al.*, 1995)]. A first guess might be to assume that each site pattern with non-zero frequency constitutes one degree of freedom; the contributions of each degree of freedom to the mean and variance of $(l_{max} - l_I)$ are 1/2 and 1/2, respectively. The overall mean and variance would be predicted to be $(n_{patt} - n_{param})/2$ and $n_{patt}/2$, respectively.

In actuality, site patterns which do not appear in the given data set still make a contribution to the mean and variance of $(l_{max} - l_I)$. Given the expected probability p_i of a given site pattern, the contribution of that site pattern to the mean and variance to $(l_{max} - l_I)$, under Poisson statistics, are explicitly

$$\begin{aligned}
\left\langle l^{(i)}_{max} - l^{(i)}_I \right\rangle &= \sum_{q=0}^{\infty} \left[\frac{(np_i)^q e^{-np_i}}{q!} \right] \left[\left(q \log(q) - q \right) - \left(q \log(np_i) - np_i \right) \right] \\
\left\langle \left(l^{(i)}_{max} - l^{(i)}_I \right)^2 \right\rangle &= \sum_{q=0}^{\infty} \left[\frac{(np_i)^q e^{-np_i}}{q!} \right] \left[\left(q \log(q) - q \right) - \left(q \log(np_i) - np_i \right) \right]^2 \qquad \text{(B1)} \\
Var\left[l^{(i)}_{max} - l^{(i)}_I \right] &= \left\langle \left(l^{(i)}_{max} - l^{(i)}_I \right)^2 \right\rangle - \left\langle l^{(i)}_{max} - l^{(i)}_I \right\rangle
\end{aligned}$$

The functions for the mean and variance are plotted in Figure B1; it is seen that both functions approach 1/2 quickly for $p_i > 1$, but are nevertheless non-vanishing $p_i < 0$. For a real

[16] It is clear, in fact, that the super-tree site pattern probabilities can be defined for any model, reversible or not. Basically, in describing the site pattern probabilities, there are $(N-1) \times 2^B$ possible pathset coefficients, where B is the number of possible bipartitions times the number of characters, and N is the number of site patterns (c^m). One can always choose the coefficients to satisfy the smaller number of $(N-1) \times 2^{[c\,(2m-5)]}$ constraints obtained in forcing the super-tree likelihood function to reduce to the usual tree likelihood functions when bipartitions not in a given tree are set to zero.

data set, one thus needs estimates of p_i for all site patterns with predicted frequencies greater than, say, 1/10; the accuracy of the estimates for $p_i > 1$ is not terribly critical. This estimation can easily be done for a given data set by optimising branch lengths and evolutionary parameters for any tree close to the expected maximum likelihood tree. The resulting parameters are then used in, e.g., PAML's *evolver* program, to simulate a data set with, say, 10 times as many sites as the data set under consideration. Then the formula (B1) is applied to non-vanishing site pattern frequencies of the simulated data set, and summed to find the overall predicted mean and variance of the $(l_{max} - l_i)$ distribution. If a number n_{param} of parameters are fitted to obtain l_i, the predicted mean should be decreased by approximately $n_{param}/2$. In this report, for DNA and amino acid data, a simulated data set 100 times and 10,000 times, respectively, the length of the real data set is constructed.

Of course, an alternative way to find the predicted mean and variance would be to actually do, say 100 simulations of data sets of the same length as the real data set, and to find $(l_{max} - l_i)$ for each replicate. See Yang (1994) for examples. This latter procedure is generally much more time-consuming, since the likelihood parameters need to be optimised for each simulated replicate. Estimates from both procedures agree (see Figure 11).

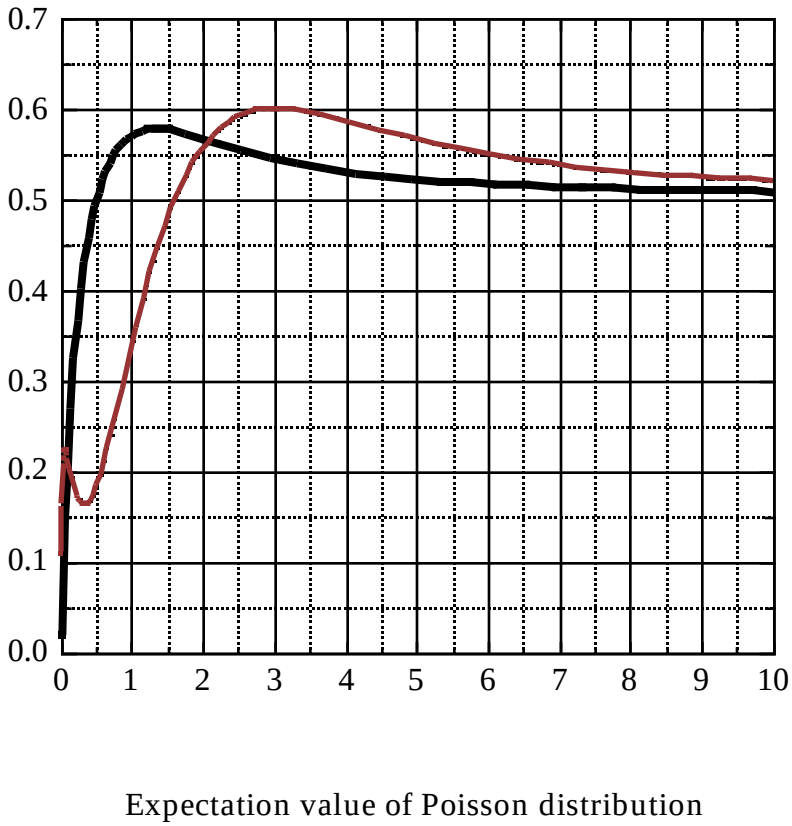


Figure B1. Mean (dark thick line) and variance (grey thin line) of $l - l_{max}$ for Poisson distributions with different expectation values.