THE ROYAL SOCIETY

Complexity of the simplest phylogenetic estimation problem

Ziheng Yang

Department of Biology (Galton Laboratory), University College London, 4 Stephenson Way, London NW1 2HE, UK (z.yang@ucl.ac.uk)

The maximum-likelihood (ML) solution to a simple phylogenetic estimation problem is obtained analytically. The problem is estimation of the rooted tree for three species using binary characters with a symmetrical rate of substitution under the molecular clock. ML estimates of branch lengths and log-likelihood scores are obtained analytically for each of the three rooted binary trees. Estimation of the tree topology is equivalent to partitioning the sample space (space of possible data outcomes) into subspaces, within each of which one of the three binary trees is the ML tree. Distance-based least squares and parsimony-like methods produce essentially the same estimate of the tree topology, although differences exist among methods even under this simple model. This seems to be the simplest case, but has many of the conceptual and statistical complexities involved in phylogeny estimation. The solution to this real phylogeny estimation problem will be useful for studying the problem of significance evaluation.

Keywords: consistency; identifiability; maximum likelihood; molecular phylogenetics; parameter space; sample space

1. INTRODUCTION

> 'I am very pleased to see that the problem offers sufficient challenge to statisticians'
>
> (Cavalli-Sforza; discussion in Edwards 1970, p. 170)

With the development of more realistic statistical models and improvement in computing power and computer programs, the maximum-likelihood (ML) method is more and more widely used in molecular phylogenetic analysis. Cavalli-Sforza & Edwards's (1967) view that phylogenetic reconstruction should best be viewed as a statistical estimation problem is now generally accepted. Given the central role of ML in statistical estimation, this point of view also stipulates that ML should be the method of choice for phylogeny estimation (Edwards 1995). It should be noted that Edwards's general likelihood framework appears to include what is often known as the Bayes method (Edwards 1970; Rannala & Yang 1996; Mau & Newton 1997; Yang & Rannala 1997).

Phylogeny reconstruction, however, is a peculiar statistical estimation problem (Yang *et al.* 1995). It provided 'sufficient challenge to statisticians' (Cavalli-Sforza; discussion in Edwards 1970), and was described as 'a source of novel statistical problems' (Neyman 1971). Some aspects of the complexity of the estimation problem were explored recently (Yang 1994, 1996, 1997; Yang *et al.* 1995). The major difficulty appears to lie in the parameter space of the problem. In Felsenstein's (1981) formulation, the likelihood is calculated separately for each tree and maximized for branch lengths in that tree. The optimum likelihood values for trees are then compared to estimate the unknown true tree. Effectively, different phylogenies have different parameter spaces and different likelihood functions (Nei 1987). As a result, it is not obvious whether ML estimate of phylogeny has the asymptotic properties (such as consistency and asymptotic efficiency) of the conventional ML method. Yang (1994) showed that ML phylogeny estimation is statistically consistent as long as the model is regular enough so that the trees are identifiable with infinite amount of data (Yang 1994). Chang (1996) and Rogers (1997) showed that even very general models used in phylogenetic analysis identify trees without problem. The asymptotic efficiency of ML is less certain (Yang 1997; Bruno & Halpern 1999). Numerous computer simulations suggest that ML performs better, or not much worse, than other methods such as parsimony or distance-matrix methods (see for example Huelsenbeck (1995) for a review). However, hypothesis testing concerning tree topologies and evaluation of the significance of the ML tree have been much more difficult. No workable parametric method has been suggested to construct a confidence interval for the ML tree or to evaluate its significance, and controversies exist concerning the interpretation of the non-parametric bootstrap method (Felsenstein 1985; Zharkikh & Li 1992; Hillis & Bull 1993; Efron *et al.* 1996).

A major difficulty of analysing the ML method of phylogeny reconstruction is that no analytical results are known even for simple cases. For example, for the case of three species with nucleotide sequences evolving under the JC69 substitution model (Jukes & Cantor 1969) and a molecular clock, ML estimates (MLEs) of branch lengths cannot be obtained analytically (Yang 1994). As a result, the ML tree cannot be determined analytically for a data outcome (a given data set) without iteration to estimate branch lengths. The lack of analytical results makes it difficult to study the properties of the method, and one has to resort to computer simulation, which typically examines a small portion of the parameter space. It is noted that the estimation problem mentioned above becomes tractable if binary characters are considered instead of nucleotides with four states. This paper describes the solution to that problem. The problem seems to be the simplest case one can imagine, and also the first for which an analytical solution to ML is obtained. Nevertheless, it has most of the complexities involved in more general cases (Yang *et al.* 1995), and the solution will be useful in studying significance tests concerning the ML tree.

Received 28 July 1999 *Accepted* 1 November 1999

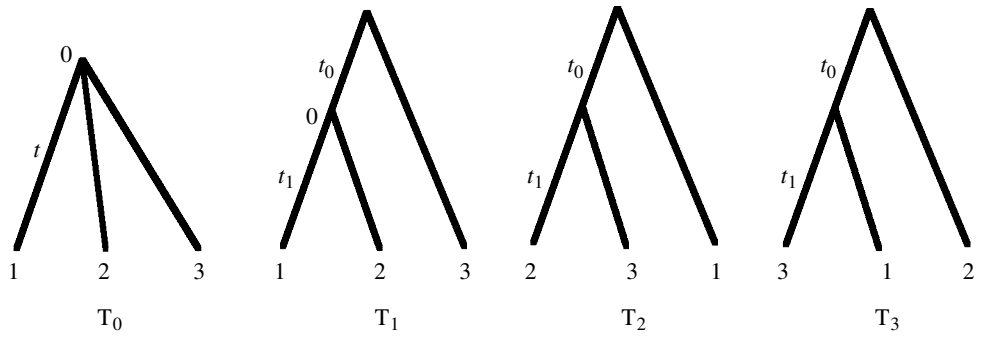

Figure 1. The star tree $T_0 = (123)$ and the three bifurcating trees for three species: $T_1 = ((12)3)$, $T_2 = ((23)1)$, and $T_3 = ((31)2)$. The branch length is defined as the expected number of substitutions (changes) per site along the branch.

2. MODEL AND PROBLEM

Consider three species 1, 2 and 3. The three (rooted) bifurcating trees are shown in figure 1: $T_1 = ((12)3)$, $T_2 = ((23)1)$ and $T_3 = ((31)2)$. The star tree $T_0 = (123)$ is chosen as the estimate in real data if none of the binary trees is any better. The objective is to estimate the true tree topology (which is one of T_1, T_2 or T_3), and evaluate the reliability (statistical significance) of the estimated tree. This paper concerns the first question (i.e. point estimation) only.

The data are three DNA sequences for the three species, each of n nucleotides long. We will consider binary characters, so that only pyrimidines (Y) and purines (R) are distinguished. The evolutionary rate is assumed to be the same over time; that is, the molecular clock holds. A stationary and homogeneous Markov process is assumed to describe nucleotide substitution, and the substitution rates are assumed to be equal in both directions. We measure time by the expected number of nucleotide substitutions, and so the instantaneous rate matrix is

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

The transition probability matrix over time t is given as

$$P(t) = \{p_{ij}(t)\} = \mathrm{e}^{Rt} = \begin{bmatrix} 1-q(t) & q(t) \\ q(t) & 1-q(t) \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(1+\mathrm{e}^{-2t}) & \frac{1}{2}(1-\mathrm{e}^{-2t}) \\ \frac{1}{2}(1-\mathrm{e}^{-2t}) & \frac{1}{2}(1+\mathrm{e}^{-2t}) \end{bmatrix}, \tag{2}$$

where $p_{ij}(t)$ is the probability that nucleotide i changes into j over time t. Note that

$$q(t) = (1-\mathrm{e}^{-2t})/2, \tag{3}$$

is the probability that two nucleotides separated by time t are different.

Data at different sites are assumed to be independently and identically distributed. There are $2^3 = 8$ possible data configurations (site patterns) at a site. Some of them (such as YYR and RRY) have equal probabilities of occurrence under any tree, and are collapsed. Four site patterns are then possible and can be represented as *xxx*, *xxy*, *yxx* and *xyx*, where x and y are any two different nucleotides (table 1). The data are the observed numbers of sites with those site patterns: n_0, n_1, n_2 and n_3. The total number of sites (the sample size) is $n = (n_0 + n_1 + n_2 + n_3)$. For the purpose of parameter estimation alone, the observed site pattern frequencies $f_i = n_i/n$ can be used.

For example, when the segment of the mitochondrial DNA of human (species 1), chimpanzee (species 2) and gorilla (species 3) published by Brown *et al.* (1982) are converted into sequences of pyrimidines and purines, the data become $n_0 = 762$, $n_1 = 54$, $n_2 = 41$, $n_3 = 38$, with $n = 895$ (table 1). This numerical example will be used in later discussions.

3. PARSIMONY AND LEAST-SQUARES METHODS

(a) *Parsimony*

The unordered parsimony method used in molecular sequence analysis does not distinguish rooted trees. However, as argued by Sober (1988), if tree T_1 is the true tree, pattern *xxy* should be more likely than patterns *yxx* and *xyx*, since the former is generated by a change over a long time-period (from node 0 to 3 in tree 1 of figure 1) while either of the latter patterns is generated by a change in a short time-period (from node 0 to 1 or 2). A parsimony-style method thus compares n_1, n_2 and n_3 and chooses the tree T_i corresponding to the largest n_i. In our data set, $n_1 > n_2$ and $n_1 > n_3$, so that tree T_1 is the estimate of the true phylogeny.

(b) *Least squares*

The least-squares (LS) method calculates pairwise distances and treats them as observed data. Branch lengths in each tree are then estimated by LS, that is, by minimizing the sum of squared differences between the observed and expected pairwise distances

$$Q = (d_{12} - \hat{d}_{12})^2 + (d_{23} - \hat{d}_{23})^2 + (d_{31} - \hat{d}_{31})^2. \tag{4}$$

The expected distance $\hat{d}_{ij}$ between two species i and j is the sum of branch lengths in the tree along the path connecting the two species. Since $q(t)$ in equation (3) is the expected proportion of different sites between two sequences separated by distance t, the sequence distance can be estimated by

$$\hat{t} = -1/2 \log\{1 - 2q\}, \tag{5}$$

where q is the proportion of different sites between the two sequences. For the formula to be applicable for all three

Table 1. *Site patterns and their probabilities of occurrence under different trees*

(The observed numbers and frequencies of site patterns are from a segment of the mitochondrial DNA of human, chimpanzee and gorilla (Brown *et al.* 1982).)

	observed			probability p_i under tree		
category (i)	pattern	numbers (n_i)	frequencies (f_i)	$T_1((12)3)$	$T_2((23)1)$	$T_3((31)2)$
0	*xxx*	$n_0 = 762$	$f_0 = 0.851397$	$p_0(t_0, t_1)$	$p_0(t_0, t_1)$	$p_0(t_0, t_1)$
1	*xxy*	$n_1 = 54$	$f_1 = 0.060335$	$p_1(t_0, t_1)$	$p_2(t_0, t_1)$	$p_2(t_0, t_1)$
2	*yxx*	$n_2 = 41$	$f_2 = 0.045810$	$p_2(t_0, t_1)$	$p_1(t_0, t_1)$	$p_2(t_0, t_1)$
3	*xyx*	$n_3 = 38$	$f_3 = 0.042581$	$p_2(t_0, t_1)$	$p_2(t_0, t_1)$	$p_1(t_0, t_1)$
sum	—	$n = 895$	1	1	1	1

pairwise comparisons, it is required that $f_1 + f_2 < 1/2$, $f_2 + f_3 < 1/2$ and $f_3 + f_1 < 1/2$. The distances are then given as

$$\left.\begin{aligned} d_{12} &= -1/2 \log\{1-2(f_2+f_3)\} = -1/2 \log\{2(f_0+f_1)-1\} \\ d_{23} &= -1/2 \log\{2(f_0+f_2)-1\} \\ d_{31} &= -1/2 \log\{2(f_0+f_3)-1\} \end{aligned}\right\}. \tag{6}$$

For the star tree T_0 (figure 1), the sum of squares $Q_0(t) = (d_{12}-2t)^2 + (d_{23}-2t)^2 + (d_{31}-2t)^2$ is minimized at $\hat{t} = (d_{12} + d_{23} + d_{31})/6$. For tree T_1, the sum of squares is

$$\begin{aligned} Q_1(t_0,t_1) = &(d_{12} - 2t_1)^2 + (d_{23} - 2t_0 - 2t_1)^2 \\ &+ (d_{31} - 2t_0 - 2t_1)^2. \end{aligned} \tag{7}$$

If $d_{12} < (d_{23} + d_{31})/2$, Q_1 is minimized at

$$\left.\begin{aligned} \hat{t}_0 &= (d_{23} + d_{31} - 2d_{12})/4 \\ \hat{t}_1 &= d_{12}/2 \end{aligned}\right\}, \tag{8}$$

with $Q_1 = (d_{23} - d_{31})^2$. If $d_{12} \geqslant (d_{23} + d_{31})/2$, tree T_1 converges to the star tree T_0. Branch lengths and sum of squares can be calculated similarly for trees T_2 and T_3. It is easy to show that T_1 minimizes Q if d_{12} is the smallest of the three distances (Saitou 1988). Note that the condition $d_{12} < \min(d_{23}, d_{31})$ is equivalent to the condition $f_1 > \max(f_2, f_3)$. Thus LS and parsimony produce the same tree if the distance formula is applicable for all pairwise comparisons.

For our numerical example, $d_{12} = 0.097118$, $d_{23} = 0.115076$, $d_{31} = 0.119313$. The estimate under the star tree T_0 is $\hat{t} = 0.055251$, with $Q_0 = 0.009436$. The estimates under tree T_1 are $\hat{t}_0 = 0.010038$, $\hat{t}_1 = 0.048559$, with $Q_1 = 0.000018$. Both T_2 and T_3 converge to the star tree T_0. Tree T_1 is the LS estimate of the true phylogeny.

4. MAXIMUM LIKELIHOOD

(a) *Estimation of branch lengths*

ML estimation of phylogeny involves optimization of branch lengths for each tree topology to calculate the optimum log likelihood for that tree and comparison of the (optimum) log-likelihood values among tree topologies (Felsenstein 1981). In the following, we obtain the MLEs of branch lengths and the log-likelihood value under each tree of figure 1. Let p_0, p_1, p_2, p_3 be the probabilities of observing the four site patterns *xxx*, *xxy*, *xyx* and *yxx*, respectively. The probability of a data outcome (n_0, n_1, n_2, n_3) is given by the multinomial distribution

$$P(n_0, n_1, n_2, n_3) = \frac{n!}{n_0! n_1! n_2! n_3!} p_0^{n_0} p_1^{n_1} p_2^{n_2} p_3^{n_3}. \tag{9}$$

The log likelihood is then

$$\ell = \sum_{i=0}^{3} n_i \log\{p_i\}, \tag{10}$$

with the constant term $\log n! - \log\{n_0! n_1! n_2! n_3!\}$ suppressed. For point estimation, it is convenient to work with the per-site log likelihood (Yang 1994):

$$\ell/n = \sum_{i=0}^{3} f_i \log\{p_i\}. \tag{11}$$

(i) *The star tree T_0*

The star tree has only one branch length t (figure 1). The branch length can also be measured by $a = (1 - e^{-2t})/2$, the probability that a nucleotide at a site in the ancestor is different from the nucleotide at that site in any current sequence. The site pattern probabilities are

$$\left.\begin{aligned} P_0(t) &= a^3 + (1-a)^3 = 1 - 3a + 3a^2 = \tfrac{1}{4} + \tfrac{3}{4} e^{-4t} \\ P_1(t) &= a^2(1-a) + (1-a)^2 a = a - a^2 = \tfrac{1}{4} - \tfrac{1}{4} e^{-4t} \\ P_2(t) &= p_3(t) = p_1(t) \end{aligned}\right\}. \tag{12}$$

The log-likelihood function is

$$\ell_0/n = f_0 \log\{1 - 3a + 3a^2\} + (1 - f_0) \log\{a - a^2\}. \tag{13}$$

The MLE of a or t can be obtained by setting $p_0 = f_0$ if a root exists. The results are summarized in table 2. Note that the MLE of t differs from the LS estimate.

It may be noted that for estimation of branch length t or a in T_0, f_0 (or $1 - f_0 = f_1 + f_2 + f_3$) is the sufficient statistic; that is, all information concerning t or a is contained in f_0. The MLE of a and the optimum likelihood is shown in figure 2. The log likelihood ranges from $-\log\{4\} = -1.386$ for random or more-divergent data ($f_0 \leqslant 1/4$) to 0 for completely identical data ($f_0 = 1$). This range holds for all four trees of figure 1.

(ii) *The binary tree $T_1 = ((12)3)$*

The branch lengths are t_0 and t_1 (figure 1). Let a be the probability that the nucleotides at a site are different at

Table 2. *MLE and optimum log likelihood under tree T_0*

data	MLE of a (or t)	optimum log-likelihood ℓ_0/n
if $f_0 > 1/4$	$\hat{a} = \frac{1}{2} - \frac{1}{2}\sqrt{(4f_0 - 1)/3}$ or $\hat{t} = -1/4 \log\{(4f_0 - 1)/3\}$	$f_0 \log\{f_0\} + (1 - f_0) \log\{(1 - f_0)/3\}$
if $f_0 \leqslant 1/4$	$\hat{a} = \frac{1}{2}$ or $\hat{t} = \infty$	$-\log\{4\} = -1.386$

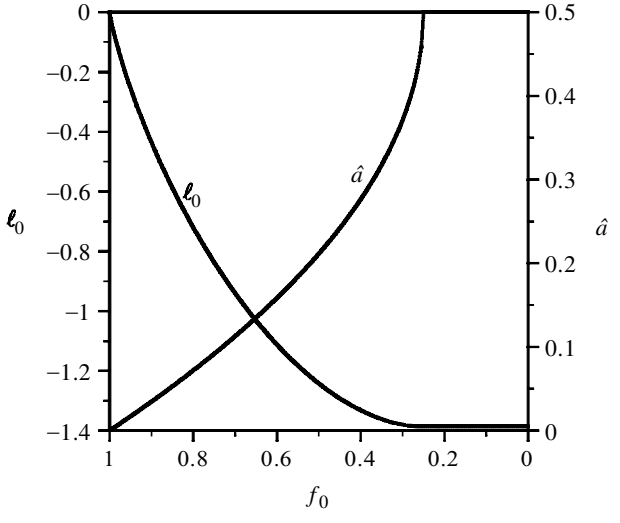

Figure 2. MLE of branch length a and per-site log-likelihood value for the star tree T_0 as a function of f_0 (see table 2).

nodes 0 and 1 in tree T_1, and b be the probability that a site is different at nodes 0 and 3 (figure 1).

$$\left.\begin{aligned} a &= (1 - e^{-2t_1})/2 \\ b &= (1 - e^{-2(2t_0 + t_1)})/2 \end{aligned}\right\}, \tag{14}$$

with $0 \leqslant a \leqslant b \leqslant 1/2$.

By using the pulley principle of Felsenstein (1981), the root of the tree can be placed at node 0 in the likelihood calculation. The probabilities of observing the four site patterns under tree T_1 are then given as follows

$$\left.\begin{aligned} p_0(t_0, t_1) &= a^2 b + (1-a)^2(1-b) = 1 - 2a - b + a^2 + 2ab \\ &= \tfrac{1}{4} + \tfrac{1}{4}e^{-4t_1} + \tfrac{1}{2}e^{-4(t_0 + t_1)} \\ p_1(t_0, t_1) &= a^2(1-b) + (1-a)^2 b = a^2 - 2ab + b \\ &= \tfrac{1}{4} + \tfrac{1}{4}e^{-4t_1} - \tfrac{1}{2}e^{-4(t_0 + t_1)} \\ p_2(t_0, t_1) &= a(1-a)(1-b) + a(1-a)b = a - a^2 \\ &= \tfrac{1}{4} - \tfrac{1}{4}e^{-4t_1} \\ p_3(t_0, t_1) &= p_2 \end{aligned}\right\}. \tag{15}$$

Note that $p_0 \geqslant \max(p_1, p_2, p_3)$ and $p_1 \geqslant p_2 = p_3$. The likelihood function, given in equation (11), is

$$\begin{aligned} \ell_1/n = {} & f_0 \log\{1 - 2a - b + a^2 + 2ab\} + f_1 \log\{a^2 - 2ab + b\} \\ & + (f_2 + f_3) \log\{a - a^2\}. \end{aligned} \tag{16}$$

MLEs of parameters a and b (or t_0 and t_1) can be found by setting $p_0 = f_0$ and $p_1 = f_1$ if a root exists, that is, if $f_0 \geqslant f_1$ and $f_1 > (1 - f_0)/3$. When that condition is not satisfied, one or both parameters will be at the boundary of the parameter space. The solution is given in table 3.

For estimation of t_0 and t_1 in T_1, f_0 and f_1 are sufficient statistics. The sample space specified by f_0 and f_1 is a triangle, since $f_0 \geqslant 0$, $f_1 \geqslant 0$ and $f_0 + f_1 \leqslant 1$ (figure 3). The space is partitioned into four regions A, B, C and D (figure 3; table 3). In region A, the MLEs are inside the parameter space ($0 < \hat{t}_0, \hat{t}_1 < \infty$), and tree T_1 has a higher likelihood than tree T_0. Note that the MLE of t_1 is the same as the LS estimate but the MLE of t_0 is different from the LS estimate. In region B, ML gives $\hat{t}_0 = 0$ and tree T_1 converges to T_0. In region C, the data are more divergent than random sequences, and $\hat{t}_1 = \infty$ and $\hat{t}_0$ is undefined, with tree T_1 converging to T_0. Region D corresponds to data in which sequences 1 and 2 are very similar and both are very different from sequence 3; in this region, $\hat{t}_1 < \infty$ and $\hat{t}_0 = \infty$, and tree T_1 has a higher likelihood score than T_0. Note also that the condition $f_1 > (f_2 + f_3)/2$ is necessary but not sufficient for T_1 to be better than T_0. In region C_2, that condition is satisfied but T_1 converges to T_0. Similarly, the condition $f_1 > \max(f_2, f_3)$ is necessary but not sufficient for tree T_1 to be the ML tree, as that condition may be satisfied in region C_2, where none of three binary trees is better than the star tree T_0. In such data, the sequences are more divergent than random sequences.

Probabilities of site patterns under trees T_2 and T_3 can be calculated similarly to equation (15) by considering the symmetry of the problem. These are summarized in table 1, where the p_0, p_1, p_2 and p_3 functions are defined in equation (15), with branch lengths t_0 and t_1 defined on the specific tree topology under consideration (see figure 1). MLEs of branch lengths and optimum-likelihood values for trees T_2 and T_3 can be obtained from table 3 by considering the symmetry of the problem. The sufficient statistics for estimation of branch lengths in tree T_2 are f_0 and f_2 and the $f_0 - f_2$ space can be partitioned for T_2 similarly to figure 3. For tree T_3, the $f_0 - f_3$ space can be similarly partitioned. For the model considered in this paper, at most two binary trees can both have higher likelihood values than the star tree T_0.

For the example data set, the estimate of branch length is $\hat{a} = 0.052266$ or $\hat{t} = 0.055205$, with $\ell_0 = -0.5835$ for the star tree T_0 (table 2). For tree T_1, the estimates are $\hat{a} = 0.046276$ and $\hat{b} = 0.064129$ or $\hat{t}_0 = 0.010036$ and $\hat{t}_1 = 0.048559$, with $\ell_1 = -0.5818$. Both T_2 and T_3 converge to the star tree T_0. Tree T_1 is the ML tree.

(b) *Estimation of tree topology and partition of sample space*

The sample space for phylogeny estimation is specified by the three variables f_1, f_2 and f_3, (since $f_0 = 1 - f_1 - f_2 - f_3$). As $f_1 \geqslant 0, f_2 \geqslant 0, f_3 \geqslant 0$ and $f_1 + f_2 + f_3 \leqslant 1$, the sample space is the tetrahedron $OABC$ in figure 4. Each data set corresponds to a point in this space. Each point in this space

Table 3. *MLEs and optimum log likelihood under tree T_1*

data	MLEs	optimum log likelihood ℓ_1/n
A: $T_1 > T_0$	$\hat{a} = \frac{1}{2} - \frac{1}{2}\sqrt{2(f_0+f_1)-1}$	$f_0 \log\{f_0\} + f_1 \log\{f_1\} + (1-f_0-f_1)\log\{(1-f_0-f_1)/2\}$
$f_0 > f_1$ and $f_1 > (1-f_0)/3$	$\hat{b} = \frac{1}{2} - \frac{1}{2}(f_0-f_1)/\sqrt{2(f_0+f_1)-1}$	
	or	
	$\hat{t}_0 = -\frac{1}{4}\log\{f_0-f_1\} + \frac{1}{4}\log\{2(f_0+f_1)-1\}$	
	$\hat{t}_1 = -\frac{1}{4}\log\{2(f_0+f_1)-1\}$	
B: $T_1 = T_0$	$\hat{a} = \hat{b} = \frac{1}{2} - \frac{1}{2}\sqrt{(4f_0-1)/3}$	$f_0 \log\{f_0\} + (1-f_0)\log\{(1-f_0)/3\}$
$f_0 > 1/4$ and $f_1 \leqslant (1-f_0)/3$	or	
B_1: $f_0 + f_1 \leqslant \frac{1}{2}$	$\hat{t}_0 = 0$, $\hat{t}_1 = -\frac{1}{4}\log\{2(f_0+f_1)-1\}$	
B_2: $f_0 + f_1 > \frac{1}{2}$		
C: $T_1 = T_0$	$\hat{a} = \hat{b} = \frac{1}{2}$	$-\log\{4\} = -1.386$
$f_0 \leqslant 1/4$ and $f_0 + f_1 \leqslant \frac{1}{2}$	or	
C_1: $f_1 \leqslant (1-f_0)/3$	$\hat{t}_0 =$ undefined, $\hat{t}_1 = \infty$	
C_2: $f_1 > (1-f_0)/3$		
D: $T_1 > T_0$	$\hat{a} = \frac{1}{2} - \frac{1}{2}\sqrt{2(f_0+f_1)-1}, \hat{b} = \frac{1}{2}$	$(f_0+f_1)\log\{(f_0+f_1)/2\}$
$f_0 \leqslant f_1$ and $f_0 + f_1 > \frac{1}{2}$	or	$+ (1-f_0-f_1)\log\{(1-f_0-f_1)/2\}$
	$\hat{t}_0 = \infty$, $\hat{t}_1 = -\frac{1}{4}\log\{2(f_0+f_1)-1\}$	

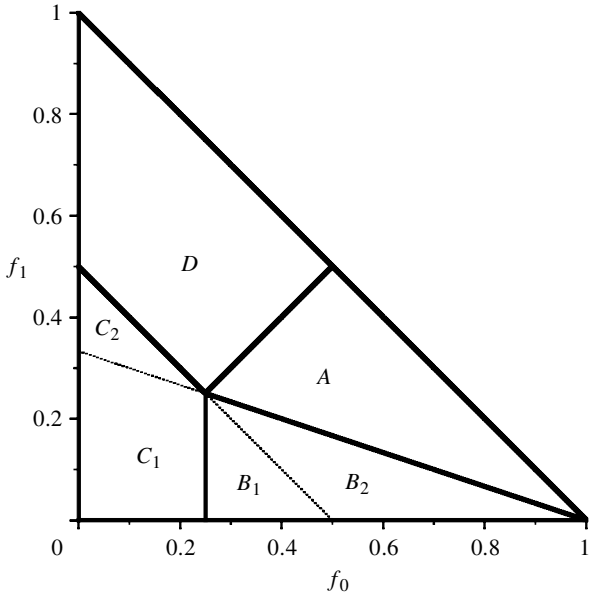

Figure 3. Partition of the sample (data) space for estimation of branch lengths in tree T_1. The data are represented by f_0 and f_1, with $f_2 + f_3 = 1 - (f_0 + f_1)$. MLEs of branch lengths and likelihood values are given in table 3. *A*, $f_0 > f_1, f_1 > (1-f_0)/3$; *B*, $f_0 > 1/4, f_1 \leqslant (1-f_0)/3$; *C*, $f_0 \leqslant 1/4, f_0 + f \leqslant 1/2$; *D*, $f_0 \leqslant f_1$, $f_0 + f_1 > 1/2$.

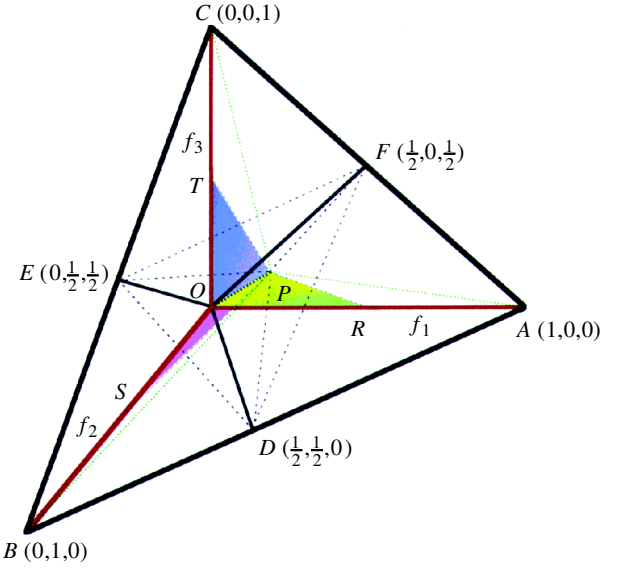

Figure 4. Partition of the sample space for tree topology estimation. The sample space is the tetrahedron *OABC*, specified by the three axes f_1, f_2 and f_3. The origin is at $O(0, 0, 0)$, with point $P(1/4, 1/4, 1/4)$ inside the tetrahedron. The sample space is partitioned into four regions (subspaces), corresponding to the four trees T_0, T_1, T_2 and T_3. If the data fall within the region for T_i, T_i will be the ML tree. The subspace for T_0 is the line segment *OP* plus the tetrahedron *PDEF*. The subspace for T_1 is a contiguous block *OPFAD*, consisting of three tetrahedrons *OPAD*, *OPAF* and *PDAF*. The subspaces for T_2 and T_3 are *OPDBE* and *OPECF*, respectively. The probability spaces are superimposed onto the sample space; line segment *OP* for T_0, triangle *OPR* for T_1, triangle *OPS* for T_2, and triangle *OPT* for T_3. They are indicated by different colours. The coordinates of points *R*, *S* and *T* are $R(1/2, 0, 0)$, $S(0, 1/2, 0)$ and $T(0, 0, 1/2)$.

also corresponds to a possible data set, apart from the discreteness of the real data due to the finite number of sites (n). Results of table 3 can be used to work out the ML tree (as well as the branch lengths and optimum-likelihood value) for any given data outcome (f_1, f_2, f_3). The results are summarized in table 4.

Estimation of the phylogeny is equivalent to partitioning or colouring the sample space of figure 4. For each point in the sample space, the ML tree is identified in table 4. Suppose we use four colours for the four trees T_0, T_1, T_2 and T_3, and colour each point in the sample space with the colour for the ML tree. Then the tetrahedron *OABC* will be partitioned into four contiguous coloured subspaces. If the data fall within the subspace for tree T_i, T_i will be the ML tree, $i = 0, 1, 2, 3$.

Table 4. *ML estimation of tree topology*

data	ML tree
if	T_0
$f_1 \geqslant \max(f_2, f_3)$ and $f_2 + f_3 \geqslant 1/2$ or	
$f_2 \geqslant \max(f_3, f_1)$ and $f_3 + f_1 \geqslant 1/2$ or	
$f_3 \geqslant \max(f_1, f_2)$ and $f_1 + f_2 \geqslant 1/2$ or	
$f_1 = f_2 = f_3$	
otherwise if	
$f_1 > f_2$ and $f_1 > f_3$	T_1
$f_2 > f_3$ and $f_2 > f_1$	T_2
$f_3 > f_1$ and $f_3 > f_2$	T_3
$f_1 = f_2 > f_3$	$T_1 = T_2$
$f_2 = f_3 > f_1$	$T_2 = T_3$
$f_3 = f_1 > f_2$	$T_3 = T_1$

The subspace for T_0 consists of the line segment *OP* and the tetrahedron *PDEF*. The subspace for T_1 is the block *OPFAD*, and consists of three tetrahedrons *OPAD*, *OPAF* and *PDAF*. The subspaces for T_2 and T_3 are the blocks *OPDBE* and *OPECF*, respectively (figure 4).

(c) *Parameter space of the tree topology estimation problem*

The parameter (probability) space for a tree topology is the space of all possible values of parameters (branch lengths) in that tree. This can be superimposed onto the f_1–f_2–f_3 space, with the observed site pattern frequencies (f_is) given by the expected site pattern probabilities (p_is) under the tree. The parameter space for the star tree T_0 is the line segment *OP* in figure 4, since $0 \leqslant p_1 = p_2 = p_3 \leqslant 1/4$. As tree T_0 has only one branch length, its parameter space is one-dimensional. The parameter space for the binary tree T_1 is the triangle *OPR* in figure 4, specified by $0 \leqslant p_2 = p_3 < p_1 < p_0 = 1 - p_1 - p_2 - p_3$. Any set of values for t_0 and t_1 in T_1 in figure 1 will generate site pattern probabilities (p_0–p_3) corresponding to a point in the triangle *OPR* in figure 4. The parameter spaces for T_2 and T_3 are the triangles *OPS* and *OPT*, respectively. The parameter space for each tree (e.g. triangle *OPR* for T_1) is fully contained within the partitioned sample space for that tree (e.g. the block *OPFAD*) (figure 4), as the ML method is consistent. As pointed out by H. Shimodaira (personal communication), ML estimation of branch lengths in each tree is equivalent to projecting the observed data (f_1, f_2, f_3) onto the probability plane of that tree. It is not clear whether the dimension of the entire parameter space for phylogeny estimation is a meaningful concept.

(d) *Distribution of data and probability of recovering the correct tree*

Suppose that the true tree is T_1, and that the true branch lengths give site pattern probabilities p_1, p_2 and $p_3 = p_2$ (from equation (15)). The point (p_1, p_2, p_3) is in the triangle *OAR* in figure 4. Then most data samples will be concentrated around that point. The probability density, that is the probability of observing any data outcome (f_1, f_2, f_3), is given by the multinomial probability (equation (9)). To plot the density onto the sample space of figure 4 would require a four-dimensional plot, but two profiles are shown in figure 5 for $a = 0.2$, $b = 0.25$ (corresponding to $t_0 = 0.0456$ and $t_1 = 0.2554$) with $n = 200$ sites in the sequence. Figure 5*a* plots the density as a function of f_0 and f_1, superimposed on the partitioned sample space for tree T_1 (see figure 3). The amount of probability density in region *A* gives the probability that tree T_1 is better than T_0 (that is, $\hat{t} > 0$ in tree T_1). Figure 5*b* plots the same

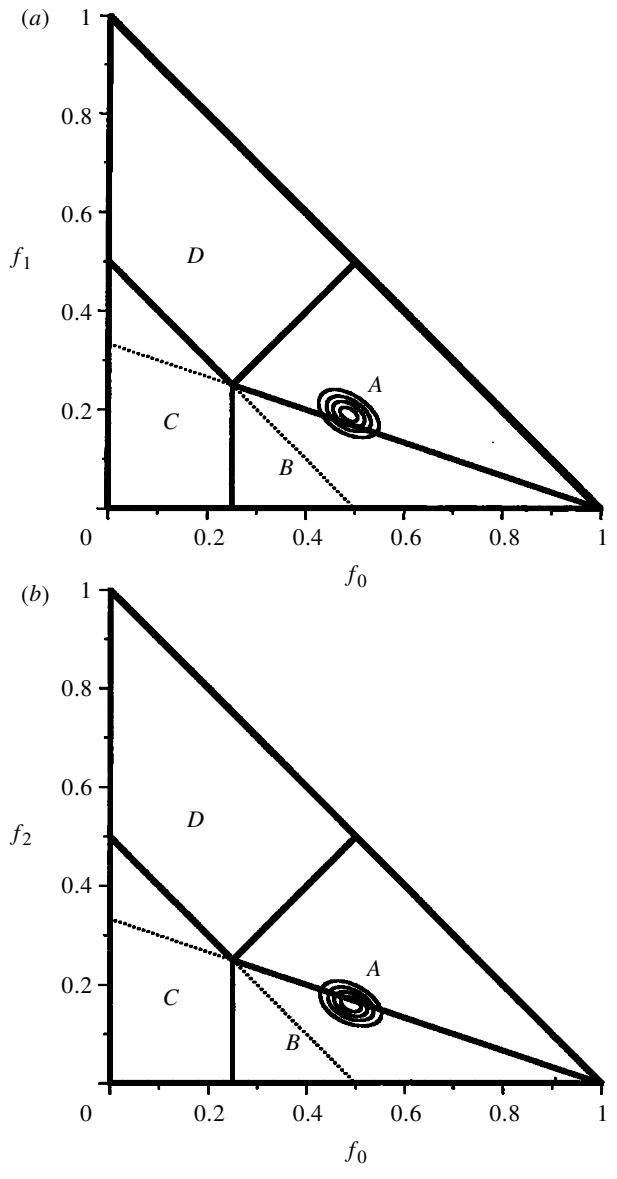

Figure 5. Probability density contours when the true tree is tree T_1 with branch lengths $a = 0.2$ and $b = 0.25$, and sample size (sequence length) $n = 200$ sites. The probabilities of the four site patterns are $p_0 = 0.49$, $p_1 = 0.19$, $p_2 = p_3 = 0.16$ (see equation (15)). The probability of observing any data outcome (n_1, n_2, n_3) or (f_1, f_2, f_3) is given by the multinomial distribution (equation (9)). Two profiles of the probability density are shown. (*a*) The density contours plotted as a function of f_0 and f_1, superimposed on the partitioned sample space for tree 1 (see figure 3). The density is centred around the point $f_0 = 0.49$, $f_1 = 0.19$. For most data samples, tree T_1 will have a higher likelihood than tree T_0. (*b*) The density contours plotted as a function of f_0 and f_2, superimposed on the partitioned sample space for tree 2. The density is concentrated around the point $f_0 = 0.49$, $f_2 = 0.16$. For a large proportion of data samples, tree T_2 will converge to tree T_0.

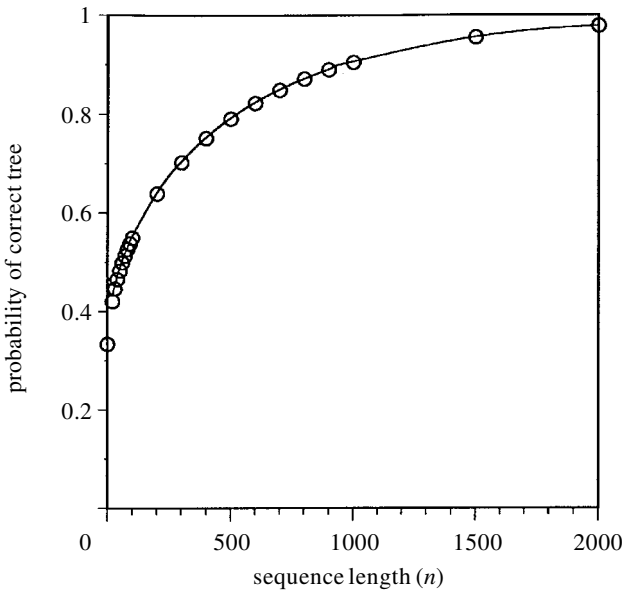

Figure 6. Probability of recovering the correct tree T_1 as a function of the sequence length (n) when the branch lengths in T_1 are $a=0.2$ and $b=0.25$. Data sets are generated by sampling from the multinomial distribution (equation (9)) and the ML tree is determined using table 4 or figure 4. Each point is obtained from 2×10^6 simulations. The curve shows the approximation by equation (17). The probability density for $n=200$ is described in figure 5.

density as a function of f_0 and f_2, superimposed on the partitioned sample space for tree T_2 (see figure 3), and the amount of density in region A gives the probability that $\hat{t}_0>0$ in tree T_2. The density is concentrated in a small area of the sample space. For shorter sequences, the distribution will be more spread out.

Figure 5 does not provide direct estimates of the proportions of data samples falling into each of the four subspaces in figure 4. For $n=200$, these proportions are $P_0=0.2\%$, $P_1=63.9\%$ and $P_2=P_3=17.9\%$, according to computer simulation. In particular, the proportion of data sets (i.e. the amount of probability density), P_1, that fall within the T_1 subspace in figure 4 is also the probability that the true tree is recovered by ML. This is shown in figure 6 for different sample sizes n. Following Zharkikh & Li (1992), that probability can be approximated by

$$P_1=\Phi\left(\frac{(p_1-p_2)\sqrt{n-[1/(p_1-p_2)]}-\sqrt{p_2/\pi}}{\sqrt{p_1+(1-1/\pi)p_2}}\right), \qquad (17)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The approximation slightly overestimates the probability, but the accuracy is high for large n. When $n=200$, the approximation gives 0.641 while the simulation result is 0.639. From equation (17), the sample size required to achieve a specified probability P of recovering the correct tree can be approximated as

$$n_P=\left[\frac{\sqrt{p_2/\pi}+z_P\sqrt{p_1+[1-(1/\pi)]p_2}}{p_1-p_2}\right]^2+\frac{1}{p_1-p_2}, \qquad (18)$$

where z_P is the one-tail standard normal variate corresponding to probability P (Zharkikh & Li 1992).

5. DISCUSSION

(a) *Generality of the problem*

The main feature that is shared by the simple problem considered in this paper and phylogeny estimation in general is that different tree topologies lie in different parameter (probability) spaces and have different likelihood functions (figure 4). Furthermore, the parameter spaces for all possible trees are embedded in a general multinomial distribution. However, large trees have many interior nodes, and the statistical support for individual nodes is of interest as well as support for the entire phylogeny. With more species, there also exist the intricate relationships among possible tree topologies.

Estimation of the tree topology is equivalent to partitioning or colouring the sample space, and different tree reconstruction methods may be considered different partitioning or colouring schemes. For more general cases, it is not entirely clear whether each tree topology has a contiguous partition of the sample space. If the partitioned subspace for the correct tree contains a larger proportion of the probability density, the reconstruction method will have a higher probability of recovering the correct tree. The problem discussed in this paper is highly symmetrical, and when the data sample falls outside the subspace for the true tree, it has equal chance of falling into the two subspaces for the two wrong trees. With more species or more complex substitution models, the partitioning may be asymmetrical, or the probability density may be highly skewed towards one particular wrong tree (for examples, see Yang 1997; Bruno & Halpern 1999).

(b) *The case of four character states*

When four nucleotides are considered under the JC69 model instead of binary characters considered above, there exist five site patterns: *xxx*, *xxy*, *yxx*, *xyx* and *xyz*, where x, y and z are any three different nucleotides (Saitou 1988; Yang 1994). Let the frequencies of those site patterns in the data be f_0, f_1, f_2, f_3, f_4. The probabilities for those site patterns (p_0–p_4) under each tree topology were obtained by Saitou (1988) and Yang (1994). It does not seem possible to obtain MLEs of branch lengths analytically, even for the single branch length in the star tree T_0 (Yang 1994). However, the same conclusion holds that if one of the binary trees (T_1, T_2, T_3) is the ML tree, it is the one corresponding to the largest of (f_1, f_2, f_3). It is not clear under what conditions a binary tree has a higher likelihood than the star tree.

A proof is given here for the statement that T_1 has a higher likelihood than T_2 if tree T_2 has a higher likelihood than T_0 and if $f_1>f_2$. The following proof uses the case of binary characters, with the likelihood calculated using equation (11) and the p_is given in equation (15) and table 1 for different trees. The proof applies to the case of four nucleotides, as indicated below, in which case the probabilities are given in Yang (1994, equation 4). Let $t_0^{(i)}$ and $t_1^{(i)}$ denote the two branch lengths in the binary tree T_i ($i=1, 2, 3$). Let ℓ_i^* be the optimum log likelihood obtained at the MLEs of branch lengths, $\hat{t}_0^{(i)}$ and $\hat{t}_1^{(i)}$, in tree T_i. Let the likelihood value for T_1 at $t_0^{(1)}=\hat{t}_0^{(2)}$ and $t_1^{(1)}=\hat{t}_1^{(2)}$ be $\ell_1^\sharp$. It

follows that $\ell_1^\sharp > \ell_2^*$; that is, the likelihood of T_1 is higher than the likelihood of T_2 when both are calculated at the MLEs of branch lengths from T_2. This is the case because equation (15) suggests that $p_1 > p_2$ holds for those branch lengths, which implies that $f_1 \log(p_1/p_2) > f_2 \log(p_1/p_2)$ or $f_1 \log p_1 + f_2 \log p_2 > f_1 \log p_2 + f_2 \log p_1$, so that $\ell_1^\sharp - \ell_2^* = f_1 \log p_1 + f_2 \log p_2 - (f_1 \log p_2 + f_2 \log p_1) > 0$. Note that when ℓ_1 and ℓ_2 are calculated using the same branch lengths, only site patterns *xxy* and *yxx* contribute differently to the two likelihoods, while other patterns (*xxx* and *xyx* in the case of binary characters (see table 1) and *xxx*, *xyx* and *xyz* in the case of four nucleotides (Yang 1994)) make the same contributions. Since the optimum branch lengths for T_2 may not be optimal for T_1, we have $\ell_1^* \geqslant \ell_1^\sharp > \ell_2^*$.

Solution to the case of binary characters is already given in table 4 and figure 4. For nucleotides with four states, the boundary conditions are not determined yet. If T_2 converges to T_0 ($\ell_2^* = \ell_0^*$) and if $f_1 > f_2$, T_1 may either converge to T_0 or have a higher likelihood than T_0. The proof above means that if a binary tree is the ML tree, it must be the one corresponding to the largest of f_1, f_2 and f_3. However, it is not known under what conditions T_0 is the ML tree. With nucleotide data, numerical calculations (not shown) suggest that it is possible for all three binary trees to have higher likelihood scores than the star tree, whether or not they are equally good. The sample space is four-dimensional and the probability space for each binary tree is two-dimensional. Partition of the sample space seems even more interesting than the case of binary characters.

This paper benefited from discussions with many colleagues over the years. In particular, I thank Peter Beerli, Nick Goldman, Hidetoshi Shimodaira and Mike Steel. This study is supported by Biotechnology and Biological Sciences Research Council grant 31/MMI09806. In addition, J. S. Rogers independently derived ML estimates of branch lengths in the binary trees under the model.

REFERENCES

Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. 1982 Mitochondrial DNA sequences of primates, tempo and mode of evolution. *J. Mol. Evol.* **18**, 225–239.

Bruno, W. J. & Halpern, A. L. 1999 Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.* **15**, 564–566.

Cavalli-Sforza, L. L. & Edwards, A. W. F. 1967 Phylogenetic analysis: models and estimation procedures. *Evolution* **21**, 550–570.

Chang, J. 1996 Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.* **134**, 189–215.

Edwards, A. W. F. 1970 Estimation of the branch points of a branching diffusion process (with discussion). *J. R. Statist. Soc.* B **32**, 155–174.

Edwards, A. W. F. 1995 Assessing molecular phylogenies. *Science* **267**, 253.

Efron, B., Halloran, E. & Holmes, S. 1996 Bootstrap confidence levels for phylogenetic trees. *Proc. Natl Acad. Sci. USA* **93**, 13 429–13 434.

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

Felsenstein, J. 1985 Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.

Hillis, D. M. & Bull, J. J. 1993 An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**, 182–192.

Huelsenbeck, J. P. 1995 The performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17–48.

Jukes, T. H. & Cantor, C. R. 1969 Evolution of protein molecules. In *Mammalian protein metabolism*, vol. 3 (ed. H. N. Munro), pp. 21–123. New York: Academic Press.

Mau, B. & Newton, M. A. 1997 Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo methods. *J. Comput. Graphical Statist.* **6**, 122–131.

Nei, M. 1987 *Molecular evolutionary genetics*. New York, NY: Columbia University Press.

Neyman, J. 1971 Molecular studies of evolution: a source of novel statistical problems. In *Statistical decision theory and related topics* (ed. S. S. Gupta & J. Yackel), pp. 1–27. New York: Academic Press.

Rannala, B. & Yang, Z. 1996 Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**, 304–311.

Rogers, J. S. 1997 On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Syst. Biol.* **46**, 354–357.

Saitou, N. 1988 Property and efficiency of the maximum likelihood method for molecular phylogeny. *J. Mol. Evol.* **27**, 261–273.

Sober, E. 1988 *Reconstructing the past: parsimony, evolution, and inference*. Cambridge, MA: MIT Press.

Yang, Z. 1994 Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.* **43**, 329–342.

Yang, Z. 1996 Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* **42**, 294–307.

Yang, Z. 1997 How often do wrong models produce better phylogenies? *Mol. Biol. Evol.* **14**, 105–108.

Yang, Z. & Rannala, B. 1997 Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.* **43**, 304–311.

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

Zharkikh, A. & Li, W.-H. 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.

As this paper exceeds the maximum length normally permitted, the author has agreed to contribute to production costs.