Goldman N 2 Yang 4 pp. 407-420, In

Proceedings of the XVII th International Biometric Conference Vol. I : Invited papers International Biometrics Society Hamilton Ontario Canada

Models of DNA Substitution and the Discrimination of Evolutionary Parameters

Nick Goldman[1] and Ziheng Yang[2,3]

[1]Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, U.K. (e-mail: n_goldma@nimr.mrc.ac.uk) and [2]Biometrics Section, Natural History Museum, Cromwell Road, London SW7 5BD, U.K. ([3]current address: College of Animal Science and Technology, Beijing Agricultural University, Beijing 100094, China)

Summary

Models of DNA nucleotide substitution are important for the estimation of phylogenetic trees and for the understanding of the evolution of DNA sequences. Statistical tests of the accuracy of commonly used models often indicate that simple models are inadequate. We have considered problems of assessing the adequacy of models and of distinguishing good models from bad ones, and questions of the levels of confidence we can have in inferences derived using the models. We conclude that it is relatively easy to assess models and to estimate their parameters. The primary aim of most researchers, however, is to reconstruct phylogenetic trees and we have much less confidence in our ability to do this. Unusual results arise, suggesting that less accurate models can give better discrimination between candidate trees.

Introduction

Mathematical models of nucleotide substitution are important for the estimation of phylogenetic trees and for understanding the evolution of DNA sequences. As more sequences are determined, attempts to refine models seem ever more worthwhile. Better models should lead to more-accurate estimates of the evolutionary history of the species concerned and to a better understanding of the forces that affected the evolution of the sequences.

The most successful and widely-used models for DNA nucleotide substitution are continuous time Markov chains. The model is defined by a 4 × 4 matrix Q whose 12 off-diagonal elements represent instantaneous rates of substitution and whose diagonal elements are fixed so that row sums are all zero, permitting solution for probabilities of all possible changes over any time period $t \geq 0$: $P(t) = \exp(tQ)$. Variants of the model are defined by the relationships between the off-diagonal elements. Common models include those denoted JC69,

where all 12 rates are equal (Jukes and Cantor, 1969); K80, with nucleotide transition (A ↔ G, C ↔ T) rates greater than transversion (A, G ↔ C, T) rates by a single factor κ (Kimura, 1980); F81, with substitution rates proportional to the frequencies of the replacement bases (π_i for bases i = A, C, G, T; Felsenstein, 1981); and HKY85, which combines the features of the K80 and F81 models (Hasegawa *et al.*, 1985).

Such models are generally applied independently to each site of DNA sequences, enabling maximum likelihood (ML) inferences to be drawn from sets of aligned sequences. Statistical tests have been devised which permit assessment of these models' accuracy in describing the evolution of DNA sequences, including effects introduced by their relationships on an evolutionary tree (Goldman, 1993). The results of such tests often indicate that the models are unacceptable (e.g. Goldman, 1993; Yang *et al.*, 1994). It has been speculated that the most worrying unrealistic assumption made by these models is that of constancy of substitution rates over all nucleotide sites. This assumption must be unrealistic, for example, for gene sequences coding for products with biological functions.

Recently there has been considerable interest in the use of the gamma distribution to describe differences in rates of evolution at different sites of DNA sequences. The rate at each site is assumed to be a random variable drawn from a gamma distribution. Yang (1993) has shown how this model may be incorporated in ML estimation of phylogenies. In Yang's (1993) formulation, the β parameter of the gamma distribution is a trivial scale factor which may be set equal to the shape parameter α, giving a distribution with mean 1 and variance $1/\alpha$. Values of α less than approximately 0.5 mean the gamma distribution has a 'reverse-J' shape and imply strong rate variation, while values of α larger than 1 or 2 imply a more or less constant rate over sites. Using different values of α, rate variation can be accommodated in a variety of real examples. The substitution models mentioned previously may each be combined with the gamma distribution model of rates across nucleotide sites, to give models denoted JC69+Γ, K80+Γ, etc.

As models become more complex, however, there are more parameters to be estimated from datasets. Naturally, we are concerned about how accurately we may hope to estimate substitution model parameters and — the usual aim of phylogenetic estimation studies — evolutionary trees and their branch lengths (measured as numbers of substitutions on each branch).

This paper looks at the questions of

- how easily and how accurately can we distinguish 'good' models from 'bad' ones?
- which models are good, and how good are they?
- how well can we estimate parameters, trees and branch lengths?

These questions are considered in the context of ML estimation of phylogenies, but some results about models are directly relevant to distance matrix based phylogenetic estimation methods, which rely on the same models.

How Easily and How Accurately Can Good Models be Distinguished from Bad Ones?

2.1 Tests of Adequacy of One Model

Navidi *et al.* (1991) proposed a statistical test of the adequacy of DNA substitution models using a likelihood ratio (LR) statistic, and assumed that (traditional) χ^2 distributions applied. This assumption is so inaccurate as to make the test unreliable, as shown by Goldman (1993) who devised Monte Carlo simulation methods that could provide estimates of the correct distributions. Unfortunately, the analysis of the simulated datasets is very time consuming, and the test procedure is impractical for alignments of more than a very few sequences.

Goldman (1993) identified two difficulties with the χ^2 approximation to the LR statistic distribution. The first is caused by the choice of trees. Goldman (1993) pointed out that it is unclear how this should contribute to calculation of the appropriate number of degrees of freedom (df). Undoubtedly, the choice amongst trees contributes to the likelihood; if, as is generally the case, the likelihood is maximized over all trees, the LR statistic cannot follow the χ^2 distribution as it could if the true tree were known.

Despite this theoretical difficulty, the problem can almost be ignored in practice. It is observed that when the tree is in doubt, the likelihoods of several reasonable trees, including the ML tree and (presumably) the true tree, are very similar. (Ironically, this is the reason why it is so difficult to estimate the true tree with confidence.) Table 1 gives an example, from the analysis of 895 base pair (bp) sequences of mtDNA of human, chimpanzee, gorilla and orangutan (Brown *et al.*, 1982) using various substitution models. Reading along each row of Table 1, columns (0)-(3) give the likelihoods for the four possible trees. It is evident that for a given model all the likelihoods are similar when, as happens in the LR test, compared with the maximum possible log-likelihood (-2104.19). The approximation of the likelihood of the ML tree (or any reasonable tree) for that of the true tree is acceptable.

The second difficulty with the χ^2 approximation concerns a peculiarity of the data. Assuming independence of nucleotide sites, the observed 'site patterns' (patterns of bases across all sequences, observed at each site) are assumed to be a sample from a multinomial distribution. Yet the number of categories of the distribution, 4^s where s is the number of sequences, is often larger than the number of data points (nucleotide sites). Furthermore, for typical aligned DNA sequences most sites are clustered in the four 'constant' categories defined by the occurrence of identical nucleotides in all the sequences. For closely related sequences these four categories can account for over 90% of the data points. A consequence is that we have very many categories with very few or no data points assigned to

them. This 'sparseness' seems to have a drastic effect on the χ^2 approximation (Reeves, 1992; Goldman, 1993).

One potential solution to this problem might be to combine the patterns into larger categories. A method for doing this has been proposed by Yang *et al.* (submitted), but its usefulness is not yet proven. At the moment the Monte Carlo simulation approach still seems necessary, although much time can be saved by not worrying about searching all possible trees for the ML tree for each simulated dataset but simply assuming a reasonable tree when there is any doubt.

2.2 *Comparison of Two Models*

Ritland and Clegg (1987) proposed that nested models could be tested in the traditional manner, using LR statistics and χ^2 distributions for performing significance tests. Goldman (1993) again used Monte Carlo techniques to derive estimates of the true distributions of LR statistics. As with tests of the adequacy of models, uncertainty over the true tree and over how to incorporate the choice amongst trees into calculation of the df seem to count against the χ^2 approximation.

Table 1
Log-likelihoods and Estimates of Parameters for Different Trees under Different Models

Model and Parameter	Tree (0): (H, C, G)	(1): ((H, C), G)	(2): ((H, G), C)	(3): (H, (C, G))	$\ell_{(1)} - \ell_0$
HKY85	-2196.96	-2187.60	-2196.96	-2194.74	9.37
$\hat{\kappa}$	11.89±2.00	12.23±2.13	11.89±2.00	11.55±1.95	
JC69+Γ	-2415.24	-2406.34	-2414.06	-2409.81	8.90
$\hat{\alpha}$	0.75±0.23	1.47±0.74	0.88±0.31	1.11±0.45	
F81+Γ	-2336.34	-2330.22	-2335.77	-2332.29	6.12
$\hat{\alpha}$	0.64±0.19	1.10±0.48	0.72±0.24	0.88±0.33	
K80+Γ (κ = 12)	-2266.70	-2262.59	-2266.70	-2266.59	
$\hat{\alpha}$	0.49±0.14	0.67±0.24	0.49±0.14	0.51±0.15	
HKY85+Γ (κ = 12)	-2176.32	-2173.69	-2176.32	-2176.29	2.63
$\hat{\alpha}$	0.39±0.09	0.46±0.12	0.39±0.09	0.39±0.09	

Data are the 895bp mtDNA sequences of human (H), chimpanzee (C), gorilla (G) and orangutan (O). ((H, C), G, O) (tree 1) is the ML tree under all the models.

We have re-examined the results of Goldman (1993), and found that the χ^2 distribution does appear to give good approximations in the comparison of nested substitution models. Fig. 1 shows an example in which F81 is compared with HKY85 using 6166bp ψη-globin pseudogenes of human (H), chimpanzee (C), gorilla (G) and orangutan (O) (Miyamoto *et al.*, 1987). This comparison is effectively a test of whether it is better to allow an unknown transition/transversion rate ratio parameter κ (HKY85), or to fix κ = 1 (F81) — a traditional test would compare twice the log-likelihood difference ($2\Delta\ell$) with a χ^2 distribution with 1 df. To test this approximation, Monte Carlo samples were generated by 'evolving' the sequences along the tree ((H, C), G, O), using estimates under F81 from the data of branch lengths for this tree. Each of 500 samples was analyzed in the same way as the original data. Fig. 1 indicates the close fit of the χ^2 approximation to the true distribution. In this example the calculated log-likelihood values under the two models are ℓ_0 = -10221.81 under F81 and ℓ_1 = -10130.14 under HKY85. Consequently, $\Delta\ell$ = ~~183.34~~ 91.67, indicating rejection of F81 in favour of HKY85 by a huge margin.

The sparseness of the data does not seem to influence tests comparing two models. The reason might be that the likelihoods under both models are affected in roughly the same way.

When hypotheses differ in only one parameter, as is the case for comparison of F81 with HKY85, the two models can be compared by examining the variance of that parameter, as

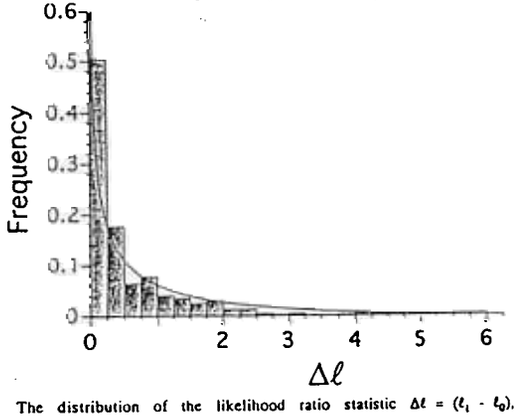

Figure 1. The distribution of the likelihood ratio statistic $\Delta\ell = (\ell_1 - \ell_0)$, for comparison of the F81 and HKY85 models, obtained from 500 Monte Carlo simulations. The 6166bp ψη-globin genes of human (H), chimpanzee (C), gorilla (G) and orangutan (O) are used, with the likelihood calculated using only one tree under both models, i.e., ((H, C), G, O), the ML tree under HKY85. (Results [not shown] obtained using the likelihood of the ML tree under each model for each sample are virtually the same.) The continuous curve is the χ^2 distribution with df = 1, scaled appropriately, which seems to be an acceptable approximation to the simulated distribution. The observed value of the test statistic is $\Delta\ell$ = -10221.81 - (-10130.14) = 91.67; the F81 model is rejected.

estimated by the curvature method (Kendall and Stuart, 1979), relying on the asymptotic Normality of ML estimates. The difficulty caused by the choice of trees can once again be ignored, the justification being the observation that parameter estimates, even more so than likelihood values, are very similar for all trees (see below). For the four-species ψη-globin genes, κ is estimated to be 5.26 ± 0.69 under the HKY85 model. A *z*-statistic can be compared to a standard normal distribution to see if the estimated κ differs from 1: $z = (\hat{\kappa} - 1)/SE(\hat{\kappa}) = 6.16$, $P < 0.0001$, and we conclude that it does.

Comparison between HKY85+Γ and HKY85 is a test of rate constancy over sites, and can be formulated as a test between H_0: $\alpha^{-1} = 0$ and H_1: $\alpha^{-1} \geq 0$ (rate constancy is the limiting case of the gamma distribution with $\alpha \to \infty$). As with tests on κ, tests based on *z*-statistics give very similar results to those based on LR statistics (results not shown).

2.3 Tests of a Molecular Clock

In phylogenetic estimation, a 'molecular clock' is taken to mean equality of substitution rates in the lineages of an evolutionary tree. This is a restriction of the general case where a different rate is permitted for each lineage. As such, tests for the existence of a molecular clock belong in the previous section as they comprise a comparison of two nested models. However, they are prone to additional complications and are considered separately here.

Tests on branch lengths, such as LR tests of the molecular clock, appear more problematic than tests on parameters common to all phylogenetic trees under consideration. Unless we are certain of the true tree, we do not even know which branches exist and tests regarding them will involve unknown uncertainties. We have examined whether the log-likelihood values for the several best trees are similar compared to the log-likelihood difference resulting from the clock assumption. The 6166bp ψη-globin pseudogenes of human (H), chimpanzee (C), gorilla (G), and orangutan (O) are analyzed, and the HKY85 model is assumed. With the assumption of a molecular clock, the position of the root of the tree can be identified (Felsenstein, 1981). The log-likelihood values of all the bifurcating trees are listed in Table 2 under both models, i.e. with and without the molecular clock assumption. As there are $2s - 3$ branch lengths in a unrooted bifurcating tree and $s - 1$ branching times in a rooted bifurcating tree for s species (Felsenstein, 1981), the LR statistic should be compared to a χ^2 distribution with $df = s - 2$. For our example in Table 2, the best unrooted tree without the clock assumption is ((H, C), G, O), with $\ell_1 = -10130.14$, while the best (rooted) tree with the clock assumption is (((H, C), G), O), with $\ell_0 = -10132.47$. This comparison gives $2\Delta\ell = 4.66$, which is not significant. Other plausible trees, such as (((C, G), H, O), give very similar results for this dataset. Removal of the molecular clock assumption therefore does not seem to significantly improve the fit of model to data, i.e., substitution rates are more or less constant along different lineages.

Similar results (not shown) are obtained for mtDNA sequences of the same species. We conclude that the test of a molecular clock can still be performed even if the true tree is unknown. However, as the likelihood values with and without the clock assumption are not very different compared to the likelihood differences caused by the choice of tree, we suggest that likelihood values of the several best trees under both models be examined, no matter whether or not the ML trees under the two models are compatible with each other.

3. What Models Are Good and How Good Are They?

Table 1 shows typical results for the comparison of different substitution models The

Table 2
Test of the Existence of a Molecular Clock

WITHOUT CLOCK			WITH CLOCK		
UNROOTED TREE			ROOTED TREE	ℓ	
(H,C,(G,O))	-10130.14	5.26	(O,(G,(H,C)))	-10132.47	5.26
			(G,(O,(H,C)))	-10167.21	5.25
			((H,C),(G,O))	-10167.21	5.25
			(C,(H,(G,O)))	-10171.47	5.24
			(H,(C,(G,O)))	-10171.47	5.24
(H,G,(C,O))	-10133.20	5.14	(O,(C,(H,G)))	-10135.71	5.14
			(C,(O,(H,G)))	-10171.70	5.11
			((H,G),(C,O))	-10171.70	5.11
			(G,(H,(C,O)))	-10173.68	5.15
			(H,(G,(C,O)))	-10173.68	5.15
(H,O,(C,G))	-10130.32	5.21	(O,(H,(C,G)))	-10134.05	5.21
			((H,O),(C,G))	-10169.56	5.20
			(H,(O,(C,G)))	-10169.56	5.20
			(G,(C,(H,O)))	-10172.07	5.19
			(C,(G,(H,O)))	-10172.07	5.19
(H,C,G,O)	-10133.48	5.16	(H,C,G,O)	-10173.67	5.15

Data are the 6166bp ψη-globin sequences of human (H), chimpanzee (C), gorilla (G) and orangutan (O). Log-likelihood values (ℓ) and estimates ($\hat{\kappa}$) of the transition/transversion rate ratio are shown for all the tree topologies. The HKY85 model is assumed either with or without a molecular clock.

sequences analysed are 895bp mtDNA sequences for human (H), chimpanzee (C), gorilla (G) and orangutan (O). Reading down the columns of Table 1 permits comparison of likelihood values under different models; in particular column (1), for the tree ((H, C), G, O), is most relevant as this is the ML tree under all models considered. Moving down the column, the models get more complex and give better fits to the data (judged by the increasing likelihoods). As JC69+Γ, F81+Γ, and K80+Γ are all special cases of HKY85+Γ, the LR test can be used to test whether these simpler models are acceptable compared to HKY85+Γ. (HKY is also a special case of HKY85+Γ, discussed in more detail below.) For these sequences, the differences in log-likelihoods are 232.65 (df = 4), 156.53 (df = 1) and 88.9 (df = 3) for the three models respectively. All three simpler models appear totally unacceptable.

We have also investigated the effect of using the gamma distribution to describe variation of substitution rates across sites. Likelihood values under two models, HKY85 and HKY85+Γ, are listed in Table 3 for a number of different datasets. The HKY85 model was chosen as previous results (e.g. see above) indicate it is generally the most acceptable model. HKY85 is a special case of HKY85+Γ, and a standard LR test is used to test whether HKY85+Γ is significantly better than HKY85, a test of rate constancy over nucleotide sites.

When we compare $\Delta\ell$ with a critical χ^2 value with df = 1, the difference is extremely significant for the mtDNA sequences, the α- and β-globin genes, the ssRNAs and the glutamine synthetase genes (all $P < 0.01$). We conclude that there is strong evidence of rate variation over the sites of these sequences. On the other hand, for the ψη-globin pseudogenes the difference is barely significant ($0.01 < P < 0.05$). This could be explained

Table 3
Maximum Log-likelihoods With and Without the Assumption of a Gamma Distribution of Rates over Sites

Data[a]	Model 0: HKY85 ℓ_0	Model 1: HKY85+Γ ℓ_1	$\ell_1-\ell_0$
five-species mtDNA[a]	-2665.42	-2632.11	33.32**
four-species mtDNA[a]	-2187.60	-2173.69	13.90**
α- and β-globin genes[b]	-1451.01	-1434.58	16.43**
ssRNA[c]	-5837.58	-5796.18	41.40**
glutamine synthetase gene[d]	-2958.05	-2948.70	9.35**
ψη-globin gene[e]	-10130.14	-10127.36	2.64*

Data are from [a]Brown *et al.* (1982), [b]Yang *et al.* (1994), [c]Navidi *et al.* (1991), [d]Pesole *et al.* (1991) and [e]Miyamoto *et al.* (1987).

* $P < 0.05$; $\chi^2_{0.05}$ (1 df) = 3.84. ** $P < 0.01$; $\chi^2_{0.01}$ (1 df) = 6.63.

by the fact that these are non-coding sequences, which are not subject to pressures of selection which might act differently at different sites, depending on the precise structure and function of a protein.

Judging simply by the log-likelihood differences (e.g. Tables 1, 3), we note that the assumptions about relative rates of different nucleotide substitutions (HKY85 *cf.* JC69, F81, K80) seem more important than the assumptions about rate variation over sites (e.g. HKY85 *cf.* HKY85+Γ). It seems clear, however, that for most sequences neither factor can be neglected.

The results presented so far relate to the relative acceptability of various models. We also consider the goodness of fit of these models, to see if the models give generally acceptable descriptions of the evolution of DNA sequences. Almost all previous studies of models that do not consider rate variation across sites have concluded that these models are inadequate (Goldman, 1993; Yang *et al.*, 1994, submitted). We present here an example where a complex model is found to give a good description of the evolution of coding DNA. The data analysed are the 895bp mtDNA sequences of human, chimpanzee, gorilla and orangutan, using the HKY85+Γ model. The Monte Carlo simulation test of Goldman (1993) was used. Using the maximum likelihood tree and branch lengths for these sequences under the HKY85+Γ model, datasets were simulated conforming to this model. Each simulated dataset was analyzed in the same manner as the original sequences, giving simulated values of the LR statistic $\Delta\ell$ whose distribution is shown in Fig. 2. For the original data, the attained value of $\Delta\ell$ is 69.5. This value falls in the middle of the distribution obtained by simulation, indicating that the HKY85+Γ model gives a good description of the evolution of these sequences.

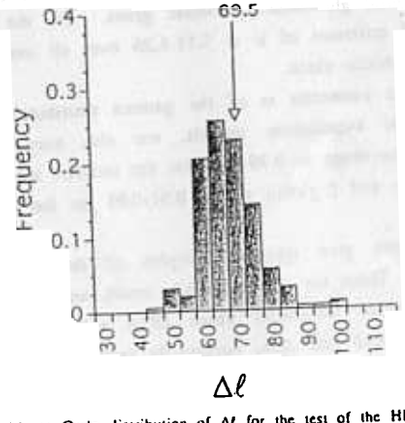

Figure 2. Monte Carlo distribution of $\Delta\ell$ for the test of the HKY85+Γ model applied to the mtDNA for four sequences. The attained value (69.5) falls well within this distribution; the HKY85+Γ model is accepted.

Due in particular to computational constraints, few other datasets have been analysed in this detailed manner. It seems likely, however, that the results presented here are general in nature and that only the most complex models, such as HKY85+Γ which allows for biologically meaningful factors such as transition/transversion bias, unequal base frequencies and rate variation across nucleotide sites, are capable of providing reasonable fits to real datasets. Certainly, each of the three specific factors mentioned is of great importance.

4. How Well Can Parameters be Estimated and Trees be Inferred?

4.1 Parameters of Substitution Processes

We concentrate here on the parameters generally estimated by ML methods in the HKY85 substitution model and in the gamma distribution used to describe the variation of rates over sites. These parameters are κ, which affects the relative rates of transition and transversion substitutions, and α, the shape parameter of the gamma distribution. Both parameters are constrained only to be greater than zero.

First we consider the variation in estimates of κ and α (and their standard errors) over different trees. Table 1 shows typical results, from the mtDNA sequences for human, chimpanzee, gorilla and orangutan. We note that estimates of κ for the HKY85 model are almost constant over all different trees. This is also the case for other datasets analyzed; the range of κ over all trees is 8.66-9.39 for mtDNA sequences of the above four species plus gibbon, 1.47-1.52 for the α- and β-globin genes, 1.78-1.80 for the ssRNAs and 0.96-0.97 for the bacterial glutamine synthetase genes. For the ψη-globin genes analysed in Table 2, the range of estimates of κ is 5.11-5.26 over all trees and both with and without the assumption of a molecular clock.

The estimates of the parameter α of the gamma distribution, though more variable over trees under less-realistic substitution models, are also very stable under the HKY85+Γ model. For example, the range is 0.39-0.46 for the mtDNA sequences for four species (Table 1), 0.22-0.29 for the α- and β-globin genes, 0.91-0.95 for the ssRNAs and 0.51-0.68 for the ψη-globin genes.

Tables 1 and 3 also give typical examples of the standard errors associated with estimates of κ and α. These are all reasonably small, and this gives us confidence that we are able to make accurate inferences about DNA substitution processes. Studies using simulated data (which, unlike real data, necessarily conform to the models under which they are analysed) confirm the findings about both the variation in estimates across trees and the size of error estimates (results not shown).

4.2 Branch Lengths of Trees

Different models most often give the same best tree for given data, and it is interesting to see whether the estimates of branch lengths are also stable under different models. Table 4 lists the estimates of branch lengths for the ML tree for the mtDNA sequence data for four species, ((human, chimpanzee), gorilla, orangutan). If we take estimates from HKY85+Γ model as the correct values, we see that all simpler models of nucleotide substitution will lead to underestimated branch lengths. The underestimation caused by ignoring variation in substitution rates over sites (HKY85) is most serious. In all cases the bias is worse for longer branches; for datasets with more-distantly related species, such errors are much more pronounced. The results here are consistent with findings from computer simulations on the estimation of sequence divergence (e.g. Gillespie, 1986; Tamura, 1992).

4.3 Estimation of Trees and the Discriminating Power of Models

Yang et al. (1994) considered the difference in log-likelihood between the ML tree and the star tree, $\ell_{(1)} - \ell_0$, and found that more complex (and reasonable) models invariably gave lower values than simpler and inappropriate models (e.g. Table 1). In particular, adding the gamma distribution causes substantial reduction in $\ell_{(1)} - \ell_0$. This appears to suggest that discriminating power decreases as models become more complex although, as noted by Yang et al. (1994) and shown by Yang (in press), $\ell_{(1)} - \ell_0$ is not a measure of the ML tree's reliability; one or more wrong trees can often be significantly better supported than the

Table 4
Maximum Likelihood Estimates of Branch Lengths under Different Models

Model	Branch Length (Ratio of Given Model to HKY85+Γ)*				
	→ C	H̄C̄ ↔ GO	→ G	→ O	
HKY85	0.0436 (0.83)	0.0522 (0.81)	0.0191 (0.84)	0.0529 (0.78)	0.1535 (0.56)
JC69+Γ	0.0438 (0.83)	0.0520 (0.81)	0.0195 (0.86)	0.0526 (0.78)	0.1585 (0.61)
F81+Γ	0.0446 (0.85)	0.0534 (0.83)	0.0192 (0.85)	0.0543 (0.81)	0.1579 (0.65)
K80+Γ	0.0482 (0.92)	0.0572 (0.88)	0.0193 (0.85)	0.0602 (0.89)	0.1915 (0.79)
HKY85+Γ	0.0525	0.0644	0.0227	0.0675	0.2416

Based on the analysis of the 895bp mtDNA sequences from human (H), chimpanzee (C) gorilla (G) and orangutan (O).

*Branch lengths represent the expected number of nucleotide substitutions pe site. Numbers in parentheses are the ratio of branch length under the present mode to that under the best model, HKY85+Γ.

star tree. In addition, it has been noted that the log-likelihood difference between the ML tree and the second best tree, $\ell_{(1)} - \ell_{(2)}$, also decreases as the model used for data analysis becomes more complex and accurate. As the likelihood method discriminates among the trees by comparing their likelihood values, this too might suggest that the ability to make a confident estimate of the true tree is reduced as more-complex models are used.

However, these tentative conclusions would be overturned if we could show that the distributions of the statistics are such that smaller values are in fact more easily distinguished from their critical values. If we denote the log-likelihood of the true tree by ℓ_1, and those of other trees by ℓ_i, $i > 1$, then we are interested in the proportion of samples in which $\ell_1 > \max_{i>1}(\ell_i)$, as these are precisely the cases in which the true tree is correctly estimated. For real data we do not know which is the true tree and so ℓ_1 is not known. This is why $\ell_{(1)} - \ell_{(2)}$ is the statistic of interest for real data, but its comparison with its distribution under some model tells us nothing directly about the question of discrimination between candidate trees: it is simply a (somewhat unusual) test of that model — certainly of no interest when the data are produced by simulation. Instead, it is appropriate to consider the distributions of $\ell_1 - \ell_i$: in particular, we choose for convenience a specific 'reasonable' alternative tree, $i = 2$, and consider $\ell_1 - \ell_2$ in the expectation that realised values of $\ell_{(1)} - \ell_{(2)}$ will often in fact equal $\ell_1 - \ell_2$.

We simulated data using the tree and branch lengths that are ML estimates when the four sequence 895bp mtDNA sequences for human, chimpanzee, gorilla and orangutan are analysed using the HKY85+Γ model. Other parameter values were also taken from this analysis: π_T = 0.254, π_C = 0.331, π_A = 0.311, π_G = 0.104; κ = 12.23; α = 0.46. Each simulated dataset was analysed under a variety of models; the whole procedure was repeated for 100 simulations. Table 5 shows some results from this simulation study.

We note that the mean values of the statistics $\ell_{(1)} - \ell_0$ and $\ell_{(1)} - \ell_{(2)}$ decrease as the model becomes more complex, in agreement with findings from real datasets (e.g. Table 1 and Yang *et al.*, 1994, submitted). Of more interest, we note also that the mean values of $\ell_1 - \ell_0$ and $\ell_1 - \ell_2$ also decrease as models become more complex.

The star tree is a special case of any more general tree (with all internal branch lengths equal to zero), and so the statistic $\ell_1 - \ell_0$ must always be non-negative; for real data, we may safely assume it will be positive and that the star tree will never be estimated using ML methods. This is further reason to discard $\ell_1 - \ell_0$ and $\ell_{(1)} - \ell_0$ as measures of discriminating power (contrary to some previous suggestions). The same argument does not apply to $\ell_1 - \ell_2$.

Means alone are not an adequate measure of discriminating power. We have tabulated the standard deviation (SD) of each distribution of $\ell_1 - \ell_2$, and x = mean/SD — assuming that the log-likelihood difference follows a Normal distribution, the probability P_+ that $\ell_1 - \ell_2$ is positive (i.e. the probability that the true tree is correctly discriminated from the

'second-choice' tree) is then given by the standard Normal cumulative distribution function $\Phi(x)$.

For the statistic $\ell_1 - \ell_2$, we see that values of mean/SD (and of Φ[mean/SD]) decrease as the complexity of the model increases. This is of some concern. It means that as the model used to analyse the data becomes more complex (becoming the correct model in the final row of Table 5), we are less and less able to discriminate between the true tree and the second-choice tree.

5. Discussion

5.1 Adequacy of Models

The results presented in this paper demonstrate that the most complex models now available are probably just about statistically acceptable, in terms of their overall 'goodness of fit'. Certainly each added level of complexity studied, modelling the biological features of unequal base frequency, transition/transversion bias and rate variation across nucleotide sites, is important. It seems that the amounts of data (i.e. number of aligned sequences

Table 5

Log-likelihood Differences Between Trees under Different Models

	Mean (SD)				Mean/SD (Φ)
Model	$\ell_{(1)}-\ell_0$	$\ell_{(1)}-\ell_{(2)}$	$\ell_1-\ell_0$	$\ell_1-\ell_2$	$\ell_1-\ell_2$
JC69	19.19 (7.24)	11.07 (8.16)	18.95 (7.52)	13.43 (8.50)	1.58 (0.943)
F81	17.68 (6.91)	10.57 (7.67)	17.41 (7.20)	12.60 (8.05)	1.57 (0.942)
K80	11.65 (5.81)	8.52 (5.99)	11.54 (5.95)	9.66 (6.10)	1.58 (0.943)
HKY85	10.23 (5.43)	7.82 (5.51)	10.11 (5.59)	8.66 (5.70)	1.52 (0.936)
JC69+Γ	9.00 (4.28)	6.44 (4.90)	8.82 (4.50)	7.32 (5.10)	1.44 (0.925)
F81+Γ	8.06 (3.97)	6.01 (4.37)	7.83 (4.21)	6.56 (4.68)	1.40 (0.919)
K80+Γ	3.12 (2.65)	2.89 (2.64)	3.08 (2.69)	2.93 (2.71)	1.08 (0.860)
HKY85+Γ	2.76 (2.40)	2.57 (2.38)	2.70 (2.45)	2.55 (2.48)	1.03 (0.849)

Results from 100 simulations using the HKY85+Γ model and the tree and other parameter estimates from analysis of 895bp mtDNA sequences of human (H), chimpanzee (C), gorilla (G) and orangutan (O). Maximum values of the log-likelihood (ℓ) were found under different models and means and SDs calculated for distributions of differences of log-likelihoods: $\ell_{(1)}$ — ML tree, ℓ_0 — star tree, $\ell_{(2)}$ — tree with second highest likelihood, ℓ_1 — true tree ((H, C), G, O), ℓ_2 — specific other tree ((C, G), H, O). Φ is the standard Normal distribution cumulative distribution function for mean/SD for $\ell_1 - \ell_2$.

and length of those sequences) that are currently available are adequate for indicating which models are best, and how good they are.

5.2 Parameter Estimation

It seems certain that we can get reasonable estimates of parameter values using the amounts of data that are currently available. The choice of phylogenetic tree for a given dataset has been shown to make little difference to estimates of parameters of the substitution process model. In the cases where it makes most difference, this difference is smallest when a good model of nucleotide substitution is used. Finally, given good estimates of phylogenetic trees (but see below) it seems that we can get reasonable estimates of the length of branches of those trees so long as a sufficiently complex model is used. As noted above, current data are generally sufficient to indicate which models are good with little difficulty.

5.3 Discrimination of Phylogenetic Trees

Previously, everything from intuition to empirical evidence via theory has promoted the use of the most complex models, but the results presented here suggest that even when a complex model is near to the truth we may be more successful at estimating the correct tree if we use over-simple models. Nothing in ML theory precludes this, but it goes against both our intuition and our experience of other applications of statistical modelling. If correct, it becomes difficult to advise practicing biologists to use the 'best' models if their primary aim is to estimate phylogenetic trees.

We note at this point a correspondence with recent studies of distance matrix based approaches to phylogenetic estimation. Simulation studies have shown circumstances in which over-simple models of nucleotide substitution used to calculate pairwise distances between sequences give higher probabilities of estimating the true tree (e.g. Schöniger and von Haeseler, 1993; Tajima and Takezaki, 1994; Schöniger and Goldman, in prep.). This effect seems related to the linearity of the distance measure over the range of distances between sequences simulated using the more complex process and the variance of the distance measure (Tajima and Takezaki, 1994; Goldstein and Pollock, in press; Schöniger and Goldman, in prep.).

A previous empirical study of the effect of models on the accuracy of tree estimation (Yang *et al.*, submitted) considered estimates (related to the bootstrap) of P_+. In that study, the estimated probabilities of selecting the correct tree increased as the complexity of the model increased (presumably towards greater accuracy). Our intuition leads us to hope that these findings are correct, but they are contradicted by the simulation results presented above. Clearly, further work is necessary to reconcile the two sets of findings.

Brown, W.M., Prager, E.M., Wang, A. and Wilson, A.C. (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution* **18**, 225-239.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368-376.

Gillespie, J.H. (1986). Rates of molecular evolution. *Annual Review of Ecology and Systematics* **17**, 637-665.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**, 182-198.

Goldstein, D.B. and Pollock, D.D. (in press). Least squares estimation of molecular distance — noise abatement in phylogenetic reconstruction. *Theoretical Population Biology*.

Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160-174.

Jukes, T.H. and Cantor, C.R. (1969). Evolution of protein molecules. *Pp.* 21-132 in: *Mammalian Protein Metabolism*, vol. 3 (ed. H.N. Munro). New York: Academic Press.

Kendall, M. and Stuart, A. (1979). *The Advanced Theory of Statistics*, vol. 2, 4th ed. London: Charles Griffin.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111-120.

Miyamoto, M.M., Slightom, J.L. and Goodman, M. (1987). Phylogenetic relations of humans and African apes from DNA sequences in the ψη-globin region. *Science* **238**, 369-373.

Navidi, W.C., Churchill, G.A. and von Haeseler, A. (1991). Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Molecular Biology and Evolution* **8**, 128-143.

Pesole, G., Bozzetti, M.P., Lanave, C., Preparata, G. and Saccone, C. (1991). Glutamine synthetase gene evolution: a good molecular clock. *Proceedings of the National Academy of Sciences U.S.A.* **88**, 522-526.

Reeves, J.H. (1992). Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *Journal of Molecular Evolution* **35**, 17-31.

Ritland, K. and Clegg, M.T. (1987). Evolutionary analysis of plant DNA sequences. *American Naturalist* **130**, S74-S100.

Schöniger, M. and von Haeseler, A. (1993). A simple method to improve the reliability of tree reconstructions. *Molecular Biology and Evolution* **10**, 471-483.

Tajima, F. and Takezaki, N. (1994). Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. *Molecular Biology and Evolution* **11**, 278-286.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution* **9**, 678-687.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10**, 1396-1401.

Yang, Z. (in press). Statistical properties of the maximum likelihood method of phylogenetic estimation, with comparison to distance matrix methods. *Systematic Biology*.

Yang, Z., Goldman, N. and Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution* **11**, 316-324.

Yang, Z., Goldman, N. and Friday, A. (submitted). Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Systematic Biology*.