J Mol Evol (1994) 39:306–314

Journal of Molecular Evolution

Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods

Ziheng Yang

Department of Zoology, The Natural History Museum, London SW7 5BD, United Kingdom

Received: 22 December 1993 / Accepted: 9 March 1994

Abstract. Two approximate methods are proposed for maximum likelihood phylogenetic estimation, which allow variable rates of substitution across nucleotide sites. Three data sets with quite different characteristics were analyzed to examine empirically the performance of these methods. The first, called the "discrete gamma model," uses several categories of rates to approximate the gamma distribution, with equal probability for each category. The mean of each category is used to represent all the rates falling in the category. The performance of this method is found to be quite good, and four such categories appear to be sufficient to produce both an optimum, or near-optimum fit by the model to the data, and also an acceptable approximation to the continuous distribution. The second method, called "fixed-rates model," classifies sites into several classes according to their rates predicted assuming the star tree. Sites in different classes are then assumed to be evolving at these fixed rates when other tree topologies are evaluated. Analyses of the data sets suggest that this method can produce reasonable results, but it seems to share some properties of a least-squares pairwise comparison; for example, interior branch lengths in nonbest trees are often found to be zero. The computational requirements of the two methods are comparable to that of Felsenstein's (1981, *J Mol Evol* 17:368–376) model, which assumes a single rate for all the sites.

Key words: Phylogeny — Maximum likelihood — Rate variation over sites — The gamma distribution — Approximate methods

Introduction

Variation of substitution rates across nucleotide sites has long been recognized as a characteristic of DNA sequence evolution, especially for sequences coding for biological products (e.g., Fitch and Margoliash 1967; Fitch and Markowitz 1970; Uzzell and Corbin 1971; Holmquist et al. 1983; Fitch 1986; Kocher and Wilson 1991; Wakeley 1993). There have been many attempts to account for such rate variation in phylogenetic analysis. Two approaches are taken. The first assumes that rates over sites are random variables drawn from a continuous distribution; for example, Nei and Gojobori (1986), Jin and Nei (1990), Li et al. (1990), and Tamura and Nei (1993) used the gamma distribution for rates over sites when they constructed estimators of the distance between two sequences. The second approach uses several categories of rates. The simplest model of this sort assumes that a proportion of sites are invariable while others are changing at a constant rate (e.g., Hasegawa et al. 1985; Palumbi 1989; Hasegawa and Horai 1991). In accounting for the extreme rate heterogeneity of the control region of the human mtDNA, Hasegawa et al. (1993) adopted a three-rate-category model, wherein some sites are assumed to be invariable while others are either moderately or highly variable.

Biologically, a continuous distribution may seem to be more reasonable, and indeed, when fitting several models to the control region of human mtDNAs, Wakeley (1993) found that a two-rate-category model could not fit the data properly, while the fit of a gamma distribution was statistically acceptable. Recently, the gamma distribution was also incorporated into a joint like-

lihood analysis by Yang (1993), a direct extension of the method of Felsenstein (1981), which assumes a single rate for all the sites. However, the algorithm of Yang (1993) involves intensive computation, making it impractical for data sets with more than a few species.

In this paper, we introduce two approximate methods. The first, called the "discrete gamma model," uses several categories to approximate the continuous gamma distribution. Rates over sites are regarded as random variables drawn from a discrete distribution. The advantage of taking a discrete distribution as an approximation to the continuous one is that only one extra parameter is needed. This appears to lead to more efficient estimation, easier interpretation of results, and also better fit of the model to data. The second method, called the "fixed-rates model," predicts rates at specific sites by assuming the star tree and the gamma distribution using the method of Yang and Wang (in press), and then combines sites into several classes according to these predicted rates. Sites in different classes are then assumed to be evolving at known (different) rates when other tree topologies are evaluated.

We will analyze three different data sets to examine the fit to data of the discrete gamma model, and its approximation to the continuous distribution, in order to determine a satisfactory number of categories in the discrete distribution. The continuous gamma model (Yang 1993), the discrete gamma model, the fixed-rates model, and a least-squares method based on pairwise distance estimates will all be applied to evaluate the possible tree topologies. The results will give us some idea about the similarities and differences among those tree estimation methods. However, the performance of these methods with regard to tree reconstruction will be examined more rigorously by using computer simulations.

Methods and Data

The Model of Nucleotide Substitution. In this paper we will use the Markov process model for nucleotide substitution that has been implemented in the DNAML program in J. Felsenstein's PHYLIP package since 1984 (Felsenstein personal communication). The rate matrix for this model is

$$Q = \begin{bmatrix} . & (1+\kappa/\pi_Y)\pi_C & \pi_A & \pi_G \\ (1+\kappa/\pi_Y)\pi_T & . & \pi_A & \pi_G \\ \pi_T & \pi_C & . & (1+\kappa/\pi_R)\pi_G \\ \pi_T & \pi_C & (1+\kappa/\pi_R)\pi_A & . \end{bmatrix} \cdot \mu \quad (1)$$

where Q_{ij} $(i \neq j)$ represents the rate of substitution from nucleotide i to j, with the nucleotides ordered T, C, A, G. The diagonals are specified by the mathematical requirement that row sums of **Q** are zero. The equilibrium distribution is given by π_T, π_C, π_A, π_G, with $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G = 1 - \pi_Y$. $\mu = 1/[4\pi_T\pi_C(1 + \kappa/\pi_Y) + 4\pi_A\pi_G(1 + \kappa/\pi_R) + 4\pi_Y\pi_R]$ is a scale factor, chosen so that the average rate of substitution is 1 when the process is in equilibrium. Time t or branch lengths in the tree are then measured by the expected number of nucleotide substitutions per site. Parameter $\kappa > \max(-\pi_Y, -\pi_R)$ adjusts for the transition/transversion rate bias; a κ larger than 0 will allow transitions to occur with higher probabilities than transversions. The rate matrix was given by Hasegawa and Kishino (1989), and the transition probability matrix, $\mathbf{P}(t) = \exp\{\mathbf{Qt}\}$, was described by Thorne et al. (1992). The model will be designated "F84."

When applied to real data, the fit of F84 is very similar to that of the model of Hasegawa et al. (1985), denoted "HKY85" (results not shown). The transition/transversion rate ratios in the two models are roughly related as $\kappa_{HKY85} = 2\kappa_{F84} - 1$, where κ_{HKY85} is equivalent to α/β in the notation of Hasegawa et al. (1985). Goldman (1993) provides a more accurate formula for this relationship. There is, however, a mathematical difference between these two models: while the rate matrix of the HKY85 model has four distinct eigenroots (Hasegawa et al. 1985), **Q** in Eq. 1 has only three, that is, $\lambda_1 = 0$, $\lambda_2 = -\mu$, $\lambda_3 = \lambda_4 = -(1 + \kappa)\mu$. The corresponding eigenvectors of **Q** are the same as those for HKY85 and are given by Hasegawa et al. (1985). The F84 model therefore involves less computation than HKY85, especially for the algorithm of Yang (1993). Another result of this difference is that a simple formula for estimating the distance between two sequences is available for F84 while there is not for HKY85, for the reasons explained by Yang (in press).

Assume that rates over sites follow a gamma distribution with (given or independently estimated) shape parameter α. Let P and Q be the proportions of sites with transitional and transversional differences respectively in the two sequences. Following, e.g., Jin and Nei (1990) or Tamura and Nei (1993), parameter κ and sequence divergence t, as defined by $t = [4\pi_T\pi_C(1 + \kappa/\pi_Y) + 4\pi_A\pi_G(1 + \kappa/\pi_R) + 4\pi_Y\pi_R] \cdot \mu t$, can be estimated as follows:

$$\hat{\kappa} = a/b - 1 \quad (2)$$

$$\hat{t} = [4\pi_T\pi_C(1 + \hat{\kappa}/\pi_Y) + 4\pi_A\pi_G(1 + \hat{\kappa}/\pi_R) + 4\pi_Y\pi_R] \cdot b \quad (3)$$

where

$$a = \overline{(1+\kappa)\mu t} = \begin{cases} -\log(A)/2, & \text{if } \alpha = \infty \\ \alpha(A^{-1/\alpha} - 1)/2, & \text{if } 0 < \alpha < \infty \end{cases} \quad (4)$$

$$b = \overline{\mu t} = \begin{cases} -\log(B)/2, & \text{if } \alpha = \infty \\ \alpha(B^{-1/\alpha} - 1)/2, & \text{if } 0 < \alpha < \infty \end{cases} \quad (5)$$

and

$$A = \frac{2(\pi_T\pi_C + \pi_A\pi_G) + 2(\pi_T\pi_C\pi_R/\pi_Y + \pi_A\pi_G\pi_Y/\pi_R) \cdot [1 - Q/(2\pi_Y\pi_R)] - P}{2(\pi_T\pi_C/\pi_Y + \pi_A\pi_G/\pi_R)} \quad (6)$$

$$B = 1 - Q/(2\pi_Y\pi_R) \quad (7)$$

The frequency parameters π_T, π_C, π_A, π_G, are estimated using the averages of the observed frequencies in the two sequences. When $\alpha = \infty$, the gamma distribution reduces to the model with a single rate over sites. In this paper, we estimate the α parameter to be used in Eqs. 2 and 3 by the method of Yang (1993) assuming the star tree. Pairwise distances estimated this way can then be used in constructing the least-squares (LS) additive tree (Cavalli-Sforza and Edwards 1967) or in other distance matrix methods. The maximum likelihood estimates of κ and t are normally found to be very similar to those obtained from Eqs. 2 and 3 (results not shown).

The Discrete Gamma Model. We use k categories to approximate the gamma distribution, with equal probability $1/k$ in each category. The density function of the gamma distribution $G(\alpha,\beta)$ is

$$g(r;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp\{-\beta r\} \cdot r^{\alpha-1},\ 0 < r < \infty \qquad (8)$$

with mean $E(r) = \alpha/\beta$ and variance $V(r) = \alpha/\beta^2$. We note that when $r \sim G(\alpha,\beta)$, $cr \sim G(\alpha,\beta/c)$. The gamma distribution with $\beta = {}^1/_2$ reduces to the χ^2 distribution, that is, $G(\alpha,{}^1/_2) \equiv \chi^2(2\alpha)$. Using these relationships, we can calculate the percentage point (the cutting point) of the gamma distribution, i.e., the value of z such that $\text{Prob}\{r < z\} = p$ where $r \sim G(\alpha,\beta)$, as follows:

$$z_G(p;\alpha,\beta) = z_{\chi^2}(p;2\alpha)/(2\beta) \qquad (9)$$

where $z_{\chi^2}(p;\nu)$ is the percentage point of the χ^2 distribution with ν degrees of freedom, which can be calculated by, say, the algorithm of Best and Roberts (1975).

The range of r, $(0,\infty)$, is cut into k categories by $k-1$ percentage points corresponding to $p = 1/k, 2/k, \ldots, (k-1)/k$. The rate in each category can then be represented by the mean of the portion of the gamma distribution falling in the category. Suppose that the two cutting points of category i are a and b. Then the rate for category i can be obtained as

$$r_i = \int_a^b r\, g(r;\alpha,\beta)\,dr \Big/ \int_a^b g(r;\alpha,\beta)\,dr$$
$$= \alpha/\beta\,[I(b\beta,\alpha+1) - I(a\beta,\alpha+1)] \Big/ (1/k) \qquad (10)$$

where $I(z,\alpha) = [1/\Gamma(\alpha)] \int_0^z \exp\{-x\} \cdot x^{\alpha-1} dx$ is the incomplete gamma ratio, which can be calculated, say, using the algorithm of Bhattacharjee (1970).

If we use the median instead of the mean to represent the average rate, r_i can be calculated as the percentage points corresponding to $p = 1/(2k), 3/(2k), \ldots, (2k-1)/(2k)$. In the current context, the scale parameter β is redundant and can be set equal to α so that the mean of the distribution is one (Yang 1993). The discrete distribution needs also to be scaled so that the mean is one if the median is used. An example is given in Fig. 1, with $\alpha = \beta = {}^1/_2$, in which case the gamma distribution is really the χ^2 distribution with one degree of freedom.

When the average rate for each category is determined, the probability of observing data x at any site can be obtained as

$$f(x) = \sum_{i=1}^{k} \frac{1}{k} \cdot f(x|r = r_i) \qquad (11)$$

The conditional probability of observing x, given that the rate for the site is $r = r_i$ is given by Yang (1993). As the postorder tree traversal algorithm of Felsenstein (1981) can be used to calculate $f(x|r = r_i)$, the computational requirement of the discrete gamma model is roughly k times that of Felsenstein's (1981) single-rate model. The continuous model was represented as, i.e., F84 + Γ (Yang et al. 1994), and we therefore represent the discrete gamma model as, i.e., F84 + dG. The discrete gamma model with $k = 4$, the value to be recommended, will be designated F84 + dG4.

It may be pointed out that the value of r_i which maximizes $f(x|r = r_i)$ in Eq. 11, can be used as the best predictor of the rate for the site.

The Fixed-Rates Model. With this model, substitution rates at sites are predicted using the method of Yang and Wang (in press), assuming the star tree and the gamma distribution for rates over sites. This method takes advantage of the observation that parameter estimates and predicted rates are more-or-less stable across tree topologies (Yang et al. 1994; Yang and Wang in press). Rates, and their corresponding sites, are then classified into $k = 4$ categories, $(0,1)$, $(1,1+\sigma)$, $(1+\sigma, 1+2\sigma)$, and $(1+2\sigma,\infty)$, with $\sigma = (1/\hat{\alpha})^{1/2}$, where $\hat{\alpha}$ is the estimated shape parameter of the gamma distribution. This scheme of classification is very poor if taken as an approximation to the gamma distribution, as the first category covers most of the sites.

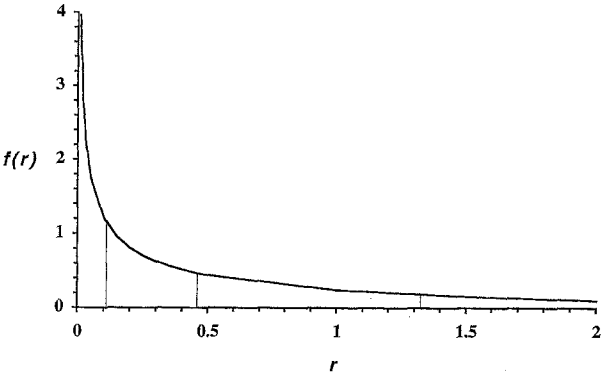

Fig. 1. Discrete approximation to the gamma distribution $G(\alpha,\beta)$, with $\alpha = \beta = {}^1/_2$. Four categories are used to approximate the continuous distribution, with equal probability for each category. The three boundaries are 0.1015, 0.4549, and 1.3233, which are the percentage points corresponding to $p = {}^1/_4, {}^2/_4, {}^3/_4$. The means of the four categories are 0.0334, 0.2519, 0.8203, 2.8944. The medians are 0.0247, 0.2389, 0.7870, 2.3535, and these are scaled to get 0.0291, 0.2807, 0.9248, and 2.7654, so that the mean of the discrete distribution is one.

It, however, reflects the discrete nature of the data; with a typical data set, the four site patterns that are represented by identical nucleotides in all the species cover most of the sites, and most often predicted rates for those sites only are less than one, the average. The rate for the ith category, $\hat{r}_i$, is obtained by averaging the predicted rates for sites in the category. The probability of observing data x at a site from the ith category is calculated as

$$f(x) = f(x|r = \hat{r}_i) \qquad (12)$$

In this formulation, rates at sites are not regarded as random variables; they are constants or parameters. Biologically, if we knew which category a site should belong to, such as in the case of the three codon positions in protein coding sequences, or if we knew whether a site was located in a highly variable region or in a very conserved region, such information could be used. When we lack such information, a good guess, as provided by the method of Yang and Wang (in press), may be used. Mathematically, the contribution to $f(x)$ from categories other than the most probable may be very small and may therefore be ignored.

Several alternatives seem possible concerning the implementation of this method. Possibilities and problems concerning the estimation of the α parameter will be discussed later. It is possible to use the rates obtained from the star tree and the continuous gamma distribution only to classify the sites, while rates for site classes can be estimated by the likelihood function based on Eq. 12. Parameter κ can also be estimated this way. It is not very clear which options will produce good performance. In this study, the average rates for site classes, $\hat{r}_i$, and parameter κ are all obtained from the star tree under the continuous gamma model, as this saves computation. Calculation of the likelihood under this model, that is, based on Eq. 12, involves roughly the same amount of computation as that of Felsenstein's (1981) single-rate model.

Data. We choose three data sets for which rates of substitution are clearly variable over sites, while other aspects, such as the amount of evolution as reflected in branch lengths in the tree, and the transition/transversion rate bias, are quite different. For sequences such as pseudogenes or "junk" DNA, for which rates are more-or-less constant over sites, the methods considered in this paper are not useful.

The mtDNA Sequences from Primates. The 895-bp mtDNA sequences of human, chimpanzee, gorilla, orangutan, and gibbon (Brown

Table 1. Likelihood values and estimates of parameters as functions of k, the number of categories in the discrete gamma model[a]

		Branch lengths								
k	ℓ	6↔5	6↔4	6↔7	7↔8	7↔3	8↔1	8↔2	$\hat{\kappa}$	$\hat{\alpha}$
1	−2,667.08	0.1385	0.0998	0.0531	0.0174	0.0578	0.0412	0.0539	4.344	NA
2	−2,625.55	0.2471	0.1627	0.0730	0.0225	0.0685	0.0504	0.0651	6.614	0.038
3	−2,620.90	0.4686	0.2842	0.1062	0.0347	0.0669	0.0559	0.0700	11.239	0.183
4	−2,621.18	0.4665	0.3010	0.1151	0.0341	0.0738	0.0576	0.0734	11.619	0.212
5	−2,621.64	0.4482	0.2953	0.1151	0.0335	0.0777	0.0586	0.0750	11.359	0.228
6	−2,621.98	0.4393	0.2917	0.1142	0.0334	0.0801	0.0594	0.0760	11.198	0.238
7	−2,622.21	0.4348	0.2897	0.1134	0.0333	0.0816	0.0602	0.0767	11.098	0.243
8	−2,622.39	0.4325	0.2887	0.1129	0.0334	0.0827	0.0608	0.0772	11.037	0.247
9	−2,622.51	0.4316	0.2884	0.1126	0.0334	0.0835	0.0613	0.0776	10.999	0.249
10	−2,622.61	0.4313	0.2884	0.1124	0.0335	0.0842	0.0617	0.0779	10.977	0.251
12	−2,622.74	0.4319	0.2890	0.1122	0.0336	0.0852	0.0625	0.0785	10.959	0.252
15	−2,622.86	0.4340	0.2904	0.1123	0.0338	0.0861	0.0632	0.0792	10.960	0.253
20	−2,622.96	0.4375	0.2928	0.1126	0.0340	0.0871	0.0640	0.0799	10.986	0.253
∞	−2,623.06	0.4532	0.3033	0.1148	0.0349	0.0892	0.0659	0.0821	11.127	0.248

[a] The mtDNA data for five species (895bp) were analyzed, assuming the best tree (Fig. 2) and the F84 model of nucleotide substitution. The averages of nucleotide frequencies across species are $\pi_T = 0.2532$, $\pi_C = 0.3289$, $\pi_A = 0.3120$, and $\pi_G = 0.1059$, with $\ell_{max} = -2{,}476.97$. The branches are specified by their two nodes in the tree shown in Fig. 2

et al. 1982) were analyzed. This segment of the mitochondrial genome codes for parts of two proteins at the two ends and three tRNAs in the middle. Rates are highly variable over sites. The transition/transversion rate bias is very high for these sequences, and only a small amount of evolution is involved.

This data set has been expanded since Brown et al. (1982), and we added sequences of this region from crab-eating macaque, squirrel monkey, tarsier, and lemur to form another, larger, data set. The sequences were aligned by A. Friday by eye. After sites involving insertions or deletions have been excluded, there are 888 nucleotides in each sequence. A much greater amount of evolution is now involved due to the addition of more distantly related sequences. These two data sets can be distinguished by their numbers of species.

The α- and β-Globin Genes from Mammals. The α- and β-globin genes of a primate (human), an artiodactyl (goat for the α-globin gene and cow for the β-globin gene), a lagomorph (rabbit), a rodent (rat), and a marsupial (the native cat for the α-globin gene and opossum for the β-globin gene) were analyzed. As the evolutionary dynamics of those two genes appear to be very similar, they were combined into one data set. Only the first and second codon positions in the coding regions were used, and there are 570 nucleotides in each combined sequence (2 × 141 for the α-globin gene and 2 × 144 for the β-globin gene). This data set is characterized by high rate variation over sites, relatively low transition/transversion rate ratio, and intermediate branch lengths.

The Small-Subunit rRNAs (ssrRNA). Small-subunit rRNA sequences of *Sulfolobus solfatarius, Halobacterium salinarium, Escherichia coli,* and *Homo sapiens* as analyzed by Navidi et al. (1991) were used. There are 1,352 nucleotides in the sequence. Rates do not appear to be highly variable over sites for those sequences, and κ is not large. The sequences are very different. Even nucleotide frequencies are quite different among species, suggesting that the substitution processes may differ among lineages. For the purpose of this study, however, this aspect of the inaccuracy of the models is ignored.

Results

The data sets are analyzed by using the discrete gamma model, assuming different numbers of categories (k), in order to find when the model produces a satisfactory fit to the data and when the model produces a good approximation to the continuous gamma distribution, as reflected by an estimate of the α parameter that is close to that obtained under the continuous gamma model. Obviously, $k = 1$ is the single-rate model, while $k = \infty$ is equivalent to the continuous gamma model. Both the mean and the median for each category have been used in the analyses, but as one might expect, the differences between them are not large. We therefore only present results obtained from the use of the mean, with comments given for the use of the median. The F84 + Γ, F84 + dG4 models, the fixed-rates model, and the pairwise least-squares method are also applied to all the tree topologies for the three data sets. Comparison of results concerning tree estimation will give us some feel about the similarities and differences among these methods. The frequency parameters in the F84 model are estimated by using the averages of the observed frequencies for all the models (methods). We do not assume the existence of a molecular clock, and as F84 is a reversible-process model, only unrooted trees can be identified (Felsenstein 1981).

The mtDNA Sequences

Results obtained from the 895-bp mtDNA sequences for five species under the discrete gamma model, F84 + dG, are presented in Table 1. The relationship among these species is probably (((human, chimpazee), gorilla), orangutan, gibbon) (e.g., Hasegawa 1991); this tree structure is assumed in the analysis (Fig. 2). The discrete gamma model with $k = 1$ is equivalent to the single-rate model. Branch lengths are severely underestimated by this model as rate variation over sites is ignored. Parameter κ is also underestimated when compared to the es-

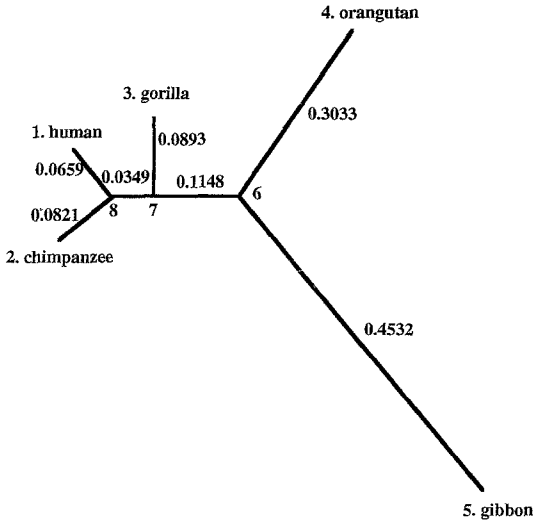

Fig. 2. The maximum likelihood tree from the 895-bp mtDNA sequences for five species. The F84 + Γ model was assumed. Branch lengths are measured by the average numbers of nucleotide substitutions per site.

timate obtained from F84 + Γ (Table 1). Both aspects were discussed by Yang et al. (1994).

As one would expect, the likelihood increases very rapidly with k when k is very small. The best fit occurs at $k = 3$, with (log-)likelihood $\ell = -2{,}620.90$, parameter estimates $\hat{\kappa} = 11.239$, $\hat{\alpha} = 0.183$. When k further increases, the likelihood decreases very slowly, until $\ell = -2623.06$ for the continuous gamma model ($k = \infty$). Estimates of parameters κ and α under F84 + Γ are $\hat{\kappa} = 11.127$ and $\hat{\alpha} = 0.248$. To get a good approximation to the continuous gamma model, as reflected by a close estimate of α, a large k seems to be needed. When the median for each category is used instead of the mean, the best fit occurs at $k = 3$, with $\ell = -2{,}620.90$, $\hat{\kappa} = 11.239$, and $\hat{\alpha} = 0.174$; the model fits the data just as well as the use of the mean, but gives a poorer approximation to the continuous gamma distribution.

All the 15 bifurcating trees and the five-species star tree are evaluated using different models (methods). The maximum likelihood tree under F84 + Γ is shown in Fig. 2. This tree can be designated HC-G, as orangutan and gibbon are the outgroups. The second and third best trees are similarly designated CG-H and HG-C respectively. Likelihood values and parameter estimates obtained from these three best trees are $\ell = -2623.06$, $\hat{\kappa} = 11.127$, and $\hat{\alpha} = 0.248$ for HC-G (Table 1), $\ell = -2{,}626.60$, $\hat{\kappa} = 11.926$, and $\hat{\alpha} = 0.224$ for CG-H, and $\ell = 2{,}626.62$, $\hat{\kappa} = 12.288$ and $\hat{\alpha} = 0.220$ for HG-C, respectively. Estimates of parameters obtained from the star tree under F84 + Γ are $\hat{\kappa} = 20.884$, and $\hat{\alpha} = 0.151$ with $\ell = -2{,}630.38$; these values of parameters are used in the fixed-rates model. The discrete gamma model with $k = 4$, F84 + dG4 produces similar results to the F84 + Γ model; the three best trees are HC-G, HG-C, and CG-H, with likelihoods −2,621.17, −2,625.33, and −2,625.54, respectively. The fixed-rates model also supports HC-G separation. When $\hat{\alpha} = 0.151$, which is obtained from the star tree under F84 + Γ, is used in Eq. 3 to calculate pairwise distances, the LS method supports the CG-H separation, although assuming a single rate for all the sites ($\alpha = \infty$) in Eq. 3 would choose the HC-G separation.

The F84 + dG model is also applied to the mtDNA data for nine species. No attempt has been made to search for the maximum likelihood tree, and the tree topology assumed in the analysis separates the species in the order human, chimpanzee, gorilla, orangutan, gibbon, crab-eating macaque, squirrel monkey, tarsier, and lemur. The best fit of the model occurs at $k = 4$, with $\ell = -5{,}044.92$, and $\hat{\kappa} = 3.855$, $\hat{\alpha} = 0.397$. The F84 + Γ model ($k = \infty$) was not fitted for computational reasons. The likelihood decreases with k when $k > 4$ increases, and for $k = 20$, the results are $\ell = -5{,}046.64$, with parameter estimates $\hat{\kappa} = 3.698$ and $\hat{\alpha} = 0.426$. When the median is used, the best fit occurs with $k = 5$, with $\ell = 5{,}044.95$, $\hat{\kappa} = 3.704$, and $\hat{\alpha} = 0.377$. Use of the median thus fits the data just as well as use of the mean, but the estimate of α is farther from that obtained at $k = 20$.

Because of the large size of the mtDNA data for nine species, we also fitted the discrete gamma model combined with the general reversible-process model of nucleotide substitution (REV, Yang in press), i.e., REV + dG. The REV model involves five rate parameters, that is, *a, b, c, d, e* instead of only one κ in the F84 model. Likelihood values and parameter estimates are listed in Table 2. Although REV + Γ is not fitted, estimates obtained from REV + dG with $k = 20$ can be expected to be very close. Apparently $k = 5$ gives the best fit, with $\ell = -5{,}031.62$ and $\hat{\alpha} = 0.462$. Using the median instead of the mean gives $k = 6$ for the best fit, with $\ell = -5{,}031.57$ and $\hat{\alpha} = 0.444$.

The α and β Globin Genes

The phylogenetic relationship among Primates (P), Artiodactyla (A), Lagomorpha (L), Rodentia (R) is not resolved although the marsupial (M) can be safely taken as the outgroup. The maximum likelihood tree under the F84 + Γ model is (((L,R),P),A) with M as the outgroup (Fig. 3). We use this tree to examine the effects of k, the number of categories, in the discrete gamma model. Because parameter estimates are quite stable over tree topologies, and the likelihoods of the several reasonable trees are very similar, our conclusion will not be biased seriously if this tree is not the true tree. The results are shown in Fig. 4. The single-rate model ($k = 1$) gives $\ell = -1{,}792.06$ with $\hat{\kappa} = 0.074$, while the continuous gamma model ($k = \infty$), F84 + Γ, gives $\ell = -1{,}761.17$, with $\hat{\kappa} = 0.116$ and $\hat{\alpha} = 0.360$. When k increases from one, the likelihood increases steadily, and the best fit appears to occur at $k = \infty$, that is, with the continuous

Table 2. Likelihood values and parameter estimates under the REV+dG model for the mtDNA sequences for nine species (888 nucleotides)[a]

k	ℓ	$\hat{a}$	$\hat{b}$	$\hat{c}$	$\hat{d}$	$\hat{e}$	$\hat{\alpha}$
1	−5,197.74	1.2552	0.2382	0.0341	0.4229	0.1479	NA
2	−5,039.27	1.1810	0.1601	0.0149	0.2633	0.1189	0.4483
3	−5,032.81	1.2367	0.1375	0.0103	0.2363	0.1025	0.4497
4	−5,031.70	1.2265	0.1310	0.0092	0.2281	0.0961	0.4539
5	−5,031.62	1.2273	0.1309	0.0090	0.2279	0.0944	0.4622
6	−5,031.73	1.2307	0.1316	0.0091	0.2287	0.0937	0.4679
7	−5,031.85	1.2339	0.1323	0.0092	0.2294	0.0935	0.4714
8	−5,031.95	1.2366	0.1328	0.0092	0.2301	0.0933	0.4736
9	−5,032.03	1.2388	0.1333	0.0093	0.2306	0.0932	0.4749
10	−5,032.10	1.2406	0.1336	0.0094	0.2310	0.0932	0.4758
12	−5,032.20	1.2431	0.1341	0.0094	0.2315	0.0931	0.4767
15	−5,032.28	1.2454	0.1345	0.0095	0.2325	0.0931	0.4770
20	−5,032.36	1.2474	0.1349	0.0096	0.2326	0.0931	0.4769

[a] The assumed (unrooted) tree separates the species in the order human, chimpanzee, gorilla, orangutan, gibbon, crab-eating macaque, squirrel monkey, tarsier, and lemur. The general reversible process model (REV) is combined with the discrete gamma model of rates over sites, where k is the number of categories. The frequency parameters are estimated by using the averages of observed values, $\pi_T = 0.2660$, $\pi_C = 0.3044$, $\pi_A = 0.3220$, and $\pi_G = 0.1076$, with $\ell_{max} = -3{,}960.997$. The rate parameters *a, b, c, d, e* in the REV model and the α parameter of the gamma distribution are estimated by iteration. Estimates of branch lengths are not shown

gamma model. The likelihood values for $k = 1, 2, 3, 4$ are −1,792.06, −1,765.13, −1,763.45, and −1,762.01, respectively. The estimates of parameters for $k = 4$ are $\hat{\kappa} = 0.106$ and $\hat{\alpha} = 0.321$. When the median instead of the mean is used for each category, the best fit also appears to occur with the continuous gamma model ($k = \infty$). At $k = 4$, the estimates are $\hat{\kappa} = 0.106$, $\hat{\alpha} = 0.280$, with $\ell = -1{,}762.04$. The fit is as good as use of the mean, but the estimate of α is further from the continuous gamma model.

The models are applied to all the possible tree topologies linking the five species. Likelihood values and parameter estimates obtained from the three best trees under the F84 + Γ model are listed in Table 3. Results obtained from the star tree are also shown. The discrete gamma model with $k = 4$, F84 + dG4 gives very similar results to those obtained under F84 + Γ (Table 3). The fixed-rates model supports the same best tree, but the second and third best trees converge to one topology as one of the two interior branch lengths approaches zero in both trees. With this model, data for sites from different rate classes are not identically distributed, and therefore the likelihoods should not be compared with ℓ_{max}, the "upper limit" obtained from the "unconstrained" model of Navidi et al. (1991) and Goldman (1993). The LS method, using $\hat{\alpha} = 0.295$ in Eq. 3 to calculate pairwise distances, also produces the same best tree. While most of the 15 bifurcating trees have strictly positive interior branch lengths under F84 + Γ or F84 + dG4, at least one interior branch length in all the nonbest trees is zero by the LS method. The fixed-rates model appears very similar to the LS method in this respect, as only two bifurcating trees have both interior branch lengths strictly positive. The same phenomenon is observed in the other two data sets.

All the models (methods) considered in this paper support the tree structure shown in Fig. 3. The relationship shown in Fig. 3 can therefore be taken as the best estimate of the phylogeny from this data set. We make no attempt to evaluate the statistical significance of this estimate.

The ssrRNA Sequences

The best tree under F84 + Γ is T_1 = ((*S. solfatarius, H. sapiens*), *H. salinarium, E. coli*), and this tree structure is assumed in examining the effect of k, the number of categories in the discrete gamma model. The results are shown in Fig. 5. The single-rate model ($k = 1$) gives $\ell = -5{,}834.11$, $\hat{\kappa} = 0.447$ while the continuous gamma model ($k = \infty$), F84 + Γ, gives $\ell = -5{,}785.45$, with $\hat{\kappa} = 0.731$, $\hat{\alpha} = 0.836$. The best fit occurs with $k = 2$, with $\ell = 5{,}785.07$, $\hat{\kappa} = 0.704$, $\hat{\alpha} = 0.722$. When k further increases, the likelihood first decreases and then increases again, but the changes can be considered to be trivial. If the median instead of the mean for each category is used, the best fit occurs at $k = 2$, with $\ell = 5{,}785.07$, $\hat{\kappa} = 0.704$, $\hat{\alpha} = 0.674$.

The second and third best trees under F84 + Γ are T_2 = ((*S. solfatarius, E. coli*), *H. salinarium, H. sapiens*), and T_3 = ((*S. solfatarius, H. salinarium*), *E. coli, H. sapiens*), with likelihood values and parameter estimates to be $\ell = -5{,}786.02$, $\hat{\kappa} = 0.723$, $\hat{\alpha} = 0.837$ for T_2, and $\ell = 5{,}786.83$, $\hat{\kappa} = 0.738$, $\hat{\alpha} = 0.825$ for T_3, respectively. Estimates from the star tree under F84 + Γ are $\hat{\kappa} = 0.746$, $\hat{\alpha} = 0.807$ with $\ell = -5{,}786.88$. Note that the likelihood values for different tree topologies are very similar, probably meaning that there is not much information in the data concerning the phylogenetic relationship of these species. The discrete gamma model with $k = 4$, that is, F84 + dG4, produces very similar results to F84 + Γ; the order of the trees is also the same. The fixed-rates model produces the same best tree

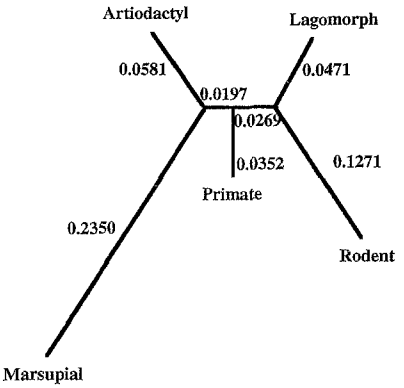

Fig. 3. The maximum likelihood tree for the five orders of mammals from the α and β globin genes (570 bp). The F84 + Γ model was assumed. Branch lengths are measured by the average numbers of nucleotide substitutions per site.

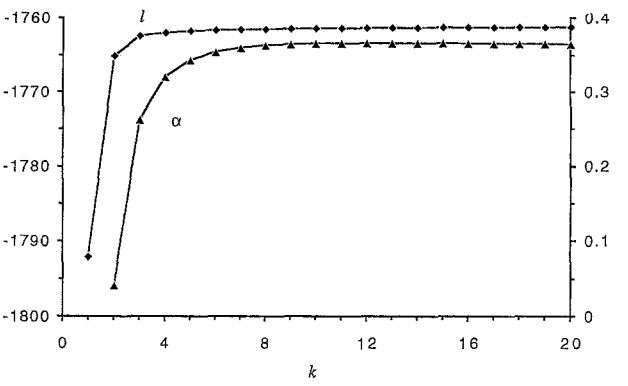

Fig. 4. Likelihood values and estimates of the α parameter as functions of k, the number of categories in the discrete gamma model. The α and β globin genes for the five mammalian orders (570 bp) are analyzed, assuming the best tree (Fig. 3) and the F84 + dG model. The average nucleotide frequencies are $\pi_T = 0.2200$, $\pi_C = 0.2449$, $\pi_A = 0.2761$, and $\pi_G = 0.2590$, with $\ell_{max} = -1{,}579.76$. When $k = \infty$, that is, with the F84 + Γ model, $\ell = -1{,}761.17$ and $\hat{\alpha} = 0.360$.

as the F84 + Γ and F84 + dG4 models; the other two bifurcating trees, T_2 and T_3, are not better than the star tree. The LS method, using $\hat{\alpha} = 0.807$ in Eq. 3 to calculate pairwise distances, chooses T_2 as the best tree, while if the distances are calculated assuming a single rate over sites ($\alpha = \infty$ in Eq. 3), the best tree is T_3. In both cases, only the best trees are strictly better than the star tree; the other two bifurcating trees have zero interior branch lengths.

As mentioned above, the nucleotide frequencies of these sequences are quite different over species, a feature that is not accommodated in any of the models (methods) addressed in this paper. As accounting for this feature in the models may well alter the order of the likelihood values for different trees, which are already very similar, we will not suggest the best estimate of the phylogenetic relationship from this data set.

Discussion

The Discrete Gamma Model

As substitution rates are definitely variable over sites for all the data sets analyzed in this study, the best fit of the discrete gamma model must occur with $k > 1$. Indeed, all possibilities appear to have been observed: the best fit occurs at $k = 2$ for the *ssr*RNA sequences, at $k = \infty$ for the α and β globin genes, and at intermediate values (three to six) for the mtDNA sequences. Different dynamics of $\hat{\alpha}$, as a function of k, are also observed. Overall, the goodness of fit of the model, as measured by the likelihood, is very similar for different values of k as long as $k \geq 3$. For data sets analyzed in this paper, the discrete gamma model with $k = 3$ gives fairly good, if not the best, fit to the data. However, for the sake of safety, we recommend using $k = 4$. When a fixed k is used in a likelihood analysis, α will be well defined, and results obtained from different analyses or from different data sets will be comparable. Using the mean or the median for each category appears to give roughly the same fit, but when k is small, the mean has been found to give a better approximation to the continuous distribution. This can be expected to be due to the mean/median ratios over the categories as determined by the density function of the gamma distribution. We thus recommend the use of the mean rather than the median. For the estimation of the phylogenetic tree, parameters κ and α do not need to be estimated for each of the tree topologies, as the estimates are quite stable across trees.

Estimating the α Parameter of the Gamma Distribution

The discrete gamma model can be used to estimate the α parameter of the continuous gamma distribution, which can then be used in estimating pairwise distances, for example, by Eq. 3 or the formula of Tamura and Nei (1993). As these equations assume a continuous gamma model, it may be important to obtain an estimate of α that is close to that obtained under the continuous gamma model. This appears to require a large value of k, for example, $k = 8$. It is noted that the improvement in accuracy of the estimation by increasing k is often very slow. For data sets with only four or five species, the approach of Yang (1993) may well be preferred.

An important confounding factor is the assumed tree structure. In theory, the estimate of α obtained from using the true tree will be best, as estimates from other tree topologies will involve systematic errors. This is seldom possible, however, as the purpose of estimating the α parameter for the distance estimators is to reconstruct the tree. Although parameter estimates are found to be quite stable over tree topologies, we do notice some patterns: given the data set, the estimate of α obtained

Table 3. Likelihood values and parameter estimates under different models and different trees[a]

Models	Trees			
	(((L,R,),P),A)	((L,R),(P,A))	(((L,R),A),P)	(L,R,P,A)
F84+Γ				
	$\ell = -1{,}761.17$	$\ell = -1{,}762.82$	$\ell = -1{,}763.02$	$\ell = -1{,}768.74$
	$\hat{\kappa} = 0.116$	$\hat{\kappa} = 0.118$	$\hat{\kappa} = 0.121$	$\hat{\kappa} = 0.126$
	$\hat{\alpha} = 0.360$	$\hat{\alpha} = 0.337$	$\hat{\alpha} = 0.333$	$\hat{\alpha} = 0.295$
F84+dG4				
	$\ell = -1{,}762.01$	$\ell = -1{,}763.40$	$\ell = -1{,}763.47$	$\ell = -1{,}768.81$
	$\hat{\kappa} = 0.106$	$\hat{\kappa} = 0.107$	$\hat{\kappa} = 0.112$	$\hat{\kappa} = 0.114$
	$\hat{\alpha} = 0.321$	$\hat{\alpha} = 0.299$	$\hat{\alpha} = 0.297$	$\hat{\alpha} = 0.264$
F84 (fixed rates)				
	$\ell = -1{,}525.26$	$\ell = -1{,}525.27$		$\ell = -1{,}527.60$

[a] The α and β globin genes (570 bp) for the five groups of mammals, that is, Primates (P), Artiodactyla (A), Lagomorpha (L), Rodentia (R), and marsupial, were analyzed. Results for the three best trees under the F84+Γ model and from the star tree are shown. The unrooted trees are shown in their rooted forms with the marsupial as the outgroup. The second and third best trees under the fixed rates model have an interior branch length approaching zero, and both converge to one tree structure: ((L,R),P,A)

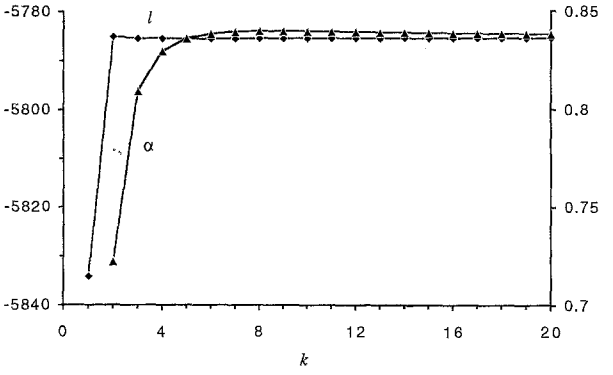

Fig. 5. Likelihood values and estimates of the α parameter as functions of k, the number of categories in the discrete gamma model. The ssrRNA genes (1,352 bp) are analyzed using the F84 + dG model. The average nucleotide frequencies are $\pi_T = 0.1847$, $\pi_C = 0.2534$, $\pi_A = 0.2396$, $\pi_G = 0.3214$, with $\ell_{max} = -5{,}591.06$. The tree assumed is ((*S. solfatarius, H. sapiens*). *H. salinarium, E. coli*), the maximum likelihood tree under F84 + Γ. When $k = \infty$ (F84 + Γ), $\ell = 5{,}785.45$ and $\hat{\alpha} = 0.836$.

from the maximum likelihood tree is almost always larger than those obtained from other tree topologies (Yang et al. 1994 and unpublished results; see also above). The estimate from the star tree is most often found to be the smallest. Therefore, using the star tree to estimate α, a practice adopted in this study, can be expected to give underestimates. This discrepancy can be large when there are many species in the data and therefore the star tree is quite different from the true tree. On the other hand, if α is always estimated from the maximum likelihood tree, α will be overestimated when the maximum likelihood tree is not the true tree. This peculiarity of the sampling properties of maximum likelihood estimates of parameters in the framework of tree estimation is discussed elsewhere. It may therefore be better to estimate α using a more-or-less reasonable tree with fewer species, that is, using a subset of the data, rather than using the star tree with all the species. Nevertheless, if distance matrix methods are only used to produce several candidate trees to be subjected to more rigorous analysis by the likelihood method, which I believe should be the case, a rough estimate of α may be acceptable.

One might expect that α could also be estimated from pairwise comparisons, using, for instance, the maximum likelihood criterion. All pairwise estimates could be averaged to produce one estimate, which could then be used, e.g., in Eq. 3, to estimate pairwise distances. However, this is found to be impossible in practice. Estimates of α obtained from comparison of only two sequences are always found to be very large; most often the gamma distribution model is not any better at all than a single rate model. It does not seem possible to reveal rate variation over sites by comparing only two sequences.

Acknowledgments. I wish to thank Clive Moncrieff and Nick Goldman for help with the implementation of the discrete gamma model. I am grateful to Clive Moncrieff for suggestions and comments on earlier versions of the manuscript. This study was supported by a grant from Department of Zoology, The Natural History Museum (London).

References

Best DJ, Roberts DE (1975) The percentage points of the χ^2 distribution. Appl Statist 24:385–388

Bhattacharjee GP (1970) The incomplete gamma integral. Appl Statist 19:285–287

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates, tempo and mode of evolution. J Mol Evol 18:225–239

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

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

Fitch WM (1986) The estimate of total nucleotide substitutions from pairwise differences is biased. Philos Trans R Soc Lond Biol 312: 317–324

Fitch WM, Margolish E (1967) A method for estimating the number of invariant amino acid coding positions in a gene, using cytochrome c as a model case. Biochem Genet 1:65–71

Fitch WM, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579–593

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

Haseqawa M (1991) Molecular phylogeny and man's place in Hominoidea. J Anthrop Soc Nippon 99:49–61

Hasegawa M, Horai J (1991) Time of the deepest root for polymorphism in human mitochondrial DNA. J Mol Evol 32:37–42

Hasegawa M, Kishino H (1989) Confidence limits on the maximum likelihood estimation of the hominoid tree from mitochondrial DNA sequences. Evolution 43:672–677

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

Hasegawa M, Rienzo AD, Kocher TD, Wilson AC (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. J Mol Evol 37:347–354

Holmquist R, Goodman M, Conry T, Czelusniak J (1983) The spatial distribution of fixed mutations within genes coding for proteins. J Mol Evol 19:137–448

Jin L, Nei M (1990) Limitations of the evolutionary parsimony method of phylogeny analysis. Mol Biol Evol 7:82–102

Kocher TD, Wilson AC (1991) Sequence evolution of mitochondrial DNA in humans and chimpanzees: Control region and a protein-coding region. In: Osawa S, Honjo T (eds) Evolution of life: fossils, molecules, and culture. Springer-Verlag, Tokyo, pp 391–413

Li W-H, Gouy M, Sharp PM, O'hUigin C, Yang Y-W (1990) Molecular phylogeny of rodentia, lagomorpha, primates, artiodactyla, and carnivora and molecular clocks. Proc Natl Acad Sci USA 87: 6703–6707

Navidi WC, Churchill GA, von Haeseler A (1991) Methods for inferring phylogenies from nucleotide acid sequence data by using maximum likelihood and linear invariants. Mol Biol Evol 8:128–143

Nei M, Gojobori T (1986) Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Palumbi SR (1989) Rates of molecular evolution and the function of nucleotide positions free to vary. J Mol Evol 29:180–187

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

Thorne JL, Kishino H, Felsenstein J (1992) Inching toward reliability: an improved likelihood model of sequence evolution. J Mol Evol 34:3–16

Uzzell T, Corbin KW (1971) Fitting discrete probability distributions to evolutionary events. Science 172:1089–1096

Wakeley J (1993) Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J Mol Evol 37:613–623

Yang Z (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401

Yang Z (in press) Estimating the pattern of nucleotide substitution. J Mol Evol

Yang Z, Wang T (in press) Mixed model analysis of DNA sequence evolution. Biometrics

Yang Z, Goldman N, Friday AE (1994) Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol Biol Evol 11:316–324