Estimating the Distribution of Selection Coefficients from Phylogenetic Data with Applications to Mitochondrial and Viral DNA

Rasmus Nielsen and Ziheng Yang†*

*Department of Biometrics, Cornell University; and †Department of Biology, University College London, London, England

The distribution of selection coefficients of new mutations is of key interest in population genetics. In this paper we explore how codon-based likelihood models can be used to estimate the distribution of selection coefficients of new amino acid replacement mutations from phylogenetic data. To obtain such estimates we assume that all mutations at the same site have the same selection coefficient. We first estimate the distribution of selection coefficients from two large viral data sets under the assumption that the viral population size is the same along all lineages of the phylogeny and that the selection coefficients vary among sites. We then implement several new models in which the lineages of the phylogeny may have different population sizes. We apply the new models to a data set consisting of the coding regions from eight primate mitochondrial genomes. The results suggest that there might be little power to determine the exact shape of the distribution of selection coefficient but that the normal and gamma distributions fit the data significantly better than the exponential distribution.

Introduction

The distribution of selection coefficients (s) of new mutations has been the focus of many population genetical studies. Major theories of molecular evolution differ in their assumptions regarding this distribution. For example, the original version of the neutral theory of molecular evolution, Kimura's (1968) strictly neutral model, assumes that all new mutations are either neutral ($s = 0$) or strongly deleterious ($s = -\infty$). In a population evolving according to this model, all segregating alleles have the same fitness assigned to them. This model is still heavily used in population genetics. Most studies aimed at estimating demographic or ancestral parameters using molecular markers, either implicitly or explicitly, assume this model, or one of its close relatives. For example, the classical coalescence model (e.g., Kingman 1982; Hudson 1990) assumes strict neutrality. Tests of neutrality such as Tajima's D test (Tajima 1989) and the HKA test (Hudson, Kreitman, and Aquade 1987) are tests of the strictly neutral model. Common methods for estimating migration rates, such as those based on F_{ST}, also assume strict neutrality. Given the importance of this assumption in these and other applications, it is no wonder that much of the theoretical part of the population genetics literature has focused on the distribution of selection coefficients.

The first important modification to Kimura's model was proposed by Ohta (1973). In her slightly deleterious mutation theory, new mutations have exponentially distributed negative selection coefficients. This model allows some mutations to be slightly deleterious, while no positive selection is allowed. A generalization of this model was provided by Kimura (1979, 1983), who suggested that the negative selection coefficients follow a gamma distribution. Kimura named this model the model of effectively neutral mutations.

There have also been many suggestions of models that involve positive selection coefficients. For example, in the classical Fisherian model (Fisher 1930*a*), the fitness effect of a new mutation is inversely related to the difference of the new allele from the ancestral allele, while both positive and negative selection coefficients are allowed in the model. Ohta (1992) proposed a modification of her original model, to allow a proportion of new mutations to have positive selection coefficients. This model is known as the nearly neutral model and is similar to some of the selection models discussed in Gillespie (1991). In fact, the current controversy regarding the distribution of selection coefficients is often reduced to a discussion of the relative importance of positive selection. Nonetheless, because of mathematical convenience, the strictly neutral model remains the most commonly assumed model in population genetic studies of demography or ancestral history.

Key words: d_N/d_S, maximum likelihood, selection coefficients, neutral theory, mitochondrial DNA, positive selection.

E-mail: rn28@cornell.edu.

Mol. Biol. Evol. 20(8):1231–1239. 2003
DOI: 10.1093/molbev/msg147
Molecular Biology and Evolution, Vol. 20, No. 8,

Codon-Based Likelihood Models

For protein-coding genes, a measure of selective pressure on amino acid replacement mutations is the nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$). Recent computational advances have made inferences regarding the distribution of ω among sites possible (Nielsen and Yang 1998; Yang et al. 2000). To estimate ω, we describe the evolutionary process in nucleotide sequences at the codon level as a continuous time Markov chain, with state space on the 61 possible sense codons in the universal genetic code (or the 60 sense codons in the vertebrate mitochondrial code). The infinitesimal rate of change from codon i to codon j in these models is (Goldman and Yang 1994; Muse and Gaut 1994):

$$q_{ij} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases} \quad (1)$$

where π_j is the stationary frequency of codon j and κ is the transition/transversion rate ratio. For multiple DNA

sequences, this process is superimposed on the lineages of a phylogeny. Using Felsenstein's (1981) algorithm and the model in equation 1, it is possible to calculate the probability of the data, given a set of parameters such as κ, ω, and the branch lengths. Therefore, it is possible to estimate ω using maximum likelihood. Parameters κ and π_j reflect mutational pressure, whereas the effect of selection at the protein level is captured in the parameter ω.

Nielsen and Yang (1998) and Yang et al. (2000) extended this model to the case in which ω varies among sites. In one model, ω was a random variable that takes the value 1 with probability p and 0 with probability $1 - p$. This model was interpreted as a strictly neutral model of evolution. If the selection coefficient is the same for all nonsynonymous mutations in a particular codon site, this distribution of ω among sites is predicted from Kimura's (1968) strictly neutral model of evolution. This model was extended by adding a proportion of sites with $\omega > 1$. Comparison between the two models using a likelihood ratio test constitutes a test of the neutral model against a positive selection alternative.

In Yang et al. (2000), several new models for variation in ω were introduced, varying in complexity from a simple gamma distribution of ω among sites to a mixture of three normal distributions. Some limited power was found in distinguishing the various models (Yang et al. 2000). For example, in all the 10 data sets analyzed, the strictly neutral model gave a significantly worse fit to the data than a model with an additional category of sites in which ω was a free parameter. However, very little power was detected to distinguish between the more parameter-rich models. In eight out of the 10 cases, the likelihood values obtained under a simple gamma distribution and under a parameter-rich mixture of three normal distributions were within 1 log-likelihood unit of each other.

As the nonsynonymous/synonymous substitution rate ratio ω is a measure of selective pressure acting on the protein, it is possible to use codon-based likelihood models to make inferences regarding the distribution of selection coefficients of nonsynonymous mutations. In this paper, we will illustrate how this can be done. We will develop some new codon-based likelihood models derived from population genetics models, and we will estimate parameters of these population genetics models and compare the fit of the models using likelihood ratio tests. Our analyses are based on relatively strong assumptions about the mutation process. In particular, we will assume that all nonsynonymous mutations in the same site have the same selection coefficient. This assumption may not be met in many cases. However, without a function mapping selection coefficients to values of ω, little or no progress can be made on estimating the distribution of selection coefficients.

Several previous methods have been suggested for estimating the distribution of selection coefficients, often based on interspecific data using allelic distributions, or frequency spectrums (e.g., Bustamante et al. 2002; Fay, Wyckoff, and Wu 2001). The new method differs from such methods in that it only considers interspecific data and does not use information regarding the allelic distribution within species.

Material and Methods

Estimation Under a Constant Population Size Among Lineages

One of the important differences between the substitution rate ratio ω and the selective coefficient s is that s is a property of a particular allele or mutation, whereas ω, as formulated in current codon models, is a property of a particular site or collection of sites in the DNA sequence. It is therefore possible to infer the distribution of s from the estimated distribution of ω only if we make additional assumptions in the mutation model. In fact under most models, it is possible to map the distribution of ω into a distribution of s. Sawyer and Hartl (1992) demonstrated how to convert an estimate of ω into an estimate of s, assuming an infinite sites model. We will apply a similar method to the finite sites models considered in this article.

Assume that there is no interference in the fixation process of multiple mutations at different sites. This will be true for interspecific data if there are not many strongly or moderately selected mutations segregating at the same time or if the level of recombination between sites is sufficiently high. If this assumption is not met, our method will tend to underestimate the selection coefficients. Additionally, we will assume that there are never more than two alleles segregating at the same nucleotide site. This is a reasonable assumption when the expected time to fixation or extinction measured in generations is short compared with the inverse of the mutation rate. This assumption should be valid for most organisms.

Under these assumptions, the fixation rate of new mutations with selection coefficients s is the product of the mutation rate (μ) per site, the chromosomal population size (N) in a haploid organism, and the probability of fixation (Kimura 1962),

$$\frac{\mu N 2s}{1 - e^{-2Ns}}, \tag{2}$$

if we assume s is small and N is large and equal to the effective population size. This result can be derived under a variety of population genetics models, such as the Wright-Fisher (e.g., Fisher 1930*b*) model and the Moran (1962) model. Likewise the rate of substitution of neutral mutations is $N\mu/N = \mu$. If we assume that all nonsynonymous mutations at the same amino acid site have the same selection coefficient and that all synonymous mutations are neutral we have

$$\omega = \frac{S}{1 - e^{-S}}, \tag{3}$$

where $S = 2Ns$. Mitochondrial genes in diploid organisms can be treated as genes segregating in haploid organisms, and nuclear genes can be treated by redefining $S = 4Ns$. By using the inverse mapping from ω to S, the distribution of selection coefficients can be obtained from the distribution of ω. For a monotone function, such as equation 3, the transformation from random variable X to $Y = g(X)$ is given by $F_Y(y) = F_X(g^{-1}[y])$, where $F_Y(a)$ and $F_X(a)$ are the CDFs for the random variables Y and X, respectively, evaluated at point a. We notice that this approach is conceptually

Table 1
Models Implemented in This Paper and Maximum-Likelihood Estimates of Parameters Obtained from the mtDNA Data Set

Model	Number of Parameters	Maximum ℓ	Parameter Estimates
1. Constant ω	1	−38,884.38	$\omega = 0.0504$
2. Lineage variation	13	−38,859.04	$S = -2.4$
3. Normal	14	−38,357.91	$\mu = -1.72$, $\sigma = 0.72$
4. Normal < 0	14	−38,357.60	$\mu = -1.72$, $\sigma = 0.77$
5. Reflected gamma	14	−38,358.32	$\alpha = 3.22$, $\beta = 1.62$
6. Reflected exponential	13	−38,501.08	$\lambda = 0.214$
7. Normal + invar.	15	−38,355.75	$\mu = -1.36$, $\sigma = 0.53$, $p_0 = 0.336$, $p_1 = 0.664$
8. Reflected gamma + invar.	15	−38,358.12	$\alpha = 3.58$, $\beta = 2.02$, $p_0 = 0.124$, $p_1 = 0.876$
9. Reflected exponential + invar.	14	−38,501.08	$\lambda = 0.214$, $p_0 = 0.0$, $p_1 = 1.0$
10. $S = 0$ or $S = -1000$	1	−39,883.91	$p_0 = 0.701$, $p_1 = 0.299$

similar to the approach used by Halpern and Bruno (1998) for calculating genetic distances.

Yang et al. (2000) have previously developed models of varying ω among codon sites. For example, in one of the models it is assumed that ω is gamma distributed among codon sites with parameters α and β (model M5 in Yang et al. 2000). The probability density function for S (using equation 3) is thus given by

$$f(S) = \frac{(\beta e^S S/h(S))^\alpha e^{-\beta e^S S/h(S)}(h(S) - S)}{Sh(S)\Gamma(\alpha)}, \quad -\infty < S < \infty, \tag{4}$$

where $h(S) = (e^S - 1)$. Here we have assumed that S and N do not vary among lineages. We will use this model to estimate the distribution of selection coefficients for two large viral DNA data sets. For such data, the model of constant N among lineages might be realistic if the virus has visited many different hosts along each lineage of the phylogeny. If we assume the viral population size follows some probability distribution among hosts, assumed to be constant in time, the mean population size will converge to the same value for all lineages. As a first approximation we will, therefore, approximate the distribution of selection coefficients using equation 4 applied to constant population sizes. However, in practice, little is known about the variation in population size among viral lineages. Also, for the analysis of data from multiple species, the assumption of constant N among lineages seems rarely to be justifiable. Therefore, one of the objectives of this article is to develop models that will relax this assumption.

Models of Varying N Among Lineages

Our goal here is to develop models that will allows us to estimate the effective population sizes in different lineages of a phylogeny jointly with the distribution of selection coefficients among sites. One of the major aims is to investigate to which degree it is possible to distinguish between different distributions of selection coefficients using phylogenetic data. The methods we develop will be applied to a previously published data set of complete mitochondrial genomes from eight primate species (Yoder and Yang 2000).

The codon-based substitution models are parameterized in terms of the population genetics parameters s and N. We make the simplifying assumption that the selection coefficient acting on new mutations at a site is constant in a particular lineage; that is, all mutations occurring at the same site have the same selection coefficient. We will also assume that the population size along each lineage of the phylogeny is a free parameter; that is, we allow variation of population size among lineages in the phylogeny. The value of ω at site i in lineage j is then

$$\omega_{ij} = \frac{2N_j s_i}{1 - e^{-2N_j s_i}}, \tag{5}$$

where N_j is the population size in lineage j and s_i is the selection coefficient acting on new mutations in site i. The rate of change in site i and lineage j is then obtained by inserting equation 5 for ω in equation 1.

Ten different models are implemented, allowing different assumptions regarding N_j and s_i. To keep the models identifiable, we may only estimate the relative population sizes. We can, for example, fix the size of the human population and estimate the population sizes in all the other lineages relative to the human population size. The parameters of the models are then the relative population sizes and those in the distribution of $S = 2Ns$. The models differ from previous approaches in that they are parameterized directly in terms of selection coefficients. Estimates of S cannot be obtained simply by transforming estimates of ω in most of these models because N is allowed to vary among lineages.

Analytical calculation of the likelihood function under the continuous distributions is not computationally possible. Instead we approximate the distributions using 10 discrete categories as in Yang et al. (2000). A summary of the models is provided in table 1.

In the first model (model 1) we assume that ω is constant among lineages and among sites (i.e., $s_i = s$ for all i and $N_j = N$ for all j). In addition to branch lengths and κ, this model has one parameter: ω. We cannot estimate s and N separately. This is the model of Goldman and Yang (1994).

In the second model (model 2), we assume that the effective population size varies among lineages but that $s_i = s$ for all i. Model 2 has 13 parameters: 12 values of N and one value of s. There are eight species, resulting in 13 different lineages in an unrooted tree. Since we can only estimate the relative population sizes, there are 12 values of N to estimate.

Model 3 assumes that s follows a normal distribution among sites with mean μ and variance σ^2. This model has 14 parameters: 12 values of N and μ and σ^2.

Model 4 is identical to model 3, except that the normal distribution has been truncated such that values of $s > 0$ are not allowed; that is, no positive selection is allowed. This model also has 14 parameters.

In model 5 it is assumed that s follows a gamma distribution reflected around the $s = 0$ axis; that is,

$$f(s) = \beta^{\alpha}\Gamma(\alpha)^{-1}e^{\beta s}(-s)^{\alpha-1}, \quad s \leq 0. \tag{6}$$

This model has 14 parameters.

In model 6 it is assumed that s follows a reflected exponential distribution reflected around the $s = 0$ axis; that is,

$$f(s) = \lambda e^{\lambda s}, \quad s \leq 0. \tag{7}$$

This model has 13 parameters.

Models 7, 8, and 9 are identical to models 2, 3, and 5, respectively, except that an extra category of sites (with proportion p_0) in which $s = -\infty$ ($\omega = 0$) is added to each model. The proportion of variable sites is then $p_1 = 1 - p_0$. Therefore, the number of parameters in models 7, 8, and 9 are 15, 15, and 13, respectively. The reason for implementing these models is that we are mostly interested in the distribution of selection coefficients in sites that may vary. If many sites are completely functionally constrained, such that all new mutations are immediately lost from the population, our parameter estimates may be heavily influenced by the presence of such sites.

In the last model (model 10), it is assumed that all new mutations at a site are either neutral ($s = 0$) or strongly deleterious ($s = -\infty$). In such a model, it is not possible to estimate population sizes for different lineages, since N does not influence the proportion of neutral sites. There is, therefore, only one parameter: p_0, the proportion of invariable sites. This model is equivalent to the neutral model of Nielsen and Yang (1998).

Note that models 3 to 9 differ from all models implemented previously, not only in the distributional assumptions but also in allowing the effect of selection to vary simultaneously among lineages and among sites. Such models are necessary to allow variation in N among lineages while inferring the distribution of selection coefficients. At the level of computation, these models are much more demanding because they require the recalculation of the transition matrices for each branch and for each site class. A computationally more efficient alternative for allowing variation among lineages and sites simultaneously was described in Yang and Nielsen (2002). However, the method considered in Yang and Nielsen (2002) allows only a few categories of sites and cannot easily be used to estimate the distribution of selection coefficients.

Viral Data

We first use the model of fixed N among lineages and gamma distributed ω to estimate the distribution of selection coefficients among sites in two large viral data sets. The first data set consists of 421 codons from each of 186 sequences from the HIV-1 *env* gene. This data set was previously analyzed by Yamaguchi-Kabata and Gojobori (2000) and Yang (2001). The second data set consists of 329 codons from each of 349 sequences from the human influenza virus H3 hemagglutinin gene. It was previously analyzed by Bush et al. (1999) and Yang (2000). A more detailed description of the two data sets can be found in these references. The data sets are analyzed assuming ω is gamma distributed among sites and that this distribution is constant among lineages. For computational and statistical details, see Yang et al. (2000). After the parameters of the gamma distribution have been estimated, the distribution of ω is transformed into a distribution of S using equation 4.

Both of the viral genes analyzed are known to undergo strong positive selection; that is, there are multiple sites for which $S > 0$ (or $\omega > 1$). They code for the coat proteins of the two viruses, which are primary targets for the host immune system. It is generally believed that the positive selection is driven by a selective pressure to avoid host immune recognition. For those data sets, it is entirely possible that more than two nucleotides may be segregating at the same time in a site. For positively selected sites, the effect will presumably be to underestimate the magnitude of the selection coefficient.

mtDNA Data

This data set consists of eight whole mitochondrial genomes from human (*Homo sapiens* [GenBank accession number X93334]), common chimpanzee (*Pan troglodytes* [GenBank accession number X93335]), bonobo (*Pan paniscus* [GenBank accession number D38116]), gorilla (*Gorilla gorilla* [GenBank accession number X93347]), Bornean orangutan (*Pongo pygmaeus p* [GenBank accession number D38115]), Sumatran orangutan (*Pongo pygmaeus abelii* [GenBank accession number X97707]), gibbon (*Hylobates lar* [GenBank accession number X99256]), and hamadryas baboon (*Papio hamadryas* [GenBank accession number Y18001]). The data set was previously analyzed by Yoder and Yang (2000) for molecular clock dating. Only the 12 protein-coding genes on the same strand of the genome are used; after alignment and removal of gaps, each sequence consists of 3,593 codons.

For each of the models discussed above, the likelihood function can be calculated as in Yang et al. (2000), which the reader can refer to for computational details. For the continuous distributions of selection coefficients, calculations are performed by discretizing the density function using 10 categories (see Yang et al. 2000). The topology of the phylogeny is assumed to be known, and the branch lengths of the phylogeny are estimated by maximum likelihood, together with parameters in the distribution of S, as summarized in table 1. The likelihood function is calculated efficiently by storing the calculated transition probability matrices for each rate category in each lineage in the computer memory. Optimization of the likelihood function takes between a few minutes to about 12 h on a PC, depending on the model.

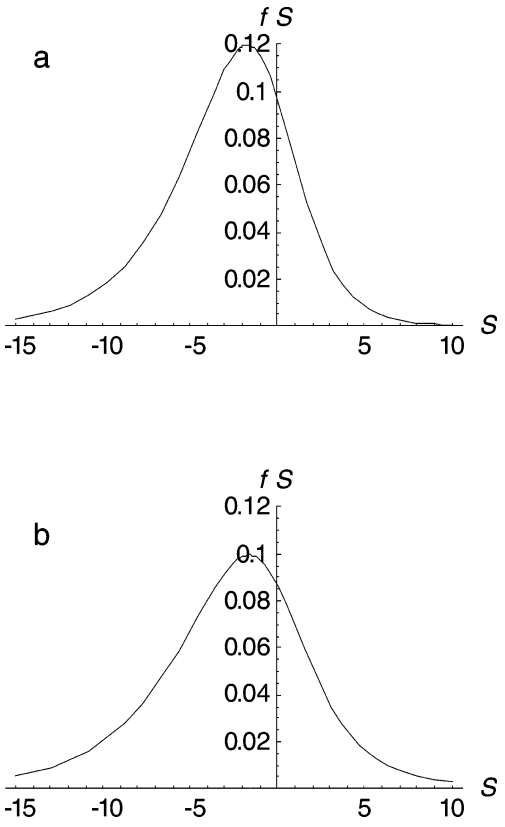

FIG. 1.—The distribution of scaled selection coefficient ($S = 2Ns$) in the HIV-1 *env* gene (*a*) and the human influenza virus hemagglutinin gene (*b*) estimated assuming a gamma distribution of ω among sites (equation 4).

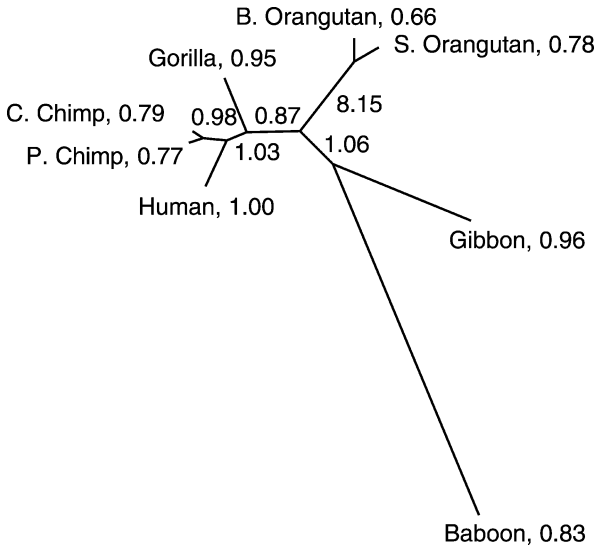

FIG. 2.—The phylogenetic tree of eight primate species for the data of 12 mitochondrial protein-coding genes. Branch lengths are proportional to the expected number of synonymous substitutions. The numbers adjacent to species names and internal branches are maximum-likelihood estimates of N for each branch, relative to the human population size, obtained under model 8 (reflected gamma + invariable).

Results

Analysis of HIV-1 and Human Influenza Viral Data Sets

The estimates of α and β in the gamma distribution of ω for sites for the HIV-1 *env* gene were $\hat{\alpha} = 0.373$ and $\hat{\beta} = 0.523$. These estimates are obtained using the methods described in Yang et al. (2000) and Yang (2000). The resulting distribution of S is plotted in figure 1a. About 22.6% of all new mutations are advantageous with positive selection coefficients. This is also the proportion of sites at which $\omega > 1$. However, since advantageous mutations have higher probabilities of going to fixation in the population than neutral or deleterious mutations, they will account for a much larger proportion of observed substitutions. In fact this proportion is given by $\int_0^\infty f(S)(S/(1-e^{-S}))dS / \int_{-\infty}^{\infty} f(S)(S/(1-e^{-S}))dS$. For the HIV-1 *env* gene, this proportion is 0.749. Therefore, the 22.6% of advantageous mutations account for about three quarters of the substitutions.

The estimates of the parameters of the gamma distribution were $\hat{\alpha} = 0.306$ and $\hat{\beta} = 0.298$ for the influenza virus hemagglutinin gene (figure 1*b*). The proportion of new mutations that are positively selected is then 0.280, slightly higher than for the HIV-1 *env* gene. The proportion of fixed mutations that are positively selected is 0.851 for the hemagglutinin gene.

We notice that the assumption of no interference among mutations may not hold for the influenza data set. The result is that we probably tend to underestimate the magnitude of the selection coefficients. The real proportion of fixed mutations that are positively selected may, therefore, be even higher than 0.851.

Analysis of Mitochondrial Protein-Coding Genes

Table 1 summarizes the results obtained from analysis of the mtDNA data. The phylogenetic tree is shown in figure 2. The maximum-likelihood value for a model with no variation in N among lineages and no variation in S among sites is −38,884.38 (model 1). If we add variation in N among lineages, the maximum-likelihood value is −38,859.04 (model 2). This difference is significant ($2\Delta\ell = 50.68$, $P \approx 10^{-6}$, df = 12), indicating a lineage specific variation in the d_N/d_S ratio. Nevertheless, given the amount of data, the likelihood difference is not very large, although statistically significant.

Model 3 assumes that the selection coefficients are normally distributed among sites. The improvement in log likelihood over a model with no variation in s is 526.47. This difference is highly significant as model 3 has only one more parameter than model 1. Evidently, variation in S among sites is important in explaining the causes of molecular evolution in mitochondrial DNA. In previous studies (e.g., Nielsen and Yang 1998; Yang et al. 2000), it was also found that accounting for variation in ω among sites greatly increases the fit of the models. As we would expect based on knowledge of protein structure and function, the selection coefficients acting on new mutations are highly site specific.

Model 4 differs from model 3 in that no sites with $s > 0$ are allowed, and the model allows only neutral or negatively selected new mutations. The likelihood under this model is almost identical to that under model 3.

Allowing the selection coefficients to follow a reflected gamma distribution among sites (model 5) gives

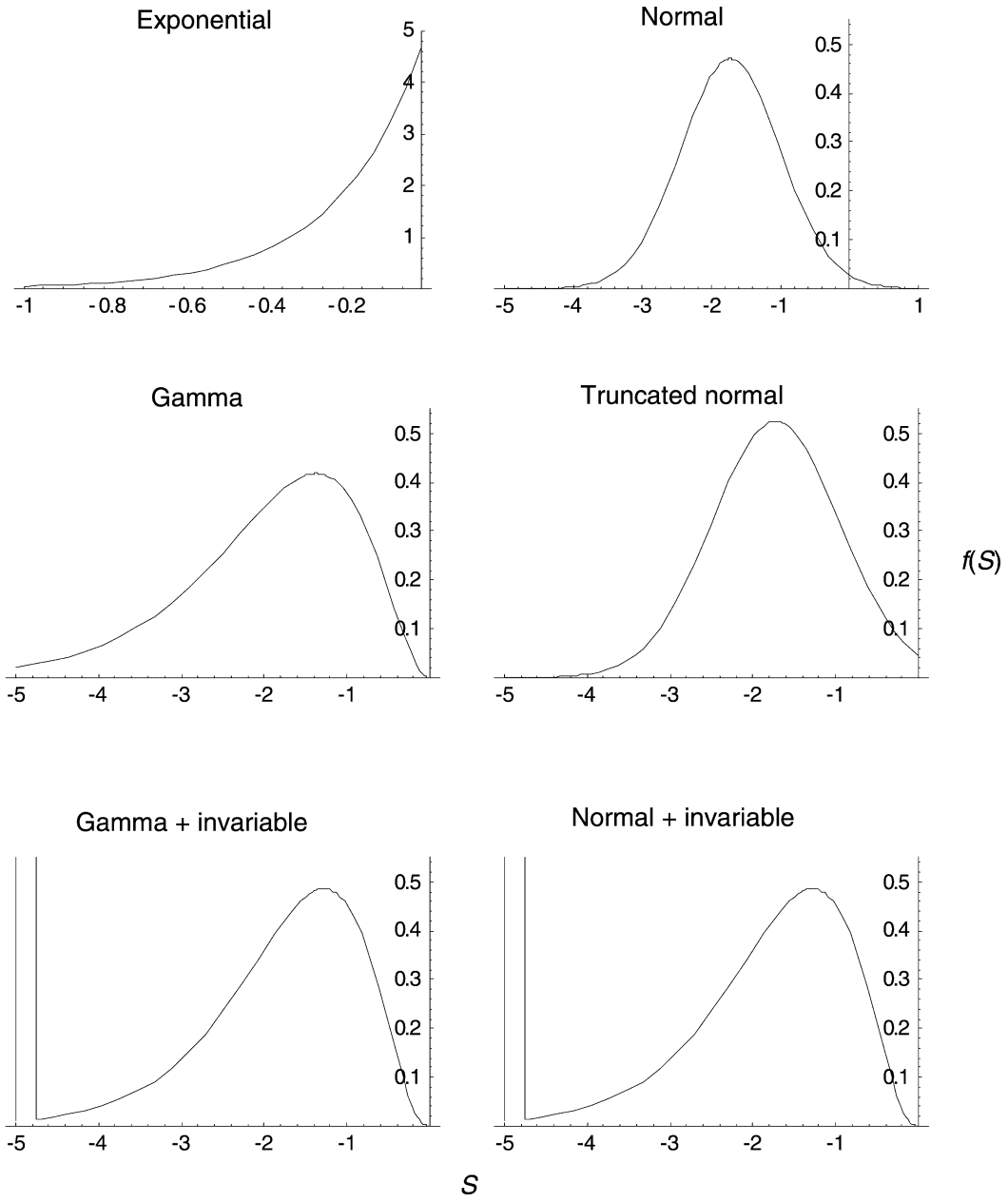

FIG. 3.—Estimated densities of scaled selection coefficients ($S = 2Ns$), where N is the human population size. Estimates are obtained from the mtDNA data for model 3 (normal), model 4 (truncated normal), model 5 (reflected gamma), model 6 (exponential), model 7 (normal + invariable), and model 8 (reflected gamma + invariable).

a maximum-likelihood value very similar to the value obtained under the normal distribution (model 4). It is not surprising that a normal and a reflected gamma distribution may give similar fit, since the gamma distribution tends to a normal distribution as α becomes large. By inspecting figure 3, we also notice that the estimated shapes of the distribution are quite similar under the two models.

The reflected exponential distribution (model 6) gives a much poorer fit to the data than the reflected gamma and the normal distributions. Even though model 6 only has one parameter less than models 3, 4, and 5, the maximum-likelihood value is about 140 likelihood units worse. This is quite a large difference. The explanation appears to be that the reflected exponential distribution cannot accommodate enough probability mass in the region $-2.5 < s < -0.5$.

Models 7, 8, and 9 are identical to models 3, 5, and 6 except that an extra category of invariable sites have been added. Adding a category of invariable sites led to only marginal increases in likelihood for all of the models, although the parameter estimates changed. The estimate of the proportion of invariable sites vary from 0.0 to 0.124. The explanation for the small increase in likelihood is probably not that invariable sites are rare, but more likely, that we do not have much power to distinguish between parameter-rich models. The same observation was made in Yang et al. (2000). The model providing the best fit, of all

the models analyzed here, is the one assuming a normal distribution and an additional category of invariable sites. This model places a lot of probability mass around $-2.5 < S < -0.5$, but also allows for some invariable sites.

Model 10 assumes that sites come in two flavors, sites that are completely invariable and sites in which all new mutations are neutral. This corresponds to the strictly neutral model in Nielsen and Yang (1998). This model performs significantly worse than all other models. It has the same number of parameters as model 1 but is nearly 1,000 likelihood units worse. Clearly, sites at which $-\infty < s < 0$ are dominant in the evolution of mtDNA.

The branch lengths in figure 2 are proportional to the expected number of synonymous mutations for the reflected gamma + invariable model of selection coefficients. In addition, the estimated population sizes, relative to the human population size, are also shown for each branch. Notice that all estimates are within the same order of magnitude. Although this is not theoretically impossible, it does appear to be somewhat unlikely. Further research is warranted to explain this observation.

Discussion

A model of normal or reflected gamma distributed selection coefficients appear to fit the data much better than a model assuming reflected exponentially distributed selection coefficients. The motivation for Kimura to suggest the gamma distribution (Kimura 1979) was that the reflected exponential distribution could not accommodate enough moderately selected mutations; that is, mutations with selection coefficients on the order of $1/N$. It seems that the mathematical arguments provided by Kimura (1979, 1981) now have found empirical support. A distribution with much of the probability mass located in the region around $1/N$ seems in fact to provide the best fit, and a reflected exponential distribution of selection coefficients can easily be rejected. Figure 3 shows the shapes of the estimated distributions. With the exception of the reflected exponential distribution, most of these shapes are similar to a lot of probability mass in the region $-2.5 < S < -0.5$. It seems that weakly selected mutations are highly important and that completely neutral and/or positively selected mutations are of only little importance in mammalian mtDNA evolution. This is also emphasized by the fact that a strictly neutral model provides a very poor fit to the data.

Very little is known from experimental data regarding the distribution of fitness effects of spontaneous mutations. Most experimental evidence regarding the fitness effects of new mutations comes from mutation accumulation experiments from Drosophila, such as the early experiments conducted by Mukai and colleagues (e.g., Mukai 1964; Mukai et al. 1972). In general, these experiments suggest that most mutations are deleterious and of small effect. Keightley (1994) reanalyzed data from Mukai et al. (1972) and Ohnishi (1974). Assuming a genome-wide mutation rate of 0.4, he obtained maximum-likelihood estimates of gamma distributions with very high variance of mutational effects ($\alpha < 1$); that is, with a majority of mutations having very little effect on fitness. This contrasts with our results, which suggest a distributions with very little probability mass centered around $S = 0$ and ($\alpha > 1$) for the gamma distribution. The data analyzed, the scaling of the parameters, and the underlying assumptions differ much between the two studies, so it is not surprising that the results differ. It is possible that our assumption of constant fitness effects of mutations occurring in the same site may increase the discrepancy between the two results.

In Yang et al. (2000), statistically significant (5% level) evidence for the existence of positively selected sites was detected in a different data set containing seven of the eight sequences analyzed here. A model in which ω is assumed to follow a beta distribution was compared with a model in which ω was assumed to follow a mixture distribution of a point mass located at $\omega > 1$ and a beta distribution. The two models were compared using a likelihood ratio test, which rejected the beta distribution model. However, in this study, it was found that a normal distribution of values of S fits the data worse than a normal distribution truncated at $S < 0$. One possible explanation for the discrepancy is the difference in model assumptions between the two studies. The test based on the beta distribution of ω may be more powerful. Also, the data differs slightly between the two studies by the inclusion of the baboon sequence in the present study.

Our major objectives in this paper were (1) to investigate whether we can use phylogenetic data to distinguish between different distributions of selection coefficients and (2) to determine which of the most common population genetics models of molecular evolution best fits the mtDNA data. It seems that we can in fact distinguish different distributions of selection coefficients. In particular, Kimura's (1979) effectively neutral model seems to fit the data much better than the exponential model of Ohta (1973), and a strictly neutral model (Kimura 1968) can be easily rejected. However, it appears that we do not have the power to distinguish between more parameter-rich distributions or between distributions with similar shapes, such as the reflected gamma and the normal distributions.

To make these inferences, we have to make some assumptions. Probably the most problematic assumption is that all new mutations at a site have the same selection coefficients. In the following, we discuss an alternative model that assumes that each of the four possible nucleotides in a site has a fixed fitness. The substitution process in a nucleotide site can then be described as a continuous time ergodic Markov chain with four states (1, 2, 3, 4) in which a transition between any of the states corresponds to a fixation event. The stationary frequencies of the Markov chain (p_1, p_2, p_3, p_4) obey the equations

$$p_i\left(\sum_{j\neq i} r_{ij}\right) = \sum_{j\neq i} p_j r_{ji} \quad p_1 + p_2 + p_3 + p_4 = 1, \qquad (8)$$

where r_{ij} is the rate of transition from state i to j. Assume that one of the nucleotides (say nucleotide 1) has a relative fitness of $1 + s$, and the remaining three nucleotides have a relative fitness of 1. Assume further that mutations between all four states occur at the same rate (μ), then

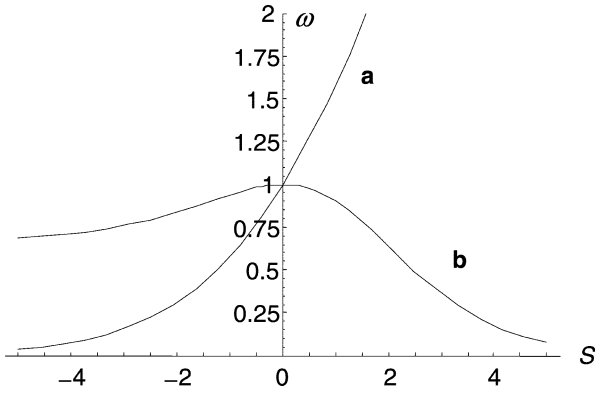

FIG. 4.—ω as a function of the selection coefficient on amino acid mutations for two models. In (*a*) all new mutations have the same selection coefficients in a haploid population with $S = 2Ns$. In (*b*) ω represents the rate of substitution in a nondegenerate site relative to the rate of substitution in a fourfold degenerate site. All four nucleotides have fixed and equal fitnesses, except that one of the nucleotides provides the individual carrying it with selective advantage *s*.

$$\begin{aligned} &p_1 \Pr(fix\ 1 \rightarrow 2,3,4) N\mu \\ &\quad = (p_2 + p_3 + p_4) \Pr(fix\ 2,3,4 \rightarrow 1) N\mu/3, \\ &\quad p_2 = p_3 = p_4, \quad p_1 + p_2 + p_3 + p_4 = 1 \end{aligned} \tag{9}$$

Where $\Pr(fix\ 2, 3, 4 \rightarrow 1)$ is the probability of fixation of a new type 1 mutant in a type 2, 3, or 4 background and $\Pr(fix\ 1 \rightarrow 2, 3, 4)$ is the probability of fixation of a new type 2, 3, or 4 mutant in a type 1 background. For small values of *s* (i.e., assuming $s \rightarrow 0$, $N \rightarrow \infty$, and $2Ns \rightarrow S$), the solution to these equations is given by

$$p_1 = \frac{e^S}{3 + e^S}, \quad p_2 = p_3 = p_4 = \frac{1}{3 + e^S}. \tag{10}$$

The nonsynonymous/synonymous rate ratio can then be calculated as

$$\omega = \frac{1}{\mu} \sum_{i=1}^{4} p_i \sum_{j \neq i} \frac{N\mu}{3} \Pr(fix\ i \rightarrow j) = \frac{2((S+1)e^S - 1)}{-3 + 2e^S + e^{2S}}. \tag{11}$$

This result could easily be generalized to cases of unequal mutation rates and more complicated selection schemes.

Equations 3 and 11 are plotted in figure 4. Notice that the pattern observed in the two models is quite different. In the first model, ω is a strictly increasing function of the selection coefficient. In the second model in contrast, the highest value of ω is obtained at $S = 0$. This model is effectively a model of constrained evolution in which a particular nucleotide represent the optimum. The difference between the two models demonstrates that any inferences will be strongly dependent on the specifics of the model.

The two models considered here represent extremes. In one case, the fitnesses associated with each allele remain constant over evolutionary time scales, and in the other case, the fitnesses are reassigned each time a new mutation occurs. More realistic evolutionary models would probably incorporate only occasional reassignments of fitnesses.

We have here explored the problem of how to estimate the distribution of selection coefficients from comparative data. We have discussed some of the necessary assumptions and some of the problems related to these assumptions. We hope this study will encourage population geneticists to seek further use of the very large amounts of available comparative data for addressing the basic questions in population genetics regarding the fitness effects of new mutations.

Acknowledgments

This research project was supported by NSF Grant DEB-0089487 to R.N., BBSRC Grants to Z.Y., and HFSP Grant RGY0055/2001-M to R.N. and Z.Y.

Literature Cited

Bush, R. M., W. M. Fitch, C. A. Bender, and N. J. Cox. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol. Biol. Evol. **16**:1457–1465.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl. 2002. The cost of inbreeding in *Arabidopsis*. Nature **416**:531–534.

Halpern, A. L., and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. **15**:910–917.

Fay, J. C., G. J. Wyckoff, and C.-I. Wu. 2001. Positive and negative selection on the human genome. Genetics **158**:1227–1234.

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

Fisher, R. A. 1930*a*. The distribution of gene ratios for rare mutations. Proc. R. Soc. Edinb. Sect. B **50**:204–219.

———. 1930*b*. The genetical theory of natural selection. 2nd edition. Dover Press, New York.

Gillespie, J. H. 1991. The causes of molecular evolution. Oxford University Press, Oxford.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* P. H. Harvey and L. Partridge, eds. Oxford surveys in evolutionary biology. Vol. 7. Oxford University Press, New York.

Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

Keightley, P. D. 1994. The distribution of mutation effects of viability in *Drosophila melanogaster*. Genetics **138**:1315–1322.

Kimura, M. 1962. On the probability of fixation of mutant genes in a population. Genetics **47**:713–719.

———. 1968. Evolutionary rate at the molecular level. Nature **217**:624–626.

———. 1979. Model of effectively neutral mutations in which selective constraint is incorporated. Proc. Natl. Acad. Sci. USA **75**:1934–1937.

———. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kingman, J. F. C. 1982. The coalescent. Stochast. Proc. Appl. **13**:235–248.

Mukai, T. 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. Genetics **50**:1–19.

Mukai, T., S. I. Chigusa, L. E. Metler, and J. F. Crow. 1972. Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. Genetics **72**:333–355.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to chloroplast genome. Mol. Biol. Evol. **11**:715–724.

Moran, P. A. P. 1962. The statistical processes of evolutionary theory. Clarendon Press, Oxford.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929–936.

Ohnishi, O. 1974. Spontaneous and ethyl methanesulfonate-induced mutations controlling viability in *Drosophila melanogaster*. Doctoral dissertation, University of Wisconsin, Madison.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature **246**:96–98.

———. 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23**:263–286.

Sawyer, S., and D. Hartl. 1992. Population genetics of polymorphism and divergence. Genetics **132**:1161–1176.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585–595.

Yamaguchi-Kabata, Y., and T. Gojobori. 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. J. Virol. **74**:4335–4350.

Yang, Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. **51**:423–432.

———. 2001. Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. Pacific Symp. Biocomp. **2001**: 226–237.

Yang, Z., and R. Nielsen. 2002. Codon substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. **19**:908–917.

Yang, Z., R. Nielsen, N. Goldman, and A.-M. K. Pedersen. 2000. Codon-substitution models for variable selection pressure at amino acid sites. Genetics **155**:431–449.

Yoder, A. D., and Z. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. **17**:1081–1090.

David Rand, Associate Editor

Accepted March 24, 2003