Copyright © 2000 by the Genetics Society of America

Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites

Ziheng Yang,* Rasmus Nielsen,† Nick Goldman‡ and Anne-Mette Krabbe Pedersen§

**Department of Biology, University College London, London NW1 2HE, United Kingdom, †Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, ‡Department of Genetics, University of Cambridge, Cambridge CB2 3EH, United Kingdom and §Department of Ecology and Genetics, University of Århus, Ny Munkegade, DK-8000 Århus C, Denmark*

Manuscript received October 26, 1999
Accepted for publication January 17, 2000

ABSTRACT

Comparison of relative fixation rates of synonymous (silent) and nonsynonymous (amino acid-altering) mutations provides a means for understanding the mechanisms of molecular sequence evolution. The nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$) is an important indicator of selective pressure at the protein level, with $\omega = 1$ meaning neutral mutations, $\omega < 1$ purifying selection, and $\omega > 1$ diversifying positive selection. Amino acid sites in a protein are expected to be under different selective pressures and have different underlying ω ratios. We develop models that account for heterogeneous ω ratios among amino acid sites and apply them to phylogenetic analyses of protein-coding DNA sequences. These models are useful for testing for adaptive molecular evolution and identifying amino acid sites under diversifying selection. Ten data sets of genes from nuclear, mitochondrial, and viral genomes are analyzed to estimate the distributions of ω among sites. In all data sets analyzed, the selective pressure indicated by the ω ratio is found to be highly heterogeneous among sites. Previously unsuspected Darwinian selection is detected in several genes in which the average ω ratio across sites is <1, but in which some sites are clearly under diversifying selection with $\omega > 1$. Genes undergoing positive selection include the β-globin gene from vertebrates, mitochondrial protein-coding genes from hominoids, the hemagglutinin (HA) gene from human influenza virus A, and HIV-1 *env*, *vif*, and *pol* genes. Tests for the presence of positively selected sites and their subsequent identification appear quite robust to the specific distributional form assumed for ω and can be achieved using any of several models we implement. However, we encountered difficulties in estimating the precise distribution of ω among sites from real data sets.

COMPARISON of synonymous (silent) and nonsynonymous (amino acid-altering) substitution rates provides an important means for studying the mechanisms of DNA sequence evolution (Kimura 1983; Gillespie 1991; Ohta 1993). As synonymous mutations are largely invisible to natural selection (but see Akashi 1995), while nonsynonymous mutations can be under strong selective pressure, comparison of the fixation rates of these two types of mutations provides a powerful tool for understanding the effect of natural selection on molecular sequence evolution. A measure that has featured prominently in such studies is the nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$), termed the "acceptance rate" by Miyata and Yasunaga (1980). Here the rates d_N and d_S are defined as the numbers of nonsynonymous and synonymous substitutions per site, and their ratio ω measures the selective pressure at the protein level. An $\omega > 1$ means that nonsynonymous mutations offer fitness advantages to the protein (individual) and have higher fixation probabilities than synonymous mutations. This is our working definition of positive selection (adaptive molecular evolution) in this article.

The ω ratio has almost always been calculated as an average over all codon (amino acid) sites in the gene and over the entire evolutionary time that separates the sequences. The criterion that this average ω be >1 is a very stringent one for detecting positive selection (*e.g.*, Sharp 1997; Akashi 1999; Crandall *et al.* 1999). In many proteins, a high proportion of amino acids may be largely invariable (with ω close to 0) due to strong functional constraints. Furthermore, most proteins appear to be under purifying selection most of the time (Li 1997). Adaptive evolution most likely occurs at a few time points and affects a few amino acids. In such cases, the ω ratio averaged over time and over sites will not be significantly >1 even if adaptive molecular evolution has occurred. For example, Endo *et al.* (1996) used this criterion to perform a database search and identified 17 proteins out of 3595 that may have been under positive selection, at a proportion of only 0.47%. The scarcity of well-established cases of molecular adaptation may be partly due to the lack of power of the detection methods.

An alternative approach is to examine the ω ratio over a short evolutionary time (for example, along particular

Corresponding author: Ziheng Yang, Department of Biology, 4 Stephenson Way, London NW1 2HE, United Kingdom.
E-mail: z.yang@ucl.ac.uk

lineages in a phylogeny) or in functionally distinct regions of the gene. Messier and Stewart (1997) used inferred ancestral sequences to identify two lineages in a phylogeny of primates that are probably under diversifying selection for the lysozyme gene. Hughes and Nei (1988) found that the ω ratio is >1 in a region of the major histocompatibility complex (MHC) that codes for the antigen recognition site, while it is <1 in other regions of the gene. When knowledge of functional domains of the protein is unavailable, or when only a few sites are expected to be undergoing positive selection, a promising approach is to devise statistical models that allow for heterogeneous ω ratios among sites (Nielsen and Yang 1998). Such models can be used to test and identify critical amino acids in a protein under positive selection.

Nielsen and Yang (1998) implemented a few simple models that allow for heterogeneous ω ratios among sites. One, the "neutral" model, assumes the existence of two classes of sites: conserved sites, at which nonsynonymous mutations are deleterious and removed ($\omega = 0$), and completely neutral sites at which $\omega = 1$. A "selection" model adds a third class of sites with the underlying ω ratio estimated from the data. These models appear too simple to capture the complexity of the substitution process of various proteins. In particular, the neutral model does not allow for sites with $0 < \omega < 1$, such as sites at which nonsynonymous mutations are "slightly deleterious." The selection model with one additional category cannot account for both sites with $0 < \omega < 1$ and sites with $\omega > 1$.

In this article, we implement a number of models (statistical distributions) for heterogeneous ω ratios among sites. We have two major objectives in fitting those models. The first is to test for the presence of positively selected sites (sites with $\omega > 1$) and to identify such sites along the gene. We find that this type of analysis seems insensitive to the exact distribution assumed for ω. We note that this test of molecular adaptation is still conservative, as it requires that positively selected sites be under diversifying selection along all lineages on the phylogeny. If positive selection affects only a few lineages, and purifying selection dominates during the rest of the evolutionary time, our test will fail.

Our second objective is to understand what distributions best describe the heterogeneous ω ratios among sites in real data. This appears to be a much harder task. Nevertheless, the distribution of ω is closely related to the fitness distribution of new mutations, and knowledge of it is important for testing competing population genetics models of molecular evolution (Kimura 1983; Gillespie 1991; Ohta 1993).

Ten data sets of genes from nuclear, mitochondrial, and viral genomes are analyzed to examine the fit of the models and to accumulate empirical knowledge of the ω distribution among sites. These analyses reveal diversifying selection in several genes for which it was not previously suspected.

DATA

Ten data sets of protein-coding genes are analyzed. Table 1 lists the number of sequences (s) and the sequence length (n) for each data set. The transition/transversion rate ratio (κ), the (average) nonsynonymous/synonymous rate ratio (ω), and the tree length (sum of all branch lengths) are estimated under the simple model of one ω ratio for all sites (Goldman and Yang 1994, model M0 below). These statistics are listed to give an indication of the sequence divergence level and the selective constraint involved in each gene. More details of the data sets follow.

D1: Mitochondrial protein-coding genes from the hominoids: The 12 protein-coding genes on the H-strand of the mitochondrial genome are concatenated and analyzed as one data set. All the 12 proteins are transmembrane proteins and appear to have similar substitution patterns (Kumar 1996). The seven species are human, common chimpanzee, pygmy chimpanzee, gorilla, Bornean orangutan, Sumatran orangutan, and common gibbon. The data are a subset of the data analyzed by Cao *et al.* (1998), where the GenBank accession numbers and references are given. The transition/transversion bias is strong, and the genes appear under strong purifying selection with ω at ~0.04.

D2: Vertebrate β-globin genes: The β-globin gene of 17 vertebrate species (human, tarsier, bush baby, hare, rabbit, cow, sheep, pig, elephant seal, rat, mouse, hamster, opossum, duck, chicken, African clawed frog, and western clawed frog) were extracted from the EMBL and GenBank databases. The sequences were aligned by hand, and the alignment appeared straightforward.

D3: Drosophila *adh* gene: The data set contains alcohol dehydrogenase (adh) gene sequences from 23 species of Drosophila. The data set was described in Nielsen (1997), where GenBank accession numbers for the sequences are given. The transition/transversion rate ratio (1.6) is relatively low for this data set.

D4: Flavivirus E-glycoprotein gene: Twenty-two dengue virus E-glycoprotein gene sequences from the alignment published by Zanotto *et al.* (1996) were used. The original alignment contains 123 sequences from more divergent groups. The 22 sequences we use are 6, 7, 7, and 2 sequences from the closely related Den 1–4 groups, respectively. The gene appears to be under strong purifying selection (with an average ω ratio of ~0.05).

D5: Human influenza virus type A HA gene: The data set is a subset of the HA1 domain of the hemagglutinin (HA) gene of human influenza viruses A analyzed by Fitch *et al.* (1997). We selected 28 sequences to represent major groups in the original data set. The HA gene encodes the major surface antigen, which is a target of

TABLE 1

Basic statistics for data sets analyzed in this article

Data set	*s*	*n*	κ	ω	*S*	PS
D1: mitochondrial gene from hominoids	7	3331	14.25	0.041	2.79	Y
D2: β-globin gene from vertebrates	17	144	2.07	0.237	7.12	Y
D3: Drosophila alcohol dehydrogenase (*adh*) gene	23	254	1.58	0.094	4.20	N
D4: flavivirus E-glycoprotein gene	22	490	3.94	0.052	12.36	N
D5: human influenza virus A hemagglutinin (HA) gene	28	329	4.62	0.391	0.85	Y
D6: HIV-1 *vif* gene	29	192	3.72	0.644	2.88	Y
D7: HIV-1 *pol* gene	23	947	4.89	0.196	1.18	Y
D8: Japanese encephalitis *env* gene	23	500	9.52	0.051	2.54	N
D9: tick-borne flavivirus NS-5 gene	18	342	2.25	0.025	26.13	N
D10: HIV-1 *env* gene V3 region	13	91	2.47	0.901	1.76	Y

s, number of sequences; *n*, number of codons in the sequence; κ, transition/transversion rate ratio (α/β in the notation of Kimura 1980); ω, nonsynonymous/synonymous rate ratio, averaged over sites (d_N/d_S); *S*, tree length, measured by the number of nucleotide substitutions along the tree per codon; PS, positive selection; Y, yes; N, no.

neutralizing antibodies produced during infection or vaccination.

D6: HIV-1 *vif* gene: A total of 29 isolates from subtype B are used. The HIV-1 *vif* gene encodes an accessory protein that is believed to be essential for pathogenic infection, but its exact function is not known (*e.g.*, Emerman and Malim 1998).

D7: HIV-1 *pol* gene: A total of 23 isolates from subtype B are used. The HIV-1 *pol* gene encodes for three proteins: the reverse transcriptase responsible for reverse transcribing the viral RNA into DNA; the polymerase responsible for cleaving the *pol* and *gag* precursor proteins into their final products; and the endonuclease (integrase) responsible for cleaving the host DNA so that the HIV DNA can be inserted. All three proteins are essential for virus replication.

D8: Japanese encephalitis *env* gene: Sequences from 23 isolates are used.

D9: Tick-borne flavivirus NS-5 gene: The data are from Kuno *et al.* (1998). The sequences are highly divergent, and the gene appears to be under strong purifying selection (with the average $\omega = 0.025$).

D10: HIV-1 *env* gene V3 region: HIV-1 *env* gene V3 region from 13 HIV-1 isolates with a known transmission history was published by Leitner *et al.* (1997).

The alignment for the influenza virus HA gene (D5) was kindly provided by Walter Fitch and those for data sets D6–D9 by Edward Holmes.

THEORY

Markov model of codon substitution: Suppose there are *n* codons (sites) in the sequence. Let the data at site *h* ($h = 1, 2, \ldots, n$) be $\mathbf{x}_h$; that is, $\mathbf{x}_h$ is a vector of codons from different sequences at that codon site. Models considered in this article allow the d_N/d_S (ω) ratio to vary among sites. Let $\omega^{(h)}$ be the ratio for site *h*. The codon substitution model specifies the relative instantaneous substitution rate from codon *i* to *j* at site *h* ($h = 1, 2, \ldots, n$) as

$$q_{ij}^{(h)} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three nucleotide positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by one synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by one synonymous transition,} \\ \omega^{(h)}\pi_j, & \text{if } i \text{ and } j \text{ differ by one nonsynonymous transversion,} \\ \omega^{(h)}\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by one nonsynonymous transition.} \end{cases} \quad (1)$$

Parameter κ is the transition/transversion rate ratio and π_j is the equilibrium frequency of codon *j*. Different assumptions can be made concerning π_j (Goldman and Yang 1994; Muse and Gaut 1994; Pedersen *et al.* 1998); in this article, they are calculated using the products of the observed nucleotide frequencies at each of the three codon positions. The transition probability matrix is calculated as $P(t) = e^{Qt}$ (Liò and Goldman 1998), where time or branch length *t* is measured by the expected number of nucleotide substitutions per codon, averaged over all sites.

Likelihood calculation under a model of heterogeneous ω ratios among sites: Using one ω parameter for each site would lead to too many parameters in the model. Instead we use a statistical distribution to account for heterogeneous ω ratios among sites (Nielsen and Yang 1998). Suppose we assume *K* classes of sites in the sequence, with the proportions and ω ratios given as

$$\begin{matrix} p_0 & p_1 & \ldots & p_{K-1} \\ \omega_0 & \omega_1 & \ldots & \omega_{K-1} \end{matrix} \quad (2)$$

We want to calculate the probability of observing data $\mathbf{x}_h$ at site *h*. Note that given the ω ratio for the site, the conditional probability of the data, $P(\mathbf{x}_h|\omega)$, can be calculated for any phylogenetic tree and branch lengths using Felsenstein's (1981) pruning algorithm (see also Goldman and Yang 1994; Muse and Gaut 1994). The (marginal) probability of the data at the site is then an

average of the conditional probability over the above distribution of ω:

$$P(\mathbf{x}_h) = \sum_{k=0}^{K-1} p_k P(\mathbf{x}_h|\omega_k). \quad (3)$$

We assume that the substitution process is independent among codon sites, and then the log-likelihood is a sum over all sites in the sequence

$$\ell = \sum_{h=1}^{n} \log\{P(\mathbf{x}_h)\}. \quad (4)$$

After maximum-likelihood (ML) estimates of parameters are obtained, an empirical Bayes approach can be used to infer which class the site is most likely from (Nielsen and Yang 1998). The posterior probability that site h with data $\mathbf{x}_h$ is from site class k (*i.e.*, that $\omega^{(h)} = \omega_k$) is

$$P(\omega_k|\mathbf{x}_h) = \frac{p_k P(\mathbf{x}_h|\omega_k)}{P(\mathbf{x}_h)} = \frac{p_k P(\mathbf{x}_h|\omega_k)}{\sum_j p_j P(\mathbf{x}_h|\omega_j)}. \quad (5)$$

The class k that maximizes the posterior probability is the most likely class for the site. When the ω values (ω_k) for some site classes are >1, this approach can be used to identify sites under positive selection. The posterior probabilities corresponding to classes with $\omega > 1$ may be summed up to give a probability $P(\omega > 1)$ for each site. Sites at which this probability is larger than a threshold value (say, 50, 95, or 99%) may be identified as potentially under positive selection.

Continuous distributions: In theory, a continuous distribution for ω, say, $f(\omega)$, can be used in the likelihood calculation. The sum in Equation 3 will then be replaced by an integral. However, we have not found a feasible way to perform this calculation and instead resort to the use of a discrete distribution as an approximation, as in Yang (1994). We use $K = 10$ categories in the discrete distribution, each with probability 1/10, and use the median value of ω within each category to represent the distribution of ω ratios within that category. The p's and ω's in Equation 2 are then calculated as functions of parameters in the continuous ω distribution. Let $F(\omega)$ be the cumulative distribution function (CDF) of the ω distribution $f(\omega)$; $F(\omega) = \int_0^\omega f(x)dx$. Let $F^{-1}(\cdot)$ be the inverse CDF; that is, if $p = F(\omega)$, then $\omega = F^{-1}(p)$. We then have $p_k = 1/K$ and $\omega_k = F^{-1}((2k + 1)/(2K))$ for $k = 0, 1, \ldots, K - 1$. The CDF of the ω distribution can be calculated in a straightforward manner for all distributions we consider, and we use a linear-search algorithm to obtain the inverse CDF.

The probability of the data $\mathbf{x}_h$ is then approximated by

$$P(\mathbf{x}_h) = \int_0^\infty f(\omega)P(\mathbf{x}_h|\omega)d\omega \approx \frac{1}{K}\sum_{k=0}^{K-1} P(x_h|\omega_k). \quad (6)$$

This may be considered a crude way of doing numerical integration.

Statistical distributions implemented: The models considered in this article are summarized in Table 2. The codes given are as used in the paml program (Yang 1997), which now implements all the models described

TABLE 2

Models of variable ω ratios among sites

Model code	p	Parameters	Notes
M0 (one-ratio)	1	ω	One ω ratio for all sites
M1 (neutral)	1	p_0	$p_1 = 1 - p_0$, $\omega_0 = 0$, $\omega_1 = 1$
M2 (selection)	3	p_0, p_1, ω_2	$p_2 = 1 - p_0 - p_1$, $\omega_0 = 0$, $\omega_1 = 1$
M3 (discrete)	$2K - 1$ ($K = 3$)	p_0, p_1, . . . , p_{K-2}, ω_0, ω_1, . . . , ω_{K-1}	$p_{K-1} = 1 - p_0 - p_1 - \ldots - p_{K-2}$
M4 (freqs)	$K - 1$ ($K = 5$)	p_0, p_1, . . . , p_{K-2}	The ω_k are fixed at 0, 1/3, 2/3, 1, and 3
M5 (gamma)	2	α, β	From $\mathcal{G}(\alpha, \beta)$
M6 (2gamma)	4	p_0, α_0, β_0, α_1	p_0 from $\mathcal{G}(\alpha_0, \beta_0)$ and $p_1 = 1 - p_0$ from $\mathcal{G}(\alpha_1, \alpha_1)$
M7 (beta)	2	p, q	From $\mathcal{B}(p, q)$
M8 (beta&ω)	4	p_0, p, q, ω	p_0 from $\mathcal{B}(p, q)$ and $1 - p_0$ with ω
M9 (beta&gamma)	5	p_0, p, q, α, β	p_0 from $\mathcal{B}(p, q)$ and $1 - p_0$ from $\mathcal{G}(\alpha, \beta)$
M10 (beta&gamma+1)	5	p_0, p, q, α, β	p_0 from $\mathcal{B}(p, q)$ and $1 - p_0$ from $1 + \mathcal{G}(\alpha, \beta)$
M11 (beta&normal>1)	5	p_0, p, q, μ, σ	p_0 from $\mathcal{B}(p, q)$ and $1 - p_0$ from $\mathcal{N}(\mu, \sigma^2)$, truncated to $\omega > 1$
M12 (0&2normal>1)	5	p_0, p_1, μ_2, σ_1, σ_2	p_0 with $\omega_0 = 0$ and $1 - p_0$ from the mixture: p_1 from $\mathcal{N}(1, \sigma_1^2)$, and $1 - p_1$ from $\mathcal{N}(\mu_2, \sigma_2^2)$, both normals truncated to $\omega > 1$
M13 (3normal>0)	6	p_0, p_1, μ_2, σ_0, σ_1, σ_2	p_0 from $\mathcal{N}(0, \sigma_0^2)$, p_1 from $\mathcal{N}(1, \sigma_1^2)$, and $p_2 = 1 - p_0 - p_1$ from $\mathcal{N}(\mu_2, \sigma_2^2)$, all normals truncated to $\omega > 1$

p, number of parameters in the ω distribution.

here. All models involve the following parameters: branch lengths in the phylogeny, the transition/transversion rate ratio κ, and base frequencies at the three codon positions. These parameters are not listed in Table 2. Model 0 (M0) assumes one ω for all sites (Goldman and Yang 1994). The neutral model (M1) assumes a proportion p_0 of conserved sites with $\omega_0 = 0$ and a proportion $p_1 = 1 - p_0$ of neutral sites with $\omega_1 = 1$. The selection model (M2) adds an additional class of sites with frequency $p_2 = 1 - p_0 - p_1$ and with ω_2 estimated from the data. M1 and M2 were implemented by Nielsen and Yang (1998). A number of new models are implemented in this article to accommodate different shapes of the ω distribution that are likely to occur in real data. The details follow.

M3 (discrete): The discrete model uses an unconstrained discrete distribution to model heterogeneous ω ratios among sites (Equation 2). The ω_k values are arranged in increasing order for unique identification. The model with K classes involves $K - 1$ frequency parameters and K parameters ω_k. In this article, $K = 3$ classes are used.

M4 (freqs): This model fixes ω at several prespecified values and estimates the corresponding frequencies for the site classes. The model with K classes involves $K - 1$ free frequency parameters p_k. We use $K = 5$, with ω_k fixed at 0, $\frac{1}{3}$, $\frac{2}{3}$, 1, and 3 for $k = 0, 1, \ldots, 4$.

M5 (gamma): This model assumes a simple gamma distribution $\mathcal{G}(\alpha, \beta)$ for ω among sites. The density function is $f(\omega) = \beta^\alpha e^{-\beta\omega}\omega^{\alpha-1}/\Gamma(\alpha)$ for $\omega > 0$. The CDF, known as the incomplete gamma function ratio, is

$$F_G(\omega; \alpha, \beta) = \int_0^\omega \beta^\alpha e^{-\beta x} x^{\alpha-1} dx/\Gamma(\alpha). \quad (7)$$

This is calculated using the algorithm of Bhattacharjee (1970).

M6 (2gamma): This model uses a mixture of two gamma distributions, $\mathcal{G}(\alpha_0, \beta_0)$ and $\mathcal{G}(\alpha_1, \beta_1)$, in the proportions p_0 and $p_1 = 1 - p_0$. The mean of the second gamma distribution is fixed at 1 ($\beta_1 = \alpha_1$), so that the model has four parameters. The CDF of the mixture distribution can be calculated from its gamma distribution components.

M7 (beta): The beta distribution $\mathcal{B}(p, q)$ has density

$$f(\omega) = \omega^{p-1}(1 - \omega)^{q-1}/B(p, q), \quad 0 \le \omega \le 1, \quad (8)$$

where $B(p, q)$ is the beta function. The beta distribution can take different shapes (*e.g.*, L, J, U, and inverted-U shapes) in the interval (0, 1). The CDF of the distribution, $F_B(\omega; p, q)$, is the incomplete beta function ratio, calculated using the algorithm of Majumder and Bhattacharjee (1973). This model does not allow for positively selected sites (with $\omega > 1$). The following four models (M8–M11) add an extra component, mainly to account for the possible occurrence of positively selected sites.

M8 (beta&ω): This model adds one extra class of sites to the beta model. A proportion p_0 of sites have ω drawn from the beta distribution $\mathcal{B}(p, q)$, and the remaining sites (proportion $p_1 = 1 - p_0$) have the same ratio ω_1. This model can be compared with the beta model (M7) to test for the presence of positive sites using a likelihood-ratio test (LRT; see below).

M9 (beta&gamma): This model uses a mixture of $\mathcal{B}(p, q)$ and $\mathcal{G}(\alpha, \beta)$, in proportions p_0 and $p_1 = 1 - p_0$. The CDF of the mixture distribution is an average of the beta and gamma CDFs.

M10 (beta&gamma+1): This model uses a mixture of a beta and a gamma. However, the gamma is shifted to the right by one unit, so that the gamma distribution accounts for positively selected sites ($\omega > 1$) only. A proportion p_0 of sites have ω from $\mathcal{B}(p, q)$ and a proportion $p_1 = 1 - p_0$ of sites have $\omega = 1 + x$, where $x \sim \mathcal{G}(\alpha, \beta)$. The CDF of ω is thus

$$F(\omega) = \begin{cases} p_0 F_B(\omega; p, q), & \text{if } \omega \le 1, \\ p_0 + p_1 F_G(\omega - 1; \alpha, \beta), & \text{if } \omega > 1. \end{cases} \quad (9)$$

M11 (beta&normal>1): This model uses a mixture of a beta distribution $\mathcal{B}(p, q)$ and a normal distribution $\mathcal{N}(\mu, \sigma^2)$, which is left-truncated at 1. Like the shifted gamma in model M10, the truncated normal accounts for positively selected sites only. The CDF for the truncated normal is $1 - \Phi((\mu - \omega)/\sigma)/\Phi((\mu - 1)/\sigma)$, where $\Phi(\cdot)$ is the familiar CDF of $\mathcal{N}(0, 1)$. Therefore the CDF of the mixed distribution is

$$F(\omega) = \begin{cases} p_0 F_B(\omega; p, q), & \text{if } \omega \le 1, \\ p_0 + (1 - p_0)\left(1 - \dfrac{\Phi((\mu - \omega)/\sigma)}{\Phi((\mu - 1)/\sigma)}\right), & \text{if } \omega > 1. \end{cases} \quad (10)$$

M12 (0&2normal>0): This model assumes a proportion p_0 of sites with $\omega = 0$ and a proportion $(1 - p_0)$ from a mixture of two normal distributions. This mixture is p_1 from $\mathcal{N}(1, \sigma_1^2)$, and $(1 - p_1)$ from $\mathcal{N}(\mu_2, \sigma_1^2)$, truncated at $\omega = 0$ to disallow values of $\omega < 0$. Consequently, the ω distribution for M12 takes value 0 with probability p_0, is drawn from a truncated normal distribution centered at $\omega = 1$ with probability $(1 - p_0)p_1$, and is drawn from a truncated normal distribution centered at $\omega = \mu_2$ with probability $(1 - p_0)(1 - p_1)$. The CDF of the normal $\mathcal{N}(\mu, \sigma^2)$ truncated at 0 is $1 - \Phi((\mu - \omega)/\sigma)/\Phi(\mu/\sigma)$. The CDF for the mixture of the two truncated normal distributions is

$$F(\omega) = 1 - p_1\Phi((\mu_1 - \omega)/\sigma_1)/\Phi(\mu_1/\sigma_1) - (1 - p_1)\Phi((\mu_2 - \omega)/\sigma_2)/\Phi(\mu_2/\sigma_2). \quad (11)$$

M13 (3normal>0): This model uses a mixture of three normal distributions truncated at $\omega = 0$: p_0 from $\mathcal{N}(0, \sigma_0^2)$, p_1 from $\mathcal{N}(1, \sigma_1^2)$, and $p_2 = 1 - p_0 - p_1$ from $\mathcal{N}(\mu_2, \sigma_2^2)$. The CDF is

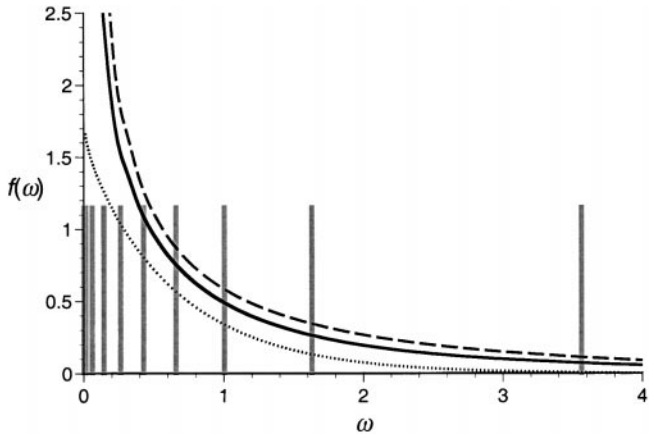

Figure 1.—Discretization of model M6 (2gamma). Parameter estimates are taken from the HIV-1 *vif* gene (data set D6; see Table 8). The dotted line represents the gamma density $\mathcal{G}(0.967, 1.452)$ and the dashed line represents the gamma density $\mathcal{G}(0.283, 0.283)$. The thick line is the mixture of the former two in proportions $p_0 = 0.383$ and $p_1 = 0.617$. The discretized version of the model uses 10 site classes, each of proportion 10%. The ω ratios for the 10 classes are 0.0003, 0.0135, 0.0598, 0.1424, 0.2621, 0.4267, 0.6569, 1.0037, 1.6282, and 3.5598.

$$F(\omega) = 1 - p_0 \times 2\Phi(-\omega/\sigma_0) - p_1\Phi((\mu_1 - \omega)/\sigma_1)/\Phi(\mu_1/\sigma_1) - p_2\Phi((\mu_2 - \omega)/\sigma_2)/\Phi(\mu_2/\sigma_2). \quad (12)$$

M13 involves six parameters and can represent a smooth distribution with as many as three modes. However, the use of many parameters in M13 was found to cause problems for the optimization algorithm.

We use 10 categories ($K = 10$) to approximate the continuous part of the ω distribution, so that 11 categories are actually used in models M8 and M12. An example is shown in Figure 1, where model M6 (2gamma) is discretized.

The continuous mixture distributions involving the beta (M9, M10, and M11) are not smooth at $\omega = 1$. Furthermore, the discretized versions of those models may not be powerful to detect positive selection. Because we use 10 categories in the discrete distribution, each of proportion 10%, there may not be any category with $\omega > 1$, if the proportion of sites in the data under positive selection is substantially <10%. In such a case, the model may not suggest positive selection, even if the parameter estimates suggest a component of positively selected sites (that is, even if $p_1 > 0$). It appears important to examine the inferred discrete distribution to determine whether the model suggests sites under selection, as is the case with other continuous mixture models (M5, M6, M12, and M13).

Phylogenetic tree topologies are obtained by ML search using simple nucleotide substitution models. For the purpose of this study, minor differences in the phylogeny do not make any significant difference in estimation of parameters or identification of positively selected sites (see below). The codon frequencies are estimated using the observed nucleotide frequencies at the three codon positions, while other parameters are estimated by numerical maximization of the likelihood function. Independent computer programs were written by Z. Yang and R. Nielsen for error checking.

Likelihood-ratio tests to compare models: The LRT may be used to compare models implemented in this article. When two models are nested, twice the log-likelihood difference will be compared with a χ^2 distribution with the degrees of freedom ν equal to the difference in the number of parameters between the two models. For example, models M2 (selection) and M3 (discrete) are more general than M1 (neutral) and can be compared with M1. When $\omega_2 > 1$ in M2 or M3, this becomes a test for the presence of positively selected sites. Similarly, M7 (beta) is a special case of models M8 (beta&ω), M9 (beta&gamma), M10 (beta&gamma+1), and M11 (beta&normal>1). Any of those more general models can be compared with M7. When the more general models indicate the presence of sites with $\omega > 1$, the comparison constitutes an LRT of positive selection. We note that the null hypothesis in those tests, M1 (neutral) or M7 (beta), corresponds to fixing one of the parameters at the boundary of the parameter space of the alternative hypothesis; that is, the proportion for the component of positively selected sites (p_2 in M2 and M3 and p_1 in M8–M11) is set to zero in the null hypothesis. In such cases, use of the simple χ^2 distribution is not reliable (Self and Liang 1987), and caution is needed when the test statistic is close to the critical value.

Although the continuous version of model M8 is nested within each of the continuous versions of models M9–M11, this is not the case with the discretized versions used in this article. Those models cannot be compared using a χ^2 approximation to the test statistic, although in theory the null distribution can be generated by Monte Carlo simulation (Goldman 1993). In comparing those models, we use the Akaike Information Criterion (Akaike 1974) as a guidance, which stipulates that one extra parameter should be counted as an improvement of one log-likelihood unit. However, many of the mixture distributions we implemented (for example, M8–M13) are found to fit data sets analyzed in this article about equally well, leaving us with little power to discriminate among different ω distributions and also rendering formal statistical tests among those models unnecessary.

RESULTS

Likelihood values and parameter estimates obtained from different data sets are listed in Tables 3–12. Estimates of branch lengths are not shown, although their sum (the tree length) under model M0 is given in Table 1. Estimates of the transition / transversion rate ratio (κ)

are quite homogeneous among models in each data set and thus are not shown in Tables 3–12. For example, $\hat{\kappa}$ ranges from 14.3 to 17.0 among the 14 models for the mitochondrial data set (D1), while for the β-globin genes (D2), $\hat{\kappa}$ ranges from 2.0 to 2.4. Estimates of branch lengths and of κ from the models of heterogeneous ω among sites (M1–M13) are usually slightly larger than estimates from the model of one ω for all sites (M0). This is probably due to the insufficient correction for multiple nonsynonymous substitutions at the same site by the one-ratio model (M0).

Our focus is the distribution of ω among sites, and we are interested in the shape of the ω distribution, as indicated by parameter estimates under different models and the fit of the models to data measured by their log-likelihood values. We discuss tests for the presence of positively selected sites and, when such sites exist, their identification. In the following, we describe results obtained from each data set. The general patterns are summarized in the discussion.

D1: Mitochondrial protein-coding genes from hominoids (Table 3): The mitochondrial genes are highly conserved. The average ω ratio ranges from 0.04 to 0.05 among all models except for M1 (neutral) and M4 (freqs), suggesting that a nonsynonymous mutation has only 4–5% as much chance as a synonymous mutation of being fixed in the population. Models M1 and M4 do not fit the data well and gave much larger and probably unreliable estimates of the average ω ratio. The selective pressure indicated by the ω ratio is highly variable among sites. For example, the model of one ω ratio for all sites (M0) is rejected by a big margin when compared with model M3 (discrete), which allows for three classes of sites with different ω ratios. The LRT statistic for this comparison is $2\Delta\ell = 2 \times [-29690.55 - (-29967.86)] = 554.62$, much greater than critical values from a χ^2 distribution with d.f. = 4.

Parameter estimates from models such as M5 (gamma) and M7 (beta) suggest highly skewed L shapes for the ω distribution, with most amino acids highly conserved or almost invariable. The strictly neutral model (M1) fits the data worse than the one-ratio model (M0). The reason appears to be that M1 does not account for sites with positive but very small ω ratios. For similar reasons, M4 (freqs) fits the data poorly, as it does not account for highly conserved sites with $0 < \omega < 1/3$ (recall that our implementation of this model fixes ω at 0, 1/3, 2/3, 1, and 3). Models M9–M11, which add a continuous component to the beta distribution, do not fit the data as well as model M8, which adds a class of sites with a constant ω ratio.

The discrete model (M3) suggests a small proportion of sites ($p_1 = 0.7\%$) under positive selection, with $\omega_2 = 1.43$. This model fits the data significantly better than M0 (one-ratio), as mentioned above, or M1 (neutral). Similarly, model M8 (beta&ω) suggests a small proportion of sites ($p_1 = 0.6\%$) under diversifying selection with $\omega_1 = 1.46$. The LRT statistic for comparing M7

TABLE 3

Likelihood values and parameter estimates for hominoid mitochondrial genes (D1)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−29967.86	0.041	$\omega = 0.041$
M1 (neutral)	−30487.41	0.203	$p_0 = 0.798$ ($p_1 = 0.202$)
M2 (selection)	−29693.68	0.050	$p_0 = 0.658$, $p_1 = 0.014$ ($p_2 = 0.328$), $\omega_2 = 0.109$
M3 (discrete)	−29690.55	0.050	$p_0 = 0.811$, $p_1 = 0.182$ ($p_2 = 0.007$) $\omega_0 = 0.009$, $\omega_1 = 0.180$, $\omega_2 = 1.435$
M4 (freqs)	−29836.60	0.082	$p_0 = 0.761$, $p_1 = 0.237$, $p_2 = 0.0$, $p_3 = 0.0$ ($p_4 = 0.001$)
M5 (gamma)	−29698.84	0.047	$\alpha = 0.204$, $\beta = 3.781$
M6 (2gamma)	−29696.67	0.047	$p_0 = 0.382$ ($p_1 = 0.618$) $\alpha_0 = 36.43$, $\beta_0 = 800.0$, $\alpha_1 = \beta_1 = 0.018$
M7 (beta)	−29699.38	0.047	$p = 0.189$, $q = 3.526$
M8 (beta&ω)	−29690.71	0.050	$p_0 = 0.994$ ($p_1 = 0.006$) $p = 0.219$, $q = 4.669$, $\omega = 1.459$
M9 (beta&gamma)	−29696.89	0.047	$p_0 = 0.951$ ($p_1 = 0.049$) $p = 0.375$, $q = 13.519$, $\alpha = 1.633$, $\beta = 0.032$
M10 (beta&gamma+1)	−29696.89	0.047	$p_0 = 0.951$ ($p_1 = 0.049$), $p = 0.378$, $q = 13.670$, $\alpha = 2.354$, $\beta = 4.697$
M11 (beta&normal>1)	−29696.89	0.047	$p_0 = 0.951$ ($p_1 = 0.049$), $p = 0.377$, $q = 13.660$, $\mu = 2.376$, $\sigma = 1.369$
M12 (0&2normal>1)	−29693.46	0.048	$p_0 = 0.624$, $p_1 = 0.054$, $\mu_2 = 0.085$, $\sigma_1 = 4.912$, $\sigma_2 = 0.001$
M13 (3normal>0)	−29696.73	0.047	$p_0 = 0.550$, $p_1 = 0.055$ ($p_2 = 0.395$), $\mu_2 = 0.051$, $\sigma_0 = 0.003$, $\sigma_1 = 1.430$, $\sigma_2 = 0.000$

(beta) and M8 (beta&ω) is $2\Delta\ell = 2 \times [-29690.71 - (-29699.38)] = 18.54$. The P value for this comparison is 0.9×10^{-4}, in comparison with the χ^2 distribution with d.f. = 2. M7 is thus rejected in favor of M8. However, the selection model (M2) does not detect positive selection in this data set. This is apparently due to the fact that the strict neutral model (M1) on which it is based is unrealistic, and the extra category added in M2 optimally accounts for deleterious mutations (with $\omega_2 = 0.11$). Models M10 (beta&gamma+1) and M11 (beta&normal>1) have the same likelihood value and the estimates of parameters under those models appear to suggest presence (at ~0.5%) of sites under positive selection. However, the discretized distributions do not have any category with $\omega > 1$ and do not suggest positive selection; neither is the fit of those two models significantly better than the beta model (M7); $2\Delta\ell = 4.98$ with $P = 0.17$ with d.f. = 3. The other continuous distributions, such as M5 (gamma), M6 (2gamma), M12 (0&2normal>1), and M13 (3normal>1), do not have site classes with $\omega > 1$ either. However, this failure is due to those models' insistence on having 10% of sites in a category with $\omega > 1$. The small proportion of positively selected sites is detected only when the proportion is a free parameter estimated from the data, that is, under models M3 (discrete) and M8 (beta&ω). In sum, the models provide consistent evidence for the presence of a small proportion of positively selected sites in the mitochondrial genes.

The Bayes approach (Equation 5) can be used to identify sites potentially under diversifying selection. Model M3 (discrete) suggested 14 positively selected sites with $P(\omega > 1) > 0.5$, while model M8 (beta&ω) located only 12, all included in the 14 found by model M3. If the stricter 95% threshold is used, only 2 sites are identified: one in ATP6 (with the posterior probabilities 0.96 under M3 and 0.94 under M8) and another in ND5 (with the probabilities 0.98 under M3 and 0.96 under M8). Overall, although the two models are constructed very differently, they produced very similar results concerning the inference of the positively selected sites.

D2: β-globin genes from vertebrates (Table 4): The β-globin genes are moderately conserved, with an average ω ratio between 0.31 and 0.36 among all but the worst-fitting models. Parameter estimates from models such as M5 (gamma), M7 (beta), and M9 (beta&gamma) suggest that the distribution of ω over sites is L-shaped, with most sites highly conserved. Although better than M0, the strictly neutral model (M1) does not fit the data well. Neither did model M4 (freqs). The continuous mixture distributions (M8–M13) fit the data almost equally well.

Parameter estimates and LRTs suggest presence of sites under diversifying selection (Table 4). For example, the discrete model (M3) indicates ~7.9% of sites are under positive selection with $\omega_2 = 1.69$. M3 fits the data significantly better than any of models M0 (one-ratio), M1 (neutral), and M2 (selection). Similarly to

TABLE 4

Likelihood values and parameter estimates for vertebrate β-globin gene (D2)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−3815.51	0.237	$\omega = 0.237$
M1 (neutral)	−3805.96	0.684	$p_0 = 0.316$ ($p_1 = 0.684$)
M2 (selection)	−3691.32	0.266	$p_0 = 0.281$, $p_1 = 0.142$ ($p_2 = 0.577$) $\omega_2 = 0.215$
M3 (discrete)	−3687.06	0.304	$p_0 = 0.386$, $p_1 = 0.535$ ($p_2 = 0.079$), $\omega_0 = 0.018$, $\omega_1 = 0.304$, $\omega_2 = 1.691$
M4 (freqs)	−3693.42	0.397	$p_0 = 0.299$, $p_1 = 0.590$, $p_2 = 0.000$, $p_3 = 0.066$ ($p_4 = 0.045$)
M5 (gamma)	−3690.82	0.326	$\alpha = 0.508$, $\beta = 1.463$
M6 (2gamma)	−3685.64	0.347	$p_0 = 0.601$ ($p_0 = 0.399$), $\alpha_0 = 2.831$, $\beta_0 = 9.829$, $\alpha_1 = 0.092$
M7 (beta)	−3697.22	0.269	$p = 0.408$, $q = 1.099$
M8 (beta&ω)	−3686.13	0.312	$p_0 = 0.943$ ($p_1 = 0.057$), $p = 0.572$, $q = 2.172$, $\omega = 2.081$
M9 (beta&gamma)	−3685.65	0.347	$p_0 = 0.667$ ($p_1 = 0.333$), $p = 1.728$, $q = 4.442$, $\alpha = 0.054$, $\beta = 0.017$
M10 (beta&gamma+1)	−3686.98	0.358	$p_0 = 0.950$ ($p_1 = 0.050$), $p = 0.550$, $q = 1.878$, $\alpha = 2.702$, $\beta = 0.257$
M11 (beta&normal>1)	−3686.98	0.358	$p_0 = 0.949$ ($p_1 = 0.051$), $p = 0.551$, $q = 1.883$, $\mu = 11.899$, $\sigma = 6.236$
M12 (0&2normal>1)	−3684.86	0.326	$p_0 = 0.231$, $p_1 = 0.181$, $\mu_2 = 0.213$, $\sigma_1 = 1.171$, $\sigma_2 = 0.150$
M13 (3normal>0)	−3685.65	0.347	$p_0 = 0.219$, $p_1 = 0.160$ ($p_2 = 0.621$), $\mu_2 = 0.205$, $\sigma_0 = 0.000$, $\sigma_1 = 1.142$, $\sigma_2 = 0.159$

the case of the mitochondrial data set, the selection model (M2) does not detect positive selection, as the strict neutral model it is based on does not allow for sites with $0 < \omega < 1$. The beta distribution (M7) is rejected when compared with any of M8 (beta&ω), M9 (beta&gamma), M10 (beta&gamma+1), or M11 (beta&normal>1). The LRT statistic is $2\Delta\ell = 22.18$ between M7 and M8 (with $P = 0.15 \times 10^{-4}$, d.f. = 2) and 20.48 between M7 and M10 or M11 (with $P = 0.13 \times 10^{-3}$, d.f. = 3). The discretized versions of models M9, M10, and M11 have one category (10%) with an ω ratio of ~1.7. Except for model M2 (selection), all other models that allow for positively selected sites (M3–M6 and M8–M13) do suggest existence of such sites.

Positively selected sites are identified using the Bayes approach (Equation 5). The different models give very similar lists of positively selected sites, although the posterior probabilities vary somewhat among models. The discrete model (M3) identified more sites under positive selection than other models. At the 99% level, eight sites are identified: 7, 42, 48, 50, 54, 67, 85, and 123. Those sites are also the top eight under model M8 (beta&ω); but at the 99% level, M8 located only sites 7, 50, and 123. To test whether the choice of tree topology has any effect, we also used the star phylogeny to analyze the data under model M3 (discrete). That analysis identified the following sites at the 99% level: 7, 42, 48, 50, 54, 67, 74, 85, 114, 123, 124, and 128. This list is very similar to those obtained using the best tree.

D3: Drosophila adh gene (Table 5): The average estimate of ω is ~0.11 for all the best-fitting models. Parameter estimates from models such as M5 (gamma) and M7 (beta) suggest that the ω distribution has a highly skewed L shape, with most sites highly conserved with very small ω ratios. Model M3 (discrete) fits the data significantly better than models M0 (one-ratio), M1 (neutral), or M2 (selection), but none of the models suggest presence of positively selected sites. The beta model (M7) provides a good fit to the data, although slightly worse than M3 (discrete), and adding an extra component to the beta (as in models M8–11) leads to no significant improvement in the log-likelihood score. Other continuous mixture models (M5, M6, M8, and M12–M13) do not suggest positive selection either.

We note that previous studies (*e.g.*, Hudson *et al.* 1987) have suggested the operation of balancing selection at one particular amino acid site at the *adh* locus in Drosophila. Our LRTs, while highlighting the extreme variation in selective pressure among sites, do not suggest existence of sites under diversifying selection. This may be due to the lack of power of our models to detect balancing selection.

D4: Flavivirus E-glycoprotein gene (Table 6): The average estimate of ω is ~0.06 for all the best-fitting mod-

TABLE 5

Likelihood values and parameter estimates for Drosophila *adh* gene (D3)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−4779.73	0.094	$\omega = 0.094$
M1 (neutral)	−4819.14	0.387	$p_0 = 0.613$
M2 (selection)	−4668.55	0.136	$p_0 = 0.547$, $p_1 = 0.069$ ($p_2 = 0.384$) $\omega_2 = 0.175$
M3 (discrete)	−4662.38	0.114	$p_0 = 0.513$, $p_1 = 0.354$ ($p_2 = 0.133$), $\omega_0 = 0.000$, $\omega_1 = 0.116$, $\omega_2 = 0.547$
M4 (freqs)	−4683.07	0.152	$p_0 = 0.594$, $p_1 = 0.381$, $p_2 = 0.000$, $p_3 = 0.025$ ($p_4 = 0.000$)
M5 (gamma)	−4663.78	0.113	$\alpha = 0.271$, $\beta = 2.163$
M6 (2gamma)	−4662.85	0.111	$p_0 = 0.402$ ($p_0 = 0.598$), $\alpha_0 = 5.006$, $\beta_0 = 34.237$, $\alpha_1 = 0.024$
M7 (beta)	−4663.82	0.114	$p = 0.225$, $q = 1.685$
M8 (beta&ω)	−4663.82	0.114	$p_0 = 1.000$ ($p_1 = 0.000$), $p = 0.225$, $q = 1.686$, $\omega = 0.651$
M9 (beta&gamma)	−4663.54	0.114	$p_0 = 0.295$ ($p_1 = 0.705$), $p = 0.008$, $q = 4.994$, $\alpha = 0.496$, $\beta = 2.806$
M10 (beta&gamma+1)	−4663.72	0.113	$p_0 = 0.974$ ($p_1 = 0.026$), $p = 0.251$, $q = 2.290$, $\alpha = 1.042$, $\beta = 0.690$
M11 (beta&normal>1)	−4663.72	0.113	$p_0 = 0.974$ ($p_1 = 0.026$), $p = 0.251$, $q = 2.290$, $\mu = 1.631$, $\sigma = 0.597$
M12 (0&2normal>1)	−4664.03	0.111	$p_0 = 0.455$, $p_1 = 0.065$, $\mu_2 = 0.000$, $\sigma_1 = 0.470$, $\sigma_2 = 0.205$
M13 (3normal>0)	−4662.86	0.111	$p_0 = 0.456$, $p_1 = 0.054$ ($p_2 = 0.500$), $\mu_2 = 0.120$, $\sigma_0 = 0.000$, $\sigma_1 = 0.266$, $\sigma_2 = 0.080$

TABLE 6

Likelihood values and parameter estimates for flavivirus E-glycoprotein gene (D4)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−9885.19	0.052	$\omega = 0.052$
M1 (neutral)	−10417.47	0.549	$p_0 = 0.451$ ($p_1 = 0.549$)
M2 (selection)	−9775.62	0.063	$p_0 = 0.383$, $p_1 = 0.013$ ($p_2 = 0.604$), $\omega_2 = 0.603$
M3 (discrete)	−9757.16	0.062	$p_0 = 0.563$, $p_1 = 0.327$ ($p_2 = 0.109$), $\omega_0 = 0.010$, $\omega_1 = 0.084$, $\omega_2 = 0.247$
M4 (freqs)	−9935.94	0.189	$p_0 = 0.434$, $p_1 = 0.566$, $p_2 = 0.0$, $p_3 = 0.0$, $p_4 = 0.0$
M5 (gamma)	−9755.06	0.058	$\alpha = 0.490$, $\beta = 7.885$
M6 (2gamma)	−9755.06	0.058	$p_0 = 1.000$ ($p_0 = 0.000$), $\alpha_0 = 0.491$, $\beta_0 = 7.890$, $\alpha_1 = 2.007$
M7 (beta)	−9755.06	0.058	$p = 0.463$, $q = 7.110$
M8 (beta&ω)	−9755.05	0.058	$p_0 = 0.990$ ($p_1 = 0.010$), $p = 0.475$, $q = 7.667$, $\omega = 0.336$
M9 (beta&gamma)	−9755.04	0.058	$p_0 = 0.986$ ($p_1 = 0.014$), $p = 0.479$, $q = 7.883$, $\alpha = 0.865$, $\beta = 0.272$
M10 (beta&gamma+1)	−9755.04	0.058	$p_0 = 0.988$ ($p_1 = 0.012$), $p = 0.479$, $q = 7.864$, $\alpha = 0.818$, $\beta = 0.683$
M11 (beta&normal>1)	−9755.04	0.058	$p_0 = 0.988$ ($p_1 = 0.012$), $p = 0.479$, $q = 7.861$, $\mu = 2.190$, $\sigma = 0.950$
M12 (0&2normal>1)	−9754.90	0.058	$p_0 = 0.239$, $p_1 = 0.053$, $\mu_2 = 0.000$, $\sigma_1 = 0.495$, $\sigma_2 = 0.075$
M13 (3normal>0)	−9754.68	0.058	$p_0 = 0.232$, $p_1 = 0.052$ ($p_2 = 0.716$), $\mu_2 = 0.000$, $\sigma_0 = 0.000$, $\sigma_1 = 4.754$, $\sigma_2 = 0.070$

els. The ω ratios are highly variable among sites and the ω distribution appears to have a highly skewed L shape, with most sites highly conserved with very small ω ratios. The discrete model (M3) fits the data much better than M0 (one-ratio), M1 (neutral), or M2 (selection), but none of the models suggest the existence of sites under diversifying selection. The beta distribution (M7) fits the data better than M3 (discrete). Since M7 also has three fewer parameters than M3, it is the preferred model for this data set. Although models M12 and M13 have marginally higher likelihood values than M7, this improvement is not more than expected given their greater numbers of parameters. Adding an additional component to the beta model to allow for positively selected sites (models M8–M11) provides no significant improvement to the model's fit to data.

D5: Human influenza A virus HA gene HA1 domain (Table 7): The average ω ranges from 0.39 to 0.41 among all models except for M1 (neutral) and M7 (beta), which do not allow for positively selected sites, fit the data badly, and give smaller estimates of ω. The average acceptance rate is <1, indicating that, on average, purifying selection dominates the evolution of the gene. The one-ratio model (M0) is easily rejected when compared with all other models, which allow the ω ratio to vary among sites. Distributions of ω among sites estimated under M5 (gamma) and M6 (2gamma) are L-shaped, with heavy tails. The beta distribution (M7) has an interesting U shape, possibly because the model is forced to attempt to account for sites with $\omega > 1$. The discrete model (M3) fits the data as well as or better than all other models considered; M8 has nearly as high a likelihood value, and uses one fewer parameter.

Models that allow for positively selected sites, that is, M2 (selection), M3 (discrete), M5 (gamma), M6 (2gamma), and M8–M13, all suggest presence of positively selected sites. For example, the selection model (M2) suggests ~1.1% of sites under positive selection with $\omega_2 = 5.8$. Model M3 (discrete) suggests a large proportion of sites (25.1%) under weak diversifying selection with $\omega_1 = 1.28$ and a small proportion of sites (0.8%) under strong diversifying selection with $\omega_2 = 6.90$. Note that the selection model (M2) suggests a large proportion of neutral sites with $\omega_1 = 1$, and the estimates from M2 and M3 are not very different. Both models have significantly higher likelihood values than models M0 and M1, providing strong evidence for adaptive evolution. Similarly, M8 (beta&ω) suggests ~1.3% of sites under positive selection with $\omega_2 = 5.2$. The LRT statistic for comparing M7 (beta) and M8 (beta&ω) is $2\Delta\ell = 2 \times 5.43 = 10.86$, $> \chi^2_{1\%} = 9.21$ with d.f. = 2. Models M10 (beta&gamma+1) and M11 (beta&normal>1) have the same likelihood value, and both models fit the data significantly better than the beta model

TABLE 7

Likelihood values and parameter estimates for human influenza virus A HA gene (D5)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−3125.61	0.391	$\omega = 0.391$
M1 (neutral)	−3083.58	0.342	$p_0 = 0.658$
M2 (selection)	−3078.20	0.401	$p_0 = 0.652$, $p_1 = 0.337$ ($p_2 = 0.011$)
			$\omega_2 = 5.815$
M3 (discrete)	−3077.73	0.412	$p_0 = 0.741$, $p_1 = 0.251$ ($p_2 = 0.008$),
			$\omega_0 = 0.049$, $\omega_1 = 1.284$, $\omega_2 = 6.898$
M4 (freqs)	−3078.66	0.396	$p_0 = 0.593$, $p_1 = 0.134$, $p_2 = 0.000$, $p_3 = 0.234$ ($p_4 = 0.039$)
M5 (gamma)	−3079.32	0.408	$\alpha = 0.238$, $\beta = 0.516$
M6 (2gamma)	−3079.29	0.409	$p_0 = 0.617$ ($p_1 = 0.383$)
			$\alpha_0 = 0.130$, $\beta_0 = 1.142$, $\alpha_1 = \beta_1 = 0.716$
M7 (beta)	−3083.63	0.317	$p = 0.011$, $q = 0.021$
M8 (beta&ω)	−3078.20	0.383	$p_0 = 0.987$ ($p_1 = 0.013$),
			$p = 0.012$, $q = 0.024$, $\omega = 5.141$
M9 (beta&gamma)	−3079.25	0.410	$p_0 = 0.934$ ($p_1 = 0.066$),
			$p = 0.133$, $q = 0.422$, $\alpha = 0.841$, $\beta = 0.085$
M10 (beta&gamma+1)	−3079.25	0.411	$p_0 = 0.925$ ($p_1 = 0.075$),
			$p = 0.139$, $q = 0.472$, $\alpha = 0.258$, $\beta = 0.008$
M11 (beta&normal>1)	−3079.25	0.411	$p_0 = 0.924$ ($p_1 = 0.076$),
			$p = 0.139$, $q = 0.473$, $\mu = 1.541$, $\sigma = 2.488$
M12 (0&2normal>1)	−3078.59	0.410	$p_0 = 0.924$, $p_1 = 0.077$,
			$\mu_2 = 0.877$, $\sigma_1 = 6.792$, $\sigma_2 = 0.180$
M13 (3normal>0)	−3079.17	0.409	$p_0 = 0.551$, $p_1 = 0.104$ ($p_2 = 0.345$),
			$\mu_2 = 0.602$,
			$\sigma_0 = 0.001$, $\sigma_1 = 2.613$, $\sigma_2 = 0.000$

(M7); the test statistic is $2\Delta\ell = 2 \times [-3079.25 - (-3083.63)] = 8.76$, with $P = 0.033$ with d.f. = 3. In addition to models M9–M11, which have components specifically designed to allow for positively selected sites, models M5 (gamma) and M6 (2gamma) also have categories with $\omega > 1$ when discretized.

Amino acid sites 135 and 226 are identified to be under positive selection at the 95% level by all models that allow for positive selection. Model M3 (discrete) suggested many more sites because the parameter estimates under this model suggest a large proportion of weakly selected sites. At the 50% level, models M5, M6, and M9–M13 suggested the same 23 positively selected sites: 15, 94, 121, 124, 133, 135, 137, 138, 155, 156, 157, 159, 163, 174, 186, 189, 196, 197, 219, 226, 246, 262, and 273.

D6: HIV-1 *vif* gene (Table 8): The pattern for this data set is rather similar to that of the influenza virus HA gene (D5; see Table 7). The average ω ratio over all sites ranges from 0.6 to 0.9 among models. The one-ratio model (M0) is easily rejected when compared with any of the more-general models that allow for variable ω ratios among sites. Model M1 (neutral) also fits the data set much better than M0. The gamma (M5) and double gamma (M6) models suggested L-shaped distributions for ω, with heavy tails. The beta model (M7) suggests a U shape, possibly because the underlying ω ratio at some sites is >1. Model M3 (discrete) fits the data as well as or better than all other models considered.

All models that allow for positively selected sites do suggest existence of such sites in this gene. For example, the selection model (M2) suggests ~8.5% of sites under positive selection with $\omega_2 = 4.2$. Model M3 (discrete) suggests a large proportion of sites (32.5%) under weak diversifying selection with $\omega_1 = 1.21$ and a small proportion of sites (7%) under strong selective pressure with $\omega_2 = 4.0$. Both models M2 and M3 have significantly higher likelihood values than models M0 (one-ratio) or M1 (neutral). Similarly, M8 (beta&ω) suggests that ~9% of sites are under positive selection with $\omega = 3.4$. The LRT statistic for comparing M7 (beta) and M8 (beta&ω) is $2\Delta\ell = 2 \times [(-3370.66) - (-3400.45)] = 59.58$, $\gg \chi^2_{1\%} = 9.21$ with d.f. = 2. LRTs comparing the beta model (M7) with any of models M9 (beta&gamma), M10 (beta&gamma+1), and M11 (beta&normal>1) give similar results. In sum, all the models provide consistent and statistically significant evidence for the existence of positively selected sites in this gene.

We plotted the posterior probability distributions, calculated using Equation 5, for sites along the HIV-1 *vif* gene. This was done for models M2 (selection), M3 (discrete), M8 (beta&ω), and M9 (beta&gamma), with results for M3 and M9 shown in Figure 2. Posterior probabilities under M2 and M3 are very similar, but as one may expect, many sites included in the neutral

TABLE 8

Likelihood values and parameter estimates for HIV *vif* gene (D6)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−3499.60	0.644	$\omega = 0.644$
M1 (neutral)	−3413.07	0.575	$p_0 = 0.425$ ($p_1 = 0.575$)
M2 (selection)	−3377.94	0.870	$p_0 = 0.404$, $p_1 = 0.511$ ($p_2 = 0.085$) $\omega_2 = 4.220$
M3 (discrete)	−3367.16	0.742	$p_0 = 0.604$, $p_1 = 0.325$ ($p_2 = 0.070$), $\omega_0 = 0.108$, $\omega_1 = 1.211$, $\omega_2 = 4.024$
M4 (freqs)	−3370.93	0.672	$p_0 = 0.317$, $p_1 = 0.323$, $p_2 = 0.000$, $p_3 = 0.259$ ($p_4 = 0.102$)
M5 (gamma)	−3369.77	0.774	$\alpha = 0.423$, $\beta = 0.507$
M6 (2gamma)	−3369.56	0.775	$p_0 = 0.383$ ($p_1 = 0.617$) $\alpha_0 = 0.967$, $\beta_0 = 1.452$, $\alpha_1 = \beta_1 = 0.283$
M7 (beta)	−3400.45	0.440	$p = 0.176$, $q = 0.223$
M8 (beta&ω)	−3370.66	0.687	$p_0 = 0.909$ ($p_1 = 0.091$), $p = 0.222$, $q = 0.312$, $\omega = 3.385$
M9 (beta&gamma)	−3369.42	0.766	$p_0 = 0.248$ ($p_1 = 0.752$), $p = 0.336$, $q = 0.270$, $\alpha = 0.336$, $\beta = 0.358$
M10 (beta&gamma+1)	−3368.48	0.787	$p_0 = 0.650$, $p = 0.635$, $q = 3.079$, $\alpha = 0.258$, $\beta = 0.211$
M11 (beta&normal>1)	−3369.65	0.760	$p_0 = 0.818$ ($p_1 = 0.182$) $p = 0.302$, $q = 0.591$, $\mu = 0.008$, $\sigma = 2.745$
M12 (0&2normal>1)	−3369.53	0.755	$p_0 = 0.256$, $p_1 = 0.205$, $\mu_2 = 0.000$, $\sigma_1 = 2.911$, $\sigma_2 = 0.789$
M13 (3normal>0)	−3367.69	0.736	$p_0 = 0.583$, $p_1 = 0.086$ ($p_2 = 0.331$), $\mu_2 = 1.145$, $\sigma_0 = 0.140$, $\sigma_1 = 4.407$, $\sigma_2 = 0.313$

class under M2 are included in the weakly selected class under M3 (results not shown). Plots of M8 and M9 are virtually identical (results not shown). At the 99% level, model M9 identified 10 positively selected sites (31, 33, 39, 63, 92, 101, 109, 122, 127, and 167). At the same level, model M2 identified 5 of the sites only: 31, 39, 122, 127, and 167, while sites 33, 63, 92, 101, and 109 are included at the 95% level. The two models used in Figure 2, M3 and M9, are constructed very differently, and yet the posterior distributions under the two models are highly similar.

D7: HIV-1 *pol* genes (Table 9): The average ω ratio over all sites ranges from 0.20 to 0.27 among models except for M1 (neutral) and M7 (beta), indicating relatively strong purifying selection. The strict neutral model (M1) and the beta model (M7) give lower estimates of the average ω ratio, as they do not allow for sites with $\omega > 1$. M1 fits the data much better than the one-ratio model (M0), but is much worse than M7 (beta). There are clearly sites with $0 < \omega < 1$. Parameter estimates under models such as M5 (gamma) and M6 (2gamma) suggest L-shaped distributions for ω, with heavy tails. The beta model (M7) suggests a U shape, and the peak at $\omega = 1$ is probably caused by sites with $\omega > 1$ in the data. The discrete model (M3) fits the data as well as or better than all other models considered. Simpler nested models (M0, M1, and M2) are all rejected when compared with M3.

The selection model (M2) does not suggest presence of positively selected sites. Similar to the cases of the mitochondrial (D1) and β-globin (D2) genes, this failure appears to be due to the unrealistic nature of the neutral null model (M1), which does not account for sites with $0 < \omega < 1$. Model M3 (discrete) suggests that a small proportion of sites (1.9%) are under strong diversifying selection with $\omega_2 = 4.7$. This model fits the data significantly better than models M0, M1, and M2. For example, the LRT statistic for the comparison between M0 and M3 is $2\Delta\ell = 2 \times [(-9363.57) - (-9619.30)] = 2 \times 255.73 = 511.46$, $\gg \chi^2_{1\%} = 13.27$ with d.f. = 4. Similarly, model M8 (beta&ω) suggests that ~2.5% of sites are under strong positive selection with $\omega = 4.1$. The LRT statistic for comparing M7 (beta) and M8 (beta&ω) is $2\Delta\ell = 2 \times [(-9365.88) - (-9405.74)] = 2 \times 39.86 = 79.72$, $\gg \chi^2_{1\%} = 9.21$ with d.f. = 2. Models M9–M11 all have the same log-likelihood value, substantially lower than that for M8 (beta&ω). Nevertheless, those models are significantly better than M7 (beta). The discretized versions of models M9–M11 all have a category of sites (10%) with $\omega \cong 1.8$. Apart from M2, which fits the data relatively poorly, all the models that allow for positively selected sites gave significant evidence for the existence of such sites. These results provide strong support for adaptive evolution in the HIV-1 *pol* gene.

All models that allow for positive selection (M3–M6 and M8–M13) pinpoint similar sets of amino acid sites as targets of diversifying selection. For example, at the

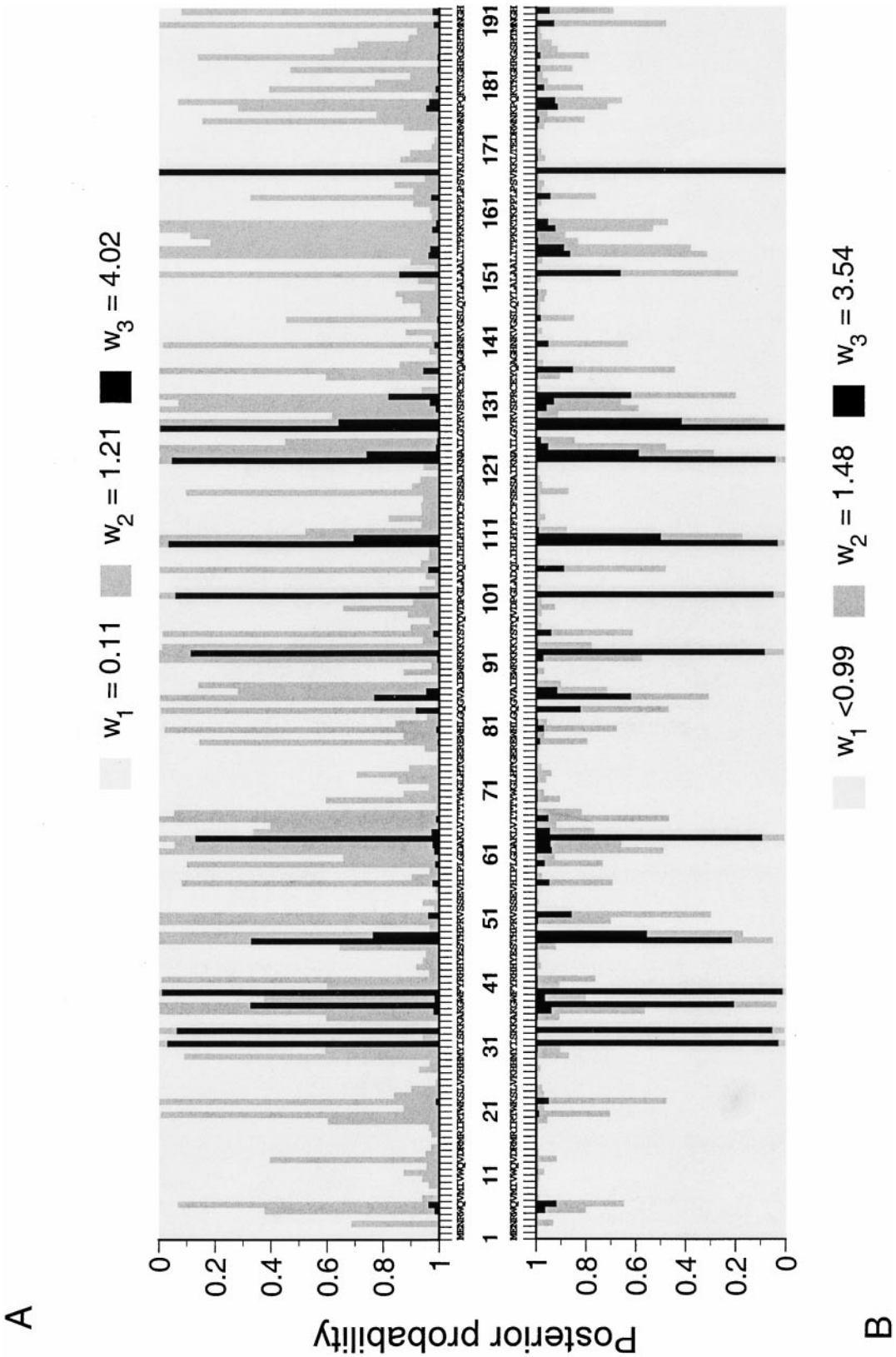

Figure 2.—Posterior probabilities of site classes along the HIV-1 *vif* gene (data set D6). (A) Model M3 (discrete) is used, which assumes three classes of sites in the gene. The estimated frequencies and ω ratios for the three classes are $p_0 = 0.604$, $p_1 = 0.325$, and $p_2 = 0.070$ and $\omega_0 = 0.108$, $\omega_1 = 1.211$, and $\omega_2 = 4.024$ (see Table 8). Those p_k are the prior distribution for each codon (amino acid) site. The data (codon configurations in different sequences) at the site change that distribution into the posterior distribution, which may be very different from the prior. For example, the posterior probabilities for the three classes at site 1 are 0.9928, 0.0072, and 0.0000, and the site is under strong purifying selection. In contrast, the posterior probabilities at site 31 are 0.0000, 0.0301, and 0.9699, and the site is almost certainly under diversifying selection. (B) Model M9 (beta&gamma) is fitted to the data with 10 categories used to approximate the continuous mixture distribution. Estimates of parameters under the model are shown in Table 8. According to those estimates, the ω ratios for the 10 categories, each of probability 10%, are 0.00036, 0.00945, 0.04338, 0.11910, 0.25502, 0.46947, 0.76134, 0.98969, 1.47748, and 3.53812. The first 8 categories are combined in the plot. The amino acid sequence of one of the variants (SF2) is superimposed on the graph.

TABLE 9

Likelihood values and parameter estimates for HIV *pol* gene (D7)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−9619.30	0.196	$\omega = 0.196$
M1 (neutral)	−9450.21	0.280	$p_0 = 0.719$ ($p_1 = 0.280$)
M2 (selection)	−9393.56	0.184	$p_0 = 0.354$, $p_1 = 0.136$ ($p_2 = 0.510$) $\omega_2 = 0.093$
M3 (discrete)	−9363.57	0.253	$p_0 = 0.838$, $p_1 = 0.142$ ($p_2 = 0.019$), $\omega_0 = 0.049$, $\omega_1 = 0.849$, $\omega_2 = 4.739$
M4 (freqs)	−9369.10	0.240	$p_0 = 0.616$, $p_1 = 0.318$, $p_2 = 0.000$, $p_3 = 0.032$ ($p_4 = 0.034$)
M5 (gamma)	−9384.32	0.264	$\alpha = 0.169$, $\beta = 0.541$
M6 (2gamma)	−9378.79	0.269	$p_0 = 0.422$ ($p_1 = 0.578$) $\alpha_0 = 2.108$, $\beta_0 = 10.202$, $\alpha_1 = \beta_1 = 0.045$
M7 (beta)	−9405.74	0.188	$p = 0.111$, $q = 0.482$
M8 (beta&ω)	−9365.88	0.250	$p_0 = 0.975$ ($p_1 = 0.025$), $p = 0.177$, $q = 0.963$, $\omega = 4.052$
M9 (beta&gamma)	−9378.79	0.269	$p_0 = 0.919$ ($p_1 = 0.081$), $p = 0.314$, $q = 2.313$, $\alpha = 0.286$, $\beta = 0.013$
M10 (beta&gamma+1)	−9378.79	0.269	$p_0 = 0.947$ ($p_1 = 0.053$), $p = 0.311$, $q = 2.216$, $\alpha = 1.116$, $\beta = 0.090$
M11 (beta&normal>1)	−9378.79	0.269	$p_0 = 0.946$ ($p_1 = 0.054$) $p = 0.312$, $q = 2.239$, $\mu = 3.571$, $\sigma = 1.360$
M12 (0&2normal>1)	−9368.00	0.270	$p_0 = 0.590$, $p_1 = 0.203$, $\mu_2 = 0.273$, $\sigma_1 = 2.389$, $\sigma_2 = 0.003$
M13 (3normal>0)	−9361.37	0.269	$p_0 = 0.726$, $p_1 = 0.033$ ($p_2 = 0.241$), $\mu_2 = 0.000$, $\sigma_0 = 0.046$, $\sigma_1 = 4.625$, $\sigma_2 = 0.652$

99% level, M8 (beta&ω) identified the following sites: 67, 347, 478, 568, 761, and 779, while at the 95% level, four additional sites are identified by that model (sites 3, 14, 41, and 379).

Seibert *et al.* (1995) examined the HIV-1 *env*, *gag*, and *pol* genes for positive selection. They used the method of Nei and Gojobori (1986) to estimate d_N and d_S and concluded that $d_N > d_S$ for the *env* gene but $d_N < d_S$ in the *gag* and the *pol* genes. Our results and those of Seibert *et al.* (1995) may be reconciled by noting that the tests presented in this article are more powerful to detect positive selection than the method used by Seibert *et al.* (1995).

D8: Japanese encephalitis *env* gene (Table 10): This gene is under strong purifying selection, as the average ω ratio over sites is ∼0.05 for all models except for M1 (neutral) and M4 (freqs). Models M1 and M4 gave much larger estimates (0.17 and 0.07, respectively), but these do not appear reliable because the two models do not fit the data well. M1 is even poorer than the one-ratio model (M0). The discrete model (M3) suggests a small proportion (0.3%) of nearly neutral sites with $\omega_2 = 1.04$, and there is no evidence for positive selection. Models such as M5 (gamma), M6 (2gamma), and M7 (beta) suggest highly skewed L-shaped distributions for ω, with most sites under strong selective pressure (with ω close to 0) and with very little probability density for the region $\omega > 1$. The discrete model (M3) fits the data as well as or better than all other models considered. Model 8 (beta&ω) fits the data better (although not significantly) than M7 (beta). However, the estimated ω for the additional category of sites is <1. Furthermore, the discretized versions of models M9–M13 do not have any category with $\omega > 1$.

D9: Tick-borne flavivirus NS-5 gene (Table 11): This gene is the most conserved among the 10 genes analyzed in this article. The average ω ratio over all sites ranges from 0.02 to 0.03 among models, except for M1 (neutral) and M4 (freqs). The latter two models produced larger but unreliable estimates as they fit the data very poorly. The strict neutral model (M1) fits the data even more poorly than the one-ratio model (M0). Although the two models have the same number of parameters and are not nested, the log-likelihood difference ($\Delta\ell = 582.48$) is huge. Neither the selection (M2) nor discrete (M3) model indicates presence of positively selected sites. The beta mixture models (M8–M11) do not fit the data any better than the simple beta model (M7). Furthermore, none of the 10 categories in the discrete distributions used to approximate the continuous beta mixture models (M9–M11) have an ω ratio >1. Models such as M5 (gamma), M6 (2gamma), and beta (M7) all suggest highly skewed L-shaped distributions for ω. The discrete model (M3) fits the data as well as the best of other models.

D10: HIV-1 *env* gene V3 region (Table 12): The ω ratio averaged over all sites ranges from 0.9 to 1.2 among models, except for models M1 (neutral) and M7 (beta).

TABLE 10

Likelihood values and parameter estimates for Japanese encephalitis *env* gene (D8)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−6886.17	0.051	$\omega = 0.051$
M1 (neutral)	−6974.78	0.168	$p_0 = 0.832$ ($p_1 = 0.168$)
M2 (selection)	−6848.06	0.055	$p_0 = 0.658$, $p_1 = 0.014$ ($p_2 = 0.329$) $\omega_2 = 0.124$
M3 (discrete)	−6845.30	0.053	$p_0 = 0.921$, $p_1 = 0.075$ ($p_2 = 0.003$), $\omega_0 = 0.022$, $\omega_1 = 0.382$, $\omega_2 = 1.043$
M4 (freqs)	−6862.14	0.070	$p_0 = 0.789$, $p_1 = 0.211$, $p_2 = 0.000$, $p_3 = 0.000$ ($p_4 = 0.000$)
M5 (gamma)	−6848.00	0.054	$\alpha = 0.198$, $\beta = 3.198$
M6 (2gamma)	−6845.91	0.054	$p_0 = 0.940$ ($p_1 = 0.060$) $\alpha_0 = 20.464$, $\beta_0 = \infty$, $\alpha_1 = \beta_1 = 2.019$
M7 (beta)	−6848.20	0.053	$p = 0.178$, $q = 2.883$
M8 (beta&ω)	−6845.38	0.052	$p_0 = 0.930$ ($p_1 = 0.070$), $p = 9.666$, $q = 400.3(\infty)$, $\omega = 0.438$
M9 (beta&gamma)	−6846.65	0.054	$p_0 = 0.950$ ($p_1 = 0.050$), $p = 0.515$, $q = 17.574$, $\alpha = 30.902$, $\beta = 0.008$
M10 (beta&gamma+1)	−6846.66	0.055	$p_0 = 0.950$ ($p_1 = 0.050$), $p = 0.521$, $q = 17.414$, $\alpha = 0.097$, $\beta = 1.100$
M11 (beta&normal>1)	−6846.67	0.055	$p_0 = 0.950$ ($p_1 = 0.050$) $p = 0.522$, $q = 17.414$, $\mu = 0.457$, $\sigma = 1.097$
M12 (0&2normal>1)	−6845.45	0.053	$p_0 = 0.334$, $p_1 = 0.054$, $\mu_2 = 0.039$, $\sigma_1 = 3.762$, $\sigma_2 = 0.000$
M13 (3normal>0)	−6848.54	0.055	$p_0 = 0.915$, $p_1 = 0.000$ ($p_2 = 0.085$), $\mu_2 = 0.492$ $\sigma_0 = 0.020$, $\sigma_1 = 29.0(\infty)$, $\sigma_2 = 0.109$

TABLE 11

Likelihood values and parameter estimates for tick-borne flavivirus NS-5 gene (D9)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−9554.63	0.025	$\omega = 0.\ 025$
M1 (neutral)	−10137.11	0.412	$p_0 = 0.588$ ($p_1 = 0.412$)
M2 (selection)	−9238.49	0.051	$p_0 = 0.571$, $p_1 = 0.028$ ($p_2 = 0.401$), $\omega_2 = 0.058$
M3 (discrete)	−9187.91	0.031	$p_0 = 0.645$, $p_1 = 0.259$ ($p_2 = 0.096$), $\omega_0 = 0.002$, $\omega_1 = 0.045$, $\omega_2 = 0.188$
M4 (freqs)	−9348.87	0.145	$p_0 = 0.587$, $p_1 = 0.402$, $p_2 = 0.00$, $p_3 = 0.011$ ($p_4 = 0.0$)
M5 (gamma)	−9188.67	0.033	$\alpha = 0.212$, $\beta = 5.596$
M6 (2gamma)	−9187.93	0.033	$p_0 = 0.499$ ($p_0 = 0.501$), $\alpha_0 = 0.847$, $\beta_0 = 21.108$, $\alpha_1 = 0.018$
M7 (beta)	−9188.80	0.033	$p = 0.203$, $q = 5.345$
M8 (beta&ω)	−9188.83	0.037	$p_0 = 1.000$ ($p_1 = 0.000$), $p = 0.203$, $q = 5.345$, $\omega = 40.436$
M9 (beta&gamma)	−9188.58	0.033	$p_0 = 0.980$ ($p_1 = 0.020$), $p = 0.220$, $q = 7.078$, $\alpha = 2.378$, $\beta = 0.006$
M10 (beta&gamma+1)	−9188.58	0.033	$p_0 = 0.980$ ($p_1 = 0.020$), $p = 0.220$, $q = 7.078$, $\alpha = 0.008$, $\beta = 1.009$
M11 (beta&normal>1)	−9188.58	0.033	$p_0 = 0.980$ ($p_1 = 0.020$), $p = 0.220$, $q = 7.078$, $\mu = 0.483$, $\sigma = 1.181$
M12 (0&2normal>1)	−9188.45	0.030	$p_0 = 0.489$, $p_1 = 0.050$, $\mu_2 = 0.000$, $\sigma_1 = 22.528$, $\sigma_2 = 0.054$
M13 (3normal>0)	−9187.76	0.032	$p_0 = 0.650$, $p_1 = 0.052$ ($p_2 = 0.298$), $\mu_2 = 0.050$, $\sigma_0 = 0.002$, $\sigma_1 = 4.062$, $\sigma_2 = 0.000$

TABLE 12

Likelihood values and parameter estimates for HIV-1 *env* gene V3 region (D10)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−1137.69	0.901	$\omega = 0.901$
M1 (neutral)	−1116.33	0.644	$p_0 = 0.356$
M2 (selection)	−1106.59	1.212	$p_0 = 0.316$, $p_1 = 0.502$ ($p_2 = 0.182$), $\omega = 3.898$
M3 (discrete)	−1105.49	1.173	$p_0 = 0.531$, $p_1 = 0.423$ ($p_2 = 0.046$), $\omega_0 = 0.175$, $\omega_1 = 1.781$, $\omega_2 = 7.141$
M4 (freqs)	−1106.43	1.069	$p_0 = 0.277$, $p_1 = 0.000$, $p_2 = 0.462$, $p_3 = 0.011$, ($p_4 = 0.249$)
M5 (gamma)	−1105.97	1.175	$\alpha = 0.557$, $\beta = 0.448$
M6 (2gamma)	−1105.90	1.169	$p_0 = 0.763$ ($p_1 = 0.237$), $\alpha_0 = 0.883$, $\beta_0 = 0.643$, $\alpha_1 = 0.101$
M7 (beta)	−1115.40	0.555	$p = 0.148$, $q = 0.118$
M8 (beta&ω)	−1106.39	1.119	$p_0 = 0.799$ ($p_1 = 0.201$), $p = 0.167$, $q = 0.149$, $\omega = 3.470$
M9 (beta&gamma)	−1105.89	1.175	$p_0 = 0.258$ ($p_1 = 0.742$), $p = 0.005$, $q = 0.007$, $\alpha = 0.702$, $\beta = 0.459$
M10 (beta&gamma+1)	−1105.88	1.176	$p_0 = 0.523$ ($p_1 = 0.477$), $p = 0.364$, $q = 1.150$, $\alpha = 0.458$, $\beta = 0.329$
M11 (beta&normal>1)	−1105.96	1.165	$p_0 = 0.694$ ($p_1 = 0.304$), $p = 0.252$, $q = 0.337$, $\mu = 0.005$, $\sigma = 2.854$
M12 (0&2normal>1)	−1105.61	1.200	$p_0 = 0.227$, $p_1 = 0.59$, $\mu_2 = 0.392$, $\sigma_1 = 29$ (∞), $\sigma_2 = 1.328$
M13 (3normal>0)	−1105.81	1.138	$p_0 = 0.611$, $p_1 = 0.194$ ($p_2 = 0.195$), $\mu_2 = 0.000$, $\sigma_0 = 1.136$, $\sigma_1 = 3.625$, $\sigma_2 = 0.000$

The latter two models give lower and unreliable estimates as they do not account for positively selected sites. This gene region has the highest overall ω ratios among the 10 data sets analyzed in this article. The HIV-1 *env* gene and in particular the V3 region are well known to be under diversifying selection. The strict neutral model (M1) fits the data much better than the model of one ω for all sites (M0). The discrete model (M3) fits the data about as well as the best of other models considered.

All models that allow for positive selection do suggest a substantial proportion (18–70%) of positively selected sites. Models such as M5 (gamma) and M6 (2gamma) suggest L-shaped distributions for ω with heavy tails. The beta model (M7) suggests a U shape, as the model uses the density at $\omega \sim 1$ to account for sites with $\omega > 1$. Many amino acids are clearly under diversifying selection. For example, the discrete model (M3) suggests ~42% of sites are under relatively weak positive selection with $\omega_1 = 1.8$ and a further 4.6% of sites are under strong positive selection with $\omega_2 = 7.1$. Similarly, M8 (beta&ω) suggests that ~20% of sites are under positive selection with $\omega_1 = 3.5$. Note that the differences between the models may not be as large as they may seem, as we expect it to be difficult to distinguish a small proportion of strongly selected sites from a large proportion of weakly selected sites. The beta distribution in M8 (beta&ω) is U-shaped with a high proportion of sites with $\omega \sim 1$. The LRT statistic for comparing M7 (beta) and M8 (beta&ω) is $2\Delta\ell = 2 \times [(-1106.39) - (-1115.40)] = 2 \times 9.01 = 18.02$, with $P = 0.00012$ compared with the χ^2 distribution with d.f. = 2. The beta mixture models (M9–M11) fit the data slightly better than M8 (beta&ω), and significantly better than the beta model (M7), again suggesting the operation of diversifying selection at some sites.

All models that allow for sites under positive selection identified sites 28, 66, and 87 with high posterior probability supports (≥99%). Model M12 (0&2normal>1) included sites 26 and 51 as well at the 99% level. At the 95% level, M12 suggested six additional sites: 22, 24, 68, 69, 76, and 83.

DISCUSSION

Effects of tree topology: Maximum-likelihood estimation under models in this article relies on the phylogenetic relationship among the sequences. To examine the effect of the phylogeny on the analysis, we used six candidate trees to fit all 14 models (M0–M13) to the vertebrate β-globin genes (D2). The six trees were either inferred from the β-globin data or based on conventional wisdom. The best tree according to the one-ratio model (M0) was used to obtain results of Table 4. Results under another tree (the worst of the six trees under model M0) are presented for four models (M0, M3, M7,

TABLE 13

Parameter estimates under an alternative tree for the β-globin gene (D2)

Model code	ℓ	d_N/d_S	Estimates of parameters
M0 (one-ratio)	−3820.81	0.237	$\omega = 0.237$
M3 (discrete)	−3688.36	0.305	$p_0 = 0.391$, $p_1 = 0.533$ ($p_2 = 0.076$), $\omega_0 = 0.019$, $\omega_1 = 0.309$, $\omega_2 = 1.752$
M7 (beta)	−3698.74	0.269	$p = 0.404$, $q = 1.090$
M8 (beta&ω)	−3687.02	0.312	$p_0 = 0.944$ ($p_1 = 0.056$), $p = 0.567$, $q = 2.152$, $\omega = 2.112$

and M8) in Table 13. The estimates under this tree are highly similar to estimates presented in Table 4. LRTs of positive selection lead to the same conclusions no matter which of the two (or six) trees is used. The inference of sites under positive selection does not seem to be sensitive to the assumed tree topology either. As mentioned before, use of even the star tree generated lists of positively selected sites for the vertebrate β-globin genes that are very similar to those obtained under the best tree. A similar analysis was performed on the HIV *vif* data set using two candidate trees. Parameter estimates, LRTs, and posterior probability calculations are all highly similar between the candidate trees (results not shown). While the correct tree should obviously be used if it is known, those results suggest that a reasonably good phylogeny may be sufficient for estimating parameters in the ω distribution and for performing LRTs of positive selection.

Computational and theoretical problems: Models implemented in this article appear very useful in testing the existence of positively selected amino acid sites and in identifying such sites when they exist. The different models also appear to produce consistent and convincing results. However, we encountered numerous practical problems. The continuous mixture models M9–M13, and especially the parameter-rich model M13 (3normal>0), were found to converge to ML estimates very slowly during the iteration. In some data-model combinations, the likelihood value was also found to be sensitive to the number of categories (K) used in the discrete approximation. While the likelihood values reported in Tables 3–12 are reliable, parameter estimates under some models may not be. In some cases, different values of the parameters gave virtually the same likelihood values, and the likelihood surface appears to be nearly flat. Those computational difficulties appear to a large extent to be caused by our use of the discrete distributions to approximate the continuous ones. Two different continuous distributions (or two sets of parameters for the same continuous distribution) may look very similar after they are discretized. We note that distinguishing statistical distributions is almost always a difficult task, but the discretization appears to have further reduced the power of the method.

We also implemented a different discretization scheme, in which $K_1 = 6$ categories are used for the region $\omega < 1$ and $K_2 = 4$ categories are used for the region $\omega \geq 1$. This scheme appears to perform better for data sets with a small proportion of positively selected sites (such as in D7), but worse when the data do not contain positively selected sites (such as in D9). When a large proportion of sites are under diversifying selection (such as in D10), both schemes appear to perform well. Since the new K_1-K_2 scheme is not consistently better than the old scheme of $K = 10$ equal-probability categories, we have not used it for analysis in this article. Future work to devise more efficient approximations to the integral of Equation 6 is highly desirable and may restore some power to distinguish those continuous distributions we implemented.

Models implemented in this article assume that the selective pressure indicated by the ω ratio is identical among evolutionary lineages. They also assume that the nonsynonymous substitution rate is independent of the amino acids being interchanged; that is, at a positively selected site, all amino acids are assumed to be acceptable and all amino acid changes are assumed to be advantageous. By the criterion we use, a site will be considered to be under positive selection only if the nonsynonymous rate, averaged over all lineages in the phylogeny and over all possible amino acid-replacement mutations at the site, is higher than the synonymous rate. Thus LRTs of positive selection implemented in this article are conservative. It appears desirable to develop models that allow the selective pressure to vary both among lineages and among sites; such models may be much more powerful for detecting adaptive molecular evolution than those implemented in this article. In this regard, it should be noted that our models, conservative as they are, identified positive selection in several genes not previously suspected of being under positive selection.

Another assumption made in this article is a constant mutation (synonymous) rate among sites. This assumption may not always be realistic. However, we suggest that synonymous rate variation is unlikely to lead to false conclusions of adaptive evolution by the methods of this article. For example, at mutational hot spots, both synonymous and nonsynonymous rates will be elevated,

and if nonsynonymous mutations do not offer a selective advantage, the underlying ω ratio at such sites will not be >1. Similarly, in our formulation, selection at silent sites for translational efficiency, which has been nicely demonstrated in Drosophila by Akashi (1995, 1999) and suggested for mammals and viruses as well, has the sole effect of changing the codon usage pattern (π_j in Equation 1). It will not lead to $\omega > 1$ if nonsynonymous mutations do not offer a selective advantage. We note that the ω ratio in our models measures the net effect of selection at the protein level (see Equation 1). Unlike many tests of neutrality suggested in population genetics (see, *e.g.*, Wayne and Simonsen 1998; Fu and Li 1999 for reviews), which may be powerful in rejecting strict neutrality but not so powerful in distinguishing among different forms of natural selection, the LRTs described in this article aim to detect molecular adaptation.

How many genes are under positive selection? Our analysis demonstrated existence of sites under diversifying selection in 6 out of the 10 genes analyzed. The HIV-1 *env* gene (data set D10) is one of the best-known examples of adaptive evolution (*e.g.*, Holmes *et al.* 1995; Mindell 1996; Yamaguchi and Gojobori 1997); the selective pressure is presumably the surveillance of the host immune system. Previous analysis (Fitch *et al.* 1997) also suggested positive selection in the human influenza virus HA gene (data set D5). Besides those two genes, our analysis also detected adaptive evolution in the HIV-1 *vif* and *pol* genes (data sets D6 and D7) and in the mitochondrial (D1) and β-globin (D2) genes. The ω ratios averaged over all sites are ⩽1 in those data sets, and our models inferred adaptive molecular evolution in spite of this overwhelming effect of purifying selection. The 10 genes analyzed in this article are not a random sample of genes in various organisms. However, it appears likely that molecular adaptation happens much more often than has been recognized. We hope that our inference of sites under diversifying selection may prompt further lab-based investigation on the structure and function of the proteins to identify the selective agents.

Recommendations concerning models implemented in this article: We note that the selection model (M2), or the LRT comparing models M1 (neutral) and M2 (selection), does not detect selection in three out of the six data sets in which positive selection is inferred by other models. These three data sets are the mitochondrial genes (D1), the β-globin gene (D2), and the HIV-1 *pol* gene (D7). The lack of power of M2 (selection) appears to be due to the same reason in all three data sets, that is, the existence of a substantial proportion of sites in the gene with $0 < \omega < 1$ and the failure of M1 (neutral) to account for them. As a result, the extra category added in M2 is forced to account for sites with $0 < \omega < 1$, and the small proportion of positively selected sites with $\omega > 1$ is incorporated into the class of neutral sites with $\omega = 1$. In such cases, model M3 (discrete) appears much more powerful. Nevertheless, model M2 does detect positive selection in the three other data sets analyzed in this article. It was also found much more powerful than *ad hoc* pairwise comparisons or sliding window analysis in a recent study of positive selection in the HIV-1 *nef* gene (Zanotto *et al.* 1999). Thus we suggest that the model be used in real data analysis, with its limitations borne in mind. Models M0–M3 all involve much less computation than other models implemented in this article and may all be successfully fitted to the data.

We also recommend LRTs based on the beta null model (M7). In particular, comparison between M7 (beta) and M8 (beta&ω) appears to provide a powerful test of positive selection. Models M9–M11 can also be compared with M7 to test for positive selection, but those models detect selection only if a substantial proportion of positively selected sites exist in the gene. They often suffer from convergence problems and seldom fit the data better than M8 despite their use of an additional parameter. In using models M9–M11 as well as other continuous distribution models not based on the beta distribution (M5, M6, M12, and M13), it is important to examine the discrete distributions to see whether there is any category with $\omega > 1$. Although we did find data sets for which M12 and M13 gave good fits, M12 often caused serious convergence problems and M13 was even worse and hardly usable.

Data and program availability and program performance: The sequence alignments, the phylogenetic trees used, and extensive lists of positively selected sites and their posterior probabilities inferred under different models will be made available at the anonymous ftp site (ftp://abacus.gene.ucl.ac.uk/pub/YNGP2000/). Models developed in this article are implemented in the codeml program in the paml program package (Yang 1997), which is distributed at the web site http://abacus.gene.ucl.ac.uk/software/paml.html.

We thank Eddie Holmes and Walter Fitch for providing some of the data sets analyzed in this article. We thank Joe Bielawski and Simon Whelan for discussions and Daniel Haydon and two anonymous referees for comments. Z.Y. is supported by Biotechnology and Biological Sciences Research Council grant 31/G10434. R.N. is supported by National Science Foundation grant 9815367 to John Wakeley. N.G. is supported by a Wellcome Trust Fellowship in Biodiversity Research. A.K.P. is supported in part by grant 9701412 from the Danish Natural Science Research Council.

LITERATURE CITED

Akaike, H., 1974 A new look at the statistical model identification. IEEE Trans. Autom. AC **19:** 716–723.

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics **139:** 1067–1076.

Akashi, H., 1999 Within- and between-species DNA sequence variation and the "footprint" of natural selection. Gene **238:** 39–51.

Bhattacharjee, G. P., 1970 The incomplete gamma integral. Appl. Stat. **19:** 285–287.

Cao, Y., A. Janke, P. J. Waddell, M. Westerman, O. Takenaka *et*

al., 1998 Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. **47:** 307–322.

Crandall, K. A., C. R. Kelsey, H. Imamichi, H. C. Lane and N. P. Salzman, 1999 Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol. Biol. Evol. **16:** 372–382.

Emerman, M., and M. H. Malim, 1998 HIV-1 regulatory/accessory genes: keys to unraveling viral and host cell biology. Science **280:** 1880–1884.

Endo, T., K. Ikeo and T. Gojobori, 1996 Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13:** 685–690.

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

Fitch, W. M., R. M. Bush, C. A. Bender and N. J. Cox, 1997 Long term trends in the evolution of H(3) HA1 human influenza type A. Proc. Natl. Acad. Sci. USA **94:** 7712–7718.

Fu, Y. X., and W. H. Li, 1999 Coalescing into the 21st century: an overview and prospects of coalescent theory. Theor. Popul. Biol. **56:** 1–10.

Gillespie, J. H., 1991 *The Couses of Molecular Evolution*. Oxford University Press, Oxford.

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

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

Holmes, E. C., L. Q. Zhang, P. Robertson, A. Cleland, E. Harvey *et al.*, 1995 The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. J. Infect. Dis. **171:** 45–53.

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

Hughes, A. L., and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex loci reveals overdominant selection. Nature **335:** 167–170.

Kimura, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, United Kingdom.

Kumar, S., 1996 Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. Genetics **143:** 537–548.

Kuno, G., G.-J. J. Chang, K. R. Tsuchiya, N. Karabatsos and C. B. Cropp, 1998 Phylogeny of the genus *Flavivirus*. J. Virol. **72:** 73–83.

Leitner, T., S. Kumar and J. Albert, 1997 Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. J. Virol. **71:** 4761–4770.

Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Liò, P., and N. Goldman, 1998 Models of molecular evolution and phylogeny. Genome Res. **8:** 1223–1244.

Majumder, K. L., and G. P. Bhattacharjee, 1973 The incomplete beta integral (AS63). Appl. Stat. **22:** 409–411.

Messier, W., and C.-B. Stewart, 1997 Episodic adaptive evolution of primate lysozymes. Nature **385:** 151–154.

Mindell, D. P., 1996 Positive selection and rates of evolution in immunodeficiency viruses from humans and chimpanzees. Proc. Natl. Acad. Sci. USA **93:** 3284–3288.

Miyata T., and T. Yasunaga, 1980 Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. J. Mol. Evol. **16:** 23–36.

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.

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

Nielsen, R., 1997 The ratio of replacement to silent divergence and tests of neutrality. J. Evol. Biol. **10:** 217–231.

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.

Ohta, T., 1993 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23:** 263–286.

Pedersen, A.-M. K., C. Wiuf and F. B. Christiansen, 1998 A codon-based model designed to describe lentiviral evolution. Mol. Biol. Evol. **15:** 1069–1081.

Seibert, S. A., C. Y. Howell, M. K. Hughes and A. L. Hughes, 1995 Natural selection on the *gag*, *pol* and *env* genes of human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. **12:** 803–813.

Self, S. G., and K.-Y. Liang, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. **82:** 605–610.

Sharp, P. M., 1997 In search of molecular Darwinism. Nature **385:** 111–112.

Wayne, M. L., and K. L. Simonsen, 1998 Statistical tests of neutrality in the age of weak selection. TREE **13:** 236–240.

Yamaguchi, Y., and T. Gojobori, 1997 Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc. Natl. Acad. Sci. USA **94:** 1264–1269.

Yang, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:** 306–314.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13:** 555–556.

Zanotto, P. M., E. A. Gould, G. F. Gao, P. H. Harvey and E. C. Holmes, 1996 Population dynamics of flaviviruses revealed by molecular phylogenies. Proc. Natl. Acad. Sci. USA **93:** 548–553.

Zanotto, P. M., E. G. Kallas, R. F. Souza and E. C. Holmes, 1999 Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics **153:** 1077–1089.

Communicating editor: W. Stephan