# Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene

**Rasmus Nielsen*[*] and Ziheng Yang*[*,†]**

*[*]Department of Integrative Biology, University of California, Berkeley, California 94720-3140 and [†]Department of Biology, University College London, London NW1 2HE, England*

Manuscript received July 21, 1997
Accepted for publication December 5, 1997

ABSTRACT

Several codon-based models for the evolution of protein-coding DNA sequences are developed that account for varying selection intensity among amino acid sites. The "neutral model" assumes two categories of sites at which amino acid replacements are either neutral or deleterious. The "positive-selection model" assumes an additional category of positively selected sites at which nonsynonymous substitutions occur at a higher rate than synonymous ones. This model is also used to identify target sites for positive selection. The models are applied to a data set of the V3 region of the HIV-1 envelope gene, sequenced at different years after the infection of one patient. The results provide strong support for variable selection intensity among amino acid sites The neutral model is rejected in favor of the positive-selection model, indicating the operation of positive selection in the region. Positively selected sites are found in both the V3 region and the flanking regions.

An excess of nonsynonymous substitutions over synonymous substitutions is an unambiguous indicator of positive natural selection at the molecular level. Estimation of synonymous and nonsynonymous substitution rates has thus provided an important tool to study the process of molecular sequence evolution. For example, positive selection has been identified this way in several systems, including the human major histocompatibility complex (Hughes and Nei 1988), primate stomach lysozymes (Messier and Stewart 1997), abalone sperm lysins (Lee *et al.* 1995), and human HIV-1 genes (Bonhoeffer *et al.* 1995; Mindell 1996; Yamaguchi and Gojobori 1997).

A number of methods have been proposed to estimate the numbers of synonymous ($d_S$) and nonsynonymous ($d_N$) substitutions per site between two sequences (*e.g.*, Li *et al.* 1985; Nei and Gojobori 1986; Ina 1995). A maximum likelihood method for estimating $d_S$ and $d_N$ based on an explicit codon substitution model was also proposed (Goldman and Yang 1994). However, these methods assume that all (codon or amino acid) sites in the sequence are under the same selection pressure with the same underlying $d_N/d_S$ ratio. This assumption is unrealistic as different sites in a protein perform different functional and structural roles, and thus must be under different selection pressure. The assumption also appears to be at variance with a strictly neutral model of protein evolution. The neutral theory (Kimura 1968) asserts that new mutations are either deleterious or neutral. A simple model of this nature may assume that some sites in a gene are invariable, at which amino acid-altering mutations are deleterious and removed by selection, while other sites are neutral, at which synonymous and nonsynonymous mutations are fixed at the same rate.

In almost all proteins where positive selection has been demonstrated to be operating, only a few amino acid sites were found to be responsible for the adaptive evolution (Hughes and Nei 1988; Yokoyama and Yokoyama 1996). In such cases, the estimate of the $d_N/d_S$ ratio for the entire sequence may be smaller than one, even if some sites are under positive selection. Furthermore, ignoring the variation in selection intensity (and thus in nonsynonymous rates) among sites leads to underestimation of nonsynonymous rates and of the $d_N/d_S$ ratio (Nielsen 1997). Therefore, current methods of estimating synonymous and nonsynonymous rates may fail to detect positive selection even when it exists.

In this paper, we develop codon-based models for the evolution of protein-coding DNA sequences that allow for variable selection intensity among sites. The nonsynonymous/synonymous substitution rate ratio ($d_N/d_S$), which reflects the selection intensity at the amino acid level, is allowed to vary among amino acid sites. The models are implemented in a maximum likelihood framework and lead to likelihood ratio tests of neutral evolution. We also develop a Bayesian approach for identifying positively selected amino acid sites. The models and methods are applied to analyze a data set of the V3 region of the HIV-1 envelope gene (Holmes *et al.* 1992).

## THEORY

**Model of codon substitution:** A simplified version of the codon substitution model proposed by Goldman and Yang (1994) will be used in this paper. Stop codons

*Corresponding author:* Rasmus Nielsen, Museum of Comparative Zoology, Harvard University, 26 Oxford St., Cambridge, MA 02138. E-mail: rasmus@mws4.biol.berkeley.edu

are not allowed in the gene sequence, and substitutions between sense codons are described by a continuous-time Markov process. The instantaneous substitution rate from codon $i$ to $j$ ($i \neq j$) is given by

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

where parameter $\kappa$ is the transition/transversion rate ratio, $\omega$ is the nonsynonymous/synonymous rate ratio, and $\pi_j$ is the equilibrium frequency of codon $j$, calculated using the nucleotide frequencies at the three codon positions. Note that the relationship holds that $\omega = d_N/d_S$. The unit of evolution under such a model is the codon and, in this paper, the term "site" refers to a codon or amino acid instead of a nucleotide.

**Neutral model:** This model assumes two categories of sites in the gene. The first category includes neutral sites where nonsynonymous mutations are neutral, with the $d_N/d_S$ ratio $\omega_1 = 1$. The substitution rate from codon $i$ to $j$ for a site in this category is

$$Q_{ij}^{(1)} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_j, & \text{for transversion,} \\ \kappa\pi_j, & \text{for transition.} \end{cases} \quad (2)$$

The second category includes the conserved sites, where nonsynonymous mutations are deleterious and eliminated by selection, and only synonymous substitutions are possible, so that $\omega_2 = 0$. The substitution rate from codon $i$ to $j$ ($i \neq j$) at such a site is thus

$$Q_{ij}^{(2)} = \begin{cases} \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Under this model, the synonymous substitution rate is constant among sites, while the nonsynonymous rate is variable. At the conserved sites, the nonsynonymous rate is zero, and at the neutral sites, the nonsynonymous rate is equal to the synonymous rate.

Let the proportions of codon sites in the two categories be $p_1$ and $p_2 = 1 - p_1$. Let $n$ be the number of sites (codons) in the sequence and the data at site $h$ be $x_h$ ($h = 1, 2, \ldots, n$). Because we do not know a priori which category a site belongs to, the probability of observing data $x_h$ is an average over the two possibilities:

$$f(x_h) = \sum_{k=1}^{2} p_k f(x_h|\omega_k) = p_1 f(x_h|\omega_1) + p_2 f(x_h|\omega_2), \quad (4)$$

where $f(x_h|\omega_k)$ is the probability of observing data $x_h$ given that site $h$ is from category $k$ ($k = 1, 2$), with nonsynonymous/synonymous rate ratio $\omega_k$. This conditional probability can be calculated for a given phylogenetic tree and branch lengths according to the method of Goldman and Yang (1994). The log likelihood is a sum over all $n$ sites in the sequence

$$l = \sum_{h=1}^{n} \log\{f(x_h)\}. \quad (5)$$

It may be noted that the structure of the model is similar to models of variable evolutionary rates among nucleotide or amino acid sites developed previously (Yang 1993, 1994, 1995). Evolutionary rates under these variable rates models, and the $d_N/d_S$ ratios ($\omega$) reflecting selection intensity under the present model, are both formulated as random variables and integrated out in the likelihood function. Calculation of the likelihood function under the codon-based model can be easily adapted from the algorithm for variable evolutionary rates among sites (Yang 1994). The matrix of transition probabilities over time $t$, $P(t) = e^{Qt}$, is calculated through diagonalization of the rate matrix $Q = \{Q_{ij}\}$; a standard numerical algorithm is used for this purpose (Goldman and Yang 1994). $P(t)$ needs to be calculated separately for each branch (with length $t$) and site category (with parameter $\omega$). Parameters in the model include the branch lengths ($t$), the transition/transversion rate ratio ($\kappa$), and the proportions of the two categories of sites ($p_1$ and $p_2$, with $p_1 + p_2 = 1$). These parameters are estimated by maximum likelihood using numerical optimization algorithms. The codon frequency parameters ($\pi_j$) are calculated using the observed nucleotide frequencies at the three codon positions.

**Positive selection model:** The neutral model can be extended by adding an extra category of positively selected sites (with $\omega_3 > 1$). Nonsynonymous mutations at such sites thus have higher probabilities of fixation than synonymous mutations. Let the proportions of sites in the three categories be $p_1$, $p_2$, and $p_3$ (with $p_1 + p_2 + p_3 = 1$), and let the corresponding nonsynonymous/synonymous rate ratios be $\omega_1 = 1$, $\omega_2 = 0$, and $\omega_3 > 1$. The probability of observing data ($x_h$) at site $h$ is then

$$f(x_h) = \sum_{k=1}^{3} p_k f(x_h|\omega_k). \quad (6)$$

The log likelihood function can be calculated similarly to that under the neutral model with two categories of sites.

The positive-selection model is an extension of the neutral model, with two more parameters. Twice the log likelihood difference between the two models can be compared with a $\chi^2$ distribution with d.f. = 2. This constitutes a likelihood ratio test of neutrality against an alternative model of positive selection.

In practice, $\omega_3$ is optimized in the entire region from zero to infinity. In this case, it is appropriate to call the model a positive-selection model only if $\omega_3 > 1$, as an estimate of $\omega_3$ smaller than one provides no evidence for positive selection.

Because one might expect a continuum of the $d_N/d_S$ rate ratio among sites even when no positive selection operates, we consider a variation of the neutral model in which the class $\omega_0 = 1$ is replaced by a truncated gamma distribution on (0, 1). A proportion $p_1$ of sites have rates from the truncated gamma distribution, while a proportion $p_2$ ($= 1 - p_1$) of sites are conserved and have $\omega_2 = 0$. The truncated gamma distribution with shape parameter $\alpha$ and scale parameter $\beta$ has the following density

$$f(w) = \frac{e^{-\beta w} w^{\alpha-1}}{\int_0^1 e^{-\beta w} w^{\alpha-1} dw}, \quad 0 < w \leq 1. \tag{7}$$

To avoid the use of too many parameters, we fix $\beta$ to be equal to $\alpha$. Use of other values for $\beta$ is found to give roughly the same likelihood. For efficient computation, the truncated gamma distribution was approximated by five rate categories with equal probabilities (see Yang 1994). We will refer to this model as the "continuous neutral model." In the corresponding positive-selection model, we add an extra category of sites with proportion $p_3$, and $d_N/d_S$ ratio $\omega_3$. This will be referred to as the "continuous positive-selection model." Notice that when $\alpha = \infty$, the two new models with continuous substitution rates are identical to the neutral and positive-selection models described above.

**Identification of positively selected amino acid sites:** When parameters of the positive-selection model are estimated, an empirical Bayes' approach can be used to infer which category the site most likely belongs to. This method is similar to the approach of Yang (1994) for estimating rate categories for nucleotide sites under a model of variable evolutionary rates among sites. The posterior probability that a site with data $x_h$ belongs to category $k$ (with rate ratio $\omega_k$) is given by

$$\text{prob}(\omega_k|x_h) = \frac{p_k f(x_h|\omega_k)}{f(x_h)} = \frac{p_k f(x_h|\omega_k)}{\Sigma_{l=1}^{3} p_l f(x_h|\omega_l)}. \tag{8}$$

The category $k$ that maximizes the posterior probability is the most likely category for the site. Positively selected sites (*i.e.*, sites belonging to the third category with $\omega_3 >$ 1) may be identified this way. The posterior probabilities provide a measure of accuracy of that inference.

## APPLICATIONS TO THE HIV-1 ENVELOPE GENE

To examine the utilities of the models developed in this paper, we analyze the DNA sequence data of the HIV-1 envelope genes published by Holmes *et al.* (1992). The sequences are from viral variants in years 3 through 7 after infection of one single patient. No sequence variation was detected in year 0, and no data were available for years 1 and 2 (Holmes *et al.* 1992). Each sequence contains 77 codon sites from the third hypervariable region (V3) and flanking regions of the external glycoprotein (gp120). There has been considerable interest in possible selective factors acting on this gene because it encodes known targets for cytotoxic T lymphocytes and neutralizing antibodies. For example, the V3 region is suggested to be under positive diversifying selection during the course of infection (Boenhoeffer *et al.* 1995). Boenhoeffer *et al.* found that the average $d_N/d_S$ rate ratios for each year, estimated using the method of Nei and Gojobori (1986) for pairwise sequence comparison, were greater than one and decreased over the years, suggesting that positive selection may be important in the diversification of the HIV-1 envelope genes and that the selection pressure may have been relaxed during the cause of the viral infection.

There are 15, 11, 23, 15, and 13 sequences for years 3, 4, 5, 6, and 7, respectively, and the number of distinct sequences is 13, 11, 17, 15, and 12, respectively. As our codon-based models involve heavy computation, which makes it unfeasible to perform an analysis on a phylogeny of all the 77 sequences simultaneously, we analyze sequences from different years as if they were separate data sets. Only distinct sequences are used. For data of each year, the likelihood method was used under both simple and sophisticated nucleotide substitution models to perform heuristic tree searches to identify candidate trees. Candidate topologies found in this way often share long interior branches, for which the statistical support is strong, while the details of the topology may be different in different analyses. Results (not shown) suggest that our codon-based analysis, to be presented below, is not sensitive to the assumed topology of the phylogenetic tree, as long as the long interior branches are preserved in the topology. Two candidate topologies for each data set are used in later analysis. They give essentially identical parameter estimates, and results obtained from only one of them are reported in this paper. The tree topologies used are not presented, but are available from the authors upon request. For each data set, several starting values were used in the iteration. This was done as a protection against the existence of multiple local optima in the likelihood function. In this study, all starting values for a given data set resulted in the same optimum.

Besides the neutral and positive-selection models of this paper, two additional codon-based models developed previously (Goldman and Yang 1994) are also used in the analysis. Both of these models assume a constant $d_N/d_S$ ratio among sites. The first, referred to as the "Goldman and Yang" model in Table 1, applies the same rate matrix ($Q$ specified by Equation 1) to all sites. The second, referred to as the "gamma-rates" model in Table 1, assumes gamma-distributed rates among codon sites (Yang 1994; Goldman and Yang 1994). In this model, the substitution rate from codon $i$ to $j$ at a codon site with rate $r$ is $rQ_{ij}$, with $Q_{ij}$ given by Equation 1. A codon with a high rate tends to have both high synonymous and high nonsynonymous rates. The shape parameter $\alpha$ of the gamma distribution is inversely related to the extent of rate variation.

**Comparison of models and tests of positive selection:** The log likelihood values under different codon-based

TABLE 1

**Parameter estimates and log likelihood values under different models for the HIV-1 envelope gene**

| Model | | Year 3 | Year 4 | Year 5 | Year 6 | Year 7 |
|---|---|---|---|---|---|---|
| Goldman and Yang | $\omega$ | 6.249 | 0.640 | 0.778 | 0.803 | 0.358 |
| | $\kappa$ | 2.539 | 4.628 | 5.556 | 2.796 | 4.163 |
| | $l$ | −537.22 | −457.07 | −483.66 | −588.89 | −555.54 |
| Gamma-rates | $\omega$ | 8.524 | 0.802 | 0.990 | 0.963 | 0.358 |
| | $\kappa$ | 2.662 | 4.500 | 5.930 | 2.730 | 4.131 |
| | $\alpha$ | 0.210 | 0.247 | 0.191 | 0.344 | 0.588 |
| | $l$ | −520.14 | −448.61 | −471.71 | −574.28 | −548.17 |
| Neutral | $p_1$ | 0.413 | 0.334 | 0.335 | 0.392 | 0.379 |
| | $p_2$ | 0.587 | 0.666 | 0.665 | 0.608 | 0.621 |
| | $\kappa$ | 2.305 | 4.562 | 5.725 | 2.752 | 4.389 |
| | $l$ | −531.94 | −450.66 | −474.93 | −576.23 | −546.09 |
| Positive-selection | $p_1$ | 0.766 | 0.394 | 0.198 | 0.369 | 0.359 |
| | $p_2$ | 0.018 | 0.560 | 0.645 | 0.547 | 0.598 |
| | $p_3$ | 0.216 | 0.046 | 0.157 | 0.084 | 0.043 |
| | $\omega_3$ | 29.537 | 14.893 | 6.920 | 6.849 | 7.709 |
| | $\kappa$ | 2.644 | 5.147 | 6.095 | 2.900 | 5.051 |
| | $l$ | −519.06 | −444.68 | −470.32 | −570.15 | −543.67 |
| Continuous neutral | $p_1$ | 0.413 | 0.334 | 0.335 | 0.392 | 0.379 |
| | $p_2$ | 0.587 | 0.666 | 0.665 | 0.608 | 0.621 |
| | $\kappa$ | 2.305 | 4.562 | 5.725 | 2.752 | 4.389 |
| | $\alpha$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| | $l$ | −531.94 | −450.66 | −474.93 | −576.23 | −546.09 |
| Continuous positive-selection | $p_1$ | 0.766 | 0.457 | 0.487 | 0.369 | 0.406 |
| | $p_2$ | 0.018 | 0.495 | 0.341 | 0.547 | 0.545 |
| | $p_3$ | 0.216 | 0.047 | 0.172 | 0.084 | 0.049 |
| | $\omega_3$ | 29.537 | 13.651 | 6.469 | 6.849 | 6.504 |
| | $\kappa$ | 2.644 | 5.160 | 6.146 | 2.900 | 5.062 |
| | $\alpha$ | $\infty$ | 10.315 | 0.513 | $\infty$ | 5.528 |
| | $l$ | −519.06 | −444.65 | −470.32 | −570.15 | −543.42 |

models are listed in Table 1 for data of each year. The gamma model reduces to the Goldman and Yang model when the gamma shape parameter $\alpha \to \infty$. Comparison of twice the log likelihood difference between the two models with a $\chi^2$ distribution with 1 d.f. suggests that the gamma model provides a significantly better fit to data of each year. This result is in accordance with previous analyses (Goldman and Yang 1994), demonstrating that rate variation is an important factor in the evolution of coding sequences.

The neutral model and the Goldman and Yang model are not nested, and so cannot be tested using a $\chi^2$ approximation. Nevertheless, their likelihood values are comparable. The two models have the same number of parameters but the log likelihood value under the neutral model is higher than that under the Goldman and Yang model by 5–12 units (Table 1). The neutral model is thus a much more realistic representation of the evolutionary dynamics of the HIV-1 envelope gene. This is also true for year 3, even though the average rate of nonsynonymous substitution is much higher than the average rate of synonymous substitution, suggesting that the neutral model should provide a poor fit to the data. Except for data of year 3, the fit of the neutral model is almost as good as, or even slightly better than, the gamma-rates model.

The positive-selection model includes two more parameters than the neutral model. These two models are nested, and twice the log likelihood difference can be compared with a $\chi^2$ distribution with d.f. = 2 to test whether the positive-selection model provides a better fit to data than the neutral model. The difference is significant ($P < 5\%$) for data of every year, except year 7 (for which $P \approx 9\%$). Clearly, the neutral model is inadequate to describe these data, and positive selection has been operating in the evolution of this viral gene. The Goldman and Yang model is also a special case of the positive-selection model with the constraint that $p_1 =$

$p_2 = 0$, and $p_3 = 1$. Twice the log likelihood difference between the two models ranges from 24 to 37 for different years. These differences are all significant ($P < 1\%$ with d.f. $= 2$), and the Goldman and Yang model with a constant $d_N/d_S$ ratio among sites is rejected in favor of the positive-selection model for data of each year.

The same conclusion is reached when rates in the range between 0 and 1 are allowed in the neutral model (continuous neutral model). Indeed, the neutral model with, and without, continuously distributed mutation rates, modeled by the truncated gamma distribution, have the same likelihood values. Similarly, the positive-selection model, with and without continuous mutation rates, have very similar likelihood values. We have also fitted another set of neutral and positive-selection models by removing the class with $\omega_2 = 0$ in the continuous mutation models. The neutral model thus constructed, which assumes that all sites have the $d_N/d_S$ rate ratio $\omega$ from the truncated gamma distribution, fits the data much more poorly than the neutral model specified by Equations 2 and 3. The corresponding positive-selection model produces likelihood values very similar to those under the simple three-class positive-selection model. Therefore, likelihood ratio tests using this set of models all suggest positive selection acting on the gene (results not shown). To summarize, the statistical support for positive selection on the envelope gene appears to be rather insensitive to the assumed distribution of selection intensity among sites.

The positive-selection model and the gamma-rates model are not nested and cannot be compared using a $\chi^2$ approximation. However, the positive selection model has higher likelihood value than the gamma-rates model for data of each year. Although the significance of the differences is uncertain, the results suggest that the positive-selection model provides the most realistic description of the evolution of the analyzed sequences. The reason appears to be that in the gamma-rates model the same distribution of rates is applied to both the synonymous and nonsynonymous substitutions. However, most of the rate variation appears to be caused by variation in the selection intensity among nonsynonymous substitutions.

**Variation of nonsynonymous substitution rates among years:** Maximum likelihood estimates of parameters obtained under different codon-based models are listed in Table 1 for data of each year. Estimates of the transition/transversion rate ratio ($\kappa$) are more variable among years than among methods, and range from two to six, indicating that transitions occur much more frequently than transversions.

Notice that the neutral model, with and without a truncated gamma distribution, provides identical parameter estimates. The reason for this may be that the true rates are highly bimodal with a mode at $\omega = 0$ and a mode at $\omega > 1$. Very little of the probability mass appears to be located in the region between zero and one. Likewise, for the selective models, almost no improvement in the likelihood is obtained by allowing nonsynonymous rates in the region between zero and one. In the following we will therefore concentrate on the results of the neutral and selection models that do not allow nonsynonymous rates between $\omega = 0$ and $\omega = 1$.

Considerable differences exist in estimates of $\omega$ between year 3 and the remaining years. For years 4 through 7 under the Goldman and Yang model, the estimate of $\omega$ is $<1$. Likewise, estimates of $p_2$ (proportion of conserved sites with $\omega_2 = 0$) under both the neutral and the positive-selection models range from 0.55 to 0.67 among years 4 through 7, suggesting that a majority of amino acid sites in the protein are conserved in years 4 through 7. Estimates of $p_3$ (proportion of positively selected sites with $\omega_3 > 1$) under the positive-selection model range from 0.04 to 0.22, indicating that only a few sites in the sequence are under positive diversifying selection at any particular time.

Boenhoeffer *et al.* (1995) calculated the average $d_N/d_S$ ratio among pairwise comparisons of sequences in the same year, and noted a decrease in the average $d_N/d_S$ ratio over the years. They suggested that relaxed selection pressure during the course of infection may be responsible. Nielsen (1997) suggested that the change in the inferred nonsynonymous/synonymous rate ratios using a model of constant selection intensity may partly be due to the presence of hypervariable nonsynonymous sites. Ignoring nonsynonymous rate variation leads to underestimation of the nonsynonymous/synonymous rate ratio at high levels of sequence divergence. The results of the present study show that hypervariable nonsynonymous sites are in fact present in the genes. However, even after the nonsynonymous rate variation has been accounted for, a change in the selection intensity is still observed between year 3 and the subsequent years. Estimates for year 3 under the positive-selection model suggest a high proportion ($p_3 =$ 21.6%) of sites under strong positive selection ($\omega_3 =$ 29.5). Estimates of both $p_3$ and $\omega_3$ for years 4 through 7 are much smaller, indicating that the effect of positive selection is stronger in year 3 than in the later years. These results are in general agreement with the conclusion of Boenhoeffer *et al.* (1995), although the standard errors of our parameter estimates (not shown) are large.

**Identification of positively selected sites:** Because the positive-selection model provides a better fit to data of every year than the neutral model, Equation 7 is used to infer the most likely site category (with the associated $d_N/d_S$ ratio) at each codon (amino acid) site. The posterior probabilities are also calculated. The results are shown in Figure 1, A–E, for years 3 through 7. Consistent with the parameter estimates (Table 1), more sites are inferred to be under positive selection in year 3 than in years 4 through 7. Only site 77 was identified with very high posterior probability to be under positive se-

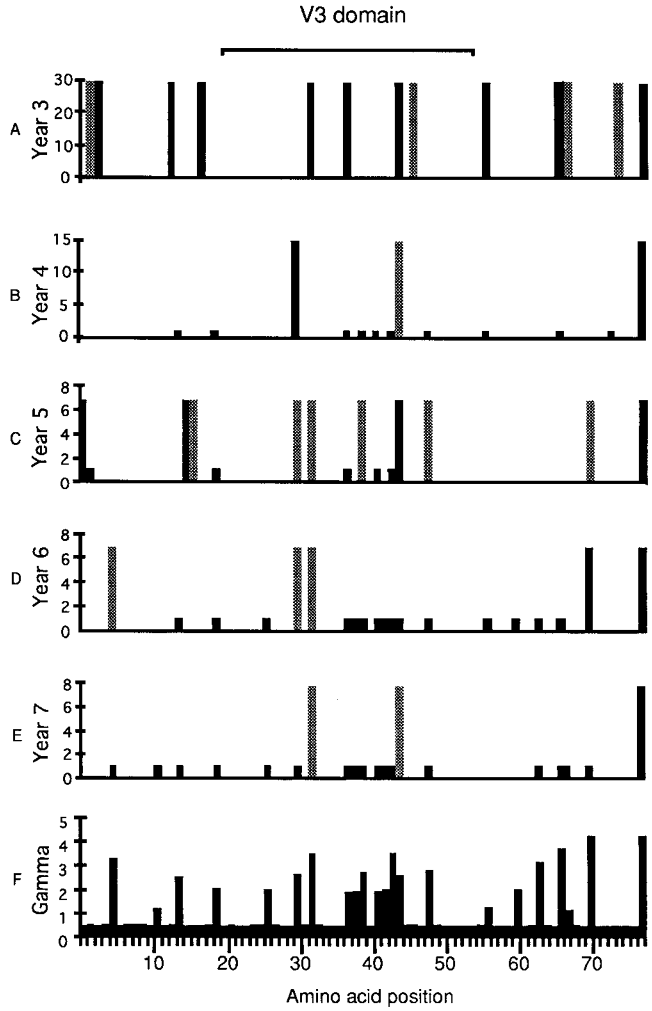

Figure 1.—(A–E) Estimates of nonsynonymous/synonymous rate ratios ($\omega$ in Equation 1) for sites in the V3 region of the HIV-1 envelope gene. Positively selected sites are labeled black if the posterior probability is larger than 95%, and gray if otherwise. Note that the scales are different among years. (F) Estimates of amino acid replacement rates under the empirical model of Jones *et al.* (1992) with gamma distributed rates among sites (Yang 1994). These are relative rates with the mean to be one. The location of the V3 region is indicated.

lection in all five years. Site 32 was identified to be under positive selection in years 3, 5, 6, and 7, while site 44 may be under positive selection in years 3, 4, 5, and 7. Apart from these sites, there seems to be considerable variation in the inferred sites for positive selection from year to year. The variation over the years may in part be due to rapid fixations of old mutations, and arrivals of new ones, as Holmes *et al.* (1992) observed drastic changes of the most abundant sequence over the years.

Our analysis suggests that there may be as many positively selected sites (sites potentially under positive selection) in the flanking regions as in the V3 region. The single site that is identified to be under positive selection in all 5 years is site 77, outside the V3 region. The results suggest that selection may be acting on a broader region of the gene sequence than previously suggested (*e.g.*, Willey *et al.* 1986). Furthermore, a considerable amount of the variability in the V3 region appears to be neutral.

It should be noted that positively selected sites identified in this way are not equivalent to highly variable sites. A positively selected site is characterized by both a high variability and an excess of nonsynonymous to synonymous substitutions. To illustrate this point, evolutionary rates at amino acid sites were estimated using the protein sequences translated from the DNA sequences of years 6 and 7 (Figure 1F). The empirical model of Jones *et al.* (1992) with gamma-distributed rates among sites was assumed. The method of Yang (1994, 1995) was then used to estimate the amino acid replacement rates at sites. Notice that there are many more sites identified as having high rates than there are sites undergoing positive selection. Our method, which assumes variable selection intensity instead of variable amino acid replacement rates, should therefore have a higher precision in identifying target sites for selection.

**Analysis of Cytochrome Oxidase II: a case of no positive selection:** To illustrate the performance of the method in a case where positive selection is not expected to have strongly influenced the evolution of the DNA sequence, we also applied the method to 10 vertebrate sequences of the mitochondrial Cytochrome Oxidase II (COII) gene. Sequences from 10 vertebrate species are used (the data set is described in Cummings *et al.* 1995). This COII gene has been extensively analyzed in the literature, and to the knowledge of the authors, there have been no suggestions of codons with elevated nonsynonymous rates. The results of the analysis are in Table 2. Note that an estimate of $\omega_3 < 1$ was obtained for this data set. This suggests that there may not be a large category of amino acid sites in this gene, where nonsynonymous substitutions occur at a higher rate

**TABLE 2**

**Parameter estimates and log likelihood values under different models for the COII gene**

| Model | $p_1$ | $p_2$ | $p_3$ | $\kappa$ | $\omega_3$ | $l$ |
|---|---|---|---|---|---|---|
| Neutral | 0.479 | 0.521 | — | 2.403 | — | −4429.049 |
| Selection | 0.454 | 0.013 | 0.532 | 2.317 | 0.027 | −3910.412 |

than synonymous substitutions. This result is in contrast to those obtained from the HIV-1 envelope gene.

## DISCUSSION

The biochemical properties of proteins suggest that the selection pressure should vary among amino acid sites. The analysis of the HIV-1 envelope genes strongly supports this assertion. A significant improvement in the fit of the model to data is achieved by allowing for variation of the selection intensity (reflected in the $d_N/d_S$ ratio) among sites. Thus, the models developed in this paper, although very simple in nature, may provide more realistic descriptions of the evolutionary processes of protein-coding DNA sequences than previous models assuming a constant selection intensity among sites. This result may be surprising and emphasizes the importance of considering variation in the selection intensity along the gene when modeling molecular evolution.

Inference on the distribution of selection intensities along the gene is an underused tool in the search for the causes of molecular evolution. It may be possible to transform population genetic models concerning the distribution of selection coefficients among alleles into distributions of the nonsynonymous/synonymous rate ratios among amino acid sites. For example, models of slightly deleterious mutations, such as the models considered by Ohta (1973), may provide quite different predictions concerning the distribution of selection intensities than the models of positive selection considered by Gillespie (1991). Such models may be fitted to protein-coding DNA sequence data using the likelihood framework developed in this paper. Different models can then be compared with molecular sequence data using rigorous likelihood ratio tests.

We also envisage the application of our likelihood ratio test of neutrality to various real data sets. It may be worthwhile to conduct a large-scale screening and apply the test to genes from different organisms and genomes. Our likelihood approach is applied to the original sequence data and accounts for the phylogenetic relationship of the sequences. The likelihood ratio test constructed this way makes full use of the information contained in the data and may be more powerful than previous methods. For example, knowledge of the protein structure helped Hughes and Nei (1988) identify the target sites for selection in the MHC molecule. Our methods do not require such knowledge in order to detect positive selection. Furthermore, when positive selection is detected, our approach enables identification of the potential target sites for positive selection in the protein sequence.

**Program performance and availability:** The codon-based models developed in this paper involve intensive numerical computation. The likelihood calculation involves manipulations of matrices of size 61 × 61, instead of size 4 × 4, for nucleotide-based models. For error checking, independent C programs were written by both authors. These were found to be computationally feasible for data of ~10–20 sequences (such as data sets analyzed in this paper) on fast workstations. The methods will be made available in the PAML program package (Yang 1997).

We thank M. Uyenoyama and two anonymous reviewers for comments. This study is supported by National Institutes of Health grant GM40282 to Montgomery Slatkin and a personal grant to R.N. from the Danish Research Council.

## LITERATURE CITED

Boenhoeffer, S., E. C. Holmes and M. A. Nowak, 1995 Causes of HIV diversity. Nature **376:** 125.

Cummings, M. P., S. P. Otto and J. Wakeley, 1995 Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. **12:** 814–822.

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.

Holmes, E. C., L. Q. Zhang, P. Simmonds, C. A. Ludlam and A. J. L. Brown, 1992 Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc. Natl. Acad. Sci. USA **89:** 4835–4839.

Hughes, A. L., and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167–70.

Ina, Y., 1995 New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. **40:** 190–226.

Jones, D. T., W. R. Taylor and J. M. Thornton, 1992 The rapid generation of mutation data matrices from protein sequences. CABIOS **8:** 275–282.

Kimura, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624–626.

Lee, Y. H., T. Ota and V. D. Vacquier, 1995 Positive selection is a general phenonemon in the evolution of abalone sperm lysin. Mol. Biol. Evol. **12:** 231–238.

Li, W.-H., C.-I. Wu and C.-C. Luo, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150–174.

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.

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.

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

Willey, R. L., R. A.. Rutledge, S. Dias, T. Folks, T. Theodore *et al.*, 1986 Identification of conserved and divergent domains within the envelope gene of the acquired immunodeficiency syndrome virus. Proc. Natl. Acad. Sci. USA **83:** 5038–5042.

Yang, Z., 1993 Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **39:** 105–111.

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., 1995 A space-time process model for the evolution of DNA sequences. Genetics **139:** 993–1005.

Yang, Z, 1997 *Phylogenetic analysis by maximum likelihood* (PAML), Version 1.3. University of California, Berkeley, California, USA.

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

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.

Yokoyama, S., and R. Yokoyama, 1996 Adaptive evolution of photoreceptors and visual pigments in vertebrates. Ann. Rev. Ecol. Syst. **27:** 543–567.

Communicating editor: M. K. Uyenoyama