

Modeling the Site-Specific Variation of Selection Patterns along Lineages

Author(s): Stéphane Guindon, Allen G. Rodrigo, Kelly A. Dyer, John P. Huelsenbeck and Joseph Felsenstein

Source: *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. 35 (Aug. 31, 2004), pp. 12957-12962

Published by: [National Academy of Sciences](#)

Stable URL: <http://www.jstor.org/stable/3373185>

Accessed: 14-11-2015 14:08 UTC

## REFERENCES

Linked references are available on JSTOR for this article:

[http://www.jstor.org/stable/3373185?seq=1&cid=pdf-reference#references\\_tab\\_contents](http://www.jstor.org/stable/3373185?seq=1&cid=pdf-reference#references_tab_contents)

You may need to log in to JSTOR to access the linked references.

---

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



National Academy of Sciences is collaborating with JSTOR to digitize, preserve and extend access to *Proceedings of the National Academy of Sciences of the United States of America*.

<http://www.jstor.org>

# Modeling the site-specific variation of selection patterns along lineages

Stéphane Guindon\*, Allen G. Rodrigo\*, Kelly A. Dyer†, and John P. Huelsenbeck\*‡§

\*Bioinformatics Institute and Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Private Bag 92019, Auckland, New Zealand;

†Department of Biology, University of Rochester, Rochester, NY 14627; and ‡Section of Ecology, Behavior, and Evolution, Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093-0116

Edited by Joseph Felsenstein, University of Washington, Seattle, WA, and approved July 23, 2004 (received for review March 29, 2004)

**The unambiguous footprint of positive Darwinian selection in protein-coding DNA sequences is revealed by an excess of nonsynonymous substitutions over synonymous substitutions compared with the neutral expectation. Methods for analyzing the patterns of nonsynonymous and synonymous substitutions usually rely on stochastic models in which the selection regime may vary across the sequence but remains constant across lineages for any amino acid position. Despite some work that has relaxed the constraint that selection patterns remain constant over time, no model provides a strong statistical framework to deal with switches between selection processes at individual sites during the course of evolution. This paper describes an approach that allows the site-specific selection process to vary along lineages of a phylogenetic tree. The parameters of the switching model of codon substitution are estimated by using maximum likelihood. The analysis of eight HIV-1 *env* homologous sequence data sets shows that this model provides a significantly better fit to the data than one that does not take into account switches between selection patterns in the phylogeny at individual sites. We also provide strong evidence that the strength and the frequency of occurrence of selection might not be estimated accurately when the site-specific variation of selection regimes is ignored.**

positive selection | codon-based model of nucleotide substitutions | phylogeny | maximum likelihood

An excess of nonsynonymous changes to synonymous changes in protein-coding DNA sequences is an unambiguous signature of adaptive molecular evolution (1). Such a pattern is best explained by a selective advantage in the past for substitutions that cause amino acid changes. Indeed, there are numerous examples where substitutions causing amino acid changes confer a selective advantage. For example, in the MHC, overdominant selection appears to be responsible for the excess of replacement substitutions in the antigen-recognition site (1). Similarly, positive selection has been detected in viral proteins subject to immune surveillance (2–4), in abalone sperm lysins (5), primate lysozymes (6), and regions involved in species-specific sperm-egg interaction (7).

A number of methods have been proposed to estimate the number of nonsynonymous and synonymous substitutions per site between two homologous sequences (e.g., refs. 8–10). These approaches aim to estimate the ratio between the number of nonsynonymous (and synonymous) substitutions per codon and the number of nonsynonymous (and synonymous) mutations per codon. Simulations have shown that some of these methods provide reliable average estimates of the nonsynonymous/synonymous rate ratio (10). However, these approaches assume that amino acids in the sequence evolve under the same selection pressure. This assumption is unrealistic because only a few amino acid sites are found to be responsible for adaptive evolution in almost all proteins that evolve under positive selection (1, 2, 11). As a result, these methods have little power in detecting positive selection when analyzing real data sets (e.g., refs. 12–14).

Nielsen and Yang (2) described a codon-based model of nucleotide substitutions that allows the selection process to vary among sites. This model is very similar in its structure to the one proposed earlier by Goldman and Yang (15). These authors describe the substitutions at the codon level as a continuous-time Markov chain, with a state space on the 61 sense codons. Under such a model, a (codon) site evolves under purifying, neutral, or positive selection. The nonsynonymous/synonymous substitution rate ratio varies across these different classes of selection. For example, an amino acid position that is evolving under strong purifying selection should have a small nonsynonymous/synonymous substitution rate ratio, whereas a position under positive selection should exhibit a nonsynonymous/synonymous rate ratio of  $>1$ . The parameters of the model explored by Nielsen and Yang (2), and later by Yang *et al.* (16), include the nonsynonymous/synonymous rate ratio for each selection class of substitution, frequency parameters that are interpreted as the prior probability that a site falls into any specific selection category, and parameters such as the equilibrium codon frequencies and transition/transversion rate ratio that are intended to describe the vagaries of the substitution process beyond the most basic ones of selection pressures on nonsynonymous and synonymous substitutions. All of these parameters can be estimated from alignments of protein-coding DNA sequences by using either a maximum-likelihood or a fully Bayesian approach. The analysis of real data sets suggests that this approach is more efficient than pairwise-based methods for detecting positive selection at the protein level (e.g., ref. 14).

In the widely used codon models described by Nielsen and Yang (2), and by Yang *et al.* (16), the selection process does not vary along lineages of the phylogeny. This constraint is probably unrealistic because several studies have demonstrated that switches between selection processes, or variation of the selection intensity, are likely to occur (e.g., refs. 4, 6, and 17–19). Yang and Nielsen (20) proposed a test for positive selection at some branches of the phylogeny, while simultaneously allowing selection to vary among sites. However, under their model, branches with a different pattern of positive selection must be specified *a priori*. This model is especially useful when one wants to test for an association between speciation (19) or gene-duplication (20, 21) events and positive selection. Similarly, Forsberg and Christiansen (22) described a codon-based model of substitution that allows nonoverlapping subtrees of the phylogeny to evolve under different selection patterns. Again, the parts of the tree that undergo different selection regimes must be specified *a priori*. The authors demonstrated that this model is well suited for analyzing how selection acts on variants of a parasite's genes in different hosts.

Whereas earlier studies (4, 6, 17–23) demonstrate that switches between selection patterns across lineages are likely to occur, little is known about the rates at which these changes

This paper was submitted directly (Track II) to the PNAS office.

§To whom correspondence should be addressed. E-mail: johnh@biomail.ucsd.edu.

© 2004 by The National Academy of Sciences of the USA

happen. Indeed, current methods make it difficult to evaluate how frequently variable selection over time leaves an identifiable footprint in alignments of homologous protein-coding sequences. Moreover, the variation of selection processes during the course of evolution may show different patterns, depending on the site to be considered. For example, independent convergent/parallel evolution events might occur in distinct lineages at different sites. In this case, models that define lineages where molecular adaptation occurred *a priori* are not well suited.

In this paper, we generalize a codon-based model of DNA substitution to allow selection to change over time. Unlike previous approaches, our model does not constrain switches among selection categories to any particular lineage *a priori*. Instead, the switch between one selection regime and another at a given codon position is a stochastic process, mediated presumably by external forces. This approach is similar to the one followed by Tuffley and Steel (24) for modeling the site-specific variation of substitution rates. Our model has been implemented in the maximum-likelihood framework. Because the traditional model is a special case of our more general model, we could perform likelihood-ratio tests of the null hypothesis that the site-specific selection regime remains constant across lineages.

The analysis of eight HIV-1 *env* sequence data sets shows that our model provides a statistically significant better description of these data than does the traditional model. We also argue that the role played by negative selection is underestimated when site-specific selection process varies across lineages and is ignored. Finally, a site-by-site analysis demonstrates that our approach can be used to detect episodic adaptive events at individual amino acid position.

## Methods

**The Model.** Our codon-based substitution model combines two processes. As with other models [e.g., Nielsen and Yang (2)], we allow substitutions between codons to occur under one of three selection regimes (purifying, neutral, and positive selection). However, we also allow changes between selection classes (or switches) to occur according to a continuous-time Markov chain. In the following description, we will first describe the assumptions we make about the substitution process, and then detail the switching process. Finally, we will describe the combined substitution/switching process.

Substitutions between codons are modeled as a continuous-time Markov process. The instantaneous rate of change from one sense codon to another is described by a  $61 \times 61$  rate matrix. Following examples in earlier work (15), we parameterize the rate matrix as:

$$Q_x(ij) = \begin{cases} 0: & \text{if codons } i \text{ and } j \text{ differ at more} \\ & \text{than one nucleotide position} \\ \omega_x \pi_j: & \text{nonsynonymous transversion} \\ \pi_j: & \text{synonymous transversion} \\ \kappa \omega_x \pi_j: & \text{nonsynonymous transition} \\ \kappa \pi_j: & \text{synonymous transition} \end{cases},$$

where  $Q_x(ij)$  is the rate of change from codon  $i$  to codon  $j$  when sequences evolve under selection pattern  $x$ ,  $\omega_x$  is the nonsynonymous/synonymous rate ratio that characterizes selection process  $x$ , and  $\pi_j$  is the stationary frequency of codon  $j$ . For each  $i$  and  $j$ ,  $\pi_i$ ,  $\pi_j$ , and  $Q_x(ij)$  remain constant through time and  $\pi_j Q_x(ij) = \pi_i Q_x(ji)$ . The matrix  $\mathbf{Q}_x = \{Q_x(ij)\}$  therefore defines a stationary, homogeneous, and time-reversible Markov process of substitution between codons.

We consider two parameterizations of the substitution model. Both parameterizations have three classes of nonsynonymous/synonymous rate ratios. The first was described by Nielsen and Yang (2) and was later designated the “M2” model by Yang *et al.*

*al.* (16). Under this model, nonsynonymous mutations are strongly deleterious and eliminated under the first selection pattern ( $x = 1$ ), so that  $\omega_1 = 0$ . Rates of nonsynonymous and synonymous substitutions are equal under the second selection process ( $x = 2$ ), so that  $\omega_2 = 1$ . Nonsynonymous mutations provide a selective advantage under the third selection process ( $x = 3$ ), so that  $\omega_3 > 1$ . The second parameterization we consider was designated the “M3” model by Yang *et al.* (16). The nonsynonymous/synonymous rate ratio is less constrained under this model; the only constraint is that  $\omega_1 < \omega_2 < \omega_3$ .

Substitutions between selection classes are governed by a three-state continuous-time Markov process with rate matrix

$$\mathbf{R} = \delta \begin{pmatrix} -(p_2 + p_3\alpha) & p_2 & p_3\alpha \\ p_1 & -(p_1 + p_3\beta) & p_3\beta \\ p_1\alpha & p_2\beta & -(p_1\alpha + p_2\beta) \end{pmatrix}.$$

The element  $R(xy)$  on the  $x$ th row and  $y$ th column of  $\mathbf{R}$  corresponds to the rate of switching between selection processes  $x$  and  $y$ . The parameter  $\delta$  is the rate of interchange between selection patterns and  $p_x$  is the equilibrium frequency of selection process  $x$ . The  $\alpha$  and  $\beta$  parameters are two relative rate ratios with important biological meanings; values of  $\alpha > 1$  indicate that switches between negative and positive selection occur more frequently than switches between negative selection and a neutral process of molecular evolution. The comparison of these relative rates might be relevant. For example, Messier and Stewart (6) have shown that adaptive episodes (i.e., positive selection) during the evolution of primate lysozyme were followed by episodes of negative selection. A value of  $\alpha > 1$  and also greater than  $\beta$  would confirm this observation. For each  $x$  and  $y$ ,  $p_x$ ,  $p_y$ , and  $R(xy)$  remain constant through time, and  $p_x R(xy) = p_y R(yx)$ . Hence, the process describing switches between selection classes follows a stationary, homogeneous, and time-reversible stochastic process.

The combined substitution and switching process may be formulated in terms of a single continuous-time Markov chain with stationary distribution  $(p_1 \pi_1, \dots, p_1 \pi_{61}, p_2 \pi_1, \dots, p_2 \pi_{61}, p_3 \pi_1, \dots, p_3 \pi_{61})$ , and rate matrix

$$\mathbf{S} = \begin{pmatrix} \mathbf{Q}_1 & 0 & 0 \\ 0 & \mathbf{Q}_2 & 0 \\ 0 & 0 & \mathbf{Q}_3 \end{pmatrix} + \delta \begin{pmatrix} -(p_2 + p_3\alpha)\mathbf{I} & p_2\mathbf{I} & p_3\alpha\mathbf{I} \\ p_1\mathbf{I} & -(p_1 + p_3\beta)\mathbf{I} & p_3\beta\mathbf{I} \\ p_1\alpha\mathbf{I} & p_2\beta\mathbf{I} & -(p_1\alpha + p_2\beta)\mathbf{I} \end{pmatrix},$$

where  $\mathbf{I}$  denotes the  $61 \times 61$  identity matrix. It is straightforward to demonstrate that  $\mathbf{S}$  is stationary, time-reversible, and homogeneous whenever  $\mathbf{Q}_1$ ,  $\mathbf{Q}_2$ ,  $\mathbf{Q}_3$ , and  $\mathbf{R}$  are. The matrix  $\mathbf{S}$  is scaled so that time is measured by an expected number of codon substitutions per amino acid position. The transition probabilities from state  $i$  to state  $j$  on a branch of length  $v$  are contained in a  $183 \times 183$  matrix designated  $\mathbf{P}_v$ . The matrix  $\mathbf{P}_v$  is calculated by using the following matrix exponentiation:  $\mathbf{P}_v = e^{\mathbf{S}v}$ . The probability of observing each site pattern (given a phylogenetic tree and the values of the parameters) under the combined substitution/switch process can be calculated by using Felsenstein's pruning algorithm (25).

Note that the model described here has only three additional parameters ( $\delta$ ,  $\alpha$ , and  $\beta$ ) as compared with the traditional model. Moreover, the models are nested: the traditional model arises as the switching rate tends to zero ( $\delta \rightarrow 0$ ). This nesting of the models allows us to apply likelihood ratio tests of the following null hypotheses:  $H_1$ , there is no switching among selection categories; and  $H_2$ , there is no bias in the switching pattern among selection categories. The likelihood ratio test statistic,



minus twice the difference in the log likelihoods under the null hypothesis and the more general alternative hypothesis, asymptotically follows a 50:50 mixture of  $\chi_0^2$  and  $\chi_1^2$  distributions for  $H_1$  and a  $\chi_2^2$  distribution for  $H_2$  [see Self and Liang (26)].

It is also worthwhile to note that the switching model of codon substitution is identical in its structure to the covarion model described by Tuffley and Steel (24). Whereas a covarion process takes into account site-specific variation of substitution rates in a phylogenetic tree, our model describes site-specific variation of selection forces along lineages.

Codon models with switches among categories are designated with “+S1” when  $\alpha = \beta = 1$  or with “+S2” when  $\alpha$  and  $\beta$  are free to vary. For example, “M2+S1” designates the model of codon substitution described by Nielsen and Yang (2), allowing switching among categories under the constraint that  $\alpha = \beta = 1$ , and “M3+S2” designates the M3 model first described by Yang *et al.* (16), while allowing for a potentially biased pattern of switching among selection categories. “M2+S” designates any switching model derived from M2. “M3+S” is defined in a similar manner.

**Detection of Positively Selected Sites.** Under M2 or M3, the posterior probability of a given selection class at some site can be estimated by using either an empirical Bayesian approach (2) or a fully Bayesian one (27). The category that maximizes the posterior probability is the most likely selection process to have acted at the corresponding site. For switching models, one cannot rely on a site remaining in any particular selection category through time. Instead, we calculate the expected fraction of time that selection process spends in a particular class. Therefore, these models can be used to detect sites in the alignment where positive selection is likely to have occurred in most of the lineages.

Let  $d_z(v, x, y)$  be the amount of time that the process dwells on selection regime  $z$  on a branch of length  $v$ , with  $x$  and  $y$  being the selection patterns at both ends of this branch. We have

$$E[d_z(v, x, y)] = \int_0^v \frac{p_{xz}(s)p_{zy}(v-s)}{p_{xy}(v)} ds,$$

where  $p_{xz}(s)$  is the probability of change from selection process  $x$  to  $z$  after  $s$  codon substitution events;  $p_{zy}(v-s)$  and  $p_{xy}(v)$  are defined in a similar manner. These probabilities can be derived from  $\mathbf{P}_v$ , or from the calculation of the matrix  $e^{\mathbf{R}v}$ . Conditional on the selection class  $x$  at one end of a branch of length  $v$  and class  $y$  at the other end, the probability of finding the switch process in state  $z$  is then  $Pr(z|x, y, v) = E[d_z(v, x, y)]/v$ .

Of course, one cannot be certain that the selection regimes  $x$  and  $y$  really occur at either end of the branch. Therefore, the expected frequency of the selection regime  $z$  is a weighted average over all possibilities of states  $x$  and  $y$  at either end of the branch, conditional on the observed codon states at the tips of the tree.

By using this approach, we are able to calculate two important quantities: (i) the expected frequency of the positive selection regime on a branch-by-branch basis, for each site of the alignment and (ii) the probability that a site is under positive selection over the entire phylogenetic history of the group by taking the expected time of occurrence of the positive selection pattern for the entire tree and dividing that number by the expected number of codon substitutions at this site (i.e., the sum of the branch lengths).

**Data.** We analyzed partial *env* sequences from HIV-1 (C2-V5) that had been obtained in a longitudinal study (28). These data sets are derived from samples that were collected on average every 8 months from eight infected patients. Sequences were aligned collectively by using CLUSTALX (29) and were then

manually adjusted within subjects. Gaps were removed in a balanced manner to preserve codon alignment (30). An earlier study (4) of the eight homologous sequence alignments using traditional codon models of substitution showed that the broad genetic diversity of HIV-1 in infected individuals is a consequence of site-specific positive selection, a likely consequence of immune recognition.

**Parameter Estimation.** An initial phylogeny was estimated for each of the eight data sets by using maximum likelihood under the general time-reversible model of nucleotide substitution (a model that allows four states) with the program PHYML (31). Among-site rate variation was modeled by using a discrete gamma distribution (32).

The tree topologies were considered fixed during the estimation of the codon model parameters. The equilibrium frequencies of codons ( $\pi_i$ ) were estimated by using the observed frequencies of the nucleotides at the three codon positions. The other parameters of the model; branch lengths, transition/transversion rate ratio ( $\kappa$ ), equilibrium frequencies of the selection classes ( $p_x$ s), switching parameters ( $\delta$ ,  $\alpha$ , and  $\beta$ ), and nonsynonymous/synonymous rate ratios ( $\omega$ s), were estimated by means of maximum likelihood by using a program written by S.G., which is available on request.

## Results

**Likelihood Analysis.** Table 1 shows the maximum values of the log likelihood ( $\ln L$ ), nonsynonymous/synonymous rate ratios ( $\omega_1$ ,  $\omega_2$ , and  $\omega_3$ ) and equilibrium frequencies of the three selection classes ( $p_1$ ,  $p_2$ , and  $p_3$ ) obtained under the six models examined in this study, for the eight HIV-1 *env* data sets. We first tested the null hypothesis that switching between selection categories did not occur during the history of these sequences.

Under the hypothesis that M2 describes perfectly the substitution process, twice the difference of log likelihoods obtained under M2 and M2+S1 asymptotically follows a 50:50 mixture of  $\chi_0^2$  and  $\chi_1^2$  distributions (see *Methods*). The same holds true for the comparison M3 vs. M3+S1. For both the M2 and M3 models, the null hypothesis that the site-specific selection pattern remains constant across lineages is rejected for each of the eight data sets. The smallest difference is obtained with the patient 2 data set by comparing the log likelihood obtained under M3 and M3+S1. The probability of such a difference or more if sequences evolved under M3 is  $p \approx 1.48 \times 10^{-4}$ . Therefore, the site-specific variation of selection regime is likely to have played an important role during the evolution of the sequences analyzed in this study.

We now examine the second null hypothesis posed above, that the pattern of switching from one category to another is not strongly biased (i.e., we test the null hypothesis:  $\alpha = \beta = 1$ ). Here, we compare the simple switching model S1 with the S2 model, which allows  $\alpha$  and  $\beta$  to freely vary. Under the hypothesis that sequences evolved under M2+S1, twice the difference of log likelihood obtained under M2+S1 and M2+S2 now asymptotically follows a  $\chi_2^2$  distribution. This finding also holds for the comparison M3+S1 vs. M3+S2. The largest differences in log likelihood values are observed in patient 8 by comparing M2+S1 vs. M2+S2 and M3+S1 vs. M3+S2, and in patient 7, by comparing M2+S1 vs. M2+S2. These three differences are statistically significant at the 5% level. However, these are the only cases where switches between selection patterns are likely to be biased. For all of the other combinations of data sets and switching models analyzed here, we cannot reject the null hypothesis that  $\alpha = \beta = 1$ .

It is also interesting to note that the differences of log likelihood obtained under M2 and M3 decrease when switches between selection regimes are taken into account. Under the hypothesis that sequences evolved under M2, twice the differ-

Table 1. Likelihood analysis of eight HIV-1 *env* gene sequence data sets

	M2	M2+S1	M2+S2	M3	M3+S1	M3+S2
P1						
lnL	-3,050.46	-3,021.78	-3,019.93	-3,036.87	-3,021.15	-3,019.13
$\omega_1 \omega_2 \omega_3$	0.00 1.00 8.31	0.00 1.00 9.40	0.00 1.00 10.01	0.15 1.22 7.50	0.04 0.91 8.62	0.04 0.71 9.43
$p_1 p_2 p_3$	0.39 0.56 0.04	0.67 0.29 0.05	0.64 0.32 0.05	0.70 0.26 0.03	0.69 0.26 0.05	0.60 0.35 0.05
P2						
lnL	-3,672.49	-3,652.61	-3,651.67	-3,658.85	-3,652.30	-3,651.23
$\omega_1 \omega_2 \omega_3$	0.00 1.00 4.39	0.00 1.00 3.86	0.00 1.00 4.47	0.15 1.14 3.85	0.06 1.36 4.23	0.03 0.49 3.98
$p_1 p_2 p_3$	0.30 0.62 0.07	0.57 0.33 0.10	0.55 0.38 0.08	0.58 0.37 0.06	0.65 0.28 0.07	0.46 0.42 0.13
P3						
lnL	-3,205.90	-3,171.99	-3,169.07	-3,184.05	-3,165.13	-3,162.90
$\omega_1 \omega_2 \omega_3$	0.00 1.00 5.20	0.00 1.00 5.07	0.00 1.00 14.17	0.19 2.10 5.95	0.00 2.92 9.99	0.00 2.83 13.82
$p_1 p_2 p_3$	0.36 0.49 0.15	0.71 0.15 0.14	0.75 0.20 0.05	0.73 0.22 0.05	0.78 0.18 0.03	0.79 0.19 0.02
P5						
lnL	-3,889.82	-3,819.30	-3,817.56	-3,838.40	-3,816.79	-3,815.98
$\omega_1 \omega_2 \omega_3$	0.00 1.00 11.88	0.00 1.00 10.01	0.00 1.00 10.44	0.14 1.04 7.34	0.05 1.71 11.51	0.05 1.39 10.80
$p_1 p_2 p_3$	0.35 0.62 0.04	0.73 0.23 0.03	0.71 0.26 0.03	0.77 0.20 0.04	0.84 0.14 0.02	0.79 0.18 0.03
P6						
lnL	-3,000.11	-2,951.67	-2,950.86	-2,977.53	-2,951.03	-2,950.77
$\omega_1 \omega_2 \omega_3$	0.00 1.00 4.33	0.00 1.00 5.95	0.00 1.00 6.43	0.08 0.84 2.93	0.00 1.23 6.22	0.00 1.05 6.24
$p_1 p_2 p_3$	0.51 0.43 0.06	0.80 0.15 0.05	0.81 0.14 0.05	0.71 0.21 0.08	0.82 0.13 0.05	0.81 0.14 0.05
P7						
lnL	-4,121.97	-4,060.46	-4,057.37	-4,084.47	-4,050.26	-4,049.37
$\omega_1 \omega_2 \omega_3$	0.00 1.00 8.40	0.00 1.00 11.61	0.00 1.00 11.81	0.32 2.70 11.84	0.19 3.29 14.56	0.17 3.07 15.09
$p_1 p_2 p_3$	0.25 0.63 0.12	0.61 0.32 0.07	0.58 0.35 0.07	0.79 0.17 0.04	0.83 0.13 0.04	0.81 0.14 0.05
P8						
lnL	-4,174.14	-4,098.80	-4,092.67	-4,136.79	-4,095.89	-4,090.22
$\omega_1 \omega_2 \omega_3$	0.00 1.00 5.34	0.00 1.00 9.20	0.00 1.00 15.05	0.10 1.03 4.17	0.03 1.41 9.93	0.05 1.06 14.85
$p_1 p_2 p_3$	0.38 0.53 0.09	0.68 0.27 0.05	0.68 0.29 0.03	0.64 0.28 0.07	0.74 0.22 0.04	0.71 0.26 0.03
P9						
lnL	-2,825.52	-2,792.29	-2,792.29	-2,805.91	-2,785.20	-2,784.66
$\omega_1 \omega_2 \omega_3$	0.00 1.00 7.23	0.00 1.00 8.17	0.00 1.00 8.16	0.22 1.78 7.56	0.12 2.07 9.33	0.12 2.03 9.22
$p_1 p_2 p_3$	0.24 0.63 0.12	0.56 0.35 0.10	0.56 0.35 0.10	0.61 0.31 0.08	0.72 0.20 0.08	0.72 0.19 0.09

M2, M2+S1, M2+S2, M3, M3+S1, and M3+S2 denote the codon-based models of substitution compared in this analysis (see text). lnL stands for log likelihood.  $\omega_1$ ,  $\omega_2$ , and  $\omega_3$  correspond to the nonsynonymous/synonymous rate ratio that characterize a negative, a (strictly or nearly) neutral, and a positive selection regime, respectively.  $p_1$ ,  $p_2$ , and  $p_3$  are the corresponding equilibrium frequencies of these three selection patterns.

ence of log likelihood obtained under M2 and M3 asymptotically follows a  $\chi^2$  distribution. This finding also holds true for the comparison M2+S1 vs. M3+S1 and M2+S2 vs. M3+S2. The null hypothesis is rejected at the 1% level for each of the eight data sets when comparing M2 with M3. This is no longer the case when comparing M2+S1 with M3+S1: the null hypothesis is rejected at the 5% level for only three data sets (P3, P7, and P9). The comparison M2+S2 vs. M3+S2 leads to the same conclusions. Hence, the description of the substitution process given by M2 is often not significantly worse than the one given by M3 if switches between selection patterns at individual sites are modeled. From a statistical point of view, it is satisfactory to note that, in most cases, taking into account switches between selection regimes allows one to put a limit on the number of free rate parameters that are needed to describe the distribution of nonsynonymous/synonymous rate ratios. The comparison of log likelihoods obtained under M2+S1 and M3 is perhaps even more convincing: M2+S1 (five rate free parameters) provides a better description of the data than M3 (six free rate parameters) for the eight data sets analyzed in this study.

**Parameter Estimates.** Estimates of  $\omega_3$  obtained under M3+S are greater than those obtained under M3 for every data set analyzed. A similar tendency is observed with M2+S vs. M2. This result is expected because M2 and M3 define  $\omega_3$  as a site-specific rate ratio. Therefore, the value of this parameter corresponds to an average of different nonsynonymous/synonymous rate ratios

over lineages. This averaging effect is less important with M2+S and M3+S because these models accommodate site-specific variation in the nonsynonymous/synonymous rate ratio. This result shows that the strength of positive selection that acts on the eight HIV-1 *env* sequences is likely to be underestimated when the site-specific variation of selection processes across lineages is not taken into account.

Note also that the negative selection and neutral process equilibrium frequency estimates ( $p_1$  and  $p_2$ ) vary greatly from M2 to M2+S, with  $p_1$  being always smaller under M2 than under M2+S. Sites at which only a few nonsynonymous substitutions occurred as compared with synonymous substitutions provide an explanation. Under M2, the probability for such sites to have been generated under a negative selection process is low because of the presence of nonsynonymous substitutions. Under M2+S, these sites have a higher probability of having been generated under a negative selection regime if the nonsynonymous substitutions are clumped. Such sites are also well described by models that allow intermediate values of the nonsynonymous/synonymous rate ratio (such as M3). This observation explains why the equilibrium frequencies of the different selection regimes obtained under M3 and M3+S are similar (Table 1). Note, however, that estimates of the frequency of negative selection given by M3 are still almost systematically smaller than those given by M3+S. Therefore, even if the tendency to underestimate this parameter is much stronger under M2, estimates obtained under M3 are likely to be biased, too, because the

site-specific variation of selection patterns is not taken into account.

**Site-By-Site Analysis.** Codon-based models of substitution provide a suitable framework for the probabilistic detection of positively selected amino acid positions (2, 27). The calculation of the posterior probability of the positive selection regime at each site is straightforward if the substitution model does not allow switches between selection patterns at individual sites. If the model allows such switches, the expected frequency of the positive selection process given the data are based on the calculation of the expected time the substitution process dwells in positive selection (see *Methods*).

When considering the eight HIV-1 *env* data sets together, 18 sites evolved under positive selection according to M3 but not M3+S1. Among these sites, four are of particular interest because they are associated with posterior probability >0.95 under M3, while positive selection is not the most likely regime to have occurred at these sites according to M3+S1.

For each of these four sites, ancestral codon states were inferred by using a joint maximum-likelihood approach (33). This method relies on a stochastic model of sequence evolution and the ancestral states were inferred under M3+S1 (M3 gave very similar results). The branches where substitutions have probably occurred in the corresponding phylogenies were then identified from the comparison of ancestral codon states between two adjacent nodes. The great majority of substitutions inferred from the four sites at which M3 and M3+S1 strongly disagree are likely to be nonsynonymous substitutions. These substitutions are generally clumped on a few branches of the phylogeny instead of being scattered on the whole tree.

Fig. 1 illustrates such an uneven distribution. The positions of the substitutions were inferred from two conflicting sites found in the data set of patient 6. Under the assumption that only one substitution occurred on branches with distinct ancestral codon states at their two ends, substitutions inferred at these two sites are all nonsynonymous substitutions. Their estimated positions are clearly not homogeneously distributed on the phylogeny. Indeed, these substitutions did not occur at the beginning and at the end of the infection for this patient. The conflict between M3 and M3+S1 is best explained by this rather obvious variation of substitution processes across lineages.

It is also striking that substitutions are located here on the same closely related lineages of the tree, regardless of the site to be considered. A detailed investigation of the role played by these substitutions is outside the scope of this paper. However, it is worth noting that the two sites that show this peculiar pattern of substitutions are located in HIV T-helper epitope regions (34). These observations suggest that the substitutions inferred at the sites where M3 and M3+S1 strongly disagree may be of biological interest.

## Conclusions

We introduce a codon-based model of nucleotide substitution that accommodates variation of natural selection processes across a tree for protein-coding DNA sequences. More precisely, site-specific switches between different values of the nonsynonymous/synonymous rate ratio are explicitly taken into account. This model extends the widely used traditional model of substitution between codons (2, 16). Because the traditional model is a special case of our model, their fit to the data can be compared in a rigorous statistical framework by using likelihood ratio tests.

Whereas the simplest switching model has only one additional parameter as compared with the standard model, the increase in the likelihood is noticeable and statistically significant for each of the eight HIV-1 *env* gene sequence data sets to be considered. Hence, switches between selection patterns at individual sites are an important evolutionary feature here. This feature is recov-



**Fig. 1.** Phylogenetic positions of substitutions inferred at two amino acid sites of patient 6 data set. M3 strongly supports the hypothesis that sequences evolved under positive selection at these sites, whereas the statistical support given by M3+S1 to the same hypothesis is less important. ★ and ● correspond to the substitutions inferred at sites 41 and 180, respectively. All of these substitutions are likely to be nonsynonymous. The leaves of the tree are labeled with the rank of the corresponding sample time (1 is the earliest sample and 10 is the latest). The position of the root was determined by using outgroup sequences collected during the earliest stages of the infection.

ered under two different hypotheses concerning the processes of substitution between codons (i.e., M2 and M3). This result demonstrates, at least for these data sets, the robustness of our model.

Yang and Nielsen (20) pointed out that "... averaging (the nonsynonymous/synonymous rate ratio) over sites is a more serious problem than averaging over lineages" and showed convincing results that confirm this assertion. However, our results indicate that ignoring the site-specific variation of selection processes may result in an underestimation of the strength of positive selection. Note that this trend should not change drastically the results of analyses that aim to identify sites that evolve under positive selection. However, more accurate estimates to measure the intensity of positive selection are obviously of great interest for deciphering the processes of molecular evolution.

M2 and M2+S models largely agree on the frequency with which positive selection acts on the HIV-1 *env* sequences. However, they clearly diverge when considering the negative selection ( $\omega_1 = 0$ ) and neutral process ( $\omega_2 = 1$ ) equilibrium



frequencies. Indeed, because the M2 model does not account for the distribution of nonsynonymous and synonymous substitutions on the phylogeny, it tends to overestimate the expected frequency of the neutral process category. This discrepancy between estimates leads to a different interpretation of the role played by purifying selection in HIV-1 *env* sequences. According to M2, these sequences mainly evolve under a neutral process, whereas the most common evolutionary force is negative selection according to M2+S. Therefore, taking into account the site-specific variation of selection processes across lineages leads to a dramatically different assessment of the roles played by distinct selection regimes.

However, it must be stressed that  $\omega_1 = 0$  and  $\omega_2 = 1$  correspond to very stringent definitions of negative selection and neutrality, respectively. M3 is a more flexible model because its three selection classes correspond to strong purifying selection ( $0 < \omega_1 \ll 1$ ), weak purifying or diversifying selection ( $\omega_2 \approx 1$ ), or strong positive selection ( $\omega_3 > 1$ ). The first class of this model is well suited for describing sites at which a few nonsynonymous substitutions occurred. As a result, differences between equilibrium frequencies estimated under M3 and M3+S are smaller than those observed when comparing M2 and M2+S (Table 1). Nonetheless, it is worth to keep in mind that any model that does not allow site-specific switches between selection regimes will underestimate the role played by negative selection in the presence of sites at which nonsynonymous substitutions are clustered due to episodic adaptive events.

We would also like to warn readers against another potential pitfall related to the use of M2. This model poorly describes the amino acid positions where the number of synonymous substi-

tutions is much larger than the number of nonsynonymous changes. Because these sites are better described by M2+S, the difference of log likelihood obtained under the two models will frequently lead to reject the hypothesis of a constant site-specific selection pattern. However, these sites could actually have been generated under a constant site-specific selection regime with a value of the nonsynonymous/synonymous rate ratio intermediate between 0 and 1. In this case, the log likelihoods obtained under M3 and M3+S will be similar. Hence, it is much safer to test the hypothesis of the constancy of the site-specific selection process by comparing M3 vs. M3+S instead of M2 vs. M2+S.

Probabilistic-based approaches have been developed during the last decades to identify codon positions that are likely to have evolved under diversifying selection. In this context, detecting site-specific variations of the intensity of positive selection or changes in selection regimes is worthwhile. A site-by-site analysis of the eight HIV-1 data sets actually reveals positions at which switches between selection processes have probably occurred. Hence, the model described in this paper is also suited to investigate switches between selection regimes at the single-site level.

**Software Availability.** A program that implements the models described in this paper is available on request from S.G.

We thank G. Ewing, N. Galtier, E. Kassardjian, P. Meintjes, H. Philippe, H. Ross, S. Plön, M. Steel, D. Welch, and two anonymous reviewers for their suggestions. This work was partially supported by grants from the U.S. Public Health Service, a postdoctoral fellowship from the Allan Wilson Centre for Molecular Ecology and Evolution (to S.G.), and National Institutes of Health Grant GM-69801 (to J.P.H.).

- Hughes, A. & Nei, M. (1988) *Nature* **335**, 167–170.
- Nielsen, R. & Yang, Z. (1998) *Genetics* **148**, 929–936.
- Haydon, D., Bastos, A., Knowles, N. & Samuel, A. (2001) *Genetics* **157**, 7–15.
- Ross, H. & Rodrigo, A. (2002) *J. Virol.* **76**, 11715–11720.
- Lee, Y., Ota, T. & Vaquier, V. (1995) *Mol. Biol. Evol.* **12**, 231–238.
- Messier, W. & Stewart, C.-B. (1997) *Nature* **385**, 151–154.
- Swanson, W., Yang, Z., Wolfner, M. & Aquadro, C. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 2509–2514.
- Li, W.-H., Wu, C.-I. & Luo, C.-C. (1985) *Mol. Biol. Evol.* **2**, 150–174.
- Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3**, 418–426.
- Yang, Z. & Nielsen, R. (2000) *Mol. Biol. Evol.* **17**, 32–43.
- Yokoyama, S. & Yokoyama, R. (1996) *Annu. Rev. Ecol. Syst.* **27**, 543–567.
- Sharp, P. (1997) *Nature* **385**, 111–112.
- Crandall, K., Kelsey, C., Imamichi, H., Lane, H. & Salzman, N. (1999) *Mol. Biol. Evol.* **16**, 372–382.
- Bielawski, J. & Yang, Z. (2001) *Mol. Biol. Evol.* **18**, 523–529.
- Goldman, N. & Yang, Z. (1994) *Mol. Biol. Evol.* **11**, 725–736.
- Yang, Z., Nielsen, R., Goldman, N. & Krabbe Pedersen, A.-M. (2000) *Genetics* **155**, 431–449.
- Ohta, T. (1993) *Genetics* **134**, 1271–1276.
- Bailly, X., Leroy, R., Carney, S., Collin, O., Zal, F., Toulmond, A. & Jollivet, D. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 5885–5890.
- Clark, A., Gnanowski, S., Nielsen, R., Thomas, P., Kejariwal, A., Todd, M., Tanenbaum, D., Civello, D., Lu, F., Murphy, B., et al. (2003) *Science* **302**, 1960–1963.
- Yang, Z. & Nielsen, R. (2002) *Mol. Biol. Evol.* **19**, 908–917.
- Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 13413–13417.
- Forsberg, R. & Christiansen, F. (2003) *Mol. Biol. Evol.* **20**, 1252–1259.
- Bazykin, G., Kondrashov, F., Ogurtsov, A., Sunyaev, S. & Kondrashov, A. (2004) *Nature* **3**, 558–562.
- Tuffley, C. & Steel, M. (1998) *Math. Biosci.* **147**, 63–91.
- Felsenstein, J. (1981) *J. Mol. Evol.* **17**, 368–376.
- Self, S. & Liang, K. (1987) *J. Am. Stat. Assoc.* **82**, 605–610.
- Huelsenbeck, J. & Dyer, K. (2004) *J. Mol. Evol.* **58**, 661–672.
- Shankarappa, R., Margolick, J., Gange, S., Rodrigo, A., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C., Learn, G., He, X., et al. (1999) *J. Virol.* **73**, 10489–10502.
- Thompson, J., Gibson, T., Plewniak, F., Jeanmougin, F. & Higgins, D. (1997) *Nucleic Acids Res.* **24**, 4876–4882.
- Rodrigo, A., Hanley, E., Goracke, P. & Learn, G. (2001) in *Computational and Evolutionary Analysis of HIV Molecular Sequences*, eds. Rodrigo, A. & Learn, G. (Kluwer, Boston), pp. 1–17.
- Guindon, S. & Gascuel, O. (2003) *Syst. Biol.* **52**, 696–704.
- Yang, Z. (1994) *J. Mol. Evol.* **39**, 306–314.
- Koshi, J. & Goldstein, R. (1996) *J. Mol. Evol.* **42**, 313–320.
- Korber, B., Brander, C., Haynes, B., Koup, R., Kuiken, C., Moore, J., Walker, B. & Watkins, D. (2002) *HIV Molecular Immunology Database 2002* (Los Alamos Natl. Lab., Theoretical Biology and Biophysics, Los Alamos, NM).