Genealogical Evidence for Positive Selection in the *nef* Gene of HIV-1

Paolo M. de A. Zanotto,* Esper G. Kallas,† Robson F. de Souza* and Edward C. Holmes‡

**Bioinformatics and Retrovirology Laboratory and †Laboratory of Immunology DIPA–Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, CEP 05508-900, Brazil and ‡The Wellcome Trust Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom*

Manuscript received March 16, 1999
Accepted for publication July 26, 1999

ABSTRACT

The pattern and process of evolution in the *nef* gene of HIV-1 was analyzed within and among patients. Using a maximum likelihood method that allows for variable intensity of selection pressure among codons, strong positive selection was detected in a hemophiliac patient over 30 mo of infection. By reconstructing the process of allele substitution in this patient using parsimony, the synapomorphic amino acid changes separating each time point were found to have high probabilities of being under positive selection, with selective coefficients of at least 3.6%. Positive selection was also detected among 39 *nef* sequences from HIV-1 subtype B. In contrast, multiple pairwise comparisons of nonsynonymous and synonymous substitution rates provided no good evidence for positive selection and sliding window analyses failed to detect most positively selected sites. These findings demonstrate that positive selection is an important determinant of *nef* gene evolution and that genealogy-based methods outperform pairwise methods in the detection of adaptive evolution. Mapping the locations of positively selected sites may also be of use in identifying targets of the immune response and hence aid vaccine design.

THE nature of the evolutionary interaction between the human immunodeficiency virus (HIV) and the human immune system has been the source of much debate, and increasingly so given the desire to understand how and why resistance appears to combinations of antiviral drugs (Leigh Brown and Richman 1997). To some it is a system governed by chance, starting with the random activation of HIV-infected cells by foreign antigens (Wain-Hobson 1994, 1996), followed by a process of allele substitution dominated by genetic drift (Leigh Brown 1997; Plikat *et al.* 1997). To others, the immune-driven positive selection of advantageous viral mutants plays the pivotal role, such that the process of within-host viral evolution is characterized by the successive appearance of escape mutants that evade the prevailing immune response (Nowak *et al.* 1996; McMichael and Phillips 1997; McMichael 1998).

Evidence for the importance of natural selection in HIV evolution comes from studies of both host and virus. On the host side it is well established that the immune response against HIV infection is mainly orchestrated by T lymphocytes, among which the cytotoxic T CD8+ cells (CTLs) play a vital role in recognizing epitopes presented by MHC class I molecules. The importance of CTLs can be inferred from the correlation between CTL activity and the control of HIV-1 viral load, with long-term nonprogressors to AIDS having particularly strong CTL responses (Musey *et al.* 1997; Ogg *et al.* 1998). More recently it was also observed that levels of viremia increased in HIV-1-infected rhesus monkeys following the removal of CD8+ lymphocytes (Schmitz *et al.* 1999). T helper (CD4+) lymphocytes have also been linked to the control of HIV viral load, and possibly influence the entire cellular and humoral immune response (Rosenberg *et al.* 1997), perhaps by enabling antigen-presenting cells to mount a stronger CTL response (Bennett *et al.* 1998; Ridge *et al.* 1998; Schoenberger *et al.* 1998).

On the virus side there is equally compelling evidence that HIV-1 is able to escape CTL recognition during infection. Several reports suggest that HIV-1 can respond to the selective pressure imposed by CTLs by fixing amino acid point mutations or deletions (Koenig *et al.* 1995; Borrow *et al.* 1997; Goulder *et al.* 1997; Price *et al.* 1997). However, the characterization of amino acid changes related to CTL escape is complex and, aside from their appearance, there is often little direct evidence that they are selectively favored.

The controversy over evolutionary mechanism is perhaps most evident with respect to *nef*, a pleiotropic gene that encodes a transactivating factor (p27), and which may reduce or increase viral replication depending on cell type (Welker *et al.* 1996; Levy 1998). Deletions in *nef* have been shown to lessen the pathogenic effects of HIV both in monkeys (Kestler *et al.* 1991) and perhaps in humans (Kirchhoff *et al.* 1995) and *nef*-deleted HIV strains have been utilized as vaccine candidates (for a review see Levy 1998). Evidence that *nef* might be sub-

Corresponding author: Paolo M. de A. Zanotto, Bioinformatics and Retrovirology Laboratory, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, CEP 05508-900, Brazil.
E-mail: pzanotto@usp.br

ject to positive selection comes from the demonstration, within a single patient, of a CTL response followed by escape mutants in a HLA B8 epitope (Price *et al.* 1997). Significantly, this sequence also showed a higher rate of nonsynonymous (d_N) to synonymous (d_S) substitution per site, an observation that is often given as evidence for positive selection (Sharp 1997). In contrast, a longitudinal study of *nef* gene evolution in another individual was claimed to provide no evidence for adaptive evolution as d_N/d_S ratios near 1.0 were observed (Plikat *et al.* 1997). Although this was taken to mean that all amino acid changes are neutral and therefore fixed by genetic drift alone, in reality it seems more likely that $d_N/d_S = 1.0$ in this case reflects the interplay of both positive and negative selection pressures (Holmes and Zanotto 1998). An analysis of *nef* gene sequences from different subtypes of HIV-1 likewise provided no evidence for positive selection (with $d_N < d_S$) at epitopes for CTLs, T-helper cells, or monoclonal antibodies, with constraints against amino acid change particularly strong within CTL epitopes (da Silva and Hughes 1998).

Such contrasting observations highlight the need to undertake more detailed investigations of the evolutionary mechanisms shaping genetic diversity in *nef*. In particular, because the commonly used pairwise methods for estimating d_N and d_S do not take full account of the genealogical information in data, and so are liable to nonindependence and pseudoreplication, it is important to test theories of evolutionary mechanism using an explicitly phylogenetic approach. Equally, it will be of value to consider genetic variation in *nef* in population genetic terms, as estimates of the rate of allele fixation and the selection coefficient of any favorable allele will be important given the possible use of *nef* in future HIV vaccines.

Herein we present a detailed examination of the evolutionary processes acting on the *nef* gene of HIV-1. We will first analyze the data set of Plikat *et al.* (1997), as well as more divergent HIV isolates, using a genealogical and likelihood-based approach for detecting positive selection (Nielsen and Yang 1998). We will then develop a cladistic model of allele substitution under positive selection from which we are able to estimate important parameters of *nef* gene evolution.

MATERIALS AND METHODS

Patient material and primary data: The analysis described in this article used four complete *nef* gene data sets. The first comprised 48 sequences from a hemophiliac infected by a contaminated batch of factor IX (Plikat *et al.* 1997). These sequences were obtained from proviral DNA amplified by PCR followed by cloning. Sequences were obtained from three time points—11, 25, and 41 mo postinfection—with CD4+ cell counts of 1204, 922, and 912, respectively. The patient received no antiviral therapy during this time. A single sequence showing a frameshift (isolate 25U52490) was excluded from the analysis. This resulted in a total of 47 nonidentical sequences of 618 bp: 15 sequences from time point 11, 16 from time point 25, and 16 from time point 41. Accession numbers for each sequence can be found in the original publication.

To examine the evolutionary process among more divergent HIV-1 isolates, three other sets of *nef* sequences were analyzed. The first contained 39 sequences of subtype B (606 bp), a viral clade of mainly European and North American origin. The second data set comprised 10 sequences (621 bp) from the larger M (main) group, thereby incorporating more divergent viral sequences from varied geographical origins—in this case three sequences from subtype A, three from B, three from D, and one from U (unassigned). The final data set contained 11 sequences (585 bp), including 9 group M sequences, one group O (outlier) sequence, and the *nef* gene sequence from a chimpanzee virus (SIV_{CPZ}), and so covering the deepest parts of the HIV-1 tree. All these sequences were collected from the 1997 release of the Los Alamos HIV database (Korber *et al.* 1997b).

Sequence alignment and phylogenetic analysis: All four *nef* data sets were aligned by hand and checked using the CLUSTALW program (Thompson *et al.* 1994). Alignments are available from the authors on request.

The phylogenetic relationships among sequences from each of these four data sets (in the hemophiliac patient the data from each time point were analyzed separately and in combination) were then reconstructed using a maximum likelihood method. The HKY85 model of nucleotide substitution was used in all cases with optimal values for the transition to transversion ratio and the shape parameter (α) of a gamma distribution of rate variation among sites (with eight categories), both determined during tree reconstruction. These parameter values are given in Tables 1 and 2. Finally, to determine the level of support for each node, 1000 bootstrap resamplings of the data were generated on neighbor-joining trees, although utilizing the maximum likelihood substitution model. All analyses were performed using the 4.0d64 test version of PAUP* kindly provided by David L. Swofford.

Analysis of selection pressures: Three maximum likelihood models were used to analyze the evolutionary processes acting on *nef*, all of which utilize gene genealogies and consider the codon, instead of the nucleotide, as the unit of evolution. The first, "invariant" model (Goldman and Yang 1994) assumes that all codons fall into a single category of sites with a fixed value of d_N/d_S—the parameter ω. The second "neutral" model allows two categories of sites (Nielsen and Yang 1998). The first category represents strictly neutral sites (p_1) that have a fixed d_N/d_S value (ω_1) of 1.0, while the second category (p_2) denotes sites where nonsynonymous changes are deleterious and so removed by negative selection, so that ω_2 is zero. The third "positive selection" model incorporates an additional category of positively selected sites (p_3) at which ω_3 can be >1, in which case nonsynonymous substitutions have higher rates of fixation than synonymous substitutions (Nielsen and Yang 1998). Additionally, individual positively selected sites can be identified by their posterior probabilities of belonging to the category of sites with $\omega_3 > 1$ using an empirical Bayesian approach: the higher the posterior probability, the more likely that a site is under positive selection. Likewise, sites belonging to the invariant or neutral categories can also be detected using posterior probabilities. All these analyses were performed using the CODEML program from the PAML package (Yang 1997).

The results of this genealogy-based analysis of selection pressures were compared to those obtained using the pairwise method of Nei and Gojobori (1986), as implemented in the MEGA sequence analysis package (Kumar *et al.* 1993). d_N/d_S values for individual codons (calculated as the mean of all pairwise comparisons) were estimated using the SNAP pro-

TABLE 1

Maximum likelihood estimates of selection pressures on HIV-1 *nef* sequences within a single hemophiliac patient

Model		11 mo	25 mo	41 mo	11 + 25 mo	25 + 41 mo	Total (11 + 25 + 41)
Input ML topology	ln L[a]	−1018.95084	−1170.141	−1093.266	−1363.078	−1463.603	−1658.001
	Ts/Tv[b]	7.024	6.773	1.221	6.699	2.985	3.535
	κ[c]	13.171	12.744	2.288	13.145	5.604	6.634
	α[d]	infinity	0.002	0.121	0.134	0.145	0.256
	taxa	16	16	15	32	31	47
	codons	206	206	206	206	206	206
1. Goldman-Yang	ω[e]	0.952	1.572	0.604	1.139	1.070	0.950
	ln L	−1003.917	−1142.659	−1081.824	−1331.687	−1431.743	−1623.403
2. Neutral	p_1[f]	0.999	0.611	0.432	0.639	0.521	0.594
	p_2[g]	0.001	0.388	0.568	0.361	0.479	0.406
	ln L	−1003.921	−1141.808	−1080.454	−1329.182	−1425.133	−1617.205
3. Positive selection	p_1	0.0001	0.000	0.035	0.4387	0.476	0.534
	p_2	0.000	0.791	0.738	0.4309	0.437	0.374
	p_3[h]	0.999	0.209	0.227	0.1305	0.086	0.088
	ω_3[i]	0.951	8.126	2.671	6.028	8.747	6.144
	ln L	−1003.914	−1128.934	−1079.531	−1320.956	−1410.861	−1607.953
Likelihood test (1 and 3)	χ^2	0.006	27.450	4.586	21.462	41.764	30.900
	P value	$P = 1$	$P < 0.001$	$P > 0.05$	$P < 0.001$	$P < 0.001$	$P < 0.001$
Likelihood test (2 and 3)	χ^2	0.0140	25.748	1.846	16.452	28.544	18.504
	P value	$P \sim 0.9$	$P < 0.001$	$P < 0.5$	$P < 0.001$	$P < 0.001$	$P < 0.001$

[a] Log likelihood.
[b] Observed transition/transversion ratio.
[c] Instantaneous transition/transversion ratio.
[d] Shape parameter of a gamma distribution of among-site rate variation.
[e] d_N/d_S ratio.
[f] Proportion of neutral codons.
[g] Proportion of deleterious codons.
[h] Proportion of positively selected codons.
[i] d_N/d_S ratio at p_3 sites.

TABLE 2

Maximum likelihood estimates of selection pressures on HIV-1 *nef* sequences from different subtypes and groups

Model		Subtype B	M group	M + O + Chimp
Input ML topology	ln L	−4713.066	−2757.589	−3986.997
	Ts/Tv	1.798	1.674	1.507
	κ	3.378	3.117	2.847
	α	0.329	0.541	0.625
	taxa	39	10	11
	codons	202	207	195
1. Goldman-Yang	ω	0.752	0.483	0.301
	ln L	−5052.875	−2761.365	−3610.225
2. Neutral	p_1	0.520	0.524	0.568
	p_2	0.480	0.476	0.432
	ln L	−4856.852	−2704.594	−3521.505
3. Positive selection	p_1	0.420	0.376	0.278
	p_2	0.466	0.000	0.382
	p_3	0.114	0.690	0.340
	ω_3	4.706	0.069	0.280
	ln L	−4750.552	−2694.304	−3488.791
Likelihood test (1 and 3)	χ^2	604.646	134.122	242.868
	P value	$P < 0.001$	$P < 0.001$	$P < 0.001$
Likelihood test (2 and 3)	χ^2	212.600	20.580	65.428
	P value	$P < 0.001$	$P < 0.001$	$P < 0.001$

Symbols as in Table 1.

gram (available at http://hiv-web.lanl.gov/SNAP/WEBSNAP/SNAP.html).

Using genealogies to represent the process of allele substitution: The process of allele substitution has an explicit phylogenetic representation. Specifically, we can assume that changes on external branches of a gene genealogy (autapomorphies) are evolutionary novelties: alleles not fixed in the population. Conversely, changes on internal branches of the genealogy (synapomorphies) represent alleles that are present in a larger (monophyletic) group of descendants. In general, therefore, the higher the frequency of an allele in a population, the deeper it will be located in the genealogy. Of most interest for our within-patient HIV sequence data are the synapomorphic changes located on the internal branches that separate each time point because these represent alleles that may have been fixed between the sampling events.

Given this framework we can study the substitution process simply by determining the most parsimonious reconstructions (MPRs) for each branch of the maximum likelihood tree linking all time points. This analysis was performed using MacClade (version 3.0, Maddison and Maddison 1992). Crucially, we can also determine whether the synapomorphic changes between time points are selectively advantageous by asking whether they reside in codons previously identified as having a high posterior probability of being positively selected. It is important to note in this context that although they use the same input tree, the MPRs and the posterior probability estimations use different optimality criteria and therefore are independent analyses.

RESULTS

Maximum likelihood analysis of positive selection in the HIV-1 *nef* gene: We first reconstructed maximum likelihood trees for samples from within the hemophiliac patient, taking each time point separately and in combination. Three codon-based maximum likelihood models were then applied to see which provided the best fit to these data. Since the positive selection model has two more parameters than the neutral model, the models are nested and their likelihoods can be compared directly using a χ^2-test with d.f. = 2. As can be seen in Table 1, the positive selection model has a better fit to the data at 25 mo ($P < 0.001$), with 20.9% falling into the selected category ($\omega_3 = 8.126$). Although positive selection was not significantly favored at 41 mo postinfection ($0.1 > P > 0.05$), a high value of ω_3 (2.671) was obtained for 22.7% of the sites. There was no evidence for positive selection at 11 mo. When successive data points were combined (*i.e.*, 11 plus 25 mo and 25 plus 41 mo) the positive selection model was significantly favored over both competing models, although with fewer positively selected sites. Those sites with high posterior probabilities of being positively selected were also determined and are plotted for the two sets of successive time points in Figure 1. A total 17 substitutions fell into this class and it is interesting that some new positively selected sites appear in the 25- plus 41-mo comparison, most notably a cluster of three at the 5′ part of the sequence. No sites with ≥90% posterior probability of evolving neutrally were identified in either comparison.

The positive selection model also had the highest

likelihood for all three time points combined, being much better than the neutral model ($P < 0.001$), although only 8.8% of sites belonged to the p_3 category with high d_N/d_S ($\omega_3 = 6.144$). Since the Goldman and Yang constant d_N/d_S model is also a special case of the positive selection model with $p_1 = p_2 = 0$ and $P_3 = 1$, twice the difference in likelihood between these two models (d.f. = 2) also constitutes a valid test statistic. For all data combinations the positive selection model gave high χ^2 values when tested against the Goldman and Yang model (Table 1).

To determine whether positive selection can be detected at greater evolutionary distances, three more *nef* data sets were examined (Table 2). For the 39 subtype B sequences the positive selection model provided a significantly better fit to the data than both competing models ($P < 0.001$), although only 11.40% of codons were selectively favored ($\omega_3 = 4.706$). The positive selection model also outperformed both the Goldman and Yang and the neutral models in the analysis of the 10 group M sequences and the 11 group M, O, and chimpanzee viruses ($P < 0.001$ in both cases). However, because the optimal values for ω_3 were both <1, we cannot formally demonstrate positive selection at these deeper phylogenetic levels.

For the subtype B data we also recorded the locations of those codons with a high posterior probability (≥90%) of being positively selected (Figure 2). A total of 22 positively selected codons were identified, 15 of which (68%) were located within known CTL epitopes. Of the seven remaining sites, two were found within targets of monoclonal antibodies and four represent contiguous amino acids, from positions 8 to 11, suggesting that this region may contain an as-yet-undescribed epitope. Intriguingly, positively selected substitutions at positions 8 and 9 were also identified in the hemophiliac patient, although no information is available regarding human leukocyte antigen (HLA) type of this individual.

Pairwise methods do not detect positive selection in *nef*: No positive selection was detected in the hemophiliac patient when sequences were analyzed using the pairwise method of Nei and Gojobori (1986). No statistically significant deviation from strict neutrality (*i.e.*, $d_N/d_S = 1.0$ at the 95% confidence interval) was observed at 11 mo ($d_N/d_S = 0.60$, $t = 1.027$, $P > 0.1$), 25 mo ($d_N/d_S = 0.70$, $t = 0.775$, $P > 0.1$), or 41 mo postinfection ($d_N/d_S = 0.50$, $t = 1.1575$, $P > 0.1$). Likewise, selection was not detected when d_N/d_S was calculated for the 11- and 25-mo time points combined ($d_N/d_S = 0.35$, $t = 0.92$, $P > 0.1$), the 25- and 41-mo third time points combined ($d_N/d_S = 0.59$, $t = 1.290$, $0.1 > P > 0.05$), nor for all time points combined ($d_N/d_S = 0.52$, $t = 1.027$, $P > 0.1$). Therefore, simple pairwise estimations of d_N/d_S provide no evidence for positive selection in these data, a conclusion also reached by Plikat *et al.* (1997). The Nei and Gojobori analysis likewise provided no evidence for adaptive evolution among the subtype B sequences, although significantly greater values of d_S over d_N were observed ($d_N/d_S = 0.60$, $t = 4.877$, $P < 0.001$), as was the case for the group M sequences ($d_N/d_S = 0.28$, $t = 7.361$, $P < 0.001$) and those sequences from groups M, O, and chimpanzee combined ($d_N/d_S = 0.22$, $t = 8.9095$, $P < 0.001$).

To investigate the discrepancy between the genealogical and pairwise methods in more detail we first subtracted d_S from d_N for each pairwise comparison estimated under the Nei-Gojobori method (Figure 3). Those pairwise comparisons suggesting positive selection (*i.e.*, $d_N > d_S$) fall to the right of the vertical line on each histogram that delineates $d_N - d_S = 0$. Although, for all time points, the distributions have a mean of $d_N < d_S$ (but very near zero), 31.66%, 38.97%, and 16.19% of pairwise comparisons fell in the positive rank for the 11-, 25-, and 41-mo time points, respectively. Thus, sequence comparisons with $d_N > d_S$ are present in the data but are lost, such that positive selection is rejected with a *t*-test, when an average of all pairwise comparisons is taken.

Next, we compared d_N/d_S values along the *nef* gene sequence using a sliding window of 20 codons, incremented 1 codon at time. Although this analysis revealed some regions where $d_N > d_S$, particularly in the 3′ part of the sequence (Figure 4), the majority of the positively selected sites identified in the maximum likelihood analysis were not detected. Even more striking is the extreme variation in d_S, with some instances of $d_N > d_S$ clearly due to regional reductions in d_S, rather than elevations in d_N. If, instead, cases are recorded in which d_N is greater than the *mean* value of d_S, two regions appear to be positively selected: the first nine codons of the sequence and codon 169, both of which were contained within the positively selected class in the maximum likelihood analysis. However, all other positively selected sites were not detected and there is no longer any evidence for positive selection in the extreme 3′ region of the gene.

Reconstructing allele substitutions in the *nef* gene: A parsimony method was next used to reconstruct the unambiguous amino acid and nucleotide changes along each branch of the maximum likelihood tree for all three time points combined (Figure 5). A very similar phylogeny was found using maximum parsimony as the initial optimality criterion.

Although the sequences from each time point do not form monophyletic groups, because those viruses present at 41 mo appear to be derived from a subset present at 25 mo, the tree is striking in that it clearly depicts a replacement of lineages through time as might be expected under natural selection, a pattern that received good bootstrap support. One silent and one amino acid change (a Arg to Lys substitution at position 105) were reconstructed on the lineage leading to the 25-mo time point, the latter of which had a very high

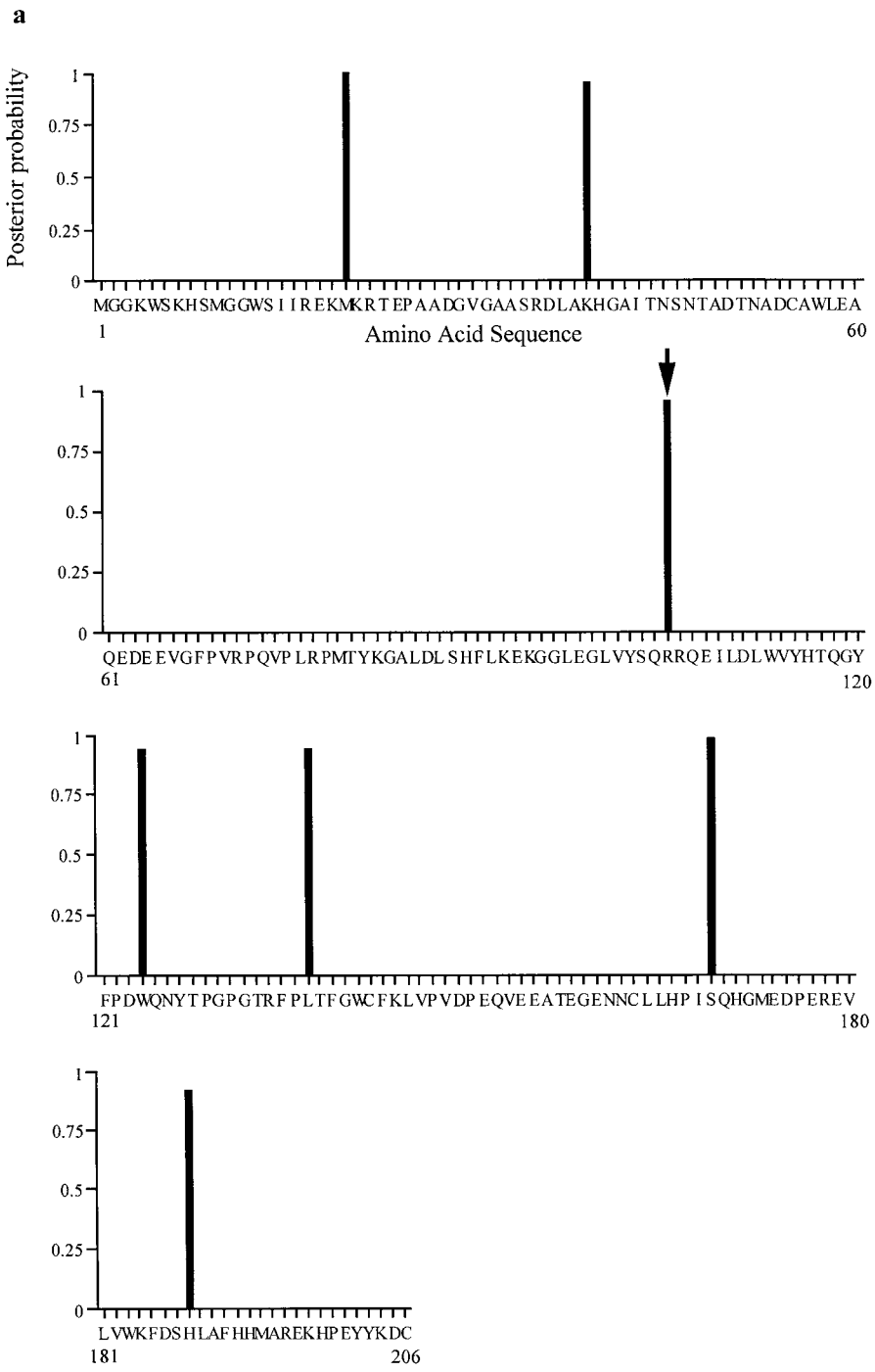

Figure 1.—Location of positively selected substitutions (with ≥90% posterior probability) within the *nef* gene of a hemophiliac patient. (a) 11- and 25-mo time points combined and (b) 25- and 41-mo time points combined. Sites estimated by the maximum likelihood analysis as evolving under positive selection are shown by bars with their posterior probabilities on the *Y* axis and the consensus Nef protein sequence for this patient along the *X* axis. No neutrally evolving sites with ≥90% posterior probability were identified. Those synapomorphic changes that separate each time point are indicated by arrows (see Figure 5).

probability (0.9564) of being positively selected. Likewise, one silent and two amino acid substitutions were reconstructed on the branch leading to the 41-mo time point and again both amino acid changes (at positions 8 and 9) had very high probabilities of being under positive selection.

Eleven more substitutions with a high probability of being positively selected were found to be synapomorphic for clusters of sequences within each time point, indicating that they represent mutations that are not yet fixed in the population or that had only a transient advantage. The remaining three putative positively selected changes were autapomorphic, which could also signify recently evolved or transiently advantageous alleles, or even recent deleterious mutations that have yet to be removed by selection (Fu and Li 1993).

Population genetic analysis of synapomorphic changes: Additional evidence for positive selection came from an analysis of various population parameters associated with allele substitution. For each time point within the hemophiliac patient, genetic diversity was quantified as θ ($2N_e\mu$), estimated using the Metropolis-

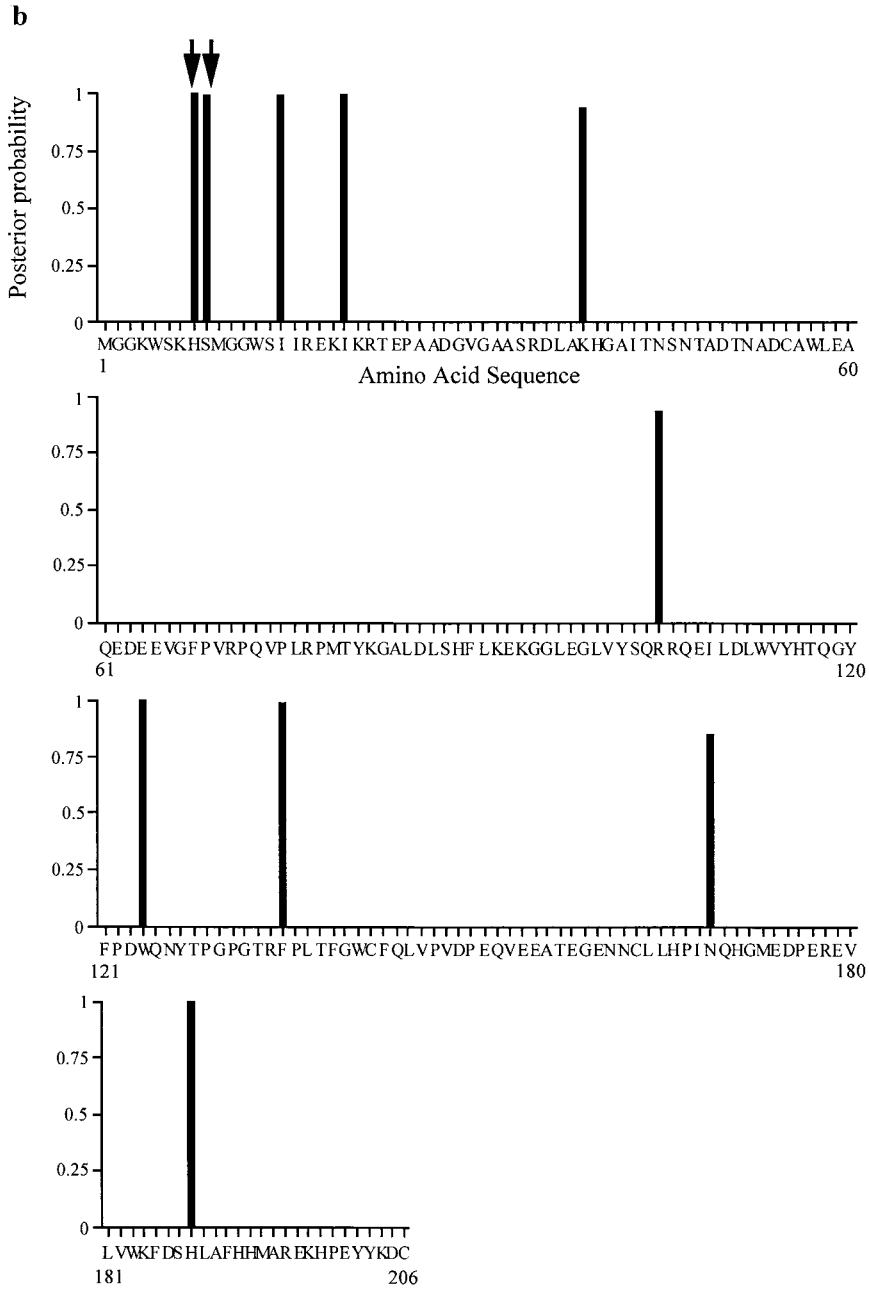

Figure 1.—*Continued.*

Hastings Monte Carlo algorithm on ultrametric trees of the data (program FLUCTUATE, version 1.1, Kuhner *et al.* 1995). These θ values—0.023587, 0.0624215, and 0.0277238—for time points 11, 25, and 41 mo postinfection, respectively, were then used to estimate values for the effective population size (N_e) of HIV-1, assuming substitution rates (μ) from 2.3×10^{-5} to 7.0×10^{-6} per genome replication (Temin 1993). This resulted in N_e estimates of 513–1685, 1357–4459, and 602–1980, for time points 11, 25, and 41 mo, respectively. Similarly low estimates of N_e have been obtained for other HIV-1-infected patients (Leigh Brown 1997).

Under neutrality, the time for a mutation to become fixed by genetic drift in a haploid population on average should be $2N_e$ generations, so that the expected times to fixation, given our range of N_e estimates, would be 1026–3370, 2714–8918, and 1204–3960 generations. Assuming a generation time for HIV-1 of ~2.6 days (Perelson *et al.* 1996) and considering the lowest value of N_e estimated (and hence the fastest fixation time), the synapomorphies we reconstruct on average would need ~89 mo (7.5 yr) to reach fixation by drift alone. That the synapomorphic changes we observed seem to be fixed far more quickly than this—our entire window of observation was only 30 mo—suggests that this substitution process was driven by positive selection.

It is also possible to estimate the selection coefficient (s) of the synapomorphies, assuming that advantageous

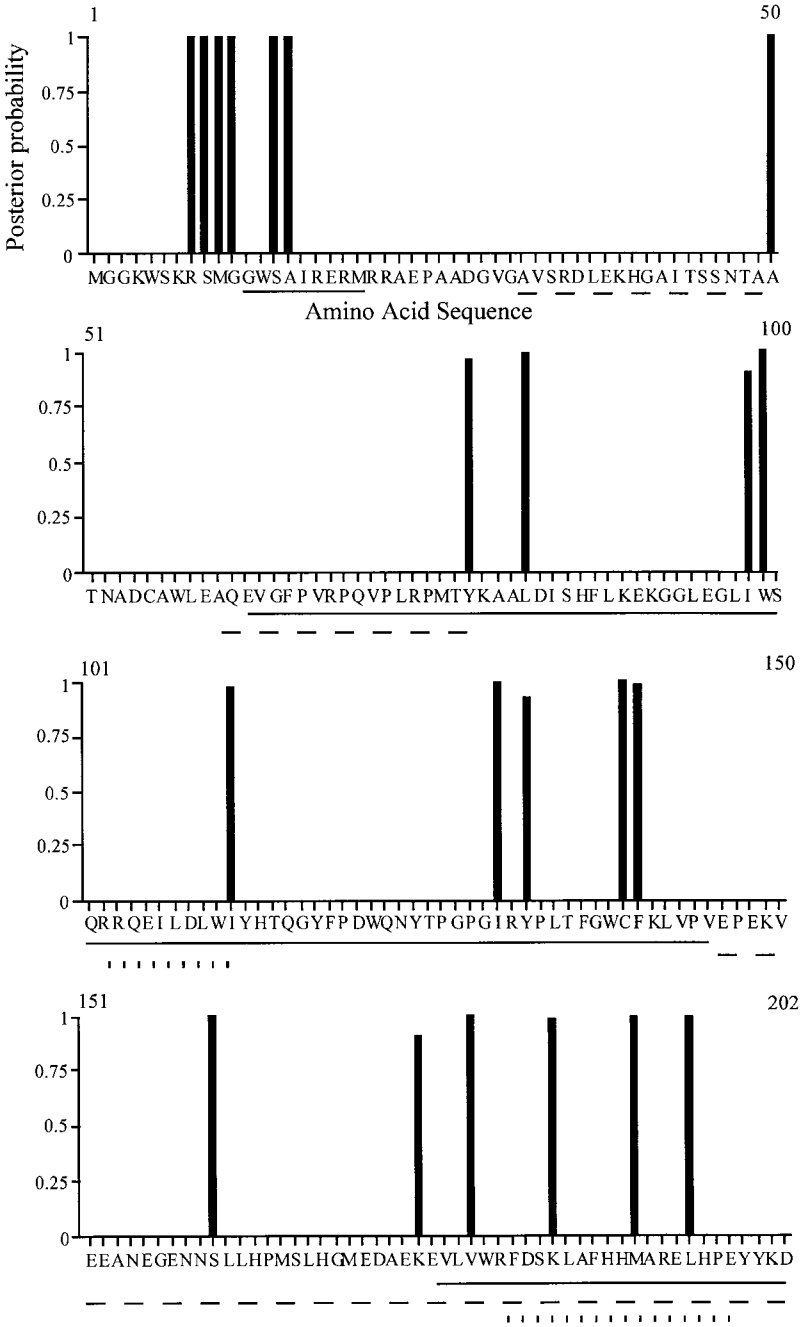

Figure 2.—Location of positively selected substitutions in 39 HIV-1 subtype B sequences. Bars represent the probability of substitutions under positive selection (with ≥90% posterior probability) and the amino acid sequence of the SF2 isolate was used as the prototype sequence for comparison (*X* axis). The continuous line beneath the plot depicts the location of known CTL epitopes from different HLA types, while dashed lines show targets of monoclonal antibodies, and dotted lines represent epitopes of T helper cells. Epitope information was taken from the Los Alamos HIV molecular immunology database (Korber *et al.* 1997a).

substitutions in a haploid population reach fixation in $\sim(2/s)\text{Log}\, e(N_e)$ generations, although with a large variance (Nei 1987). If we assume that the two selected synapomorphies at 41 mo first appeared at 25 mo, then they reached fixation in 16 mo, which, conservatively assuming that N_e takes the lower value for this period (602), would make $s = 0.069$. Of course, if any of these synapomorphies were present before they were first sampled, then lower values of *s* would be obtained, although it is equally likely that these synapomorphic changes first appeared more recently than we assume (*i.e.*, at later intervals along the branches linking time points), in which case selection coefficients would increase. The synapomorphy at 25 mo is slightly harder to interpret because, although it is reconstructed as occurring on the lineage leading to the 25-mo sample, it is not present in all isolates from this time point because there appears to have been a reversion to Arg in some sequences. However, it is found in all sequences at 41 mo, so we may assume it has been fixed by this time. We can therefore conservatively estimate that this mutant took 30 mo to reach fixation, which would mean $s = 0.036$, assuming the lowest value of N_e calculated during this time (513).

DISCUSSION

Positive selection on *nef* genes: Our genealogical study of HIV-1 *nef* gene evolution within and among

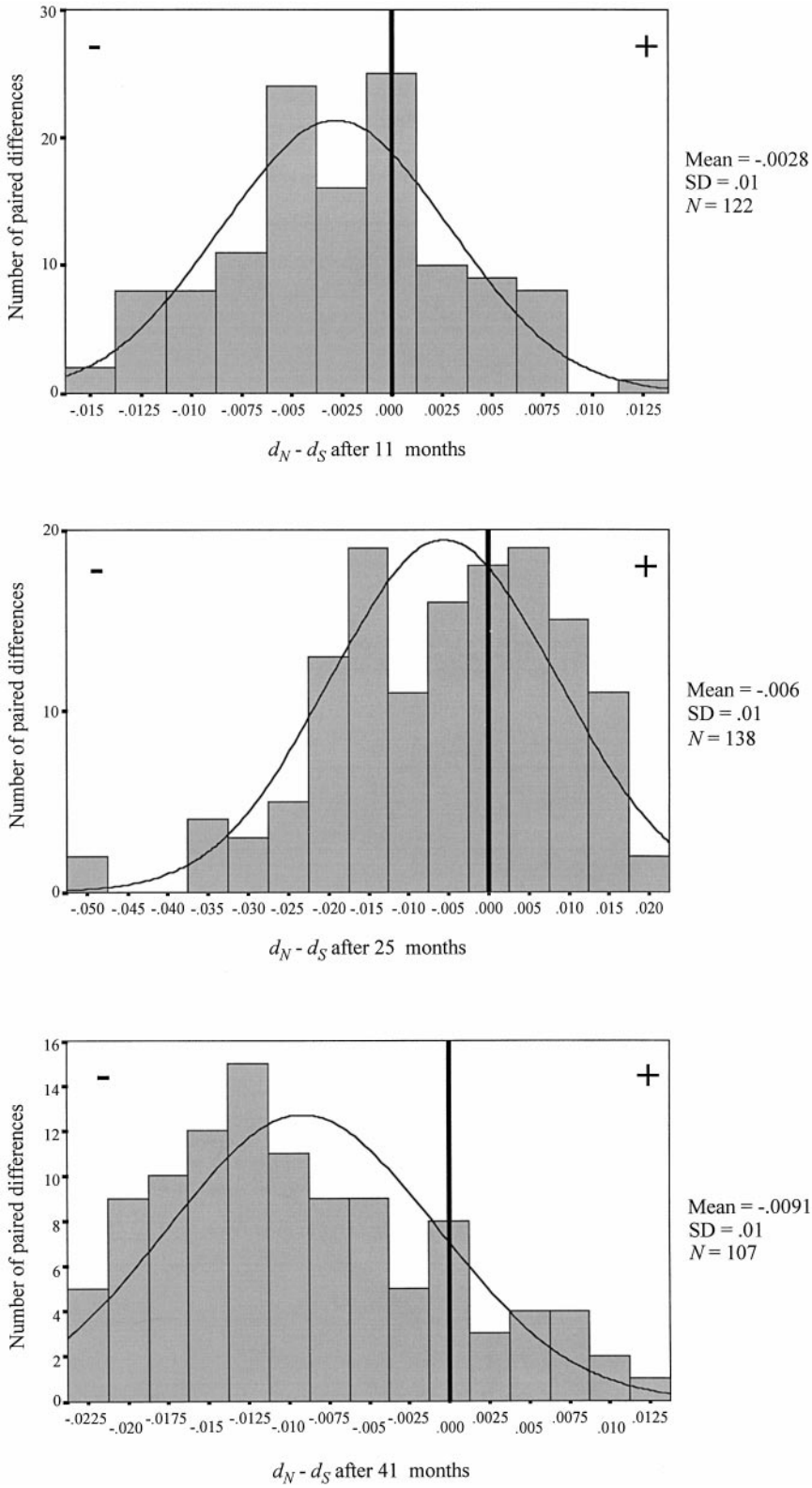

Figure 3.—Histograms showing the values of $d_N - d_S$ for each pairwise comparison in the hemophiliac patient. Although the mean values for each time point are negative (*i.e.*, $d_N < d_S$), many of the individual pairwise comparisons provide estimates in which $d_N > d_S$, so that they are assigned a positive rank and fall on the right-hand side of the histogram. SD, standard deviation.

patients has revealed an important role for positive selection, with high d_N/d_S values at some codons. Within the hemophiliac patient some of these selected codons were also found to be synapomorphic for samples taken from successive time points and thus fall along the "backbone" of the tree, itself strong evidence that they represent the successful (fixed) alleles from which all other mutants are derived. A similar finding comes from the analysis of strains of influenza A virus collected over many years (and representing many epidemics) where those sites under positive selection are likewise found at antigenically important residues and are located on

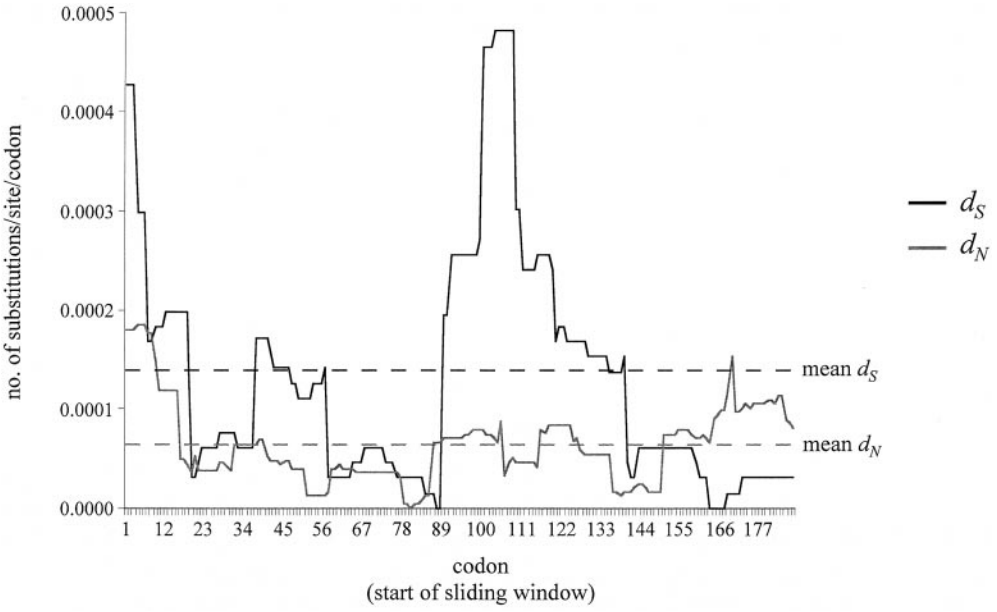

Figure 4.—Sliding window analysis of the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site, per codon along the *nef* gene sequence of the hemophiliac patient. Numbers (*Y* axis) are calculated as the mean of all pairwise comparisons. Mean d_S and d_N values across all codons are shown as hatched lines.

the main trunk of the tree (Fitch *et al.* 1991). Finally, the location of the amino acid sites under positive selection in the hemophiliac patient changes with time, consistent with the notion that the immune system may shift its attention among epitopes following the appearance of escape mutants (Nowak *et al.* 1995), a process that has been observed in *nef* (Price *et al.* 1997).

Taken together we believe that these observations represent compelling evidence for the immune-driven positive selection of nucleotide substitutions in *nef*. A possible alternative explanation is that our "positively selected" substitutions are in fact nearly neutral and fixed by genetic drift when the viral population is small (Ohta 1992). However, the rate of fixation of nearly neutral mutations under drift is the same as that of strictly neutral changes, and we have already shown that

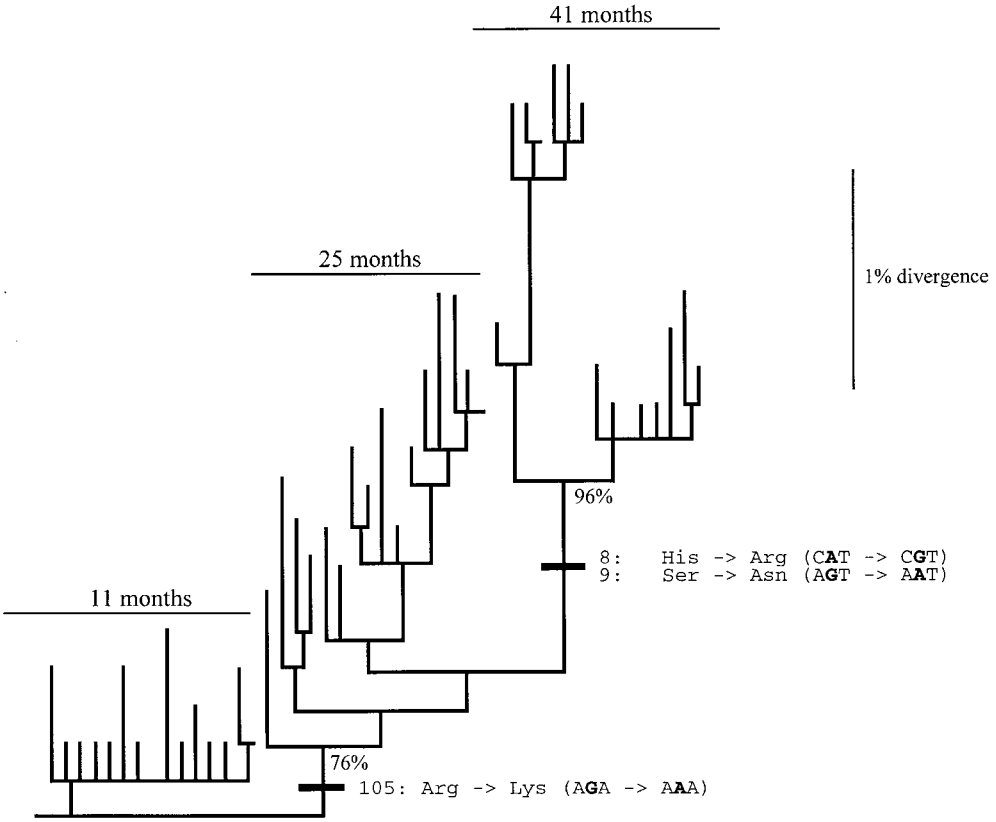

Figure 5.—Maximum likelihood phylogenetic tree of 30 mo of *nef* gene evolution in the hemophiliac patient. Branch lengths are proportional to the amount of change at the DNA level, with the most parsimonious reconstruction (MPR) of unambiguous synapomorphic amino acid changes, and corresponding nucleotide substitutions, shown next to the branches leading to the 25-mo and 41-mo time points (sequence coordinates refer to amino acids). The number of bootstrap replications supporting the separation of these time points is also shown. All three synapomorphic changes had high posterior probabilities of being under positive selection (see Figure 1).

this rate is too low for the substitution dynamics we observe.

Limitations of pairwise methods: Our study also indicates that genealogy-based methods provide a much more sensitive description of selection pressures than those using multiple pairwise comparisons, even when sliding windows are incorporated. Although the various methods for estimating d_S and d_N based on pairwise comparisons are useful when the sites under positive selection are known *a priori* (Hughes and Nei 1988; Zanotto *et al.* 1995; Yokohama and Yokohama 1996) or when there is an overwhelming excess of nonsynonymous changes in particular regions (Bonhoeffer *et al.* 1995; Price *et al.* 1997; Karlsson *et al.* 1998), all are limited by their oversampling of distances associated with deeper branches, the movement of sites between the synonymous and nonsynonymous categories, and the fact that estimations of the two rates are not independent (Muse 1996). Furthermore, the pairwise methods currently available assume a constant selection pressure among sites and so tend to underestimate nonsynonymous rates (Nielsen 1997). Finally, although the Nei and Gojobori method works well, given low levels of sequence divergence and equal rates of substitution among bases (Muse 1996), this evidently limits its applicability to a rapidly evolving organism with a very biased substitution process like HIV-1. As a recent case in point, pairwise comparisons of d_N/d_S failed to detect positive selection in HIV-1 sequences that were obviously selected for antiviral resistance (Crandall *et al.* 1999).

Our study further confirms that adaptive evolution often occurs at a small number of residues in a polypeptide, in this case most likely CTL epitopes. As a consequence, methods that take average d_N/d_S values among many sites are necessarily coarse and may miss evidence for very localized selection pressure (Sharp 1997). For example, Endo *et al.* (1996) considered cases in which $d_N > d_S$ in 50% of pairwise comparisons to provide good evidence for positive selection, yet the *nef* gene sequences analyzed here would have clearly been excluded under this criterion, as would some other notable cases of adaptive evolution such as primate lysozymes (Messier and Stewart 1997). The camouflaging of positively selected sites will obviously be most acute when more divergent sequences are compared, and even in our maximum likelihood analysis d_N/d_S decayed through time such that, although the "positive selection" codon evolution model had the highest likelihood in comparisons among HIV-1 group M sequences, the increase in the number of silent changes meant that no codons in which $d_N > d_S$ could be identified. A lack of sensitivity is also a limitation for sliding window analyses, which, while providing more evidence for positive selection in *nef*, still failed to detect the majority of the positively selected changes. Furthermore, there are no objective criteria by which to choose either the window or increment sizes and in our study different analyses based on the sliding window gave different interpretations of which sites might be selected. Future studies of positive selection will evidently be most fruitful if they consider closely related sequences where the footprint of adaptive evolution may still be uncovered and if they utilize analytical methods that take account of the phylogenetic relationships of the sequences in question. Ultimately such methods should also be able to recognize the selective advantage conferred by individual mutations.

Cladistic representation of the substitution of *nef* alleles: The clear phylogenetic separation of *nef* sequences from the three time points in the hemophiliac patient was instrumental in our study of evolutionary processes as it allowed us to estimate a number of important population parameters. For example, the synapomorphic changes at 41 mo, if they first appeared at 25 mo, took no more than about 185 generations to reach fixation, some 5.5 times faster than expected under neutrality, given the lowest values of N_e estimated. Even if fixation took the entire 30 mo of the sampling period this substitution process is still 3.0 times faster than the neutral expectation. Likewise, these fixed substitutions have very high selection coefficients, with s at least 0.036 under the most conservative assumptions, and are greater than those estimated for wild-type reverse transcriptase alleles in the absence of treatment with the drug AZT (s = 0.004 to 0.023; Goudsmit *et al.* 1996) and for balancing selection at loci of the human major histocompatibility complex (s = 0.0007 to 0.042; Satta *et al.* 1994).

Of course, the estimates of N_e (and hence s) that we present assume neutrality and we have shown here that positive selection has acted on these sequences. However, our point is that even with low values of N_e many more generations than observed are required to explain the rapid substitution of *nef* alleles by drift alone. Furthermore, larger values of N_e would increase values of s so that the selection coefficients we present are likely to represent lower bounds. Finally, if N_e really is as low as we estimate then our analysis suggests that this is due to the purging action of selectively driven population bottlenecks, rather than high variation in the number of viral progeny produced by infected cells (Leigh Brown 1997).

One questionable assumption we do make is that the synapomorphic changes for each time point have truly undergone fixation during the period of sampling, especially since the viral population within hosts may be partitioned by tissue type [although this is debated—see, for example, Delwart *et al.* (1998)]. However, given the enormous census population size of HIV-1, with some 10^{10} virions produced each day (Perelson *et al.* 1996), the fact that all sequences at the 41-mo time point had these synapomorphic substitutions at least argues for their high frequency in the population. Furthermore, each synapomorphy has a high posterior probability of being subject to positive selection. It there-

fore remains to be seen how different frequencies of these mutations would affect our reconstruction of their substitution dynamics.

Using selection analysis to locate epitopes: The identification of CTL epitopes is essential if we are to better characterize the cellular response to viral infection. This task, however, is complex. Mathematical models have been used to predict the likelihood of putative CTL peptide sequences within different viral proteins, applying scores before screening with CTL assays using ^{51}Cr (Falk *et al.* 1991). Novel approaches using ELISPOT or peptide-loaded HLA-tetramers in flow cytometry are potentially easier, although more accurate and flexible methods for the identification of candidate peptides are still desirable. We suggest that examining those amino acid sites under positive selection may be a useful way to identify possible epitope regions, as many of the positively selected sites we detect correspond to CTL epitopes or highlight regions where others may reside. Not only may such an evolutionary approach shorten the time, labor, and cost of these studies, but it may ultimately assist our understanding of the immuno-pathogenesis of AIDS and other infectious diseases.

We thank Rasmus Nielsen, Ziheng Yang, Yun-Xin Fu, and Takashi Gojobori for their suggestions and comments. Two anonymous referees also made useful suggestions concerning an earlier version of this manuscript. P.M.A.Z. was funded by a Conselho Nacional de Pesquisa (CNPq) productivity grant (300188/98-6) and by Programa Nacional de Excelência (PRONEX) grant 139/96. E.C.H. was funded by The Royal Society (U.K.) and The Wellcome Trust. R.F.S. was funded by Coordenação de Aperfeiçoamento de Pesquisa e Ensino Superior (CAPES) and E.G.K. by PRONEX grant 139/96.

LITERATURE CITED

Bennett, S. R., F. R. Carbone, F. Karamalis, R. A. Flavell, J. F. Miller *et al.*, 1998 Help for cytotoxic-T-cell responses is mediated by CD40 signaling. Nature **393:** 478–480.

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

Borrow, P., H. Lewiki, X. Wei, M. S. Horwitz, N. Peffer *et al.*, 1997 Antiviral pressure exerted by HIV-1 specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. Nat. Med. **3:** 205–211.

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.

da Silva, J., and A. L. Hughes, 1998 Conservation of cytotoxic T lymphocyte (CTL) epitopes as a host strategy to constrain parasite adaptation: evidence from the *nef* gene of human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. **15:** 1259–1268.

Delwart, E. L., J. I. Mullins, P. Gupta, G. H. Learn, Jr., M. Holodniy *et al.*, 1998 Human immunodeficiency virus type 1 populations in blood and semen. J. Virol. **72:** 617–623.

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.

Falk, K., O. Rotzschke, S. Stevanovic, G. Jung and H. G. Rammensee, 1991 Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature **351:** 290–296.

Fitch, W. M., J. M. E. Leiter, X. Li and P. Palese, 1991 Positive Darwinian evolution in human influenza A viruses. Proc. Natl. Acad. Sci. USA **88:** 4270–4274.

Fu, Y. X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

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

Goudsmit, J., A. de Ronde, D. D. Ho and A. S. Perelson, 1996 Human immunodeficiency virus in vivo: calculations based on a single zidovudine resistance mutation at codon 215 of reverse transcriptase. J. Virol. **70:** 5662–5664.

Goulder, P. J. R., R. E. Phillips, R. A. Colbert, S. McAdam, G. Ogg *et al.*, 1997 Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. Nat. Med. **3:** 212–217.

Holmes, E. C., and P. M. de A. Zanotto, 1998 Genetic drift of human immunodeficiency virus type 1? J. Virol. **72:** 886–887.

Hughes, A. L., and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals positive selection. Nature **335:** 367–370.

Karlsson, A. C., S. Lindback, H. Gaines and A. Sonnerborg, 1998 Characterization of the viral population during primary HIV-1 infection. AIDS **12:** 839–847.

Kestler, H. W., D. J. Ringler, K. Mori, D. L. Panicali, P. K. Sehgal *et al.*, 1991 Importance of the *nef* gene for maintenance of high virus loads and for development of AIDS. Cell **65:** 651–662.

Kirchhoff, F., T. C. Greenough, D. B. Brettler, J. L. Sullivan and R. C. Desrosiers, 1995 Absence of intact *nef* sequences in a long-term survivor with nonprogressive HIV-1 infection. N. Engl. J. Med. **332:** 228–232.

Koenig, S., A. J. Conley, Y. A. Brewah, G. M. Jones, S. Leath *et al.*, 1995 Transfer of HIV-1 specific cytotoxic T lymphocytes to an AIDS patient leads to selection for mutant HIV variants and subsequent disease progression. Nat. Med. **1:** 330–336.

Korber, B., C. Brander, B. F. Haynes, J. P. Moore, R. Koup *et al.*, 1997a *HIV Molecular Immunology Database 1997*. Los Alamos National Laboratory, Los Alamos, NM.

Korber, B., B. Foley, T. Leitner, F. McCutchan, B. Hahn *et al.*, 1997b *Human Retrovirus and AIDS 1997*. Los Alamos National Laboratory, Los Alamos, NM.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

Kumar, S., K. Tamura and M. Nei, 1993 *MEGA: Molecular Evolutionary Genetics Analysis*, version 1.01. The Pennsylvania State University, University Park, PA.

Leigh Brown, A. J., 1997 Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population. Proc. Natl. Acad. Sci. USA **94:** 1862–1865.

Leigh Brown, A. J., and D. D. Richman, 1997 HIV-1: gambling on the evolution of drug resistance? Nat. Med. **3:** 268–271.

Levy, J. A., 1998 *HIV and the Pathogenesis of AIDS*, Ed. 2. ASM Press, Washington, DC.

Maddison, W. P., and D. R. Maddison, 1992 *MacClade: Analysis of Phylogeny and Character Evolution*. Version 3.0. Sinauer Associates, Sunderland, MA.

McMichael, A., 1998 T cell responses and viral escape. Cell **93:** 673–676.

McMichael, A. J., and R. E. Phillips, 1997 Escape of human immunodeficiency virus from immune control. Annu. Rev. Immunol. **15:** 271–296.

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

Muse, S. V., 1996 Estimating synonymous and nonsynonymous substitution rates. Mol. Biol. Evol. **13:** 105–114.

Musey, L., Y. Hu, L. Eckert, M. Christensen, T. Karchmer *et al.*, 1997 HIV-1 induces cytotoxic T lymphocytes in the cervix of infected women. J. Exp. Med. **185:** 293–303.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

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.

Nowak, M. A., R. M. May, R. E. Phillips, S. Rowland-Jones, D. G. Lalloo *et al.*, 1995 Antigenic oscillations and shifting immunodominance in HIV-1 infections. Nature **375:** 606–611.

Nowak, M. A., R. M. Anderson, M. C. Boerlijst, S. Bonhoeffer, R. M. May *et al.*, 1996 HIV-1 evolution and disease progression. Science **274:** 1008–1010.

Ogg, G. S., X. Jin, S. Bonhoeffer, P. R. Dunbar, M. A. Nowak *et al.*, 1998 Quantitation of HIV-1 specific cytotoxic T lymphocytes and plasma load of viral RNA. Science **279:** 2103–2106.

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

Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard and D. D. Ho, 1996 HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science **271:** 1582–1586.

Plikat, U., K. Nieselt-Struwe and A. Meyerhans, 1997 Genetic drift can determine short-term human immunodeficiency virus type 1 *nef* quasispecies evolution in vivo. J. Virol. **71:** 4233–4240.

Price, D. A., P. J. R. Goulder, P. Klenerman, A. K. Sewell, P. J. Easterbrook *et al.*, 1997 Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. Proc. Natl. Acad. Sci. USA **94:** 1890–1895.

Ridge, J. P., F. di Rosa and P. Matzinger, 1998 A conditioned dendritic cell can be a temporal bridge between a CD4+ T-helper and a T-killer cell. Nature **393:** 474–478.

Rosenberg, E. S., J. M. Billingsley, A. M. Caliendo, S. L. Boswell, P. E. Sax *et al.*, 1997 Vigorous HIV-1-specific CD4+ T cell responses associated with control of viremia. Science **278:** 1447–1450.

Satta, Y., C. O'Huigin, N. Takahata and J. Klein, 1994 Intensity of natural selection at the major histocompatibility complex loci. Proc. Natl. Acad. Sci. USA **91:** 7184–7188.

Schmitz, J. E., M. J. Kuroda, S. Santra, V. G. Sasseville, M. A. Simon *et al.*, 1999 Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. Science **283:** 857–860.

Schoenberger, S. P., R. E. Toes, E. I. van der Voort, R. Offringa and C. J. Melief, 1998 T-cell help for cytotoxic T lymphocytes is mediated by CD40-CD40L interactions. Nature **393:** 480–483.

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

Temin, H. M., 1993 The high rate of retrovirus variation results in rapid evolution, pp. 219–233, in *Emerging Viruses*, edited by S. S. Morse. Oxford University Press, Oxford.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Wain-Hobson, S., 1994 Is antigenic variation of HIV important for AIDS and what might be expected in the future? pp. 185–209, in *The Evolutionary Biology of Viruses*, edited by S. S. Morse. Raven Press, New York.

Wain-Hobson, S., 1996 Running the gamut of retroviral variation. Trends Microbiol. **4:** 135–141.

Welker, R., H. Kottler, H. R. Kalbitzer and H.-G. Kräusslisch, 1996 Human immunodeficiency virus type 1 Nef protein is incorporated into virus particles and specifically cleaved by the viral proteinase. Virology **219:** 228–236.

Yang, Z., 1997 Phylogenetic Analysis by Maximum Likelihood (PAML), Version 1.4. Department of Integrative Biology, University of California, Berkeley.

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

Zanotto, P. M. de A., G. F. Gao, T. Gritsun, M. S. Marin, W. R. Jiang *et al.*, 1995 An arbovirus cline across the Northern hemisphere. Virology **210:** 152–159.

Communicating editor: J. Hey