MOLECULAR BIOLOGY AND EVOLUTION

A Statistical Characterization of Consistent Patterns of Human Immunodeficiency Virus Evolution Within Infected Patients

Scott Williamson,† Steven M. Perry,† Carlos D. Bustamante,* Maria E. Orive,† Miles N. Stearns,† and John K. Kelly†*

*Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York; †Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence

Within-patient HIV populations evolve rapidly because of a high mutation rate, short generation time, and strong positive selection pressures. Previous studies have identified "consistent patterns" of viral sequence evolution. Just before HIV infection progresses to AIDS, evolution seems to slow markedly, and the genetic diversity of the viral population drops. This evolutionary slowdown could be caused either by a reduction in the average viral replication rate or because selection pressures weaken with the collapse of the immune system. The former hypothesis (which we denote "cellular exhaustion") predicts a simultaneous reduction in both synonymous and nonsynonymous evolution, whereas the latter hypothesis (denoted "immune relaxation") predicts that only nonsynonymous evolution will slow. In this paper, we present a set of statistical procedures for distinguishing between these alternative hypotheses using DNA sequences sampled over the course of infection. The first component is a new method for estimating evolutionary rates that takes advantage of the temporal information in longitudinal DNA sequence samples. Second, we develop a set of probability models for the analysis of evolutionary rates in HIV populations in vivo. Application of these models to both synonymous and nonsynonymous evolution affords a comparison of the cellular-exhaustion and immune-relaxation hypotheses. We apply the procedures to longitudinal data sets in which sequences of the *env* gene were sampled over the entire course of infection. Our analyses (1) statistically confirm that an evolutionary slowdown occurs late in infection, (2) strongly support the immune-relaxation hypothesis, and (3) indicate that the cessation of nonsynonymous evolution is associated with disease progression.

Introduction

Describing and quantifying the forces that shape HIV evolution within patients is critical to both (1) a mechanistic understanding of the interaction between the virus population and the immune system and (2) the design and implementation of clinical intervention strategies. Viral adaptation to the host environment is thought to play a causal role in HIV pathogenesis (Tersmette et al. 1989; Nowak et al. 1991; Wodarz, Klenerman, and Nowak 1998; Wolinsky and Learn 1999). Therefore, precisely characterizing the adaptive changes that occur during chronic infection should help elucidate the series of evolutionary events that lead to disease progression (e.g., Ross and Rodrigo 2002; Williamson 2003). Viral evolution also confounds medical treatment efforts because HIV populations quickly evolve resistance to antiviral drugs (see Shankarappa [1999] for review). Only those strategies that directly account for viral evolution in response to the intervention have been successful. For instance, highly active antiretroviral therapy (HAART) has proved to be a very successful long-term treatment strategy, primarily because the introduction of "drug cocktails" greatly slows the evolution of drug resistance (Finzi et al. 1997; Wong et al. 1997). Clearly, a detailed description of within-patient viral evolution is necessary to understanding the dynamics of HIV infection.

The most detailed information regarding viral evolution in vivo comes from longitudinal studies, in which samples of viral DNA or RNA are taken from the same infected patient over time (e.g., Balfe et al. 1990; Wolfs et al. 1991; Holmes et al. 1992; Zhang et al. 1993). We use "longitudinal data set" to refer to the set of all sequences sampled from a single patient over time. The temporal structure of longitudinal data sets allows researchers to determine the immediate consequences of specific events in viral evolution. Examples include the evolution of drug resistance (e.g., Larder, Darby, and Richman 1989), the evolution of viral phenotypes that escape the current immune response (McMichael and Phillips 1997; Goulder et al. 2001), and disease progression (Viscidi 1999; Ross and Rodrigo 2002; Williamson 2003).

Shankarappa et al. (1999) published an extensive collection of longitudinal data sets, including data from nine infected individuals. Sequences of the C2-V5 region of the envelope gene (*env*) were sampled over the entire course of infection in each patient. Samples were taken at approximately 10-month intervals with 10 to 20 sequences determined per sample. The authors assert that the most important pattern evident in their data is "consistent stages" of viral evolution (figure 6 in Shankarappa et al. [1999]). In the first stage, both *env* nucleotide diversity and divergence from the founding population increase linearly. During the next stage, which lasts, on average, approximately 2 years, nucleotide diversity either stabilizes or declines, but divergence continues to accumulate. In the final stage, divergence also stabilizes—that is, *env* sequence evolution nearly comes to a halt. This final stage usually begins less than 1 year before the onset of clinical AIDS ($CD4^+$ T cell counts <300 cells/μl). Further, Shankarappa et al. (1999) found close associations with these patterns and the emergence of the X4 viral genotype, an important indicator of disease progression (Schuitemaker et al. 1992). They noted that the time of peak diversity coincided with the initial emergence of X4 viruses, and divergence stabilization coincided with the X4 viruses reaching their highest frequency. In this paper, we investigate the causal factors leading to these consistent

Key words: HIV, divergence rate, longitudinal study.

E-mail: sw292@cornell.edu.

Mol. Biol. Evol. 22(3):456–468. 2005
doi:10.1093/molbev/msi029
Advance Access publication October 27, 2004

patterns of *env* evolution. In particular, we address what causes the evolutionary rate to slow so dramatically toward the end of infection.

We consider two alternative hypotheses to explain divergence stabilization. The first hypothesis, which we refer to as cellular exhaustion, proposes that the viral generation time increases as the primary target cell populations ($CD4^+$ T cells) become depleted. As the generation time increases, the evolutionary rate slows and divergence stabilizes. Several factors could contribute to the increase in generation time coincident with target cell depletion. First, as $CD4^+$ T cells are depleted late in infection, a greater proportion of replication cycles might go through other infectable cell types, such as macrophages and $CD4^+$ memory T cells. The intracellular stage of the viral life cycle occurs much more slowly in these alternative cell types (Finzi et al. 1997, 1999), and a shift to increasing usage of these cell types can have a major impact on the viral replication rate (Kelly 1994, 1996; Kelly et al. 2003). Also, the average time of the virion stage of the viral life cycle could increase simply because target cells are more rare and encountered less frequently by virions. This would also increase the overall generation time.

The second hypothesis, which we refer to as immune relaxation, is that the evolutionary rate slows because infection disrupts immune function. This, in turn, diminishes positive selection pressure to escape immune response (Bonhoeffer, Holmes, and Nowak 1995). Immune-mediated positive selection is thought to be an important force driving evolution in the HIV genome, especially in the *env* gene (e.g., Wolfs et al. 1990; Holmes et al. 1992; Bonhoeffer, Holmes, and Nowak 1995; Wolinsky et al. 1996; Richman et al. 2003). As the immune system is disrupted, the evolutionary rate could fall back closer to the expected neutral rate. This hypothesis, if correct, would have important implications for HIV evolution and dynamics. First, it would suggest that the primary selective force acting on *env* is the immune system itself. This result would not be surprising as it is generally assumed that the immune system is the primary source of positive selection pressure. However, the importance of immune-mediated selection relative to other selective pressures has been difficult to establish. Second, and perhaps more importantly, the immune-relaxation hypothesis suggests that the parts of the immune system disrupted by infection (most notably, $CD4^+$ T cells) play an indispensable role in responding to epitopes coded for by *env* and in driving *env* sequence evolution. Thus, the immune-relaxation hypothesis implies a feedback loop, with the virus directly impairing HIV-specific responses. Immunological evidence points to such an interaction between the virus and HIV-specific $CD4^+$ memory T cells (Douek et al. 2002). However, the evolutionary significance of this interaction has not yet been established.

We can differentiate between cellular exhaustion and immune relaxation by contrasting patterns of divergence at nonsynonymous and synonymous sites within *env*. The cellular-exhaustion hypothesis, because it is based on an increase in the viral generation time, predicts that both nonsynonymous and synonymous divergence will slow down at the time of divergence stabilization. In contrast, the immune-relaxation hypothesis predicts a slow-down in only the nonsynonymous sites because the synonymous sites are not subject to immune-mediated selection. In this paper, we measure the initial divergence rate and the time of divergence stabilization for both nonsynonymous and synonymous sites in the Shankarappa et al. (1999) data sets in hopes of distinguishing between the two alternative hypotheses of divergence stabilization.

When Shankarappa et al. (1999) initially identified the "consistent patterns," the times of peak diversity and divergence stabilization were estimated by eye. Given the stochastic nature of viral evolutionary dynamics, this subjective procedure can be problematic. Mayer-Hamblett and Self (2001) addressed this problem for changes in viral diversity by using regression models to estimate the time of peak diversity and to test hypotheses regarding changes in the level of diversity. Here, we develop a statistical procedure for characterizing changes in divergence rates based on a simple probability model. This method not only allows us to test the cellular-exhaustion and immune-relaxation hypotheses but also allows more accurate association between divergence stabilization and other aspects of infection (e.g., disease progression) and more accurate measures of the divergence rate.

In the next section, we develop a statistical framework for the analysis of divergence in longitudinal data sets and apply it to the longitudinal data from eight of the HIV-infected patients studied by Shankarappa et al. (1999). We derive a diffusion model to predict the likelihood of the various evolutionary trajectories observed in the data, either with or without divergence stabilization. This model serves as the underpinning for a maximum-likelihood estimation routine, which allows estimation of evolutionary parameters such as the mean and variance of divergence rate. It also provides a rigorous statistical test (a likelihood ratio) for the existence of divergence stabilization within patients. We use a parametric bootstrapping routine to evaluate the significance of the likelihood ratio tests. Finally, we test the assumptions of our approach using forward population-genetic simulations.

Methods

Measuring Divergence

We require an accurate method for measuring divergence to detect potentially subtle changes in the rate of divergence within a longitudinal data set. The typical approach for measuring divergence between sequences is to apply a nucleotide-substitution model that corrects for saturation; that is, a prevalence of multiple substitutions at the same nucleotide position (chapter 3 in Graur and Li [2000]). This correction is important, because saturation itself will lead to an apparent reduction in the rate of divergence through time. Unfortunately, the correction applied can be very sensitive to both the type of substitution model chosen and the method used (if any) to account for rate variation. Furthermore, the nucleotide substitution process in the HIV genome is fairly complex; that is, the nucleotide frequency profile is A-rich ($A \approx 40\%$), and the A→C transversion is more common than some transitions (Hillis, Huelsenbeck, and Cunningham 1994). To account for

these peculiarities, it is necessary to apply complicated substitution models (Moriyama et al. 1991), which only increases the uncertainty caused by model selection. For instance, Anderson et al. (2001) found that the best approximating substitution model for the gp120 region of *env* is a general time-reversible model (GTR) with eight parameters plus one parameter for gamma-distributed rate variation.

We take an alternative approach for measuring divergence. The temporal structure of longitudinal data allows us to infer multiple substitutions at individual sites by simple inspection of the time series. This approach is sound as long as sampling times are frequent enough so that multiple substitutions do not typically occur in between timepoints. We have developed a simple algorithm for determining when multiple substitutions have occurred between a given sequence and the progenitor sequence. Also, we have developed an algorithm for using the temporal information in the data to more accurately delineate between nonsynonymous and synonymous substitutions when more than one change has occurred in the same codon. These methods are described in detail in the *Supplementary Material*.

Data and Analysis Scheme for Each Patient

Nucleotide sequences in longitudinal data sets were obtained for all patients from the Shankarappa et al. (1999) study, except for patient 11, which was excluded because of smaller sample sizes and infrequent sampling. First, the progenitor sequence in each patient was approximated as the consensus sequence of the earliest sample. Next, the progenitor sequence and all of the sequences in the remaining timepoints were aligned using the default parameters of ClustalX (Thompson et al. 1997) and then hand corrected. In data sets from most patients, there were regions that could not be aligned for the whole data set; these regions were excluded from the analysis. In terms of the nucleotide position in the progenitor sequence (excluding gaps), these regions are p1 (574 to 588), p2 (382 to 411 and 577 to 597), p5 (573 to 579), p8 (382 to 387 and 565 to 585), and p9 (392 to 414 and 558 to 594). After alignment, nonsynonymous, synonymous, and total nucleotide divergence values were tabulated for each sequence using the methods outlined in the *Supplementary Material*. Our estimates of mean divergence (synonymous and nonsynonymous) within each patient are given as a function of time in figure 1. The individual sequence values were used in the series of statistical procedures described below. Nucleotide sequence alignments and the source code for the programs that implement the divergence tabulation are available from the corresponding author.

Probability Model for Divergence

We assume that the mean divergence of the entire population is governed by a diffusion process (Brownian motion) with drift parameter φ (the divergence rate) and diffusion parameter σ^2 (the variance in the divergence rate). Hereafter, we will refer to the mean divergence of the entire population as the "population divergence." In a small interval of time (Δt), the change in population divergence is a normally distributed random variable with expectation $\varphi^*\Delta t$ and variance $\sigma^2*\Delta t$. This model allows random fluctuations in the realized rate of evolution, even when the underlying rate is constant. The population divergence can even decline in some time intervals, although the whole process is subject to the boundary condition that divergence must be greater than or equal to zero for $t > 0$. Also, note that the variance parameter σ^2 is not directly related to any measure of within-population genetic variation; rather, it describes the variance in population divergence among replicate populations.

Our constant-rate model assumes that the divergence rate (φ) is constant throughout infection. This serves as a null model for testing divergence stabilization. Let $u(y, t)$ be the probability density function of the population divergence, y, at time t. u is governed by the following partial differential equation:

$$\frac{\partial u}{\partial t} = -\varphi \frac{\partial u}{\partial y} + \sigma^2 \frac{1}{2} \frac{\partial^2 u}{\partial y^2} \quad (1)$$

subject to the initial condition $u(y,0) = \delta(0)$, where δ is the Dirac delta function, and a reflecting boundary at $y = 0$. Our divergence-stabilization model involves one additional parameter, τ, which is the time of divergence stabilization. For $t < \tau$, the population divergence (y) follows a diffusion with drift parameter φ and variance σ^2. At time τ however, the drift parameter changes from φ to 0, whereas the diffusion parameter remains unchanged.

If we ignore the boundary condition, the solution of equation (1) is a normal density function. For the constant-rate model, the mean is φt and the variance is $\sigma^2 t$. The same variance is obtained for the divergence-stabilization model, but the mean is φt for $t < \tau$ and $\varphi\tau$ for $t > \tau$. We use this approximation to develop the maximum-likelihood framework described below. However, the boundary is explicitly included in the parametric bootstrapping simulations that we use to evaluate our likelihood-ratio tests.

Two sorts of stochasticity must be considered: evolutionary fluctuations and sampling variation. Let $i = 0, 1, 2, \ldots k$ denote the samples taken at different timepoints, in chronological order. In each patient, we use sample 0 to infer the progenitor sequence of all subsequent samples. Early in infection, the viral population is genetically homogeneous within the *env* gene (Zhang et al. 1993). Therefore, we approximate the progenitor sequence as the consensus sequence of sample 0. For samples 1 through k, let t_i denote the time (since seroconversion) that the sample was taken. Let y_i denote the (unobserved) population divergence at time t_i, measured as the number of differences per site from the progenitor. Let n_i be the sample size of each timepoint. Let x_{ij} denote the divergence of sequence j in sample i, in terms of the number of differences per site from the progenitor. Conditional on each y_i, we assume that the x_{ij} are also normally distributed:

$$x_{ij} \mid y_i \sim N(y_i, s_i^2) \quad (2)$$

where s_i^2 is the variance in divergence among viruses within the population at the ith sampling time. To obtain the unconditional distribution of x_{ij}, we integrate over all y_i:

$$P[x_{ij}] = \int_{-\infty}^{\infty} P[x_{ij} \mid y_i] P[y_i] dy_i \sim N(\varphi t_i, \sigma^2 t_i + s_i^2) \quad (3)$$

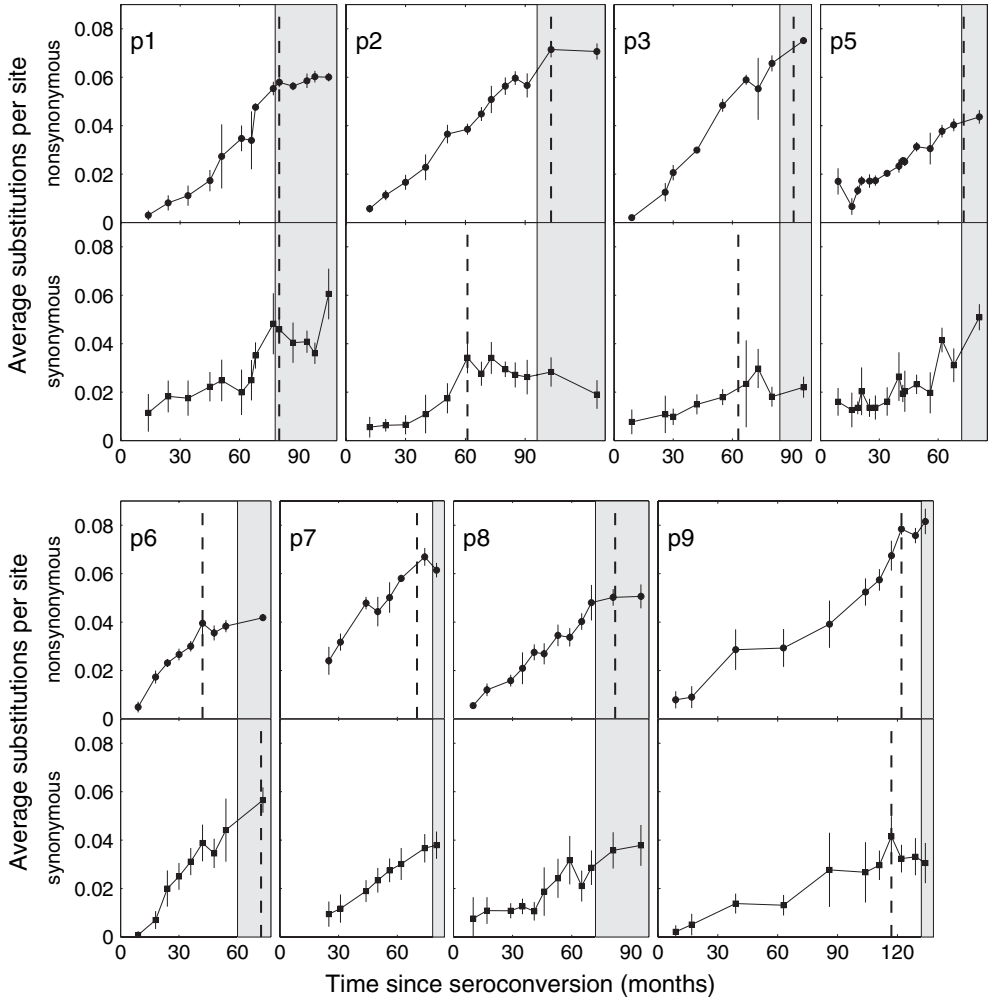


FIG. 1.—Average nonsynonymous and synonymous divergence as a function of time since seroconversion for viral populations from nine patients. Vertical dashed lines mark maximum-likelihood estimates (MLEs) for the time of divergence stabilization (τ); plots without a vertical dashed line indicate that the MLE of τ is greater than the time of the last sample. Error bars represent 95% confidence intervals on the mean, obtained using the sample variances at each timepoint. Regions shown in gray indicate the time of disease progression; i.e.' the time after the $CD4^+$ T cell count drops below 300 cells/ml (clinical AIDS). For comparison, all plots are drawn on the same scale.

The joint distribution of all x_{ij} from a single longitudinal data set follows a multinormal distribution. The covariance between two observations in the same sample is

$$\mathrm{Cov}[x_{ij_1}, x_{ij_2}] = \mathrm{Var}[y_i] = \sigma^2 t_i \qquad (4)$$

where $j_1 \neq j_2$. The covariance between observations at different timepoints is

$$\mathrm{Cov}[x_{i_1 j_1}, x_{i_2 j_2}] = \mathrm{Cov}[y_{i_1}, y_{i_2}] = \begin{cases} \sigma^2 t_{i_1} & i_1 < i_2 \\ \sigma^2 t_{i_2} & i_2 < i_1 \end{cases} \qquad (5)$$

There is a well-developed theory for estimation of parameters from multinormal data (Searle, Casella, and McCulloch 1992). Let $\mathbf{x}$ denote the vector of the x_{ij}'s, with the observations from different timepoints concatenated sequentially. The constant-rate model contains $k+2$ parameters: k population variances (s_1^2, s_2^2,..., s_k^2) and the drift (φ) and diffusion (σ^2) parameters of equation 1. Under this model, the log-likelihood is

$$\ell_{CR}(\varphi, \sigma^2; s_1^2, s_2^2, \ldots, s_k^2 \mid \mathbf{x}) = -\frac{n}{2}\mathbf{ln}(2\pi) - \frac{1}{2}\mathbf{ln}|\mathrm{V}| - \frac{1}{2}(\mathbf{x}-\mathbf{q})^{\mathrm{T}}\mathrm{V}^{-1}(\mathbf{x}-\mathbf{q}) \qquad (6)$$

where $\mathbf{q}$ is the vector of expected values of each of the x_{ij}'s ($E[x_{ij}] = \varphi t_i$), V is the variance-covariance matrix containing the terms described in equations (4) and (5), n is the total number of sequences sampled in all of the timepoints, and $|*|$ denotes the determinant of a matrix.

The divergence-stabilization model contains $k+3$ parameters: k population variances, φ, σ^2, and τ. The only effect this new parameter has on the likelihood function is within the vector of expected values. Call this new vector, $\mathbf{p}$:

$$p_{ij} = E[x_{ij}] = \begin{cases} \varphi t_i & t_i \leq \tau \\ \varphi \tau & t_i > \tau \end{cases} \qquad (7)$$

Again, the vectors of expected values are concatenated sequentially. The log-likelihood function for the divergence-stabilization model is

$$\ell_{DS}(\varphi, \sigma^2, \tau; s_1^2, s_2^2, \ldots, s_k^2 \mid \mathbf{x}) = -\frac{n}{2}\mathbf{ln}(2\pi) - \frac{1}{2}\mathbf{ln}|\mathrm{V}| - \frac{1}{2}(\mathbf{x}-\mathbf{p})^{\mathrm{T}}\mathrm{V}^{-1}(\mathbf{x}-\mathbf{p}) \quad (8)$$

Numerical values for the log-likelihoods (equations 6 and 8) were determined using a program written in the C programming language and executed on a series of Dell Pentium 4 computers. We used the standard estimator for the variance of a normal population for the s^2 value of each sample (page 52 in Sokal and Rohlf [1995]). For fitting the constant-rate model to the data, we conducted a guided search for the maximum-likelihood estimates of φ and σ^2 using the method of scoring (chapter 8 in Searle, Casella, and McCulloch [1992] and Kelly and Arathi [2003]).

To fit the divergence-stabilization model for a single data set, we conducted a series of optimizations over the range of possible values for τ. We started with the largest possible value for τ, which is t_k, then maximized the log-likelihood function with a fixed value of τ; that is, the profile log-likelihood. From this optimization, we recorded the maximum-likelihood estimates for φ and σ^2 and the maximum log-likelihood value. We then reduced the value of τ by 1 month and maximized the profile log-likelihood again by finding optimal values for φ and σ^2. This process was repeated with each successive iteration considering a lower value of τ. An example of this procedure, the actual analysis of nonsynonymous evolution in patient 1, is given in figure 2. Thus, for both synonymous and nonsynonymous evolution within each patient, the model-fitting procedure yields a set of log-likelihood values defined for each possible value of τ.

For both the constant-rate and divergence-stabilization models, our search procedure rapidly converged on the same maximum-likelihood estimates from different start values. However, it may be possible to develop more efficient search algorithms (perhaps necessary for larger data sets) by noting that both models are essentially "Gaussian state-space models" (Durbin and Koopman 2001). Such models have been used extensively for the analysis of time series data in economics and engineering. The Kalman filter might be adapted to provide a general framework for both prediction-based and likelihood-based parameter estimation of models such as those described here (chapter 3 in Harvey [1989]). The fact that we did not incorporate the boundary condition in equations 6 and 8 implies that our log-likelihood values are only approximate. However, this should not impact the hypothesis testing described in the next section. The same assumption applies to both the constant-rate and divergence-stabilization models, and the relative fits of these models depend on sequence data from late in infection, by which time the boundary is irrelevant.

Hypothesis Testing

It is possible to construct a test within each patient by comparing the maximum likelihoods associated with each model. An evolutionary slowdown is indicated if the maximum likelihood for the divergence-stabilization model is sufficiently greater than the maximum likelihood for the constant-rate model. Because the constant-rate model is nested within the divergence-stabilization model (the divergence-stabilization model reduces to the constant-rate model for $\tau = t_k$), a likelihood-ratio test (LRT) is appropriate. The LRT statistic is $2\{\ell_{DS}(\hat{\varphi}_{DS},\hat{\sigma}^2_{DS},\hat{\tau}_{DS};s_1^2,s_2^2,\ldots,s_k^2 \mid \mathbf{x}) - \ell_{CR}(\hat{\varphi}_{CR},\hat{\sigma}^2_{CR};s_1^2,s_2^2,\ldots,s_k^2 \mid \mathbf{x})\}$, where $\hat{\varphi}_{DS}$, σ^2_{DS}, and $\hat{\tau}_{DS}$ are the maximum-likelihood estimates (MLEs) under the divergence-stabilization models, and $\hat{\varphi}_{CR}$ and σ^2_{CR} are the MLEs under the constant-rate model. We calculated this statistic for both nonsynonymous and synonymous evolution within each patient.

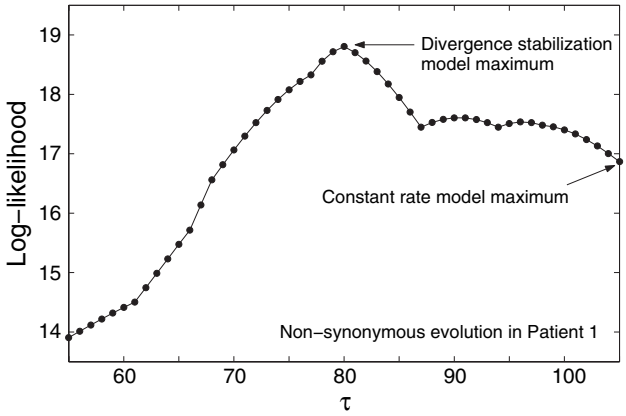


FIG. 2.—The maximum profile log-likelihood over a range of values for τ, the time of divergence stabilization. This plot depicts nonsynonymous divergence in patient 1.

Both evolutionary variability and small sample sizes limit the power of tests based on individual patients. Thus, to distinguish immune relaxation from cellular exhaustion, we also performed two analyses that combine data across patients. These combined analyses are possible because evolution in different patients is independent. As a consequence, we can sum the (maximum) log-likelihoods across patients for the different evolutionary models. This allows us to compare the constant-rate and divergence-stabilization models with combined LRTs using all the nonsynonymous or synonymous data simultaneously. If divergence stabilization occurs in all patients, these combined LRTs will have more power to detect it. In the combined analyses, the log-likelihood value for the constant-rate model is simply the sum of the maximum log-likelihood values for this model within each patient.

We consider two different versions of divergence stabilization as alternatives to the constant-rate model. The first version posits that divergence stabilization is directly related to disease progression. In each patient, divergence stabilization occurs at time $T_p + \alpha$, where T_p is the time of disease progression, and α is constant across all patients. We define T_p as the first time at which the patient's $CD4^+$ T cell count dropped below 300 cells/ml. Hereafter, we will refer to this model as the single-α divergence-stabilization model. The maximum log-likelihood for the model is obtained by finding the value for α that maximizes the sum of patient-specific log-likelihoods with $\tau = T_p + \alpha$ within each patient. The second variant of the divergence-stabilization model, which we refer to as the free-τ model, allows the time of divergence stabilization (τ) to vary among patients without constraint.

If the constant-rate model were true, we might expect the likelihood-ratio test statistic to follow a chi-square distribution, with the number of degrees of freedom equal to the difference in the number of parameters between the models that are being compared. In the present context, the comparison between the single-α divergence-stabilization model and the constant-rate model would involve a single degree of freedom because the more complicated model adds only one parameter (α). The degrees of freedom for the free-τ model would be equal to the number of patients, because τ is estimated separately in each patient. Unfortunately, the asymptotic theory that justifies use of the chi-square distribution for evaluating LRT statistics is not applicable in this case. The parameters σ^2 and τ are close to the edges of their respective ranges, and the sample sizes in each timepoint are relatively small ($\sim$10 to 20 sequences). Therefore, we use parametric bootstrapping to approximate the null distributions of our LRT statistics (Davison and Hinkley 1997).

It is straightforward to simulate divergence under either the constant-rate or the divergence-stabilization model for any particular set of parameter values. First, we simulated the population divergence recursively over small time steps ($\Delta t = 0.1$ months): starting at $y = 0$ at time 0, the population divergence at the next time step, $y(t + \Delta t)$, is drawn from a normal distribution with mean $y(t) + \varphi * \Delta t$ and variance $\sigma^2 * \Delta t$. To incorporate the boundary, negative values were reset to 0. Next, for the sampling times (t_i) in each particular patient, the divergence values for individual sequences were simulated following equation 2 using the simulated population divergences, where the sampling variance was estimated from the data. For the analyses that consider data from each patient separately, we used our maximum-likelihood estimates for φ and σ^2 from the fitted constant-rate model to produce a set of 500 simulated data sets for each patient in the Shankarappa et al. (1999) study. We then fitted both models to each simulated data set, recorded the maximum log-likelihoods for each model, and calculated the LRT (see Kelly [2003] for an application of the parametric bootstrap to multinormal data). The null distributions for our single-α and free-τ LRTs were obtained by combining simulation results from the individual patients. We repeatedly reconstituted whole data sets (involving eight patients) by randomly selecting simulated data sets from each patient. The combined LRTs were calculated for these reconstituted data sets and the values recorded. This procedure was repeated 10,000 times to estimate the distributions.

Forward Simulations

In developing our constant-rate and divergence-stabilization models, we have made two main assumptions: (1) the mean divergence of the entire population is accurately described by a diffusion process, and (2) the boundary condition at $y = 0$ does not strongly affect our method of inference. To assess the validity of these assumptions under different evolutionary scenarios, we performed forward population-genetic simulations of sequence evolution. These simulations incorporated natural selection, recombination, and mutation.

Each iteration proceeded as follows: we started with a genetically homogeneous population of $2N$ sequences, each of length L. This formed the ancestral population. Reproduction proceeded according to a Wright-Fisher model for diploids—that is, with nonoverlapping generations, random mating, and a constant population size. Mutation occurred at rate μ per sequence (not per base pair) per generation, and the mutation rate was uniform over the whole sequence. When a mutation occurred at a previously mutated site, it reverted back to the ancestral state. Recombination occurred at rate r per sequence per generation, and recombination was modeled as a crossover process occurring only at homologous sites within diploid individuals. We did not allow insertions or deletions. For the simulations with natural selection, a certain proportion of new mutations, p_s, were subject to selection, and the remainder were neutral. The selection coefficient s describes the relative advantage of the new mutation over the ancestral allele at that site. The selective effects of multiple mutations within a sequence were combined multiplicatively. At 10-month intervals (the approximate interval between sampling in the Shankarappa et al. [1999] study), the mean divergence of the entire population was tabulated relative to the ancestral population, not correcting for back mutation. This was repeated up to 100 months. For each parameter combination, the diploid population size and sequence length were held constant ($N = 1000$, $L = 650$), and we assumed that one generation occurred every 2 days (Rodrigo et al. 1999, Fu 2001, Drummond et al. 2002). The mutation rate, μ, was selected such that the sequences diverged by approximately 1% per year, which is typical of the Shankarappa et al. (1999) data. The simulations were iterated 500 times for each parameter combination.

Figure 3 shows the results of the forward simulations for several different combinations of recombination and selection parameters. The (unbounded) diffusion models presented above predict that, among replicate populations (1) both the mean and variance of population divergence should increase linearly with time, (2) at any timepoint, the population divergence should be normally distributed, and (3) because of the memoryless property of the diffusion, the covariance in population divergence among two different timepoints should equal the variance of the earlier timepoint. Figures 3*c* and *d* indicate that prediction (1) holds for different recombination rates and selection regimes: both the mean and variance increase linearly. Figure 3*a* shows the distribution of population divergence at different timepoints for the case of neutral evolution with a high recombination rate, along with a fitted normal distribution. The normality prediction seems to hold very well. The same pattern is observed for the other parameter combinations, including those with selection. Finally, figure 3*b* plots the covariance between timepoints as a function of the variance of the earlier timepoint. The agreement with prediction (3) is quite good. Taken together, our forward simulations indicate that, for a range of evolutionary models, mean population divergence is very well approximated by a diffusion, and our assumption regarding the boundary does not have a strong affect on our probability model.

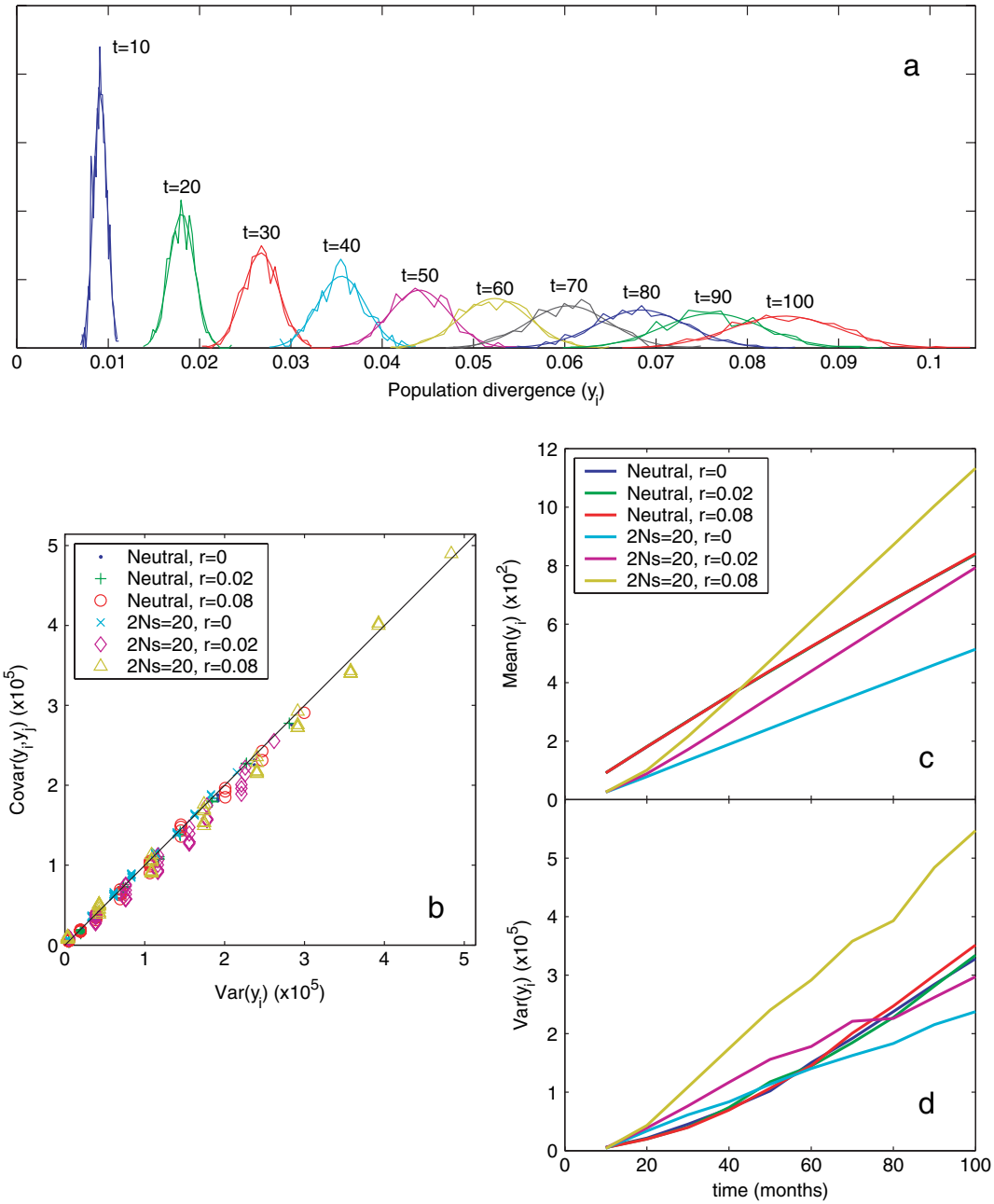


FIG. 3.—Forward population-genetic simulations of sequence evolution and the predictions of the diffusion model for population divergence. (*a*) The distribution of population divergence (mean divergence of the entire population from a progenitor) at different timepoints, along with fitted normal distributions, for the neutral simulation with a high recombination rate. (*b*) The covariance in population divergence between two timepoints, plotted as a function of the variance of the earlier timepoint. The diffusion models predict that these measures should be equal. (*c*) The mean population divergence across replicates as a function of time. (*d*) The variance in population divergence across replicates as a function of time. Under the constant-rate diffusion model, mean population divergence and variance in population divergence should increase linearly with time.

Results

Figure 1 suggests that (1) in some patients, nonsynonymous divergence seems to stabilize, whereas synonymous divergence does not, (2) nonsynonymous divergence stabilization seems to coincide with disease progression, and (3) the rate of nonsynonymous divergence tends to be higher than the rate of synonymous divergence. These observations are confirmed by our statistical analysis. Parameter estimates, LRT statistics, and *P* values from our combined analyses (free-τ and single-α) are shown in table 1. Our most striking result is that, in the free-τ and single-α analyses, we find strong evidence for divergence stabilization at nonsynonymous, but not synonymous, sites. For instance, in our free-τ analysis, we reject the constant-rate model ($P = 0.0014$) for nonsynonymous sites, but we find little evidence for divergence stabilization at the synonymous sites ($P = 0.11$). Results from individual analyses are shown in table 2. When the data were not combined across patients, we lacked sufficient power to reject the

Table 1
Analyses of Synonymous and Nonsynonymous Divergence Using Combined Likelihood Ratio Tests That Compare Divergence-Stabilization Models to the Null, Constant-Rate Model Using Data from All Patients

Site	Model	Log-Likelihood	LRT	*P*
syn	CR	−650.02		
	single-α DS ($\hat{\alpha} = 27$)[a]	−649.92	0.20	0.55
	free-τ DS	−641.21	17.63	0.11
non	CR	22.19		
	single-α DS ($\hat{\alpha} = 7$)[a]	34.03	23.67	0.0002
	free-τ DS	41.45	38.52	0.0014

NOTE.—CR = constant rate, DS = divergence stabilization, LRT = likelihood-ratio test, non = nonsynonymous, syn = synonymous.

[a] $\hat{\alpha}$ is the maximum-likelihood estimate of the time of divergence stabilization relative to the time of disease progression across patients.

constant-rate model in most data sets. However, in six of eight patients, the LRT statistic of the nonsynonymous data set is higher than that of the synonymous data set.

We use our single-α model to explore the relationship between divergence stabilization and disease progression. We find strong evidence that disease progression and divergence stabilization at nonsynonymous sites are tightly associated but no evidence of a similar relationship at synonymous sites. We reject the constant-rate model ($P = 0.0002$) in favor of the single-α model for nonsynonymous sites, but we find no evidence ($P = 0.55$) for divergence stabilization at the synonymous sites. Furthermore, when we estimate the nonsynonymous divergence stop time (τ) in each patient individually, we notice a striking, and apparently curvilinear, relationship between our estimates and the time of disease progression (fig. 4); however, there is no discernable relationship between the synonymous stop time and disease progression (fig. 4).

Because our drift parameter, φ, represents the rate of evolution per site, we can use our estimates of φ for nonsynonymous and synonymous sites to estimate the d_N/d_S ratio—that is, the ratio of nonsynonymous changes per nonsynonymous site to synonymous changes per synonymous site. Assuming that synonymous changes are neutral, the d_N/d_S ratio is a statistic that summarizes the relative contributions of deleterious, neutral, and adaptive mutations to the substitution process. The value $d_N/d_S > 1$ is considered a clear indicator of widespread adaptive evolution (e.g., Nielsen and Yang 1998), whereas $d_N/d_S < 1$ implies at least some sites are selectively constrained. Estimates of the d_N/d_S ratios for viral populations in each patient are shown in table 3. For most data sets, $d_N/d_S > 1$, indicating widespread positive selection. In six out of the eight patients, estimates of d_N/d_S obtained under the divergence-stabilization model were higher than estimates obtained assuming a constant evolutionary rate. Also, we observe less variation across patients when d_N/d_S is estimated under the divergence-stabilization model. Notably, our estimates of the d_N/d_S ratio are highly correlated with the time of disease progression across patients ($r = 0.5997$, $P = 0.0003$), i.e., the signal of positive selection is stronger in patients with longer progression times. This result is fully consistent with previous analyses that found a similiar pattern using phylogenetic (Ross and Rodrigo 2002) and population genetic (Williamson 2003) methods for detecting adaptive evolution, each of which involves a very different set of biological assumptions than the present analysis.

Table 2
Maximum-Likelihood Estimates of the Mean Divergence Rate (φ), the Variance in Divergence (σ^2), and the Divergence Stop Time (τ) in Each Patient Under the Constant-Rate Model and the Divergence-Stabilization Model, and Likelihood-Ratio Tests Comparing the Two Models

Patient	Site	Model	$\hat{\varphi}$ ($\times 10^4$)	$\hat{\sigma}^2$ ($\times 10^6$)	$\hat{\tau}$[a]	Log-Likelihood	LRT	*P*
p1	syn	CR	5.01	5.27		−87.67		
		DS	5.01	5.27	t_k	−87.67	0	0.57
	non	CR	5.73	2.18		16.87		
		DS	7.21	1.48	80	18.80	3.86	0.11
p2	syn	CR	1.64	2.36		−126.37		
		DS	4.77	0.86	61	−122.99	6.75	0.07
	non	CR	5.68	1.31		−22.36		
		DS	6.86	0.16	103	−16.57	11.59	0.009
p3	syn	CR	2.39	0.42		−44.79		
		DS	3.38	0	63	−40.45	8.69	0.03
	non	CR	7.82	1.54		22.76		
		DS	8.26	1.38	91	23.04	0.56	0.30
p5	syn	CR	6.17	3.81		−121.02		
		DS	6.17	3.81	t_k	−121.02	0	0.54
	non	CR	5.45	0.90		46.30		
		DS	5.98	0.38	73	47.11	1.61	0.21
p6	syn	CR	7.75	1.89		−59.55		
		DS	7.84	1.89	72	−59.55	0.01	0.52
	non	CR	5.74	1.52		35.64		
		DS	8.80	0.51	42	39.27	7.25	0.07
p7	syn	CR	4.70	0		−63.69		
		DS	4.70	0	t_k	−63.69	0	0.47
	non	CR	7.79	3.29		−27.41		
		DS	9.35	1.08	70	−24.91	4.99	0.09
p8	syn	CR	3.82	0		−87.53		
		DS	3.82	0	t_k	−87.53	0	0.51
	non	CR	6.00	0		−6.78		
		DS	6.13	0	82	−2.64	8.28	0.002
p9	syn	CR	2.66	0		−59.42		
		DS	2.77	0	117	−58.33	2.16	0.11
	non	CR	6.04	3.00		−42.84		
		DS	6.37	2.85	122	−42.65	0.36	0.44

NOTE.—CR = constant rate, DS =divergence stabilization, LTR = likelihood-ratio test, non = nonsynonymous, syn = synonymous.

[a] $\hat{\tau} = t_k$ indicates that the maximum-likelihood estimate of τ was the time of the last sample.

Discussion

Our analyses suggest that divergence stabilization—that is, the reduction in evolutionary rate coincident with disease progression—is caused primarily by relaxation of positive selection rather than by an increase in the average generation time. We reject the constant-rate model for nonsynonymous sites but not for synonymous sites. Both the single-α analysis and the separate analyses of individual patients suggest a strong relationship between the time of disease progression and the time of divergence stabilization for nonsynonymous sites but not for synonymous sites. These results are predicted by the immune-relaxation

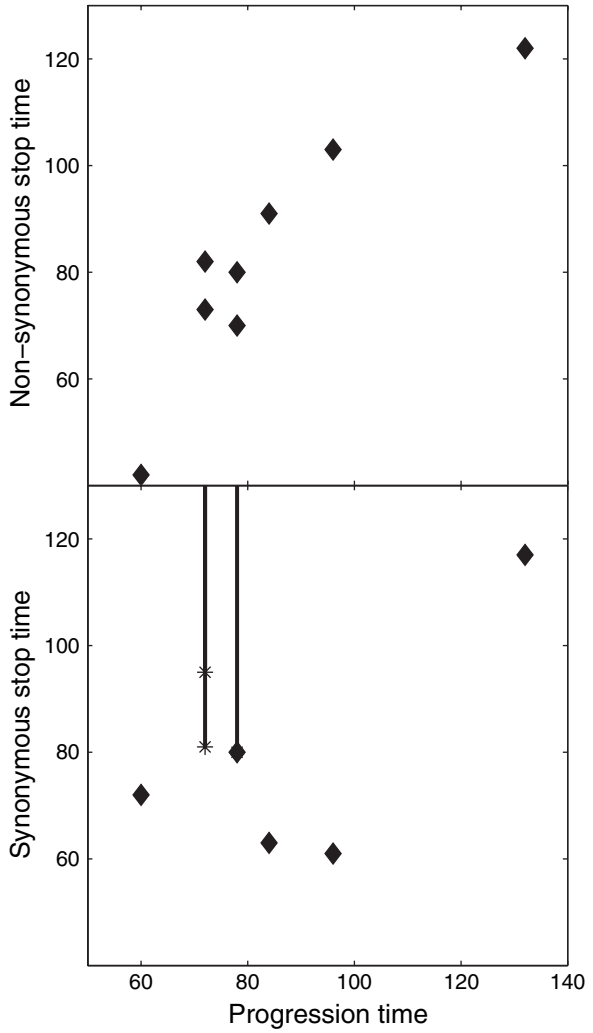


FIG. 4.—The relationship between the time of disease progression (T_p) and our maximum-likelihood estimates for the time of divergence stabilization (τ) at both nonsynonymous and synonymous sites. There is a strong and apparently curvilinear relationship between disease progression and divergence stabilization at nonsynonymous sites, whereas there is no obvious relationship between progression and divergence stabilization at synonymous sites. Vertical lines connected to stars indicate that the MLE for the stop time is greater than the time plotted at the star—that is, divergence stabilization had not occurred by the time of the last sample.

hypothesis, and they are inconsistent with the cellular-exhaustion hypothesis.

The support for the immune-relaxation hypothesis in the present set of analyses corroborates several common ideas about HIV evolution. First, our results are fully consistent with the notion that the immune system is the primary agent of positive selection in the C2-V5 region of *env* (Bonhoeffer, Holmes, and Nowak 1995; Wolinsky et al.1996; Ross and Rodrigo 2002). Second, our results suggest that HIV disrupts those agents of the immune system that play a critical role in controlling infection. Immunological studies support the existence of such a feedback loop. Douek et al. (2002) found that the virus preferentially infects HIV-specific $CD4^+$ T cells. Because HIV-specific $CD4^+$ T cells are thought to play an important role in controlling viremia (e.g., Rosenberg et al. 1997), Douek et al. (2002) proposed that this viral specificity contributes to the uncontrolled viral replication that coincides with progression to AIDS. Our analyses further suggest that the interaction between the virus and HIV-specific immune effectors, such as HIV-specific $CD4^+$ T cells, alters the course of viral evolution.

Table 3
Ratios of d_N/d_S Estimated As the Ratio of Estimated Divergence Rate Parameters ($\hat{\varphi}_N/\hat{\varphi}_S$) for the Constant-Rate and Divergence-Stabilization Models

Patient	d_N/d_S (CR)	d_N/d_S (DS)
1	1.15	1.44
2	3.46	1.44
3	3.28	2.44
5	0.88	0.97
6	0.74	1.12
7	1.66	1.99
8	1.57	1.61
9	2.27	2.30

NOTE.—CR = constant rate, DS =divergence stabilization.

In a previous analysis, Bonhoeffer, Holmes, and Nowak (1995) observed a decrease in the d_N/d_S ratio of the *env* gene V3 loop over time within a single infected patient. They attributed this to a reduction in selection intensity over the course of infection. However, Nielsen (1999) pointed out that such a change could also be caused by the nonequilibrium dynamics of an evolving virus population. Because the population is initially genetically homogeneous, it takes some time for new mutations to spread to detectable frequencies, where they could be counted as substitutions. This transition time will be longer for neutrally evolving synonymous mutations than for positively-selected nonsynonymous mutations, so the d_N/d_S ratio would decrease over time to some equilibrium, without a change in selection intensity. In contrast, we do not expect nonequilibrium dynamics to have a large effect on our analyses. We follow divergence at each type of site separately. The founder effect of initial infection will cause the divergence rate at each type of site to increase monotonically until it reaches some equilibrium. It will not cause the divergence rate to drop back to 0, as is the case in our divergence-stabilization model. Therefore, our test for a decrease in the divergence rate is conservative with respect to the founder effect.

Relation to Other Methods

Our diffusion-based analysis is complementary to three other approaches that have been taken in analyzing longitudinal data sets: phylogenetic methods, coalescence theory, and regression models. Phylogenetic methods allow powerful inferences regarding natural selection and other population processes. Indeed, a phylogenetic tree has become the de facto method for summarizing the information in a longitudinal data set (e.g., Wolinsky et al. 1996; Ganeshan et al. 1997; Markham et al. 1998; Shankarappa et al. 1999; Yamaguchi-Kabata and Gojobori 2000), and phylogenetic methods based on codon-substitution models (Nielsen and Yang 1998) have been

used to identify widespread positive selection in longitudinal data sets (e.g., Zanotto et al. 1999, Ross and Rodrigo 2002). A potential difficulty in applying phylogenetic methods to longitudinal data sets is that recent studies suggest that the recombination rate of HIV populations in vivo is high (Jung et al. 2002, Wain-Hobson et al. 2003). Phylogenetic methods generally assume no recombination. Although the effects of violating the no-recombination assumption are not fully understood, simulations have shown that even small amounts of recombination have a large effect on tree statistics and can exaggerate of the degree of substitution-rate heterogeneity (Schierup and Hein 2000). Also, unrecognized recombination can cause codon-based phylogenetic methods to falsely identify positive selection (Shriner et al. 2003; Anisimova, Yang, and Nielsen 2003). In general, results from phylogenetic analyses of longitudinal data sets can be difficult to interpret in the presence of recombination.

A second approach taken in analyzing longitudinal data sets is coalescence theory (Kingman 1982; Hudson 1983), which has been adapted specifically for longitudinal data sets ("the serial coalescent" of Rodrigo and Felsenstein [1999]). Because the serial coalescent is based on a population process, it can be used to estimate important population parameters, such as the within-patient effective population size (Leigh-Brown 1997, Seo et al. 2002, Drummond et al. 2002) and the generation time (Rodrigo et al. 1999, Fu 2001, Drummond et al. 2002). The coalescent approach makes two assumptions that may be violated in the *env* gene. First, current implementations of the serial coalescent assume no recombination. Second, the serial coalescent assumes neutral evolution, whereas natural selection is thought to play an important role in *env*. Modifications to the coalescent that allow for selection (e.g., Neuhauser and Krone 1997, Barton and Etheridge 2004) are not applicable to the type of selection common in HIV: strong positive and diversifying selection at many linked sites.

Regression modeling is a third approach that has been taken in analyzing longitudinal data from HIV-infected patients. Mayer-Hamblett and Self (2001) describe a set of regression models for characterizing genetic diversity within and among samples of sequences. Applying these models to the Shankarappa et al. (1999) data set, they conclude that viral diversity (variation among sequences within a single sample) increases over most of the course of infection but declines at the end. However, in contrast to the analysis presented here, they failed to confirm a stabilization of divergence towards the end of infection. The differing conclusions of our study and that of Mayer-Hamblett and Self (2001) probably stem, in part, from differences in our respective modeling approaches. However, the most important difference is likely that we distinguish synonymous and nonsynonymous changes. Only the latter seems to provide compelling evidence of divergence stabilization.

By following the mean population divergence from the progenitor sequence rather than the genealogy of all sequences, our analysis is fairly permissive in terms of population processes such as recombination and natural selection. Our null model requires only that the underlying rate of substitution is approximately constant through time. For example, consider the case of two linked loci subject to neutral evolution, sampled in a single individual. In this situation, past recombination events will not affect the number of substitutions at either locus (or the total number of substitutions), simply because each locus is expected to experience the same number of generations whether or not recombination has occurred. Also, consider the case of frequent selective sweeps—that is, rapid fixation events of advantageous mutations. If the rate of advantageous mutation is constant, then the rate of substitution will also be constant. Population-genetic simulations support these arguments: the diffusion model seems to provide an adequate description of mean population divergence through time even when both recombination and natural selection are operating.

The sufficiency of the diffusion models stems largely from the fact that we are considering only one aspect of the complex data structure from longitudinal studies. The models make no prediction about the pattern of variation among sequences within a single sample, either in the extent to which they differ from the progenitor or from each other. Such measures of viral diversity exhibit large fluctuations over the course of infection, even within a single patient (Shankarappa et al. 1999; Mayer-Hamblett and Self 2001). Thus, one advantage of our approach is that it is relatively insensitive to these large-scale fluctuations. One disadvantage, however, is that such fluctuations in viral diversity may play an important role in disease progression (Nowak et al. 1991), and our models cannot address this problem. A major challenge in the analysis of longitudinal data sets lies in developing methods that characterize changes in divergence, diversity, and the relationship between the two.

Acknowledgments

We thank Mark Jensen, Dan Shriner, and three anonymous reviewers for helpful comments and suggestions on the manuscript. This work was supported by NIH grant 1 R01 GM60792-01A1 to J.K.K. and M.E.O. and a grant from the Cornell Genomics Initiative to C.D.B.

Supplementary Material

To correct for saturation, we use the temporal information in longitudinal data to infer when multiple substitutions have occurred at the same site. We propose the following algorithm:

1. *Identify sites at which more than one derived (non-ancestral) nucleotide is observed in the entire longitudinal data set.*
2. *At each of those sites, if one of the derived nucleotides is first observed before the other(s)*:

   a. *Count all sequences containing the first derived nucleotide as one substitution away from the progenitor at that site.*
   b. *Let t_{init} be the time that the second derived nucleotide is first observed, and let s_2 be the set of all sequences*

from t_{init} that have the second derived nucleotide. Then find the minimum number of pairwise differences between s_2 and all the sequences in all the previous timepoints. Call the minimum-distance sequence from the previous timepoints m.

c. *If m contains the first derived nucleotide at the site in question, count each subsequent sequence that contains the second derived nucleotide as two substitutions from the progenitor at the site.*

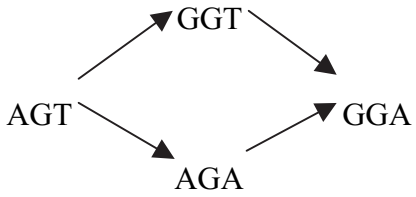
d. *Otherwise, count each sequence with the second derived nucleotide as one substitution from the progenitor at the site.*

e. *If a third derived nucleotide is present in the longitudinal data set, repeat* b to d *for this character state.*

3. *Alternatively, if both* (*or all three*) *derived nucleotides are first observed at the same sampling time, all sequences containing each derived nucleotide are counted as just one substitution away from the progenitor at that site.*

To summarize, when the second derived nucleotide is first observed, we infer that it has arisen from whatever character state is found in the closest (in terms of pairwise differences) previous sequence. This routine does not account for back mutation.

In addition to correcting for multiple substitutions, we can also use the temporal information in longitudinal data sets to more accurately delineate between nonsynonymous and synonymous substitutions. A common problem in DNA sequence analysis lies in counting the numbers of nonsynonymous and synonymous substitutions when there have been multiple changes in a single codon. Consider the following example for the changes in a particular codon: AGT (Ser) → GGA (Gly). Assuming that only two changes have occurred, the potential intermediaries are:



If AGA is the true intermediate, then there have been two nonsynonymous changes, but if GGT is the true intermediate, then there have been one synonymous and one nonsynonymous change. There are several possible ways to weight the likelihood of the different paths, and, consequently, different sequence analysis programs can produce different counts for the numbers of nonsynonymous and synonymous sites. However, with longitudinal data sets, when multiple changes have occurred in the same codon, we can simply look back at previous timepoints and infer which path was taken. We propose the following algorithm for counting nonsynonymous and synonymous changes when there have been two changes in the same codon and when the different intermediaries produce different counts:

1. *For each codon, identify sequences that differ from the progenitor at two positions within the codon AND the different intermediate codons produce different counts.*
2. *In the sequences in previous timepoints, look for the potential intermediaries at those codons.*
3. *If only one of the two potential intermediaries are present in all of the previous timepoints, infer that it is the intermediate character state, and count nonsynonymous changes accordingly.*
4. *If none or both of the potential intermediaries are present, weight the two different paths as equally likely.*

Literature Cited

Anderson, J. P., A. G. Rodrigo, G. H. Learn, Y. Wang, H. Weinstock, M. L. Kalish, K. E. Robbins, L. Hood, and J. I. Mullins. 2001. Substitution model of sequence evolution for the human immunodeficiency virus type 1 subtype B gp120 gene over the C2-V5 region. J. Mol. Evol. **53**:55–62.

Anisimova, M., R. Nielsen, and Z. Yang. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164**:1229–1236.

Balfe, P., P. Simmonds, C. A. Ludlam, J. O. Bishop, and A. J. Brown. 1990. Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. J. Virol. **64**:6221–6233.

Barton, N. H., and A. M. Etheridge. 2004. The effect of selection on genealogies. Genetics **166**:1115–1131.

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

Davison, A. C., and D. V. Hinkley. 1997. Bootstrap methods and their application. Cambridge University Press, Cambridge, UK.

Douek, D. C., J. M. Brenchley, M. R. Betts et al. (15 co-authors). 2002. HIV preferentially infects HIV-specific $CD4^+$ T cells. Nature **417**:95–98.

Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, and W. Solomon. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161**:1307–1320.

Durbin, J., and S. J. Koopman. 2001. Time series analysis by state space methods. Oxford University Press, New York.

Finzi, D., J. Blankson, J. D. Silicano et al. (16 co-authors). 1999. Latent infection of CD4 T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. Nat. Med. **5**:512–517.

Finzi, D., M. Hermankova, T. Pierson et al. (15 co-authors). 1997. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. Science **278**:1295–1300.

Fu, Y.-X. 2001. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. Mol. Biol. Evol. **18**:620–626.

Ganeshan, S., R. E. Dickover, B. T. M. Korber, Y. J. Bryson, and S. M. Wolinsky. 1997. Human immunodeficiency virus type 1 genetic evolution in children with different rates of development of disease. J. Virol. **71**:663–677.

Goulder, P. J. R., C. Brander, Y. Tang et al. (19 co-authors). 2001. Evolution and transmission of stable CTL escape mutations in HIV infection. Nature **412**:334–338.

Graur, D., and W.-H. Li. 2000. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Mass.

Harvey, A. C. 1989. Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge, UK.

Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham. 1994. Application and accuracy of molecular phylogenies. Science **264**:671–677.

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

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23**:183–201.

Jung A, R. Maier, J. P. Vartanian, G. Bocharov, V. Jung, U. Fischer, E. Meese, S. Wain-Hobson, and A. Meyerhans. 2002. Multiply infected spleen cells in HIV patients. Nature **418**:144.

Kelly, J. K. 1994. An application of population genetic theory to synonymous gene sequence evolution in the human immunodeficiency virus (HIV). Genet. Res. **64**:1–9.

Kelly, J. K. 1996. Replication rate and evolution in the human immunodeficiency virus. J. Theor. Biol. **180**:359–364

Kelly, J. K. 2003. Deleterious mutations and the genetic variance of male fitness components in *Mimulus guttatus*. Genetics **164**:1071–1085.

Kelly, J. K., and H. S. Arathi. 2003. Inbreeding and the genetic variance in floral traits of *Mimulus guttatus*. Heredity **90**: 77–83.

Kelly, J. K., S. Williamson, M. E. Orive, M. S. Smith, and R. D. Holt. 2003. Linking dynamical and population genetic models of persistent viral infection. Am. Nat. **162**:14–28.

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

Larder, B. A., G. Darby, and D. D. Richman. 1989. HIV with reduced sensitivity to zidovudine (AZT) isolated during prolonged therapy. Science **243**:1731–1734.

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.

Markham, R. B., W.-C. Wang, A. E. Weisstein et al. (11 coauthors) 1998. Patterns of HIV-1 evolution in individuals with differing rates of CD4 T cell decline. Proc. Natl. Acad. Sci. USA **95**:12568–12573.

Mayer-Hamblett, N, and S. Self. 2001. A regression modelling approach for describing patterns of HIV genetic variation. Biometrics **57**:449–460.

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

Moriyama, E. N., Y. Ina, K. Ikeo, N. Shimizu, and T. Gojobori. 1991. Mutation pattern of human immunodeficiency virus gene. J. Mol. Evol. **32**:360–363.

Neuhauser, C., and S. M. Krone. 1997. The genealogy of samples in models with selection. Genetics **145**:519–534.

Nielsen, R. 1999. Changes in d_s/d_n in the HIV-1 env gene. Mol. Biol. Evol. **16**:711–714.

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. Anderson, A. R. McLean, T. F. Wolfs, J. Goudsmit, and R. M. May. 1991. Antigenic diversity thresholds and the development of AIDS. Science **254**:963–969.

Richman, D. D., T. Wrin, S. J. Little, and C. J. Petropoulos. 2003. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc. Natl. Acad. Sci. USA **100**:4144–4149.

Rodrigo, A. G., and J. Felsenstein. 1999. Coalescent approaches to HIV population genetics. Pp. 233–272 *in* K. A. Crandall, ed. The evolution of HIV. John Hopkins University Press, Baltimore.

Rodrigo, A. G., E. G. Shaper, E. L. Delwart, A. K. Iversen, M. V. Gallo, J. Brojatsch, M. S. Hirsch, B. D. Walker, and J. I. Mullins. 1999. Coalescent estimates of HIV-1 generation time in vivo. Proc. Natl. Acad. Sci. USA **96**:2187–2191.

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

Ross, H. A., and A. G. Rodrigo. 2002. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. J. Virol. **76**: 11715–11720.

Schierup, M. H., and J. Hein. 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics **156**:879–891.

Schuitemaker, H., M. Koot, N. A. Kootstra, M. W. Dercksen, R. E. de Goede, R. P. van Steenwijk, J. M. Lange, J. K. Schattenkerk, F. Miedema, and M. Tersmette. 1992. Biological phenotype of human immunodeficiency virus type 1 clones at different stages of infection: progression of disease is associated with a shift from monocytotropic to T-cell-tropic virus population. J. Virol. **66**:1354–1360.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. Variance components. John Wiley and Sons, New York.

Seo, T.-K, J. L. Thorne, M. Hasegawa, and H. Kishino. 2002. Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. Genetics **160**: 1283–1293.

Shankarappa, R. 1999. Evolution of HIV-1 resistance to antiviral agents. Pp. 469–490 *in* K. A. Crandall, ed. The evolution of HIV. John Hopkins University Press, Baltimore.

Shankarappa, R., J. B. Margolick, S. J. Gange et al. (12 coauthors). 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. **73**:10489–10502.

Shriner, D., D. C. Nickle, M. A. Jensen, and J. I. Mullins. 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet. Res. **81**:115–121.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry, 3rd ed. W. H. Freeman and Company, New York.

Tersmette, M., R. A. Gruters, F. de Wolf, R. E. de Goede, J. M. Lange, P. T. Schellekens, J. Goudsmit, H. G. Huisman, and F. Miedema. 1989. Evidence for a role of virulent human immunodeficiency virus (HIV) variants in the pathogenesis of acquired immunodeficiency syndrome: studies on sequential HIV isolates. J. Virol. **63**:2118–2125.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **24**:4876–4882.

Viscidi, R. P. 1999. HIV evolution and disease progression via longitudinal studies. Pp. 346–389 *in* K. A. Crandall, ed. The evolution of HIV. John Hopkins University Press, Baltimore.

Wain-Hobson, S., C. Renoux-Elbé, J.-P. Vartanian, and A. Meyerhans. 2003. Network analysis of human and simian immunodeficiency virus sequence sets reveals massive recombination resulting in shorter pathways. J. Gen. Virol. **84**: 885–895.

Williamson, S. 2003. Adaptation in the *env* gene of HIV-1 and evolutionary theories of disease progression. Mol. Biol. Evol. **20**:1318–1325.

Wodarz, D., P. Klenerman, and M. A. Nowak. 1998. Dynamics of cytotoxic T-lymphocyte exhaustion. Proc. R. Soc. Lond. B Biol. Sci. **265**:191–203.

Wolfs, T. F., J. J. de Jong, H. Van den Berg, J. M. Tijnagel, W. J. Krone, and J. Goudsmit. 1990. Evolution of sequences encoding the principal neutralization epitope of human immunodeficiency virus 1 is host-dependent, rapid, and continuous. Proc. Natl. Acad. Sci. USA **87**:9938–9942.

Wolinsky, S. M., B. T. M. Korber, A. U. Neumann et al. (11 coauthors). 1996. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. Science **272**:537–542.

Wolinsky, S. M., and G. H. Learn. 1999. Levels of diversity within and among host individuals. Pp. 275–314 *in* K. A. Crandall, ed. The evolution of HIV. John Hopkins University Press, Baltimore.

Wong, J. K., M. Hezareh, H. F. Günthard, D. V. Havlir, C. C. Ignacio, C. A. Spina, and D. D. Richman. 1997. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. Science **278**:1391–1300.

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

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

Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Brown, and P. Simmonds. 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J. Virol. **67**:3345–3356.

Edward Holmes, Associate Editor

Accepted October 20, 2004