Letter to the Editor

Changes in $d_s/d_n$ in the HIV-1 env Gene

Rasmus Nielsen

Department of Integrative Biology, University of California at Berkeley

Positive selection in the human immunodeficiency virus type 1 (HIV-1) env gene has been demonstrated by comparisons of the abundance of nonsynonymous ($d_s$) and synonymous ($d_n$) pairwise variability in the DNA sequence (Bonhoeffer, Holmes, and Nowak 1995; Seibert et al. 1995; Yamaguchi and Gojobori 1997). An increase in $d_n$ above the neutral expectation of $d_n = d_s$ provides an unambiguous indication of the action of positive selection (Kimura 1983). The observed increase in $d_n$ in the V3 loop of the env gene is hypothesized to be caused by immune mediated selection for sequence diversity (Brown and Monaghan 1988; Simmonds et al. 1990; Holmes et al. 1992). According to this hypothesis, cytotoxic T lymphocytes (CTLs) are important in limiting the viral population. Viral variants carrying new CTL epitopes may avoid immune recognition and thereby spread in the viral population. The consequence at the molecular level is a decrease in $d_s/d_n$ to below one (Bonhoeffer, Holmes, and Nowak 1995; Seibert et al. 1995).

The most compelling evidence for positive selection is provided by Bonhoeffer, Holmes, and Nowak (1995), who observed an increase in $d_n$ over $d_s$ in the V3 loop of the env gene by a factor of 10 for sequences obtained from an infected individual after 3 years of infection. Bonhoeffer, Holmes, and Nowak (1995) also observed an increase in $d_s/d_n$ among samples taken from the infected individual at different stages of the infection. This result was interpreted as evidence for a decrease in the selection pressure during the course of infection. It was hypothesized that the change in selection pressure was caused by a decline in the CD4+ CTL population. Although there has been some discussion regarding the method used to estimate $d_s$ and $d_n$ (Rodrigo and Mullins 1996; Nielsen 1997), it remains the case that there is a significant increase in the $d_s/d_n$ ratio between the third year of infection and the subsequent years in the data analyzed by Bonhoeffer, Holmes, and Nowak (1995) (Rodrigo and Mullins 1996; Nielsen 1997; Nielsen and Yang 1998). Previous analyses have assumed that a change in $d_s/d_n$ is evidence for a change in the selection pressure and is caused by changes in the dynamics of the immune system. However, this article will show that even if $d_n$ and $d_s$ can be estimated with 100% accuracy, an increase in $d_s/d_n$ similar to the increase observed by Bonhoeffer, Holmes, and Nowak (1995) will arise under a constant selection pressure due to the population genetic dynamics of natural selection. The ratio $d_s/d_n$ is a function both of the time since infection and of the viral population size ($N$), and therefore changes in the $d_s/d_n$ ratio may not reflect changes in the strength of selection. These new effects have been ignored in previous studies. Problems regarding the estimation of $d_n$ and $d_s$ have been dealt with elsewhere (Rodrigo and Mullins 1996; Nielsen 1997; Nielsen and Yang 1998) and will not be discussed in this article.

To investigate the properties of the statistic $d_s/d_n$, it is necessary to develop an appropriate population genetic model for the effect of selection on the HIV env gene. Such a model can be established by assuming that a haplotype with a new epitope may almost entirely avoid immune inactivation until the immune system develops responses to the new haplotype. For definiteness, an invading epitope is assumed to be new. Haplotypes coding for the new epitope will have selective advantage ($s$) until an appropriate immune response has been developed. To parameterize such a model, it is necessary to make some assumptions regarding the factors determining the time it takes to develop an immune response. Intuitively, both the frequency of the epitope in the population and the time it has been in existence should be of importance. One way to model this is to assume that the time until an immune response occurs depends deterministically on how much the immune system has been exposed to the epitope. The fitness of a haplotype coding for epitope $i$ in generation $j$ since the epitope first occurred by mutation is then given by

$$w_{ij} = \begin{cases} 1 + s \cdot \frac{N f_{ik}}{c} & \text{if } \sum_{k=0}^{k=j} N f_{ik} < c, \\ 1 & \text{else} \end{cases} \quad (1)$$

where $f_{ik}$ is the total frequency of the $i$th epitope in the $k$th generation, $s$ is the selection coefficient, and $c$ is a constant that determines how much the immune system must be exposed to a new epitope before appropriate immune responses are developed. Obviously, this model does not capture all of the dynamics of selection in HIV. However, since little is known with certainty about the population genetical dynamics of selection in HIV, this model provides an appropriate trade-off between simplicity and realism and is sufficiently complex to demonstrate all of the selective dynamics of interest in this article. The conclusions of this article are not dependent on the exact details of the assumed fitness function.

In the following, the model (eq. 1) will be applied to examine whether a change in $d_s/d_n$ such as the change observed by Bonhoeffer, Holmes, and Nowak (1995) will occur under natural circumstances when the selection coefficient remains constant, i.e., when no extrinsic...
factors act to change the strength of selection. However, the time-dependent expectation of \( d/d_e \) cannot easily be evaluated analytically. It is therefore necessary to perform computer simulations to investigate this question.

Computer simulations of the evolution of viral sequences under the selection model (eq. 1) can be performed using standard population genetical techniques. To describe the viral sequences, we will assume that they mutate according to an infinite-sites model (Watson 1975). This implies that only one substitution may happen in each nucleotide site. The effect of genetic drift and selection will be modeled by a Wright-Fisher model of evolution (Fisher 1958). The evolution of the population is simulated by keeping track of the identity of the viral sequences in a population of size \( N \). The frequency of each haplotype and the mutational makeup of each haplotype is stored in the computer memory. For example, if a haplotype is scored as \( \{12, 18, 23\} \), this would indicate that the haplotype is carrying mutations 12, 18 and 23. When a mutation arises, it is decided whether the mutation is nonsynonymous or synonymous, and this information is stored in the computer memory. Only nonsynonymous mutations create haplotypes with a new CTL epitope. For the sake of computational convenience, it is assumed that the population evolves in discrete generations. In each generation, the population experiences mutation, genetic drift, and selection. It is assumed that a new mutation occurs in an individual viral sequence in each generation with probability \( \mu \); i.e., a binomially distributed number of mutations occur in the population in each generation. After a round of mutations have occurred, selection and drift are allowed to operate. The probability that a viral sequence in generation \( j + 1 \) is of type \( i \) is

\[
    h_{ij} w_{ij} / \sum_k h_{kj} w_{kj},
\]

where \( h_{ij} \) is the frequency of haplotype \( i \) in generation \( j \) and \( w_{ij} \) is the relative fitness of haplotype \( i \) in generation \( j \). The population in generation \( j + 1 \) consists of \( N \) viral sequences sampled according to these probabilities. Selection, mutation, and genetic drift have thereby been allowed to operate.

At any given point in time, a sample can be taken from the population, and the average number of nonsynonymous pairwise differences and the average number of synonymous pairwise differences in the sample can be calculated. This procedure mimics the common method for estimating \( d/d_e \). However, the expectation of \( d_e \) (\( E(d_e) \)) and the expectation of \( d_i \) (\( E(d_i) \)) for a sample of sequences equals \( E(d_e) \) and \( E(d_i) \) for the entire population. Since \( d_d \) and \( d_i \) for the entire population have smaller variances than the equivalent measures for a sample, only the results for the entire population are reported here. By repeating the simulations, \( E(d_i/d_e) \) and \( E(d_i)E(d_e) \) can be estimated. Because of the possibility of obtaining samples in which \( d_i = 0 \), only results for \( E(d_i)E(d_e) \) will be reported.

In the simulations, we will assume that an individual is infected by a single viral type that may subsequently undergo mutation, selection, and genetic drift. The ratio of synonymous to nonsynonymous mutations before the action of positive selection is \( \omega \). This is roughly equivalent to assuming that \( 1 - 1/(4\omega) \) of all nonsynonymous mutations are immediately lethal to the virus. It will also be assumed that \( \mu = 0.01 \) and \( N = 1,000 \). This population size is large enough to capture all of the relevant population genetical dynamics. One hundred replicate simulations are performed for each data series, and \( E(d_n) \) and \( E(d_i) \) at a given time point are estimated as the averages of the values observed in the 100 time series. It is assumed that an individual is only infected by one viral variant. In the case of the data presented by Bonhoeffer, Holmes, and Nowak (1995), no variation in the viral sequence was observed in the samples obtained immediately after infection.

Notice (fig. 1) that the qualitative pattern is identical in all simulations. Initially, the model is effectively neutral, because the immune system has not evolved responses to any of the epitopes. At this phase, \( E(d_i)/E(d_e) = \omega \). When an immune response to the original epitope has developed (after approximately \( c \) generations), selection starts to operate, \( d_i \) increases fast, and \( E(d_i)/E(d_e) \) drops to a very low value. Shortly after the drop, \( E(d_i)/E(d_e) \) starts increasing again until an equilibrium is established. There is a simple intuitive explanation for this pattern. Synonymous variation, caused solely by mutation and genetic drift (including hitchhiking effects), accumulates at a slower rate than nonsynonymous variation. After the initial decrease in \( d/d_e \), there will therefore be a long period in which \( d/d_e \) increases before reaching an equilibrium. In other words, \( d/d_e \) will change through time, because \( d_i \) and \( d_n \) approach their equilibrium values at different rates. An increase in \( d_i/d_e \) is therefore expected even in models of constant selection pressure. Observed changes in \( d_i/d_e \) (e.g., Bon-
The described effect is not a particularity of the chosen model, but may be typical of a larger class of selective models. To illustrate this, evolution was simulated under a classical model of frequency-dependent selection. In this model, it is assumed that a haplotype coding for epitope $i$ has fitness

$$w_i = 1 + s(1 - f_i),$$

where $f_i$ is the total frequency of haplotypes coding for epitope $i$. Simulations were performed under an infinite-sites model as described above and assuming $N = 1,000$. Again, 100 replicate simulations were performed for each data series. Notice (fig. 2) that a similar increase in $d/s$ is observed in this model. However, there is no latency period, since selection immediately operates on all new epitopes.

The effect of selection is not only a product of the selection coefficient ($s$), but is also a product of the population size. This problem has been explored in many contexts within population genetics (Kimura 1983). In general, selection acts more effectively in large populations. Population genetics theory predicts that the effect of $N$ is largest when $Ns$ is on the order of one (Kimura 1983). Since little is known about $Ns$ in real HIV viral populations, it cannot be excluded that the $d/s$ ratio will change when the population size changes. An increased immune response is expected to lower the viral population size. Curiously, an increased immune response may, under positive selection, therefore result in an increase in $d/s$. Consequently, an observed increase in $d/s$ may in fact be caused by a stronger immune response. This prediction is quite opposite to the traditional interpretation (Bonhoeffer, Holmes, and Nowak 1995), in which $d/s$ is negatively correlated with the strength of the immune response. For example, in the model described by equation (1) when $s = 0.02$, $c = 10$, $\mu = 0.01$, $\omega = 2.0$, and $N = 100$, the equilibrium value of $d/s$ is approximately 1.29, but when $N = 1,000$, the equilibrium value of $d/s$ is approximately 0.96. Nonetheless, if a collapse in the immune system has a relatively stronger effect on $s$ than on the viral population size, one would still expect to observe an increase in $d/s$ when the immune system collapses.

While the ratio of nonsynonymous to synonymous variation can be applied effectively to demonstrate the existence of positive selection, differences in $d/s$ between samples from different individuals or from one infected individual obtained at different times should not alone be interpreted as evidence for changes in the selection intensity or disease progression. Observed changes in the $d/s$ ratio are expected even under a model assuming a constant selection coefficient. Likewise, changes in the strength of the immune response may not result in predictable changes in the $d/s$ ratio if the selection coefficient is on the same order of magnitude as the effective population size. Observations of temporal changes in the $d/s$ ratio alone may therefore provide only a little information about the status of the immune system.

### Acknowledgments

This work was greatly improved by the comments of reviewing editor S. A. Sawyer, who corrected several errors in the original manuscript. This work was supported in part by National Institute of Health grant GM40282 to Montgomery Slatkin and by a fellowship to R.N. from the Danish Research Council.

### Literature Cited


STANLEY A. SAWYER, reviewing editor

Accepted January 13, 1999