*Genet. Res., Camb.* (1994), **64**, *pp.* 1–9 *With 2 text-figures*  

# An application of population genetic theory to synonymous gene sequence evolution in the human immunodeficiency virus (HIV)

JOHN K. KELLY
*Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637*

(*Received 25 January 1994 and in revised form 15 April 1994*)

**Summary**

A population genetic model is developed and then applied to the synonymous gene sequence variation observed in samples of the Human Immunodeficiency Virus Type 1 (HIV-1). The samples, which were taken from several previous studies, contain sequences of the envelope glycoprotein gene (gp 120) of HIV-1. This analysis suggests that the viral population within an infected patient at any specific time is likely to be composed of close relatives. The viruses in a sample are likely to share a recent common ancestor probably due to consistent positive selection for non-synonymous mutations coupled with low recombination in this region of the genome. There is no substantial difference in synonymous evolutionary rate between samples of sequences obtained from Peripheral Blood Mononucleate Cells (PBMCs) and samples taken from blood plasma. This is likely to be due to the high rate of migration between these 2 HIV sub-populations. The mutation rate for the genetic region examined is estimated at $9{\cdot}20 \times 10^{-4}$ per site per month. Under the assumptions of the estimation procedure, this estimate can be bounded between 8·50 and $9{\cdot}91 \times 10^{-4}$ with 95% confidence. When coupled with direct estimates of mutation rate, the rate of synonymous evolution suggests that the mean number of generations per month for HIV-1 *in vivo* is between 1 and 4.

## 1. Introduction

Populations of HIV-1 within single patients harbour remarkable levels of genetic variation (Hahn *et al.* 1986; Fisher *et al.* 1988; Balfe *et al.* 1990; Simmonds *et al.* 1991; Wolfs *et al.* 1991). This variation is likely to have important consequences both for HIV transmission and pathogenesis. Strains of HIV vary in their ability to infect different cell types *in vitro* and also in their cytopathic tendencies (Chesebro *et al.* 1991; Fouchier *et al.* 1992). In addition, genetic changes in the envelope glycoprotein gene (gp 120) have been shown to alter the susceptibility of HIV isolates to immune recognition (Looney *et al.* 1988; MacKeating *et al.* 1989; Tersmette *et al.* 1988).

Longitudinal studies of HIV populations, in which multiple samples are taken from the same individual over the course of an infection (e.g. Balfe *et al.* 1990; Simmonds *et al.* 1991; Wolfs *et al.* 1991; Holmes *et al.* 1992), have established that much of this variation develops *de novo* over the course of a single infection. Little or no variation is found in the envelope gene when samples of viruses are obtained from individuals that have become infected with HIV very recently (Zhang *et al.* 1993). However, when samples are taken from the same individuals several years later, extensive gene sequence variation is observed. This variation results from the evolutionary divergence of the HIV population from the original sequence (Wolfs *et al.* 1991; Holmes *et al.* 1992).

Here, I develop a model that predicts the expected rate and pattern of synonymous gene sequence evolution in an HIV population under the assumption that synonymous changes are selectively neutral. The model predicts the value of several summary statistics of synonymous evolution. Longitudinal studies of gene sequence evolution conducted by Balfe *et al.* (1992), Wolfs *et al.* (1991), and Holmes *et al.* (1992) provide sufficient data to establish the values of these summary statistics from a series of patients. I apply the model to these data by inverting the theoretical predictions and using the observed values of the summary statistics to estimate, or at least make inferences about, the parameters of the model. I argue that the values of these parameters provide important information about the interaction between HIV and the immune system of an infected patient.

This paper uses a population genetic model as an inferential structure. Two aspects of the longitudinal studies of HIV sequence evolution greatly facilitate

this exercise. First, it is possible to compare sequences directly to their ancestors instead of just comparing them to the other sequences in the same sample. The expected difference between a sequence and its ancestor can be established without making many of the assumptions necessary to establish the expected difference between sequences in the same sample. Secondly, longitudinal data from multiple patients is available. In effect, this provides multiple realizations of the same evolutionary process. Comparing the pattern of evolutionary change in different HIV populations allows us to estimate the amount of stochasticity in the evolutionary processes of HIV and provides a means to statistically justify parameter estimates.

This study focuses on the demographic structure of HIV populations. Demographic factors, such as viral generation time, have important effects on the dynamics of amino acid evolution. I test whether there are significant differences in the mean number of generations per unit time between different sub-populations of HIV within an infected individual; whether there is significant variation in viral generation time within a single HIV population; and whether there is significant common ancestry in the geneological structure of intra-patient HIV populations. I use the observed rate of synonymous gene sequence evolution to produce a statistically bounded estimate for the mutation rate per site per month and to suggest upper and lower limits for the mean number of HIV generations per month *in vivo*. These estimates may be useful for setting limits on the rate of adaptive HIV evolution within patients and for determining whether different clinical practices may affect this rate.

## 2. Theory

The complex retroviral replication cycle is initiated when a virion infects a host cell (Fig. 1). A viral enzyme, reverse transcriptase, generates a DNA copy of the viral RNA template. The resulting minus strand of DNA is converted into a double stranded molecule and then integrated into the host cell genome. The viral sequence, in DNA form within the host genome, is referred to as provirus. RNA progeny are then generated by transcription via RNA polymerase II and packed into virions. The virions are released to infect other cells. Point mutations may occur at several stages of the replication cycle but are especially likely during reverse transcription.

The duration of each phase in the replication cycle may be variable, especially the period of integration. This has the important consequence that a random samples of viruses may contain individuals that have different numbers of ancestors between themselves and the viruses that founded the population. These viruses will have experienced different lengths of 'evolutionary time'. A discrete generation population genetic model, in which all individuals have the same number of ancestors, is therefore probably inappropriate for HIV evolution. A second important consideration in the development of a synonymous evolution model is that non-synonymous mutations are often under strong selection in HIV. These selected mutations may have an important effect of the dynamics of neutral variation at linked sites.

I assume that the HIV population that eventually becomes established within an infected patient evolves from a single progenitor sequence. This does not require that a single virus founds the entire population, but merely that all viruses that contribute descendants to the future samples from the population share a single sequence (within the region of the genome with which we are concerned). This single sequence is referred to as the progenitor sequence. The actual infecting particle that an individual virus is descended from is referred to as its progenitor virus.

I assume that the introduction of synonymous point mutations over the course of a replication cycle is a Poisson process. This assumption is justified when the per site mutation rate is small and the number of sites is sufficiently large (Kimura 1969). I denote $\mu_s$ as the total synonymous mutation rate per replication cycle (summed across all sites in the sequence and all stages of the replication cycle). I assume that all sites within the focal sequence are completely linked and that all new mutations occur at sites that had not previously undergone mutation in the ancestry of this particular virus. Non-synonymous mutations may also occur within the sequence, but I assume that they occur independently of synonymous mutations. I assume that synonymous mutations are selectively neutral. The selection regime on non-synonymous mutations is arbitrary.

I denote $S(T)$ as the number of synonymous differences between a randomly sampled virus and the progenitor sequence at time $T$ (number of time units after infection) and $k(T)$ as the number of generations between that virus and its progenitor virus. Because synonymous mutations are introduced independently each generation, $S(T)$ is the sum of $k(T)$ independent Poisson random variables where $k(T)$ is itself a random variable. The probability distribution of $S(T)$ is:

$$\text{Prob}[S(T) = i] = \sum_n \text{Prob}[k(T) = n]\,\text{Prob}[S(T) = i \mid k(T) = n], \quad (1)$$

where

$$\text{Prob}[S(T) = i \mid k(T)] = \frac{(k(T)\mu_s)^i\, e^{-k(T)\mu_s}}{i!}. \quad (2)$$

The mean and variance of this distribution are:

$$E[S(T)] = \mu_s E[k(T)], \quad (3)$$

$$\text{Var}[S(T)] = \mu_s E[k(T)] + \mu_s^2 \text{Var}[k(T)]. \quad (4)$$

These relations are useful because they can be used to derive the expected values for several sample

https://doi.org/10.1017/S0016672300032481 Published online by Cambridge University Press

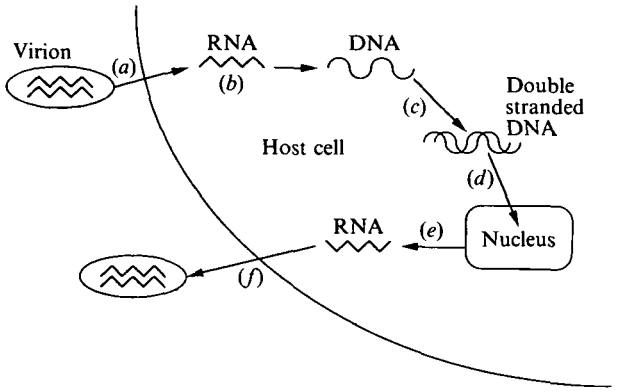


Fig. 1. The replication cycle of HIV proceeds through a series of stages that are represented by letters in the figure. The cycle begins with (*a*) entry of viral RNA into a host cell, followed by (*b*) reverse transcription, (*c*) replication of the resulting DNA into a double stranded molecule, (*d*) incorporation of the viral DNA into the host genome, (*e*) transcription of the proviral DNA, and finally (*f*) the production and release of new virions.

statistics that can be calculated directly from gene sequence data. The mean number of synonymous substitutions per virus, $\alpha$, and the variance in the number of substitutions per virus within an HIV population, $\gamma$, can be estimated when the progenitor sequence is known by using the following formulae:

$$\alpha = \frac{1}{n}\sum_{j=1}^{n} s_j \tag{5}$$

and

$$\gamma = \frac{1}{n-1}\sum_{j=1}^{n} s_j^2 - \alpha^2, \tag{6}$$

where $s_j$ is the number of synonymous differences between the $j$th sequence and the progenitor in a sample of size $n$. The number of generations in the ancestry of a virus, $k(T)$, can be expressed as the product of $g$, the number of generations per unit time in the ancestry of that virus, and $T$, the amount of time since infection. Using eqns 3 and 4, the expected values of $\alpha$ and $\gamma$ can be shown to equal:

$$E[\alpha] = \mu_s \bar{g} T, \tag{7}$$

$$E[\gamma] = (\mu_s \bar{g} T + \mu_s^2 \sigma_g^2 T^2)(1-\rho), \tag{8}$$

where $\bar{g}$ and $\sigma_g^2$ are the mean and variance of the distribution of $g$ in the viral population and $\rho$ is the correlation of the number of synonymous differences from the progenitor among viruses within the population.

In principle, each of the four parameters in eqns 7 and 8 may have 'infection specific' values. The synonymous mutation rate, $\mu_s$, is determined by the total mutation rate in a sequence and the fraction of all changes that are synonymous. Since both may depend on the initial sequence, $\mu_s$ will be specific to the progenitor sequence. If progenitor sequences differ substantially among different HIV infections, then some portion of the interpatient variation in $\alpha$ and $\gamma$ will be attributable to variation in $\mu_s$.

The distribution of $g$ in an HIV population, from which $\bar{g}$ and $\sigma_g^2$ are obtained, is determined by the infection rate of free virions into new cells and especially the duration of reproductive activity by proviral DNA. The number of ancestors between a given virus and its progenitor virus should be inversely related to the lengths of reproductive lifespan of its ancestors. The length of reproductive lifespan depends on the pathogenicity of the virus, i.e. how quickly the production of new viruses from proviral DNA kills the host cell. The value of $\bar{g}$ should thus be positively related to viral pathogenicity. If pathogenicity is related to rate of replication, the distribution of $g$ could also be expressed as a function of the distribution of proviral replication rates.

The variance in $g$, $\sigma_g^2$, should parallel the variance in pathogenicity and replication rate. There are likely to be both genetic and environmental sources of variation in these quantities. An environmental component of variation is the range of cellular environments experienced by HIV. Incorporation of proviral DNA into different parts of the host cell genome may cause variation in replication rate due to the variable activity of surrounding genes within that cellular type (Zhang *et al.* 1993). In addition, several studies have identified genetic strains of HIV with variable pathogenicity (Fisher *et al.* 1988; Hwang *et al.* 1991; Westervelt *et al.* 1992; and references therein). When there is genetic variation for pathogenicity, the generation times of viruses within lineages will be correlated. This will increase $\sigma_g^2$. At present, it is unknown whether $\bar{g}$ and $\sigma_g^2$ have constant values across different infections. They may vary among patients or over time within a single patient or even among sub-populations of HIV at the same time within a single patient.

The correlation parameter, $\rho$, depends on the genealogical structure of the HIV population within a patient and specifically on the extent of shared ancestry by viruses. The correlation of $s_j$ is shown, in Appendix 1, to equal:

$$\rho = \frac{\mu_s E[\hat{k}] + \mu_s^2 \mathrm{Var}[\hat{k}]}{\mu_s E[k] + \mu_s^2 \mathrm{Var}[k]}, \tag{9}$$

where $\hat{k}$ is the number of generations of shared ancestry for two randomly sampled viruses. When the mutation rate is small, $\rho$ reduces approximately to the ratio of average number of generations that a virus shares with another randomly sampled virus to the average number of total generations:

$$\rho \approx \frac{E[\hat{k}]}{E[k]}. \tag{10}$$

The amount of shared ancestry is a function of the genealogical structure of the population. Because the genealogy of an HIV population probably changes stochastically over the course of an infection, a consistent value for $\rho$ across infections is unlikely.



https://doi.org/10.1017/S0016672300032481 Published online by Cambridge University Press

Table 1. *Summary of longitudinal sequence studies*

| Patient | Source | Sample type | $M$ | $T$ | $\theta$ | $\alpha$ | $\gamma$ | $\alpha/T$ | $\alpha/(TK\theta)$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 p74 | Zhang *et al.* (1993) Balfe *et al.* (1990) | DNA | 35 | 62 | 0·59 | 1·17 | 0·157 | 0·0189 | $9{\cdot}14\times10^{-4}$ |
| 2 p77 | Zhang *et al.* (1993) Balfe *et al.* (1990) | DNA | 35 | 62 | 0·63 | 1·17 | 0·157 | 0·0189 | 8·56 |
| 3 p82 | Zhang *et al.* (1993) Balfe *et al.* (1990) | DNA | 33 | 62 | 0·61 | 1·43 | 0·527 | 0·0231 | 10·80 |
| 4 p82 | Holmes (1992) | RNA | 35 | 86 | 0·61 | 1·54 | 0·764 | 0·0179 | 8·39 |
| 5 W1 | Wolfs *et al.* (1991) | RNA | 35 | 60 | 0·61 | 1·13 | 0·696 | 0·0188 | 8·78 |
| 6 W495 | Wolfs *et al.* (1991) | RNA | 35 | 57 | 0·63 | 1·25 | 0·500 | 0·0219 | 9·95 |

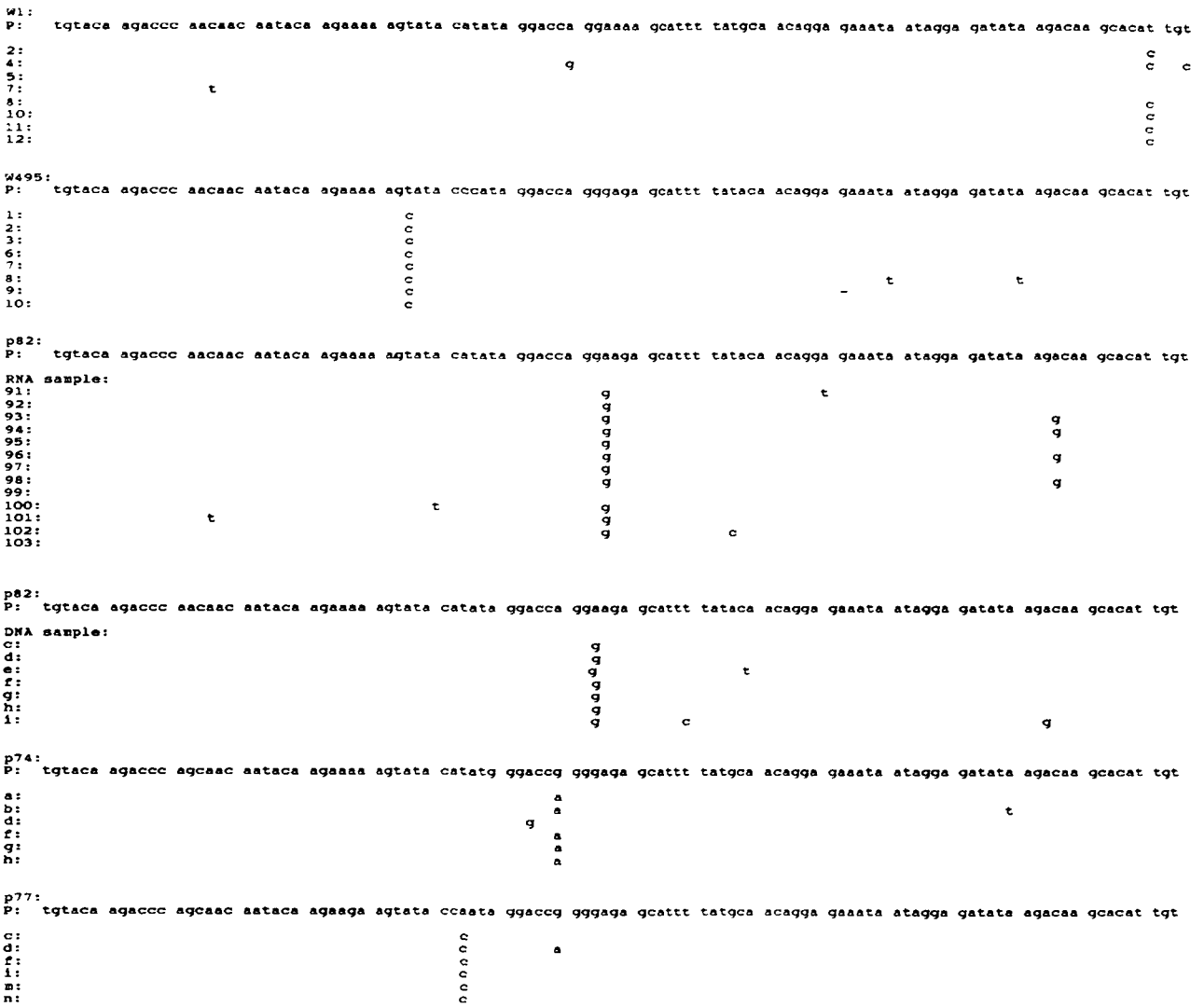


Fig. 2. The set of nucleotide sequences from each patient. The first sequence in each set (preceded by the *P*) is the progenitor. Under each progenitor sequence is the set of sequences taken later. All positions at which synonymous replacements occurred are denoted. Nonsynonymous substitutions are not shown.

## 3. The studies

Sequence data was extracted from 4 published papers: Balfe *et al.* (1990), Zhang *et al.* (1993), Holmes *et al.* (1992), and Wolfs *et al.* (1991). Each of these studies presents sequence data from the V3 hypervariable region of the envelope glycoprotein gene. I focus on a 35 amino acid section of the V3 area called the 'loop region'. This sequence was chosen for practical reasons. The sequence is shared among these studies,

https://doi.org/10.1017/S0016672300032481 Published online by Cambridge University Press

its length is conserved across and within patients, and it has a highly conserved progenitor sequence.

Balfe *et al.* (1990) present a set of sequence samples obtained in 1989 from several haemophiliacs infected via blood transfusion in 1984. The sequences were extracted from proviral DNA incorporated in Peripheral Blood Mononucleate Cells (PBMCs). The loop region progenitor sequences for the patients p74, p77, and p82 in Balfe *et al.* (1990) are presented in Zhang *et al.* (1993). These three samples are numbers 1, 2 and 3 in Table 1. Holmes *et al.* (1992) present a temporal series of HIV sequences extracted from virions (RNA) in the blood plasma of one of the same patients (p82). I contrast the final sample from this study taken in 1991 to the p82 progenitor sequence to obtain values for $\alpha$ and $\gamma$ (number 4 in Table 1). Wolfs *et al.* (1991) present temporal series of HIV sequences (RNA) obtained from blood plasma samples from 2 different patients (W1 and W495). I use the last sample taken from each patient to obtain values of $\alpha$ and $\gamma$ (numbers 5 and 6 in Table 1).

The initial samples obtained from patients p74, p77, and p82 were genetically homogenous. The initial samples from patients W1 and W495 were not completely homogenous but were dominated by a single clone with a few other sequences that differed at one or two sites. The dominant sequence in each initial sample from W1 and W495 was chosen as the progenitor. Each sequence taken later from each patient was denoted as to the number of third codon position synonymous site differences from its progenitor sequence. Substitutions were classified as either synonymous or non-synonymous based on their effect on the amino acid sequence in the original codon of the progenitor sequence. Figure 2 gives the progenitor nucleotide sequence from each patient and the synonymous replacements observed in each sequence in the later sample. The number of months between infection and the final samples ($T$ in Table 1) was estimated from information provided in the original studies. Finally, I calculated $\theta$, the fraction of all third position changes that are synonymous for each progenitor sequence.

## 4. Analysis

A contrast of $\alpha$ and $\gamma$ values from the same HIV population allows a comparison of predictions based on the same parameters. If both $\sigma_g^2$ and $\rho$ are zero, then the expected value of $\alpha-\gamma$ is zero. The expected difference can be positive if, and only if, $\rho$ is positive. The expected value of $\alpha-\gamma$ can be negative if, and only if, $\sigma_g^2$ is positive. Thus, consistent differences between $\alpha$ and $\gamma$ across patients indicates that either $\rho$ and $\sigma_g^2$ is generally positive. This comparison does not require constancy of $\mu_s$ or $\bar{g}$ across patients or an accurate estimate of $T$ for each sample because paired values are used.

Values of $\alpha$ and $\gamma$ from each patient are presented in Table 1. In every case, $\alpha$ is greater than $\gamma$. In 5 of 6 cases, $\gamma$ is less than half $\alpha$. A paired sample $t$-test indicates that $\alpha-\gamma$ is significantly positive ($t = 9{\cdot}08$; D.F. $= 5$; $P < 0{\cdot}0003$). Two alternative non-parametric tests, the sign test ($Z = 2{\cdot}04$; $P = 0{\cdot}04$) and the Wilcoxon matched pair test ($Z = 2{\cdot}20$; $P = 0{\cdot}028$) also indicate a significant difference between $\alpha$ and $\gamma$. The positive values of $\alpha-\gamma$ suggests that $\rho$ is consistently non-zero, indicating substantial shared history in the genealogical structure of HIV populations.

The $t$-test of differences between $\alpha$ and $\gamma$ assumes that the distribution of $\alpha-\gamma$ under the null hypothesis is approximately normal. The exact distribution of this quantity, or of either $\alpha$ or $\gamma$ alone, is unknown. In fact, it cannot be determined without specifying the distribution of $k(T)$ in eqn 1. However, because the numerator of the $t$-statistic is calculated by averaging multiple independent random variables (the $\alpha-\gamma$ values from different patients), its distribution should converge on normality as a consequence of the central limit theorem. This convergence is also exploited below to determine confidence intervals for the mutation rate per site per month.

Each of the statistical tests applied to the $\alpha-\gamma$ values assumes independence of the paired samples. The two different pairs from p82 can be treated as independent only if $\rho = 0$. This assumption is part of the null hypothesis being tested here and it is therefore valid to treat the two pairs of values from p82 as independent for this analysis. However, as this analysis indicates that $\rho$ is generally positive, we cannot treat these samples as independent random variables in subsequent analyses.

The expectation of $\alpha/T$ is equal to the product of $\bar{g}$ and $\mu_s$ (eqn 3). I now assume that $\mu_s$ is approximately constant across the patients in Table 1. Constancy of $\mu_s$ implies that the only parametric source of variation in $\alpha/T$ is the differences in $\bar{g}$ among HIV populations. A significant difference in $\alpha/T$ values between viral RNA and proviral DNA samples would indicate that viruses in different HIV sub-populations within patients have different mean generation times. Two $t$-tests contrasting RNA and DNA samples (each excluding one of the $\alpha/T$ values from p82) found no evidence for differences between these two sample types.

If the relation between $\mu_s$ and the total mutation rate per replication cycle can be determined, the data in Table 1 can be used to obtain an unbiased and statistically bounded estimate of the mutation rate per site per month. I now employ a simple mutational model to obtain such an estimate. I assume that all 3rd positions sites in the sequence are equally mutable and that all base changes at these sites are equally likely. Under this condition,

$$\mu_s = M\theta u, \tag{11}$$

where $M$ is the number of sites in the sequence, $\theta$ is the

https://doi.org/10.1017/S0016672300032481 Published online by Cambridge University Press

fraction of all changes that are synonymous, and $u$ is the mutation rate per base pair per replication cycle.

For the data in Table 1, $M = 35$ and $\theta$ varies between 0·59 and 0·63. The estimated value of $u * \bar{g}$ from each patient is given in the last column of Table 1. The two estimates from p82 were averaged and then combined with the values from the other 4 patients to give an estimated mutation rate per site per month of $9{\cdot}20 \times 10^{-4}$. This estimate can be bounded between $8{\cdot}50 \times 10^{-4}$ and $9{\cdot}91 \times 10^{-4}$ with 95% confidence (assuming normality of the sample mean).

## 5. Discussion

In this study, the mean number of synonymous mutations per virus, $\alpha$, was consistently greater than the estimated intra-population variance in the number of synonymous mutations per virus, $\gamma$. In most cases $\gamma$ was less than half $\alpha$ indicating that most of the viruses isolated in each sample share a recent common ancestor with other viruses in the sample. This result is supported by the pattern of synonymous nucleotide replacements observed in Fig. 2. Sequences within the same sample share many of the same base substitutions (as opposed to simply the same number of site differences from their progenitor). This is unlikely to have occurred if the lineages of the viruses were distinct since initial infection.

This finding is important because high values of $\rho$ are a signature of natural selection. Natural selection will tend to reduce the amount of neutral variation in areas of low recombination (Hill & Robertson, 1966; Kreitman, 1991; Charlesworth *et al.* 1993). There is strong evidence of selection on non-synonymous mutations in the HIV-1 envelope gene. Simmonds *et al.* (1990) report that non-synonymous mutations are almost twice as likely to become fixed as synonymous mutations in the hyper-variable regions of this gene. It thus seems likely that 'selective sweeps' (*sensu* Kreitman 1991) are largely responsible for the close relatedness of HIV isolates from the same sample. It is interesting that a similar mechanism is probably responsible for the initial homogeneity of HIV populations in the envelope gene. Zhang *et al.* (1993) suggest that selection for specific sequences in the V3 region of the envelope gene results in an 'initial virus population [that is] essentially clonal ... in the selected region'. Sequence homogeneity early during infection is also observed in other regions of the envelope gene, presumably because of their linkage to the V3 area. The selective processes responsible for sample homogeneity (relatively speaking) at different stages of infection may be quite different however.

There was no significant difference in the synonymous substitution rates per month ($\alpha/T$) between plasma RNA and PBMC proviral DNA samples. At first, this result is somewhat surprising because PBMCs may harbour dormant proviral DNA for long periods of time (Holmes *et al.* 1992). This long dormancy suggests that isolates from PBMCs should have relatively low values for $\bar{g}$. Conversely, the virion population in the blood plasma may be dominated by the progeny of actively replicating viruses, presumably with shorter generation times (Holmes *et al.* 1992). However, it is important to realize that $\bar{g}$ is determined by the total number of generations in the ancestry of viruses and not just the generation time of their most recent ancestors. The plasma and PBMC populations are probably not distinct. Gene flow among HIV sub-populations will mix viruses with variable numbers of ancestors and reduce any variation among sub-populations. The existence of gene flow between plasma and PBMC sub-populations of HIV is supported by Simmonds *et al.* (1991) finding that specific haplotypes that are dominant in a plasma sample at one time often appear in later PBMC samples.

A final analysis yielded a bounded estimate for the *in vivo* mutation rate per site per month. This estimate is specific to the V3 loop region of the envelope gene. The confidence interval on this estimate is surprisingly narrow, placing $u * \bar{g}$ between 8·50 and $9{\cdot}91 \times 10^{-4}$ with 95% confidence. This high level of precision is due to the low level of variation in the rate of synonymous evolution among patients. An accurate estimate of the mutation rate per month is directly relevant to models HIV pathogenesis (e.g. Nowak *et al.* 1990; Simmonds *et al.* 1991). Nowak and his colleagues (Nowak *et al.* 1990; Nowak *et al.* 1991) have hypothesized that the immune system of an infected individual and the resident HIV population undergo a kind of 'evolutionary arms race'. These researchers have analysed models in which the HIV population is able to produce 'escape mutants', new variants that can evade the current immune surveillance of the patient. The escape mutant increases in abundance until eventually, the immune system responds and suppresses it. Their models exhibit a strange threshold behaviour however, such that once the number of antigenically distinct viruses reaches a certain number, the immune system collapses and the total density of HIV increases dramatically. Nowak *et al.* (1991) suggest that when an HIV population reaches this genetic diversity threshold, an HIV positive individual passes from the latent phase of the disease to the development of immunodeficiency. The present analysis is relevant to this model because the rate at which an HIV population produces escape mutations will be limited by the overall mutation rate per unit time, $u * \bar{g}$. The mutation rate per month may also influence other evolutionary processes that occur in HIV populations such as the rate of adaptation to different cell types.

The average number of generations per month for HIV *in vivo* could be estimated from the present data if the mutation rate per replication cycle within the loop region of V3 were known. A substantial number of studies have examined the fidelity of HIV-1 reverse

Table 2. *HIV-1 reverse transcriptase fidelity assays*

| Source | Template | Error rate* ($\times 10^{-4}$) |
|---|---|---|
| Preston *et al.* (1988) | DNA | 2·50 |
| Roberts *et al.* (1988) | DNA | 5·88 |
| Weber & Grosse (1989) | DNA | 1·35 |
| Riccheti & Buc (1990) | DNA | 1·90 |
| Bakhanashvili & Hizi (1992*a*) | DNA | 1·57 |
| Bakhanashvili & Hizi (1992*b*) | RNA | 3·82 |
| Ji & Loeb (1992) | DNA | 1·69 |
| Ji & Loeb (1992) | RNA | 1·45 |
| Yu & Goodman (1992) | DNA | 1·08 |
| Yu & Goodman (1992) | RNA | 3·53 |
| Bakhanashvili & Hizi (1993) | DNA | 4·56 |

Most of the error rate estimates are averages from multiple tests reported in these studies.

transcriptase (RT) *in vitro*. Table 2 summarizes some of these studies and contains data on error rates per base pair by RT during both reverse transcription (RNA to DNA) and DNA replication (DNA to DNA). HIV-1 RT seems to introduce errors in about equal frequency during both processes when the same sequence is used as a template (Bakhanashvili & Hizi, 1993). If RT is the primary source of point mutations and both stages of replication are equally mutable, then a plausible range for $\bar{g}$ can be established by dividing the estimated value of $u * \bar{g}$ obtained here by twice the minimum and maximum RT error rates in Table 2. This suggests that the mean number of generations per month for HIV-1 *in vivo* is between 0·78 and 4·26.

Several assumptions of the analysis merit further comment. The model is based on the assumption that the entire sample of sequences taken at time $T$ has evolved from a single progenitor sequence. This assumption is based on the empirical finding that HIV populations are genetically homogenous within the envelope gene at seroconversion (Zhang *et al.* 1993). It is supported by Wolfs *et al.* (1991) finding that the amount of divergence in samples taken several years after infection was intermediate between the amount of divergence in early and late samples. This suggests that the genetic variation in the final samples (which was the focus the analyses presented here) was the result of a progressive divergence of the HIV populations from initially homogenous founder populations and not the sudden appearance of genetic variants that existed at the time of sero-conversion but were hidden. In a similar kind of study, Holmes *et al.* (1992) show how all of the various sequences from p82 coalesce back to a single progenitor sequence. Finally, Wolinsky *et al.* (1992) obtained samples of HIV sequences from both the mother and infant of three cases of placental HIV transmission. By performing phylogenetic analyses on their data, these authors were able to conclude that the genotypic population within each infant was derived from a single form present in the mother.

Taken together, these results support a monoclonal origin for the envelope gene sequences observed later in infection. It is important to note, however, that all of this data is limited to the envelope gene. Genetic homogeneity at sero-conversion is not observed in the gag gene (Zhang *et al.* 1993). Thus, an alternative model is necessary to describe synonymous evolution in this region of the HIV genome.

The model presented here also assumes that all sites within the focal sequence are completely linked. Recombination frequently occurs in retroviruses and there is some evidence that it occurs among the different hypervariable regions within the envelope gene of HIV (Simmonds *et al.* 1991). This study focused on a very small sequence within one of the hypervariable regions and thus recombination probably had little effect on the results. However, it is an important consideration for the application of the model to larger sequences. I show in Appendix 2 that recombination does not change the expected number of synonymous mutations per virus and therefore does not change the expected value of $\alpha$. Recombination can reduce the variance in $S(T)$, however, and thus lower the expected value of $\gamma$.

The estimate of $u * \bar{g}$ obtained here should be treated with caution because neither of the assumptions used to derive eqn 11 are likely to be exactly correct. Roberts *et al.* (1988) found that sites within an HIV sequence may vary quite substantially in their mutability and it is commonly found that, at any specific site, some nucleotide substitutions are more likely than others. The validity of the suggested range for $\bar{g}$ is even more uncertain. Its accuracy depends on the sufficiency of *in vitro* studies of RT fidelity to measure the mutation rate per replication cycle. Riccheti & Buc (1990) and Bakhanashvili & Hizi (1993) provide strong evidence that the mutation rate is sequence (template) dependent. Since none of the RT fidelity studies in Table 2 use the loop region of V3 as a template, it is difficult to be certain that the average mutation rate per site within the loop region lies between the extreme values in Table 2. Some evidence suggests that mutability is higher in the hypervariable regions of the HIV genome (Doi, 1991). If the mutation rate per site in the loop region of V3 is greater than $11{\cdot}76 \times 10^{-4}$ (the upper limit suggested by Table 2), then the mean number of generations per monty by HIV-1 *in vivo* may be less than 0·78.

Despite the estimation difficulties and the other caveats regarding population origination and recombination, several aspects of the analysis are very promising. Most important is the low variance among patients in the rate of synonymous evolution. This indicates that hypothesis tests and estimation procedures based on synonymous variation should have substantial statistical power. For example, despite the limited number of patients, the data strongly suggest

that $\rho$ is significantly positive across infections. In light of these results, I suggest that population genetic analyses of synonymous variation may prove a useful tool for making inferences about important aspects of HIV infection.

This paper benefitted from reviews by Alan Molumby and Mike Wade and from discussions with B. Charlesworth, N. Johnson, S. Orzack, C. Grimsly, B. Obo and M. Amba. The Rileys generously provided computer facilities. The author is supported by a National Science Foundation predoctoral fellowship.

## References

Bakhanashvili, M. & Hizi, A. (1992). Fidelity of reverse transcriptase of human immunodeficiency virus type 2. *FEBS. Letters* **306**, 151–156.

Bakahanashvili, M. & Hizi, A. (1992). Fidelity of the RNA-dependent DNA synthesis exhibited by the reverse transcriptases of human immunodeficiency virus types 1 and 2 and of murine leukemia virus: mispair extension frequencies. *Biochemistry* **31**, 9393–9398.

Bakhanashvili, M. & Hizi, A. (1993). The fidelity of the reverse transcriptases of human immunodeficiency viruses and murine leukemia virus, exhibited by the mispair extension frequencies, is sequence dependent and enzyme related. *FEBS Letters* **319**, 201–205.

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

Charlesworth, B., Morgan, M. & Charlesworth, D. (1994). The effect of deleterious neutral mutations on neutral molecular evolution. *Genetics* (in the press).

Chesebro, B., Nishio, J., Perryman, S., Cann, A., O'Brien, W., Chen, I. S. & Wehrly, K. (1991). Identification of human immunodeficiency virus envelope gene sequences influencing viral entry in CD4 positive HeLa cells, *T*-leukemia cells, and macrophages. *Journal of Virology* **65**, 5782–5789.

Doi, H. (1991). Importance of purine and pyrimidine content of local nucleotide sequences (6 bases long) for evolution of the human immunodeficiency virus type 1. *Proceedings of the National Academy of Sciences U.S.A.* **88**, 9282–9286.

Fisher, A. G., Ensoli, B., Looney, D., Rose, A., Gallo, R. C., Saag, M. S., Shaw, G. M., Hahn, B. H. & Wong Staal, F. (1988). Biologically diverse molecular variants within a single HIV-1 isolate. *Nature* **334**, 444–447.

Fouchier, R. A. M., Groenink, M., Kootstra, N. A., Tersmette, M., Huisman, G. H., Miedema, F. & Schuitemaker, H. (1992). Phenotype associated sequence variation in the third variable domain of the human immunodeficiency type 1 gp120 molecule. *Journal of Virology* **66**, 3138.

Hahn, B. H., Shaw, G. M., Taylor, M. E., Redfield, R. R., Markham, P. D., Salahuddin, S. Z., Wong-Stall, F., Gallo, R. C., Parks, E. S. & Parks, W. (1986). Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. *Science* **232**, 1548–1553.

Hill, W. G. & Robertson, A. (1966). The effects of linkage on the limits to artificial selection. *Genetic Research* **8**, 269–294.

Holmes, E. C., Zhang, L. Q., Simmonds, P., Ludlam, C. A. & Brown, A. J. L. (1992). Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proceedings of the National Academy of Sciences U.S.A.* **89**, 4835–4839.

Hwang, S. S., Boyle, T. J., Lyerly, H. K. & Cullen, B. R. (1991). Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science* **253**, 71–74.

Ji, J. & Loeb, L. A. (1992). Fidelity of HIV-1 reverse transcriptase in copying RNA *in vitro*. *Biochemistry* **31**, 954–958.

Kimura, M. (1969). The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population due to Steady Flux of Mutations. *Genetics* **61**, 893–903.

Kreitman, M. (1991). Detecting selection at the level of DNA. In *Evolution at the Molecular Level* (ed. R. K. Selander, A. G. Clark and T. S. Whittam).

Looney, D. J., Fisher, A. G., Putney, S. D., Rusche, J. R., Redfield, R. R., Burke, S. D., Gallo, R. C. & Wong Staal, F. (1988). Type restricted neutralization of molecular clones of HIV. *Science* **241**, 357–359.

MacKeating, J. A., Gow, J., Goudsmit, J., Pearl, L. H., Mulder, C. & Weiss, R. (1989). Characterisation of HIV-1 neutralization escape mutants. *AIDS* **3**, 777–784.

Nowak, M. A., Anderson, R. M. & May, R. M. (1990). The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease. *AIDS* **4**, 1095–1103.

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

Preston, B. D., Poiesz, B. J. & Loeb, L. A. (1988). Fidelity of HIV reverse transcriptase. *Science* **242**, 1168–1171.

Ricchetti, M. & Buc, H. (1990). Reverse transcriptases and genomic variability: the accuracy of DNA replication is enzyme specific and sequence dependent. *EMBO Journal* **9**, 1583–1593.

Roberts, J. D., Bebenek, K. & Kunkel, T. A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science* **242**, 1171–1173.

Simmonds, P., Balfe, P., Ludlam, C. A., Bishop, J. O. & Brown, A. J. L. (1990). Analysis of sequence diversity in the hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. *Journal of Virology* **64**, 5840–5850.

Simmonds, P., Zhang, L. Q., McOmish, F., Balfe, P., Ludlam, C. A. & Brown, A. J. L. (1991). Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 *env* sequences in plasma viral and lymphocyte-associated proviral populations *in vivo*: implications for models of HIV pathogenesis. *Journal of Virology* **65**, 6266–6276.

Tersmette, M., Gruers, R. A., de Wolf, F., de Groede, R. E., Lange, J. M., Schellekens, P. T., Goudsmit, J., Huisman, H. G. & Miedema, F. (1988). Evidence for a role of virulent human immunodeficiency virus (HIV) variants in the pathogenesis of acquired immunodeficiency syndrome: studies on sequential HIV isolates. *Journal of Virology* **63**, 2118–2125.

Weber, H. & Grosse, F. (1989). Fidelity of human immunodeficiency virus type 1 reverse transcriptase in copying natural RNA. *Nucleic Acid Research* **17**, 1379–1393.

Westervelt, P., Trowbridge, D. B., Epstein, L. G., Blumberg, B. M., Li, Y., Hahn, B. H., Shaw, G. M., Price, R. W. & Ratner, L. (1992). Macrophage tropism determinants of human immunodeficiency virus type 1 *in vivo*. *Journal of Virology* **66**, 2577–2582.

Wolfs, T. F. W., Zwart, G., Valk, M., Kuiken, C. L. & Goudsmit, J. (1991). Naturally occurring mutations

within HIV-1 V3 genomic RNA lead to antigenic variation dependent on a single amino acid substitution. *Virology* **185**, 195–205.

Wolinsky, S. M., Wike, C. M., Korber, B. T. M., Hutto, C., Parks, W. P., Rosenblum, L. A., Kunstman, K. J., Furtado, R. & Munoz, J. L. (1992). Selective transmission of human immunodeficiency virus type 1 variants from mothers to infants. *Science* **255**, 1134–1137.

Yu, H. & Goodman, M. F. (1992). Comparison of HIV-1 and avian myeloblastosis virus reverse transcriptase fidelity on RNA and DNA templates. *Journal of Biological Chemistry* **267**, 10888–10896.

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

## Appendix 1

The correlation of $S$ for two randomly sampled viruses ($s_i$ and $s_j$) can be expressed explicitly in terms of the amount of shared history by expressing $s_i$ and $s_j$ as the sum of two components:

$$\left.\begin{aligned} s_i &= s(k_{ij}) + s(k_i - k_{ij}), \\ s_j &= s(k_{ij}) + s(k_j - k_{ij}), \end{aligned}\right\} \tag{A 1}$$

where $k_{ij}$ is the number of generations since initial infection that viruses $i$ and $j$ shared a common ancestor, $s(k_{ij})$ is the number of mutations occurring in the lineage during this period of shared history, $s(k_i - k_{ij})$ is the number of mutations occurring in the lineage of virus $i$ after the most recent common ancestor, and $s(k_j - k_{ij})$ is the number of mutations occurring in the lineage of virus $j$ after the most recent common ancestor. The conditional covariance of $s_i$ and $s_j$ is

$$\begin{aligned} &\mathrm{Cov}[s_i, s_j \mid k_{ij}, k_i, k_j] \\ &\quad = \mathrm{Cov}[s(k_{ij}), s(k_{ij})] + \mathrm{Cov}[s(k_{ij}), s(k_j - k_{ij})] \\ &\quad + \mathrm{Cov}[s(k_i - k_{ij}), s(k_{ij})] + \mathrm{Cov}[s(k_i - k_{ij}), s(k_j - k_{ij})]. \end{aligned} \tag{A 2}$$

Since evolutionary changes in distinct lineages are independent, all terms except the first are zero. Thus,

$$\mathrm{Cov}[s_i, s_j \mid k_{ij}, k_j, k_i] = \mathrm{Cov}[s(k_{ij}), s(k_{ij})] = \mathrm{Var}[S \mid k_{ij}]. \tag{A 3}$$

By analogy with eqn 4,

$$\mathrm{Cov}[s_i, s_j] = \mu_s E[k_{ij}] + \mu_s^2 \mathrm{Var}[k_{ij}] \tag{A 4}$$

and after dividing this expression by eqn 4,

$$\rho = \frac{\mu_s E[k_{ij}] + \mu_s^2 \mathrm{Var}[k_{ij}]}{\mu_s E[k] + \mu_s^2 \mathrm{Var}[k]}, \tag{A 5}$$

which, after a slight modification of notation, gives eqn 9.

## Appendix 2

Let the sequence contain $L$ linkage groups which recombination may occur among but not within. Let $v_i$ be the number of synonymous substitutions within the $i$th linkage group. Thus, the total number of synonymous mutations on a randomly sampled sequence, $S$, is:

$$S = \sum_{i=1}^{L} v_i \tag{A 6}$$

and

$$E[S] = \sum_{i=1}^{L} E[v_i] = \sum_{i=1}^{L} \mu_{s_i} E[k_i], \tag{A 7}$$

where $\mu_{s_i}$ is the synonymous mutation rate within the $i$th linkage group and $k_i$ is the number of ancestors of the $i$th linkage group. Because the average number of ancestors is the same for all linkage groups,

$$E[S] = E[k] \sum_{i=1}^{L} \mu_{s_i} = \mu_s E[k], \tag{A 8}$$

which is equivalent to the result for no recombination (eqn 3).

The variance in $S$ when there is recombination is:

$$\mathrm{Var}[S] = \sum_{i=1}^{L} \sum_{j=1}^{L} \mathrm{Cov}[v_i, v_j], \tag{A 9}$$

where

$$\mathrm{Cov}[v_i, v_j] = \mu_{s_i} \mu_{s_j} \mathrm{Cov}[k_i, k_j]. \tag{A 10}$$

Denoting $\pi_{ij}$ as the correlation in the number of ancestors between the $i$th and $j$th linkage unit on a randomly sampled virus, we find that:

$$\mathrm{Var}[S] = \mu_s E[k] + \mathrm{Var}[k] \sum_{i=1}^{L} \sum_{j=1}^{L} \mu_{s_i} \mu_{s_j} \pi_{ij}. \tag{A 11}$$

The correlation coefficient, $\pi_{ij}$, is determined by the level of recombination between the $i$th and $j$th linkage groups. If the two linkage groups are completely linked, then the number of ancestors per group on a single virus will inevitably be the same. In this case, $\pi_{ij} = 1$. When $\pi_{ij}$ is 1 for all $i$ and $j$, then eqn A 11 reduces to eqn 4. If there is free recombination between the $i$th and $j$th linkage groups (such that parental contributors of these groups to the sampled virus were different), then the number of ancestors per group will be unrelated. In this case, $\pi_{ij} = 0$ and $\mathrm{Var}[S]$ is reduced relative to eqn 4.

https://doi.org/10.1017/S0016672300032481 Published online by Cambridge University Press