Available online at www.sciencedirect.com

SCIENCE DIRECT

Virology 329 (2004) 101–108

VIROLOGY

www.elsevier.com/locate/yviro

Molecular clock-like evolution of human immunodeficiency virus type 1

Yi Liu[a,*], David C. Nickle[a], Daniel Shriner[a], Mark A. Jensen[a], Gerald H. Learn Jr.[a], John E. Mittler[a], James I. Mullins[a,b,c]

[a]*Department of Microbiology, University of Washington School of Medicine, Seattle, WA 98195, United States*
[b]*Department of Medicine, University of Washington School of Medicine, Seattle, WA 98195, United States*
[c]*Department of Laboratory Medicine, University of Washington School of Medicine, Seattle, WA 98195, United States*

Received 19 April 2004; returned to author for revision 22 June 2004; accepted 16 August 2004
Available online 18 September 2004

Abstract

The molecular clock hypothesis states that the rate of nucleotide substitution per generation is constant across lineages. If generation times were equal across lineages, samples obtained at the same calendar time would have experienced the same number of generations since their common ancestor. However, if sequences are not derived from contemporaneous samples, differences in the number of generations may be misinterpreted as variation in substitution rates and hence may lead to false rejection of the molecular clock hypothesis. A recent study has called into doubt the validity of clock-like evolution for HIV-1, using molecular sequences derived from noncontemporaneous samples. However, after separating their within-individual data according to sampling time, we found that what appeared to be nonclock-like behavior could be attributed, in most cases, to noncontemporaneous sampling, with contributions also likely to derive from recombination. Natural selection alone did not appear to obscure the clock-like evolution of HIV-1.

Keywords: Molecular clock; HIV-1; Natural selection; Recombination

Introduction

The assumption of evolution progressing in a predictable, clock-like manner is widely used in studies of viral evolution, including HIV-1 (Drummond and Rodrigo, 2000; Korber et al., 2000; Kuhner et al., 1995; Leitner and Albert, 1999). The molecular clock hypothesis was introduced in comparative studies of the hemoglobin (Zuckerkandl and Pauling, 1965) and cytochrome *c* (Margoliash, 1963) proteins, in which the rates of amino acid replacement were shown to be approximately equal among various mammalian lineages. According to the neutral theory of evolution, for any given gene (or protein), mutations arise according to a Poisson process in which the average rate of nucleotide (or amino acid) substitution is approximately constant over time in all lineages, and is equal to the neutral mutation rate (Kimura, 1983). Furthermore, the variance of the substitution rate is expected to be equal to the mean (Kimura, 1983).

However, studies of different proteins and species have shown that the variance of the substitution rate is often greater than the mean, a phenomenon called overdispersion (Gillespie, 1984, 1986). Studies of the constancy of substitution rates of glycerol-3-phosphate dehydrogenase, superoxide dismutase, and xanthine dehydrogenase (Ayala, 2000; Rodriguez-Trelles et al., 2001) and of various genes from three species of *Drosophila* (Takano, 1998) have revealed erratic substitution patterns. Natural selection has been suggested as an explanation for these inconsistencies (Ayala, 2000; Gillespie, 1984, 1986; Rodriguez-Trelles et al., 2001; Takano, 1998).

* Corresponding author. Department of Microbiology, University of Washington School of Medicine, P.O. Box 358070, Seattle, WA, 98195-8070. Fax: +1 206 732 6167.
E-mail address: yiliu197@u.washington.edu (Y. Liu).

0042-6822/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2004.08.014

Natural selection is critical to the pathogenesis of HIV-1 and SIV (Burns and Desrosiers, 1994; Koenig et al., 1995; Mayers et al., 1992; Moore et al., 2002; Overbaugh and Bangham, 2001; Ross and Rodrigo, 2002; Trachtenberg et al., 2003). Diversifying selection favors mutants that are different from their ancestors and is commonly associated with immune escape (Borrow et al., 1997; Soudeyns et al., 1999). Directional selection increases the frequency of one genotype at the expense of others, regardless of the relative frequencies in the population. The latter has been associated with HIV drug resistance (Frost et al., 2001), cell tropism, and coreceptor usage (Hoffman and Doms, 1998). Frequency-dependent selection takes place when fitness is dependent on genotype frequencies. In vitro competition studies between isogenic strains of HIV-1 indicate that the replicative fitness of a particular strain is dependent on the proportion of that strain in the culture (Yuste et al., 2002). Thus, frequency-dependent selection may also contribute to the development of genetic variation in HIV-1 in vivo (Yuste et al., 2002).

Most investigations to date have supported the hypothesis that HIV-1 evolves in a clock-like manner. The effect of natural selection on the molecular clock is not clear. In a study of the evolutionary rates and base substitution patterns of HIV, hepatitis B virus, and influenza A virus, it was shown that the evolution of the analyzed viral genes could readily be explained by the neutral theory of molecular evolution (Gojobori et al., 1990). By analyzing the *env* V3 and *gag* p17 regions of HIV-1 in a known transmission chain, Leitner and Albert (1999) showed that neither the V3 nor the p17 mutational processes were overdispersed, which indicated that the introduction of nucleotide substitutions could be described adequately by a simple stochastic Poisson process, consistent with the molecular clock hypothesis. However, Jenkins et al. (2002), using the same data set (Leitner and Albert, 1999), rejected the existence of the molecular clock using a likelihood-ratio test (LRT). A recent paper by Posada and Crandall (2001) also called into doubt the validity of a molecular clock in describing the evolution of HIV-1. They studied viral variation at a variety of levels: within individuals, within a subtype, within an HIV group, and among groups. Three genes, *env*, *gag*, and *pol*, were analyzed at each of these levels. Their LRT analyses rejected the existence of a molecular clock for all three genes at all four levels.

HIV-1 has high replication and mutation rates that generate high levels of genetic variation, and its *env* gene diverges at a rate of about 1% per year (Shankarappa et al., 1999). Thus, significant levels of evolutionary change can occur in a matter of weeks or months. The LRT used by Posada and Crandall assumed that all lineages had equal time to evolve. Leitner and Albert's HIV-1 data set and most of Posada and Crandall's data sets are from different individuals. It is unlikely that viruses sampled from different hosts have evolved for the same number of generations. However, a nearly homogeneous population at the *env* locus has been shown to exist early in infection (Delwart et al., 1994; Zhang et al., 1993; Zhu et al., 1993). Thus, contemporary samples obtained after early infection within an individual are more likely to consist of viruses that have evolved for equal numbers of generations (see below for a discussion of when this might not be the case). Consequently, the hypothesis of a molecular clock is less likely to be rejected for such data sets. Upon inspection of the within-individual data sets analyzed by Posada and Crandall, we found that they contained sequences sampled over a 5-year period. After separating the sequences into two groups according to the time of their sampling (1985 or 1990), we retested the molecular clock hypothesis for the *env*, *gag*, and *pol* genes. We found that the conclusion of nonclock-like behavior was due primarily to noncontemporaneous sampling. However, the molecular clock hypothesis was rejected with marginal significance for some of the genes examined, within the same sampling time. We therefore examined other factors that may cause rejection of the molecular clock hypothesis, such as recombination and natural selection. We found that recombination might contribute to rejection of the molecular clock hypothesis in these data, while natural selection alone was unlikely to cause rejection of the clock according to our simulations.

Results

LRT of clock-like behavior in contemporaneous samples

Table 1 shows that when the 1985 and 1990 within-individual data from Posada and Crandall were combined, the null hypothesis of a molecular clock was rejected with extremely low *P* values in all three genes. When we separated the data into two groups according to the time of sampling, for the 1985 *gag* data set, the molecular clock was still strongly rejected (Table 1). However, the null hypothesis of a molecular clock was not rejected for the 1990 *pol* data set. Whereas for all other data sets, the null hypothesis was rejected with weak significance, and the *P* values were at least six orders of magnitude higher than those obtained when evaluating the combined data. Similar results were obtained using parametric bootstrapping (Table 1). Thus, the previous rejection of the null hypothesis of a molecular clock was largely associated with noncontemporaneous sampling.

Potential effect of recombination or natural selection

We next investigated the cause of the departure from the molecular clock in these separated data sets. To identify regions potentially having experienced recombination or natural selection, we used the program PLATO (Grassly and Holmes, 1997). We found that the overall phylogeny and mutation model fit the data poorly in the p2 and p7 regions of *gag* (*gag* 975–1217, corresponding

Table 1
Results of the LRT and the parametric bootstrap test of the molecular clock hypothesis[a]

Data	Sampling time	Nonclock-$\ln L_1$	Clock-$\ln L_0$	P (LRT)	P (parametric bootstrapping)
gag	1985 and 1990	2676	2714	10^{-13}	–
gag	1985	2329.1773	2349.4811	10^{-9}	<0.001
gag	1990	2294.6247	2299.3755	0.023	0.027
pol	1985 and 1990	5246	5322	10^{-29}	–
pol	1985	4407.2657	4412.4731	0.015	0.027
pol	1990	4747.4854	4750.5420	0.106	0.100
env	1985 and 1990	5337	5363	10^{-8}	–
env	1985	4384.3980	4388.8647	0.030	0.045
env	1990	4806.1814	4810.2296	0.044	0.047

[a] All data are from Posada and Crandall's within-individual data sets (Posada and Crandall, 2001). The -$\ln L_1$ and -$\ln L_0$ values for the combined 1985 and 1990 samples are taken from their Table 6, and the P values by LRT were calculated from the log likelihood values they reported.

position in HIV-1_{HXB2}) in the 1985 data set, and the p6 region (*gag* 1378–1503, corresponding position in HIV-1_{HXB2}) in the 1990 data set, suggesting that these regions may have experienced recombination or natural selection. Indeed, when these regions were excluded, the molecular clock was not rejected (Table 2). In contrast, random deletion of equal amounts of data outside these regions resulted in rejection rates by the LRT of 100% and 97% for 1985 and 1990 *gag* data sets, respectively. Thus, support of the molecular clock hypothesis in Table 2 was not likely to be due to a loss of power by the exclusion of data. PLATO did not detect anomalous regions in the 1985 or 1990 *env* or the 1985 *pol* data sets.

Potential effect of natural selection

To examine the effect of natural selection on clock-like behavior, simulations under directional, diversifying, and frequency-dependent selection were performed and the existence of clock-like behavior was examined by the LRT. As shown in Table 3, in the neutral case, the null hypothesis of a molecular clock was rejected in 10% of the simulations. In simulations of directional selection, the rejection rates were 7%, 7%, and 12% for values of the selection coefficient s equal to 0.001, 0.01, and 0.1, respectively. Furthermore, none of the distributions of the LRT test statistic, $-2 \log \Lambda$, from simulations of directional selection was significantly different from the neutral distribution. Under diversifying selection, the rejection rates were 7%, 13%, and 8% for s values of 0.001, 0.01, and 0.1, respectively. As in the simulations of directional selection, none of the three distributions of $-2 \log \Lambda$ was significantly different from the neutral distribution.

In our simulations of neutral evolution, the null hypothesis of the molecular clock was rejected for 10 of 100 simulated data sets by the LRT (Table 3). It is known that for many nested phylogenetic hypotheses, the χ^2 distribution is not appropriate, and LRT are likely to reject null hypotheses, including the null hypothesis of the molecular clock, too readily (Goldman, 1993; Huelsenbeck et al., 1996). Parametric bootstrapping has been suggested as a better method for estimating the null distribution of the test statistic in question (Goldman, 1993; Huelsenbeck et al., 1996). When tested by parametric bootstrapping, 4 of the 10 neutral replicates that rejected the molecular clock hypothesis by the LRT now did not reject. By a one-sample Kolmogorov–Smirnov test, the distribution of $-2 \log \Lambda$ of our neutral simulation is marginally different from the χ^2 distribution with 18 degrees of freedom, $P = 0.075$. Thus, the relatively higher rejection rate by LRT may be because the χ^2 distribution is not an accurate estimate of the true distribution of the statistic in this case.

In simulations of frequency-dependent selection, the rejection rate was 15% with two possible genotypes at the site experiencing selection and 9% with 20 possible genotypes (Table 3). Neither distribution of $-2 \log \Lambda$ was significantly different from the neutral distribution.

To test the effect of multiple sites independently experiencing positive selection, simulations were done on a locus of 1000 bp containing 10 or 100 amino acid sites under directional or diversifying selection, with all of the gains in fitness being additive. As shown in Table 3, with 10 sites experiencing selection, the rejection rate was 11% under directional and 10% under diversifying selection. With 100 sites experiencing selection, the rejection rate was 12% under directional and 11% under diversifying selection (Table 3). Again, none of these four distributions of $-2 \log \Lambda$ was significantly different from the neutral distribution.

Before the occurrence of selectively advantageous mutations and after their fixation, all sequences evolve neutrally. In samples collected between these two events, positive selection is in action, and the site under selection

Table 2
Results of the LRT and the parametric bootstrap test for the *gag* gene[a]

Sampling Time	Nonclock-$\ln L_1$	Clock-$\ln L_0$	P (LRT)	P (parametric bootstrapping)
1985	1889.2401	1891.8474	0.157	0.147
1990	2006.0351	2008.4212	0.190	0.172

[a] Regions identified by PLATO as potentially having experienced recombination or natural selection were excluded from the *gag* gene of Posada and Crandall's within-individual data sets (Posada and Crandall, 2001).

Table 3
Results of simulations of different types of selection

Types of selection[a]			Rejection rate[b]	P[c]
Neutral		$s = 0$	0.10	–
Directional selection	Single site	$s = 0.001$	0.07	0.443
		$s = 0.01$	0.07	0.677
		$s = 0.1$	0.12	0.894
	Multiple sites	10 sites, $s = 0.01$	0.11	0.677
		100 sites, $s = 0.01$	0.12	0.261
Diversifying selection	Single site	$s = 0.001$	0.07	0.556
		$s = 0.01$	0.13	0.344
		$s = 0.1$	0.08	0.961
	Multiple sites	10 sites, $s = 0.01$	0.10	0.344
		100 sites, $s = 0.01$	0.11	0.894
Frequency-dependent selection		2 genotypes, $s = 0.01$	0.15	0.794
		20 genotypes, $s = 0.01$	0.09	0.443

[a] s:selection coefficient.
[b] False-positive rejection rate by LRT.
[c] P values of Kolmogorov–Smirnov tests of the probability distributions of $-2 \log \Lambda$ between simulations of selection and the neutral distribution.

is polymorphic as both ancestral and mutant states cosegregate. Thus, rejection of the molecular clock hypothesis may be more likely to occur during this time interval. To examine this hypothesis, 100 simulations with $s = 0.1$ were conducted and, to ensure high polymorphism, samples were collected when the frequency of the advantageous mutation first exceeded 50%. Using the LRT, the molecular clock hypothesis was rejected 8% of the time under directional selection and 12% of the time under diversifying selection (data not shown). These results suggest that the molecular clock hypothesis would generally not be rejected for samples containing mutations in the process of becoming fixed due to either directional or diversifying selection.

Discussion

The molecular clock hypothesis assumes that, on average, lineages with equal time to evolve will accumulate equal amounts of variation. HIV-1 has genes with divergence rates as high as 1% per year (Shankarappa et al., 1999). Therefore, HIV-1 lineages with significantly different amounts of time to evolve may accumulate significantly different amounts of variation. The within-individual data used by Posada and Crandall were sampled over a period of 5 years, which is the estimated mean fixation time of a new mutation in HIV-1 (Shriner et al., 2004b). Hence, we might expect to see a significant accumulation of substitutions in the period of time between the 1985 and 1990 samples. Such noncontemporaneous sampling could result in the rejection of the hypothesis of a molecular clock. Our reexamination of a within-individual data set used by Posada and Crandall (2001) demonstrated that their rejection of the hypothesis of a molecular clock was predominantly due to noncontemporaneous sampling, with likely contributions from recombination. Furthermore, natural selection alone had generally little effect on the clock-like evolution of HIV-1.

Using PLATO, we identified regions in the *gag* gene data sets potentially having experienced recombination or natural selection. Upon exclusion of these regions, the hypothesis of a molecular clock was no longer rejected for the contemporaneous *gag* gene data sets. Recombination causes different regions of a sequence to reflect different phylogenetic histories. The reconstructed maximum-likelihood phylogeny is therefore not an accurate reflection of the real history of all sites. Simulation studies have found that even low levels of unacknowledged recombination can cause rejection of the molecular clock hypothesis (Posada, 2001; Schierup and Hein, 2000). A recombination rate, ρ (the number of recombination events between two sequences per $2N$ generations, where N is the effective population size), as low as 1 can lead to a rejection rate of approximately 30% (Schierup and Hein, 2000). When $\rho > 8$, the null hypothesis is rejected in almost all cases (Schierup and Hein, 2000). Given an estimated recombination rate for HIV-1 of at least 2.8 crossovers per genome per generation in vitro (Levy et al., 2004; Zhuang et al., 2002), which approximates or exceeds the point mutation rate in vivo (Shriner et al., 2004a), recombination is very likely to have contributed to the rejection of the hypothesis of molecular clock in the *gag* gene data sets described here.

Our simulation results indicate that without recombination and migration, evolution is consistent with a molecular clock when both the mutation rate and the effective population size are constant, regardless of whether positive selection is active. Therefore, the weak departures from the molecular clock hypothesis noted for the *env* and *pol* genes were unlikely to be due to natural selection alone. Other processes, such as the existence of reservoirs and compartmentalization (Nickle et al., 2003), are likely to contribute to rejections of the molecular

clock hypothesis. Viral sequences emerging from latently infected cells or compartments can be derived from multiple time points in the past, thereby leading to a collection of noncontemporaneous sequences and the rejection of a molecular clock. In addition, even contemporaneous samples could have gone through different numbers of replication cycles. When different lineages have different generation times, such lineage effects (Li et al., 1987) reflect the fact that samples collected at the same calendar time will have evolved for different numbers of generations. Thus, the distances from the contemporaneous lineages to their common ancestor will be different, resulting in the false rejection of clock-like behavior.

The assumption of evolution progressing in a predictable, clock-like manner is widely used in studies of viral evolution to date transmission events (Korber et al., 2000; Pybus and Rambaut, 2002; Pybus et al., 2000; Rambaut et al., 1997), estimate generation times (Drummond and Rodrigo, 2000; Rambaut, 2000), and estimate population genetics parameters such as population-scaled rates of mutation (Kuhner et al., 1995), recombination (Kuhner et al., 2000), migration (Beerli and Felsenstein, 1999, 2001), and growth (Kuhner et al., 1998). Our simulation results suggest that clock assumption in such studies is robust to violation by selection.

Materials and methods

Likelihood-ratio testing of the null hypothesis of a molecular clock

To test the molecular clock hypothesis, LRTs were performed as described by Posada and Crandall (2001). Maximum-likelihood trees were inferred using PAUP* (Swofford, 1999). Under the null hypothesis of substitution-rate constancy across lineages, trees were inferred under the constraint of a molecular clock (Fig. 1A). Under the alternative hypothesis of substitution-rate heterogeneity across lineages, trees were inferred without the constraint of a molecular clock (Fig. 1B). Maximum-likelihood values under the null and alternative hypotheses were compared using the following statistic:

$$\Lambda = \frac{\max[L(\text{null hypothesis}|\text{data})]}{\max[L(\text{alternative hypothesis}|\text{data})]}$$

When nested hypotheses are examined (i.e., the null hypothesis is a special case of the alternative hypothesis), $-2\log\Lambda$ is approximately χ^2 distributed with q degrees of freedom, where q is the difference in the number of free parameters between the null and alternative hypotheses. In this case, the free parameters refer to branches that can vary independently in length. For a bifurcating, unrooted tree of n sequences without the constraint of a molecular clock, there are $2n-3$ branches, each of which can vary independently in length. For a bifurcating, rooted tree of n sequences with the constraint of a molecular clock, there are $2n-2$ branches, $n-1$ of which can vary independently in length. Thus, there are $(2n-3)-(n-1)=n-2$ degrees of freedom (Felsenstein, 1981). The appropriateness of this approximation can be assessed by comparing it to a simulated distribution of the likelihood statistic generated by parametric bootstrapping with selection coefficient $s=0$; we used the one-sample Kolmogorov–Smirnov test for this comparison. The significance level α for the LRT was chosen as 5%.

Parametric bootstrap testing of the null hypothesis of a molecular clock

Parametric bootstrapping (Goldman, 1993; Huelsenbeck and Rannala, 1997) was used to generate the null distribution of the test statistic ($-2\log\Lambda$). One thousand data sets were simulated using the program Seq-Gen (Rambaut and Grassly, 1997) under the constraint of a molecular clock, using the parameters of the substitution model estimated by maximum-likelihood from the original data. The $-2\log\Lambda$ values were calculated as above and the $-2\log\Lambda$ value obtained from the original data was compared to this distribution.

PLATO

This program uses a sliding window method to identify regions of an alignment that poorly support a given overall phylogenetic topology and model of nucleotide substitution

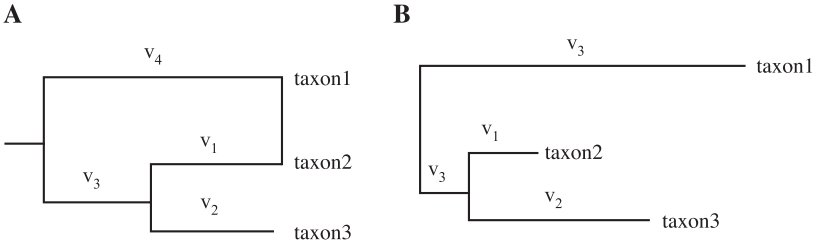

Fig. 1. Idealized tree structures under the null and alternative hypotheses for LRT of a molecular clock. "v" indicates branch lengths. (A) Under the null hypothesis with the constraint of a molecular clock, the substitution rate is constant across lineages. All contemporaneously sampled sequences in the tree have the same distance from the tip to the root. In the rooted case of $n=3$ sequences, $v_1=v_2$ and $v_1+v_3=v_4$, such that there are $n-1=2$ free branch length parameters. (B) Under the alternative hypothesis without the constraint of a molecular clock, the rate is heterogeneous across lineages. In the unrooted case of $n=3$ sequences, there are $2n-3=3$ free branch length parameters.

(Grassly and Holmes, 1997). We used this program to identify regions potentially having experienced recombination or natural selection.

Forward simulation

To examine the effect of natural selection on clock-like behavior, we used a forward simulator based on the Wright–Fisher model (Fisher, 1930; Wright, 1931). According to this model, generations are non-overlapping, and each offspring is randomly derived from a parent of the previous generation. To address the effect of overlapping generations, we assumed that in the short term a steady state is reached between viral clearance and production. Assuming a simple exponential decay according to $N(t) = N(0)e^{-ct}$, in which the population of size $N(0)$ decays to size $N(t)$ after t days, with a clearance rate constant of plasma virions, c, of 23/day (Ramratnam et al., 1999) and a generation time of 2.0 days (Markowitz et al., 2003), virtually 100% of plasma virions are new each generation. Thus, the model of non-overlapping generations is adequate in describing HIV-1 evolution.

A model of haploid organisms with a constant population size of N was used. Fig. 2 shows the probability of an individual of the current generation being the parent of a particular offspring in the next generation. Under neutrality, this probability is $\frac{1}{N}$ (Fig. 2A). With natural selection, this probability will change from generation to generation. Suppose that there are two competing genotypes, *adv* (for advantageous genotype) and *dis* (for disadvantageous genotype), with respective relative fitness w_{adv} and w_{dis}. Suppose that in generation t the frequency ratio of *adv* to *dis* is $\frac{N_{adv}}{N_{dis}}$. In generation $t+1$, the expectation of the frequency ratio will be $\frac{w_{adv}N_{adv}}{w_{dis}N_{dis}}$. Thus, the expected frequency of *adv* will be $\frac{w_{adv}N_{adv}}{w_{adv}N_{adv}+w_{dis}N_{dis}}$ and that of *dis* will be $\frac{w_{dis}N_{dis}}{w_{adv}N_{adv}+w_{dis}N_{dis}}$. The probability of any individual *adv* being the parent of a particular offspring in the next generation will be $\frac{w_{adv}}{w_{adv}N_{adv}+w_{dis}N_{dis}}$ and the probability of any individual *dis* being the parent will be $\frac{w_{dis}}{w_{adv}N_{adv}+w_{dis}N_{dis}}$ (Fig. 2B). Under directional and diversifying selection, if $w_{dis} = 1$, then $w_{adv} = 1 + s$, where s is the selection coefficient and is greater than zero.

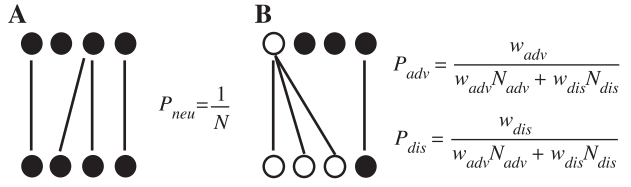

Fig. 2. The probability of an individual of the current generation becoming the parent of a particular offspring of the next generation. (A) Under neutrality, the probability, p_{neu}, is $(1/N)$. (B) Under natural selection, the probabilities for an *adv* (advantageous genotype; open circle), p_{adv}, and for a *dis* (disadvantageous genotype; filled circle), p_{dis}, are given. The circles on the top row represent parents and those on the bottom represent offspring. w_{adv} and w_{dis} are fitness values for *adv* and *dis*, respectively. N is the total population size, N_{adv} is the population size of *adv*, and N_{dis} is the population size of *dis*.

Under frequency-dependent selection, the rare genotype is favored: the higher the frequency, the lower the fitness. Suppose w_i $(i = 1,\ldots,n)$ is the relative fitness value of each of the n competing genotypes. In our model of frequency-dependent selection, fitness of a genotype is linearly dependent on its frequency: $w_i = 1+s(1-p_i)$, in which p_i is the frequency of the i^{th} genotype in the population. The equilibrium mean frequency of any genotype is $1/n$.

For all of the parameters in the forward simulations, we used values based on previous studies of HIV-1 evolution. The effective population size of HIV-1 within an individual has in most cases been estimated to be on the order of 10^3 (Leigh Brown, 1997; Seo et al., 2002; Shriner et al., 2004a,b). The point mutation rate of HIV-1 has been estimated to be 2.5×10^{-5} per site per generation (Mansky, 1996). A nearly homogeneous population at the *env* locus has been shown to exist early in infection (Delwart et al., 1994; Zhang et al., 1993; Zhu et al., 1993). Thus, we simulated a homogeneous population with a constant size of 1000, experiencing a mutation rate of 2.5×10^{-5} per site per generation, with a locus of 1000 nucleotides. Recombination and migration were not considered. The existence of a molecular clock is determined by the number of mutational events that have accumulated on each lineage, and is relatively independent of the identities of the bases involved. Therefore, the choice of substitution model will have little to no effect on our conclusions. Thus, for simplicity, we assumed that mutations occur with equal probability among all four nucleotides and that the four nucleotides occur with equal frequency. Samples of 20 sequences were collected at the 1000th generation. Four selective regimes were simulated: $s = 0$ (strictly neutral), $s = 0.001$ ($sN = 1$, nearly neutral), $s = 0.01$ (weakly selected), and $s = 0.1$ (strongly selected). Sites under selection were randomly chosen. For each selective regime, 100 simulations for both directional and diversifying selection were conducted. Under directional selection, mutation to one particular amino acid at the queried position is advantageous over all the other genotypes. Under diversifying selection, mutations are advantageous when they change the amino acid at the queried position from the original amino acid in the first generation. Frequency-dependent selection on a randomly chosen amino acid site with $s = 0.01$ was simulated, with 2 or 20 genotypes, to assess the extreme possibilities of one mutant allele form, and all possible amino acid changes at that site. Using the initial sequence as an outgroup, the null hypothesis of a molecular clock was tested by the LRT. The two-sample Kolmogorov–Smirnov test was used to determine if the probability distributions of $-2 \log \Lambda$ from simulations of selection were different from that from neutral evolution.

Acknowledgments

We thank Katherine Davis for editorial assistance. This work was supported by grants from the US Public Health

Services including support from the University of Washington Center for AIDS Research.

References

Ayala, F.J., 2000. Neutralism and selectionism: the molecular clock. Gene 261, 27–33.

Beerli, P., Felsenstein, J., 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152, 763–773.

Beerli, P., Felsenstein, J., 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. U.S.A. 98, 4563–4568.

Borrow, P., Lewicki, H., Wei, X., Horwitz, M.S., Peffer, N., Meyers, H., Nelson, J.A., Gairin, J.E., Hahn, B.H., Oldstone, M.B., Shaw, G.M., 1997. Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. Nat. Med. 3, 205–211.

Burns, D.P., Desrosiers, R.C., 1994. Envelope sequence variation, neutralizing antibodies, and primate lentivirus persistence. Curr. Top. Microbiol. Immunol. 188, 185–219.

Delwart, E.L., Sheppard, H.W., Walker, B.D., Goudsmit, J., Mullins, J.I., 1994. Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays. J. Virol. 68, 6672–6683.

Drummond, A., Rodrigo, A.G., 2000. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Mol. Biol. Evol. 17, 1807–1815.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Clarendon, Oxford.

Frost, S.D., Gunthard, H.F., Wong, J.K., Havlir, D., Richman, D.D., Leigh Brown, A.J., 2001. Evidence for positive selection driving the evolution of HIV-1 env under potent antiviral therapy. Virology 284, 250–258.

Gillespie, J.H., 1984. The molecular clock may be an episodic clock. Proc. Natl. Acad. Sci. U.S.A. 81, 8009–8013.

Gillespie, J.H., 1986. Natural selection and the molecular clock. Mol. Biol. Evol. 3, 138–155.

Gojobori, T., Moriyama, E.N., Kimura, M., 1990. Molecular clock of viral evolution, and the neutral theory. Proc. Natl. Acad. Sci. U.S.A. 87, 10015–10018.

Goldman, N., 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36, 182–198.

Grassly, N.C., Holmes, E.C., 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol. Biol. Evol. 14, 239–247.

Hoffman, T.L., Doms, R.W., 1998. Chemokines and coreceptors in HIV/SIV-host interactions. Aids 12 (Suppl A), S17–S26.

Huelsenbeck, J.P., Hillis, D.M., Nielsen, R., 1996. A likelihood-ratio test of monophyly. Syst. Biol. 45, 546–558.

Huelsenbeck, J.P., Rannala, B., 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276, 227–232.

Jenkins, G.M., Rambaut, A., Pybus, O.G., Holmes, E.C., 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J. Mol. Evol. 54, 156–165.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge Univ. Press, Cambridge.

Koenig, S., Conley, A.J., Brewah, Y.A., Jones, G.M., Leath, S., Boots, L.J., Davey, V., Pantaleo, G., Demarest, J.F., Carter, C., Wannebo, C., Yannelli, J.R., Rosenberg, S.A., Lane, C.H., 1995. Transfer of HIV-1-specific cytotoxic T-lymphocytes to an AIDS patient leads to selection for mutant HIV variants and subsequent disease progression. Nat. Med. 1, 330–336.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., Bhattacharya, T., 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288, 1789–1796.

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

Kuhner, M.K., Yamato, J., Felsenstein, J., 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149, 429–434.

Kuhner, M.K., Yamato, J., Felsenstein, J., 2000. Maximum likelihood estimation of recombination rates from population data. Genetics 156, 1393–1401.

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

Leitner, T., Albert, J., 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc. Natl. Acad. Sci. U.S.A. 96, 10752–10757.

Levy, D.N., Aldrovandi, G.M., Kutsch, O., Shaw, G.M., 2004. Dynamics of HIV-1 recombination in its natural target cells. Proc. Natl. Acad. Sci. U.S.A. 101, 4204–4209.

Li, W.H., Tanimura, M., Sharp, P.M., 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J. Mol. Evol. 25, 330–342.

Mansky, L.M., 1996. Forward mutation rate of Human Immunodeficiency Virus Type 1 in a T lymphoid cell line. AIDS Res. Hum. Retroviruses 12, 307–314.

Margoliash, E., 1963. Primary structure and evolution of cytochrome *c*. Proc. Natl. Acad. Sci. U.S.A. 93, 672–679.

Markowitz, M., Louie, M., Hurley, A., Sun, E., Di Mascio, M., Perelson, A.S., Ho, D.D., 2003. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. J. Virol. 77, 5037–5038.

Mayers, D.L., McCutchan, F.E., Sanders-Buell, E.E., Merritt, L.I., Dilworth, S., Fowler, A.K., Marks, C.A., Ruiz, N.M., Richman, D.D., Roberts, C.R., et al., 1992. Characterization of HIV isolates arising after prolonged zidovudine therapy. J. Acquired Immune Defic. Syndr. 5, 749–759.

Moore, C.B., John, M., James, I.R., Christiansen, F.T., Witt, C.S., Mallal, S.A., 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science 296, 1439–1443.

Nickle, D.C., Shriner, D., Mittler, J.E., Frenkel, L.M., Mullins, J.I., 2003. Importance and detection of virus reservoirs and compartments of HIV infection. Curr. Opin. Microbiol. 6, 410–416.

Overbaugh, J., Bangham, C.R., 2001. Selection forces and constraints on retroviral sequence variation. Science 292, 1106–1109.

Posada, D., 2001. Unveiling the molecular clock in the presence of recombination. Mol. Biol. Evol. 18, 1976–1978.

Posada, D., Crandall, K.A., 2001. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. 18, 897–906.

Pybus, O.G., Rambaut, A., 2002. GENIE: estimating demographic history from molecular phylogenies. Bioinformatics 18, 1404–1405.

Pybus, O.G., Rambaut, A., Harvey, P.H., 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155, 1429–1437.

Rambaut, A., 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16, 395–399.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.

Rambaut, A., Harvey, P.H., Nee, S., 1997. End-Epi: an application for inferring phylogenetic and population dynamical processes from molecular sequences. Comput. Appl. Biosci. 13, 303–306.

Ramratnam, B., Bonhoeffer, S., Binley, J., Hurley, A., Zhang, L., Mittler, J.E., Markowitz, M., Moore, J.M., Perelson, A.S., Ho, D.D., 1999. Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. Lancet 354, 1782–1785.

Rodriguez-Trelles, F., Tarrio, R., Ayala, F.J., 2001. Erratic overdispersion of three molecular clocks: GPDH, SOD, and XDH. Proc. Natl. Acad. Sci. U.S.A. 98, 11405–11410.

Ross, H.A., Rodrigo, A.G., 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., Hein, J., 2000. Recombination and the molecular clock. Mol. Biol. Evol. 17, 1578–1579.

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

Shankarappa, R., Margolick, J.B., Gange, S.J., Rodrigo, A.G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C.R., Learn, G.H., He, X., Huang, X.L., Mullins, J.I., 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. 73, 10489–10502.

Shriner, D., Rodrigo, A.G., Nickle, D.C., Mullins, J.I., 2004a. Pervasive genomic recombination of HIV-1 in vivo. Genetics (in press).

Shriner, D., Shankarappa, R., Jensen, M.A., Nickle, D.C., Mittler, J.E., Margolick, J.B., Mullins, J.I., 2004b. Influence of random genetic drift on HIV-1 env evolution during chronic infection. Genetics 166, 1155–1164.

Soudeyns, H., Paolucci, S., Chappey, C., Daucher, M.B., Graziosi, C., Vaccarezza, M., Cohen, O.J., Fauci, A.S., Pantaleo, G., 1999. Selective pressure exerted by immunodominant HIV-1-specific cytotoxic T lymphocyte responses during primary infection drives genetic variation restricted to the cognate epitope. Eur. J. Immunol. 29, 3629–3635.

Swofford, D.L., 1999. PAUP* 4.0: Phylogenetic Analysis Using Parsimony (*and Other Methods) 4.0b2a. Sinauer Associates, Inc., Sunderland, MA.

Takano, T.S., 1998. Rate variation of DNA sequence evolution in the *Drosophila* lineages. Genetics 149, 959–970.

Trachtenberg, E., Korber, B., Sollars, C., Kepler, T.B., Hraber, P.T., Hayes, E., Funkhouser, R., Fugate, M., Theiler, J., Hsu, Y.S., Kunstman, K., Wu, S., Phair, J., Erlich, H., Wolinsky, S., 2003. Advantage of rare HLA supertype in HIV disease progression. Nat. Med. 9, 928–935.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.

Yuste, E., Moya, A., Lopez-Galindez, C., 2002. Frequency-dependent selection in human immunodeficiency virus type 1. J. Gen. Virol. 83, 103–106.

Zhang, L.Q., MacKenzie, P., Cleland, A., Holmes, E.C., Leigh Brown, A.J., Simmonds, P., 1993. Selection for specific sequences in the external envelope protein of HIV-1 upon primary infection. J. Virol. 67, 3345–3356.

Zhu, T., Mo, H., Wang, N., Nam, D.S., Cao, Y., Koup, R.A., Ho, D.D., 1993. Genotypic and phenotypic characterization of HIV-1 in patients with primary infection. Science 261, 1179–1181.

Zhuang, J., Jetzt, A.E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B.D., Dougherty, J.P., 2002. Human immunodeficiency virus type 1 recombination: rate fidelity, and putative hot spots. J. Virol. 76, 11273–11282.

Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. In: Bryson, V., Vogel, H.J. (Eds.), Evolving Genes and Proteins. Academic Press, New York, pp. 97–166.