

Estimating Mutation Rate and Generation Time from Longitudinal Samples of DNA Sequences

Yun-Xin Fu

Human Genetics Center, University of Texas at Houston

We present in this paper a simple method for estimating the mutation rate per site per year which also yields an estimate of the length of a generation when mutation rate per site per generation is known. The estimator, which takes advantage of DNA polymorphisms in longitudinal samples, is unbiased under a number of population models, including population structure and variable population size over time. We apply the new method to a longitudinal sample of DNA sequences of the *env* gene of human immunodeficiency virus type 1 (HIV-1) from a single patient and obtain 1.62×10^{-2} as the mutation rate per site per year for HIV-1. Using an independent data set to estimate the mutation rate per generation, we obtain 1.8 days as the length of a generation of HIV-1, which agrees well with recent estimates based on viral load data. Our estimate of generation time differs considerably from a recent estimate by Rodrigo et al. when the same mutation rate per site per generation is used. Some factors that may contribute to the difference among different estimators are discussed.

Introduction

Mutation rate per nucleotide per year is a fundamental quantity for studying the molecular evolution of an organism. Mutation rates are usually estimated by one of two approaches. The first approach is to use homologous DNA sequences from two species with divergence time calibrated by an independent source, usually paleontological data. The second approach is to directly examine the number of mutations over one or a few generations. The first approach is simpler and more economical but is applicable only when a reliable estimate of the divergence time is available. In comparison, the second approach is more widely applicable in principle, but it usually requires examination of either a large number of individuals or a large DNA segment to obtain a reasonable number of changes. The genetic state of the progenitor also needs to be known. For large animals, it is costly to use the second approach because of the lengthy generation time. For rapidly evolving organisms, DNA polymorphisms in longitudinal samples, that is, samples taken at a series of time points, provide another way to estimate mutation rates.

The generation time, or the length of a generation, of an organism is the average length of time between two identical and successive stages in the life cycle of the organism. For example, the generation time of animals of large body size can be defined as the average length of time for an adult to produce another adult; for a virus such as human immunodeficiency virus type 1, the generation time can be defined as the average length of time from the release of the virion until it infects another cell and causes the release of another virion. Generation time not only is part of the biological properties of an organism, but also plays an essential role in analyzing polymorphism data from a population, because population genetic models are usually developed

with units of time corresponding to a certain number of generations, rather than days or years.

The life cycle of large animals can be observed easily, and it is usually not a problem to derive a generation time. For example, 20 years is widely used for one human generation. It is difficult to observe the life cycle of small organisms, such as viruses, in vivo, so there is a need to estimate the generation time. DNA polymorphisms in longitudinal sample provide an opportunity to do so when an independent estimate of the mutation rate per generation is available. The purpose of this paper is to present a simple method for estimating mutation rate per site per year which also yields an estimate of generation time when the mutation rate per generation is known. We apply the method to a longitudinal sample from an HIV patient both to illustrate the method and to obtain an estimate of mutation rate and an estimate of generation time for HIV-1. The differences between the new method and the method of Rodrigo and Felsenstein (1999) will be discussed.

The Theory

Suppose a sample of n sequences is taken at time t from a population of a haploid organism. The choice to consider a haploid population is entirely for the convenience of later discussion, and the theory is almost identical for a diploid population.

Let d_{kl} be the number of nucleotide substitutions per site between sequences k and l ($k \neq l$).

$$E(d_{kl}) = \theta_t, \quad (1)$$

where E is expectation and θ_t is a quantity whose value is determined by the dynamics of the population. For example, it is well known from the coalescent theory (Kingman 1982a, 1982b; see, e.g., a recent review by Li and Fu [1999]) that under the neutral Wright-Fisher model with a constant effective population size N , the value of θ_t is $2N\mu$ (e.g., Tajima 1983), where μ is the mutation rate per site per generation. In general, θ_t is dependent on the time t at which the sample is taken.

Suppose that a second sample of m sequences is taken T days later from the same population. For se-

Key words: mutation rate, generation time, longitudinal sample, HIV, coalescent process.

Address for correspondence and reprints: Yun-Xin Fu, Human Genetics Center, University of Texas at Houston, 6901 Bertner Avenue S222, Houston, Texas 77030. E-mail: fu@hgc.sph.uth.tmc.edu.

Mol. Biol. Evol. 18(4):620–626. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

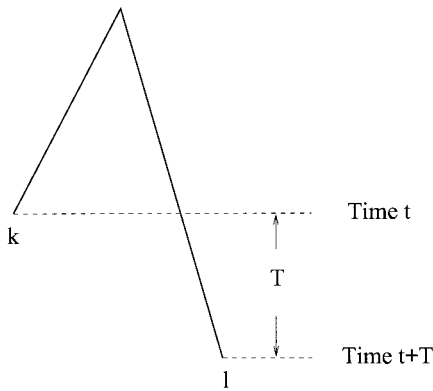


FIG. 1.—Schematic relationship between two sequences taken T days apart.

quences k and l ($k \neq l$) from the second sample, we have, similar to equation (1), that

$$E(d_{kl}) = \theta_{t+T}. \quad (2)$$

Again under the neutral Wright-Fisher model with a constant effective population size, $\theta_{t+T} = 2N\mu$. In addition to the pairwise number of nucleotide substitutions within each sample, we can examine nucleotide substitutions between two sequences, one from each sample. Let d'_{kl} be the number of nucleotide substitutions per site between sequence k from sample 1 and sequence l from sample 2. Then,

$$E(d'_{kl}) = \theta_t + \mu \times (G \times T), \quad (3)$$

where μ is the mutation rate per site per generation and G is the number of generations per day. This relationship can be seen clearly from figure 1. The first term to the right of the equality is due to the fact that at time t , the ancestral sequence of sequence l is just another random sequence from the population at time t ; thus, the expected number of nucleotide substitutions between this ancestral sequence and sequence k from sample 1 is thus θ_t . The second term to the right of the equality is the expected number of mutations per site occurring in the ancestral sequence of sequence k in T days, assuming that the number of mutations occurring in a sequence in a given number GT of generations is a Poisson variable with mean equal to μGT . It should be noted that although not every sequence in sample 2 has experienced the same number of generations in the past T days, equation (3) still holds because the mean number of generations for each sequence is $G \times T$. Let ν be the mutation rate per site per day. Then, $\nu = \mu G$. Therefore, equation ([REF:d12]) can be written as

$$E(d'_{kl}) = \theta_t + \nu T. \quad (4)$$

Let Π_i be the mean number of nucleotide substitutions per site between two sequences from sample i , and let Π_{12} be the mean number nucleotide substitutions per site between two sequences, one from sample 1 and one from sample 2. Then, we have

$$E(\Pi_1) = \theta_t \quad (5)$$

$$E(\Pi_2) = \theta_{t+T} \quad (6)$$

$$E(\Pi_{12}) = \theta_t + \nu T. \quad (7)$$

From these expectations, it is easy to see that an unbiased estimator of ν is

$$\hat{\nu} = \frac{\Pi_{12} - \Pi_1}{T}. \quad (8)$$

With the further assumption that θ_t is the same for different values of t , another unbiased estimator of ν is

$$\hat{\nu}' = \frac{\Pi_{12} - (\Pi_1 + \Pi_2)/2}{T}. \quad (9)$$

Although the estimator $\hat{\nu}'$ should have a smaller variance than $\hat{\nu}$, the assumption for ensuring its unbiasedness is likely questionable in many situations. Therefore, $\hat{\nu}$ is generally preferable to $\hat{\nu}'$.

If the mutation rate μ per site per generation is known or has been estimated independently, an unbiased estimator of G is

$$\hat{G} = \frac{\hat{\nu}}{\mu}. \quad (10)$$

The length L of a generation, i.e., the generation time, can be obtained as

$$\hat{L} = \hat{G}^{-1} = \frac{\mu}{\hat{\nu}} \quad (11)$$

Combining Estimates

Suppose that samples from r different time points were taken. For each pair of samples, one can obtain an estimate of G . Let G_{ij} be the estimate from samples i and j . Then, there are $r(r-1)/2$ estimates of G . How to combine these estimates to obtain an overall estimate of G is not only an interesting theoretical issue, but also of great practical importance. Since each G_{ij} is an unbiased estimate of G , for any weights $\alpha_{ij} \geq 0$ and $\sum_{ij} \alpha_{ij} = 1$, the quantity

$$G_\alpha = \sum_{i < j} \alpha_{ij} G_{ij} \quad (12)$$

is also an unbiased estimate of G . Therefore, there are many ways to combine pairwise estimates.

The simplest method is to take an unweighted average. That is, let $\alpha_{ij} = 2/[r(r-1)]$, resulting in

$$\bar{G}_s = \frac{2}{r(r-1)} \sum_{i < j} G_{ij}. \quad (13)$$

A problem with this estimator is that terms with large variances may dominate the final estimate. Therefore, a better way to combine estimates is to take the variance of each estimate into consideration. Let T_{ij} be the time length between samples i and j . Then, $\Pi_{ij} - \Pi_i$ reflects on average the number of mutations in a sequence during a period of length T_{ij} , which is approximately Poisson distributed. Therefore, the variance of $\Pi_{ij} - \Pi_i$ should be proportional to T_{ij} . In fact, it can be shown that

Table 1
The Mean Numbers (Π_{ij}) ($\times 10^2$) of Nucleotide Substitutions

SAMPLE	SIZE	TIME	SAMPLE					
			1	2	3	4	5	
1	13	0	3.92, 4.06					
2	15	214	4.31, 4.48	3.67, 3.79				
3	15	671	5.54, 5.79	4.72, 4.90	4.41, 4.58			
4	9	699	4.51, 4.68	4.25, 4.40	4.80, 4.99	4.24, 4.39		
5	8	1,005	4.67, 4.85	4.41, 4.57	4.72, 4.91	4.44, 4.60	3.92, 4.05	

NOTE.—The first and second values in each entry correspond, respectively, to the value of Π without correction and the value of Π with correction using Kimura's two-parameter model.

$$\text{Var}(\Pi_{ij} - \Pi_i) = c_{ij} + vT_{ij}, \quad (14)$$

where c_{ij} is a complex quantity which is a function of the sample sizes, T_{ij} , and parameters that affect population dynamics, such as population size, growth rate, and generation time. The reason why c_{ij} is even dependent on generation time is that it depends on the number of ancestral sequences the j th sample has at the time i th sample was taken, and this number is dependent on generation time. Since generation time is what we intend to estimate, an overall estimator G_α which is dependent on generation time is undesirable (although the variance of any G_α is dependent on generation time). Because of the complexity of c_{ij} , a reasonable strategy for guiding our choice of weights is to ignore both c_{ij} and the correlations among estimates. This simple strategy results in optimal weights being $\alpha_{ij} = T_{ij}/\sum_{i<j} T_{ij}$ and the corresponding estimators of v and G being

$$v_T = \frac{\sum_{i<j} (\Pi_{ij} - \Pi_i)}{\sum_{i<j} T_{ij}} \quad (15)$$

$$G_T = \mu^{-1}v_T, \quad (16)$$

where μ is the mutation rate per site per generation. The length L of a generation, i.e., the generation time, can thus be estimated as $L_T = G_T^{-1}$.

Variance

The variance of any of the G_α estimators is extremely complex even under the simple situation of constant population size. Not only is the formula for individual $\text{Var}(G_{ij})$ intractable, but the differences between G_{ij} values are correlated to each other. We therefore propose the use of bootstrap samples for estimating the variance of L_T (or G_T).

Let n_i be the number of sequences in sample i ($i = 1, \dots, r$). A bootstrap sample will consist of r subsamples, with the i th subsample obtained by selecting n_i sequences with replacement from the n_i sequences. This stratified bootstrap is reasonable in the absence of detailed knowledge of the dynamics of the population being studied. The bootstrap estimate of the variance of L_T is obtained by the following steps: (1) carry out bootstrap sampling for each of the r samples, (2) compute the value of L_T from the bootstrap samples, and (3) repeat steps 1 and 2 many times; the resulting sampling variance of L_T is then an estimate of the variance of L_T .

It should be noted that there are two levels of variance associated with L_T . One is due to the stochasticity of evolution, which leads to differences among replicates of the same evolutionary process. The other is due to the effect of sampling. The bootstrap estimate of variance described above only accounts for the variance due to sampling. Although one needs to be cautious in interpreting this component of variance, it nevertheless gives a lower bound of the total variance. When multiple populations are available which can be considered replicates of the same evolutionary process (e.g., the HIV populations in different human hosts), a bootstrap resampling procedure consisting of sampling from both within a replicate and among replicates will give an estimate of the total variance.

Application to HIV

A number of studies involving within-host longitudinal samples of DNA sequences of human immunodeficiency virus type 1 (HIV-1) have been reported (e.g., Balfe et al. 1990; Simmonds et al. 1991; Wolfs et al. 1991; Holmes et al. 1992; Zhang et al. 1997; Rodrigo et al. 1999). We will use the case analyzed by Rodrigo et al. (1999) for illustration of our method and for the purpose of comparison of estimates. This longitudinal sample was obtained from a homosexual Caucasian male who was diagnosed as HIV-1 seropositive following an episode of aseptic meningitis in 1985 (see Rodrigo et al. [1999] and references therein). The first sample was taken in April 1989, and subsequent samples were taken 7, 22, 23, and 34 months later. The patient started treatment with zidovudine at month 13 after the first sample and continued until the end of the study.

The 0.65-kb region of the *env* gene spanning the third to the fifth variable regions was sequenced for each virus in the sample. We will utilize the same sequence alignment as that in Rodrigo et al. (1999). For simplicity, we will consider only nucleotide substitutions. Three methods for computing the number of nucleotide substitutions between two sequences are considered. One is the number of nucleotide differences, the second is the distance using Jukes-Cantor correction, and the third is the distance corrected using Kimura's two-parameter model. The sample size and time lengths between samples, as well as the Π values by the first and third methods, are given in table 1.

From table 1, the mutation rate per site per day is estimated by v_T to be 4.71×10^{-5} without correction

Table 2
Pairwise Estimates of G Using Kimura's Two-Parameter Model with $\mu = 2.5 \times 10^{-5}$

SAMPLE	SAMPLE			
	1	2	3	4
2	0.789			
3	1.032	0.970		
4	0.357	0.499	5.820	
5	0.317	0.395	0.388	0.269

and 4.45×10^{-5} with correction using Kimura's two parameter model. These values correspond, respectively, to 1.71×10^{-2} and 1.62×10^{-2} per site per year. The estimate using Jukes-Cantor correction lies between these two values.

As we pointed out earlier, one can obtain an estimate of the generation time when the mutation rate per site per generation is known. Mansky (1996) estimated that the overall mutation rate per site per generation was 4×10^{-5} , which includes base substitution, frameshift, deletion, and insertion. Since we only consider base substitutions in our analysis, it is necessary to use the mutation rate for base substitution only. Using Mansky's data (table 4 in Mansky 1996) that there are 15 base substitutions in 5,272 shuttle vector proviruses with a target segment of 114 nt, we obtain a nucleotide substitution rate of $15/(5,272 \times 114) = 2.5 \times 10^{-5}$ per site per generation. The pairwise estimates of the number G of generations per day are given in table 2.

The estimates of generation time L using the original distance, the distance with Juke-Cantor correction, and the distance with correction by the Kimura two-parameters model are given in table 3, together with bootstrap estimates of standard errors. For comparison, estimates using Rodrigo and Felsenstein (1999) (see discussion in the next section) are also included.

It is clear from table 3 that differences in the estimates of L among correction methods are rather minor, but there are considerable differences between estimates based on G_s and G_T . The reason why G_s is nearly twice as large as G_T —and thus G_s^{-1} is nearly half of G_T^{-1} —is that G_s is dominated by a single large estimate of G from samples 3 and 4. Because there were only 28 days separating these two samples, the resulting estimate of G has a large variance. In comparison, G_T gives a smaller weight to the estimate from these two samples, which results in a smaller value and, consequently, a larger estimate of generation time. Table 3 also shows that the bootstrap standard error of G_s^{-1} is slightly larger than that of G_T^{-1} , which further supports our hypothesis that G_T^{-1} is a better estimator than G_s^{-1} . We thus conclude that the generation time for the HIV population in this patient is about 1.78 ± 0.25 days. In comparison, combined estimates from pairwise estimates of G (table 4) by the Rodrigo and Felsenstein (1999) method are less than half of our estimates.

Alternative Methods

There are several existing methods for estimating the generation time. Here, we will focus on a method

Table 3
Estimation of Generation Time L

COMBINED ESTIMATE	OUR METHOD			RODRIGO AND FELSENSTEIN'S (1999) METHOD
	No Correction	Jukes and Cantor	Kimura Two-Parameter	
G_s^{-1}	0.97 (0.30)	0.93 (0.30)	0.92 (0.29)	0.42
G_T^{-1}	1.88 (0.25)	1.79 (0.25)	1.78 (0.25)	0.69

NOTE.—Numbers in parentheses are estimates of standard error based on 3,000 bootstrap samples.

proposed by Rodrigo and Felsenstein (1999), but another entirely different approach will be discussed as well. A brief description of Rodrigo and Felsenstein's (1999) method is as follows.

Consider a sample of m sequences from a haploid population with a constant effective population size N . If one examines the history of these m sequences by tracing backward in time, one will find that there is a period in which there are m ancestral sequences, a period in which there are $m - 1$ ancestral sequences, and so on. The time length t_i of the period in which there are i ancestral sequences has an exponential distribution with a mean equal to $2N/i(i - 1)$ (Kingman 1982a). The number of generations from time $t + T$ back to the first sampling time t at which there are l ancestral sequences is $G_l = t_m + \dots + t_{l+1}$, whose expectation is equal to

$$E(G_l) = \frac{2N(m - l)}{lm}, \tag{17}$$

where $m - l$ is the number of coalescent events among the m sequences during the time period T . For a pair of samples taken T days apart, one can estimate the number of generations between the two sampling times using the above equation if both N and l are known. Similar to our estimator, one can convert G_l to an estimate of the generation time L as T/G_l .

The effective population size N can be estimated from an estimate of $\theta = 2N\mu$, where μ is the mutation rate per generation per site. There are a number of methods available for estimating θ (see review by Li and Fu 1999). For example, since Π_2 has a mean equal to θ under the assumption of a constant effective population size, one can estimate N as $\hat{N} = \Pi_2/(2\mu)$. Rodrigo et al. (1999) used a more complex method and obtained estimates of N similar to those by Brown and Richman (1997). The value of $m - l$ was estimated from a rooted phylogeny of the sequences from both samples as the

Table 4
Pairwise estimates of G by Rodrigo et al. (1999) with $\mu = 2.5 \times 10^{-5}$

SAMPLE	SAMPLE			
	1	2	3	4
2	2.19			
3	2.46	2.16		
4	1.36	1.87	10.0	
5	0.54	0.58	1.39	1.60

number of coalescent events among the sequences in the second samples. Rodrigo et al. (1999) used the neighbor-joining method (Saitou and Nei 1987) for phylogeny reconstruction and used an outgroup sequence to root the tree. Note that another tree-based approach for estimating mutation rate is proposed by Rambaut (2000).

Using the approach described above, Rodrigo et al. (1999) obtained an estimate of the generation time for HIV-1 of 1.2 days. At first glance, their estimate appears to be comparable with our estimate of 1.78 days, but the two estimates cannot be directly compared for two reasons. First, Rodrigo et al. (1999) used a different approach than ours to combine pairwise estimates. Second, we use here a different mutation rate. Although only nucleotide substitutions were considered, Rodrigo et al. (1999) nevertheless used the mutation rate 4×10^{-5} compiled by Mansky (1996), which includes both insertions and deletions. When only nucleotide substitutions are counted, the mutation rate from Mansky's data becomes 2.5×10^{-5} per site per generation. With this mutation rate, pairwise estimates of G computed from table 2 of Rodrigo et al. (1999) become those in table 4. Comparison of table 4 with table 2 reveals that Rodrigo et al.'s (1999) estimate of G is considerably larger than ours in every case, and is on average more than twice as large as ours. Although some of the differences must be due to the variances in both estimates, it is highly unlikely that random errors alone can result in such systematic differences. Some possible causes for this discrepancy will be discussed later.

A very different technique for estimating generation time using within-host longitudinal viral load data has been developed (Coffin 1995; Wei et al. 1995; Perelson et al. 1996). This approach is based on the principle that when a potent drug—such as Ritonavir—which is a protease inhibitor, is administered to a patient, the rate of loss of virions in plasma can be modeled by a set of differential equations with a few parameters, which can be estimated from the longitudinal viral load data. The values of these parameters can then be used to estimate the generation time. The estimates of generation time from this technique vary from about 4 days by Wei et al. (1995) to 2.6 days by Perelson et al. (1996). The latter group have recently revised their estimate to 1.8 days (Rodrigo et al. 1999), which agrees well with our estimate of 1.78 days.

Discussion

Very often, a statistical method for analyzing a population sample is developed under a specific model, such as the constant effective population size assumed by Rodrigo et al. (1999). When the population in question evolves in a manner that is significantly different from the model, the statistical analysis and the resulting conclusions can be misleading. Therefore, it is important to understand how an estimator behaves under various situations. The estimator of mutation rate proposed in this paper has the distinct feature of being unbiased in a variety of situations, which deserves further discussion.

In the case of population growth or, in general, varying effective population size, it is easy to see that \bar{v} is an unbiased estimator of v because equations (5) and (7) hold regardless of the value of effective population size. This property of our estimator is particularly important for its application to fast-changing viral populations such as HIV-1, because a within-host population can change dramatically in size over a short period of time.

The estimator \bar{v} is also unbiased in the presence of population structure, because regardless of population structure, every sequence in the second sample experiences the same amount of time since the time at which the first sample was taken. As long as a consistent sampling strategy is used from different samples, equations (5) and (7) hold regardless of population structure.

It is also obvious that recombination does not introduce bias in our estimate of \bar{v} either, because equations (5) and (7) hold in the presence of recombinations. Of course, this is not to say that nonconstant effective population size, population structure, and recombination have no effect on our estimate, because they do affect the variance of the estimator.

Natural selection is an important factor to consider when analyzing samples from viral populations such as HIV-1. When the DNA region under study is not directly involved in natural selection, our estimator should remain nearly unbiased. This includes the situation in which the region under study is tightly linked to a locus that is under strong natural selection. For example, if natural selection has led to the fixation of a favorable mutation before sampling starts, then its effect is very similar to that of a growing population and thus will not lead to bias in our estimator. When many mutations in the samples are not selectively neutral, the accumulation of nucleotide changes in a sequence in a given period may deviate from Poisson distribution, and the substitution rate can be elevated or reduced depending on the type of natural selection. In the case of deleterious mutations, the mutation rate per year estimated from equation (13) is likely to be smaller than that extrapolated by mutation rate per site per generation and generation time. This will result in an overestimate of the generation time. On the other hand, if positive selection is involved, the substitution rate per site per year will be elevated, which will result in an underestimate of the generation time. One way to minimize the effect of natural selection is to conduct analyses on synonymous substitutions only. With more and more data available, such analyses should be very informative. Since our analysis in this paper is mainly for the purpose of illustration, and also because of the relative small samples, we do not pursue the more detailed analysis.

Since the number of mutations that substantially enhance viral survival should be small compared with the total number of mutations, the bias in our estimate of v due to positive selection is unlikely to be substantial. Nevertheless, since positive selection is likely operating on the *env* gene of HIV-1 (e.g., Bonhoeffer, Holmes, and Nowak 1995; Yamaguchi and Gojobori 1997; Zhang et al. 1997), our estimate of the generation

time may be slightly affected. Another potential source of error in the estimate of generation time is the mutation rate per site per generation. For example, if the mutation rate is underestimated, then it will result in an underestimate of the generation time. With these caveats, it is encouraging that our estimate of generation time agrees well with the recent estimate of 1.8 days from viral load data (see Rodrigo et al. 1999). It will be interesting to see if the agreement continues to hold with increasing data.

What causes Rodrigo et al.'s (1999) estimates of G to be consistently larger than ours? Although the variances in both estimators may lead to fluctuation, the discrepancy is likely due to some fundamental differences between the two estimators. One similarity between the two estimators of generation time is that both rely on estimates of the numbers of generations per day. However, our method is more direct and is unbiased for estimating the number of generations per day, while Rodrigo and Felsenstein's (1999) method is indirect, relying on estimates of both the effective population size and the number of coalescent events among the sequences in sample 2 in the period that separates the two samples. Counting the coalescent events directly from an estimated phylogeny will likely overestimate this number even if the phylogeny is perfectly reconstructed. The simple analysis below will reveal why this is so.

Consider two random samples with sizes n and m , respectively, taken at the same time. Their coalescences will be like that for a single sample of $n + m$ sequences. A coalescent event will be counted as a coalescence between sequences from sample 2, or simply a coalescence within sample 2, if and only if neither of the two coalescing sequences has a descendant in sample 1. Let $p(n, m)$ be the probability that there is no coalescence within sample 2. Each time a coalescence occurs, each pair of sequences has the same probability to be chosen. There are $n(n - 1)/2$ ways to coalesce two sequences from sample 1, and there are nm ways to coalesce one sequence from sample 1 and one sequence from sample 2. Therefore, we have the following recurrence equation:

$$p(n, m) = \frac{n(n - 1)}{(n + m)(n + m - 1)}p(n - 1, m) + \frac{2nm}{(n + m)(n + m - 1)}p(n, m - 1). \quad (18)$$

When there is only one sequence in the second sample, there is no chance of having coalescence within sample 2. Therefore, the initial condition for solving the above recurrence equation is

$$p(k, 1) = 1, \quad k = 1, \dots \quad (19)$$

The probability that there is at least one coalescence within sample 2 is $1 - p(n, m)$, and its numerical values for a number of sample size combinations are given in table 5. It is clear from table 5 that for those sample sizes in our longitudinal samples, this probability is quite high. For example, when $n = 15$ and $m = 13$, the probability is 0.94. This analysis suggests that even when there is no time ($T = 0$) separating the two sam-

Table 5
The Probability ($1 - p(n, m)$) of Having at Least One Coalescence Among the Sequences in the Second Sample when $T = 0$

m	n						
	8	9	13	15	30	50	100
8 ..	0.84	0.81	0.70	0.66	0.44	0.30	0.17
9 ..	0.90	0.87	0.78	0.74	0.52	0.36	0.21
13 ..	0.99	0.98	0.95	0.94	0.78	0.62	0.39
15 ..	1.00	1.00	0.98	0.97	0.87	0.72	0.49
30 ..	1.00	1.00	1.00	1.00	1.00	0.99	0.93

ples, it is very likely to observe coalescent events within the second sample, so that counting these events as the estimate of $m - l$ and then using this estimate as the basis for estimating G will result in an overestimation of G . This is likely one of the reasons why Rodrigo et al.'s (1999) estimates of G (table 4) are consistently larger than our estimates (table 2). Such a discrepancy will also be observed if the effective population size N is underestimated in Rodrigo et al. (1999).

Table 5 also shows that when n is large, $1 - p(n, m)$ becomes small, which suggests that it is possible to reduce bias in Rodrigo and Felsenstein's (1999) method by using a much larger sample size for the first sample. However, when more than two samples are taken, such as the samples analyzed in this paper, it is not easy to persuade an experimenter to, say, halve the previous sample size whenever a new sample is taken, because samples are usually not collected entirely for a single purpose. In general, increasing all of the sample sizes will improve the accuracy of the final estimate, which is also true for Rodrigo and Felsenstein's (1999) method because coalescent time is smaller when there are many sequences; thus, the error due to an incorrect estimate of l is not as severe as that for a small sample.

It is worth emphasizing that the longitudinal samples required for estimating mutation rate and generation time do not have to come from within single host. The estimator is applicable to samples in which each sequence comes from a different host. This feature is valuable for studying the evolution of a pathogen that does not stay in a single host for a long period. On the other hand, if longitudinal samples are available from multiple populations which can be considered replicates of the same evolutionary process, accuracy in the estimation of generation time can be substantially improved because samples from different hosts should be more or less independent. Furthermore, the total variance of the estimator can be obtained by bootstrapping both within-population and among-populations samples. Longitudinal samples from within-host HIV populations are being accumulated, and it will be interesting to see how variable the generation time can be among different hosts.

We have so far considered two ways of utilizing the pairwise number of nucleotide substitutions. Another possible use of equations (5) and (7) is to estimate the time length T separating two samples when both the generation time and the mutation rate per generation are known. An estimate of T is

$$\hat{T} = \frac{\Pi_{12} - \Pi_1}{\mu G}. \quad (20)$$

A potential use of such an estimator is to date the ancestral population from which an ancient DNA sample is obtained. This, of course, requires that the modern sample is taken from a population that shared the same ancestral population from which the ancient DNA sample was derived.

Acknowledgments

I thank Dr. Stanley Sawyer and reviewers for their comments, and Dr. Allen G. Rodrigo for kindly providing me his sequence alignment. This work was supported by NIH grants R29 GM50428 and R01 HG01708 and a fellowship from the Japan Society for the Promotion of Science. Special thanks go to Dr. Naruya Saitou for hosting me in Mishima.

LITERATURE CITED

- BALFE, P., P. SIMMONDS, C. A. LUDLAM, J. O. BISHOP, and A. J. BROWN. 1990. Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. *J. Virol.* **64**:6221–6233.
- BONHOEFFER, S., E. C. HOLMES, and M. A. NOWAK. 1995. Causes of HIV diversity. *Nature* **376**:125.
- BROWN, A. J. L., and D. D. RICHMAN. 1997. HIV-1: gambling on the evolution of drug resistance? *Nat. Med.* **3**:268.
- COFFIN, J. M. 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* **267**:483–489.
- HOLMES, E. C., L. Q. ZHANG, P. SIMMONDS, C. A. LUDLAM, and A. J. BROWN. 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proc. Natl. Acad. Sci. USA* **89**:4835–4839.
- KINGMAN, J. F. C. 1982*a*. On the genealogy of large populations. *J. Appl. Prob.* **19A**:27–43.
- . 1982*b*. The coalescent. *Stochast. Processes Applications* **13**:235–248.
- LI, W. H., and Y. X. FU. 1999. Coalescent theory and its applications in population genetics. Pp. 45–79 in E. HALLORAN and S. GEISSER, eds. *Statistics in genetics*. Springer, New York.
- 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.
- PERELSON, A. S., A. U. NEUMANN, M. MARKOWITZ, J. M. LEONARD, and D. D. HO. 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**:1582–1586.
- RAMBAUT, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenetics. *Bioinformatics* **16**:395–399.
- RODRIGO, A. G., and J. FELSENSTEIN. 1999. Coalescent approaches to HIV population genetics. In K. A. CRANDALL, ed. *The evolution of HIV*. Johns Hopkins University Press, Baltimore, Md.
- RODRIGO, A. G., E. G. SHPAER, E. L. DELWART, A. K. N. IVERSEN, M. V. GALLO, J. BROJATSCH, M. S. HIRSCH, B. D. WALKER, and J. I. MULLINS. 1999. Coalescent estimates of HIV-1 generation time in vivo. *Proc. Natl. Acad. Sci. USA* **96**:2187–2191.
- SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406–425.
- SIMMONDS, P., L. Q. ZHANG, F. MCOMISH, P. BALFE, C. A. LUDLAM, and A. J. BROWN. 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. *J. Virol.* **65**:6266–76.
- TAJIMA, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**:437–460.
- WEI, X., S. K. GHOSH, M. E. TAYLOR et al. (12 co-authors). 1995. Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**:117–122.
- WOLFS, T. F., G. ZWART, M. BAKKER, M. VALK, C. L. KUIKEN, and J. GOUDSMIT. 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.
- YAMAGUCHI, Y., and T. GOJOBORI. 1997. Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. *Proc. Natl. Acad. Sci. USA* **94**:1264–1269.
- ZHANG, L., R. S. DIAZ, D. D. HO, J. W. MOSLEY, M. P. BUSCH, and A. MAYER. 1997. Host-specific driving force in human immunodeficiency virus type 1 evolution in vivo. *J. Virol.* **71**:2555–2561.

NARUYA SAITOU, reviewing editor

Accepted December 11, 2000