Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 2187–2191, March 1999
Evolution

Coalescent estimates of HIV-1 generation time *in vivo*

ALLEN G. RODRIGO*†, EUGENE G. SHPAER‡, ERIC L. DELWART§, ASTRID K. N. IVERSEN¶, MICHAEL V. GALLO‖, JÜRGEN BROJATSCH**, MARTIN S. HIRSCH††, BRUCE D. WALKER††, AND JAMES I. MULLINS*‡‡

*Department of Microbiology, University of Washington, Seattle, WA 98195; ‡Perkin-Elmer, Applied Biosystems, Inc., Foster City, CA 94404; §Aaron Diamond AIDS Research Center, New York, NY 10016; ¶Department of Infectious Diseases, The National University Hospital, Copenhagen, Denmark; ‖Ontogeny, Inc., Cambridge, MA 02138; **Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA 02115; ††Infectious Diseases Unit, Massachusetts General Hospital, Boston, MA 02129; and ‡‡Department of Medicine, University of Washington, Seattle, WA 98195

Edited by M. T. Clegg, University of California, Riverside, CA, and approved January 4, 1999 (received for review October 20, 1998)

ABSTRACT **The generation time of HIV Type 1 (HIV-1) *in vivo* has previously been estimated using a mathematical model of viral dynamics and was found to be on the order of one to two days per generation. Here, we describe a new method based on coalescence theory that allows the estimate of generation times to be derived by using nucleotide sequence data and a reconstructed genealogy of sequences obtained over time. The method is applied to sequences obtained from a long-term nonprogressing individual at five sampling occasions. The estimate of viral generation time using the coalescent method is 1.2 days per generation and is close to that obtained by mathematical modeling (1.8 days per generation), thus strengthening confidence in estimates of a short viral generation time. Apart from the estimation of relevant parameters relating to viral dynamics, coalescent modeling also allows us to simulate the evolutionary behavior of samples of sequences obtained over time.**

The integration of mathematical modeling and experimental approaches has led to a deeper understanding of HIV-1 viral dynamics *in vivo*. In particular, these studies suggest that the viral population in the peripheral blood (and the secondary lymphatics) turns over rapidly, with generation times estimated to be on the order of one to two days on average (ref. 1; A. Perelson, personal communication). However, because these studies are based on abstract and simple models of a complex biological system, it is not obvious how accurate these estimates are. One way to address this is to try to estimate the same parameters by using different methods and different types of data.

Here, we apply a new method developed by Rodrigo and Felsenstein (2) which is based on a mathematical construct introduced by Kingman (3, 4) called the *n*-coalescent (or coalescent for short). The coalescent relies on the fundamental notion that all individuals in a population have a genealogy, so that if we begin with a sample of *n* individuals, each drawn randomly from a population, and we reconstruct the genealogy of these *n* individuals, we will begin to see lineages coalescing as we move further back in time. Each coalescent event represents the split of two lineages from a common ancestor. Obviously, if we go back far enough, we arrive at a point where all lineages have coalesced, and this point represents the most recent common ancestor of all sequences in the sample. The mathematics of coalescent theory gives us the distribution of times, measured as the number of generations, between coalescent events, as one moves back in time along the genealogy. The distribution of times itself is contingent on the dynamics of the population in question. For instance, the expected coalescence time of two individuals drawn at random, each from a compartment of a subdivided population, depends on the rate of migration between the compartments. Similarly, two individuals drawn from a growing population will have a different expected time to coalescence than two individuals drawn from a population at equilibrium (5). In fact, any population process that affects the relatedness of randomly sampled individuals will also affect the distribution of coalescent times in predictable ways, so that it becomes possible to infer the magnitude of these processes by simple models that relate population dynamics to genealogy.

One of the major problems facing HIV molecular evolutionary biologists is sampling: with 10^{10} virions produced daily in an infected individual, and 10^6–10^7 infected cells present, it still remains logistically infeasible to sample any more than tens or hundreds of HIV sequences per individual. How, then, can inferences be made about the evolutionary processes that modulate the genetic variation of viral populations? Fortunately, this is precisely what the coalescent allows us to do, that is, make inferences about population processes based on the genealogy of a small sample of sequences drawn from a much larger population.

Recently, Rodrigo and Felsenstein (2) extended the coalescent framework to handle the genealogy of samples of sequences drawn serially in time. Among other things, they showed that the added information afforded by serial samples allows us to estimate the average length of time of each generation. It is informative to compare estimates of generation time obtained by using the coalescent to that derived by using mathematical models of viral population dynamics. If these estimates are similar, they serve to strengthen our confidence in both the estimates and the methods themselves.

In this study, we use the coalescent method to estimate the expected generation time of the HIV-1 population in an infected individual who remained asymptomatic over the 4 years when samples were taken. Our estimated expected generation time is consistent with values obtained by Perelson *et al.* (1) using a mathematical model of viral dynamics, thus offering independent support for a rapid turnover of the viral population.

Theory. The following is a description of the coalescent and the method proposed by Rodrigo and Felsenstein (2). If a nonrecombining DNA sequence is obtained from each of two haploid individuals drawn at random from a population of constant size *N*, where each member of the population has an equal propensity to reproduce, the probability that the two sequences do not share a common ancestor in the previous generation is

$$\frac{N(N-1)}{N^2} = \left(1 - \frac{1}{N}\right). \quad [1]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

This paper was submitted directly (Track II) to the *Proceedings* office. Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. U00804–U00822, U00831–U00850, and U00873–U00888).

†To whom reprint requests should be addressed. e-mail: rodrigo@u.washington.edu.

If n individuals are sampled, the probability that there are no shared ancestors in the previous generation, $P(t=1)$ or $P(1)$ for convenience, is

$$P(1) = \prod_{i=1}^{n-1}\left(1 - \frac{i}{N}\right). \quad [2]$$

Assuming $n \ll N$, and N is large, and ignoring all terms of order $1/N^2$ or smaller, Eq. **2** is approximated by

$$P(1) \approx 1 - \frac{1}{N} - \frac{2}{N} - \frac{3}{N} - \cdots - \frac{(n-1)}{N} = 1 - \frac{n(n-1)}{2N}. \quad [3]$$

The probability that the first coalescent event occurs in generation t some time in the past is the probability that no coalescent events occur from generation 1 to $t-1$ and that one coalescent event occurs in generation t, that is:

$$P(t) = [1 - P(1)]P(1)^{t-1}. \quad [4]$$

Eq. **4** is the density function of a geometric distribution. For continuous-valued generation times, the exponential distribution can be used as an approximation, so that Eq. **4** can be reexpressed as

$$P(t) \approx \int_{t-1}^{t} \frac{n(n-1)}{2N} \exp\left(-\frac{n(n-1)}{2N}s\right)ds$$

$$\approx \frac{n(n-1)}{2N} \exp\left(-\frac{n(n-1)}{2N}t\right). \quad [5]$$

The expected time for the first coalescent event, i.e., the time taken for n lineages to coalesce to $n-1$ lineages, is

$$E[t_{n\to n-1}] = \frac{2N}{n(n-1)} \text{ generations} \quad [6]$$

with variance

$$V[t_{n\to n-1}] = \frac{4N^2}{n^2(n-1)^2} \text{ generations.} \quad [7]$$

The expected number of generations required to move from n to l lineages can be obtained by summing

$$E[t_{n\to n-1}] + E[t_{n-1\to n-2}] + \cdots + E[t_{l+1\to l}].$$

This is

$$E[t_{n\to l}] = \frac{2N(n-l)}{nl} \text{ generations.} \quad [8]$$

Since $(n-l) = c$ is the number of coalescent events that have already occurred, Eq. **8** can be rewritten

$$E[t_c] = \frac{2Nc}{n(n-c)} \text{ generations.} \quad [8a]$$

Eqs. **1**–**8** are part of the basic theory of the coalescent (3, 4). To obtain the genealogy of n individuals, one typically reconstructs the phylogeny of gene sequences obtained from each of the individuals in the sample, or if two or more samples are available, the joint phylogeny of all sequences.

Generation time, i.e., the average time taken for a viral genome (or infected cell) to produce another of its kind in a replication cycle, can be estimated if the number of generations that has elapsed between two samples collected serially is known. A joint phylogeny of the sequences collected from all serial samples allows us to estimate this. To illustrate, consider Fig. 1, in which the joint phylogeny is shown for two sets of HIV DNA partial envelope (*env*) nucleotide sequences obtained from peripheral blood mononuclear cells in blood drawn from the same individual 214 days apart. Sequences from the later sample (indicated by ▲) are typically more closely related to sequences from the same sample. Eleven coalescent events have occurred between lineages associated with the later sample over the 214-day period, with only 4 lineages remaining when sequences from the earlier sample appear on the joint phylogeny. The number of days per generation, τ, can be estimated as

$$\tau = \frac{d}{E[t_c]} = \frac{dn(n-c)}{2Nc}, \quad [9]$$

where d is the number of chronological units (in this case, days) between samples and N is the effective size of the viral population. For the example above, this gives us an estimate of $584.7/N$.

The method outlined above makes the following assumptions: (*i*) the population from which the samples are drawn remains at equilibrium, i.e., N remains constant, (*ii*) our phylogeny is a good representation of the true genealogy of the sampled individuals, and (*iii*) each individual has the same potential to produce offspring. The second of these assumptions implicitly presupposes that the processes of recombination, migration, and selection do not interfere with our ability to correctly reconstruct the genealogy of the samples.

METHODS

All samples were obtained from a homosexual Caucasian male who was diagnosed as HIV-1 seropositive following an episode of aseptic meningitis in February of 1985, when he was 23 years old. Two previous reports on the clinical course of this patient

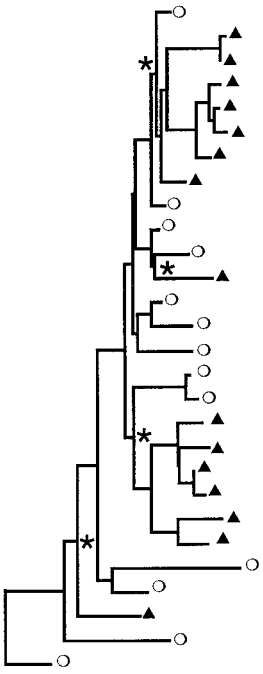

FIG. 1. The genealogy of partial HIV *env* sequences obtained from sample 1 (○) and a later sample 2 (▲). Sequences from the later sample cluster together more than one would expect by chance alone, and 11 of the 14 possible coalescent events occur between lineages that exclusively include sequences from the later time point only. * marks the nodes where sequences from different time points share close common ancestors.

have been published (patient 2 in ref. 6 and patient 1 in ref. 7). Over the course of 3 years beginning in 1989, blood was obtained at time points 7, 22, 23, and 34 months after the first specimen (obtained in April 1989, and reported in ref. 8). He started treatment with zidovudine at month 13 and continued this until after the period reported here (month 34). The $CD4^+$ cell count during the period of study has ranged between 264 and 467 per μl.

Between 8 and 15 HIV DNA sequences of a 0.65-kb region of the *env* gene spanning the third to fifth variable regions were obtained from each of 5 peripheral blood mononuclear cell specimens using standard techniques (8). Nucleotide sequences (GenBank accession nos. U00804–U00822, U00831–U00850, and U00873–U00888) were aligned by using the Multiple Alignment Sequence Editor, MASE (9). The evolutionary distances between pairs of sequences were estimated by using the general maximum-likelihood method (10) with γ-distributed substitution rates across sites (γ parameter $\alpha = 0.5$) implemented in the computer program PAUP* (D. Swofford, Smithsonian Institution). A phylogenetic tree (Fig. 2) was constructed by using the neighbor-joining method (11) with the estimated evolutionary distance matrix. The reference sequence HIVRF (GenBank accession no. M17451) was used to root the neighbor-joining tree.

To validate the assumption that there is no detectable selection acting on the molecular evolution of HIV *env* sequences in this individual, the proportion of nonsynonymous (*dn*) and synonymous substitutions (*ds*) per potential nonsynonymous and synonymous site were calculated for sequences obtained from each time point by using MEGA Ver. 1.02 [Molecular Evolutionary Genetic Analysis (12)]. Only sequences that were nonidentical over the entire length were used in these analyses. A Jukes–Cantor correction for multiple substitutions was applied (13). The standard errors of *ds* and *dn* were calculated by using the method described by Nei and Gojobori (14) and implemented in MEGA Ver. 1.02. Student's *t*-tests for statistical differences between *dn* and *ds* from each sample failed to detect any evidence that the proportion of nonsynonymous substitutions was different from the proportion of synonymous substitutions. We assume, therefore, that even if selection is acting over the region of the envelope sampled, its effects are not likely to be strong enough to distort our phylogeny (below, we discuss in more detail how selection can influence our estimates). We also visually examined the multiple sequence alignment for patterns of recombination and found none.

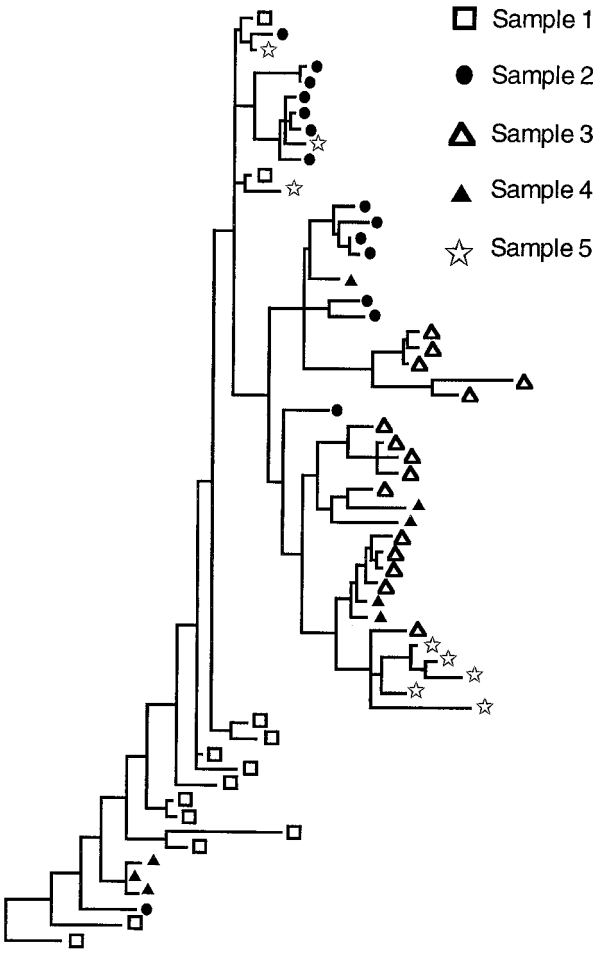

FIG. 2. Joint phylogeny of sequences obtained from all five samples. See text for details on phylogenetic tree reconstruction. The tree was rooted using the reference sequence HIVRF (data not shown).

Generation times were estimated by counting the number of coalescent events that occurred in the genealogy of sequences from a later sample when compared with an earlier sample (for every possible pair of samples). However, because the estimated phylogenetic tree alone does not take into account the uncertainty that is inherent in phylogenetic reconstruction, 100 replicate trees were generated by bootstrap-resampling the nucleotide sequence data (15). For each replicate tree, all sequences not associated with the pair of samples under consideration were pruned from the joint phylogeny. The generation time was estimated as

$$\tau = \frac{1}{100}\sum_{i=1}^{100}\frac{dn(n-c_i)}{2Nc_i}, \qquad \textbf{[10]}$$

which is simply the average value of Eq. **9** over all bootstrap replicates. As is typical with population genetic methods, the effective population size, N, is estimated as part of the composite parameter $\Theta = 2N\mu$, where μ is the mutation rate (note: Θ is a fundamental population genetics parameter and may be loosely characterized as a measure of genetic diversity). Θ was estimated for each sample of sequences by using the coalescent-likelihood estimation program FLUCTUATE (16). The results are shown in Table 1. To obtain N, Θ was divided by 2μ, where $\mu = 4 \times 10^{-5}$ mutations per site per generation (17). The average effective population size was thus estimated as $n = 1{,}260 \pm 136$. This value is of the same order of magnitude as those obtained recently by Leigh Brown (18), and the relatively constant value of N (Table 1) supports our first assumption that the population size remains in equilibrium over the course of the study.

RESULTS AND DISCUSSION

The joint phylogenetic tree of HIV partial *env* sequences obtained from five sampling occasions is shown in Fig. 2. There

Table 1. Summary statistics for each sequence sample set

Sample	Days from first sample	No. of sequences	Average pairwise diversity, %	θ	N
1	0	13	3.6	0.088	1100
2	214	15	3.9	0.106	1325
3	671	15	5.0	0.074	925
4	699	9	4.2	0.144	1800
5	1005	8	4.1	0.092	1150

Average pairwise diversity was estimated by calculating Felsenstein-81 maximum-likelihood evolutionary distances (36) between every possible pair of sequences and averaging these. Effective population size, N, was estimated by dividing θ by 2μ where $\mu = 4 \times 10^{-5}$ mutations per site per generation.

is some temporal structure to the tree, evidenced by the clustering of sequences from single time points into related groups, e.g., the majority of sequences from sample 5 cluster as a single clade, and sequences from sample 3 cluster as two closely related clades. Nonetheless, viruses from a given time point do not necessarily cluster together but may be related to viruses from earlier time points.

In Table 2, we show the average number of lineages (taken over all bootstrap replicates) of the genealogy of a later sample that intermingle with those of an earlier sample, for all possible pairs of samples. The average estimates of generation time are in **bold**, and the estimated number of generations that have elapsed are in parentheses. The overall mean generation time is 1.2 days per generation (median = 1.1 days per generation; interquartile range: 0.7–2 days per generation). By using average values of infected-cell clearance rate (i.e., $\delta = 0.7\ d^{-1}$; ref. 19) and viral clearance rate ($c = 3\ d^{-1}$; ref. 1), one finds that the average viral generation is 1.8 days (A. Perelson, personal communication), which is close to the estimate obtained here. It is encouraging that two independent methods, each using very different data, should arrive at similar estimates of viral generation times on the order of one to two days. As noted earlier, corroboration from different sources strengthens our confidence that the estimate of generation times is correct.

The coalescent method we use here is one of a number of molecular evolutionary techniques that are available (or are becoming available) to phylogeneticists and molecular evolutionary biologists. Kelly (37), for instance, developed a population genetic method for estimating the synonymous mutation rate and the number of generations per unit time of an HIV population *in vivo*. Kelly's method requires that the progenitor sequence be known; the coalescent method developed here does not have this requirement, and may therefore be more widely applicable. Coalescent models also allow us to simulate the genealogy of samples of sequences under a variety of conditions, including changes in effective viral population size (16, 20), compartmentalization (21–23), recombination (24–27), and selection (28–30). These approaches are reviewed in Rodrigo and Felsenstein (2). The extension of these models to incorporate serial samples is a continuing effort (31), and is particularly pertinent for HIV research because it is not unusual to obtain several samples of sequences from HIV-infected individuals over the course of infection.

Under the simplest population model, i.e., one with no change in population size, and without recombination, selection, or migration, the analysis above indicates that the number of coalescent events that occur with sequences from a later sample is a function of the time interval separating the two samples. If this interval, measured in terms of the number of generations that have elapsed, is small, then only a few coalescent events, if any, will have occurred before sequences from the earlier sample are introduced into the joint phylogeny. If, on the other hand, the sampling interval is long, then most of the coalescent events would have occurred, so that sequences from each time point would cluster discretely on the joint phylogeny. It is tempting to attribute differences in phylogenetic patterns, i.e., discrete clusters of sequences from different time points versus intermingling of sequences, to differences in the biology of the virus, the host, or both. Explanations for such patterns typically invoke selection and adaptive evolution (32). Although these hypotheses may be true, it is probably fruitful in the first instance to consider the most parsimonious explanation for such patterns, i.e., that they may arise as a simple consequence of the separation of samples in time.

Table 2. Coalescence data

	Sample			
Sample	**1**	**2**	**3**	**4**
2	5.2 (295)			
	0.73			
3	2.0 (1038)	3.1 (620)		
	0.65	**0.74**		
4	2.7 (596)	3.5 (414)	5.5 (175)	
	1.17	**1.17**	**0.16**	
5	3.7 (337)	4.1 (384)	3.9 (289)	3.7 (306)
	2.97	**2.78**	**1.15**	**1.00**

Average number of lineages that have yet to coalesce, estimated number of generations, and average estimated generation times based on comparisons between all possible pairs of sequence sets over 100 bootstrap replicate trees. Values are the number of lineages of the later sample (designated by the row label) that remain as one moves back in time to when the earlier sample (designated by the column label) was obtained. Values in parentheses indicate the estimated number of generations (calculated by dividing the number of days between each pair of samples and the average generation time for that pair). Values in **bold** are the estimated number of days per generation, calculated using Eq. **10**.

It is instructive to examine how many coalescent events occur in the genealogy of a sample of sequences in the interval of time separating two sampling occasions. If we assume, for instance, that the later sample has 20 sequences, the expected number of generations it takes for half the number of coalescent events in the genealogy of that sample to occur can be estimated by substituting $l = (n - 1)/2$ in Eq. **8** to obtain $0.22N$ generations. If $n = 1{,}000$, this translates into 220 generations, or 264 days, assuming a generation time of 1.2 days per generation. However, on average, the final coalescent event does not occur until approximately $2N$ generations have elapsed (we can obtain this result by letting $l = 1$ and substituting this into Eq. **8**). If there are approximately 300 generations per year, as our results suggest, the expected sampling interval required to see two sets of sequences completely separated on a phylogenetic tree is $2N$; with $n = 1{,}000$, this corresponds to just under 7 years. Therefore, although we may expect to see significant degrees of clustering with sampling intervals as short as 8 to 9 months, there is likely to be intermingling of at least some lineages for up to 6 years and beyond. Once again, explanations for such long-lived lineages include hypotheses of "hidden" compartments of virus that subsequently reseed the peripheral blood and infected-cell latency. These hypotheses should be considered in parallel with simple genealogical models of viral population dynamics described here, possibly using the latter as null hypotheses in appropriately constructed statistical analyses.

At this point, we revisit the issue of selection and how this may affect our estimate of generation time. Incorporating selection into the coalescent framework has been a longstanding problem. Neuhauser and Krone (28, 29) have recently described an approach that allows selection to be incorporated into genealogical models. They show that under very weak selection or very strong selection, the coalescent can still reliably describe the genealogy of a sample of sequences (the latter because fixation occurs so rapidly that after a few generations, all variants in the population are equally fit). Preliminary results by Golding (33) also appear to indicate that genealogies are resilient to relatively strong purifying selection. Although the selective forces on the HIV population *in vivo* are likely to vary from individual to individual and from one stage of disease to another, these results offer some circumstantial evidence that coalescent approaches may be reasonably robust in the face of selection.

In this paper, we have tried to address the issue of whether the sequences we have sampled are under any selective pressure by comparing the proportions of nonsynonymous and synonymous substitutions in *env*. This is, perhaps, a less than satisfactory approach: even if *dn* equals *ds*, it does not necessarily imply the absence of selection, because these values are

composite quantities averaged over all sites and thus fail to take account of selective effects at individual positions. If, as has been suggested by many researchers, there is positive selection for *env* to diversify and evade the host's own immune defenses, then we envisage several possibilities. First, selection may be weakly positive so that the probability of fixation of advantageous variants is not substantially greater than what is expected by genetic drift. If this were true, then the methods we have applied, and indeed the coalescent approach in general, will work, because under such weak selection, *env* evolution would be effectively neutral. However, most researchers would argue in support of a stronger selective effect. Possibly, positive selection results in differential reproductive potential, so that the variance of burst-sizes is greater than expected under a Poisson birth–death process. This would result in an effective population size that is smaller than the census population size (34, 35), and indeed with HIV this appears to be the case (18). However, because our estimate of generation time is a function of the estimated effective population size, it implicitly takes this type of selective effect into account.

Types of selective processes likely to affect our coalescent estimate of generation time include selective sweeps that occur periodically, or strongly positive selection, possibly at only a few sites. Under these conditions, it is conceivable that coalescent-based approaches will not apply, and our estimate of HIV-1 generation time will be correct by coincidence alone. Nonetheless, in the absence of any evidence that this is indeed the case, it is premature to discount our estimate and the applicability of the method.

The issue of selection is part of a broader problem, of course: that of the utility of simple models with assumptions that are almost certainly never satisfied. Indeed, mathematical models of HIV-1 viral dynamics *in vivo* have their own set of simplifying assumptions. Nonetheless, construction of the first, simple model that serves as the framework for later theoretical work is, arguably, the most difficult part of the process leading up to the development of more realistic models. Here, we have attempted to define such a framework, one based on an extension of the coalescent that will allow estimates of population parameters to be made on the basis of serially sampled sequence genealogies. Coalescent models are likely to enhance the inferential repertoire of HIV researchers, because the very process of mutation that makes the virus difficult to control also provides an alternative way to explore the dynamics of the virus in its own microcosm.

The authors thank Dr. Joe Felsenstein, whose involvement in initial discussions on the problem of longitudinal samples and the coalescent was invaluable, Drs. Jerry Learn and Yang Wang for comments on the manuscript, and Amy Berson and Richard Oh for technical support. The comments of two anonymous reviewers contributed significantly to the final version of this manuscript. This research was supported by grants from the U. S. Public Health Service.

1. Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. (1996) *Science* **271,** 1582–1586.
2. Rodrigo, A. G. & Felsenstein, J. (1999) in *Coalescent Approaches to HIV Population Genetics*, ed. Crandall, K. (Johns Hopkins Univ. Press, Baltimore).
3. Kingman, J. F. C. (1982) *J. Appl. Probability* **19A,** 27–43.
4. Kingman, J. F. C. (1982) *Stochastic Processes Appl.* **13,** 235–248.
5. Slatkin, M. & Hudson, R. R. (1991) *Genetics* **129,** 555–562.
6. Ho, D. D., Rota, T. R., Schooley, R. T., Kaplan, J. C., Allan, J. D., Groopman, J. E., Resnick, L., Felsenstein, D., Andrews, C. A. & Hirsch, M. S. (1985) *N. Engl. J. Med.* **313,** 1493–1497.
7. Ho, D. D., Sarngadharan, M. G., Resnick, L., Dimarzoveronese, F., Rota, T. R. & Hirsch, M. S. (1985) *Ann. Intern. Med.* **103,** 880–883.
8. Kusumi, K., Conway, B., Cunningham, S., Berson, A., Evans, C., Iversen, A. K. N., Colvin, D., Gallo, M. V., Coutre, S., Shpaer, E. G., *et al.* (1992) *J. Virol.* **66,** 875–885.
9. Faulkner, D. V. & Jurka, J. (1988) *Trends Biochem. Sci.* **13,** 321–322.
10. Kishino, H. & Hasegawa, M. (1989) *J. Mol. Evol.* **29,** 170–179.
11. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
12. Kumar, S., Tamura, K. & Nei, M. (1994) *Comput. Appl. Biosci.* **10,** 189–191.
13. Jukes, T. H. & Cantor, C. R. (1969) in *Evolution of Protein Molecules*, ed. Munro, H. N. (Academic, New York), pp. 21–132.
14. Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3,** 418–426.
15. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
16. Kuhner, M. K., Yamato, J. & Felsenstein, J. (1998) *Genetics* **149,** 429–434.
17. Mansky, L. M. (1996) *AIDS Res. Hum. Retroviruses* **12,** 307–314.
18. Leigh Brown, A. J. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 1862–1865.
19. Perelson, A. S., Essunger, P., Cao, Y., Vesanen, M., Hurley, A., Saksela, K., Markowitz, M. & Ho, D. D. (1997) *Nature (London)* **387,** 188–191.
20. Griffiths, R. C. & Tavare, S. (1994) *Phil. Trans. R. Soc. Lond. B* **344,** 403–410.
21. Slatkin, M. & Maddison, W. P. (1989) *Genetics* **123,** 603–613.
22. Slatkin, M. & Maddison, W. P. (1990) *Genetics* **126,** 249–260.
23. Beerli, P. & Felsenstein, J. (1999) *Genetics*, in press.
24. Griffiths, R. C. & Marjoram, P. (1996) *J. Computat. Biol.* **3,** 479–502.
25. Hudson, R. R. & Kaplan, N. L. (1985) *Genetics* **111,** 147–164.
26. Hudson, R. R. (1987) *Genet. Res.* **50,** 245–250.
27. Hudson, R. R. & Kaplan, N. L. (1988) *Genetics* **120,** 831–840.
28. Krone, S. M. & Neuhauser, C. (1997) *Theoret. Popul. Biol.* **51**.
29. Neuhauser, C. & Krone, S. M. (1997) *Genetics* **145,** 519–534.
30. Kaplan, N. L., Darden, T. & Hudson, R. R. (1988) *Genetics* **120,** 819–829.
31. Felsenstein, J., Kuhner, M. K., Yamamoto, J. & Beerli, P. (1999) *IMS Lecture Series* **33,** in press.
32. Wolinsky, S. M., Korber, B. T. M., Neumann, A. U., Daniels, M., Kuntsman, K. J., Whetsell, A. J., Furtado, M. R., Cao, Y., Ho, D. D., Safrit, J. T. & Koup, R. A. (1996) *Science* **272,** 537–542.
33. Golding, B. (1997) in *Progress in Population Genetics and Human Evolution*, eds. Donnelly, P. & Tavaré, S. (Springer, New York), Vol. 87.
34. Wright, S. (1938) *Science* **87,** 430–431.
35. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
36. Felsenstein, J. (1981) *J. Mol. Evol.* **17,** 368–376.
37. Kelly, J. K. (1994) *Genet. Res.* **64,** 1–9.