Reconstructing Genealogies of Serial Samples Under the Assumption of a Molecular Clock Using Serial-Sample UPGMA

Alexei Drummond and Allen G. Rodrigo

School of Biological Sciences, University of Auckland, Auckland, New Zealand

Reconstruction of evolutionary relationships from noncontemporaneous molecular samples provides a new challenge for phylogenetic reconstruction methods. With recent biotechnological advances there has been an increase in molecular sequencing throughput, and the potential to obtain serial samples of sequences from populations, including rapidly evolving pathogens, is fast being realized. A new method called the serial-sample unweighted pair grouping method with arithmetic means (sUPGMA) is presented that reconstructs a genealogy or phylogeny of sequences sampled serially in time using a matrix of pairwise distances. The resulting tree depicts the terminal lineages of each sample ending at a different level consistent with the sample's temporal order. Since sUPGMA is a variant of UPGMA, it will perform best when sequences have evolved at a constant rate (i.e., according to a molecular clock). On simulated data, this new method performs better than standard cluster analysis under a variety of longitudinal sampling strategies. Serial-sample UPGMA is particularly useful for analysis of longitudinal samples of viruses and bacteria, as well as ancient DNA samples, with the minimal requirement that samples of sequences be ordered in time.

Introduction

It is well known that some of the more pernicious human viral pathogens evolve rapidly. Indeed, it is their evolution that stymies attempts to battle infection with antiviral drugs—resistance evolves too quickly. With HIV-1, for instance, 10^{-5}–10^{-4} substitutions accumulate at each site in each generation, and there are an estimated 140–300 generations per year (Perelson et al. 1996; Rodrigo et al. 1999). Parts of the HIV genome have been shown to accumulate substitutions at a rate of 0.92% per year (Shankarappa et al. 1999). There is some thought in the research community that understanding how these viruses evolve is the key to understanding how one may control disease. Recent results give us cause to think that this may be true: a study by Shankarappa et al. (1999) found that in nine individuals infected with HIV, the pattern of viral evolution within each patient was strikingly similar, with certain features that appeared predictive of progression to AIDS. If such commonality of pattern is universal, then generalizations can be made about the process of evolution that such patterns suggest, and this, in turn, may lead to a strategy to control progression.

The study by Shankarappa et al. (1999) involved repeated sampling of the viral population from each individual over several years, but such sampling schemes are not uncommon for such rapidly evolving pathogens (Holmes et al. 1992; Wolinsky et al. 1996; Rodrigo et al. 1999). A starting point for many evolutionary and population genetic methods is a reconstructed phylogeny of sampled sequences (Felsenstein 1992; Fu 1994; Nee et al. 1995; Pybus, Rambaut, and Harvey 2000), often under the assumption of a molecular clock, but until now, there has been no method for reconstructing evolutionary trees of serially sampled sequences under this assumption. In this paper, we present such a method. The serial-sample unweighted pair grouping method with arithmetic means (sUPGMA) is a fast, flexible phylogenetic reconstruction method that can be used whenever samples have been obtained at different times. These samples may be of sequences from a rapidly evolving viral population obtained from within a patient over the course of infection or from cohorts of individuals sampled over time. We demonstrate the efficiency of sUPGMA at recovering the true topology and describe accessory analyses that allow the estimation of population parameters and mutation rate. Finally, we discuss various extensions of sUPGMA and its associated analyses.

Serial-Sample UPGMA

Consider the following sampling scheme. A population is sampled several times over the course of a study period, and at each sampling time a number of sequences are obtained. If these sequences have evolved so that all lineages accumulate substitutions at the same rate over the same period of time (i.e., according to a molecular clock), then the best representation or model of their phylogeny will look something like that shown in figure 1*E*. Here, six sequences were sampled, two at each of three time points. One would expect, if clocklike evolution were occurring, that sequences from the same time point would terminate at identical times. One method for reconstructing phylogenies of sequences according to a molecular clock is the unweighted paired group method with arithmetic means (UPGMA; see Sneath and Sokal 1973). However, with UPGMA, all tips on the tree terminate at the same time (i.e., the tree is ultrametric). What is required to reconstruct the phylogeny shown in figure 1*E* is a method that allows the tips to terminate at different times but constrains tips sampled at the same time to terminate at identical distances from the root. Serial-sample UPGMA allows for this. The method consists of four sequential steps.

Key words: longitudinal samples, serial samples, ancient DNA, phylogenetic reconstruction, UPGMA, sUPGMA.

Address for correspondence and reprints: Allen G. Rodrigo, School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand. E-mail: a.rodrigo@auckland.ac.nz.

Mol. Biol. Evol. 17(12):1807–1815. 2000

A

	A1	A2	B1	B2	C1	C2
A1		*a*	*b*	*c*	*d*	*e*
A2	0.0271		*f*	*g*	*h*	*i*
B1	0.0317	0.005		*j*	*k*	*l*
B2	0.0547	0.0153	0.0582		*m*	*n*
C1	0.0693	0.0875	0.0089	0.0736		*o*
C2	0.0327	0.0584	0.0383	0.0352	0.0512	

B

	$d(m_i, n_j)$	Θ	$\delta_{A\to B}$ (δ_1)	$\delta_{B\to C}$ (δ_2)
d(A1, A2)	0.0271	1	0	0
d(B1, B2)	0.0582	1	0	0
d(C1, C2)	0.0512	1	0	0
d(A1, B1)	0.0317	1	1	0
d(A1, B2)	0.0547	1	1	0
d(A2, B1)	0.005	1	1	0
d(A2, B2)	0.0153	1	1	0
d(A1, C1)	0.0693	1	1	1
d(A1, C2)	0.0327	1	1	1
d(A2, C1)	0.0875	1	1	1
d(A2, C2)	0.0584	1	1	1
d(B1, C1)	0.0089	1	0	1
d(B1, C2)	0.0383	1	0	1
d(B2, C1)	0.0736	1	0	1
d(B2, C2)	0.0352	1	0	1

(estimated values: Θ=0.0326, δ_1= 0.00368, δ_2=0.016)

C

	A1	A2	B1	B2	C1	C2
A1		*a*	$b+\delta_1$	$c+\delta_1$	$d+\delta_1+\delta_2$	$e+\delta_1+\delta_2$
A2	0.0271		$f+\delta_1$	$g+\delta_1$	$h+\delta_1+\delta_2$	$i+\delta_1+\delta_2$
B1	0.0354	0.0087		$j+2\delta_1$	$k+2\delta_1+\delta_2$	$l+2\delta_1+\delta_2$
B2	0.0584	0.0190	0.0656		$m+2\delta_1+\delta_2$	$n+2\delta_1+\delta_2$
C1	0.0890	0.1072	0.0323	0.0970		$o+2\delta_1+2\delta_2$
C2	0.0524	0.0781	0.0617	0.0586	0.0906	

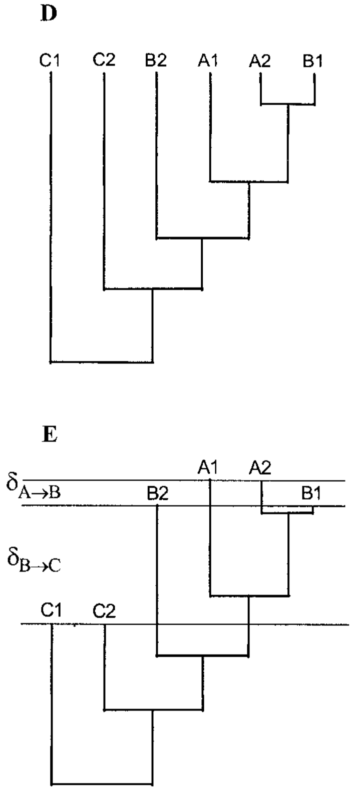

FIG. 1.—The main steps involved in the sUPGMA procedure. *A,* First, a distance matrix of the sequences sampled must be collected. *B,* A matrix is constructed that relates each observed distance to the parameters to be estimated. Each row in *B* corresponds to an instance of equation (2), and the binary values in the columns correspond to the *X*'s in equation (3). For convenience, only a single Θ is estimated in this example. Once this matrix is constructed, the least-squares solution (eq. 4) can be used to estimate the parameters. *C,* The estimated values of δ are then used to correct the original distance matrix (eq. 6). *D,* A standard UPGMA tree is constructed from these corrected distances. *E,* The branches in the UPGMA tree are then trimmed using the estimated δ's to produce the serially sampled genealogy.

Step 1. Estimation of δ's

Simply, step one involves estimating the expected number of substitutions per site accumulating between sampling times. It has been shown how this may be done for pairs of samples (Y.-X. Fu, personal communication). The expected distance between a pair of sequences, one from a later time point and the other from an earlier time point, is

$$E[\mathrm{dist}(S_{\mathrm{early}}, S_{\mathrm{late}})] = E[\mathrm{dist}(S^{(1)}_{\mathrm{early}}, S^{(2)}_{\mathrm{early}})] + \delta_{\mathrm{early}\to\mathrm{late}}. \qquad (1)$$

The first term on the right-hand side is simply the expected average distance between any two sequences from the earlier time point. To obtain an estimate of δ, we substitute the average pairwise distance between early and late sequences calculated from our sample for the term on the left and the average pairwise distance between pairs of early sequences for the first term on the right and solve. The problem becomes tricky when there are more than two time points, because now it becomes possible to calculate δ's for every possible pair of sampling times. The problem with this approach is that it may happen that, for any three time points *A, B,* and *C* (where *C* is earlier than *B,* which is earlier than *A*),

$\hat{\delta}_{CA} \neq \hat{\delta}_{CB} + \hat{\delta}_{BA}$ (where $\hat{\delta}$ is the estimated value), when, in fact, under any reasonable model, the equivalent equality must be true. To overcome this problem, we adopted a general regression approach to estimate δ, as follows. Consider a data set of p samples, with sample i obtained more recently than sample $i + 1$ ($i \in 1, \ldots, p$). Let $d(m_i, n_j)$ be the evolutionary distance between the ith sequence of the mth sample and the jth sequence of the nth sample; by convention, we will assume that $m \geq n$; i.e., we will only consider elements in the diagonal and lower triangular matrix of pairwise distances.

We can model each $d(m_i, n_j)$ by its expectation $E[d(m_i, n_j)]$, and from equation (1), we obtain

$$E[d(m_i, n_j)] = E[d(m_i^{(1)}, m_i^{(2)})] + \delta_{m\to n}. \qquad (2)$$

For reasons that will become obvious below, we will designate $E[d(m_i^{(1)}, m_i^{(2)})] = \Theta_m$. In addition, $\delta_{m\to n}$ can be written as the sum of $\delta_{m\to m-1}, \delta_{m-1\to m-2}, \ldots, \delta_{n+1\to n}$. Thus, we can write the linear equation relating the distances to the parameters as

$$\begin{aligned} d(m_i, n_j) = \sum_{k=1}^{p} \Theta_k X_k &+ X_{(2\to1)m,j}\delta_{2\to1} \\ &+ X_{(3\to2)m,j}\delta_{3\to2} + \cdots + X_{(p\to p-1)m,j}\delta_{p\to p-1} \\ &+ \epsilon_{m_i,n_j}, \end{aligned} \qquad (3a)$$

where $\delta_{k\to k-1}$ is the expected number of substitutions that have accumulated between the kth and the $(k - 1)$th samples,

$$X_k = \begin{cases} 1 & \text{if } k = m, \\ 0 & \text{otherwise} \end{cases} \qquad (3b)$$

$$X_{(k\to k-1)m,n} = \begin{cases} 1 & \text{if } m \geq k \quad \text{and} \quad n \leq k - 1, \\ 0 & \text{otherwise,} \end{cases} \qquad (3c)$$

and ϵ_{m_i,n_j} is the error due to natural variation, measurement, and sampling.

The vector of estimated parameters $\hat{\mathbf{a}} = \{\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_p, \hat{\delta}_{2\to1}, \ldots, \hat{\delta}_{p-1\to p}\}$ is obtained by the standard least-squares solution:

$$\hat{\mathbf{a}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{d}, \qquad (4)$$

where $\mathbf{d}$ is a vector of pairwise distances. With this approach, the estimate of the δ's satisfies the condition $\hat{\delta}_{CA} = \hat{\delta}_{CB} + \hat{\delta}_{BA}$. One additional constraint that we make to the δ's is to set any value of δ that has been estimated as a negative value to 0.

For the estimation approach above, it is not essential to know the actual sampling times, only the order in which the samples were drawn. If the actual sampling times are known, then an alternative approach to estimating δ is to estimate a single constant, ω, effectively the number of substitutions per unit time, and multiply this by the time interval between two sampling occasions, i.e., $\omega(t_1 - t_2)$.

Once again, we estimate ω using a regression procedure. In this case,

$$d(m_i, n_j) = \sum_{k=1}^{p} \Theta_k X_k + \omega(t_m - t_n) + \epsilon_{m_i,n_j}, \qquad (5)$$

where t_k is the time at which the kth sample was obtained. Note that ω is not the mutation rate per generation unless time is expressed in generation units. However, ω can be converted to the mutation rate (i.e., the number of substitutions per site per generation) if the generation time is known.

Step 2. Correction of Pairwise Distances

Each pairwise distance d_{ij} in the distance matrix is now transformed to a corrected distance, $c(m_i, n_j)$, as follows:

$$c(m_i, n_j) = d(m_i, n_j) + \hat{\delta}_{m\to1} + \hat{\delta}_{n\to1}, \qquad (6)$$

where $\hat{\delta}_{m\to1}$ and $\hat{\delta}_{n\to1}$ are the δ's associated with the divergence between samples m and n and the most recent sampling occasion (labeled "1"). What this does, in effect, is extend the distances of sequences sampled earlier to a value that approximates the expected divergences of sequences obtained most recently.

Step 3. Cluster Using UPGMA

In step 3, UPGMA or WPGMA (weighted PGMA; Sneath and Sokal 1973) is applied to the corrected distance matrix.

Step 4. Trim Back branches

Once the UPGMA tree has been constructed, for any terminal lineages which extend to sequences in sample m, $\hat{\delta}_{t(i)\to0}$ is subtracted from the branch length. The sUPGMA tree has the topology recovered by UPGMA (on corrected distances), with tips terminating in the appropriate order of sampling.

Estimation of Population Parameters and Mutation Rate

As described in step 1 above, a vector of parameters is estimated as part of the tree-building algorithm. This vector takes the form $\hat{\mathbf{a}} = \{\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_p, \hat{\delta}_{2\to1}, \ldots, \hat{\delta}_{p-1\to p}\}$ when the order of samples is known and $\hat{\mathbf{a}} = \{\hat{\Theta}_1, \ldots, \hat{\Theta}_p, \hat{\omega}\}$ when exact times are known. Of course, within this framework, there is no need to specify a model with different values of Θ; instead, we could estimate a single parameter, Θ_0, such that $\hat{\mathbf{a}} = \{\hat{\Theta}_0, \hat{\omega}\}$. In this case, the average pairwise diversity at each time point is effectively a random variable with expectation Θ_0. Setting Θ_0 as a constant is equivalent to assuming a population model with constant effective size: under such a model, $\Theta_0 = 2N_e\mu$, where N_e is the effective population size and μ is the mutation rate per site per generation (Tajima 1983).

Although the interpretation of a single Θ_0 is easily accommodated within a simple constant-sized population model, this is not the case when multiple Θ's are estimated. Multiple Θ's should not be taken as (independent) estimates of different $2N_e\mu$ values, because the

Table 1
Sampling Strategies Under Which Phylogenetic Reconstruction Was Tested

Total Sequences	Sampling Strategies[a]						
20	2 × 10	4 × 5	5 × 4	10 × 2			
40	2 × 20	4 × 10	5 × 8	8 × 5	10 × 4	20 × 2	
80	2 × 40	4 × 20	5 × 16	8 × 10	10 × 8	16 × 5	20 × 4

[a] Sampling strategies are represented by the number of time points multiplied by the number of sequences per time point.

overlap in genealogies from one sample to the next affects the pairwise distances of the sequences in a complex way. The simple assignment of different Θ's in our model does not incorporate these complexities.

However we choose to define our model, the variance of the estimates cannot be easily calculated analytically. However, for a constant-sized population model at least, a parametric bootstrap method for obtaining the variance of these estimates can be implemented. For a given set of parameter estimates, a large number (typically >1,000) of serially sampled genealogies can be simulated using the estimated parameters (and assuming a constant population size) to generate pseudoreplicate data sets. For each generated pseudoreplicate, the sUPGMA procedure is then repeated, resulting in a range of estimates for Θ's and δ's or ω. For a 95% confidence interval and 1,000 replicates, the 26th and 975th estimates (when ranked) are taken as the upper and lower 95% confidence limits of the original estimate.

Efficiency of Tree Reconstruction

To test the efficiency of sUPGMA, simulated data sets were created for which the real phylogenetic tree was known. Rodrigo and Felsenstein (1999) described how Kingman's (1982*a*, 1982*b*) *n*-coalescent, essentially a diffusion approximation of the times of $n - 1$ coalescent events on an *n*-taxon tree, could be extended to coalescent trees with noncontemporaneous tips. One of the novel properties of coalescent trees of serial samples is that sampling a direct descendant of a sequence sampled at an earlier time point becomes possible (although unlikely when N_e is very large). The probability of a single lineage from a later time point having a direct ancestor in an earlier sample is equal to the fraction of the total population size sampled at the earlier time ($n_{t(\mathrm{earlier})}/N_e$) (Epperson 1999). This possibility was also permitted in the simulations performed, representing an extension of the original description of the serial-sample coalescent of Rodrigo and Felsenstein (1999). It should be noted that this inclusion results in the possibility of multiple coalescent events occurring at the same time point when more than one direct ancestor is sampled at one time point. However, this happens at an appreciable rate only when the assumption of a very large population size is broken (i.e., when $n^2 \geq N_e$). At this point, the diffusion approximation of coalescent intervals itself is no longer valid. To avoid this problem, values of N_e were selected so that n^2 was always smaller than N_e. Therefore, the simulations were performed under the assumption of a constant population size, and N_e was set to 10,000, which is large enough to fulfill the requirement that $n^2 \ll N_e$. The mutation rate was set to 5×10^{-6} mutations per site per generation. This resulted in an overall Θ value of 0.1 (for a haploid population), comparable to published values for HIV evolution (Leigh Brown 1997; Rodrigo et al. 1999). The model of evolution used in the simulations was a simple Jukes-Cantor substitution model (Jukes and Cantor 1969). The simulated genealogies were drawn from populations with no selection, recombination, or subdivision.

The serial-sample coalescent algorithm was implemented in a small Java program for the purpose of generating coalescent trees under a variety of different sampling strategies (table 1). This allowed an appraisal of the effect of different sampling strategies on the accuracy of tree-building algorithms. For each sampling strategy tested, a range of 0.5% to 10% intersample divergence was tested, with an increment of 0.5%. For each sampling strategy and each divergence, 1,000 simulated genealogies were constructed. All simulations resulted in time-ordered DNA sequences of 1,000 nt in length. This length was comparable with lengths of many gene loci available for phylogenetic study and is not so long that assuming no recombination is untenable. For each simulation, a pairwise Jukes-Cantor distance matrix was constructed. The ability of sUPGMA and UPGMA to correctly reconstruct the simulated genealogies using the pairwise distances was evaluated. The reconstructed trees of each method were compared with the real tree using the symmetric difference index (SDI) tree comparison metric (Robinson and Foulds 1981). This metric counts the number of clades in each tree that are not present in the other tree.

Figure 2 shows the performance of sUPGMA and UPGMA on serially sampled data sets with four serial samples. Essentially the same pattern was seen for all sampling strategies. The performance of sUPGMA generally increases with divergence, while the performance of UPGMA generally decreases. The graphs in figure 2 indicate that once some low threshold of intersample divergence is exceeded, sUPGMA reconstructs the genealogy more accurately than UPGMA. Table 2 shows the approximate threshold values for a variety of sampling strategies. Each threshold value was found by picking the lowest divergence for which sUPGMA performed better on average than UPGMA. In general, our simulations indicated that the divergence threshold decreases with an increase in the size of each sample. Therefore, collecting more sequences within each time

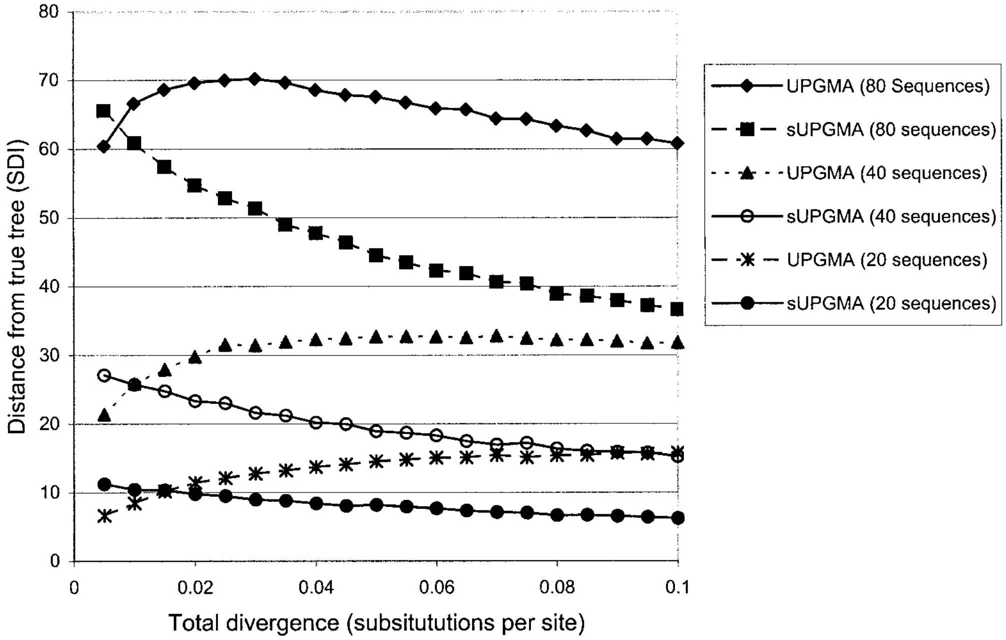

FIG. 2.—Phylogenetic reconstruction performances of sUPGMA and UPGMA on four samples.

point improves the ability of the least-squares procedure to detect small divergences.

Efficiency of Parameter Estimation

The efficiency of parameter estimation of Θ, ω, and δ's was measured by simulating two sets of 1,000 serially sampled genealogies, one parameterized in accordance with equation (3a), and the other in accordance with equation (5). One thousand genealogies of four samples, each with five sequences, were simulated under the Jukes-Cantor model of substitution, resulting in time-ordered sequences of 1,000 nucleotides. Figure 3 shows the distribution of estimates of Θ (true value = 0.1) for the 1,000 simulations with a total divergence over the four samples of 6%. The mean estimate of Θ was 0.0986, with a skewness statistic of 1.753, showing that the least-squares procedure produces estimates of Θ that are unbiased but have a positively skewed distribution (tables 3 and 4). The least-squares estimators of δ_1, δ_2, δ_3, and ω are also unbiased, although once again, the distributions of the estimates are skewed. Figures 4–7 show frequency distributions for estimates of δ_1, δ_2, δ_3, and ω, respectively.

Table 2
Threshold Values for Total Divergence (expected substitutions per site) Over Which sUPGMA Outperforms UPGMA

Total Sequences	No. of Sampling Occasions[a]						
	2	4	5	8	10	16	20
20	0.01	0.02	0.02	—	0.035	—	—
40	0.005	0.01	0.015	0.02	0.025	—	0.05
80	0.005	0.01	0.01	0.015	0.02	0.035	0.04

[a] The number of sequences at each time point is equal and can be obtained from this table by dividing the number of time points by the total number of sequences.

An Example Data Set

In this section, we illustrate the use of sUPGMA with a data set of serially sampled partial envelope (*env*) gene sequences of cell-associated HIV DNA obtained from a long-term asymptomatic individual over five sampling occasions. These samples and the patient history have previously been described (Rodrigo et al. 1999). In total, there were 60 sequences in this data set. Pairwise distances were constructed using a general time-reversible model allowing for unequal nucleotide frequencies and relative rates of substitutions. Substitution and frequency parameters of the substitution model were estimated with PAUP*, version 4.0b4 (D. Swof-

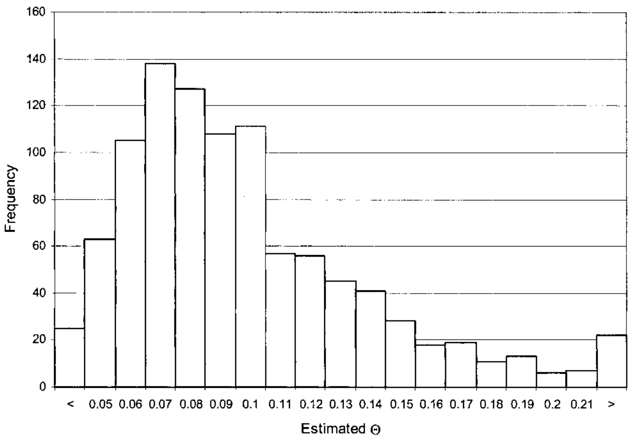

FIG. 3.—Θ estimates for four samples of five sequences for 1,000 simulated trees (real value of Θ = 0.1, total divergence = 0.06 expected substitutions).

Table 3
Statistics of Estimated Θ and δ's for 1,000 Simulated Data Sets of Four Samples of Five Sequences

	Θ	δ_1	δ_2	δ_3
True value	0.1	0.02	0.02	0.02
Mean	0.0986	0.0203	0.0189	0.0205
Standard Deviation	0.04232	0.0229	0.0269	0.0477
Skewness	1.753	1.807	2.356	2.452
97.5th percentile	0.207189	0.079969	0.094122	0.141634
2.5th percentile	0.045227	−0.01973	−0.01408	−0.03477

ford, Smithsonian Institution). sUPGMA was applied to the pairwise distance matrix to reconstruct the serial genealogy of the sequences, allowing different values of Θ and δ's. We also reconstructed the genealogy by assuming a constant Θ and mutation rate, ω, and used parametric bootstrapping of 1,000 simulated trees to obtain 95% confidence intervals for these parameters. The reconstructed trees are shown in figure 8, and the associated parameter estimates are given in table 5.

It is instructive to consider some of the main points of these results. When Θ and δ are allowed to vary, sUPGMA is unable to distinguish between samples 2 and 3; i.e., for this interval, $\delta = 0$. In fact, these two samples were obtained only 1 month apart, so this result is reasonable. When Θ is held constant and ω is estimated, the values obtained are $\Theta = 0.0446$ (95% confidence interval [0.0184, 0.1016]) and $\omega = 7.8 \times 10^{-6}$ substitutions per site per day (95% confidence interval $[-3.47 \times 10^{-6}, 3.87 \times 10^{-5}]$). This estimate of ω translates into an annual substitution rate of 0.3%. This is certainly lower than other HIV-1 *env* gene substitution rate estimates that have previously been obtained, which are on the order of 1% per year (Shankarappa et al. 1999). It is not clear why our estimate of substitution rate is three times as low as other estimates. It is pertinent to note that with the patient from whom the samples were obtained, antiretroviral therapy was initiated at an early stage of the study, and this, in turn, may have lengthened the average generation time of infected cells (see below) and consequently lowered the substitution rate. When a varying substitution rate was allowed, the average rate obtained over the entire 1,005 days of the study was 1.53×10^{-5} substitutions per site per day (0.6% per year), which is closer to previously obtained results. However, this mean rate is still deflated by the very slow substitution rate observed in the last 306 days of the study (see table 5).

Table 4
Statistics of estimated Θ and ω for 1,000 Simulated Data Sets of Four Samples of Five Sequences

	Θ	ω
True value	0.1	5×10^{-6}
Mean	0.09996	4.95×10^{-6}
Standard deviation	0.0454	3.88×10^{-6}
Skewness	1.797884	2.186
97.5th percentile	0.2224	1.56×10^{-5}
2.5th percentile	0.0440	2.71×10^{-7}

Interestingly, the 95% confidence interval of our estimate of mutation rate encloses 0. While this can mean that there is no evidence that there has been a detectable substitution accumulation over time, it can also mean that there were some sequences obtained at a later time point that appear more closely related to those from an earlier time point. In fact, in the original tree published by Rodrigo et al. (1999), this appears to be the case.

Discussion

Serial-sample UPGMA is a variant of UPGMA which constructs genealogies of samples of sequences obtained at different times under the assumption of a molecular clock. Serial-sample UPGMA is a two-step procedure. The first step involves estimating the expected sequence divergence between samples obtained at different times. The second step requires the construction of a corrected distance matrix adjusted to take account of these expected divergences, and subsequent clustering using UPGMA. Given a more accurate estimation procedure for the divergences, the accuracy of sUPGMA tree reconstruction can be improved. For example, given a perfect estimate of divergences, the sUPGMA procedure will perform better than UPGMA under all sampling strategies and divergences (simulations not shown). Therefore, the threshold divergences required for sUPGMA to outperform UPGMA will be reduced by the use of better estimators of δ's and/or ω.

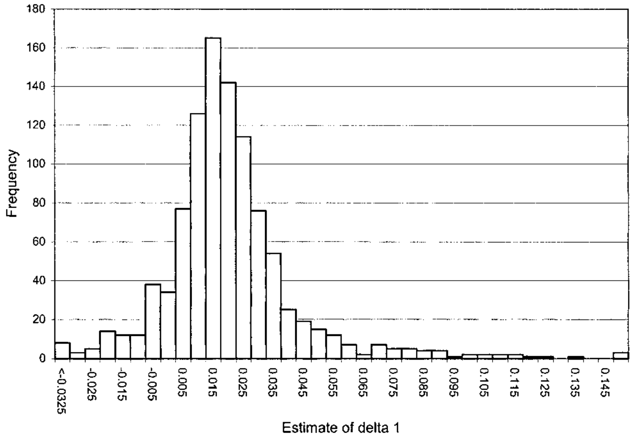

FIG. 4.—δ_1 estimates for four samples of five sequences for 1,000 simulated trees (real value of $\delta_1 = 0.02$).

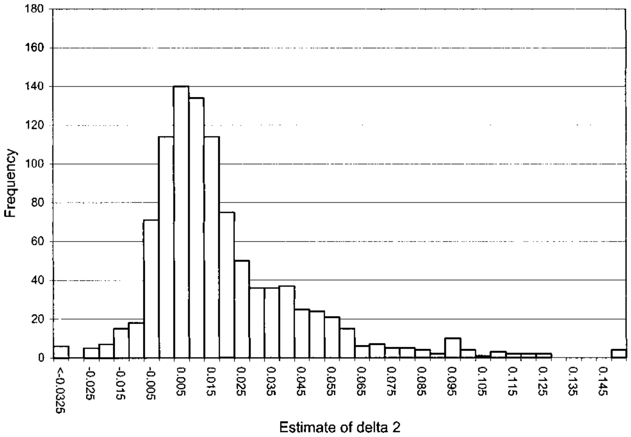

FIG. 5.—δ_2 estimates for four samples of five sequences for 1,000 simulated trees (real value of $\delta_2 = 0.02$).

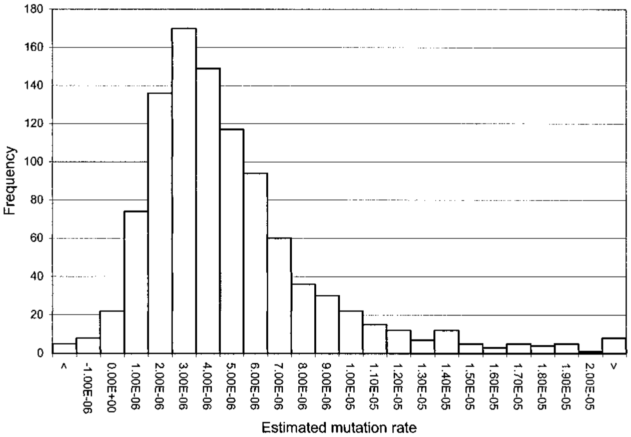

FIG. 7.—Estimated mutation rate, ω, from 1,000 simulations of four samples of five sequences (real value = 5 × 10^{-6} substitutions per site per generation).

When a molecular clock does not apply, UPGMA is known to perform poorly as a tree reconstruction method. However, in the case of clocklike data that have experienced large amounts of evolution, the accuracy of UPGMA in reconstructing clocklike genealogies has been favorably compared with methods such as maximum-likelihood phylogenetic reconstruction (Pybus, Rambaut, and Harvey 2000). Our results demonstrate that the accuracy of UPGMA for phylogenetic reconstruction can be improved by modifying the distances between longitudinally sampled sequences to correct for the extra divergence expected between earlier time points and the most recent time point. The rationale behind using sUPGMA as a basis for a tree reconstruction procedure for serial samples is to provide an (1) accurate and (2) rapid estimation of a serially sampled genealogy. Both the criteria for large divergences and clocklike evolution are fulfilled in at least some virus populations (Gojobori, Moriyama, and Kimura 1990; Leitner and Albert 1999; Shankarappa et al. 1999). However, perhaps most importantly, the speed of sUPGMA allows very large data sets (with hundreds or thousands of sequences) to be analyzed with relative ease. This is an important feature when taking into account the sizes of genealogies already under consideration (e.g., Shankarappa et al. 1999). The distance-corrected matrix that is constructed as part of sUPGMA can also be used with other members of the family of hierarchical algorithmic clustering methods such as WPGMA, complete-linkage and single-linkage clustering.

As part of our parameter estimation procedures, we also introduce two parameterizations of expected intersample sequence divergence. In one case—ω parameterization—divergence is expressed as a product of the sampling interval and mutation rate (with the latter scaled to the same units of time as the sampling interval). A second parameterization that we use, δ parameterization, is less constrained. With δ parameterization, the *i*th interval between two sampling occasions

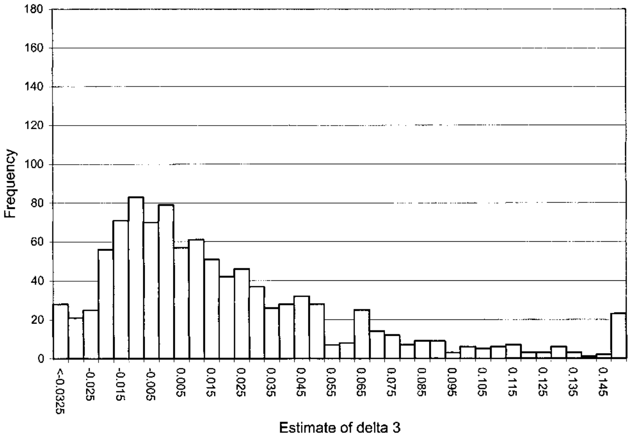

FIG. 6.—δ_3 estimates of four samples of five sequences for 1,000 simulated trees (real value of $\delta_3 = 0.02$).

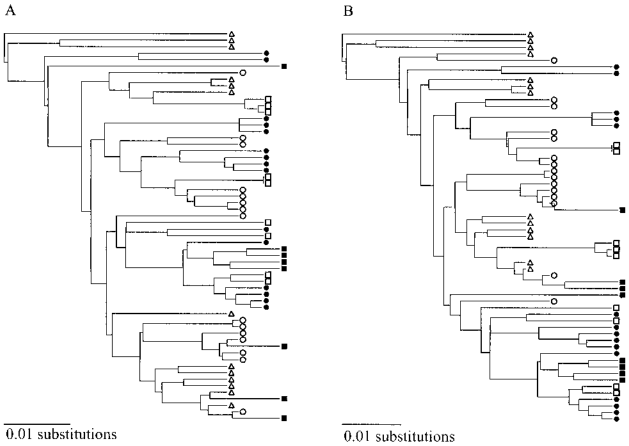

FIG. 8.—Two sUPGMA trees constructed from an example data set. Tree A was constructed under the assumption of a constant population size and a constant mutation rate. Tree B was constructed allowing a different population size at each sampling point and allowing the varying-rate model, in which each time interval has a different mutation rate.

Table 5
Estimated Θ and ω for 1,000 Simulated Data Sets of Four Samples of Five Sequences

Sample	Days from First Sample	No. of Sequences	Θ Estimates	δ^a Estimates
1	0	13	0.0410	0.00386 (1.80×10^{-5})[b]
2	214	15	0.0388	0.01054 (2.31×10^{-5})
3	671	15	0.0519	0.0 (0.0)
4	699	9	0.0452	9.54×10^{-4} (3.12×10^{-6})
5	1,005	8	0.0410	n/a

[a] Measured in expected substitutions per site between the given sample and the sample immediately following it.

[b] Corresponding mutation rates are shown in parentheses in mutations per site per day.

is effectively allowed to have its own mutation rate, ω_i, so that $\delta_i = \omega_i t_i$, where t_i is the length of the interval. In a sense, δ parameterization provides a new intermediate model of evolution between the two extremes of a strict molecular clock and the absence of a molecular clock. We call this intermediate model the varying-clock model. With HIV, for instance, the application of antiretroviral therapy leads to changes in the relative frequencies of different infected cell types (Perelson et al. 1996). Since each cell type has a different mean generation time, a change in population structure will lead to a change in mean generation time and, consequently, a change in the average mutation rate. This was already alluded to above when we analyzed our example data set. Under such conditions, a varying-clock model may be appropriate. (Note that the varying-clock model we propose is different from lineage-specific models of variable mutation rates. In the latter, the mutation rate is assumed to change independently along different branches of the tree [Thorne, Kishino, and Painter 1998; Huelsenbeck, Larget, and Swofford 2000].)

Although we focused on rapidly evolving viral populations in this paper, it should be obvious that sUPGMA and its associated procedures of parameter estimation apply equally well to eukaryotic populations from which ancient and/or archival DNA is available. We anticipate that the search for better methods to analyze such populations will only become more important with the increasing frequency of longitudinal sampling strategies and the acquisition of DNA samples from ancient or archival material.

A computer program called PEBBLE that implements sUPGMA and other related methods, written in the Java programming language, can be obtained from *www.cebl.auckland.ac.nz.* This software will run on all computer platforms that support the Java Virtual Machine version 1.1 (JVM 1.1). This includes Microsoft Windows, Linux, and MacOS.

Acknowledgments

We thank Matthew Goode and Greg Ewing for assistance in developing the PEBBLE software package and the distributed programming techniques used for the simulations. For helpful comments and discussion, we thank Roald Forsberg and David Nickle. This work was supported by NIH grant 59174.

LITERATURE CITED

EPPERSON, B. K. 1999. Gene genealogies in geographically structured populations. Genetics **152**:797–806.

FELSENSTEIN, J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59**:139–147.

FU, Y. X. 1994. A phylogenetic estimator of effective population size or mutation rate. Genetics **136**:685–692.

GOJOBORI, T., E. N. MORIYAMA, and M. KIMURA. 1990. Molecular clock of viral evolution, and the neutral theory. Proc. Natl. Acad. Sci. USA **87**:10015–10018.

HOLMES, E. C., L. Q. ZHANG, P. SIMMONDS, C. A. LUDLAM, and A. J. LEIGH BROWN. 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of HIV-1 within a single infected patient. Proc. Natl. Acad. Sci. USA **89**:4835–4839.

HUELSENBECK, J. P., B. LARGET, and D. SWOFFORD. 2000. A compound Poisson process for relaxing the molecular clock. Genetics **154**:1879–1892.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KINGMAN, J. F. C. 1982*a*. On the genealogy of large populations. J. Appl. Probability **19A**:27–43.

———. 1982*b*. The coalescent. Stochastic Processes Appl. **13**: 235–248.

LEIGH BROWN, A. J. 1997. Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc. Natl. Acad. Sci. USA **94**:1862–1865.

LEITNER, T., and J. ALBERT. 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc. Natl. Acad. Sci. USA **96**:10752–10757.

NEE, S., E. C. HOLMES, A. RAMBAUT, and P. H. HARVEY. 1995. Inferring population history from molecular phylogenies. Philos. Trans. R. Soc. Lond. B Biol Sci. **349**: 25–31.

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.

PYBUS, O. G., A. RAMBAUT, and P. H. HARVEY. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155**: 1429–1437.

ROBINSON, D. F., and L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–148.

RODRIGO, A. G., and J. FELSENSTEIN. 1999. Coalescent approaches to HIV population genetics. Pp. 233–272 *in* K. 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.

Shankarappa, R., J. B. Margolick, S. J. Gange et al. (12 co-authors). 1999. Consistent viral evolutionary dynamics associated with the progression of HIV-1 infection. J. Virol. **73**:10489–10502.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy. W. H. Freeman, San Francisco.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. **15**:1647–1657.

Wolinsky, S. M., B. T. M. Korber, A. U. Neumann, M. Daniels, K. J. Kuntsman, A. J. Whetsell, M. R. Furtado, Y. Cao, D. D. Ho, and J. T. Safrit. 1996. Adaptive evolution of HIV-1 during the natural course of infection. Science **272**:537–542.

Edward Holmes, reviewing editor

Accepted August 7, 2000