Inferring Evolutionary Rates Using Serially Sampled Sequences from Several Populations

Allen G. Rodrigo,† Matthew Goode,*† Roald Forsberg,‡ Howard A. Ross,*† and Alexei Drummond*[1]*

*Computational and Evolutionary Biology Laboratory, School of Biological Sciences, and the †Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand; and ‡Bioinformatics Research Center, Department of Ecology and Genetics, University of Århus, Århus, Denmark

The estimation of evolutionary rates from serially sampled sequences has recently been the focus of several studies. In this paper, we extend these analyzes to allow the estimation of a joint rate of substitution, ω, from several evolving populations from which serial samples are drawn. In the case of viruses evolving in different hosts, therapy may halt replication and therefore the accumulation of substitutions in the population. In such cases, it may be that only a proportion, p, of subjects are nonresponders who have viral populations that continue to evolve. We develop two likelihood-based procedures to jointly estimate p and ω, and empirical Bayes' tests of whether an individual should be classified as a responder or nonresponder. An example data set comprising HIV-1 partial envelope sequences from six patients on highly active antiretroviral therapy is analyzed.

Introduction

Recently, there has been an increased interest in the analysis of serial nucleotide sequence samples that are gathered from the same population, each sample obtained at a different time. This includes samples from rapidly evolving viral populations such as HIV and Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) (Leitner and Albert 1999; Forsberg et al. 2001) and ancient DNA samples obtained from preserved or fossilized tissue (Leonard, Wayne, and Cooper 2000; Barnes et al. 2002; Lambert et al. 2002). Several methods have been developed to estimate the value of some time-dependent evolutionary parameter when serially sampled sequences are available. Rodrigo et al. (1999) and, later, Fu (2001) developed methods to estimate the generation time of a population. Maximum-likelihood and least-squares estimators of single or multiple substitution rates have also been developed (Drummond and Rodrigo 2000; Rambaut 2000; Drummond, Forsberg, and Rodrigo 2001). Drummond and Rodrigo (2000) also described a method to reconstruct serial genealogies using serial sample UPGMA (sUPGMA). Most recently, Seo et al. (2002*a*) have explored optimal experimental designs for serial sampling, when the aim is to estimate substitution rate and/or divergence times. In addition, Seo et al. (2002*b*) and Drummond et al. (2002) have described more sophisticated methods for the estimation of substitution rates and effective population size.

The estimation methods developed to date use only sequences sampled serially from a single population. However, certainly with viruses, it is quite common to sample viral sequences from several different hosts, and within each host, at different timepoints (e.g., Gunthard et al. 1999; Holmes et al. 1992; Rodrigo et al. 1999; Shankarappa et al. 1999). If we assume, as is frequently done for viruses such as HIV-1, that there is little likelihood of multiple transmission events, then viruses in each host are part of an isolated and unique population, evolving independently from a single founding variant.

In this paper, we describe two likelihood-based methods for jointly estimating a substitution rate using serially sampled sequences, when these are obtained from different populations. These methods are analogous to those developed by Gu (2001) for the analysis of functional divergence in protein families, and we use Gu's terminology in this paper. In the first of these procedures, the subtree of sequences from each population is treated as an unrelated phylogeny. This "subtree likelihood" (STL) approach uses the likelihoods of the subtrees as independent contributors to the total likelihood of all samples. In an alternative approach, a phylogeny of all sequences is constructed and the "whole-tree likelihood" (WTL) is then used as a basis for estimation.

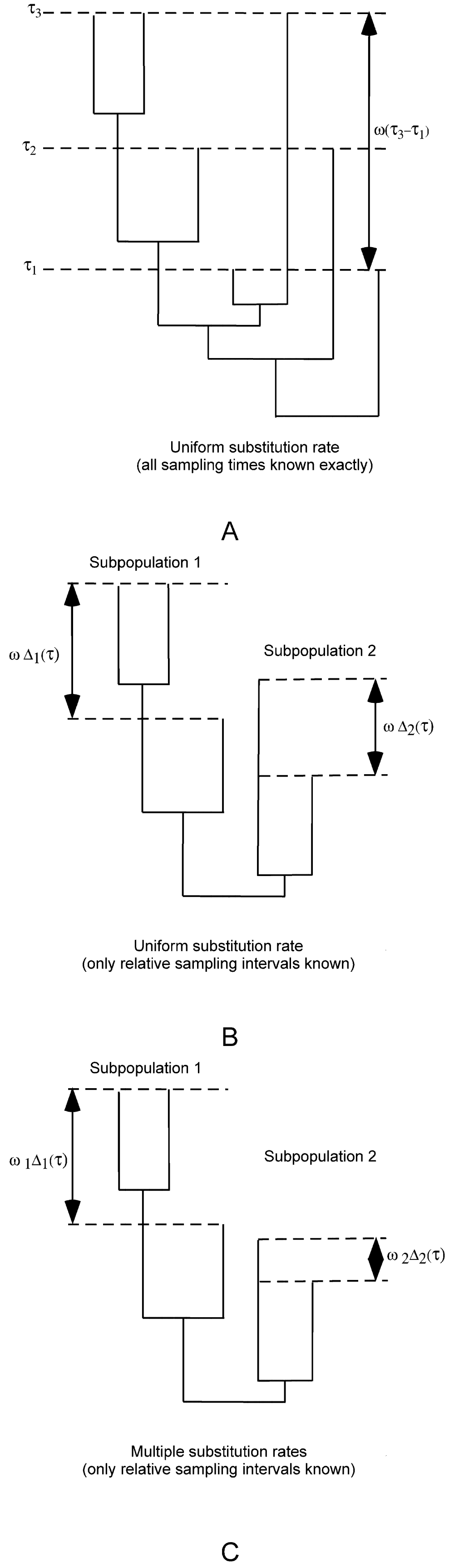
The joint estimation of substitution rate is a reasonably simple extension to work previously done (Rambaut 2000) under both approaches. There is, however, an interesting problem that provides a more challenging application of the STL or WTL procedure. Gunthard et al. (1999) described a study in which HIV-1 partial envelope (*env*) gene sequences were obtained from individuals just before and 2 years after the commencement of combination antiretroviral therapy. The aim of the study was to determine if antiretroviral therapy effectively controlled viral replication, as had previously been suggested by several workers (Finzi et al. 1997; Wong et al. 1997). In the event that the patient responds to therapy and viral replication is halted, there would be no measurable (or statistically significant) accumulation of substitutions in *env* sequences sampled before and after therapy. In a study such as this, the aim is to quantify and test whether the virus population continues to evolve within a host over the period of the study. Gunthard et al. (1999) analyzed each patient separately, but such an analysis runs the risk of inflating the probability of a type I error. Here, we apply the STL and WTL procedures to provide joint estimators of both the proportion of individuals who do not respond to therapy (i.e., whose viral population continues to replicate and accumulate

[1] Present address: Department of Statistics, and Department of Zoology, University of Oxford, Oxford, United Kingdom.

Key words: serial samples, substitution rate, subtree likelihood, whole-tree likelihood, maximum likelihood.

E-mail: a.rodrigo@auckland.ac.nz.

Mol. Biol. Evol. 20(12):2010–2018. 2003
DOI: 10.1093/molbev/msg215
Molecular Biology and Evolution, Vol. 20, No. 12,



Uniform substitution rate
(all sampling times known exactly)

Uniform substitution rate
(only relative sampling intervals known)

Multiple substitution rates
(only relative sampling intervals known)

substitutions) and the rate of ongoing viral substitution in these patients. Finally, we show how each patient may be assigned to the class of nonresponders or responders using empirical Bayes' classifiers.

Methodology

Consider the case of sequences sampled serially from a single population for which there is exact information on sampling times and a known phylogeny. In a model where it is assumed that there is a uniform rate of substitution (the single-rate with dated tips [SRDT] model, Rambaut 2000), total branch lengths from the root of the tree to the tips are no longer required to be equal. Instead, branch lengths are determined by the number of sampling intervals the branches traverse and the substitution rate (fig. 1). The parameters of the tree are the substitution rate, ω, the vector of times, τ, corresponding to the dated tips and the ($n-1$ for a bifurcating tree) internal node heights (h) measured in units of substitutions (following Rambaut 2000; note that the tip times may be measured either in generations or in some calendar unit, and a simple rescaling allows one to move between the two). As described previously by Drummond, Forsberg, and Rodrigo (2001), ω is only estimated within the interval bounded by the first and last samples. Specifically, no assumptions are made with regard to the rate between the earliest sampling time and the root of the tree. This is because the branch lengths, l, between the root and the earliest sampling time can only be optimized jointly as $l_i = \omega t_i$. Setting $\omega=0$ is equivalent to terminating all tips an equal distance from the root and assuming that all sequences in the sample are contemporary, as is done under a standard molecular clock model.

For a given phylogeny, T, for which only the topology is known, we may estimate the joint likelihood of ω and $\mathbf{H}$, the vector of internal node heights on T, as the conditional probability of obtaining the sequence data, $\mathbf{S}$, given ω, T, $\mathbf{H}$, and τ, the vector of sampling times, as well as the instantaneous substitution rate matrix, $\mathbf{M}$ (also assumed to be known):

$$L(\omega, \mathbf{H}) = Prob(\mathbf{S} \mid \omega, T, \mathbf{H}, \tau, \mathbf{M}) \qquad (1)$$

Since T, τ, and $\mathbf{M}$ are fixed, we will write $L(\omega, \mathbf{H}) = Prob(\mathbf{S} \mid \omega, \mathbf{H})$ without loss of generality. This likelihood is calculated in the standard manner (Felsenstein 1981; Goldman 1990; Rodriguez et al. 1990) for phylogenetic trees; the addition of ω and τ enters the calculations as constraints on the branch tip positions (fig. 1). The MLEs of the rate, $\hat{\omega}$, and elements of the vector of node heights, $\hat{\mathbf{H}}$, are constrained to be greater than or equal to zero and are

←

FIG. 1.—Alternative models of phylogenies with serially sampled sequences. (*A*) A "single-rate with dated tips" (SRDT) tree, with the times of serial sequence samples sequences known precisely. Under a molecular clock, sequences from each timepoint terminate at the same distance from the root of the tree. The branch lengths are extended by the product of a single substitution rate, ω, and the sampling interval. (*B*) A partially constrained phylogeny of sequences from two subpopulations. Only the lengths of the sampling intervals, Δ_1 and Δ_2, are known for the samples from both populations. A common rate, ω, is assumed. (*C*) A partially constrained phylogeny with sampling intervals known and with different rates, ω_1 and ω_2, for each population.

chosen such that $L(\hat{\omega}, \hat{\mathbf{H}})$ is maximized. It is worth noting, at this point, that since we are only interested in ω, $\mathbf{H}$ is a nuisance parameter, and we estimate it only because it is necessary to do so. (Note: Although we are assuming $\mathbf{M}$ to be fixed, it is also possible to estimate $\mathbf{M}$).

Using Subtree Likelihoods (STLs)

We wish to extend this model to the case where there are n serially sampled data sets, $\mathbf{S}_1,\ldots,\mathbf{S}_n$, each from a different population. Associated with each is a fixed tree, T_i, possibly a different model of evolution, $\mathbf{M}_i$, and a different set of sampling times, τ_i. When the aim is to estimate a common substitution rate, the likelihood function can be written as

$$\begin{aligned} &L(\omega, \mathbf{H}_1, \ldots, \mathbf{H}_n) \\ &\quad = Prob(\mathbf{S}_1, \ldots, \mathbf{S}_n \mid \omega, T_1, \ldots, T_n, \\ &\qquad\qquad \mathbf{H}_1, \ldots, \mathbf{H}_n, \tau_1, \ldots, \tau_n, \mathbf{M}_1, \ldots, \mathbf{M}_n) \end{aligned} \tag{2}$$

As above, we will write $L(\omega, \mathrm{H}_1, \ldots, \mathrm{H}_n) = Prob(\mathbf{S}_1, \ldots, \mathbf{S}_n \mid \omega, \mathrm{H}_1, \ldots, \mathrm{H}_n)$ since $T_1, \ldots, T_n$, $\tau_1, \ldots, \tau_n$ and $\mathbf{M}_1, \ldots, \mathbf{M}_n$ are fixed.

Here, we assume that $\mathbf{S}_1, \ldots, \mathbf{S}_n$ are drawn from independent populations, so that for a sample of aligned sequences, $\mathbf{S}_i$, from the ith population,

$$Prob(\mathbf{S}_i \mid \mathbf{S}_1, \ldots, \mathbf{S}_{i-1}, \mathbf{S}_{i+1}, \ldots, \mathbf{S}_n) = Prob(\mathbf{S}_i) \tag{3}$$

That is, sequences obtained from the ith population do not depend on sequences obtained from any other population. For this assumption to be true, the sequences from other populations must not have any influence on the prior probabilities of obtaining the sequences in the ith population. If this condition is met, then any estimate of ω that is derived using a set of sequences, $\mathbf{S}_i$, is totally uninfluenced by estimates of ω obtained with other samples. Such a situation would apply if the subtrees are connected by very long branches on the joint phylogeny or by very short branches with equal prior probabilities at the roots of all subtrees. As Gu (2001) points out, this approach is computationally tractable and, as we will show, appears to give very similar results to the WTL approach.

Given equation 3, it follows that

$$L(\omega, \mathbf{H}_1, \ldots, \mathbf{H}_n) = \prod_{i=1}^{n} Prob(\mathbf{S}_i \mid \omega, \mathbf{H}_1, \ldots, \mathbf{H}_n) \tag{4}$$

For the ith population, $\mathbf{S}_i$ depends on the evolutionary history of that population only and consequently on the node heights, $\mathbf{H}_i$, associated with topology, T_i. Therefore, equation 4 can be rewritten as

$$L(\omega, \mathbf{H}_1, \ldots, \mathbf{H}_n) = \prod_{i=1}^{n} Prob(\mathbf{S}_i \mid \omega, \mathbf{H}_i) = \prod_{i=1}^{n} L_i(\omega, \mathbf{H}_i) \tag{4a}$$

where $L_i(\omega, \mathbf{H}_i)$ is the likelihood defined in equation 1 of ω and $\mathbf{H}$ for the ith population. The joint MLE of ω and $\mathbf{H}_i, \ldots, \mathbf{H}_n$ is chosen to maximize equation 4. Asymptotic $(1 - \alpha)\%$ profile confidence limits of ω can be derived by locating upper and lower values, ω^*, such that

$$\ln L(\hat{\omega}, \hat{\mathbf{H}}_i, \ldots, \hat{\mathbf{H}}_n) - \ln L(\omega^*, \mathbf{H}'_1, \ldots, \mathbf{H}'_n) = \chi^2_{1,\alpha}/2 \tag{5}$$

where $\hat{\omega}$ and $\hat{\mathbf{H}}_1, \ldots, \hat{\mathbf{H}}_n$ are the MLEs of ω and $\mathrm{H}_1, \ldots, \mathrm{H}_n$, respectively, and $\mathbf{H}'_1, \ldots, \mathbf{H}'_n$ are the MLEs of $\mathrm{H}_1, \ldots, \mathrm{H}_n$ when $\omega = \omega^*$. For 95% confidence limits, $\chi^2_{1,0.05}/2 = 1.92$ (Rambaut 2000; Drummond, Forsberg, and Rodrigo 2001; Ota et al. 2001).

How do we modify equation 5 to allow groups of subpopulations to have different values of ω? As discussed above, we want to extend equation 4 to allow for the possibility that there is no measurable accumulation of substitutions in some populations so that for these, $\omega = 0$. The following description applies specifically to this case, but it is general enough to be applied to other values of ω as well. More importantly, whereas we focus exclusively on two rate categories (i.e., $\omega > 0$ and $\omega = 0$), these methods can also be generalized to data with more than two rate categories.

We define a Bernoulli random variable, R, where $R = 0$ classifies a population for which $\omega = 0$, and $R = 1$, a population where $\omega > 0$. Let $\mathbf{R} = (R_1, \ldots, R_n)$ represent the vector of population states. We can define the joint likelihood of $\mathbf{R}$ and a common positive-valued ω (for those populations for which $R = 1$) as

$$\begin{aligned} &L(\mathbf{R}, \omega, \mathbf{H}_1, \ldots, \mathbf{H}_n) \\ &\quad = Prob(\mathbf{S}_1, \ldots, \mathbf{S}_n \mid \mathbf{R}, \omega, \mathbf{H}_1, \ldots, \mathbf{H}_n) \end{aligned} \tag{6}$$

The condition of independence given in equation 3 needs to be extended as follows:

$$\begin{aligned} &Prob(\mathbf{S}_i \mid \mathbf{S}_1, \ldots, \mathbf{S}_{i-1}, \mathbf{S}_{i+1}, \ldots, \mathbf{S}_n, \mathbf{R}_1, \ldots, \mathbf{R}_i, \ldots, \mathbf{R}_n) \\ &\quad = Prob(\mathbf{S}_i \mid \mathbf{R}_i) \end{aligned} \tag{7}$$

This means that the evolution of sequences sampled from any given population depends on the status of that population only and not on that of any other population. Therefore,

$$\begin{aligned} L(\mathbf{R}, \omega, \mathbf{H}_1, \ldots, \mathbf{H}_n) &= \prod_{i=1}^{n} Prob(\mathbf{S}_i \mid R_i, \omega, \mathbf{H}_i) \\ &= \prod_{i} L_i(R_i, \omega, \mathbf{H}_i) \end{aligned} \tag{8}$$

$L_i(\mathbf{R}, \omega, \mathbf{H}_i)$ is either the likelihood of the tree (topology, $\boldsymbol{T}_i$, and node heights, $\mathbf{H}_i$) with all terminal tips equidistant from the root (when $R_i = 0$; ω is included for completeness but does not feature in the likelihood calculations) or the likelihood of $\boldsymbol{T}_i$ with tips terminating according to the sampling times and substitution rate, ω, (when $R_i = 1$). This is equivalent to finding the particular configuration of population states, $R_1 \ldots R_n$, and the value of ω associated with those populations for which $R = 1$, such that $L(\mathbf{R}, \omega)$ is maximized.

The value of this approach is that it identifies the populations that show an accumulation of substitutions and those that do not. However, frequently what is required is an estimate of the proportion of populations that are classified as either $R = 0$ or $R = 1$. Of course, this can be estimated simply from $\mathbf{R}$ after maximizing equation 7. Ideally, however, if the intention is to obtain an MLE of this proportion, the likelihood function needs to be recast.

Let the probabilities associated with $R = 0$ and $R = 1$ be $(1 - p)$ and p respectively. MLEs of p and a positive-valued ω can be obtained jointly by maximizing the following likelihood function:

$$\begin{aligned} L(\omega, p, \mathbf{H}_1, \ldots, \mathbf{H}_n) &= \prod_{i=1}^{n} \left[\sum_{j \in (0,1)} Prob(\mathbf{S}_i \mid R_i = j, \omega, p, \mathbf{H}_{ij}) \right. \\ &\qquad \left. \times Prob(R_i = j \mid \omega, p, \mathbf{H}_{ij}) \right] \\ &= \prod_{i} [(1-p)L_i(R_i = 0, \omega, \mathbf{H}_{i0}) + pL_i(R_i = 1, \omega, \mathbf{H}_{i1})] \end{aligned} \tag{9}$$

where $L(\mathbf{R} = 0, \omega, \mathbf{H}_{i0})$ is the ith likelihood of $\boldsymbol{T}_i$ with node heights, $\mathbf{H}_{i0}$, optimized under a standard molecular clock, and $L(\mathbf{R} = 1, \omega, \mathbf{H}_{i1})$ is a dated-tips tree with optimized node heights, $\mathbf{H}_{i1}$, and ω common to all populations.

Asymptotic bivariate $(1 - \alpha)$%-profile confidence envelopes may be obtained by locating pairs of (ω^*, p^*) such that $\ln \mathrm{L}(\hat{\omega}, \hat{p}, \hat{\mathbf{H}}_i, \ldots, \hat{\mathbf{H}}_n) - \ln L(\omega^*, p^*, \mathbf{H}'_1, \ldots, \mathbf{H}'_n) = \chi^2_{2,\alpha}/2$; here, $\mathbf{H}'_1, \ldots, \mathbf{H}'_n$ are the node heights that give the highest likelihood for ω^* and p^*. Alternatively, a profile confidence likelihood interval may be obtained for each parameter (either ω or p). For p, for instance, locate upper and lower values, p^*, such that

$$\begin{aligned} &\ln L(\hat{\omega}, \hat{p}, \hat{\mathbf{H}}_i, \ldots, \hat{\mathbf{H}}_n) - \ln L(\omega', p^*, \mathbf{H}'_1, \ldots, \mathbf{H}'_n) \\ &\quad = \chi^2_{1,\alpha}/2 \end{aligned} \tag{10}$$

The same procedure can be used to find upper and lower confidence values for ω.

Joint estimation of p and ω does not specifically identify which populations are classified as $R = 1$ or $R = 0$. It is however possible to use an empirical Bayes' procedure to classify the ith population according to their relative posterior probabilities after fixing $\omega = \hat{\omega}$, $p = \hat{p}$, and $\mathbf{H}_i$ to $\hat{\mathbf{H}}_{i0}$ or $\hat{\mathbf{H}}_{i1}$ depending on whether $R_i = 0$ or $R = 1$, respectively. To implement this, the ratio

$$\begin{aligned} \Lambda_i &= \frac{Prob(R_i = 1 \mid \mathbf{S}_i, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{i1})}{Prob(R_i = 0 \mid \mathbf{S}_i, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{i0})} \\ &= \frac{Prob(\mathbf{S}_i \mid R_i = 1, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{i1}) Prob(R_i = 1 \mid \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{i1})}{Prob(\mathbf{S}_i \mid R_i = 0, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{i0}) Prob(R_i = 0 \mid \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{i0})} \end{aligned}$$

$$\Lambda_i = \frac{\hat{p} Prob(\mathbf{S}_i \mid R_i = 1, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{i1})}{(1 - \hat{p}) Prob(\mathbf{S}_i \mid R_i = 0, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{i0})} \tag{11}$$

is calculated. If $\Lambda_i > 1$, then it is more probable that the ith population has a nonzero rate of substitution over the period of sampling.

It is worthwhile noting that identifying the precise configuration of $\mathbf{R} = (R_1, \ldots R_n)$, as we do in equation 8, and deriving $\hat{p}$ by calculating a posteriori the proportion of subpopulations classified as $R_i = 1$ (or 0) may lead to an inconsistent estimate of p, (i.e., there exists a small ε such that $Prob(|\hat{p} - p| < \varepsilon) \to 0$, as $n \to \infty$). This is because as n increases, the probability of incorrectly classifying subpopulations increases; this affects our estimate of p when it is calculated a posteriori. For this reason, it may be more defensible theoretically to estimate p directly.

For k ($k > 2$) categories of rates, there will be a vector $\mathbf{p} = \{p_1, \ldots, p_k\}$ where p_i corresponds to the proportion of subpopulations in rate category i. Equation 11 will be inapplicable in such a case. Nonetheless, for each subpopulation, it is easy enough to calculate the posterior probability of each rate category (for two categories, these correspond to the numerator and denominator of equation 11). These posterior probabilities can be thought of as "classification probabilities;" a subpopulation is assigned to the category with the highest classification probability.

Using Whole-Tree Likelihoods (WTL)

An alternative to the STL methods described above is to build and use a tree that represents the joint phylogeny of all sequences sampled from all populations. If the real sampling times from different populations are known, it is possible to build a serial phylogenetic tree of the entire set of sequences. In this circumstance, the complete phylogeny can be used to estimate a single mutation rate under the SRDT model. This would be analogous to what Gu (2001) refers to as the "whole-tree likelihood" approach. However, even if these times are available, it is still not obvious that a single tree and the SRDT analysis would be the appropriate approach, because it assumes that the rates of substitution of the virus between individuals are the same as those within individuals. Of course, this may not be true, since the accumulation of substitutions between individuals is subject to the evolutionary dynamics operating as a result of transmission from host to host.

An alternative is to construct a partially-constrained serial phylogenetic tree that allows the sequences within each population to evolve according to the SRDT model but also allows the lengths of branches connecting the subtrees of sequences from the different populations to vary freely (fig. 1*B*). In this case, the likelihood of the partially-constrained serial tree, $\boldsymbol{T}$, with a single model of evolution, $\mathbf{M}$, and node heights, $\mathbf{H_T}$, some of which are free to vary, is

$$\begin{aligned} &L(\omega, \mathbf{H_T}) \\ &\quad = Prob(\Phi = \{\mathbf{S}_1, \ldots, \mathbf{S}_n\} \mid \omega, T, \mathbf{H_T}, \tau_1, \ldots, \tau_n, \mathbf{M}) \end{aligned} \tag{12}$$

In this case, there is a single rate, ω, estimated for all populations regardless of the sampling intervals. It is perhaps more interesting to modify equation 12 to allow the populations to have different rates (e.g., some populations to have rate $\omega > 0$ and others have rate $\omega = 0$). In principle, this is straightforward: we need only constrain the node heights of the respective samples on a tree appropriate to their assigned rates (fig. 1*C*).

If we are interested in estimating the numbers of individuals with rates $\omega > 0$ or $\omega = 0$, it is possible to choose a partially constrained tree with the particular assignment of samples to each rate group that has the highest likelihood. So, by cycling through all 2^n possible combinations of rate assignments, we are able to identify the ML combination.

The approach above is equivalent to that applied using the subtree likelihood method, in which a particular combination of population states $R_1 \ldots R_n$ (and the value of ω associated with those populations for which $R = 1$) that maximizes $L(\mathbf{R}, \omega)$ is found. As with that approach, the disadvantage is that we do not estimate a proportion, p, of the number of populations that have state $R = 1$.

It is possible, albeit tedious, to estimate both p and ω using the WTL approach. Let $\mathbf{R}_i = (R_{i1}, \ldots R_{in})$ represent the ith combination of states assigned to the n samples, $i = 1, \ldots, 2^n$. Let k_i be the number of samples assigned state $R = 1$ and $(n - k_i)$ the number assigned state $R = 0$ in $\mathbf{R}_i$. The joint likelihood for ω and p is

$$L(\omega, p) = \sum_{i=1}^{2^n} p^{k_i}(1-p)^{(n-k_i)} L(\mathbf{R}_i, \omega) \qquad (13)$$

and is analogous to an expansion of equation 9. Obtaining equation 13 actually involves cycling through 2^n possible instances of the fixed topology, each of which has a different configuration of subclades assigned to the two rate categories.

Finally, we are interested in assigning subpopulations to different rate categories. As with the STL approach, we do this using an empirical Bayes' classifier. For the jth subpopulation, $j = 1, \ldots, n$, with rate assignment $R_{ij} \in \mathbf{R}_i$ in the ith combination of rate assignments,

$$\Lambda_j = \frac{Prob(R_{ij}=1 \mid \Phi, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=1)})}{Prob(R_{ij}=0 \mid \Phi, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=0)})}$$
$$= \frac{Prob(R_{ij}=1 \mid \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=1)}) Prob(\Phi \mid R_{ij}=1, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=1)})}{Prob(R_{ij}=0 \mid \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=0)}) Prob(\Phi \mid R_{ij}=0, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=0)})} \qquad (14)$$

where $\hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=1)}$ and $\hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=0)}$ indicate node heights of ML topologies for which the sequences associated with the jth subpopulation when it has rate assignments 1 and 0, respectively. The terms $Prob(\Phi \mid R_{ij} = 1, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=1)})$ and $Prob(R_{ij} = 0 \mid \Phi, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=0)})$ are, in fact, the likelihoods of the jth subpopulation having rate assignments 1 and 0, respectively, after fixing ω and p to their ML estimates. There are 2^{n-1} combinations in which the jth subpopulation has rate assignment 1 and the same number of combinations in which it has rate assignment 0. Therefore,

$$Prob(\Phi \mid R_{ij}=1, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=1)})$$
$$= L(R_{ij}=1) = \sum_{i=1}^{2^n} \frac{R_{ij}}{\hat{p}} \hat{p}^{k_i}(1-\hat{p})^{(n-k_i)} L(\mathbf{R}_i, \hat{\omega}) \qquad (15)$$

The multiplier $R_{ij}/\hat{p}$ needs a little explanation. The numerator, R_{ij}, ensures that the only terms that are used are those that correspond to combinations of $\mathbf{R}_i$ for which $R_{ij} = 1$. The denominator, $\hat{p}$, corrects the product $\hat{p}^{k_i}(1-\hat{p})^{(n-k_i)}$ because we are fixing the jth subpopulation to have rate $\mathrm{R}_{ij} = 1$. Similarly

$$Prob(\Phi \mid R_{ij}=0, \hat{\omega}, \hat{p}, \hat{\mathbf{H}}_{\mathbf{T}(R_{ij}=0)})$$
$$= L(R_{ij}=0) = \sum_{i=1}^{2^n} \frac{(1-R_{ij})}{(1-\hat{p})} \hat{p}^{k_i}(1-\hat{p})^{(n-k_i)} L(\mathbf{R}_i, \hat{\omega}) \qquad (16)$$

Therefore, substituting equation 15 and equation 16 into equation 14,

$$\Lambda_j = \frac{\hat{p} L(R_{ij}=1)}{(1-\hat{p}) L(R_{ij}=0)}$$
$$= \frac{\sum_{i=1}^{2^n} R_{ij} \hat{p}^{k_i}(1-\hat{p})^{(n-k_i)} L(\mathbf{R}_i, \hat{\omega})}{\sum_{i=1}^{2^n} (1-R_{ij}) \hat{p}^{k_i}(1-\hat{p})^{(n-k_i)} L(\mathbf{R}_i, \hat{\omega})} \qquad (17)$$

As with the STL approach, if $\Lambda_j > 1$ then the jth population is classified in rate category 1, that is, with rate $\omega = \hat{\omega} > 0$.

Example: Estimating the Proportion of Individuals Responding to Antiretroviral Therapy

Gunthard et al. (1999) studied the evolution of partial (regions C2–C3) HIV-1 *env* sequences over 2 years of combination antiretroviral therapy in six individuals. Viral RNA sequences were obtained just before therapy began ("early" sequences) and cell-associated viral DNA sequences were obtained 2 years later ("late" sequences). As mentioned previously, if therapy is successful at halting viral replication, there is no opportunity for the virus to accumulate mutations, since this only happens when viral RNA is reverse transcribed to cDNA after infection of host cells. Therefore, one expects that successful therapy will leave behind a population of viral "fossils" embedded in the genomes of host cells infected before therapy began. This means that when therapy begins, the mutation rate, ω, becomes zero. Note that setting $\omega = 0$ only makes sense if serial sequence samples are available. In the absence of serially sampled data, $\omega = 0$ implies that there can be no differences between sequences, but with serial samples, $\omega = 0$ only implies that over some period between sampling events, there was no accumulation of substitutions.

Gunthard et al. (1999) reconstructed the phylogenies of each set of sequences from each subject. They then measured the evolutionary distance of each sequence to the root of each tree and compared the distances of early and late sequences using a nonparametric test. There are obvious problems with this approach, principally the genealogical dependence of evolutionary distances. In effect, this analysis assumes that each sequence terminates a lineage that evolved independently from the most recent common ancestor. Here, we reanalyze the data obtained by Gunthard et al. (1999). We used PAUP* (Swofford 1999) to construct individual maximum-likelihood phylogenies for sequences from each subject with a common GTR model of evolution that we had previously estimated for all subjects simultaneously. We applied the analyses described above for both the STL-based and WTL-based approaches. For the WTL analyses, an unrooted tree of the entire data set was used. As expected, sequences from each subject clustered together. For the STL analyses, the phylogeny of each set of sequences was rooted using sequences from other subjects as outgroups. Once the trees were rooted, the outgroup

Table 1
Maximum-Likelihood Estimate of Substitution Rates[a] and Log-Likelihoods Associated with Different Values of ω

Subject	MLE of ω (−log-likelihood)	−log-likelihood when ω = 0	−log-likelihood when ω = 0.017	Λ_{STL}	Λ_{WTL}
Patient A	0.0000 (905.27)	905.27	914.58	2.26×10^{-5}	1.36×10^{-5}
Patient B	0.0000 (525.91)	525.91	673.52	1.96×10^{-65}	2.17×10^{-67}
Patient C	0.0032 (798.34)	799.30	804.25	0.002	0.001
Patient K	0.0000 (585.32)	585.32	585.35	0.242	0.01
Patient L	0.0000 (707.41)	707.41	817.18	5.31×10^{-49}	1.54×10^{-50}
Patient M	0.0176** (868.14)	875.32	869.75	65.609	492.749

NOTE.—** Statistically different from ω = 0 ($p < 0.01$). The last two columns provide empirical Bayes' ratios, calculated under the STL (Λ_{STL}) and WTL (Λ_{WTL}) approaches of the probabilities of ω > 0 versus ω = 0 for each population, as discussed in the main text. A value greater than 1 signifies that it is more probable that a population has a positive-valued substitution rate.

[a] Expressed as number of substitutions per site per year.

sequences were pruned from the trees so that the rooted topology contained only the sequences for that subject.

STL Analyses

First, for each subject, we derived a nonnegative MLE of ω by setting the times between early and late sequences at 2 years, using the estimated phylogeny and common GTR model of evolution. Interestingly, the MLEs of ω for four of the six subjects were, in fact, 0. The MLEs of ω for sequences obtained from Patient M was 1.8% per year, and that of Patient C was 0.3% per year. Using the asymptotic likelihood ratio test (LRT) described by Drummond, Forsberg, and Rodrigo (2001), we found that ω was statistically different from 0 only for Patient M ($P < 0.01$ [table 1]). This result is interesting because we were only able to find evidence for the continued accumulation of substitutions in one subject. In contrast, using the nonparametric approach and treating the distance-to-root of each sequence as an independent measurement, Gunthard et al. (1999) found that the viral populations in three subjects continued to evolve. This is not surprising, because the assumption that each sequence in a given sample sits at the tip of an independently evolving lineage falsely inflates both the degrees of freedom of a test (broadly defined, the apparent number of replicates) and our confidence that any estimated difference is statistically significant.

Next, we searched for the combination of six population states, $\mathbf{R} = (R_1, \ldots, R_6)$, representing sequence sets with statistically detectable increases in substitutions between sampling times ($R = 1$) and those without ($R = 0$). As described in equation 8, for any configuration in which $R = 1$ was assigned to more than one set of sequences, a common ω was estimated. The configuration that had the highest log-likelihood (–4,391.35) of all 64 possible configurations of **R** was one in which only Patient M had a non-0 ω (1.8% per year). This, of course, agrees with the result obtained above: only sequences from Patient M contained sufficient signal to detect a statistically significant non-0 substitution rate between sampling times.

At this point, it is worth noting that the log-likelihood of one other configuration was only very slightly different from that of the ML configuration. The configuration in which Patients K and M have states $R = 1$, and a common value of ω = 0.017 (1.7% per year) has a log-likelihood of –4,391.37. At first glance, this is a curious result—an examination of table 1 indicates that for Patient K, the MLE of ω = 0. Why then should Patient K be assigned a state that signals a detectable accumulation of substitutions? The reason becomes obvious when we examine the topology of the sequences for Patient K (fig. 2). The tree is reciprocally monophyletic, with early sequences (labeled with the prefix "KV") and late sequences ("KP") clustering on different clades. This means that simply by moving the position of the root on the branch connecting the two clades, it is possible to get estimates of ω that range from 0 to some positive value without changing the log-likelihood.

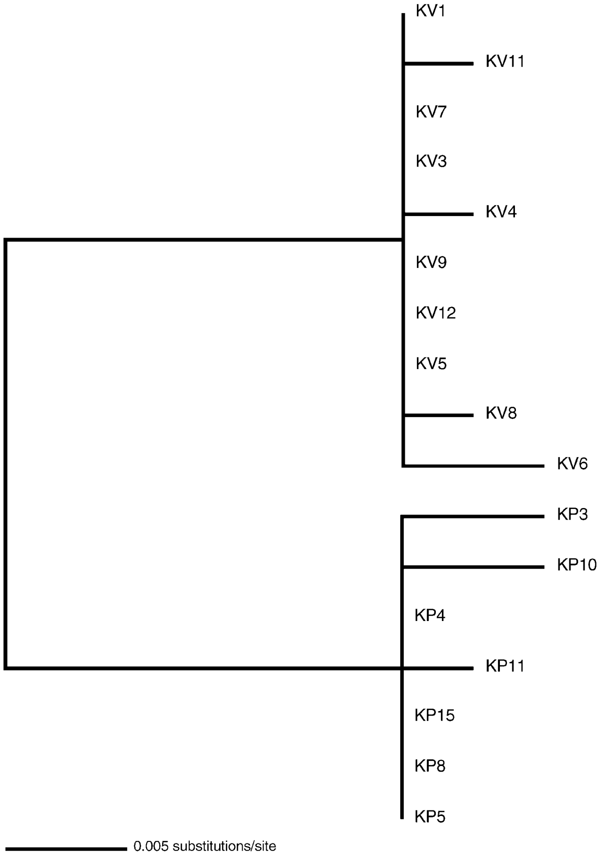


FIG. 2.—The midpoint-rooted phylogeny of sequences sampled from Patient K. Sequences with the prefix KV were obtained before therapy and those with the prefix KP were obtained 2 years after therapy.

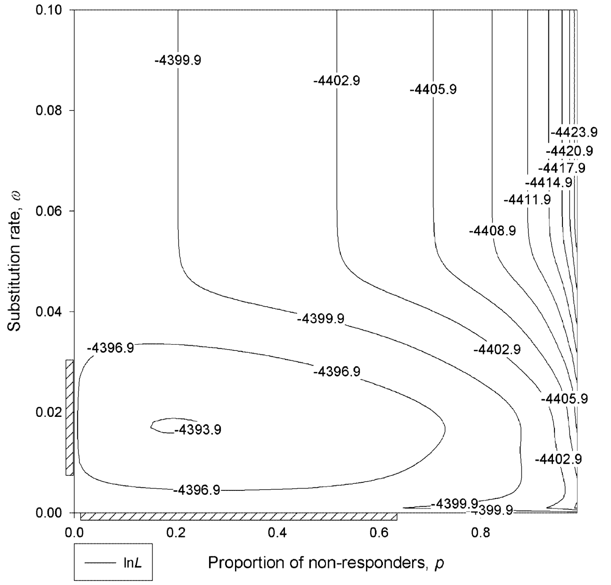


FIG. 3.—Contour plot of $\ln L$, obtained using equation 9, as a function of the proportion of responders, p, and substitution rate, ω. The contour corresponding to $\ln L$ = –4,396.9 is the joint 95% profile confidence envelope of ω and p. The hatched bars on the horizontal and vertical axes correspond to the 95% profile confidence intervals of p and ω, respectively.

Finally, we estimated the proportion of responders and a common ω by maximizing the likelihood given in equation 9. This was done using a grid search with values of p between 0 and 1 at an interval of 0.01 and values of ω between 0 and 0.1 and with an interval of 0.001. The resulting surface plot of log-likelihoods is given in figure 3. The joint ML estimates of ω and p are 0.017 and 0.20, respectively. The value of ω agrees with estimates obtained above, and the estimate of p is also consistent with those results. Only one of six subjects (Patient M) had a rate that was statistically non-0, and if we discount Patient K because the sequence data is uninformative about rates, then we are left with only one in five patients (or 20%) showing statistical evidence of continued evolution and, by implication, nonresponsiveness to therapy. The bivariate 95% profile confidence envelope corresponds to the contour that represents $\ln L = –4{,}396.9$, and the hatched bars on the horizontal and vertical axes represent the 95% profile confidence intervals for p and ω, respectively. Finally, using the ML values of ω and p to classify the status of each subject with an empirical Bayes' procedure (table 1), we found that, as expected, only Patient M had a value of Λ that was greater than 1, signifying a non-0 rate of evolution. It is worth noting that for Patient K, the likelihoods associated with ω = 0 and ω = 0.017 are effectively identical; consequently, the value of Λ is determined solely by the ratio of p to $(1 - p)$, since these are the probabilities of belonging to the two rate categories.

WTL Analyses

When a partially constrained tree was fitted to the data and a single rate allowed for all subpopulations on the tree (see equation 12), the ML estimate of ω was, in fact, 0 ($\ln L$= –2,717.5). However, if some populations were allowed to have a common ω > 0 and others, ω = 0, the combination of rate categories that had the highest likelihood ($\ln L$ = –2,709.7) was one in which only Patient M had a non-0 substitution rate, ω = 0.017 or 1.7% per year (fig. 4). This result is identical to that obtained above using STLs. Interestingly, the combination in which Patients K and M both have non-0 rates (ω = 0.012) has a log-likelihood that is very close ($\ln L$ = –2,711.2). This also agrees with the results obtained with the STL-based analysis.

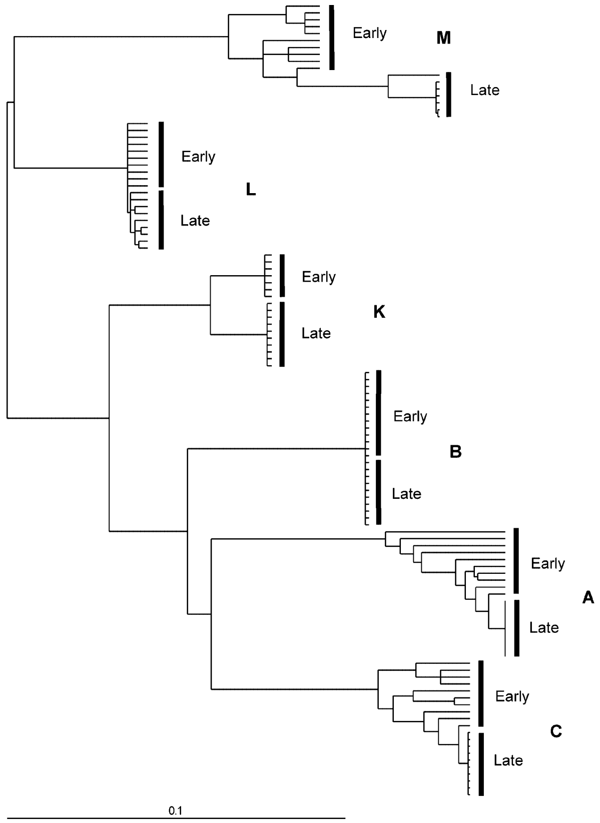


FIG. 4.—The partially-constrained joint phylogeny of sequences from all subjects indicating the ML combination of rate assignments. The ML combination only has Patient M assigned as a nonresponder.

We next used the WTL analysis to jointly obtain the ML estimates of ω and p. We did this using a grid search over ω and p with the same dimensions as done for the equivalent STL analysis. The contour plot of the log-likelihoods is shown in figure 5. The ML estimates of ω and p are 0.0175 and 0.17, respectively. Once again, these results are almost identical to those obtained using the equivalent STL analysis.

Finally, we applied the empirical Bayes' classifier given in equation 17 to each of the subpopulations and obtained the results shown in table 1. These results confirm our previous analyzes in showing that only Patient M can be classified as a nonresponder. The values of Λ are not markedly different from that obtained using the STL classifier.

Discussion

In this paper, we describe methods to estimate substitution rates of homologous genes from several

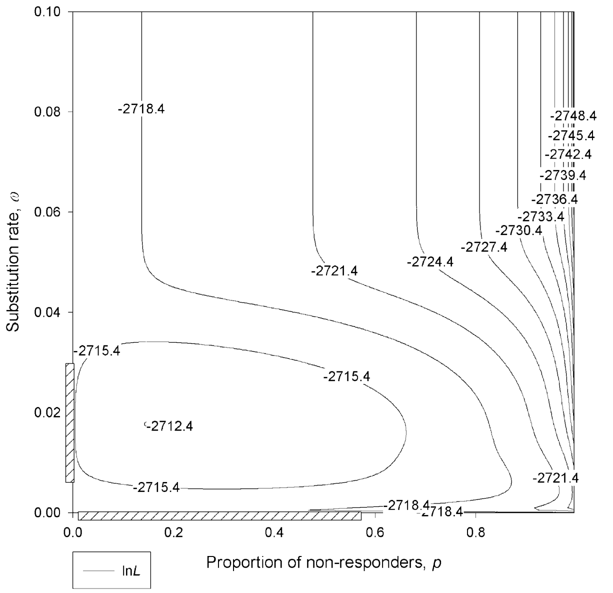


FIG. 5.—Contour plot of $\ln L$, obtained using equation 13, as a function of the proportion of responders, p, and substitution rate, ω. The contour corresponding to $\ln L = -2{,}715.4$ is the joint 95% profile confidence envelope of ω and p. The hatched bars on the horizontal and vertical axes correspond to the 95% profile confidence intervals of p and ω, respectively.

independently sampled populations. We make one principal assumption in constructing these methods: a single substitution rate applies to all sequences and all populations or, as in the case of subjects undergoing antiretroviral therapy, to those populations that continue to accumulate substitutions. Departures from this assumption can, nonetheless, be accommodated in the same framework. For instance, it is possible to restate equation 4a to differentiate the populations into two or more sets, each with its own substitution rate. In fact, this is exactly what equation 9 does, except that it constrains one of the rates to be 0. At the extreme, it is possible to allow each population to have its own substitution rate. However, in this case, the solution is trivial, because the likelihood is maximized when the ML rate for each population is determined.

The approaches we apply here based on subtree likelihoods and whole-tree likelihoods are equivalent to those used by Gu (2001). In our example, there appears to be very little difference in the results. It is worth reiterating the points that Gu (2001) makes in his comparison of the subtree likelihood versus whole-tree likelihood approaches. In essence, if there are very long or very short internal branch lengths between clades, the whole-tree method offers little improvement over the subtree method. This is because the value of the whole-tree approach depends on the extent to which the joint phylogeny influences the prior probabilities of nucleotide states at the roots of each of the individual subtrees and hence the likelihood of the tree and its attendant parameter estimates. If these prior probabilities and the resultant likelihood are unaffected or only marginally affected by combining the subtrees into a single joint phylogeny, then there is little added value in the whole-tree approach, particularly when we consider the computational overheads of the method. These can be quite substantial; for instance, the grid search to generate the likelihood contour plot using the STL method, programmed with JAVA version 1.4, took an average of 71 s on a PC with an AMD Athlon 1400+ processor and 512 Mb RAM. In contrast, the WTL grid search, on the same machine, took 8,195 s, or just over 2 h.

The STL methods we have developed in this paper can also be applied to serial sequence samples drawn from different, unlinked loci. In this case, we may be interested in testing whether the different loci are evolving at the same rate. Alternatively, the sampled loci may be partitioned into groups each evolving with its own rate. The methods we have described here work as well with such data.

Obviously, these methods should only be applied if each subpopulation is evolving in a clocklike manner. It should be routine to validate this assumption first using the tests described by Rambaut (2000) and Drummond, Forsberg, and Rodrigo (2001).

Our analyses, as we have described them, rely on the assumption of a given topology. To relax this assumption, we need to allow for uncertainty in the evolutionary relationships of the sampled sequences. Two of us (A.D. and A.G.R.) have been involved in recent work on the use of Bayesian analysis of serially sampled sequences using Markov chain Monte Carlo (MCMC) methods (Drummond et al. 2002). This approach allows different topologies to be sampled in proportion to their contribution to the joint posterior probability of all unknown quantities. The plan for the immediate future is to incorporate the analyses, and different options, described here into the MCMC framework already available.

Acknowledgments

This work was supported by NIH grant RO1-GM59174. A.D. was supported by a New Zealand FRST Bright Futures Scholarship. R.F. was supported by a NIH grant R01-GM60729, by grant number 9901522 from the Danish Agricultural and Veterinary Research Council, and by grant numbers 51-00-0392 and 21-02-0206 from the Danish Natural Science Research Council. We thank two anonymous reviewers for comments that helped us improve the manuscript.

Literature Cited

Barnes, I., P. Matheus, B. Shapiro, D. Jensen, and A. Cooper. 2002. Dynamics of Pleistocene population extinctions in Beringian brown bears. Science **295**:2267–2270.

Drummond, A., R. Forsberg, and A. G. Rodrigo. 2001. Estimating stepwise changes in substitution rates using serial samples. Mol. Biol. Evol. **18**:1365–1371.

Drummond, A., G. K. Nicholls, A. G. Rodrigo, and W. Solomon. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161**:1307–1320.

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

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

Finzi, D., M. Hermankova, T. Pierson et al. (15 co-authors). 1997. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. Science **278**:1295–1300.

Forsberg R., M. B. Oleksiewicz, A. M. K. Petersen, J. Hein, A. Bøtner, and T. Storgaard. 2001. A molecular clock dates the common ancestor of European-type porcine reproductive and respiratory syndrome virus at more than 10 years before the emergence of disease. Virology **289**: 174–179.

Fu, Y. X. 2001. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. Mol. Biol. Evol. **18**:620–626.

Goldman, N. 1990. Maximum likelihood inferences of phylogenetic trees, with special reference to the Poisson process model of DNA substitutions and to parsimony analysis. Syst. Zool. **39**:345–361.

Gu, X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. **18**:453–464.

Gunthard, H. F., S. D. Frost, A. J. Leigh Brown et al. (12 co-authors). 1999. Evolution of envelope sequences of human immunodeficiency virus type 1 in cellular reservoirs in the setting of potent antiviral therapy. J. Virol. **73**:9404–9412.

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.

Lambert, D. M., P. A. Ritchie, C. D. Millar, B. Holland, A. J. Drummond, and C. Baroni. 2002. Rates of evolution in ancient DNA from Adelie penguins. Science **295**:2270–2273.

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.

Leonard, J.A., R. K. Wayne, and A. Cooper. 2000. Population genetics of Ice Age brown bears. Proc. Natl. Acad. Sci. USA. **97**:1651–1654.

Ota, R., P. J. Waddell, M. Hasegawa, H. Shimodaira, and H. Kishino. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol. Biol. Evol. **17**:798–803.

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

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.

Rodriguez, F., J. F. Oliver, A. Marin, and J. R. Medina. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. **142**:485–501.

Seo, T. K., J. L. Thorne, M. Hasegawa, and H. Kishino. 2002*a*. A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times. Bioinformatics **18**:115–123.

———. 2002*b*. Estimation of effective population size of HIV-1 within a host. A pseudomaximum-likelihood approach. Genetics **160**:1283–1293.

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.

Swofford, D. L. 1999. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.

Wong, J. K., M. Hezareh, H. F. Gunthard, D.V. Havlir, C. C. Ignacio, C. A. Spina, and D. D. Richman. 1997. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. Science **278**:1291–1295.

Diethard Tautz, Associate Editor

Accepted July 6, 2003