The Inference of Stepwise Changes in Substitution Rates Using Serial Sequence Samples

Alexei Drummond, Roald Forsberg,† and Allen G. Rodrigo**

*School of Biological Sciences, University of Auckland, Auckland, New Zealand; and †Department of Ecology and Genetics, University of Århus, Århus, Denmark

It is frequently true that molecular sequences do not evolve in a strictly clocklike manner. Instead, substitution rate may vary for a number of reasons, including changes in selection pressure and effective population size, as well as changes in mean generation time. Here we present two new methods for estimating stepwise changes in substitution rates when serially sampled molecular sequences are available. These methods are based on multiple rates with dated tips (MRDT) models and allow different rates to be estimated for different intervals of time. These intervals may correspond to the sampling intervals or to a priori–defined intervals that are not coincident with the times the serial samples are obtained. Two methods for obtaining estimates of multiple rates are described. The first is an extension of the phylogeny-based maximum-likelihood estimation procedure introduced by Rambaut. The second is a new parameterization of the pairwise distance least-squares procedure used by Drummond and Rodrigo. The utility of these methods is demonstrated on a genealogy of HIV sequences obtained at five different sampling times from a single patient over a period of 34 months.

Introduction

Although molecular sequences accumulate substitutions over time, the rate at which this occurs may not be constant. The rate of substitution is dependent on biological processes, including the intensity of selection, changes in effective size (when selection is present), and changes in the dynamics of the population, say, a shift in mean generation time. These effects can change substitution rate (1) over time, (2) in different lineages, and (3) at different positions along the sequence. In this paper, we present methods that model the substitution rate of molecular sequences obtained serially from individuals within a population or between species (and higher taxa) by allowing the rate to change over time in a stepwise fashion.

Population genetics studies that utilize molecular sequences typically rely on samples of sequences that have been obtained contemporaneously (Felsenstein 1992; Fu 1994; Nee et al. 1995; Pybus, Rambaut, and Harvey 2000). However, there has recently been increased interest in the analysis of samples that are gathered serially, each at a different time. This includes samples from rapidly evolving viral populations such as HIV (Leitner and Albert 1999; Rodrigo et al. 1999) and samples of ancient DNA from fossilized remains (Leonard, Wayne, and Cooper 2000). It is our aim to derive estimates of substitutional parameters from this type of data using biologically relevant models.

Recently, two papers have independently described methods to estimate substitution rate, μ, from serial samples under the assumption of a molecular clock. Rambaut (2000) shows how a phylogeny-based maximum-likelihood estimate (MLE) of the constant substitution rate, μ, expressing the divergence between dated sequences as a product of μ and the time interval, can be obtained (fig. 1*a*). Drummond and Rodrigo (2000), using a distance-matrix least-squares (LS) approach, parameterize intersample divergence in two ways. First, analogous to Rambaut's single rate with dated tips (SRDT) model, ω-parameterization estimates only a single substitution rate, ω, using ωt_i as the intersample divergence for the ith interval with elapsed time t_i (ω is the number of substitutions per unit time, but since time may be measured in chronological units rather than in generations, Drummond and Rodrigo use ω instead of μ). Second, with δ-parameterization, each intersample interval is allowed to have its own substitution rate, ω_i; i.e., for the ith interval with elapsed time t_i, $\omega_i t_i = \delta_i$ (fig. 1*b*). In keeping with Rambaut's terminology, we will refer to this as the ''multiple rates with dated tips'' (MRDT) model. Drummond and Rodrigo (2000) go on to use these substitution rate estimates in a phylogenetic reconstruction procedure called serial-sample unweighted pair grouping method with arithmetic means (sUPGMA) which recovers a tree with lineages that terminate in the order of sampling.

In this paper, we extend Rambaut's (2000) tree-based SRDT likelihood estimation procedure to include the MRDT model. In addition, we show that there are two forms of the MRDT model, one where the rates are estimated differently for each sampling interval (corresponding to Drummond and Rodrigo's [2000] δ parameterization, above), and another where the rates are different for different a priori–defined intervals that do not necessarily coincide with sampling intervals (fig. 1*c*). Maximum-likelihood (ML) and LS estimators can be constructed for both forms of the MRDT model. Finally, we illustrate the use of these methods on an example data set of HIV-1 partial envelope (*env*) sequences obtained serially from an individual who was treated with Zidovudine midway through the sampling program.

Likelihood Model

Let us consider the case of sequence data for which there is exact time information and a known phylogeny.

Key words: serial samples, stepwise rate changes, maximum likelihood, least squares, substitution rate.

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. 18(7):1365–1371. 2001

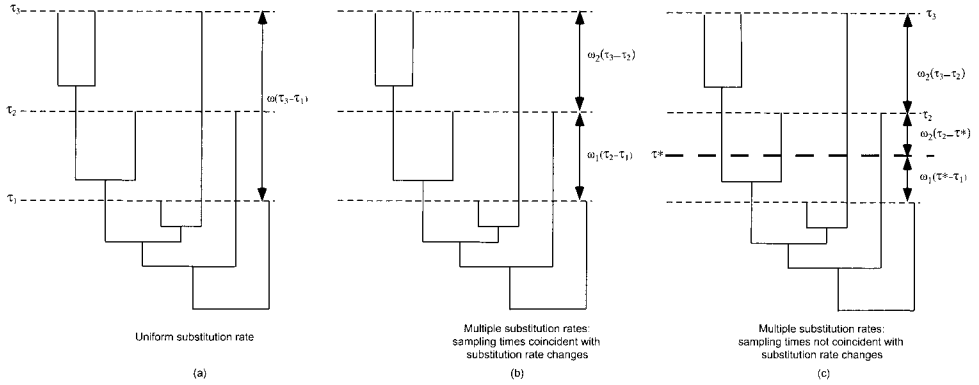

FIG. 1.—Three different models of substitution: (*a*) the SRDT model, with a uniform substitution rate; (*b*) the MRDT model, where each sampling interval has a different substitution rate; and (*c*) the MRDT model, where the substitution rate change point does not coincide with a sampling occasion.

Our generalization allows for the rate of substitution to have stepwise changes over time and gives rise to a multiple-rate model. The MRDT model is constructed by dividing the one substitution rate of the SRDT model (Rambaut 2000) into a vector $\boldsymbol{\Omega} = \{\omega_1, \omega_2, \ldots, \omega_k\}$, where ω_i is the *i*th substitution rate in the model (fig. 1). Hence, this ω-parameterization allows the substitution rate to have a number of stepwise changes between the most recent and the most ancient sampling times. As in the SRDT model, branch lengths from the root of the tree are no longer required to be equal. Instead, branch lengths must sum to values determined by the temporal spacing of the tip in question and the different substitution rates of the time periods that the tip traverses. Since the information about substitution rates comes from the relative positioning of tips in the tree, it is evident that rate parameters can only be estimated for time intervals for which there exists at least one sequence sample. Hence, the maximum number of ω parameters is given by the number of sampling points minus one, as one time point is needed for reference. However, this maximum number of rate parameters cannot be estimated for every tree topology. Take, for example, the simplest case of two sequences sampled at different times. In this situation, the uncertainty of the root confounds rate and time parameters, and the sequence data only hold information about the upper limit of the rate (set by the branch length between the two sequences).

The parameters of the tree are thus the substitution rates $\boldsymbol{\Omega}$ and the vector of times 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). Our framework estimates a series of substitution rates only 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. Over this interval, there is no chronological information, and the branch lengths are free to be optimized in the standard manner as composite parameters of time and substitution rate. This rate may, of course, be of interest, for example, in dating the most recent common ancestor (MRCA). In this case, additional assumptions must be made: a natural assumption in the case of stepwise changes is that the earliest estimated rate remains constant when extrapolated back in time to the root.

For a given tree T, the likelihood of $\boldsymbol{\Omega}$ is the conditional probability of obtaining the sequence data $\mathbf{S}$ given $\boldsymbol{\Omega}$, T, and $\boldsymbol{\tau}$, the vector of times, as well as the instantaneous substitution rate matrix $\mathbf{M}$ (also assumed to be known):

$$L(\boldsymbol{\Omega}) = \text{Prob}(\mathbf{S} \,|\, \boldsymbol{\Omega}, T, \boldsymbol{\tau}, \mathbf{M}). \quad (1)$$

This likelihood is calculated in the standard manner (Felsenstein 1981; Goldman 1990; Rodriguez et al. 1990) for phylogenetic trees; $\boldsymbol{\Omega}$ and $\boldsymbol{\tau}$ enter the calculations as constraints on the branch tip positions (fig. 1*b* and *c*). The MLEs of the rates, $\hat{\omega}_i$, are jointly chosen such that $L(\hat{\boldsymbol{\Omega}})$ is maximized. The only remaining constraint in place is that each estimated rate cannot be less than zero.

Confidence interval estimation in the case of multiple ω's is less straightforward. There are at least two ways of computing confidence intervals for multiple rates. First, multivariate upper and lower $(1 - \alpha)$% confidence limits may be obtained by locating rates that correspond to log likelihood values differing from the maximum log likelihood value by $\chi^2_{k,\alpha}/2$, where k is the number of rates estimated. If unbiased, these confidence intervals have a probability of $(1 - \alpha)$ of enclosing the true $\boldsymbol{\Omega}$. Second, a profile confidence likelihood interval may be obtained for each ω as follows. Over a range of ω_i, locate the upper and lower values of ω_i such that

$$\begin{aligned} -2\big|&\ln L(\omega_1^*, \omega_2^*, \ldots, \omega_{i-1}^*, \omega_i, \omega_{i+1}^*, \ldots, \omega_k^*) \\ &- \ln L(\hat{\omega}_1, \hat{\omega}_2, \ldots, \hat{\omega}_{i-1}, \hat{\omega}_i, \hat{\omega}_{i+1}, \ldots, \hat{\omega}_k)\big| \\ &= \chi^2_{1,\alpha}/2, \end{aligned} \quad (2)$$

where $\hat{\omega}_j$ is the MLE of the *j*th rate, and $\hat{\omega}_j^*$ is the ML estimate of the *j*th rate when ω_i is fixed at a given value.

In the case where all elements of $\boldsymbol{\Omega}$ are equal, the MRDT model collapses to the SRDT model of a uni-

form molecular clock. If all ω parameters are set to 0, the MRDT model reduces to the standard contemporaneous tips clock model (the single-rate [SR] model; Goldman 1993; Rambaut 2000). In fact, under the likelihood framework, one is able to test whether the MRDT model is a significantly better model for the data than the SRDT model. Since the SRDT model is simply a constrained MRDT model, the standard asymptotic likelihood ratio test can be applied. In this case, the test statistic,

$$\Delta = 2[\ln L(\boldsymbol{\Omega}, \text{not all } \omega \in \boldsymbol{\Omega} \text{ equal}) - \ln L(\boldsymbol{\Omega}, \text{all } \omega \in \boldsymbol{\Omega} \text{ equal})], \quad (3)$$

is asymptotically distributed χ^2 with $k - 1$ degrees of freedom under the null hypothesis that the two models are not significantly different, where k is the number of ω parameters.

When testing the SRDT model against the SR model, the null and alternative hypotheses are of the form

$$H_0: \quad \omega = 0 \qquad H_1: \quad \omega > 0.$$

The test is a one-tailed test. If α is chosen as the level of significance, then the null hypothesis should be rejected if

$$\Delta = 2[\ln L(\omega > 0) - \ln L(\omega = 0)] > \chi^2_{1,2\alpha}. \quad (4)$$

Incidentally, this result can also be derived by treating the constraint that ω has to be greater than or equal to zero as a boundary value problem (Ota et al. 2000).

Least-Squares Model

With the distance matrix LS estimate of $\boldsymbol{\Omega}$ described by Drummond and Rodrigo (2000), the expected evolutionary distance $d(m_i, n_j)$ between a pair of sequences m_i (of the ith sample; assume this is the earlier time point) and n_j, is equal to the expected pairwise distance Θ_i for sequences from sample i plus the added substitutions accruing between sequences from sample i and sample j in the interval $\tau_j - \tau_i$. If there exist times $\tau_{i+1}, \tau_{i+2}, \ldots, \tau_{j-1}$ in this interval that correspond to changes in substitution rate, then

$$d(m_i, n_j) = \Theta_i + \omega_{i\rightarrow i+1}(\tau_{i+1} - \tau_i) + \omega_{i+1\rightarrow i+2}(\tau_{i+2} - \tau_{i+1}) + \cdots + \omega_{j-1\rightarrow j}(\tau_j - \tau_{j-1}) + \epsilon_{m_i,n_j}. \quad (5)$$

The parameter estimates $\hat{\mathbf{P}} = \{\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Omega}}\}$ are obtained by the standard LS solution:

$$\hat{\mathbf{P}} = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{d}, \quad (6)$$

where $\mathbf{d}$ is the vector of the pairwise distances, and $\mathbf{t}$ is the matrix of time intervals and [0, 1] values signifying the absence or presence, respectively, of the Θ's associated with each of the samples. Unlike the MLE, LS rate estimates obtained using equation (6) are not constrained to be nonnegative. Such a constraint can be added with appropriate linear programming strategies.

The standard error of the LS estimates of ω cannot be calculated easily because of the nonindependence of the pairwise distances. Drummond and Rodrigo (2000) advocate the use of the parametric bootstrap (Efron and Tibshirani 1993; Goldman 1993) to generate confidence intervals of the estimates. Parametric bootstrapping requires specification of a model and subsequent simulation of pseudoreplicate data sets with the same number of sequences and sites as the original data, assuming that the estimates recovered using the observed data are the ''true'' values of the parameters. With the SRDT model and an assumption of a constant Θ over time, it is easy to generate pseudoreplicate data sets under a coalescent model in which population size is held constant (Drummond and Rodrigo 2000). However, under the MRDT model, parametric bootstrapping is not simple, since any resampling procedure must accommodate changing substitution rates and Θ's. This is a drawback of the distance-based LS method—procedures for variance estimation are often elusive.

Example

In this section, we illustrate the use of the MRDT model on an HIV data set previously published by Rodrigo et al. (1999), where the onset of drug therapy is shown to coincide with a significant reduction in substitution rate.

Before the advent of potent combination therapy against HIV, drugs were less effective in lowering viral load and hindering progression toward AIDS. To investigate the effect of a one-drug therapy regime on the evolutionary progression of HIV, we analyzed previously published data consisting of serially sampled partial HIV-1 envelope (*env*) sequences from an infected individual who began Zidovudine treatment partway through the sampling period (Rodrigo et al. 1999). Complete details of the data set are given in Rodrigo et al. (1999); briefly, the data set contains an initial sample followed by additional samples after 7 (day 214), 22 (day 671), 23 (day 699), and 34 months (day 1005). Monotherapy with Zidovudine was initiated after 13 months (day 409; J. Mullins, University of Washington, personal communication) and continued during the remaining time of study. Therefore, the data set contains two samples from before and three samples from after treatment began.

It has been suggested that highly active combination antiretroviral therapy leads to a cessation of viral replication (Finzi et al. 1997; Wong et al. 1997). A natural question is whether monotherapy with Zidovudine had the effect of slowing or halting viral replication in the particular individual from whom samples were available. If viral replication does, in fact, cease (or slow down), this will be reflected in the rate at which substitutions accumulate, since it is during the process of viral replication that this occurs. This corresponds to testing whether a, MRDT model, which allows for a change in

Table 1
Maximum-Likelihood and Least-Squares (LS) Estimates of Substitution Rates Under the Single-Rate (SR), Single Rate with Dated Tips (SRDT), and Multiple Rates with Dated Tips (MRDT) Models

Dataset	Model	$-\ln L$	MLE (substitutions/site/day)	Hypothesis Tests	Δ	df	P	LS Estimates (substitutions/site/day)
Complete	SR	4,082.50						
	SRDT	4,080.83	1.36×10^{-5} $[7.14 \times 10^{-6}, 2.09 \times 10^{-5}]$	SRDT vs. SR	3.34	1	0.034	7.8×10^{-6} $[-3.47 \times 10^{-6}, 3.87 \times 10^{-5}]$[b]
	MRDT	4,075.41	$\omega_{before} = 4.15 \times 10^{-5}$ $[2.6 \times 10^{-5}, 5.8 \times 10^{-5}]$[a] $\omega_{after} = 0.0$ $[0.0, 0.8 \times 10^{-5}]$[a]	MRDT vs. SRDT	10.84	1	0.001	$\omega_{before} = 3.87 \times 10^{-5}$ $\omega_{after} = -3.35 \times 10^{-6}$
Before therapy ..	SR	2,441.16	—					
	SRDT	2,430.90	5.03×10^{-5} $[2.81 \times 10^{-5}, 7.49 \times 10^{-5}]$	SRDT vs. SR	20.52	1	3.0×10^{-6}	2.69×10^{-5} $[-8.28 \times 10^{-6}, 1.22 \times 10^{-4}]$[b]
After therapy ..	SR	2,542.34	—					
	SRDT	2,542.10	5.8×10^{-7} $[0.0, 2.12 \times 10^{-5}]$	SRDT vs. SR	0.48	1	0.24	4.51×10^{-6} $[-2.51 \times 10^{-5}, 6.59 \times 10^{-5}]$[b]

NOTE.—MLE = maximum-likelihood estimate. Confidence intervals are presented in brackets.
[a] Profile likelihood confidence intervals.
[b] The 95% confidence intervals were obtained by a parametric bootstrap procedure using 1,000 replicates. Parameter Θ was kept constant.

substitution rate after the onset of therapy, provides a better fit to the data than an SRDT model and, if so, whether the estimated substitution rate after drug therapy is significantly different from zero.

The data set consisted of 60 sequences from five time points. The length of the alignment was 660 nt. Gapped columns were included in the analysis. To begin with, the data set was first split into two subsets, one containing all sequences before therapy (28 sequences), and the other containing all sequences after therapy commenced (32 sequences). For each of these data sets, a neighbor-joining tree was built and an ML general time-reversible (REV) model was estimated using PAUP* 4.0b4 (Swofford 1999).

Each tree was used to estimate a uniform substitution rate using the SRDT likelihood model as implemented in the computer program TIPDATE (Rambaut 2000). TIPDATE was also used to find the ML roots for the two trees. This was done by rooting the tree at every branch on the unrooted topology and optimizing the branch lengths in accordance with the dated tips. The rooted topology that maximizes the likelihood was used to estimate the substitution rate. All estimated rates are reported in table 1. A rate (ω_{before}) of 5.034×10^{-5} substitutions per site per day (1.84% per year, 95% confidence limit = [1.02%, 2.73%]) was obtained for the sequences before therapy, and a rate (ω_{after}) of 5.8×10^{-7} substitutions per site per day (0.021% per year, 95% confidence limit = [0.0%, 0.77%]) for the sequences after therapy. As ω_{after} has a confidence interval that encloses 0, we cannot show that significant substitutions have occurred since therapy commenced.

The complete data set, consisting of sequences obtained before and after therapy, was then used to obtain an unconstrained and unrooted neighbor-joining tree, once again using the REV substitution model. Once again, an SRDT model was fitted to the tree (after the ML root was found), and a uniform substitution rate of 1.346×10^{-5} substitutions per site per day (0.49% per year, 95% confidence limit = [0.26%, 0.76%]) was estimated. An MRDT model was then fitted to the full data set, allowing two substitution rates, the first up to the time of therapy (i.e., 409 days from the first sample), and the second after this time. Rates of 4.145×10^{-5} substitutions per site per day (1.51%) and 0.0 substitutions per site per day were estimated simultaneously for ω_{before} and ω_{after}, respectively. Trees reconstructed using MRDT and SRDT models are shown in figure 2. To obtain the 95% confidence intervals for both substitution rates, a grid search of the two parameters was undertaken. The rate ω_{before} was allowed to vary from 0 to 10^{-4} substitutions per site per day, while ω_{after} was allowed to vary from 0 to 5×10^{-5} substitutions per site per day, both in steps of 10^{-6} substitutions per site per day. The likelihood surface resulting from this search is shown in figure 3 as a contour plot. The resulting 95% profile confidence intervals were obtained by taking the maximum and minimum values of ω_{before} and ω_{after} on the contour demarcating $\chi^2_{1,0.05}$ (=1.92) log likelihood units from the maximum log likelihood. For ω_{before}, the profile likelihood confidence interval is $[2.6 \times 10^{-5}, 5.8 \times 10^{-5}]$, whereas for ω_{after}, it is $[0, 0.8 \times 10^{-5}]$. The bivariate confidence interval for $\hat{\mathbf{\Omega}} = \{\hat{\omega}_{before}, \hat{\omega}_{after}\}$ is also outlined on the likelihood surface contour plot by the contour demarcating $\chi^2_{2,0.05}/2$(=2.99) log likelihood units from the maximum log likelihood. The upper and lower values of ω_{before} and ω_{after} on this bivariate confidence interval contour are $[2.1 \times 10^{-5}, 6.1 \times 10^{-5}]$ for ω_{before} and $[0, 1.1 \times 10^{-5}]$ for ω_{after}. Of course, these intervals are larger than the profile likelihood confidence intervals, but only marginally so.

Table 1 gives the log likelihood scores obtained using the different models described above. For the complete data set with samples before and after therapy included, the most general clocklike model is the MRDT model. As explained above, the SRDT model is constrained so that all ω's are equal. The SR model with contemporaneous tips is a further constraint on the SRDT with all ω's equal and set to zero. In Table 1,

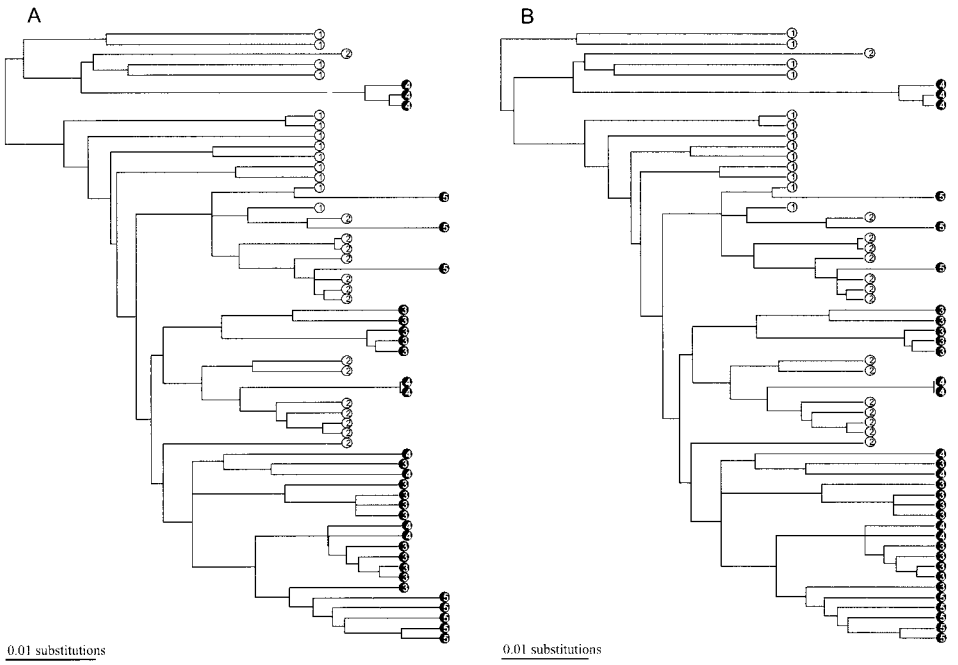

FIG. 2.—The maximum-likelihood SRDT tree (*A*) and the maximum-likelihood MRDT tree (*B*) for the full example data set. Open and filled circles represent before- and after-therapy sequences, respectively. Sample numbers are given within the circles.

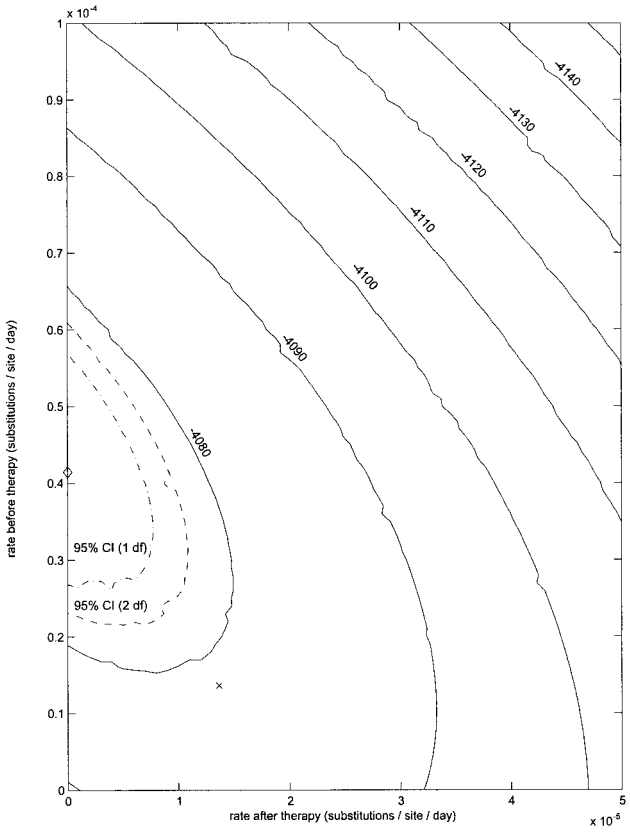

FIG. 3.—The likelihood surface of the ω parameters (substitution rates before and after therapy) for the example data set. Both the 95% profile confidence region and the 95% bivariate confidence region are shown. A cross marks the maximum-likelihood equal rates point (SRDT) on the surface, located outside both confidence regions. A diamond marks the peak of the likelihood surface.

likelihood ratio test statistics have been computed for MRDT versus SRDT and for SRDT versus SR models. The SRDT model is significantly better than the SR model ($P < 0.05$), and the MRDT model is significantly better than the SRDT model ($P < 0.01$).

Similar analyses were performed for before- and after-therapy samples, except that in these instances, the only comparison made was between the SRDT and SR models. For the before-therapy samples, the SRDT model has a statistically better fit to the data than the SR model ($P < 0.01$). However, for the after-therapy sequence subset, the SRDT model cannot be distinguished statistically from the SR model. Taken on its own, this suggests that there is little or no accumulation of substitutions over this period. (Caution must be taken with this interpretation: as we discuss in the next section, the MRDT model is significantly worse than a model that assumes no consistent clocklike pattern of evolution among the sequences).

Equivalent estimates were also derived with the LS method. Table 1 summarizes the results. Both the likelihood and the least-squares procedures consistently estimated a higher rate of substitution before therapy, about an order of magnitude greater than the estimated rate after therapy.

Discussion

The framework presented allows for the modeling of complex evolutionary scenarios such as the evolution of HIV sequences undergoing drug therapy. Application of the MRDT model to samples obtained from an individual treated with zidovudine appears to indicate a reduction in substitution rate after the commencement of therapy. Our results are consistent with those obtained

elsewhere (Chun et al. 1997; Wong et al. 1997). Independent rate estimates from samples before and after therapy have nonoverlapping 95% confidence intervals and are therefore significantly different at $\alpha = 0.05$. This poses a problem for any SRDT estimation procedure. TIPDATE, for instance, returns a rate of 0.5% per year for the entire genealogy. This rate is lower than previously published rates of HIV evolution (0.93% per year; Shankarappa 1999). However, it is similar to other published estimates for this data set that assume a single rate (0.3% per year; Drummond and Rodrigo 2000). In this paper, we outlined a likelihood framework that addresses this discrepancy, in addition to providing a pairwise distance least-squares estimation approach. There are, nonetheless, several features of these analyses that bear mention and indicate that more work in this area is required.

First, for any analysis that involves the inference of some kind of clocklike behavior, whether it be a constant clock or a changing clock, a first step should be a test of whether such a model is significantly worse than an unconstrained nonclock model (also called a different-rate [DR] model). The DR model is the standard used in phylogenetic tree reconstruction and effectively allows every branch to have its own substitution rate. Thus, the length of the *i*th branch is an estimate of the composite parameter $\omega_i t_i$. If an SRDT model or an MRDT model is significantly worse than the DR model, it means that at least some lineages are not evolving in a clocklike manner. In fact, where appropriate, we recommend a hierarchy of nested likelihood ratio tests: DR versus MRDT, MRDT versus SRDT, and SRDT versus SR. For our example data set, the DR model was always significantly better than the SRDT and MRDT models (data not shown). Our primary intention with the use of the data set was simply to illustrate the methods described in this paper, rather than to make substantive statements about the effects of monotherapy on substitution rates. It is important, however, to note this here for completeness.

Also, the likelihood estimation procedures presented here (and by Rambaut [2000]) assumes that the evolutionary history of the sequences, i.e., the topology of the genealogy, is known or can be reconstructed correctly. The bias introduced into parameter estimation and hypothesis-testing procedures by using incorrect genealogies is largely unknown. On the other hand, the least-squares estimation procedure is not based on a reconstructed topology and therefore may not suffer from this possible source of bias. For example, for a single-rate model, the least-squares estimator has been shown to be an unbiased estimator (Drummond and Rodrigo 2000). However, distance-based LS methods do not take into account the correlations induced by shared history, thus making variance estimation difficult.

Ultimately, the best approach would be to incorporate the uncertainty of the genealogy explicitly into a probabilistic framework. One way of taking the uncertainty of the topology into consideration in the likelihood model is to integrate over a number of topologies. A natural way to do this is to use a Markov chain Monte Carlo (MCMC) sampling procedure to sample tree space in proportion to the likelihood of the data (Kuhner, Yamato, and Felsenstein 1995). This approach has been used, for example, to incorporate the uncertainty in the tree topology into estimates of population size and growth rates (Kuhner, Yamato, and Felsenstein 1995, 1998). This method has a natural extension to the estimation of substitution rates and can also be used to find confidence intervals in topology space under the SRDT and MRDT models of evolution.

One of the interesting observations of this study is that different models (SR, SRDT, MRDT) can have different ML tree topologies. This may turn out to be a common occurrence. For example, for a 45-sequence subset of the data, 729 strictly bifurcating ML tree topologies were found. Although these trees had identical likelihoods under an unconstrained (nonclock) model, they had a range of likelihoods under the SR, SRDT, and MRDT models. Furthermore, no single strictly bifurcating topology represented the ML topology under all three models. If one chooses to use different topologies for each model, then the asymptotic approximation to the likelihood ratio test cannot be used. Instead, some alternative procedure (say, a parametric simulation procedure [Goldman and Whelan 2000]) should be used. A sampling method such as MCMC would also be useful in this case, as the sampling procedure integrates over tree space in proportion to the likelihood of the data. Thus, for two competing models, a null and alternative distribution can be compared.

In the previous section, we also alluded to the fact that different rootings of an unrooted tree can have different likelihoods under a given model of substitution. By extension, this also means that different models may require the tree to be rooted differently. This does not change the mechanics of any likelihood ratio test, since no new free parameters are added to the model. However, if the root of the tree is not known, an extra step needs to be added to any analysis to find the appropriate root.

Serial molecular samples add a new dimension to population genetics studies. Since it is possible to estimate substitution (or mutation) rate independently of other parameters, it is also possible to decouple composite population parameters like $\Theta = 2N_e\mu$ (where N_e is the effective population size) into their component parts. The models we introduce here go one step further and allow these parameters to be expressed as functions of time. Although we have only spoken of stepwise changes in substitution rates, these models can be generalized to allow substitution rate to vary as any predefined function of time. With viral populations such as HIV, this becomes especially interesting, since it allows us to study changes in average generation time and substitution rate during disease progression or under different therapeutic regimes. In conjunction with the estimation of demographic functions of time (Pybus, Rambaut, and Harvey 2000), it also means that we can decompose $\theta(t) = 2N_e(t)\mu(t)$ into the component functions of $N_e(t)$ and $\mu(t)$, where $\mu(t)$ is a stepwise function of time.

In this paper, we have assumed that the times corresponding to changes in substitution rates are fixed either to the sampling times or to some time point known a priori. Similarly, we have also assumed that the phylogeny is known. However, since these times and the phylogeny are parameters embedded in the model, they can also be jointly estimated within the likelihood framework.

The models we have described in this paper apply to any set of molecular sequences of sufficient length, or obtained sufficiently far apart in time, that an appreciable amount of substitutions has accumulated. These include ancient DNA sequences, as well as rapidly evolving viral sequences. In conjunction with efforts to model lineage-specific rates (Thorne, Kishino, and Painter 1998; Huelsenbeck, Larget, and Swofford 2000) and other time- or lineage- dependent processes, the models presented here go some way toward a more realistic description of the evolution of molecular sequences.

MLE and LS estimates under the SR, SRDT, and MRDT models can be obtained using the computer program PEBBLE, available from the website http://www.cebl.auckland.ac.nz or from the authors.

Acknowledgments

We thank Matthew Goode and Greg Ewing for assistance in developing the PEBBLE software package and the distributed programming techniques used for likelihood surface calculations. We thank Jim Mullins for providing information on the timing of therapy and other aspects of the example data set. For helpful comments and discussion, we thank David Nickle, Andrew Rambaut, and another anonymous reviewer. This research was funded by NIH grant GM59174. A.D. was supported by a New Zealand FRST Bright Futures Scholarship.

LITERATURE CITED

CHUN, T. W., L. STUYVER, S. B. MIZELL, L. A. EHLER, J. A. MICAN, M. BASELER, A. L. LLOYD, M. A. NOWAK, and A. S. FAUCI. 1997. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. Proc. Natl. Acad. Sci. USA **94**:13193–13197.

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.

EFRON, B., and R. TIBSHIRANI. 1993. An introduction to the bootstrap. Chapman and Hall, London.

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

———. 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.

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.

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

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.

———. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

GOLDMAN, N., and S. WHELAN. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol. Biol. Evol. **17**:975–978.

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

KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140**:1421–1430.

———. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149**:429–434.

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.

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.

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.

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.

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.

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). Version 4.0b4. Sinauer, Sunderland, Mass.

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.

WONG, J. K., M. HEZAREH, H. F. GÜNTHARD, 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**:1291–1300.

EDWARD HOLMES, reviewing editor

Accepted March 23, 2001