BIOINFORMATICS

Vol. 18 no. 1 2002
Pages 115–123

A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times

Tae-Kun Seo[1, 2], Jeffrey L. Thorne[3], Masami Hasegawa[1, 2] and Hirohisa Kishino[4,*]

[1]Department of Biosystems Science, The Graduate University for Advanced Studies, Hayama, Kanagawa, 240-0193, Japan, [2]The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu Minato-ku, Tokyo 106-8569, Japan, [3]Bioinformatics Research Center, Box 7566, North Carolina State University, Raleigh NC 27695-7566, USA and [4]Laboratory of Biometrics, Graduate School of Agriculture and Life Sciences, University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo 113-8657, Japan

Received on January 30, 2001; revised on May 28, 2001; accepted on July 23, 2001

ABSTRACT

Motivation: The high pace of viral sequence change means that variation in the times at which sequences are sampled can have a profound effect both on the ability to detect trends over time in evolutionary rates and on the power to reject the Molecular Clock Hypothesis (MCH). Trends in viral evolutionary rates are of particular interest because their detection may allow connections to be established between a patient's treatment or condition and the process of evolution. Variation in sequence isolation times also impacts the uncertainty associated with estimates of divergence times and evolutionary rates. Variation in isolation times can be intentionally adjusted to increase the power of hypothesis tests and to reduce the uncertainty of evolutionary parameter estimates, but this fact has received little previous attention.

Results: We provide approximations for the power to reject the MCH when the alternative is that rates change in a linear fashion over time and when the alternative is that rates differ randomly among branches. In addition, we approximate the standard deviation of estimated evolutionary rates and divergence times. We illustrate how these approximations can be exploited to determine which viral sample to sequence when samples representing different dates are available.

Contact: seo@ism.ac.jp; thorne@statgen.ncsu.edu; hasegawa@ism.ac.jp; kishino@wheat.ab.a.u-tokyo.ac.jp

INTRODUCTION

The process by which viral RNA is copied into DNA by the viral enzyme reverse transcriptase is error-prone and forms the basis for high genetic variability and for a high rate of evolution. Several authors have reported that viruses with RNA genomes mutate at a very high speed (reviewed by Mansky, 1998). As a result, the speed of viral evolution can exceed more than one million times the rate at which mammals evolve (Suzuki and Gojobori, 1997). A viral sequence from an isolate obtained just a few years ago may be substantially more similar to a common ancestral sequence than would be a viral sequence isolated today. Thus, isolation dates should often be considered when analyzing viral data. Viral sequences that are sampled at different times, which are termed 'serially sampled data' or 'non-contemporaneous data' (Rambaut, 2000), have previously been analyzed with linear regression of pairwise distances (Leitner and Albert, 1999) and with maximum likelihood methods (Rambaut, 2000) that assume a constant rate of evolution. When all sequences are isolated at exactly or almost the same time, supplemental information such as fossil data is needed to calibrate the rate of the molecular clock (Hasegawa *et al.*, 1985; Waddell *et al.*, 1999) because, although their product can be estimated, time and rate cannot be inferred separately. In the case of serially sampled data, this external calibration is not essential because the difference in sampling times generates information regarding the evolutionary rate (Rambaut, 2000).

The advantages of serially sampled viral data for divergence time estimation have been nicely illustrated in a recent report on HIV evolution (Korber *et al.*, 2000). In this study, the ancestral date of the HIV-1 M group was inferred to substantially precede the development of oral polio vaccine in the 1950s. The estimates for this ancestral date thereby cast doubt on the hypothesis (Hooper, 1999) that the pandemic HIV outbreak was initiated by administration of oral polio vaccine.

*To whom correspondence should be addressed.

Many tests for the constancy of evolutionary rates are available for contemporaneous data (e.g. Felsenstein, 1981; Wu and Li, 1985; Muse and Weir, 1992; Hasegawa *et al.*, 1993; Tajima, 1993; Takezaki *et al.*, 1995). We note that the majority of tests that purportedly test for constancy of evolutionary rates actually test for rate homogeneity among lineages. Specifically, the hypothesis that rates are constant over time is a special case of the hypothesis that all lineages evolve at a common rate at a given time but that this common rate may change over time. When sequences are all isolated at the same time, these two hypotheses are difficult to distinguish. Here, we show that these hypotheses can be distinguished when sequences are isolated at different times and when the truth is a linear trend over time in the common rate. In a related work, Drummond and Rodrigo (2000) recently presented a distance-based method for evolutionary reconstruction from serially sampled data that allows different time intervals to be associated with different rates of sequence evolution.

There is ample biological evidence that rates of evolution may simultaneously change in several lineages. The antiviral drug AZT can increase the retroviral mutation rate by up to 10-fold (Julias *et al.*, 1997). Likewise, the Manganese cation can accelerate the mutation rate of HIV (Vartanian *et al.*, 1999). A possible acceleration of evolutionary rates of influenza seems to have occurred at around 1992 (Fitch *et al.*, 1997). However, the alternative explanation that the putative rate acceleration is due to intensive sampling of sequences between 1992 and 1996 cannot be excluded. Furthermore, population genetic theory shows that the rate of fixation of slightly advantageous or slightly deleterious mutations depends on the effective population size (Ohta, 1987); this means that a change in population size can affect rates of evolution in all lineages. Therefore, it is very plausible that rates of evolution may simultaneously change in a population or in all viral lineages within a host.

In this paper, experimental design techniques for serially sampled data are employed. These techniques allow an informed choice to be made about which isolates should be sequenced and added to an existing data set in order to test the Molecular Clock Hypothesis (MCH) or estimate divergence times. A few previous studies have applied experimental design techniques to population genetics and phylogenetics. Pluzhnikov and Donnelly (1996) examined the optimal choice of sample size and sequence length for the estimation of genetic diversity. Goldman (1998) quantified the information content of a data set and introduced experimental design techniques to phylogenetics. In contrast to Goldman's work, we focus here on choosing which samples to sequence on the basis of their isolation time. We imagine the situation in which rates and internal node times have already been estimated from a viral data set. Next, one more sample among those isolates available will be sequenced. The question is how to choose the isolation date for the next sample to be sequenced in order to yield a more powerful test of the MCH or in order to improve estimates of evolutionary rates or divergence times.

METHODS

Assumption

We assume that the numbers of evolutionary events along branches follow a Poisson distribution and that this number can be directly observed. In other words, we do not correct for the multiple hits that a single site may experience. We will show that the Poisson distribution is a good approximation to the Jukes–Cantor model for a moderate range of rates and sampling times.

Power to reject the MCH with serially sampled data

In this section, we explore the power to reject the MCH of a constant rate of evolution for two different alternative hypotheses.

Detecting a deterministic trend of the evolutionary rate. First, we compare the null hypothesis of a molecular clock to the simple alternative hypothesis that all lineages evolve at a common rate $r(t)$ but that this common rate changes over time in a deterministic linear fashion. The null and alternative hypotheses cannot be differentiated with contemporaneous data but they can be separated with serially sampled data. Consider the linear trend

$$r(t) = a(t - t_1) + r, \tag{1}$$

where t_1 is the time of the root node. Between time t_i and t_j, the average rate is

$$\overline{r(t)} = r + \tfrac{1}{2}a(t_i + t_j - 2t_1). \tag{2}$$

The likelihood function is

$$L = \prod_{k=1}^{p} \mathrm{e}^{-\bar{r}_k \tau_k N} \frac{(\bar{r}_k \tau_k N)^{x_k}}{x_k!},$$

where τ_k is the time duration of branch k, $\bar{r}_k$ is the mean rate of branch k, x_k is the number of evolutionary events on branch k, N is the sequence length, and p is the number of branches. The null and alternative hypotheses are

$$H_0 : a = 0$$
$$H_1 : a \neq 0.$$

As the number of branches increases, the distribution of $2\Delta \log L = 2 \log \frac{L(\mathbf{X}|\hat{r},\hat{a},\hat{\mathbf{t}})}{L(\mathbf{X}|\hat{\hat{r}},0,\hat{\hat{\mathbf{t}}})}$ tends under H_0 to a χ^2 distribution with 1 degree of freedom (df), and tends under

H_1 to a non-central χ^2 distribution with 1 df and with non-central parameter

$$\lambda = -a^2 \mathbf{V}_{aa}^{-1}, \tag{3}$$

where $\mathbf{V}$ is the inverse of Fisher information matrix (Stuart and Ord, 1999). Here, the single circumflex ($\hat{\ }$) and double circumflex ($\hat{\hat{\ }}$) respectively denote Maximum Likelihood Estimators (MLEs) under H_1 and H_0. The power function of the likelihood ratio test is

$$\begin{aligned} P &= \int_{\chi'^2_{0.05}(1,0)}^{\infty} \mathrm{d}\chi'^2(1,\lambda) \\ &\simeq \int_{\frac{1+\lambda}{1+2\lambda}\chi^2_{0.05}(1)}^{\infty} \mathrm{d}\chi^2\left(1 + \frac{\lambda^2}{1+2\lambda}\right) \end{aligned} \tag{4}$$

where $\chi'^2(1,\lambda)$ is the non-central χ^2 distribution with 1 df and non-central parameter λ and $\chi'^2_{0.05}(1,0)$ is the same as the 95% point of the central χ^2 distribution with 1 df (Stuart and Ord, 1999).

Detecting random fluctuations of the evolutionary rate. If the rate of evolution is constant over time and if sites evolve independently, then the number of evolutionary events that occur on a branch will have a Poisson distribution and therefore will have an identical mean and variance. Furthermore, the expected number of changes that occur on a branch will be linearly proportional to the time duration of the branch. The possibility has been heavily investigated that evolution proceeds so that the variance in the number of evolutionary events among branches with the same time duration exceeds the mean (e.g. Takahata, 1987, 1991; Gillespie, 1991). We can consider two situations that cause this overdispersion. In the first, the rate varies among lineages. In the second, the rate is constant, but the evolutionary process is not Poisson and its variance is higher than its mean. Overdispersion of this second type can be directly integrated into an evolutionary model by explicitly including dependence among sites (e.g. Goldman and Yang, 1994; Muse and Gaut, 1994; Knudsen and Hein, 1999) or its statistical effects can be added without explicitly specifying its biological source (Cutler, 2000). We consider only the first sort of overdispersion in this paper, but the second situation can also be studied with the approach used here.

For a branch k, denote the rate per unit time by r_k and the time interval by τ_k. Although rates are actually likely to be autocorrelated over time, we assume here that the rates of different branches are independent realizations from a gamma distribution. With autocorrelation or with a distribution that differs from a gamma, the recommended sampling design is expected to be qualitatively unchanged. Also, assume there is no rate variation among sites. For a Poisson substitution process and given r_k and τ_k, the probability of observing x_k substitutions in time duration τ_k is

$$p(x_k | r_k, \tau_k) = \mathrm{e}^{(-r_k \tau_k N)} \frac{(r_k \tau_k N)^{x_k}}{x_k!}. \tag{5}$$

If r_k is a random variable from a gamma distribution with Probability Density Function (PDF)

$$g(r|\mu,\delta) = \frac{1}{\Gamma\left(\frac{1}{\delta}\right)(\mu\delta)^{\frac{1}{\delta}}} r^{\frac{1}{\delta}-1} \mathrm{e}^{-\frac{r}{\mu\delta}}, \tag{6}$$

then we can construct the likelihood relating the data to the hyperparameters μ and δ

$$\begin{aligned} & f(x_k|\mu,\delta,\tau_k) \\ & \quad := \int_0^{\infty} p(x = k | r_k, \tau_k) g(r_k|\mu,\delta)\, \mathrm{d}r_k \\ & \quad = \frac{\Gamma\left(x_k + \frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta}\right) x_k!} \left(\frac{1}{\mu\delta\tau_k N + 1}\right)^{\frac{1}{\delta}} \left(\frac{\mu\delta\tau_k N}{\mu\delta\tau_k N + 1}\right)^{x_k}. \end{aligned} \tag{7}$$

This is the PDF of a negative binomial distribution. Its mean and variance are

$$\begin{aligned} E(X) &= \mu\tau_k N \\ \mathrm{Var}(X) &= \mu\tau_k N + \delta(\mu\tau_k N)^2. \end{aligned} \tag{8}$$

With p total branches, the likelihood function is

$$\begin{aligned} & L(\mu,\delta,\tau,|\mathbf{X}) \\ & \quad = \prod_{k=1}^{p} \frac{\Gamma\left(x_k + \frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta}\right) x_k!} \left(\frac{1}{\mu\delta\tau_k N + 1}\right)^{\frac{1}{\delta}} \left(\frac{\mu\delta\tau_k N}{\mu\delta\tau_k N + 1}\right)^{x_k}. \end{aligned} \tag{9}$$

When $\delta = 0$ in equation (8), the negative binomial distribution is reduced to a Poisson distribution. However, as δ gets larger than 0, the variation of rates among branches increases and the evolutionary process departs from a Poisson distribution. Therefore, we can consider the following hypothesis testing problem,

$$\begin{aligned} H_0 &: \delta = 0 \\ H_1 &: \delta \neq 0. \end{aligned}$$

It is important to consider boundaries of the parameter space in a likelihood ratio test (Self and Liang, 1987). When the true parameters are on a boundary, the distributions of estimated parameters can be approximated by a truncated normal distribution with a point mass on the boundary. Therefore, the likelihood ratio distribution is obtained from a mixture of χ^2 distributions. Boundary problems in phylogenetics have been investigated by several authors (e.g. Whelan and Goldman, 1999; Ota *et al.*, 2000). When $\delta = 0$, the gamma distribution (equation 6) is not defined and its PDF is reduced to one point (μ), but the negative binomial distribution (equation 7) is still well defined and is obtained by $\lim_{\delta \to 0} f(x|\mu, \delta)$. In fact, equation (7) is also defined for δ on the set of reciprocals of negative integers, in which case it is reduced to the binomial distribution (Collings and Margolin, 1985; Yanagimoto, 1992). Therefore, $\delta = 0$ is not a boundary point, and the above testing problem is well defined and free from boundary constraint. As in the previous subsection,

$$2\Delta \log L = 2 \log \frac{L(\mathbf{X}|\hat{\mu}, \hat{\delta}, \hat{\boldsymbol{\tau}})}{L(\mathbf{X}|\hat{\hat{\mu}}, 0, \hat{\hat{\boldsymbol{\tau}}})} \sim \chi'^2(1, \lambda) \text{ for large } p, \tag{10}$$

with non-central parameter

$$\lambda = -\delta^2 \mathbf{V}_{\delta\delta}^{-1} = -\delta^2 E\left\{\frac{\partial^2}{\partial \delta^2} \log L(\mu, \delta, \boldsymbol{\tau}|\mathbf{X})\right\}, \tag{11}$$

where the last equality is due to the orthogonality of δ and other parameters. We can determine the power to reject H_0 with equation (4).

Estimation of evolutionary rate and time of origin

When the MCH is not rejected, investigators often estimate rates or divergence times assuming rate constancy. For θ being a column vector representing rates and times of internal nodes, the asymptotic distribution of the MLE ($\hat{\theta}^{\mathrm{MLE}}$) is a normal distribution with mean θ. The asymptotic variance of the MLE is

$$\mathrm{Var}(\hat{\theta}^{MLE}) \longrightarrow \left\{-E\frac{\partial^2}{\partial\theta\partial\theta^T}\log L\right\}^{-1} \text{ as } N \to \infty. \tag{12}$$

Therefore, the relationship between the dispersion of sampling times and the variance of a certain element of the MLE (e.g. Var($\hat{r}$), the variance of the MLE of the rate) can be examined.

The estimated rate is not a function of the estimated root time. Figure 1 shows a tree with three taxa that are sampled at known times T_1, T_2, and T_3 ($T_1 < T_2 < T_3$). If evolutionary events in each branch occur following a Poisson process and the numbers of changes are denoted with x_1, x_2 and x_3, the MLEs are

$$\begin{aligned}\hat{t}_2 &= \frac{T_3 x_2 - T_2 x_3}{x_2 - x_3}\\ \hat{r} &= \frac{x_2 + x_3 - \frac{T_2 x_3}{T_3 - \hat{t}_2} + \frac{T_3 x_3}{T_3 - \hat{t}_2}}{2(T_3 - \hat{t}_2)N}\\ \hat{t}_1 &= \frac{1}{2}\left\{T_1 + \hat{t}_2 - \frac{x_1}{\hat{r}N}\right\},\end{aligned} \tag{13}$$

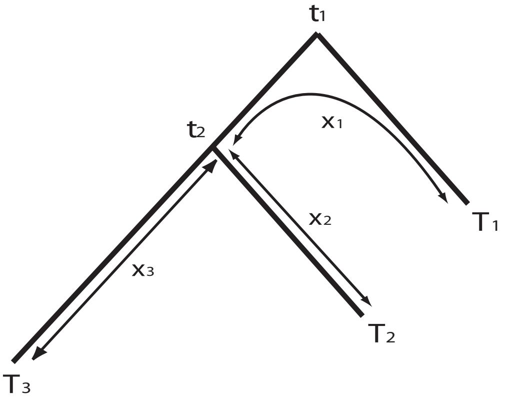

Fig. 1. A three-taxon tree with MLEs that are analytically feasible. Sampling times T_1, T_2 and T_3 are known. The number of evolutionary events x_1, x_2 and x_3 are assumed to be directly observed Poisson random variables. Analytic MLEs are obtained via equation (13).

as long as x_2 is less than x_3. From equation (13), it is clear that $\hat{r}$ is not affected by $\hat{t}_1$. In other words, the parameter $\hat{t}_1$ is estimated only after the estimation of the rate ($\hat{r}$) and the time of the descendant node ($\hat{t}_2$). This can be generalized to cases with more than three taxa and only one outgroup taxon to demonstrate that the outgroup is important in estimating the time of the root but not in estimating the rate. This does not mean that the outgroup is useless in the estimation of the rate. In Figure 1, without T_1, we cannot know x_2 and x_3 separately but can only measure $x_2 + x_3$. The outgroup enables the number of evolutionary events to be separately measured on the different ingroup branches. We note that, even with an accurately estimated rate, the root time is still not guaranteed to be accurately estimated because the estimate of the root time also depends on the number of changes that happen to occur on the branch to the outgroup.

There is a positive probability that the estimated time since the root is infinite. When $x_2 \geqslant x_3$, $\hat{t}_1 = \hat{t}_2 = -\infty$ and $\hat{r} = 0$. Because there is always a finite positive probability that $x_2 \geqslant x_3$, the mean of $\hat{t}_1$ is $-\infty$ and the variance of $\hat{t}_1$ is ∞ even though the probability that $\hat{t}_1 = -\infty$ is very small when there are a large number of taxa. For this reason, we use a 95% Confidence Interval (CI) instead of variance to examine the precision of the estimated date of the origin.

RESULTS AND DISCUSSION

Consider the situation in which some samples have already been sequenced and in which there are many other available samples that could be sequenced. For some reason such as cost, only one more sample will be sequenced. The question will be which of the available sequences should be sequenced. If the goal is to evaluate

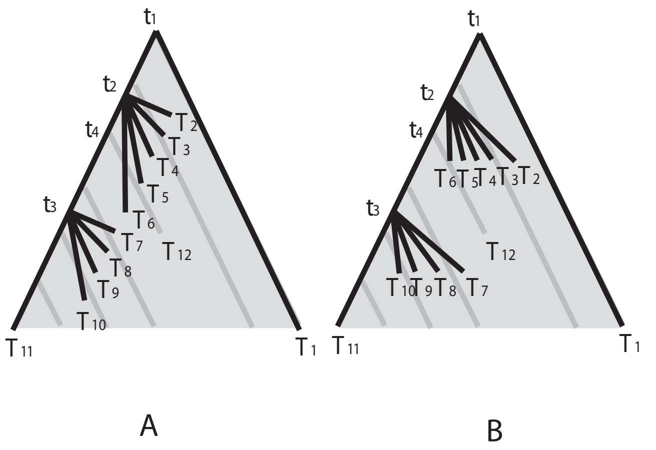

Fig. 2. Trees employed in the numerical analyses and the simulation study. (a) Tree A. (b) Tree B.

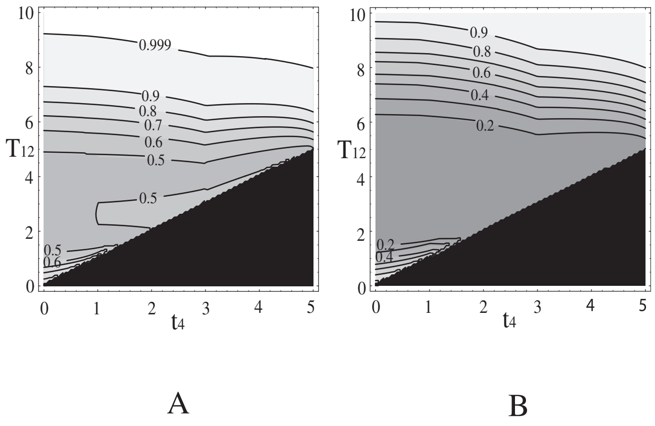

Fig. 3. Power to reject molecular clock with a linear trend over time in rates. The linear trend has slope $a = 0.005$ and y-intercept $r = 0.01$ substitutions per site per unit time (see equation 1). (a) Power for Tree A, as a function of sampling time T_{12} and bifurcation time t_4. (b) Power for Tree B, as a function of sampling time T_{12} and bifurcation time t_4.

the MCH, the preceding power results can assist in selecting the isolation time of the sample to be sequenced. Similarly, the preceding results on estimation error enable an intelligent decision as to selecting the isolation time of the sample to be sequenced when the goal is to estimate divergence dates or evolutionary rates.

We illustrate these experimental design issues with the trees depicted in Figures 2a and b. The trees in Figures 2a and b will be respectively referred to as Trees A and B. Each contains ten ingroup taxa and one outgroup taxon. Furthermore, the internal node times and the sum of the time durations represented by the branches are identical for Tree A and Tree B. To provide a reference, the time of the root node t_1 is set to 0. To lessen any possible effects of dispersion of internal node times, the trees have multifurcating nodes at times $t_2 = 1.0$ and $t_3 = 3.0$. We note that star-like topologies often arise in studies of serially sampled viral data (e.g. Bush *et al.*, 1999). In Tree A, ingroup sequences are sampled every 0.4 time units from T_2 (1.4) to T_{11} (5.0). In contrast, Tree B has five taxa ($T_2, \ldots, T_6$) sampled at time 2.2, four ($T_7, \ldots, T_{10}$) sampled at time 4.0 and two (T_{11} and the outgroup T_1) sampled at 5.0. One additional sample with isolation date T_{12} will be sequenced. For simplicity, we will assume the lineage leading to the sample will join the tree at a node t_4 that is somewhere on the lineage leading from t_1 to T_{11}. How should the date T_{12} of this additional sample be selected? We will show that this choice depends not only on t_4 and T_{12}, but also on the dispersion of sampling times among ingroup taxa. For example, we show below that Tree A provides more power to reject the MCH and more accuracy for estimating evolutionary parameters than does Tree B.

In all scenarios involving Tree A and Tree B, sequence lengths are assumed to be 1000 nucleotides. Also, based on the diverse previous estimates of rates of viral evolution (Fitch *et al.*, 1997; Leitner and Albert, 1999; Rambaut, 2000), one unit time for Tree A or Tree B corresponds to a value somewhere in the range from 1.5 to 30 years.

Asymptotic power to reject the MCH

The asymptotic power to reject the MCH was examined for Tree A and Tree B in two cases: (i) the rate increases linearly over time; and (ii) the rate varies randomly.

Linear trend. By applying equations (3) and (4) to Trees A and B, we can examine the power to detect a linear trend. We observe that the power to detect a trend for Tree A is greater than for Tree B so long as the bifurcation points(t_4s) and the sampling points(T_{12}s) are the same in the two trees (see Figure 3). In both trees, the general pattern is that power is higher when the sampling point T_{12} is soon after or long after the bifurcation point t_4 than when an intermediate amount of time elapses between the bifurcation and sampling points. Figure 3 has two important implications. First, the range of sampling times should be as large as possible to get higher power. Second, the additional sample should predate the oldest sequence sampled (T_2, in this case) or it should have an isolation date that is as recent as possible.

Random fluctuation. By applying equations (4) and (11), we can investigate the power to reject the MCH when the truth is that rates on branches randomly fluctuate according to a gamma distribution. In this case, the power to reject the MCH is not very different between

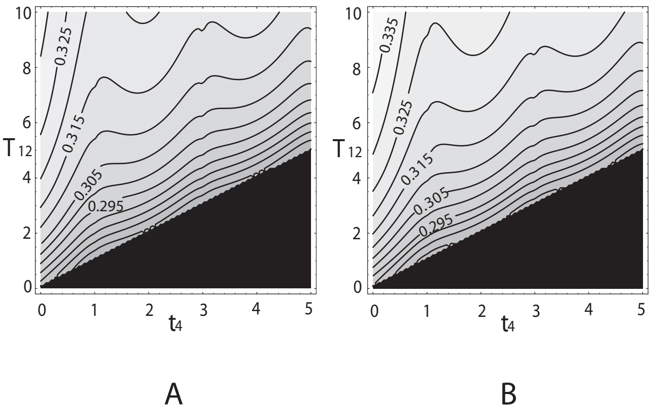

Fig. 4. Power to reject molecular clock when rate fluctuates randomly according to a gamma distribution. The gamma distribution was assigned parameters $\mu = 0.01$ and $\delta = 0.1$ (see equation 6). The coefficient of variation for this gamma distribution is about 0.32 and the lower and upper five percentiles of this distribution are approximately 0.004 796 and 0.017 08. (a) Power for Tree A, as a function of sampling time T_{12} and bifurcation time t_4. (b) Power for Tree B, as a function of sampling time T_{12} and bifurcation time t_4.

Trees A and B for the scenarios that we have explored (see Figure 4). In other words, the power does not seem to be greatly affected by the dispersion of sampling times when the rate varies randomly and independently among the branches. Instead, the power is mainly decided by total branch length. Since Trees A and B have the same branch length, the power is mainly a function of the length of additional branch between t_4 and T_{12}. Irrespective of the position of t_4, the power increases as the length of the branch increases. This is quite reasonable, because as branch length increases, the ratio of the variance to the mean of the negative binomial distribution increases linearly (equation 8).

Estimation of the root time by assuming a clock

The standard error of parameter estimates can be approximated via equation (12). As a whole, the 95% CI for the root time of Tree A is narrower than that for Tree B (see Figure 5). Sampling additional data close to the root and before the oldest already sequenced isolate (T_2) or sampling close to the most recent time is recommended. This is consistent with the results for the power to detect a linear trend. Also, the contour lines are bent when $t_4 = 1.0$ or $t_4 = 3.0$, the times when multifurcations occurred. This shows that an additional sample would improve the accuracy more if it originated near a multifurcating internal node.

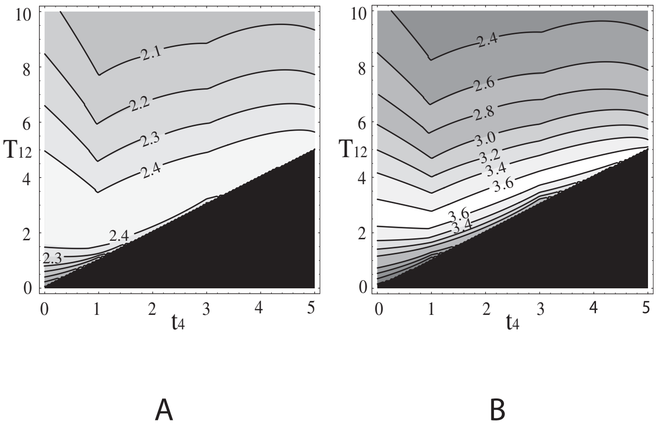

Fig. 5. Width of 95% CI for the root time when the MCH is true. For this example, $r = 0.01$ (see equation 12). (a) Width for Tree A, as a function of sampling time T_{12} and bifurcation time t_4. (b) Width for Tree B, as a function of sampling time T_{12} and bifurcation time t_4.

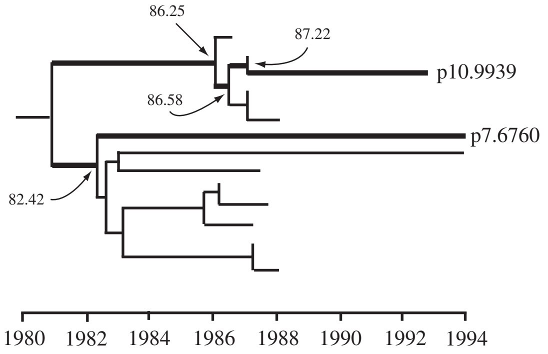

Fig. 6. Simplified and redrawn tree from Figure 1 in Leitner and Albert (1999). Among all the transmission times from Leitner and Albert (1999), only the four that are related to p10.9939 or p7.6760 are shown.

Estimation of the rate of HIV evolution

We applied our approach to the HIV-1 data of Leitner and Albert (1999). They investigated thirteen viral sequences from nine HIV-1 infected individuals whose transmission history is known. Figure 6 is a simplified and redrawn version of the tree from Figure 1 in Leitner and Albert (1999). For the estimated rate of V3 being $r = 0.0067$ substitutions per site per year (Leitner and Albert, 1999) and for the tree shown in Figure 6 along with the transmission history, we calculated the standard error ($\sigma_{\hat{r}}$) of $\hat{r}$ as 0.001 19 using equation (12).

Now, we compare two types of sampling problems: one involves introducing an additional sequence to the branch connecting the root and p7.6760 (Figure 7a) and the other involves adding to the branch connecting the root and

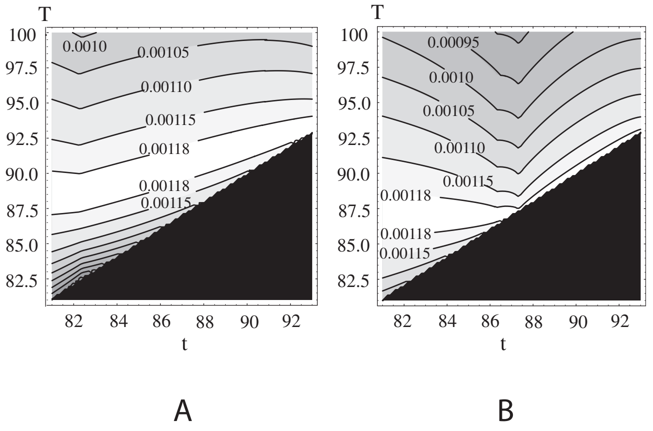

Fig. 7. $\sigma_{\hat{r}}$ when an additional sequence is added to the simplified Leitner and Albert tree. Sequence lengths were set to 300 nucleotides. (a) The additional sequence is sampled at time T and joins the tree at a point t on the branch connecting the root and p7.6760. (b) The additional sequence is sampled at time T and joins the tree at a point t on the branch connecting the root and p10.9939.

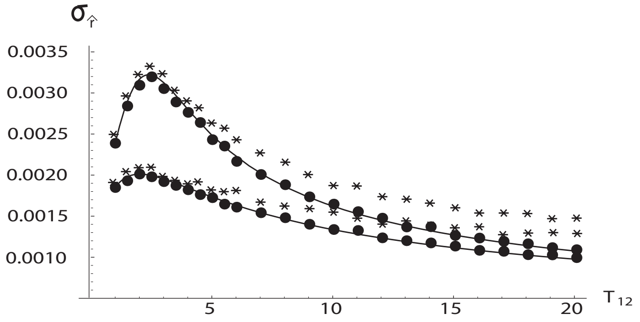

Fig. 8. Examination of approximate standard errors of $\hat{r}(\sigma_{\hat{r}})$. Simulations were performed by setting $t_4 = 0.5$ and by setting $r = 0.02$ substitutions per site per unit time. Simulations were performed for both Trees A and B and for many different values of T_{12} (1, 2, . . . , 19, 20 and 1.5, 2.5, 3.5, 4.5, and 5.5). Solid circles (●) denote estimates from 5000 data sets obtained by sampling Poisson distributions. Asterisks (*) denote estimates from 2000 simulated data sets of sequences. The lower parts correspond to Tree A and the upper parts to Tree B.

p10.9939 (Figure 7b). The time at which the additional sequence is sampled is denoted by T and the time at which it joins the rest of the tree is denoted by t. There is a zone ($\sigma_{\hat{r}} \geqslant 0.001\,18$) in which the additional sequence does not greatly improve the accuracy of estimation. As for the earlier example with Tree A and Tree B, the additional sequence should predate the oldest ingroup sequence or should have been isolated as recently as possible. The contour lines in Figure 7 are bent and show a lower $\sigma_{\hat{r}}$ around $t = 82.3$ (Figure 7a) or around $t = 86.3$ and $t = 87.3$ (Figure 7b). There are several very short internal branches near these time points (we refer to such situations as quasi-multifurcations). If an additional sequence joins the rest of the tree near a quasi-multifurcation, then it is likely to more substantially improve the accuracy of estimation. This is again consistent with the earlier Trees A and B examples.

Checking the effects of assumptions

For simplicity, we have assumed that the numbers of evolutionary events along branches are directly observable. This assumption should have little impact for short branches. When branches are longer and likely to experience multiple events at a site, this assumption may be problematic. To examine its effects, we performed simulations with Trees A and B. Two sets of simulations are considered here. In the first, the number of events on each branch was sampled from a Poisson distribution. In the second, evolution of DNA sequences of length 1000 was simulated according to the Jukes–Cantor model (Jukes and Cantor, 1969) of nucleotide substitution. The first simulation set examines the quality of the asymptotic approximations for r. The second set examines both the asymptotic approximations for r and the effect of multiple substitutions at a site. In the second set of simulations, we used the Tipdate software (Rambaut, 2000) to estimate the evolutionary rate.

The results of the Poisson-based simulations are quite similar to the asymptotic result obtained by equation (12). Although there is more deviation from the asymptotic results in sequence-based simulations than in Poisson-based simulations (see Figure 8), the standard deviation of $\hat{r}$ obtained from sequence-based simulations is close to the asymptotic result when T_{12} is less than 5.0. When T_{12} is larger than 5.0, the standard deviation of $\hat{r}$ obtained by sequence-based simulation is greater than the asymptotic standard deviation. This is probably attributable to multiple substitutions at sites introducing errors in estimating the number of changes that have occurred. Despite the slight differences, the asymptotic results are still useful because the purpose is not to obtain the exact value of the variance or the CI of unknown parameters but to select additional samples to be sequenced.

General experimental design for the study of viral evolution

Usually, we do not know the times of internal nodes or the true evolutionary rates. In these cases, we suggest using inferred rates and times to determine the optimal isolation time of an additional sequence. After adding an additional taxon, one strategy would be to use re-estimated rates and times when finding the next optimal sampling time. It would be interesting to compare this strategy to one that simultaneously chooses several additional samples to be sequenced.

It is important to construct an experimental design which enables estimation of rates and times with high precision. Although viral samples can be obtained that are several decades old (Taubenberger *et al.*, 1997) or even more than one hundred years old (Fraile *et al.*, 1997), the source of old virus is limited. Given an outgroup, ingroup sequences from at least two different time points are necessary for the estimation of the rate under the MCH. When there is a deterministic trend in evolutionary rate, a pair of distances from two sequences to a common outgroup gives an estimate of mean evolutionary rate between the times at which the sequences are sampled. Testing the MCH against a deterministic trend requires comparison between multiple rate estimates for different time intervals. Because each of the rates needs two sampling times in the ingroup, at least three sampling times are necessary for this test.

Although further research is warranted, regression analysis may provide a useful analogy for interpreting the numerical results in this paper. In experimental design for simple regression analysis, the slope can be estimated with high precision when the explanatory variable has a large variance. In this case, sequence isolation times are analogous to the explanatory variable. Sampling from either close to the root or from the latest available time is recommended, because this increases the range of isolation times and thereby increases the variance of the times at which rate is estimated. As a consequence, the variance of the estimated slope for the trend in rates is reduced. The power to reject the MCH against the alternative hypothesis of a linear trend in rates is not solely a function of the range of sequence isolation times or the variance of isolation times. The power is also influenced by the distribution of lengths of intervals that separate isolation times. A short interval naturally provides less information about the average rate during that interval than does a long interval. The exact relationship between isolation times and power is a complicated function of the phylogenetic tree structure and the evolutionary process. However, we emphasize that a large range of time between the earliest and latest isolation dates is desirable. This large range is similarly desirable when the goal is to estimate the evolutionary rate under the MCH.

Recent findings regarding the viral adaptation process indicate that more refined methods for statistical characterization of evolutionary rates are warranted. For example, Shankarappa *et al.* (1999) studied the evolution of the C2–V5 region of the HIV-1 *env* gene and observed linearly increasing divergence followed by stabilization. This stabilization occurred a mean of 0.46 years earlier than the failure of T-cell homeostasis. The result implies a somewhat constant evolutionary rate during the early and intermediate phase of an asymptomatic infection period followed by a decreased rate afterwards. To examine this viral adaptation process and its variability among hosts, it would be valuable to estimate the change point of evolutionary rates and its variability among hosts. When obtaining a large sample of sequences from a host, the experimental design of sampling times becomes crucial. With rough prior information on the minimum value s_0 of a change point that a host might have, the rate before s_0 would be reliably estimated by concentrating the sampling times soon after seroconversion and just before s_0. Intensive sampling after the time s_0 would be needed to reliably detect the position of the change point.

In the framework considered here, changes in the evolutionary rate are assumed to affect the entire sequence. However, biological phenomena such as an alteration of gene function can have more localized effects. In fact, changes in evolutionary rates at specific sites can be exploited to detect those sites important for gene function (Gu, 2001). Other localized phenomena, including unusually high ratios of nonsynonymous to synonymous substitution (Yang *et al.*, 2000), can be used to identify positions that may be especially important for protein function. Although it has not been a focus here, we emphasize that experimental design techniques are also relevant when the goal is to study site-specific patterns of evolution. Because viruses tend to be quickly evolving, the ability to collect noncontemporaneous sequence data means that studies of viral evolution are more amenable to thoughtful experimental design than are studies of evolution in other biological systems.

ACKNOWLEDGEMENTS

We thank A.Rambaut for his Tipdate 1.0 source code, T.Leitner for offering detailed information about his data, T.Yanagimoto for technical advice, and two anonymous reviewers. T-K.S. and H.K. were supported by JSPS grants 12554037 and 13308013, M.H. and H.K. by JSPS grant BSAR-497, and J.L.T. by NSF grants DBI-0077503 and INT-990934. The first author was supported by the Japanese government scholorships program for foreign students.

REFERENCES

Bush,R.M., Bender,C.A., Subbarao,K., Cox,N.J. and Fitch,W.M. (1999) Predicting the evolution of human influenza A. *Science*, **286**, 1921–1925.

Collings,B.J. and Margolin,B.H. (1985) Testing goodness of fit for the Poisson assumption when observations are not identically distributed. *J. Am. Statist. Assoc.*, **80**, 411–418.

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

Cutler,D.J. (2000) Estimating divergence times in the presence of an overdispersed molecular clock. *Mol. Biol. Evol.*, **17**, 1647–1660.

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

Fitch,W.M., Bush,R.M., Bender,C.A. and Cox,N.J. (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl Acad. Sci. USA*, **94**, 7712–7718.

Fraile,A., Escriu,F., Aranda,M.A., Malpica,J.M., Gibbs,A.J. and Garacia-Arenal,F. (1997) A century of tobamovirus evolution in an Australian population of *Nicotiana glauca*. *J. Virol.*, **71**, 8316–8320.

Gillespie,J.H. (1991) *The Causes of Molecular Evolution*, Chapter 3, Oxford University Press, New York.

Goldman,N. (1998) Phylogenetic information and experimental design in molecular systematics. *Proc. R. Soc. Lond.* B, **265**, 1779–1786.

Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.

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

Hasegawa,M., Kishino,H. and Yano,T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.

Hasegawa,M., Rienzo,A.D., Kocher,T.D. and Wilson,A.C. (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. *J. Mol. Evol.*, **37**, 347–354.

Hooper,E. (1999) *The River: A Journey to the Source of HIV and AIDS*. Little Brown, Boston, pp. 1070.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–132.

Julias,J.G., Kim,T., Arnold,G. and Pathak,V.K. (1997) The antiretrovirus drug 3′-azido-3′-deoxythymidine increases the retrovirus mutation rate. *J. Virol.*, **71**, 4254–4263.

Knudsen,B. and Hein,J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.

Korber,B., Muldoon,M., Theiler,J., Gao,F., Gupta,R., Lapedes,A., Hahn,B.H., Wolinsky,S. and Bhattacharya,T. (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science*, **288**, 1789–1796.

Leitner,T. and Albert,J. (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc. Natl Acad. Sci. USA*, **96**, 10 752–10 757.

Mansky,L.M. (1998) Retrovirus mutation rates and their role in genetic variation. *J. Gen. Virol.*, **79**, 1337–1345.

Muse,S.V. and Weir,B.S. (1992) Testing for equality of evolutionary rates. *Genetics*, **132**, 269–276.

Muse,S.V. and Gaut,B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, **11**, 715–724.

Ohta,T. (1987) Very slightly deleterious mutations and the molecular clock. *J. Mol. Evol.*, **26**, 1–6.

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

Pluzhnikov,A. and Donnelly,P. (1996) Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics*, **144**, 1247–1262.

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

Self,S.G. and Liang,K-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Statist. Assoc.*, **82**, 605–609.

Shankarappa,R., Margolick,J.B., Gange,S.J., Rodrigo,A.G., Upchurch,D., Farzadegan,H., Gupta,P., Rinaldo,C.R., Learn,G.H., He,X., Huang,X.L. and Mullins,J.I. (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.*, **73**, 10 489–10 502.

Stuart,A. and Ord,J.K. (1999) *Kendall's Advanced Theory of Statistics*, vol. 2, Arnold, London, pp. 245–249.

Suzuki,Y. and Gojobori,T. (1997) The origin and evolution of Ebola and Marburg viruses. *Mol. Biol. Evol.*, **14**, 800–806.

Tajima,F. (1993) Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, **135**, 599–607.

Takahata,N. (1987) On the overdispersed molecular clock. *Genetics*, **116**, 169–179.

Takahata,N. (1991) Statistical models of the overdispersed molecular clock. *Theor. Pop. Biol.*, **39**, 329–344.

Takezaki,N., Rzhetsky,A. and Nei,M. (1995) Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.*, **12**, 823–833.

Taubenberger,J.K., Reid,A.H., Krafft,A.E., Bijwaard,K.E. and Fanning,T.G. (1997) Initial genetic characterization of the 1918 'Spanish' influenza virus. *Science*, **275**, 1793–1796.

Vartanian,J-P., Sala,M., Henry,M., Wain-Hobson,S. and Meyerhans,A. (1999) Manganese cations increase the mutation rate of human immunodeficiency virus type 1 *ex vivo*. *J. Gen. Virol.*, **80**, 1983–1986.

Waddell,P.J., Cao,Y., Hasegawa,M. and Mindell,D.P. (1999) Assessing the Cretaceous superordinal divergence times within birds and placental mammals by using whole mitochondrial protein sequences and an extended statistical framework. *Syst. Biol.*, **48**, 119–137.

Whelan,S. and Goldman,N. (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.*, **16**, 1292–1299.

Wu,C.-I. and Li,W.-H. (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl Acad. Sci. USA*, **82**, 1741–1745.

Yanagimoto,T. (1992) The Mantel–Haenszel statistics for the extended odds ratio in the negative binomial distribution. *J. Japan Statist. Soc.*, **22**, 7–17.

Yang,Z., Nielsen,R., Goldman,N. and Pedersen,A-M.K. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.