

Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot

Korbinian Strimmer and Oliver G. Pybus

Department of Zoology, University of Oxford

We present an intuitive visual framework, the generalized skyline plot, to explore the demographic history of sampled DNA sequences. This approach is based on a genealogy inferred from the sequences and provides a nonparametric estimate of effective population size through time. In contrast to previous related procedures, the generalized skyline plot is more applicable to cases where the underlying tree is not fully resolved and the data is not highly variable. This is achieved by the grouping of adjacent coalescent intervals. We employ a small-sample Akaike information criterion to objectively choose the optimal grouping strategy. We investigate the performance of our approach using simulation and subsequently apply it to HIV-1 sequences from central Africa and mtDNA sequences from red pandas.

Introduction

Contemporary DNA sequences contain information about the demographic history of the population from which they were sampled. As a result, the inference of demographic parameters from genetic data has become an important topic in statistical genetics, with applications in fields as diverse as anthropology, conservation biology, epidemiology, and virology (Harvey et al. 1996). Estimation of effective population size, as well as its rate of change through time, can provide useful information about the evolutionary and demographic history of a population.

Methods for estimating demographic history from gene sequences are mostly based on coalescent theory (Kingman 1982*a*, 1982*b*; Hudson 1990; Nordborg 2001). They usually rely on a simple parametric model $N(t)$ which describes effective population size through time. Time t is zero at present and increases into the past, hence $N(0)$ is the effective population size at present. Two simple demographic models are frequently used: constant population size $N(t) = N(0)$, with one parameter $N(0)$ and exponential growth $N(t) = N(0)e^{-rt}$, with two parameters r and $N(0)$. Often, however, there is no prior reason to assume a specific model of demographic history for the data in question. Moreover, the available models may be too simplistic. Hence, nonparametric and model selection tools can play a useful role in the inference of population history from gene sequence data.

Nee et al. (1995) proposed the lineage through time (LTT) plot to graphically investigate the demographic history of gene sequences. LTT plots display the rate of coalescence through time in a genealogy which has been reconstructed from an alignment of homologous sequences. Pybus, Rambaut, and Harvey (2000) described a simple transformation that converts this rate of coalescence into a plot of estimated effective population size against time, which we call here the classic skyline

plot. The LTT and classic skyline plot approaches are closely related and both assume that a fully resolved phylogeny with reliable estimates of divergence times is available. As a consequence, these approaches can only be applied to data that exhibit a strong phylogenetic signal and are not appropriate for alignments which contain identical sequences. In addition, neither method provides an assessment of coalescent error. This is the error that results from the randomness inherent in the coalescent process.

In this paper we introduce the generalized skyline plot, a simple framework for exploring the demographic signal in a sample of DNA sequences. This method extends the classic skyline plot by allowing multiple coalescent events (for which little divergence time information is available) to be grouped together. The classic plot is a special case of the generalized plot, which arises when no coalescent events are grouped. The generalized plot can be applied to data sets which contain identical sequences and has the added benefit of smoothing the classic plot, which typically displays stochastic noise. We show that the most appropriate amount of smoothing can be determined by using a penalized likelihood approach. Furthermore, we derive the skyline plot as a simple method of moments estimator based on standard coalescence distributions, which enables us to compute estimates of the coalescent error. To illustrate our approach, we analyze HIV-1 sequences from central Africa and investigate the demographic history of red pandas using mtDNA sequences.

Methods

The coalescent describes the relationship between the shape of an intrapopulation genealogy (representing the ancestry of randomly-sampled, nonrecombining, neutrally evolving sequences) and the demographic history of the sampled population (Kingman 1982*a*, 1982*b*). The coalescent process arises as an approximation to a general class of population genetics models (including the Wright-Fisher reproduction model) and is valid when the effective population size is large. In the coalescent model, the sequences sampled at present are traced back in time to a single common ancestor, with coalescent events among lineages occurring according

Key words: coalescent process, corrected Akaike criterion, HIV-1, model selection, likelihood, red panda, skyline plot.

Address for correspondence and reprints: Oliver G. Pybus, South Parks Road, Oxford, OX1 3PS, UK. E-mail: oliver.pybus@zoo.ox.ac.uk.

Mol. Biol. Evol. 18(12):2298–2305. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

to a nonhomogeneous Poisson process. For a constant effective population size N the rate of coalescence is

$$\lambda_n = \frac{\binom{n}{2}}{N},$$

where n is the number of lineages before the coalescence and where time is measured in units of substitutions per site. Note that the rate λ_n changes after each coalescent event. Thus, the waiting time w_n until the next coalescent event is exponentially distributed according to

$$\Pr(w_n) = \lambda_n e^{-\lambda_n w_n}, \tag{1}$$

with expectation

$$E(w_n) = \frac{1}{\lambda_n} = \frac{N}{\binom{n}{2}} \tag{2}$$

and variance

$$\text{var}(w_n) = \frac{1}{\lambda_n^2} = \frac{N^2}{\binom{n}{2}^2}. \tag{3}$$

The accumulated waiting time $w_{n,k} = \sum_{i=1}^k w_{n-i+1}$ until k coalescent events have occurred is the sum of k different exponential variables ($\lambda_i \neq \lambda_j$ for $i \neq j$) and thus follows a hypo-exponential distribution (e.g., Ross 1997)

$$\Pr(w_{n,k}) = \sum_{i=1}^k c_{i,k} \lambda_{n-i+1} e^{-\lambda_{n-i+1} w_{n,k}}, \tag{4}$$

where $c_{i,k} = \prod_{j=1; j \neq i}^k \lambda_{n-j+1} / (\lambda_{n-j+1} - \lambda_{n-i+1})$. This distribution has expectation

$$E(w_{n,k}) = \sum_{i=1}^k E(w_{n-i+1}) = N \frac{2k}{n(n-k)} \tag{5}$$

and variance

$$\begin{aligned} \text{var}(w_{n,k}) &= \sum_{i=1}^k \text{var}(w_{n-i+1}) \\ &= N^2 \sum_{i=1}^k \binom{n-i+1}{2}^{-2} \end{aligned} \tag{6}$$

Deterministic changes of N through time can be introduced in the coalescent by a nonlinear scaling factor (Hudson 1990; Griffith and Tavaré 1994; Donnelly and Tavaré 1995; Kuhner, Yamato, and Felsenstein 1998). If selection, recombination, or noncontemporary sequences are present then further adjustments to the coalescent are necessary (e.g., Rodrigo and Felsenstein 1999; Nordborg 2001).

The Classic Skyline Plot

Suppose that we have a fully resolved genealogy \hat{G} with m tips, estimated from a given sequence align-

ment in such a way that \hat{G} 's internal nodes are dated according to a given time scale. This requires a molecular clock, or more generally, a model of rate correlation among different branches in the tree (Gillespie 1991; Sanderson 1997; Thorne, Kishino, and Painter 1998; Huelsenbeck, Larget, and Swofford 2000). \hat{G} defines $m - 1$ ordered internode intervals I_m, I_{m-1}, \dots, I_2 where the subscript indicates the number of lineages present during each interval. The length of interval I_n is denoted by \hat{w}_n . A simple demographic model can then be constructed as follows. During each interval I_n we assume that population size is a local constant, M_n , but between different intervals the population size is allowed to change. Hence, for a set of $m - 1$ intervals, we approximate the demographic history $N(t)$ by a piecewise constant function with $m - 1$ independent variables M_m, M_{m-1}, \dots, M_2 .

A method of moments estimator for the population size during each interval I_n is then constructed by setting the expected waiting time (eq. 2) for the next coalescent event equal to \hat{w}_n , and solving the resulting equation for M_n . This gives the classic skyline plot estimate

$$\hat{M}_n = \hat{w}_n \frac{n(n-1)}{2} \tag{7}$$

for the population size during time interval I_n . Pybus, Rambaut, and Harvey (2000) derived this simple result using an alternative argument based on the variable population size coalescent.

The Generalized Skyline Plot

Generally, we expect the accuracy of the observed intervals \hat{w} (obtained from a reconstructed genealogy) to be adversely affected by limited genetic variation. The number of substitutions occurring in an internode interval is often modeled by a Poisson distribution. Consequently, the observed number of substitutions is proportional to the time elapsed when either the substitution rate or the internode interval is large. However, this approximation breaks down when the product of interval length and substitution rate is small. Under such circumstances it would be beneficial to pool small intervals together so that all intervals are large enough for time to be proportional to the number of substitutions. Zero-length intervals always occur if the alignment contains identical sequences, and also arise when the branch lengths of a genealogy are estimated using maximum likelihood under a molecular clock. The disadvantage of pooling intervals is that some (but not all) of the temporal structure in the data is lost. When the sequences contain very little or no genetic variation, a Bayesian approach employing prior distributions for the substitution and coalescent parameters is required (Tavaré et al. 1997). However, in these cases a single-tree estimator such as the skyline plot is inappropriate.

Allowing pooled intervals in the skyline plot leads to the derivation of the generalized skyline plot. Consider a composite time interval $I_{n,k}$ where n denotes the number of lineages at the start of the interval, and k is the total number of coalescent events taking place dur-

ing this interval. $I_{n,k}$ has observed length $\hat{w}_{n,k} = \hat{w}_n + \hat{w}_{n-1} + \dots + \hat{w}_{n-k+1}$. If we assume a locally constant population size $M_{n,k}$ during this composite interval we can construct a method of moments estimator for $M_{n,k}$ using equation (5), and arrive at

$$\hat{M}_{n,k} = \hat{w}_{n,k} \frac{n(n-k)}{2k}. \tag{8}$$

Note that the generalized skyline plot (eq. 8) contains the classic skyline plot (eq. 7) as a special case when each interval contains only a single coalescent event ($k = 1$). If there is only a single composite interval $I_{m,m-1}$ that contains all $m - 1$ coalescent events in the genealogy, then equation (8) collapses to $\hat{M}_{m,m-1} = \hat{w}_{m,m-1}m/[2(m-1)]$. This is the standard population genetic relationship between effective population size and the time to the most recent common ancestor of a sample of size m .

Grouping Intervals and Model Selection

In order to choose which intervals in genealogy \hat{G} should be pooled we adopt the following convention. First, the set of standard internode intervals I_m, I_{m-1}, \dots, I_2 is determined from \hat{G} . Next, if an interval is smaller than a certain threshold ϵ then the interval is considered as small. Proceeding from I_m to I_2 , each small interval is pooled with the neighboring interval closer to the root. If the neighboring interval is also small, then pooling continues until the composite interval is larger than ϵ . Note that this approach prevents the occurrence of zero-length intervals at present. Thus ϵ determines how much temporal structure in the data is retained and hence controls the degree to which the skyline plot is smoothed. The choice of ϵ is guided by two opposing objectives. On the one hand, ϵ should be large enough to remove the noise in the data which arises from the randomness of the mutational process. On the other hand, ϵ should be small enough to preserve the actual demographic signal in the data.

How should the most appropriate value of ϵ be chosen? Visual inspection of skyline plots calculated under various ϵ values is helpful, but an objective approach based on statistical model selection would be preferable. Here we outline one possible approach which penalizes skyline plots that overfit the data. As skyline plots represent specific hypotheses of demographic history, we can calculate the likelihood of a skyline plot using standard approaches, given the observed internode interval lengths (Griffith and Tavaré 1994; Pybus, Rambaut, and Harvey 2000). For a skyline plot derived from a genealogy with m sequences the log-likelihood $\log L$ reduces to

$$\log L = \sum_{i=2}^m \log \frac{\binom{i}{2}}{\hat{M}} - \frac{\binom{i}{2}}{\hat{M}} \hat{w}_i. \tag{9}$$

Note that the estimated population size \hat{M} for any subinterval in a composite interval $I_{n,k}$ is $\hat{M}_{n,k}$. Now let

K be the number of inferred parameters (=number of composite intervals in the skyline plot) and let $S = m - 1$ be the sample size (=number of coalescent events in the genealogy). We can compare skyline plots with different ϵ values by penalizing the log-likelihood of each plot using the AIC_c correction

$$\log L_{AIC_c} = \log L - K - \frac{K(K+1)}{S-K-1}. \tag{10}$$

(Hurvich and Tsai 1989). The AIC_c approach is a second-order extension of Akaike's well-known first-order AIC correction, $\log L_{AIC} = \log L - K$ (Akaike 1974). However, Akaike's AIC is valid only for large samples with $S/K > 40$, whereas AIC_c is also valid for small sample sizes (Burnham and Anderson 1998). As K depends on ϵ , we can use equation (10) to obtain an optimal generalized skyline plot, by choosing the value of ϵ which maximizes $\log L_{AIC_c}$.

Statistical Properties and Simulations

Here, we investigate the statistical properties of the skyline plot and study its performance using sequence data simulated under known demographic scenarios.

First, we analytically calculate the coalescent variance σ_c^2 of the skyline plot. For the classic skyline plot we use equation (3) and obtain

$$\sigma_c^2(\hat{M}_n) = \binom{n}{2} \sigma_c^2(\hat{w}_n) = \hat{M}_n^2. \tag{11}$$

The coalescent variance for the generalized skyline plot can be computed similarly,

$$\begin{aligned} \sigma_c^2(\hat{M}_{n,k}) &= \frac{\sigma_c^2(\hat{w}_{n,k})}{\left(\sum_{i=1}^k 1 / \binom{n-i+1}{2}\right)^2} \\ &= \hat{M}_{n,k}^2 \frac{\sum_{i=1}^k 1 / \binom{n-i+1}{2}}{\left(\frac{2k}{n(n-k)}\right)^2}, \end{aligned} \tag{12}$$

using the variance of the hypo-exponential distribution (eq. 6). The last factor in equation (12) equals 1 if $k = 1$ (i.e., the classic skyline plot result) and is smaller than 1 otherwise. Note that the coalescent error of the skyline plot is large. This is probably due to the nonparametric nature of the plot, that is, the sample size S is small with respect to the number of parameters K . The variance of the generalized skyline plot (eq. 12) becomes smaller as more intervals are pooled because the ratio of data points to parameters increases.

To investigate the bias of the skyline plot we conducted a small simulation study. For various settings of m and k (see table 1), simulations were performed as follows: (1) 1,000 genealogies with m tips were simulated using the demographic model $N(t) = 0.1$, (2) The first k internode intervals were grouped together and the

Table 1
Bias and Variance of the Generalized Skyline Plot

m	k	$E(\hat{M})$	$b(\hat{M})$	$\text{var}(\hat{M})$	$\sigma_c^2(\hat{M})$
10	1	0.09753	-0.002473	0.009993	0.009511
10	2	0.10210	0.002101	0.005184	0.005277
10	5	0.09911	-0.000895	0.002318	0.002272
30	1	0.10201	0.002010	0.010848	0.010406
30	10	0.10090	0.009029	0.001049	0.001074
30	20	0.10081	0.000809	0.000739	0.000732
50	1	0.09782	-0.002184	0.009751	0.009568
50	10	0.10074	0.000735	0.000989	0.001031
50	20	0.09958	-0.000424	0.000540	0.000540

skyline plot estimate \hat{M} was calculated using equation (8) for each of the 1,000 simulated gene trees, and (3) The expectation $E(\hat{M})$ and the bias $b(\hat{M}) = E(\hat{M}) - M$ were computed along with the observed variance $\text{var}(\hat{M})$ and the theoretical variance $\sigma_c^2(\hat{M})$.

The results are summarized in table 1. They indicate that the generalized skyline plot is an unbiased estimator of the effective population size during an interval $I_{n,k}$ (when the coalescent intervals \hat{v} are known without error). As expected from the earlier analytical results, the variance of this estimate is large but declines quickly when intervals are pooled ($k > 1$). Note that the

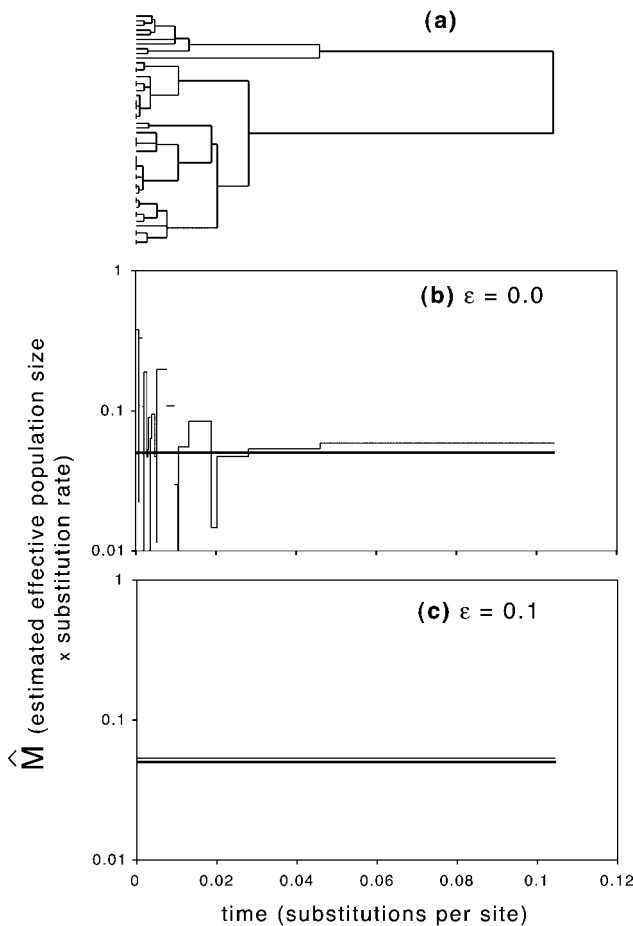


FIG. 1.—Skyline plots for DNA sequences simulated under constant population size: *a*, estimated tree; *b*, classic skyline plot ($\epsilon = 0$); and *c*, generalized skyline plot (AIC_c estimate of $\epsilon = 0.1$). The thick line shows the true demographic history.

skyline plot (and the above simulation) assumes that the effective population size is locally constant during an interval. If the population size changes within an interval then the skyline plot (as a piecewise constant estimator of $N(t)$) is, by definition, biased. However, in this case the classic skyline plot provides an estimate of the harmonic mean of $N(t)$ during each interval (Pybus, Rambaut, and Harvey 2000).

Next, we studied the performance of the classic and generalized plots using sequence data simulated under known demographic scenarios. The purpose of these simulations was to determine whether the generalized plot is more reliable than the classic plot when the DNA sequences used are not highly variable. The simulations were performed as follows: (1) Expected coalescent trees, which contain no coalescent error, were obtained under two demographic models, $N(t) = 0.05$ (constant) and $N(t) = e^{-1000t}$ (exponential). These models were chosen to approximately represent the history of animal mtDNA sequences. Note that time is measured in substitutions per site, (2) Sequences were simulated down these trees using the HKY (Hasegawa, Kishino, and Yano 1985) model (transition-transversion ratio = 10; nucleotide frequencies $\pi_A = 0.3$, $\pi_C = 0.25$, $\pi_G = 0.15$, and $\pi_T = 0.3$) and no rate heterogeneity. The constant-model alignment contained 500 bp and the exponential-model alignment contained 1,500 bp, (3) Genealogies were estimated from the simulated sequences using the TBR search heuristic in PAUP* (Swofford 1998). The substitution model specified earlier was used, and (4) Classic and generalized skyline plots were obtained from the estimated genealogies. The ϵ value was found by optimizing the AIC_c corrected log-likelihood (see eq. 10).

Figures 1 and 2 show the simulation results for the constant and exponential models, respectively. Under the constant-size model, many of the simulated sequences are not unique and many of the internode intervals in the estimated tree are very small (fig. 1*a*). Thus the number of observed substitutions provides little information about the true coalescent interval lengths and consequently the classic skyline plot is very noisy (fig. 1*b*). In contrast, the generalized skyline plot estimate is smooth and almost identical to the true demographic history (fig. 1*c*). The optimal ϵ was 0.1, which resulted in all the observed intervals being pooled into a single composite interval. This should be expected, as the true demographic history contains no changes in population size.

Under the exponential model, only two sequences were identical (fig. 2*a*) and both the classic and generalized plots provide a good estimate of the true demographic history, although the generalized plot is less noisy (fig. 2*b* and *c*). The optimal ϵ was 0.00115. Interestingly, both plots appear to slightly overestimate population size in the past, which suggests that, in this set of sequences, the estimated branch lengths near the root of the genealogy are too long.

Results

We illustrate our framework by analyzing two previously published data sets. We investigate the demo-

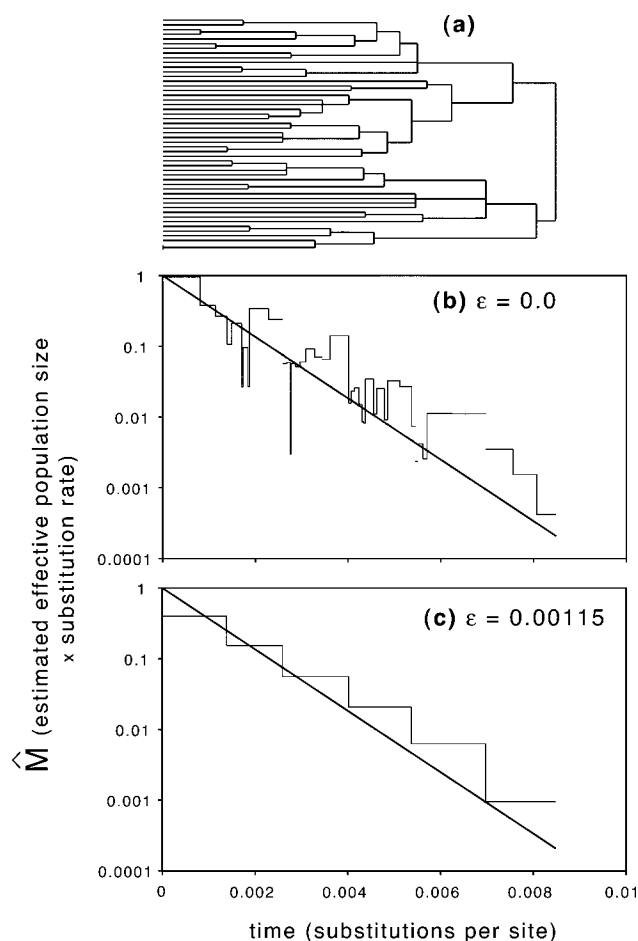


FIG. 2.—Skyline plots for DNA sequences simulated assuming an exponentially growing population: *a*, estimated tree; *b*, classic skyline plot ($\epsilon = 0$); and *c*, generalized skyline plot (AIC_c estimate of $\epsilon = 0.00115$). The thick line shows the true demographic history.

graphic history of HIV-1 using sequences sampled from Central Africa, and we also analyze mtDNA sequences from red pandas (*Ailurus fulgens*). These examples were chosen for two reasons. First, these data sets have been previously studied using other coalescent methods, so alternative results are available for comparison. Second, the genealogies inferred from these sequences contain a number of short or zero-length branches, which allow us to compare the performance of the generalized and classic plots.

HIV-1 in Central Africa

HIV-1 group M contains the viruses which cause the global HIV pandemic and appears to have arisen in Central Africa during the last 100 years. Vidal et al. (2000) investigated the genetic diversity of HIV-1 group M in this region by obtaining viral gene sequences (*env* gene, V3-V5) in 1997 from 197 infected individuals living in the Democratic Republic of Congo. Yusim et al. (2001) used a customized maximum likelihood approach to estimate a phylogeny for this large data set, and it is this phylogeny which we use here (fig. 3*a*). Detailed interpretation of the HIV tree and further analysis of this

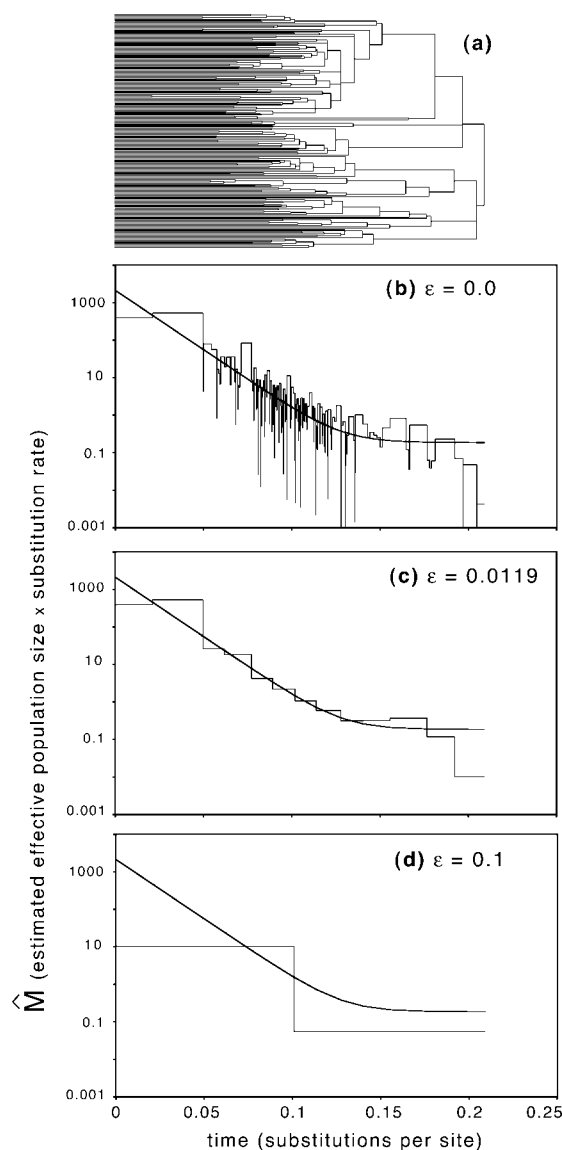


FIG. 3.—The HIV-1 data set. *a*, The estimated genealogy. *b–d*, Generalized skyline plots. *b*, $\epsilon = 0.0$ (the classic skyline plot); and *c*, $\epsilon = 0.0119$ (the AIC_c estimate); *d*, $\epsilon = 0.1$. The thick black curve on each plot is a maximum likelihood parametric estimate obtained from the same data (see text).

data set can be found in Rambaut et al. (2001) and Yusim et al. (2001).

The classic skyline plot for the tree of Yusim et al. (2001) is shown in figure 3*b*. This plot corresponds to the case where $\epsilon = 0$. The tree contains many internode intervals which are zero or near zero in length. Consequently, the plot contains gaps (where the estimated effective population size \hat{M}_n is zero) and spikes (where \hat{M}_n is close to zero). Figure 3*c* and *d* show other generalized plots for the same tree. As ϵ is increased, the generalized plot becomes less noisy than the classic plot, but also becomes less finely resolved. If ϵ is very large then too many intervals are grouped and, as a result, information about demographic history is lost (see fig. 3*d*).

The thick curves in figure 3*b–d* show a maximum likelihood estimate of population size obtained from the

HIV tree using a specific parametric model, $N(t) = N(0)(\alpha + [1 - \alpha]e^{-rt})$, called the expansion model. The parameters of this model were estimated using maximum likelihood (see Yusim et al. 2001). Figure 3c shows the generalized plot with the highest AIC_c value ($\epsilon = 0.0119$). This plot is neither noisy nor oversimplified, and corresponds closely to the maximum likelihood parametric estimate.

We note that it is very unlikely that this HIV-1 data set has been evolving according to the molecular clock and without recombination. Therefore, statistical estimates of population parameters from these data based on the standard neutral coalescent model must be treated with caution. The quantitative effects of recombination on coalescent-based estimates of demographic history have yet to be determined.

Red Pandas in Southwestern China

The red panda, which inhabits southwestern China, is an endangered species. To investigate the genetic diversity of this species, Su et al. (2001) obtained a data set of 53 homologous sequences, 250 bp in length, from the 5' end of the mtDNA control region. The alignment contains only 25 haplotypes, and thus many sequences are identical. We estimated a genealogy for these sequences by maximum likelihood, using the TBR search heuristic in PAUP (Swofford 1998). The HKY substitution model was used (estimated transition-transversion ratio = 36.5; nucleotide frequencies $\pi_A = 0.28$, $\pi_C = 0.26$, $\pi_G = 0.14$, and $\pi_T = 0.32$) under the assumption of a molecular clock. Clock-like evolution could not be rejected using a likelihood ratio test (Felsenstein 1981).

Figure 4 shows the classic skyline plot and the optimal generalized skyline plot (AIC_c estimate of $\epsilon = 0.0008$) obtained from the panda mtDNA genealogy. The generalized skyline plot (fig. 4b) suggests that the effective population size of red pandas has followed a logistic growth. Su et al. (2001) analyzed the same data using pairwise difference distributions and concluded that the red pandas had undergone recent population growth. In contrast, figure 4c suggests an approximately constant population size at present, with growth in the distant past. Pairwise difference distributions do not explicitly incorporate phylogenetic structure and are therefore expected to be less powerful than methods which do, such as the skyline plot (Felsenstein 1992).

The classic skyline plot (fig. 4a) for the same tree gives a different picture of demographic history, as it suggests that effective population size has increased approximately exponentially in the recent past. This conclusion is a result of the limited phylogenetic signal in the data, which does not permit accurate estimation of the short internode intervals near the tips of the genealogy (as discussed earlier for fig. 1).

For a comparison, we also obtained a maximum likelihood estimate of effective population size using the program FLUCTUATE, which assumes a model of exponential growth (Kuhner, Yamato, and Felsenstein 1998). This estimate is shown as a thick line in figure 4b and c. Although the FLUCTUATE estimate only par-

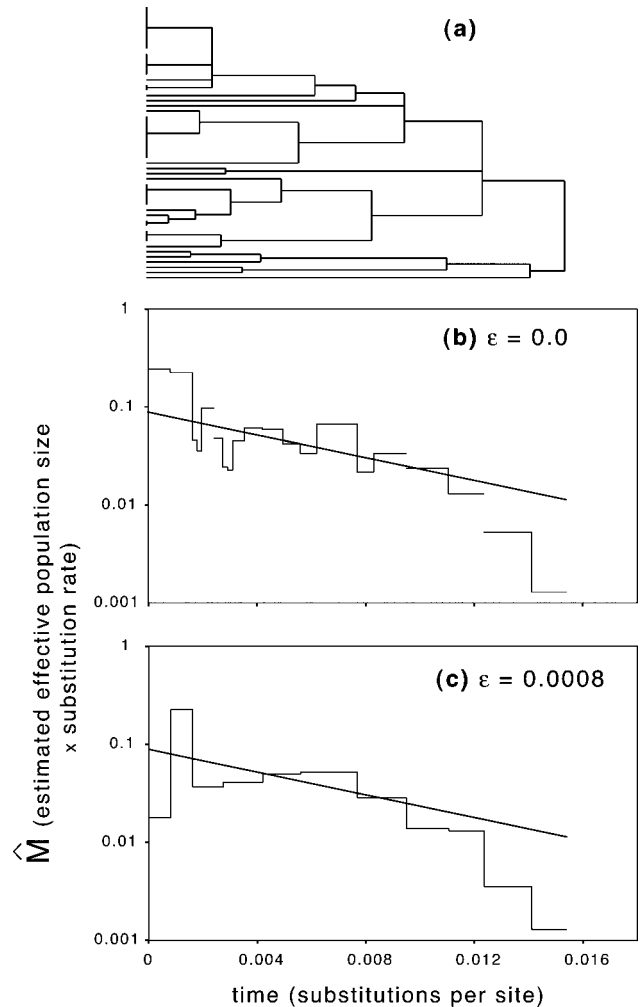


FIG. 4.—The red panda data set. *a*, The estimated genealogy. (*b*–*c*) Generalized skyline plots for the red panda genealogy: *b*, $\epsilon = 0.0$ (the classic skyline plot); and *c*, $\epsilon = 0.0008$ (the AIC_c estimate). The thick black line on each plot is a maximum likelihood parametric estimate obtained from the same data (see text).

tially matches the skyline plot estimates, it does clearly illustrate the effectiveness of the skyline plot as a model selection tool. If a logistic growth model was implemented in the FLUCTUATE package, then we would expect it to provide a better fit to the red panda data than the exponential model used here.

Discussion

The generalized skyline plot offers a flexible framework for exploring the demographic history of a sample of DNA sequences, and provides an estimate of effective population size which explicitly incorporates phylogenetic structure. It has three main advantages over the LTT plot and the classic skyline plot, (1) it can be applied to data containing a weaker phylogenetic signal or identical sequences (or both), (2) it provides an estimate of the coalescent error, and (3) it enables the stochastic noise present in the classic plot to be reduced.

The present approach is thus particularly useful as a rapid model selection tool, that is, the generalized sky-

line plot provides insights with respect to which parametric models may be suitable for a given data set. In the case of the HIV-1 data set (fig. 3), it indicates a model of exponential growth with a growth rate that increases through time. For the red panda mtDNA data set, a model of logistic growth appears to be most appropriate (fig. 4).

Our method is computationally fast and algorithmically straightforward. Tree estimation is separated from the problem of demographic inference, thus the underlying tree reconstruction method can be adapted to the particular data set in question. If an unusual or complicated substitution model is required, or if a model which permits variation in evolutionary rates among lineages is warranted (e.g., Gillespie 1991; Sanderson 1997; Thorne, Kishino, and Painter 1998; Huelsenbeck, Larget, and Swofford 2000), then these models can be used without altering the skyline plot method.

On the other hand, our approach requires that at least some of the divergence times in a gene tree can be reliably inferred, so it cannot be used on data containing very little variation. It is also important to realize that our approach is a single-tree method. It is therefore complementary to computationally intensive approaches which treat the tree as an unknown nuisance variable and effectively use a collection of trees to infer effective population size (Kuhner, Yamato, and Felsenstein 1995, 1998; Stephens and Donnelly 2000).

In addition to the coalescent error σ_c^2 estimated here, the skyline plot also carries an error introduced by the uncertainty of the phylogenetic estimates of coalescent times. This error has been ignored here and we are currently investigating ways of estimating its effect on the skyline plot.

Acknowledgments

We thank Andrew Rambaut and Peter Donnelly for discussion, and Bing Su for providing the red panda sequence alignment. We would also like to thank the editor and referees for helpful comments. One referee pointed out the useful simplification of equation (5). This work was supported by an Emmy-Noether-Fellowship of the Deutsche Forschungsgemeinschaft (K.S.) and by grant 50275 from the Wellcome Trust (O.G.P.).

APPENDIX

Computer Programs

Several computer programs are available for skyline plot analysis. The approach is implemented in the C++ program GENIE by O.G.P., available from <http://evolve.zoo.ox.ac.uk>, and in Java in the PAL library (Drummond and Strimmer 2001), available from <http://www.pal-project.org>. A web interface for skyline plot analysis written by Andrew Rambaut will be online at <http://evolve.zoo.ox.ac.uk>.

LITERATURE CITED

- AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19**:716–723.
- BURNHAM, K. P., and D. R. ANDERSON. 1998. Model selection and inference: a practical information-theoretic approach. Springer, New York.
- DONNELLY, P., and S. TAVARÉ. 1995. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**:401–421.
- DRUMMOND, A., and K. STRIMMER. 2001. PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics* **17**:662–663.
- 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.
- GILLESPIE, J. H. 1991. The causes of molecular evolution. Oxford University Press, Oxford.
- GRIFFITH, R. C., and S. TAVARÉ. 1994. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B* **344**:403–410.
- HARVEY, P. H., A. J. LEIGH BROWN, J. MAYNARD SMITH, and S. NEE, eds. 1996. New uses for new phylogenies, Oxford University Press, Oxford.
- HASEGAWA, M., H. KISHINO, and K. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:160–174.
- HUDSON, R. R. 1990. Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **9**:1–44.
- HUELSENBECK, J. P., B. LARGET, and D. SWOFFORD. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* **154**:1879–1892.
- HURVICH, C. M., and C. L. TSAI. 1989. Regression and time series model selection in small samples. *Biometrika* **76**:297–307.
- KINGMAN, J. F. C. 1982a. The coalescent. *Stoch. Proc. Applns.* **13**:235–248.
- . 1982b. On the genealogy of large populations. *J. Appl. Probab.* **19A**:27–43.
- 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.
- NEE, S., E. C. HOLMES, A. RAMBAUT, and P. H. HARVEY. 1995. Inferring population history from molecular phylogenies. *Philos. Trans. R. Soc. Lond. B* **349**:25–31.
- NORDBORG, M. 2001. Coalescent theory. Pp. 179–212 in D. BALDING, M. BISHOP, and C. CANNINGS, eds. *Handbook of statistical genetics*. Wiley, Chichester, England.
- 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., D. L. ROBERTSON, O. G. PYBUS, M. PEETERS, and E. C. HOLMES. 2001. Phylogeny and the origin of HIV-1. *Nature* **410**:1047–1048.
- RODRIGO, A. G., and J. FELSENSTEIN. 1999. Coalescence approaches to HIV population genetics. Pp. 233–272 in K. A. CRANDALL, ed. *The evolution of HIV*. John Hopkins University Press, Baltimore.
- ROSS, S. M. 1997. Introduction to probability models. 6th edition. Academic Press, San Diego.

- SANDERSON, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**:1218–1231.
- STEPHENS, M., and P. DONNELLY. 2000. Inference in molecular population genetics. *J. R. Statist. Soc. B* **62**:605–655.
- SU, B., Y.-X. FU, Y.-X. WANG, L. JIN, and R. CHAKRABORTY. 2001. Genetic diversity and population history of the red panda (*Ailurus fulgens*) as inferred from mitochondrial DNA sequence variations. *Mol. Biol. Evol.* **18**:1070–1076.
- SWOFFORD, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, Mass.
- TAVARÉ, S., D. J. BALDING, R. C. GRIFFITHS, and P. DONNELLY. 1997. Inferring coalescence times from DNA sequence data. *Genetics* **145**:505–518.
- 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.
- VIDAL, N., M. PEETERS, C. MULANGA-KABEYA, N. NZILAMBI, D. ROBERTSON, W. ILUNGA, H. SEMA, K. TISHIMANGA, B. BONGO, and E. DELAPORTE. 2000. Unprecedented degree of HIV-1 group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J. Virol.* **74**:10498–10507.
- YUSIM, K., M. PEETERS, O. G. PYBUS, T. BHATTACHARYA, E. DELAPORTE, C. MULANGA, M. MULDOON, J. THEILER, and B. KORBER. 2001. Using HIV-1 sequences to infer historical features of the AIDS epidemic and HIV evolution. *Philos. Trans. R. Soc. Lond. B* **356**:855–866.

KEITH CRANDALL, reviewing editor

Accepted August 27, 2001