Extinction rates can be estimated from molecular phylogenies

SEAN NEE, EDWARD C. HOLMES, ROBERT M. MAY AND PAUL H. HARVEY

A.F.R.C. Unit of Ecology and Behaviour, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, U.K.

SUMMARY

Molecular phylogenies can be used to reject null models of the way we think evolution occurred, including patterns of lineage extinction. They can also be used to provide maximum likelihood estimates of parameters associated with lineage birth and death rates. We illustrate: (i) how molecular phylogenies provide information about the extent to which particular clades are likely to be under threat from extinction; (ii) how cursory analyses of molecular phylogenies can lead to incorrect conclusions about the evolutionary processes that have been at work; and (iii) how different evolutionary processes leave distinctive marks on the structure of reconstructed phylogenies.

1. INTRODUCTION

Molecular phylogenies can both describe the hierarchical relationships among taxa by defining clades (monophyletic groups), and provide a time axis by utilizing molecular clocks. Even when those clocks have not been calibrated against real time, molecular phylogenies can still describe the temporal orderings of nodes. We have shown elsewhere how many properties of the past, present and even the future of a clade leave their signatures in its molecular phylogeny, even if that phylogeny is based only on information from extant organisms (Harvey *et al.* 1991, 1994*b*; Nee *et al.* 1992, 1994*b*). As a consequence, it is possible to use theoretical modelling, statistical analysis and biological knowledge of the clade under investigation to make inferences about many aspects of the tempo and mode of its evolutionary history. Here we describe how such inferences can be made, and produce illustrative examples to show how extinction rates can be estimated and other components of biodiversity can be measured.

We first describe the relationship between the *actual* phylogeny of a group, as would have been recorded in a perfect fossil record, and its molecular phylogeny, which we will refer to as the *reconstructed* phylogeny as it is based solely on extant species. We show how the reconstructed phylogeny can be used to determine which clades may have enjoyed unusually high rates of cladogenesis (lineage splitting) and to identify biological correlates of diversification. We then introduce a simple model of cladogenesis, which has a good pedigree in palaeontology (Raup *et al.* 1973; Gould *et al.* 1977; Stanley 1979), the constant rate birth–death process, and explore its implications for the analysis of the reconstructed phylogeny of a complete clade. We extend that model to describe situations in which the molecular phylogeny is based on only a random sample of extant species from the clade, and show how that extended model can provide important information for estimating biodiversity.

2. ACTUAL AND RECONSTRUCTED PHYLOGENIES

We first distinguish between actual and reconstructed phylogenies. Consider figure 1 which shows an actual phylogeny with lineages that give rise to descendants in the present day picked out in bold. The bold-lined phylogeny, with kinks removed, is the reconstructed phylogeny. We note four points when examining the actual and reconstructed phylogenies. First, both phylogenies have the same number of taxa in the present day. Second, at any point in the past, the reconstructed phylogeny generally has fewer lineages present (and never more) than does the actual phylogeny. Third, whereas the number of lineages can decrease towards the present in the actual phylogeny, that cannot happen in the reconstructed phylogeny. Finally, the reconstructed phylogeny provides timings for when each pair of species last shared a common ancestor and, therefore, commences at that point in the past when all present-day species shared their most recent common ancestor.

The first observation that strikes people with an interest in biodiversity is the enormous extent to which diversity varies amongst taxa: the huge difference between the number of species of beetles and butterflies would be a good example. In making a decision whether or not the causes of a clade's apparently high diversification should be investigated, it is desirable to

© 1994 The Royal Society and the authors

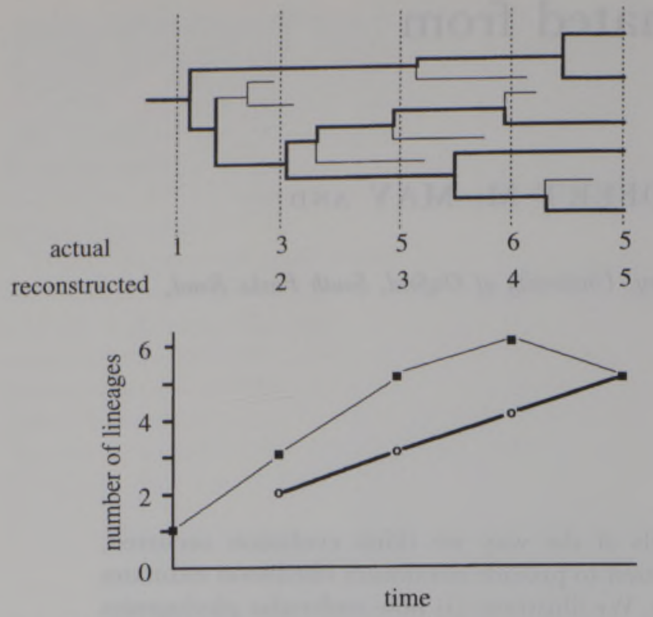

Figure 1. A hypothetical phylogeny as it would appear in a perfect fossil record. The bold lines represent those lineages which have some descendants at the present day and, so, would appear in a phylogeny reconstructed from molecular data. The numbers of lineages through time for the actual and reconstructed phylogenies are plotted at the bottom of the figure.

know whether or not the diversification really is remarkable with reference to some null model.

Fortunately, such an analysis is very simple. Consider a time window, say that between the middle and the right-hand dotted lines on figure 1. Our analysis is restricted to those lineages existing at the earlier time in the window (ancestral lineages) that will give rise to progeny lineages existing at the later time. For a broad class of models, the number of progeny lineages of any particular ancestral lineage in a reconstructed phylogeny has a geometric distribution (Nee *et al.* 1994*b*). We now ask the question, how many progeny lineages does each ancestral lineage give rise to? And does that distribution of progeny lineages fit our geometric expectation?

An an example of this procedure, figure 2 shows the results of an analysis of a molecular phylogeny of the birds (Harvey *et al.* 1991; Nee *et al.* 1992). The histogram shows the numbers of ancestral lineages with 1, 2, 3, etc. progeny lineages, and the curve shows the fitted geometric distribution. The two lineages that gave rise to 15 and 19 progeny lineages, the Passeri and Ciconiiformes, respectively, are anomalous: the probability in this case of any lineage giving rise to more than 14 progeny lineages is less than 0.005.

In addition to identifying clades which are unusually diverse, we can also exploit molecular phylogenies to construct tests of hypotheses concerning biological correlates of diversification. One way to do this is as follows. Suppose we wish to test whether small body size promotes cladogenesis, an hypothesis that has been put forward to account for the radiation of the passerines. Having reconstructed ancestral character states, we can ask, for each node in a phylogeny whether the branch leading to the smaller-bodied clade is longer or shorter than the branch leading to the larger bodied-clade (Nee *et al.* 1992). Under the null hypothesis that the branch lengths leading to the lighter and heavier daughter clades have the same expected length, we found no evidence to support the hypothesis that body size promotes cladogenesis (Nee *et al.* 1992).

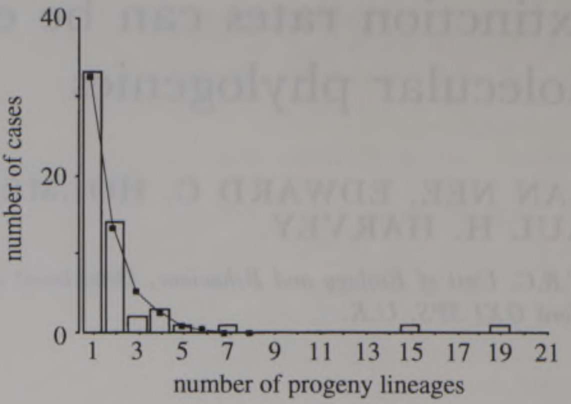

Figure 2. The results of the time window analysis, described in the text, as applied to the Sibley and Ahlquist's molecular phylogeny of the birds. The histogram shows the numbers of ancestral lineages with one, two, three, etc. progeny lineages, and the line is the expectation from the fitted geometric distribution.

Another approach to testing for correlates of diversification uses sister taxon comparisons, comparing the sizes of the two clades on each side of a node against expectations which are, again, based on the geometric distribution (Slowinski & Guyer 1993).

3. THE CONSTANT RATE BIRTH–DEATH MODEL

To make further inferences about various quantities of interest from the evidence provided by reconstructed phylogenies, it is necessary to have a variety of explicit theoretical models which describe how these phylogenies may have been generated. One of the simplest, the constant rate birth–death process, supposes that each lineage, at each point in time, has the same probability of giving rise to a new lineage, or of going extinct, as any other lineage, and that these probabilities do not change over time. The growth of reconstructed phylogenies under this model, and its generalizations, is analysed elsewhere (Nee *et al.* 1994*b*). To describe its properties in broad terms, we represent the information contained in a reconstructed phylogeny simply as a lineage-through-time plot (figure 3), in which points are plotted when the second, third, fourth, etc. lineages appear and we then connect-the-dots (Harvey *et al.* 1994*b*).

If there was no extinction, i.e. the phylogeny grew as a pure birth process, then the curve representing the actual number of lineages through time in figure 3 would be exactly coincident with the curve representing the number through time in the reconstructed phylogeny, and this would be a straight line on the semilogarithmic plot. As the death rate d becomes larger relative to the birth rate b (but still $d < b$), the two lines pull apart, remaining joined solely at the

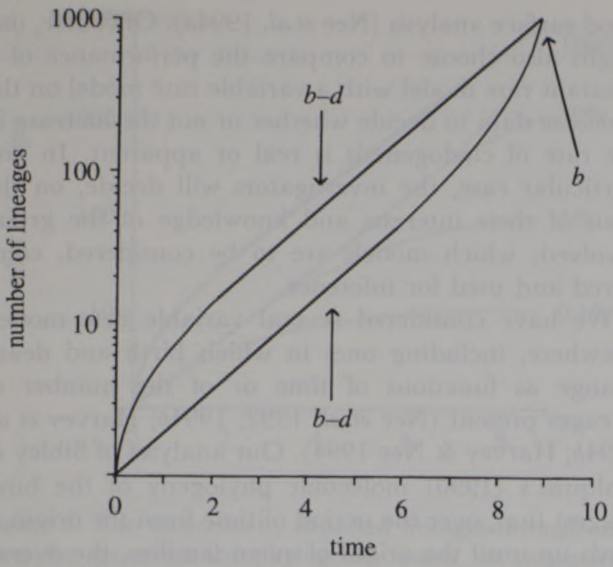

Figure 3. The theoretically expected growth in the numbers of lineages through time for an actual (top line) and reconstructed phylogeny growing according to a constant rate birth–death process. The slopes of both curves are *b–d*, the speciation rate minus the extinction rate, over most of the history of the clade, and the slope of the reconstructed phylogeny asymptotically approaches the speciation rate towards the present day. The two curves pull apart further the greater the ratio of the extinction rate to the speciation rate.

origin and the present day. The slopes of the lines are the same $(b–d)$ most of the time. Each line has a period of curvature. 'The push of the past', the apparently higher rate of cladogenesis at the beginning of the growth of the actual phylogeny, results from the fact that we are only considering those clades which survived to the present day, and these are the ones which on average got off to a flying start. 'The pull of the present', the apparent increase in the rate of cladogenesis in the recent past in the reconstructed phylogeny, results from the fact that lineages which arose in the more recent past have had less time to go extinct and, so, are more likely to be represented in our reconstructed phylogeny. The slope of the reconstructed curve asymptotically approaches the birth rate as we get closer to the present (Harvey *et al.* 1994*b*). Both the pull of the present and the push of the past become larger as d/b increases towards unity.

In fact, the parameters of most interest for phylogenies which grow according to this model are not *b* and *d* separately, but functions of *b* and *d*; specifically, $(b–d)$ and d/b. The first, $(b–d)$, controls the rate of growth of the phylogeny and the second, d/b, controls both the magnitude of the pull of the present and the vulnerability of small clades to extinction through 'demographic stochasticity' (*sensu* May 1973; see MacArthur & Wilson 1967).

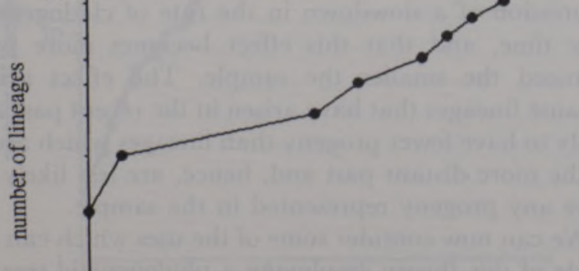

Figure 4. The lineages-through-time plot for species of the *Drosophila melanogaster* subgroup.

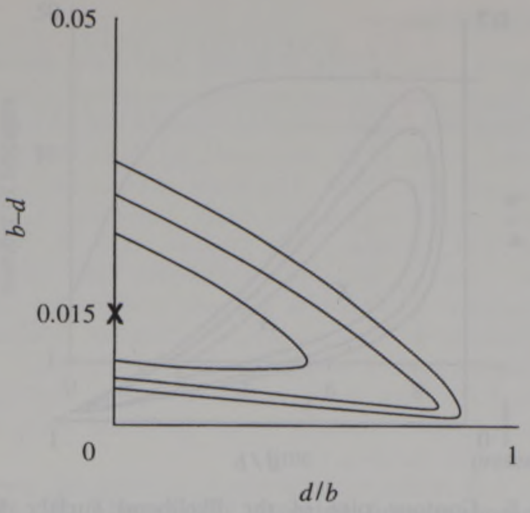

Figure 5. Contour plot of the likelihood surface for the *Drosophila* data. The maximum likelihood estimate, the peak of the surface, is marked by an *X*.

Consider, now, figure 4, which shows the increase through time in the number of lineages in a molecular phylogeny of the *Drosophila melanogaster* species subgroup (after Caccone *et al.* 1988). The graph starts when there are two lineages, because we have no information about the time of origin of the first lineage, and the time axis is in arbitrary units. The curve appears to be a straight line, with stochastic wobbling, and simple inspection suggests that a constant rate birth–death model is a reasonable assumption for the data, and that the extinction rate for this group is zero, as there is no upward curve towards the present. Using the underlying probability model, we can construct a log likelihood surface for the parameters (figure 5), from which we may make inferences about which parameter values are supported by the data. The peak of this surface, marked by an X, is the maximum likelihood estimate. The maximum likelihood estimate of d/b is zero (i.e. a zero rate of extinction). The contour lines in the figure correspond to one, two and three units of 'support'. The maximum likelihood estimate of the parameters is about seven times more likely to produce the observed data than any point on the second contour line, and about twenty times more likely to produce the data than any point on the third contour line.

The shallowness of the likelihood surface in the d/b direction tells us quite clearly that we cannot exclude the possibility that, in fact, *Drosophila* has a substantial value of d/b. Although *Drosophila* give the appearance of having a zero extinction rate, this analysis shows us that we cannot have any great degree of confidence in that conclusion on the basis of these data alone. Such uncertainty about the parameters for small clades is not entirely a weakness of the evidence provided by reconstructed phylogenies as opposed to actual phylo-

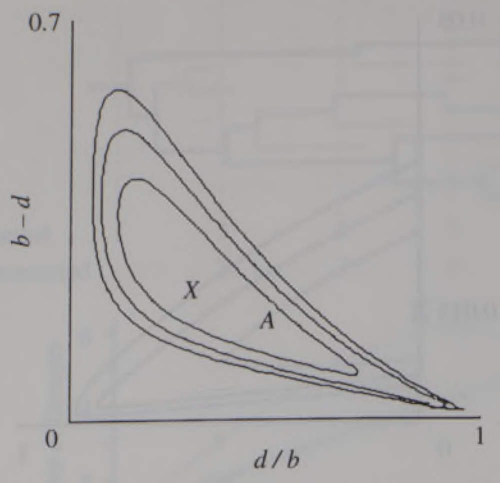

Figure 6. Contour plot of the likelihood surface for a simulated actual phylogeny of a clade that grows to about the same size as the *Drosophila* clade. The parameters chosen for the simulation were as follows: $b=0.4$, $d=0.2$, and the simulation was run for twenty time units. The letter *A* denotes the actual parameter values used and the *X* denotes the maximum likelihood estimate.

genies: figure 6 shows the likelihood surface for a simulated actual phylogeny with about the same number of species at the present day as the *Drosophila* subgroup. One benefit of this likelihood surface approach, as opposed to simply identifying an estimate of a quantity of interest, is that it quantifies our uncertainty: knowledge of what is uncertain is knowledge gained.

Consider the increase through time in the number of lineages of salamanders of the genus *Plethodon* (figure 7, derived from Highton & Larson's (1979) phylogeny). It might be tempting to conclude from the rapid acceleration in the rate of cladogenesis in the recent past that this genus is enjoying its salad days and that its future looks rosy. But appearances can deceive. In fact, we know that an apparent rapid acceleration in the rate of cladogenesis in the recent past is a property of reconstructed phylogenies that grow according to a birth–death process with a death rate that is large relative to the birth rate. So, in fact, it may be more reasonable to conclude that *Plethodon* has a large d/b, and, so, is highly vulnerable to extinction. This impression is confirmed by the likelihood surface analysis (Nee *et al.* 1994*a*). Of course, one might also choose to compare the performance of a constant rate model with a variable rate model on the *Plethodon* data to decide whether or not the increase in the rate of cladogenesis is real or apparent. In any particular case, the investigators will decide, on the basis of their interests and knowledge of the group involved, which models are to be considered, compared and used for inference.

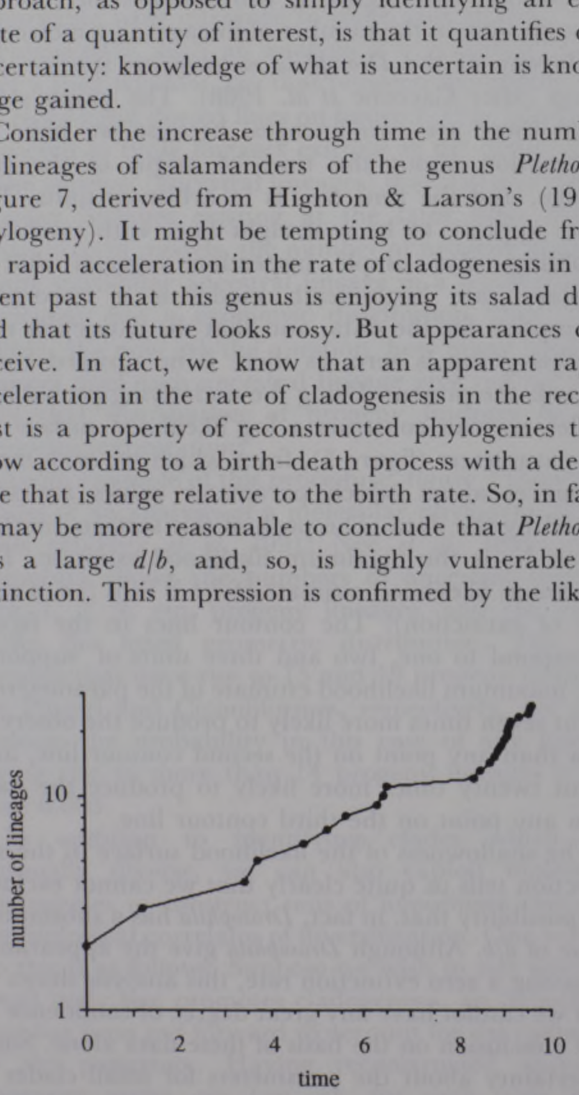

Figure 7. Lineages-through-time plot for salamanders of the genus *Plethodon*.

We have considered several variable rate models elsewhere, including ones in which birth and death change as functions of time or of the number of lineages present (Nee *et al.* 1992, 1994*b*; Harvey *et al.* 1994*b*; Harvey & Nee 1994). Our analyses of Sibley & Ahlquist's (1990) molecular phylogeny of the birds suggest that, over the period of time from the origin of birds up until the origin of avian families, the overall rate of cladogenesis (defined as $b-d$) was slowing down: either the lineage birth rate was decreasing through time or the lineage death rate was increasing through time (or both): possibly as a consequence of a niche-filling process. Subsequent analyses by Robert M. Zink and Joseph Slowinski (personal communication) on mitochondrial DNA sequence data suggest that the rate of cladogenesis has also been slowing in nine out of ten avian genera studied. We have also asked whether a reconstructed molecular phylogeny, such as that of the birds, could provide evidence for mass extinction events. The answer is that it could, but even if there was an 80% mass extinction of birds at the end of the Cretaceous, it is unlikely that we should detect clear evidence for it from the molecular phylogeny, particularly if background extinction rates were high relative to speciation rates (Harvey and Nee 1994).

4. WHEN ONLY A SAMPLE OF LINEAGES HAS BEEN ANALYSED

The theory we have been using is appropriate for reconstructed phylogenies that are based on all the members of a clade. We now explore the consequences of relaxing this assumption. Consider the simplest relaxation, and suppose that the reconstructed phylogeny consists of a set of species which have been chosen at random, with respect to their phylogenetic relationships, from a clade which has grown according to a constant rate birth–death process. Figure 8 shows that the effect of such sampling is to create the impression of a slowdown in the rate of cladogenesis over time, and that this effect becomes more pronounced the smaller the sample. The effect arises because lineages that have arisen in the recent past are likely to have fewer progeny than lineages which arose in the more distant past and, hence, are less likely to have any progeny represented in the sample.

We can now consider some of the uses which can be made of this theory employing a phylogenetic tree of the retrovirus HIV-1, the 24 sequences isolated primarily from Americans in the mid to late 1980s (Harvey *et al.* 1994*a*). It is not a very great leap from macroscopic lineages to viral lineages: the simple birth–death model we are using is a very simple model

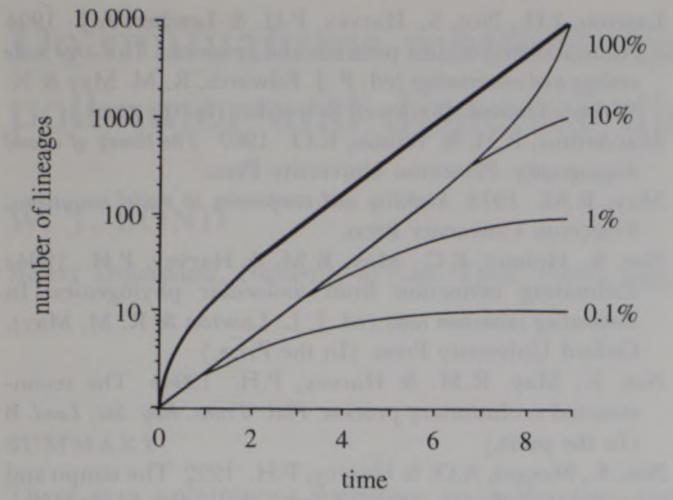

Figure 8. The theoretically expected lineages-through-time plots for reconstructed phylogenies based on successively smaller samples from the actual phylogeny described by the top line.

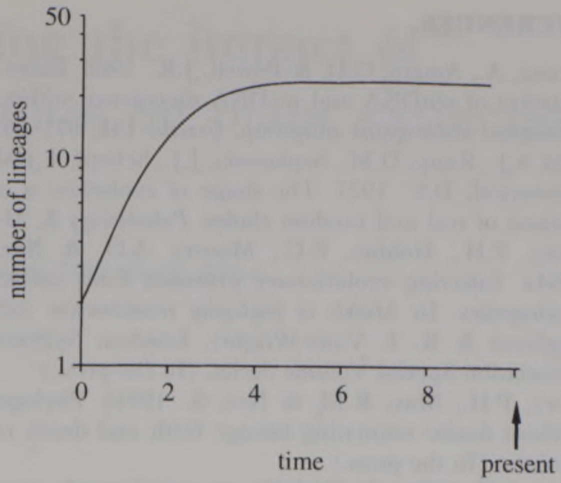

Figure 10. Theoretically expected lineages-through-time plot if 24 taxa (lineages) represent only one in ten thousand of the taxa in a clade.

of the growth of a clade, whether the clade consists of species of macroscopic organisms or lineages of an epidemically spreading disease organism.

Following Harvey *et al.* (1994*a*), figure 9 shows the increase through time in the actual number of lineages in the molecular phylogeny and figure 10 shows the theoretically expected increase through time in the number of lineages if the 24 people from whom the virus was isolated represent a very small fraction of the total number of people infected (the fraction 1 in 10000 was chosen for this figure). The similarity between these two figures gives us some confidence that the theoretical model is not wildly inappropriate, and we have used it to construct an estimate of the number of people infected (Nee *et al.* 1994*b*).

The sampling theory has other interpretations. Suppose that we wished to estimate the number of species in a large, little known clade. If we had a random sample of species in the clade then this could be done using the same analysis, where the fraction is now the fraction of the clade represented in our sample. Such an interpretation raises interesting questions about the relationships between those features of species which are likely to affect our sampling – such as abundance, body size and range size – and the phylogenetic relationships among the species (Lawton *et al.* 1994). Or suppose that we had reason to believe that a clade had recently suffered a mass extinction (as a result of human activities or otherwise), then, if we had all the extant species in our phylogeny, the same analysis could be used to estimate how many species have gone extinct. Again, interesting questions arise about the interpretation of 'phylogenetically random sampling' when we are discussing extinction. Of course, as we emphasize repeatedly, the particular models we are discussing here do not have any special status: phylogeny interpretation, as a discipline, will consider a large number of different models.

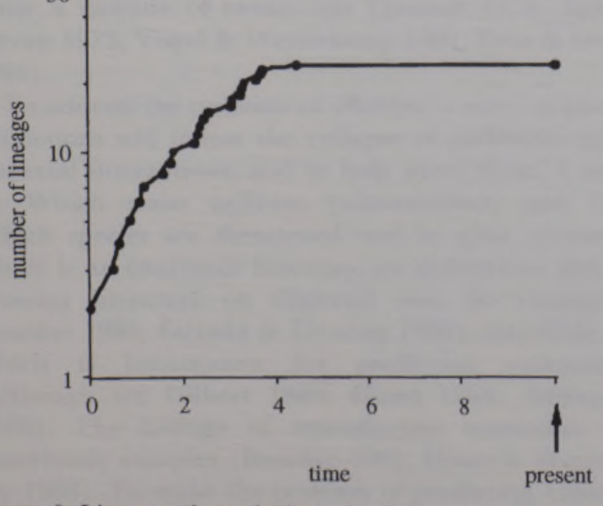

Figure 9. Lineages-through-time plot for a phylogenetic tree of 24 HIV-1 lineages.

An alternative class of models, known generically as the 'coalescent' (Kingman 1982), retain the number of lineages as constant so that whenever one lineage goes extinct, another immediately divides to replace it. Such models could describe density-dependent cladogenesis, for example, in which a niche space is saturated by a clade and the use of such models in the present context has been advocated by Hey (1992). When we compare the properties of the lineages-through-time plots from such constant number models with those derived from the constant rate birth–death process, we find marked differences (Nee *et al.* 1994*b*). For example, the plots for small samples taken from constant rate birth-death process decelerate towards the present (figure 8) while those for constant number models accelerate toward the present. If we had not considered the constant number models, we should mistakenly interpret a steepening towards the present in a small sample as good evidence for an increase in the rate of cladogenesis. Different processes do, indeed, leave different signatures on molecular phylogenies but it is important to interpret those signatures correctly.

We are grateful to AFRC (S.N.), The Wellcome Trust (grant 038468 to P.H.H.), SERC (grant GR/H53655 to P.H.H.), and the Royal Society (R.M.M.) for supporting the work described in this contribution.

REFERENCES

Caccone, A., Amato, G.D. & Powell, J.R. 1988 Rates and patterns of scnDNA and mrDNA divergence within the *Drosophila melanogaster* subgroup. *Genetics* **118**, 671–683.

Gould, S.J., Raup, D.M., Sepkowski, J.J., Schopf, T.J.M. & Simberloff, D.S. 1977 The shape of evolution: a comparison of real and random clades. *Paleobiology* **3**, 23–40.

Harvey, P.H., Holmes, E.C., Mooers, A.O. & Nee, S. 1994*a* Inferring evolutionary processes from molecular phylogenies. In *Models in phylogeny reconstruction* (ed. P. Eggleton & R. I. Vane-Wright). London: Systematics Association Special Volume Series. (In the press.)

Harvey, P.H., May, R.M. & Nee, S. 1994*b* Phylogenies without fossils: estimating lineage birth and death rates. *Evolution* (In the press.)

Harvey, P.H. & Nee, S. 1994 Comparing real with expected patterns from molecular phylogenies. *Biol. J. Linn. Soc.* (In the press.)

Harvey, P.H., Nee, S., Mooers, A.Ø. & Partridge, L. 1991 These hierarchical views of life: phylogenies and metapopulations. In *Genes in ecology* (ed. R. J. Berry & T. J. Crawford), pp. 123–137. Oxford: Blackwell Scientific.

Hey, J. 1992 Using phylogenetic trees to study speciation and extinction. *Evolution* **46**, 627–640.

Highton, R. & Larson, A. 1979 The genetic relationships of the salamanders of the genus *Plethodon*. *Syst. Zool.* **28**, 579–599.

Kingman, J.F.C. 1982 The coalescent. *Stoch. Process. Appl.* **13**, 235–248.

Lawton, J.H., Nee, S., Harvey, P.H. & Letcher, A.J. 1994 Animal distributions: patterns and processes. In *Large scale ecology and conservation* (ed. P. J. Edwards, R. M. May & N. Webb). Oxford: Blackwell Scientific. (In the press.)

MacArthur, R.H. & Wilson, E.O. 1967 *The theory of island biogeography*. Princeton University Press.

May, R.M. 1973 *Stability and complexity in model ecosystems*. Princeton University Press.

Nee, S., Holmes, E.C., May, R.M. & Harvey, P.H. 1994*a* Estimating extinction from molecular phylogenies. In *Estimating extinction rates* (ed. J. L. Lawton & R. M. May). Oxford University Press. (In the Press.)

Nee, S., May, R.M. & Harvey, P.H. 1994*b* The reconstructed evolutionary process. *Phil. Trans. Roy. Soc. Lond.* B (In the press.)

Nee, S., Mooers, A.O. & Harvey, P.H. 1992 The tempo and mode of evolution revealed from molecular phylogenies. *Proc. natn. Acad. Sci. U.S.A.* **89**, 8322–8326.

Raup, D.M., Gould, S.J., Schopf, T.J.M. & Simberloff, D.S. 1973 Stochastic models of phylogeny and the evolution of diversity. *J. Geol.* **81**, 525–542.

Sibley, C.G. & Ahlquist, J.E. 1990 *Phylogeny and classification of birds*. New Haven: Yale University Press.

Slowinski, J.B. & Guyer, C. 1993 Testing whether certain traits have caused amplified diversification: an improved method based on a model of random speciation and extinction. *Am. Nat.* **142**, 1019–1024.

Stanley, S.M. 1979 *Macroevolution: pattern and process*. San Francisco: Freeman.