The reconstructed evolutionary process

SEAN NEE, 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

Phylogenies reconstructed from contemporary taxa do not contain information about lineages that have gone extinct. We derive probability models for such phylogenies, allowing real data to be compared with specified null models of evolution, and lineage birth and death rates to be estimated.

1. INTRODUCTION

The simplest null model for the growth of a phylogenetic tree is a birth–death process in which the rates at which lineages either give birth to new lineages, or die, remain constant through time. This process has been studied before, both analytically and by simulation, for comparison with the changes in the numbers of taxa through a fossil record (Raup *et al.* 1973). Here we describe a null model appropriate for molecular phylogenies, which only record lineages which have given rise to at least one contemporary descendant. This stochastic process, which, for convenience, we call the 'reconstructed process', is what Kendall (1948*a*,*b*) calls a generalized birth process and is, in some ways, simpler than the birth–death process on which it is based. We derive the geometric distribution for the number of lineages existing at any particular time in the reconstructed process, and the distribution of waiting times between birth events, and generalize the results to varying birth and death rates. We construct the likelihood function for a reconstructed phylogeny which provides the basis for the estimation of birth and death rates. These latter can be distinguished, even in molecular phylogenies that do not contain information about lineages which have gone extinct. We also show how the theory can be generalized to phylogenies that contain only a sample of the extant members of a clade. Jagers (1991) provides a recent perspective on birth–death population models. Throughout this manuscript, we assume that molecular phylogenies are accurate; we discuss the influence of particular inaccuracies in such phylogenies elsewhere (e.g. Mooers *et al.* 1994).

2. THE BIRTH–DEATH PROCESSES

Consider a continuous time birth–death process which starts at time 0 with a single lineage. Lineages give rise to new lineages at a per-lineage rate λ and go extinct at a per-lineage rate μ. Even if $\lambda > \mu$, such a process can go extinct because all its lineages have gone extinct. This allows us to distinguish four related processes; figure 1 is a visual guide to the relevant distinctions. The first is all-encompassing, and is the simple birth–death process which may or may not survive to some arbitrary time t between its origin at time 0 and time T, which is the present day. The second process is a subset of the realizations of the first and consists of those realizations which survive to time t between times 0 and the present, but may or may not go extinct before the present day. The third process is the subset of these latter realizations which *do* survive to the present. The fourth is the reconstructed process, derived from the third by pruning the historical record of those lineages which do not have contemporary descendants. This corresponds to an ideal molecular phylogeny. Our primary interest is the third and fourth processes.

Let $\Pr\{i,t\}$ denote the probability that a process has i lineages at time t. The $\Pr\{i,t\}$ have geometric (or, strictly speaking, modified geometric) distributions for all but the third process, whose distribution is that of the sum of two independent random variables, each of which has a geometric distribution. Hence, the $\Pr\{i,t\}$ are elementary. To help avoid confusion, we subscript the $\Pr\{i,t\}$, 1 to 4, thereby emphasizing which of the four processes, described in the previous paragraph, is being referred to.

The $\Pr\{i,t\}$ can all be defined in terms of two functions of time, u_t and $P(t,T)$:

$$u_t \equiv \frac{\lambda(1-\exp(-(\lambda-\mu)t))}{\lambda-\mu\exp(-(\lambda-\mu)t)}, \tag{1}$$

$$P(t,T) \equiv \frac{\lambda-\mu}{\lambda-\mu\exp\{-(\lambda-\mu)(T-t)\}}. \tag{2}$$

$P(t,T)$ is the probability that a single lineage alive at time t has some descendants, i.e. has not gone extinct, at the later time T (Kendall 1948*a*). So $P(0,T)$ is the probability that a birth–death process which starts at time 0 with a single lineage is not extinct at time T.

For the simple birth–death process (Kendall 1948*a*),

$$\begin{cases} \Pr_1\{0,t\} = 1 - P(0,t) \\ \Pr_1\{i,t\} = P(0,t)(1-u_t)u_t^{i-1}, \qquad i > 0. \end{cases} \tag{3}$$

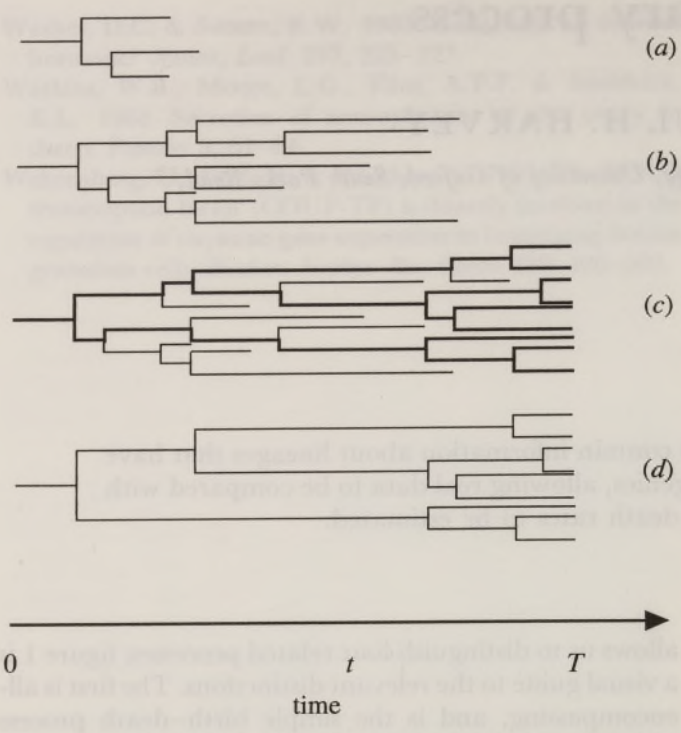

Figure 1. (*a*) A birth–death process which goes extinct before time *t*; (*b*) survives to *t*, but goes extinct before the present time *T*; the process (*c*) survives to the present. The bold lines are those lineages in (*c*) which have some descendants at the present. In (*d*) we have simply redrawn the bold lines in (*c*), removing the kinks, to construct an ideal reconstructed phylogeny.

For a birth–death process that is not extinct at time *t*, we have immediately, from distribution (3), the conditional probability $\Pr_2\{i,t\}$:

$$\Pr_2\{i,t\} = (1-u_t)u_t^{i-1}, \qquad i>0. \tag{4}$$

It is straightforward to derive from this the further conditional probability, $\Pr_3\{i,t;T\}$, for a birth–death process that survives to T. To do this, we compound distribution (4) with the probability that at least one of the i lineages existing at time t has some descendants at time T, and normalize appropriately:

$$\Pr_3\{i,t;T\} = \frac{(1-u_t)u_t^{i-1}(1-(1-P(t,T))^i)}{\sum_{i=1}^{\infty}(1-u_t)u_t^{i-1}(1-(1-P(t,T))^i)}, \qquad i>0. \tag{5}$$

This ugly expression disguises a simple underlying structure. The generating function of the probabilities (5) is:

$$\sum_{i=1}^{\infty}\Pr_3\{i,t;T\}s^i = \left\{\frac{s(1-u_t)}{1-u_ts}\right\}\left\{\frac{1-u_t(1-P(t,T))}{1-u_t(1-P(t,T))s}\right\}. \tag{6}$$

We recognize the generating function (6) as the product of the generating functions of two random variables, say X and Y, where the distribution of X is given by (4) and the distribution of Y is given by:

$$\Pr\{i,t;T\} = (1-u_t(1-P(t,T)))(u_t(1-P(t,T)))^i, \qquad i\geqslant 0, \tag{7}$$

that is, Y has a geometric distribution with parameter $u_t(1-P(t,T))$. Hence, the number of lineages existing at time t for a birth–death process which will survive to the later time T can be treated as the sum of two independent random variables, each with a geometric distribution, but with different parameters. This distribution is, of course, not itself geometric. But we will now see that the distribution of the process reconstructed from this one, by pruning all those lineages from the tree that do not have contemporary descendants is, once again, characterized by a geometric distribution for the $\Pr\{i,t\}$. There are two ways to derive this distribution. The first continues our onward march and compounds distribution (5) with a modified binomial distribution. The second way ignores the underlying birth–death process and identifies a generalized birth process which generates the reconstructed process. We will do both, as each technique can be more useful than the other for addressing different questions.

Denote the probabilities (5) by z_i. Of the k lineages existing at time t, i will have some progeny at the present time T, where i has a binomial distribution with parameter $P(t,T)$ and no zero term (since at least one will survive to the present). So, for the reconstructed process:

$$\Pr_4\{i,t;T\} = \sum_{k=1}^{\infty}\frac{z_k\binom{k}{i}P^i(t,T)(1-P(t,T))^{k-i}}{1-(1-P(t,T))^k}. \tag{8}$$

This simplifies to:

$$\Pr_4\{i,t;T\} = \left(1-u_t\frac{P(0,T)}{P(0,t)}\right)\left(u_t\frac{P(0,T)}{P(0,t)}\right)^{i-1}, \qquad i>0, \tag{9}$$

that is, a geometric distribution with parameter $u_tP(0,T)/P(0,t)$. This simplification is most readily achieved by calculating the generating function of distribution (8). Notice that $P(0,T)/P(0,t)$ is the overall probability that a birth–death process will survive to time T given that it has survived to time t.

We now identify a generalized birth process which generates distribution (9). Letting $n(t)$ be the number of lineages at time t, we grow a reconstructed phylogenetic tree from time 0, with $n(0)=1$, as follows. Each lineage gives rise to daughter lineages at a rate $\lambda P(t,T)$, so after the small time interval $\mathrm{d}t$,

$$n(t+\mathrm{d}t) = n(t)+1 \text{ with probability } n(t)\lambda P(t,T)\mathrm{d}t,$$
$$n(t+\mathrm{d}t) = n(t) \text{ with probability } 1-n(t)\lambda P(t,T)\mathrm{d}t. \tag{10}$$

This is a generalized birth process, with birth rate $\lambda P(t,T)$, and the formulae in Kendall (1948*b*) give us:

$$\Pr_4\{i,t;T\} = (1-\eta_{t,T})\eta_{t,T}^{i-1}, \qquad i>0, \tag{11}$$

where

$$\eta_{t,T} = \frac{\lambda(1-\exp(-(\lambda-\mu)t))}{\lambda-\mu\exp(-(\lambda-\mu)T)}. \tag{12}$$

It is easily confirmed that $\eta_{t,T} = u_t P(0,T)/P(0,t)$. We see that $\eta_{t,T}$ is the same as u_t, except that the denominator is a function of T rather than t. The fact that the progeny distribution for the reconstructed process is geometric was exploited in Nee *et al.* (1992) to identify episodes of 'species selection', that is to identify clades with a surprisingly large (in the statistical sense) number of lineages.

3. TIMES BETWEEN BIRTH EVENTS IN THE RECONSTRUCTED PROCESS

There are two ways to derive the distribution of waiting times between birth events in a reconstructed phylogeny. The first is to extend the probability model (10) as follows. Given that we have n lineages at time t_n,

$$\Pr\{\text{time until next lineage} > t + \mathrm{d}t; 0 < t < T - t_n\} =$$
$$\Pr\{\text{time until next lineage} > t;$$
$$0 < t < T - t_n\}\{1 - n(t)\lambda P(t,T)\mathrm{d}t\}. \quad (13)$$

Deriving and solving the differential equation gives us

$$\Pr\{\text{time until next lineage} > t; 0 < t < T - t_n\} =$$
$$\exp(-\lambda n \int_{t_n}^{t_n+t} P(s,T)\mathrm{d}s), \quad (14)$$

$$= \mathrm{e}^{-n(\lambda-\mu)t} \times$$
$$\left(\frac{1 - \frac{\mu}{\lambda}\exp(-(\lambda-\mu)(T-t_n-t))}{1 - \frac{\mu}{\lambda}\exp(-(\lambda-\mu)(T-t_n))}\right)^n. \quad (15)$$

The more direct approach is to realize that the $\Pr\{\text{time until next lineage} > t; 0 < t < T - t_n\}$ is simply the probability that each of the n lineages which we observe at time t_n has only one progeny (itself) an amount of time t later. Hence

$$\Pr\{\text{time until next lineage} > t; 0 < t < T - t_n\} =$$
$$(1 - \eta_{t,(T-t_n)})^n. \quad (16)$$

The probability that another lineage does not appear at all is simply $(1 - \eta_{(T-t_n),(T-t_n)})^n$. This approach makes it easy to generalize to varying birth and death rates.

From (15) we can derive the probability density of t, the waiting time for a birth:

$$n(\lambda-\mu)\mathrm{e}^{-n(\lambda-\mu)t} \times$$
$$\frac{(1 - \frac{\mu}{\lambda}\exp(-(\lambda-\mu)(T-t_n-t)))^{n-1}}{(1 - \frac{\mu}{\lambda}\exp(-(\lambda-\mu)(T-t_n)))^n}, \quad (17)$$

and, as noted above, the probability that another lineage does not appear at all is $(1 - \eta_{(T-t_n),(T-t_n)})^n$. The term before the quotient in density (17) is an exponential density with parameter $n(\lambda-\mu)$. As discussed in Harvey *et al.* (1994*b*), the reconstructed process behaves as a pure birth process with birth rate $(\lambda-\mu)$ over much of its early history. (The earliest information provided by molecular phylogenies is the time of the first bifurcation. If we wish to guess how long before that the first lineage arose, $1/(\lambda-\mu)$, the mean of an exponential distribution with parameter $(\lambda-\mu)$, is a rational guess.)

The times between birth events in a reconstructed phylogeny have been analysed for the pure birth process by Hey (1992) and by Sanderson & Bharathan (1993) for the situation in which we have all contemporary species, and by Slatkin & Hudson (1991) for the situation in which we have only a small sample of the contemporary species. For a birth–death process, Hey (1992) derived the distribution of waiting times for a model in which whenever a lineage is born another one simultaneously goes extinct, so the overall number of species is kept constant through time. If the constant number of species is N, then, Hey shows, the waiting time for the birth of the nth species is exponentially distributed with parameter $\lambda n(n-1)/(N-1)$. It is tempting to compare his results with ours for $\lambda = \mu$, when

$$\Pr\{\text{time until next lineage} > t; 0 < t < T - t_n\} =$$
$$\left\{1 - \frac{\lambda t}{1 + \lambda(T - t_n)}\right\}^n. \quad (18)$$

This is nonsense, however. Our model, as it stands, exhibits qualitatively different behaviour (extinction is asymptotically certain) when $\lambda = \mu$ and, so, cannot be used for comparison with Hey in this way. As an example, consider $< n(t) >$, the average number of lineages in the reconstructed phylogeny at time t, in the limit $\lambda = \mu$:

$$< n(t) > = \frac{1 + \lambda T}{1 + \lambda(T - t)}, \quad (19)$$

which says that the average number of lineages increases to infinity towards the present! Hey's model may usefully be viewed as an analytically tractable model of density-dependent cladogenesis, and applications of it can be found in Hey (1992) and Nee *et al.* (1994). (Density-dependent cladogenesis is a process in which birth and death rates vary as a function of the number of other lineages present. In this paper, we only allow birth and death rates to vary as functions of time: see below.) A somewhat unattractive feature of Hey's model is that the N lineages may have been around for an arbitrary length of time, and this possibility is reflected in the probability distributions of the times between nodes.

Hey notes the intimate relationship between his model and coalescent models in population genetics. Among other things, these models analyse the distribution of the times between nodes in a phylogeny (genealogy) of alleles from a population which has had a constant population size throughout its history. Apart from a difference in timescale, Hey's waiting time is the same as that for the canonical coalescent model (Kingman 1982) which studies the genealogy of alleles produced in a world with no selection or recombination (see Hudson (1990) for a review). This is not surprising, since Hey's model is, in fact, Moran's model (see e.g. Watterson 1984) with exponentially distributed times between 'generations'.

4. INFERRING BIRTH AND DEATH RATES FROM MOLECULAR PHYLOGENIES

Given a molecular phylogeny, our data set is the set $\{t_2, t_3, \ldots t_N\}$ of times when the second, third, fourth ... Nth lineages first appear, where N is the total

number of lineages in the phylogeny. We can take t_2 to be the origin. Define $x_n \equiv T - t_n$, so x_n is the length of time between the present and the birth of the nth lineage (figure 2). We will describe the construction of the likelihood for the simple case illustrated in figure 2. The generalization will then come naturally.

We arbitrarily designate one of the branches from each node as the daughter branch. The first contribution to the likelihood comes from the birth events at t_3 and t_4. (The birth of the second lineage does not contribute to the likelihood. This had to happen or we would not be looking at a tree!) These events have probabilities proportional to $(i-1)\lambda P(t_i, T)$ (see model (10)). The second contribution comes from the total amount of time that lineages do *not* give birth. As is evident from the broken tree, there are two lineages that have a single progeny (themselves) after an amount of time x_2, one lineage with a single progeny after an amount of time x_3, and one with a single progeny after an amount of time x_4. The probability of a single progeny after an amount of time x_i is $(1-u_{x_i})$. Multiplying together these probabilities, and inserting N into the formula instead of '4', we construct the likelihood for the general case of N lineages;

$$\text{lik} = (N-1)!\lambda^{N-2} \times \left\{\prod_{i=3}^{N} P(t_i, T)\right\}(1-u_{x_2})^2 \prod_{i=3}^{N}(1-u_{x_i}). \quad (20)$$

In a different context, and by a somewhat different route, Thompson (1975, pp. 54–58) derived a likelihood which differs from this in one relevant respect. In terms of our symbolism, her likelihood is this one multiplied by the additional term $(P(t_2, T))^2$, which is the probability that the two lineages from the first node in the tree are not extinct at the present time. We can derive likelihood (20) from Thompson's likelihood by dividing her likelihood by $(P(t_2, T))^2$, that is by conditioning on the fact that they are, indeed, not extinct (we *are* looking at a tree). Likelihood (20) can also be derived by multiplying together the densities (17) of the waiting times between births in the reconstructed phylogeny.

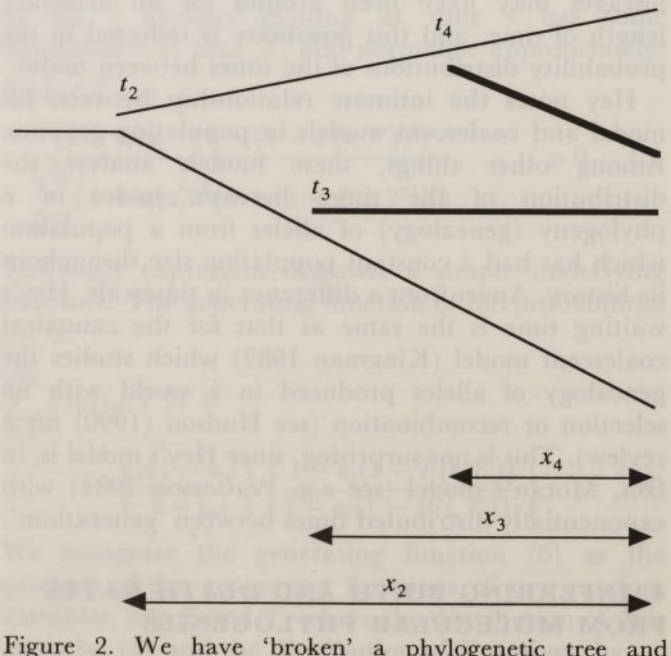

Figure 2. We have 'broken' a phylogenetic tree and arbitrarily designated daughter branches (bold lines). The t_i are the actual dates of the nodes and the x_i are the length of time elapsed between the nodes and the present day.

Defining the new parameters $a \equiv \mu/\lambda$ and $r \equiv (\lambda - \mu)$, likelihood (20) becomes

$$\text{lik} = (N-1)!r^{N-2}\exp\left(r\sum_{n=2}^{N-1} x_{n+1}\right)(1-a)^N \times \prod_{n=2}^{N}\frac{1}{(\exp(rx_n)-a)^2}, \quad (21)$$

and one can inspect the likelihood surface in the convenient space $\{0 < a < 1, 0 < r\}$. Elsewhere, we have applied this and other theory described in this paper (Nee *et al.* 1994).

5. TOPOLOGICAL FEATURES OF TREES

Although tree topologies are not our primary concern here, there is one point we wish to make which may not be widely appreciated. If different processes give rise to trees with the same topological features, then one can use whatever process is convenient to make inferences about these features. For example, the simplest coalescent model for a genealogy of alleles in population genetics supposes that, as we look back in time, any two lineages are as likely to fuse into a single lineage as any other two. Looking forward, this means that when a lineage divides in two, this node is as likely to occur on any particular lineage as any other. But this is the defining feature of the tree topologies generated by, in the simplest case, a pure birth process. This means that, as long as we are interested in purely topological features of the tree, any inferences based on a pure birth process are valid for the genealogical trees of population genetics.

This can be quite useful for generating results. Consider, for example, the 'remarkable fact' to which Felenstein (1992, p. 143) draws attention: 'If we consider the two lineages that result from the earliest fork, and wait until a total of n lineages exist, the distribution of the number of descendants of the left lineage is uniform on $1, 2, \ldots, n-1$.' The generalization of this would be as follows. If there are k lineages at any particular time and n lineages sometime later, then all vectors of progeny numbers, $(x_1, x_2, \ldots x_k)$, such that the sum total of the elements is n, are equally probable. This generalization, which was used in Nee *et al.* (1992), is true and follows immediately from the geometric distribution of progeny number under the pure birth process, as the probability of any vector of progeny numbers is proportional to $(1-u)^k u^{n-k}$.

We will now give an example of how further results can be easily achieved with simple combinatorial formulae (e.g. Feller 1968, Chapter 1). There are $(n-1)!/(k-1)!(n-k)!$ distinguishable and equiprob-

able progeny vectors, and $n!/x_1!\ldots x_k!$ arrangements of progeny into each such vector. So, for a given ancestor ordering, the probability of a vector is $(k-1)!(n-k)!x_1!\ldots x_k!/n!(n-1)!$. Multiplying this by $k!$, the number of ancestor orderings, we have the probability of the familial relationships defined by a vector. This is another derivation of probability (2.3) of Theorem 1 in Kingman (1982).

6. WHEN BIRTH AND DEATH RATES VARY OVER TIME

So far, our development has been in terms of the functions u_t and $P(t,T)$. Let $\mathrm{r}(t)=-(\lambda-\mu)t$. It is readily confirmed, from (1) and (2), that

$$u_t = 1 - P(0,t)\exp[\mathrm{r}(t)]. \tag{22}$$

One consequence of this is that the results presented so far can be expressed entirely in terms of the conceptually meaningful functions $P(t,T)$ and $\exp[-\mathrm{r}(t)]$, the latter being the expected number of progeny at time t for the simple birth–death process. A more important consequence is that this formulation allows a ready generalization to the case in which birth and death rates vary as functions of time.

We will now designate the birth and death rates as $\lambda(t)$ and $\mu(t)$ to denote their time dependence. Define

$$\rho(\tau,t) \equiv \int_t^\tau \{\mu(s)-\lambda(s)\}\mathrm{d}s, \tag{23}$$

which is the natural generalization of $\mathrm{r}(t)$.

In the general case (Kendall 1948*b*),

$$P(t,T) = \left[1 + \int_t^T \mu(\tau)\exp(\rho(\tau,t))\mathrm{d}\tau\right]^{-1}, \tag{24}$$

and it is still true that

$$u_t = 1 - P(0,t)\exp[\rho(t,0)], \tag{25}$$

where u_t is the parameter of the geometric distribution of progeny of a birth–death process when birth and death rates vary through time.

The earlier derivations of generating function (6), distribution (9) and likelihood (20) do not assume that birth and death rates are constant. They depend simply on the existence of meaningful u_t and $P(t,T)$. This means that all the distributions, and any quantities derived from them, are valid in the general case, and that one simply inserts the appropriate functions into the formulae.

Consider the expressions for $< n(t) >$, the average number of lineages existing at time t, for each of the four birth–death processes we have defined. For the simple birth–death process, we have from distribution (3)

$$< n(t) > = \frac{P(0,t)}{1-u_t} = \exp[-\rho(t,0)]. \tag{26}$$

For the process which survives to at least t,

$$< n(t) > = \frac{1}{1-u_t} = \frac{\exp[-\rho(t,0)]}{P(0,t)}. \tag{27}$$

For the process which survives to T, $< n(t) >$ is the sum of the means of the two geometric distributions discussed after generating function (6):

$$< n(t) > = \frac{1}{1-u_t} + \frac{u_t(1-P(t,T))}{1-u_t(1-P(t,T))}, \tag{28a}$$

$$= \frac{\exp[-\rho(t,0)]}{P(0,T)} - \frac{P(0,t)-P(0,T)}{P(0,t)P(t,T)}. \tag{28b}$$

Finally, we have for the reconstructed process:

$$< n(t) > = \frac{1}{1-u_t\dfrac{P(0,T)}{P(0,t)}}, \tag{29a}$$

$$= \frac{\exp[-\rho(t,0)]P(t,T)}{P(0,T)}. \tag{29b}$$

(28b) and (29b) can be derived from (28a) and (29a), respectively, using the identity

$$\exp[-\rho(t,0)]\left(\frac{1}{P(0,T)} - \frac{1}{P(0,t)}\right) = \frac{1}{P(t,T)} - 1, \tag{30}$$

which can be readily confirmed by substituting the explicit equations (23) and (24) into equation (30).

We emphasize that these expressions are valid for arbitrarily time-dependent $\lambda(t)$ and $\mu(t)$. In practice, this means that one need only derive the appropriate $P(\cdot)$ and $\rho(\cdot)$ functions for the problem of interest, from equations (23) and (24).

7. EXAMPLES

Experience with simulation shows that the increase in the average number of lineages through time is a useful and reliable guide to general features of the behaviour of the processes (Harvey and Nee 1994; Harvey *et al.* 1994*a*). In Harvey *et al.* (1994*b*) we discuss this increase under various evolutionary scenarios. We will consider two new ones here. First, a model of changing birth rates suggested by Strathmann & Slatkin (1983). Second, we will show how to deal with a situation in which the reconstructed phylogeny consists of only a random sample of the extant species, rather than all the members of the clade, and see how this sampling affects the picture.

Strathmann & Slatkin (1983) asked whether the observed numbers of phyla with small numbers of species are improbable under a birth–death model with constant birth and death rates. They concluded that they are. They then suggested a model in which birth rates are initially higher than death rates but subsequently decline. Such a situation could be generated by an unusual event, such as the break up of Gondwanaland. We model this as follows. We suppose that the death rate is constant. Initially, the birth rate is higher than the death rate and then decays hyperbolically with time, i.e. $\lambda(t) = k\mu/(1+at)$. k and a are chosen so that initially the birth rate is four times the death rate and becomes equal to the death rate half way towards the present, i.e. at $T/2$. There is no explicit expression for $P(t,T)$ in this case and we must resort to numerical integration. The

results of this analysis are depicted in figure 3. The contrast between what actually occurred and what we see in the reconstructed phylogeny is quite stark in this case. A smooth deceleration in birth rate manifests itself as a quick radiation followed by a long period of stasis, with another burst of cladogenesis in the recent past.

So far, our models have assumed that the reconstructed phylogeny is based on *all* the extant members of the clade. There is a large number of ways that this assumption may be violated. One way is that the species chosen for study are a random set with respect to their phylogenetic relationships. Even with this restriction there is still a variety of ways in which this random set may be constructed. We will now consider the simplest. Suppose that each species has a probability, f, of being included in our analysis. We can pretend that a mass extinction happened a moment before the analysis and all surviving species are included in the analysis. Because the mass extinction is an entirely notional event, we will use model (10) which deals directly with the reconstructed phylogeny. We will suppose that birth and death rates are constant.

Kendall's derivations are in continuous time and, so, do not allow for simultaneous extinction or, to put it another way, a finite probability of death at any particular instant in time. We can circumvent this problem by exploiting the reasoning that leads to Dirac delta functions. Let the death rate be the following function of time:

$$\mu(t,T) = \mu - g(t,T)\ln f. \tag{31}$$

Here μ is a constant, f is the probability of surviving the mass extinction, and $g(t,T)$ is essentially the Dirac delta function: a continuous function of time which is very close to zero everywhere except at $t = T$, the present, where it is very sharply peaked. Thus,

$$\int_0^s g(t,T)\mathrm{d}t = \begin{cases} 0, & s < T, \\ 1, & s > T. \end{cases} \tag{32}$$

By using such functions we can validly exploit continuous time models and then, when we have our results, pass to the limit representing an instantaneous force of mortality at time T which is survived with probability f. Proceeding in this way we find the following expression for $< n(t) >$, the average number of lineages at time t for the reconstructed process (compare equation (29b)):

$$< n(t) > = \frac{\exp[(\lambda - \mu)t]P_s(t,T)}{P_s(0,T)}, \tag{33}$$

where

$$P_s(t,T) = \frac{f(\lambda - \mu)}{f\lambda + (\lambda(1-f) - \mu)\exp(-(\lambda - \mu)(T - t))}. \tag{34}$$

$P_s(t,T)$ has the same meaning as $P(t,T)$, and we have subscripted it solely to denote that it is the appropriate function for the postulated sampling régime. Figure 4 shows that the effect of this sampling is to create a spurious impression of a decline in birth rate and/or increase in death rate through time. This sampling theory has numerous biological interpretations and applications (Nee *et al.* 1994).

We thank the AFRC (SN), the Royal Society (RMM), the SERC (PHH: GR/H53655) and The Wellcome Trust (PHH: 38468) for supporting this work. We thank Professor Sir David Cox for helpful discussions.

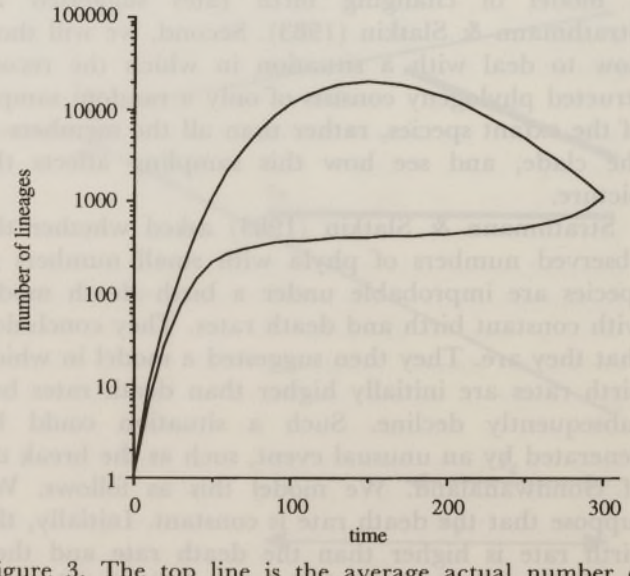

Figure 3. The top line is the average actual number of lineages through time, from equation (28b), and the bottom line is the average number of lineages at each time in the reconstructed phylogeny, from equation (29b). As described in the text, $P(t,T)$ is found by numerical integration. $\rho(\tau,t) = \mu(\tau - t) + (\ln(1 + at) - \ln(1 + a\tau))\mu k/a$. For this figure we chose $T = 300$, $\mu = 0.075$, $k = 4$, $a = 0.2$.

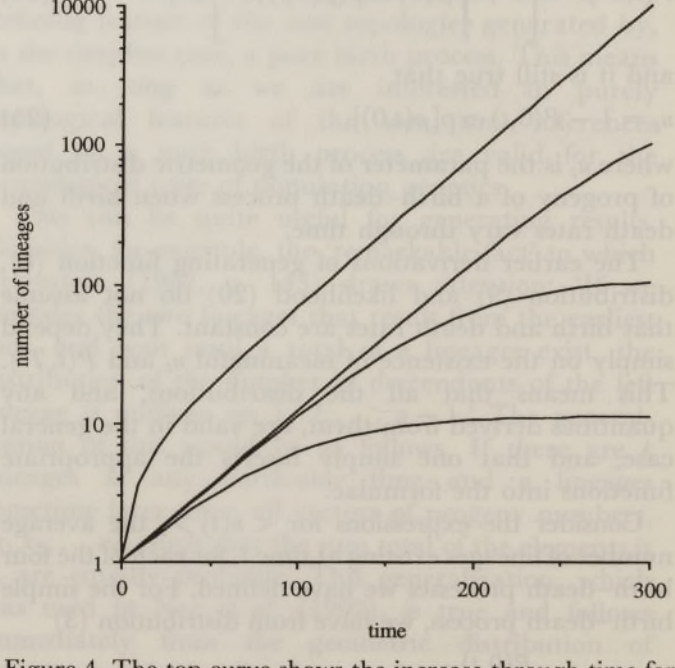

Figure 4. The top curve shows the increase through time for the average number of lineages that actually existed. This curve is generated by equation (28b) with $\lambda = 0.139$ and $\mu = 0.114$. Of the 10 000 species alive today, a fraction f is randomly chosen to construct the reconstructed phylogeny. Counting from the top, the second, third and fourth curves are generated by equation (33) with $f = 0.1$, 0.01 and 0.001, respectively, with λ and μ as for the top curve.

REFERENCES

Feller, W. 1970 *An introduction to probability theory and its applications: volume 1*. New York: John Wiley and Sons.

Felenstein, J. 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res., Camb.* **59**, 139–147.

Harvey, P.H., Holmes, E.C., Mooers, A.Ø. & Nee, S. 1994*a* Inferring evolutionary processes from molecular phylogenies. In *Models in phylogeny reconstruction*. Systematics Association Special Volume Series. (In the press.)

Harvey, P.H., May, R.M. & Nee, S. 1994*b* Phylogenies without fossils. *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.)

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

Hudson, R.R. 1990 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–44.

Jagers, P. 1991 The growth and stabilization of populations. *Stat. Sci.* **6**, 269–283.

Kendall, D.G. 1948*a* On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika* **35**, 6–15.

Kendall, D.G. 1948*b* On the generalized birth-and-death process. *Ann. Math. Stat.* **19**, 1–15.

Kingman, J.F.C. 1982 On the genealogy of large populations. *J. Appl. Prob.* **19A**, 27–43.

Mooers, A.Ø., Nee, S. & Harvey, P.H. 1994 Biological and algorithmic correlates of phenetic tree pattern. *Biol. J. Linn. Soc.* (In the press.)

Nee, S., Holmes, E.C., May, R.M. & Harvey, P.H. 1994 Extinction rates can be estimated from molecular phylogenies. *Phil. Trans. R. Soc. Lond.* B **344**, 77–82.

Nee, S., Mooers, A.Ø. & 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.

Sanderson, M.J. & Bharathan, G. 1993 Does cladistic information affect inferences about branching rates? *Syst. Biol.* **42**, 1–17.

Slatkin, M. & Hudson, R.R. 1991 Pairwise comparison of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562.

Strathmann, R.R. & Slatkin, M. 1983 The improbability of animal phyla with few species. *Paleobiology* **9**, 97–106.

Thompson, E.A. 1975 *Human evolutionary trees*. Cambridge University Press.

Watterson, G.A. 1984 Lines of descent and the coalescent. *Theor. Pop. Biol.* **26**, 77–92.

Received 23 November 1993; accepted 1 February 1994