BioMed Central

Research

Theoretical size distribution of fossil taxa: analysis of a null model

William J Reed*[1] and Barry D Hughes[2]

Address: [1]Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia V8W 3P4, Canada and [2]Department of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010, Australia

Email: William J Reed* - reed@math.uvic.ca; Barry D Hughes - hughes@ms.unimelb.edu.au

* Corresponding author

Abstract

Background: This article deals with the theoretical size distribution (of number of sub-taxa) of a fossil taxon arising from a simple null model of macroevolution.

Model: New species arise through speciations occurring independently and at random at a fixed probability rate, while extinctions either occur independently and at random (background extinctions) or cataclysmically. In addition new genera are assumed to arise through speciations of a very radical nature, again assumed to occur independently and at random at a fixed probability rate.

Conclusion: The size distributions of the pioneering genus (following a cataclysm) and of derived genera are determined. Also the distribution of the number of genera is considered along with a comparison of the probability of a monospecific genus with that of a monogeneric family.

Background

Mathematical modelling of the evolution of lineages goes back at least to Yule[1] who developed the eponymous *Yule process* (homogeneous pure birth process) in which speciations occur independently and at random. Yule's model did not include extinctions *per se*, because he believed that they resulted only from cataclysmic events. This issue was discussed at greater length by Raup[2], who distinguished between background and episodic extinctions. Raup started from a homomogeneous birth-and-death process model (in which background extinctions occur, like speciations, independently and at random) for which he presented mathematical results, and described more complex models of extinction including episodic extinctions and a mixture of episodic and background extinctions. However he gave no mathematical results for these models. Stoyan[3] considered a time in-homogeneous birth-and death process, in which speciation and

background extinction rates varied with time, based on the idea that younger paraclades have higher speciation rates, while older ones have higher background extinction rates.

There has been considerable discussion (*e.g.* Raup[2]; Patzkowsky[4]; Przeworski and Wall[5]) about the suitability of the null birth-and-death process model (with constant birth and death rates) as a macroevolutionary model of species diversification. In order to truly assess the validity of such a model it is necessary to have a full understanding of its properties which can then be compared with the fossil record. Specifically analysis is needed to generate hypotheses, which can be tested against available data. To date such an analysis is incomplete, relying on the partial analytic results of Raup[2] and the simulation results of Patzkowsky[4] and Przeworski and Wall[5].

Analytic results are clearly superior to simulation ones. In particular with analytic results for the size distribution of a clade one can fit the model via a multinomial likelihood, using observed size distributions, and thence test the adequacy of the underlying birth-and-death model using a statistical goodness-of-fit test. In addition analytic results are preferable to simulation ones, in that it is much easier to interpret a parametric formula than a collection of simulation results; and one does not have to distinguish between sampling variation due to a finite number of runs (noise) and signal.

It is the purpose of this paper to conduct a more thorough analysis of the birth-and-death model than that previosly carried out by Raup[2]. In particular we obtain results for size distributions of taxa and probabilities of monotypic taxa. In this paper we confine attention to obtaining analytic results and defer actual fitting and testing of the fit, using observed fossil data, to a future paper.

We develop the mathematical model presented by Raup[2] (and used in simulations by the above authors) to include the possibility of episodic, cataclysmic extinctions in which complete lineages are destroyed. We consider a hiearchy of models, which can include both cataclysmic and background extinctions of species and examine the resulting size distributions of extinct genera. We start (following section), as did Yule, by considering cataclysmic extinction only. Furthermore like Patzkowsky[4] and Przeworski and Wall [5], we assume that at any time an existing species can split, yielding a new species so radically different from existing ones that it becomes the founding member of a new genus. Thus we assume that the probability of a new genus being formed in an infinitesimal interval $(t, t + dt)$ is proportional to the total number of species in existence at time t. We derive results for the size distribution of extinct genera.

In the third and fourth sections we do the same assuming only background extinctions (but no cataclysmic extinction); and both cataclysmic and background extinctions (although the results here are limited). The fifth section is devoted to the distribution of the number of genera derived from the pioneering species and in the final section the probability of a monotypic genus is compared with that of a monogeneric family.

Cataclysmic extinctions only
Yule[1] considered the evolution of a genus begining with one species at time $t = 0$, which thenceforth evolves as a homogeneous pure birth process (Yule process) with speciation rate (birth parameter) λ. He then showed that N_t, the number of species alive at time t, follows a geometric distribution with probability mass function (pmf)

$$p_n(t; 1) = \Pr\{N_t = n | N_0 = 1\} = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1} \quad (1)$$

for $n = 1,2,....$ If instead there are initially n_0 species then from standard results (*e.g.* Bailey, 1964) the distribution of N_t is negative binomial with pmf

$$p_n(t;n_0) = \Pr\{N_t = n | N_0 = n_0\} = \binom{n-1}{n_0-1} e^{-\lambda n_0 t}(1-e^{-\lambda t})^{n-n_0} \quad (2)$$

for $n = n_0, n_0 + 1,....$

We now consider evolution of genera, and of species within genera, over an epoch between cataclysmic events. Let the time origin be the time of the previous cataclysm, and suppose only a single genus (containing n_0 species) survived that cataclysm. Let τ be the time of the succeeding cataclysm. Yule assumed that new genera were formed from old in a process analogous to that of speciation, thereby establishing that the time in existence of any genus would follow a truncated exponential distribution, with parameter equal to the rate at which new genera are formed from old. But it is more realistic to assume that a new genus is formed when a speciation within an existing genus is of such a radical form as to qualify the new species as belonging to a completely new genus. Thus the probabilty of a new genus being formed in an infinitesimal interval $(t, t + dt)$ should be proportional to *the existing number of species in all existing genera in the family* (and not to the existing number of genera in the family). We let

K_t denote the number of genera at time t, evolved from the pioneeering n_0 species;

L_t denote the number of species at time t in all genera, evolved from the pioneeering n_0 species; and

N_t denote the number of species in the pioneering genus at time t.

We assume that speciations (within a genus) occur at the rate λ and new genera are formed from existing species at the rate γ. Then to order $o(dt)$ the following state transitions (of K_t, L_t, N_t) can occur in $(t, t + dt)$:

$(k, l - 1, n - 1) \to (k, l, n)$ with probability $\lambda(n - 1)dt$

$(k, l - 1, n) \to (k, l, n)$ with probability $\lambda(l - 1 - n)dt$

$(k - 1, l - 1, n) \to (k, l, n)$ with probability $\gamma(l - 1)dt$

$(k, l, n) \to (k, l, n)$ with probability $1 - (\lambda + \gamma)ldt$.

Letting $p_{k,l,n}(t) = P(K_t = k, L_t = l, N_t = n)$, the following differential-difference equations can be established from the above:

$$\frac{d}{dt}p_{k,l,n}(t) = \lambda(n-1)p_{k,l-1,n-1}(t) + \lambda(l-1-n)p_{k,l-1,n}(t)$$
$$+\gamma(l-1)p_{k-1,l-1,n}(t) - (\lambda+\gamma)lp_{k,l,n}(t). \quad (3)$$

Using the generating function

$$\Phi(x,\gamma,z;t) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\sum_{n=1}^{\infty}p_{k,l,n}(t)x^k\gamma^l z^n, \quad (4)$$

multiplying (3) by $x^k \gamma^l z^n$ and summing yields the following partial differential equation

$$\Phi_t = \gamma(\lambda\gamma + \chi\gamma - (\lambda+\gamma))\,\Phi_\gamma + \lambda\gamma z(z-1)\,\Phi_z, \quad (5)$$

which can be solved by the method of characteristics (*e.g.* Bailey,[6]) with initial condition $\varphi(x, \gamma, z; 0) = x\gamma^{n_0}z^{n_0}$. From the solution the generating functions of K_t, L_t and N_t can be derived. They are

$$\Phi_K(x,t) = E(x^{K_t}) = x\left\{\frac{p(t)}{1-x[1-p(t)]}\right\}^{n_0}, \quad (6)$$

$$\Phi_L(\gamma,t) = E(\gamma^{L_t}) = \left\{\frac{\gamma e^{-(\lambda+\gamma)t}}{1-\gamma[1-e^{-(\lambda+\gamma)t}]}\right\}^{n_0}, \quad (7)$$

$$\Phi_N(z,t) = E(z^{N_t}) = \left\{\frac{z e^{-\lambda t}}{1-z(1-e^{-\lambda t})}\right\}^{n_0}, \quad (8)$$

where

$$p(t) = \frac{(\lambda+\gamma)e^{-(\lambda+\gamma)t}}{\gamma+\lambda e^{-(\lambda+\gamma)t}}. \quad (9)$$

From this it is clear that both the total number of species, L_t, and the number of species in the pioneering genus, N_t, have negative binomial distributions (with parameters n_0 and $e^{-(\lambda+\gamma)t}$ and n_0 and $e^{-\lambda t}$ respectively); while the number of genera K_t has a distribution related to the negative binomial – precisely $K_t + n_0 - 1$ has a negative binomial distribution with parameters n_0 and $p(t)$. The expected number of genera at time t is

$$E(K_t) = 1 + \frac{n_0\gamma}{\lambda+\gamma}\left[e^{(\lambda+\gamma)t}-1\right]. \quad (10)$$

It can be shown (see Appendix) that the times of formation of derived genera constitute an *order statistic process*. This means that they can be considered as the order statistics of a collection of independent, identically distributed (iid) random variables. From this it is shown that at any

fixed time τ, the times t_1, t_2,...,t_k that the derived genera have been in existence are iid random variables with probability density function (pdf)

$$f_k(t) = \frac{(\lambda+\gamma)e^{-(\lambda+\gamma)t}}{1-e^{-(\lambda+\gamma)\tau}}, \quad 0 < t < \tau. \quad (11)$$

By summing (3) over k and l one can show that N_t is a pure birth process with birthrate λ; and by summing over k and n that L_t is a pure birth process with birthrate $\lambda + \gamma$. From the fact that a pure birth process is an order statistic process it can be shown (see Appendix) that at time τ the times since establishment of all non-pioneering species in the pioneering *genus* are independently distributed random variables, with a truncated exponential distribution with pdf

$$f_N(t) = \frac{\lambda\, e^{-\lambda t}}{1-e^{-\lambda\tau}}, \quad 0 < t < \tau; \quad (12)$$

and that the times since establishment of all non-pioneering species in the pioneering *family* are independently distributed random variables, with a truncated exponential distribution with pdf

$$f_L(t) = \frac{(\lambda+\gamma)e^{-(\lambda+\gamma)t}}{1-e^{-(\lambda+\gamma)\tau}}, \quad 0 < t < \tau. \quad (13)$$

Note the fact that $f_L(t) \equiv f_K(t)$ *i.e.* the marginal distribution of the time since establishment of a derived genus in the family is the same as that of a derived species in the family.

Consider now the case when τ is the time of the first cataclysm since the appearance of the pioneering genus. The size distribution of all derived (non-pioneering) genera at the time of the cataclysm can be obtained by integrating the geometric pmf $p_n(t; 1)$ in (1) with respect to the truncated exponential distribution $f_K(t)$ between 0 and τ. This yields the pmf

$$q_n^{\text{deriv}} = \int_0^\tau p_n(t;1)f_K(t)dt$$
$$= \frac{1+\gamma/\lambda}{1+e^{-(\lambda+\gamma)\tau}}[B(2+\gamma/\lambda,n) - B_{e^{-\lambda\tau}}(2+\gamma/\lambda,n)], \quad (14)$$

where

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \quad B_x(a,b) = \int_0^x z^{a-1}(1-z)^{b-1}dz$$

are the *beta function* and *incomplete beta functions*, respectively. Alternatively the term in square brackets can be expressed in terms of the cumulative distribution function (cdf) $F(x; a, b)$ of the *beta distribution* with parameters a and b leading to

$$q_n^{\text{deriv}} = \frac{(1+\gamma/\lambda)B(2+\gamma/\lambda,n)}{1-e^{-(\lambda+\gamma)\tau}}\left[1 - F(e^{-\lambda\tau};2+\gamma/\lambda,n)\right]. \qquad (15)$$

This can be readily computed using standard statistical software.

The distribution of the size of the pioneering genus at time τ has pmf $q_n^{\text{pion}} = p_n(\tau, n_0)$ where p_n is negative binomial pmf given by (2). The distribution of the size of all existing genera at time τ is simply a mixture of q_n^{pion} and q_n^{deriv}. Precisely

$$q_n = \pi_K(\tau)q_n^{\text{pion}} + [1-\pi_K(\tau)]q_n^{\text{deriv}}, \qquad (16)$$

where $\pi_K(\tau)$ is the probability that a genus in existence at time τ is the pioneering genus, *i.e.*

$$\pi_K(\tau) = \text{E}\left(\frac{1}{K_\tau}\right) = \int_0^1 \frac{\Phi_K(s,\tau)}{s}ds, \qquad (17)$$

which can be evaluated as

$$\pi_K(\tau) = \frac{(\lambda+\gamma)e^{-(\lambda+\gamma)\tau}}{\gamma[1-e^{-(\lambda+\gamma)\tau}]}\log\left(\frac{\gamma+\lambda e^{-(\lambda+\gamma)\tau}}{(\lambda+\gamma)e^{-(\lambda+\gamma)\tau}}\right). \qquad (18)$$

Note that as $\tau \to \infty$, $\pi_K(\tau) \to 0$ and

$$q_n \to \frac{(\gamma/\lambda+1)\Gamma(\gamma/\lambda+2)\Gamma(n)}{\Gamma(\gamma/\lambda+n+2)}. \qquad (19)$$

This distribution was obtained by Yule[1] and is now known as the *Yule distribution*; for this distribution q_n behaves asymptotically like a power-law, *i.e.*,

$$q_n \sim (\gamma/\lambda + 1)\Gamma(\gamma/\lambda + 2) \times n^{-(2 + \gamma/\lambda)}$$

as $n \to \infty$, yielding the asymptotic straight line when q_n is plotted against n on logarithmic axes. We note in passing that setting $\gamma = 0$ in (19) does *not* yield the size distribution (as $\tau \to \infty$) of a single genus, since when $\gamma = 0$, $\pi_K \equiv 1$. In this case $N_\tau \to \infty$ with probability one.

Figure 1 shows the size distribution of pioneering and derived genera, along with the mixed distribution of all genera, calculated from the above formulae, for different values of n_0 and τ. They show how the results of Yule [1] need to be modified to take into account the effects of: (a) the evolution of new genera ; (b) pioneering genera of size (n_0) greater than one; and (c) the time, τ, until cataclysmic extinction. Large values of τ (right-hand panels), resulting in straight-line plots on the log-log scale, correspond most closely to the situation considered initially by Yule. In this

case approximate power-law (fractal) distributions occur. The deviations from such a power-law distribution are greatest when cataclysmic extinction occurs earlier (smaller τ) and when the number of species in the pioneering genus (n_0) differs greatly from one (lower panels). The distribution of derived genera (dotted lines) is unaffected by the initial size (n_0) of the pioneering genus. However the overall size distribution is affected (especially at values immediately above n_0) because of the fact that the pioneering genus size has support on $\{n_0, n_0 + 1,...\}$ while that of derived genera is on $\{1, 2,...\}$. This effect becomes less important when a long time elapses before the cataclysmic extinction event (because when τ is large, $\pi_K(\tau)$ is small–derived genera will in probability outnumber the pioneering one).

Background extinctions only

In this section we consider the size distribution of a fossil genus, starting with a single species (the case of a genus beginning with n_0 species is considered later in this section), subject to speciations at rate λ and background (individual) extinctions occurring independently and at random, at rate μ.

Thus N_t, the number of species alive t time units after the origin of the genus, follows a homogeneous birth and death process. Let M_t denote the total number of species in the genus that have existed by time t (*i.e.* $M_t = 1$ + number of speciations). The size of an extinct genus is a random variable M_T, where T itself is a random variable, denoting the time of extinction. Since no speciations can occur in a genus once it is extinct, we have that for $t \geq T$, $M_t \equiv M_T$. However T may not be finite ($N_t > 0$ for all t). Thus finding the distribution of the size of an extinct genus will involve conditioning on $T < \infty$ (or $N_\infty = 0$). Clearly it is given by the distribution of M_∞ conditional on $N_\infty = 0$.

Now let

$$p_{m,n}(t) = \Pr(M_t = m, N_t = n). \qquad (20)$$

It was shown by Kendall[7] that $p_{m,n}$ satisfies the differential-difference equations

$$\frac{d}{dt}p_{m,n}(t) = -(\lambda+\mu)np_{m,n}(t) + \lambda(n-1)p_{m-1,n-1}(t) + \mu(n+1)p_{m,n+1}(t) \qquad (21)$$

with initial condition

$p_{m,n}(0) = 1$ if $m = n = 1$; $p_{m,n}(0) = 0$ otherwise.

Let

$$\Psi(s,z;t) = \sum_{m=1}^{\infty}\sum_{n=0}^{\infty} p_{m,n}(t)s^m z^n \qquad (22)$$

be the generating function for M_t, N_t. Muliplying both sides of (21) by $s^m z^n$ and summing over $m = 1,... \infty$; $n = 0,...,\infty$ yields the partial differential equation

$$\Psi_t = (sz^2\lambda - (\lambda + \mu)z + \mu)\Psi_z. \quad (23)$$

This equation was derived and solved by Kendall[7], using the method of characteristics. The solution is (for $\lambda \neq \mu$)

$$\Psi(s,z;t) = \frac{\beta(sz-\alpha)\exp(\lambda\alpha t)+\alpha(\beta-sz)\exp(\lambda\beta t)}{(sz-\alpha)\exp(\lambda\alpha t)+(\beta-sz)\exp(\lambda\beta t)}, \quad (24)$$

where $\alpha = \alpha(s)$, $\beta = \beta(s)$ are the two (positive) roots of the quadratic equation

$$\lambda x^2 - (\lambda + \mu)x + \mu s = 0. \quad (25)$$

These roots are distinct for $0 \leq s \leq 1$, except when $\lambda = \mu$, where the roots are distinct for $0 \leq s \leq 1$, but coincide for $s = 1$. We select $\beta(s)$ to be the smaller root, so that

$$\beta(s) = \frac{1}{2}\left\{1 + \frac{\mu}{\lambda} - \sqrt{\left(1+\frac{\mu}{\lambda}\right)^2 - \frac{4\mu s}{\lambda}}\right\} \quad (26)$$

and note that $\alpha(1) = \max\{\lambda, \mu\}/\lambda$, $\beta(1) = \min\{\lambda, \mu\}/\lambda$ and $\lambda[\alpha(1) - \beta(1)] = |\lambda - \mu|$.

From (24) the individual generating function $\psi_M(s; t) = E(s^{M_t})$ of M_t (and similarly that of N_t) can be derived. Specifically

$$\Psi_M(s;t) = E(s^{M_t}) = \frac{\beta(s-\alpha)+\alpha(\beta-s)e^{-\lambda(\alpha-\beta)t}}{(s-\alpha)+(\beta-s)e^{-\lambda(\alpha-\beta)t}}. \quad (27)$$

Expanding this in a power-series expansion will yield the size distribution of the number of species which have existed by a finite time t. Simple closed-form expressions are not obtainable, but the expansion can be done numerically for specified parameter values using a computer mathematics program such as Maple VII[8]. It is easy to show that

$$E(M_t) = \Psi'_M(1) = 1 + \frac{\lambda}{\lambda - \mu}\left[e^{(\lambda-\mu)t} - 1\right]. \quad (28)$$

Note that for $\lambda > \mu$, $E(M_t) \to \infty$ as $t \to \infty$; while for $\lambda < \mu$, $E(M_t) \to \mu/(\mu - \lambda)$.

To find the distribution of the size of an extinct genus we consider the distribution of M_t conditional on $N(t) = 0$.

This has generating function $\Omega(s; t) = E(s^{M_t}|N_t = 0)$ given by

$$\Omega(s;t) = \sum_{m=1}^{\infty} \text{pr}(M_t = m \mid N_t = 0)s^m = \frac{\Psi(s,0;t)}{\text{pr}(N_t = 0)}. \quad (29)$$

The probabilty of extinction by time t in the denominator can be evaluated as $\Psi(1, 0; t)$ (or from standard results on birth and death processes) yielding

$$\Omega(s;t) = \left[\frac{\alpha\beta(1-e^{-\lambda(\alpha-\beta)t})}{\alpha - \beta e^{-\lambda(\alpha-\beta)t}}\right]\left[\frac{\max\{\lambda,\mu\}-\min\{\lambda,\mu\}e^{-|\lambda-\mu|t}}{\mu(1-e^{-|\lambda-\mu|t})}\right], \quad (30)$$

for $\lambda \neq \mu$, and

$$\Omega(s;t) = \left[\frac{\alpha\beta(1-e^{-\lambda(\alpha-\beta)t})}{\alpha - \beta e^{-\lambda(\alpha-\beta)t}}\right]\left[\frac{\lambda t + 1}{\lambda t}\right] \quad (31)$$

when $\lambda = \mu$.

Since once a genus is extinct it remains extinct forever, the size distribution

$$q_m^\dagger \overset{\text{def}}{=} \Pr\{M_\infty = m \mid N_\infty = 0\} \quad (32)$$

of an extinct fossil genus can be found by letting $t \to \infty$ in the generating function $\Omega(s; t)$ above. Since $\alpha(s) \geq \beta(s)$, with the inequality strict for $0 \leq s < 1$, we have $e^{-\lambda(\alpha-\beta)t} \to 0$ as $t \to \infty$. Thus if we let $t \to \infty$ in the generating function above, we deduce that for all $\lambda > 0$ and $\mu > 0$,

$$\Omega(s;\infty) = \sum_{m=1}^{\infty} q_m s^m = \frac{\max\{\lambda,\mu\}\beta(s)}{\mu} \quad (33)$$

$$= \frac{(\lambda + \mu)}{2\min\{\lambda,\mu\}}\left\{1 - \sqrt{1 - \frac{4\lambda\mu s}{(\lambda+\mu)^2}}\right\}. \quad (34)$$

Using the binomial theorem to expand the square root in (34) yields the pmf $q_m^\dagger$ for the size of an extinct fossil genus. Where $m \geq n_0 = 1$,

$$q_m^\dagger = \frac{(\lambda + \mu)}{\min\{\lambda,\mu\}}\frac{(2m-2)!}{(m-1)!m!}\frac{(\lambda\mu)^m}{(\lambda+\mu)^{2m}} \quad (35)$$

$$\sim \frac{(\lambda+\mu)}{4\pi^{1/2}\min\{\lambda,\mu\}m^{3/2}}\left[\frac{4\lambda\mu}{(\lambda+\mu)^2}\right]^m. \quad (36)$$

We observe that asymptotically q_m decays faster than a power-law, except in the case when $\lambda = \mu$ when it follows a power law with exponent -3/2.

The expected size of an extinct genus can be found by evaluating the derivative $\Omega_s(1; \infty)$, yielding
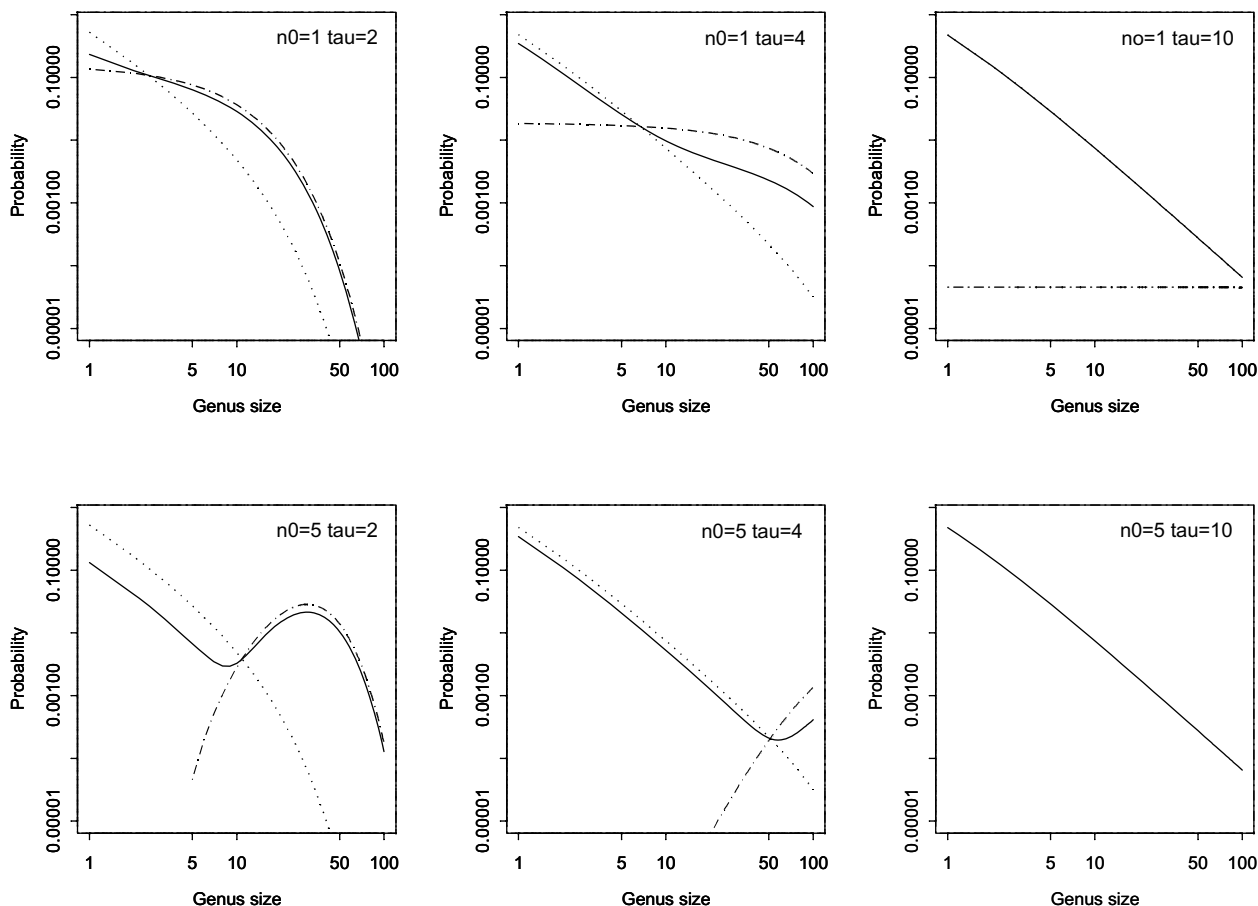
Figure 1
Logarithmic plots (both scales logarithmic) of the size distribution of genera, assuming only cataclysmic extinctions. The top row corresponds to $n_0 = 1$ and the bottom row to $n_0 = 5$. The three columns (from left to right) correspond to $\tau = 2, 4$ and 10. In all cases $\lambda = 1$ and $\gamma = 0.1$. For the sake of display the points of the probability mass function have been joined by lines:- dotted for derived genera; dot-dash for the pioneering genus and solid for the mixed distribution of all genera. The distribution of the pioneering genus (dot-dash) does not appear in the lower right-hand panel because the pmf assumes values less than 0.0001 for all sizes up to 100. In consequence the mixed distribution (solid line) is overlaid on that of derived genera (dotted line). Similarly in the upper right-hand panel the dotted and solid lines are overlaid.

$$E(M_\infty \mid N_\infty = 0) = \begin{cases} \lambda/(\lambda - \mu), & \lambda > \mu; \\ \infty & \lambda = \mu; \\ \mu/(\mu - \lambda), & \lambda < \mu. \end{cases} \quad (37)$$

The case $\lambda = \mu$ represents a phase transition analogous to the percolation phase transition (Hughes[9], Grimmett[10]). For this case although with probability one the genus goes extinct (*i.e.* $N_\infty = 0$, w.p.1), the expected time for this to happen is infinite.

If there were initially n_0 species in the genus, the expressions for the generating functions (24), (27) and (34) need to be modified by raising the expressions on the right-hand side to the n_0th power. In particular, if we denote the pmf for the size of an extinct genus by $q_m^\dagger(n_0)$ we have

$$\sum_{m=n_0}^{\infty} q_m^\dagger(n_0)s^n = \left\{ \frac{(\lambda + \mu)}{2\min\{\lambda, \mu\}} \left[1 - \sqrt{1 - \frac{4\lambda\mu s}{(\lambda + \mu)^2}} \right] \right\}^{n_0}. \quad (38)$$

We deduce at once from Eq. (38) that
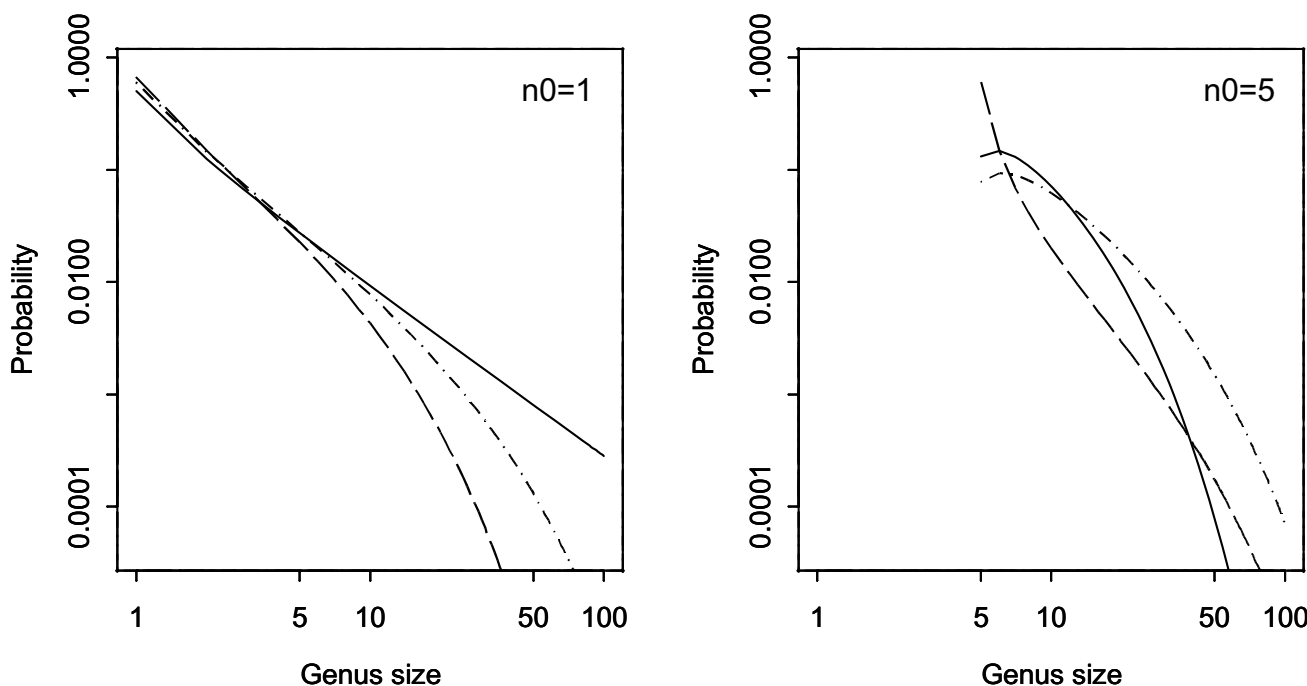
Figure 2
Logarithmic plots (both scales logarithmic) of the size distribution of genera, assuming only background extinctions. The left-hand plot is for n_0 = 1 and the right-hand one for n_0 = 5. For both plots λ = 1. For the sake of display the points of the probability mass function have been joined by lines:- solid (μ = 1); broken (μ = 1.5) and dot-dash (μ = 0.5).

$$q_m^\dagger(n_0) = \left[\frac{(\lambda+\mu)}{2\min\{\lambda,\mu\}}\right]^{n_0}\left[\frac{4\lambda\mu}{(\lambda+\mu)^2}\right]^m Q_m(n_0) \quad (m \geq n_0), \qquad (39)$$

where

$$\sum_{m=n_0}^{\infty} Q_m(n_0)z^m = [1-(1-z)^{1/2}]^{n_0}. \qquad (40)$$

The extraction of numerical values for the coefficients $Q_m(n_0)$ for a modest fixed value of n_0 is not difficult in practice. Alternatively, $Q_m(n_0)$ can be found by a contour integral argument that we shall not write out here, leading to the formula

$$Q_m(n_0) = \frac{1}{\pi}\sum_{\substack{j=1 \\ j\,\text{odd}}}^{n_0}\binom{n_0}{j}\sin(j\pi/2)\frac{\Gamma(j/2+1)\Gamma(m-j/2)}{\Gamma(m+1)} \quad (m \geq n_0). \qquad (41)$$

In particular, the following simple formula holds for n_0 = 1, 2, 3 or 4:

$$Q_m(n_0) = \frac{n_0(2m-2)!}{2^{2m-1}(m-1)!\,m!}\{1-\frac{(n_0-1)(n_0-2)}{4(m-3/2)}\}, \quad m \geq n_0.$$

From Eqs (39) and (41) we see that for arbitrary fixed $n_0 \geq 1$,

$$q_m^\dagger(n_0) \sim \left[\frac{\lambda+\mu}{2\min\{\lambda,\mu\}}\right]^{n_0}\left[\frac{4\lambda\mu}{(\lambda+\mu)^2}\right]^m \frac{n_0}{2\pi^{1/2}m^{3/2}}$$

as $m \to \infty$. The right-hand side of this differs from that of (36) only by a multiplicative constant, and for all $n_0 \geq 1$ asymptotically $q_m^\dagger(n_0)$ decays faster than a power law except in the case $\lambda = \mu$, when it follows a power law with exponent -3/2.

Fig. 2 shows the distribution of the size of an extinct genus plotted on logarithmic axes, for two values of n_0 and three values of μ with λ = 1. In the case n_0 = 1 (left-hand panel), an approximate power-law distribution (straight-line plot) can be seen in the case of equal birth and death rates ($\lambda = \mu$, the solid line). When the birth and death rates differ ($\lambda \neq \mu$) there is departure from the power-law with faster decay in probabilities as genus size increases both when $\lambda > \mu$ and when $\lambda < \mu$. In the case when the initial size n_0 of the pioneering genus exceeds one (right-hand panel), similar results pertain asymptotically (large genus sizes), but perturbations in the size distribution occur at the lower end (around n_0).

Both background and cataclysmic extinctions
We have very limited results in the case. The difficulty lies in the fact that at the time (τ, say) at which the cataclysmic extinction event occurs, different genera will have been in existence for different lengths of time. Unlike the case discussed in an earlier (no background extinctions) where we established that the times of establishment of new genera formed an order-statistic process, whence it followed that at time τ, the times in existence of distinct genera constituted iid random variables with a truncated exponential distribution, in the present case (with background extinctions) we have not been able to establish that the times of establishment of new genera constitute an order-statistic process. Thus it has not been possible to determine the size distribution of derived genera, destroyed in the cataclysm, since their time in existence is unknown. This is particularly unfortunate, since it seems that in fact for many fossil families both background and cataclysmic extinctions have occurred (Raup and Sepkoski [11]).

The only genus for which the time in existence is known is the pioneering genus. The pgf of the size of this genus is given by $\left[\Phi_M \left(s; \tau \right) \right]^{n_0}$ where Φ_M is defined in (27). This cannot be expanded in terms of simple functions to obtain explicit probabilities for sizes, although of course it can always be done numerically for specific parameter values. The expected size of the pioneering genus is

$$E(M_{pion}) = n_0 \left(1 + \frac{\lambda}{\lambda - \mu} \left[e^{(\lambda - \mu)\tau} - 1 \right] \right). \qquad (42)$$

1 Size distribution of families
In this section we consider the number of *genera* in the *family* derived from the pioneering species, assuming (as in the second section) that new genera are created by extreme speciations (at probabilistic rate γ) and (as in the third section) that background extinctions occur at the rate μ.

It can be shown (see Appendix) that the number of genera, G_t, which have existed up to time t has a generating function $\Phi_G(s; t) = E(s^{G_t})$ given by

$$\Phi_G(s; t) = s \left[\frac{(\lambda + \gamma)}{\lambda + \gamma s} \tilde{\Psi} \left(\frac{\lambda + \gamma s}{\lambda + \gamma}; t \right) \right]^{n_0}, \qquad (43)$$

where $\tilde{\Psi}$ is the same as Ψ_M in (27), but with λ replaced by $\lambda + \gamma$. This can be verified directly in the case $\mu = 0$ (only cataclysmic extinctions) for which $G_t \equiv K_t$ (see second section) with $G_t + n_0 - 1$ having a negative binomial distribu-

tion. In the more general case the proof is somewhat technical and is relegated to the Appendix. The expected number of genera in the family can easily be determined from (43) as

$$E(G_t) = 1 + n_0 \frac{\gamma}{\lambda + \gamma - \mu} \left(e^{(\lambda + \gamma - \mu)t} - 1 \right). \qquad (44)$$

If, following a cataclysmic event from which n_0 species survived, a subsequent cataclysm occurred τ time units later, the size distribution of the family (number of genera) derived from these n_0 pioneering species, would have pgf $\Phi_G(s; \tau)$. While no simple expansion of this is possible it can be done numerically. Some examples are shown in Fig. 3. The distributions show considerable deviation from a power law (straight line in logarithmic plots). They appear similar to the corresponding distributions of number of species in a genus (Fig. 1, top row) for smaller values of τ, but are further from the power-law form for larger τ. Thus it would appear that under the birth-and-death model power-law (fractal-like) size distributions are less likely to occur at higher taxonomic levels.

Monotypic taxa
One characteristic of interest in the empirical study of lineages is the proportion of monotypic taxa. Przeworski and Wall[5] compared the proportions of monospecific genera and of monogeneric families observed in the fossil record with results from a simulation of a birth-and-death process model. In this section we compute probabilities of such monotypic taxa. We consider the cases of (1) only background extinctions; and (2) only cataclysmic extinctions.

Only background extinctions
For a genus in existence for t time units, the probability of it having only ever contained one species by that time is

$$\Pr(M_t = 1) = \Psi'_M(0; t) = \lim_{s \to 0} \frac{\Psi_M(s; t)}{s} = \frac{\lambda e^{-(\lambda + \mu)t} + \mu}{\lambda + \mu} \qquad (45)$$

where Ψ_M is as in (27). Since all extinct fossil genera are finite in size, the probability of such a genus being monospecific is (from the results in fourth section)

$$\Pr(\text{monospecific genus}) = \Pr(M_\infty = 1 \mid M_\infty < \infty) = \begin{cases} \dfrac{\mu}{\lambda + \mu}, & \lambda \leq \mu, \\[2ex] \dfrac{\lambda}{\lambda + \mu}, & \lambda > \mu. \end{cases} \qquad (46)$$

Note that this is never less than one half (with this minimum value occurring when $\lambda = \mu$), so in the absence of cataclysmic extinctions, one should expect at least half of all extinct genera to be monospecific.
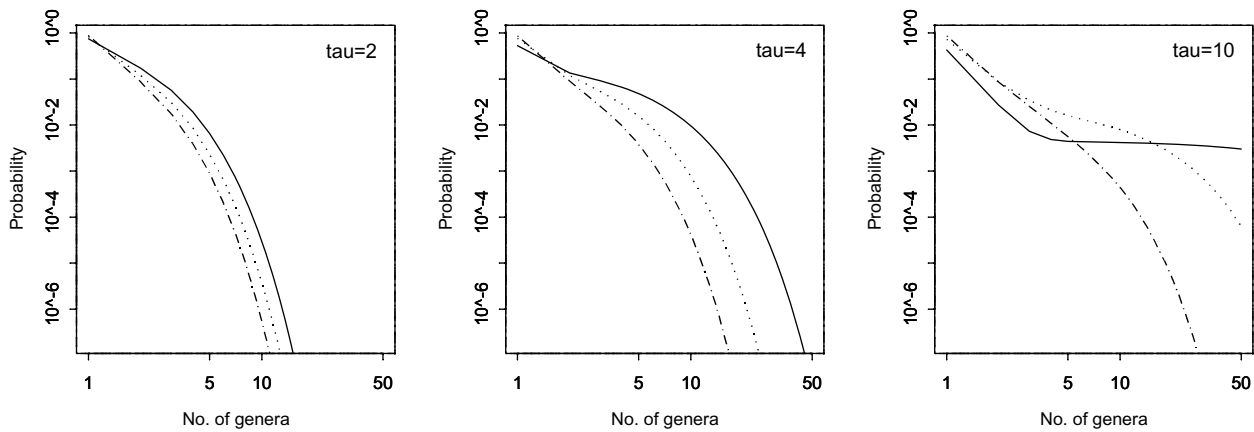
Figure 3
Logarithmic plots (both scales logarithmic) of the distribution of the number of genera in a family, assuming background and cataclysmic extinctions. The three panels (from left to right) correspond to τ = 2,4 and 10. In all cases λ = 1; γ = 0.1; n_0 = 1. For the sake of display the points of the probability mass function have been joined by lines:- solid (μ = 1); dotted (μ = 1.5) and dot-dash (μ = 0.5).

Consider now the distribution of the number of *genera* derived from a pioneering genus with n_0 species. Again since all observed extinct families will be of finite size, the probability of such a fossil family being monogeneric is

$$\Pr(G_\infty = 1 \mid G_\infty < \infty) = \begin{cases} \left(\dfrac{\lambda + \gamma}{\mu} \right)^{n_0} \Phi'_G(0,\infty), & \lambda + \gamma > \mu, \\ \Phi'_G(0,\infty), & \lambda + \gamma \le \mu. \end{cases} \quad (47)$$

where

$$\Phi'_G(0,\infty) = \frac{\partial}{\partial s} \Phi_G(s,\infty) \mid_{s=0} = \left[\left(\frac{(\lambda + \gamma)}{\lambda} \right) \Psi_L \left(\frac{\lambda}{\lambda + \gamma}; \infty \right) \right]^{n_0}$$

using (43). Thus, using (34), when $\lambda + \gamma > \mu$

$$\Pr(\text{monogeneric family}) = \left[\frac{\lambda + \gamma}{2\lambda\mu} \left(\lambda + \gamma + \mu - \sqrt{(\lambda + \gamma + \mu)^2 - 4\lambda\mu} \right) \right]^{n_0};$$

and when $\lambda + \gamma \le \mu$, the right hand side is modified by the fraction $(\lambda + \gamma)/(2\lambda\mu)$ being replaced by $1/(2\lambda)$.

Comparing the probability of a monospecific genus with that of a mono-generic family is complicated in general because of the number of parameters. But one can show that with $n_0 = 1$, the probability of a monogeneric family always exceeds that of a monospecific genus if the rate of formation of new genera is suitably small - *i.e.* if $0 < \gamma < \gamma_0$, for some positive γ_0 (depending on λ and μ). In this case

of course the probability of a monogeneric family will also exceed 0.5.

Only cataclysmic extinctions
If a cataclysmic extinction event occurs at time τ, the probabilities of a monotypic genus and of a monogeneric family can be found easily from the results of the second section using the explicit expressions for the generating functions of the number of species N_τ (8); and for the number of genera L_τ (6). Specifically if there is initially a single species in the genus the probability that it is monospecific at the time of extinction is

$$\Pr(\text{monospecific genus}) = \Pr(N_\tau = 1) = e^{-\lambda\tau}, \quad (48)$$

which is simply the probabilty of no speciations in $(0, \tau)$. In contrast the probabilty of a monogeneric family is

$$\Pr(\text{monogeneric family}) = \Pr(K_\tau = 1) = [p(\tau)]^{n_0} = \left[\frac{(\lambda + \gamma)e^{-(\lambda + \gamma)\tau}}{\gamma + \lambda e^{-(\lambda + \gamma)\tau}} \right]^{n_0}. \quad (49)$$

Comparing the right-hand sides of the above two equations, one can show that provided $\gamma < \lambda/n_0$ then Pr(monogeneric family) > Pr(monospecific genus) for τ less than some threshold value τ_0, say; but for $\tau > \tau_0$ the inequality is reversed. Thus as with the case of only background extinctions, monogeneric fossil families should be more common than monospecific fossil genera when the inter-cataclysm period is short. However if the inter-cataclysm period is longer the situation may be reversed.

Concluding remarks

In the paper a number of analytic results on the size distributions of genera and families and on the probabilities of monospecific taxa have been derived under the assumption of a simple homogeneous birth-and-death model and various extinction scenarios. The results are incomplete due to the complexity of the analysis, especially in the case when both cataclysmic and background extinctions can occur. However it is hoped that there are sufficient results to enable testing of the birth-and-death model using empirical taxon size distributions obtained from the fossil record.

Undoubtedly more complex plausible extinction scenarios than the two extremes discussed in this paper could be considered. For example one could consider major extinction events which resulted in the destruction of a significant proportion (but not all) of species within a genus. However realistically formulating a model for this, not to mention its subsequent analysis, seems to present a formidable task.

One could also consider the size distribution of taxa existing over more than one inter-cataclysmic epoch. In this case one would need to consider mixtures of the distributions, using different (but assumed known) values of τ. In principle this is not difficult to do. If the durations of inter-cataclysmic epoch were not known one could consider τ as a random variable and consider the resulting infinite mixture. As a null model for catclysmic extinction events, it seems reasonable to assume that they occur independently at random, so that the time between two successive events would have an exponential distribution. An overall distribution for the size of a taxon could then be obtained by integrating the results obtained in the earlier sections with respect to an exponential density. This has been considered in another paper (Hughes and Reed[12]) where it is shown that, under certain conditions, the resulting size distributions exhibit fractal-like behaviour.

Appendix

A point process $\{X_t, t \geq 0\}$ is said to be an *order statistic process* (Feigin[13]) if conditional on $X_\tau - X_0 = k$ the successive jump times (times of events) $T_1, T_2,...,T_k$ are distributed as the order statistics of k independent, identically distributed random variables with support on $[0, \tau]$. The simplest example is when $\{X_t\}$ is a Poisson process, for which conditional on $X_\tau - X_0 = k$, it is well known that the event times $T_1, T_2,..., T_k$ have the same distribution as the order statistics of of k independent, uniformly distributed random variables on $[0, \tau]$.

For a given order statistic process the order statistic distribution can be shown (Feigin[13](Theorem 2)) to have cdf

$$F(t) = \frac{m(t) - m(0)}{m(\tau) - m(0)}, \ 0 \leq \gamma \leq \tau, \qquad (50)$$

where $m(t) = E(X_t)$.

Puri[14] (Theorem 8) gives conditions for a non-homogeneous birth process, with birth rates $\theta_i(t)$, to be an order statistic process. For the process $\{K_t\}$ (the number of genera) in second section, the birth rates $\theta_k(t)$ are given by

$$\theta_k(t)dt = \Pr\left(K(t + dt) = k + 1 | K(t) = k\right)dt + o(dt). \qquad (51)$$

If we sum over l and n in (3) we find that with $p_k(t) = \Pr\{K_t = k\}$,

$$\frac{d}{dt}p_k(t) = \gamma E(L_t \mid K_t = k-1)p_{k-1}(t) - \gamma E(L_t \mid K_t = k)p_k(t) \qquad (52)$$

$$= \theta_{k-1}(t)p_{k-1}(t) - \theta_k(t)p_k(t) \qquad (53)$$

so that K_t does evolve under a non-homogeneous birth process, with birth rates

$$\theta_k(t) = \gamma E(L_t | K_t = k). \qquad (54)$$

We now calculate $\theta_k(t)$ explicitly. From Eq. (6),

$$p_k(t) = \frac{(n_0)_{k-1}p(t)^{n_0}}{(k-1)!}[1 - p(t)]^{k-1} \qquad (55)$$

with $p(t) = [(\lambda + \gamma)e^{-(\lambda + \gamma)t}]/[\gamma + \lambda e^{-(\lambda + \gamma)t}]$ and we note for later use that

$$\frac{p'(t)}{p(t)} = \frac{\gamma(\lambda + \gamma)}{\gamma + \lambda e^{-(\lambda+\gamma)t}}.$$

Since $p_0(t) = 0$, we have

$$\theta_1(t) = -\frac{p_1'(t)}{p_1(t)} = -\frac{n_0 p'(t)}{p(t)} = \frac{\gamma(\lambda + \gamma)n_0}{\gamma + \lambda e^{-(\lambda+\gamma)t}}. \qquad (56)$$

For $k \geq 1$ we have from (53) a difference equation to solve for $\theta_k(t)$:

$$(k - 1)\theta_{k-1}(t) - [1 - p(t)](n_0 + k - 2)\theta_k(t) = (n_0 + k - 2)\{n_0[1 - p(t)] - (k - 1)p(t)\}\frac{p'(t)}{p(t)}.$$

By inspection, a solution of this equation is given by

$$\theta_k(t) = -\frac{p'(t)}{p(t)}(n_0 + k - 1), \ k \geq 1.$$

As this solution gives the correct result (56) for $k = 1$ and a first-order linear difference equation needs only one boundary condition to uniquely determine the solution, we have proved that the birth rate is

$$\theta_k(t) = \frac{\gamma(\lambda + \gamma)(n_0 + k - 1)}{\gamma + \lambda e^{-(\lambda + \gamma)t}}, \ k \geq 1.$$

Puri's [[14], Theorem 8] condition for an order-statistic process on $(0, \tau)$ requires the existence of a positive, continuous and integrable function, $h(t)$ and positive constants $L(k)$ for $k = 1, 2,...$, with $L(1) = 1$ such that

$$\theta_k(t) \exp\left\{ \int_0^t [\theta_{k+1}(u) - \theta_k(u)] du \right\} = \frac{h(t)L(k+1)}{L(k)},$$

In the present case this is satisfied with

$$h(t) = n_0 \gamma e^{(\lambda + \gamma)t}$$

and $L(k) = (n_0 - 1)_k / n_0^k$. Also from Puri's Theorem 8,

$$E(K_t) = 1 + \int_0^t h(u) du = 1 + \frac{n_0 \gamma}{\lambda + \gamma}[e^{(\lambda + \gamma)t} - 1].$$

This agrees with the direct derivation of the expectation from the pgf of K_t (10) and enables the computation of the joint distribution of the times of establishment of derived genera as that of the order statistics of a random sample of size k from a distribution with cdf

$$F(t) = \frac{e^{(\lambda + \gamma)t} - 1}{e^{(\lambda + \gamma)\tau} - 1}, \quad 0 \leq t \leq \tau. \tag{57}$$

Thus it follows that at time τ the times since establishment of all derived genera are independent random variables with the truncated exponential distribution with pdf $f_K(t)$ given in (11).

To establish the truncated exponential nature of the distributions ($f_N(t)$ and $f_L(t)$, given in second section) of the times since establishment of species in respectively the pioneering genus and the pioneering family, is much easier. From the facts (established in the second section) that $\{N_t\}$ and $\{L_t\}$ are pure birth processes with both N_t and L_t having negative binomial distributions with $E(N_t) = n_0 e^{\lambda t}$ and $E(L_t) = n_0 e^{(\lambda + \gamma)t}$, and the well-known fact that a pure birth process is an order statistic process (Feigin[13]), one can easily establish (using (50)) the cdfs of the times since establishment of non-pioneering species in respectively the pioneeing genus and family. The pdfs, $f_N(t)$ and $f_L(t)$ given in (12) and (13) follow.

To establish the relationship (43) between the generating functions of G_t (the number of genera which have existed by time t) and L_t (the number of species which have existed by t), first let

$$Y_t = L_t - n_0 \text{ and } Z_t = G_t - 1 \tag{58}$$

denote the numbers of derived species and genera respectively. Since any speciation could have given rise to a new genus with probability $p = \gamma/(\lambda+\gamma)$, independently of other speciations, it follows that $Z_t | \gamma \sim \text{Bin}(\gamma, p)$ and hence that

$$
\begin{aligned}
\Pr(G_t = g \mid L_t = l) &= \Pr(Z_t = g - 1 \mid Y_t = l - n_0) \\
&= \binom{l - n_0}{g - 1} p^{g-1} q^{l - n_0 - g + 1} \\
&= \frac{p^{g-1}}{(g-1)!} D_q^{(g-1)} q^{l - n_0}
\end{aligned}
\tag{59}
$$

where $q = 1 - p$ and D_q is the differential operator $\frac{d}{dq}$. Multiplying by the pmf $f_l = P(L_t = l)$ and summing from $l = n_0$ to ∞ yields the marginal pmf of G_t, which can be written

$$
\begin{aligned}
\Pr(G_t = g) &= \frac{p^{g-1}}{(g-1)!} D_q^{(g-1)} q^{-n_0} \sum_{l=n_0}^{\infty} q^l f_l \\
&= \frac{p^{g-1}}{(g-1)!} D_q^{(g-1)} \left[\frac{\tilde{\Psi}(q)}{q^{n_0}} \right]
\end{aligned}
$$

where $\tilde{\Psi}(\cdot)$ is the pgf of L_t which is the same as the pgf of M_t (see (27)), but with λ replaced by $\lambda + \mu$. Thus

$$\Pr(G_t = g) = \frac{p^{g-1}}{(g-1)!} D_q^{(g-1)} \left[\frac{\tilde{\Psi}(q;t)}{q} \right]^{n_0}, \tag{60}$$

where (using (27))

$$\tilde{\Psi}(q;t) = \frac{\beta'(q - \alpha') + \alpha'(\beta' - q)e^{-(\lambda + \gamma)(\alpha' - \beta')t}}{(q - \alpha') + (\beta' - q)e^{-(\lambda + \gamma)(\alpha' - \beta')t}}. \tag{61}$$

with α' and β' being the roots of (25) with λ replaced by $\lambda+\mu$. The generating function of $\Psi_G(s; t)$ can be obtained by multiplying (60) above by s^g and summing from $g = 1$ to ∞:

$$
\begin{aligned}
\Psi_G(s;t) &= \sum_{g=1}^{\infty} s^g \Pr(G_t = g) \\
&= s \sum_{g=1}^{\infty} \frac{[sp]^{g-1}}{(g-1)!} \frac{d^{g-1}}{d\gamma^{g-1}} \left[\frac{\tilde{\Psi}(\gamma;t)}{\gamma} \right]^{n_0} \Bigg|_{\gamma=q} \qquad (62)\\
&= s \left[\frac{\tilde{\Psi}(q+sp;t)}{q+sp} \right]^{n_0}
\end{aligned}
$$

since the penultimate line is simply a Taylor series expansion about q of the last line.

Thus we conclude that

$$
\Psi_G(s;t) = s \left[\frac{(\lambda+\gamma)}{\lambda+\gamma s} \tilde{\Psi}\left(\frac{\lambda+\gamma s}{\lambda+\gamma} ; t \right) \right]^{n_0} . \qquad (63)
$$

References

1. Yule GU: **A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. F.R.S.** *R Soc Lond, Philos Trans (B)* 1924, **213**:21-87.
2. Raup DM: **Mathematical models of cladogenesis.** *Paleobiology* 1985, **11**:42-52.
3. Stoyan D: **Estimation of transition rates of inhomogeneous birth-death processes with a paleontological application.** *Elektronische Informationsverabeitung u. Kybernetic* 1980, **16**:647-649.
4. Patzkowsky ME: **A hierarchical branching model of evolutionary radiations.** *Paleobiology* 1995, **21**:440-460.
5. Przeworski M, Wall JD: **An evaluation of a hierarchical branching process as a model for species diversification.** *Paleobiology* 1998, **24**:498-511.
6. Bailey NTJ: *The Elements of Stochastic Processes* J Wiley and Sons: New York; 1964.
7. Kendall DG: **On the generalized "birth-and-death" process.** *Ann Math Stats* 1948, **19**:1-15.
8. Maple VII: Waterloo Maple Inc: Waterloo, Ontario; 2001.
9. Hughes BD: *Random Walks and Random Environments, Random Environments Volume 2*. Oxford University Press; 1966.
10. Grimmett G: *Percolation* 2nd edition. Springer-Verlag: Berlin; 1999.
11. Raup DM, Sepkoski JJ: **Periodicity of extinction in the geologic past.** *Proc Nat Acad Sci USA* 1984, **81**:801-805.
12. Hughes BD, Reed WJ: **A problem in paleobiology.** *arXiv:physics/0211090* 2002.
13. Feigin PD: **On the characterization of point processes with the order statistic property.** *J Appl Prob* 1979, **16**:297-304.
14. Puri PS: **On the characterization of point processes with the order statistic property without the moment condition.** *J Appl Prob* 1982, **19**:39-51.