Coherent Stochastic Models for Macroevolution

David Aldous and Lea Popovic

Department of Statistics
University of California
367 Evans Hall # 3860
Berkeley, CA 94720-3860

April 19, 2004

Abstract

We give a mathematician's view of evolutionary biology literature concerning stochastic models for phylogenetic trees. We spotlight a model for the tree on n extant species that would be observed if macroevolution were purely random. The model can be extended in two ways – to time series of observed taxa in a fossil record, and to different levels of the taxonomic hierarchy – and provides a logically consistent ("coherent") framework for simultaneous study thereof. We illustrate with a variety of theoretical calculations and simulations, and propose a variety of real-data projects suggested by our analysis.

`xxx This is a draft summarizing current state of an ongoing research project, not intended for publication in this form.`

1 Introduction

There is a substantial literature on comparing data on different aspects of *biodiversity* or *macroevolution* – the evolutionary history of speciations and extinctions – with the predictions of simple "pure chance" stochastic models. Available data includes
(a) fossil time series – fluctuations in number of taxa over time;
(b) shapes of phylogenetic trees on extant species (Mooers and Heard [27] provide an extensive survey);
(c) the distribution of number of species per genus.
The fit of simple models, and of more elaborate models incorporating conjectured biological process, have been studied in these contexts. While data-motivated models are scientifically natural, a mathematical aesthetic suggests a somewhat different approach: start with a "pure chance" model which encompasses simultaneously all the kinds of data that one might hope to find. Here are two instances of what one would like such a *coherent* model to provide.
(i) Joint description of the phylogenetic tree on an extant clade of species, its extension to the tree on an observed small proportion of extinct species, and the (unobserved) entire tree on all extinct species.
(ii) Joint description of fossil time series at different levels of the taxonomic hierarchy.
We emphasize (ii) because paleontology literature tends to assume that a model can be applied at any level, without enquiring whether this assumption is logically self-consistent.

Our purpose is to present what is arguably the mathematically fundamental such model. The underlying model is simple – a critical branching process conditioned to have n lineages at the present time. The model extends to higher-order taxa by assuming each new species has some probability of founding a new higher-order taxon; we consider several more detailed classification schemes emphasizing desiderata such as monophyletic groups. Model fossil species by assuming each extinct species has some probability of being seen in the fossil record. Though hardly new in concept, our focus on conditioning to have n lineages (for comparison with real clades on n extant taxa) makes our results somewhat new in detail.

Conceptually, this is a *neutral* model which does not incorporate conjectured biological process such as intrinsic tendency for species numbers to increase, differential speciation or extinction rates, or ecological constraints on numbers of species. For well-understood mathematical reasons (see end of Section 1.2) neutral models like our are implausible for large clades. In

a sense, the model seems most appropriate as a "null hypothesis" for small clades, at the recent fringe of the Tree of Life, or for a geological period free of mass extinctions and their aftermath.

The recent paper of Tavaré et al [49] is an exemplar of how to study a particular clade (primates) by combining different sources of data – phylogenetic trees on extant species, fossil record – with macroevolutionary models. Our model is not intended to be realistic enough for such applications, but rather, as a logical starting platform for building more realistic models.

1.1 Plan of paper

Section 2 reviews standard models, and Section 3 reviews the various biological uses of such models. Section 4 describes our basic model at the species level, and Section 5 presents its mathematical properties, encompassing both analytic formulas (some developed in detail elsewhere) and simulation results. Section 6 describes how our model is extended to model higher-order taxa. Section 7 (`xxx under construction!`) presents mathematical properties of the model for higher level taxa.

This paper has two different sets of conclusions. First, we obtain a suite of mathematical results within our model – Section 1.2 provides an overview. As mathematicians we are reticent to claim that simplistic mathematical models lead to biological insight. But we do think that our broader-ranging approach provides a somewhat different conceptual framework for asking questions about quantitative aspects of biodiversity, which could be studied by evolutionary biologists looking at real data. We outline five such questions in Section 1.3.

1.2 Overview of mathematical results for our model

1.2.1 Orders of magnitude

We draw attention to the following order[1] results for a clade with n extant species.

- The n^2 *law* : that the number of extinct species is order n^2 (Section 5.5).
- The n *law* : that the time since clade origin or since last common ancestor is order n times the mean species lifetime (Section 5.4).

[1] *order* means order of magnitude: $\frac{1}{2}n^2$ and $2n^2$ are both order n^2

- The $1/r$ *law* : that with probability $1/r$ there was some past time at which the number of species was at least r times the present number (Section 5.5).
- The $(\log n)/n$ *law* : that the probability a given extinct species is ancestor to some extant species is order $(\log n)/n$ (Section 5.6).
- The *constant law* : that the probability that a given extant species is descendant of some other extant species has non-zero limit as $n \to \infty$ (Section 5.3).

1.2.2 Formulas

Within our model for a clade on n extant species we develop fomulas for various quantities; these are derived as $n \to \infty$ limits, but one can check via simulation that they are usually reasonably accurate for moderate values of n.

- A "local" description of the probability structure of lineages, which permits easy calculations (Section 5.1).
- A "loss of evolutionary history under random extinctions" calculation (Section 5.2).
- A formula for the joint distribution of time back to origin of clade; time back to last common ancestor; number of species at that time (Section 5.4).
- Number of extinct species (Section 5.5).
- Chance that a fossil is ancestral to some extant species (Section 5.6).

1.2.3 Properties of the model for higher level taxa

`xxx This is work in progress. Section 7.1 outlines what we hope to do.`

In this version we present only the following fragments. Throughout the paper we write "genus" for an arbitrary higher level taxon.

- Shape of phylogenetic tree on extant genera (Section 7.2).
- Fluctuation rates for time series of number of genera (Section 7.3).
- Number of species per genus (Section 7.4).

1.2.4 Comments

Our fundamental purpose in this project is to illustrate the wide range of calculations possible within a coherent model. We re-emphasize that we are using a model which deliberately lacks biologically-motivated assumptions; but there would be no difficulty in principle in repeating calculations within models claimed to be more biologically realistic.

Note that the *n law* (Section 1.2.1) makes our model unrealistic for large extant clades: with the usual "a few million years" estimate for mean species lifetime [39], a 200-species clade originating less than 30 Myr ago would be implausible within the model.

Finally, we record a simulation study (section 5.7) of estimates of past within-clade speciation and extinction rates based only on the phylogenetic tree of lineages of extant species; our study casts doubts on the ability of such data to provide even crude estimates reliably.

1.3 Specific projects in quantitative macroevolution

Here are five specific questions suggested by our work, which we would like to draw to the attention of experts.

1.3.1 Fluctuations at different hierarchical levels

Sepkoski's compendia [43] (see also Benton [7]) are justly celebrated for providing raw data for the statistical study of long-term evolutionary history. Because of the difficulty of resolving fossils to the species level, this data is typically presented as time series for numbers of genera and families. Consider summary statistics
$\mu_g[\mu_f]$ = mean lifetime of a genus [family]
$\mathrm{var}_g[\mathrm{var}_f]$ = fluctuation rate for genera [families]
G = mean number of genera per family.
Within any probability model there will be theoretical relationships between these quantities.

Project. Compare the relationships between these quantities observed in data with the predictions of a probability model.

Of course, such questions have previously been studied (see section 7.5 and the quote in section 3.2), in part because deviations from randomness may be relevant to issues such as *competitive or expansive?* (Section 3.3). However, the mathematical models previously used seem rather haphazard; our model provides a more coherent framework for making predictions. Some

preliminary results are given in Section 7.3. We study a notion of *normalized* fluctuation rate for a given taxonomic level, where the normalization is such that the rate would equal 1 if we were modeling that level directly as a "pure chance" process. Thus in our model the normalized rate for *species* equals 1 because we do model species fluctuations as being pure chance. For higher taxa the rate is not 1 because fluctuations in taxons of species numbers are derived from fluctuation numbers via a scheme for assigning species to (say) genera. Table 5 shows that, under one plausible classification scheme for fossil taxa, the normalized fluctuation rates for taxa of average size 10 species drops to 0.68. Conceptually, the point is that such reduced fluctuation rates are predicted as an artifact of hierarchical classification rather than as a consequence of some biological effect.

1.3.2 Phylogenetic tree shape and hierarchical level

It is a longstanding puzzle [27] that real phylogenetic trees seem more "imbalanced" than predicted by a natural *Markov model*, though more balanced than a (less natural) *uniform* or *PDA (proportional to distinguishable arrangements)* model. Many possible biological explanations for imbalance have been proposed, including "artifact of higher-level classification". Our model provides a framework for studying such questions, coherently with study of fluctuations described above.

Two studies [17, 26] of published small trees discussed in [27] conclude there is no such hierarchical trend in imbalance, but do indicate that more complete trees tend to be more balanced than less complete trees. These, and most other, studies used a summary statistic to measure imbalance of each tree. A different method, used in [4] on a few large trees, seems less arbitrary and more powerful. Each branchpoint of a binary tree splits a clade of size m (say) into subclades of sizes a and $m - a$, where we take $a \leq m/2$ as the size of the smaller daughter clade. Given a collection of trees, take all the splits in all the trees, and then calculate the function

$$a(m) = \text{median size of smaller daughter clade in split of size-}m\text{ clade.}$$

This function provides a measure of "balance" in a collection of trees which has three advantages over using summary statistics (uses more within-tree structure; avoids arbitrary choice of summary statistic; avoids issues of normalization required to compare different size trees). As an illustration of what can be done within our model, in section 7.2 we study the shape of the phylogenetic tree on extant genera arising from a plausible classification scheme for extant species. Table 4 indicates that imbalance (as measured by

$a(m)$) does increase with average number of species per genus, though this increased imbalance is more prominent within smaller clades. This analysis suggests that observed imbalance in trees on higher level taxa may be an artifact of classification.

Project. Use published data to compare the functions $a(m)$ for complete trees on species, on genera, on families, etc.

1.3.3 Extant ancestral species

Our model predicts (6) that for about 63% of extant species, some ancestral species should be extant. While this numerical value depends on rather arbitrary details of the model, general mathematical principles show that for any model incorporating extinctions and speciations which are not "tightly coupled" in some way (see example below), the model will predict that some constant percentage (not close to 0%) will have extant ancestors.

Project. Find data to estimate, within well-studied extant clades, the proportion α of extant species with some extant ancestor.

Anecdotally, biologists regard α as small, though we have been unable to find useful data, perhaps in part because cladistics dogma discourages asking this question. Models like ours assume a species is a well-defined entity with a time of origin and a time of extinction (this idea is *cladogenesis*), in contrast to *anagenesis*, meaning change along an unbranching lineage. If data indeed confirms that α is small, then it is evidence in favor of at least moderate prevalence of anagenesis. Moreover, any kind of statistical study of question such as "what was the speciation rate of this particular clade during its radiation" implicitly depends on models of the type which predict non-small α, and so all such work would appear less convincing if data reveals α to be small.

To make a cladogenesis model with small α, one needs a model such as the following. Take a small parameter $\theta < \frac{1}{2}$, Suppose that for each species, the following events occur with relative chances

extinction	θ
speciation	θ
replacement by a daughter species	$1 - 2\theta$.

Such models will give a small value of α (specifically $\alpha = (1 - e^{-2\theta})2\theta$), precisely because they are effectively interpolating between cladogenesis and anagenesis (smaller θ giving a larger contribution of anagenesis).

1.3.4 Past fluctuations in sizes of extant clades

There is an intuitively appealing biological explanation of readily-identifiable clades. A successful clade begins with a *key innovation* in one species, followed by a rapid *adaptive radiation* of species sharing that innovation; clade size increases until a level set by ecological constraints, and stays at roughly this maximum level (while individual species arise and disappear) until some extrinsic factor upsets the equilibrium. Notwithstanding textbook examples of clades (horse, rhinoceros) which were much larger in the past, some version of this "logistic" picture is often taken to be self-evident, as the following quote (our emphasis added) indicates.

> [We study] theoretical clades that have either been growing exponentially throughout their history or have been of constant size, such that each time a new lineage has appeared by speciation another lineage has gone extinct. *These extremes bracket the plausible dynamical histories of real clades.* .... Logistic growth, in which diversity rises to some maximum, is a convenient model for macroevolutionary clade expansion .... In this framework, exponential growth is the early phase of logistic growth, and the constant size model describes a clade that has been at its maximum size for some time. (Nee and May [30])

This may be a perfectly reasonable view of large clades (flowering plants, birds, mammals), but what about small clades? A particular way to think about past fluctuations of clade size is to consider the quantity

$$R = \frac{\text{max number of species at any past time}}{\text{current number of species}}.$$

Here $R \geq 1$ because we include "current time" in "any past time". The standard view, as quoted above, is that typically R will be close to 1. In contrast, our model predicts (11) the $1/r$ law:

$$P(R \geq r) = 1/r$$

so that R would vary widely between clades, with a median value of 2. This prediction may seem unrealistic to biologists, but is it less realistic than predicting $R = 1$?

Project. Look at small extant clades (size $5 - 40$, say) with extensive fossil records, and attempt to estimate the distribution of R from the fossil record.

The next project addresses a similar issue in a different way.

1.3.5 Variability of realizations, and conservative analysis of biological significance

Though the intrinsic variability of realizations of stochastic models of macroevolution has often been noted, the classical statistical practice of comparing *averages* of quantities derived from data with averages predicted by models often makes it hard to keep variability in mind. Figure 5 later provides a dramatic illustration. That figure shows three quantities associated with our model (time of clade origin; time of last common ancestor of extant species; number of species at time of last common ancestor) and shows 10 realizations; each of these quantities varies by a factor of 10 over the realizations.

Space constraints of print journals used to make it impractical to show pictures of multiple realization of stochastic models, but the Web has no constraint, and authors of new models should routinely post simulations. Our site [1] shows, for a selection of values of n, 10 realizations of the phylogenetic tree on n extant species derived from our model. Figure 3 later shows three out of 10 realizations for $n = 20$. If we saw three real trees with such radically different radiation patterns and times, then we would surely be inclined to attribute biological significance to the differences. But for our model, no one of the three trees is particularly unlikely. Our model, with its intrinsic greater variability, therefore provides a *more conservative* approach to assessing significance of observed features of phylogenetic trees.

More concretely, consider the problem of estimating past speciation and extinction rates within a clade using only the phylogentic tree on extant species. Our simulation study (section 5.7) shows that, if our model were the true underlying model, then estimating parameters in a standard birth-and-death model would give wildly variable and unreliable estimates. This casts doubt on the whole prospect of estimating rates from such data, in the context of a single clade. But in contrast, one can hope that a statistical study of many clades would provide some insight into typical patterns of macroevolution.

Project. Assemble a database of phylogenetic trees (with relative time scale) on extant species, for statistical study of macroevolutionary process (as a complement to the fossil compendia [7, 43]).

Existing databases such as TreeBASE [50] record cladograms, but these are far less useful for the studies we envisage, such as

Project. Use such a database for a careful study of whether, in the context of extant small clades, stochastic models designed to exhibit logistic/exponential growth provide a better fit than alternate models with the same number of parameters.

2 Standard stochastic models

2.1 The basic picture

Our conceptual "basic picture" for macroevolution is that different species are different entities; each species originates at some time as a "daughter" of an existing species; each species survives until some extinction time (or the present time). See figure 1; each species is represented as a vertical line running downwards from time of origin to time of extinction, with the parent-daughter relation indicated by a horizontal line. This basic picture can of course be criticized in many ways (over details of speciation, since it assumes speciation is relatively rapid; and for ignoring *anagenesis*, that is change along an unbranching lineage) but as a reasonable conceptual simplification it is uncontroversial. We study probability models imposed on this basic picture. Our viewpoint is that anything one does involving phylogenetic trees and probability models (trees on higher-level taxa; inference in the context of unseen extinct species) should be consistent with the basic picture. This should also be uncontroversial, though is rarely emphasized in the literature.

2.2 The Yule model

Yule [53] proposed the basic model for speciations without extinctions. Initially there is one species; in a time interval $[t, t+dt]$ each species has chance dt to give rise to a daughter species. In this model the number of species grows exponentially with time. So for given n one can get a model for an n-species tree by taking the present as a random time at which the number of species equals n.

Though this species model is familiar nowadays, the main point of Yule's work is invariably overlooked. He superimposed a model of *genera* by supposing that, from within each existing genus, a new species of new genus arises at some constant stochastic rate λ. This leads to a one-parameter family of long-tailed distributions for number of species per genus (see [4] for brief description). Yule's model perhaps foreshadows "hierarchical selection above the species level" [15]; in contrast, our model for higher-level taxa (Section 6) doesn't involve separate genus-level biological effect, but rather combines species-level novelty with conventions about how systematicians construct genera.

2.3 The Moran/coalescent model

These models, developed and extensively used in population genetics, can also be applied to macroevolution (see e.g. [19]). In the Moran model ([12] sec. 3.3) the number of coexisting species is fixed at n. At successive discrete times, one randomly-chosen species goes extinct and another randomly chosen one speciates. Implicit in this model (run from the indefinite past until the present) is a model for the phylogenetic tree on the n extant species; for large n, with suitable rescaling of the time unit, the phylogenetic tree approximates the continuous-time *coalescent* model. To describe the coalescent model, we run time backwards from the present, starting with n "lines of descent"; in a time interval dt, each *pair* of lines of descent has chance dt to merge ("coalesce") into one line, and we continue until reaching a single *last common ancestor*. See [25] for a recent survey.

2.4 Conventions: trees, time units, parameters

In the paper we use the term *phylogenetic tree*, or just *tree*, to mean a tree with a time scale; speciations and extinctions occur at definite times. Published data on extant species typically just shows the topology of such a tree, without times at which lineages diverged. We use the term *cladogram* for such trees – it seems helpful to emphasize the distinction. We are concerned with models for phylogenetic trees, but of course such a model automatically induces a model on cladograms.

To compare different models, one needs a convention which relates the time unit within the model to real-world time. A natural convention is to take an "evolutionary time unit" (ETU) to be the mean lifetime of a species until extinction. With this convention, the discrete steps in the Moran model take $\frac{1}{n}$ ETU, and the time unit in the coalescent model is 1 ETU. We will use this convention (for the Yule model, without extinctions, one has to fudge, but this is not important).

The models above are zero-parameter models[2].

2.5 Critical branching process

Raup et al. [16, 40] proposed a model for extinct clades which is essentially what mathematicians call the *continuous time critical binary branching process* (CBP). In this model, each lineage becomes extinct at constant rate λ

[2]More precisely, the real-world parameter "mean species lifetime" is absorbed into the ETU. In physicists jargon, there are zero *dimensionless* parameters.

(that is, with chance $\lambda\ dt$ in each interval of length dt), and also gives rise to daughter species at the same rate λ. So $\lambda = 1$ per ETU. Starting with a single lineage, the clade is likely to become extinct quickly, but one can study the large clades which sometimes arise by chance. When the number of lineages n is large, the CBP behaves over short times like the Moran model, though of course the difference is that in the CBP the number of lineages fluctuates randomly with time. Note that CBP is also a zero-parameter model.

2.6 Birth-and-death models

What mathematicians call *birth-and-death models* are conceptually natural for modeling the fluctuations of number of lineages with time. The general (infinite-parameter) such model has parameters $(\lambda_i, \mu_i,\ i \geq 1)$. When the total number of lineages is i, then at rate λ_i some lineage splits, and at rate μ_i some lineage dies; to model a tree, in each case we choose the lineage uniformly at random. Thus the Yule model has $\lambda_i = ci,\ \mu_i = 0$ and the CBP has $\lambda_i = ci,\ \mu_i = ci$. The *linear* birth-and-death model has $\lambda_i = \lambda i,\ \mu_i = \mu i$. This model (with $\lambda > \mu > 0$) is often used to model growing clades, though its mathematical property of either rapid extinction or ultimate exponential growth is of course unrealistic. Note that $\mu = 1$ per ETU, so this is a one-parameter model.

Many variants of birth-death models have been proposed, though ironically the key debate between logistic and exponential growth (Section 3.3) has been generally discussed via deterministic models (Sepkoski [46]) rather than stochastic ones.

2.7 Cladogram shape

If we use any of these models to produce a tree on n extant species, and then reduce to the cladogram linking these n species, the probability model on n-species cladograms is the same *Markov model* regardless of the original model on trees.[3] Essentially, starting with one species one can let the number of species fluctuate with time in an arbitrary deterministic or random way to end with n species; as long as the model has the *exchangeability* property

[3] Using the CBP model of a large *extinct* clade, and then randomly sampling a relatively small number n of species and looking at the cladogram on sampled species, gives a different distribution on cladograms: the *uniform* distribution. This fact, and a related simple description of the phylogenetic tree on the sampled species, play a central role in recent mathematical study [3] of random trees outside the context of biological evolution.

when some species goes extinct, it is a uniform random species;
when some species speciates, it is a uniform random species.

we get the Markov model on cladograms. See [4] for recent discussion.

2.8 Other models

We have listed above only the mathematically simplest models; one can invent an unlimited number of further models incorporating conjectured biological process. One can think of speciation and extinction rates as being time-dependent rather than size (of clade)-dependent. Or these rates could be variable and partially inherited down lineages [18].

There is a rather disjoint literature by statistical physicists (e.g. [32]), designing models of macroevolution to exhibit behavior (scaling laws, self-organized criticality) found in other physical settings.

3 Biological use of stochastic models of macroevolution

The uses to which these stochastic models have been put are surprisingly varied, but can roughly be fitted into three categories.

(a) *Inference about macroevolutionary process.* When we see some notable feature in the evolutionary record of speciations and extinctions, we are inclined to assume it must have some biological significance. But if the same feature would likely be observed in a hypothetical model of "purely random" macroevolution, then the evidence for biological significance is weak.

(b) *A convenient way of doing speculative calculations.* What proportion of identified extinct species might be direct ancestors of extant species? What effect would extinction of some proportion of extant species have on the overall diversity of extant species?

(c) *As a component in algorithms for reconstructing trees from data using statistical methods.* Practical tree reconstruction is a huge discipline [33, 42]. In brief, parsimony methods pay no attention to any probability model for trees; maximum likelihood methods implicitly assume (from a Bayesian viewpoint) that all possible trees are *a priori* equally likely; and Bayesian methods use an explicit prior distribution on trees. See [20] for recent discussion of the Bayesian methodology.

While (c) is an important area it is remote from our concern in this paper; we are envisaging that the literature shows "true trees" with which to compare the predictions of models. What we are actually doing in this

paper is (b), as summarized in Section 1.2. But our underlying motivation comes from (a), and so we will briefly recount some of the issues for which our simple models, or more realistic modifications thereof, might provide insight.

3.1 Assessing goodness of fit of stochastic models

To what extent stochastic models fit real data is of course the $64,000 question. We do not attempt to answer it here, or survey the literature thoroughly, but confine ourselves to brief remarks. To fit time-series data from a particular clade (whether total number of taxa in an extinct clade, or number of lineages of an extant clade) to one of these birth-and-death type models by estimating parameters is a well-understood aspect of inference for stochastic processes [6]; for representative biological literature see e.g. [29, 34, 48] and see [28] for further citations. A *statistical test of significance* is a tool for answering a very particular question: does data provide strong evidence to overturn an artificial presumption that a null chance model is precisely true? While tests of significance have often been done in the context of diversity models ([22, 52] and many others), we regard them as a misplaced emphasis. For biodiversity, all we should expect is that a model might be a crudely accurate representation for some clades while being grossly inaccurate for others, and formal tests of significance are simply not designed to compare qualitative "goodness of fit". To argue convincingly that (say) the logistic model is the best explanation of clades of a particular size or duration, one should compare that model with another model *with the same number of parameters* representing some biological alternative explanation, and this apparently has never been done.

3.2 Long term macroevolutionary process

Jablonski [21] gives a concise account of paleontologists' views of the resolved and unresolved issues in long-term macroevolutionary process. Many of these issues (major mass extinctions; the apparently non-random origin in time and location of major evolutionary innovations) are not amenable to simple stochastic modeling of the kind we consider, though our kind of model, maybe realistic for shorter periods, seems relevant to certain aspects:

> The complex trajectory of taxonomic diversity through [600 Myr] has proved robust to continued sampling and, as shown by simulations, to very different phylogenetic approaches to grouping

> species into higher taxa. But diversity time series become increasingly jagged and disparate at lower taxonomic levels and on regional scales, both because sampling is less complete and because *lower-diversity lineages really are almost inevitably more volatile*. [21] (our emphasis added).

Note that, despite obvious possible sources of bias in the fossil record (older groups may be less frequently preserved; taxonomic boundaries may be drawn more broadly for ancient groups than for recent groups), paleontologists are adamant (e.g. quote above and [10, 44]) that the large-scale pattern of diversity through time provided by the fossil record is generally accurate.

Kirchner and Weil [5] unearthed the surprising result that fossil data shows that the cross-correlation between the time series of extinctions and originations is maximized at a time lag of of about 10 Myr (originations later than extinctions), whether the "Big Five" mass extinctions are included or excluded.

3.3 Logistic versus exponential growth

An ongoing debate, nicely summarized in Benton [8], concerns whether a typical radiation of a group is primarily *competitive* (replacing existing groups) or *expansive* (occupying new niches). The former suggests a logistic-type curve for species diversity as time increases, whereas the latter suggests an exponential-type curve. To quote [8]

> It is hard to provide a clear test of which of these kinds of curves fits the [very long-term] data best. In all cases, investigators accept that the curve fits are not perfect, since the patterns of generally increasing diversity are offset by many drops in diversity, some associated with major mass extinctions, others with extinction events of more local scale, or affecting only certain taxa. When such perturbations are excluded, proponents of the exponential and logistic models claim to have found curves that fir the empirical data well ..... Large-scale plots of the diversification of life seemingly cannot yet distinguish between patterns of unfettered expansion (exponential curves) and those of long-term steady-state conditions (logistic curves). This is an important problem to resolve, because it goes to the heart of our understanding of evolution: do species evolve within a tight

straightjacket imposed by their interactions with other organisms (the equilibrium view), or has much of evolution been limited only by the capacity of organisms to enter new ecospace (the expansionist view)?

3.4 Short term macroevolutionary process

An elegant use of shorter term stochastic models has been made by McKee [23, 24]. The fossil record of large mammals in southern and eastern Africa over the last 3 My shows apparent "pulses" of extinctions and speciations. Is this real, or could it be an artifact arising from the limited number of different dates of sites yielding the fossils? McKee studies this by comparing data with simulations from a Moran-type model (as a null hypothesis) and with simulations from a variant assuming pulses.

4 The model: species level

Here are three different trees one can associate with an extant clade. Call the tree linking all extant and all extinct species the *complete tree*. Call the subtree recording only the extant species and their ancestor species the *ancestral tree*. Call the tree showing divergence times of extant species, without identifying linages at past times with particular extinct species, the *lineage* tree.

We now propose coherent models for such trees with a given number n of extant species. We will take for the underlying model is the critical binary branching process (CBP), from Section 2.5 : each species becomes extinct at rate 1 and produces daughter species at rate 1. We will measure time $t > 0$ backwards from the present time $t = 0$: "t decreases" is equivalent to "time runs forwards".

4.1 The model for clades on n extant species: definition and simulations

We define the model as follows.
(a) The clade originates with one species at a random time in the past, whose prior distribution is uniform on $(0, \infty)$.
(b) As time runs forward, each species becomes extinct or speciates at rate 1, as in the CBP model.
(c) Condition on the number of species at the present time $t = 0$ being exactly equal to n.

The "posterior distribution" on the evolution of lineages given this conditioning is then a mathematically completely defined random tree on n extant species, which we write as $c-\mathrm{TREE}_n$ (here c is mnemonic for *complete*) [4]. A realization of this tree then also determines a realization of the ancestral tree on extant species, which we write as $a-\mathrm{TREE}_n$ (where a is mnemonic for *ancestral*), and a realization of the lineage tree on extant species, which we write as $l-\mathrm{TREE}_n$ (where l is mnemonic for *lineage*),

Associated with these random trees are a variety of numerical-valued random quantities, in particular:

$$\begin{aligned} C_n(t) &= \text{total number of species at time } t \\ A_n(t) &= \text{number of species at time } t \text{ in the ancestral tree} \\ &\quad \text{(equivalently: within the lineage tree)} \\ T_n^{\mathrm{origin}} &= \text{time of origin of clade} \\ T_n^{\mathrm{lca}} &= \text{time of last common ancestor of extant species.} \end{aligned}$$

Figure 1 shows a realization of the complete tree $c-\mathrm{TREE}_n$ for $n=20$, together with the quantities defined above. Figure 2 shows the restriction of that realization to the ancestral tree $a-\mathrm{TREE}_{20}$. (These trees are drawn in a particular objective way described in the legend; usual ways of drawing large trees with a time scale, as opposed to cladograms, seem to involve more arbitrary layout decisions to ensure that small subclades are drawn near their parents. Repeats of Figures 1, 2 and 7 with different realizations can be found on the web site [1].)

The underlying idea of the model – critical branching, which is equal-rates speciation and extinction – has of course been used before. Raup et al ([40] and subsequent papers) explored such models for extinct clades via simulation. Basic mathematical analysis of models for extant clades has been done in Hey [19], Nee et al. [31] and subsequent papers. Our particular $c-\mathrm{TREE}_n$ model has been considered via simulation in Wollenberg et al [52], though we regard their work as somewhat flawed in detail for reasons explained in Section 8.2.

[4] In (a) we use an *improper* [total probability is infinite] prior distribution, but after conditioning the posterior distribution of $c-\mathrm{TREE}_n$ is *proper* [total probability is 1].

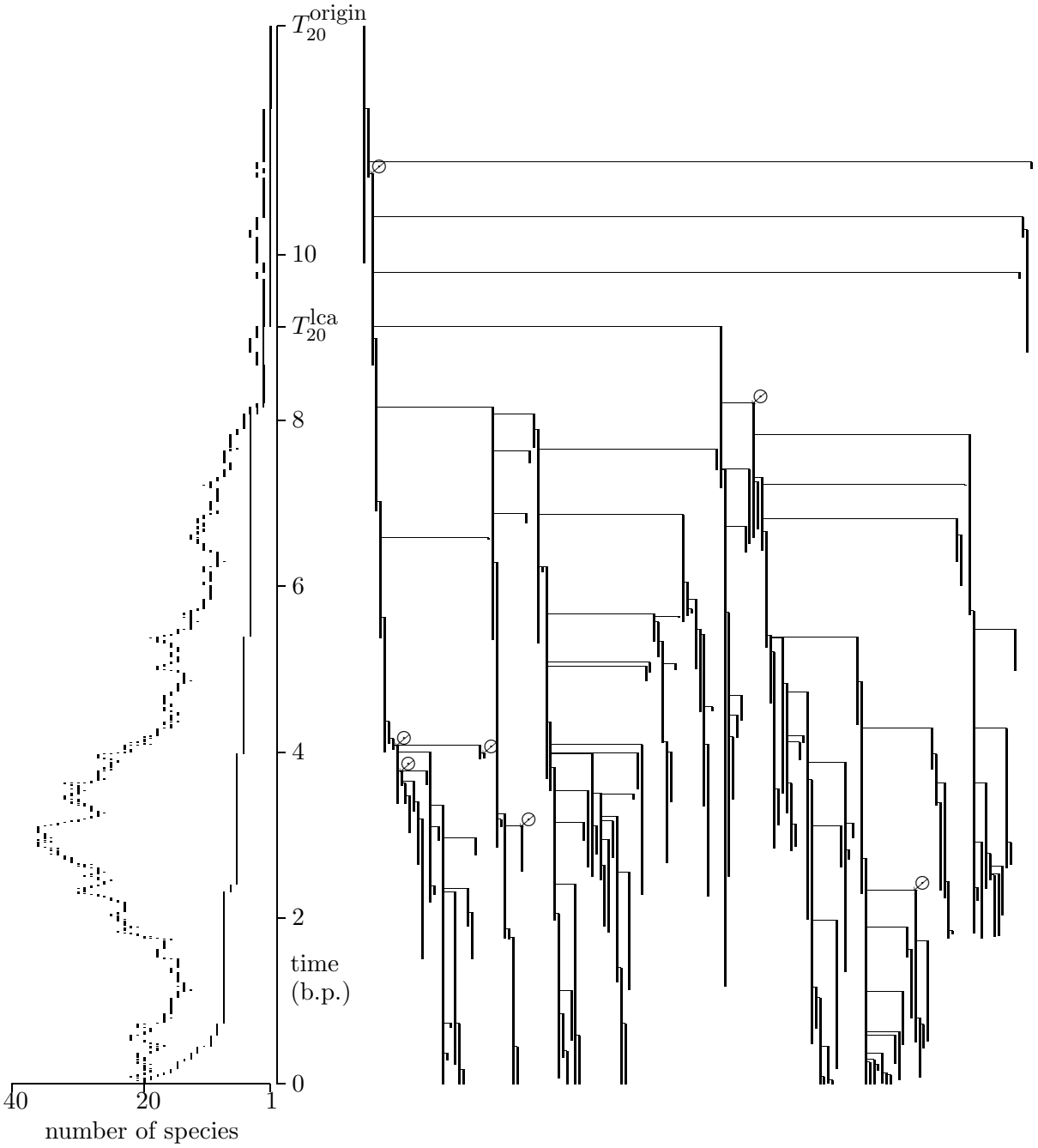


Figure 1. A realization of $c-\text{TREE}_{20}$, a complete clade on 20 extant species. The figure is drawn so that each species occupies a vertical line (from time of origin to time of extinction (or present)), different species evenly spaced left-to-right (so that each subclade is a consecutive series), using the convention: daughters are to right of parents, earlier daughters rightmost. On the left is a time series of numbers of species: the outer line is total number of species, the inner line is number of ancestors of extant species. The $\odot$ are marks used later to construct genera. In this realization there were a total number 142 of extinct species, with a maximum of 38 species at one time; $T^{\text{lca}} = 9.05$ and $T^{\text{origin}} = 12.75$.

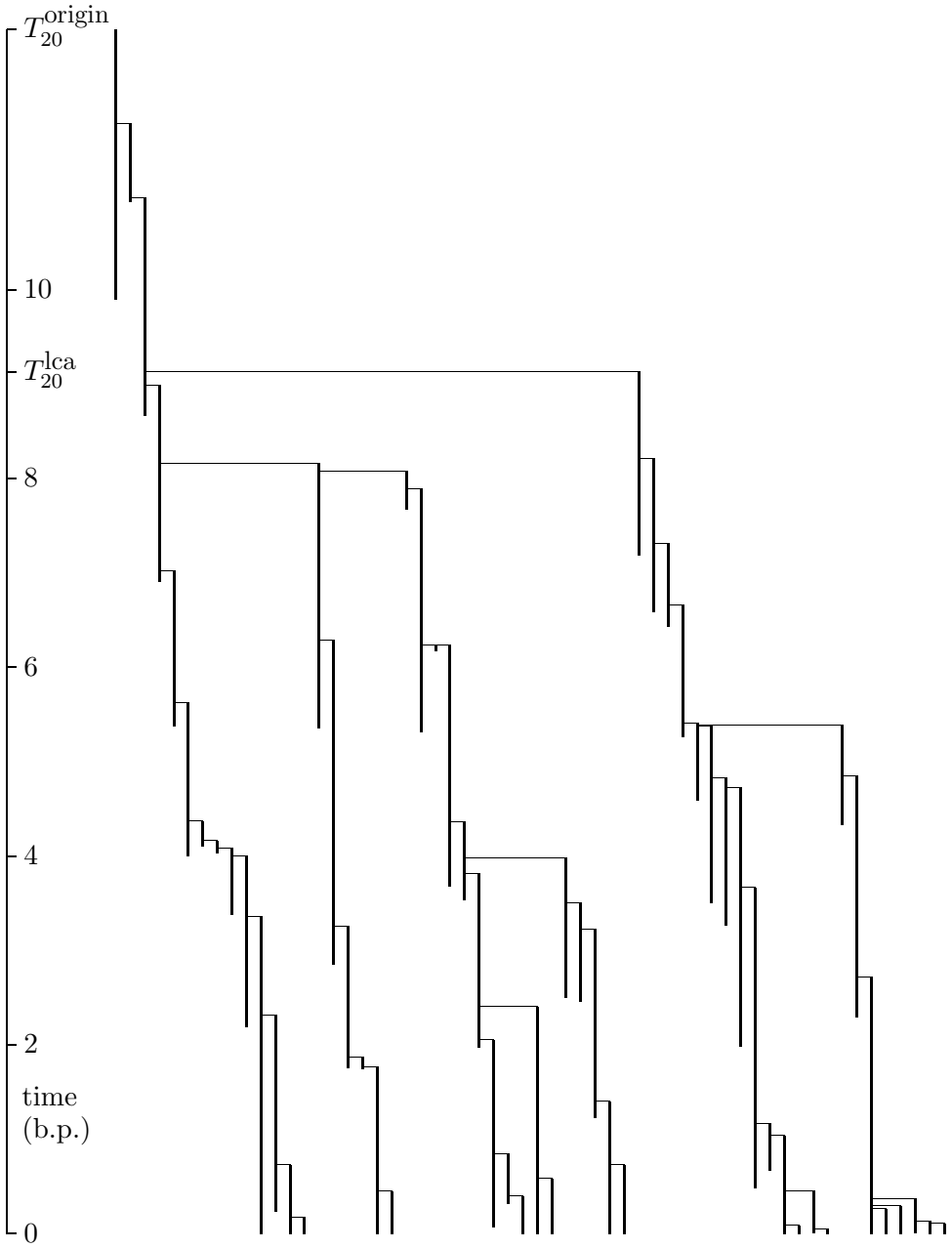


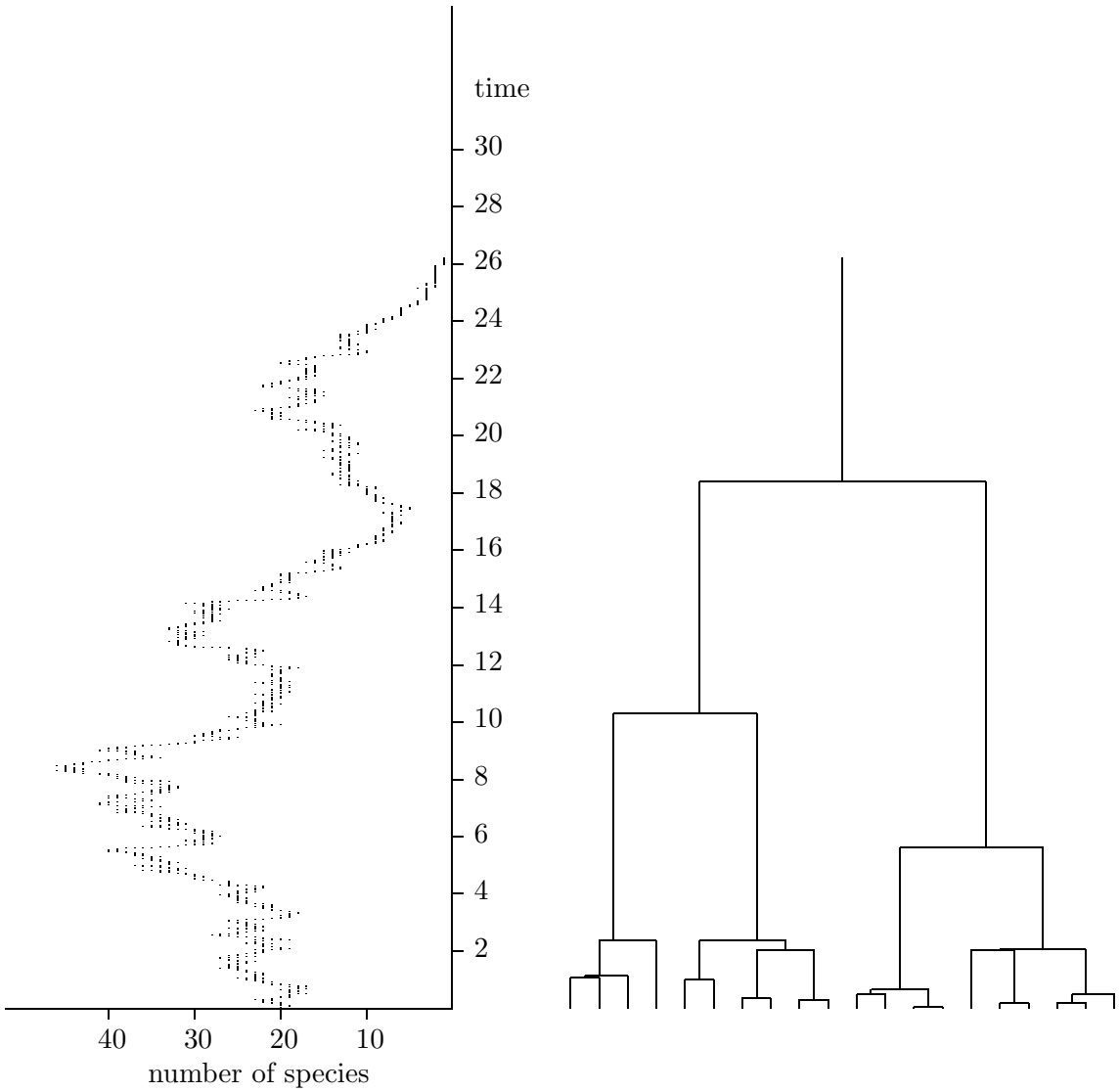
Figure 2. The realization of $a - \text{TREE}_{20}$ derived from figure 1, showing only extant species and their ancestors. There were 38 extinct ancestral species.

4.2 The lineage tree on extant species

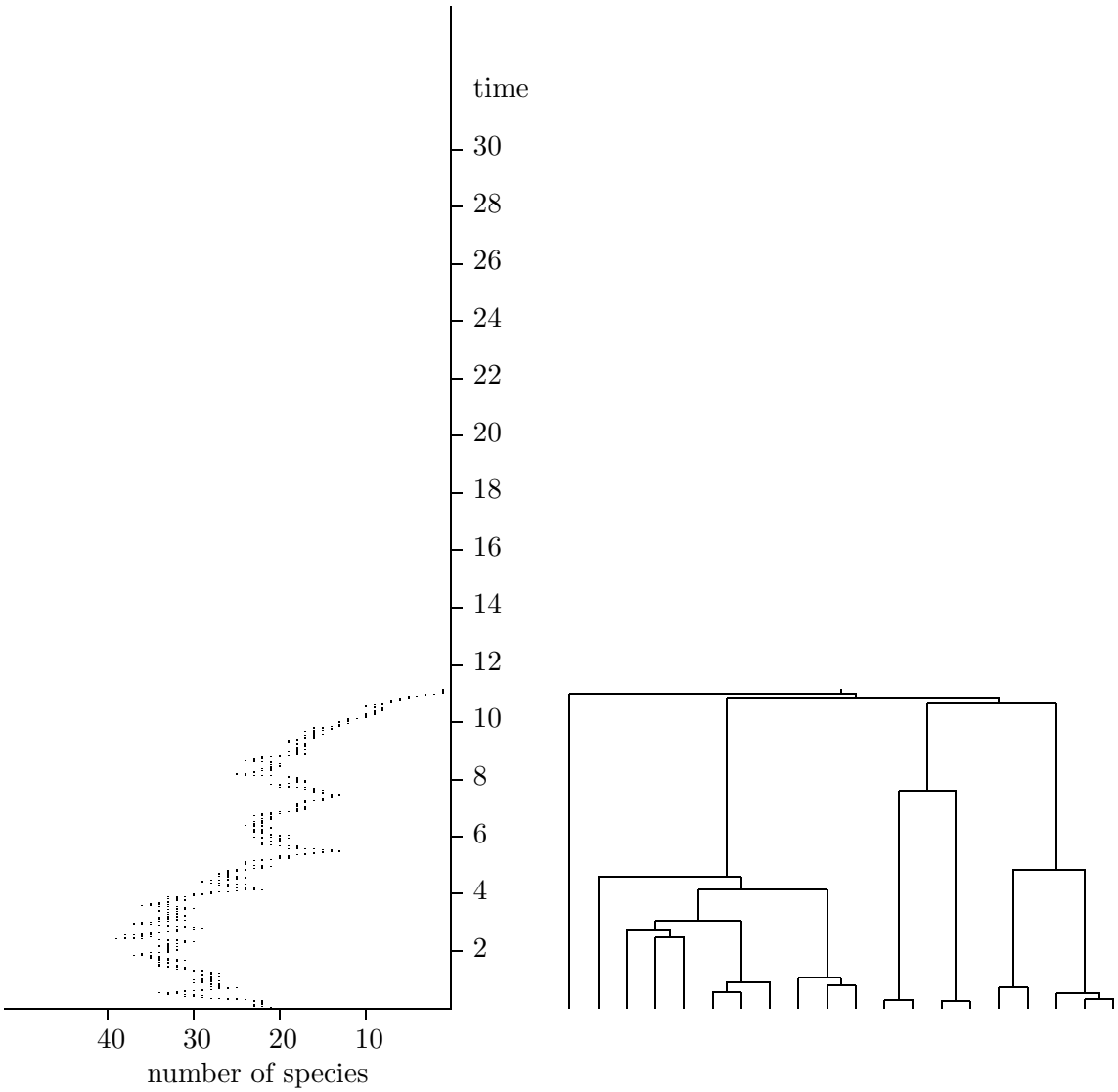
The trees above distinguish different ancestral species and their lifetimes; in the *lineage tree* we just record lineages of extant species. This is the familiar phylogenetic tree obtained from molecular data on extant species.

In Figure 1 we deliberately chose a realization which was "typical", in that various quantities of interest are close to their median values. But part of our general message is that there is no such thing as a "typical" realization, because realizations vary dramatically. Our web site [1] shows 10 realizations of the lineage tree on n species, for each of $n = 8, 12, 20$. Figure 3 shows three of the $n = 20$ realizations.

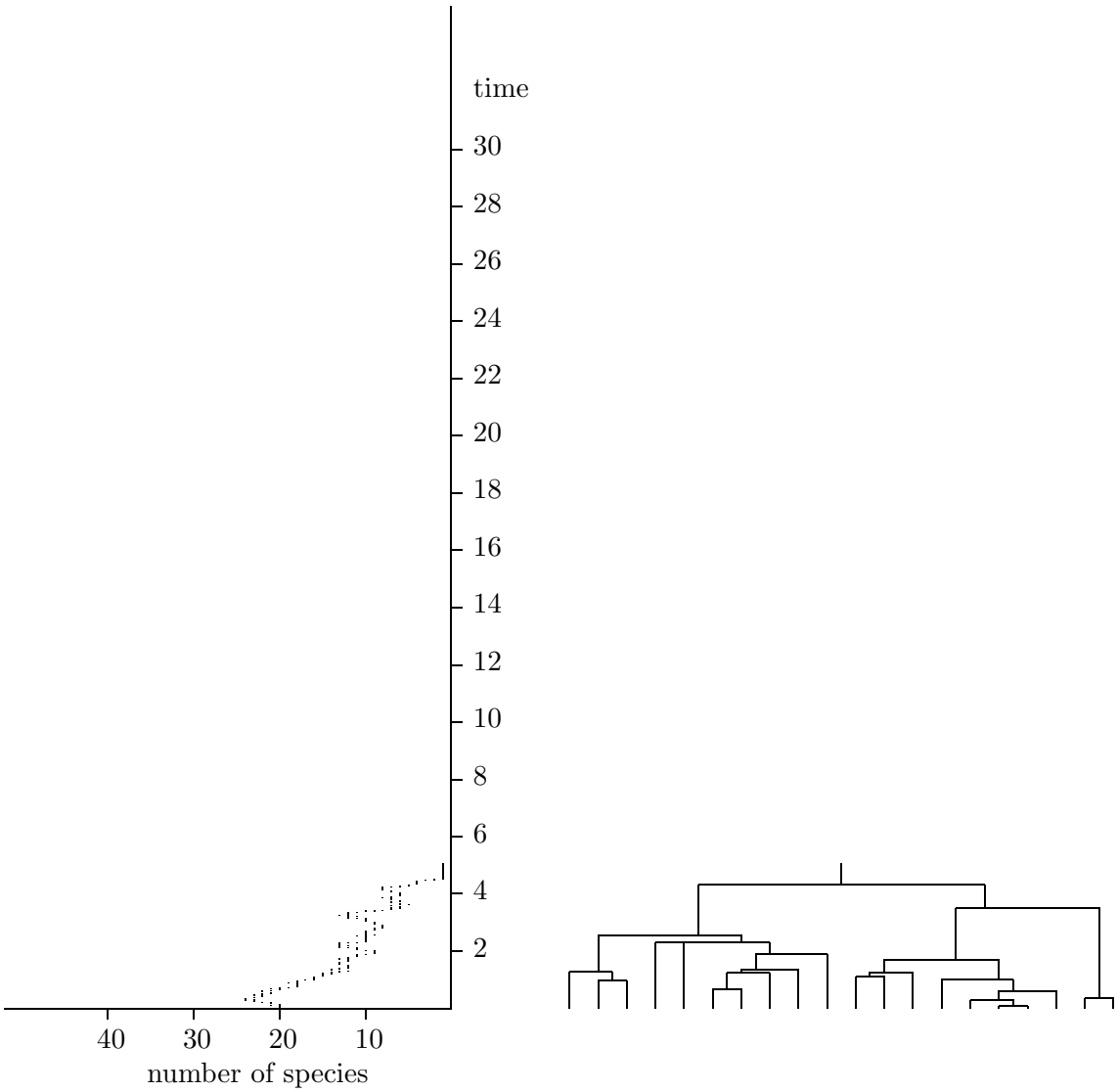
Figure 3. Three realizations of the lineage tree in our model, with $n = 20$. These are realizations $1, 5, 6$ from [1].



Number of extant species	20
Time of last common ancestor	18.40
Time of origin of clade	26.21
max number of species at one time	46
R = (max number species)/(current number species)	2.30
Number of extinct species	532



Number of extant species	20
Time of last common ancestor	10.99
Time of origin of clade	11.15
max number of species at one time	39
R = (max number species)/(current number species)	1.95
Number of extinct species	255



Number of extant species	20
Time of last common ancestor	4.33
Time of origin of clade	5.06
max number of species at one time	24
R = (max number species)/(current number species)	1.20
Number of extinct species	39

5 Mathematical properties: species level

Our methodology will be to write down approximate formulas, derived from $n \to \infty$ asymptotic results. Of course such "large n" results may seem inappropriate, since as mentioned in the introduction we are envisaging our model as realistic only for comparatively small clades. But they do serve to give qualitative insight about a wide range of features; and for small n one could give numerical calculations or simulations for any particular features of biological interest.

5.1 Local structure of $l - \text{TREE}_n$

A rigorous mathematical result (see Theorem 5 of [36]) shows that the "posterior distribution" of our model has an asymptotic stochastic limit as the number of extant species increases. This result implies a simple approximate description of the local[5] structure of the lineage tree on extant species[6] $l - \text{TREE}_n$, for "large n". One can also think of this result as the exact description of the local structure of the lineage tree of a hypothetical infinite clade $l - \text{TREE}_\infty$.

Construction of $l - \text{TREE}_\infty$ [36]. First put the extant species at positions $\ldots, -2, -1, 0, 1, 2, 3, \ldots$ on the horizontal axis. At each midpoint $\ldots, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5, \ldots$ we put a branchpoint $\times$ at a random height, these heights being independent with probability density function

$$f(t) = (1+t)^{-2}, \quad t > 0.$$

These branchpoints determine the phylogenetic tree; one recipe is that the lineages between species i and species $i+1$ diverge at the time of the branchpoint above position $i + \frac{1}{2}$. Figure 4 shows an example of 20 species in a realization of $l - \text{TREE}_\infty$; the species are labeled $\{-9, \ldots, 10\}$.

Elementary calculations. While some calculations within $l - \text{TREE}_\infty$ are implicitly done in the existing literature as asymptotics within other models, our construction makes many calculations quite easy. Here are some examples.

[5] *Local* is math jargon for "within order 1 ETU time of the present". Similarly *global* means "over time-scale of order n", i.e. back to the origin of the clade, and *intermediate* means intermediate between these time scales.

[6] The same process $l-\text{TREE}_\infty$ arises as the limit in the Moran model; this is a reflection of the fact that the Moran model and CBP are similar over short and intermediate time periods.

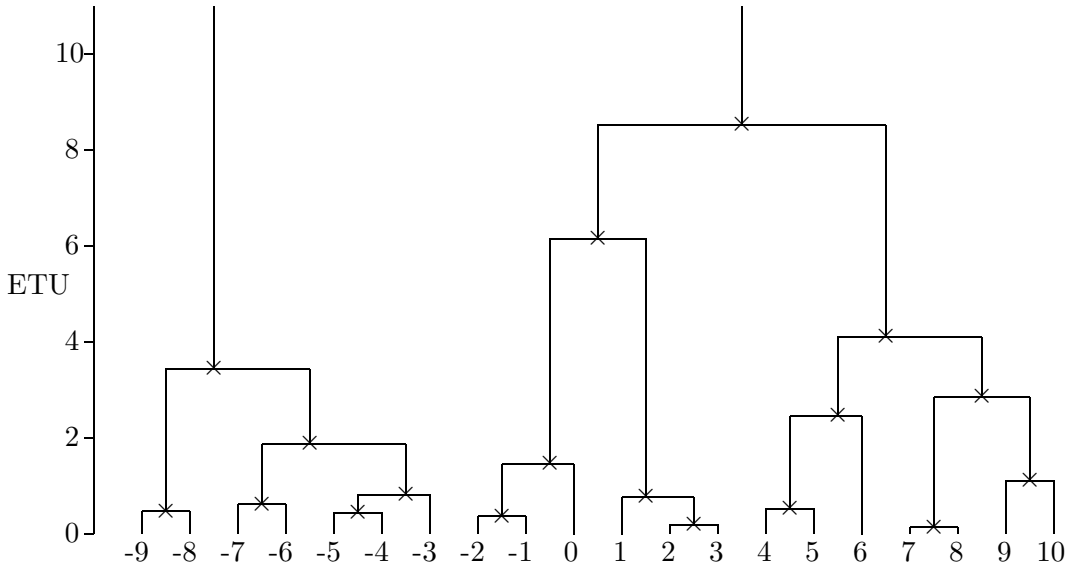


Figure 4. A realization of part of $l - \mathrm{TREE}_\infty$, approximating the local structure of $l - \mathrm{TREE}_n$ for large n. The 2 visible ancestral lineages diverged at around 16 ETU.

(a) The density of lineages (number of lineages relative to number of extant species) at time t is just the density of branchpoints at height greater than t, that is

$$G(t) = \int_t^\infty f(s)ds = (1+t)^{-1}.$$

The size of (that is, number of extant species descended from) a typical lineage at time t has geometric($\frac{1}{1+t}$) distribution

$$p_t(i) = \left(\tfrac{1}{1+t}\right)\left(\tfrac{t}{1+t}\right)^{i-1},\ i \geq 1 \tag{1}$$

because this is the distribution of distances between branchpoints at heights greater that t.

(b) As t increases (time runs backwards) a lineage merges with some other lineage at stochastic rate

$$\mathrm{merge-rate}(t) = 2\tfrac{f(t)}{G(t)} = \tfrac{2}{1+t}$$

because such a merger occurs in $[t, t+dt]$ when one of the two branchpoints separating the given lineage from its neighboring lineages, which must be

at height $\geq t$, occurs during $[t, t+dt]$, and this has chance $f(t)dt/G(t)$ for each branchpoint. Moroever, if a lineage merges at t then (independent of the size of the first lineage) the size of the second lineage has the geometric distribution $p_t(\cdot)$ at (1).

(c) As t decreases (time runs forwards) a lineage of size m at time t splits at rate

$$\text{split} - \text{rate}_m(t) = \frac{m-1}{t(1+t)} \tag{2}$$

and the size of the left subclade has

$$\text{uniform distribution on } \{1, 2, \ldots, m-1\}. \tag{3}$$

To verify this, note that the unconditional rate of mergers of clades of sizes m_1, m_2 at time t (per unit time, relative to number of species) equals

$$G(t)(1-G(t))^{m_1-1} f(t) (1-G(t))^{m_2-1} G(t)$$

by considering the required heights of branchpoints for this event to occur. Similarly the number of size $m_1 + m_2$ lineages at time t, relative to number of species, equals

$$G(t)(1-G(t))^{m_1+m_2-1} G(t).$$

Thus the rate of splitting of a size-$m_1 + m_2$ lineage into subclades of sizes m_1, m_2 equals

$$\frac{G(t)(1-G(t))^{m_1-1} f(t) (1-G(t))^{m_2-1} G(t)}{G(t)(1-G(t))^{m_1+m_2-1} G(t)} = \frac{1}{t(1+t)}$$

implying (2,3).

5.2 Loss of evolutionary history

Nee and May [30] interpret total edge-length of a phylogenetic tree as an indicator of "total evolutionary history" of a clade. One can then ask what proportion of total evolutionary history (call this proportion LEH(ρ), say) would be lost if a proportion ρ of extant species are suddenly made extinct. We complement calculations of [30] by giving an exact calculation within $l - \text{TREE}_\infty$, as an approximation to $l - \text{TREE}_n$ for large n. In this setting the proportionate loss LEH(T, ρ) depends on the duration of time T over which we track evolutionary history. Assume the species to be made extinct are chosen randomly. The formula is

$$\text{LEH}(T, \rho) = 1 - \frac{\log(1 + (1-\rho)T)}{\log(1+T)}. \tag{4}$$

As the table indicates, this is broadly in line with the conclusion of [30] that "about half the history is preserved by saving 20% of species[7]".

		ρ		
		0.5	0.8	0.9
	10	0.25	0.54	0.71
T	30	0.19	0.43	0.60
	100	0.15	0.34	0.48

Table 1. Proportion of edge-length ("evolutionary history over time duration T") lost in $l-\mathrm{TREE}_\infty$ when proportion ρ of species go extinct.

Derivation of formula (4). Because the number of lineages (relative to number of extant species) at time t is $G(t) = 1/(1+t)$, we see

$$\mathrm{LEH}(T,\rho) = \frac{\int_0^T G(t)q(t,\rho)\ dt}{\int_0^T G(t)\ dt}$$

where $q(t,\rho)$ is the chance that, in a lineage at time t, all the extant species are chosen (by the random choices) to go extinct. Using (1) and a standard calculation,

$$q(t,\rho) = \sum_{i=1}^{\infty} p_t(i)\rho^i = \frac{\rho}{1+t-\rho t}$$

and a routine integration exercise gives (4).

5.3 Local structure of $c-\mathrm{TREE}_n$

Parallel to the discussion (Section 5.1) of $l-\mathrm{TREE}_\infty$ as an approximation to the local structure of $l-\mathrm{TREE}_n$ for large n, there is an infinite tree $c-\mathrm{TREE}_\infty$ which is the $n\to\infty$ limit of the local structure of the complete tree $c-\mathrm{TREE}_n$. Such limits arise for different models of random trees in different applications of probability; see [2] for a survey. We first give a mathematical description of the limit object, and then describe in what sense it is a limit.

Construction of $c-\mathrm{TREE}_\infty$: similar to [2]. The construction is specified relative to a reference species v_* and a reference time t_*; think of t_* as a long

[7] Since T_n^{lca} is order n (see Section 5.4), to apply our calculation roughly to $l-\mathrm{TREE}_n$ we set $T\approx n$ and our formula (4) is consistent with equation (1) of [30] in the Moran model.

time ago, and v_* as an extinct species. Write $v_1, v_2, v_3, \ldots$ for the ancestors of v_*, and declare that the lengths of time backwards from t_* to the origin of v_1, from that time until the origin of v_2, and so on, are independent exponential(1) random times. For each vertex v_j, its lifetime extends beyond the origin of v_{j-1} for an independent exponential(1) random time; during its lifetime it has other daughter species at random times at stochastic rate 1; in turn these daughter species behave as in the CBP process of rate $\lambda = 1$. And for the reference species v_* itself, its lifetime extends beyong t_* for an independent exponential(1) random time; and its daughter species behave as above.

The limit procedure. In the context of the finite model $c-\mathrm{TREE}_n$, pick at random[8] a pair (v_*, t_*) where t_* is a past time and v_* is a species existing at time t_*. Consider $c-\mathrm{TREE}_n$ relative to (v_*, t_*); that is, consider the portion of $c-\mathrm{TREE}_n$ within a "window" of some arbitrary extent t_0, a vertex being within this window if the path linking it to v_* stays within the time interval $t_* \pm t_0$. Then the sense in which $c - \mathrm{TREE}_n$ converges to $c - \mathrm{TREE}_\infty$ is that the part of $c-\mathrm{TREE}_n$ within an arbitrary fixed size window relative to the randomly-chosen reference (v_*, t_*) converges in distribution to the part of $c - \mathrm{TREE}_\infty$ within the same sized window relative to its reference point (v_*, t_*).

Convergence relative to an extant species. The discussion above relates to a reference species chosen at random from $c-\mathrm{TREE}_n$, which (with probability $\to 1$ as $n \to \infty$) is an extinct species. There is a parallel result for a randomly-chosen *extant* species. In this setting the limit is simply the part of $c - \mathrm{TREE}_\infty$ at times *before* t_*, where t_* is now the present time 0.

Elementary calculations. The above construction makes it easy to do $n \to \infty$ limit calculations for $c - \mathrm{TREE}_n$ by doing exact calculations within $c - \mathrm{TREE}_\infty$. For instance, for a random extant species,

$$P(\text{parent is extant}) = 1/2 \tag{5}$$

$$P(\text{some ancestor is extant}) = 1 - e^{-1} = 0.63... \tag{6}$$

Indeed, the probability in (5) is $P(\xi_1 < \xi_2)$, where ξ_1 is the time since origin of v_*, and ξ_2 is the subsequent lifetime of its parent v_1; because ξ_1 and ξ_2 are independent exponential(1) random times, the probability equals 1/2 by symmetry. For (6), the times at which some ancestor originates form a Poisson process of rate 1, and an ancestor originating at time t before present has chance e^{-t} to be extant, so the random number of extant ancestors has

[8] Precisely, we pick (v_*, t_*) uniformly from the set of possible (v, t) pairs.

Poisson distribution with mean $\int_0^\infty e^{-t} \times 1 \; dt = 1$, and thus takes value 0 with probability e^{-1}.

5.4 Times of origin and last common ancestor, and number of contemporaneous species

In Section 4.1 we defined, in terms of our model $c - \text{TREE}_n$,

$$\begin{aligned} T_n^{\text{origin}} &= \text{time of origin of clade} \\ T_n^{\text{lca}} &= \text{time of last common ancestor of extant species} \end{aligned}$$

and we can also consider

$$C_n(T_n^{\text{lca}}) = \text{number of species at time } T_n^{\text{lca}}.$$

All three quantities scale linearly with n, and there is a limit rescaled joint distribution; that is, we may write

$$\tfrac{1}{n}(T_n^{\text{origin}}, T_n^{\text{lca}}, C_n(T_n^{\text{lca}})) \approx (T_*, S_*, R_*) \text{ for large } n. \tag{7}$$

The limit distribution is specified (see (2.16) of [37], p.22) by the joint probability density function

$$f_{T_*,S_*,R_*}(t,s,r) = s^{-4}(t-s)^{-2} \exp\left(-\frac{1}{s} - \frac{tr}{s(t-s)}\right), \quad 0 < s < t, \; 0 < r. \tag{8}$$

Figure 5 illustrates this distribution via a scatter diagram. One can make the mathematical structure of this distribution more interpretable as follows. The joint density of (T_*, S_*) is

$$f_{T_*,S_*}(t,s) = s^{-2}t^{-2} \exp(-\tfrac{1}{s}), \quad 0 < s < t. \tag{9}$$

The joint density (9) coincides with the joint density obtained using two independent exponential(1) variables ξ_1, ξ_2 via

$$(T_*, S_*) \stackrel{d}{=} (\tfrac{1}{\xi_1}, \tfrac{1}{\xi_1+\xi_2}).$$

The marginal densities of T_* and S_* are

$$f_{T_*}(t) = t^{-2} \exp(-\tfrac{1}{t}), \quad 0 < t,$$

$$f_{S_*}(s) = s^{-3} \exp(-\tfrac{1}{s}), \quad 0 < s.$$

The conditional distribution of T_* given S_* is

$$P(T_* > us | S_* = s) = 1/u, \quad u \geq 1. \tag{10}$$

Now given $(T_* = t, S_* = s)$, the conditional distribution of R_* is

$$f_{R_*|t,s}(r) = \lambda^2(s,t) r \exp(-r\lambda(s,t)), \quad 0 < r; \quad \text{for } \lambda(s,t) = \tfrac{t}{s(t-s)}.$$

This is the $Gamma(2, \lambda(s,t))$ density. Integrating (8) out over (s,t) gives the marginal density of R_* to be

$$f_{R_*}(r) = 2(1+r)^{-3}, \quad 0 < r.$$

In particular, note that R_* has mean value $E(R_*) = 1$ but its variance is $\text{var}(R_*) = \infty$. The cumulative density of R_* is given simply by

$$P(R_* > r) = (1+r)^{-2}, \quad 0 < r$$

so that numerically its 10th percentile is 0.054 and its 90th percentile is 2.16. Note R_* is very variable.

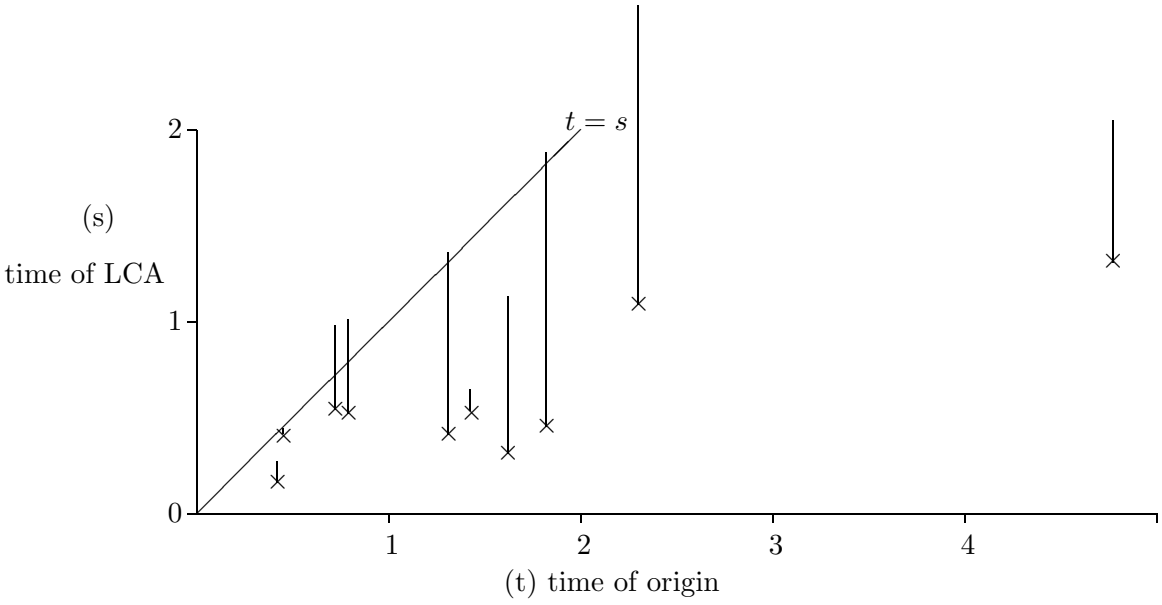


Figure 5. Scatter diagram of 10 realizations of the standardized joint distribution (T_*, S_*, R_*). Points $\times$ give the (t,s)-values, and the height of line segment is the r-value. The time unit equals n ETU, in the approximation (7) for $c - \text{TREE}_n$. Note the extreme variability of R_*; the smallest value was 0.03 and the largest was 1.56.

5.5 Numbers of extinct species

The underlying CBP model with time run forwards is the birth-and-death process (Section 2.6) with birth rate $\lambda_i = i$ and death rate $\mu_i = i$. A simple consequence of our Bayesian construction of $c-\text{TREE}_n$ is that the process $(C_n(t), 0 \le t \le T_n^{\text{origin}})$ giving the total number of species existing at time t as t increases (time runs backwards), is distributionally the same CBP, started with n species and continued until the first time T_n^{origin} it reaches 0 (see Theorem 1 for the proof). This observation has several simple consequences. First, classical results on hitting probabilities for simple random walk (or "the martingale property") imply

$$P(\max_t C_n(t) \ge c) = \tfrac{n}{c}, \quad c = n, n+1, n+2, \ldots \tag{11}$$

which we can rephrase as *the* $1/r$ *rule* mentioned in Section 1.2: *The chance that at some past time the number of coexisting species was at least* r *times the current number, equals* $1/r$.

A second consequence is that we can obtain large-n approximations for the total number ($\mathcal{G}_n$, say) of species in $c-\text{TREE}_n$. This number scales as n^2 and we have

$$\mathcal{G}_n \approx n^2 \mathcal{G}_*$$

where the limit $\mathcal{G}_*$ has probability density function

$$(4\pi g^3)^{-1/2} \exp(-\tfrac{1}{4g}), \quad 0 < g < \infty.$$

Theorem 3 gives details and proof.

5.6 The chance that a fossil species is ancestral

Textbooks (e.g. [33] page 24) often say

> the probability that a given fossil is actually part of an ancestral lineage [of some extant species] is actually rather remote.

Various calculations relevant to this issue can be done within our model.

(a). Consider a species that originated at time t (t ETU before present). Then the chance that some descendant species (or the species itself) is extant at present equals (in the $n \to \infty$ limit)

$$1/(1+t). \tag{12}$$

This formula comes from the construction of the tree $c-\text{TREE}_\infty$ which gives the local description of the structure of the complete tree $c-\text{TREE}_n$.

As stated in Section 5.3 the descendants of a species v evolve, as time runs forwards, as in an ordinary CBP process. If v originates at time t before the present, then the chance that some descendant of v (or v itself) is extant at present equals the chance of the survival of a CBP for t or more time. For a CBP the population size at any time τ has a shifted geometric distribution

$$P(N(\tau) = 0) = \frac{\tau}{1+\tau}; \quad P(N(\tau) = k) = \frac{\tau}{(1+\tau)^{k+1}}, \quad k \geq 1,$$

and since the survival of CBP for time t or more is the complement of the event that the population size of CBP at t is equal to 0, the chance of some extant descendant is precisely $1 - t/(1+t)$.

(b). Within our model $c-\mathrm{TREE}_n$ of clades with n extant species, let $\hat{N}_n^{\mathrm{anc}}$ and N_n^{anc} be the numbers of extinct species ancestral to the extant species, during the time since the last common ancestor of the extant species (for $\hat{N}_n^{\mathrm{anc}}$) or the time since origin of clade (for N_n^{anc}). It can be shown[9] that in the $n \to \infty$ limit both $\hat{N}_n^{\mathrm{anc}}$ and N_n^{anc} are approximately $n \log n$. Precisely,

$$P((1-\varepsilon)n\log n \leq \hat{N}_n^{\mathrm{anc}} \leq N_n^{\mathrm{anc}} \leq (1+\varepsilon)n\log n) \to 1, \quad \text{any } \varepsilon > 0. \quad (13)$$

Note that in contrast to most quantities we study, these quantities are asymptotically *non-random* to first order. However, as the numerical results in Table 2 show, for small and moderate values of n there is considerable variability in these quantities. Numerically we see a good fit to

$$\mathrm{median}(\hat{N}_n^{\mathrm{anc}}) \approx n(\log n - 1.4). \quad (14)$$

	percentiles of $\hat{N}_n^{\mathrm{anc}}$			approx	percentiles of $\hat{p}_n^{\mathrm{anc}}$			approx
n	25th	50th	75th	(14)	25th	50th	75th	(16)
10	4	9	19	9.0	0.093	0.151	0.212	0.147
20	18	32	55	31.9	0.095	0.147	0.204	0.146
40	59	91	143	91.6	0.064	0.110	0.164	0.110

Table 2. The distribution of $\hat{N}_n^{\mathrm{anc}}$ = number of extinct ancestral speecies, and the distribution of $\hat{p}_n^{\mathrm{anc}}$ = proportion of ancestral species amongst all extinct species (both during the time since last common ancestor); and the approximations (14,16) to the medians. "Log" is natural logarithm.

[9] Here is a rough order-of-magnitude calculation. In our model, such a clade has existed for time of order n, and during most of that time there have been order n species existing. So the number of extinct species with extant descendants is, using (12), $\approx \int_0^n n(1+t)^{-1}dt \approx n \log n$.

(c). Returning to the original question – what is the chance that a random extinct species is ancestral to some extant species – there are some subtleties involved in choosing how to formulate this as a precise mathematical question. Here is one approach. Within our model $c-\mathrm{TREE}_n$ of clades with n extant species, we have the "numbers of extinct ancestral species" quantities $\hat{N}_n^{\mathrm{anc}}, N_n^{\mathrm{anc}}$ above. We also have the corresponding total numbers of extinct species, say $\hat{\mathcal{G}}_n - n$ and $\mathcal{G}_n - n$ (where the $-n$ terms remove extant species, for consistency with $\mathcal{G}_n$ in section 5.5). Thus within a clade we can consider proportions of extinct species which are ancestral

$$\hat{p}_n^{\mathrm{anc}} = \frac{\hat{N}_n^{\mathrm{anc}}}{\hat{\mathcal{G}}_n - n}; \quad p_n^{\mathrm{anc}} = \frac{N_n^{\mathrm{anc}}}{\mathcal{G}_n - n} \tag{15}$$

where the former considers time since last common ancestor and the latter considers time since origin of clade. We saw in section 5.5 that $\mathcal{G}_n \approx n^2 \mathcal{G}_*$ for a random limit $\mathcal{G}_*$; similarly $\hat{\mathcal{G}}^n \approx n^2 \hat{\mathcal{G}}_*$ for a different random limit $\hat{\mathcal{G}}_*$. It follows from this and (13) that both $\hat{p}_n^{\mathrm{anc}}$ and p_n^{anc} grow as $n^{-1} \log n$ with random limits:

$$\hat{p}_n^{\mathrm{anc}} \approx n^{-1} \log n \times 1/\hat{\mathcal{G}}_*; \quad p_n^{\mathrm{anc}} \approx n^{-1} \log n \times 1/\mathcal{G}_*.$$

Table 2 illustrates the distribution of $\hat{p}_n^{\mathrm{anc}}$ and shows a reasonable fit, for small and moderate n, to an approximation

$$\mathrm{median}(p_n^{\mathrm{anc}}) \approx \frac{2.1(\log n - 1.6)}{n}. \tag{16}$$

5.7 Estimating rates in birth-and-death processes: a simulation study

The linear birth-and-death process (section 2.6) is a standard model which can be used to try to estimate past speciation and extinction rates within a clade. Consider the setting where the only data we have is the lineage tree for the n extant species (as usual we are supposing we know the true lineage tree; we are seeking to study aspects of unobserved extinct species). We regard the linear birth-and-death model as having 3 parameters (t_*, λ, μ), where

$$\begin{aligned} t_* &= \text{time before present of clade origin} \\ \lambda i &= \text{total speciation rate, when } i \text{ species} \\ \mu i &= \text{total extinction rate, when } i \text{ species.} \end{aligned}$$

It is routine[10] to calculate numerically maximum likelihood estimates (MLEs) of the parameters, based on a lineage tree as data.

We studied what happens if one applies this procedure – estimating parameters assuming the underlying model of species diversity is a linear birth-and-death process – to simulated data from our $c - \text{TREE}_n$ model. Table 3 shows the MLEs derived from 10 typical realizations (pictured at [1]) with $n = 8$. Of course, in our model we really have $\lambda = \mu = 1$ in each realization, and realization-dependent values of T^{lca} and T^{origin}.

realization	1	2	3	4	5	6	7	8	9	10
MLE of λ	0.4	0.6	1.3	1.5	0.6	1.3	0.3	0.3	0.9	1.2
MLE of μ	0.1	0.4	0.2	1.0	0.3	0.3	0.2	0.1	0.4	0.2
MLE of t_*	9.1	6.5	1.7	3.1	4.7	2.5	11.4	13.8	4.2	1.5
T^{lca}	8.4	6.0	1.5	2.8	4.3	2.4	11.3	13.3	3.9	1.4
T^{origin}	37.1	13.3	1.7	8.2	8.9	11.9	36.2	154.7	21.8	1.7

Table 3. MLE estimates of linear birth-death parameters based on realizations from our model ($n = 8$).

So in this setting the estimated values of λ and μ in the linear model are very misleading. Not only does "the pull of the recent" make the estimated λ larger than the estimated μ, but also the estimated values are varying widely between realizations[11]. As a secondary point, observe that in most realizations the estimated time of clade origin is about 10% greater than the observed time of last common ancestor of extant species, regardless of further information about the shape of the lineage tree; under our model the true time of clade origin varies greatly.

As mentioned in Section 1.3.5, there has apparently been no large-scale statistical study of extant clades to determine which stochastic models are realistic; our model, with its intrinsic variability, provides a conservative approach to inference questions. So, while our results are not directly comparable with those in existing literature (Nee et al [31, 29, 28], Paradis [34, 35]) they do cast serious doubt on the ability to reconstruct at any level of detail the history of a clade from the phylogeny of extant species.

[10] The only subtle issue is that one should compute the likelihood *without* conditioning on n. To see why, note that when μ/λ is large the process is a priori unlikely to reach n species; this is a real effect which would incorrectly be factored out by conditioning.

[11] The variability is not a "small n" effect: as emphasized throught the paper, variability in our model persists in the $n \to \infty$ limit.

6 The model for higher-order taxa

6.1 Thinking about higher-order taxa

The tree-based classification of species emphasized in *cladistics* is widely recognized as the logically desirable classification scheme, while the traditional Linnean hierarchy is useful in practice for human-interpretable discussions of different clades, and for classifying fossil species where there is too little information for reliable tree-classification. The difficulties of reconciling tree-based classifications with the Linnean hierarchy are well recognized in the systematics literature, though how this issue interacts with stochastic modeling has not been investigated so thoroughly. In particular, the inevitable subjectivity of a hierarchical scheme – how widely to cast the net of a single genus or family – and the fact that fossil time series are typically presented at the genus or family level [43] have led to concerns [47] that such time series may give a biased picture of species diversity. These concerns have been discussed via data and model simulations using hierarchical schemes defined in a somewhat ad-hoc manner ([45], [41]: see Section 7.5) ; we seek to give a more detailed analysis within our stochastic model.

Our goal is to make a model of phylogenetic trees of genera (we write *genus* for concreteness, to represent an arbitrary higher-order level) which is coherent with an underlying model for species. This involves two issues. One issue, not involving randomness, is how one reconciles a phylogenetic tree on species with the Linnean hierarchy in such a way that one can define a phylogenetic tree of genera. The second issue is how to make a probability model of this "novel genera" process which reflects "pure chance" rather than some particular conjectured biological mechanism. We address these issues separately in the next two sections.

6.2 Defining genera in terms of "new types" of species

Suppose we have a complete tree on a clade of species, and suppose certain species are distinguished as "new type" due to some characteristic judged biologically significant which is expected to persist in descendant species. A scheme for using such "new types" to define genera should, as a minimum requirement, have the following property.

> A genus cannot contain both a species which is a descendant of some "new type" species s and a species which is not a descendant of s

or in other words

> A "new type" species and its descendants comprise one or more genera.

To a mathematician, there are three reasonable ways to define genera satisfying this requirement, which we call the *coarse, medium* and *fine* schemes, because they produce successive smaller genera. Discussions of schemes broadly like these can be found in the systematics literature, but strike mathematicians as somewhat ad-hoc and imprecise.

Before giving detailed definitions, let us emphasize a conceptual point. Given a complete tree on the species in a clade, and given a subset of these species which are "new type", choose one of the three schemes below to define genera. Each genus has an extinction time (or is extant), an origination time, and (if not the originating genus of the clade) is a daughter of some other genus. Thus one can draw (Figure 6) a tree on genera in the same style as the tree on species; this is part of *coherence*.

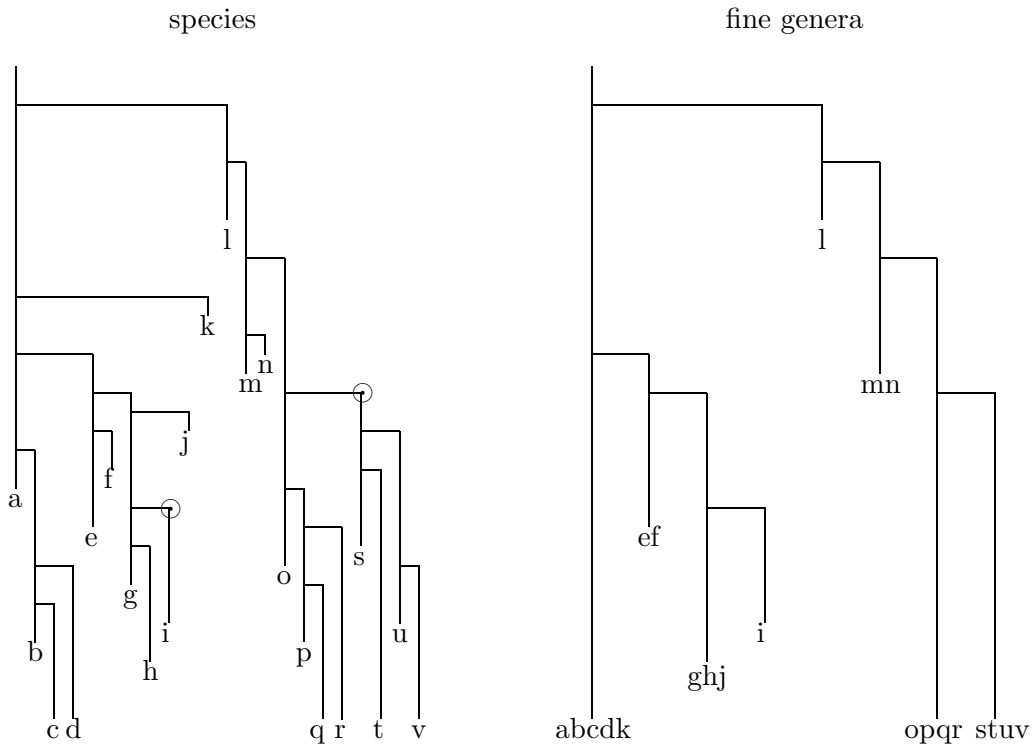
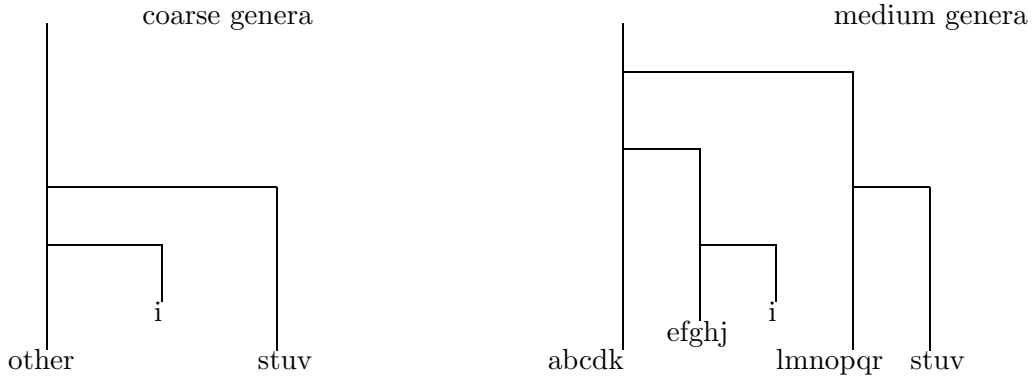


Figure 6. Illustration of our schemes for defining genera in terms of new types. Above left is a complete clade of 6 extant and 16 extinct species ($abcd\cdots uv$), with two species designated as new types and marked $\odot$. In the fine scheme, this induces 8 genera (3 extant), whose tree is shown above right. The other schemes are shown below, with compressed time scale.



The coarse scheme. Here we imagine that each "new type" species s creates a genus, consisting of itself and all descendants which are not descendants of (or identical to) some other "new type" descendant of s. Equivalently, we have the following rule :

> two species are in the same *coarse genus* provided the path between them involves no "new type" species (except perhaps their last common ancestor).

If we declare that the originating species of the entire clade is "new type", then the number of coarse genera is just the number of "new type" species. In this and the other schemes, each genus has a *founder* (chronologically first) species, and here we have :

> the founder species of the coarse genera are precisely (only) the "new type" species.

In Figure 6 (which has two "new types" in addition to the originating species) we get three genera: $\{i\}$, $\{stuv\}$, and the remaining species.

An unsatisfactory feature of the coarse scheme can be seen in Figure 6 : species h, r are put in the same genus even though h is more closely related to i (in a different genus) that to r, while r is more closely related to s (in a different genus) than to h. To remedy this we may consider the following requirement :

> If α, β are in the same genus and γ in a different genus, then α cannot be strictly closer related to γ than to β; that is, the last common ancestor of $\{\alpha, \gamma\}$ cannot be a strict descendant of the last common ancestor of $\{\alpha, \beta\}$.

The fine scheme. With this scheme we construct genera which satisfy the above requirement by using the following rule:

> two species are in the same genus provided the path between them involves no species (except perhaps their last common ancestor) which is of "new type" or has a descendant of "new type".

In contrast to the coarse scheme in this scheme we have many more founders of new genera, namely :

> the founder species of the fine genera are precisely the "new type" species and all of their ancestors.

In Figure 6, we obtain eight fine genera ($\{abcdk\}$, $\{ef\}$, $\{ghj\}$, $\{i\}$, $\{l\}$, $\{mn\}$, $\{opqr\}$, and $\{stuv\}$).

The fine scheme has its own drawback, again as seen in Figure 6 : it creates $\{ef\}$ and $\{ghj\}$ as distinct genera, as well as $\{l\}$ and $\{mn\}$ as distinct genera, even though it might be difficult to make the distinction from a fossil record. Here is a different property one might desire.

> Taking a representative species from each genus and drawing the cladogram on these species gives the same cladogram regardless of the choices of species.

The medium scheme. The idea for this scheme is to find the coarsest scheme that will satisfy the above property. We construct such genera by using the rule:

> two species are in the same genus provided no species on the path between them (except perhaps their last common ancestor) is a "new type" species, and provided no species on the path between them (including their last common ancestor) is an "essential branchpoint" species.

Here an *essential branchpoint* species is one which is the last common ancestor of some two "new type" species. Here is the description of founder species.

> The founder species of the medium genera are precisely the "new type" species, the essential branchpoint species, and the daughters of branchpoint species which are ancestors of new types.

For mathematical justification of the criteria for building the schemes given their desired properties see Section 9.3.

Monophyletic groups It is generally regarded as desirable that groups such as genera be *monophyletic*, that is consist of some founder species and all its descendants. Whether this desideratum should be regarded as essential is controversial: see e.g. Benton [9]. Within our underlying detailed picture of macroevolution at the level of species, it is impossible to define genera usefully so that this constraint is literally satisfied. For instance, if in Figure 6 we declare $\{stuv\}$ to be a genus, then the parent species o needs to be put into some genus, which is therefore not monophyletic. However, modifying the definition of monophyletic to allow all descendants of some daughters of the founder to be excluded, the fine genera do satisfy the modified definition.

6.3 The probability model for higher-order taxa

Reconsider our model (Section 4.1) for the complete tree $c - \mathrm{TREE}_n$ on a clade with n extant species. Introduce a parameter $0 < \theta < 1$, and suppose that each species (extinct or extant) has chance θ to be a new type. Then any of the three schemes from the previous section can be used to define an induced tree on genera, which we shall call $(n, \theta) - fine$GENERA etc. Figure 7 shows a realization derived from the 162-species clade in Figure 1. Decreasing the parameter θ will increase the average number of species per genus: alternatively, regard decreasing θ as moving up the taxonomic hierarchy.

This stochastic model differs from that in Yule (as already mentioned in Section 2.2) and from that in Gould et al [16] who specify higher-level groups using size constraints. The model is intended to capture the "neutral" idea that a clade is defined by a heritable character but that this character has no "selective advantage", i.e. that the species in the clade have unchanged speciation and extinction rates.

Note that conceptually it is simple to model simultaneously two or more higher levels such as {genus, family} by using two probabilities $\theta_{\mathrm{family}} < \theta_{\mathrm{genus}}$.

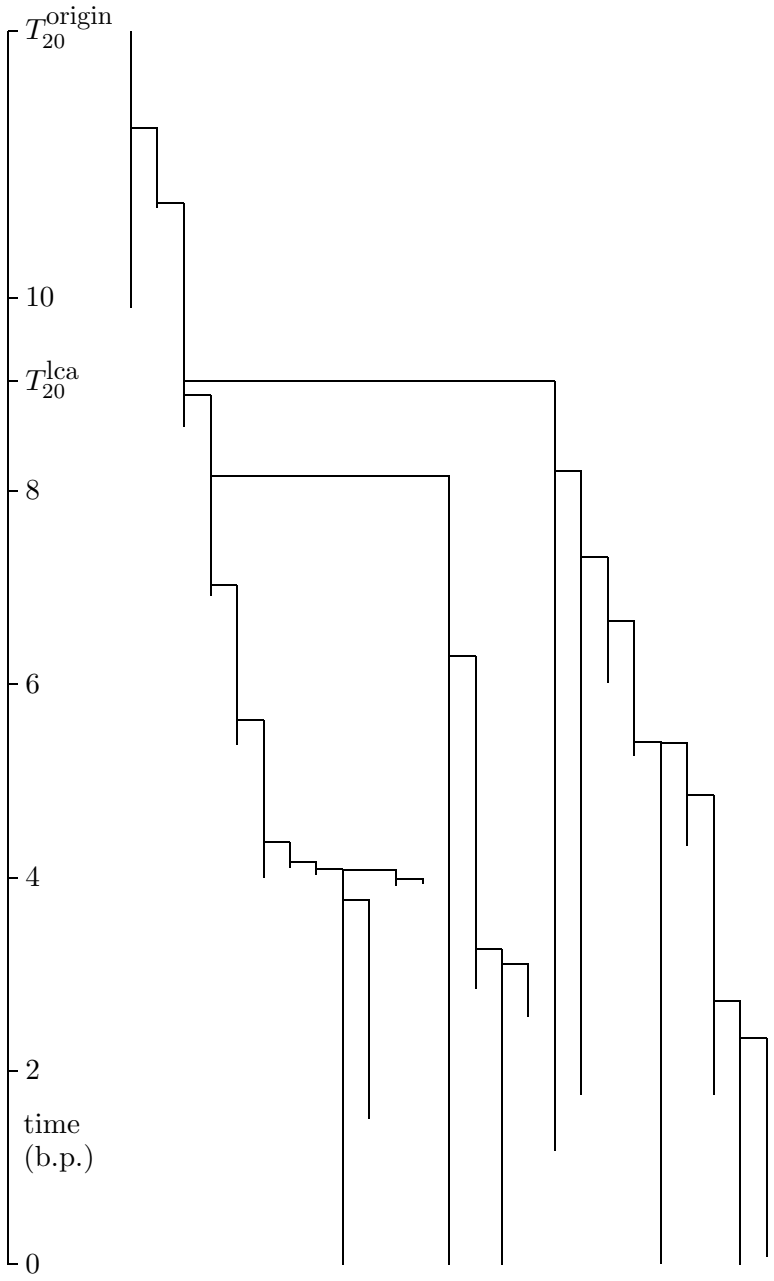


Figure 7. A realization of the tree $(n, \theta) - fine$GENERA on extant and extinct genera, with $n = 20$ species and $\theta = 0.04$. It was derived from the realization of $c-\text{TREE}_{20}$ in Figure 1, with the "new type" species there indicated by $\odot$. In this realization, there were 7 such "new type" species, producing 25 genera, of which 5 were extant.

7 Mathematical properties: higher order taxa

7.1 Overview

We now turn to properties of our model involving higher level taxa, as usual writing "genera" for concreteness. Recall that the conceptual point is that our model for (for instance) the phylogenetic tree on genera is derived from the underlying model for species and from the schemes for defining genera in terms of "new type" species (section 6.2). The theoretical parameter θ (probability that a new species is a "new type") determines average number of species per genus (see formulas (20,23)) which can be regarded as a data-derived parameter.

We seek to study the following kind of quantities.

- Number of species per (extinct) genus
- Number of species per (extant) genus
- Lifetime of extinct genera
- Lifetime (until present) of extant genera
- Fluctuation rates for number of genera
- Shape of the tree on extinct genera
- Shape of the tree on extant genera.

We need to distinguish the extinct and extant cases because, of course, by virtue of being extant a genus is likely to be larger and longer-lived than a typical extinct genus.

The seven quantities above, and the three schemes for genera, make twenty-one problems, some of which are amenable to theoretical calculation and others seeming to require simulation.

`xxx Work in progress!  This version gives just a brief snapshot.`

7.2 Shape of tree on extant fine genera

The setting is an extant clade, where we use the "fine" scheme for defining genera in terms of "new type" species, but where we only have data from extant species so that we notice only the "new type" species that are ancestral to extant species. We seek to study the shape of the tree on genera. As described in section 1.3.2, we measure "shape" in terms of splits of a parent

clade (being a set of *genera* here) into daughter clades. For an n-genus parent clade, the smaller daughter clade may consist of 1 or 2 or 3 or $\dots n/2$ genera, and we summarize this distribution by recording
(a) the chance that smaller daughter clade is size 1;
(b) the median size of smaller daughter clade.
Table 4 shows numerical values for several choices of average number of species per genus, and several values of parent clade size. The first row gives this data for species, where our model coincides with the usual Markov model and predicts uniformly distributed splits.

mean number species per genus	parent clade size						θ
	5		10		15		
	$p(1)$	med	$p(1)$	med	$p(1)$	med	
1	0.50	1.5	0.22	2.8	0.14	4.0	1
5	0.73	1.2	0.32	2.6	0.21	3.8	0.047
10	0.79	1.1	0.44	1.9	0.25	3.7	0.0145
20	0.85	1.1	0.54	1.4	0.34	3.5	0.0045

Table 4. Shape of tree on extant fine genera. For the given size (number of genera) in a parent clade, the table shows the probability $p(1)$ that smaller daughter clade size equals 1, and the median size *med* of smaller daughter clade. Results from Monte Carlo simulations of model with 400 extant species.

Increased imbalance is indicated by the first probability increasing and by the second median decreasing. The table indicates
(i) imbalance increases with size of genus, i.e. as we go up the taxonomic hierarchy,
(ii) measured by median, the increase in imbalance is more prominent within smaller clades than within larger clades. Measured by $p(1)$ there seems little clade-size dependence.

7.3 Standardized fluctuation rates

Within our basic CBP model, write $N(t)$ for number of species at time t and $G(t)$ for number of genera at time t, using one of our schemes for defining genera. Here we are imagining t as a typical time in the past, so we are considering extinct species. A basic mathematical property of the CBP model is that the stochastic fluctuations, measured by variance of changes in the time series, have a simple form over times t of order 1 ETU. Given $N(0) = n(0)$ we have

$$\mathrm{var}(N(t) - n(0)) \approx 2n(0)t.$$

In other words the ratio

$$\frac{\mathrm{var}(N(t)) - n(0)}{2n(0)t} \tag{17}$$

is approximately 1 regardless of the value of $n(0)$ (assumed not too small) or the value of t (assumed of order 1 ETU). This motivates studying fluctuations in number of genera in the same way. Given $G(0) = g(0)$ we anticipate that the ratio

$$\frac{\mathrm{var}(G(t)) - g(0)}{2g(0)t}$$

should not depend much on $g(0)$ or t. Now the ratio in (17) is really 1/(mean species lifetime), becoming 1 ETU by definition. Similarly, if one were to model genera directly as a CBP, then the ratio above would be 1/(mean genus lifetime). Thus we finally come to define *normalized fluctuation rate* as

$$(\text{mean genus lifetime}) \times \frac{G(t) - g(0)}{2g(0)t}. \tag{18}$$

Note this could be estimated from data in the natural way, by averaging $(G(t_{i+1}) - G(t_i))^2/G(t_i)$ over intervals $[t_i, t_{i+1}]$ of length t.

Rephrasing the discussion above, the interpretation of *normalized* fluctuation rate is as the ratio of actual observed fluctuation rate to what the rate would be if genera themselves behaved as a CBP. Within our model, where species behave as a CBP and genera are defined by one of our schemes and random "new types", we get theoretical predictions for normalized fluctuation rates. Table 5 gives numerical values in one case.

mean number species per genus	1	5	10	20	40
normalized fluctuation rate	1.00	0.84	0.68	0.57	0.45

Table 5. Normalized fluctuation rates for coarse genera. Results from Monte Carlo simulations of model started with 400 species. Ratio (18) estimated with t = mean genus lifetime.

7.4 Number of species per extinct genus

`xxx Paragraph below to be polished later.`

As stated at the start of Section 5, our methodology for analytic results is derive approximations by appealing to exact formulas in the $n \to \infty$ limit. Such arguments depend on the "local weak convergence" ideas of section 5.3. That is, relative to a typical (= extinct) species the future process is just CBP; so we can argue "forwards in time". A typical extinct genus comprises some descendants of a species conditioned on that species being a founder;

can do calculations without needing to look backwards in time. Calculations are harder with extant genera because we need to argue backwards in time.

Note that in the extinct case, there is a general relation, for any definition of genus along with founder species:

$$E \text{ (number of species in genus)} = 1/P(\text{typical species is a founder species}). \tag{19}$$

Of course various details of the underlying model are arbitrary, e.g. species lifetime assumed Exponential so s.d. equals mean. But qualitative features hopefully are robust.

7.4.1 Number of species per extinct coarse genus

It turns out to be easy to find the distribution of the number $\mathcal{G}$ of species in a typical coarse genus.

$$\begin{aligned} \text{(mean) } E\mathcal{G} &= \theta^{-1} && (20)\\ \text{(variance) var}\mathcal{G} &= \theta^{-1} - 3\theta^{-2} + 2\theta^{-3} && (21)\\ \text{(generating function) } E(z^{\mathcal{G}}) &= \frac{1}{2} + \frac{1 - \sqrt{(2-\theta)^2 - 4(1-\theta)z}}{2(1-\theta)}. && (22) \end{aligned}$$

xxx Lea: give exact formula also! Certainly known, e.g. from lengths of excursions of biased RW.

Recall that for coarse genera, the founder species of each genus are exactly the "new type" species. So a coarse genus consists of its "new type" founder and its descendant species, with the modification that any "new type" descendant species are discarded (and so don't have descendants). The chance of a species being a founder is θ, so (19) implies (20). Because the relative chances of a species to first (become extinct; have daughter species which is not "new type") are $(1; 1-\theta)$, it is clear that the coarse genus has the distribution of a Galton-Watson process whose offspring distribution D is shifted geometric$(p = 1/(2-\theta))$;

$$P(D = d) = \tfrac{1}{2-\theta} \left(\tfrac{1-\theta}{2-\theta}\right)^d, \ d \geq 0.$$

By classical theory, the probability generating function $g(z) = E(z^{\mathcal{G}})$ of the total size $\mathcal{G}$ of the Galton-Watson process is determined by the probability generating function $f_D(z) = E(z^D)$ as the unique positive solution of the equation $g(z) = z f_D(g(z))$ ([13] XII.5.). When the offspring distribution is shifted geometric(p) we have $f_D(z) = p/(1-(1-p)z)$, and hence $g(z) = (1 - \sqrt{1 - 4p(1-p)z})/2(1-p)$. Setting $p = 1/(2-\theta)$ gives (22). From the generating function formula we can derive the formula for variance.

θ	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
0.5	0.667	0.148	0.066	0.037	0.023	0.015	0.011	0.008	0.006	0.004
0.25	0.571	0.140	0.068	0.042	0.029	0.021	0.016	0.013	0.010	0.009
0.09	0.524	0.131	0.065	0.041	0.028	0.021	0.017	0.013	0.011	0.009
0.04	0.510	0.127	0.064	0.040	0.028	0.021	0.016	0.013	0.011	0.009
0.01	0.502	0.126	0.063	0.039	0.027	0.021	0.016	0.013	0.011	0.009
0.0025	0.501	0.125	0.063	0.039	0.027	0.021	0.016	0.013	0.011	0.009
0.001	0.500	0.125	0.062	0.039	0.027	0.020	0.016	0.013	0.011	0.009

Table 6. Probability distribution $\{p_k\}_{k\geq 1}$ of the "number of species in a typical *coarse genus* ".

7.4.2 Number of species per extinct fine genus

Recall that a founder species of a fine genus is a species such that it itself, or some descendant species, is "new type". A standard formula for CBP is that the number N_s of descendants of a random species s, including s itself, has probability generating function

$$G(z) = \sum_{n=1}^{\infty} P(N_s = n) z^n = 1 - \sqrt{1-z}.$$

(This is the $\theta = 0$ case of (22).) Consider

$$\begin{aligned} q(\theta) &= \text{probability that } s \text{ is the founder of a genus} \\ &= \text{probability that } s \text{ or a descendant is a "new type".} \end{aligned}$$

Because $1-q(\theta)$ is the chance that each of the N_s species is not "new type",

$$1 - q(\theta) = G(1-\theta) = 1 - \sqrt{\theta}.$$

Thus we get the formula

$$q(\theta) = \sqrt{\theta}.$$

As a corollary, note by (19)

$$\text{mean number of species per fine genus} = 1/\sqrt{\theta}. \tag{23}$$

Another corollary is that the proportion of genera in which the founder is a "new type" equals $\theta/\sqrt{\theta} = \sqrt{\theta}$.

A more elaborate calculation (Section 9.2) shows that the probability generating function of $\mathcal{G}$, the number of species in a genus, is

$$E(z^{\mathcal{G}}) = \frac{z}{\sqrt{\theta}} \left(\frac{1}{1 - \sqrt{\theta} + \sqrt{1 - z(1-\theta)}} - \frac{1-\theta}{1 + \sqrt{1 - z(1-\theta)}} \right). \quad (24)$$

From this formula we get that

> the variance of number of species per fine genus equals $\sqrt{\theta}/2\left(1/\theta^2 + 1\right) - 1$.

We can also calculate the probability distribution of $\mathcal{G}$ for different values of the θ. The following table shows $p_k = P(\mathcal{G} = k)$, for $k = 1, \ldots, 10$, and a few select values of $\theta = 1/2, \ldots, 1/1000$.

θ	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
0.5	0.740	0.167	0.056	0.021	0.008	0.004	0.002	0.001	0.000	0.000
0.25	0.583	0.193	0.093	0.051	0.030	0.018	0.011	0.007	0.005	0.003
0.09	0.444	0.180	0.103	0.067	0.046	0.033	0.025	0.019	0.015	0.012
0.04	0.378	0.165	0.099	0.067	0.049	0.037	0.029	0.023	0.019	0.016
0.01	0.313	0.146	0.090	0.063	0.047	0.036	0.029	0.024	0.020	0.018
0.0025	0.281	0.136	0.084	0.059	0.044	0.035	0.028	0.023	0.020	0.017
0.001	0.256	0.127	0.079	0.056	0.042	0.033	0.027	0.022	0.019	0.016

Table 7. Probability distribution $\{p_k\}_{k \geq 1}$ of the "number of species in a typical *fine genus* ", with θ the chance of a species being of "new type".

7.5 Previous work

xxx Lea: what you write here is good, but needs a little more detail (e.g. what does "better" mean?)

- found/cited papers: Sepkoski [45] used simulation of CBP sampled at some rate to form a tree on base taxa, then group into higher order taxa according to several different schemes (some monophyletic, some paraphyletic) to show paraphyletic taxa can capture underlying base diversity better (especially if sampling rate is poor). Robeck et al. [41] improve/refine the type of different schemes used above adding randomized (non-cladistic) schemes as control for the distribution on the size of higher order taxa; they argue simulation results show it is the number of higher taxa and their sizes that influences how well they capture underlying base taxa diversity. Common

conclusion: sometimes paraphyletic schemes do better, sometimes the monophyletic ones do.

8 Final remarks

The main conclusions of this paper were outlined in Sections 1.2 and 1.3 and will not be recapitulated here. Let us instead emphasize some slightly more technical matters.

8.1 Tree and hierarchical classifications

As we wrote in Section 6, the difficulties of reconciling tree-based classifications with the Linnean hierarchy are well recognized in the systematics literature, though how this issue interacts with stochastic modeling has not been investigated so thoroughly. Let us reiterate that the treatment in Section 6.2 of the classification issues is separate from our particular stochastic model, and so could be incorporated into more realistic models.

8.2 Comments on simulations

The viewpoint

> Because one can easily simulate stochastic models, whether simple or complex, on a computer, the mathematical study of simple models is unnecessary

has some practical merit, but let us illustrate some pitfalls. Another consequence of the observation above (11) is that it is easy to simulate $c-\mathrm{TREE}_n$ by first simulating the "time run backwards" process $(C_n(t), 0 \le t \le T_n^{\mathrm{origin}})$ and then superimposing the branching structure (each daughter species arises from a uniform random parent). Wollenberg et al [52] sought to give a simulation study of the model (critical branching conditional on n extant species) which we have formalized as $c - \mathrm{TREE}_n$; their purpose was to assess goodness of fit to three actual phylogenetic trees. By not knowing the observation, they needed to rely on rejection sampling (discarding realizations which didn't have the required n extant species), and fixed the time of origin rather than allowing it to vary; moreover (unaware of the $1/r$ law) they deleted realizations in which the maximum number of species at one time exceeded $\max(50, n + 20)$, which for a typical value $n = 17$ would remove (where time of origin is not fixed) about 1/3 of the realizations.

Each of the latter modifications reduces variability of realizations, and in particular has dramatic effect on likelihood of extreme realizations. So we guess that in $c-\mathrm{TREE}_n$ the P-values would be substantially different from those reported in [52], though perhaps not enough to affect their conclusions about their particular three examples.

8.3 External constraints on time of origin

We have emphasized the variability of realizations of our model, but we should admit a sense in which this is artifactual. Suppose our data is the lineage tree on extant species in a clade, so that we know T^{lca}. Then our model predicts a probability distribution (10) for T^{origin}, whose median is $2T^{\mathrm{lca}}$. This sounds biologically unrealistic; the point is that we will typically have extra data – molecular phylogenies linking this clade with some other clade, or fossil evidence – which tells us that T^{origin} must be less than some known value t_*. For making inference about evolutionary history of the clade using a model like ours, one should include the knowledge of t_*, and this will reduce variability which is in part associated with *a priori* possible large values of T^{origin}.

8.4 Incorporating the fossil record

To compare a model involving extinct species with fossil data obviously requires one to pay attention to incompleteness of the fossil record. The simplest way to extend our model to this setting would be to assume that each extinct species has chance p to be recognized in the fossil record, independently for different species. This leads to a model of a clade with n extant species and m observed fossil species (out of some unknown total number of extinct species). Mathematical treatment of this model has been given in [36]. One could extend these ideas in several ways, e.g. to consider whether an extinct genus contains a recognized fossil species, but we have not pursued such questions.

References

[1] D. J. Aldous. Stochastic models for phylogenetic trees. Web site www.stat.berkeley.edu/users/aldous/Phylo/Pindex.html.

[2] D.J. Aldous. Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.*, 1:228–266, 1991.

[3] D.J. Aldous. The continuum random tree III. *Ann. Probab.*, 21:248–289, 1993.

[4] D.J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, 16:23–34, 2001.

[5] J. W. Kirchner amd A. Weil. Delayed biological recovery from extinctions throughout the fossil record. *Nature*, 404:177–180, 2000.

[6] I.V. Basawa and B.L.S. Prakasa Rao. *Statistical Inference for Stochastic Processes*. Academic Press, 1980.

[7] M.J. Benton. *The Fossil Record 2*. Chapman and Hall, 1993.

[8] M.J. Benton. The history of life: Large databases in palaeontology. In D.A.T. Harper, editor, *Numerical Palaeobiology*, pages 249–283. Wiley, 1999.

[9] M.J. Benton. Stems, nodes, crown clades, and rank-free lists: is Linnaeus dead? *Biological Reviews*, 75:633–648, 2000.

[10] M.J. Benton. Finding the tree of life: Matching phylogenetic trees to the fossil record through the 20th century. *Proc. Roy. Soc. London Ser. B*, 268:2123–2130, 2001.

[11] J. Alroy et al. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proc. Nat. Acad. Sci.*, 98:6261–6266, 2001.

[12] W.J. Ewens. *Mathematical population genetics*. Springer, 1979.

[13] W. Feller *An Introduction to Probability Theory and its Applications*. Vol.1, 3rd ed. John Wiley and Sons, 1968.

[14] M. Foote and J.J. Sepkoski Jr. Absolute measures of the completeness of the fossil record. *Nature*, 398:415–417, 1999.

[15] S. J. Gould. *The Structure of Evolutionary Theory*. Harvard Univ. Press, 2002.

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

[17] S.B. Heard. Patterns in tree balance among cladistic, phenetic and randomly generated phylogenetic trees. *Evolution*, 46:1818–1826, 1992.

[18] S.B. Heard. Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evolution*, 50:2141–2148, 1996.

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

[20] J.P. Huelsenbeck, B. Larget, R.E. Miller, and F. Ronquist. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, 51:673–688, 2002.

[21] D. Jablonski. The future of the fossil record. *Science*, 284:2114–2116, 1999.

[22] M. Kirkpatrick and M. Slatkin. Searching for evolutionary pattern in the shape of a phylogenetic tree. *Evolution*, 47:1171–1181, 1993.

[23] J. K. McKee. Turnover patterns and species longevity of large mammals from the late Pliocene and Pleistocene of Southern Africa: A comparison of real and simulated clades. *J. Theoret. Biol.*, 172:141–147, 1995.

[24] J.K. McKee. Faunal turnover rates and mammalian biodiversity of the late Pliocene and Pleistocene of Eastern Africa. *Paleobiology*, 27:500–511, 2001.

[25] M. Möhle. Ancestral processes in population genetics. *J. Theor. Biol.*, 204:629–638, 2000.

[26] A.Ø. Mooers. Tree balance and tree completeness. *Evolution*, 49:379–384, 1995.

[27] A.Ø. Mooers and S. B. Heard. Inferring evolutionary process from phylogenetic tree shape. *Quarterly Rev. Biology*, 72:31–54, 1997.

[28] S. Nee. Inferring speciation rates from phylogenies. *Evolution*, 55:661–668, 2001.

[29] S. Nee, E.C. Holmes, R.M. May, and P.H. Harvey. Extinction rates can be estimated from molecular phylogenies. *Philos. Trans. Roy. Soc. London Ser. B*, 344:77–82, 1994.

[30] S. Nee and R. May. Extinction and the loss of evolutionary history. *Science*, 278:692–694, 1997.

[31] S. Nee, R.M. May, and P.H. Harvey. The reconstructed evolutionary process. *Philos. Trans. Roy. Soc. London Ser. B*, 344:305–311, 1994.

[32] M.E.J. Newman. Self-organized criticality, evolution and the fossil extinction record. *Proc. Roy. Soc. London Ser. B*, 263:1605–1610, 1996.

[33] R.D.M. Page and E.C. Holmes. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, 1998.

[34] E. Paradis. Assessing temporal variations in diversification rates from phylogenies: Estimation and hypothesis testing. *Proc. Roy. Soc. London Ser. B*, 264:1141–1147, 1997.

[35] E. Paradis. Detecting shifts in diversification rates without fossils. *American Naturalist*, 152:176–187, 1998.

[36] L. Popovic. Asymptotic genealogy of a critical branching process. Technical Report 628, U.C. Berkeley Statistics Dept., 2002.

[37] L. Popovic. Asymptotic genealogy of a branching process and a model of macroevolution. PhD Thesis, U.C. Berkeley Statistics Dept., 2003.

[38] D.M. Raup. Biases in the fossil record of species and genera. *Bull. Carnegie Museum Natural History*, 13:85–91, 1979.

[39] D.M. Raup. *Extinction: Bad Genes or Bad Luck*. W.W. Norton, New York, 1991.

[40] D.M. Raup, S. J. Gould, T.J.M. Schopf, and D.S. Simberloff. Stochastic models of phylogeny and the evolution of diversity. *J. Geology*, 81:525–542, 1973.

[41] H.E. Robeck, C.C. Maley, and M.J. Donoghue. Taxonomy and temporal diversity patterns. *Paleobiology*, 26:171–187, 2000.

[42] S. Semple and M. Steel. *Phylogenetics*. Oxford Univ. Press, to appear.

[43] J.J. Sepkoski Jr. A compendium of fossil marine animal families. *Milwaukee Publ. Mus. Contrib. Biol. Geol.*, 83:1–156, 1992.

[44] J.J. Sepkoski Jr. Ten years in the library: New data confirm paleontology patterns. *Paleobiology*, 19:43–51, 1993.

[45] J.J. Sepkoski Jr. and D.C. Kendrick. Numerical experiments with model monophyletic and paraphyletic taxa. *Paleobiology*, 19:168–184, 1993.

[46] J.J. Sepkoski Jr. Competition in macroevolution: the double wedge revisited. In D. Jablonksi et al, editor, *Evolutionary Paleobiology*, pages 211–255. University Chicago Press, 1996.

[47] A.B. Smith and C. Patterson. The influence of taxonomic method on the perception of patterns of evolution. In M.K. Hecht and B. Wallace, editors, *Evolutionary Biology, volume 23*, pages 127–216. Plenum Press, 1988.

[48] D. Stoyan, H. Stoyan, and Th. Fiksel. Modelling the evolution of the number of genera in animal groups (Yule's problem revisited). *Biometrical Journal. Journal of Mathematical Methods in Biosciences.*, 25:443 – 451, 1983.

[49] S. Tavaré, C. R. Marshall, O. Will, C. Soligo, and R.D. Martin. Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature*, 416:726–729, 2002.

[50] TreeBASE. A database of Phylogenetic Knowledge. www.treebase.org.

[51] C. Tudge. *The Variety of Life.* Oxford Univ. Press, 2000.

[52] K. Wollenberg, J. Arnold, and J.C. Avise. Recognizing the forest for the trees: Testing temporal patterns of cladogenesis using a null model of stochastic diversification. *Mol. Biol. Evol.*, 13:833–849, 1996.

[53] G.U. Yule. A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis. *Philos. Trans. Roy. Soc. London Ser. B*, 213:21–87, 1924.