Five Statistical Questions about the Tree of Life

David Aldous∗
Department of Statistics
367 Evans Hall # 3860
U.C. Berkeley, CA 94720-3860
aldous@stat.berkeley.edu

Maxim Krikun
Institut Elie Cartan
Université Henri Poincare
Nancy, France
krikun@iecn.u-nancy.fr

Lea Popovic
Dept of Mathematics and Statistics
Concordia University
Montreal Canada H3G 1M8
e-mail: lpopovic@mathstat.concordia.ca

January 6, 2008

Abstract

Recent work on stochastic modeling of phylogenies raises five questions to which we would like to draw the attention of quantitatively inclined biologists. (1) For a clade on \( n \) extant species, write \( m \geq n \) for the maximum number of coexisting species at any past time; how large do we expect a priori that the ratio \( m/n \) might be? (2) From a correct phylogeny of the extant species of a clade, what can we deduce about past speciation and extinction rates? (3) What proportion of extant species are in fact descendants of still-extant ancestral species, and how does this compare with predictions of models? (4) The “unbalanced” nature of phylogenetic trees has often been noted. When one moves from trees on species to trees on sets of species (whether traditional higher order taxa, or clades within PhyloCode) does one expect trees to become more unbalanced as a purely logical consequence of tree structure, without signifying any real biological phenomenon? (5) Time series of taxa counts from the distant past fossil record indicate numbers of observed genera or families. How do we expect that fluctuation rates for counts of higher order taxa should compare to fluctuation rates for number of species?

1 Introduction

There is a long tradition, dating back at least to Yule [35], of study of simple stochastic models for aspects of biodiversity or macroevolution – the evolutionary history of speciations and extinctions. Such models have several distinct potential uses, for instance:

- Amongst several approaches to reconstructing phylogenies on extant species from molecular data, the Bayesian approach [14] requires a prior distribution on possible trees.

- As null models when seeking to attribute biological significance to data. Is a particular clade expansion an instance of adaptive radiation [8]? Does a typical clade expansion represent expansion into novel niches, or replacement of other clades [7]? Were speciation rates unusually high during the Cambrian radiation [18]? Are the apparent “pulses” of extinctions and speciations seen in the fossil record of large mammals in southern and eastern Africa over the last 3 Myr real, or could they be an artifact arising from the limited number of different dates of sites yielding the fossils [19, 20]?

- Detecting analogs of the regression fallacy [11] in which some observed effect (predictable on statistical grounds) is incorrectly thought to require some casual explanation.

- xxx more

The authors of this paper are mathematicians, and make no claim to answer biology questions, but following the logic of mathematics encourages us to pose questions that are perhaps framed differently from those that a professional biologist might pose. From our recent mathematical work on a particular model for a clade with \( n \) extant species [5] and on induced models for phylogenetic trees on genera [4] (the parts relevant to this paper will be reviewed briefly in sections 2.1 and 5.3), we have distilled for this paper five topics on which we have observations and questions that we hope will interest some biologists. Let us jump right into these topics (sections 2 - 6), postponing further general discussion to a final section 7.

2 Past fluctuations in sizes of extant clades

Readers will be familiar with the following 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 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 [25]: our emphasis added)

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 statistic

\[ R = \frac{\text{maximum number of species at any one past time}}{\text{current number of species}}. \]

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 ([5] Corollary 6) the \( 1/r \) law:

\[ P(R \geq r) = 1/r, \quad 1 \leq r < \infty \]

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 \( R = 1 \)?

**Suggested project.** Look at small extant clades (size 5 – 20 species, say) with extensive fossil records, and attempt to estimate the distribution of \( R \) from the fossil record.

Having quoted a result within our model, we should now describe the model.
2.1 The species-level model

The following model [5] for the phylogenetic tree on a clade with \( n \) extant species (for given \( n \)) is intended to capture the intuitive idea of “purely random macroevolution”. In (b) the phrase “rate 1” means “with probability \( dt \) in each time interval of length \( dt \)”. So the time unit in the model equals mean species lifetime. The Bayesian terminology (prior, posterior) is convenient but remember we are just making a mathematical definition, not doing statistical analysis.

(a) The clade originates with one species at a random time before present, whose prior distribution is uniform on \((0, \infty)\).

(b) As time runs forward, each species may become extinct or may speciate, each at rate 1.

(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 now yields a mathematically defined phylogenetic random tree on \( n \) extant species and a random number of extinct species. Note that this is a “complete” tree in which lineages of extant species pass through explicit extinct species, but of course removing this information gives what we will call the “lineage tree” on the \( n \) extant species (and this is what is usually called the “phylogenetic tree” on extant species).

3 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. Let us illustrate of one aspect of variability. Consider three quantities associated with our species model: the time \( T_{\text{origin}}^n \) of clade origin; the time \( T_{\text{mrca}}^n \) of the most recent common ancestor of extant species; and the number \( Z_n \) of species at the time of most recent common ancestor (recall the time unit equals mean species lifetime). It turns out that these quantities grow linearly with \( n \) but there is no “law of averages”; instead as \( n \to \infty \) there is a limit joint distribution, say \((T_{\text{origin}}^n, T_{\text{mrca}}^n, Z)\), for the normalized statistics \((n^{-1}T_{\text{origin}}^n, n^{-1}T_{\text{mrca}}^n, n^{-1}Z_n)\) and an explicit formula for this limit joint distribution can be found ([5] Corollary 8). Figure 1 illustrates this joint
distribution by showing 10 realizations. The key point to note is that each of these quantities varies by a factor of 10 over the realizations.

\[ t = s \]

Figure 1. Scatter diagram of 10 realizations of the standardized joint distribution \((T_{\text{origin}}, T_{\text{mrca}}, Z)\). Points \(\times\) give the \((T_{\text{origin}}, T_{\text{mrca}})\)-values, and the length of line segment is the \(Z\)-value. Note the extreme variability of \(Z\); the smallest value was 0.03 and the largest was 1.56.

Our site [2] shows, for a selection of values of \(n\), 10 realizations of the lineage tree on \(n\) extant species derived from our model, demonstrating vividly that different realizations can look very different. Three of the 10 realizations for \(n = 20\) are shown here in Figure 2. 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. Implications of this variability are discussed in the next two subsections.
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
\[ R = \frac{\text{max number species}}{\text{current number species}} \]
\[ R = 2.30 \]
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
$R = \frac{\text{max number species}}{\text{current number species}}$ 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
$R = \frac{\text{max number species}}{\text{current number species}}$ 1.20
Number of extinct species 39

**Figure 2.** Tree realizations of the lineage tree in our model, $n = 20$. These are realizations 1, 5, 6 from [2].
3.1 Qualitative assessments of nonrandomness

Wollenberg et al [34] show three published examples of lineage trees (columbines, cranes, *Drosophila virilis*: comments refer to the B examples) which look quite different. After doing a statistical test of significance with reference to a probability model similar to ours, they conclude

\[
\ldots \text{consistent with the original authors’ impressions, the columbines and the cranes do indeed display nonrandom phylogenetic patterns of diversification, with the columbines showing recent and the cranes showing ancient significant clustering of speciation events \ldots} \ldots \text{[also consistent were] the results for the *D. virilis* group, where no evidence of temporal nonrandomness [could be identified].}
\]

But the visual difference between these three empirical trees is no larger than the visual difference between realizations of our model – indeed, the three realizations in Figure 2 resemble the (recent, ancient, steady) diversification visible for (columbines, cranes, *D. virilis*) – and we suspect that any reasonable test statistic would indicate that each data tree from [34] might plausibly arise “by chance” within our model. The point is that our model, with its intrinsic greater variability, provides a *more conservative* approach to assessing significance of observed features of phylogenetic trees.

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

More concretely, consider the problem of estimating past speciation and extinction rates within a clade using only the lineage tree (assumed correct) on extant species. An often used model is the linear birth-and-death model, which we regard as having 3 parameters \((t^*, \lambda, \mu)\), where

\[
\begin{align*}
    t^* &= \text{time before present of clade origin} \\
    \lambda_i &= \text{total speciation rate, when } i \text{ species} \\
    \mu_i &= \text{total extinction rate, when } i \text{ species}.
\end{align*}
\]

It is routine\(^1\) to calculate numerically maximum likelihood estimates (MLEs) of the parameters, based on a lineage tree as data.

\(^1\)The only subtle issue is that one should compute the likelihood *without* conditioning on \(n\). To see why, note that when \(\mu/\lambda\) 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.
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 model (from section 2.1). Of course, in our model we really have $\lambda = \mu = 1$ in each realization, and realization-dependent values of $T^{\text{mrca}}$ and $T^{\text{origin}}$. Table 1 shows the MLEs derived from each of 10 typical realizations of the model.

<table>
<thead>
<tr>
<th>realization</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE of $\lambda$</td>
<td>0.4</td>
<td>0.6</td>
<td>1.3</td>
<td>1.5</td>
<td>0.6</td>
<td>1.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.9</td>
<td>1.2</td>
</tr>
<tr>
<td>MLE of $\mu$</td>
<td>0.1</td>
<td>0.4</td>
<td>0.2</td>
<td>1.0</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>MLE of $t^*$</td>
<td>9.1</td>
<td>6.5</td>
<td>1.7</td>
<td>3.1</td>
<td>4.7</td>
<td>2.5</td>
<td>11.4</td>
<td>13.8</td>
<td>4.2</td>
<td>1.5</td>
</tr>
<tr>
<td>$T^{\text{mrca}}$</td>
<td>8.4</td>
<td>6.0</td>
<td>1.5</td>
<td>4.3</td>
<td>4.3</td>
<td>2.4</td>
<td>11.3</td>
<td>13.3</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>$T^{\text{origin}}$</td>
<td>37.1</td>
<td>13.3</td>
<td>1.7</td>
<td>8.2</td>
<td>8.9</td>
<td>11.9</td>
<td>36.2</td>
<td>154.7</td>
<td>21.8</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Table 1. MLE estimates of linear birth-death parameters based on each of 10 realizations from our model ([2], $n = 8$).

So in this setting the estimated values of $\lambda$ and $\mu$ in the linear model are very misleading. Not only does a “pull of the recent” effect make the estimated $\lambda$ larger than the estimated $\mu$, but also the estimated values are varying widely between realizations. As a secondary point, observe that in most realizations the estimated time $t^*$ of clade origin is about 10% greater than the observed time of most recent 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.

### 3.3 Discussion

While the results above are not directly comparable with those in existing literature (e.g. Nee et al [26, 24, 23], Paradis [28, 29]) they do cast serious doubt on the ability to reconstruct at any level of detail the history of a single clade from the phylogeny of extant species. In contrast, a statistical study of phylogenetic trees (with relative time scale) from many clades might provide some insight into typical patterns of recent speciation and extinction rates.

xxx maybe our model not realistic, but is it LESS realistic than others?

---

2The variability is not a “small $n$” effect: as emphasized throughout the paper, variability in our model persists in the $n \to \infty$ limit.
4 Extant ancestral species

Polar bears are often said to be a recent (200 Kyr) daughter species of brown bear [9]. This is an unusual (amongst familiar animals) instance where an ancestral species is extant, and suggests the question

within well-studied extant clades, what proportion $\alpha$ of extant species have some extant ancestor?

Anecdotally, biologists regard $\alpha$ as small, though we have been unable to find useful data, perhaps in part because cladistics dogma discourages asking this question. Our model predicts ([5] Corollary 11) that for about 63% of extant species, some ancestral species should be extant. While this numerical value depends on rather arbitrary details of the particular 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%) of extant species will have extant ancestors.

This contrast between models and data is striking to us, though biologists with whom we have discussed this issue tend to dismiss it as simply reflecting the practical difficulty in distinguishing similar species. But our point is that all existing statistical study of questions involving past speciation or extinction rates implicitly depends on models of the type which predict non-small $\alpha$, and so all such work would appear less convincing if data really shows $\alpha$ to be small.

4.1 Models predicting small $\alpha$

Of course it is not hard to invent models which do predict small values for $\alpha$; here is one such. Recall that our basic model assumes that 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. (One possible interpretation of “$\alpha$ is small” is as evidence in favor of at least moderate prevalence of anagenesis, but we refrain from entering such controversies). To make a cladogenesis model with small $\alpha$, 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 the stated relative chances:

\[
\begin{align*}
\text{extinction} & : \theta \\
\text{speciation} & : \theta \\
\text{replacement by a daughter species} & : 1 - 2\theta.
\end{align*}
\]
Such models will give a small value of $\alpha$ (specifically $\alpha = 2\theta(1 - e^{-2\theta})$), precisely because they are effectively interpolating between cladogenesis and anagenesis (smaller $\theta$ giving a larger contribution of anagenesis). Thus one possible explanation of small $\alpha$ is that the dominant mode of macroevolution is replacement of species by daughter species rather than branching phylogenies.

5 Phylogenetic tree shape and hierarchical level

It is a longstanding puzzle [22] 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. Imbalance in a particular clade is often taken as evidence of some particular biological effect, as illustrated by the following quote.

We combine statistical and phylogenetic approaches to test the hypothesis that adaptive radiation and key innovation have contributed to the diversity of the order Araneae. The number of unbalanced araneid clades (those whose species numbers differ by 90% or more) exceeds the number predicted by a null Markovian model. The current phylogeny of spider families contains 74 bifurcating nodes, of which 31 are unbalanced. As this is significantly more than the 14.8 expected unbalanced nodes, some of the diversity within the Araneae can be attributed to some deterministic cause (e.g., adaptive radiation). [8]

In the different context of “deep phylogeny” relating extinct and extant groups, one frequently sees “comb” cladograms such as that in Figure 3, in which single leaves split off one after another.
Figure 3. A phylogeny of terrestrial vertebrates [17].

One can think of three interpretations of such cladograms. One possibility is that they are just wrong, i.e. do not show the correct relationship between monophyletic terminal clades. An opposite possibility is that they are correct and show some significant biological effect, such as xxx. An intermediate possibility is that such cladograms are simultaneously correct and artifactual, in the sense that a species-level tree would be more balanced, but the procedure of drawing cladograms whose terminal taxa are clades tends to produce more unbalanced trees for purely mathematical reasons without biological significance. This is a possibility that we can study mathematically, using three variant theoretical schemes developed in detail in [4] for extending the section 2.1 species-level model to trees whose terminal taxa are clades. (For concreteness of language we write genera, but we mean any reasonable way of partitioning species into groups.)

We review the schemes in section 5.3 but the general idea can be seen from the illustrative example in Figure 4. We imagine a systematist is given a correct phylogeny on species, and chooses a number of apomorphies (a trait which characterises an ancestral species and its descendants; we call the ancestral species a new type) judged significant. Figure 4 illustrates three ways in which the same underlying data (a phylogeny on 22 species, of which 2 are new type) can be used to define genera and hence a tree on
genera. We call the three schemes fine, medium, coarse. Very roughly, one can interpret the fine scheme as attempting to conform with cladistic principles by making taxa be monophyletic, whereas the coarse scheme is what a traditional taxonomist given only fossil data showing these two apomorphies (and not knowing the true tree on species) might devise. Each scheme can be defined by an algorithm, or characterized abstractly via certain intrinsic properties, and such mathematical characterizations (stated in section 5.3) provide the motivation for using these particular schemes.

xxxx not what systs actually do, of course
Figure 4. Illustration [4] 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 · · · uv), with two species \{i, s\} designated as new types and marked •. 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, and giving 3 genera (coarse scheme) or 4 genera (medium scheme).
5.1 Measuring tree balance

There is a sizable literature (see e.g. [1]) on summary statistics $T_n$ measuring “balance” of an $n$-leaf tree and their distribution under the usual Markov model as a null hypothesis. While this is the natural way to study individual trees, it is not satisfactory for studying the overall statistical properties of a collection of trees of different sizes, because to compare different $n$ one needs to standardize, and the standardization depends on the null model, which is generally regarded [22] as unrealistic.

A different method, used in [3] on a few large trees, seems less arbitrary and more powerful way of analyzing individual large trees or collections of small trees. 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). A companion function, useful in the context of studying occurrence of comb-like cladograms, is

$$p(m) = \text{proportion of splits of size-}m \text{ clades where smaller daughter clade size = 1.}$$

5.2 Predictions of our model

Let’s review what we’re doing. On the base model (section 2.1) for “purely random” macroevolution of species, we superimpose a scheme (one of three schemes to be specified in section 5.3) for defining groups based on randomly arising apomorphies which do not confer advantage, that is the speciation and extinction rates do not change. This requires a parameter $\theta$ indicating rate of new apomorphies, which determines the mean number of species per genus. The result is a model of trees on groups of species, intended as a null model for “purely random” macroevolution as seen in data which we are unable to resolve to the level of species. Our use of this model in this section 5 (section 6 describes another use) is to study whether the procedure of drawing cladograms whose terminal taxa are clades tends to produce more unbalanced trees for purely mathematical reasons without biological significance.
In studying the model, there are statistical differences between extinct and extant clades, and between the three schemes for constructing genera. We consider first the case of extant clades and fine genera.

Table 2 shows the measures $a(m)$ and $p(m)$ of tree balance within our model. Recall we are talking about trees whose terminal taxa are the genera. Within such trees, we look at clades consisting of $m = 5, 10, 15$ genera, and for the root split of such clades we record the median size $a(m)$ of smaller subclade and the proportion $p(m)$ of smaller subclades with size 1. This is repeated for different values of the parameter $\theta$ chosen to make the mean number of species per genus be $\mu = 1, 5, 10, 20$. The first row ($\mu = 1$) is just the species-level model, where our model coincides with the usual Markov model and predicts uniformly distributed splits.

<table>
<thead>
<tr>
<th>$\mu = \text{mean number species per genus}$</th>
<th>parent clade size $m$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(5)$</td>
<td>$a(5)$</td>
<td>$p(10)$</td>
</tr>
<tr>
<td>1</td>
<td>0.50</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>0.73</td>
<td>1.2</td>
</tr>
<tr>
<td>10</td>
<td>0.79</td>
<td>1.1</td>
</tr>
<tr>
<td>20</td>
<td>0.85</td>
<td>1.1</td>
</tr>
</tbody>
</table>

**Table 2. Shape of tree on extant fine genera.** For the given size (number of genera) $m$ in a parent clade, the table shows the probability $p(m)$ that smaller daughter clade size equals 1, and the median size $a(m)$ of smaller daughter clade. Results from Monte Carlo simulations of model with 400 extant species. We used a continuized median, being more informative than the discrete median.

Increased imbalance is indicated by $p(m)$ increasing and by $a(m)$ 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.

This analysis suggests the possibility that observed imbalance in trees on higher level taxa may be an artifact of classification. On the other hand, two studies [12, 21], based on summary statistics, of published small trees discussed in [22] conclude there is no such hierarchical trend in imbalance. It would be interesting to repeat such data analysis on larger trees.
5.3 The three genera classification schemes

Given a correct phylogeny on species, and certain of these species designated “new type”, what is a sensible way to partition the species into genera? We treated this [4] as a mathematics question, by first writing down what logical properties would we like such a classification scheme to possess. It turns out that, as with voting systems [33], no scheme for assigning species to genera has all the properties one might like, which is why we end up with three different schemes.

Property 1. A genus cannot contain both a species \( a \) which is a descendant of some “new type” species \( s \) and also a species \( b \) which is not a descendant of \( s \).

Here “descendant” includes \( s \) itself, so in particular a “new type” species and its parent must be in different genera.

Next note that if we required every genus to be a clade (monophyletic) then we could never have more than one genus, because otherwise some parent-daughter pair \( \{a, b\} \) would be in different genera and then the genus containing \( a \) is not a clade. We will consider a weaker property. Any two distinct species \( \{a, b\} \) have a most recent common ancestor \( \text{MRCA}(a, b) \), which is some species (maybe \( a \) or \( b \)). Given three distinct species \( \{a, b, c\} \), say \( (a, b) \) are more closely related than \( (a, c) \) if \( \text{MRCA}(a, b) \) is a descendant of \( \text{MRCA}(a, c) \). Here again we allow \( \text{MRCA}(a, b) = \text{MRCA}(a, c) \).

Property 2. Given three distinct species \( \{a, b, c\} \), with \( a \) and \( b \) in the same genus and \( c \) in a different genus, then \( (a, b) \) are more closely related than \( (a, c) \).

As another kind of desirable property, one would like to be able to draw a tree or cladogram on genera in some unique way, and the next property (for a classification scheme) provides one formalization of this idea.

Property 3. Choosing one representative species from each genus and drawing the cladogram on these species gives a cladogram which does not depend on choice of representative species.

Note that the trivial scheme of assigning each species to a different genus possesses all these properties, so the issue is to find the coarsest schemes with such properties. It turns out that we can define three genera classification schemes (coarse, medium, fine) as the coarsest schemes with the following properties.

<table>
<thead>
<tr>
<th>Coarse</th>
<th>Property 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td>Properties 1 and 3</td>
</tr>
<tr>
<td>Fine</td>
<td>Properties 1 and 2</td>
</tr>
</tbody>
</table>

Theorem 1 of [4] gives simple algorithmic descriptions of the constructions
of the genera in each scheme.
6 Fluctuations at different hierarchical levels

Sepkoski’s compendia [32] (see also Benton [6]) 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, raising the issue of how reliable is it as a proxy for time series for numbers of species. Paleontologists tend to regard it as reliable, as indicated in the first sentence below.

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. ([15], our emphasis added).

We focus on the emphasized assertion: is this inevitability a biological or a mathematical effect? Suppose we have data, for a sequence of times $\tau(0), \tau(1), \ldots, \tau(k)$, on number of families at each time, and numbers of genera in each family at each time. How should we measure volatility of “number of families” and of “number of genera in a family” to ascertain at which level there is greater volatility? After deciding that, our model provides one way of thinking about how we should expect volatilities to vary between levels, as a purely mathematical effect.

6.1 Standardized fluctuation rates

We take the basic species-level model of section 2.1, but now (thinking of extinct clades) do not condition on number of extant species. 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, and recalling the time unit is mean species lifetime, say $\mu_s$. A basic mathematical property of our model is that the stochastic fluctuations, measured by variance of changes in the time series, have a simple form over times intervals of the same order as mean species lifetime. Given $N(t_0) = n(0)$ we have

$$\text{var}(N(t_0 + t) - n(0)) \approx 2n(0)t.$$
Translating into real time units $\tau$ (measured in Myear, say) gives

$$\text{var}(N(\tau_0 + \tau) - n(0)) \approx 2n(0)\tau/\mu_s.$$ 

In other words the ratio

$$\frac{\mu_s \text{var}(N(\tau_0 + \tau)) - n(0)}{2n(0)\tau}$$

is approximately 1 regardless of the value of $n(0)$ (assumed not too small) or the value of $\tau$ (assumed of order $\mu_s$).

This analysis suggests a way to define standardized fluctuation rates for time series of genera. Write $\mu_g$ for mean genus lifetime. Given $G(\tau_0) = g(0)$ define

$$\text{standardized fluctuation rate for genera} = \frac{\mu_g \text{var}(G(\tau_0 + \tau)) - g(0)}{2g(0)\tau}$$

for $\tau$ of order $\mu_g$. The point is that, if we were to model genera directly as behaving statistically like species, then the standardized fluctuation rate for genera would equal 1, and so a different fluctuation rate within our model is indicating a mathematical effect of using genera classification schemes.

Table 3 gives numerical values in one case.

<table>
<thead>
<tr>
<th>mean number species per genus</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>standardized fluctuation rate</td>
<td>1.00</td>
<td>0.84</td>
<td>0.68</td>
<td>0.57</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 3. Standardized fluctuation rates for coarse genera. Results from Monte Carlo simulations of model started with 400 species. Ratio (2) estimated with $\tau = \text{mean genus lifetime}$.

Returning to the original issue (data presented as time series for numbers of genera and families), if the data showed an average of 8 genera per family, and if we were willing to guess an average of 5 species per genus, then Table 5 gives a prediction that (under our basic “neutral” model and the coarse genera scheme) the fluctuation rate for families relative to the fluctuation rate for genera would be $0.45/0.84 = 0.54$. Of course it would be interesting to try to estimate such ratios in paleontological data.
Of course, such questions have previously been studied

- found/cited papers: Sepkoski [16] 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. [30] 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.
7 General discussion

Our five topics obviously have connections with a large body of literature in evolutionary theory, and we lack both space and competence to describe all related existing literature.

The monograph of Schluter [31] starts with

Adaptive radiation is a spectacular feature of evolution. It is also widespread, more so than the list of familiar cases . . . . . . would suggest. Much of life’s diversity, perhaps even most of it, has arisen during similar episodes of speciation and phenotypic and ecological divergence.

xxx compares AR to null model, not alternate; individual clades.

xxx suprising lack of study of big question ”how much”

xxx puzzle that no one seems to have tried to tackle the big picture of looking statistically at all the data for evolution of species – xxx extant and fossil – – in signal-to-noise sense: what done, how to develop further? Monograph?

xxx it is prudent to be suspicious of any scientific theory that purports to explain everything in its field.

xxx view our stuff as a “thought experiment” rather than realistic modeling. xxx amateur status; ask the unanswerable (cf. R).

We are skeptical about the use of such models as formal null hypotheses in formal statistical tests of significance, (typical such uses are xxx Yule xxx Gould xxx adaptive radiation)

xxx square root law means neutral unrealistic

xxx “coherent” – tree balance and extinct/extant and time series

Finally we should mention two somewhat different bodies of literature. One, from statistical physics (exemplified by [27]), studies a variety of more elaborate models to determine which models predict power-law distributions for quantities such as sizes of extinction events, lifetimes of genera or number of species per genus. The second [13] studies models which combine speciation with ecological theories of island biogeography and relative species abundance.

xxx census analogy

xxx closest we know is the survey [22] on the particular issue of inferring evolutionary process from phylogenetic tree shape.

xxx Sep library?

xxx big Q: is everything adaptive radiation
xxx big Q: clade expansion is typically expansion into novel niches, or replacement of other clades
xxx age of size n clade – relate to population/lifetimes
xxx somewhere; our theory is $n = \infty$; what about $n = 4$ for e.g. $T_{\text{origin}}$?
xxx overstating variability because really know something about time of origin
xxx models where speciation and extinction rates vary between lineages can of course explain any tree shape xxx so not useful.
References


XXX - I looked at the supertree of Dinosauria [10], and made a scatterplot of splits: (parent clade size, smaller daughter clade size)

![Scatterplot of splits in the supertree of Dinosauria](image)

**Figure 8.** Scatterplot of splits in the supertree of *Dinosauria:* points * give the \((n, m)\)-values, where \(n\) is the size of the clade and \(m\) is the size of its smaller daughter clade. The Coless’ index of imbalance for this supertree is \(I_c \approx 0.09\).

XXX - according to \(I_c\) this is not so badly imbalanced; I didn’t dare plot the line through the graph and get an estimate for \(a(n)\). For Markov model on a full \(n\)-clade \(E(I_c) \approx \frac{\log(n/2)}{n/2}\), but without estimates of \(\sigma^2(I_c)\) it may not be so useful to know that under the Markov model the Dinosauria supertree has \(E(I_c) \approx 0.4\).