Probability Distributions on Cladograms*

DAVID ALDOUS†

August 22, 1996

in: Random Discrete Structures, ed. D. Aldous and R. Pemantle, p. 1-18. Springer (IMA Volumes Math. Appl. 76), 1996.

Abstract

By analogy with the theory surrounding the Ewens sampling formula in neutral population genetics, we ask whether there exists a natural one-parameter family of probability distributions on cladograms ("evolutionary trees") which plays a central role in neutral evolutionary theory. Unfortunately the answer seems to be "no" – see Conjecture 2. But we can embed the two most popular models into an interesting family which we call "beta-splitting" models. We briefly describe some mathematical results about this family, which exhibits qualitatively different behavior for different ranges of the parameter β.

1 Probability distributions on partitions and neutral population genetics

The first few sections give some conceptual background. The reader wishing to "get right to the point" should skim these and proceed to section 3.

For each n there is a finite set of partitions of $\{1, 2, \ldots, n\}$ into unordered families $\{A_1, A_2, \ldots, A_k\}$ of subsets. A one-parameter family $(P_\theta^{(n)})$ of probability distributions on this set of partitions can be defined by

$$P_\theta^{(n)}\{A_1, \ldots, A_k\} = \frac{\Gamma(\theta)}{\Gamma(n+\theta)} \prod_{i=1}^{n}((i-1)!\ \theta)^{m_i}$$

where m_i is the number of A's with exactly i elements. The right side is a slightly disguised statement of the *Ewens sampling formula* in neutral population genetics, which is central to a mathematically rich and elegant theory which has made an impact in non-mathematical genetics. See e.g. [16, 26, 11, 9]. Our

*Research supported by N.S.F. Grant DMS92-24857.

†Department of Statistics, University of California, Berkeley CA 94720.

purpose in this paper is to ask whether there is an analogous theory for neutral evolutionary trees. In particular, we study whether the following two results have analogs for evolutionary trees.

(a) *Consistency characterization.* A family $(P^{(n)}; 1 \leq n < \infty)$ satisfies the following three conditions iff it is one of the $(P_\theta^{(n)})$.

(i) Exchangeability. For each n, the distribution is invariant under permutations of the labels $\{1, 2, \ldots, n\}$.

(ii) Sampling invariance. For each n, $P^{(n)}$ induces a distribution on partitions of $\{1, 2, \ldots, n-1\}$ by the action of deleting n: this distribution is $P^{(n-1)}$.

(iii) Subset deletion. For each $j < n$, given that $P^{(n)}$ has $\{j+1, j+2, \ldots, n\}$ as a set in the partition, the remaining partition of $\{1, 2, \ldots, j\}$ has distribution $P^{(j)}$.

(b) *Interpretation via time-evolution.* Suppose there are k neutral alleles (an *allele* is a possible "type" of a gene; *neutral* means to confer no selective advantage or disadvantage). Count the proportions $(X_i(t); 1 \leq i \leq k)$ of a population with allele i in generation t. Then under natural models there is a k-dimensional diffusion $(\tilde{X}_i(t))$ representing the limit (as population size tends to infinity and time is rescaled), where randomness comes from the random number of copies of an individual allele which appear in the next generation. If we also allow random mutations to produce new alleles, we get an infinite-dimensional diffusion ("the infinitely-many-alleles model"). A random sample of n individuals from the population can be partitioned into subsets with identical alleles, and this random partition has distribution $P_\theta^{(n)}$, where the parameter θ is related to the mean number of mutations per generation.

2 Phylogenetic trees

A *phylogenetic tree* is a visual representation of an assertion about relationships between species $A, B, C, \ldots$. There are many varieties of such tree, differing in the details of what exactly is being asserted – see Eldredge and Cracraft [10] for an extensive discussion. Figure 1 is a *cladogram.* The basic interpretation is the obvious one: amongst species $\{C, D, E\}$, the most closely related are D and E, and so on. The species are distinguished, i.e. switching A with D gives a different cladogram. But there is no distinction between right and left edges, i.e. switching A with B gives the same cladogram. And there is no explicit time scale.

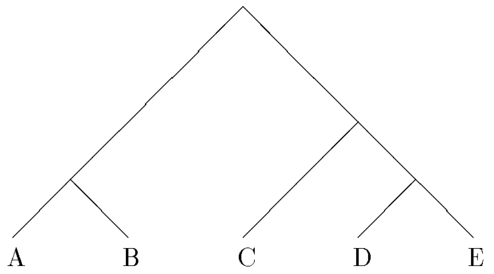


Figure 1

Biologists believe in evolution, and so implicitly believe there is a true "evolutionary tree" linking all living and extinct species, and that classifications of species should be consistent with evolutionary history. The classical Linnaean hierarchy (originally *species, genus, order, class, kingdom* but subsequently extended to many more ranks) remains in practical use, but theoreticians have conducted a vigorous debate about how classification ought to be done.

Figure 2 is one way to picture a true evolutionary tree. Species are represented by vertical lines, from their time of origin to their time of extinction, with dotted horizontal lines indicating the origin of a species from its parent species. Implicit in such a picture are a set of generally-held beliefs about evolution (e.g. that species arise comparatively quickly and then remain largely unchanged until extinction) which I won't go into.

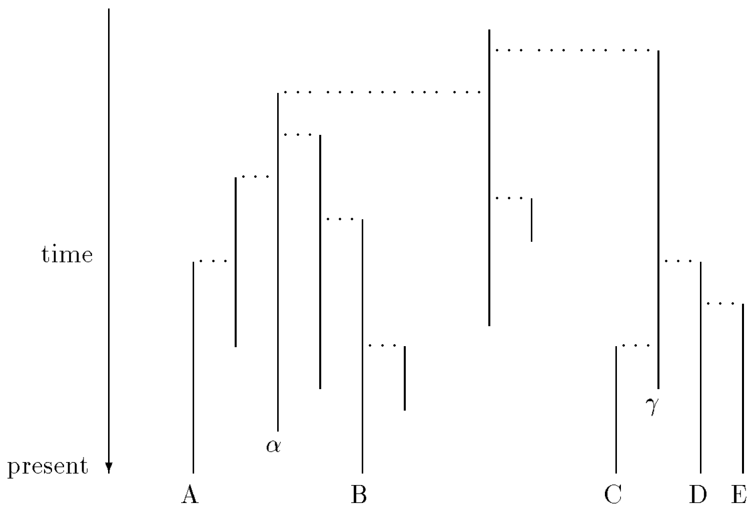


Figure 2

In practice one seldom has enough information about extinct species to be able to draw a tree as detailed as Figure 2, but it is useful to envisage such a tree in order to avoid drawing unwarranted inferences from other representations. To illustrate, consider the cladogram in Figure 1, which is consistent with the evolutionary tree of Figure 2. ¿From the cladogram one might think, loosely speaking, that A and B are more closely related to each other than are C and D. But the evolutionary tree indicates the opposite is possible, if we measure closeness by either time of divergence or number of intermediate species. More dramatically, a cladogram does not indicate ancestor-descendant pairs. In Figure 1, A and B might be "cousins" (as Figure 2 shows), or one might be an offspring of the other (as Figure 2 shows E to be an offspring of D).

A third type of picture, a *phenogram*, is often used, in particular in the context of molecular genetics analysis of living species.

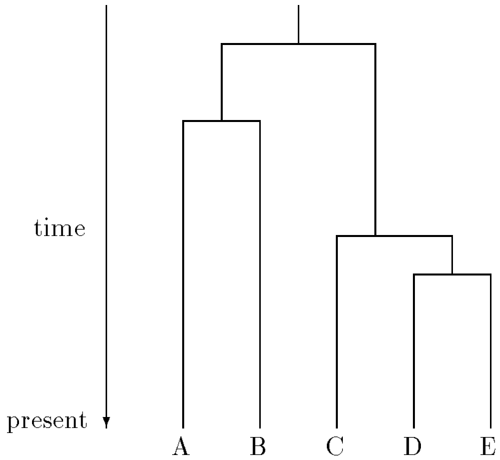


Figure 3

A phenogram contains more information than the cladogram, in that there is an absolute time scale. But the previous warning about casual inferences still holds. ¿From Figure 3 one might assume that the common ancestor α of $\{A, B\}$ lived earlier than the common ancestor γ of $\{C, D, E\}$, whereas Figure 2 shows it is possible that α both originated later than γ and became extinct later than γ.

Minor points. (a) We're going to assume cladograms have only binary splits, although actual cladograms exhibited in the literature usually have some higher-order splits also. One can interpret a higher-order split as a collection of binary splits which cannot be resolved with the data available.

(b) Drawing cladograms with diagonal lines is just a convention, but is useful for distinguishing them from phenograms or other kinds of evolutionary tree.

2.1 Why consider probability models?

There are two quite different reasons for considering probability models of phylogenic trees. The first reason concerns technical aspects of reconstructing trees from data. Molecular biologists in general have eschewed probability models in favor of parsimony (deterministic best-fit) methods, which have the advantage of telling you what tree to write down (up to often-serious non-uniqueness and computational tractability issues), but the disadvantage of not indicating quantitative confidence assertions. To implement a more classical statistical methodology involves a complex array of modeling problems, one of which is to specify an *a priori* model of evolutionary history. A related issue is testing the actual algorithms used: given a hypothetical true phylogenetic tree and a model for the mutation process underlying the observed data, how well does the algorithm reconstruct the tree? Since these procedures are computationally intensive anyway, it would seem better to use a "good" *a priori* model (if we could agree on one!) rather than a model chosen purely for mathematical simplicity.

My own motivation comes from the more conceptual question

> If we had the true evolutionary tree of all species, what could we infer about the relative roles of selectivity and neutrality in the pattern of speciations and extinctions?

Several disparate lines of relevant research appear in the literature. Gould, Raup et al [24, 12, 23] compared paleontological data with random models (essentially critical branching processes). [1] The recent book by Kauffman [15] contains a wide-ranging study of mathematical models of selectivity effects. See also the conference proceedings [22] and the work cited in the next section. But the bottom line is that (in contrast to neutral population genetics) there is no accepted definite notion, at either the conceptual or mathematical level, of "neutral evolution of species".

One could discuss models of any of the varieties of tree discussed in section 2. For the technical issues mentioned above it is most natural to use phenograms, whereas for our purpose of extracting patterns from published trees we shall use cladograms, which are becoming the most common form of published phylogeny.

2.2 Two particular probability distributions on phylogenetic trees

There is a scattered (and mostly mathematically unsophisticated) biological literature on *a priori* models of random phylogenetic trees. Brief surveys are contained in [14, 19]. We describe below the two models which have been most discussed in the biological literature. As noted later these models (under a bewildering variety of different names) have also been extensively studied in other contexts.

The number of different cladograms on n species is

$$c_n = (2n-3)!! = (2n-3)(2n-5)\cdot \quad \cdot 3 \cdot 1.$$

One way to see this is to note that a cladogram on n species has $2n-1$ edges (for this purpose we add an edge upwards from the root) and that each choice of edge at which to add a $n+1$'st species leads to a different cladogram on $n+1$ species, so $c_{n+1} = (2n-1)c_n$. In the <u>uniform model</u> we assume each cladogram is equally likely. The <u>Yule model</u>, considered as a phenogram, is just the elementary continuous-time pure birth process started with one lineage. That is, each lineage persists for a random, exponential(1) time and then splits into two lineages. Continue the process until there are n lineages. The resulting

[1] Their work focused on number of species as a function of time, whereas ours focuses on the combinatorial structure of phylogenetic trees.

random cladogram has an equivalent description as "random joinings", where we count time backwards from the present. Starting with n lines of descent, we choose uniformly at random one of the $n(n-1)/2$ pairs and join the pair, to make $n-1$ lines of descent, and continue until there is a unique line. In either case, throwing away the time scale leaves a random cladogram.

2.3 Search trees

Some probability distributions on cladograms can be associated with well-studied random search trees in computer science. Let us explain the connection briefly.

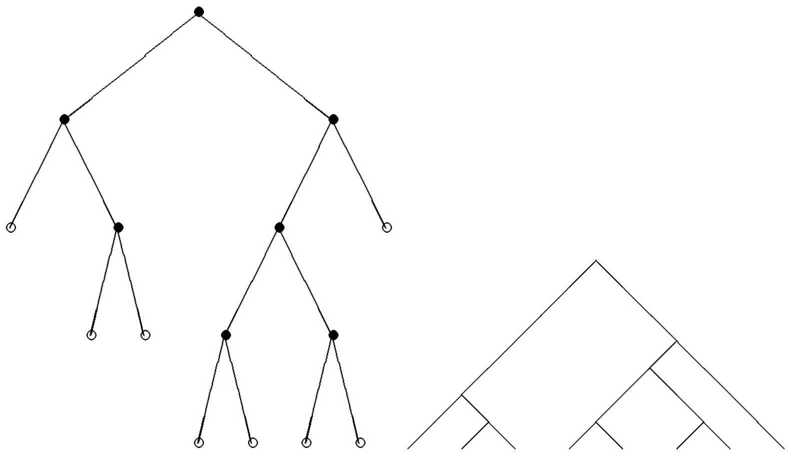


Figure 4

Figure 4 shows a subtree of the infinite binary tree, i.e. a tree in which a node has either zero children, a left child, a right child, or both a left and a right child. Such a "binary tree" can be described using either the $n-1$ *internal* nodes $\bullet$ or, as is customary in computer science, via the n *external* nodes $\circ$, i.e. those nodes outside the subtree whose parent is in the subtree. Such a tree can be mapped to a cladogram on n leaves, as shown in Figure 5. Thus a probability distribution on binary trees induces a probability distribution on cladograms, by randomly (uniformly) labeling the leaves $1, 2, \ldots, n$.

The recent book Mahmoud [20] treats several models of random binary tree from the viewpoint of search trees in computer science. Some of these models – precisely, the induced models on cladograms – occur later.

A wide-ranging abstract study of trees as proximity indicators (but not emphasizing probability models) can be found in Barthelemy and Guenoche [6].

3 Axiomatizing properties of random cladograms

We seek probability models for "neutral evolution of species". By analogy with the results in section 1 for neutral population genetics, it is natural to hope there exists a one-parameter family of probability distributions on cladograms for which

(a) random cladograms $(T_n; n \geq 2)$ are in the family iff they satisfy some specified intrinsic compatibility conditions.

(b) These random cladograms arise from some natural model of species evolving with time.

Let's start with idea (b). Consider the following class of models.

> At each time t there are a finite number of species alive (starting with one species at time 0). ¿From time to time there is an "event" which is either an extinction or a speciation, i.e. either some species B becomes extinct or some species A splits into species A and A'. The time from t until the next event, and the chance the next event is an extinction rather than a speciation, may depend on the past in an arbitrary way. But if the next event is an extinction then each species is equally likely to be the one to become extinct, and if the next event is an speciation then each species is equally likely to be the one to speciate.

At first sight the arbitrariness should allow us to get a family of models, with a parameter representing (say) the ratio of speciation rate to extinction rate. But this is false, because it is easy to show

Lemma 1 *For any model of the class above which ends with n living species, the cladogram of those species is distributed as the Yule model described in section 2.2.*

Turning to idea (a), the following two compatibility conditions for a family $(T_n; n \geq 2)$ of random cladograms seem to be the natural analogs of those in section 1.

(i) Exchangeability. For each n, the random cladogram is exchangeable in the labels of the n species, i.e. invariant under permutations.

(ii) Group elimination. For each $1 \leq k < n$, conditional on $\{k+1, k+2, \ldots, n\}$ being a group in T_n (i.e. being the set of descendants of some internal vertex), the cladogram restricted to $\{1, 2, \ldots, k\}$ is distributed as T_k.

It is easy to check that the Yule model and the uniform model (described in section 2.2) satisfy these conditions, as does the family of *combs*, i.e. the family with the deterministic "maximally unbalanced" shape below, and with the n species uniformly randomly distributed amongst positions.

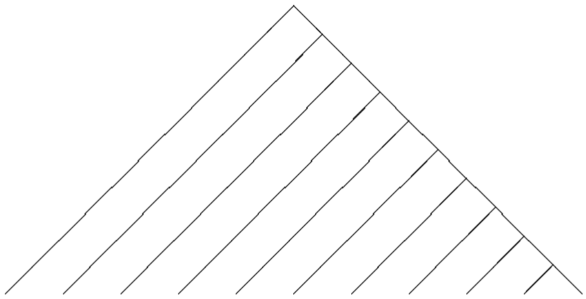


Figure 5

Unfortunately we have been unable to construct any more examples, so

Conjecture 2 *The three families above are the only families satisfying (i)-(ii).*

Another open question in this "axiomatic" spirit will be mentioned in section 6.3, but the bottom line is that our attempt to define models of "neutral evolution of species" by some close analogy with neutral population genetics seems completely unsuccessful. In the next section we resort to pulling a model out of thin air.

4 The beta-splitting model

Suppose that for each $n \geq 2$ we are given a probability distribution $\mathbf{q}_n = (q_n(i); i = 1, 2, \ldots, n-1)$ which is symmetric ($q_n(i) = q_n(n-i)$). Then we can define probability distributions on cladograms in the obvious way: the root-split has i elements in the left branch and $n-i$ elements in the right branch, where i is chosen at random according to the distribution q_n and where each of the $\binom{n}{i}$ choices of elements for the left branch are equally likely. Repeat recursively in each branch. Interpret the resulting tree as a cladogram by removing the left/right markers. Call these *Markov branching* models.

To specialize this construction, consider a probability density f on $(0,1)$ which is symmetric (that is, $f(x) = f(1-x)$), and define

$$q_n(i) = a_n^{-1} \binom{n}{i} \int_0^1 x^i (1-x)^{n-i} \, f(x) dx \tag{1}$$

for normalizing constant

$$\begin{aligned} a_n &= \int_0^1 (1 - x^n - (1-x)^n) f(x) dx \qquad (2) \\ &= 1 - 2 \int_0^1 x^n f(x) dx \; . \end{aligned}$$

This specialization has a simple interpretation in terms of splitting intervals (a topic discussed from a different viewpoint by Brennan and Durrett [7, 8]). Start with n uniform random "particles" on the unit interval. Split the interval at a random point with density f. Repeat recursively on subintervals, splitting each interval $[a, b]$ at a point $a+X(b-a)$ where the X's are independent with density f, stopping when a subinterval contains only one particle.

Figure 6 illustrates the construction and its interpretation as a cladogram. Note that a subinterval split in which all particles go into one side of the split is suppressed.

Note that for (1, 2) to make sense it is not necessary for f to be a probability density. It is enough to have $f \geq 0$ be symmetric and satisfy $\int_{0+} xf(x)dx < \infty$.

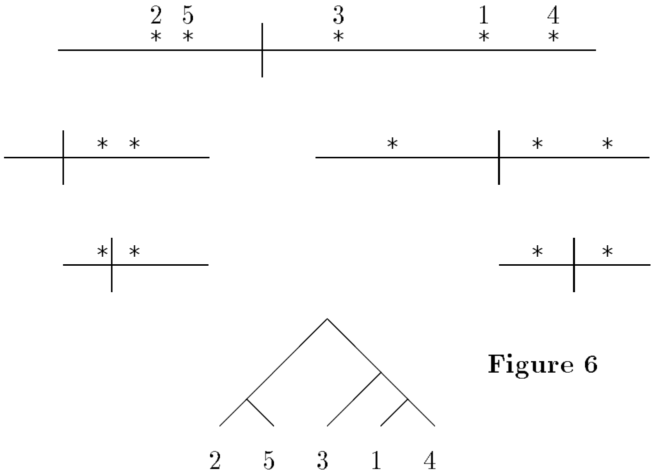


Figure 6

We now specialize further to define a one-parameter family $(T_\beta^{(n)})$ of random cladograms, parametrized by $-2 \leq \beta \leq \infty$. For $-1 < \beta < \infty$ these are obtained by the interval-splitting construction above with the beta density

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^\beta (1-x)^\beta, \ 0 < x < 1. \tag{3}$$

Applying (1),

$$q_n(i) = \frac{1}{a_n(\beta)} \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)}, \ 1 \leq i \leq n-1 \tag{4}$$

where as at (2) $\alpha_n(\beta)$ is the normalizing constant. For $\beta = \infty$ we get the interval-splitting construction where intervals are split deterministically at their center. For $-2 < \beta \leq -1$ the recursive construction with q_n defined at (4) still makes sense, corresponding as mentioned above to the function $f(x) = x^\beta(1-x)^\beta$ for which $\int_{0+} xf(x)dx < \infty$. Finally, for fixed n the $\beta \to -2$ limit is the "comb" described in section 3, so we take the comb as the $\beta = -2$ model.

4.1 Special cases

Three special case have been studied in the literature.

$\boldsymbol{\beta = \infty}$. This is the "symmetric binary trie" studied in computer science, and surveyed in Chapter 5 of [20]. It has been briefly considered in the biological literature [19].

$\boldsymbol{\beta = 0}$. Here we have

$$q_n(i) = \frac{1}{n-1}, \; 1 \leq i \leq n-1.$$

This is the Yule model of section 2.2. It arises from the "binary search tree" in computer science, surveyed in Chapter 2 of [20]. It also arises from the "coalescent" in mathematical population genetics ([16, 26]). And as mentioned in section 2.2 it arises from the Yule process, i.e. the linear pure birth process. All these processes are different as processes indexed by time, but the induced random cladograms are identical.

$\boldsymbol{\beta = -3/2}$. This is the uniform model from section 2.2. To verify, a counting argument shows that the uniform model corresponds to a recursive construction with

$$q_n(i) = \frac{1}{2} \binom{n}{i} \frac{c_i c_{n-i}}{c_n}. \tag{5}$$

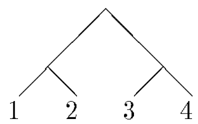
where $c_n = (2n-3)!!$ is the number of cladograms on n leaves. But we can write $c_n = 2^{n-1}\Gamma(n-\frac{1}{2})/\Gamma(\frac{1}{2})$, which leads to

$$q_n(i) = \frac{\Gamma(n+1)}{\Gamma(n-\frac{1}{2})\Gamma(\frac{1}{2})} \frac{\Gamma(i-\frac{1}{2})\Gamma(n-i-\frac{1}{2})}{\Gamma(i+1)\Gamma(n-i+1)}$$

which is indeed (4) for $\beta = -3/2$.

This is the model of cladograms associated (in the sense of section 2.3) with the uniform random binary tree with n external leaves, which has been studied extensively in combinatorics and computer science (e.g. Knuth [17]). Its asymptotics can be studied via the continuum tree set-up of [2, 3].

The beta-splitting model certainly satisfies the exchangeability condition (i) of section 3, but in general does not satisfy the group elimination property (ii). As a specific example, for $\beta = \infty$ the chance of the cladogram



equals 1/7. But if we take the cladogram on 6 leaves and condition on $\{5, 6\}$ being a group, then (by an elementary but tedious calculation) the probability that the cladogram restricted to $\{1, 2, 3, 4\}$ is the cladogram above equals $\frac{5 \cdot 13}{3 \cdot 157}$.

The special case $\beta = -1$ has apparently not been studied before, but turns out to have interesting properties. Note that here we have $q_n(i) = a_n \frac{1}{i(n-i)}$. Because

$$\sum_{i-1}^{n-1} \frac{1}{i(n-i)} = \sum_{i=1}^{n-1} \frac{1}{n} \left(\frac{1}{i} + \frac{1}{n-i} \right) = \frac{2h_{n-1}}{n}$$

where h_{n-1} is the harmonic sum, we can write

$$q_n(i) = \frac{n}{2h_{n-1}} \, \frac{1}{i(n-i)}, \ 1 \leq i \leq n-1. \tag{6}$$

4.2 Is there an underlying process?

We are introducing the beta-splitting models as the mathematically most natural way to embed the Yule model and the uniform model into a one-parameter family. To make this convincing for biological applications one would like an underlying continuous-time process of speciation and extinction for which the general beta-splitting model was the associated cladogram. I do not know a natural candidate for such a process.

It has often been asserted in the biological literature [14, 19] that the uniform model is unsatisfactory because there is no such underlying process. This is not entirely correct. Consider conditioned critical branching processes, which have been studied in the biological literature ([24, 12]) as models for neutral evolution of large groups of species. One of the results from [2, 3] is that, for a random sample of n species from a large such group, the cladogram on the sampled species will follow the uniform model.

This observation makes it slightly more plausible that there might be some subtle process underlying the general beta-splitting model.

4.3 A data-set

The most famous datum in the subject (discussed in [14] and by many other writers) is $(9672, 21)$, the split between bird species and crocodilian species. Under the Yule model ($\beta = 0$) the chance of a more unbalanced split is 0.004, whereas under the uniform model ($\beta = -1.5$) the chance is 0.878. Guyer and Slowinski ([14] table 1) consider the sizes of the smaller branch in the root split in 30 large cladograms. It is clear from the raw data (and confirmed in [14] by a test of significance) that the splits are more unbalanced than predicted under the Yule model, but more balanced than predicted under the uniform model. Figure 7 gives a visual display of the data under each model using the quantile transform, (i.e. the bird-crocodile split would be plotted at 0.878 under the uniform model and at 0.004 under the Yule model.) So under a true model we would see 30 independent uniform points.

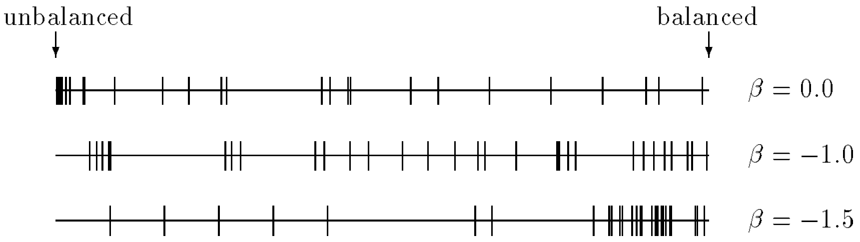


Figure 7

The fit to the $\beta = -1$ model is strikingly better than to the usual models. Is this just a fluke?

Let me also mention that two papers (Savage [25], Guyer and Slowinski [13]) analyze data on small cladograms ($4 \leq n \leq 7$ and $n = 5$). [25] concludes that the Yule model gives a better fit than the uniform model, whereas [13] concludes the opposite.

5 Some asymptotics for the beta-splitting model

This section outlines some asymptotics for the beta-splitting family. Details of some of the more interesting results may be given elsewhere. In the three special cases ($\beta = -3/2,\ 0,\ \infty$) these results (and much more) are either explicitly known or can be proved by routine methods.

5.1 Asymptotics of the root-split distribution

The top-most split in a cladogram splits it into two branches of sizes i and $n-i$. If we randomly call these "left" and "right" then the size L_n of the left branch has the distribution q_n at (4). And if we write B_n for the size of the branch containing a specified leaf, then by exchangeability

$$P(B_n = i) = \frac{2i}{n}\, q_n(i);\ 1 \leq i \leq n-1. \tag{7}$$

It is straightforward to obtain the following asymptotics as $n \to \infty$.

Lemma 3 $\boldsymbol{\beta = \infty}$. *$n^{-1}L_n \xrightarrow{d} \frac{1}{2}$; $n^{-1}B_n \xrightarrow{d} \frac{1}{2}$.*

$\boldsymbol{-1 < \beta < \infty}$. *$n^{-1}L_n \xrightarrow{d} X_\beta$ and $n^{-1}B_n \xrightarrow{d} Y_\beta$, where X_β has the beta distribution (3) and where Y_β has density*

$$f(x) = \frac{\Gamma(2\beta+3)}{\Gamma(\beta+1)\Gamma(\beta+2)}\, x^{\beta+1}(1-x)^{\beta}. \tag{8}$$

$\beta = -1$.

$$\frac{\log \min(L_n, n - L_n)}{\log n} \xrightarrow{d} U \text{ and } \frac{\log(n - B_n)}{\log n} \xrightarrow{d} U$$

where U has the uniform distribution on $(0,1)$.

$-2 < \beta < -1$. *For each fixed $i \geq 1$,*

$$P(\min(L_n, n - L_n) = i) \to \gamma_\beta(i) \text{ and } P(B_n = n - i) \to \gamma_\beta(i)$$

where the limit distribution γ_β is given by

$$\gamma_\beta(i) = \kappa_\beta^{-1} \int_0^\infty \frac{e^{-t} t^i}{i!} t^\beta \, dt = \kappa_\beta^{-1} \frac{\Gamma(i + 1 + \beta)}{\Gamma(i + 1)}$$

for normalizing constant

$$\kappa_\beta = \int_0^\infty (1 - e^{-t}) t^\beta \, dt = \frac{\pi}{\Gamma(b) \sin(\pi(-\beta - 1))}. \tag{9}$$

And if $i \sim xn$ for $0 < x < 1$,

$$q_n(i) \sim \frac{1}{2} \kappa_\beta^{-1} x^\beta (1 - x)^\beta n^\beta. \tag{10}$$

Note that the distribution γ_β can also be expressed as

$$\gamma_\beta(i) = \kappa'_\beta \prod_{j=2}^{i} \left(1 + \frac{\beta}{j}\right), \ i \geq 1$$

for different normalizing constants.

5.2 Depth and height statistics

The *depth* of a leaf in a cladogram is the number of branchpoints on the path from that leaf to the root, where we include the root as a branchpoint. Thus in Figure 1, species A has depth 2 and species E has depth 3. We can define the following three numbers for a cladogram.

- $\bar{d}$, the average depth of the leaves.
- $\tilde{d}$, the depth of the leaf found by starting at the root and recursively choosing the larger branch at each branchpoint (averaging over possibilities if an even split is encountered).
- d^*, the maximal depth of a leaf.

So $\bar{d} \leq \tilde{d} \leq d^*$. In many settings, statistics like these are called *heights*, and in particular one could call d^* the height of the cladogram.

Proposition 4 *Define $\bar{D}_n, \tilde{D}_n, D^*_n$ to be the random values of the statistics above for the beta-splitting model. Note that $E\bar{D}_n = ED_n$, where D_n is the depth of species 1. As $n \to \infty$,*
$\boldsymbol{-1 < \beta \leq \infty}$.

$$(E\bar{D}_n, E\tilde{D}_n, ED^*_n) \sim (\bar{\rho}(\beta), \tilde{\rho}(\beta), \rho^*(\beta)) \ \log n$$

and $D'_n/ED'_n \ \stackrel{d}{\to}\ 1$ for each of the four statistics, where (for X_β, Y_β as in Lemma 3 and $\beta < \infty$)

$$\frac{1}{\bar{\rho}(\beta)} = -E\log Y_\beta = \frac{\Gamma(2\beta+3)}{\Gamma(\beta+1)\Gamma(\beta+2)} \int_0^1 x^{\beta+1}(1-x)^\beta \log(1/x)dx.$$

$$\frac{1}{\tilde{\rho}(\beta)} = -E\log\max(X_\beta, 1-X_\beta) = \frac{2\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} \int_{1/2}^1 x^\beta(1-x)^\beta \log(1/x)dx.$$

$$\rho^*(\beta) = \frac{-\theta}{\log 2EX^\theta_\beta}, \text{ where } \theta = \theta(\beta) \text{ is the solution of}$$

$$EX^\theta_\beta \log X_\beta = \theta^{-1}(EX^\theta_\beta)\log(2EX^\theta_\beta).$$

$\boldsymbol{\beta = -1}$.

$$(E\bar{D}_n, E\tilde{D}_n) \sim (3\pi^{-2}, 6\pi^{-2}) \ \log^2 n.$$

$\boldsymbol{-2 < \beta < -1}$.

$$(E\bar{D}_n, E\tilde{D}_n, ED^*_n) \sim (\bar{\rho}(\beta), \tilde{\rho}(\beta), \rho^*(\beta)) \ n^{-\beta-1}$$

where

$$\frac{1}{\bar{\rho}(\beta)} = \kappa_\beta^{-1} \int_0^1 x^{\beta+1}(1-x)^\beta(1-x^{-\beta-1}) \ dx$$

$$\frac{1}{\tilde{\rho}(\beta)} = \kappa_\beta^{-1} \int_{1/2}^1 x^{\beta}(1-x)^\beta(1-x^{-\beta-1}) \ dx$$

for κ defined at (9), and where we do not have a simple expression for $\rho^*(\beta)$. For each of the four statistics we have a non-degenerate limit distribution for D'_n/ED'_n.

Some numerical values are tabulated below. The unexplained decimal numbers were obtained by numerical evaluation of the formulae above, and the others (except for $\rho^*(-1.5)$, discussed later) by exact evaluation.

β	-0.5	0	1	∞
$\bar{\rho}(\beta)$	$\frac{1}{2\log 2-1} = 2.59$	2	$\frac{12}{7} = 1.71$	$\frac{1}{\log 2} = 1.44$
$\tilde{\rho}(\beta)$	4.55	$\frac{1}{1-\log 2} = 3.26$	$\frac{1}{\frac{13}{12}-\log 2} = 2.56$	$\frac{1}{\log 2} = 1.44$
$\rho^*(\beta)$	6.38	4.31	3.19	$\frac{1}{\log 2} = 1.44$

β	-2	-1.75	-1.5	-1.25
$\bar{\rho}(\beta)$	$1/2$	0.952	$\pi^{1/2} = 1.77$	8.13
$\tilde{\rho}(\beta)$	1	1.71	$\frac{2\pi^{1/2}}{2^{3/2}-\log(3+2^{3/2})} = 3.33$	15.5
$\rho^*(\beta)$	1	?	$2\pi^{1/2} = 3.54$	?

Outline of proof.

$\boldsymbol{-1 < \beta \leq \infty}$. Consider the process of splitting the unit interval into subintervals. For these first-order results the issue is when the "relevant subinterval" has length $\approx 1/n$. For D_n the relevant subinterval is the one containing a prespecified point, and the lengths of such subintervals behave as an i.i.d. product of Y_β's. Similarly for $\tilde{D}_n$, the lengths of the relevant subintervals behave as an i.i.d. product of $\max(X_\beta, 1-X_\beta)$'s. For D^*_n we are concerned with the maximal-length subinterval after m splits, but here log (length) behaves as (a slight variation of) branching random walk, and the result comes from the usual large deviation analysis of the rightmost walker in branching random walk. (This is an elaboration of the known result ([20] section 2.7) that $1/\rho^*(0)$ is the solution of $2x = \exp(x-1)$).

$\boldsymbol{\beta = -1}$. The recurrence for $t_n = ED_n$ is (for any β)

$$t_n = 1 + \sum_{i=1}^{n-1} \frac{2i}{n} q_n(i) t_i. \tag{11}$$

In the case $\beta = -1$ this becomes

$$0 = 1 + \frac{1}{h_{n-1}} \sum_{i=1}^{n-1} \frac{t_i - t_n}{n-i}.$$

Writing $t_n \sim c \log^2 n$ and approximating the sum by an integral,

$$0 = 1 + 2c \int_0^1 \frac{\log y}{1-y}\, dy,$$

and the integral equals $\pi^2/6$.

The analysis of $E\tilde{D}_n$ is similar, leading to an equation

$$0 = 1 + 2c \int_{1/2}^1 \frac{\log y}{y(1-y)}\, dy,$$

and the integral here equals $\pi^2/12$.

To explain intuitively the order of magnitude in this case, note that D_n is distributed as the number of steps of the following Markov chain, started at n and run until absorption at 1.

$$P(j,i) = \frac{1}{h_{j-1}}\ \frac{1}{j-i};\ 1 \leq i \leq j-1.$$

The mean position after one step from j is about $j(1-\frac{1}{\log j})$, so it takes $O(\log n)$ steps to go from n to around $n/2$, and so it requires $O(\log^2 n)$ steps until absorption.

$\mathbf{-2 < \beta < -1}$. In the recurrence (11), writing $t_n \sim cn^{-\beta-1}$ and approximating $q_n(i)$ by the asymptotic values (10),

$$0 \approx 1 + \kappa_\beta^{-1} cn^\beta \sum_{i=1}^{n-1} \left(\frac{i}{n}\right)^{\beta+1} \left(\frac{n-i}{n}\right)^\beta (i^{-\beta-1} - n^{-\beta-1}).$$

Approximating the sum by an integral gives the result for $\bar{\rho}(\beta)$, and the argument for $\tilde{\rho}(\beta)$ is similar. Finally, finiteness of ρ^* and the existence of non-degenerate limit distributions are consequences of the process approximation indicated in section 6.1.

Remarks on Proposition 4 for $\beta = -3/2$. The $\beta = -3/2$ case fits into a long-studied line of work in probabilistic combinatorics, the "simply-generated trees" of Meir and Moon [21]. Many asymptotic results for these models have been proved by generating function methods, in particular the values of $\bar{\rho}(-3/2)$ and $\rho^*(-3/2)$. See [2] for a brief discussion and the interpretation of the limits in terms of Brownian excursion. Curiously, $\tilde{\rho}(-3/2)$ has only recently been investigated in that literature – see the preprints by Vatutin [27] and Luczak [18].

6 Directions for further study

6.1 More on the beta-splitting model

Two features of the results outlined in section 5 seem sufficiently interesting (from the mathematical, rather than biological, point of view) to warrant more careful study. In all models of random n-vertex trees known to me which deal with "combinatorial" trees (rather than trees with vertices in d-dimensional space), the height statistics (c.f. Proposition 4) grow as either $\Theta(\log n)$ or as $\Theta(n^{1/2})$. So our beta-splitting family exhibits two types of novel behavior. For $\beta = -1$ the mean depth grows as $\Theta(\log^2 n)$, and this immediately raises a host of questions whose answers cannot be immediately guessed by analogy: is the mean height also $\Theta(\log^2 n)$? what are the spreads of the various statistics $D_n, \bar{D}_n, \tilde{D}_n, D_n^*$? are there other models for which these statistics are $\Theta(\log^2 n)$ with different constants? I do not see any elegant probabilistic way of studying this case, but obviously one can try analytic techniques. Secondly, in the case $-2 < \beta < -1$ the mean depth grows as $\Theta(n^{-\beta-1})$. Randomly call edge-splits "left" and "right", thereby inducing a left-to-right ordering on the leaves. Define $H_n(t), 0 \le t \le 1$, by

$$H_n(i/n) = n^{\beta+1}(\text{ depth of leaf } i \text{ in the left-to-right ordering})$$

with linear interpolation. Then one can see heuristically that as $n \to \infty$ there is some limiting stochastic process $H_\infty(t), 0 \leq t \leq 1$ and that the quantities D_n considered in Proposition 4 have non-degenerate limit distributions expressible in terms of H_∞. For $\beta = -3/2$ the limit process is Brownian excursion – this is part of the circle of ideas discussed in [2, 3]. But for general $-2 < \beta < -1$ the limit processes H_∞ seem novel and interesting stochastic processes. Informally, they can be constructed as processes in which intervals split and shrink continuously, as opposed to the discrete-step splitting of section 4 for $\beta > -1$. Details will be given in [4].

6.2 Other one-parameter families

We introduced the beta-splitting models as a mathematically natural way to embed the Yule model and the uniform model into a one-parameter family. One can of course invent other such families. For instance, under the Yule model the chance of a particular cladogram t equals

$$\frac{2^{n-1}}{n!} \prod_{i=3}^{n} (i-1)^{-d_i(t)}$$

where $d_i(t)$ is the number of internal nodes with exactly i descendant species. Thus we can define a family $(T_\gamma^{(n)})$ for which

$$P(T_\gamma^{(n)} = t) = a_n(\gamma) \prod_{i=3}^{n} (i-1)^{\gamma d_i(t)}.$$

So $\gamma = 0$ is the uniform model and $\gamma = -1$ is the Yule model. This is a different family, because (for instance) the $\gamma \to -\infty$ limit is the uniform distribution on maximally-balanced n-cladograms.

The only one-parameter family in the literature which exhibits qualitative change in behavior as the parameter varies are the randomly-growing binary trees discussed in [1] and [5]. These contain the Yule model but not the uniform model.

6.3 Another characterization question

Consider the "Markov branching" models $(P^{(n)}; n \geq 1)$ defined at the start of section 4. These are automatically exchangeable, but do not necessarily have the property (c.f. property (ii) in section 1)

Sampling consistency. For each n, $P^{(n)}$ induces a distribution on cladograms on $\{1, 2, \ldots, n-1\}$ by the action of deleting n: this distribution is $P^{(n-1)}$.

So it is natural to ask

Open Problem 1 *Characterize the subclass of Markov branching models which satisfy the sampling consistency property.*

This subclass contains the beta-splitting models, but I suspect it is much larger.

Acknowledgements. I thank John Kingman, Warren Ewens and an anonymous referee for helpful suggestions. The numerical calculations used MATHEMATICA, and I thank Mike Steele for assistance therewith.

References

[1] D. Aldous and P. Shields. A diffusion limit for a class of randomly-growing binary trees. *Probab. Th. Rel. Fields*, 79:509–542, 1988.

[2] D.J. Aldous. The continuum random tree II: an overview. In M.T. Barlow and N.H. Bingham, editors, *Stochastic Analysis*, pages 23–70. Cambridge University Press, 1991.

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

[4] D.J. Aldous. A family of self-similar continuous interval-splitting processes. In Preparation, 1995.

[5] M.T. Barlow, R. Pemantle, and E. Perkins. Diffusion-limited aggregation on a tree. Technical report, U. British Columbia, 1994.

[6] J.-P. Barthelemy and A. Guenoche. *Trees and Proximity Representations*. Wiley, 1991.

[7] M.D. Brennan and R. Durrett. Splitting intervals. *Ann. Probab.*, 14:1024–1036, 1986.

[8] M.D. Brennan and R. Durrett. Splitting intervals II: Limit laws for lengths. *Probab. Th. Rel. Fields*, 75:109–127, 1987.

[9] P. Donnelly and P. Joyce. Weak convergence of population genealogical processes to the colescent with ages. *Ann. Probab.*, 20:322–341, 1992.

[10] N. Eldredge and J. Cracraft. *Phylogenic Patterns and the Evolutionary Process*. Columbia University Press, New York, 1980.

[11] W. J. Ewens. Population genetics theory: The past and the future. In S. Lessard, editor, *Mathematical and Statistical Developments of Evolutionary Theory*, pages 177–227, 1990.

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

[13] C. Guyer and J.B. Slowinski. Comparisons between observed phylogenetic topologies with null expectations among three monophyletic lineages. *Evolution*, 45:340–350, 1991.

[14] C. Guyer and J.B. Slowinski. Adaptive radiation and the topology of large phylogenies. *Evolution*, 47:253–263, 1993.

[15] S.A. Kauffman. *The Origins of Order: Self Organization and Selection in Evolution*. Oxford University Press, 1993.

[16] J.F.C. Kingman. *Mathematics of Genetic Diversity*. S.I.A.M., Philadelphia PA, 1980.

[17] D.E. Knuth. *The Art of Computer Programming*, volume 1. Addison-Wesley, 1968.

[18] T. Luczak. A greedy algorithm for estimating the height of random trees. Technical Report 1190, I.M.A., Minneapolis MN, 1993.

[19] W.P. Maddison and M. Slatkin. Null models for the number of evolutionary steps in a character on a phylogenic tree. *Evolution*, 45:1184–1197, 1991.

[20] H. M. Mahmoud. *Evolution of Random Search Trees*. Wiley, 1992.

[21] A. Meir and J.W. Moon. On the altitude of nodes in random trees. *Canad. J. Math.*, 30:997–1015, 1978.

[22] M.H. Nitecki and A. Hoffman, editors. *Neutral Models in Biology*. Oxford University Press, 1987.

[23] D.M. Raup. Mathematical models of cladogenesis. *Paleobiology*, 11:42–52, 1985.

[24] 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.

[25] H.M. Savage. The shape of evolution: Systematic tree topology. *Biological J. Linnean Soc.*, 20:225–244, 1983.

[26] S. Tavare. Line-of-descent and genealogical processes and their applications in population genetics models. *Theoret. Population Biol.*, 26:119–164, 1984.

[27] V.A. Vatutin. On the height of the primary path of random rooted trees. Technical report, Mathematics, Chalmers Univ., 1993.