Tree Models for Macroevolution and Phylogenetic Analysis

GRAHAM R. JONES*

21e Balnakeil, Durness, Lairg IV27 4PT, UK;
*Correspondence to be sent to: 21e Balnakeil, Durness, Lairg IV27 4PT, UK;
E-mail: arr@gjones.name.

Received 23 July 2009; reviews returned 14 October 2009; accepted 9 August 2011

Associate Editor: Mark Holder

Abstract.—It has long been recognized that phylogenetic trees are more unbalanced than those generated by a Yule process. Recently, the degree of this imbalance has been quantified using the large set of phylogenetic trees available in the TreeBASE data set. In this article, a more precise analysis of imbalance is undertaken. Trees simulated under a range of models are compared with trees from TreeBASE and two smaller data sets. Several simple models can match the amount of imbalance measured in real data. Most of them also match the variance of imbalance among empirical trees to a remarkable degree. Statistics are developed to measure balance and to distinguish between trees with the same overall imbalance. The match between models and data for these statistics is investigated. In particular, the use of the usual proportional to distinguishable sampling consistency gives an interpretation. When a speciation occurs, the two sub-trees will each develop a number of tips

It is well known that phylogenetic trees are more unbalanced than those generated by a Yule process (Mooers and Heard 1997; Aldous 2001). Two recent papers (Ford 2006; Blum and François 2007) have quantified the degree of this imbalance for the large set of trees available in the TreeBASE data (Sanderson et al. 1994). This paper presents a more detailed analysis of the same data set and two others, compiled by McPeek and Brown (2007) and Heath et al. (2008), and considers more models.

Balance is important for Bayesian phylogenetic analysis, as is shown by the following example. Consider a rooted tree with \( n + 2 \) tips, in which there is a clade \( C \) of size \( n \) and two tips \( x \) and \( y \) not in \( C \). Two possible topologies are \( (C, (x, y)) \) and \( ((C, x), y) \), as shown in Figure 1. Two common priors for Bayesian phylogenetic analysis are the constant rate birth–death model (Edwards 1970; Yang and Rannala 1996, 1997), referred to here as the CRBD model and the proportional to distinguishable arrangements (PDA) models. The ratio of the probabilities of these two trees under the CRBD model is \( n \), whereas under the PDA model it is 1. (This is proved later.) This shows that a small change in topology near the root of a large tree can make a big difference to the prior probability. In practical terms, using BEAST 1.4 (Drummond and Rambaut 2007) with its usual CRBD prior would give an \( n : 1 \) advantage to \( (C, (x, y)) \) over \( ((C, x), y) \) compared with using MrBayes 3.1 (Huelsenbeck and Ronquist 2001) with the usual PDA prior.

I show that several models can fit both the overall degree of imbalance in the TreeBASE data and the variance of the imbalance in the data. It is not surprising that they can match the overall degree of imbalance because they each have a parameter that can be adjusted to achieve this, but it is remarkable that the variance in the models is so similar to that in the data. I introduce two statistics that distinguish between trees with very similar balance and demonstrate that trees can vary independently in all three "directions." I investigate the differences between the data sets and how various models fit the data as measured by these statistics. The models considered include those of Ford (2006) and Blum and François (2007) and several based on age-dependent (Bellman–Harris) branching processes. The latter are processes in time in which each species behaves independently of its parent and all other species, but where the rates of extinction and speciation depend on how old the species is.

All trees considered here are binary and rooted. A model of tree formation provides a probability distribution over tree topologies for a given number of tips. This distribution may be defined by describing the process of tree generation, or explicitly as a formula for calculating the probability of a given topology (or both). Because evolution is a process in time, models that can be described as a process of diversification over time seem preferable to models that lack any process-based interpretation. When a speciation occurs, the two sub-trees will each develop a number of tips \( x \) and \( y \), say, with a total of \( n = x + y \). Many of the trees I consider here can be characterized by a set of probability distributions \( q(x, y) \) on the \( n - 1 \) pairs of integers \( \{(1, n - 1), (2, n - 2), \ldots, (n - 1, 1)\} \), one distribution for each positive integer \( n \). Such a distribution is known as a splitting distribution (Ford 2006). I will also use (or abuse) the term to refer to a set of distributions for all \( n \). Because speciations are supposed to happen independently in all models considered here, a splitting distribution implies a distribution over tree topologies. The probability of a tree is obtained by multiplying the probabilities of the splits at each internal node.

There are more subtle characteristics of tree shape that can distinguish between trees having the same overall amount of balance. Aldous (1996) defined a property that a model for tree formation may possess, which he called sampling consistency. A model has this property if the deletion of a randomly chosen tip from a tree of
size \( n \) leaves a tree of size \( n - 1 \) belonging to the same model. It follows that one can remove \( k \) tips at random and end up with a tree from the same model too. Ford (2006) calls this property deletion stability. The latter term is used in ecology to mean something different, so I will use “sampling consistency.” The process of removing some tips at random is called pruning. Sampling consistency is a convenient property for phylogenetic analysis because if true, the probability of the topology is not affected by whether species are densely or sparsely sampled from extant species (as long as the sampling is random). The models of Ford (2006) and Aldous (1996) are both sampling consistent. However, it seems difficult to find models that are processes in time, produce the observed amount of imbalance, and are sampling consistent. (See Proposition 2 for a negative result.)

Process-based models often yield a probability distribution for divergence times in addition to tree topologies. However, information on divergence times will not be compared with real data here, because doing so would require good estimates of them for the real data sets that are examined. The distribution of divergence times is also very sensitive to the extinction rate in the model.

The Tree Models

The CRBD, PDA, and Aldous Models

In all the models considered here, the event of extinction or speciation will be called a “termination,” and the life of a species is regarded as ending at termination: its descendants, if any, are two new species. In the CRBD model, all species in the tree behave independently of one another and all are assumed to have the same constant speciation rate, and the same constant extinction rate. A special case is the Yule model in which the extinction rate is zero. The topology of the CRBD model can be characterized as a tree in which \( q(x, y) = 1/(n - 1) \) for all \( x, y \) satisfying \( x + y = n \), where \( n \) is the tree size. The PDA model is one where every distinct topology is given an equal probability. It is sometimes known as a “uniform” model. It is used as a prior in much Bayesian phylogenetic analysis. It can be given a biological interpretation (Steel and McKenzie 2001; Kontoleon 2006), but it is not as simple as the CRBD model when viewed as a process in time. Real phylogenetic trees are generally more balanced than PDA but less balanced than CRBD. The PDA model can be characterized by a splitting distribution as a special case of the beta-splitting model described next.

Aldous (1996) described a family of models parameterized by a single parameter \( \beta \) taking values in \([-2, \infty)\), which includes the CRBD and PDA models. The balance increases with \( \beta \), with the PDA model appearing at \( \beta = -1.5 \) and the CRBD model at \( \beta = 0.0 \). As \( \beta \to \infty \), even more balanced trees are produced. The beta-splitting models are defined via a splitting distribution

\[
q(x, y) = s_n(\beta)^{-1} \frac{\Gamma(x + 1 + \beta) \Gamma(y + 1 + \beta)}{\Gamma(x + 1) \Gamma(y + 1)},
\]

where \( s_n(\beta) \) is a normalization constant and \( \Gamma() \) is the gamma function. They have recently been given an interpretation in terms of diversification rates (McCullagh et al. 2008), but this does not seem to be one that is biologically realistic because the diversification rates depend on the size of the tree to be generated. Of particular interest is the value \( \beta = -1 \) because it has some special mathematical properties and appears to be a good fit to the available data. I will call this the Aldous Branching or AB model. The following result is useful in calculations involving the beta-splitting model.

**Proposition 1** Suppose that a tree is formed according to the beta-splitting model of Aldous. Let \( x_n = q(1, n - 1) \) and let

\[
\Delta(i, n - i, \beta) = \frac{\Gamma(i + 1 + \beta) \Gamma(n - i + 1 + \beta)}{\Gamma(i + 1) \Gamma(n - i + 1)}
\]

and

\[
s_n = s_n(\beta) = \sum_{i=1}^{n-1} \Delta(i, n - i, \beta)
\]

be the normalization constant. Then for \( n \geq 3 \), these recursive formulas hold:

\[
s_{n+1} = \frac{1}{n + 1} \left( n + 2 + 2\beta + \frac{2(n + \beta)}{n} x_n \right) s_n
\]

\[
x_{n+1} = \frac{(n + \beta) x_n s_n}{s_{n+1}} = \frac{(n + \beta)(n + 1) x_n}{n(n + 2 + 2\beta + 2(n + \beta) x_n)}.
\]

A proof is in the Appendix. As an immediate application of this Proposition, it is easy to show by induction that if \( \beta = -1.5 \), then

\[
x_n = n/(4n - 6),
\]

so that the probability of an extreme split \((n - 1, 1)\) or \((1, n - 1)\) is \(n/(2n - 3)\) for \( n > 2 \) tips in the PDA model. As another example, here is a calculation for the situation described in the introduction. Assume the beta-splitting model and let \( R_{n+2}(\beta) \) be the ratio of the probabilities of \((C, (x, y))\) and \((C, x, y)\) appearing in Figure 1. Then, \( R_{3}(\beta) = 1 \) and for \( n \geq 2 \)

\[
R_{n+2}(\beta) = \frac{\Pr((C, (x, y)))}{\Pr((C, x, y))} = \frac{q(n, 2) q(1, 1)}{q(n + 1, 1) q(n, 1)}.
\]

**Figure 1**. Two similar trees, which are rooted by the outgroups.
Note that the probability of the topology inside $C$ cancels out in the above. Now

$$R_{n+2}(\beta) = \frac{\Delta(n, 2)}{\Delta(n + 1, 1)x_{n+1}}$$

because $q(1, 1) = 1$, $q(n, 1) = x_{n+1}$, $q(n, 2) = \Delta(n, 2)/s_{n+2}$, and $q(n + 1, 1) = \Delta(n + 1, 1)/s_{n+2}$. It then follows that

$$R_{n+2}(\beta) = \frac{(2 + \beta)n}{2(n + \beta)x_{n+1}}. \quad (4)$$

Using this together with Proposition 1 allows $R_{n+2}(\beta)$ to be calculated recursively for any $\beta$. For the special cases CRBD, AB, PDA, it follows that

$$R_{n+2}(0) = n, \quad (5)$$

$$R_{n+2}(-1) = \sum_{i=1}^{n} (1/i) \approx 0.577 + \log(n), \quad (6)$$

$$R_{n+2}(-1.5) = 1.\quad (7)$$

**The Alpha Model of Ford**

Ford (2006) described another family of models parameterized by a single parameter $\alpha$ taking values in $[0, 1]$, which interpolates between the CRBD and PDA models. It is not a process in time. Instead, trees can be “grown” by starting with a single branch and then randomly choosing at each step whether to add a new branch at a tip or in the middle of an internal branch, the parameter $\alpha$ controlling the proportion of each type. They can also be characterized by a splitting distribution, which is different from that of the Aldous model. In this model, balance decreases with $\alpha$, with $\alpha = 0$ giving the CRBD model, and $\alpha = 1/2$ the PDA model. The Ford model cannot produce trees more balanced than the CRBD model.

**Blum–François Models**

Some models have an interpretation as a time process using diversification rates, such that when a speciation occurs, the sum of the diversification rates of the two descendant species is equal to that of the parent species. The parent thus “shares” its diversification rate between its descendants, and the total diversification rate, summed over all species, stays constant. This means that the behavior of the species in the tree becomes slower and slower as they become more numerous, and some nonlinear function is required to map “tree time” to real time. The simplest such model was described by Kirkpatrick and Slatkin (1993). Let $p$ be a fixed value in the interval $(0, 1)$. If a species with diversification rate $\kappa$ speciates, then the descendants have diversification rates $pk$ and $(1 - p)\kappa$. As one would expect, this produces unbalanced trees if $p$ is close to 0 or 1.

The splitting distribution is a mixture of two “mirrored” binomials

$$q(x, y) = \frac{1}{2} \left( \frac{x + y - 2}{x - 1} \right) \left( p^{x-1}(1-p)^{y-1} + p^{y-1}(1-p)^{x-1} \right). \quad (8)$$

Blum and François (2007) extended this model by allowing $p$ to be random variable chosen from a symmetric distribution on $[0, 1]$. They used a symmetric beta distribution, resulting in a “beta-binomial” model, but other distributions could be used. I will call all such models “Blum and François (BF)” models.

A BF model can only be sampling consistent under a restrictive condition.

**Proposition 2** Suppose that a tree is formed as a process in time in which speciation of a species with diversification rate $\kappa$ produces two descendant species with diversification rates $p\kappa$ and $(1 - p)\kappa$, where $p$ is chosen from a symmetric distribution on $[0, 1]$. Assume that the process is sampling consistent. Then for odd $n$, the splitting distribution satisfies $q(1, n - 1) = 1/(n - 1)$.

A proof is in the Appendix. It may be noted that the CRBD model satisfies this condition on $q(1, n - 1)$. This condition rules out BF models as plausible models and implies that sampling consistency should not be enforced upon BF models.

**Age-Dependent Models**

Age-dependent processes (Athreya and Ney 1972) are well-studied extensions of the CRBD model. They were first studied by Bellman and Harris (1952) and are sometimes known as Bellman–Harris models. They have been previously considered as models for evolutionary trees (Losos and Adler 1995; Chan and Moore 1999; Agapow and Purvis 2002; Gernhard et al. 2008; Hartmann 2008). In such a process, each species behaves independently of every other species. The rate at which termination of a species occurs is dependent on its age—the time since its birth.

In an age-dependent model, a species lives for a time, then either goes extinct or produces exactly two descendant species. The ratio of the extinctions to speciations is given by a fixed number $p$, which will always be less than 1 here, and is set to either 0.1 or 0.9 in the simulations. The lifetime of a species $i$ is given by a function $S$ such that $Pr(i$ lives for at least time $t) = S(t)$. By necessity, $S$ is a decreasing function of $t$ because a species cannot live until $t$ unless it already lived to $s$ if $s < t$. I will assume that $S(0) = 1$, that is, that there is zero probability that the species has zero lifetime. I will also assume that $S$ tends to zero as $t$ tends to infinity, that is, that no species lives for an infinite time.

The function $f = -S'$, where $S' = dS/dt$ gives the density of lifetimes on $[0, \infty)$. The function $h = f/S = -S'/S$ gives the conditional rate of termination for a species that has lived until time $t$. In the language of survival analysis, $S$ is the survival function, $f$ is the event density function, and $h$ is the hazard function (often denoted by $\lambda$). The method
used for sampling from age-dependent models is in the Appendix. Experiments with many possibilities for $f$ showed that an “L”-shaped event density function appears to be necessary to produce unbalanced trees. If the event density is exponential, say $f(t) = e^{-t}$, then $S = e^{-t}$ as well, and one has the standard birth–death model with constant rates of extinction and speciation, and thus balanced trees. It is also known that the very unbalanced PDA model can be produced in some cases when the survival function $S(t)$ does not tend to zero as $t$ tends to infinity (Pinelis 2003; Kontoleon 2006), so that there is a nonzero probability that a species persists indefinitely. These can be described as even more “L” shaped than the ones used here.

There is some biological justification for these models with “L”-shaped event density functions. It has been suggested that when a species is young, there is likely to be an increased chance of further speciations (Steel and Mckenzie 2001). This would apply to the descendants of a species that is able to occupy a new niche due to a key innovation or to the descendants of a species that invades a new region. It may also be that geological processes that fragment ranges (e.g., changing sea levels) will often continue to fragment ranges, resulting in further speciations quickly after the first. It also seems likely that the early life of a new species is often a delicate one. This could be due to competition with its sibling species or other closely related species, or the fact that a new habitat is likely to contain perils as well as opportunities, or the fact that a fragmented range is likely to result in small populations. The age-dependent model does not represent any of these processes explicitly, but a spike near zero in the event density function does capture the overall effect. Losos and Adler (1995) provide biological justification for the opposite case, where a species has a lower rate of speciation when it is young. This produces more balanced trees than the CRBD model.

The theory of branching processes (Athreya and Ney 1972, chapter IV) describes the long-term behavior of age-dependent models. If $f$ is the event density function and $m = 2/(1 + p)$ is the expected number of offspring at the termination of a species, then the solution of

$$m \int_0^\infty e^{-\alpha y} f(y) dy = 1$$

is the Malthusian parameter $\alpha$. The long-term expected number of species is then proportional to $e^{\alpha t}$, where $t$ is the time from the origin.

Understanding the behavior of these models is quite difficult, especially in the presence of extinctions, and they can seem counterintuitive at times. It is worth remarking that to produce imbalance in the presence of extinctions, it not sufficient that one subtree starting at a node grows much bigger than the other. It is also necessary that the smaller subtree actually survives to the present. I am not aware of results from branching theory that relate particularly to balance. There are asymptotic results which describe the moments of the number of species as time tends to infinity (Athreya and Ney 1972, section IV.5). It seems reasonable to suppose that large values of high-order moments would be associated with large amounts of imbalance.

### Measuring Balance

Many ways have been suggested for measuring the balance of a tree. Most of the measures, such as the popular Colless statistic, are not independent of tree size. Thus, trees of different sizes cannot be directly compared. Ford (2006) and Blum and François (2007) estimated parameter values from trees taken from TreeBASE. I will present results from fitting the beta-splitting model to three data sets. This model was chosen because it is sampling consistent and can generate trees more balanced than Ford’s model.

When Blum and François (2007) compared the TreeBASE trees to Aldous’s beta-splitting model, they found very good agreement with the AB model ($\beta = -1.0$). Their estimate for $\beta$ was $-0.95$, very close to $-1.0$. However, there is a problem with their estimator. They used an maximum likelihood (ML) estimator for $\beta$. Over the main range of interest, $\beta \in [-1.5, 0]$, this estimator is biased, producing higher (less negative) estimates for $\beta$ than it ideally should. Table 1 shows mean values for the ML estimate $\beta_{\text{MLE}}$ of $\beta$ for the CRBD, AB, and PDA models, based on 10,000 simulated trees for each mean, for tree sizes 10, 20, 50, and 100. Any model that produces perfectly balanced trees (where all splits are as even as possible) with a nonzero probability will result in an ML for $\beta$ that will sometimes be infinite.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$B$</th>
<th>Means of $\beta_{\text{MLE}}$</th>
<th>Means of $\hat{\beta}$</th>
<th>$\beta$ values from $\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>-1.50</td>
<td>0.333</td>
<td>-0.97</td>
<td>-1.34</td>
<td>-1.46</td>
</tr>
<tr>
<td>-1.00</td>
<td>0.5</td>
<td>-0.22</td>
<td>-0.59</td>
<td>-0.89</td>
</tr>
<tr>
<td>0.00</td>
<td>0.667</td>
<td>0.54</td>
<td>0.46</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Notes: The first two columns show true values. $\beta_{\text{MLE}}$ is the ML estimator for $\beta$. $\hat{\beta}$ is a Bayesian estimator for $B$. Means from many simulated samples are shown for these two estimators for various tree sizes. The rightmost section shows the means for $\hat{\beta}$ translated into $\beta$ values for direct comparison with $\beta_{\text{MLE}}$. 

**TABLE 1.** The behavior of two estimators of balance for the four tree sizes (in the second row). $\beta$ and $B$ are alternative parameters controlling balance in the beta-splitting model of Aldous (1996)
Although Blum and François (2007) counted any estimates above 2.0 as 2.0, in order to remove such infinities, it should not be surprising that it remains biased.

Figure 2 shows a scatterplot of ML estimates for β of trees from TreeBASE, plotted against tree size. It is similar to figure 2 of Blum and François (2007), and any estimate above 2.0 is counted as 2.0, but the data are somewhat different, and the algorithm for local regression may differ too. As well as the scatterplot, Figure 2 also shows a local regression curve (solid line) for these trees, and for comparison, a curve (dashed line) showing the mean of ML estimates for β for trees made from a large number of simulations of the AB model. The main deviation of the TreeBASE trees from the β = −1.0 line is similar to that for the AB model, that is, the deviation is mainly due to the properties of the statistic used. In particular, the comment of Blum and François (2007) that “the data shows that tree shape undergoes a rapid change from the smaller to the intermediate-sized and larger trees” is not supported when this behavior of the statistic is taken into account. Thus, the model actually fits the data better than was claimed.

The nature of the problem with the ML estimator of β suggests reparameterizing. Instead of estimating β, I estimate \( B = (\beta + 2)/(\beta + 3) \), which can be inverted to give \( \beta = (3B - 2)/(1 - B) \). This B has a range from 0 to 1, and Table 2 shows some key values. Other mappings \( \beta \rightarrow B \) could be used, but this one is convenient to think about when choosing prior and loss functions for B in a Bayesian context, and it turns out to produce approximately symmetrical distributions on the data and models. I use a Bayesian estimator for \( B \) based on a uniform prior on \([0, 1]\) and squared-error loss function.

I will denote this estimator as \( \hat{B} \) and will use it as the basic measure of balance. The choice of prior and loss is a matter of subjective judgment. The prior can be based on estimates of probabilities such as \( \Pr(B > 2/3) \), which is the probability that a randomly selected tree is more balanced than a CRBD tree. If no data about tree balance were available at all, one might well set this to be 1/2, but the fact that empirical trees are usually less balanced than a CRBD tree was known long before the data sets here were collected. One could argue that the prior should be more concentrated, with \( \Pr(B \in [1/3, 2/3]) \) larger than 1/3, but the uniform prior is more robust against the occurrence of very balanced or unbalanced trees in the data. The squared-error loss function is symmetric and treats errors in \( \hat{B} \) evenly over the range \([0, 1]\), so that, for example, an estimate of 0.2 when the true value is 0.3 is regarded as equally bad as estimating 0.6 when the true value is 0.5. This seems reasonable to me in the context of phylogenetic trees. For example, equations (5–7) give some idea of the consequences of getting \( \hat{B} \) wrong in the context of phylogenetic analysis.

The squared-error loss function results in \( \hat{B} \) being the posterior mean of B (Berger 1985, section 4.4.2).

It should be emphasized that the choice of prior and loss are mainly important in estimating the balance of small trees. The data will dominate for large trees. The case of a single tree of size 4 will illustrate how \( \hat{B} \) differs from the MLE for B. The likelihood given a single observation of a tree that splits 2 : 2 at the root can be calculated as \( f(B) = 3B/(4 + 3B) \), which is a curve, not far from a straight line, which rises from 0 at \( B = 0 \) to 3/7 at \( B = 1 \). The MLE of B is thus 1, whereas \( \hat{B} \) is the mean of the posterior density that is proportional to \( f(B) \); and this mean can be calculated as approximately 0.636. The MLE of B corresponds to a process that always produces perfectly balanced trees, whereas \( \hat{B} \) corresponds to a process that produces slightly less balanced trees than CRBD. The frequentist behavior of \( \hat{B} \) under the CRBD, AB, and PDA models for various tree sizes is shown in Table 1. This shows mean values for \( \hat{B} \) for these models, based on 10,000 simulated trees for each mean, for tree sizes 10, 20, 50, and 100. It also shows these values translated back into \( \beta \) values for comparison with \( \beta_{MLE} \).

It can be seen that \( \hat{B} \) gives accurate estimates for smaller tree sizes than \( \beta_{MLE} \).

To distinguish between trees having very similar values of \( B \), further statistics are required. I define a statistic \( D_{n,m} \) to be the difference between \( B \) for the whole tree of size \( n \) and the average value of \( B \) for the trees obtained by removing all but \( m \) tips. More formally, let \( T \) be a tree with \( n \) tips and let \( m < n \). Define

\[
D_{n,m} = \left( \begin{array}{c} n \\ m \end{array} \right)^{-1} \sum B_S - B_T,
\]

where \( B_S \) and \( B_T \) are the average values of \( B \) for the \( m \) trees obtained by removing all but \( m \) tips from \( T \) and for the whole tree of size \( n \), respectively.

Table 2. Relation between \( B \), Aldous’s \( \beta \), Ford’s \( \alpha \), and tree type

<table>
<thead>
<tr>
<th>( B )</th>
<th>( \beta )</th>
<th>( \alpha )</th>
<th>Tree type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>−2</td>
<td>1</td>
<td>Comb</td>
</tr>
<tr>
<td>1/3</td>
<td>−3/2</td>
<td>1/2</td>
<td>PDA</td>
</tr>
<tr>
<td>1/2</td>
<td>−1</td>
<td>—</td>
<td>AB</td>
</tr>
<tr>
<td>2/3</td>
<td>0</td>
<td>0</td>
<td>CRBD</td>
</tr>
<tr>
<td>1</td>
<td>∞</td>
<td>—</td>
<td>Very balanced</td>
</tr>
</tbody>
</table>
where the sum runs over all subtrees \( S \) of \( T \) obtained by deleting \( n - m \) tips from \( T \), and where \( B_S \) and \( B_T \) are the \( B \) statistic as defined earlier and calculated for the relevant trees. In practice, when \( n \) is large the values of \( D_{n,m} \) will be estimated from a random subset of the subtrees. \( D_{n,m} \) is a measure of the deviation from sampling consistency. It has a theoretical value of 0 for the beta-splitting model of Aldous (1996). Note that because it is based on \( B \), which is in turn based on the beta-splitting model, it is not guaranteed to be zero for every sampling consistent model. The following may give an intuitive understanding of what \( D_{n,m} \) measures. When tips are pruned, the general effect is remove “recent growth” but leave the branches nearer the root mostly untouched. Thus, the values for \( B_S \) reflect the balance nearer the root, and \( D_{n,m} \) measures the difference in balance between the “older” and “newer” parts of the tree. For example, if a tree is constructed by first making a comb with 32 tips (i.e., as unbalanced as possible), and then putting 32 balanced trees of size 4 on each of these tips, to make a tree of size 128, \( B \) is about 0.54 for the whole tree but drops to about 0.39 when 64 of the tips are removed, so that \( D_{128,64} \) is about −0.15 for this tree. Likewise, it is possible to construct trees which are balanced near the root and unbalanced near the tips that have \( D_{n,m} > 0 \).

A tree may be unbalanced because nearly all its splits are fairly uneven, or because a few of its splits are extremely uneven. A simple measure of this tendency is to count the fraction of splits that are as extreme as possible, of form \((1, n - 1)\) or \((n - 1,1)\). It is not very meaningful to compare fractions for very different \( n \), so unless the trees are very large, it is not useful to compare two trees this way. However, given a collection of trees one can collect all the splits of size \( n \) for various \( n \) and thus compare two collections of trees. Under the beta-splitting model, the values of this can be found from Proposition 1.

**Comparison of Models and Data Sets**

Three data sets of empirical trees were used, the first from TreeBASE. The TreeBASE trees were downloaded from http://www.phylo.org/treebase/trees/in August 2007. The index numbers ran from 705 through 3148. Trees containing multifurcations were ignored. Many sets of taxa contain an outgroup that has been deliberately added, and this of course affects the balance at the root. Blum and François (2007) used an automatic method for detecting this situation by looking for extreme splits at the root, but it is difficult to make exact comparisons between simulated and real data if this is done. Many of the models considered here will produce trees with a \((1, n - 1)\) or \((n - 1,1)\) split at the root around 20% of the time, even when \( n \) is large, although of course no outgroup is involved in the simulations. Instead, I use a method like that of Ford (2006). The root itself was ignored for all the TreeBASE trees, and the two subtrees at the root were considered. All such subtrees (which I shall call trees from now on) with at least four tips formed the basic data set used here. It consists of 739 trees with sizes from 4 to 268. About half were of size at least 20; 104 had size at least 50; 18 had size at least 100. These data were used for the scatter plot in Figure 2 discussed previously. For the main balance results presented below, only the 140 trees of size at least 40 were used. This is to ensure there are no artifacts arising from the choice of statistic. From Table 1, it can be seen that \( \hat{B} \) shows very little bias for trees of this size. The TreeBASE data have not been subjected to any quality control, it is very eclectic, and many of the trees are sparsely sampled. It is, however, the largest set of data available.

The second set of empirical trees was compiled by McPeek and Brown (2007) and is available from http://www.enallagma.com/data And Software.php. These trees were manually extracted from the literature and consist of species trees in which sampling is at least 50%, and many are believed to be 100% complete. The data set is of high quality, but fairly small. Outgroups have already been removed, so the roots were left in. Some contain multifurcations, but ignoring these would omit too much of this precious data, so multifurcations were resolved randomly using a method like that of Blum and François (2007), with the AB model used to generate the missing branches. I set the minimum size threshold to 30 tips instead of 40 to keep more of them. There were then 38 trees to analyze. Results using only the binary trees, only the trees with complete sampling, and only those above size 40 were very similar to those presented below.

The third set of empirical trees was compiled by Heath et al. (2008). These trees were extracted from the literature, and consist of species trees in which sampling is at most 10%. The data set is of high quality, and outgroups have been removed, but the low sampling rates mean that bias may well have been introduced by the way the taxa were chosen. Some of the trees were inferred using ML or Bayesian techniques, others used maximum parsimony. The results reported here use just the ML/Bayesian trees, and were treated like the McPeek–Brown data, except that the minimum size threshold was 40. This resulted in 49 trees that I will refer to as “Heath trees”. Results for the parsimony trees were similar.

The number and sizes of the trees sampled from each model were chosen to match the TreeBASE data, so there are 140 trees in each sample. CRBD, Ald85, and PDA are Aldous beta-splitting models with \( \beta = 0, -0.85, -1.5 \), respectively. F18 is Ford’s model with \( \alpha = 0.18 \). KS12 is the Kirkpatrick–Slatkin model with \( p = 0.12 \), so that its splitting probabilities are given by equation (8). BF50 is the Blum–François model from Blum and François (2007) with parameter \( \alpha = 0.50 \), meaning that \( p \) is chosen from a beta distribution on \([0,1]\) with shape parameter 0.50. The others are age-dependent models with survival functions as follows. AdG1 and AdG9 use a gamma density with shape parameter 4. AdW1 and AdW9 use a Weibull density with shape parameter 0.57, so that \( S(t) = \exp(-t^{0.57}) \). AdP1 and AdP9 use a survival
function with \( S(t) = (1 + t)^{-\alpha} \), corresponding to a Pareto density function \( f(t) = \alpha(1 + t)^{-1-\alpha} \). Here \( \alpha \) is set to 1.5. The final digit in each case indicates the ratio \( \rho \) of the extinction rate to speciation rate. “1” means that \( \rho \) is 0.1 (i.e., 10 speciations for every extinction), and “9” means that \( \rho \) is 0.9. In each case, the single parameter was chosen to give \( \hat{B} \) close to 0.53, which is approximately the mean balance of all the empirical data, except that for the age-dependent models, the parameter was chosen so that both choices for \( \rho \) gave a reasonable match. \( \rho \) is not a free parameter in a model with extinctions because the rate of extinctions could in principle be determined by other means (although this is very difficult in practice and was not attempted here).

Because changing \( \rho \) changes \( \hat{B} \), the match here can only be done roughly.

**Balance**

I will discuss the three statistics in turn, and for each one deal with comparisons between data sets, then comparisons between models and data. Figure 3 shows violin plots of balance for TB, which is derived from the TreeBASE data, McPB for the McPeekBrown data, and Heath for the Heath data, and the models. The McPeek–Brown data set is more balanced than the other two data sets in which taxa have been more sparsely sampled. It is surprising that the eclectic and mixed-quality TreeBASE data set shows no higher variance than the high quality and much more homogeneous McPeek–Brown data set. The TreeBASE data set is larger, so a larger range is to be expected.

Apart from the CRBD and PDA models that are included for comparison, all the models can match the data well. It is hardly surprising that the models can match the balance of the data because they all have a parameter or a function that can be adjusted to achieve this. It is surprising that the variance is so well matched by most of the models. This is not what one expects when simple models are matched to the results of a very complex process. It is not just the evolutionary process itself; the choice of taxa by researchers, and the methodology used to produce the trees are also possible sources of variance not captured in any of the models. Note that although biases can cancel, with some factors tending to increase and others tending to decrease the balance, variances always add. I cannot think of any plausible process that might “censor” the data by removing, or tending to remove, extreme values of balance. It is also noteworthy that nearly all the models show such similar variance to one another. Only the KS12 shows markedly smaller variance.

Age-dependent models can behave differently for different extinction: speciation ratios. Of those presented here, the balance of AdW model was least variable. Several other possibilities for event density functions were investigated (results not shown). The general situation is that the right amount of imbalance can be produced by a spike near zero or a long tail when extinctions are low, but a long tail is needed to produce imbalance when extinction is high.

**Deviation from Sampling Consistency**

Figure 4 shows violin plots for the \( D = D_{n,m} \) statistic measuring deviation from sampling consistency, with \( m \) is set to an integer close to \( n/2 \). The same data sets and models are used as for balance. The TreeBASE and Heath data both show a reduction in balance when the trees are sampled, but the McPeek–Brown data set is almost sampling consistent. Note that the variance is not larger in the TreeBASE data set. The CRBD, PDA, Ald85, and F18 models all show little deviation from sampling consistency, as expected from theory. The two BF models KS12 and BF50 show that these can produce either more or less balance when subsampled. It is possible, given particular \( n \) and \( m \), to find a density on \([0, 1]\) so that models of this type appear sampling consistent (results not...
shown) but given the negative result of Proposition 2, this did not seem worth pursuing. The age-dependent models all show somewhat increased balance when sub-sampled. This seems to be a general feature and applies to several other event density functions not shown here.

**Extreme Splits**

Figure 5 shows proportions of extreme splits for the TreeBASE data and various models. The same set of trees of size at least 40 is used as for the balance measurements, but in this case, the trees were treated as a collection rather than as individual trees. Internal nodes with subtrees in certain size ranges (20–40, 40–80, 80 and more) were found in the collection, and the proportion of those nodes with extreme splits is shown in the graphs. Because this statistic is in general expected to vary with tree size, it does not make much sense to compare values for trees of very different sizes. The ranges represent a compromise between avoiding bias due to different tree sizes and obtaining enough samples to reduce the variance of the calculated proportions. The two other data sets are too small to give reliable estimates for the individual ranges. The overall values for all trees of size above 20 in these sets was 0.14 for the Heath data and 0.22 for the McPeek–Brown data.

All the empirical data show considerable proportions of extreme splits. Remember that outgroups were effectively removed from the TreeBASE data by ignoring the root, and that these represent splits of internal nodes. In the TreeBASE data, the proportion of extreme splits actually increases as the tree size grows from 20–40 to 40–80, a fact which is not easy to explain except by chance.

The CRBD trees have very few extreme splits and the PDA trees have too many; the estimated values in both cases are close to the theoretical values. This statistic shows that the property proved in Proposition 2 for sampling consistent BF models is not seen in real data. The KS12 model does not produce enough extreme splits to even roughly approximate the data, but the others are reasonably close given the variance of the estimates.

**DISCUSSION**

**Models**

I have not used any formal model selection procedure to rank the models. One reason is that the models fall into three groups: (1) those that model only topology; (2) those that model a process of diversification in (distorted) time, but do not model extinctions; (3) the age-dependent ones that model a process in real time and include explicit extinctions. These three groups also increase in mathematical complexity. It seems better to regard these groups complementing one another, and choosing between them will depend on the use to which they are put as much as which provides the best fit. Another problem with model selection is that those in groups (2) and (3) require a choice of a density, not just a finite set of parameters.
The CRBD, age-dependent, and BF models can be compared in general terms. In the CRBD model, each species knows nothing at all about the rest of the tree, and nothing about its own history. In an age-dependent process, each species knows its birth date, but no more; it does not inherit anything from its parent species. In the BF models, species inherit something from their parents, so that “fast” species tend to have fast descendants, but their individual rate is then constant throughout their life. BF models do not model extinctions, whereas the CRBD model and age-dependent models may. Including extinctions in the CRBD model does not affect balance, but in age-dependent models, extinctions will affect balance in general.

Clearly the CRBD model is too balanced and the PDA model is too unbalanced, as reported by others. In group (1), the models of Aldous (1996) and Ford (2006) with appropriate parameter values both match the data well. In group (2) are the models of BF type, which can match the data with a suitable choice of density on [0, 1], with the beta density of Blum and François (2007) being better overall than that of Kirkpatrick and Slatkin (1993). The main defect of the age-dependent models is that they do not match the deviation from sampling consistency seen in the data. The Weibull event density function seems the best among those presented here because it produces similar balance over a wide range of extinction rates. Because the age-dependent models attempt to model more of the evolutionary process, I will discuss how one might evaluate them in other ways.

Age-dependent models quite often produce a tree which is very unbalanced at the root, despite the fact that both species starting at the root follow an identical process. This can happen because a moderate probability is assigned to a “flurry” of quick speciations, and a moderate probability is also assigned to a species, or small subtree, which takes a long time to “get going.” Once both subtrees have reached a modest size, their growth will be approximately exponential, with the same Malthusian parameter controlling their rate of growth, so that the ratio between their sizes tends to a constant. This makes quite different predictions about the size of the subtrees at different stages of their growth, as compared with, say, a standard birth–death model that is fitted separately to each subtree. For the age-dependent model, most of the relative difference in size (as a ratio) will emerge early on, whereas two standard birth–death models will show this ratio increasing exponentially, that is, slowly to begin with, but increasing without limit. It may be that fossil evidence could distinguish these possibilities.

It seems that long tails are needed to match the data as well as possible. One way of assessing whether this is plausible, is to look at the probability a species lasts much longer than the median length. Table 3 shows some values for three event density functions: the exponential that produces the standard birth–death model with constant rates; and the gamma and Weibull distributions with the same parameters as used earlier. I have not compared these with any data, but it does seem that the exponential distribution makes very long-lived species too unlikely. In a recent study, Alfaro et al. (2009) found that the CRBD model gave a poor fit to the low-diversity lineages.

Models can be useful even if they offer no insight into underlying mechanisms. If a model characterizes what a typical phylogenetic tree looks like, one can investigate whether particular traits occur among species, which are associated with phylogenetic trees that do not look typical. The low variance of the TreeBASE data despite its apparent shortcomings makes this approach more promising than one might expect.

### Table 3. Approximate probabilities of a species lasting much longer than the median for three distributions for lifetimes

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Multiple of median</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Exponential</td>
<td>1 in 1000</td>
</tr>
<tr>
<td>Gamma</td>
<td>1 in 15</td>
</tr>
<tr>
<td>Weibull</td>
<td>1 in 13</td>
</tr>
</tbody>
</table>

Notes: The gamma has a shape parameter of 0.4 and the Weibull has a shape parameter of 0.57.
Implications for Bayesian Phylogenetic Analysis

The usual priors for tree topology in Bayesian phylogenetic analysis, namely PDA and CRBD, are poor matches for the balance in empirical trees. The PDA model is the worst and appears especially unsuitable if the sampling is nearly complete. Of course, if there is a strong enough signal in the alignment the prior will be overwhelmed, but it is not easy to tell how strong the signal is. The main danger that I see is that a posterior clade probability such as 0.99 is interpreted as due to a strong signal, when it might be due to a weak signal plus bias from the prior. If a Bayesian analysis is working as it should, with a model and prior that are both good enough, then posterior probabilities can be interpreted simply, but poor choices of prior or model can make the posterior probabilities quite meaningless. It is good that simply, but poor choices of prior or model can make the posterior probabilities quite meaningless. It is good that the PDA and CRBD models do at least bracket the degree of balance observed in real data, so if a tree is supported by two analyses, one with a PDA prior and one with an CRBD prior, there is much less to worry about.

The TreeBASE data leave much to be desired when attempting to draw conclusions about macroevolutionary processes, and the Heath data are sparsely sampled, but these data sets do seem appropriate for designing a tree topology prior for many Bayesian phylogenetic analyses, in the sense that whatever biases researchers introduce by sampling, they are likely to be present in analyses for some time to come. If sampling is nearly complete, then the McPeek–Brown data forms a better basis. A tree topology prior based on the Aldous’s beta-splitting model seems a good choice, leaving the choice of β to the user. I think this would be straightforward to implement in a Bayesian phylogenetic analysis program that uses rooted trees, and rather more complex in one that uses unrooted trees. The choice of prior for node times or branch lengths is a topic that deserves further research.

ACKNOWLEDGMENTS

I thank Tracy Heath for trees, Anne-Marie Fiore-Donno for discussions, and Klaas Hartman, Arne Mooers, and Mark Holder for their constructive comments on earlier versions of this article.

REFERENCES


APPENDIX

Proof of Proposition 1. The key is to rewrite the sum of \( n - 1 \) terms of form \( \Delta(i, n - i) \) as a sum of \( n \) terms, \( n - 2 \) of which are of form \( \Delta(i, n + 1 - i) \).

\[
(n + 2 + 2\beta)s_n = (n + 2 + 2\beta)(\Delta(1, n - 1) + \Delta(2, n - 2) + \cdots + \Delta(n - 1, 1))
\]

\[
= (n + \beta)\Delta(1, n - 1) + (2 + \beta)\Delta(1, n - 1) + (n - 1 + \beta)\Delta(2, n - 2) + (3 + \beta)\Delta(2, n - 2) + (n + \beta)\Delta(3, n - 3) + \cdots + (n + \beta)\Delta(n - 1, 1)
\]

From (2), it follows that

\[
(i + 1 + \beta)\Delta(i, n - i) + (n - i + \beta)\Delta(i + 1, n - i - 1)
\]

\[
= (n + 1)\Delta(i + 1, n - i)
\]

for \( 1 \leq i \leq n - 2 \), so

\[
(n + 2 + 2\beta)sn = 2(n + \beta)\Delta(1, n - 1) + (n + 1)\sum_{i=2}^{n-1} \Delta(i, n + 1 - i)
\]

\[
= 2(n + \beta)\Delta(1, n - 1) + (n + 1)s_{n+1} - 2(n + 1)\Delta(1, n).
\]

From (2), the equation \( (n + \beta)\Delta(1, n - 1) = n\Delta(1, n) \) is easily verified, so it follows that

\[
(n + 2 + 2\beta)s_n = (n + 1)s_{n+1} - 2\Delta(1, n)
\]

\[
= (n + 1)s_{n+1} - 2(n + 1)\Delta(1, n).
\]

It is shown in the supplementary material of Blum and François (2007) that under the above conditions,

\[
q(x, y) = \frac{1}{2} \left( y + \frac{x - 1}{x - 1} \right) \int_0^1 (p^{x-1} (1 - p)^{y-1}) + p^{y-1} (1 - p)^{x-1} f(p)dp.
\]

(11)

(This results from integrating (8). Blum and François only consider the case of a Beta distribution, but the generalization is obvious.) Note that \( f \) and therefore \( q \) is symmetric (i.e., \( q(x, y) = q(y, x) \)). Now Proposition 41 of Ford (2006) gives a necessary and sufficient condition for sampling consistency in terms of the \( q(x, y) \), namely that for all \( x, y \geq 1 \), the following holds:

\[
(x + y + 1 - 2q(1, x + y))q(x, y)
\]

\[
= (x + 1)q(x + 1, y) + (y + 1)q(x, y + 1).
\]

(12)

Putting \( x = y \) and substituting for \( q(x, x) \) and \( q(x, x + 1) \) using (11) gives

\[
(2x + 1 - 2q(1, 2x))(2x - 1) \int_0^1 (p^{x-1} (1 - p)^{x-1} f(p)dp
\]

\[
= (x + 1)(2x - 1) \int_0^1 (p^{x-1} (1 - p)^{x-1}) f(p)dp.
\]

But the two integrals are identical because \( p^{x-1} (1 - p)^{x-1} + p^{x-1}(1 - p)^x = p^{x-1}(1 - p)^{x-1} (p + 1 - p) = p^{x-1} (1 - p)^{x-1} \). They therefore cancel regardless of \( f \) and it follows that

\[
(2x + 1 - 2q(1, 2x)) \left( \frac{2x - 1}{x - 1} \right) = (x + 1) \left( \frac{2x - 1}{x - 1} \right).
\]

This can be simplified to give \( q(1, 2x) = 1/2x \) as required.

\[\square\]

Notes on Algorithms Used

The programs for generating sample trees are written in C++. The graphs and tables were produced using R (R Development Core Team 2008). Sampling from the models with splitting distributions is straightforward. The age-dependent models are more difficult (Stadler 2008).

I assume that the survival function \( S \) is given. Suppose that at a particular time, there are \( n \) extant species in the tree with birth times \( b_i \) (\( 1 \leq i \leq n \)). The termination times \( d_i \) can be found by sampling from the appropriate density and adding to \( b_i \). A speciation or extinction is chosen according to the value of \( \rho \) in the model. The time of the next event in the process is the smallest value of \( d_i \).

In order to generate trees with a given numbers of tips matching the TreeBASE data, the following was done. It is similar to the method of Stadler (2008) though
developed independently. The process was repeatedly run from a starting point of a single species until either it becomes extinct or it becomes so large that the probability that the number of tips will ever become equal to the largest tree required (of size 268) is small enough to ignore. The process was run until 280 (twice the number of trees in the TreeBASE data of size at least 40) runs each contained at least one tree of size 40. In practice, nearly all the runs ended with a big tree not extinction, so they nearly all contained at least one tree of all sizes 40–268. Each occurrence of a tree is accompanied by a duration, the time that the process spends in this state before an extinction or speciation changes it into another tree. Trees were then sampled in proportion to these durations. This produces unbiased samples under the assumption that the root time has a (improper) uniform distribution over all times in the past. The samples are not independent: it is possible that the same tree is sampled twice, and trees within one run are likely to be similar to one another.