# Tree Models for Macro-Evolution and Phylogenetic Analysis

Graham Jones

5th February 2009

email: art@gjones.name
web site: www.indriid.com

**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 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. The TreeBASE data is compared to simulated data from a range of models, and it is shown that several simple models generate trees which match the amount and spread of imbalance in real data to a remarkable degree. Two statistics are developed which distinguish between trees with the same overall imbalance and the match between models and data for these statistics is investigated. In particular age-dependent (Bellman-Harris) branching process are studied in detail.

# 1 Introduction

It has long been recognized that phylogenetic trees are more unbalanced than those generated by a Yule process [2], [14]. Two recent papers [9], [5] have quantified the degree of this imbalance for the large set of trees available in the TreeBASE data [19]. This paper presents a more detailed analysis of the same data set, and considers more models.

## 1.1 Why balance matters for phylogenetic analysis

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 ERM (Equal Rates Markov) and the PDA (Proportional to Distinguishable Arrangements) models. The ratio of the probabilities of these two trees under the ERM model is $n$, while under the PDA model it is 1. (See sections 2.1, 2.2 and 2.3 for more details.) 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 [6] with its usual ERM prior would give an $n : 1$ advantage to $(C, (x, y))$ over $((C, x), y)$ compared

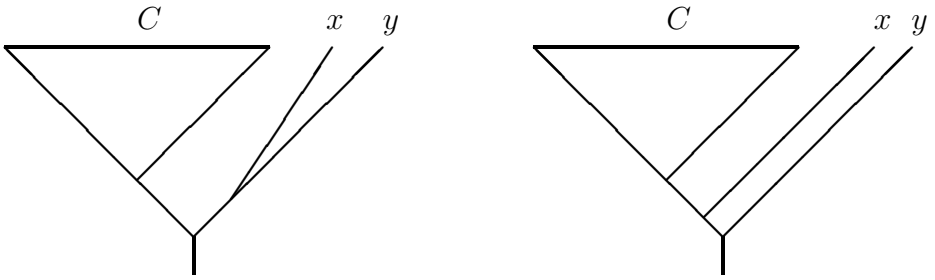


Figure 1: Two similar trees.

to using MrBayes 3.1 [12] with the usual PDA prior, when everything else in the analyses was the same.

The degree of balance is closely related to the distribution of node times or branch lengths. It may be that the commonly used priors based on the ERM model for node times, or the exponential distribution for branch lengths, make very short and very long branches too unlikely.

### 1.2 Outline of this article

I show that the model considered in [9] fits the data *better* than the article claims, and that the models considered in [9], [5] as well as some others, 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 since they each have a parameter which 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 which distinguish between trees with very similar balance, and demonstrate that trees can vary independently in all three 'directions'. I investigate how various models fit the TreeBASE data as measured by these statistics. Finally I concentrate on models based on age-dependent (Bellman-Harris) branching process. These 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.

I start with a discussion of the general setup, and the beta-splitting model of Aldous [2], which is the most mathematically tractable and which forms the basis of much of the statistical analysis of other models.

## 2 Fundamentals and the beta-splitting model

All trees considered here are binary and rooted. By a model of tree formation I mean a probability distribution over rooted tree topologies, and possibly over node times as well, for trees with a given number of tips. The probability distribution may be defined implicitly via a process which generates the trees, or explicitly as a formula for calculating the probability of a given tree under the model (or both).

In some cases the trees are defined via what happens to diversification rates (or speciation rates and extinction rates) at nodes or along branches, so that they can be viewed as developing in time. In this case, there will also be a probability distribution for node times. It may be the case that the 'time' in such trees is a distorted version of real time. The main focus here is on topologies but evolution is a process in time and so such a model seems preferable.

When a speciation occurs, the two subtrees 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 $\{(1, n-1), (2, n-2), ...(n-1, 1)\}$, one distribution for each positive integer $n$. Such a distribution is known as a *splitting distribution* [9]. I will also use (or abuse) the term to refer to a set of distributions for all $n$. Since speciations are supposed to happen independently in all models considered here, a splitting distribution implies a distribution over tree topologies: to find the probability of a tree you multiply the probabilities of the splits at each internal node.

The rest of this section concerns the beta-splitting model of Aldous [1], starting with two well known special cases.

## 2.1 The Yule or ERM model

The Yule model is the oldest and best-known. It is the distribution you get if all species in the tree behave independently of one another and all are assumed to have the same constant diversification rate. Its topology can be characterized as a tree in which $q(x, y) = 1/(n-1)$ for all $x$, $y$ and $n = x + y$. It is also known as the 'Equal Rates Markov' distribution or ERM model, and that is what I shall call it.

## 2.2 The PDA model

The PDA (Proportional to Distinguishable Arrangements) 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 [2], [22], [21], [16] but it is not as simple as the ERM model when viewed as a process in time. See section 6 for further discussion. The trees are more unbalanced than those of the ERM model. Real phylogenetic trees are generally more balanced than PDA but less balanced than ERM. The PDA model can be characterized by a splitting distribution as a special case of the beta-splitting model described next.

## 2.3 The beta-splitting model of Aldous

Aldous [1] described a family of models parameterized by a single parameter $\beta$ taking values in $[-2, \infty]$. The balance increases with $\beta$, with the PDA model appearing at $\beta = -1.5$ and the ERM 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:

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

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 [13], but this does not seem to be one that is biologically realistic (the diversification rates depend on the size of the tree to be generated). Of particular interest is the value $\beta = -1$ since it has some special mathematical properties and appears to be a very good fit to the available data. I will call this the AB model.

The following result is useful in calculations involving this 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*

$$\Delta(i, n-i, \beta) = \frac{\Gamma(i+1+\beta)\Gamma(n-i+1+\beta)}{\Gamma(i+1)\Gamma(n-i+1)} \tag{2}$$

*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:*

$$s_{n+1} = \frac{1}{n+1}\Big(n+2+2\beta+\frac{2(n+\beta)}{n}x_n\Big)s_n$$

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

*Proof.* 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 $\Delta(i, n+1-i)$.

$$\begin{aligned}
(n+2+2\beta)s_n &= (n+2+2\beta)(\Delta(1,n-1)+\Delta(2,n-2)+...+\Delta(n-1,1)) \\
&= (n+\beta)\Delta(1,n-1)+ \\
&\quad (2+\beta)\Delta(1,n-1)+(n-1+\beta)\Delta(2,n-2)+ \\
&\quad (3+\beta)\Delta(2,n-2)+(n-2+\beta)\Delta(3,n-3)+ \\
&\quad \cdots \\
&\quad (n-1+\beta)\Delta(n-2,2)+(2+\beta)\Delta(n-1,1)+ \\
&\quad (n+\beta)\Delta(n-1,1)
\end{aligned}$$

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

$$\begin{aligned}(n+2+2\beta)s_n &= 2(n+\beta)\Delta(1,n-1)+(n+1)\sum_{i=2}^{n-1}\Delta(i,n+1-i)\\ &= 2(n+\beta)\Delta(1,n-1)+(n+1)s_{n+1}-2(n+1)\Delta(1,n).\end{aligned}$$

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

$$\begin{aligned}(n+2+2\beta)s_n &= (n+1)s_{n+1}-2\Delta(1,n)\\ &= (n+1)s_{n+1}-2s_{n+1}x_{n+1} \qquad (3)\end{aligned}$$

and

$$\begin{aligned}n(n+2+2\beta)s_n &= n(n+1)s_{n+1}-2(n+\beta)\Delta(1,n-1)\\ &= n(n+1)s_{n+1}-2(n+\beta)x_n s_n \qquad (4)\end{aligned}$$

Dividing (3) through by $s_{n+1}$, and (4) by $ns_n$ leads to the desired result. □

As an immediate application of this Proposition, it is easy to show by induction that if $\beta=-1.5$ then

$$x_n = n/(4n-6), \qquad (5)$$

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 $n \geq 2$

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

Note that the probability of the topology inside $C$ cancels out in the above. Now

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

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

$$R_{n+2}(\beta) = \frac{(2+\beta)n}{2(n+\beta)x_{n+1}} \tag{6}$$

Using this together with Proposition 1 allows $R_{n+2}(\beta)$ to be calculated recursively for any $\beta$. For the special cases ERM, AB, PDA it follows, either by induction using (6), or by direct calculation, that

$$R_{n+2}(0) = n \tag{7}$$

$$R_{n+2}(-1) = \sum_{1}^{n}(1/i) \approx 0.577 + \log(n) \tag{8}$$

$$R_{n+2}(-1.5) = 1 \tag{9}$$

The next section develops some statistics related to $\beta$ and the probability of extreme splits $x_n$.

# 3 Measuring balance

Many ways have been suggested for measuring the balance of a tree. Most of the proposed measures such as the popular Colless statistic are not independent of tree size, so they cannot be used to compare a tree of size 80 with one of size 50. In principle, they could be calibrated, but one needs a null model to calibrate against.

A more useful approach, at least for the purposes of this paper, and which was adopted by [9] and [5] in their studies of the TreeBASE data set, is to estimate the parameter of a model which can generate trees of varying balance. Ford [9] used his alpha model in this way, while [5] used Aldous' beta-splitting model, as I will. However, considerable care is still needed, as I show in the next two sub-sections.

## 3.1 Ford's p-values

The model itself is described in more detail in section 4.2. Here I just note that it has a single parameter $\alpha$ which controls the degree of balance. Ford [9] estimated $\alpha$ for trees of size at least 10 from TreeBASE using a maximum likelihood (ML) estimator. The estimates had a median at 0.22, but varied considerably from tree to tree. He then went on to calculate various measures of fit, by taking each tree, finding an estimate $\hat{\alpha}$ of $\alpha$, generating many trees from the model with $\alpha$ set equal to $\hat{\alpha}$, and thus estimating the distribution of some statistic (such as the Colless index) under the model and finally producing a p-value for each tree. In theory, the p-values should have a uniform distribution on $[0,1]$ if the model is correct.

This procedure is flawed. The model is being overfitted to the individual trees. Figure 2 replicates (closely, not exactly) one of Ford's results and also shows the result if $\alpha$ is set to 0.25 for all

trees. Despite 0.25 being an apparently bad estimate for some of the trees, the p-values show good agreement with a uniform distribution. In other words, by this measure, *all* the trees in TreeBASE are well matched by Ford's alpha model, with $\alpha = 0.25$. In Figure 4 of [5] a similar graph is shown for the AB model and a different shape statistic which also shows good agreement. The key point is that, as far as these statistical tests go, both models fit very well.

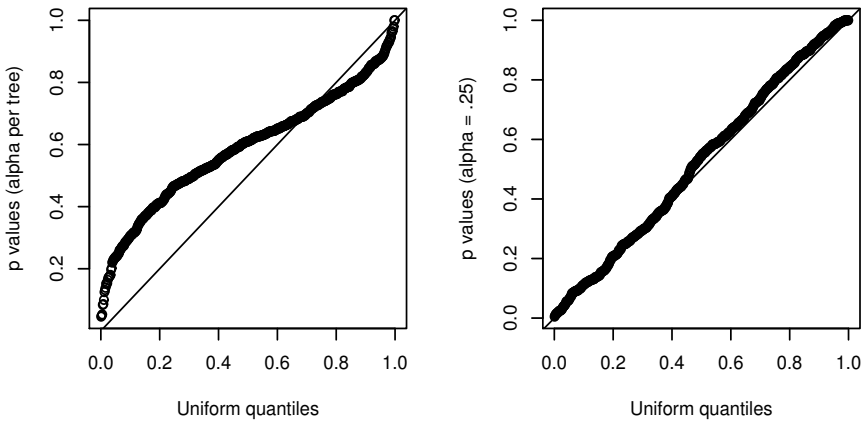


Figure 2: p-values for Colless statistic.

## 3.2 The maximum likelihood estimator of Blum and François

Another study [5] compared the TreeBASE trees to Aldous's beta-splitting model. They found very good agreement with the model, when $\beta = -1.0$, that is, the AB model. Their estimate for $\beta$ was -0.95, very close to -1.0. However, there are minor problems with their estimator and with their choice of data. I will leave discussion of the data until section 5. 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 $\hat{\beta}$ for the ERM, AB, and PDA models, based on 10000 simulated trees for each mean, for tree sizes 10, 20, 50, 100. Also note that any model which produces perfectly balanced trees (where all splits are as even as possible) with a nonzero probability will produce a distribution for $\hat{\beta}$ with infinite mean, since a perfectly balanced tree has a ML estimate for $\beta$ of infinity. I have followed [5] and counted any estimates above 2.0 as 2.0.

Figure 3 shows a scatterplot of ML estimates for beta of trees from TreeBASE (obtained as described in section 5), plotted against tree size. It is similar to Figure 2 of [5], but the data is somewhat different, and the algorithm for local regression may differ too. As well as the scatterplot, Figure 3 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 beta for trees made from a large number of simulations of the AB model. The main deviation of the TreeBASE trees from the $\beta = -1.0$ 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 in [5] 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 behaviour of the statistic is taken into account.

| | Tree size | | | |
|---|---|---|---|---|
| $\beta$ | 10 | 20 | 50 | 100 |
| -1.50 | -0.97 | -1.34 | -1.46 | -1.48 |
| -1.00 | -0.22 | -0.59 | -0.89 | -0.95 |
| 0.00 | 0.54 | 0.46 | 0.28 | 0.16 |

Table 1: Means of ML estimates of beta under AB model.

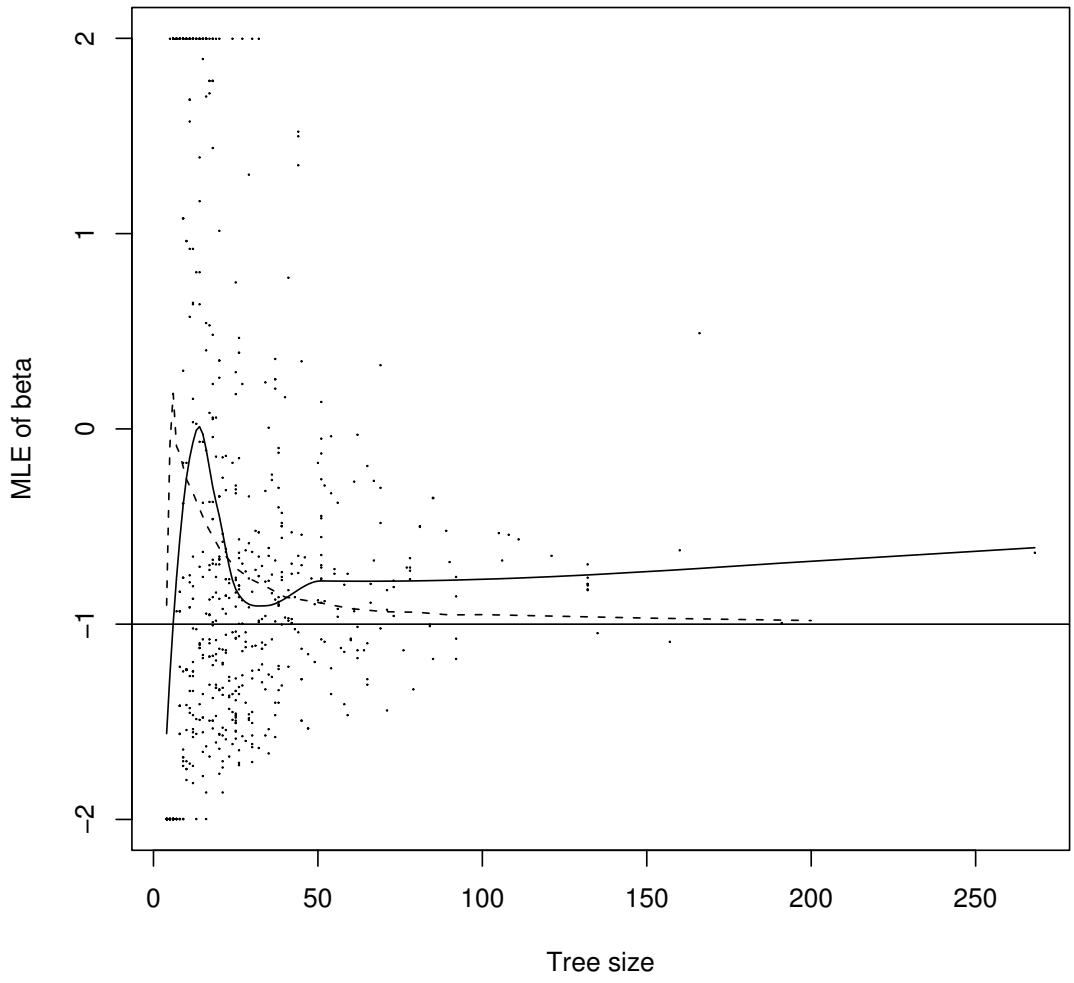


Figure 3: Scatter plot of ML estimates of beta for TreeBASE trees. The solid line shows a local regression curve for these trees. The dashed line shows means of ML estimates of beta predicted by the AB model.

So again the model actually fits the data better than was claimed, although in this case the mismatch was not large in the first place.

## 3.3 A new estimator

I now propose a new balance estimator based on Aldous's $\beta$, which reduces the problems evident with the ML estimates of beta. Instead of estimating $\beta$, 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.

| $B$ | $\beta$ | $\alpha$ | Tree type |
|---|---|---|---|
| 0 | -2 | 1 | comb |
| 1/3 | -3/2 | 1/2 | PDA |
| 1/2 | -1 | | AB |
| 2/3 | 0 | 0 | ERM |
| 1 | $\infty$ | | very balanced |

Table 2: Relation between $B$, Aldous's $\beta$, Ford's $\alpha$ and trees.

I use a Bayesian estimator for $B$ based on a uniform prior on $[0, 1]$ and squared-error loss function, which is therefore the posterior mean of $B$. I will denote this estimator as $\hat{B}$ and will use it as the basic measure of balance. I think an argument could be made for both the prior and loss function being reasonable in the context of phylogenetic trees, but I will content myself with a demonstration of its frequentist behaviour under the ERM, AB, PDA, models for various tree sizes. Table 3 shows mean values for the $\hat{B}$ for these models, based on 10000 simulated trees for each mean, for tree sizes 10, 20, 50, 100. Table 4 shows these values translated back into beta values for comparison with Table 1. It can be seen that $B$ gives accurate estimates for smaller tree sizes than the ML estimate of $\beta$.

| | Tree size | | | |
|---|---|---|---|---|
| $B$ | 10 | 20 | 50 | 100 |
| 0.333 | 0.43 | 0.37 | 0.34 | 0.34 |
| 0.5 | 0.54 | 0.52 | 0.51 | 0.50 |
| 0.667 | 0.62 | 0.66 | 0.67 | 0.67 |

Table 3: Means of $\hat{B}$ estimates.

| | Tree size | | | |
|---|---|---|---|---|
| $\beta$ | 10 | 20 | 50 | 100 |
| -1.50 | -1.25 | -1.41 | -1.48 | -1.49 |
| -1.00 | -0.83 | -0.90 | -0.97 | -0.99 |
| 0.00 | -0.35 | -0.06 | 0.03 | 0.02 |

Table 4: Beta values derived from $\bar{B}$.

I now turn to more subtle characteristics of tree shape which can distinguish between trees having the same overall amount of balance.

## 3.4 The effect of pruning on balance

Aldous [1] 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 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 and Aldous are both sampling consistent. However, it seems difficult to find models which are processes in time, produce the observed amount of imbalance, and are sampling consistent. (See 4.3 for a negative result.)

A more pragmatic goal is to measure the deviation from sampling consistency. In particular one can ask how much the balance changes when a tree is pruned. I define a particular statistic $D_{m,n}$ 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

$$D_{n,m} = \binom{n}{m}^{-1} \sum B_S \quad - \quad B_T$$

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 in 3.3 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.

### 3.5 Extreme splits

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.

## 4 Models

This section lists the models considered as candidates for fitting the data.

### 4.1 ERM, AB, PDA

These three are special cases of the beta-splitting model of Aldous, as previously described. ERM and PDA are included for comparisons. It is already known that they do not match the data.

## 4.2 The alpha model of Ford

Ford [9] described another family of models parameterized by a single parameter $\alpha$ taking values in $[0, 1]$ which interpolates between the ERM and PDA models. They 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. In Ford's model, balance decreases with $\alpha$, with $\alpha = 0$ giving the ERM model, and $\alpha = 1/2$ the PDA model. Unlike the beta-splitting model of Aldous it cannot produce trees more balanced than ERM. I will use this model with $\alpha = 0.25$, the value being chosen by fitting to the whole TreeBASE data set by consideration of the results previously described in section 3.1.

## 4.3 Blum-François models

I now turn to some models which have an interpretation as a time process using diversification rates. 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 behaviour 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 in [15]. 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 $p\kappa$ 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:

$$q(x, y) = \frac{1}{2}\binom{x + y - 2}{x - 1}(p^{x-1}(1 - p)^{y-1} + p^{y-1}(1 - p)^{x-1}). \tag{10}$$

Blum and François [5] 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 others are possible. I will call such models 'BF models'. The advantage of these models over more complex time processes is that the splitting distribution is readily calculated. The following result shows that if a BF model is sampling consistent, then it satisfies a fairly restrictive property.

**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).*

*Proof.* I will only prove this in the case where the distribution on [0,1] has a continuous density, $f(p)$. The proof for a discrete distribution is similar. It is shown in the Supplementary Material to [5] that under the above conditions,

$$q(x, y) = \frac{1}{2}\binom{x + y - 2}{x - 1}\int_0^1 (p^{x-1}(1 - p)^{y-1} + p^{y-1}(1 - p)^{x-1})f(p)dp. \tag{11}$$

(This results from integrating (10). Blum and François only consider the case of a Beta distribution, but the generalization is obvious.) Note that $f$ and therefore $q$ is symmetric (ie $q(x,y) = q(y,x)$). Now Proposition 41 of [9] 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.

$$(x+y+1-2q(1,x+y))q(x,y) = (x+1)q(x+1,y) + (y+1)q(x,y+1). \tag{12}$$

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

$$(2x+1-2q(1,2x))\binom{2x-2}{x-1}\int_0^1 p^{x-1}(1-p)^{x-1}f(p)dp =$$

$$(x+1)\binom{2x-1}{x}\int_0^1 (p^x(1-p)^{x-1} + p^{x-1}(1-p)^x)f(p)dp.$$

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

$$(2x+1-2q(1,2x))\binom{2x-2}{x-1} = (x+1)\binom{2x-1}{x}.$$

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

It may be noted that the ERM model satisfies this condition on $q(1,n-1)$. I think that one could show more about the models of the type considered in this proposition, but as will be shown in section 5.3, this single condition pretty much rules them out as models for real phylogenetic trees. Thus, sampling consistency will have to be abandoned for BF models.

## 4.4 Age dependent models

In the mathematical field of branching processes, a well studied extension of the constant rate birth-death process (ERM model) is that of age dependent processes [3]. They were first studied by Bellman and Harris in 1952 and are sometimes known as Bellman-Harris models. They have not received much attention as models for evolutionary trees, but have been recently considered in [10]. In such a process, each species behaves independently of the other species, and independently of its parent, but its lifetime is dependent on its age, that is, on the time since its birth. The process can include extinctions. (In fact the mathematical model can include any number of offspring.)

One can compare it to the ERM model and the BF models of sections 2.1 and 4.3. In the ERM 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, while the ERM model and age dependent models may do. Including extinctions in the ERM model does not affect balance, but in age dependent models, extinctions will affect balance in general.

Here is a more detailed description of the models considered here. A species lives for a time, then either goes extinct or produces exactly two descendant species. I will give the event of extinction or speciation the name 'termination', and I regard the life of a species as ending at termination: its descendants, if any, are two new species. 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 $t$, since 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 lasts forever. The function $f = -S'$ where $S' = \mathrm{d}S/\mathrm{d}t$ gives the density of lifetimes on $[0, \infty)$. The function $h = f/S = -S'/S$ gives the conditional rate of termination for a species which 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 $\lambda$).

# 5 Fitting models to the data

The 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 which has been deliberately added, and this of course affects the balance at the root. [5] 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 [9]. 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 4 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. This data was used for the scatter plot in Figure 3 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 artefacts arising from the choice of statistic. From table 3 it can be seen that $B$ shows very little bias for trees of this size. The same number of trees of the same sizes was generated for each of the models.

## 5.1 Balance

Figure 4 shows box plots of balance for **TB** which is derived from the TreeBASE data and the models. The details of the models are as follows. **ERM**, **AB**, and **PDA** are Aldous beta splitting models with $\beta = 0, -1, -1.5$ respectively. ERM and PDA have other interpretations as described earlier. **F25** is Ford's model with $\alpha = 0.25$. **KP01** is the Kirk-Patrick model with $p = 0.1$, so that its splitting probabilities are given by equation (10). **BF58** is the Blum-Francois model from [5] with parameter $\alpha = 0.58$, meaning that $p$ is chosen from a beta distribution on $[0, 1]$ with shape parameter $0.58$.

The others are age-dependent models with survival functions as follows. **AdC1** and **AdC9** use a $\chi^2$ density with one degree of freedom as the event density function. Equivalently it is a gamma

density with shape parameter 1/2. The survival function is thus 1 minus the cdf for this density, which can be expressed in terms of the cdf $\Phi$ for a standard normal as $S(t) = 2 - 2\Phi(\sqrt{t})$. **AdB1** and **AdB9** use a Burr density. The Burr distribution is a flexible distribution on $[0, \infty)$ with cdf $F(x) = (1 - x^{-c})^{-b/c}$ where $b$ and $c$ control the shape. The density is approximately $bx^{-b-1}$ for $x$ near zero, and approximately $bx^{-c-1}$ for large $x$. This means that the behaviour for small and large $x$ can be controlled independently by varying $b$ and $c$. Here $b = 0.4$ and $c = 2.5$ resulting in a survival function $S(t) = 1 - (1 - t^{-2.5})^{-.16}$. (Note that there are several distributions known as a 'Burr' distribution, and they can be parameterized in various ways.) **AdP1** and **AdP9** use a survival function with $S(t) = (1 + t)^{-1/2}$, corresponding to a Pareto density function $f(t) = (1/2)(1+t)^{-3/2}$. **AdE1** and **AdE9** use a mixture of exponentials such that $f(t) = (1/2)e^{-t} + (15/2)e^{-15t}$ and $S(t) = (1/2)e^{-t} + (1/2)e^{-15t}$. The final digit in each case indicates the ratio $\rho$ of the extinction rate to speciation rate. "1" means that $\rho$ is 0.1 (ie ten speciations for every extinction), and "9" means that $\rho$ is 0.9. I will call these 'low extinction' and 'high extinction' models.

Apart from the ERM and PDA models which are included for comparison, and the high extinction age dependent model AdP9, all the models match the data well. The AdP model is the one with the longest tail in it event density function, and it becomes very unbalanced when there are many extinctions. It is hardly surprising that the models can match the mean balance of the TreeBASE data since they all have a parameter or a function which can be adjusted to change the balance. 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. I cannot think of any plausible process which might 'censor' the data by removing, or tending to remove, extreme values of balance. I consider it remarkable that not one of the 140 TreeBASE trees shows a 'strange' value. It is also surprising that nearly all the models show such similar variance. Only the KP01 shows markedly smaller variance.

Age-dependent models behave differently for different extinction:speciation ratios. While AdP9 is much more imbalanced than AdP1, both AdC and AdE show, to a lesser extent, the opposite trend. 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.

## 5.2 Sampling consistency

I now turn to the secondary statistics relating to balance. Figure 5 shows box plot for the $D_{n,m}$ statistic measuring deviation from sampling consistency. The same set of trees of size at least 40 is used as for the balance measurements; the subsampled trees have sizes down to 20. $m$ is set to $\lfloor n/2 \rfloor$, an integer close to $n/2$. The TreeBASE data shows a reduction in balance when the trees are sampled. the reduction is not large and the explanation may lie in the way researchers tend to choose taxa rather than any evolutionary process. The ERM, PDA, AB and F25 models all show little deviation from sampling consistency, as expected from theory. The small deviationa are due to the finite sample size. The two BF models KP01 and BF58 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

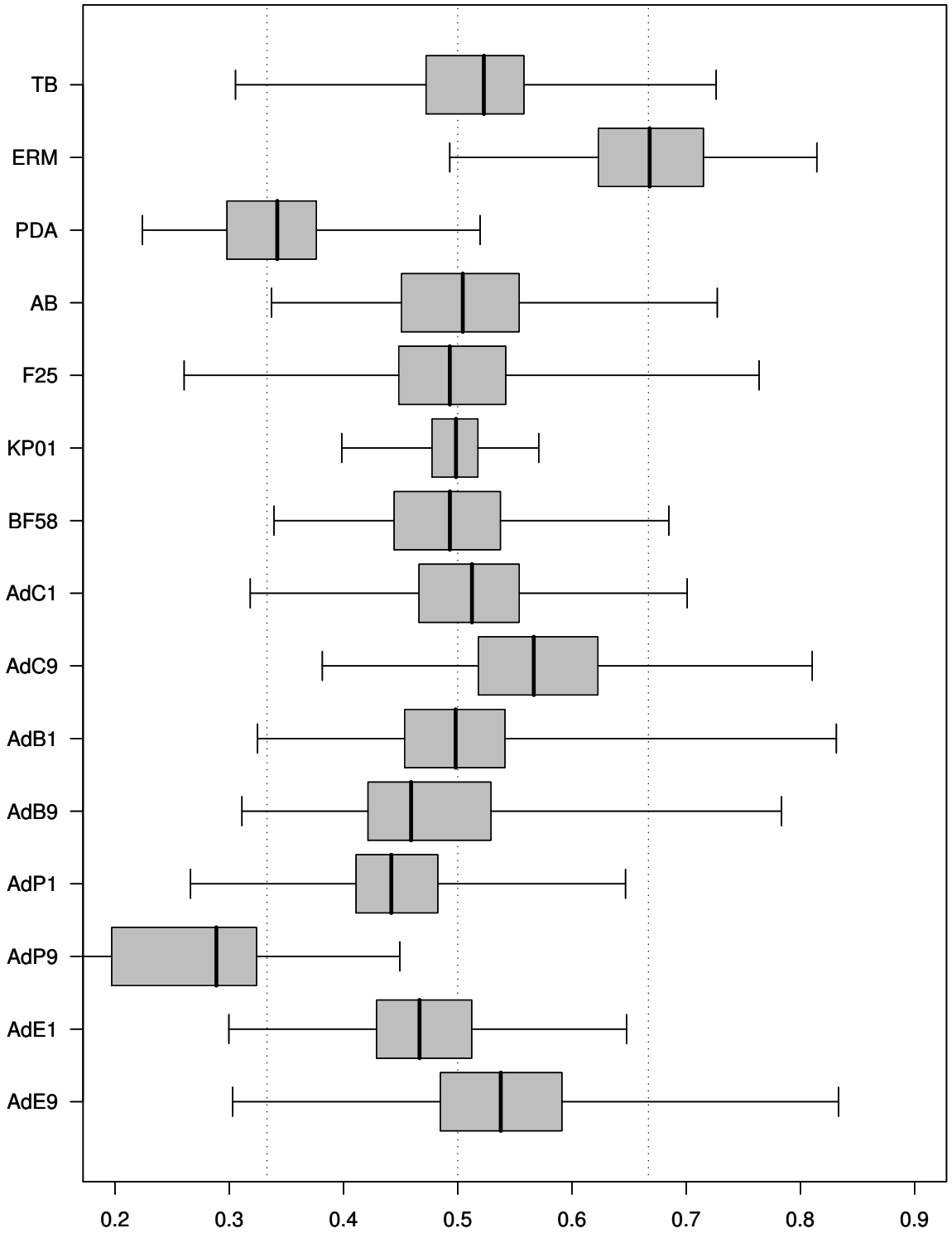


Figure 4: Box plots of balance for TreeBASE data and various models. The boxes show the first and third quartiles and the median values. The whiskers show minimum and maximum values. See text for details of the models.

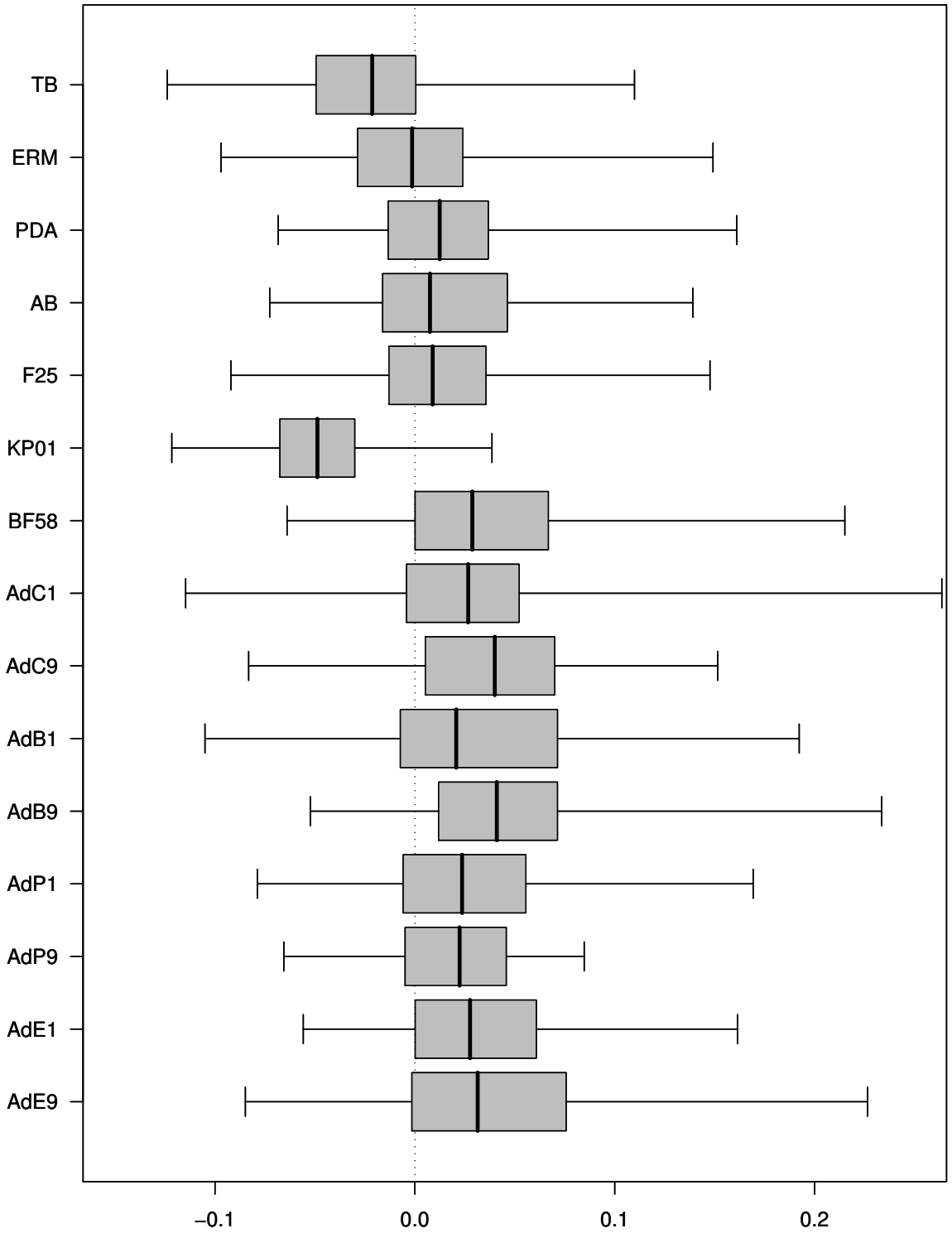


Figure 5: Box plots of $D_{n,\lfloor n/2 \rfloor}$ for TreeBASE data and various models. The boxes show the first and third quartiles and the median values. The whiskers show minimum and maximum values. See text for details of the models.

negative result of Proposition 2, this did not seem worth pursuing.

The age-dependent models all show somewhat increased balance when subsampled. This seems to be a general feature, and applies to several other event density functions not shown here.

## 5.3 Extreme splits

Figure 6 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. Since 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 TreeBASE data shows considerable proportions of extreme splits with no tendency to decrease as the tree size grows. Indeed the proportion increases as the tree size grows from 20-40 to 40-80, a fact which is not easy to explain except by chance. Remember that outgroups were effectively removed from the data by ignoring the root, and that these represent splits of internal nodes. The ERM tree has very few extreme splits: the theoretical number is $2/(n-1)$ for a split of $n$ tips, and the estimated values are close to this. This statistic alone is sufficient to rule out the ERM model. It also shows that the property proved in Proposition 2 for sampling consistent BF models is not seen in real data. As expected from equation (5) the PDA model has a high number of extreme splits, close to 0.5.

The AdP9 model has very high proportions of extreme splits, considerably more than the PDA model. The KP01 model does not produce enough extreme splits to even roughly approximate the data, and the AdC9 and AdE9 models are on the low side, especially for large tree sizes. Most of the other models produce proportions of extreme splits which are in the right general area, and it seems unwise to draw any firm conclusions based on the differences. As can be seen from the general pattern across the three plots, this statistic is not much affected by the size of the splits.

## 5.4 Where the models fail

Clearly the ERM model is too balanced and the PDA model is too unbalanced, as reported by others.

The models of Aldous and Ford with appropriate parameter values match the data well. Their main defect is that they are not processes in time, and provide no probabilities for node times. They do not include extinctions, and it is hard to see how they might be incorporated.

A model of BF type can match the data with a suitable choice of density on $[0, 1]$. Although they can be interpreted as a process in time, the mapping from 'tree time' to real time is not known. They do not include extinctions, and while they could be incorporated, the attractive mathematical tractability would be lost.

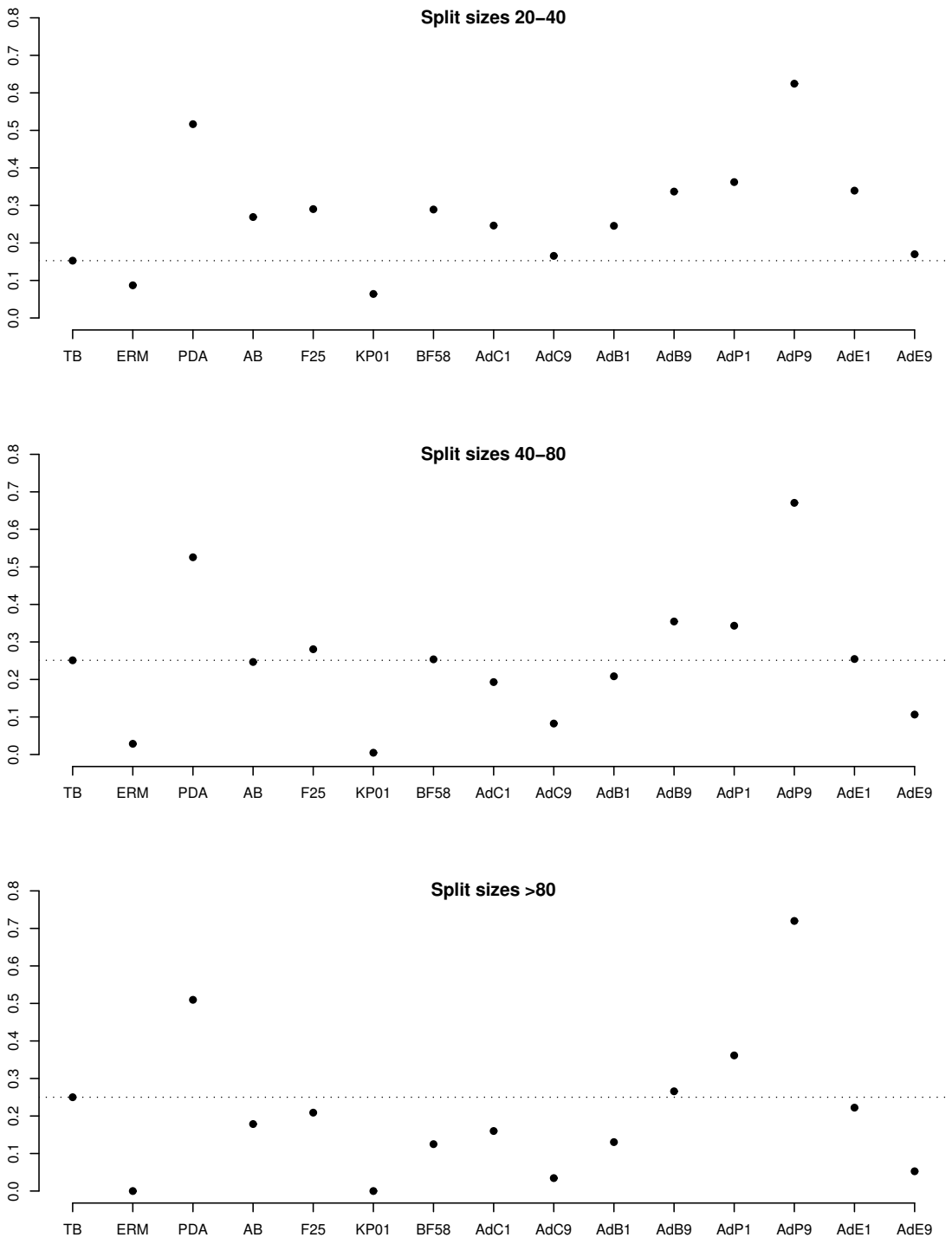


Figure 6: Proportions of extreme splits for TreeBASE data and various models, for three size ranges, 20-40, 40-80, and 80+. See text for details of the models.

The age-dependent models are the most realistic, being processes in real time with a natural way of including extinctions. They do not match the deviation from sampling consistency seen in the TreeBASE data, but this not seem sufficient to rule them out at this time. In other respects they seem good matches to the data, particularly the Burr distribution with its spike at zero and long tail.

# 6 Further discussion of age-dependent models

It may be noted that all the age-dependent models considered here have an 'L' shaped event density function. AdC has a spike near zero. AdP has a very thick tail and infinite mean. The other two have spikes near zero and long tails. An 'L' shaped event density function appears to be necessary to produce unbalanced trees. It is the case of course that 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 ([21], [17], [16]) so that there is a non-zero probability that a species lasts forever. These can be described as even more 'L' shaped than the ones used here.

The theory of branching processes [3], Chapter IV, describes the long term behaviour of age-dependent models. If $f$ is the event density function, and $m$ is the expected number of offspring at the termination of a species, then the solution of

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

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 root. I am assuming here that $m > 1$, that is, that there are more speciations than extinctions. In terms of the ratio $\rho$ used earlier, and set to $.1$ and $.9$ in the simulations, $m = 2/(1 + \rho)$.

Understanding the behaviour of these models is quite difficult, especially in the presence of extinctions, and they can seem counter-intuitive 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 ([3], section IV.5). It seems reasonable to suppose that large values of high order moments would be associated with large amounts of imbalance.

## 6.1 How realistic are they in other ways?

Although the age dependent models require specifying a event density function, which at first sight seems to make them infinitely flexible, my impression is that the requirement to match the data, particularly the balance over a wide range of extinction rates, may be enough to pin down the general shape of the event density function quite well. One can then ask if such an event density function, or the model it implies, makes sense 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 not-too-tiny probability is assigned to a 'flurry' of quick speciations, and a not-too-tiny 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 to, say, a standard birth-death model which 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, while 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 has been suggested that when a species is young, there is likely to be an increased chance of further speciations [22]. This would apply to the descendants of a species which is able to occupy a new niche due to a key innovation, or to the descendants of a species which invades a new region. It may also be that geological processes which fragment ranges (eg changing sea levels) will often continue to fragment ranges, resulting in further speciations quickly after the first. It also seems likely that the early period of a species is a dangerous 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.

It seems that long tails are needed to match the TreeBASE 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 5 shows some values for three event density functions: the exponential which produces the standard birth-death model with constant rates; the $\chi^2$ with one degree of freedom; and the Burr distribution with the same parameters as used earlier.

| | Multiple of median | | |
|---|---|---|---|
| model | 10 | 20 | 30 |
| exp | 1 in 1000 | 1 in $10^6$ | 1 in $10^9$ |
| $\chi^2$ | 1 in 30 | 1 in 400 | 1 in 4500 |
| Burr | 1 in 30 | 1 in 150 | 1 in 400 |

Table 5: Approximate probabilities of a species lasting much longer than median.

I have not compared these to any data, but it does seem that the exponential distribution makes very long-lived species (eg 'living fossils') far too unlikely. When choosing a prior for a Bayesian analysis it is generally improves robustness to use ones with long tails [4], so the age-dependent models seem more appropriate.

# 7 Notes on algorithms used

The programs for generating sample trees are written in C++. The graphs and tables were produced using R [18]. Sampling from the models with splitting distributions is straightforward. The age-

dependent models are more difficult [20].

## 7.1 Generating trees from age dependent process

I assume that the survival function $S$ is given. Suppose that at a particular time $x$, there are $n$ species with birth times $b_i$ $(1 \leq i \leq n)$. Let $d_i$ denote the time that species $i$ terminates. From the usual formula for conditional probabilities, it follows that

$$\begin{aligned} \Pr(d_i \geq y \mid d_i \geq x) &= \frac{\Pr(d_i \geq y \ \wedge \ d_i \geq x)}{\Pr(d_i \geq x)} \\ &= \frac{\Pr(d_i \geq y)}{\Pr(d_i \geq x)} \\ &= \frac{S(y - b_i)}{S(x - b_i)}. \end{aligned}$$

So the probability that no species terminate before $y$ is

$$S^\star(y) = \Pr(\forall i : d_i \geq y \mid \forall i : d_i \geq x) = \prod_i \frac{S(y - b_i)}{S(x - b_i)}. \tag{13}$$

since the species behave independently. $S^\star()$ is thus a survival function for all species existing at time $x$. In order to sample the time to the next event, a uniform random number $u$ in $[0, 1]$ is drawn, and the equation $S^\star(t) = u$ solved for $t$ using binary search. Since I am assuming that $S(t) \to 0$ as $t \to \infty$, a time $T$ can be chosen such that the probability of survival beyond $T$ is small enough to ignore, so $[0, T]$ provides the starting point for the binary search.

Having sampled a time $t$ for a new event, the species which terminates must be chosen, and this is done using the hazard function $h = S'/S$. The termination rate of species $i$ at time $t$ is $h(t - b_i)$, so a species is chosen in proportion to these values.

Finally, a speciation or extinction is chosen according to the model. By definition of the model, the ratio between these is a constant.

## 7.2 Conditioning on extant number of tips

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 [20] 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 one another.

# References


[1] Aldous D J, 1996. "Probability distributions on cladograms", Pages 1-18 in Random Discrete Structures (D. J. Aldous, and R. Pemantle, eds.) Springer IMA Volumes Math. Appl. 76.

[2] Aldous D J, 2001. "Stochastic models and descriptive statistics for phylogenetic trees, from Yule to Today", Stat. Sci. 16:23-34.

[3] Athreya K B, Ney P E, *Branching Processes*, Springer-Verlag 1972, Dover 2004.

[4] Berger J O, *Statistical Decision Theory and Bayesian Analysis, 2nd Edition*, Springer, 1985.

[5] Blum M G B, François O, 2007. "Which Random Processes Describe the Tree of Life? A Large-Scale Study of Phylogenetic Tree Imbalance", Syst Biol. 2006 Aug; 55(4): 685-91

[6] Drummond A J, Rambaut A, 2007. "BEAST: Bayesian evolutionary analysis by sampling trees", BMC Evolutionary Biology 7:214

[7] Drummond A J, Ho S Y W, Phillips M J, Rambaut A, 2006. "Relaxed Phylogenetics and Dating with Confidence", PLoS Biology 4, e88.

[8] Felsenstein J, *Inferring Phylogenies*, Sinauer Associates, Inc., 2004

[9] Ford D, 2005. "Probabilities on cladogram: Introduction to the alpha model", Arxiv preprint math-0511246, november 2005.

[10] Gernhard T, Hartmann K, Steel M, 2008. "Stochastic properties of generalised Yule models, with biodiversity applications", J. Math. Biol. 57:713-735

[11] Gnedenko B V, Kolmogorov A N, *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley, 1954.

[12] Huelsenbeck J P, Ronquist F, "MRBAYES: Bayesian inference of phylogenetic trees", Bioinformatics 2001, 17:754-755.

[13] Peter McCullagh, Jim Pitman, Matthias Winkel. "Gibbs Fragmentation trees". arXiv:0704.0945v1

[14] Mooers A, Heard S B, 1997. "Inferring evolutionary process from phy-logenetic tree shape", Quart. Rev. Biol. 72:31-54.

[15] Kirkpatrick M, Slatkin M, 1993. "Searching for evolutionary patterns in the shape of a phylogenetic tree". Evolution 47:1171-1181.

[16] Kontoleon N, 2006. "The Markovian Binary Tree: A Model of the Macroevolutionary Process", PhD Thesis, The University of Adelaide.

[17] Pinelis I, 2003. “Evolutionary models of phylogenetic trees”, Proc. R. Soc. Lond. B 270, 1425-1431

[18] R Development Core Team, 2008. “R: A language and environment for statistical computing”. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[19] Sanderson M J, Donoghue M J, Piel W, Eriksson T, 1994. “TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life”. Am. J. Bot. 81:183-189,

[20] Stadler, T, 2008. “Evolving Trees - Models for Speciation and Extinction in Phylogenetics”, PhD Thesis, Technische Universität München, Zentrum Mathematik

[21] Steel M, McKenzie A, 2001. “The ‘shape’ of phylogenies under simple random speciation models”. Proceedings of Biological Evolution and Statistical Physics conference, Max Plank Institute (Dresden, 2000), (Book Chapter P. 165-185).

[22] Steel M, McKenzie A, 2001. “Properties of phylogenetic trees generated by Yule-type speciation models”. Math Biosci. 2001 Mar; 170(1):91-112.