A Study on the Empirical Support for Prior Distributions on Phylogenetic Tree Topologies

Lin M. Himmelmann[*] and Dirk Metzler[†]
Institut für Informatik
Johann Wolfgang Goethe-Universität
Postfach 11 19 32
D-60054 Frankfurt am Main

Abstract: There are several models for the evolutionary process forming a species tree. We examine the Birth-and-Death model (BDM), the Proportional-to-Distinguishable Arrangements (PDA) model, the Kirkpatrick and Slatkin (KS) model, the Beta-Splitting (BS) model and a model where birth rates evolve according to a Geometric Brownian Motion Process (GBM). For testing and calibrating the models, we evaluate tree topologies from TreeBASE and a large tree provided by the Tree of Life project. In a simulation we compare the distribution of tree topologies generated by the models with tree topologies of observed trees. For describing the distribution of topologies we use the tree imbalance statistics B_1, Colless'C and Shao and Sokal's N, and calculate the maximum-likelihood estimate of β from the BS model. Further we explore the splitting pattern of the generated trees. From the observed trees we show that trees generated by the BDM are too balanced and trees generated by the PDA model are too imbalanced. The BS model with $\beta = -1$, the KS model with ratio $1:2$ and an adjusted GBM model represent better fitted models for reproducing the imbalance in observed tree topologies.

1 Introduction

A frequently used assumption in evolutionary biology is that a set of species can be related by a common tree. This implies that the species under consideration must have a unique common ancestor, representing the root of the tree. From this ancestor, branching events in the tree correspond to the speciation of a lineage in consecutive lineages. Several models are suggested to imitate the macro-evolutionary process forming such a species tree ([Pin03]). These models can be used as priors in Bayesian methods for reconstructing phylogenetic trees from sequence data. Even if no prior distribution on the space of trees is explicitly modeled, typically a uniform distribution on trees is implicitly assumed, which might be a cause of unwanted bias. Furthermore there is a need to have an appropriate null model in statistical tests ([CM02]).

Two often used models for the topology of a tree, are the Birth-and-Death model (BDM)

[*] linhi@cs.uni-frankfurt.de
[†] metzler@cs.uni-frankfurt.de

with equal rates and the Proportional-to-Distinguishable Arrangements (PDA) model. The BDM model ([Ken48]) is a branching process. Each lineage l undergoes speciation of rate λ_l and extinction of rate μ_l. Here we restrict on a special case of the general Birth-and-Death model, sometimes called Equal-Rates Markov model, where the birth rate λ and death rate μ are constant among lineages and in time. The PDA model was introduced by [Ros78]. Every leaf-labeled tree topology has the same probability under the PDA model. A class of tree-generating processes inducing the same probability distribution on tree topologies as the PDA model is discussed in [MS01] and [Pin03]. The topology of a tree is the unlabeled topological branching pattern, excluding temporal information, also called shape. It is repeatedly reported that the distribution of tree topologies generated by the BDM tends towards too balanced tree topologies and the PDA tends towards too unbalanced tree topologies ([Hea96], [Ald01] and [Pin03]). To correct this, a couple of models were proposed to reproduce the typical balance in observed trees ([KS93] and [Ald96]).

There are only a few studies taking into account the huge amount of todays reconstructed trees. Blum and Francois ([BF06]) use the collection of tree topologies stored in the TreeBASE, which is a public online database for phylogenetic trees[1] ([SDPE94] and [Mor96]). They compare the observed topologies with tree topologies resolved by the BDM, PDA and Beta-Splitting (BS) model with parameter $\beta = -1$ ([Ald96]), using a tree topology statistic whose P-values should be normally distributed under the correct tree generating mechanism. Their study supports that the BS model with $\beta = -1$ shows a good fit to the observed trees. Ford ([For05]) introduces a tree topology generating model with one parameter, called alpha model, which encompasses the BDM and the PDA model for appropriate parameters. From the TreeBASE trees he estimates the parameter for the observed trees, which lies between the parameters of the BDM and the PDA model. Matsen ([Mat06]) presents a genetic algorithm to receive an optimal statistic over a class of tree shape statistics to differentiate between the observed trees in the TreeBASE and the BS model with $\beta = -1$. He finds that the BS model with $\beta = -1$ is highly significant rejected.

For our study we also use tree topologies in TreeBASE. Additionaly we use subtrees from the large tree given by the Tree of Life project[2]. This project aims to provide information about all living organisms and their evolutionary history ([MS06]).

2 Tree Imbalance Statistics

Every tree-generating mechanism produces tree topologies with a certain distribution. Because the space of tree topologies grows overexponentially with the number of leaves ([Hol98]), it is customary to use only specific features of the tree topologies and evaluate the distribution of this feature under the tree generating process. The most widely used feature for describing the topology of a tree is its imbalance. The balance of a tree topology can be quantified by some real function on the space of tree topologies, deriving a

[1] see http://www.treebase.org

[2] see http://www.tolweb.org

statistic for the underlying tree-generating model. The imbalance statistics used here were chosen with regard of their explanatory power and their correlation structure, examined in a simulation study by Agapow and Purvis [AP02].

A rooted binary tree topology with n leaves has $(n-1)$ inner nodes. The size of a tree or subtree rooted at some inner node i is the number of leaves that originate from the inner node i. Colless'C ([Col82]) is defined as

$$\bar{C} := \frac{2}{(n-1)(n-2)} \sum_{i=1}^{n-1} |r_i - l_i|, \tag{1}$$

where r_i is the number of leaves in the right subtree and l_i is the number of leaves in the left subtree rooted at the inner node i. The more unbalanced the tree is, the bigger is the difference of the leave size in the daughter species at the inner nodes and the statistic yields higher values. The statistic is normalised by $\frac{2}{(n-1)(n-2)}$ in order to get values between 0 and 1.

The measure B_1 ([SS90]) is defind as

$$B_1 := \sum_{i=1}^{n-2} \frac{1}{M_i}, \tag{2}$$

where M_i is the maximal number of edges from the inner node i to the leaves of the subtree rooted in i. The summation is taken over all inner nodes, except for the root. This definition is taken from [SS90] but it differs from the definition in [AP02], where summation is taken over all inner nodes i with subtree size greater than 3. Since the most balanced tree minimises the maximal number of nodes to traverse from the root to the leaves, this statistic yields higher values as the tree becomes more balanced.

Shao and Sokal's N ([SS90]) gives the mean path length from the leaves to the root of the tree. It is defined as

$$\bar{N} := \frac{1}{n} \sum_{i=1}^{n} N_i, \tag{3}$$

where N_i is the number of edges along the path connecting the leave i with the root. The interpretation is similary to B_1, however in this case the contributive part is in the nominator, thus a higher value represents more imbalance in a tree.

Finally, we calculate the maximum-likelihood value for the parameter β in Aldous's BS model ([Ald96]). The BS model provides a one-parameter family of probability distributions $q_{\beta,n}$ on $\{1, 2, \ldots, n-1\} \subset \mathbb{N}$. Given a node v in a binary tree from which n leaves descend, $q_{\beta,n}(i)$ is the probability that v's outgoing edges split the set of descending leaves into subsets of i and $(n-i)$ leaves. The split is denoted as $\{i|(n-1)\}$. The parameter β is in the range between -2 and ∞. A higher value of β will generate more balanced tree topologies. Aldous [Ald96] shows that for $\beta = 0$ the BS model generates tree topologies with the same distribution as the BDM, and for $\beta = -1.5$ it generates tree topologies with

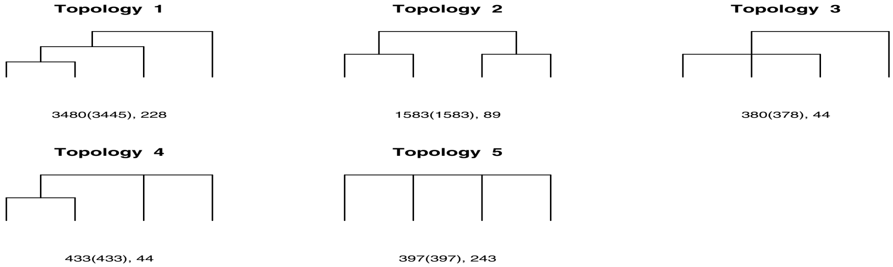

Figure 1: Topologies of size 4. In the caption are the counts from TreeBASE, outgroup-corrected values from TreeBASE in brackets, and counts from the Tree of Life.

the same distribution as the PDA model. As suggested in [BF06] and [Ald96] we consider the BS model with fixed parameter $\beta = -1$.

3 Evaluation

We scan all entries in the TreeBASE, included in May 2007. Inner nodes of degree two are removed in a preprocessing step. We create two data sets containing all obtained trees. In the first set, we automatically remove from every tree a possible outgroup, if the root node has a binary split and one of the subtrees contains only one leaf. In that case, the subtree with one leaf is removed. Because this procedure may incorporate bias, the second set is not outgroup-corrected, for control. From each dataset we respectively extract all subtree topologies of sizes 4 and 5. Similarly we use the current XML structure from the Tree of Life project, encoding the current Tree of Life (May 2007), and extract all subtrees of sizes 4 and 5. We use only those subtrees, with the highest confidence level, where all leave nodes were marked as leaves and which contained no extinct subtree or species. The resulting counts for the multifurcating tree topologies are given in Fig.1 and Fig.2.

In the following we examine the fraction of binary tree topologies to compare and calibrate our tree topology generating models. Omitting multifurcating tree topologies is not satisfactory for several reasons. First, we do not want to waste information in the data, in order to reduce variability in the estimates. Second, multifurcations often reflect uncertainty in reconstructing the correct evolutionary history. From this point of view it should in principle be possible to resolve most multifurcations by binary splits if enough information is available and if branch lengths can be arbitrarily small. Third, by ignoring multifurcating subtrees we get a biased answer, because in general we will not get a representative sample by picking bifurcating subtrees. The assignment of mutlifurcating trees to their binary counterparts will influence the distribution of topologies.

We count binary topologies of sizes 4 and 5, and in a second step we assign the multifurcating tree topologies by resolving the multifurcations under the hypotheses of the BDM and the PDA model. The resulting fractions did not differ much whether we outgroup-corrected the tree topologies or not. The fraction of topologies of size 4 from the outgroup-

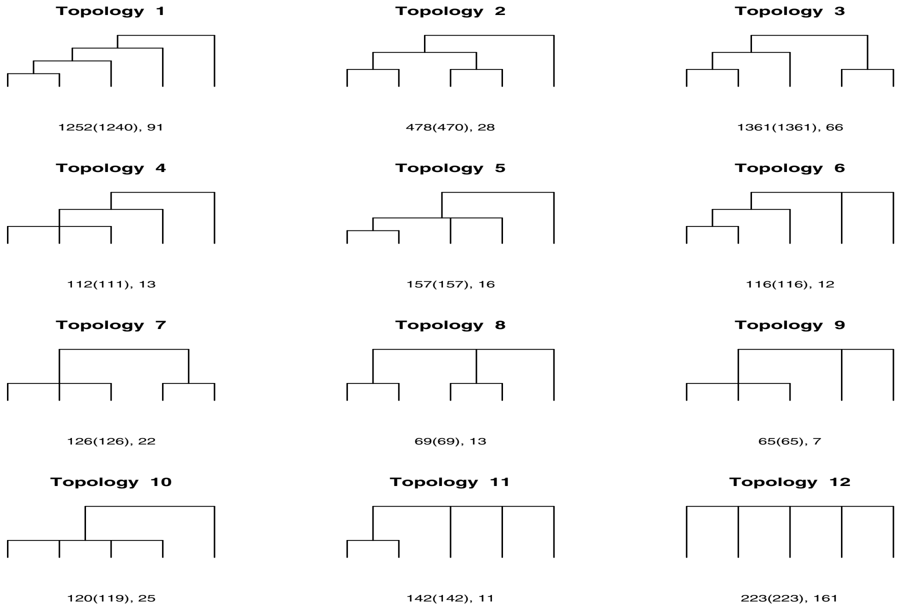

Figure 2: Topologies of size 5. In the caption are the counts from TreeBASE, outgroup-corrected values from TreeBASE in brackets, and counts from the Tree of Life.

corrected trees and the Tree of Life are given in Tab.1, with multifurcating tree topologies resolved by the BDM. Note that there is only one multifurcating tree topology of size 4 which is informative, and it is equivalent resolving it by the BDM or PDA model. Tab.2 shows the fractions of topologies of size 5 from the outgroup-corrected trees and the Tree of Life, with multifurcating tree topologies resolved by the BDM and the PDA model.

4 Calibration and Simulation Results

The BDM, PDA and BS model do not need to be adjusted, because there are no free parameters. For the calibration of the other models we use the fractions of observed trees of sizes 4 and 5. In the Kirkpatrick and Slatkin (KS) model ([KS93]) branching events in lineages may occur at different rates. Every time a branching event appears, the rate of branching is inherited to the daughter lineages with a deterministic assignment, such that the sum of the rates in the offspring lineages is twice the rate of the parent lineage. The only parameter in this model controls the partitioning of the birth rate to the offspring lineages. From Tab.1 we observe an approximate proportion of unbalanced to balanced trees of size 4 with ratio $7:3$ in the TreeBASE. The KS model was calibrated with this fraction, resulting in a ratio of rates in the daughter lineages of $1:2$.

We calibrate a model where birth and death rates for each lineage evolve according to a geometric Brownian Motion process ([Hea96]), and call it GBM model. A geometric Brown-

	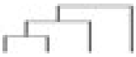	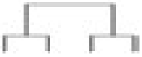
TreeBASE	0.707	0.293
Tree of Life	0.754	0.247
BDM	$\frac{2}{3} = 0.667$	$\frac{1}{3} = 0.333$
Beta-Splitting ($\beta = -1$)	$\frac{8}{11} = 0.727$	$\frac{3}{11} = 0.273$
PDA	$\frac{4}{5} = 0.800$	$\frac{1}{5} = 0.200$
Kirkpatrick Slatkin (1 : 2)	$\frac{7}{10} = 0.700$	$\frac{3}{10} = 0.300$
3/4 BDM + 1/4 PDA	0.700	0.300
2/3 BDM + 1/3 PDA	0.711	0.289
1/3 BDM + 2/3 PDA	0.756	0.244
GBM model (σ=2) sim.	0.778	0.222

Table 1: Fraction of tree topologies for size 4.

	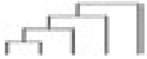	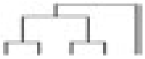	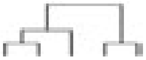
TreeBASE (BDM resolved)	0.422	0.153	0.426
TreeBASE (PDA resolved)	0.441	0.143	0.416
Tree of Life (BDM res.)	0.478	0.157	0.365
Tree of Life (PDA res.)	0.510	0.138	0.352
BDM	$\frac{2}{6} = 0.333$	$\frac{1}{6} = 0.167$	$\frac{3}{6} = 0.500$
Beta-Splitting ($\beta = -1$)	$\frac{48}{110} = 0.436$	$\frac{18}{110} = 0.164$	$\frac{44}{110} = 0.400$
PDA	$\frac{4}{7} = 0.571$	$\frac{1}{7} = 0.143$	$\frac{2}{7} = 0.286$
Kirkpatrick Slatkin (1 : 2)	0.447	0.195	0.357
3/4 BDM + 1/4 PDA	0.393	0.161	0.446
2/3 BDM + 1/3 PDA	0.417	0.159	0.429
1/3 BDM + 2/3 PDA	0.492	0.151	0.357
GBM model (σ=2) sim.	0.542	0.163	0.296

Table 2: Fraction of tree topologies for size 5.

ian Motion has continuous paths which evolve according to $g(t)=\exp((\mu-\sigma^2/2)t+\sigma B_t)$, where B_t is a standard Brownian Motion process, $\mu>0$ is the mean value, and $\sigma>0$ controls the volatility of the process. We only consider a process with death rate $=0$, which is started with an initial birth rate $=1$, and mean value μ fixed to be 1. With $\sigma=0$ the model has the same distribution on tree topologies as the BDM. To adjust the GBM model we simulate small tree topologies with different volatility parameters σ. By comparison with the observed fractions in Tab.1 and Tab.2 it becomes obvious that the trees are too balanced for $\sigma<1$ and too imbalanced for $\sigma>4$. We calibrate the model with $\sigma=2$.

Finally we model a mixture (Mix) of the BDM and the PDA model, where we determine one parameter to specify the probability of choosing between the BDM and the PDA model in advance. We adjust three different Mix models for comparison by fitting the fractions given in Tab.1 and Tab.2.

We perform a goodness of fit analysis with the adjusted models. Using the models we simulate trees with the same size as all binary trees in the TreeBASE. For every set of trees τ we evaluate B_1, Colless'C and Shao and Sokal's N and estimate the maximum-likelihood estimator of β in the BS model. Because of the dependence of the imbalance statistics on the tree size, we generate for each model and each statistic a plot in dependence of the size of the leaves. The outgroup-correction has no remarkable influence on the results. The plots without outgroup-correction are given in the supplementary material ([HM07]).

For a visual inspection among the models, we compare the quantiles of B_1, Colless'C and Shao and Sokal's N, and β for the simulated trees against the corresponding quantiles for the TreeBASE trees. For all statistics, all models lie between the BDM and the PDA model, where the BDM always generates too balanced trees and the PDA model generates too imbalanced trees. The Q-Q plot for Colless'C is shown in Fig.3, the other statistics are shown in the supplementary material ([HM07]).

For a direct comparison between the trees generated by the adjusted models and the TreeBASE trees, we extract for every inner node v, the size of the subtree originating from v and the size of its greater daughter subtree. From these data we estimate, separately for every set of trees τ, the distribution of possible splits $\hat{s}_n^\tau$ on $\{[\frac{n}{2}],\ldots,(n-1)\}$ for $n=2,\ldots,100$, with $[x]:=\min\{k\in\mathbb{N}_0 : k\leq x\}$. $\hat{s}_n^\tau(i)$ is the empirical probability of observing the split $\{i|(n-i)\}$ or the split $\{(n-i)|i\}$ at an inner node which gives rise to a subtree of size n. For every τ and $n\in\mathbb{N}$ define I_n^τ to be the expected number of splits in the path from the root of a tree of size n towards the leaves, if always the biggest subtree is chosen. I_n^τ can be calculated by the following recursion: $I_1^\tau=0$, $I_2^\tau=1$ and $I_n^\tau=\sum_{i=[\frac{n}{2}]}^{n-1}\hat{s}_n^\tau(i)(1+I_i^\tau)=1+\sum_{i=[\frac{n}{2}]}^{n-1}\hat{s}_n^\tau(i)I_i^\tau$, for $n\geq 3$. In Fig.4 the values for I_n^τ are plotted for each set of trees. The KSM, GBM and BS model show good accordance with the observed trees.

We build from each set of trees τ a 100×100-Matrix M^τ containing the empirical distributions $\hat{s}_n^\tau$. The entries of the matrix M^τ are defined to be $M_{n,i}^\tau:=\hat{s}_n^\tau(i)$, indexed by the number of leaves n and all possible bigger subtrees $i\in\{[\frac{n}{2}],\ldots,(n-1)\}$. The other entries in the matrix M^τ are set equal to 0. We calculate the L_1-Norm ($||X||_{L_1}=\sum_{i,j}|x_{ij}|$) of the difference between the submatrices of the first $n\times n$ entries from the simulated models and the observed trees, for $n=3,\ldots,100$. The resulting values for the models are

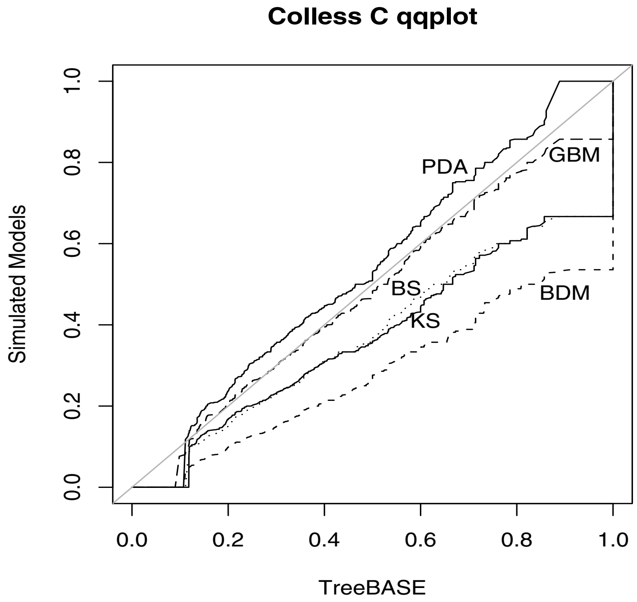

Figure 3: Q-Q plot Colless'C (see text for explanation).

shown in the supplementary material ([HM07]). Again the KSM, GBM and BS model show good accordance with the observed trees.

5 Conclusion

We explored the TreeBASE database and a large tree provided by the Tree of Life project, to supply tree topologies of size 4 and 5 for the calibration and testing of macro-evolutionary models. The observed distribution of small tree topologies, the evaluation of imbalance statistics and the splitting pattern comparison, indicates that the BDM generates too balanced tree topologies and the PDA model generates too unbalanced tree topologies. The imbalance of a typical tree lies between these two standard models. This observation agrees with those of [Hea96], [Ald01], [Pin03] and [BF06]. Our simulation confirms the good fitting of the BS model with $\beta = -1$, first supposed by Aldous [Ald01] and supported by the study of Blum and Francois [BF06]. The simulation of tree topologies with the KSM with ratio $1:2$ and the GBM model with $\sigma = 2$ produce more reasonable tree topologies than the BDM and PDA model. The splitting pattern of the adjusted BS, KSM and GBM model shows good consistence with the observed imbalance in trees. In the comparison of the statistics these models are all located between the BDM and PDA model.

If we restrict on markovian branching processes ([Ald96]) as an adequate model describ-

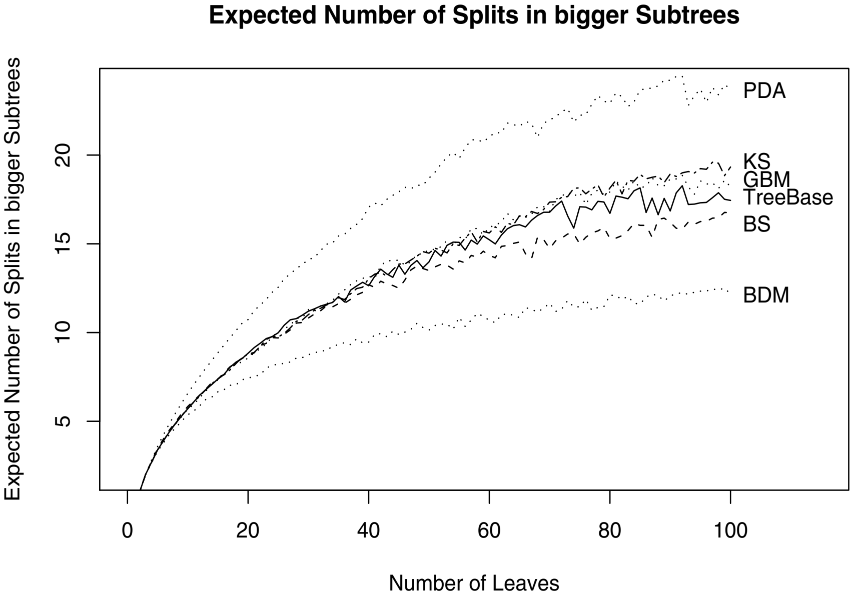

Figure 4: Expected Number of Splits in bigger Subtrees (see text for explanation).

ing the splitting structure of a tree, the BS model with $\beta = -1$ seems to be an appropriate candidate, even if Matsen [Mat06] shows significant discrepancies from this model and the TreeBASE trees. It would be interesting to investigate if the observed trees bear any evidence for violating the markovian assumption. As more and more trees with reconstructed branch lengths are getting available, it should be possible to create and validate better fitted models by incorporating temporal information.

Acknowledgment

We thank Johannes Bergs for converting the Tree of Life into newick format and an anonymous referee for helpful comments and hints to literature.

References

[Ald96] David Aldous. Probability distributions on cladograms. In David Aldous and R. Pemantle, editors, *The IMA Volumes in Mathematics and its Applications*, volume 76 of *Random discrete structures*, pages 1–18. Springer Verlag, 1996.

[Ald01] David Aldous. Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today. *Statistical Science*, 16:23–34, 2001.

[AP02] Paul-Michael Agapow and Andy Purvis. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Systematic Biology*, 51(6):866–872, 2002.

[BF06] Michael G. B. Blum and Olivier Francois. Which Random Processes Describe the Tree of Life? A Large-Scale Study of Phylogenetic Tree Imbalance. *Systematic Biology*, 55(4):685–691, August 2006.

[CM02] Kai M. A. Chan and Brian R. Moore. Whole tree methods for detecting differential diversification rates. *Systematic Biology*, 51(2):855–865, 2002.

[Col82] D. H. Colless. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*, 31:100–104, 1982.

[For05] Daniel J. Ford. Probabilities on cladograms: introduction to the alpha model. *Preprint: arXiv: math.PR/0511246*, 2005.

[Hea96] Stephen B. Heard. Patterns in Phylogenetic Tree Balance with Variable and Evolving Speciation Rates. *Evolution*, 50(6):2141–2148, 1996.

[HM07] Lin M. Himmelmann and Dirk Metzler. Supplementary Material for A Study on the Empirical Support for Prior Distributions on Phylogenetic Tree Topologies. *http://www.informatik.uni-frankfurt.de/∼linhi/treeevaluation*, 2007.

[Hol98] Susan Holmes. Phylogenies: An Overview. In S. Geisser and B. Halloran, editors, *IMA Volumes in mathematics and its applications*, volume 112 of *Statistics in Genetics*, pages 81–119. Springer Verlag, New York, 1998. STRESS - Indizee, S.105.

[Ken48] David G. Kendall. On the generalized Birth-and-Death process. *The annals of mathematical statistics*, 19(1):1–15, 1948.

[KS93] Mark Kirkpatrick and Montgomery Slatkin. Searching for evolutionary pattern in the shape of a phylogenetic tree. *Evolution*, 47:1171–1181, 1993.

[Mat06] Frederick A. Matsen. Optimization over a class of tree shape statistics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2006.

[Mor96] V. Morell. TreeBASE: the roots of phylogeny. *Science*, 273(5275):568–570, 1996.

[MS01] Andy McKenzie and Mike A. Steel. Properties of phylogenetic trees generated by yule-type speciation models. *Mathematical Bioscience*, 170:91–112, 2001.

[MS06] D.R. Maddison and K.-S. Schulz. The Tree of Life Web Project. Internet address: http://tolweb.org, 1996–2006.

[Pin03] Iosif Pinelis. Evolutionary models of phylogenetic trees. *Proceedings of the royal society London B*, 270:1425–1431, 2003.

[Ros78] Donn E. Rosen. Vicariant Patterns and Historical Explanation in Biogeography. *Systematic Zoology*, 87(2):159–188, 1978.

[SDPE94] M. J. Sanderson, M. J. Donoghue, W. H. Piel, and T. Eriksson. TreeBASE: A prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *American Journal of Botany*, 81(6), 1994.

[SS90] Kwang-Tsao Shao and Robert R. Sokal. Tree balance. *Systematic Zoology*, 39(3):266–276, 1990.