Shapes of Phylogenetic Trees: Age Model and Likelihoods

Stephanie Keller-Schmidt[1], Murat Tuğrul[2], Victor M. Eguíluz[2],
Emilio Hernández-García[2], Konstantin Klemm[1]

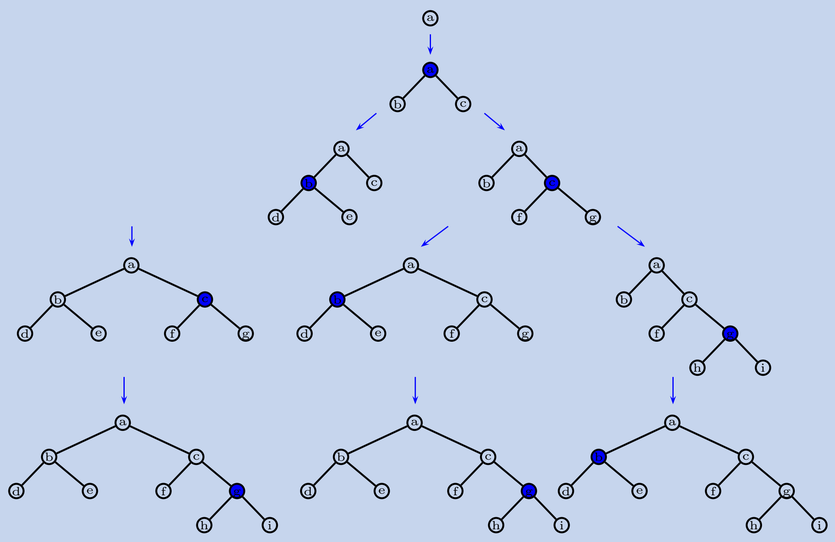
[1] *Bioinformatics Group, Department of Computer Science, University Leipzig, Germany*
[2] *IFISC, UIB − CSIC Campus Universitat de les Illes Balears, Palma de Mallorca, Spain*
E − Mail : stephanie@bioinf.uni − leipzig.de

Abstract

Phylogenetic trees serve to represent phylogenetic relationships arising from evolutionary events such as speciation and extinction. In these binary trees, leave nodes represent observed (extant) species while inner nodes stand for events of speciation. The shape of a tree is its pure graph structure neglecting the annotation of evolutionary time and species. Tree shapes based on empirical deviate significantly from those predicted by completely uncorrelated speciation processes such as the Equal Rates Markov (ERM) model. In this model, the depth (average distance of leaves from root) scales logarithmically with the number of leaves. A faster depth scaling with system size is observed in the real trees as a sign of systematic imbalance. Here we introduce a potential explanation of tree imbalance in terms of a speciation rate that decreases with the age of a species. Trees are grown by iterative branching. The branching leaf is chosen with probability inversely proportional to the time that passed since the leaf's creation. This model produces trees with depth scaling as log_2 of tree size, in good agreement with trees in the databases TreeBASE and PANDIT. Furthermore, the likelihood of this model with respect to real trees is typically larger or equal to that of the AB model [1]. The latter model is known to account for the observed imbalance but is not based on any evolutionary mechanism [2]. Along with the modeling results, we introduce a general efficient Monte-Carlo method for likelihood estimation of growing tree models that do not factorize over branches. The method will facilitate verification and comparison of many complex growth models using large real trees.

Age Model

- Initialization $n = 1$: create a single node (root).
- Choose leaf l with probability suppressed by age
$$p_l \propto (n - t_l)^{-1}$$
and replace l by a cherry.
- $n =$ number of leaves = time
- $t_l =$ creation time of leaf l



$$L_{\text{age}}(T) = p\left((b,c,g),T\right) + p\left((c,b,g),T\right) + p\left((c,g,b),T\right)$$

Data

For the analysis and the calculation of the likelihood of the Age model we used the data of the two major databases.

database	TreeBASE	PANDIT
leaves are	species	proteins
number of trees	5212	46428
number of leaves	4 . . . 960	2 . . . 5121

In the case of TreeBASE polytomic bids were replaced by binary splits randomly. Furthermore monotomies were solved for trees of both databases.

Likelihood vs. Probabilty

Probability
→ What is the probability $Pr_A(T)$ for a model A to obtain the given data T?

Likelihood
→ With which probability $L_A(T)$ was data T generated using a certain model A?

When is a model A better than model B?

$$L_A(T) > L_B(T)$$

We calculated $L_{age}(T)$ for the Age model exactly by adding up probabilities of all sequences of branchings generating a tree T.

$$L_{\text{age}}(T) = \sum_{s \in S_c(t)} p(s,T)$$

with

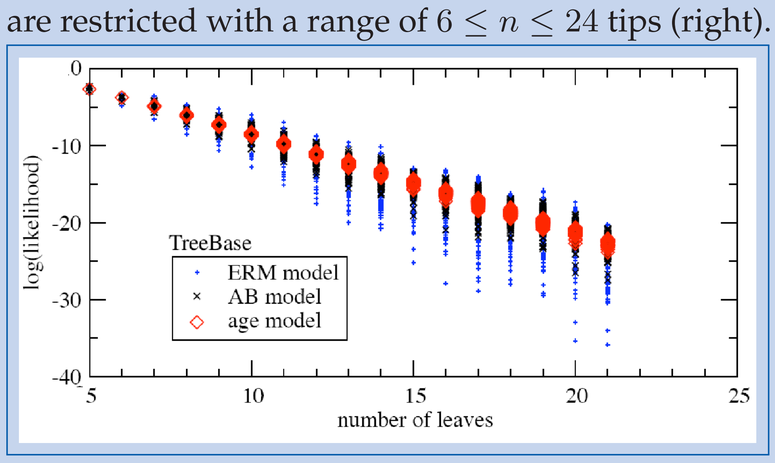
$$p(s,T) = \prod_{i=2}^{n-1} \frac{(s(i) - s(m(i))^{-1}}{\sum_{j \in B(s,s(i))} (s(i) - s(m(j))^{-1}}$$

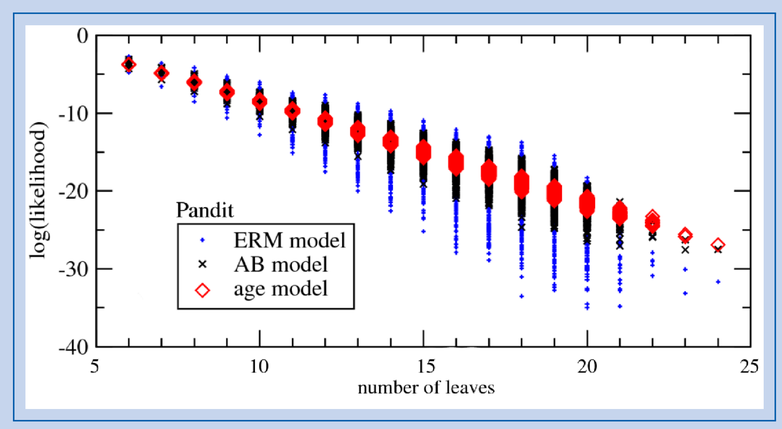
and

$$\begin{aligned} B(s,t) &= \{j \in I \setminus \{1\} \mid s(m(j)) < t < s(j)\} \\ &\quad \cup \{j \in A \setminus I \mid s(m(j)) < t\} \end{aligned}$$

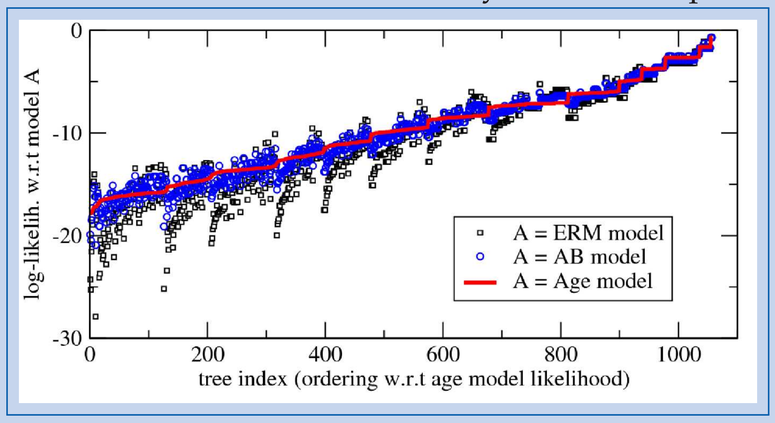
Results

In the following figures the likelihood analysis of empirical trees for the Age, AB and ERM models is shown. Each contributes one data point for each of the three models. For the scatter plot of the log-likelihoods for the TreeBASE database (left) all trees with $5 \leq n \leq 21$ were used. For the the PANDIT database the trees are restricted with a range of $6 \leq n \leq 24$ tips (right).

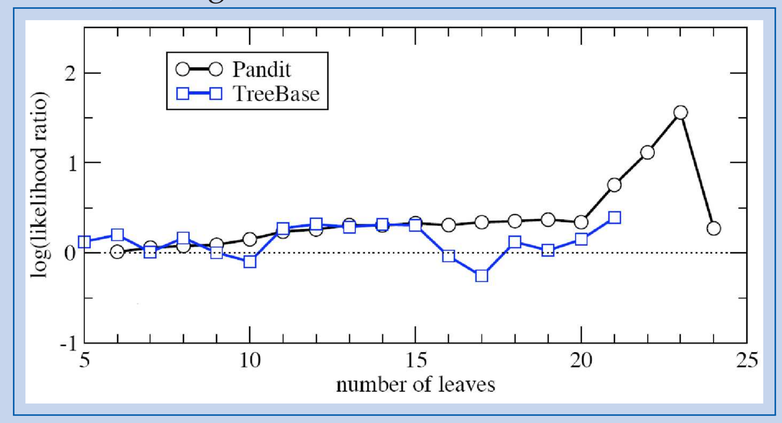




Log-likelihood for the Age, AB and ERM model for each tree in TreeBASE ordered by amount of tips n.



Logarithm of likelihood ratios between Age and AB model, averaged over all trees of the same size n.



In comparision to the AB model the Age model holds the advantage of a macroevolutionary motivation. It could be shown that the correlation of log-likelihoods between the AB model and ERM model is stronger than correlation of log-likelihoods between Age model and ERM model. Furthermore distances in the tree data reproduced by the Age model are slightly better than by the AB model.

References

[1] D. Aldous. Probability Distributions on Cladograms In *Random Discrete Structures 1996*
[2] M. Blum and O. Francois. Which Random Processes Describe the Tree of Life? A Large-Scale Study of Phylogenetic Tree Imbalance In *Systematic Biology 2006*

UNIVERSITÄT LEIPZIG
printed by Universitätsrechenzentrum Leipzig