An Age-Dependent Branching Model for Macroevolution

Emilio Hernández-García[1], Stephanie Keller-Schmidt[2], Murat Tuğrul[3], Víctor M. Eguíluz[1], Konstantin Klemm[2]

[1] IFISC (CSIC-UIB), Palma de Mallorca – Spain.
[2] Bioinformatics Institute, University of Leipzig – Germany
[3] IST Austria, Klosterneuburg – Austria

emilio@ifisc.uib-csic.es

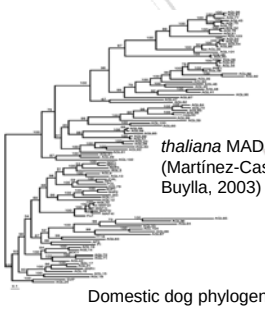
Summary

■ **Phylogenetic trees** are reconstructions of the **evolutionary history** of organisms or genes (or their proteins) based on present-day genomic information. **Branching patterns** in phylogenetic trees help to identify and distinguish different **evolutionary mechanisms**.

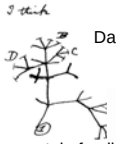
■ The **imbalance** of phylogenetic trees (i.e. the amount of **asymmetry** between the two subtrees arising in a branching event) exhibits a systematic deviation from the expectation of a purely **random tree growth process** (such as provided by the Yule or the ERM models). Random tree branching leads to a **scaling of the depth of the trees** (the mean distance of tips from root) with tree size *n* as ***d ~ log n***, whereas true phylogenies display a faster depth scaling with size [Herrada2008]. Some models [Ford2005,Hernandez2010] have been already proposed to fit such behavior, but without a clear biological meaning.

■ Here **[Keller2011]** we introduce an **age-dependent stochastic branching model** based on the hypothesis that speciation rate is a decreasing function of the **waiting time** since the last speciation. We find that the depth grows as ***d ~ (log n)²*** in leading order with tree size *n*. This result is in good agreement with the trend observed by exhaustive analysis of the phylogenetic databases **TreeBASE** and **PANDIT**. **Exact likelihood computation** of the model on the trees up to 20 tips contained in the databases is performed. Higher likelihoods values are found when compared with a previously suggested model [Aldous1996].

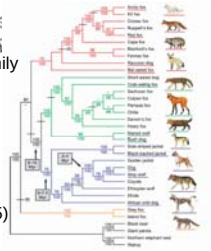
1.- Motivation: finding a branching model consistent with available phylogenetic data



thaliana MADS-box protein family (Martínez-Castilla and Alvarez-Buylla, 2003)



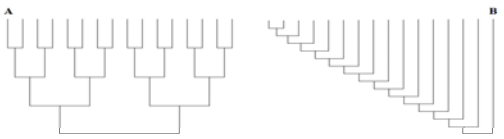
Darwin sketch (1837)



Domestic dog phylogeny (Pollinger, 2005)

2.- Balance, imbalance and the scaling of the tree depth

The most **balanced** (A) and the most **imbalanced** (B) trees with 16 leaves.



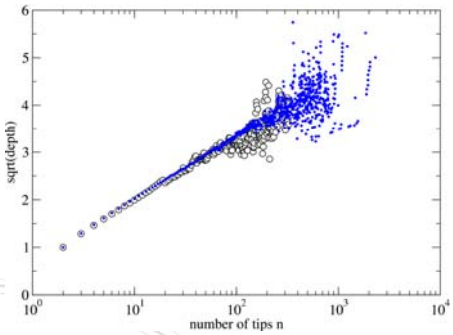
We characterize **balance** with the dependence of the **tree depth** *d* with the number of leaves *n*, for large *n*.

$$d = n^{-1}\sum_{i=1}^{n} d_i$$

d_i is the number of edges between the leave *i* and the root.

For a fully balanced tree, ***d ~ log n***
This is also the scaling for the **random tree** and for virtually any model of **uncorrelated branching**

For a fully imbalanced tree, ***d ~ n***



The square-root of the mean depth vs. size of phylogenetic trees contained in databases for species (**TreeBASE**; empty circles) and proteins (**PANDIT**; filled circles). The mean depth is averaged for all trees having the same number of tips. In this scale (log-linear), the behavior ***d ~ (log n)²*** is a straight line. Data from **TreeBASE** were downloaded from **http://www.treebase.org** on June, 2007 containing 5,212 phylogenetic trees; data from **PANDIT** were downloaded from **http://www.ebi.ac.uk/goldman-srv/pandit** on May 2008 and contains 7,738 protein families.

3.- An age-dependent branching model

Since **temporal correlations** are needed to break the simplest logarithmic scaling, we propose a model in which **at each time step one of the species branches**, which is selected with a **probability inversely proportional to its age** (time since birth): **Species just speciated are more likely to speciate again than older ones**

MEAN DEPTH CALCULATION:

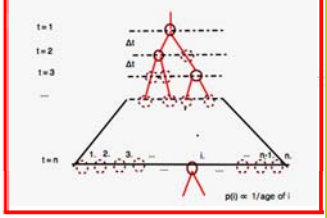
The premise: mean depth path is made of branches whose ages are approximated by expected value at time *t*:

$$t(d) - t(d-1) = \langle\tau\rangle_{t(d)}$$

which yields to

$$\frac{dd}{dn} = [\langle\tau\rangle_n]^{-1}.$$

$p_i = c\ (age\ of\ i)^{-1}$; *c is normalization constant*



Assuming a mean value for normalization constant $c(n)$, one can show

$$\langle\tau\rangle_n = n\ c(n)$$

Considering $c^{-1}_{\text{slowest}}(n) \leqslant c^{-1}(n) \leqslant c^{-1}_{\text{fastest}}(n)$
where c_{slowest} and c_{fastest} refer to the normalization constant of the slowest and fastest realization of the branching process:

$$c^{-1}_{\text{slowest}}(n) = 1 + \sum_{i=1}^{n-1} i^{-1}, \quad c^{-1}_{\text{fastest}}(n) = 2\sum_{i=1}^{\lfloor n/2 \rfloor} i^{-1} + r(n)$$

where $r(n) = ((n+1)/2)^{-1}$ if n is odd, and $r(n) = 0$ otherwise,

$$c(n) \to (\log n)^{-1} \text{ as } n \to \infty$$

and $\langle\tau\rangle_n \to \frac{n}{\log n}$ as $n \to \infty$.

Thus, $d \sim (\log n)^2$ as $n \to \infty$.

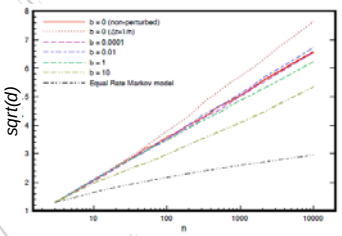
There are indications that the present model is at the **critical point** in a family of branching models with $p \propto (age)^{-q}$, giving a **phase transition** between **random branching behavior** ($d \sim \log n$) for $q<1$ and **power law scaling** ($d \sim n^{q-1}$) for $q>1$.

4.- Model variations

Simulation results confirm the $d \sim (\log n)^2$ behavior.

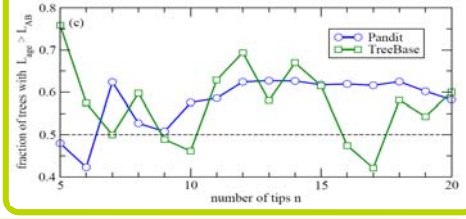
The modeling is robust under some modifications:

- from $p_i \propto (age)^{-1}$ to $\propto (age + const.)^{-1}$
- from $\Delta t = 1$ to $\Delta t = 1/n$.



5.- Comparison with data: model likelihood

We have calculated (see [Keller2011]) the probability that each tree *T* in our datasets (up to 20 leaves) has been obtained from the age model (**likelihod**: $L_{age}(T)$), and compared it with the likelihood $L_{AB}(T)$ under the so-called AB model [Aldous1996], which is a branching model without clear biological meaning but that also leads to $d \sim (\log n)^2$. The age model typically describes better the data (for both species and proteins).



References

[Aldous1996] – D. Aldous, *Probability distributions on Cladograms*. In Random Discrete Structures. Ed. by Aldous and Pemantle (1996).
[Ford2005] – D.J. Ford. *Probabilities on cladogram: Introduction to the alpha model.* PhD Thesis, Stanford University (2005).
[Hernandez2010] – E. Hernández-García, M. Tugrul, E.A. Herrada, V.M. Eguíluz, K. Klemm. *Simple models for scaling in phylogenetic trees*. Int. J. Bif. Chaos 20, 805 (2010).
[Herrada2008] – E.A. Herrada, C.J. Tessone, K. Klemm, V.M. Eguíluz, E. Hernández-García, C.M. Duarte, *Universal scaling in the branching of the tree of life*. PLoS ONE 3, e2757 (2008).
[Herrada2011] – E.A. Herrada, V.M. Eguíluz, E. Hernández-García, C.M. Duarte. *Scaling properties of protein family phylogenies,* BMC Evolutionary Biology 11, 155 (2011).
[Keller2011] – S. Keller-Schmidt, M. Tugrul, V.M. Eguíluz, E. Hernández-García, K. Klemm. *An age-dependent branching model for macroevolution*. Submitted to Syst. Biol. (2011).