Statistics of phylogenetic tree structures

Stephanie Keller-Schmidt

Group of Bioinformatics
Group of Parallel Computing and Complex Systems
Departement of Computer Science
University of Leipzig

Bled, 17. February 2009

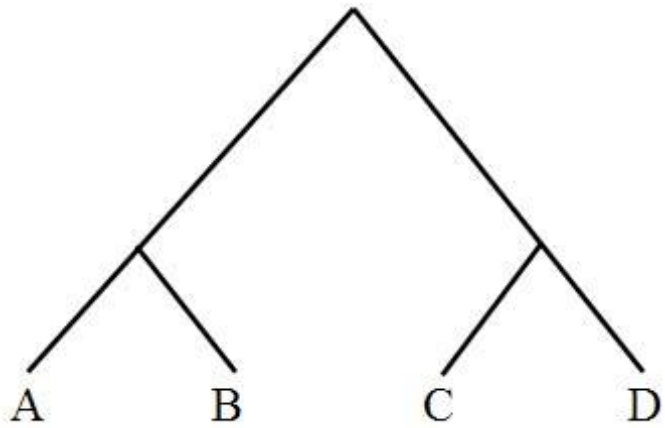
Outline

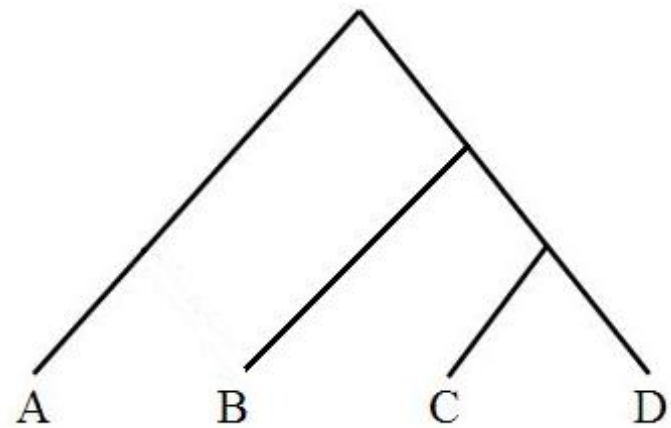
Reasons for considering probability models of phylogenetic trees and generate random trees with models :

- Understand speciation and extinction.
- Do predictions that models make about tree shape which can be used to test hypothesis concerning speciation.
- Useful for exploring biases in tree reconstruction methods.
- Testing algorithms: how well does it reconstruct a tree.

Aim: infer how diversity has arisen.
How: fitting stochastic models to tree data.

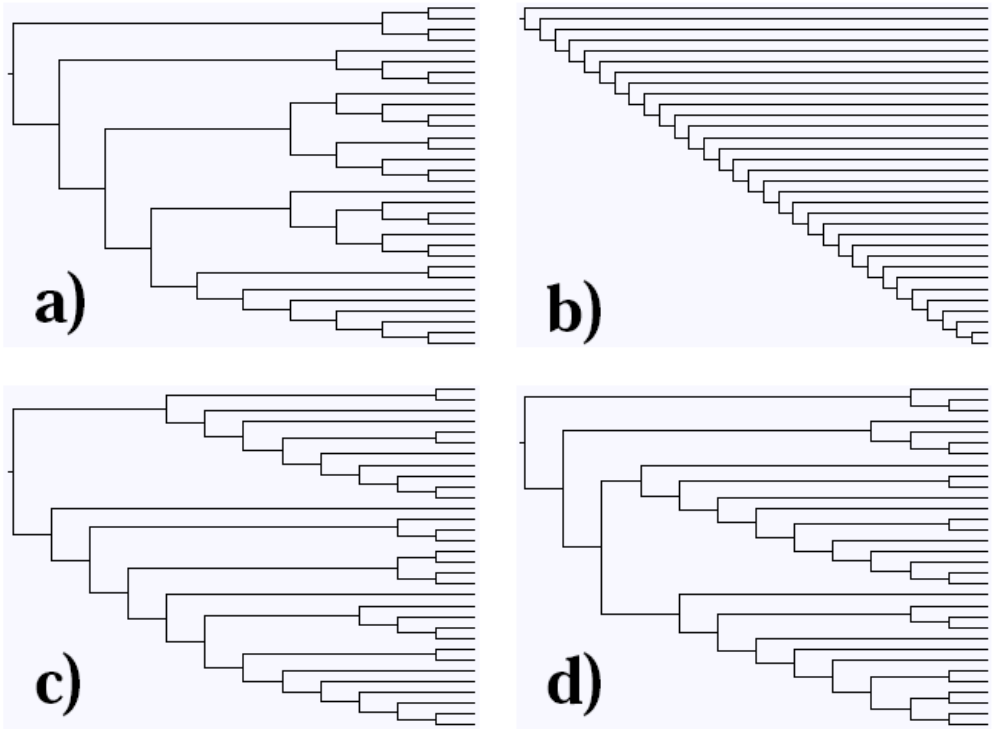
Tree Balance





- Degree to which daughter subtrees of internal nodes are of similar or different size.
- Refers to topological structure of tree, not considering the branch length.

Tree Balance



Data

5211 trees of major database TreeBASE (polytomic bids replaced by binary splits).

Early studies (Guyer and Slowinski(1991),Heard(1992)): reconstructed phylogenies are more imbalanced than predicted by Equal Rates Markov (ERM) model. $\Rightarrow$ Reasons?

Imbalance visible by distance scaling:

$$\langle d \rangle = \text{average distance from root} \propto (\log n)^2$$

for a subtree with n leaves.

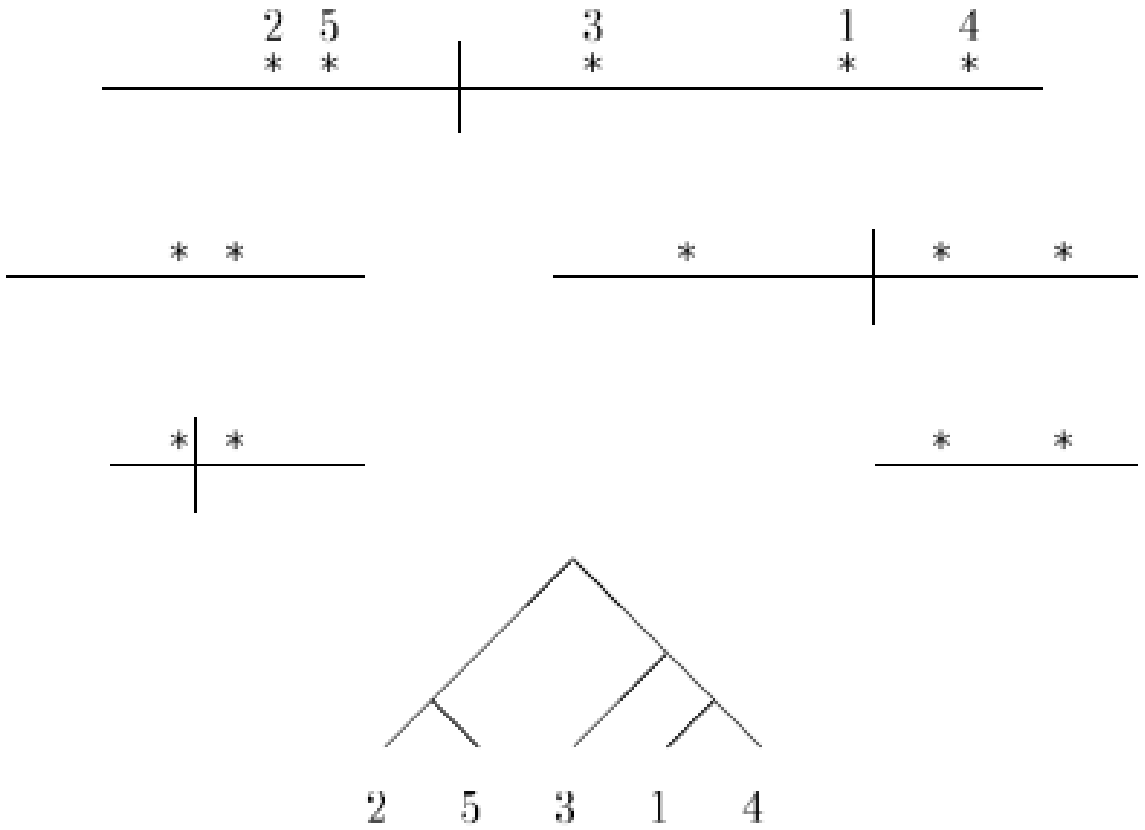
AB-model

The AB-model

- assumes that the splitting in a subtree is independent of what has happened above this subtree.
- is a simple probability distributions on trees, where amount of element in left branch is chosen at random according to the distribution.
- is not intended to model any evolutionary process.
- is the only "approved" model for the treebase data so far.

$\Rightarrow$ Instance of “beta-splitting” model might approximate the distribution of macroevolutionary phylogenetic tree reconstructed from sequence data.

AB-model



Aldous(1996)

AB-model

Idea: recursively split the taxa into subclades using a distribution derived from the beta distribution.

Assume: clade has n taxa, probability of the split being between subclades of size i and $n-i$ is:

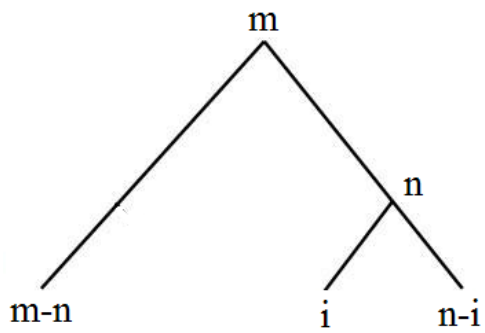
$$p(i|n-1) \propto \frac{1}{i(n-i)} \quad \textit{for} \quad i \in 1,2,\ldots,n-1$$

Question

- Do trees from the database fullfill the Markov-property?

$$p(i|n) \propto \frac{1}{i(n-i)}$$

$$p(i|m,n) \stackrel{?}{=} p(i|n)$$



Mutual Information MI

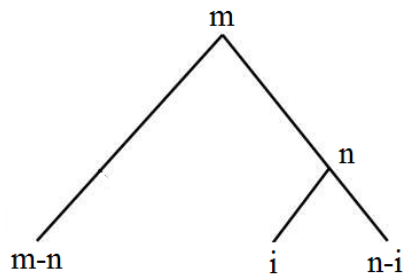
- MI of two random variables X, Y= quantity that measures the mutual dependence of the two variables

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right)$$

- $MI(X;Y) = 0$ iff X and Y are independent random variables
- $p(x,y)$ joint probability distribution function of X, Y
- $p_1(x), p_2(y)$ marginal probability distribution functions of X, Y

Mutual Information MI

	$n < \frac{m}{2}$	$n > \frac{m}{2}$	
$i < \mu(n)$			
$i > \mu(n)$			



Results Probabilities and MI

TREEBASE	$n < \frac{m}{2}$	$n > \frac{m}{2}$	
$i < \mu(n)$	0.239	0.246	0.485
$i > \mu(n)$	0.261	0.255	0.516
	0.500	0.501	

$MI \approx 10^{-4}$ (over 104643 trees)

ABMODEL	$n < \frac{m}{2}$	$n > \frac{m}{2}$	
$i < \mu(n)$	0.234	0.233	0.467
$i > \mu(n)$	0.266	0.267	0.533
	0.500	0.500	

$MI \approx 10^{-5}$ (over 302006 trees)

AGEMODEL	$n < \frac{m}{2}$	$n > \frac{m}{2}$	
$i < \mu(n)$	0.221	0.198	0.419
$i > \mu(n)$	0.279	0.301	0.580
	0.500	0.499	

$MI \approx 10^{-3}$ (over 252502 trees)

Age model

Idea: The longer species i has not been involved in speciation, the less likely it is to do so now.

Initialize: Set time $t = 0$, generate root node.

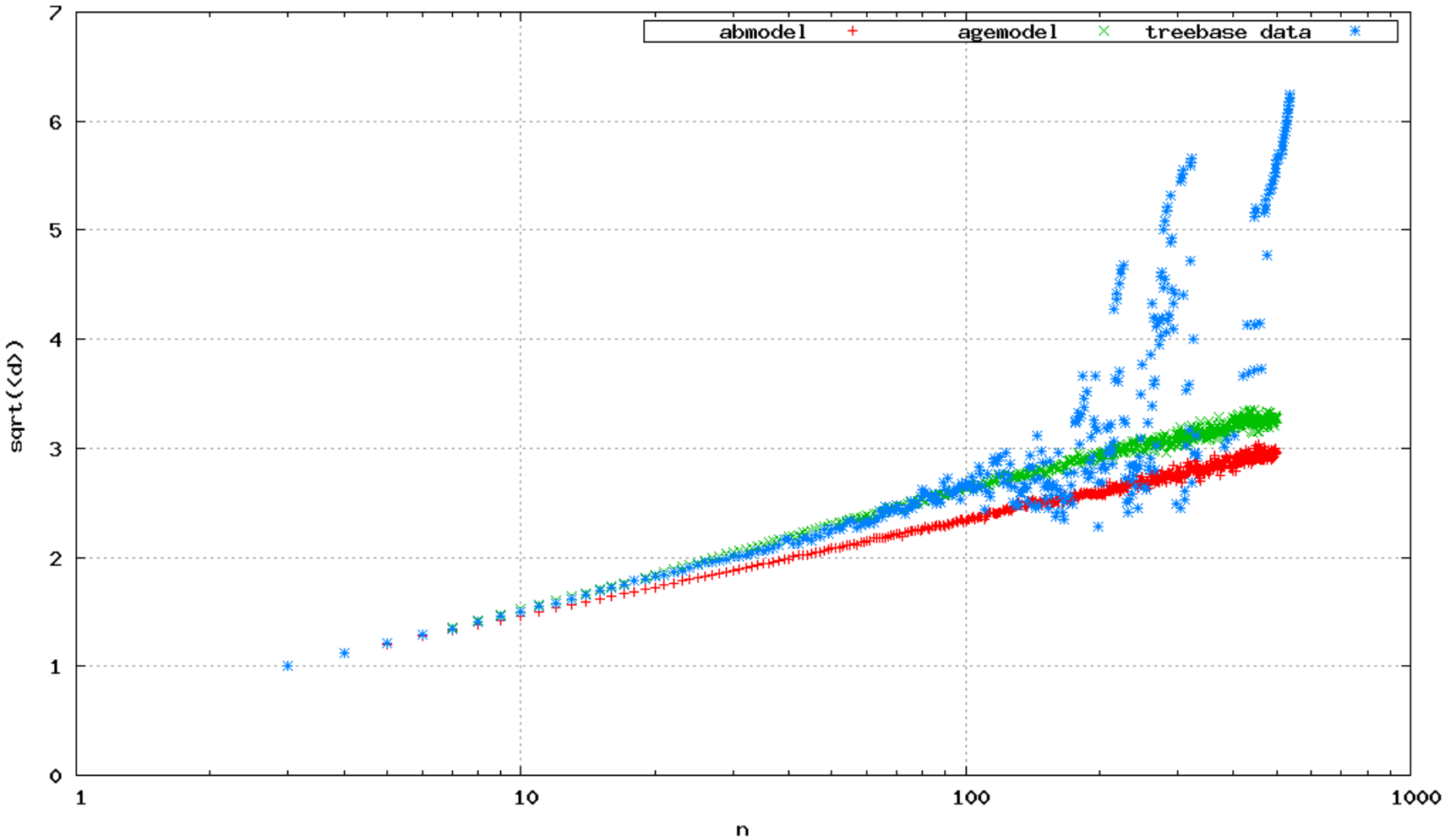
Iterate:

- Increment time t.
- From the set of leaves, choose leave i with probability

$$p_i \propto (t - t_i)^{-1}$$

  (inversely proportional to age) where t_i is the time node i was generated.
- Chosen leave i splits into two subclades.

Tree Imbalance: $\langle d \rangle \sim (\log n)^2$



Summary

- Comparison of trees generated by model vs. treebase data.
- Stochastic independence between nested subtree structures reproduced by models: AB model and age model.
- AB-model, however, not motivated by real macroevolution.
- Distances in the tree data reproduced by age model, slightly better than AB-model.

Thank You