

Wellcome trust advanced course  
“Molecular Evolution”

## **Bayesian Methods for Molecular Phylogenetics**

Manolo Gouy  
CNRS - Université de Lyon

The ML phylogenetic tree-building method considers parameters:

- tree topology  $\tau$
- branch lengths  $\nu$
- across-sites rate variation  $\alpha$
- rate substitution matrix  $\theta$

and estimates the parameter values  $(\tau_o, \nu_o, \alpha_o, \theta_o)$  that maximize

$$\Pr(\text{sequences} \mid \tau_o, \nu_o, \alpha_o, \theta_o).$$

Bayesian phylogenetics go one step further in using probabilities:

We assume there exists (and we know them) **prior probabilities** for all values of all of our parameters :

- tree topology  $\tau$  : some probability for each tree shape
- branch lengths  $\nu$  : some probability for each possible length
- across-sites rate variation  $\alpha$  : some probability for each  $\alpha$  value
- rate substitution matrix  $\theta$  : some probability for each value of each matrix term.

It thus becomes possible (and interesting) to consider new quantities, called **posterior probabilities**

$$\Pr(\tau_o, \nu_o, \alpha_o, \theta_o | \text{sequences})$$

These posterior probabilities mean:

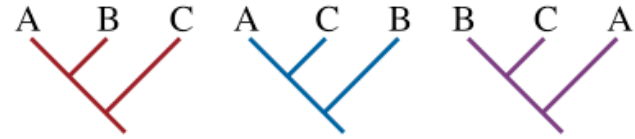
“Given my sequences, what parameter values (e.g., tree shape, branch length) have high probability ? What values have low probabilities ?”

Posterior probabilities can be expressed using Bayes' formula

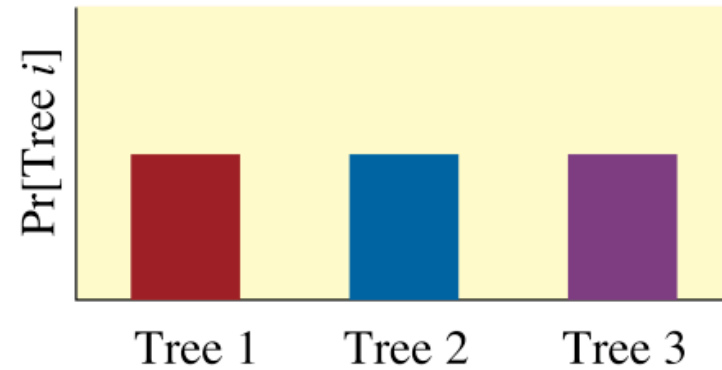
$$\Pr(A | B) = \frac{\Pr(B | A) \cdot \Pr(A)}{\Pr(B)}$$

We obtain :

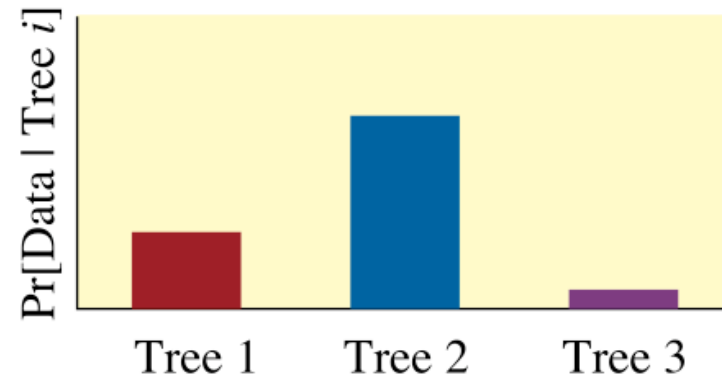
$$\underbrace{\Pr(\tau_o, \eta_o, \alpha_o, \theta_o | \text{Sequences})}_{\text{posterior probabilities}} = \frac{\overbrace{\Pr(\text{Sequences} | \tau_o, \eta_o, \alpha_o, \theta_o)}^{\text{likelihood}} \cdot \overbrace{\Pr(\tau_o, \eta_o, \alpha_o, \theta_o)}^{\text{prior probas}}}{\Pr(\text{Sequences})}$$



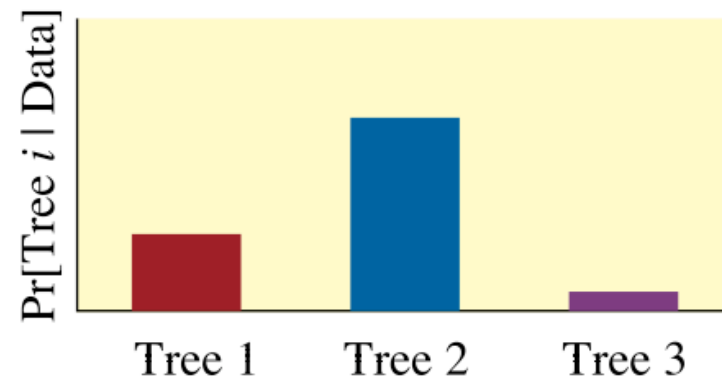
The **prior probability** of a tree represents the probability of the tree before the observations have been made. Typically, all trees are considered equally probable.



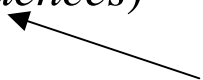
The **likelihood** is proportional to the probability of the sequence data set given the parameters (tree shape, branch lengths, ... ).



The **posterior probability** of a tree is the probability of a tree conditional to the sequence data. It is obtained by combining the prior and the likelihood using Bayes formula.



$$\underbrace{\Pr(\tau_0, \eta_0, \alpha_0, \theta_0 \mid Sequences)}_{\text{posterior probabilities}} = \frac{\overbrace{\Pr(Sequences \mid \tau_0, \eta_0, \alpha_0, \theta_0)}^{\text{likelihood}} \cdot \overbrace{\Pr(\tau_0, \eta_0, \alpha_0, \theta_0)}^{\text{prior probas}}}{\Pr(Sequences)}$$


 not computable

Our problem is that we know how to compute the numerator of this expression, but not the denominator because it would require integration over all possible values of all parameters.

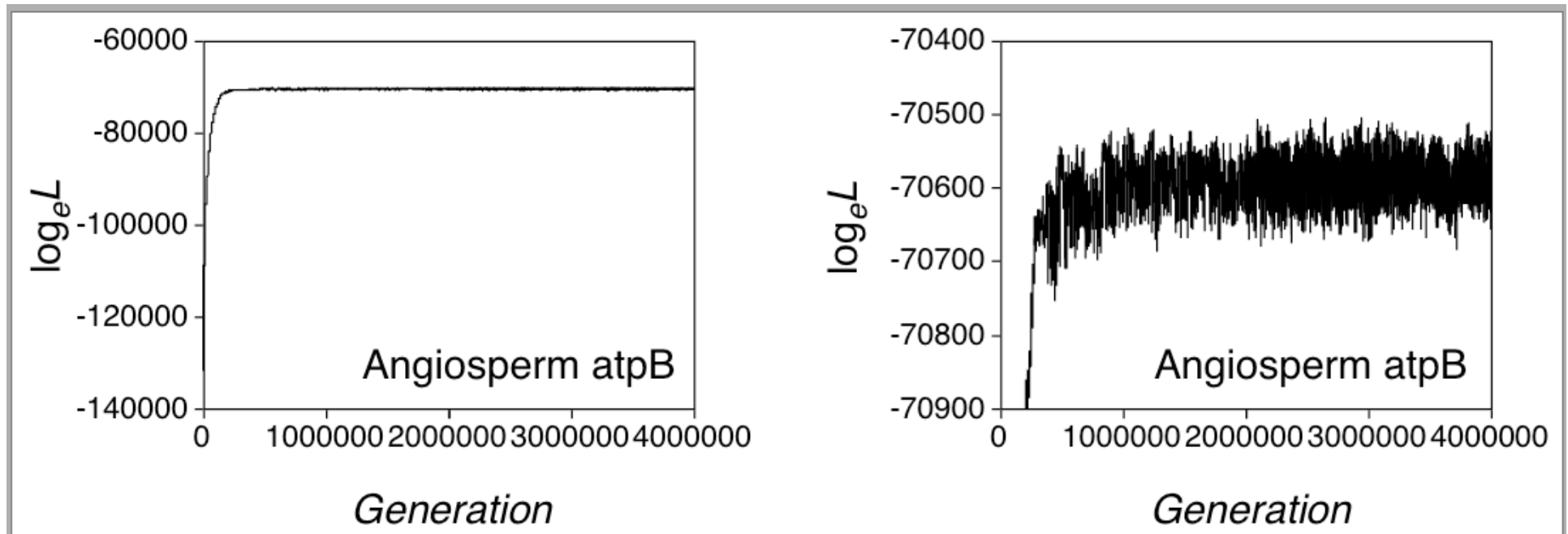
This is where the **Markov Chain Monte Carlo** method (MCMC) comes in.

MCMC allows to sample a (potentially very complex) distribution even if we don't know how to compute  $\Pr(\text{point})$  but only  $\Pr(\text{point}_1) / \Pr(\text{point}_2)$ .

$$\frac{\Pr(\tau_1, \eta_1, \alpha_1, \theta_1 \mid Sequences)}{\Pr(\tau_2, \eta_2, \alpha_2, \theta_2 \mid Sequences)} = \frac{\Pr(Sequences \mid \tau_1, \eta_1, \alpha_1, \theta_1) \cdot \Pr(\tau_1, \eta_1, \alpha_1, \theta_1)}{\Pr(Sequences \mid \tau_2, \eta_2, \alpha_2, \theta_2) \cdot \Pr(\tau_2, \eta_2, \alpha_2, \theta_2)}$$

How does an MCMC method run ?

Starting from an initial set of parameter values  $(\tau_0, \nu_0, \alpha_0, \theta_0)$  it builds iteratively a very long series of parameter value sets  $(\tau_1, \nu_1, \alpha_1, \theta_1), (\tau_2, \nu_2, \alpha_2, \theta_2), (\tau_3, \nu_3, \alpha_3, \theta_3), \dots, (\tau_n, \nu_n, \alpha_n, \theta_n), \dots$  whose equilibrium distribution (*i.e.*, far into the chain) equals the posterior distribution sought for.



# MrBayes : a Bayesian tool for molecular phylogeny

[ Huelsenbeck & Ronquist (2001) *Bioinformatics* 17:754 ]

A Bayesian implementation must completely specify:

- what are the parameters used ( $\tau, \nu, \theta, \alpha$ ) ?
- what are their prior distributions?

## $\theta$ Substitution rate matrices

Three evolutionary models are possible.

1. J&C : all information is in branch lengths (parameter  $\nu$ ).
2. K2P : TR/TV ratio follows a beta( $x, y$ ) where  $x/y$  = average TR/TV ratio; (default: beta(1,1)).
3. GTR : the 6 model rates (A $\leftrightarrow$ C, A $\leftrightarrow$ G, A $\leftrightarrow$ T, C $\leftrightarrow$ G, C $\leftrightarrow$ T, G $\leftrightarrow$ T) each follows a gamma distribution with user-chosen relative mean (default: equal means); the 3 equilibrium frequency parameters are set to average sequence frequencies.

$\alpha$  across-sites rate variation: choice between no variation, discretized gamma distribution with  $\alpha$  parameter, with or without invariable sites.

*prior*:  $\alpha$  can be uniformly distributed on [min , max] or exponentially distributed (default Unif [0.1 , 50]); fraction of invariable sites Unif [0,1].

$\nu$  tree branch lengths.

1. Unrooted case

- Unif [min, max]
- Exponential( $\lambda$ ) (default exp(10))

2. Rooted case with molecular clock

$\tau$  tree topologies

1. Uniform distribution (default)

2. Also possible to express topological constraints



## Typical results of a MrBayes run

A large sample ( $10^6$  points) from the parameter space  $(\tau, \nu, \alpha, \theta)$  is built. Its initial part ( $10^4$ ) is ignored, and the remaining part is sub-sampled (one out of 100 points is retained).

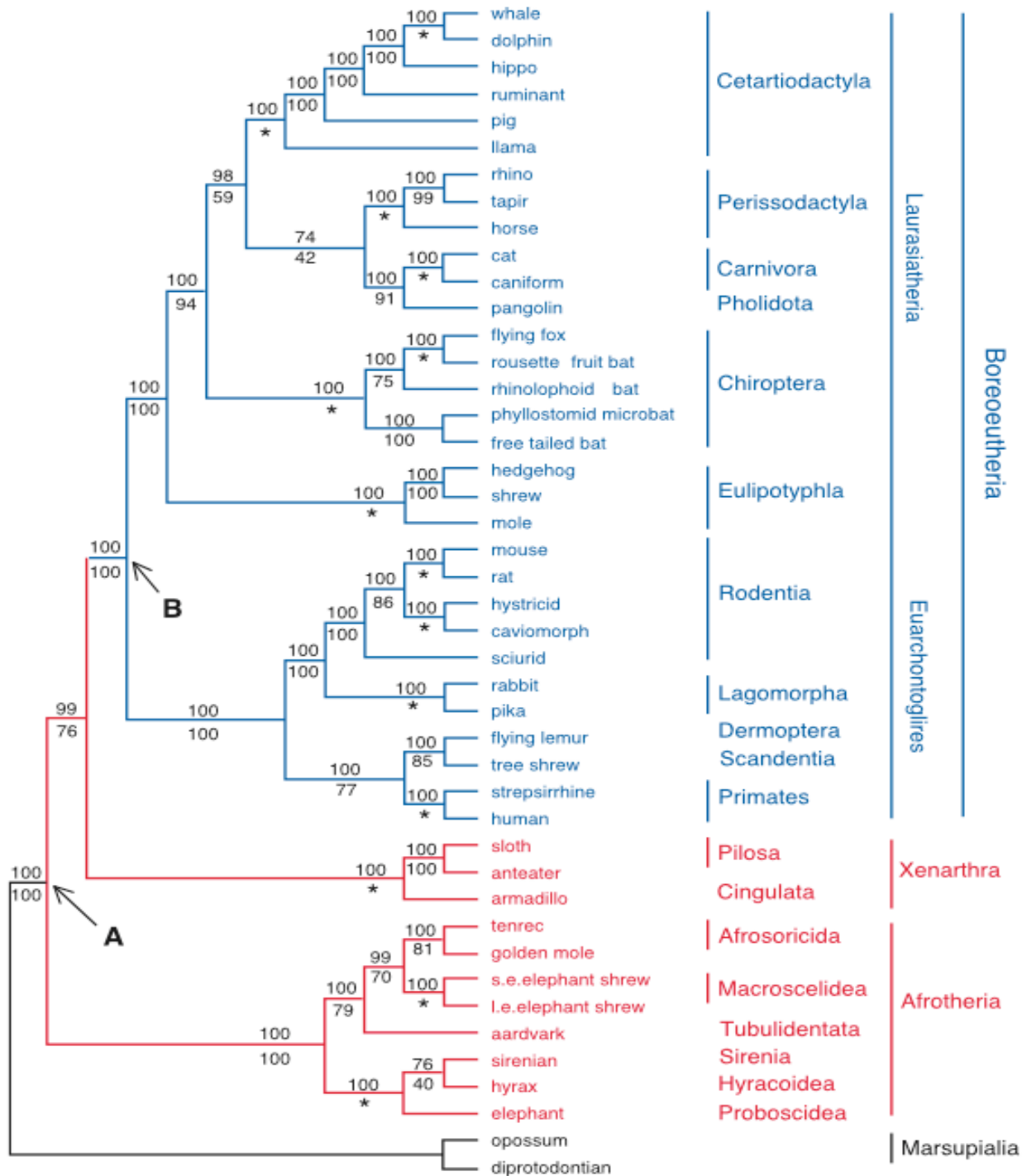
One gets  $\approx 10^4$  trees.

Typical expression of results:

majority rule consensus tree,

and

posterior probability of its clades: fraction of the  $10^4$  trees containing this clade.



## Bayesian phylogenetic analysis of placental mammals.

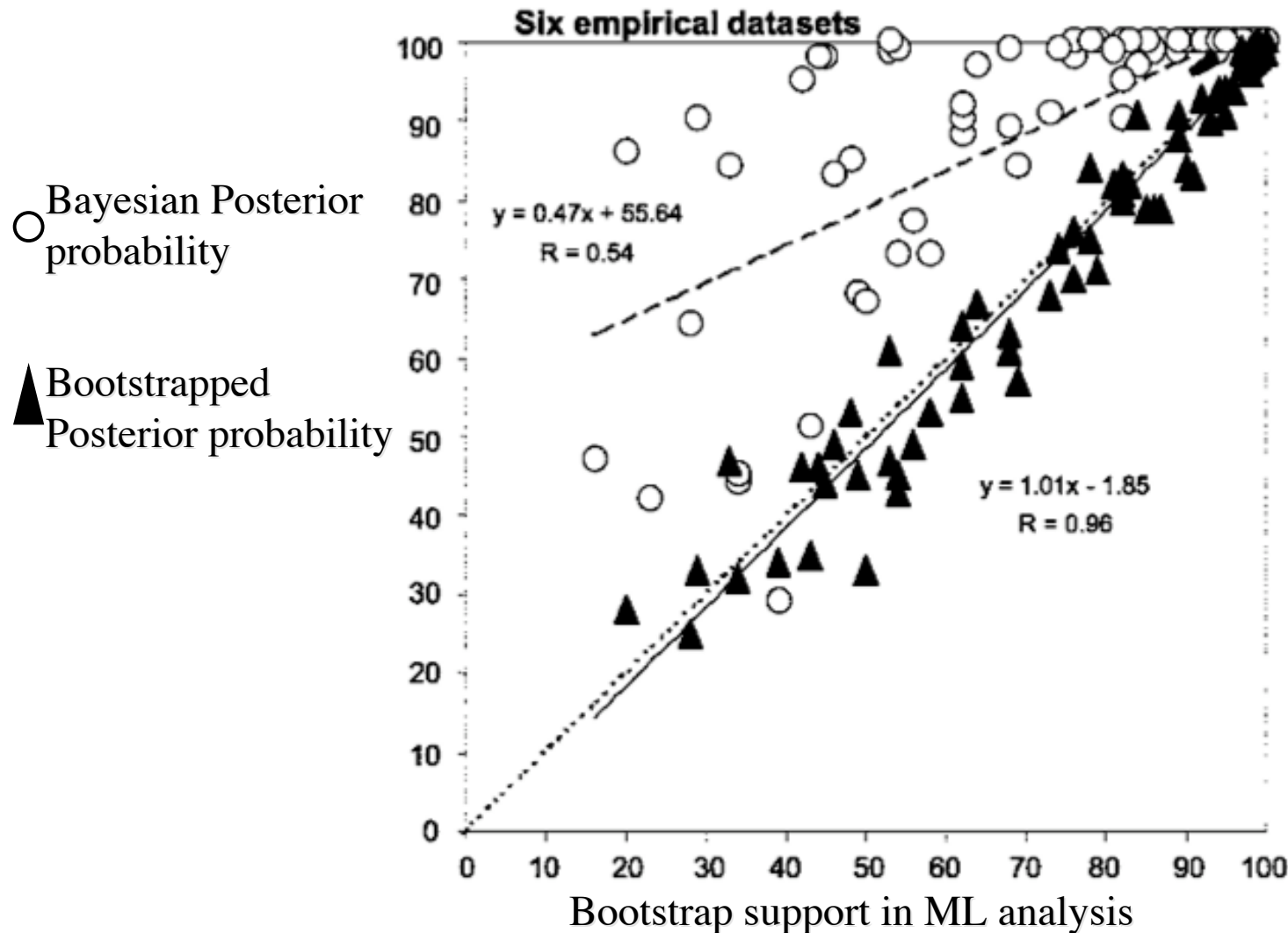
Tree without branch lengths.

above branches: posterior probability of clades;

below branches: ML bootstrap support.

# Bayesian overestimation of clade support ?

Bayesian clade support is often much higher than bootstrap support



Douady et al. (2003)  
Mol. Biol. Evol.  
20:248–254

## Advantages of the Bayesian approach for phylogenetics

- Statistically sound and well founded.
- Explicit assumptions.
- Can use very complex evolutionary models.
- Does not yield a single tree as does ML, but a distribution of trees. The variation in shape of these trees expresses the phylogenetic uncertainty present in sequences.
- Posterior probabilities nicely express statistical support for internal branches (or clades).
- Posterior probabilities of internal branches are often very high.

## Drawbacks of Bayesian methods

- Prior probabilities must be specified, most often totally arbitrarily.
- Convergence of the MCMC chain may be difficult to ensure.
- Posterior branch support may be too optimistic.
- Computation time longer than with fast ML algorithms.