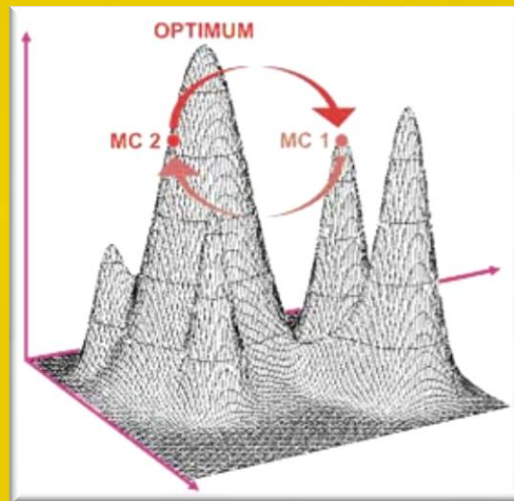


# Bayesian inference of phylogeny

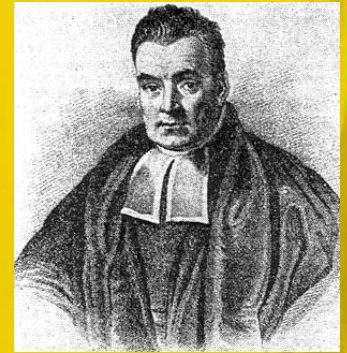
Βαλειαση ιμφειρενσε οφ βμλιοδεμλ



MrBayes  
Phylobayes

# Thomas Bayes (1702 - 1761)

Mathématicien & Philosophe



**Théorème de Bayes** publier en 1763:

Essay towards solving a problem in the doctrine of chances.

*Philosophical Transactions of the Royal Society of London.*

**Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference**

**Bruce Rannala, Ziheng Yang**

J Mol Evol (1996) 43:304–311

JOURNAL OF **MOLECULAR  
EVOLUTION**

© Springer-Verlag New York Inc. 1996

# Théorème de Bayes

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

A and B: 2 events that are NOT independent



→ P(1/2)

9 fair dice  
1 biased dice




Let's pick 1 dice in the box: **P(1/10)** to take the biased dice


We roll twice the dice and get 2 : do we know more about which dice was taken out of the box ?

# Théorème de Bayes

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

$$P(\text{biased dice} \mid \text{two } \img alt="die icon" data-bbox="308 500 355 565"/>) = \frac{\overbrace{P(\text{two } \img alt="die icon" data-bbox="485 455 532 520" \mid \text{biased dice})}^{1/2 \times 1/2 = 0.25} \times \overbrace{P(\text{biased dice})}^{1/10}}{P(\text{two } \img alt="die icon" data-bbox="640 540 687 605"/>)}$$
$$= \frac{(1/10 \times 1/2 \times 1/2) + (9/10 \times 1/6 \times 1/6)}{0.5}$$

**= 0.5**  Posterior probability

The prior probability of choosing the biased dice was 0.1, but knowing that two  were obtained in two rolls with this dice, the posterior probability of having indeed chosen the biased dice is now 0.5

# Théorème de Bayes en phylogénétique

What is the probability that our hypothesis **H** is correct, given the data **D** we have observed ?

Given: **D** a data set = an alignment

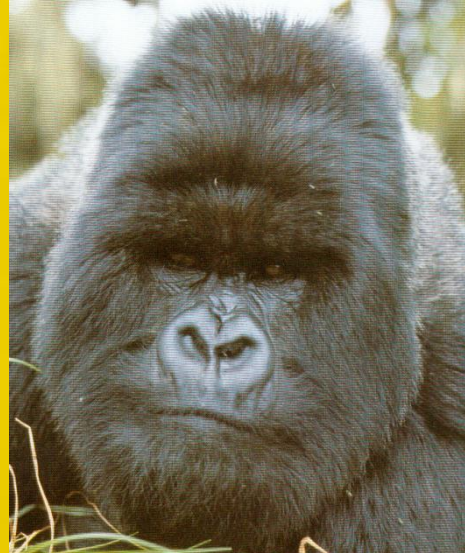
**H** an hypothesis = a tree topology, branch length,  
and model of evolution

likelihood of the tree      prior probability (prob. of our hyp.  
*before* the data were collected)

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}$$

Total likelihood of the data given all possible hypotheses (i.e. disregarding which hyp. is correct)

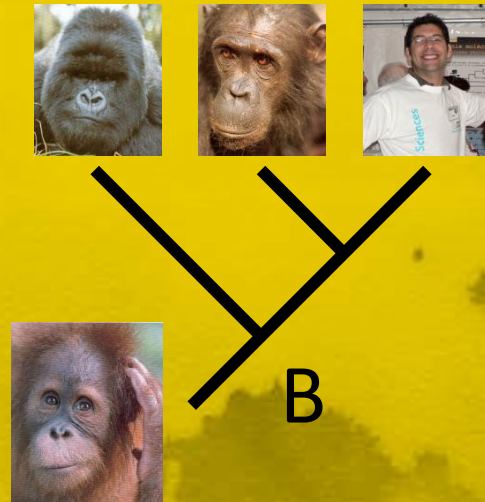
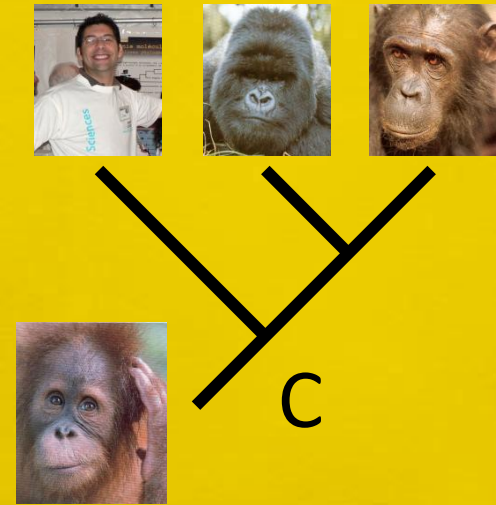
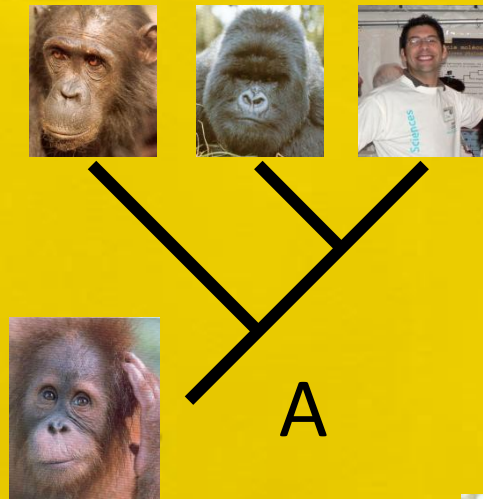
# Infer relationships among three species

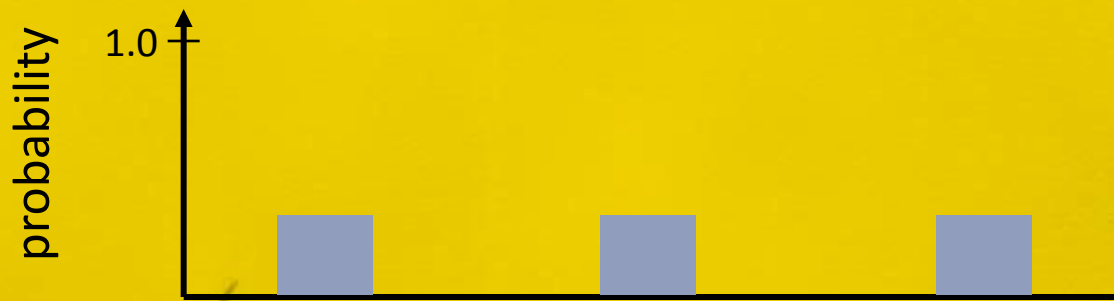
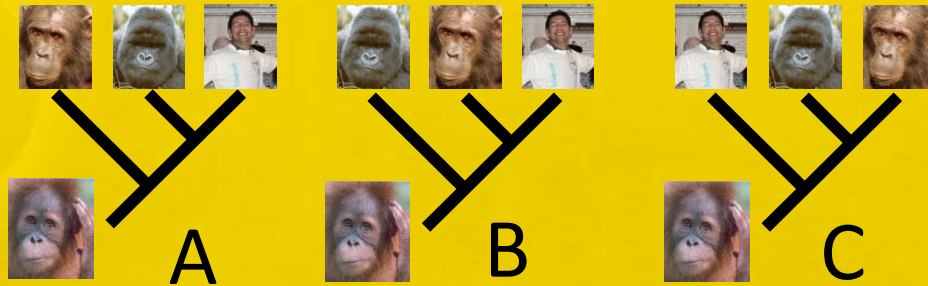


Outgroup:



# Three possible trees (topologies)

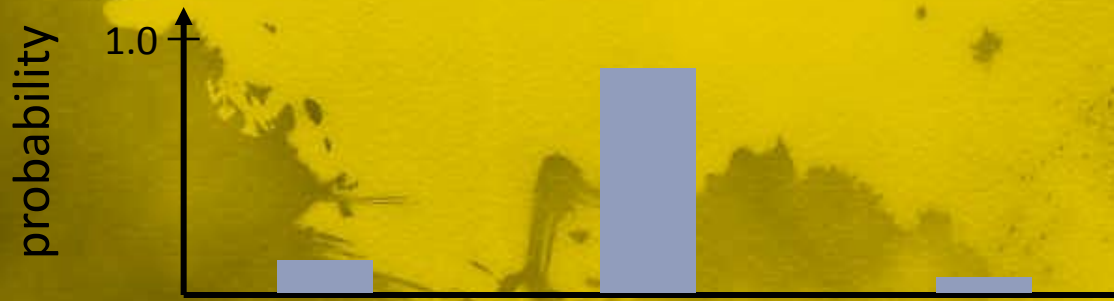




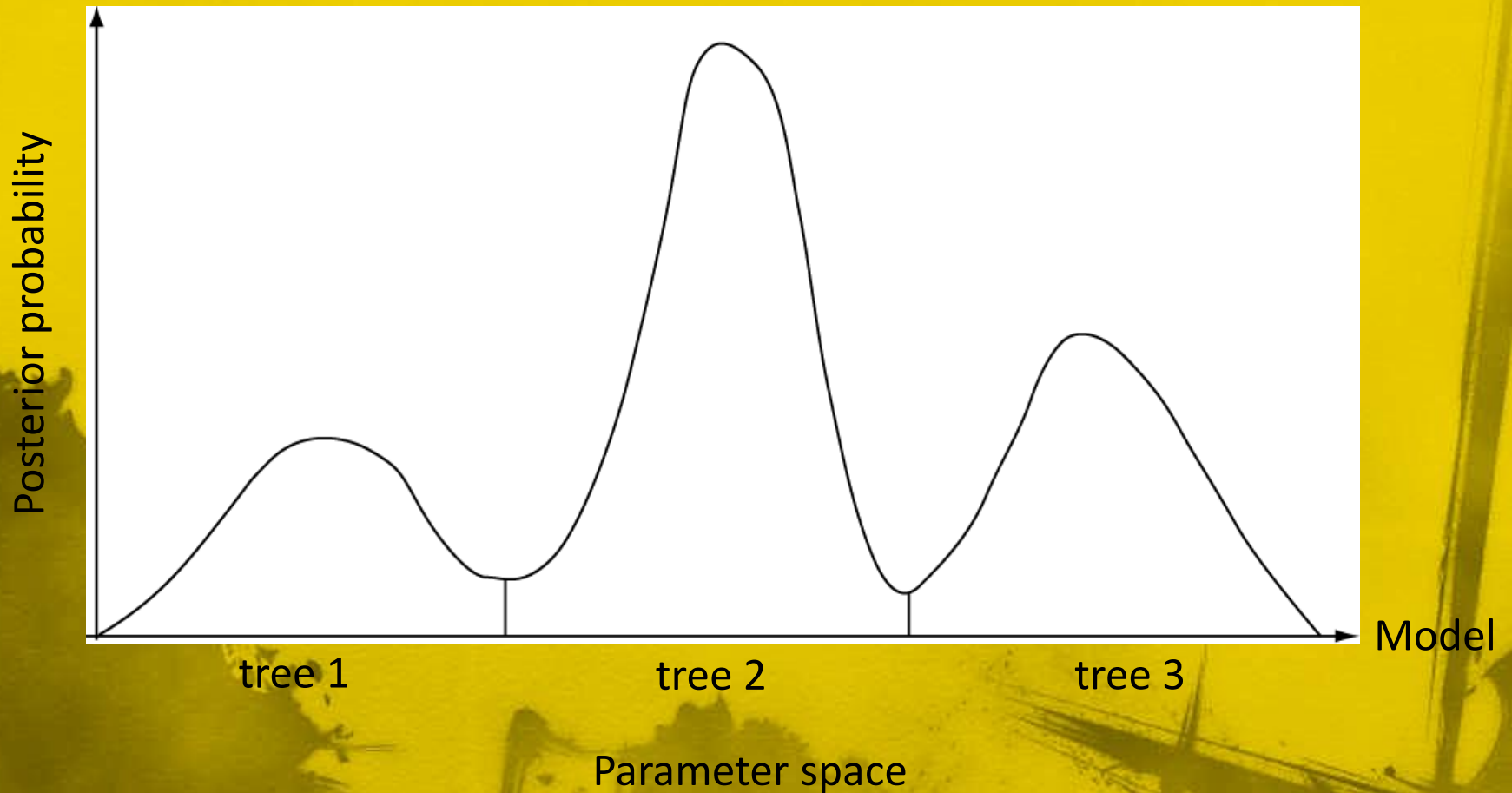
Data (observations)



$P(H|D)$



# Posterior probability distribution



# Markov Chain Monte Carlo (MCMC) algorithm

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}$$

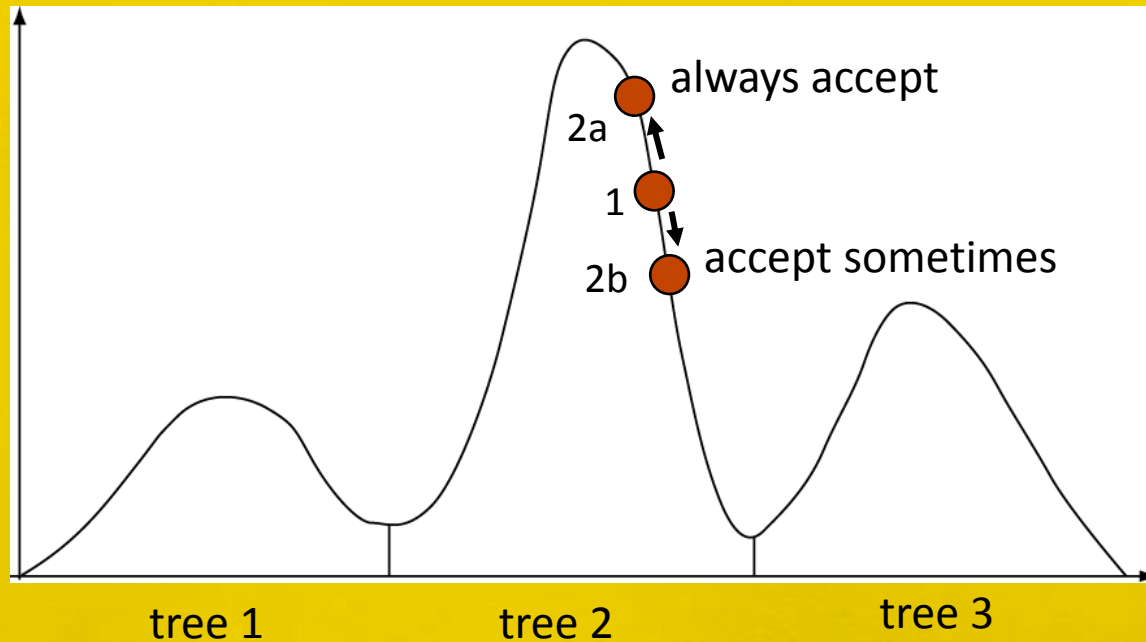
$P(D)$

En pratique, cette valeur est impossible à calculer

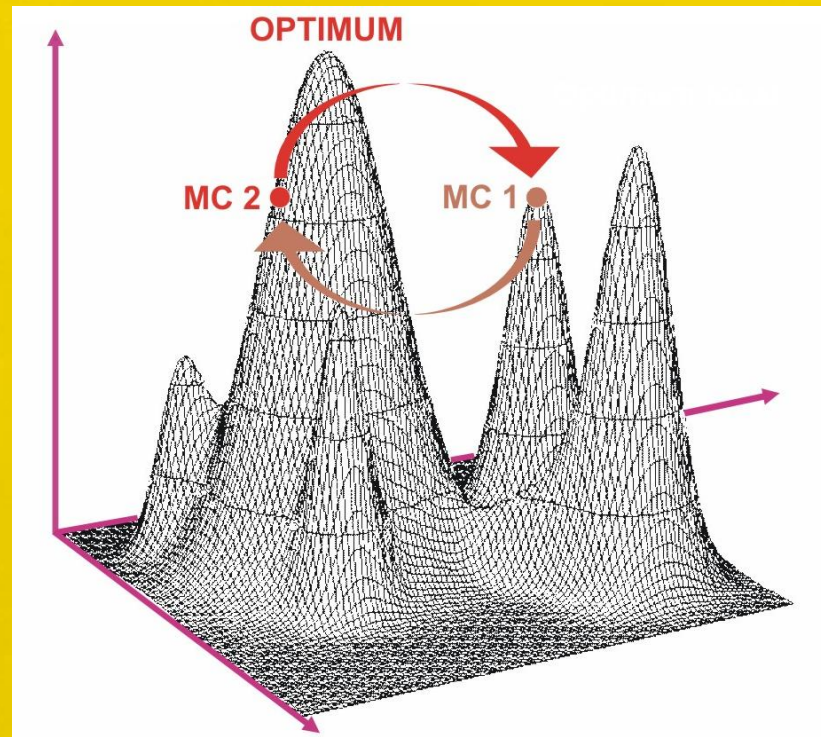
Total likelihood of the data given all possible hypotheses

➔ MCMC pour approximer la probabilité postérieure d'une phylogénie en échantillonnant des arbres au hasard à partir de leur distribution postérieure

# MCMC: le principe



- ⇒ On modifie aléatoirement un des paramètres de  $H$  (paramètre  $\alpha$  de la distribution Gamma, longueur d'une branche de l'arbre, ...).
- ⇒ Les valeurs de likelihood sont comparées (si nouvel arbre meilleur, accepté comme nouvelle probabilité *à priori*, passages à des valeurs plus faibles tolérées).
- ⇒ La séquence des arbres visités forme une chaîne MCMC.
- ⇒ Visiter le plus grand nombre d'arbres possibles (en général 1'000'000 générations).



This allows the exploration of the space of all possible values for all parameters of  $H$ , centered around the best values for each of these parameters, but not strictly converging to an optimum as it would be the case with the ML approach

The analysis must be run long enough so that the number of visited trees is sufficient i.e. the space of all possible values for all parameters is exhaustively explored

# MrBayes 3.1.2

<http://mrbayes.csit.fsu.edu/>

## Demonstration

# Phylobayes

## A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process

*Nicolas Lartillot and Hervé Philippe*

*Mol. Biol. Evol.* 21(6):1095–1109. 2004

DOI:10.1093/molbev/msh112

Advance Access publication March 10, 2004

Nouveau programme utilisant une approche bayésienne

Ne fonctionne que pour des alignements en protéine

Version actuelle: bons résultats pour < 50 espèces et entre 1000 et 10'000 acides aminés. Au-delà, le temps nécessaire pour une analyse robuste devient extrêmement long

Très intéressant car implémente un nouveau modèle d'évolution: le modèle **CAT**

Disponibilité: [http://www.lirmm.fr/mab/article.php3?id\\_article=329](http://www.lirmm.fr/mab/article.php3?id_article=329)

# Phylobayes: le modèle CAT

Heterogeneity of amino acid replacement process across sites

Bien qu'il existe 20 acides aminés, seulement 2 à 4 résidus différents sont généralement observés à chaque site → chaque site est caractérisé par une grande spécificité biochimique

➔ La plupart des positions ont subies des substitutions multiples parmi un sous-set de l'alphabet complet

➔ Importantes conséquences sur l'homoplasie générale car il est plus probable d'observer une évolution convergente vers un même acide aminé (matrices empiriques classiques assument à chaque site les 20 aa avec une prob. égale aux fréquences d'équilibre)

Le modèle CAT permet de grouper chaque position d'un alignement en des catégories biochimiques particulières, chacune étant décrite par son propre profile de fréquences d'équilibre