

# Bayesian phylogenetic inference

## **Comparative Genomic Analyses**

Rui Borges  
Vetmeduni Vienna

# Bayesian inference

Bayesian inference: way to reason about probabilities.

- ▶ nothing more than a probability analysis
- ▶ a mathematical formalization of a decision process
- ▶ constitutes a different interpretation of probability

Thomas Bayes (1702-1761)



# Bayesian inference

Bayesian approach to probability has some unique aspects.

- ▶ prior beliefs
- ▶ information from the data
- ▶ the idea of updated probability

# Bayesian inference

The Bayes' theorem or Bayes' rule is the fundamental formula of Bayesian inference.

$$p(\theta|D) \propto p(\theta)p(D|\theta)$$

- ▶  $p(\theta)$ : **prior distribution**
- ▶  $p(D|\theta)$ : **likelihood**
- ▶  $p(\theta|D)$ : **posterior distribution**

The posterior distribution specifies the the probability after the prior has been updated with the available data.

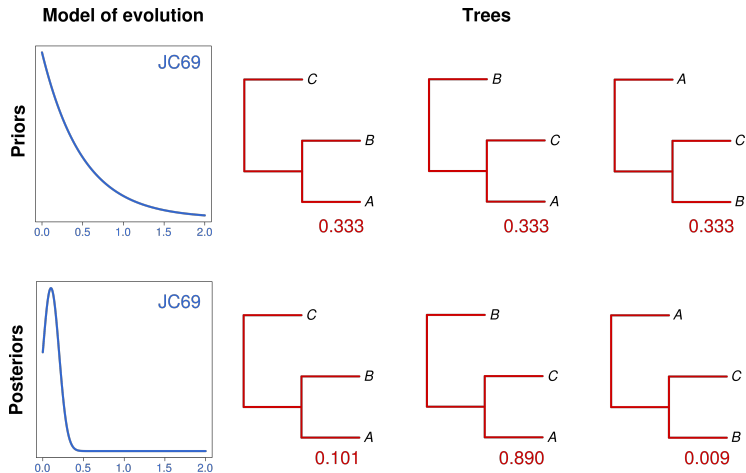
# Bayesian phylogenetic inference

The Bayes' theorem translates straightforwardly to tree inference problems.

$$p(\tau, \theta | D) \propto p(\tau, \theta) p(D | \tau, \theta)$$

- ▶  $D$  is the molecular sequence alignment
- ▶  $\tau$  and  $\theta$  represent the tree and the model of evolution parameters

# Bayesian phylogenetic inference



# Bayesian phylogenetic inference

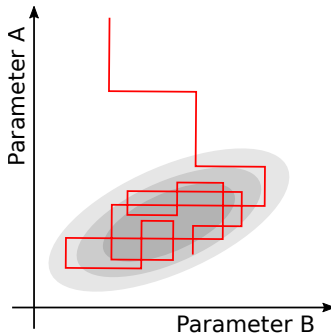
Estimating the posterior distribution in a phylogenetic context can be difficult.

- ▶ impossible to derive  $p(\tau, \theta|D)$  analytically
- ▶ concentrated in a small part of a vast parameter space

# Markov chain Monte Carlo

The posterior distribution is estimated using Markov chain Monte Carlo (or **MCMC**) sampling.

- ▶ set up a Markov chain that converges to the posterior probability distribution
- ▶ MCMC represents random samples from the posterior

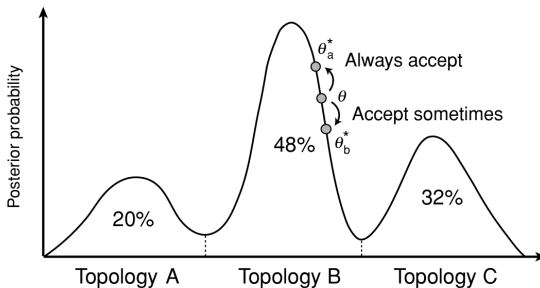




# Markov chain Monte Carlo

Metropolis-Hastings algorithm is an MCMC method.

- ▶ make small random changes on the parameter values
- ▶ accept or reject those changes according to the appropriate probabilities



# Markov chain Monte Carlo

An MCMC run is a random sample of the posterior distribution.

- ▶ the amount of time it spends sampling a particular region is proportional to the posterior probability of that region given that it converged to the target distribution
- ▶ convergence needs to be monitored

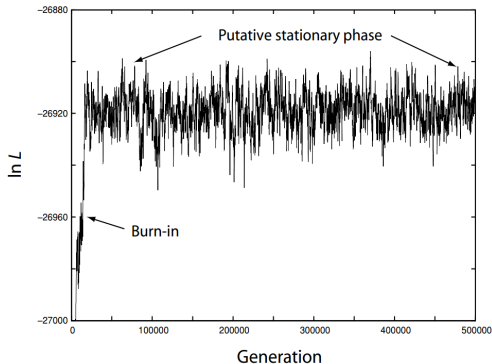
# Markov chain Monte Carlo

## Burn-in:

- ▶ early phase of the run
- ▶ heavily influenced by the starting points
- ▶ likelihood increases very rapidly

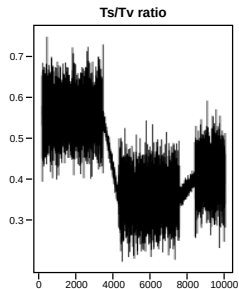
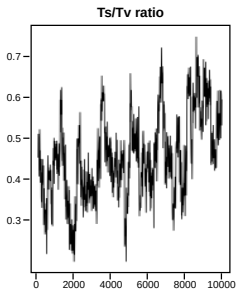
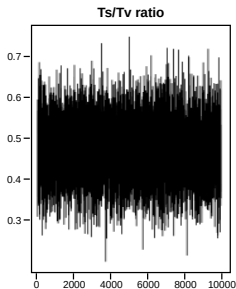
## Stationary phase:

- ▶ the chain approaches its stationary distribution
- ▶ the likelihood values reach a plateau



# Markov chain Monte Carlo

Checking the convergence of MCMC with trace plots:



# Summarizing the posterior distribution

The model parameters and the tree are summarized differently:

- ▶ model parameters are usually continuous and can be summarized as any usual statistics: means, median, the credibility interval
- ▶ trees are more difficult to summarize
- ▶ **posterior clade probabilities**: the sum of the posterior probabilities of all trees that contain that clade

# Summarizing the posterior distribution

## Exercise

Bayesian phylogenetic inference in a sequence alignment with five species returned the three topologies with the following posterior probabilities (P.p.):

Topology	P.p.
(( (Human,Dog) , (Chicken,Lizard) ), Frog)	0.76
(( ((Human,Dog), Chicken) , Lizard) , Frog)	0.17
(( (Human,Dog), Chicken) , (Lizard,Frog))	0.07

What is the posterior probability of the following clades: (Chicken,Frog), (Chicken lizard), ((Human,Dog),Chicken) and (Human,Dog)?

# Bayesian *versus* maximum likelihood trees

## Maximum-Likelihood trees

- ▶  $p(D|\tau, \theta)$
- ▶ Maximum likelihood tree
- ▶ ignores pre-existing information
- ▶ bootstrapping
- ▶ resample characters

## Bayesian trees

- ▶  $p(\tau, \theta|D)$
- ▶ Maximum a-posteriori tree
- ▶ considers pre-existing information
- ▶ MCMC
- ▶ resample parameters

# Literature

**The Phylogenetic Handbook** by Lemey, Salemi and Vandamme (2009)  
Cambridge University Press

- ▶ Chapter 7: sections 7.1, 7.2 and 7.3, 7.4, 7.6, 7.7 and 7.9