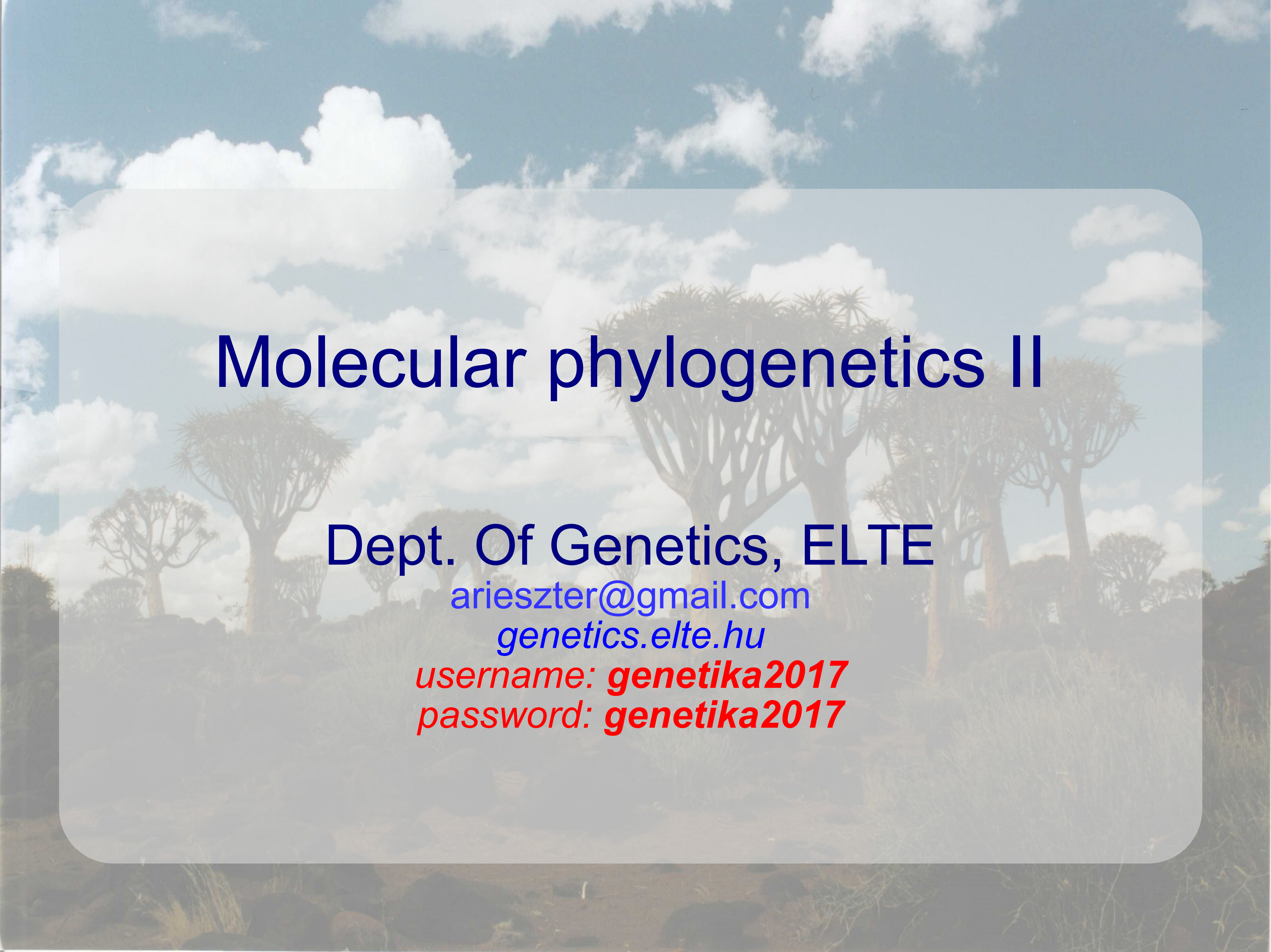


Molecular phylogenetics II

Dept. Of Genetics, ELTE
arieszter@gmail.com
genetics.elte.hu

username: ***genetika2017***
password: ***genetika2017***

What will we talk about?

- Introduction to probability modeling
- Character based methods 2:
  - Maximum likelihood
  - Bayesian inference
    - MCMC

Steps of molecular phylogenetic analyses

1. Input data: multiple aligned sequences
2. Phylogenetic methods:

    Choosing the best substitution model

    Distance based method OR

    Character based methods

3. Find or calculate the best tree
4. Estimate the reliability, robustness of the tree

Molecular phylogenetic methods

- Distance based method

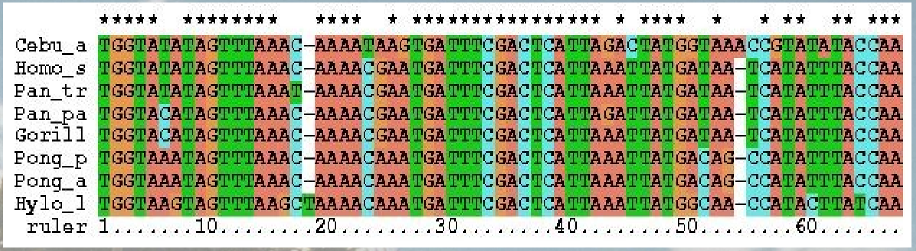
  Neighbor-joining
  (Saitou & Nei, 1987)

- Character based methods

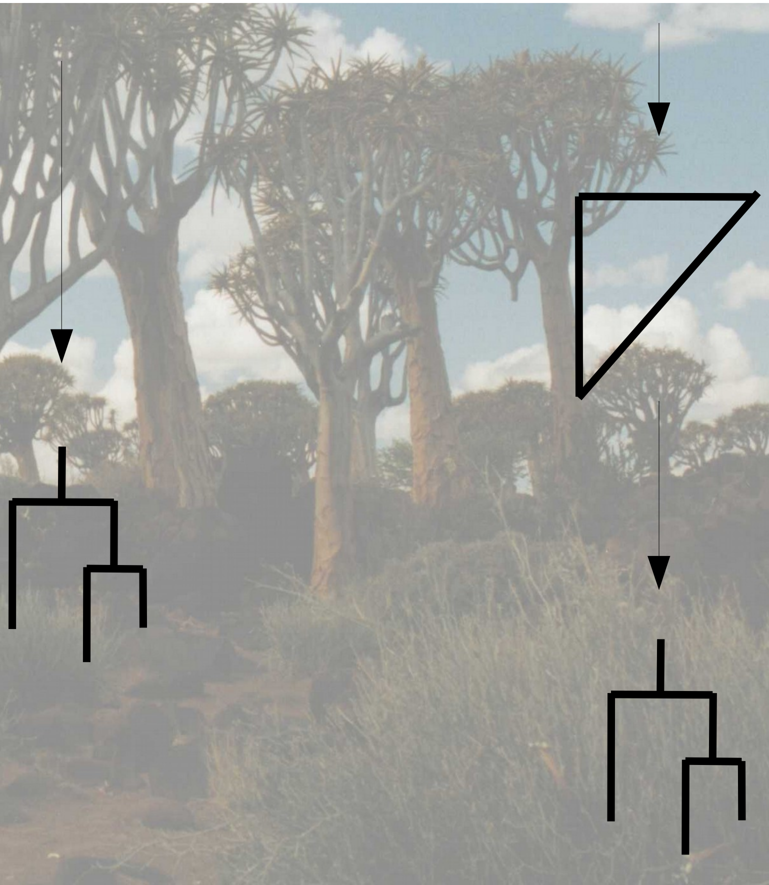
  Maximum parsimony
  (Fitch, 1971)

  Maximum likelihood
  (Felsenstein, 1981)

  Bayeian inference
  (Rannala & Yang, 1996)

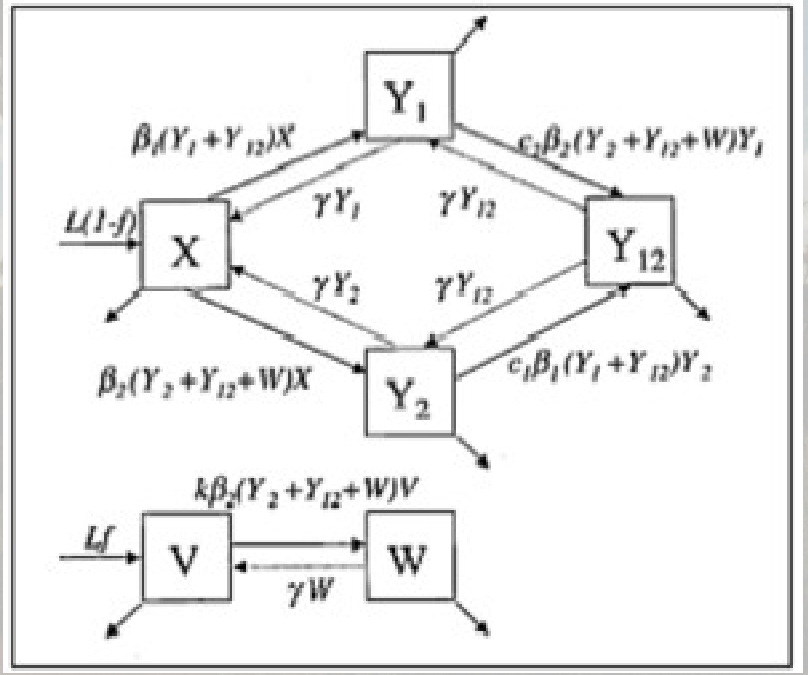


```
       *****  *********  ****  * **************** * ****  *   * **  ** ***
Cebu_a TGGTATATAGTTTAAAC-AAAATAAGTGATTTCGACTCATTAGACTATGGTAAACCGTATATACCAA
Homo_s TGGTATATAGTTTAAAC-AAAACGAATGATTTCGACTCATTAAATTATGATAA-TCATATTTACCAA
Pan_tr TGGTATATAGTTTAAAT-AAAACGAATGATTTCGACTCATTAAATTATGATAA-TCATATTTACCAA
Pan_pa TGGTACATAGTTTAAAC-AAAACGAATGATTTCGACTCATTAGATTATGATAA-TCATATTTACCAA
Gorill TGGTACATAGTTTAAAC-AAAACGAATGATTTCGACTCATTAAATTATGATAA-TCATATTTACCAA
Pong_p TGGTAAATAGTTTAAAC-AAAACAAATGATTTCGACTCATTAAATTATGACAG-CCATATTTACCAA
Pong_a TGGTAAATAGTTTAAAC-AAAACAAATGATTTCGACTCATTAAATTATGACAG-CCATATTTACCAA
Hylo_l TGGTAAGTAGTTTAAGCTAAAACAAATGATTTCGACTCATTAAATTATGGCAA-CCATACTTATCAA
 ruler 1........10........20........30........40........50........60......
```



What is a model?

- Mathematical models are:
  - Incomprehensible
  - Useless
  - No fun at all

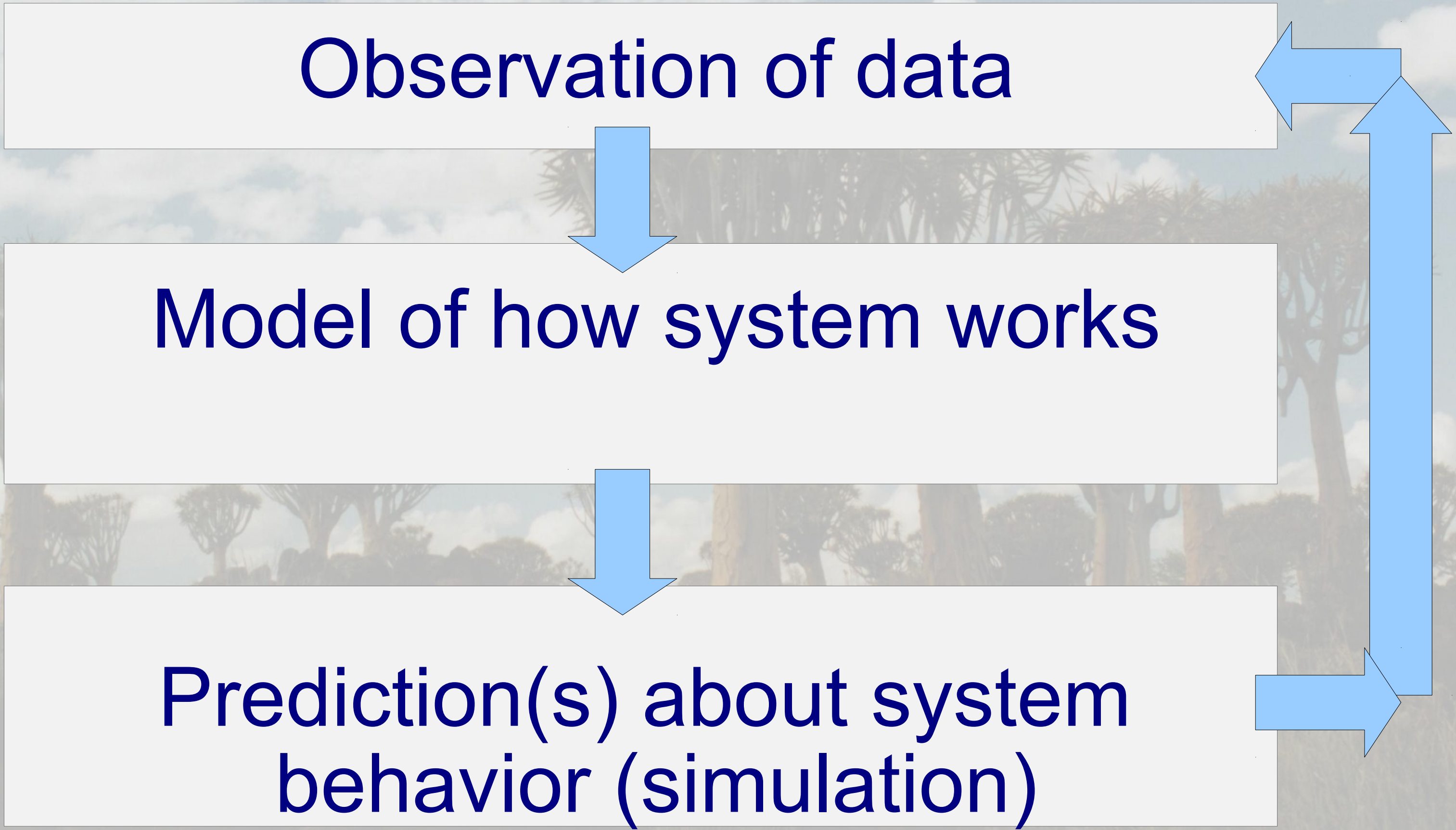


What is a model?

- A matematikai modellek:
  - ~~Felfoghatatlanok~~
  - ~~Hiábavalóak~~
  - ~~Unalmasak~~

- Model = hypothesis !!!
- Hypothesis (as used in most biological research):
  - Precisely stated, but qualitative
  - Allows you to make qualitative predictions
- Arithmetic model:
- Mathematically explicit (parameters)
- Allows you to make quantitative predictions

The Scientific Method



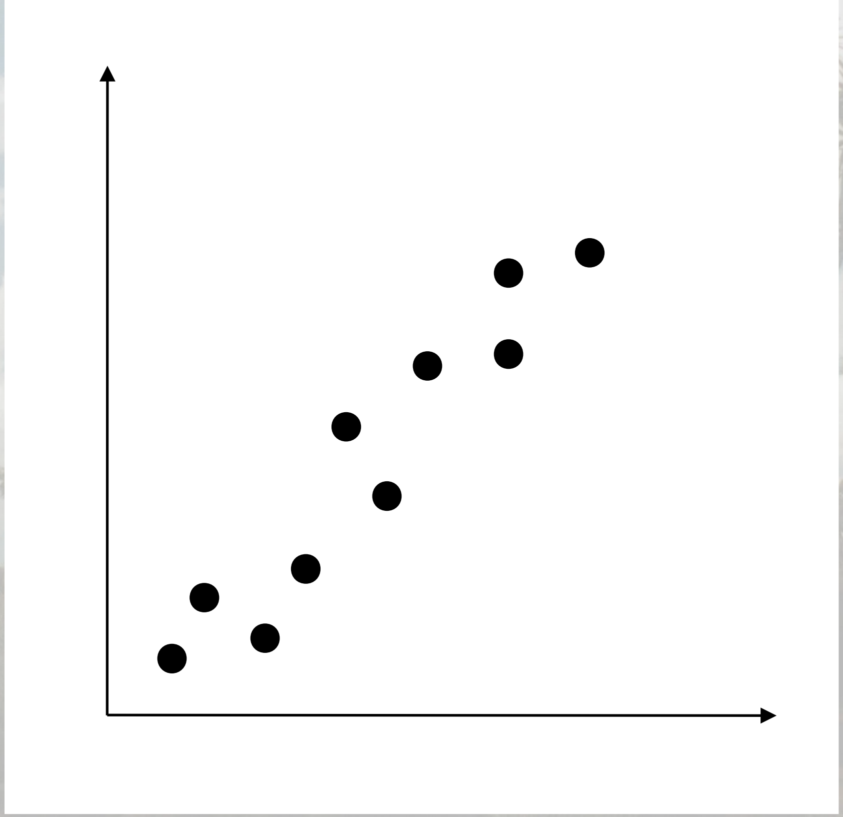
Observation of data

Model of how system works

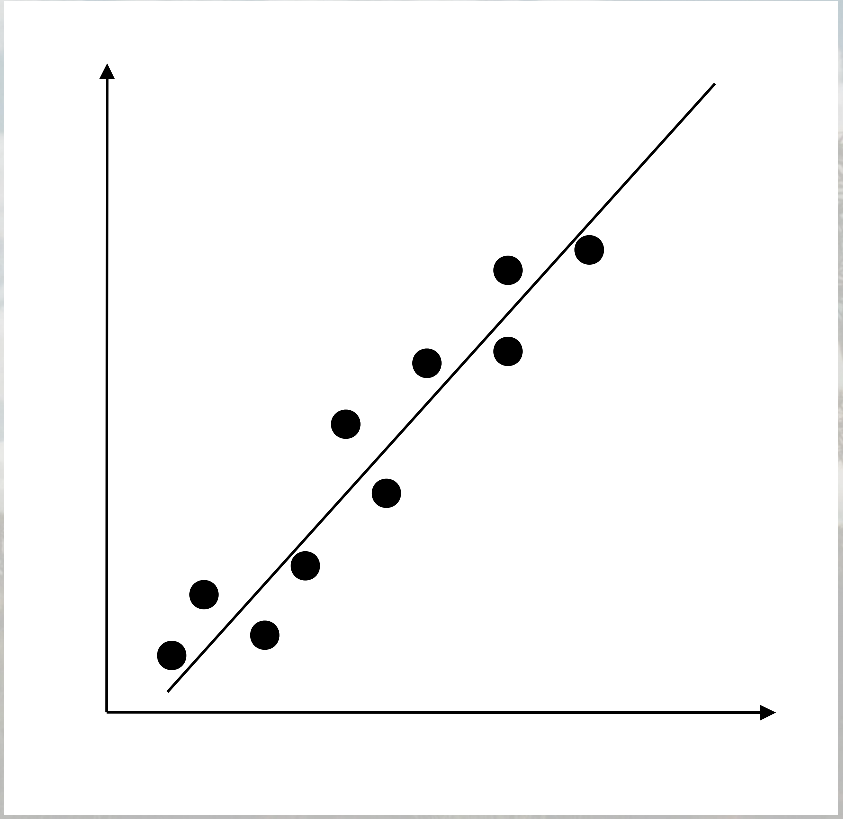
Prediction(s) about system behavior (simulation)

Modeling: An example

Observed data:



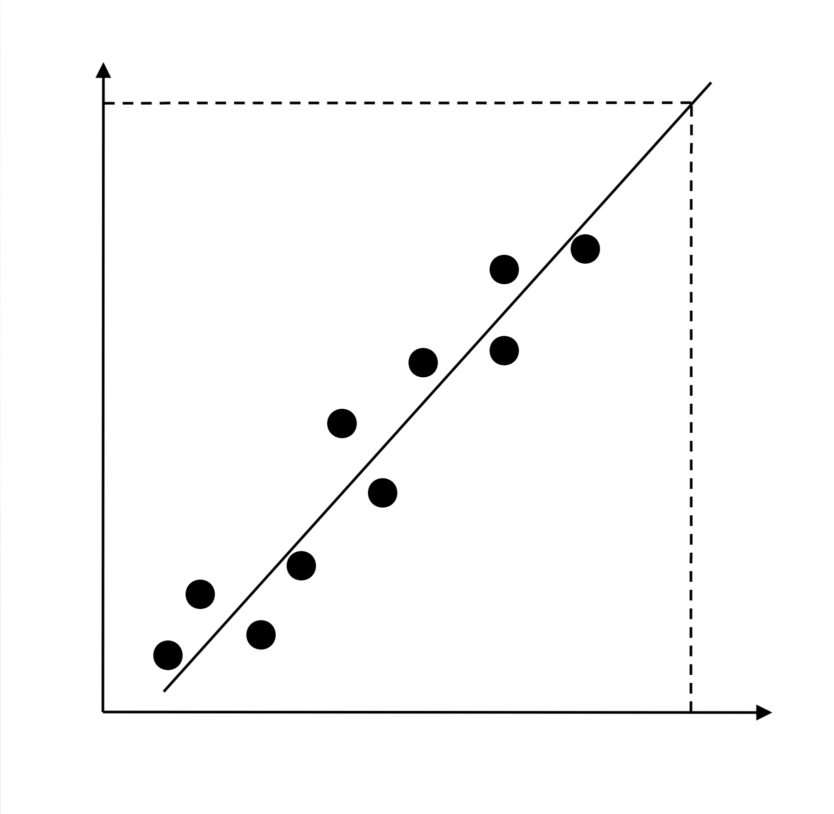
Modeling: An example



$y = ax + b$

Simple 2-parameter model

Modeling: An example



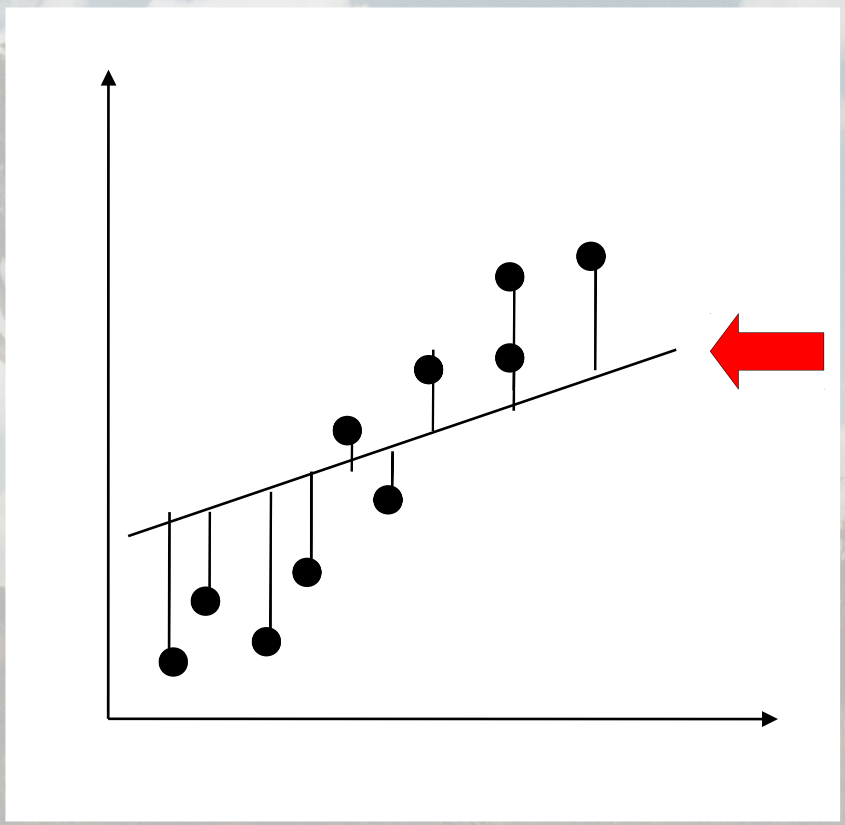
$y = ax + b$

Predictions based on model

Model Fit, parameter estimation

Measure of how well the model fits the data: sum of squared errors (SSE)

- Best parameter estimates: those that give the smallest SSE (least squares model fitting)

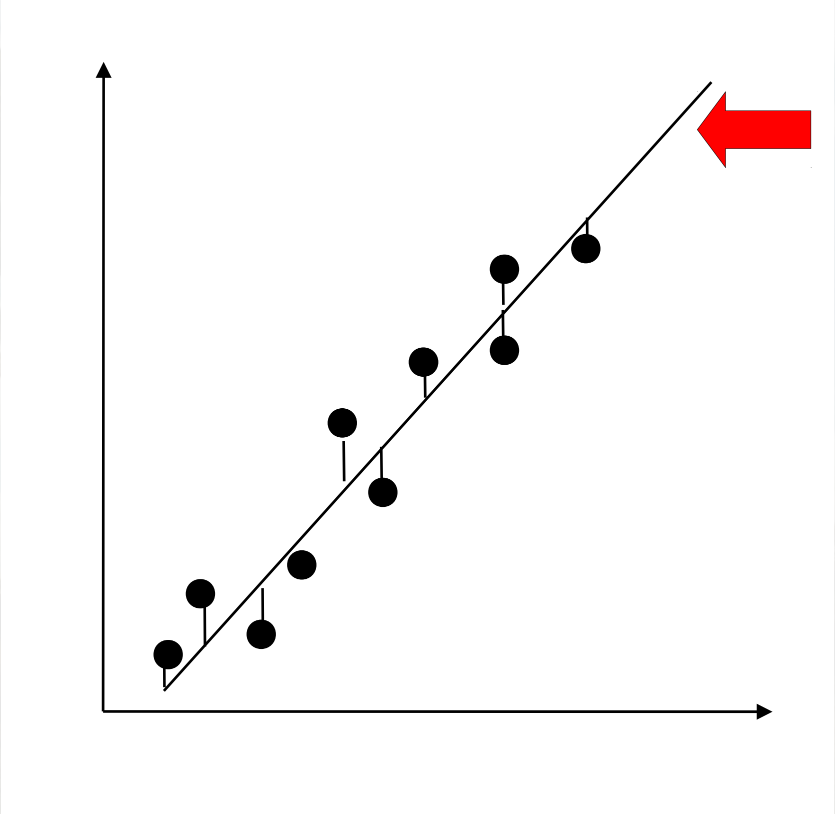


y = ax + b

Model Fit, parameter estimation

Measure of fit between model and data: sum of squared errors (SSE)

- Best parameter estimates: those that give the smallest SSE (least squares)



y = 1,24x - 0,56

The Maximum likelihood estimation

In statistics, maximum likelihood (ML) estimation is a method of **estimating the parameters of a statistical model given observations**, by finding the parameter values that maximize the likelihood of making the observations given the parameters.

Likelihood = P (Data | Model) = Probability of the data given the model. (| : conditional probability)

Maximum likelihood: Those parameters of the model which give the highest likelihood value to the data.

For a fixed set of data and underlying statistical model, the method of maximum likelihood selects the set of values of the model parameters that maximizes the likelihood function.

Maximum likelihood – coin tossing

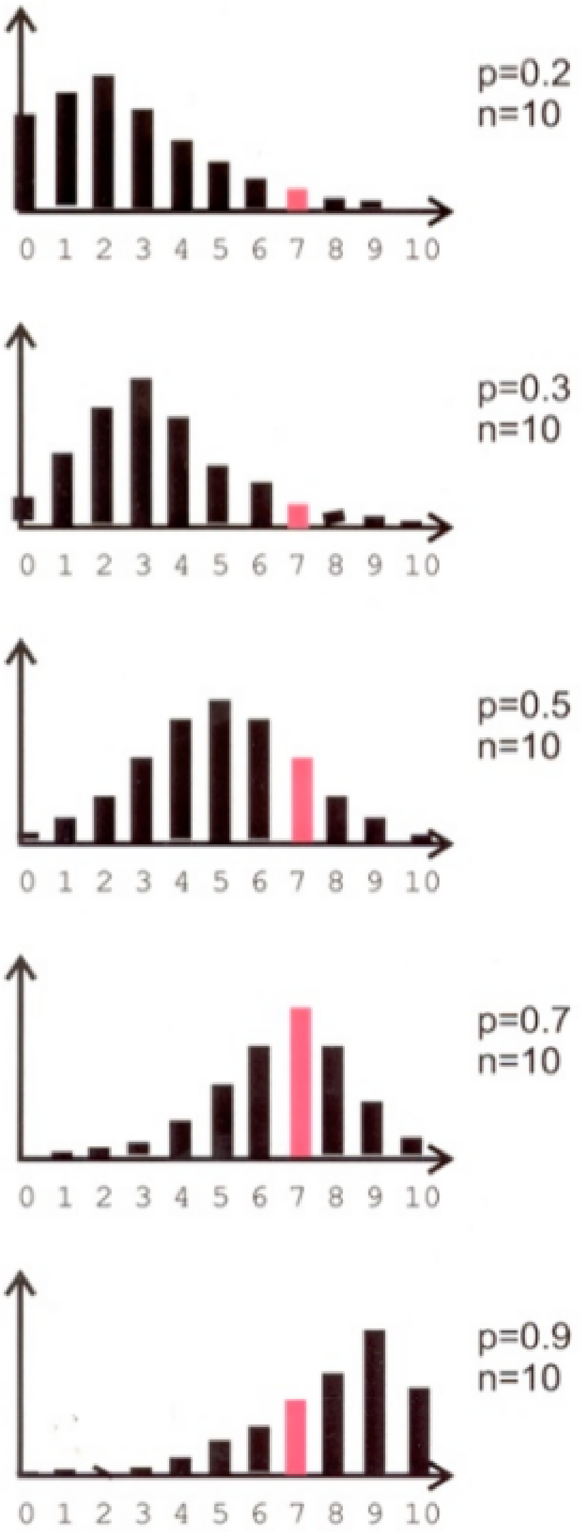
Starting point:

- You have some observed data and a probabilistic model for how the observed data was produced

- Example:
  - Data: result of tossing coin 10 times - 7 heads, 3 tails
  - Model: coin has probability p for heads, 1-p for tails.
  - The probability of observing h heads among n tosses is:

$$P(h \text{ heads}) = \binom{h}{n} p^h (1-p)^{n-h}$$

- Goal:
  - You want to find the best estimate of the (unknown) parameter value based on the observations.
    - (here the only parameter is “p”)



→ Probability distribution for possible outcomes when value of p-parameter=0.2 and n=10 tosses of coin.

Probabilities sum to 1.

Likelihood of p having the value 0.2 given that we observed x=7 heads:

$L(p=0.2 \mid x=7)$ =

$Pr(x=7 \mid p=0.2) = 0.001$

→ p=0.7 is the maximum likelihood estimate of p given that we observed x=7 heads.

Note that the likelihoods $L(p \mid x=7)$ do not necessarily sum to 1

Probabilistic modeling applied to phylogeny

Observed data: multiple alignment of sequences

```
H.sapiens globin: A G G G A T T C A
M.musculus globin:A C G G T T T - A
R.rattus globin:  A C G G A T T - A
```

Probabilistic model parameters (simplest case):

- Tree topology and branch lengths
- Nucleotide-nucleotide substitution rates (or substitution probabilities)
- Nucleotide frequencies: πA, πC, πG, πT

The Maximum likelihood method

Likelihood provides probabilities of the sequences given a model of their evolution on a particular tree.

- The more probable the sequences given the tree, the more the tree is preferred.
- The algorithm will choose that tree (topology and branch lengths) and model parameter values where the likelihood was the greatest.
- All (or a lot) possible trees are considered; computationally intense → *Heuristic methods are applied*
- Because the user can choose a model of evolution, the method can be useful for widely divergent groups or other difficult situations.

The Maximum likelihood method

Not like in the case of Maximum parsimony, Maximum likelihood will considere all sites of the aligned sequences.

The method requires a substitution model to assess the probability of particular mutations.

It will choose the best tree based upon its likelihood value.

The Maximum likelihood method

The likelihood of hypothesis *H*: $L_H = P(D|H)$

D: probability of data given hypothesis H

Likelihood of a tree (here the tree is the hypothesis):

$$L = P(D \mid T) = \prod_{i=1}^{m} P\left(D^{(i)} \mid T\right)$$

$D^{(i)}$: Data at site *i*

We take the log or ln of the Likelihood (it is easier to handle).

The ML of 2 sequences

Sites	1	2
sequence X:	A	C
sequence Y:	G	C

$$L_{AG(t)} = f_A P_{AG(t)}$$

$$L_{XY(t)} = \prod_{i=1}^{s} f_{x_i} P_{x_i,y_i}(t)$$

$P_{AG(t)}$: Probabiliy that *A* become *G* in time *t*.

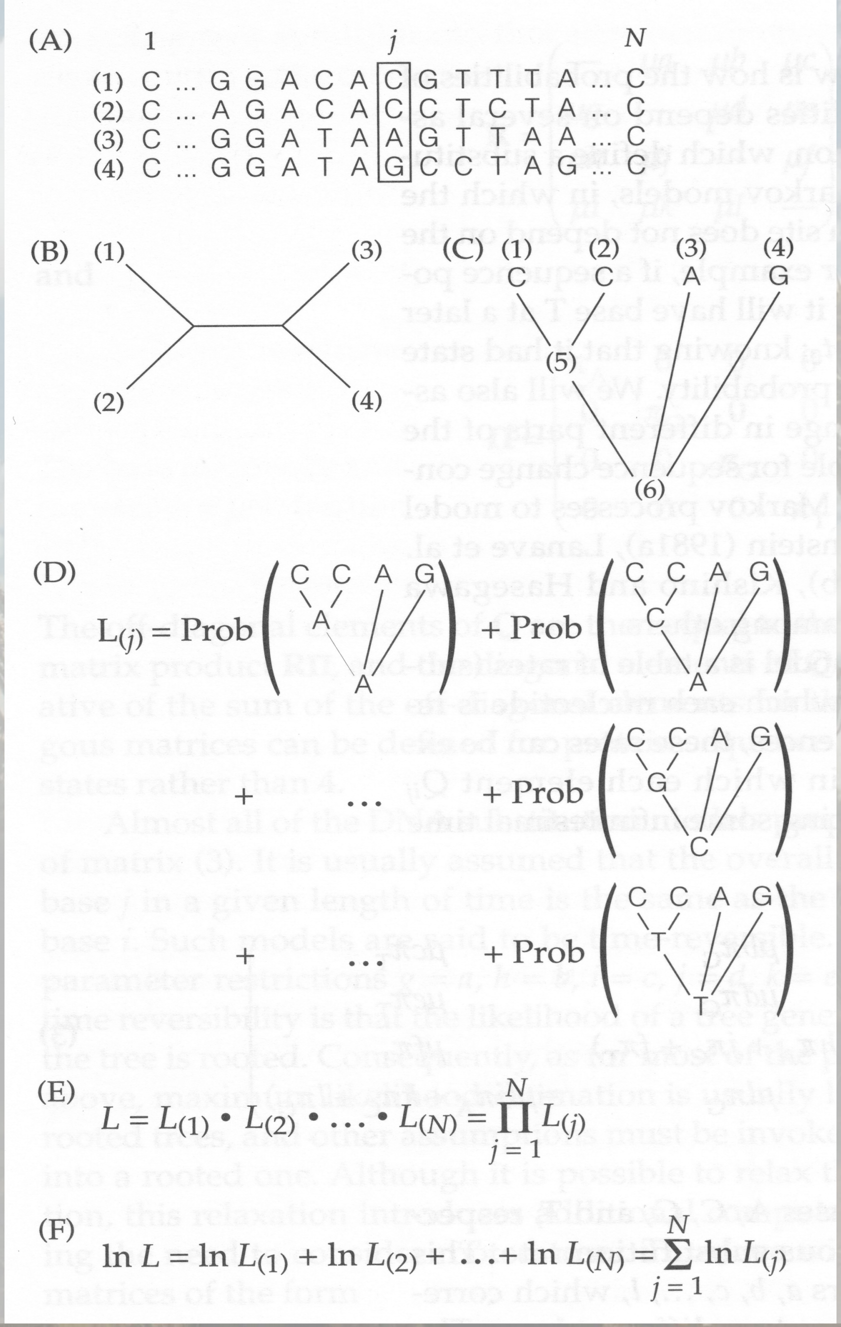
f_A: The frequency of *A*

$L_{XY(t)}$: Likelihood that *sequence X* become *sequence Y* in time *t*.

$P_{X_iY_i(t)}$: Probabiliy that the nucleotide of the *i*th site of *sequence X* become nucleotide of the *i*th site of *sequence Y* in time *t*.

s: length of sequences

Computing the probability of an entire alignment given tree topology and other parameters



(1) C ... G G A C A C G T T T A ... C
(2) C ... A G A C A C C T C T A ... C
(3) C ... G G A T A A G T T A A ... C
(4) C ... G G A T A G C C T A G ... C

$$L = L_{(1)} \cdot L_{(2)} \cdot \ldots \cdot L_{(N)} = \prod_{j=1}^{N} L_{(j)}$$

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ldots + \ln L_{(N)} = \sum_{j=1}^{N} \ln L_{(j)}$$

Probability must be summed over all possible combinations of ancestral nucleotides.

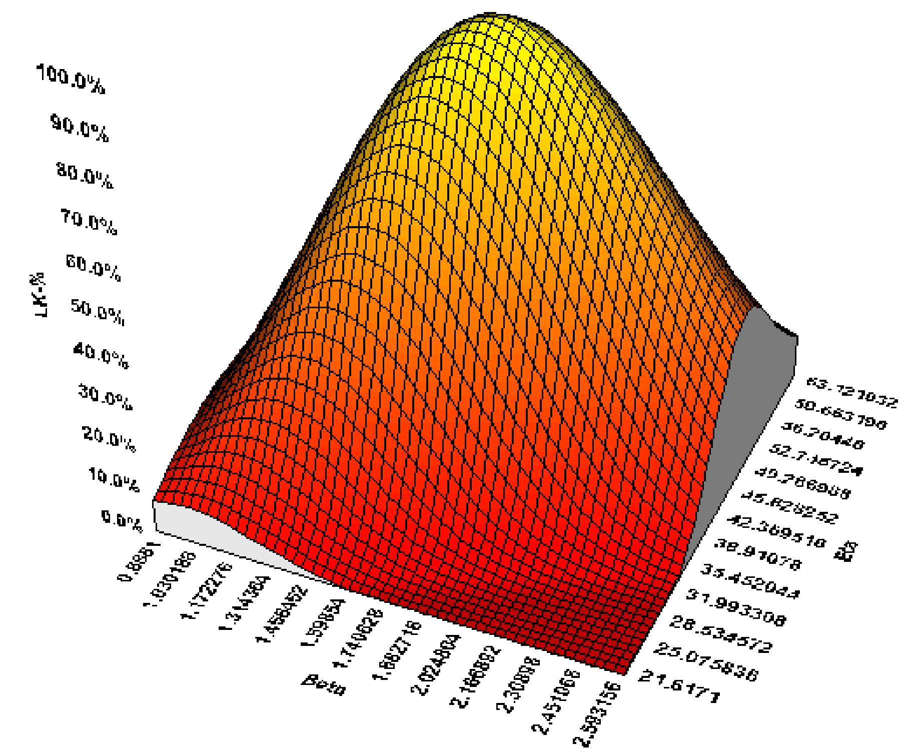
(Here we have two internal nodes giving 16 possible combinations)

Probability of individual sites are multiplied to give the overall probability of the alignment, i.e., the likelihood of the model.

Often the log of the probability is used (log likelihood)

ML phylogeny: heuristic tree search

- Data: sequence alignment
- Model parameters: nucleotide frequencies, nucleotide substitution rates, tree topology, branch lengths.



1. Choose random initial values for all parameters, compute likelihood
2. Change parameter values slightly in a direction so likelihood improves
3. Repeat until maximum found

Results:

(1) ML estimate of tree topology
(2) ML estimate of branch lengths
(3) ML estimate of other model parameters
(4) Measure of how well model fits data (likelihood).

Advantages of ML method

lower variance than other methods (i.e. estimation method least affected by sampling error)

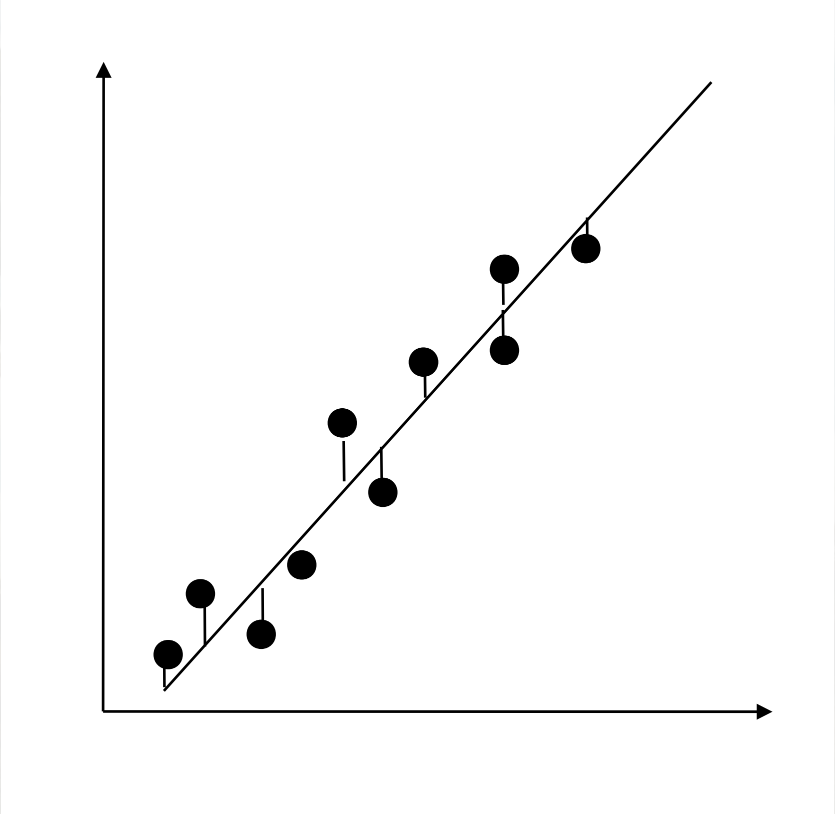
- robust to many violations of the assumptions in the evolutionary model, even with very short sequences
  - it may outperform alternative methods such as parsimony or distance methods.
- has explicit model of evolution that you can make fit the data
- evaluate different tree topologies (vs. NJ)
- use all the sequence information itself (vs. Distance)
- better accounting for branch lengths, e.g. incorporates “multiple hits” thereby providing more realistic branch length
- Also, information is derived from sites that would be uninformative under parsimony

Disadvantages of ML method

- very computationally intensive and so slow (though this is becoming much less of an issue)
- Misleading results of likelihood-based phylogenetic analyses in the presence of missing data.
- the result is dependent on the model used
- questionably applicable to complex data like morphology given the difficulty of modeling the numerous processes
- philosophically less well established, especially in terms the applicability of probabilities and statistical measures of unique historical events (vs. Parsimony as a general principle).

Model selection?

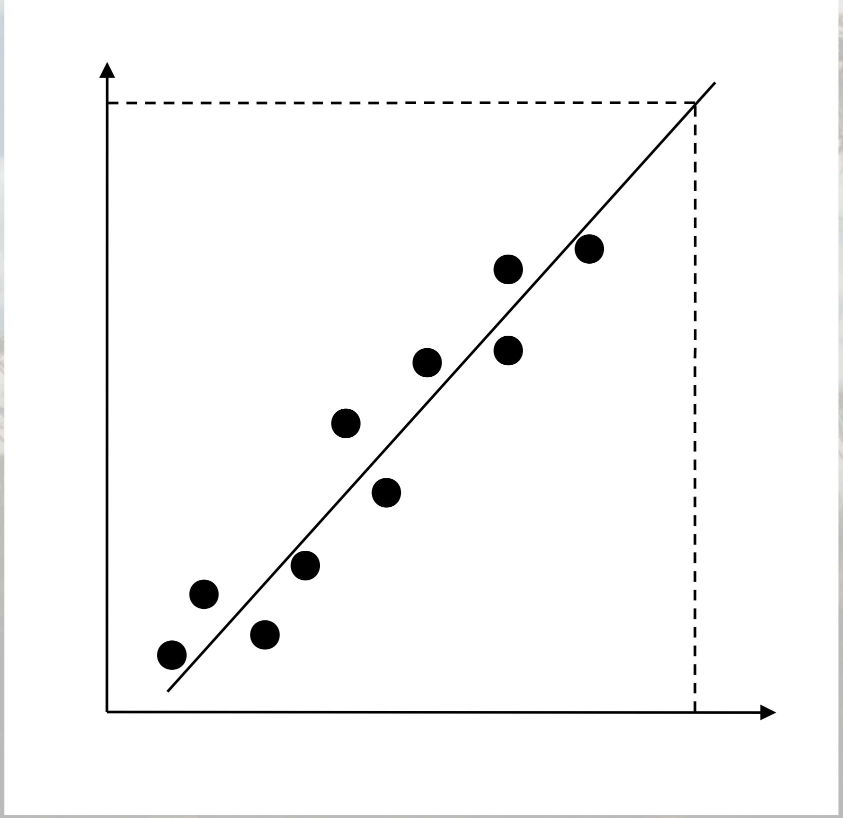
- Measure of fit between model and data (e.g., SSE, likelihood, etc.)
- How do we compare different types of models?



y = 1,24x - 0,56

Model selection?

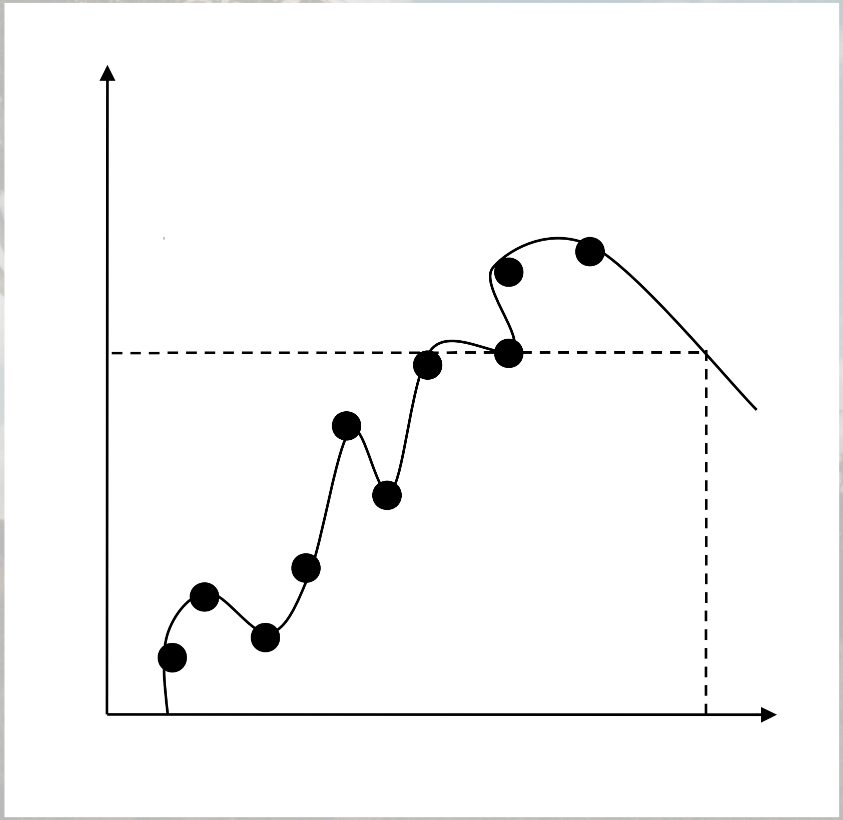
Over-fitting: More parameters always result in a better fit to the data, but not necessarily in a better description



$y = ax + b$

2 parameter model:

Good description, poor fit



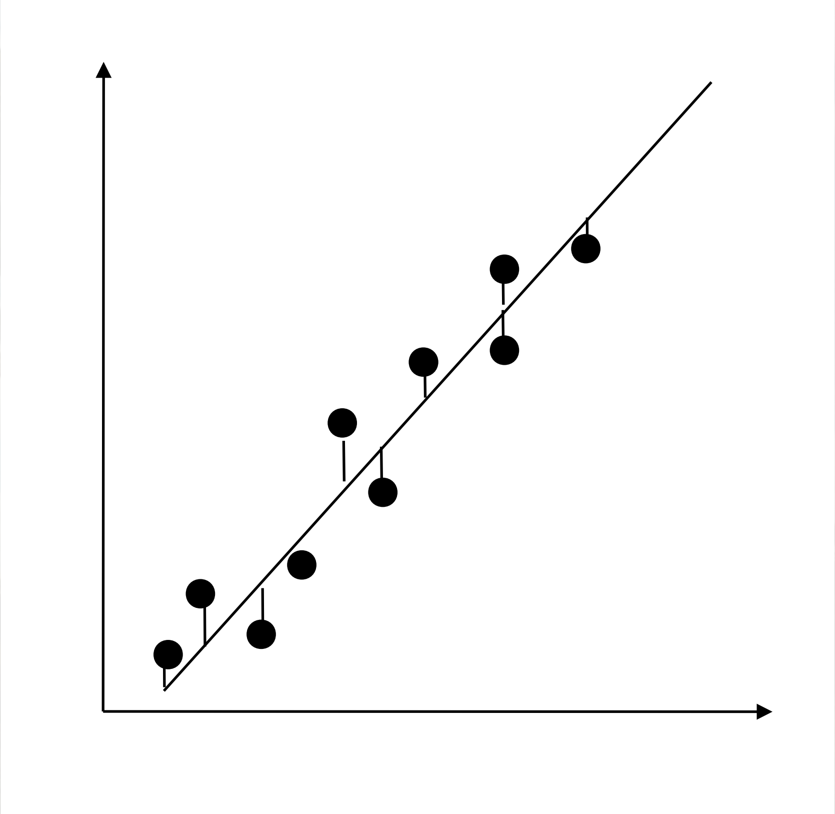
$y = ax^6+bx^5+cx^4+dx^3+ex^2+fx+g$

7 parameter model:

Poor description, good fit

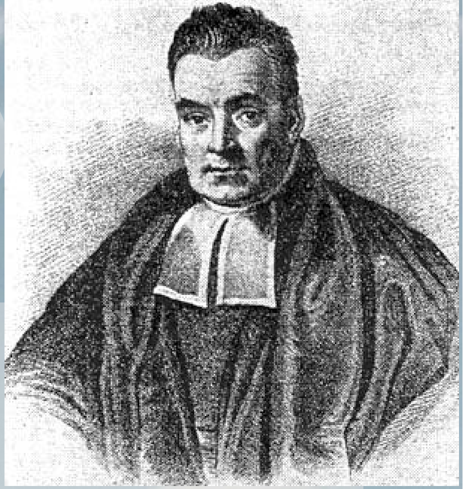
Selecting the best model

- How to compare different models?
- The model describes our data better but uses less parameters should be chosen: ie. by → *Likelihood ratio test*



y = 1,24x - 0,56

The Bayesian inference



Bayesian inference of phylogeny uses a likelihood function to create a quantity called the **posterior probability of trees**

- using a model of evolution, based on some prior probabilities
- producing the most likely phylogenetic tree for the given data.
- The Bayesian approach has become popular due to advances in computing speeds and the integration of Markov chain Monte Carlo (MCMC) algorithms.
- Bayesian inference has a number of applications in molecular phylogenetics and systematics.
- There is a popular free software to calculate Bayesian phylogenetic trees: **MrBayes** of Fredrik Ronquist, John Huelsenbeck & Paul van der Mark

The Bayes' theorem

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

- P(A|B) is the conditional probability of observing A given B is TRUE
- P(B|A) is the conditional probability of observing B given A is TRUE
- P(A) and P(B) are probabilities of A and B without conditioning on each other

Bayes – „coin tossing”

Let«s assume:

- We hava a bag of coins
- 90% of the coins are normal (50% heads, 50% tails)
- 10% of the coins are loaded (unfair; 80% heads, 20% tails)

What is the probability that we grab a loaded coin if we pull one out from the bag?

If we have no more information: the answer will be 10% (0,1).

This is the ***prior*** probability.

If it is allowed to toss the coin 10 times we may change our answer about the probability of the loaded coin.

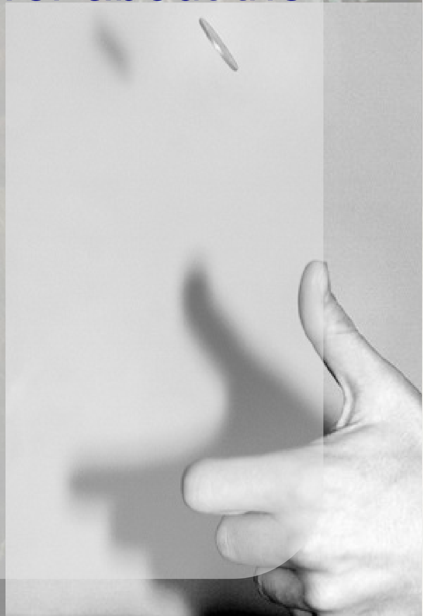
This is how we get the ***posterior*** probability.

With which we can make more precise predictions.

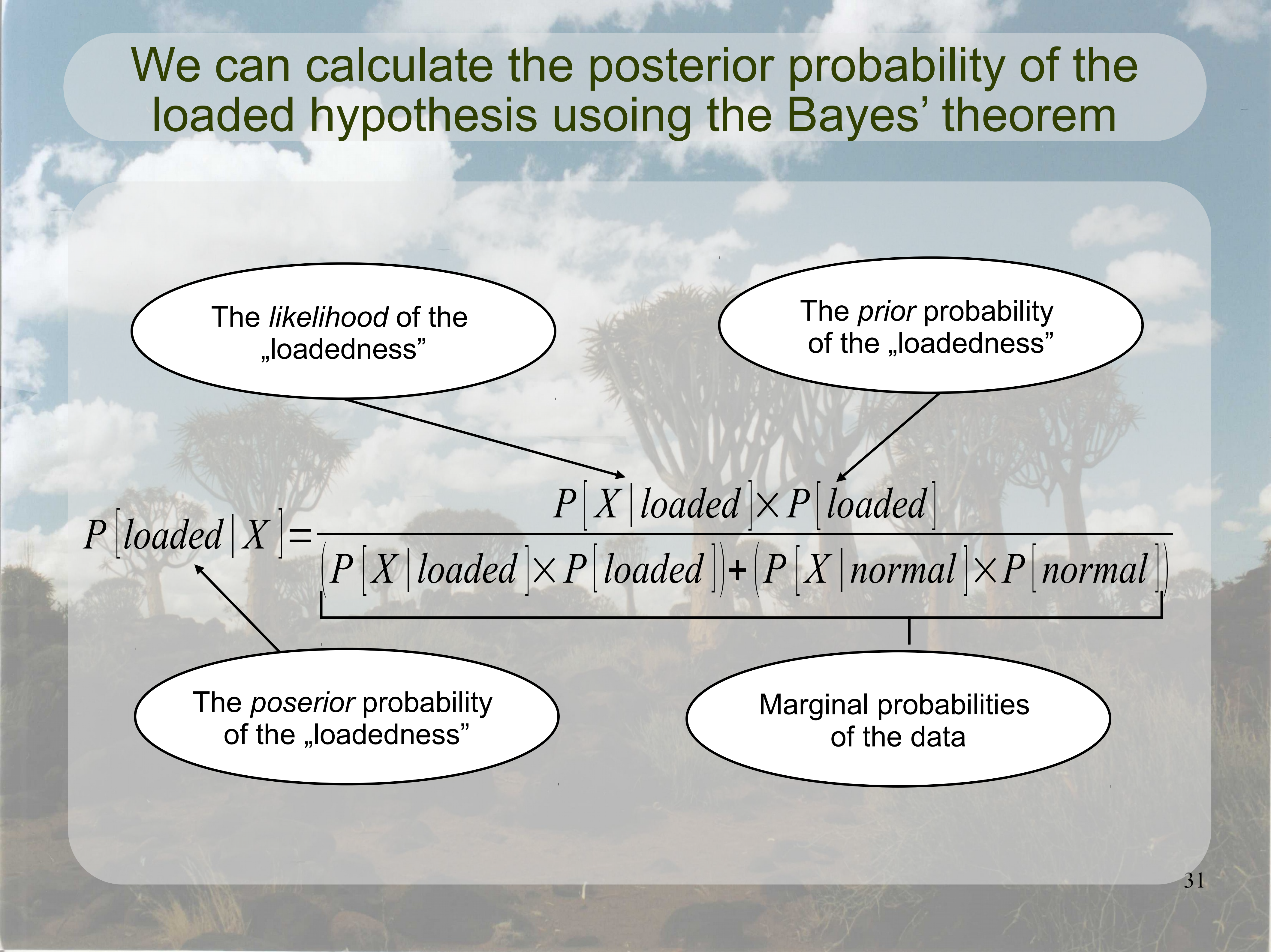
If the result of 10 tossing is: X: HHIHHIIHHH

$$P [X \mid \text{normal}] = 0.5^{10} = 9.76 \times 10^{-4}$$

$$P [X \mid \text{loaded}] = 0.8^{7} \times 0.2^{3} = 1.67 \times 10^{-3}$$



We can calculate the posterior probability of the loaded hypothesis usoing the Bayes’ theorem



The *likelihood* of the „loadedness”

The *prior* probability of the „loadedness”

$$P\left[loaded\,|\,X\right]=\frac{P\left[X\,|\,loaded\right]\times P\left[loaded\right]}{\underbrace{\left(P\left[X\,|\,loaded\right]\times P\left[loaded\right]\right)+\left(P\left[X\,|\,normal\right]\times P\left[normal\right]\right)}}$$

The *poserior* probability of the „loadedness”

Marginal probabilities of the data

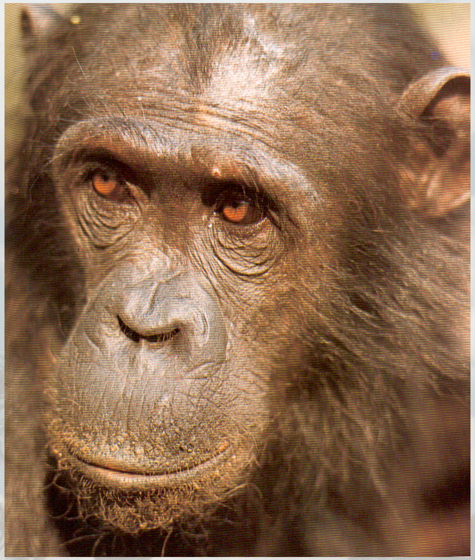
Result

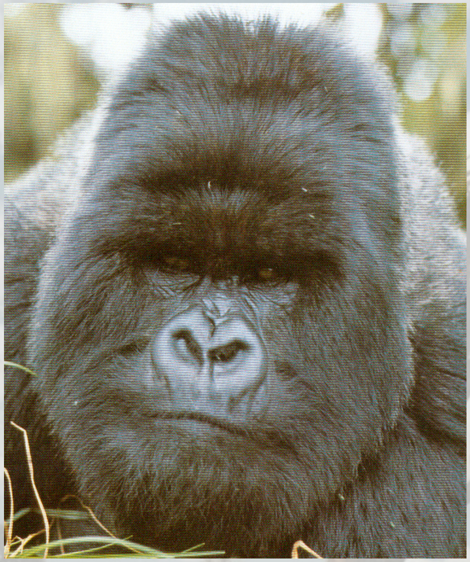
The probability that the coin was loaded given the data (series of heads and tails of 10 tossing):

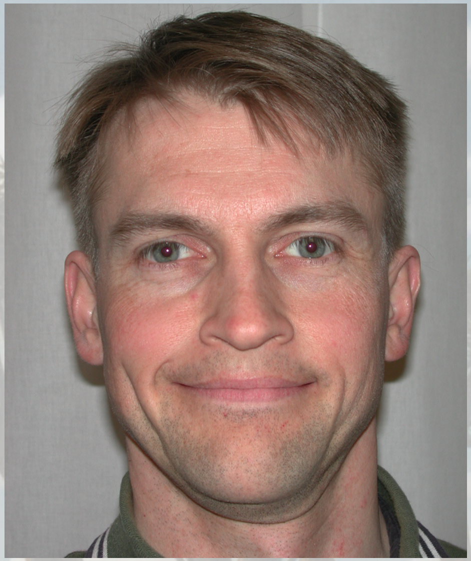
$$P\left[loaded \mid X\right] = \frac{1{,}67 \times 10^{-3} \times 0{,}1}{\left(1{,}67 \times 10^{-3} \times 0{,}1\right) + \left(9{,}76 \times 10^{-4} \times 0{,}9\right)} = 0{,}13$$

If we try to translate this to phylogeny:
X is the sequence alignment,
„loaded” are the model parameters: tree topology, branch lengths, substitution model parameters

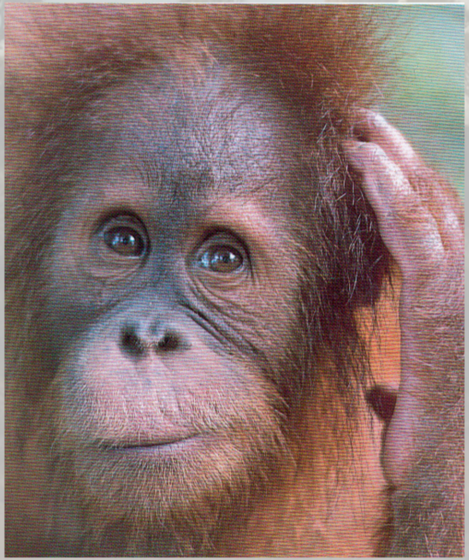
Infer relationships among three species



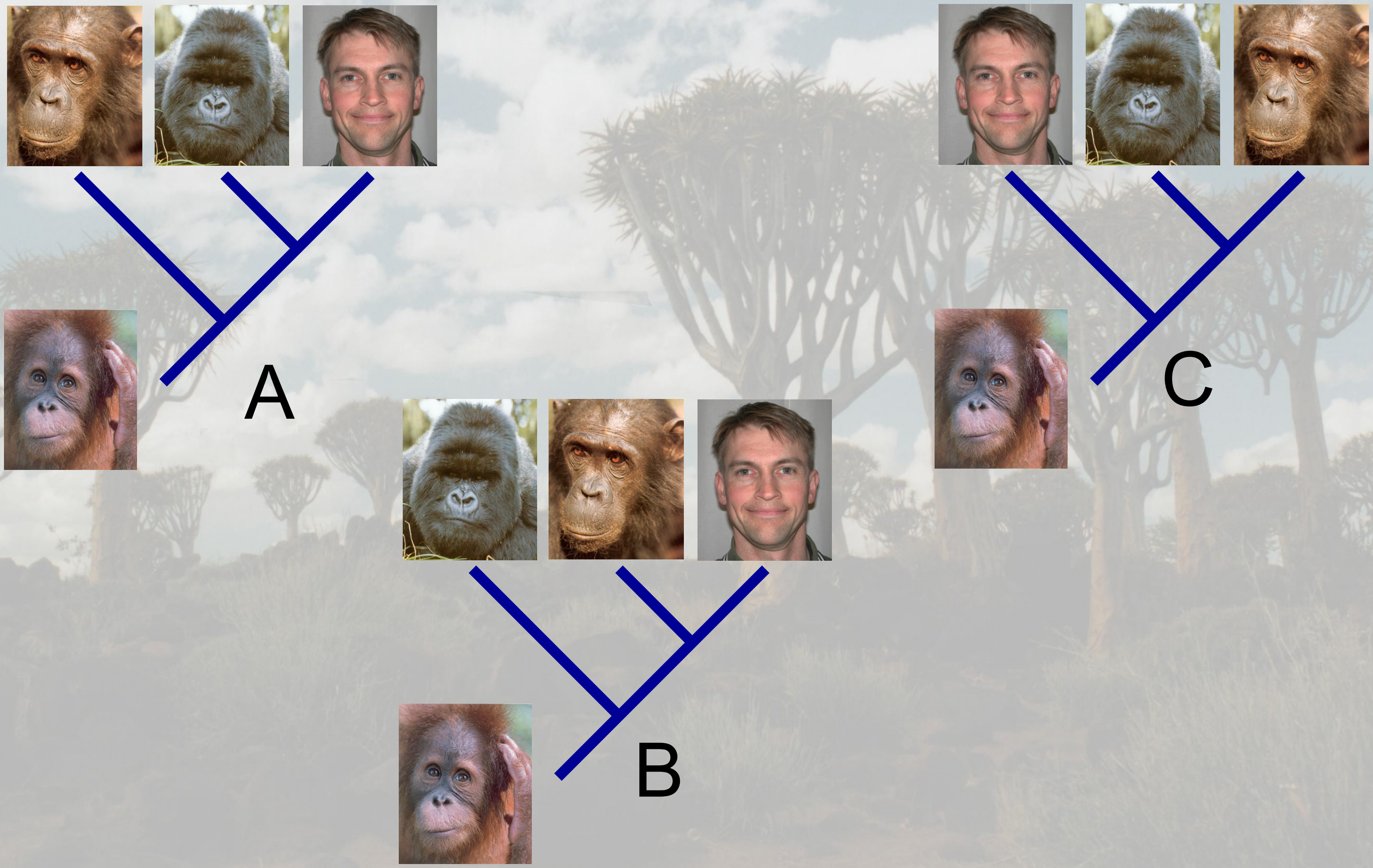


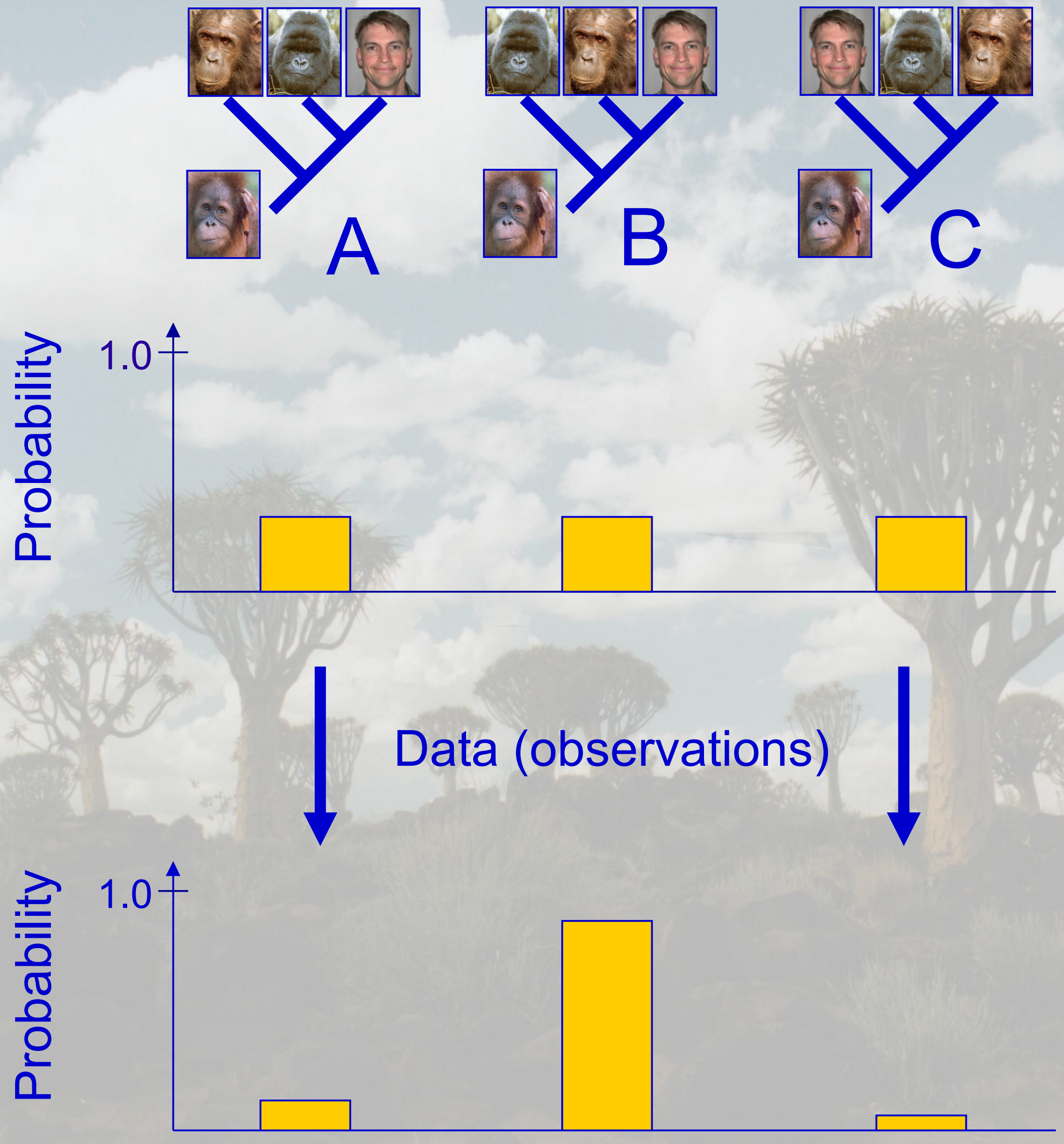


Outgroup:



Three possible trees (topologies):



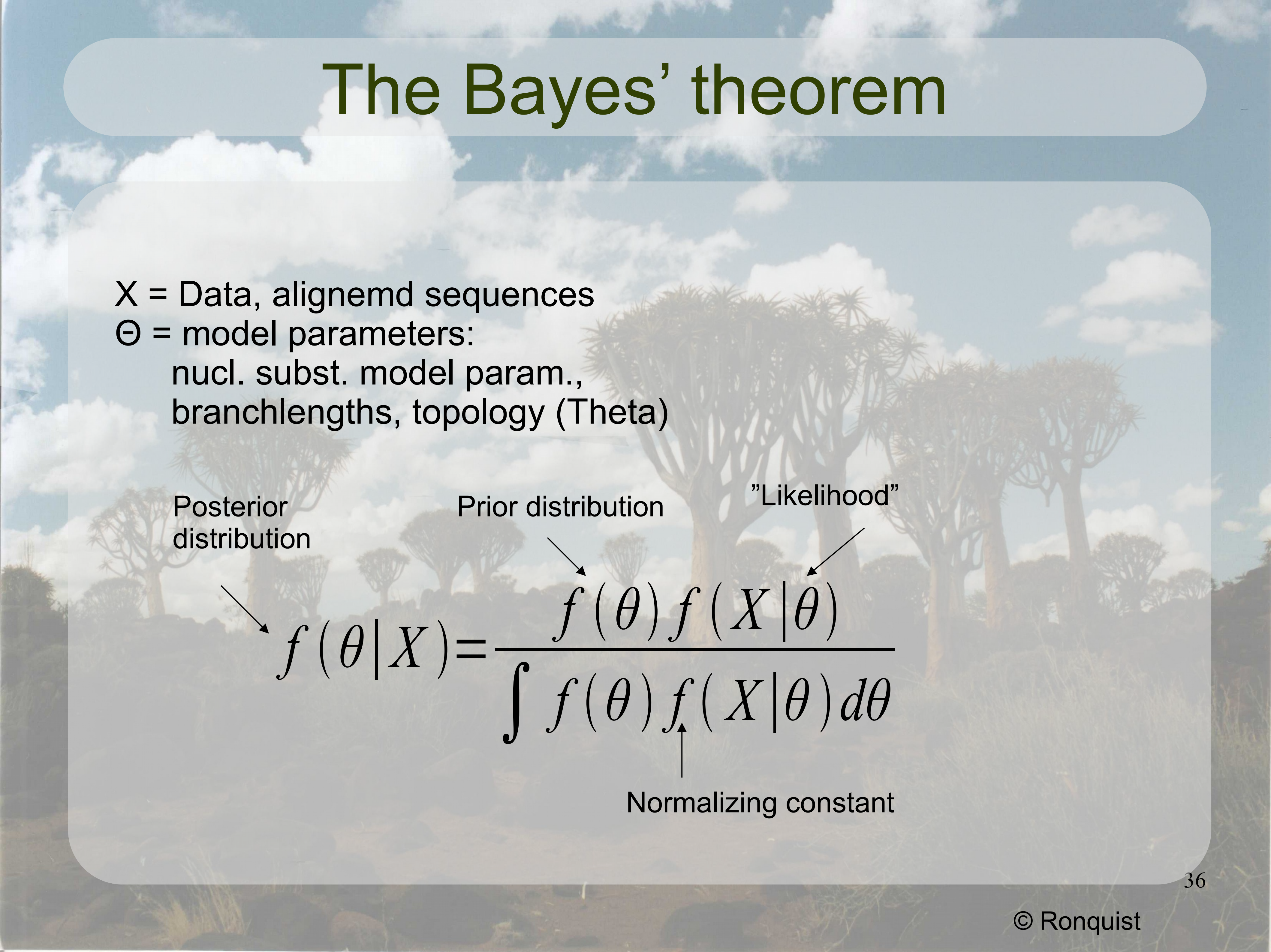


Model (Hypothesis)

Prior distribution

Data (observations)

Posterior distribution



The Bayes' theorem

X = Data, alignemd sequences
Θ = model parameters:
nucl. subst. model param.,
branchlengths, topology (Theta)

Posterior distribution

Prior distribution

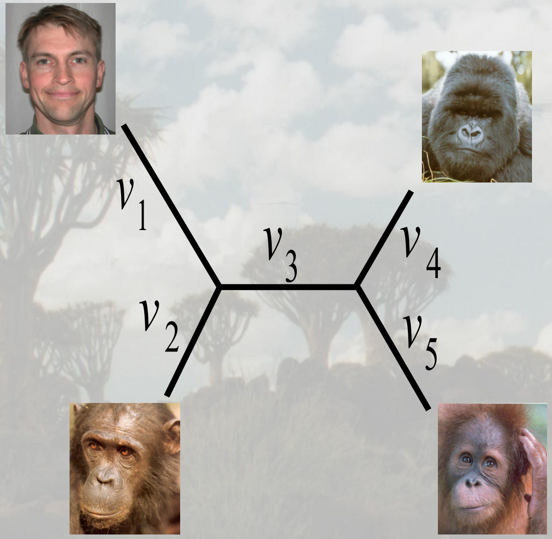
”Likelihood”

$$f(\theta|X) = \frac{f(\theta)\,f(X|\theta)}{\int f(\theta)\,f(X|\theta)\,d\theta}$$

Normalizing constant

Model: topology AND branch lengths

θ Parameters

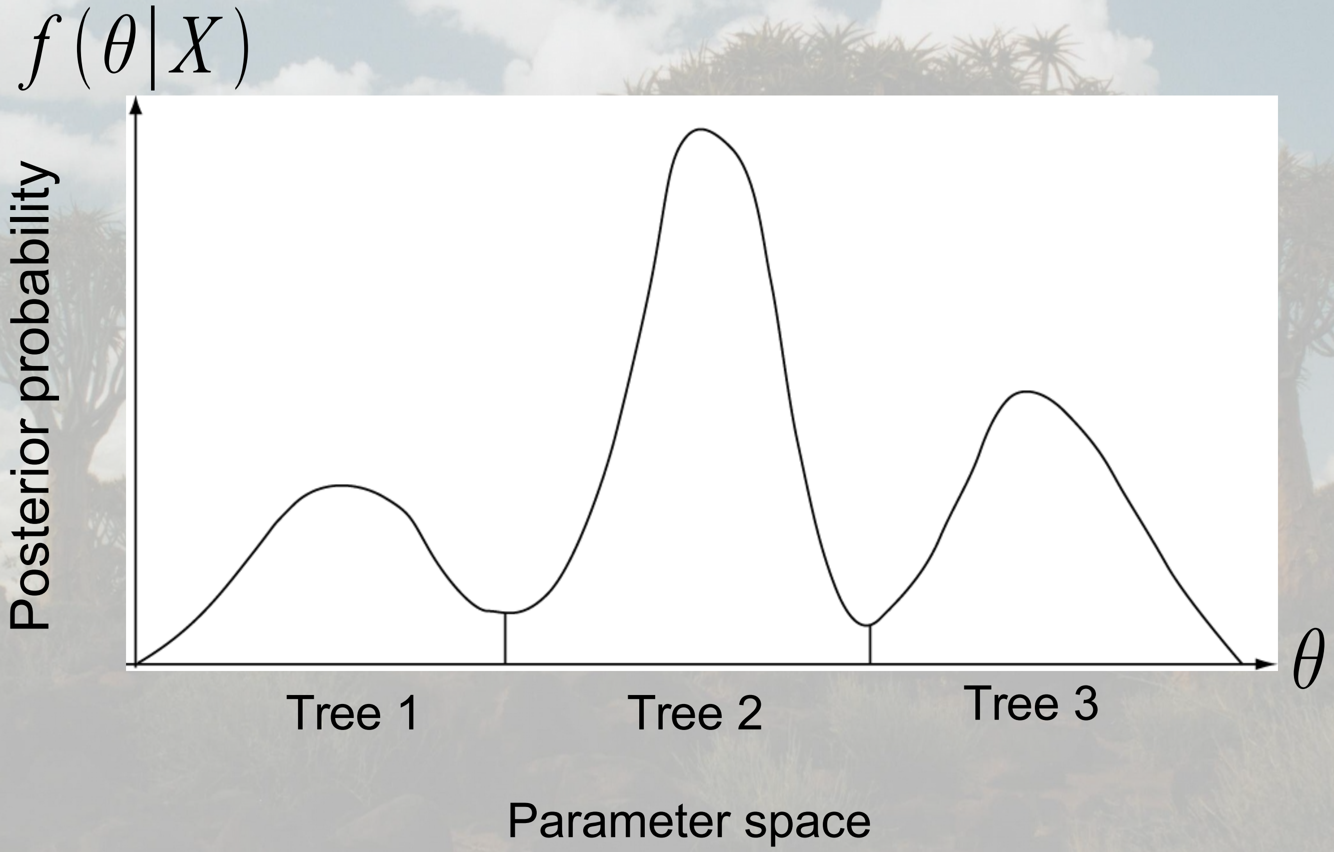


topology (Tau) $\left(\tau\right)$

Branchlengths (expected amount of change) $\left(v_i\right)$

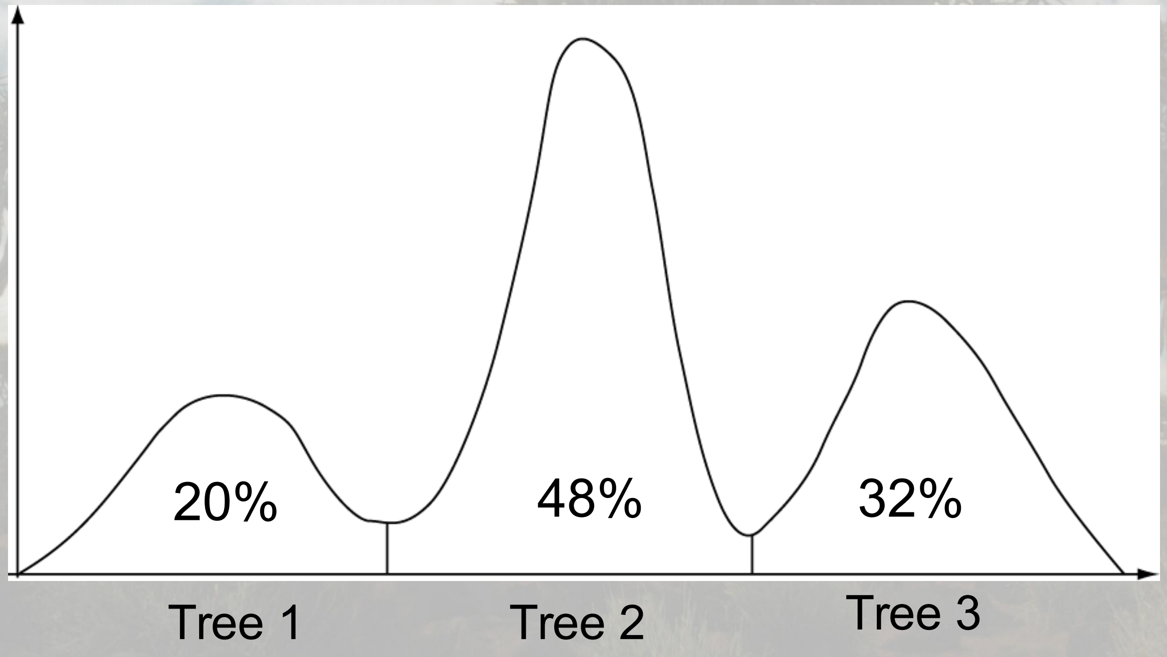
$$\theta=\left(\tau, v\right)$$

Posterior probability distribution



© Ronquist

- We can focus on any parameter of interest (there are no nuisance parameters) by marginalizing the posterior over the other parameters (integrating out the uncertainty in the other parameters)



Percentages denote marginal probability distribution on trees.

© Ronquist

Why is it called marginalizing?

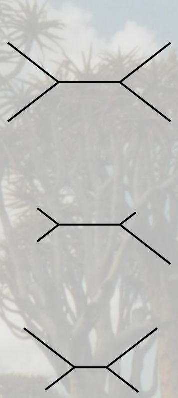
Trees

Branch length vectors

	τ_1	τ_2	τ_3	
v^1	0.10	0.07	0.12	0.29
v^2	0.05	0.22	0.06	0.33
v^3	0.05	0.19	0.14	0.38
	0.20	0.48	0.32	

joint probabilities

marginal probabilities



Markov chain Monte Carlo MCMC

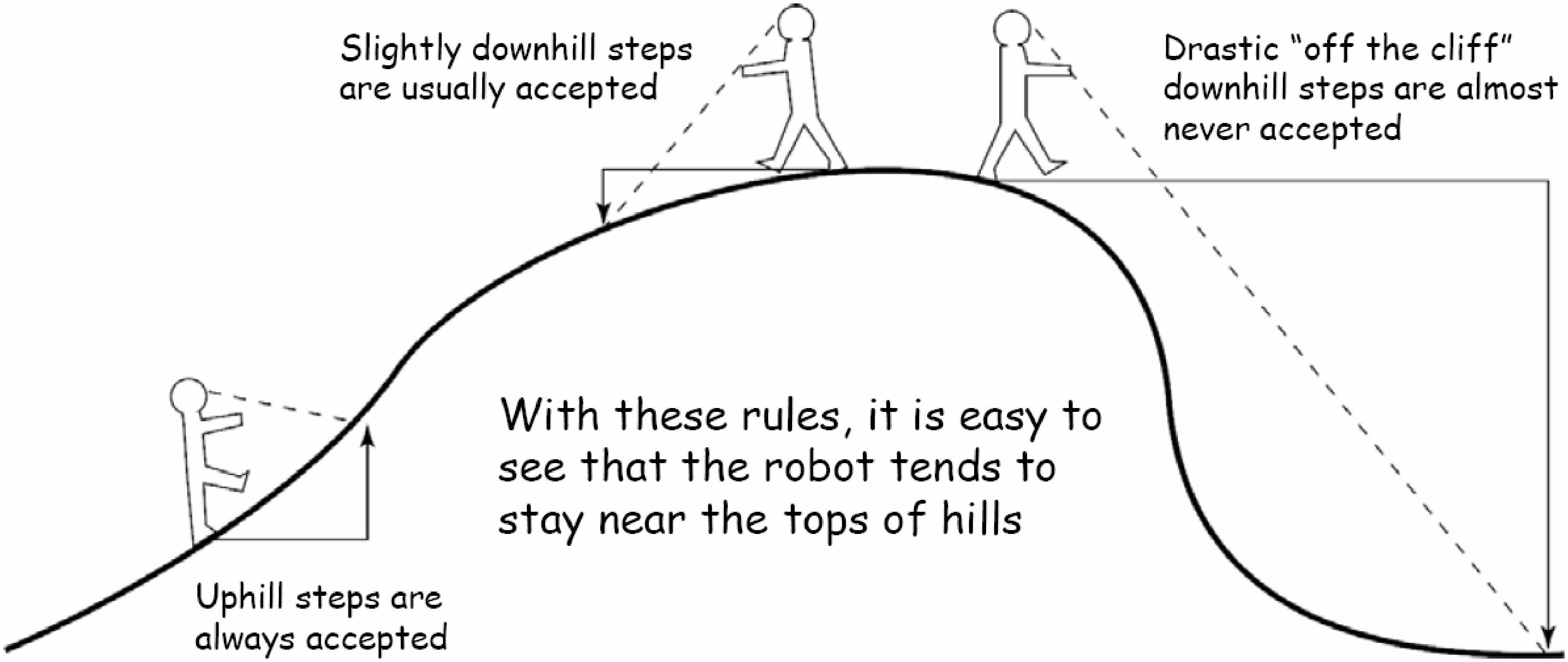
Markov chain Monte Carlo methods are a class of algorithms for **sampling from a probability distribution.** The state of the chain after a number of steps is used as a **sample of the desired distribution**. The quality of the sample improves as a function of the number of steps.

It searches the tree with the ML given the sequences.

The likelihoods can be converted to real probabilities using the Bayes's theorem (sum to 1).

It doesn't looks for one best tree but sum up (consensus) all good trees.

Az MCMC robot



Slightly downhill steps are usually accepted

Drastic "off the cliff" downhill steps are almost never accepted

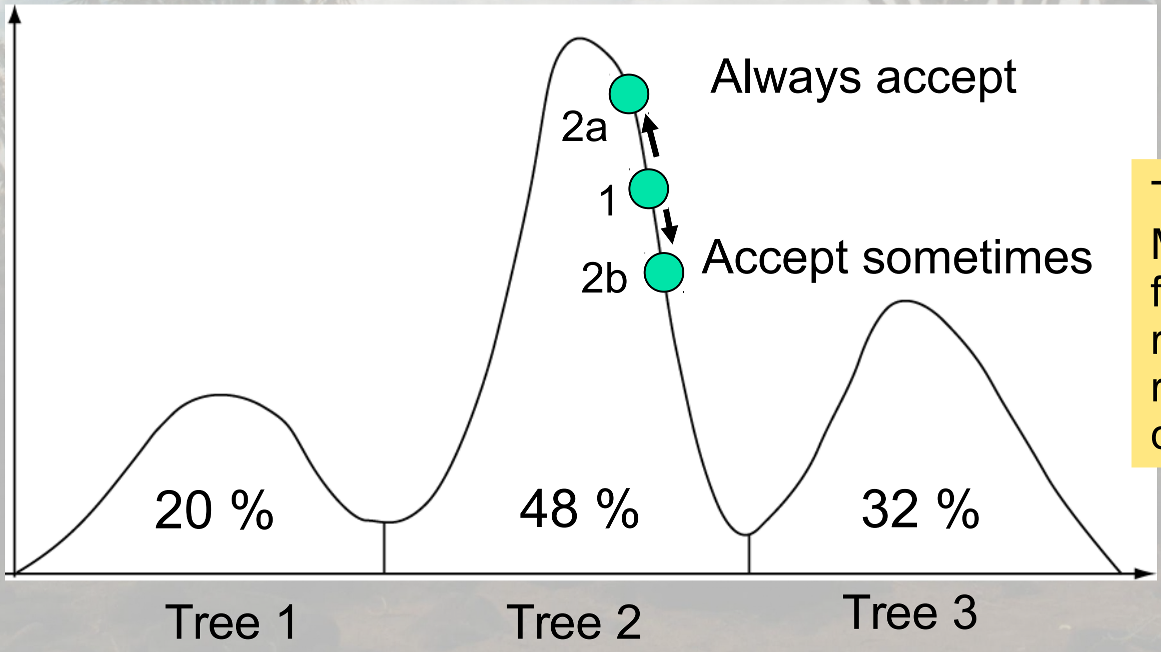
With these rules, it is easy to see that the robot tends to stay near the tops of hills

Uphill steps are always accepted

Paul Lewis©

MCMC

1. Start at an arbitrary point
2. Make a small random move
3. Calculate height ratio (r) of new state to old state:
   - r > 1 -> new state accepted
   - r < 1 -> new state accepted with probability r . If new state not accepted, stay in the old state
4. Go to step 2



The proportion of time the MCMC procedure samples from a particular parameter region is an estimate of that region's posterior probability density.

Metropolis-coupled Markov chain Monte Carlo = MCMCMC or MC^3

It runs not 1 but 4 robots (chains) parallel.

The main robot is the „cold chain”, the other 3 are the heated chains.

The cold chain is the main chain and behaves as before.

The heated chains explore a landscape that is flatter than the landscape explored by the cold chain.

They make bigger steps and accept wrong values easier than cold chain.

After some generations the cold chain change place with one of the heated chains.

Therefore, it is easier for a heated chain to cross deep valleys in the landscape and not get stuck in local optima.

MCRobot Program

Paul Lewis©:

http://hydrodictyon.eeb.uconn.edu/people/plewis/downloads/mcrobot21z.exe

http://web.uconn.edu/gogarten/bioinf/mcrobot.html

Hány generáció szükséges?

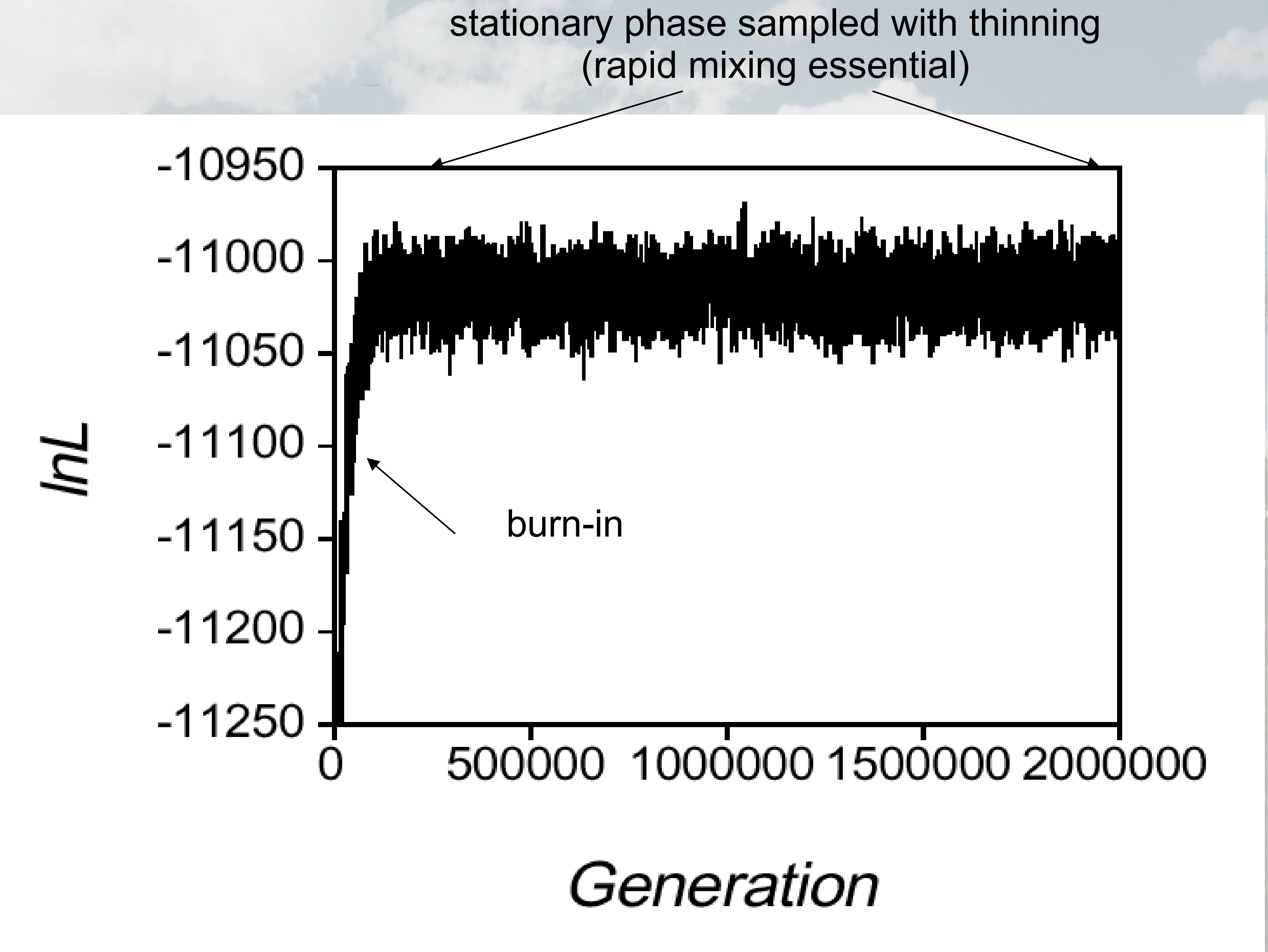
Ha elég sok generáción át léptetgetjük a robotokat elmondható, hogy egy idő után elég jól felderítik a fa és paraméter érték-teret és már csak a legnagyobb csúcsok közelében időznek.

Erről úgy győződhetünk meg, ha megnézzük magukat a likelihood értékeket.

Ha likelihoodokban elég rég óta nem találunk növekvő tendenciát, csak kiegyenlített fluktuációt, akkor bízhatunk benne, hogy a jelenleg megtalált csúcsoknál magasabbak már nincsenek a fa-térben és elkészíthetjük a meglévő legvalószínűbb fákból a konszenzusunkat.

Ezt átlagos esetben 5millió generáció után érhetjük el.

Number of generations (steps)



stationary phase sampled with thinning
(rapid mixing essential)

burn-in

The most important MC^3 parameters

Number of generations (millions)

sampling density (ie. from every 100th or 1000th generations)

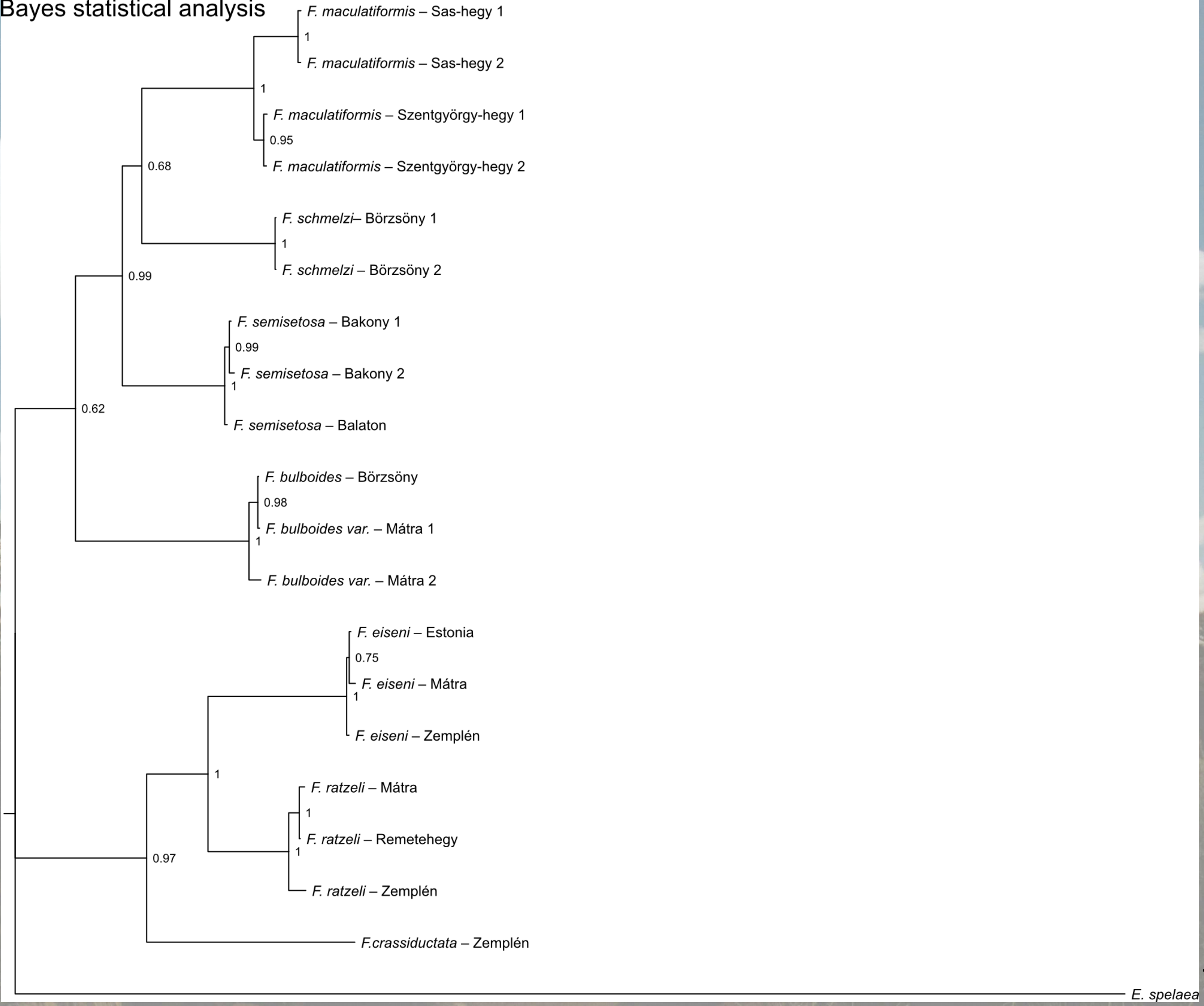
burnin (the first 25%of the samples should be considered in the consensus tree)

step length (no need to change)

number of robots (1 cold and 3 heated chain)

number of independent runs (2-4)

Bayes statistical analysis



Sources

Anders Gorm Pedersen

Fredrik Ronquist

Wikipedia

Thanks for the authors!

Thanks for the attention!