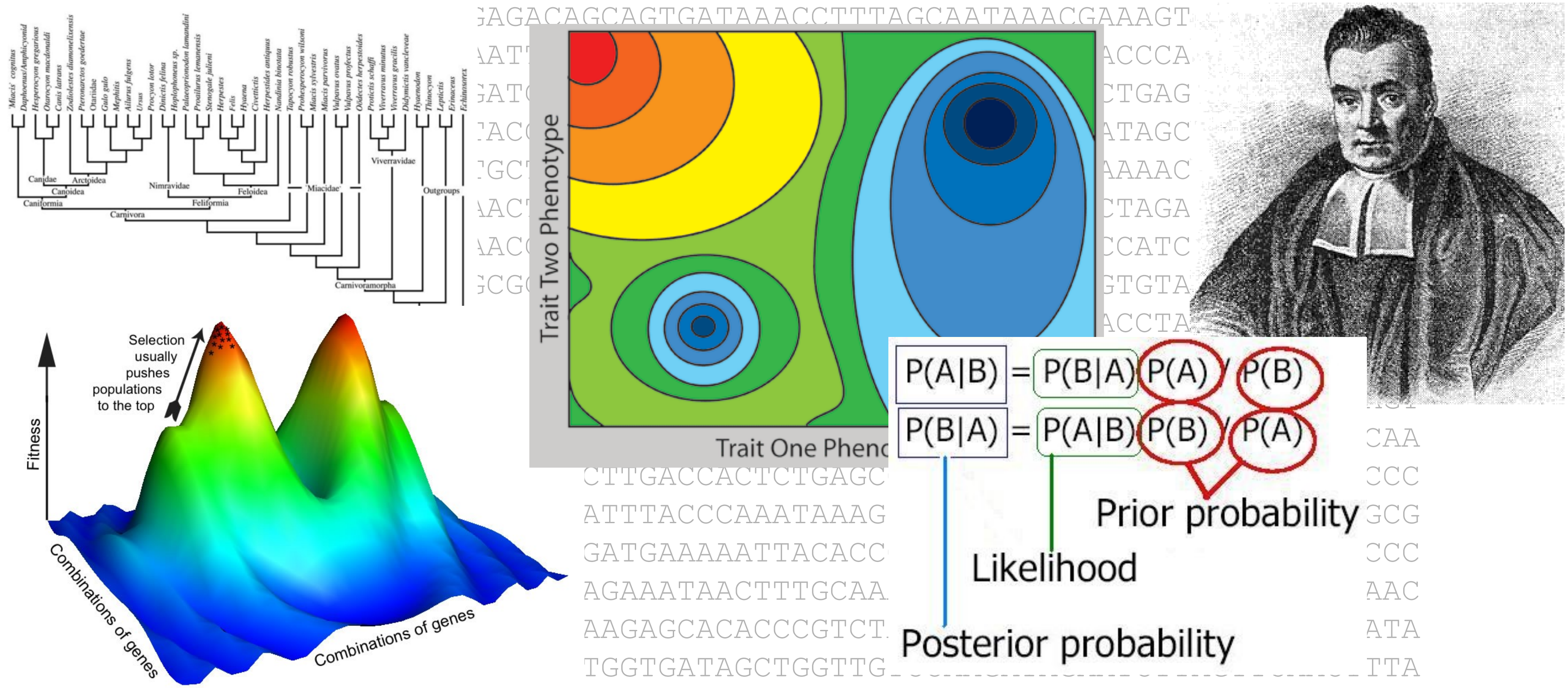
MODULARIZACE VÝUKY EVOLUČNÍ A EKOLOGICKÉ BIOLOGIE
CZ.1.07/2.2.00/15.0204

PHYLOGENETIC ANALYSIS II.



P(A|B) = P(B|A) P(A) / P(B)

P(B|A) = P(A|B) P(B) / P(A)

Prior probability

Likelihood

Posterior probability

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

MAXIMUM LIKELIHOOD, ML (maximální věrohodnost)

15 coin tosses:

→ score TTHHHTHTTTHTHHT
tj. 7× head (H), 8× tail (T)

Likelihood = conditional probability of data (final score) given the hypothesis:

L = Pr(D | H) = Pr(7× head, 8× tail | hypothesis)

Probability of head = p, tail = $(1 – p)$

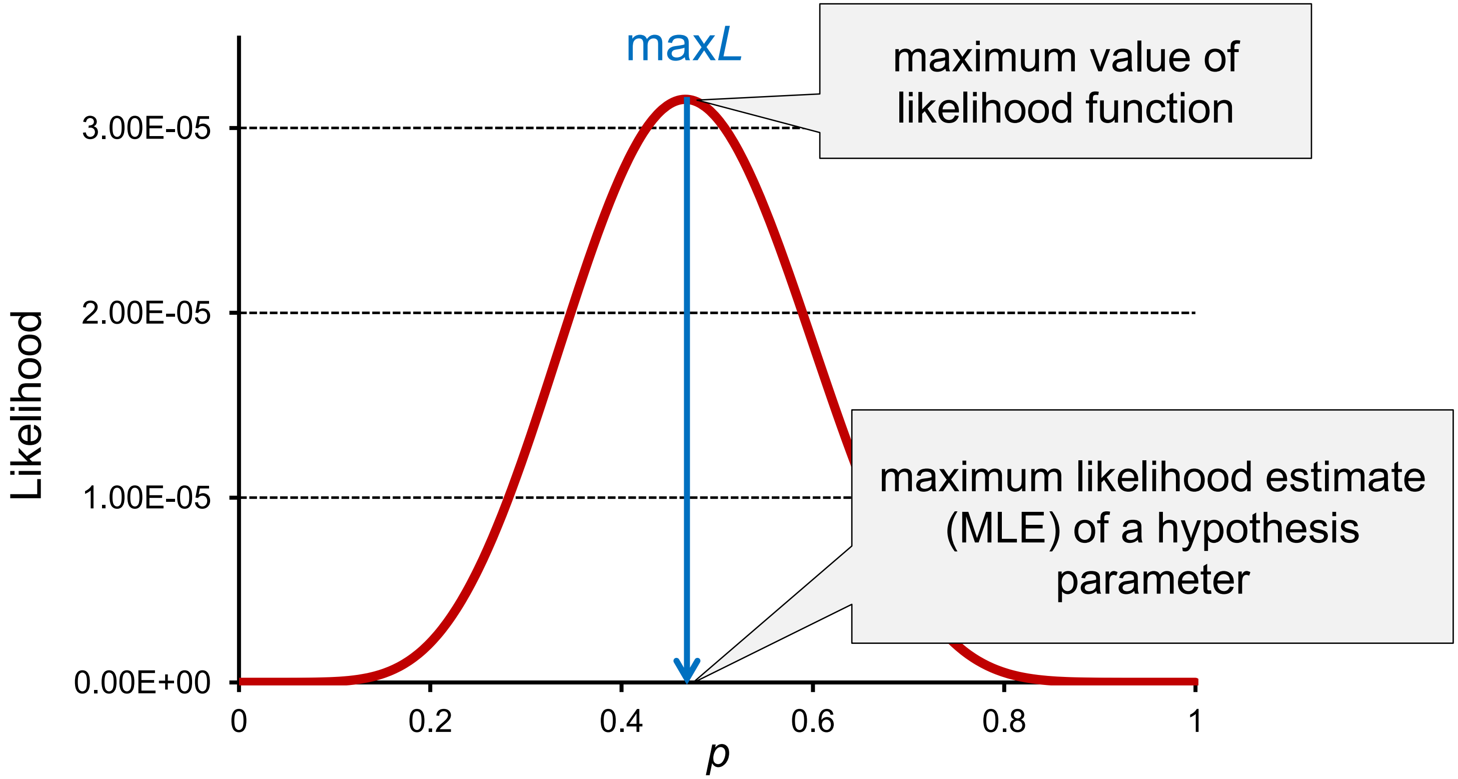
score TTHHHTHTTTHTHHT [7× head (H), 8× tail (T)]

Because tosses independent ⇒ probability of final score =

$(1 - p)\times(1 - p)\times p\times p\times p\times(1 - p)\times p\times(1 - p)\times(1 - p)\times(1 - p)\times p\times(1 - p)\times p\times p\times(1 - p) =$

$= p^7(1\text{-}p)^8$

maximum = 0,4666 ≈ 7/15



maximum value of likelihood function

maximum likelihood estimate (MLE) of a hypothesis parameter

Hypothesis?

Eg. H = coin is not „biased“, ie. $p = 1/2 \Rightarrow L = 3{,}0517.10^{-5}$

If the coin is biased so that we get tail in 2/3 cases:

$p = 1/3 \Rightarrow L = 1{,}7841.10^{-5}$

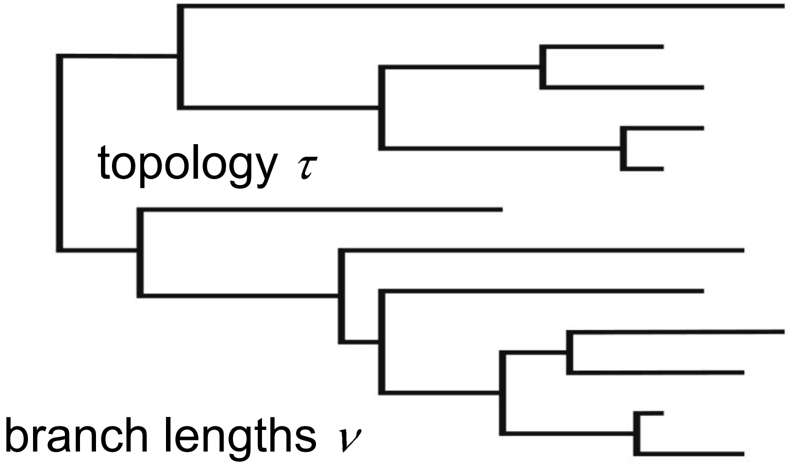
⇒ result of tosses 1,7× more probable with unbiased coin

Maximum likelihood in phylogenetic analysis

data:

```
1          TCAAAAATGGCTTTATTCGCTTAATGCCGTTAACCCTTGCGGGGGCCATG
2          TCCGTGATGGATTTATTTCCGCAATGCCTGTCATCTTATTCTCAAGTATC
3          TTCGTGATGGATTTATTGCAGGTATGCCAGTCATCCTTTTCTCATCTATC
4          TTCGTGACGGGTTTATCTCGGCAATGCCGGTCATCCTATTTTCGAGTATT
```

tree:

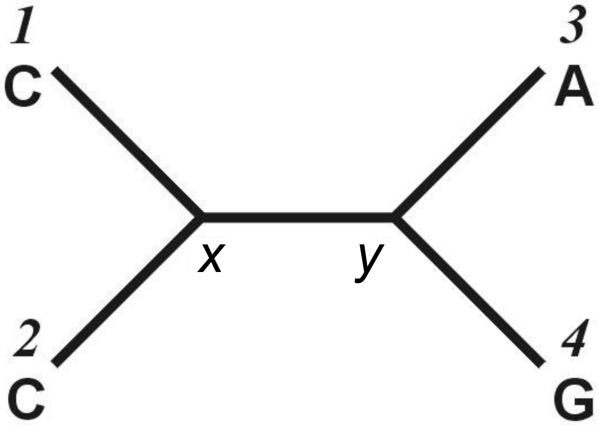


+ evolutionary model θ

= hypothesis

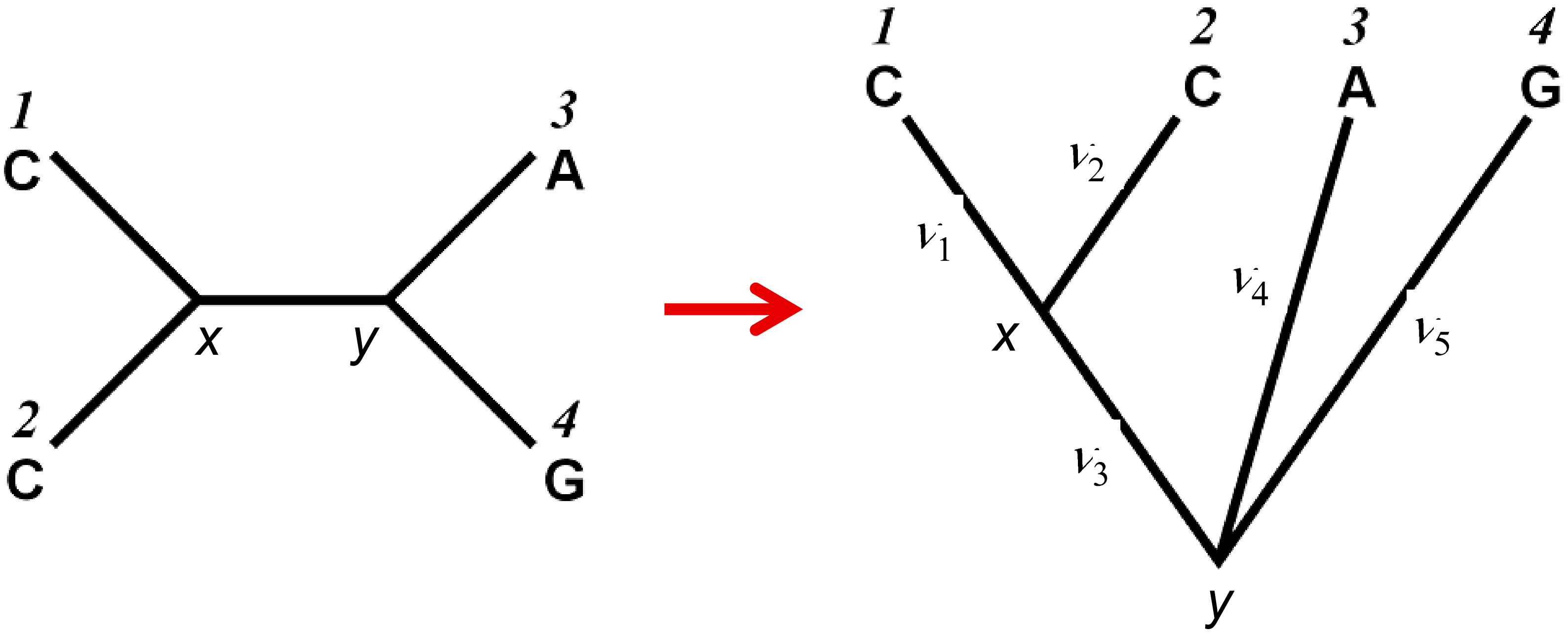
L = P(D | H): D = sequence matrix (data), H = τ (topology) + ν (branch lenghts) + θ (model)

```
     1                  j                             N
1    TCAAAAATGGCTTTATTCGCTTAATGCCGTTAACCCTTGCGGGGGCCATG
2    TCCGTGATGGATTTATTTCCGCAATGCCTGTCATCTTATTCTCAAGTATC
3    TTCGTGATGGATTTATTGCAGGTATGCCAGTCATCCTTTTCTCATCTATC
4    TTCGTGACGGGTTTATCTCGGCAATGCCGGTCATCCTATTTTCGAGTATT
```

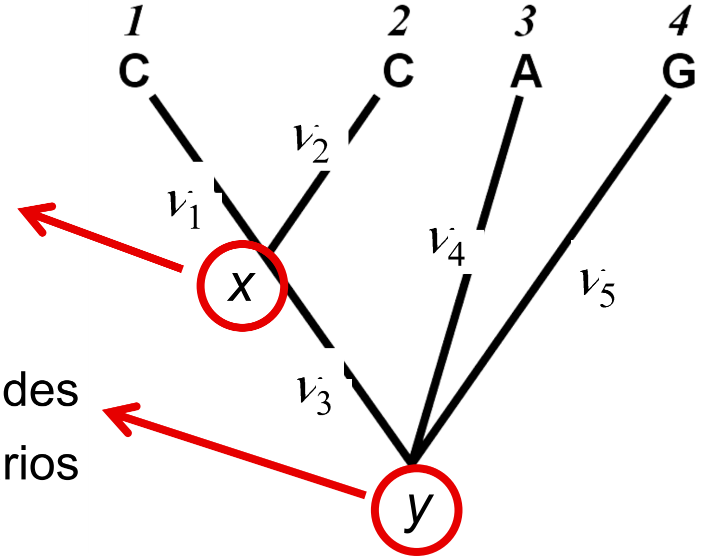


```
                  1                  j                              N
1                 TCAAAAATGGCTTTATTCGCTTAATGCCGTTAACCCTTGCGGGGGCCATG
2                 TCCGTGATGGATTTATTTCCGCAATGCCTGTCATCTTATTCTCAAGTATC
3                 TTCGTGATGGATTTATTGCAGGTATGCCAGTCATCCTTTTCTCATCTATC
4                 TTCGTGACGGGTTTATCTCGGCAATGCCGGTCATCCTATTTTCGAGTATT
```

ν_i = branch lengths



```
      1                  j                              N
1     TCAAAAATGGCTTTATTCGCTTAATGCCGTTAACCCTTGCGGGGCCATG
2     TCCGTGATGGATTTATTTCCGCAATGCCTGTCATCTTATTCTCAAGTATC
3     TTCGTGATGGATTTATTGCAGGTATGCCAGTCATCCTTTTCTCATCTATC
4     TTCGTGACGGGTTTATCTCGGCAATGCCGGTCATCCTATTTTCGAGTATT
```



v_i = branch lengths

x: 4 nucleotides

y: 4 nucleotides

⇒ 4 × 4 = 16 possible scenarios

$$L(1) = P(y)\times P(y\rightarrow x)\,v_3\times P(x\rightarrow \mathrm{C})\,v_1\times P(x\rightarrow \mathrm{C})\,v_2\times P(y\rightarrow \mathrm{A})\,v_4\times P(y\rightarrow \mathrm{G})\,v_5$$

$$L(j) = P(\text{scenario } 1) + \ldots. + P(\text{scenario } 16)$$

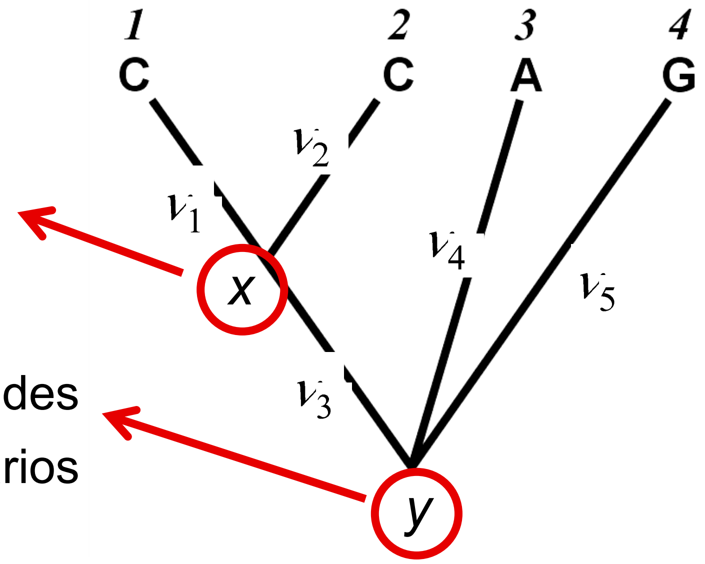
```
    1                  j                              N
1   TCAAAAATGGCTTTATTCGCTTAATGCCGTTAACCCTTGCGGGGGCCATG
2   TCCGTGATGGATTTATTTCCGCAATGCCTGTCATCTTATTCTCAAGTATC
3   TTCGTGATGGATTTATTGCAGGTATGCCAGTCATCCTTTTCTCATCTATC
4   TTCGTGACGGGTTTATCTCGGCAATGCCGGTCATCCTATTTTCGAGTATT
```

v_i = branch lengths

x: 4 nucleotides

y: 4 nucleotides

$\Rightarrow$ 4 × 4 = 16 possible scenarios

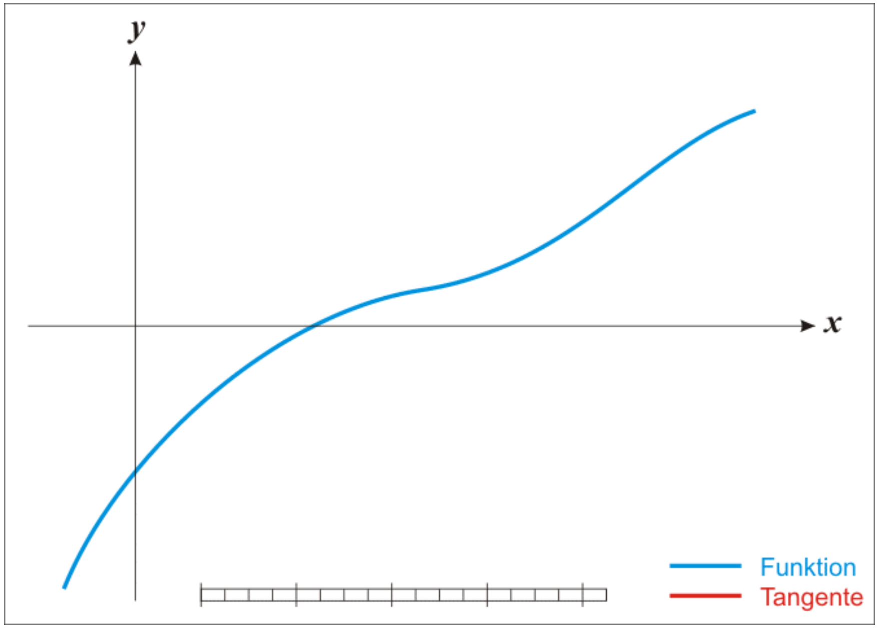


all sites: $L = L(1) \times L(2) \times \ldots \times L(j) \times \ldots \times L(N) = \prod_{j=1}^{N} L_j$

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

Search for maximum likelihood of the tree

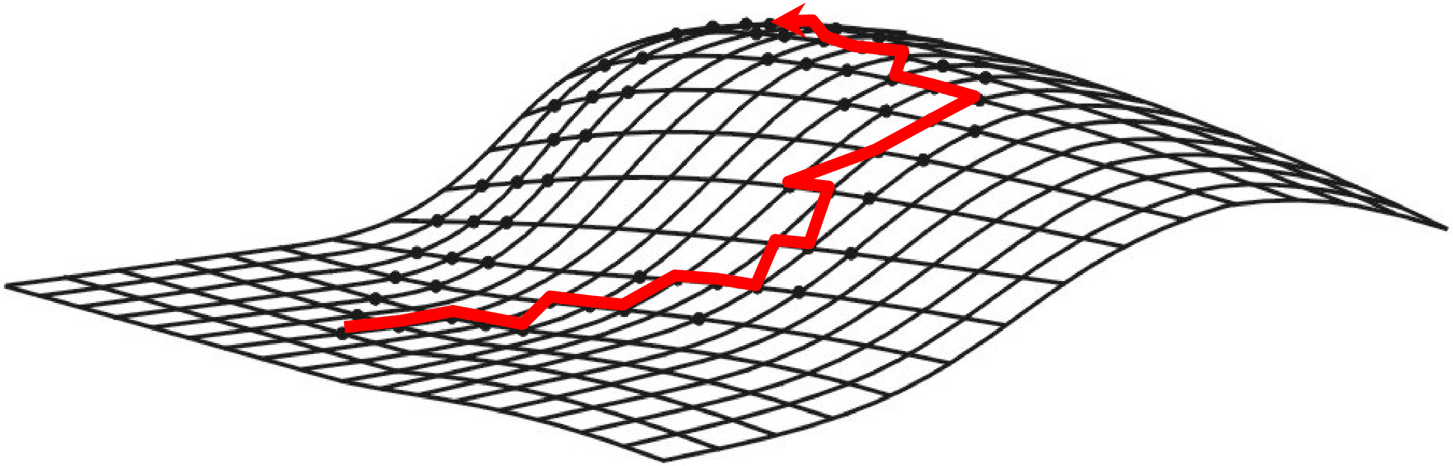
→ eg. Newton (Newton-Raphson) method



https://upload.wikimedia.org/wikipedia/commons/e/e0/NewtonIteration_Ani.gif

Maximum likelihood tree search: heuristic search

Heuristic search

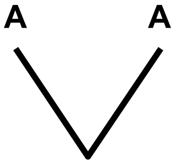


stepwise addition ... eg. PHYLIP

star decomposition ... eg. MOLPHY; neighbor-joining tree

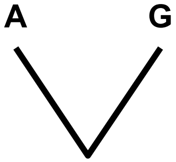
branch swapping

Likelihood (ML) and parsimony (MP)

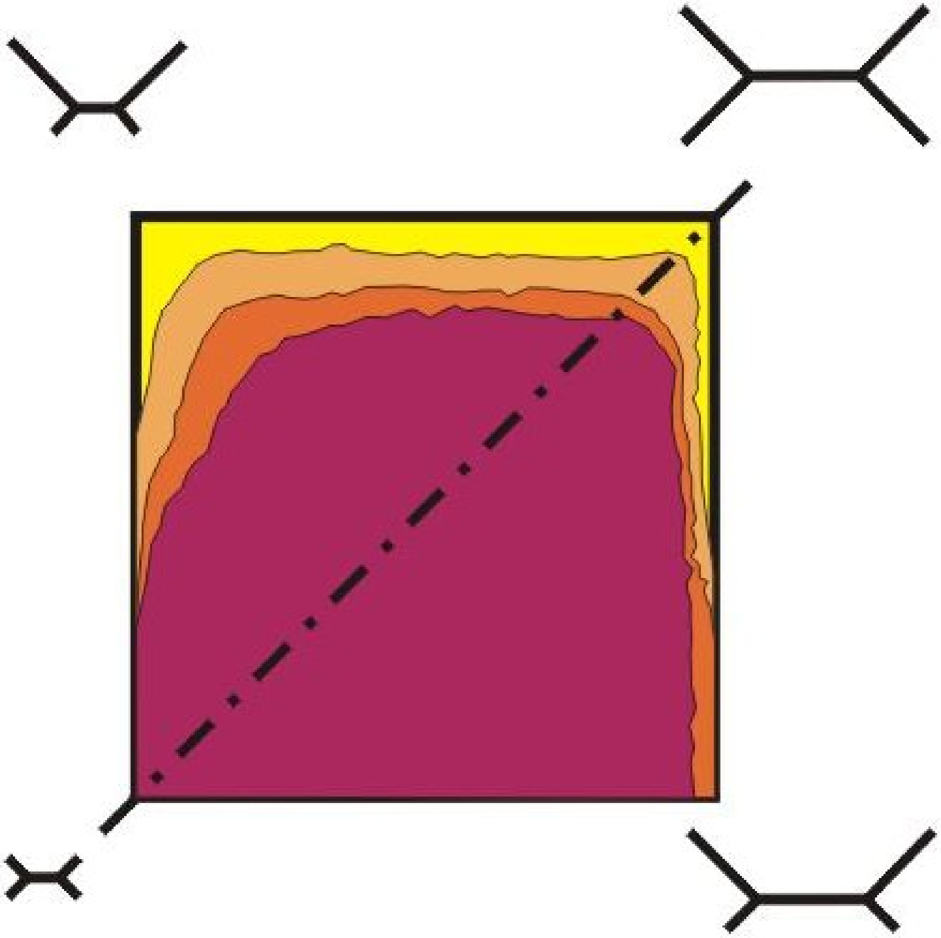


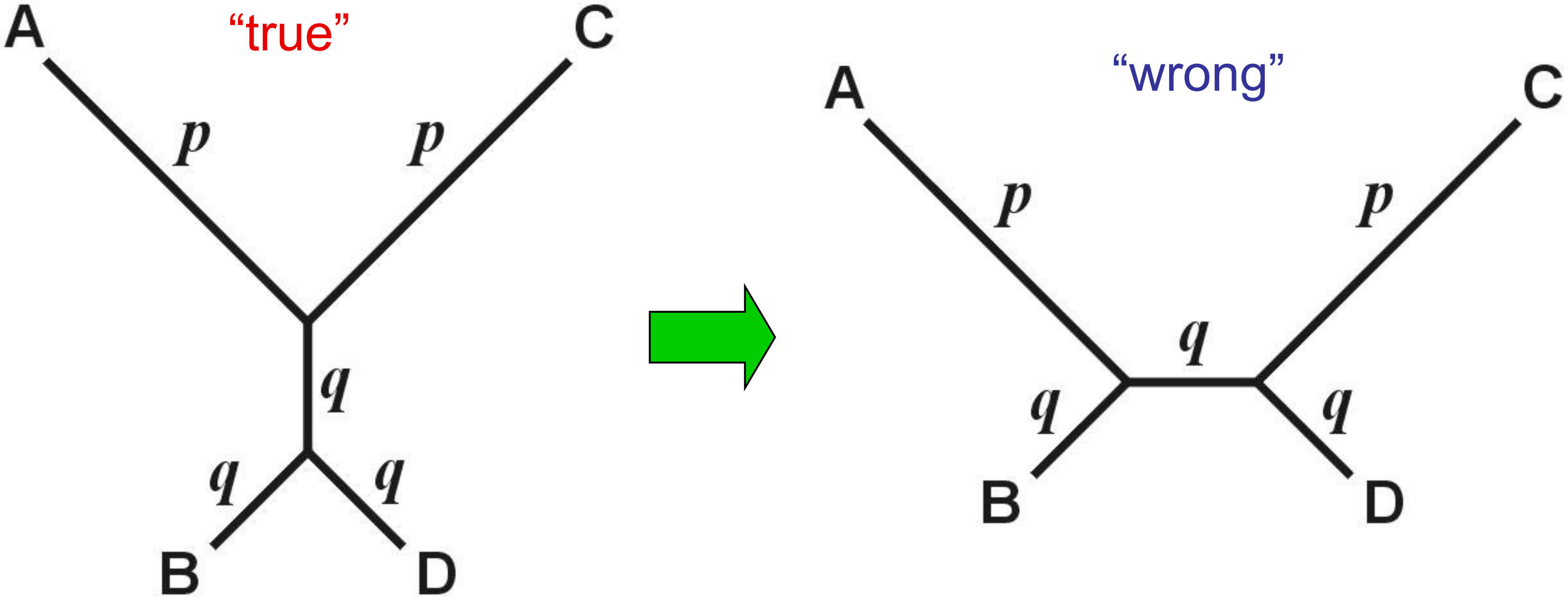
No. changes	Parsimony	$\nu = 0,01$	$\nu = 0,10$	$\nu = 0,20$	$\nu = 1,00$
		(0,2475)	(0,2266)	(0,20611)	(0,11192)
0	100	99,99	99,83	99,31	82,17
1	0	0,00	0,00	0,00	0,00
2	0	0,0011	0,11	0,44	9,13
3	0			0,034	3,55
4	0				0,0027

No. changes	Parsimony	$\nu = 0,01$	$\nu = 0,10$	$\nu = 0,20$	$\nu = 1,00$
		(0,00083)	(0,00786)	(0,01462)	(0,04602)
0	0	0,00	0,00	0,00	0,00
1	100	99,66	96,64	92,36	66,54
2	0	0,33	3,22	6,22	21,19
3	0		0,12	0,48	8,61
4	0		0,003	0,023	2,05
5	0			0,0037	0,42



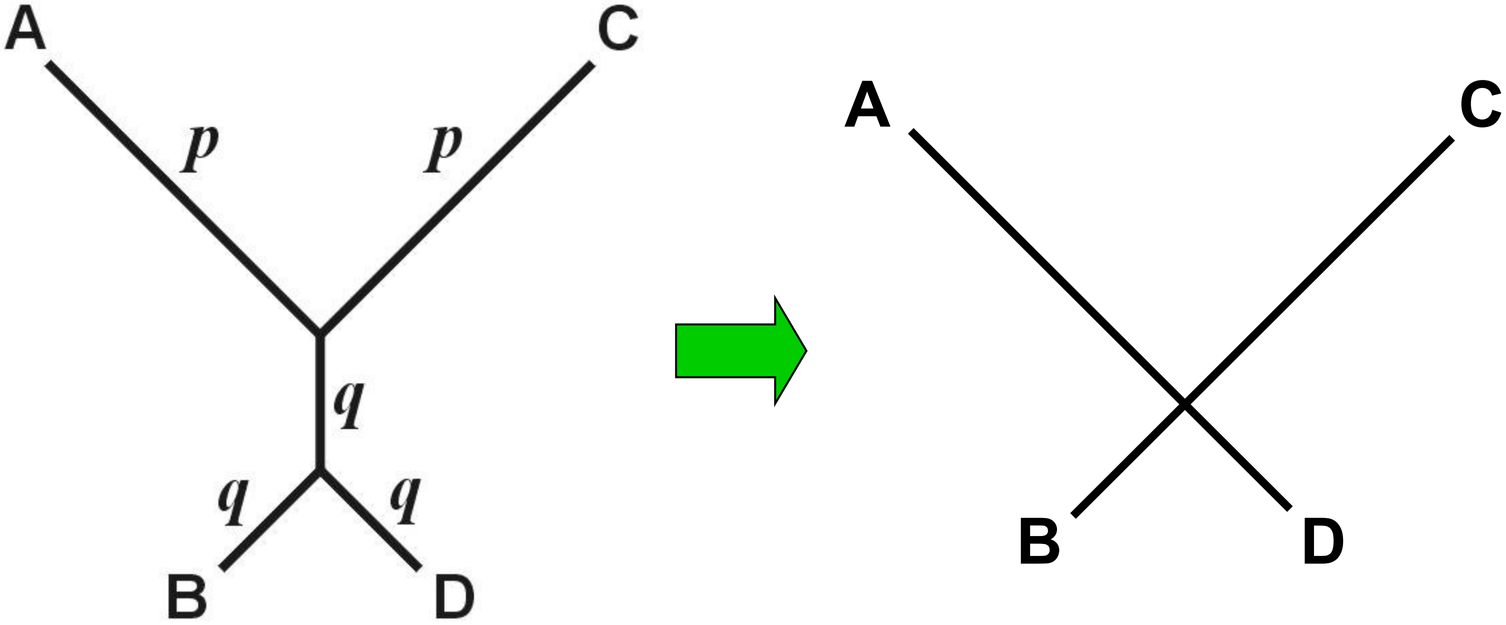
Likelihood and consistency





Farris
(anti-Felsenstein,
inverse Felsenstein)
zone

“long-branch repulsion”


A
C
p
p
q
q
q
B
D
A
C
B
D

BAYESIAN ANALYSIS
(Bayesovská analýza)

ML: Probability of data given hypothesis

Bayesian approach:

Conditional probability of hypothesis given data

$P(\mathrm{H} \mid \mathrm{D})$

Example.: set of 100 dice, from which we choose one

we know that of 100 dice, 80 are ‘fair’ and 20 biased for 6

2 throws: 1. throw = ⚁ 2. throw = ⚅

What is the probability our dice is biased?

probability of individual results:

all the same in unbiased dice, varied in biased dice:

	unbiased	biased
⚀	1/6	1/21
⚁	1/6	3/21
⚂	1/6	3/21
⚃	1/6	4/21
⚄	1/6	4/21
⚅	1/6	6/21

$P(H|D)$ is called posterior probability (aposteriorní pravděpodobnost)

posterior probability is a function of likelihood $L = P(D|H)$

and prior probability (apriorní pravděpodobnost) reflecting our a priori expectation or knowledge

Posterior probability that the coin is biased is given by the Bayes equation:

likelihood

prior probability

$$P(H|D) = \frac{P(D|H) \times P(H)}{\Sigma[P(D|H_i) \times P(H_i)]}$$

sum of numerators across all alternative hypotheses

Thomas Bayes

For our example of 2 dice throws:

	unbiased	biased
⚀	1/6	1/21
⚁	1/6	3/21
⚂	1/6	3/21
⚃	1/6	4/21
⚄	1/6	4/21
⚅	1/6	6/21

prior probability (biased) = 0,2
(20/100 biased dice in the set)

Pr. of getting ⚂ ⚅ with unbiased dice:

$P = 1/6 \times 1/6 = 1/36$

Pr. of getting ⚂ ⚅ with biased dice:

$P = 3/21 \times 6/21 = 18/441$

$$\mathrm{P}(\text{biased}|\text{ ⚂ ⚅}) = \frac{\mathrm{P}(\text{⚂ ⚅ }|\text{biased}) \times \mathrm{P}(\text{biased})}{\mathrm{P}(\text{⚂ ⚅ }|\text{biased}) \times \mathrm{P}(\text{biased}) + \mathrm{P}(\text{⚂ ⚅ }|\text{fair}) \times \mathrm{P}(\text{fair})}$$

$$= \frac{18/441 \times 2/10}{18/441 \times 2/10 + 1/36 \times 8/10} = \underline{0{,}269}$$

Bayesian method in phylogenetic analysis:

posterior probability

likelihood

prior probability

$$P(\boldsymbol{\tau}, \nu, \boldsymbol{\theta}|\mathbf{X}) = \frac{P(\mathbf{X}|\boldsymbol{\tau}, \nu, \boldsymbol{\theta})P(\boldsymbol{\tau}, \nu, \boldsymbol{\theta})}{\sum_{i=1}^{B(s)}[P(\mathbf{X}|\boldsymbol{\tau}, \nu, \boldsymbol{\theta})P(\boldsymbol{\tau}, \nu, \boldsymbol{\theta})]}$$

sum across all hypotheses
(= marginal likelihood)

Parameters of Bayesian analysis mostly continuous ⇒
P → probability density functions

either ML estimates → empirical BA

or all combinations → hierarchical BA

$$P(\mathbf{X}|\boldsymbol{\tau}, \nu, \boldsymbol{\theta}) = \int P(\mathbf{X}|\boldsymbol{\tau}, \nu, \boldsymbol{\theta})\, dF(\nu, \boldsymbol{\theta})$$

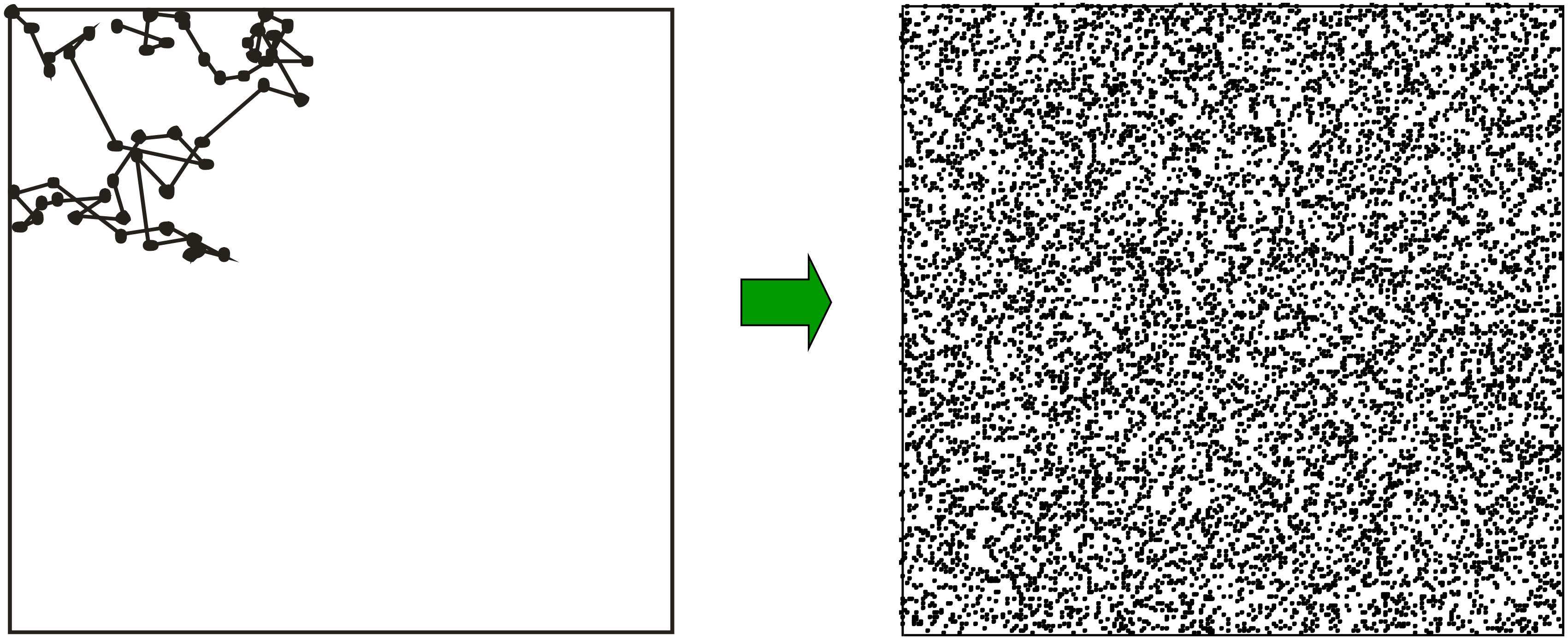
Problem: calculations too complex ⇒ impossible to solve analytically,
only numerically

solution: Monte Carlo methods
random sampling, approximation of reality when sample size high

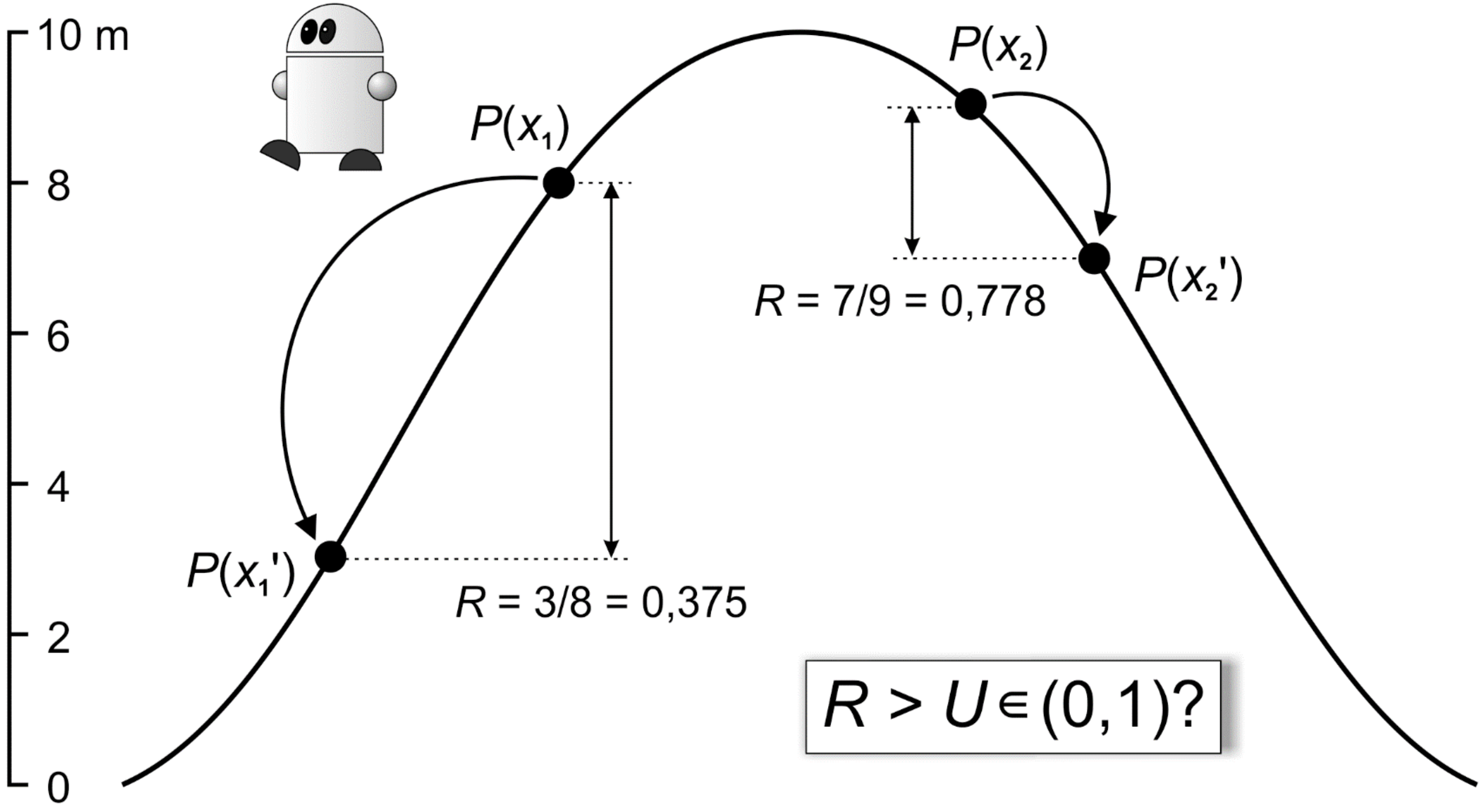
Markov chain Monte Carlo (MCMC)

Markov process: t_{-1}: A → t_0: C → t_{+1}: G

… *P* same across the whole phylogeny = homogenous Markov process



Metropolis-Hastings algorithm:



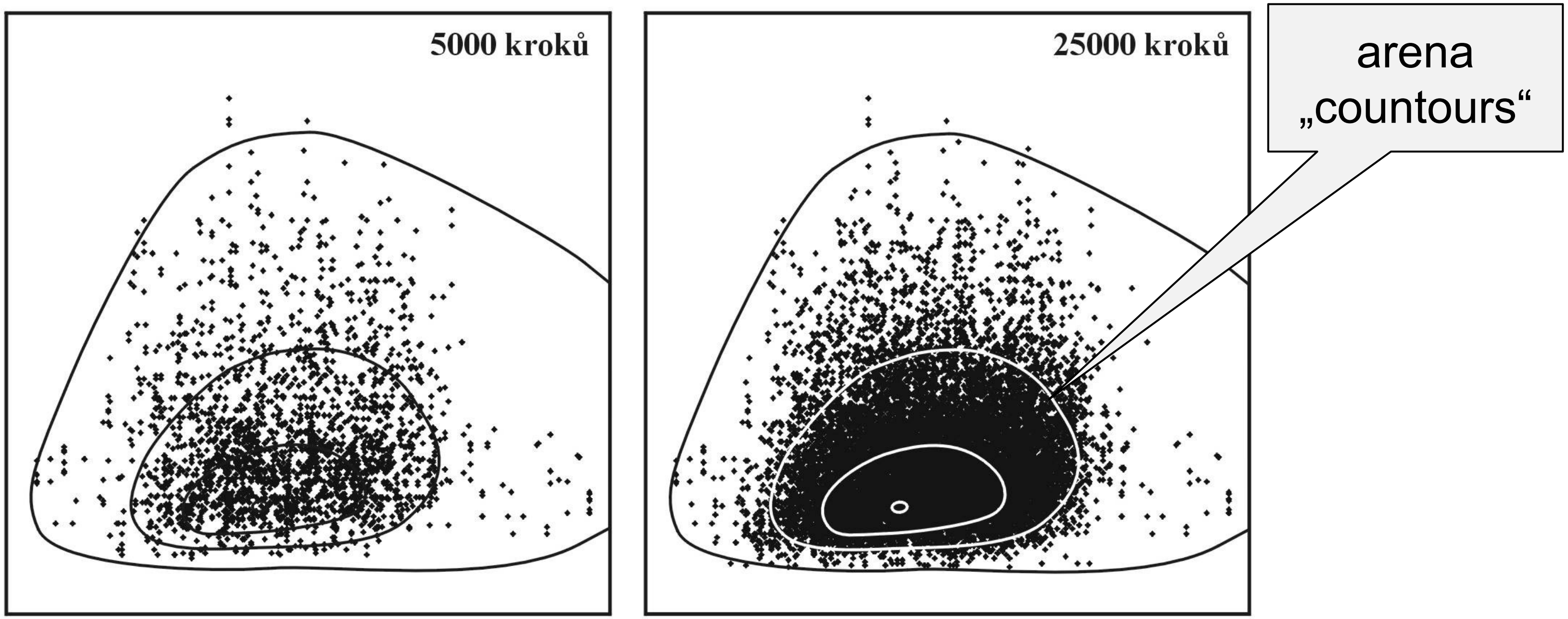
$R > U \in (0,1)?$

Metropolis-Hastings algorithm:

Change of parameter $x \rightarrow x'$

1. if $P(x') > P(x)$, accep x'
2. if $P(x') \leq P(x)$, calculate $R = P(x')/P(x)$
   since $P(x') \leq P(\mathrm{x})$, R must be ≤ 1
3. generate random number U from uniform distribution from interval (0, 1)
4. if $R \geq U$, accept x', if not, retain x

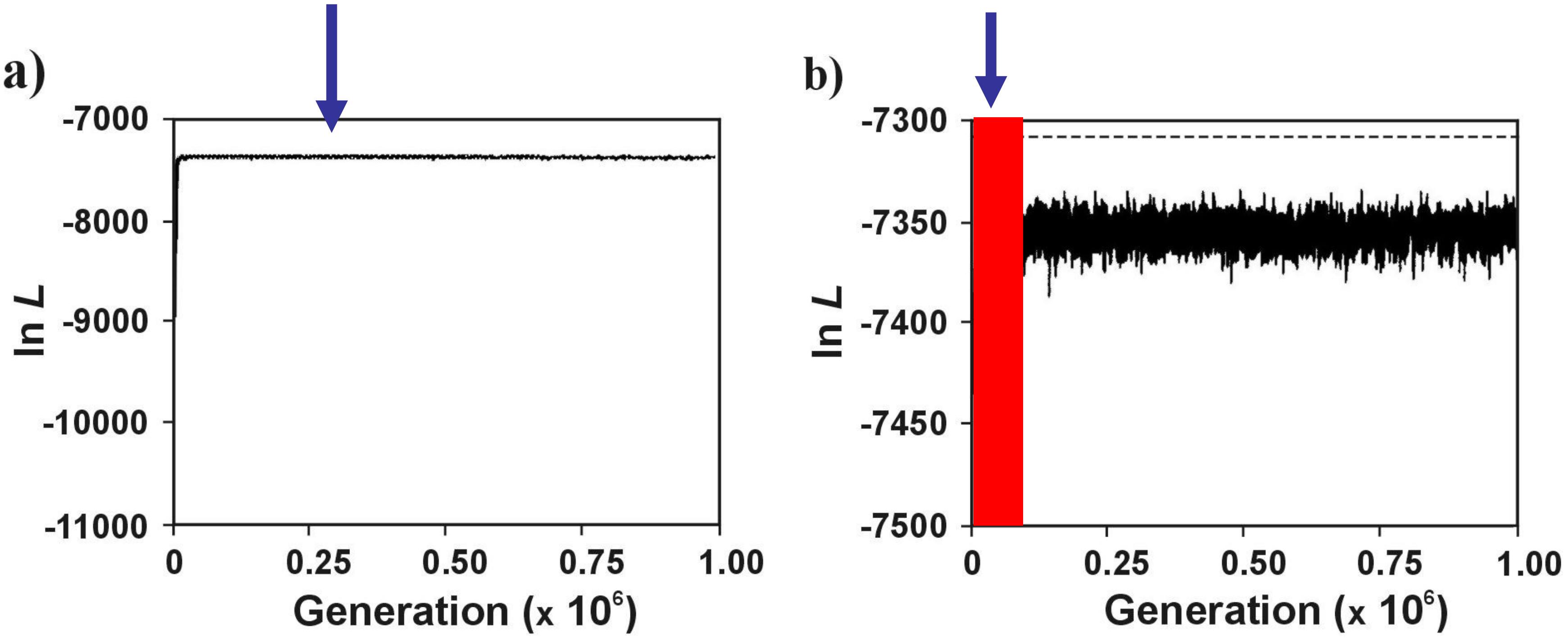
directed movement of robot across arena:



arena „countours“

stationary phase (plateau)

“burn-in”



Reversible jump MCMC:

allows changing number of parameters in each MC step

we can use eg. for modelling variation of evolution between sites in sequences, for choosing models or for making non-homogenous substitution models (eg. different base composition along branches)

Metropolis coupled MCMC (MCMCMC, MC^3):

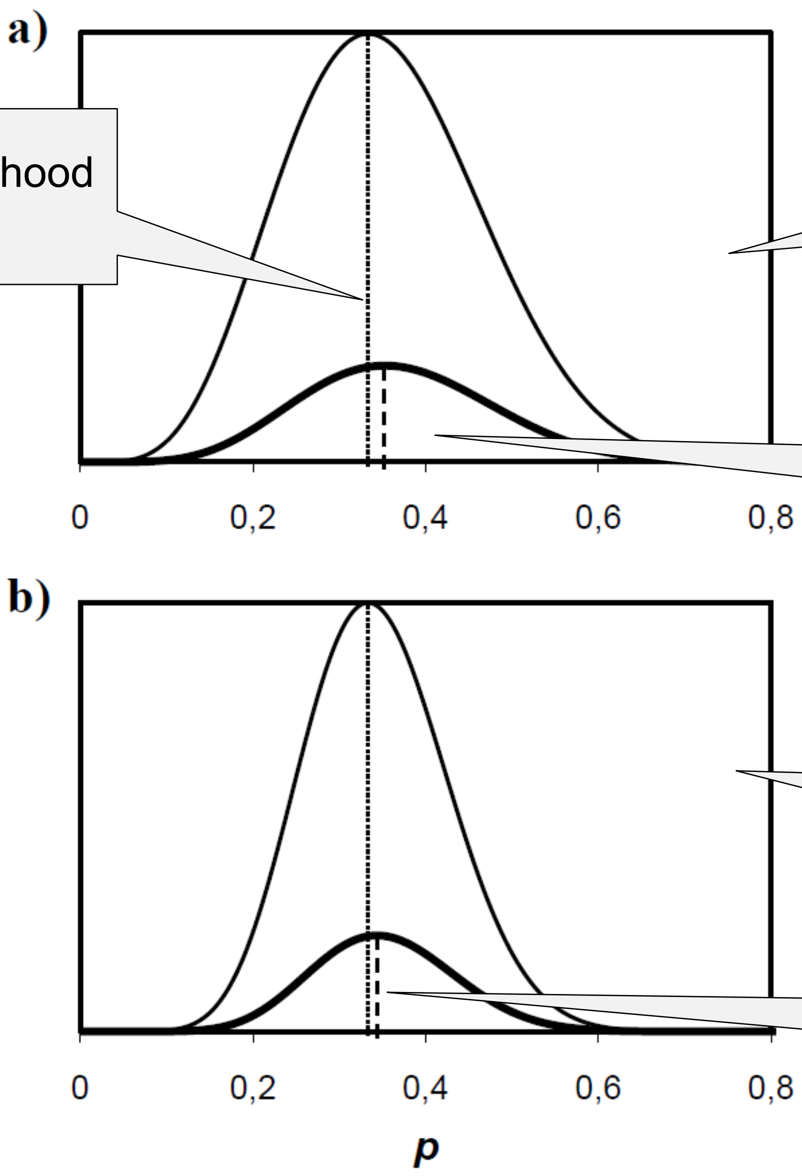
1 „cold“ chain, 3 „heated“ chains

same starting point, due to stochasticity rapid divergence of „robots“

MrBayes: *http://morphbank.ebc.uu.se/mrbayes/*

Problem with priors

prior = 0,5



maximum likelihood
= 0,333

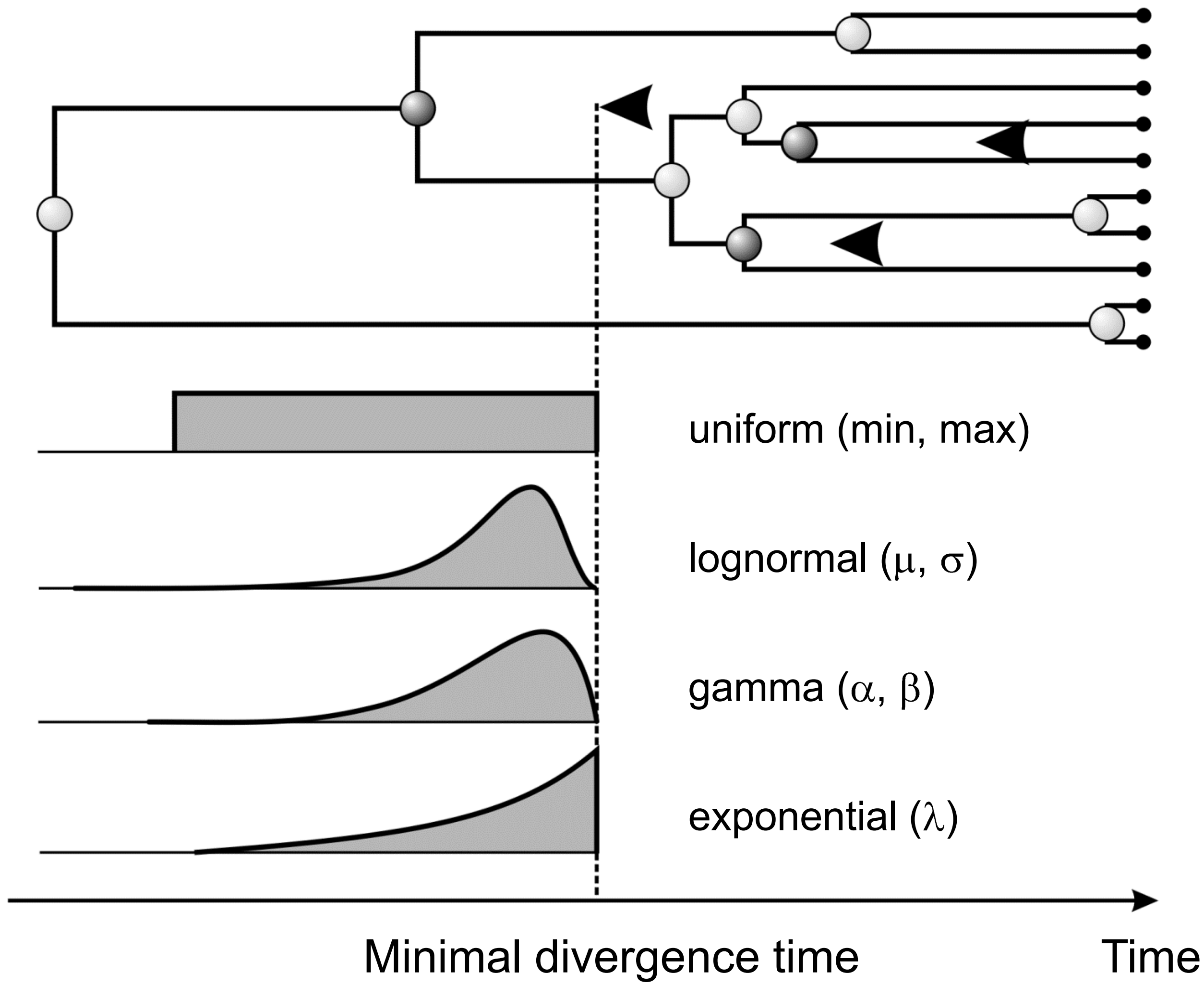
15 coin tosses
score 5 H : 10 O

due to prior,
posterior pr.
shifted to the right

30 coin tosses
score 10 H : 20 O

difference from
ML smaller

Problem with priors



uniform (min, max)

lognormal (μ, σ)

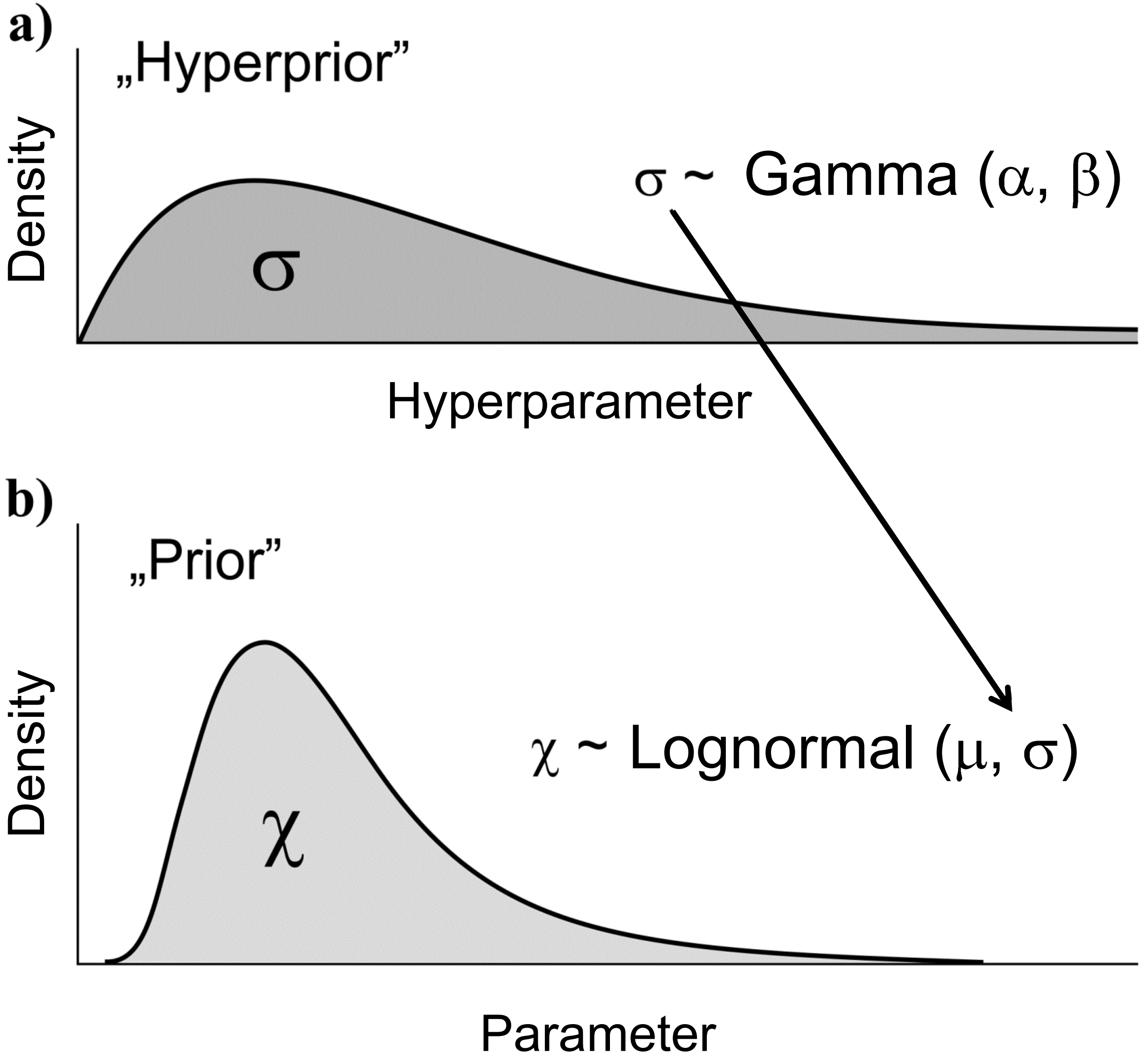
gamma (α, β)

exponential (λ)

Minimal divergence time

Time

Setting priors:



$\sigma \sim$ Gamma (α, β)

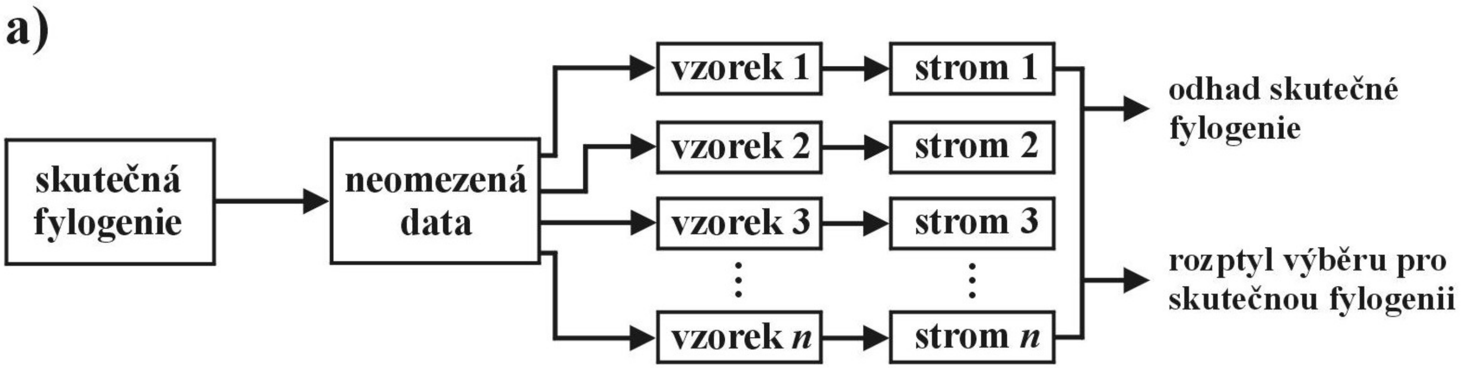
$\chi \sim$ Lognormal (μ, σ)

Measuring tree reliability

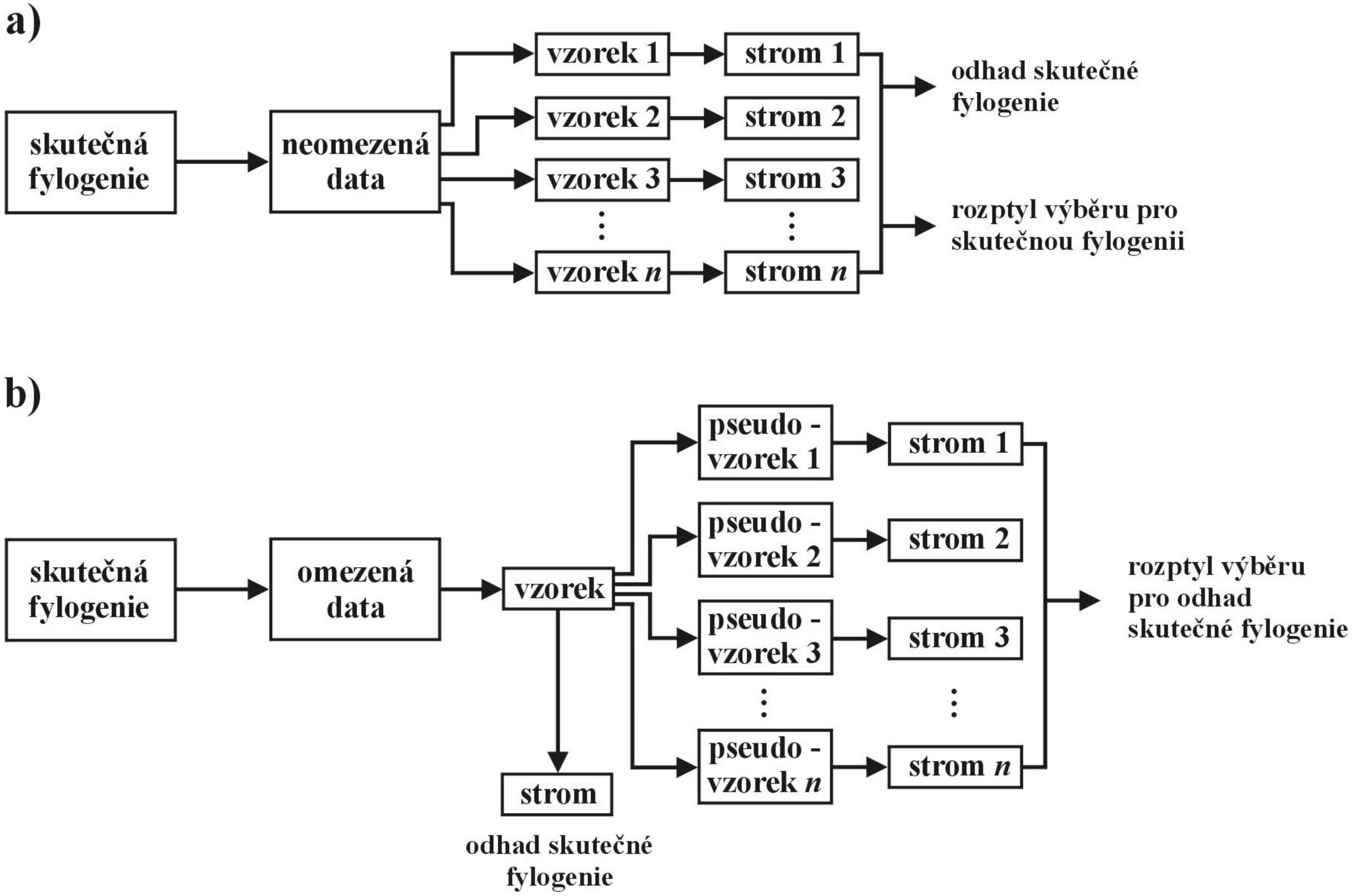
Resampling methods

without replacement = jackknife
with replacement = bootstrap

bootstrap:


a)
skutečná fylogenie
neomezená data
vzorek 1
vzorek 2
vzorek 3
vzorek *n*
strom 1
strom 2
strom 3
strom *n*
odhad skutečné fylogenie
rozptyl výběru pro skutečnou fylogenii

bootstrap:



a)

skutečná fylogenie

neomezená data

vzorek 1

vzorek 2

vzorek 3

⋮

vzorek *n*

strom 1

strom 2

strom 3

⋮

strom *n*

odhad skutečné fylogenie

rozptyl výběru pro skutečnou fylogenii

b)

skutečná fylogenie

omezená data

vzorek

strom

odhad skutečné fylogenie

pseudo - vzorek 1

pseudo - vzorek 2

pseudo - vzorek 3

⋮

pseudo - vzorek *n*

strom 1

strom 2

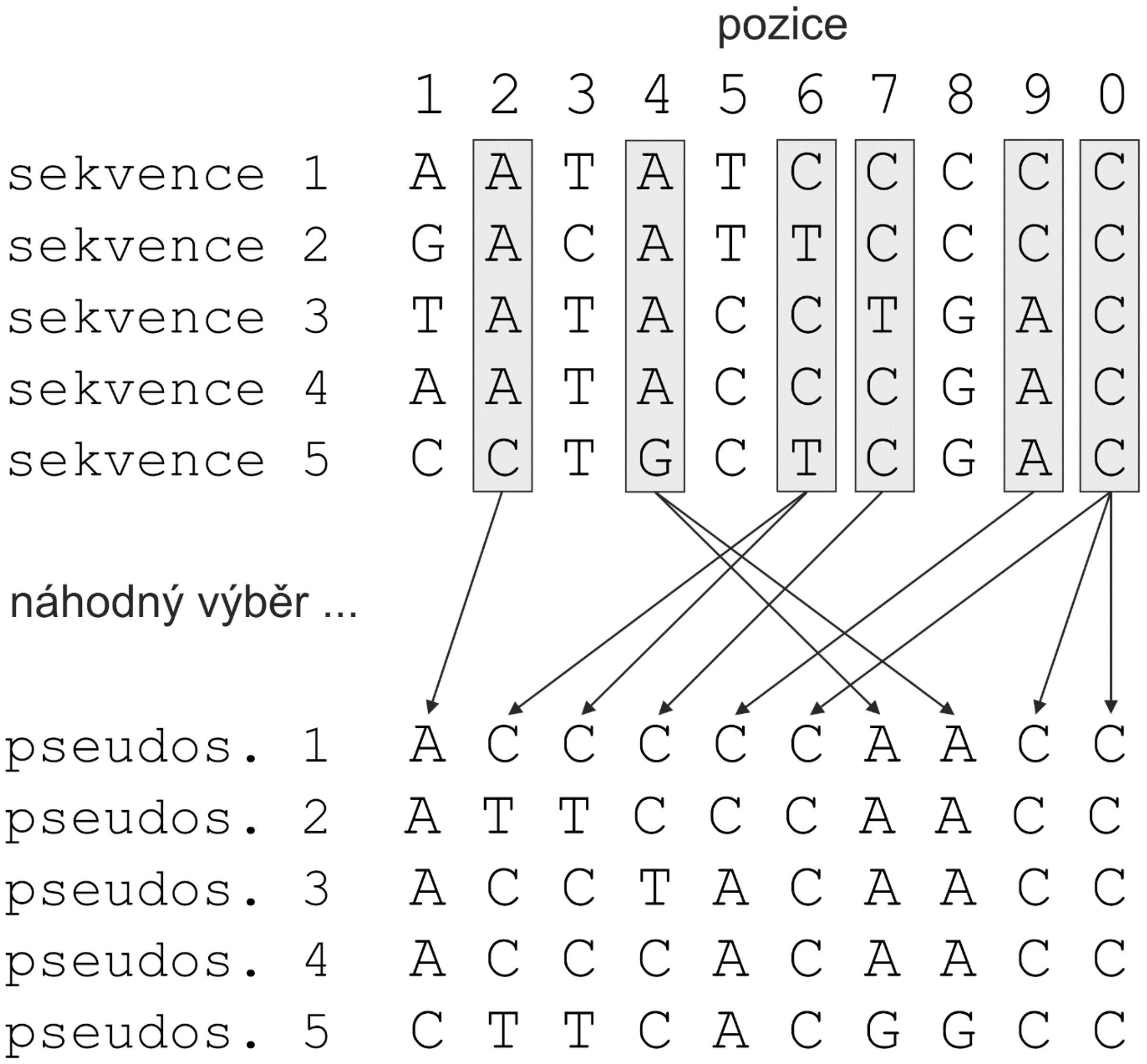
strom 3

⋮

strom *n*

rozptyl výběru pro odhad skutečné fylogenie

bootstrap:



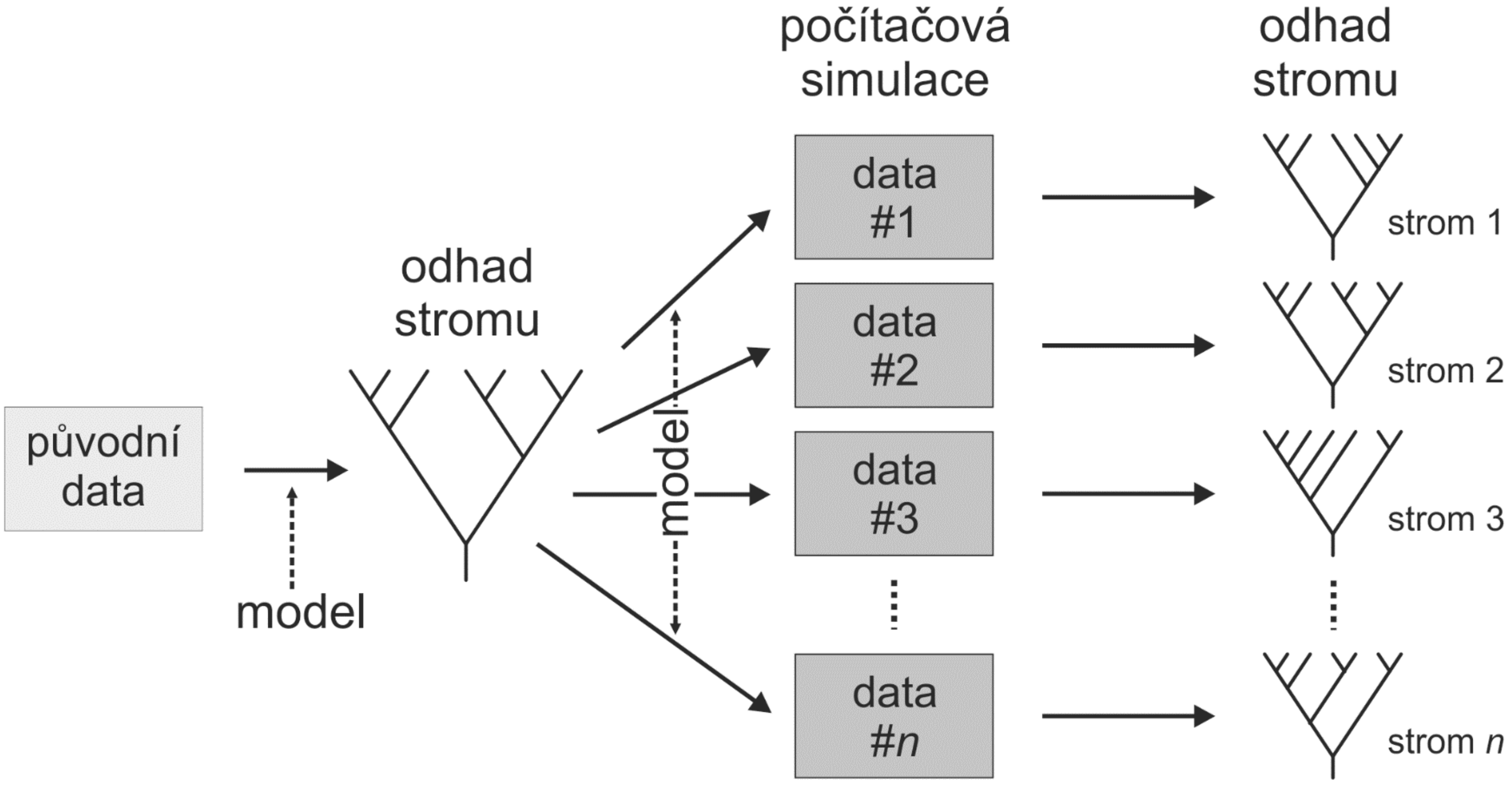
pozice

	1	2	3	4	5	6	7	8	9	0
sekvence 1	A	A	T	A	T	C	C	C	C	C
sekvence 2	G	A	C	A	T	T	C	C	C	C
sekvence 3	T	A	T	A	C	C	T	G	A	C
sekvence 4	A	A	T	A	C	C	C	G	A	C
sekvence 5	C	C	T	G	C	T	C	G	A	C

náhodný výběr ...

pseudos. 1	A	C	C	C	C	C	A	A	C	C
pseudos. 2	A	T	T	C	C	C	A	A	C	C
pseudos. 3	A	C	C	T	A	C	A	A	C	C
pseudos. 4	A	C	C	C	A	C	A	A	C	C
pseudos. 5	C	T	T	C	A	C	G	G	C	C

parametric bootstrap: evolutionary model

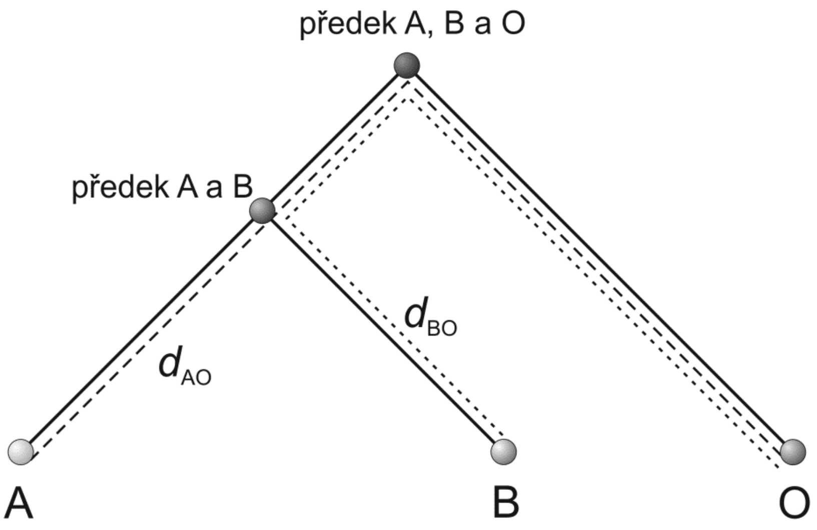


Bayesian analysis: posterior probabilities

Hypothesis testing

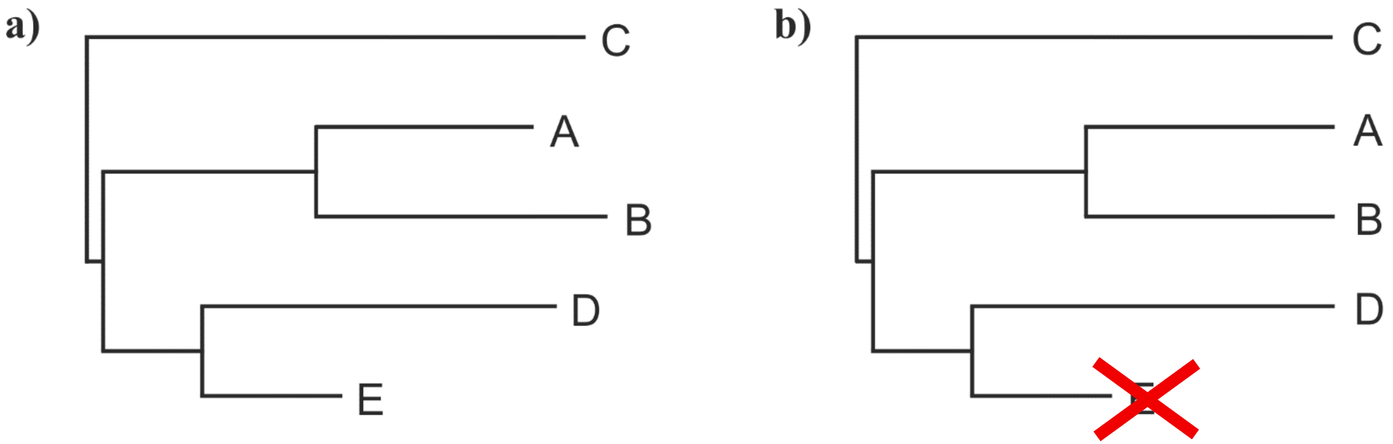
Test of molecular clock:

Relative rate test (RRT): AC=BC?



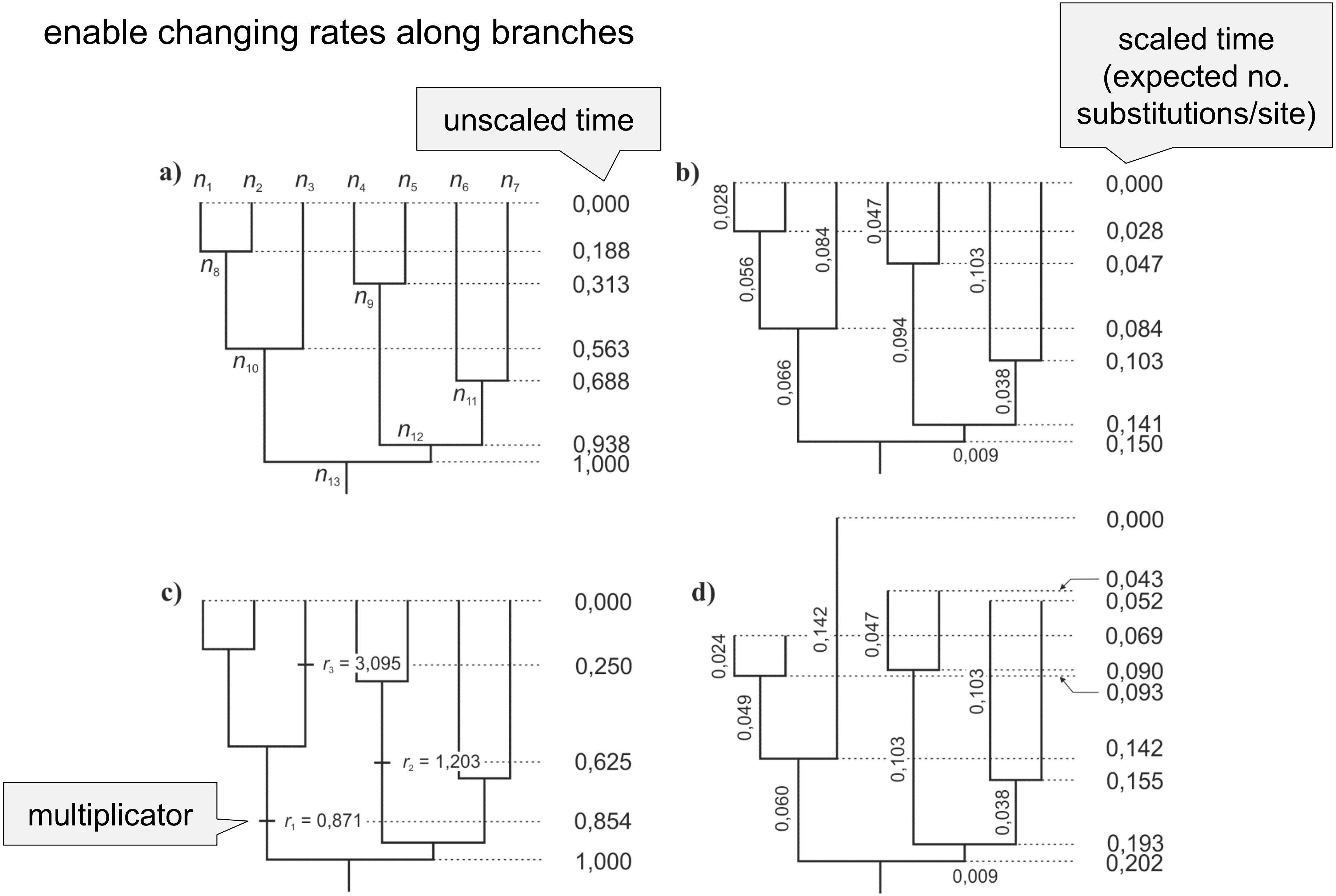
Linearized trees

removing significantly different taxa



Relaxed molecular clock

enable changing rates along branches



unscaled time

scaled time (expected no. substitutions/site)

multiplicator

Tree comparison

Are two trees significantly different?

Tests of paired positions:

winning sites test

Felsenstein’s *z* test

Templeton’s test

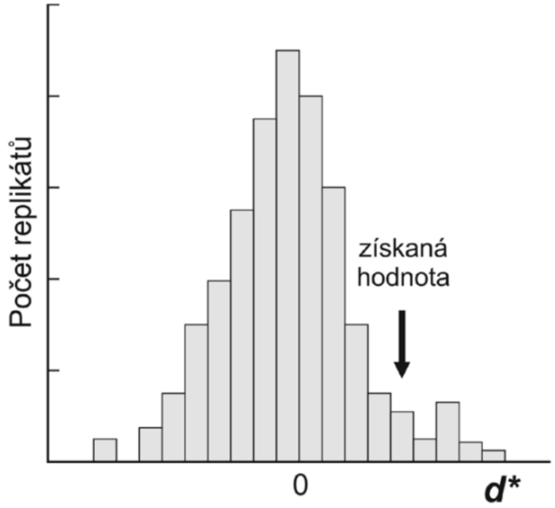
Kishino-Hasegawa test (KHT, RELL)

a)

$d^*_i = \mathrm{lnL}^*_{T1} - \mathrm{lnL}^*_{T2}$,
kde *i* je bootstrapový replikát

$d^*_1 = \mathrm{lnL}^*_{T1} - \mathrm{lnL}^*_{T2}$
$d^*_2 = \mathrm{lnL}^*_{T1} - \mathrm{lnL}^*_{T2}$
$d^*_3 = \mathrm{lnL}^*_{T1} - \mathrm{lnL}^*_{T2}$
...

$d^*_n = \mathrm{lnL}^*_{T1} - \mathrm{lnL}^*_{T2}$



For more than two trees:

Shimodaira-Hasegawa (SH) test

Tree comparison

To what degree are two trees different?

Tree distances:

partition metric

quartet metric

path difference metric

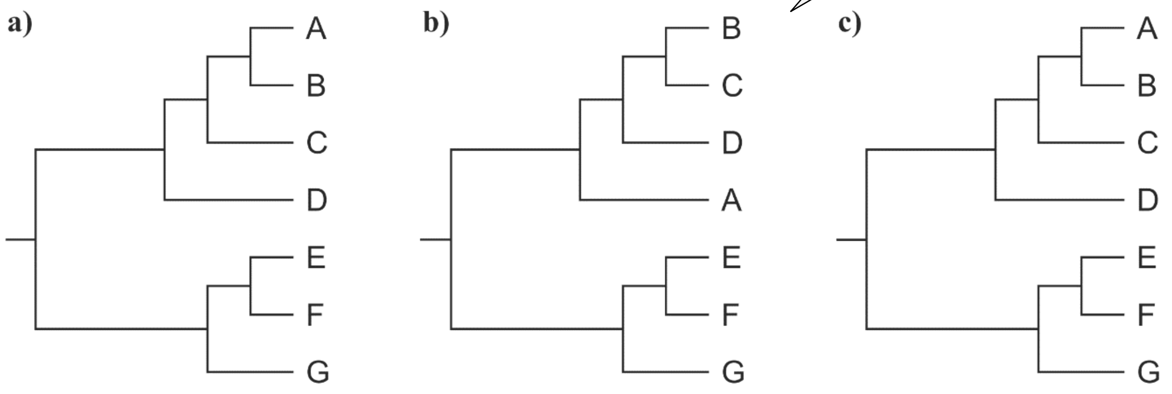
methods incorporating branch lengths

Problems with tree distances

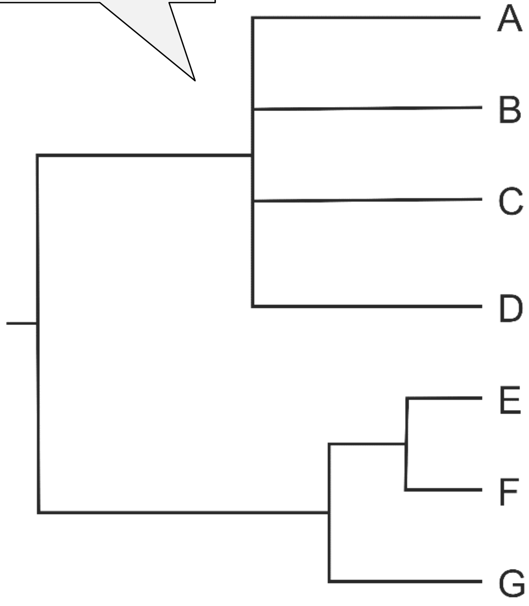
Consensus trees

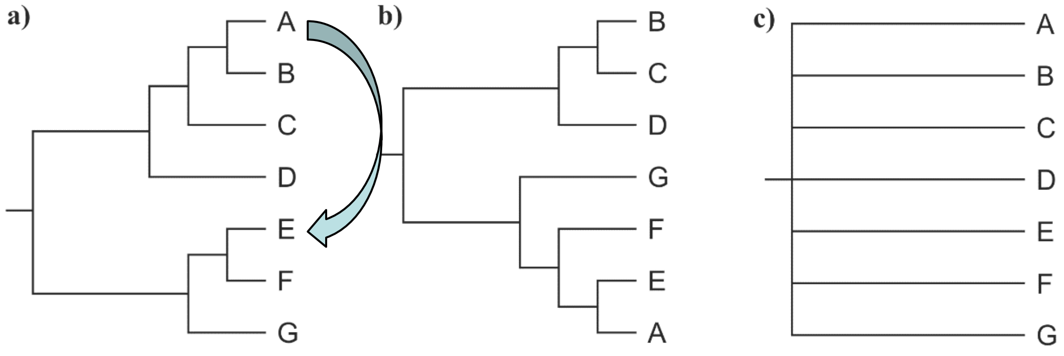
source trees

strict consensus



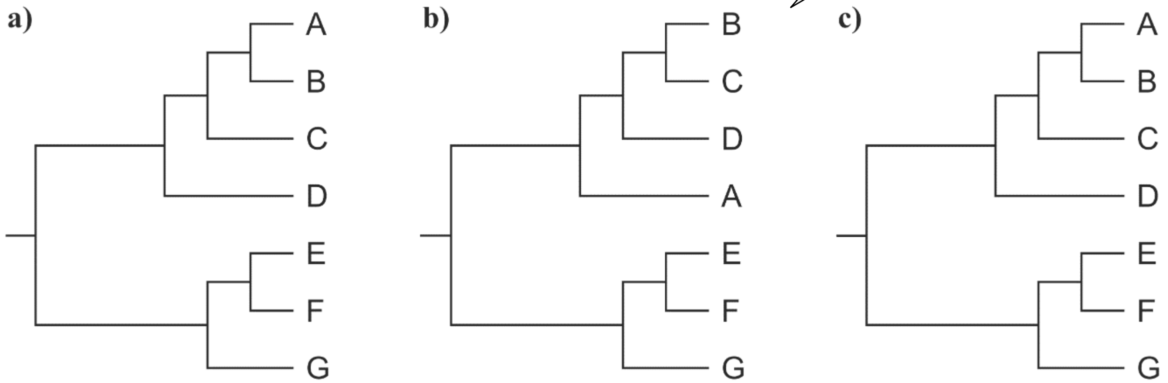
strict consensus tree

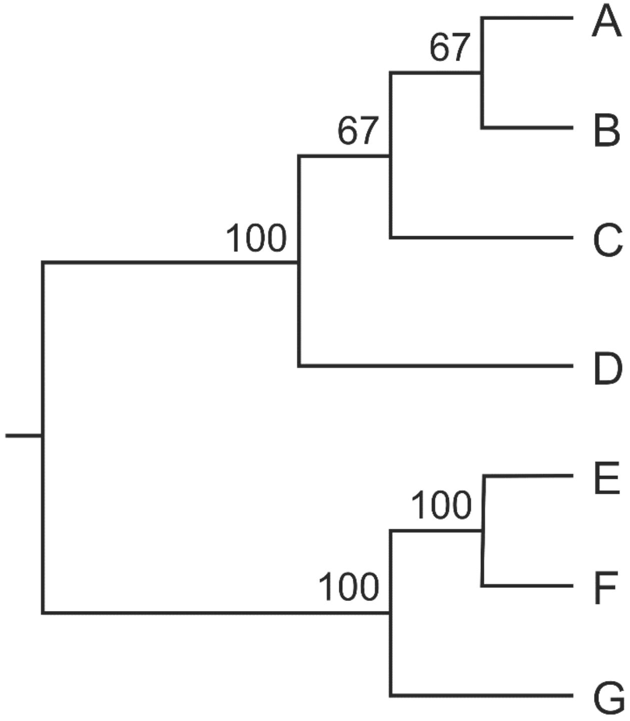




majority-rule

source trees





majority-rule consensus

Consensus trees

problem with consensus trees – combined vs. separate analysis, supermatrix vs. supertree

consensus trees in resampling methods, Bayesian analysis

Phylogenetic programs

alignment:

ClustalX *http://inn-prot.weizmann.ac.il/software/ClustalX.html*

phylogeny inference:

http://evolution.gs.washington.edu/phylip/software.html

PAUP*
PHYLIP
McClade ... MP
MOLPHY, PHYML, TREE-PUZZLE ... ML
MrBayes ... BA

managing trees:

TreeView *http://taxonomy.zoology.gla.ac.uk/rod/treeview.html*