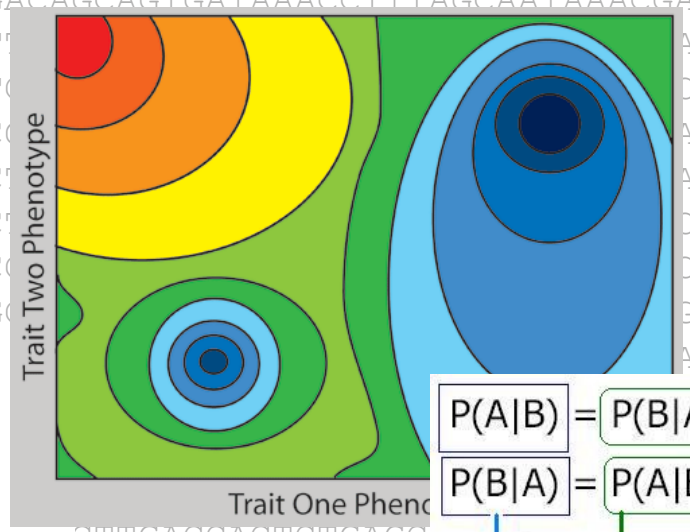
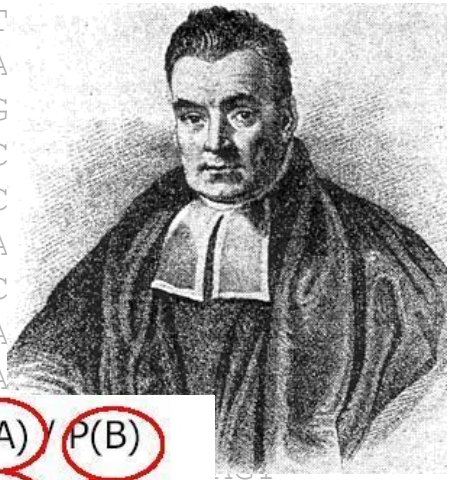
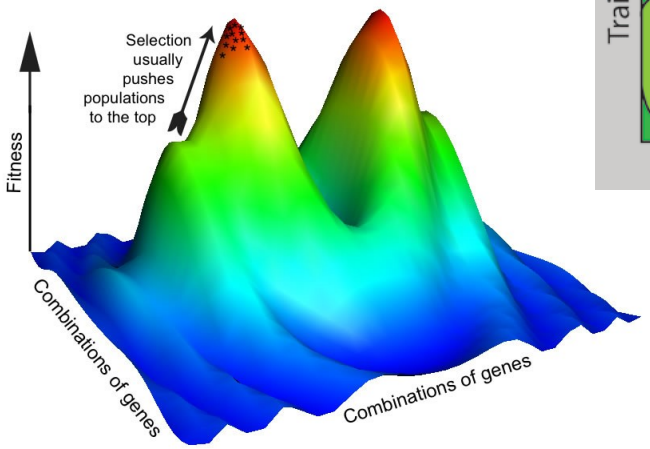
MODULARIZACE VÝUKY EVOLUČNÍ A EKOLOGICKÉ BIOLOGIE
CZ.1.07/2.2.00/15.0204

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

esf evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

MAXIMUM LIKELIHOOD, ML (maximální věrohodnost)

15 coin tosses:

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

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

$$L = \Pr(D \mid H) = \Pr(7\times \text{ head}, 8\times \text{ tail} \mid \text{hypothesis})$$

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

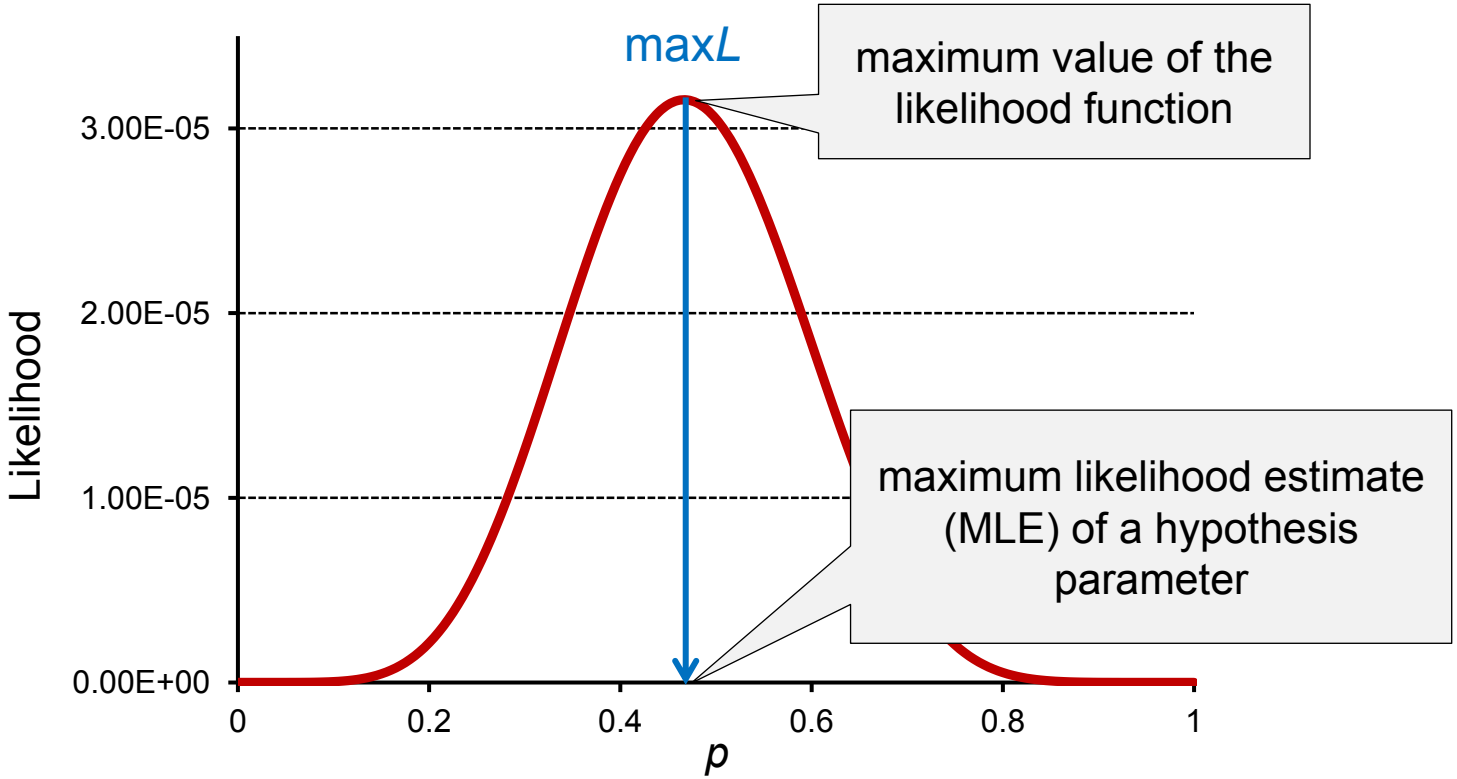
score TTHHHTHTTTHTHHT [7× head (H), 8× tail (T)]

Because tosses independent $\Rightarrow$ 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 $\approx$ 7/15



Hypothesis?

Eg. H1 = coin is not „biased“, ie. $p = 1/2 \Rightarrow L = 3{,}0517.10^{-5}$

H2 = coin is biased so that we get tail in 2/3 cases:

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

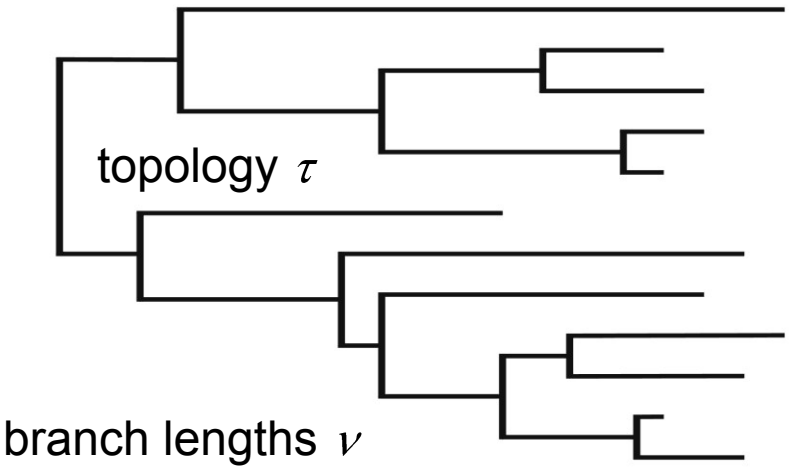
$\Rightarrow$ result of tosses 1,7$\times$ more probable with unbiased coin

Maximum likelihood in phylogenetic analysis

data:

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

tree:

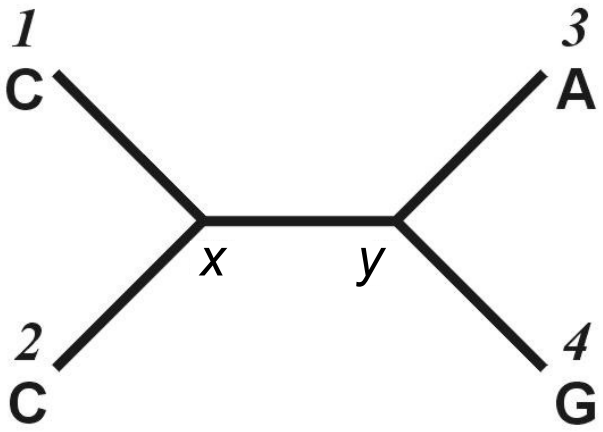


+ evolutionary model θ

= hypothesis

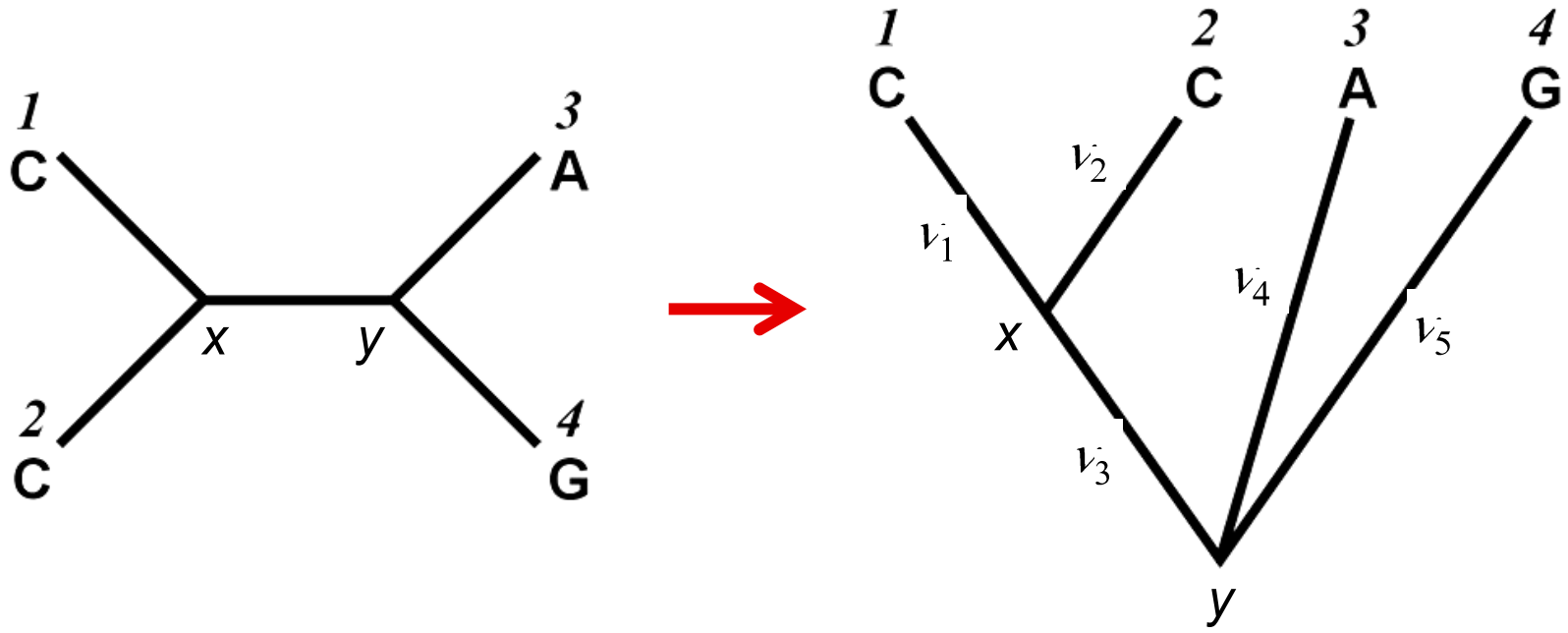
$L = P(D \mid H)$: D = sequence matrix (data), H = τ (topology) + ν (branch lenghts) + θ (model)

	1	j	N
1	TCAAAAATGGCTTTATTCG	**C**	TTAATGCCGTTAACCCTTGCGGGGGCCATG
2	TCCGTGATGGATTTATTTC	**C**	GCAATGCCTGTCATCTTATTCTCAAGTATC
3	TTCGTGATGGATTTATTGC	**A**	GGTATGCCAGTCATCCTTTTCTCATCTATC
4	TTCGTGACGGGTTTATCTC	**G**	GCAATGCCGGTCATCCTATTTTCGAGTATT



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

ν_i = branch lengths



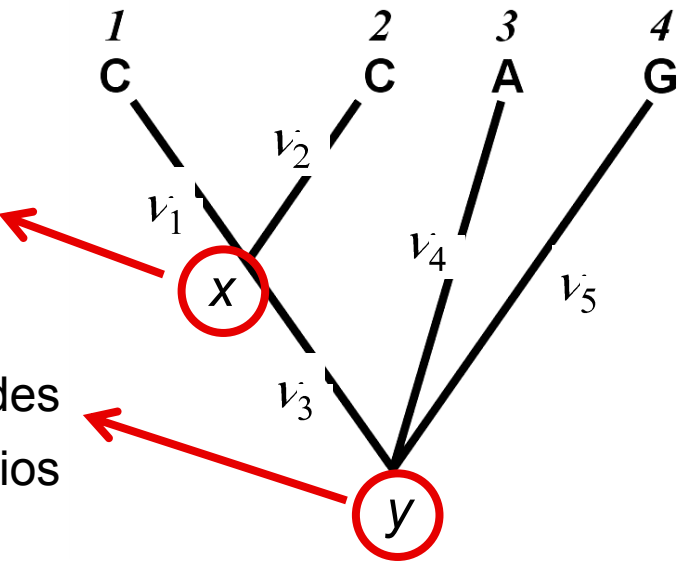
	1	*j*	*N*
1	TCAAAAATGGCTTTATTCG	**C**	TTAATGCCGTTAACCCTTGCGGGGGCCATG
2	TCCGTGATGGATTTATTTC	**C**	GCAATGCCTGTCATCTTATTCTCAAGTATC
3	TTCGTGATGGATTTATTGC	**A**	GGTATGCCAGTCATCCTTTTCTCATCTATC
4	TTCGTGACGGGTTTATCTC	**G**	GCAATGCCGGTCATCCTATTTTCGAGTATT

ν_i = branch lengths

x: 4 nucleotides

y: 4 nucleotides

$\Rightarrow 4 \times 4 = 16$ possible scenarios



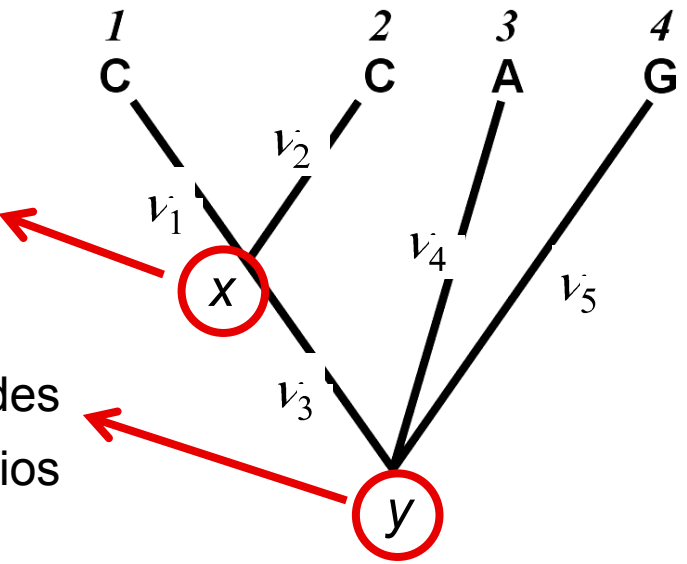
$$L(1) = P(y) \times P(y \to x)\,\nu_3 \times P(x \to \mathrm{C})\,\nu_1 \times P(x \to \mathrm{C})\,\nu_2 \times P(y \to \mathrm{A})\,\nu_4 \times P(y \to \mathrm{G})\,\nu_5$$

E.g. $L(1;\ x{=}\mathrm{A}, y{=}\mathrm{T}) = P(\mathrm{T}) \times P(\mathrm{T} \to \mathrm{A})\,\nu_3 \times P(\mathrm{A} \to \mathrm{C})\,\nu_1 \times P(\mathrm{A} \to \mathrm{C})\,\nu_2 \times P(\mathrm{T} \to \mathrm{A})\,\nu_4 \times P(\mathrm{T} \to \mathrm{G})\,\nu_5$

$L(j)$ = *P*(scenario 1) + …. + *P*(scenario 16)

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

ν_i = branch lengths



x: 4 nucleotides

y: 4 nucleotides

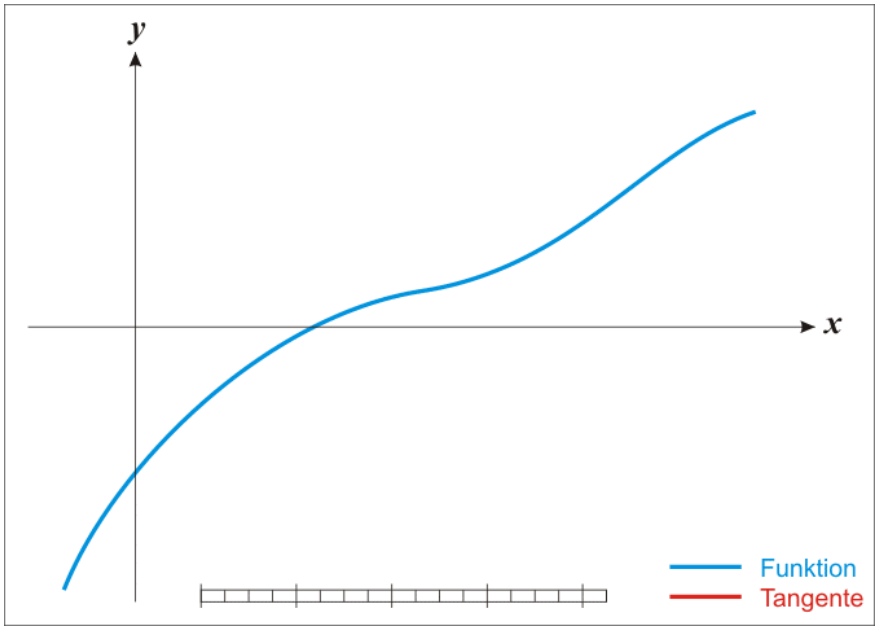
$\Rightarrow 4 \times 4 = 16$ possible scenarios

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

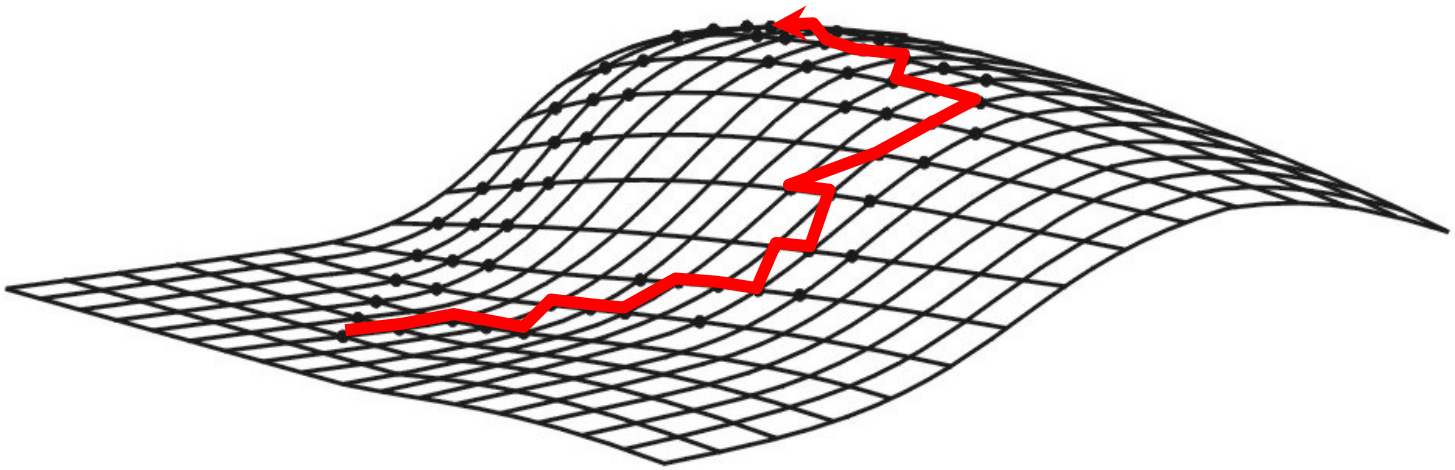
→ eg. Newton (Newton-Raphson) method



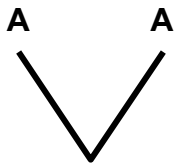
https://upload.wikimedia.org/wikipedia/commons/e/e0/NewtonIteration_Ani.gif

Maximum likelihood tree search: heuristic search

Heuristic search



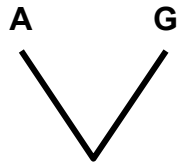
stepwise addition ... eg. PHYLIP
star decomposition ... eg. MOLPHY; neighbor-joining tree
branch swapping



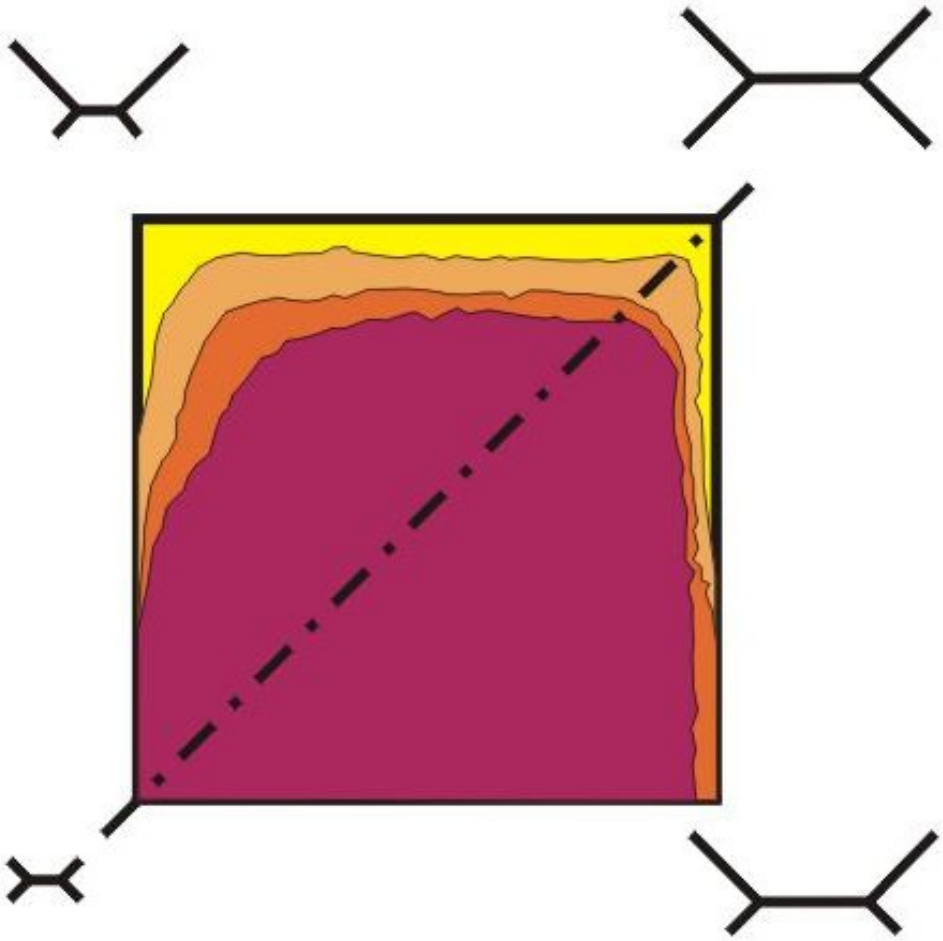
Likelihood (ML) and parsimony (MP)

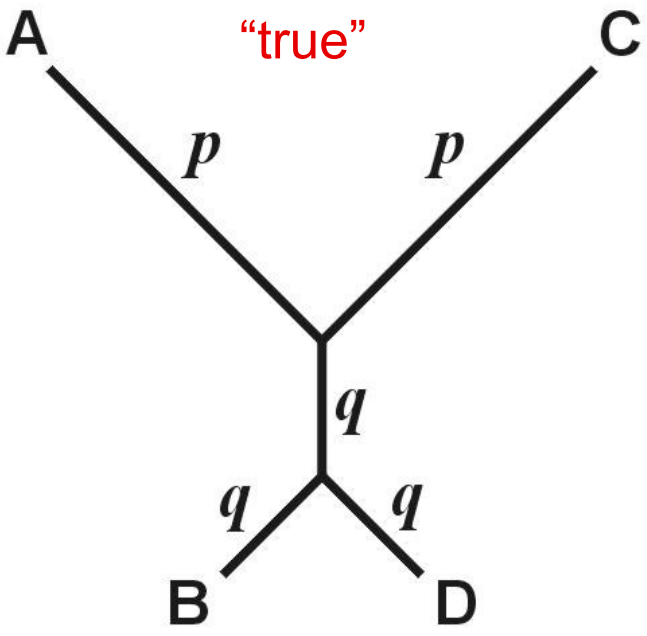
No. changes Parsimony

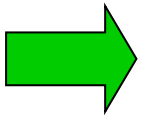
No. changes Parsimony

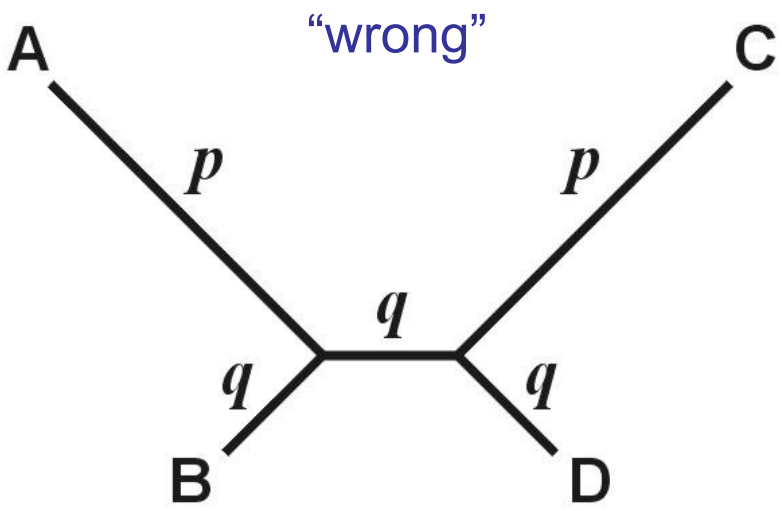


Likelihood and consistency



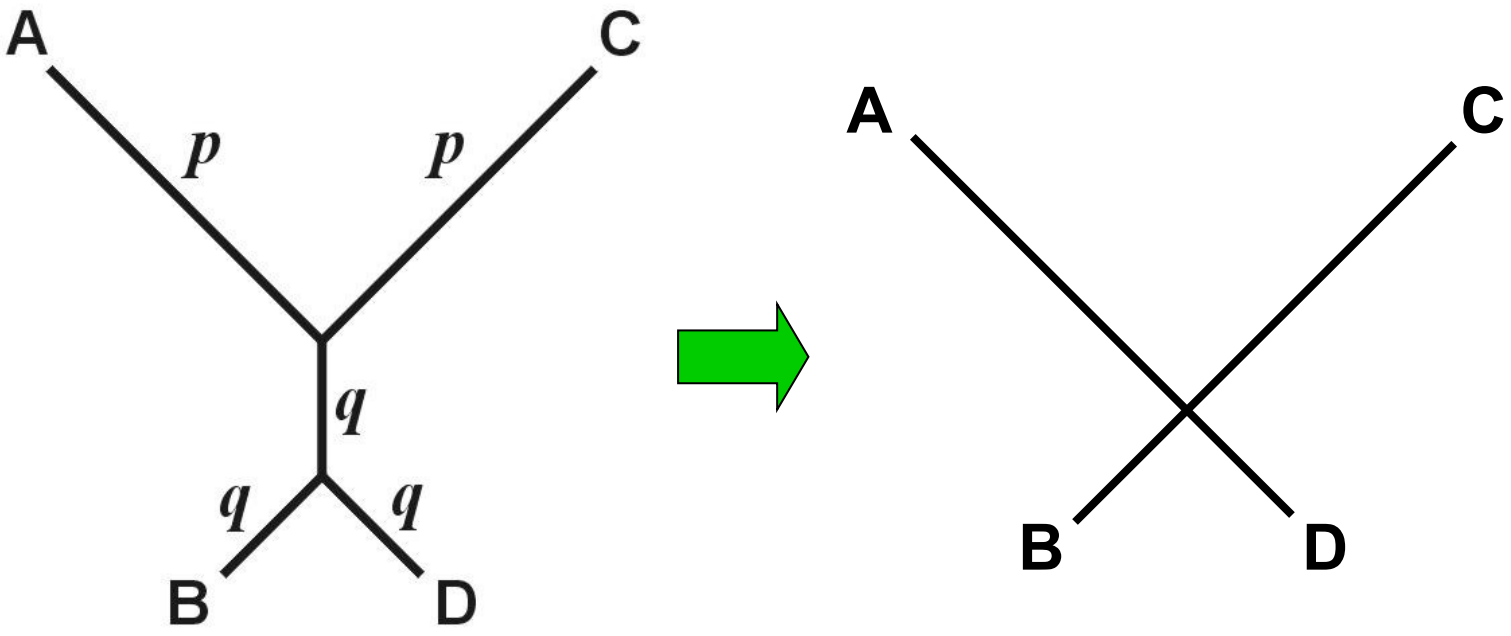






Farris (anti-Felsenstein, inverse Felsenstein) zone

"long-branch repulsion"

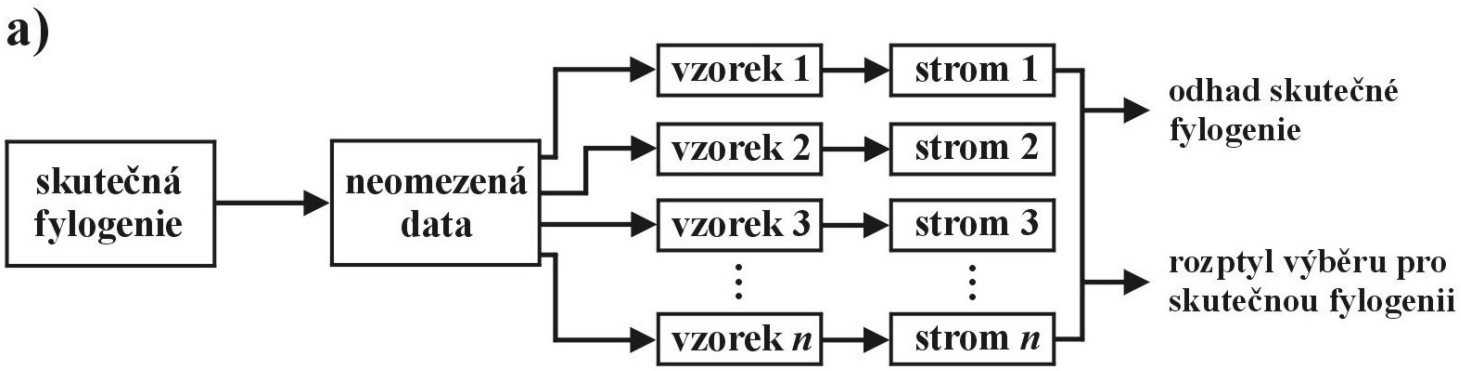

A
C
p
p
q
q
q
B
D
A
C
B
D

Measuring tree reliability

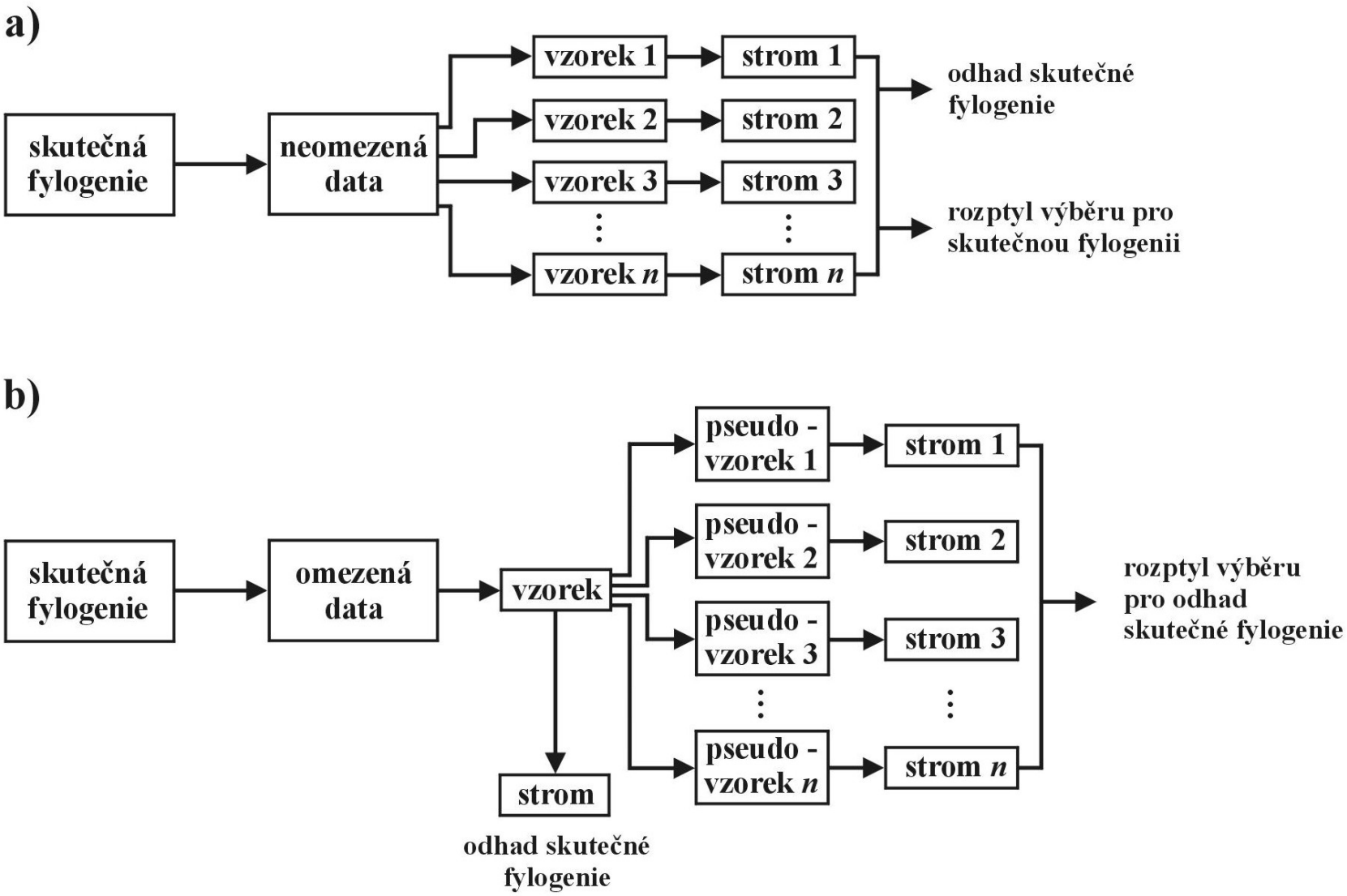
Resampling methods

without replacement = jackknife
with replacement = bootstrap

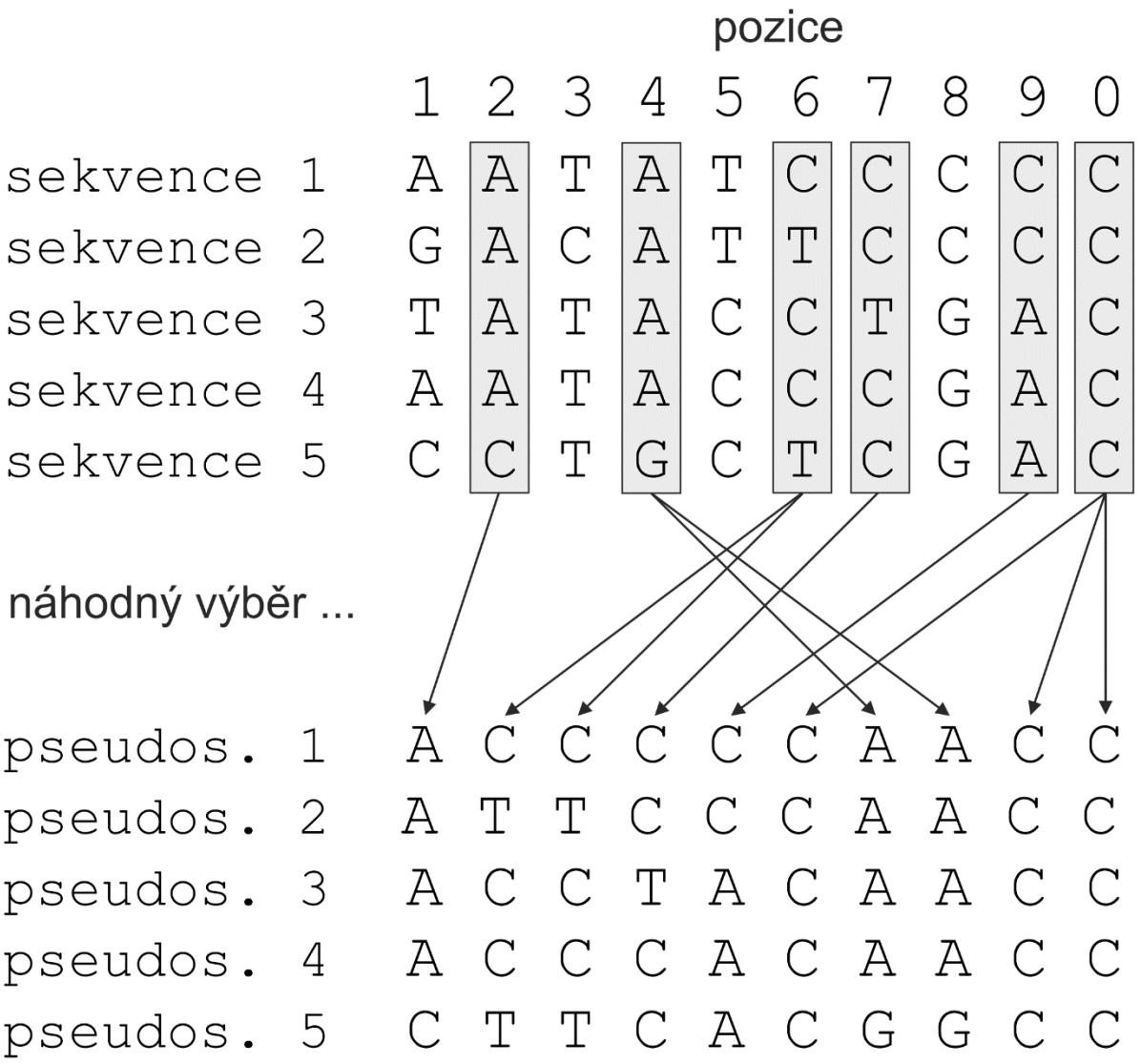
bootstrap:


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

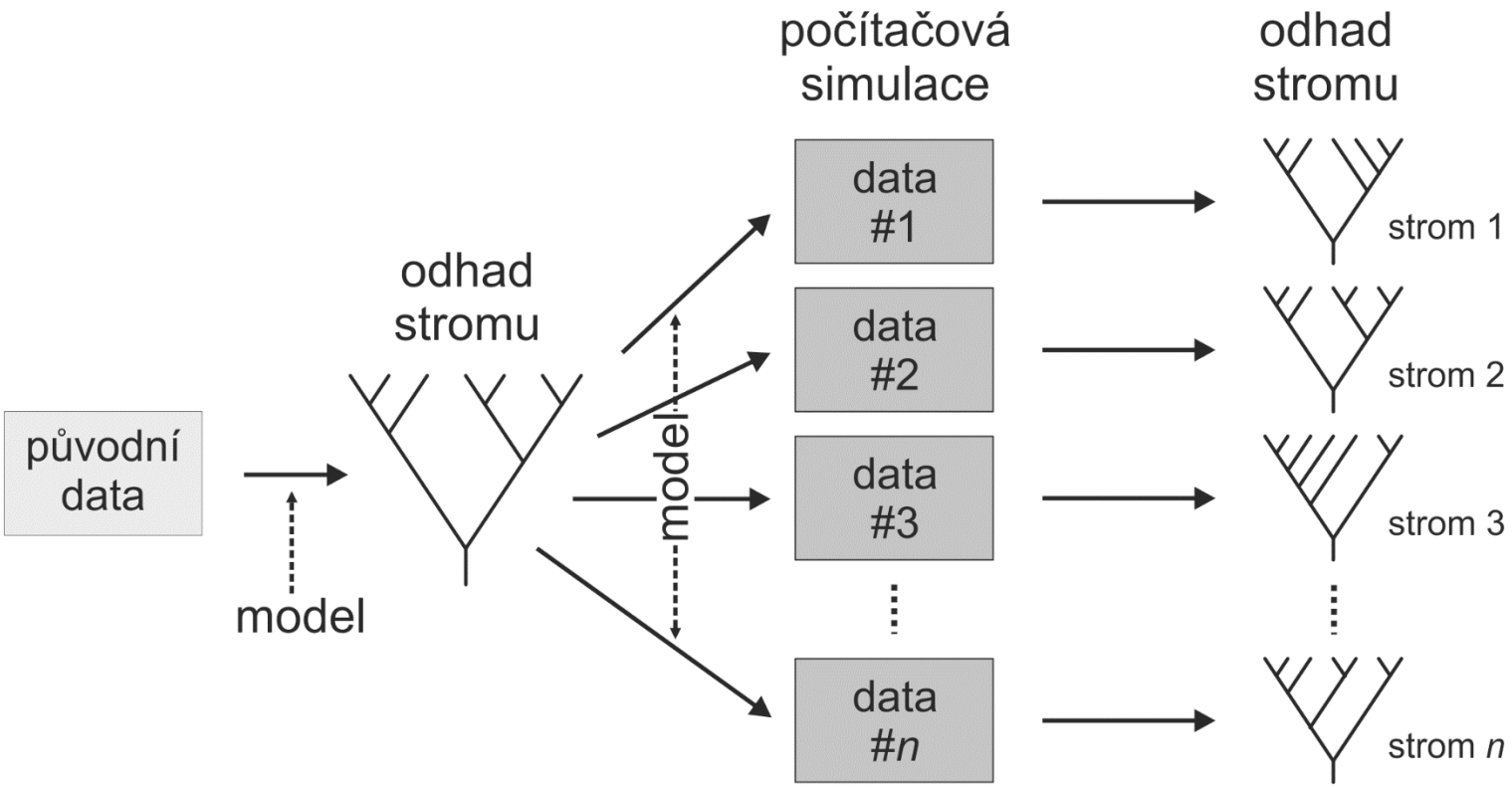
bootstrap:


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

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

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

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

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

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

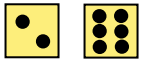
sum of numerators across all alternative hypotheses

Thomas Bayes

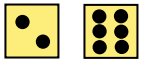
For our example of 2 dice throws:

	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

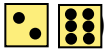
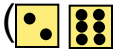
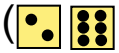
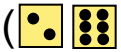
prior probability (biased) = 0,2
(20/100 biased dice in the set)

Pr. of getting  with unbiased dice:

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

Pr. of getting  with biased dice:

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

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

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

Bayesian method in phylogenetic analysis:

posterior probability

likelihood

prior probability

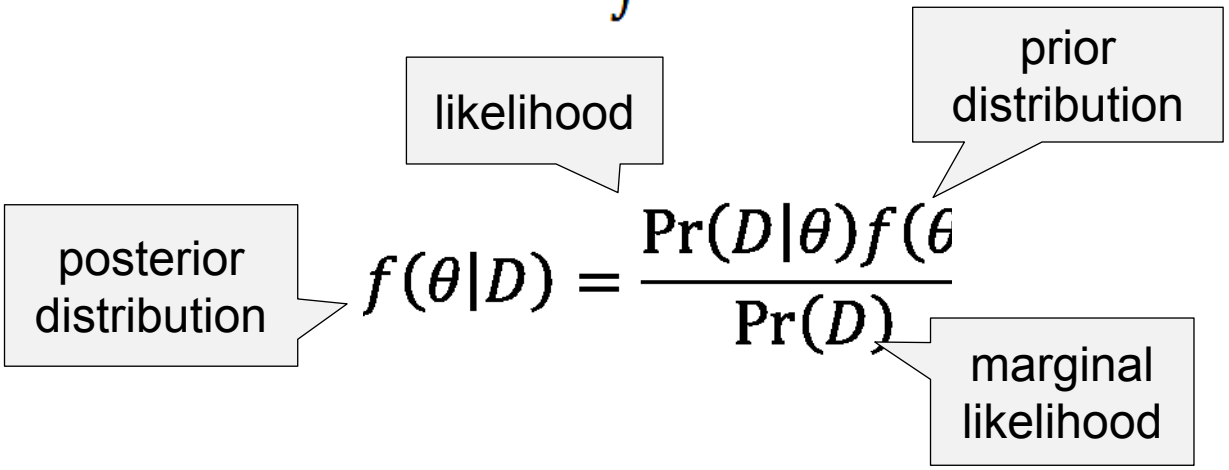
$$P(\boldsymbol{\tau}, \boldsymbol{\nu}, \boldsymbol{\theta}|\mathbf{X}) = \frac{P(\mathbf{X}|\boldsymbol{\tau}, \boldsymbol{\nu}, \boldsymbol{\theta})P(\boldsymbol{\tau}, \boldsymbol{\nu}, \boldsymbol{\theta})}{\sum_{i=1}^{B(s)}[P(\mathbf{X}|\boldsymbol{\tau}, \boldsymbol{\nu}, \boldsymbol{\theta})P(\boldsymbol{\tau},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

sum across all hypotheses (= marginal likelihood)

Parameters of Bayesian analysis mostly continuous ⇒

$P \rightarrow$ probability density functions

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

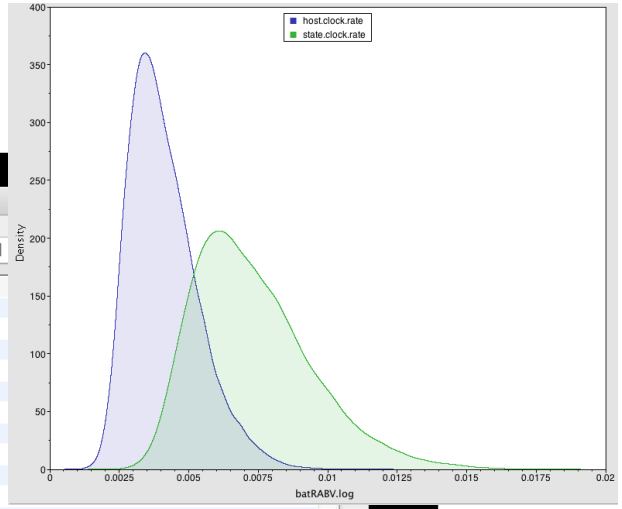


θ = set of (continuous) parameters in the model, including the tree, substitution model parameters, clock rates, etc.

either ML estimates → empirical BA

or all combinations → hierarchical BA

outcome = probability distribution



Tracer

Trace Files:

Trace File	States	Burn-In
batRABV_short.log	114820000	11482000

Reload

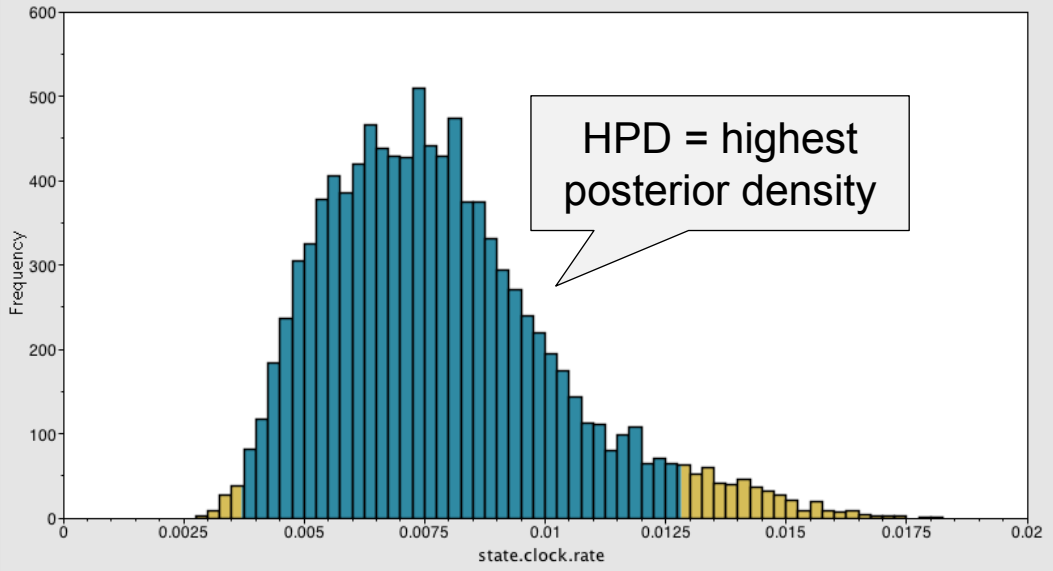
Traces:

Statistic	Mean	ESS	...
skygrid.logPopSize42	5.667	42	R
skygrid.logPopSize43	5.547	44	R
skygrid.logPopSize44	5.4	47	R
skygrid.logPopSize45	5.266	50	R
skygrid.logPopSize46	5.108	54	R
skygrid.logPopSize47	4.917	58	R
skygrid.logPopSize48	4.698	63	R
skygrid.logPopSize49	4.464	70	R
skygrid.logPopSize50	4.198	82	R
skygrid.cutOff	250	1	R
kappa	11.1	7622	R
frequencies1	0.292	4174	R
frequencies2	0.217	4908	R
frequencies3	0.212	4392	R
frequencies4	0.279	4428	R
alpha	0.214	8285	R
batRABV.clock.rate	2.48E-4	24	R
host.clock.rate	4.339E-3	28	R
state.clock.rate	7.824E-3	27	R
host.nonZeroRates	21.985	1926	I
state.nonZeroRates	19.285	114	I
root.host	n/a	-	C
root.state	n/a	-	C
host.rates.Ap.Ef	0.945	310	R
host.rates.Ap.Lb	0.954	467	R
host.rates.Ap.Lbl	0.974	457	R
host.rates.Ap.Lc	0.759	824	R
host.rates.Ap.Li	1.003	464	R
host.rates.Ap.Ln	0.975	552	R
host.rates.Ap.Ls	1.081	510	R
host.rates.Ap.Lx	1.041	492	R
host.rates.Ap.Ma	0.929	494	R
host.rates.Ap.Mc	1	513	R
host.rates.Ap.Ml	1.012	520	R
host.rates.Ap.My	1.083	265	R

Type: (R)eal (I)nt (C)at

Estimates | Marginal Density | Joint-Marginal

Summary Statistic	state.clock.rate
mean	7.824E-3
stderr of mean	4.6621E-4
stdev	2.414E-3
variance	5.8276E-6
median	7.4921E-3
value range	[2.864E-3, 0.0181]
geometric mean	7.4749E-3
95% HPD interval	[3.7032E-3, 0.0128]
auto-correlation time (ACT)	3.855E6
effective sample size (ESS)	26.8
number of samples	10335



Setup...

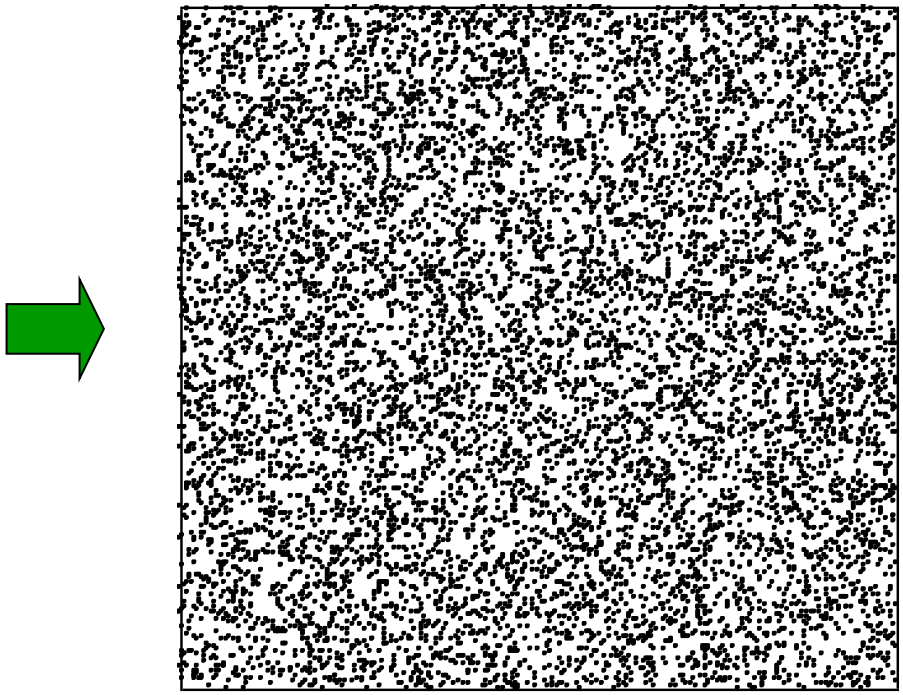
Problem: calculations too complex $\Rightarrow$ impossible to solve analytically, only numerically

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

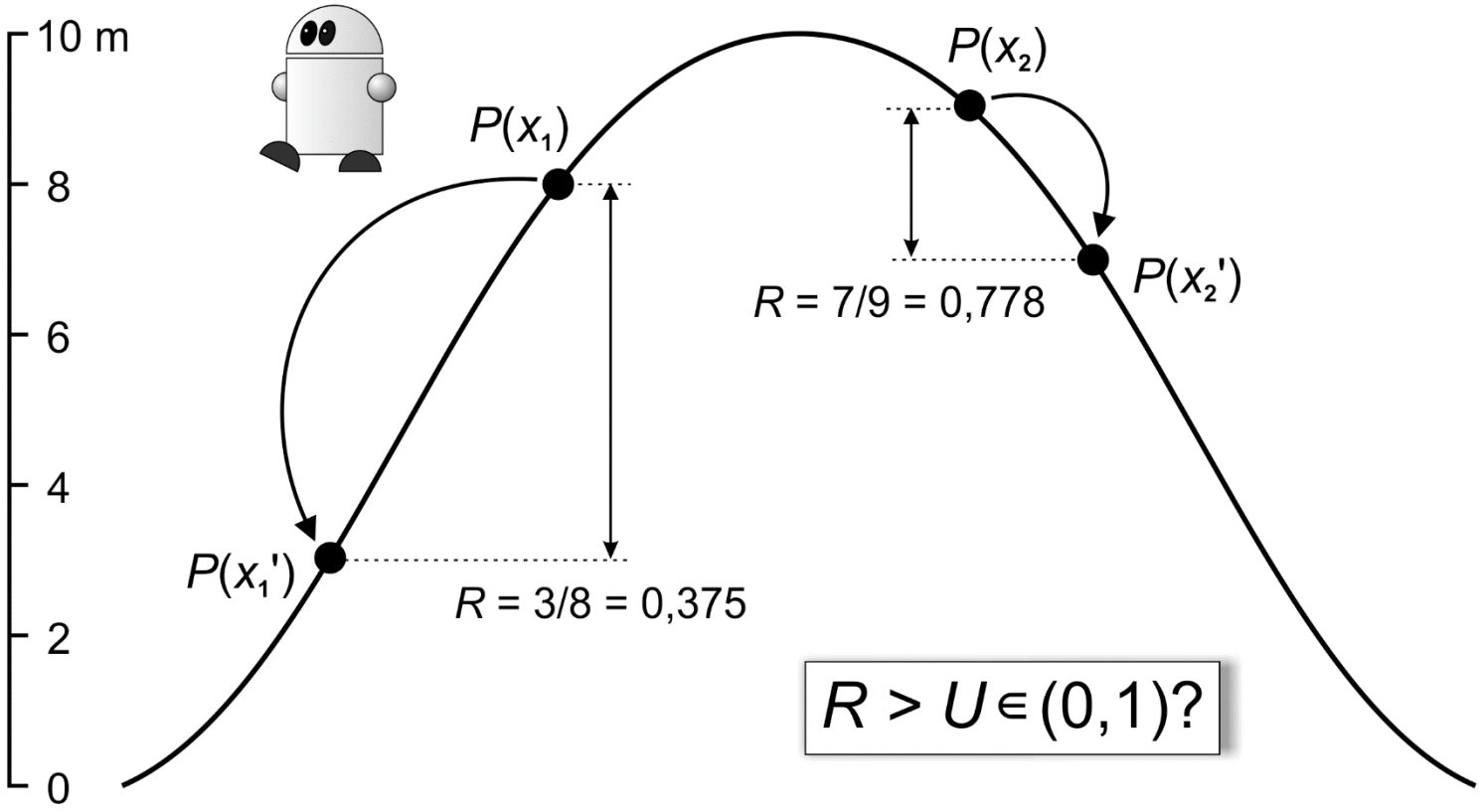
Markov chain Monte Carlo (MCMC)

Markov process: t_{-1}: A $\rightarrow$ t_0: C $\rightarrow$ t_{+1}: G
Markov property = „memorylessness“

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



Metropolis-Hastings algorithm:

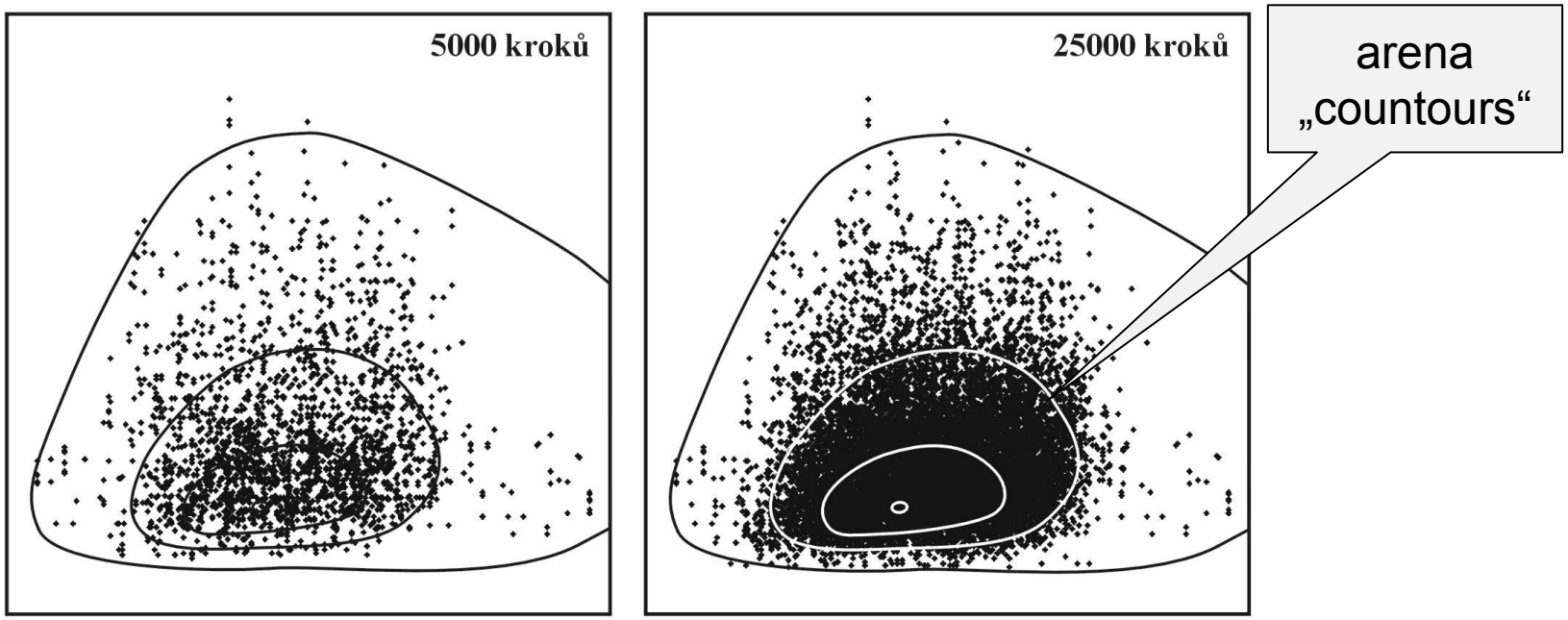


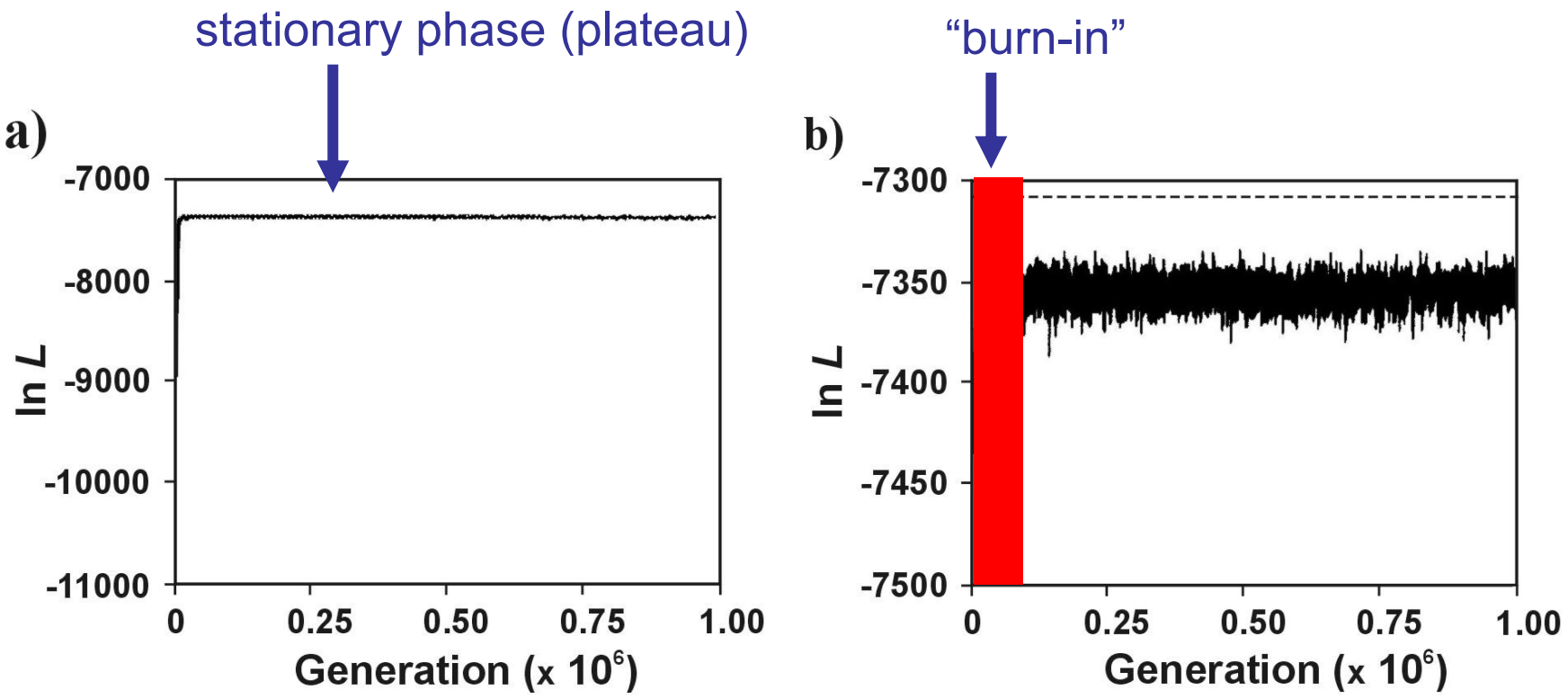
Metropolis-Hastings algorithm:

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:




stationary phase (plateau)
a)
-7000
-8000
-9000
-10000
-11000
ln L
0
0.25
0.50
0.75
1.00
Generation (x 10^6)
"burn-in"
b)
-7300
-7350
-7400
-7450
-7500
ln L
0
0.25
0.50
0.75
1.00
Generation (x 10^6)

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)

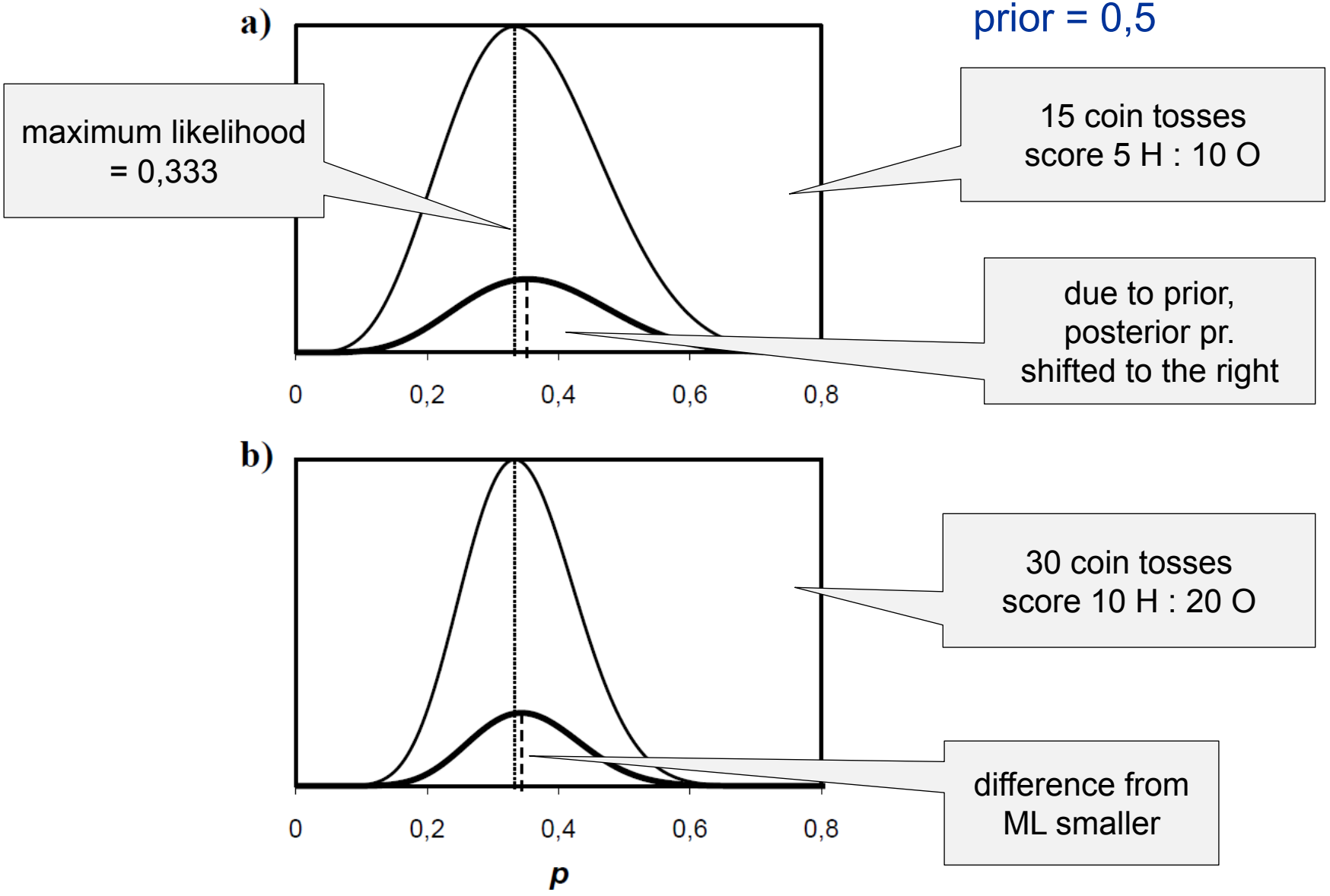
Metropolis coupled MCMC (MCMCMC, MC^3):

1 „cold“ chain, 3 „heated“ chains

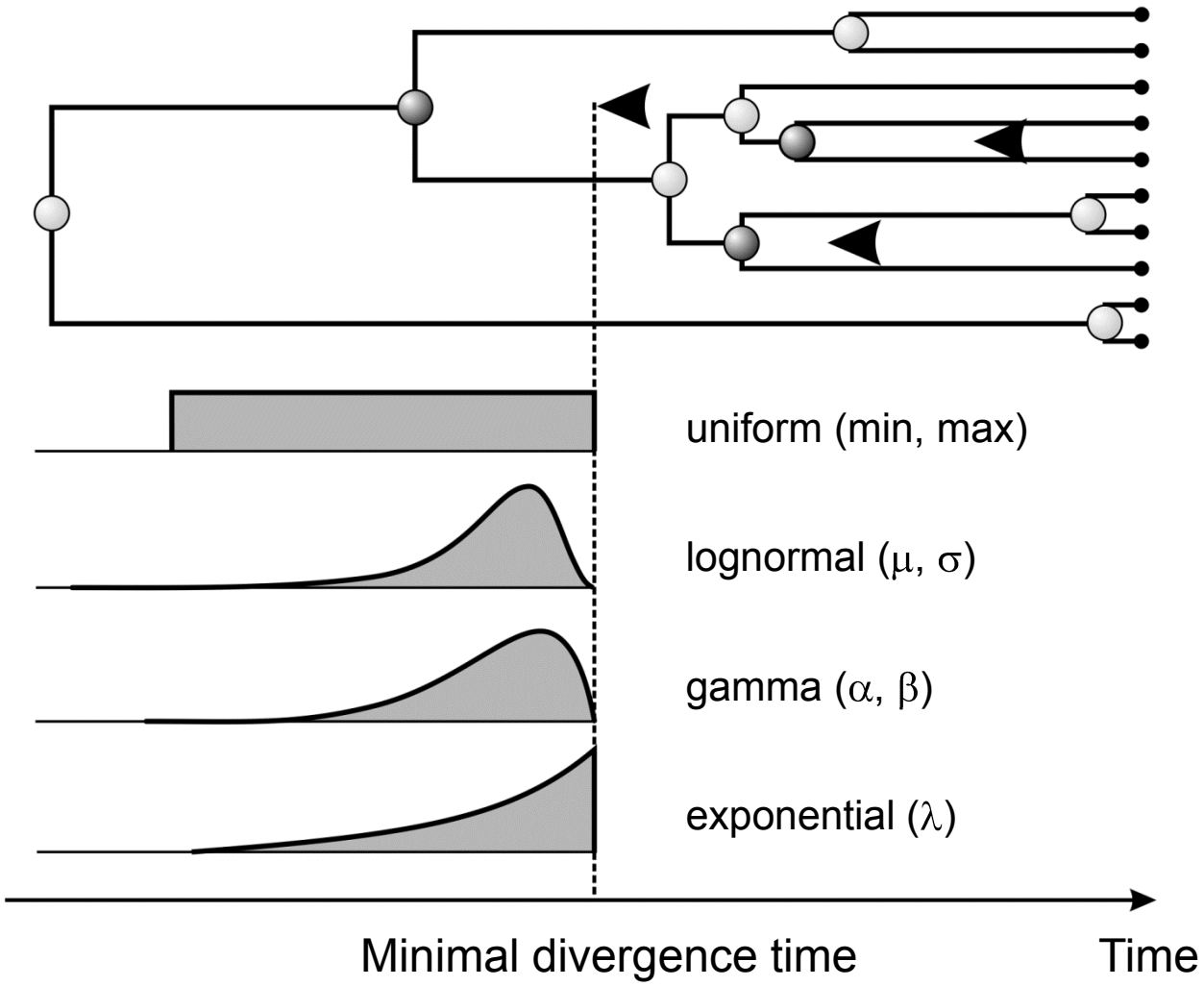
same starting point, due to stochasticity rapid divergence of „robots“

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

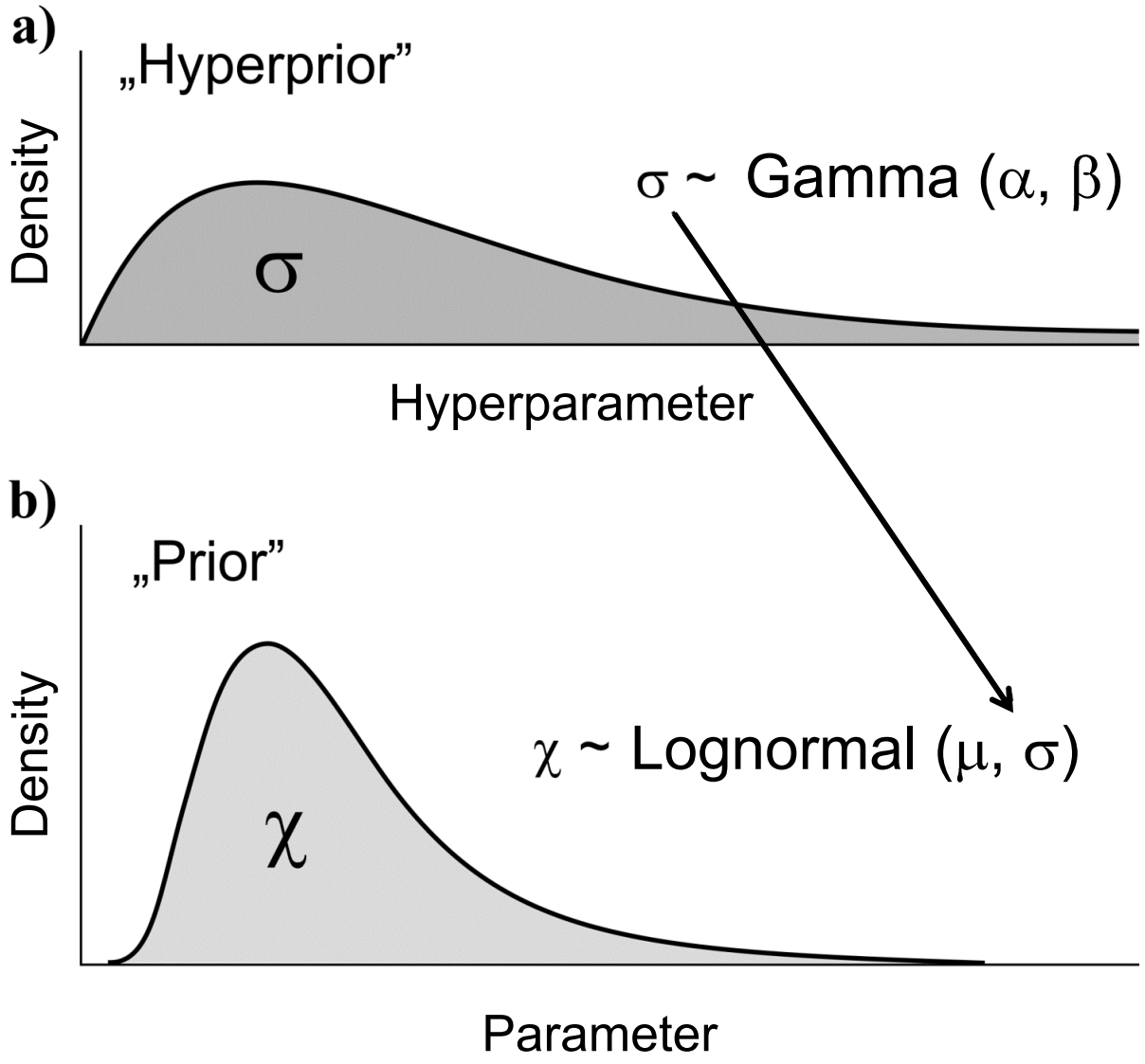
Problem with priors

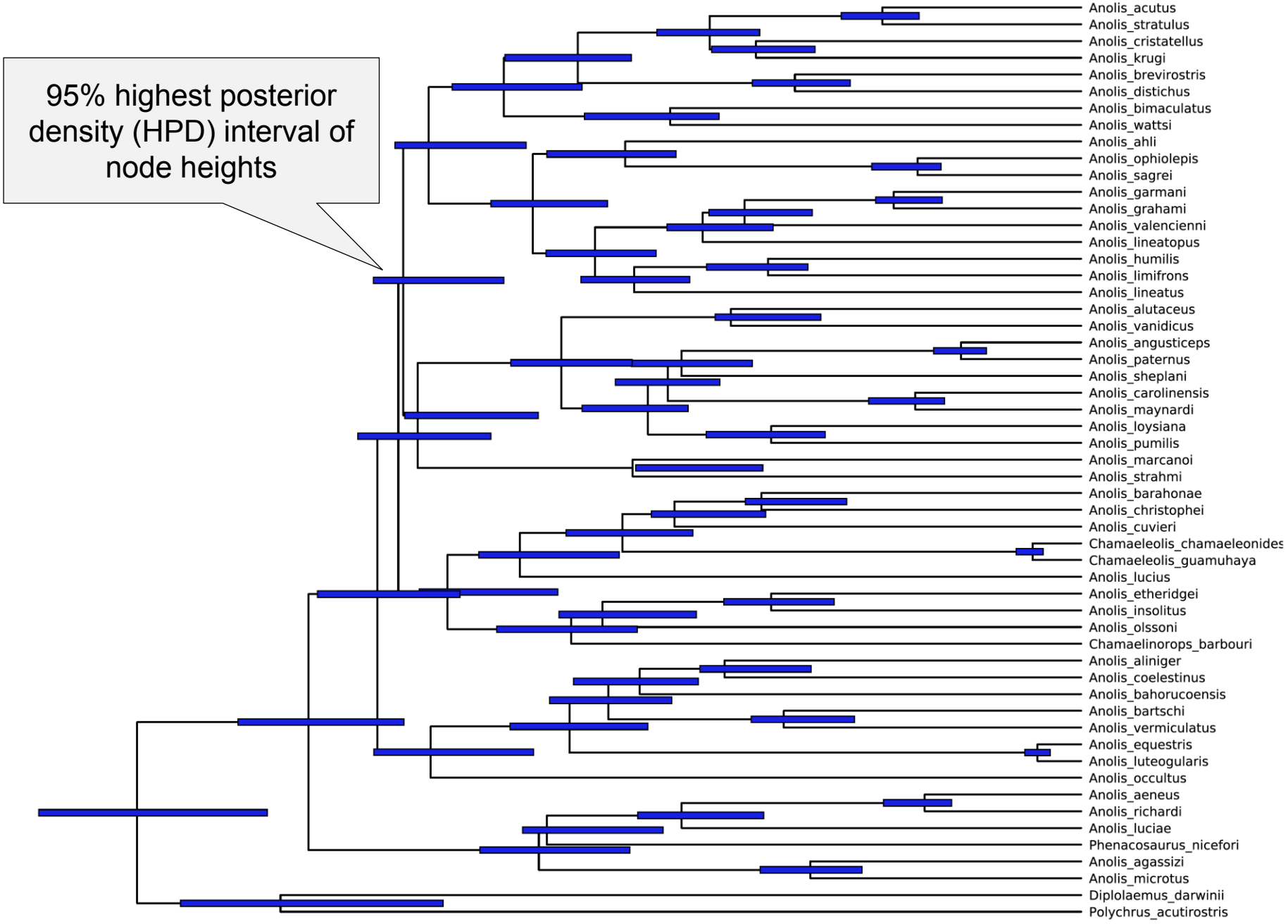


Which priors to choose?



Setting priors:




95% highest posterior density (HPD) interval of node heights
Anolis_acutus
Anolis_stratulus
Anolis_cristatellus
Anolis_krugi
Anolis_brevirostris
Anolis_distichus
Anolis_bimaculatus
Anolis_wattsi
Anolis_ahli
Anolis_ophiolepis
Anolis_sagrei
Anolis_garmani
Anolis_grahami
Anolis_valencienni
Anolis_lineatopus
Anolis_humilis
Anolis_limifrons
Anolis_lineatus
Anolis_alutaceus
Anolis_vanidicus
Anolis_angusticeps
Anolis_paternus
Anolis_sheplani
Anolis_carolinensis
Anolis_maynardi
Anolis_loysiana
Anolis_pumilis
Anolis_marcanoi
Anolis_strahmi
Anolis_barahonae
Anolis_christophei
Anolis_cuvieri
Chamaeleolis_chamaeleonides
Chamaeleolis_guamuhaya
Anolis_lucius
Anolis_etheridgei
Anolis_insolitus
Anolis_olssoni
Chamaelinorops_barbouri
Anolis_aliniger
Anolis_coelestinus
Anolis_bahorucoensis
Anolis_bartschi
Anolis_vermiculatus
Anolis_equestris
Anolis_luteogularis
Anolis_occultus
Anolis_aeneus
Anolis_richardi
Anolis_luciae
Phenacosaurus_nicefori
Anolis_agassizi
Anolis_microtus
Diplolaemus_darwinii
Polychrus_acutirostris

Time-trees and phylodynamics

Phylodynamics = synthesis of mathematical epidemiology and statistical phylogenetics

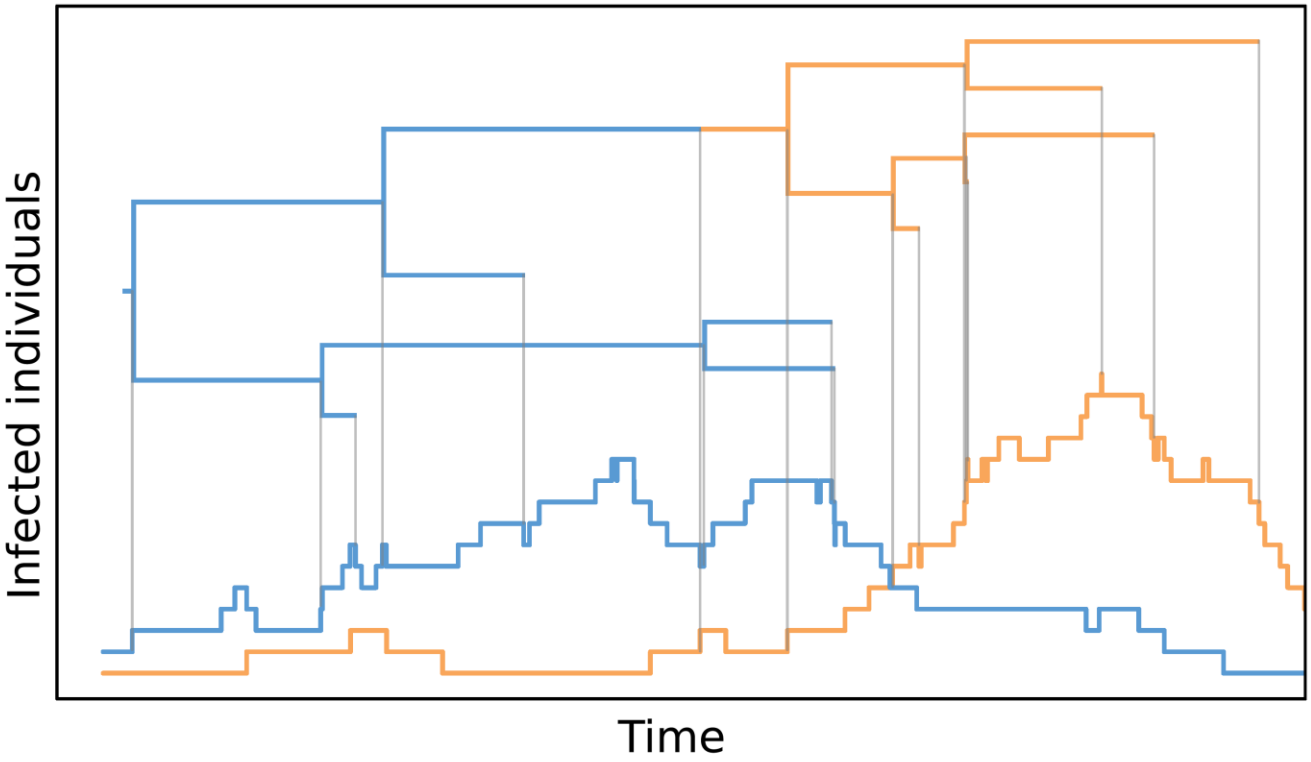
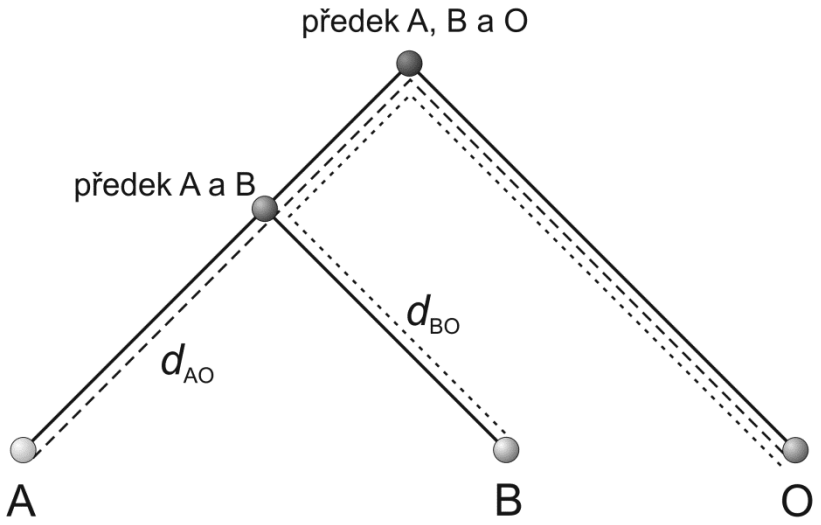


Figure 5.4 A two-deme phylodynamic time-tree with associated stochastic dynamics of infected compartments. (With thanks to Tim Vaughan for producing this figure).

Hypothesis testing

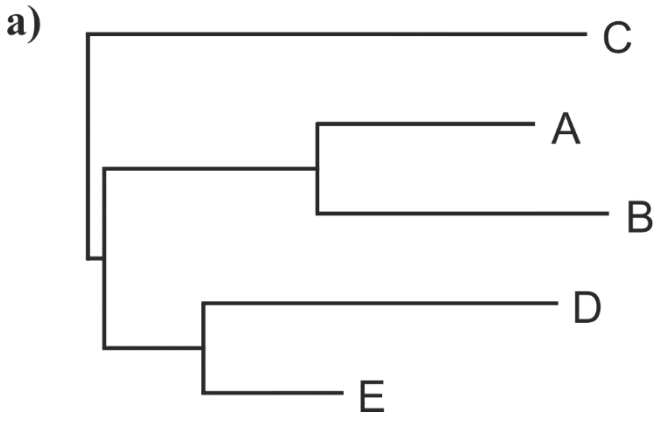
Test of molecular clock:

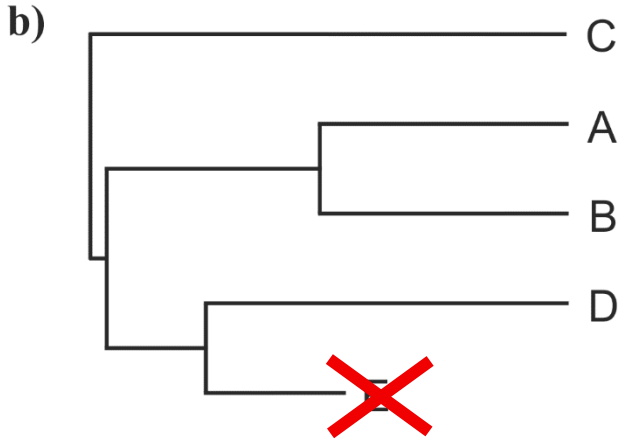
Relative rate test (RRT): AC=BC?



Linearized trees

removing significantly different taxa



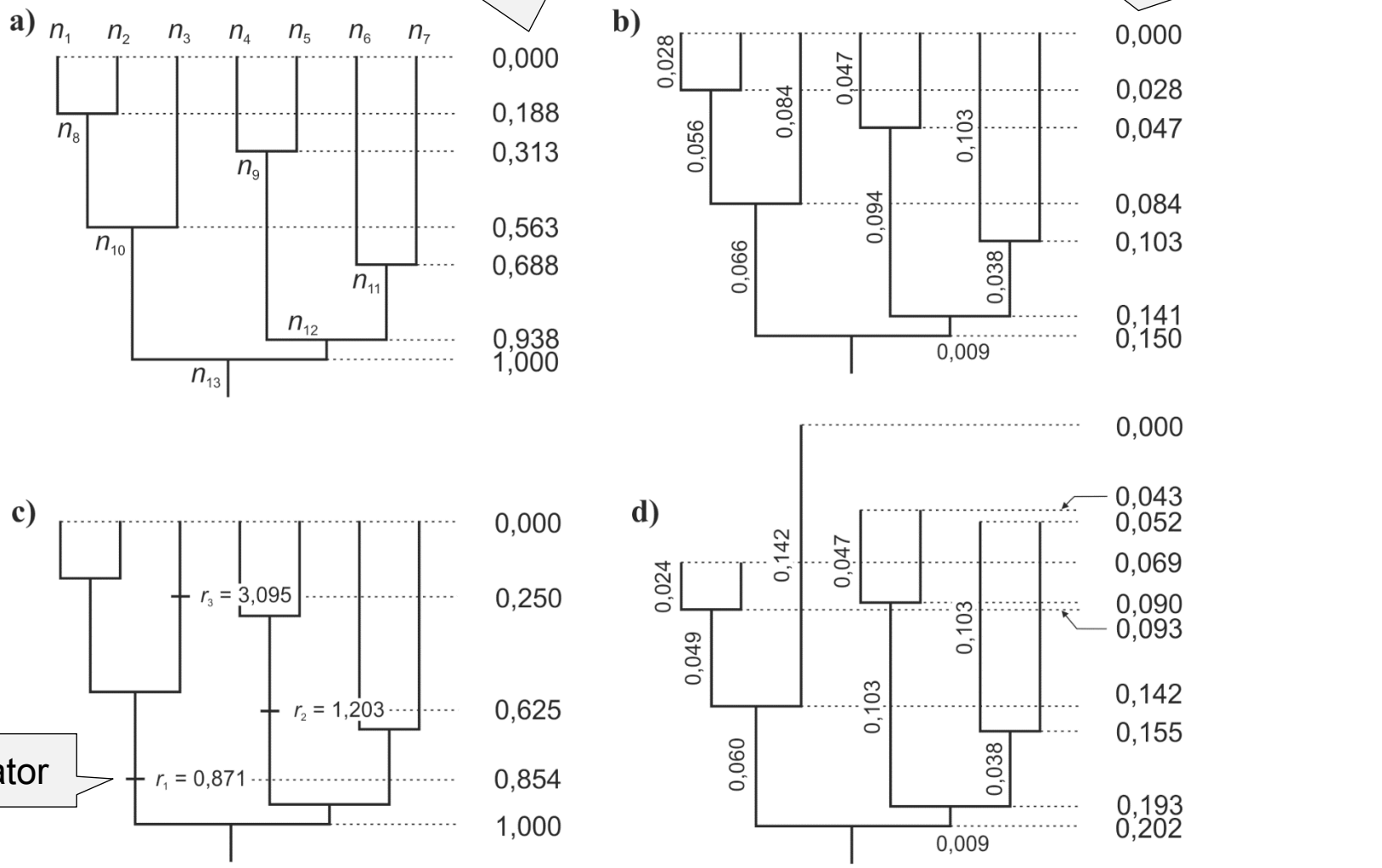


Relaxed molecular clock

enable changing rates along branches

unscaled time

scaled time (expected no. substitutions/site)

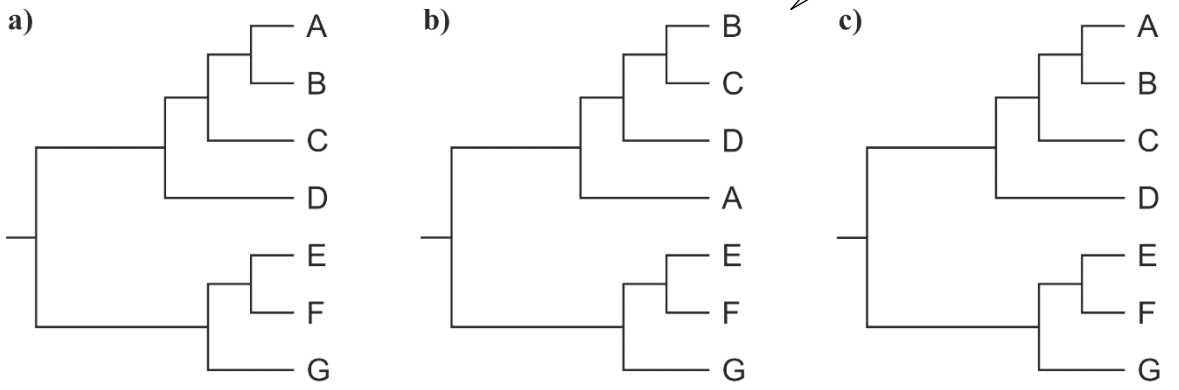


multiplicator

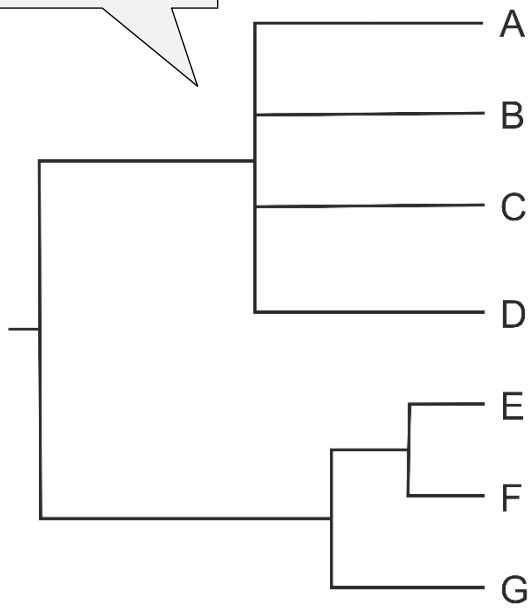
Consensus trees

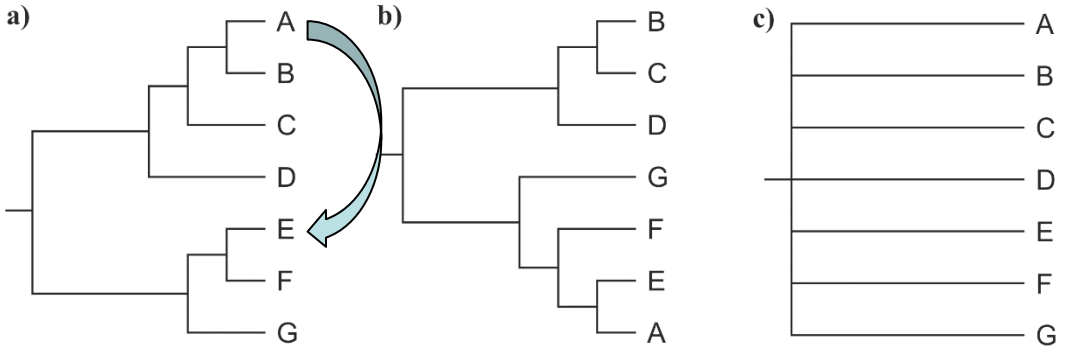
strict consensus

source trees

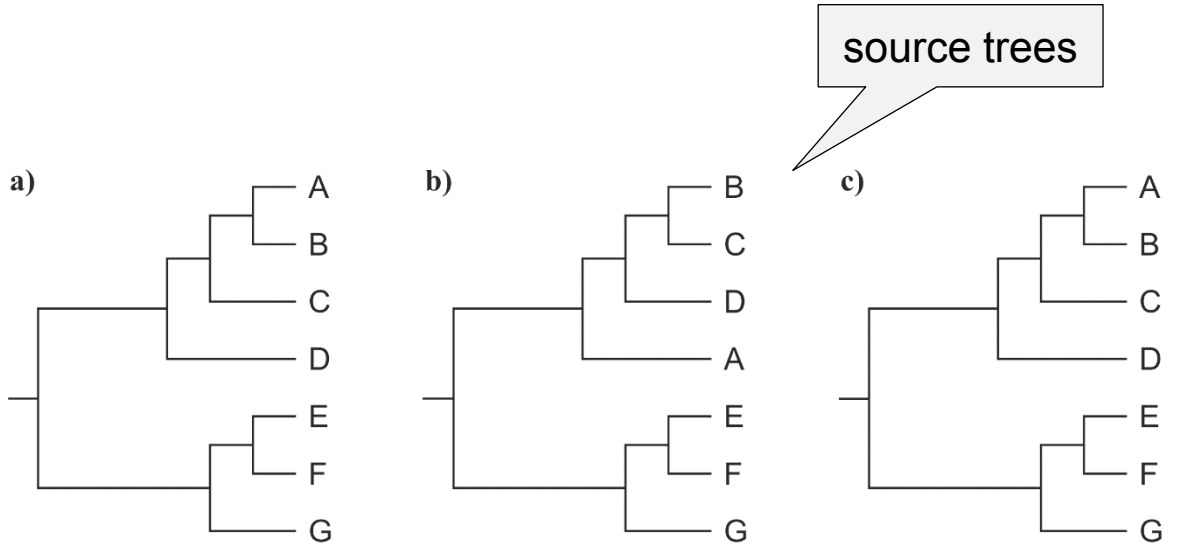


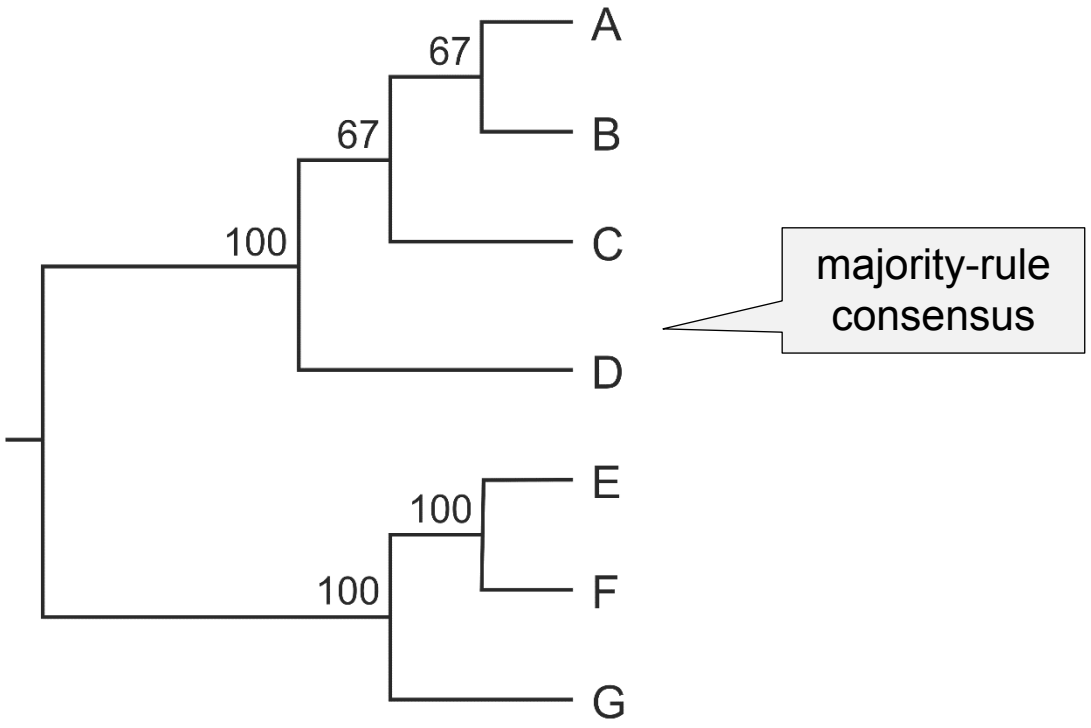
strict consensus tree





majority-rule





Consensus trees

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

consensus trees in resampling methods

Bayesian analysis:

consensus tree

maximum *a posteriori* tree = tree with greatest posterior probability (i.e. was sampled most often in the MCMC)

maximum credibility tree = tree with the maximum *product* of the posterior clade probabilities (BEAST, TreeAnnotator)

Phylogenetic programs

phylogeny inference:

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

PAUP*
PHYLIP
MOLPHY, PHYML, MEGA ... ML
MrBayes, BEAST ... BA

managing trees:

TreeView

FigTree