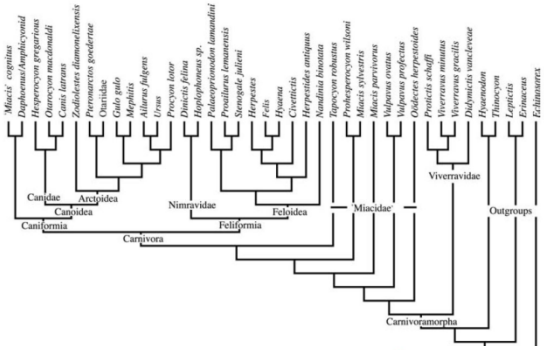
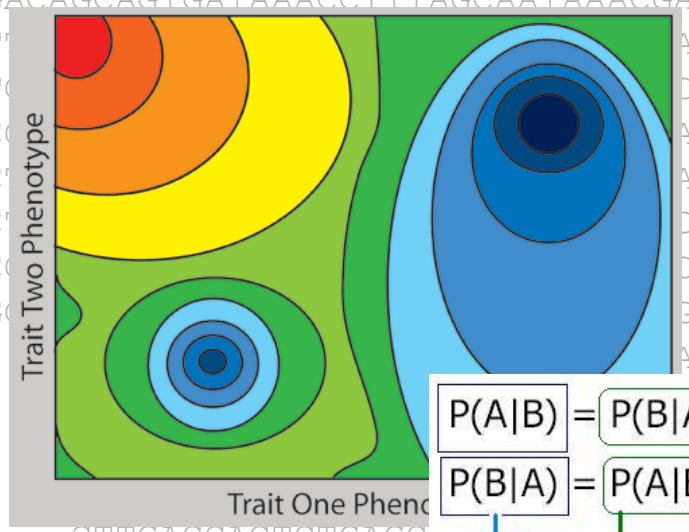
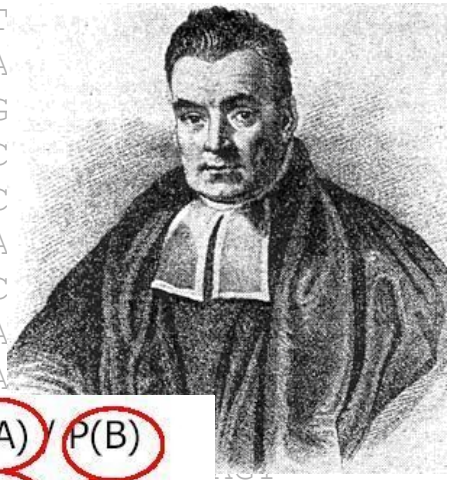
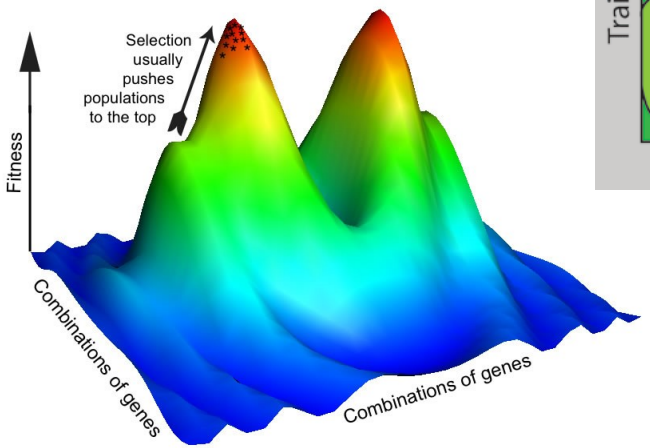
MODULARIZACE VÝUKY EVOLUČNÍ A EKOLOGICKÉ BIOLOGIE
CZ.1.07/2.2.00/15.0204

FYLOGENETICKÁ ANALÝZA 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

15 hodů mincí:

→ skóre OOHHHOHOOOHOHHO
tj. 7× panna (hlava, H), 8× orel (O)

Věrohodnost = podmíněná pravděpodobnost dat (výsledného skóre) při dané hypotéze:

$$L = \Pr(D \mid H) = \Pr(7\times \text{ hlava}, 8\times \text{ orel} \mid \text{hypotéza})$$

Pravděpodobnost, že padne hlava = p, orel = $(1 - p)$

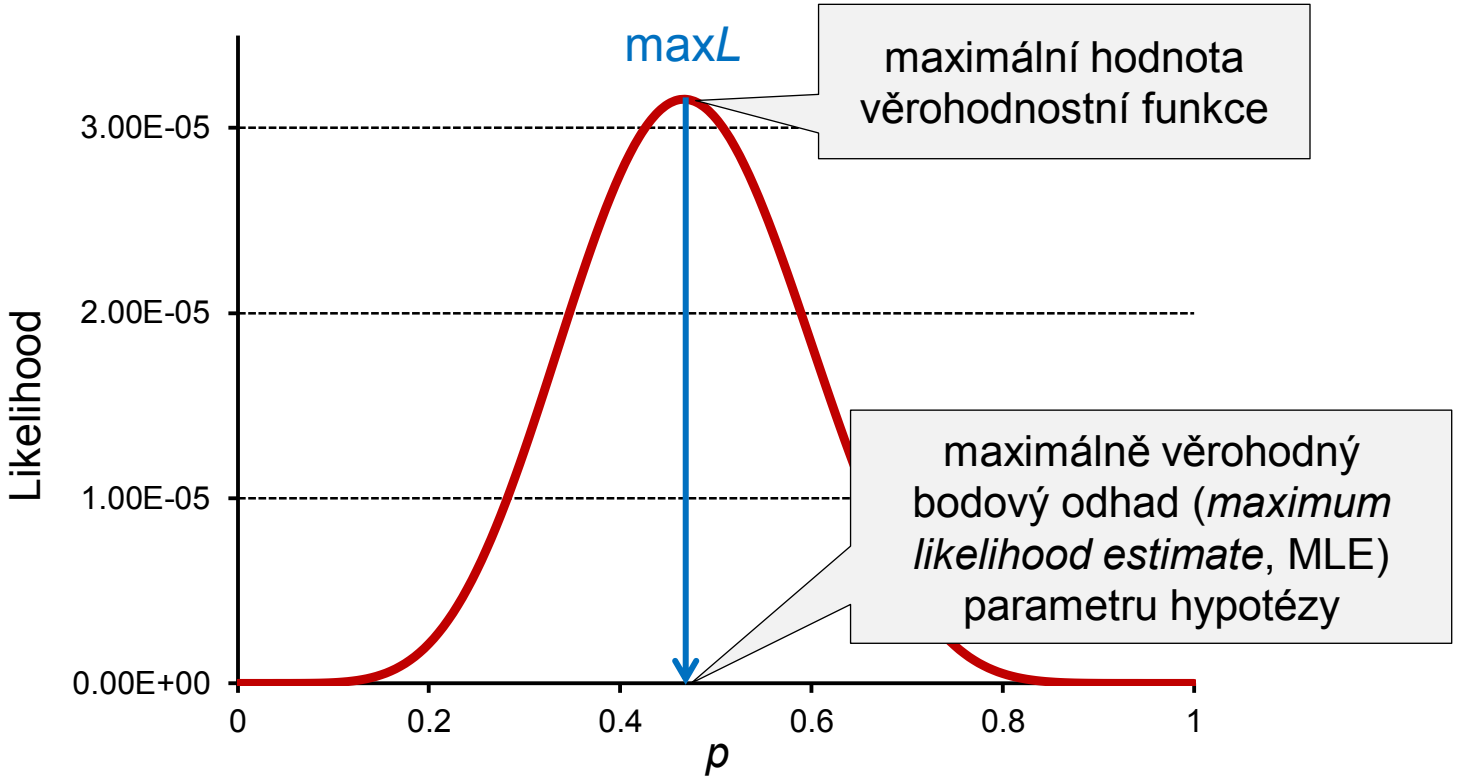
skóre OOHHHOHOOOHOHHO [7× panna (hlava, H), 8× orel (O)]

Protože hody nezávislé $\Rightarrow$ pravděpodobnost výsledného skóre =

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

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

maximum = 0,4666 $\approx$ 7/15



Hypotéza?

Např. H = mince není „cinknutá“, tj. $p = 1/2 \Rightarrow L = 3{,}0517.10^{-5}$

Je-li mince upravena tak, aby ve 2/3 případů padl orel:

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

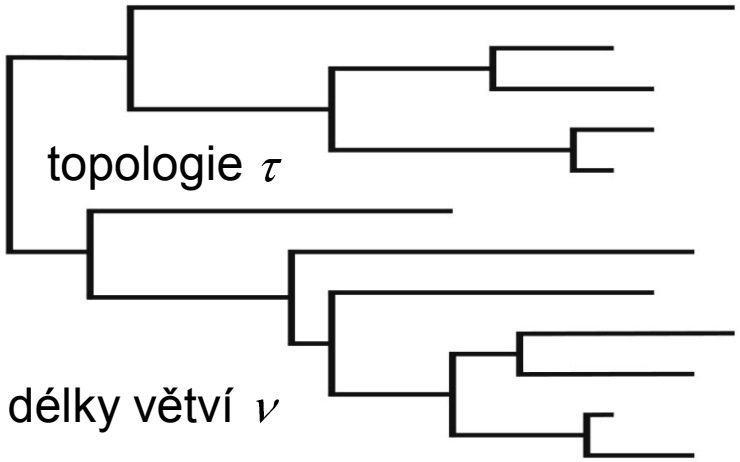
$\Rightarrow$výsledek hodů 1,7$\times$ pravděpodobnější s pravou mincí

Maximální věrohodnost ve fylogenetické analýze

data:

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

strom:



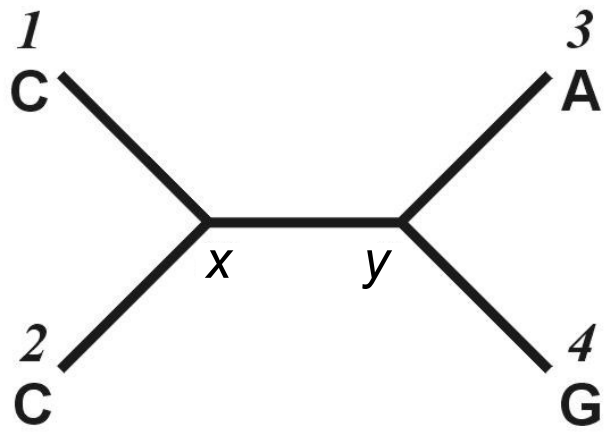
+ evoluční model θ

= hypotéza

nuisance parameters

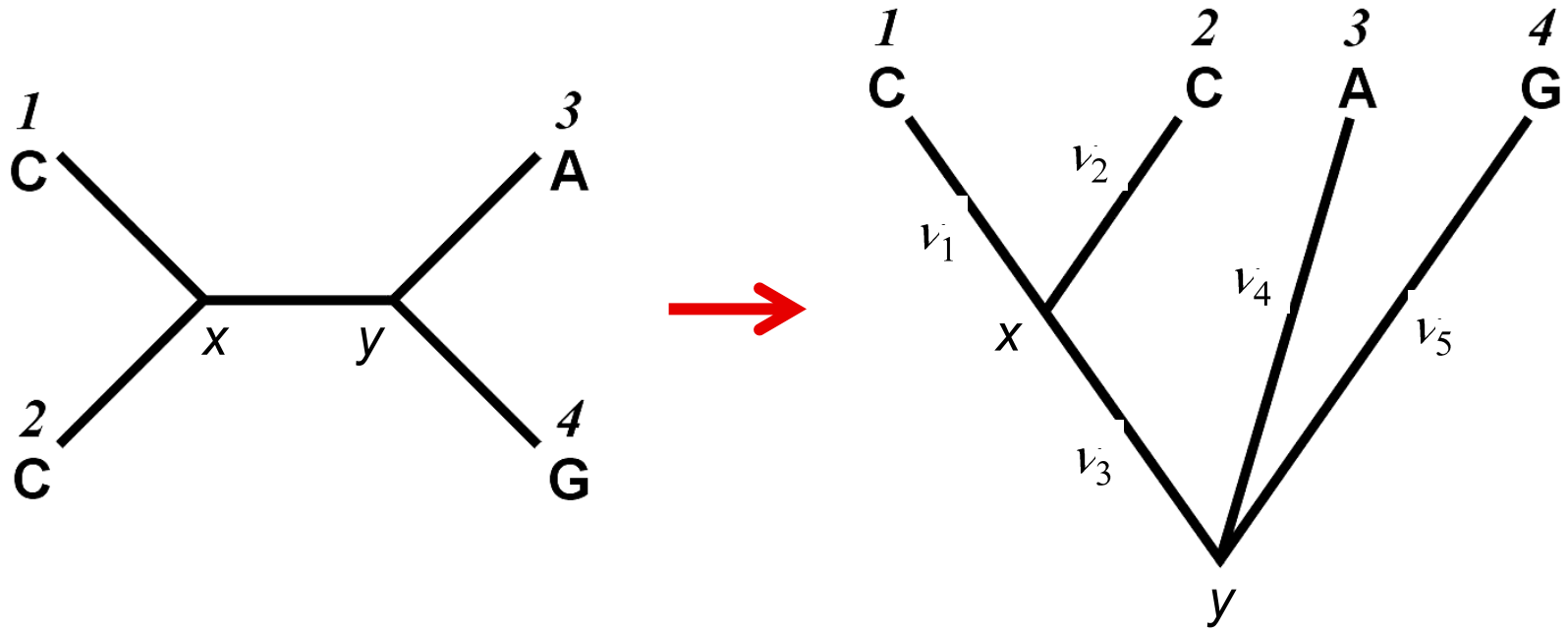
$L = P(D \mid H)$: D = matice sekvencí (dat), H = τ (topologie) + ν (délky větví) + θ (model)

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



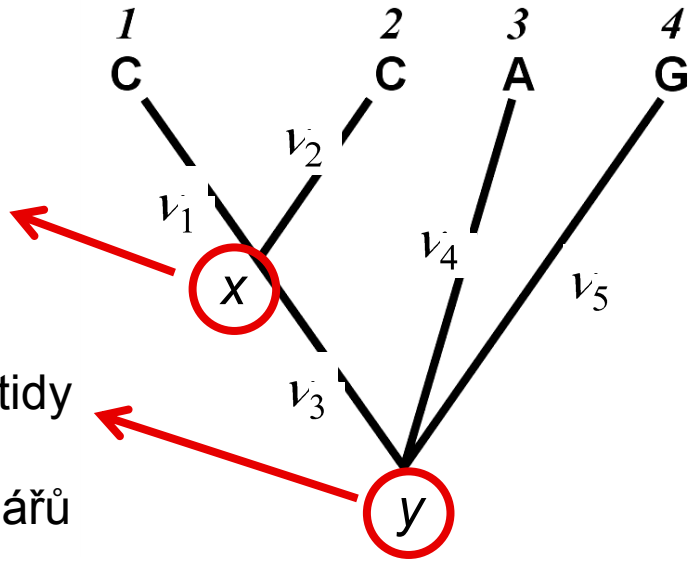
```
      1                          j                          N
1     TCAAAAATGGCTTTATTCGCTTAATGCCGTTAACCCTTGCGGGGGCCATG
2     TCCGTGATGGATTTATTTCCGCAATGCCTGTCATCTTATTCTCAAGTATC
3     TTCGTGATGGATTTATTGCAGGTATGCCAGTCATCCTTTTCTCATCTATC
4     TTCGTGACGGGTTTATCTCGGCAATGCCGGTCATCCTATTTTCGAGTATT
```

ν_i = délky větví



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

ν_i = délky větví



x: 4 nukleotidy

y: 4 nukleotidy

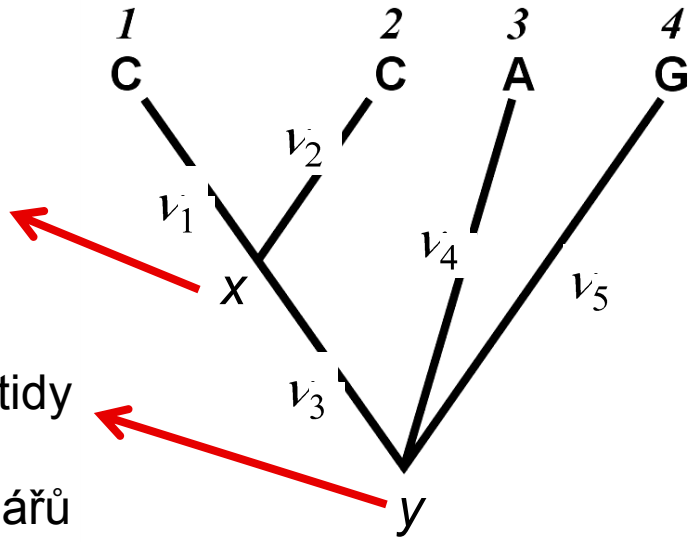
$\Rightarrow 4 \times 4 = 16$ možných scénářů

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

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

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

ν_i = délky větví



x: 4 nukleotidy

y: 4 nukleotidy

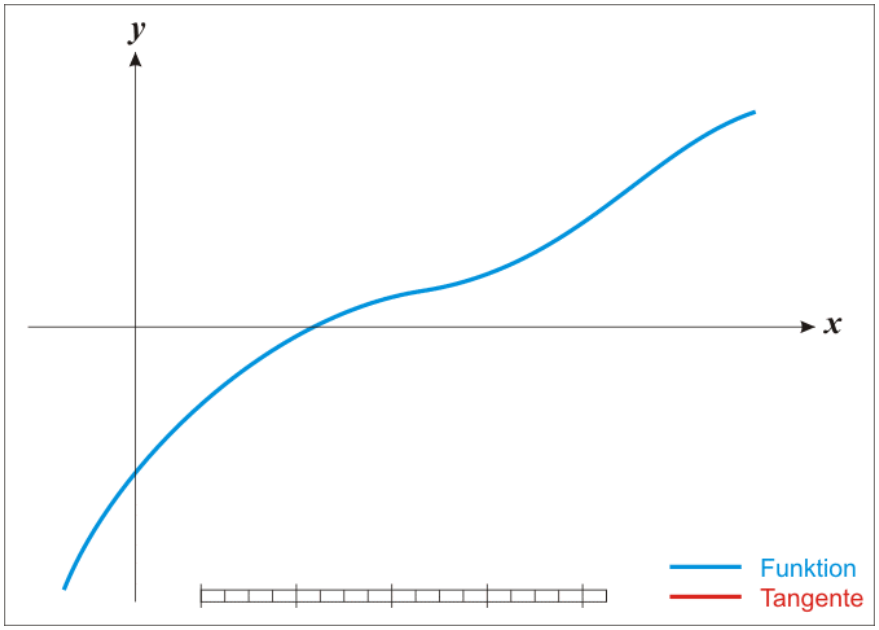
$\Rightarrow 4 \times 4 = 16$ možných scénářů

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

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

Hledání maximální věrohodnosti daného stromu

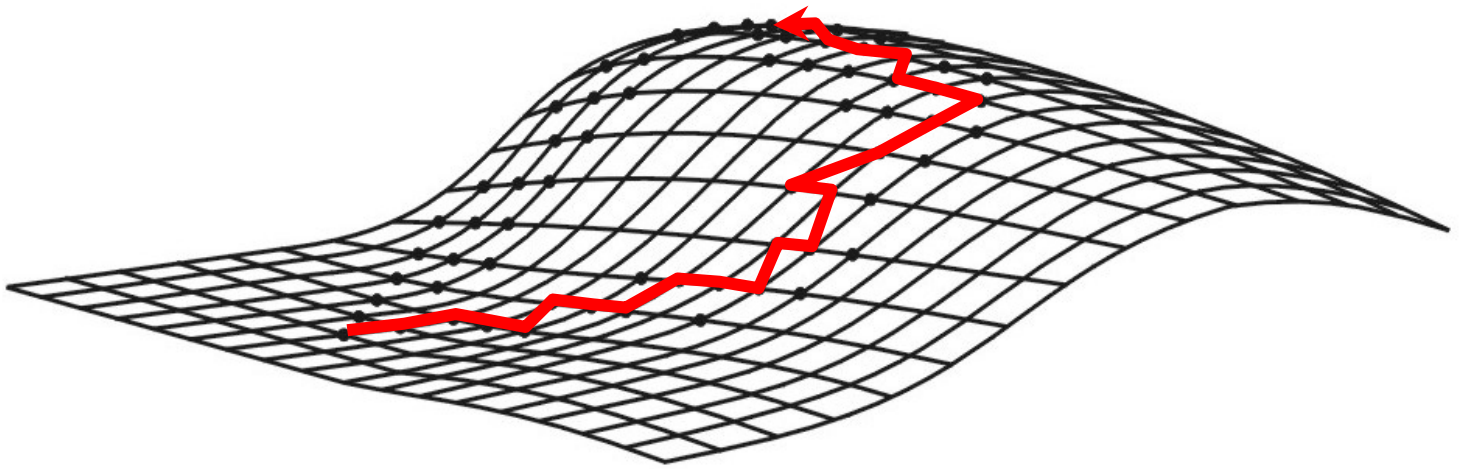
→ např. Newtonova (Newtonova-Raphsonova) metoda



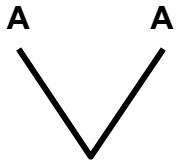
https://upload.wikimedia.org/wikipedia/commons/e/e0/NewtonIteration_Ani.gif

Hledání nejvěrohodnějšího stromu: heuristické hledání

Heuristické hledání



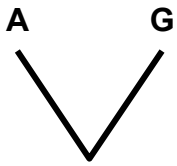
stepwise addition ... např. PHYLIP
star decomposition ... např. MOLPHY; *neighbor-joining* tree
branch swapping



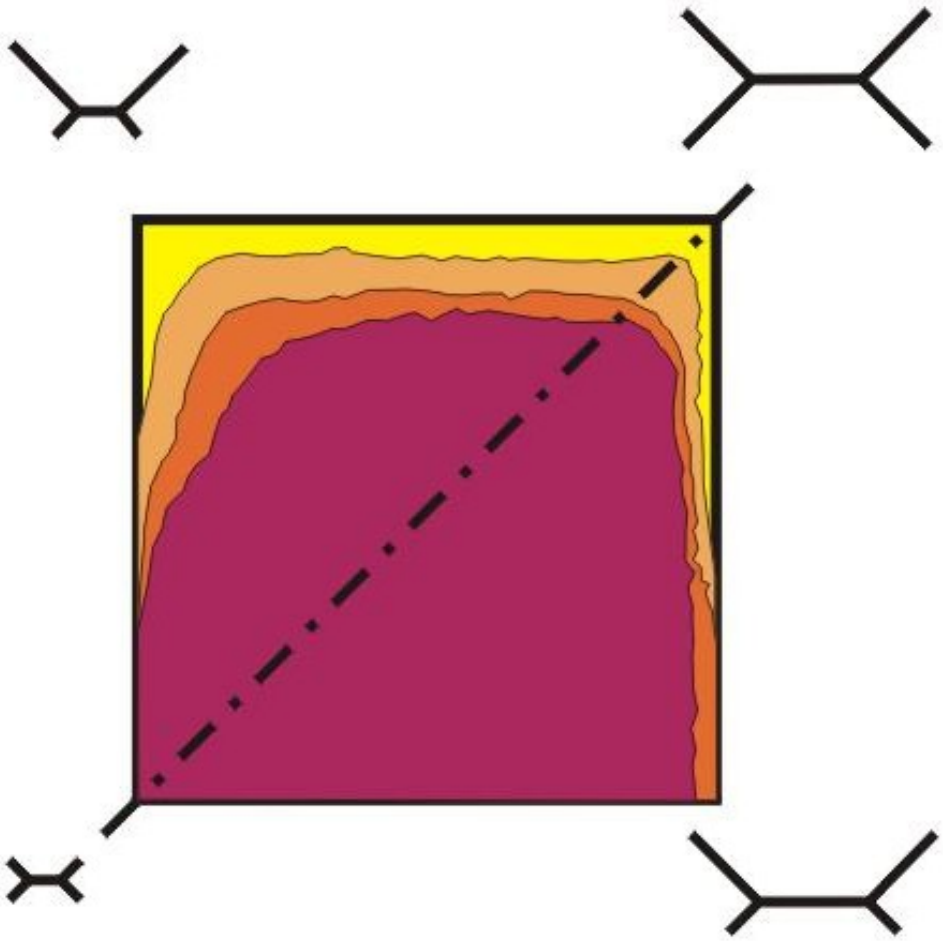
Věrohodnost (ML) a úspornost (MP)

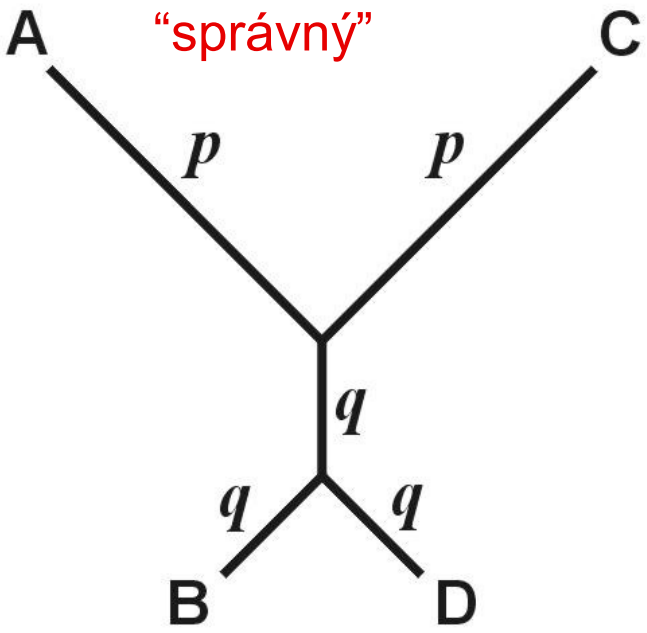
Počet změn	Parsimonie	$\nu = 0{,}01$ (0,2475)	$\nu = 0{,}10$ (0,2266)	$\nu = 0{,}20$ (0,20611)	$\nu = 1{,}00$ (0,11192)
0	100	99,99	99,83	99,31	82,17
1	0	0,00	0,00	0,00	0,00
2	0	0,0011	0,11	0,44	9,13
3	0			0,034	3,55
4	0				0,0027

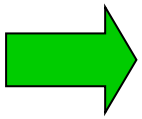
Počet změn	Parsimonie	$\nu = 0{,}01$ (0,00083)	$\nu = 0{,}10$ (0,00786)	$\nu = 0{,}20$ (0,01462)	$\nu = 1{,}00$ (0,04602)
0	0	0,00	0,00	0,00	0,00
1	100	99,66	96,64	92,36	66,54
2	0	0,33	3,22	6,22	21,19
3	0		0,12	0,48	8,61
4	0		0,003	0,023	2,05
5	0			0,0037	0,42

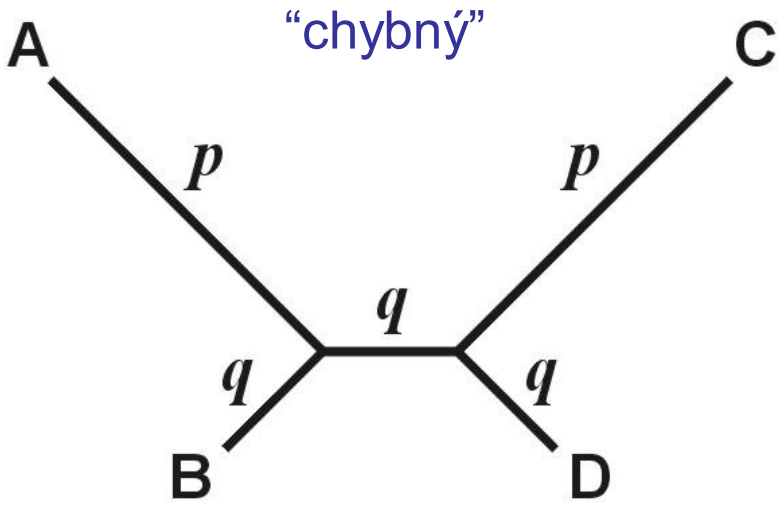


Věrohodnost a konzistence









Farrisova (anti-Felsensteinova, inverzní Felsensteinova) zóna

“long-branch repulsion”

BAYESOVSKÁ ANALÝZA

ML: Jaká je pravděpodobnost dat při dané hypotéze?

bayesiánský přístup:

Jaká je pravděpodobnost hypotézy při daných datech?

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

Př.: soubor 100 kostek, ze kterých máme vybrat jednu

víme, že ze 100 kostek je 80 v pořádku, ale 20 je upraveno tak, aby padala 6

2 hody: 1. hod =  2. hod = 

Jaká je pravděpodobnost, že naše kostka je falešná?

pravděpodobnosti jednotlivých výsledků:

u pravých kostek stejné, u falešných se liší:

	pravá	falešná
	1/6	1/21
	1/6	3/21
	1/6	3/21
	1/6	4/21
	1/6	4/21
	1/6	6/21

Pravděpodobnost $P(H \mid D)$ se nazývá aposteriorní (*posterior probability*)

aposteriorní pravděpodobnost je funkcí věrohodnosti $L = P(D \mid H)$

a apriorní pravděpodobnosti (*prior probability*), která vyjadřuje náš apriorní předpoklad nebo znalost

Aposteriorní pravděpodobnost, že naše kostka je falešná, je dána Bayesovou rovnicí:

věrohodnost

apriorní pravděpodobnost

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

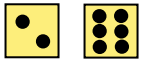
suma čitatelů pro všechny alternativní hypotézy

Thomas Bayes

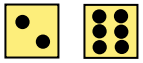
Pro náš příklad se 2 hody kostkou:

	pravá	falešná
	1/6	1/21
	1/6	3/21
	1/6	3/21
	1/6	4/21
	1/6	4/21
	1/6	6/21

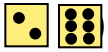
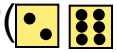
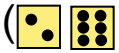
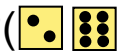
apriorní pravděpodobnost (falešná) = 0,2
(20/100 falešných kostek v souboru)

Pr., že dostaneme  s pravou kostkou:

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

Pr., že dostaneme  s falešnou kostkou:

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

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

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

Bayesovská metoda ve fylogenetické analýze:

aposteriorní pravděpodobnost

marginální věrohodnost

apriorní pravděpodobnost

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

suma přes všechny možné stromy

Parametry pro bayesovskou analýzu většinou kontinuální $\Rightarrow$

$P \rightarrow$ pravděpodobnostní hustotní funkce (*probability density functions*)

buď ML odhady $\rightarrow$ empirická BA

nebo všechny kombinace $\rightarrow$ hierarchická BA

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

Problém: výpočty příliš složité $\Rightarrow$ nelze řešit analyticky, pouze numericky aproximovat

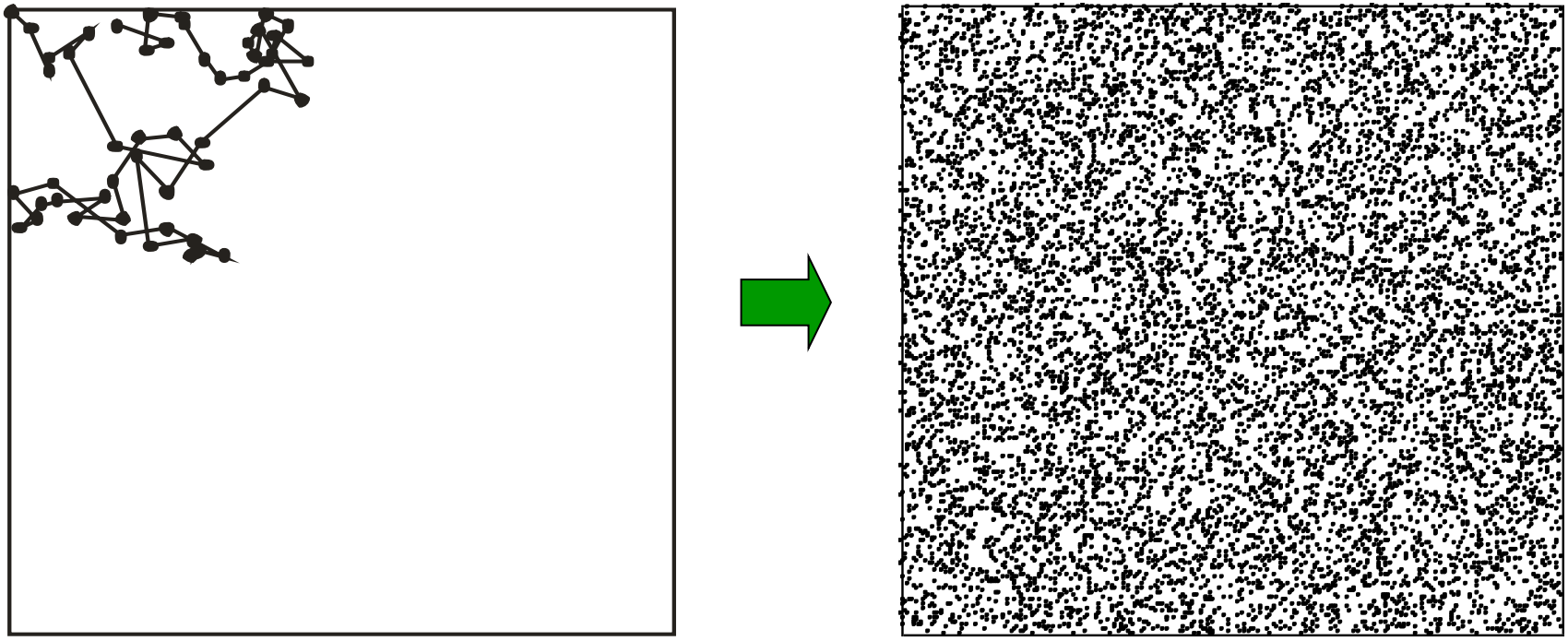
řešení: metody Monte Carlo

náhodný výběr vzorků, při velkém množství aproximace skutečnosti

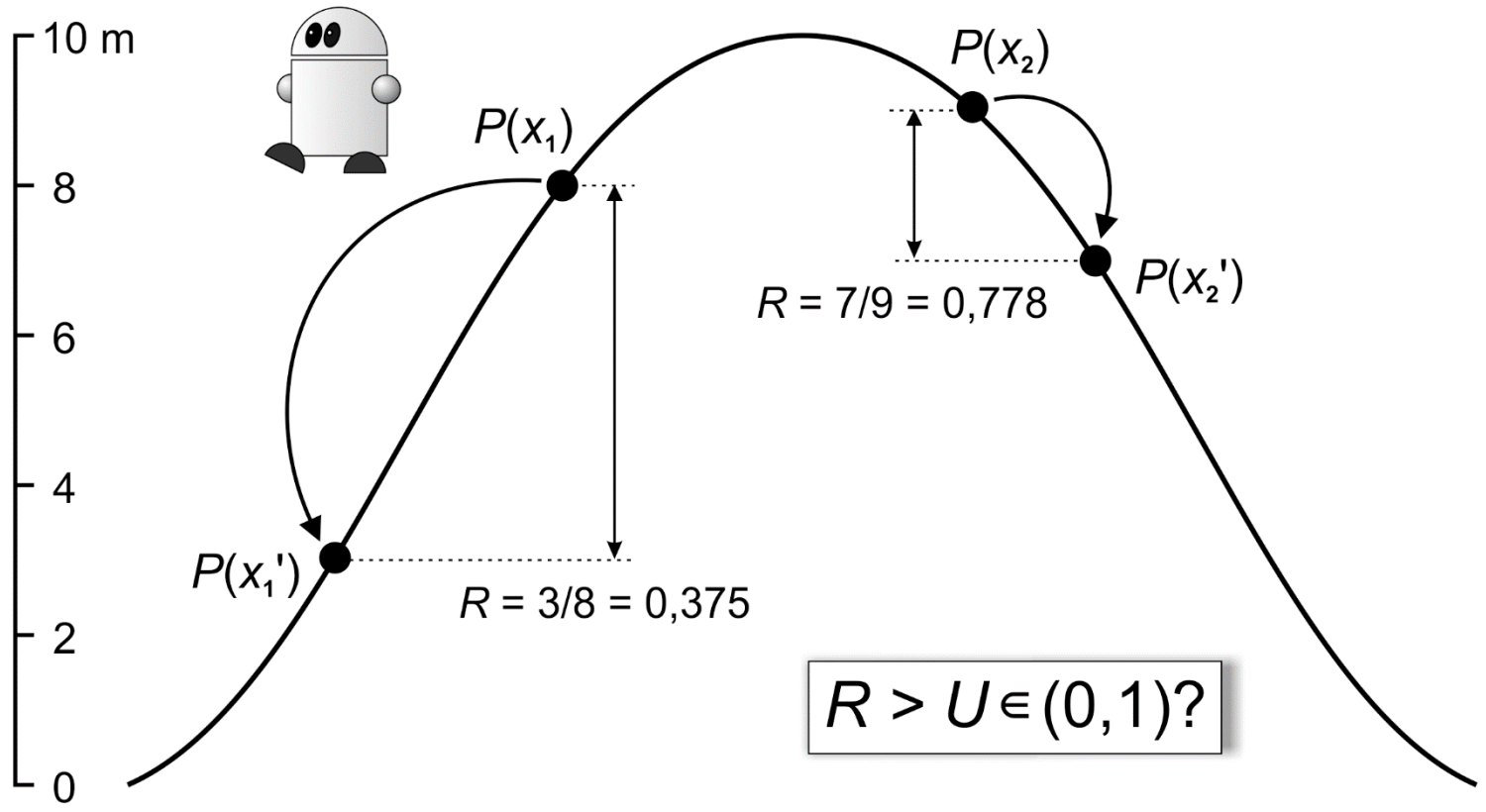
Markovovy řetězce: Markov chain Monte Carlo (MCMC)

Markovův proces: t_{-1}: A $\rightarrow$ t_0: C $\rightarrow$ t_{+1}: G

… *P* stejná po celé fylogenii = homogenní Markovův proces



Metropolisův-Hastingsův algoritmus:

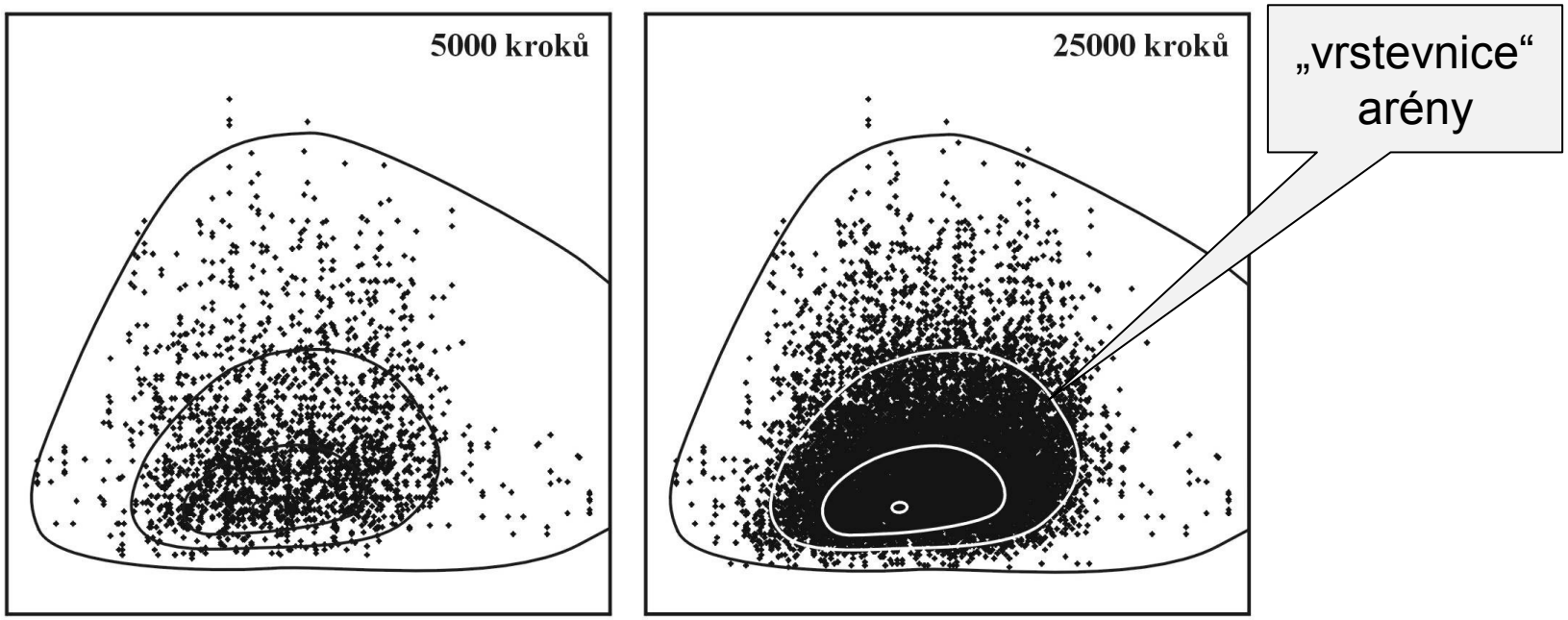


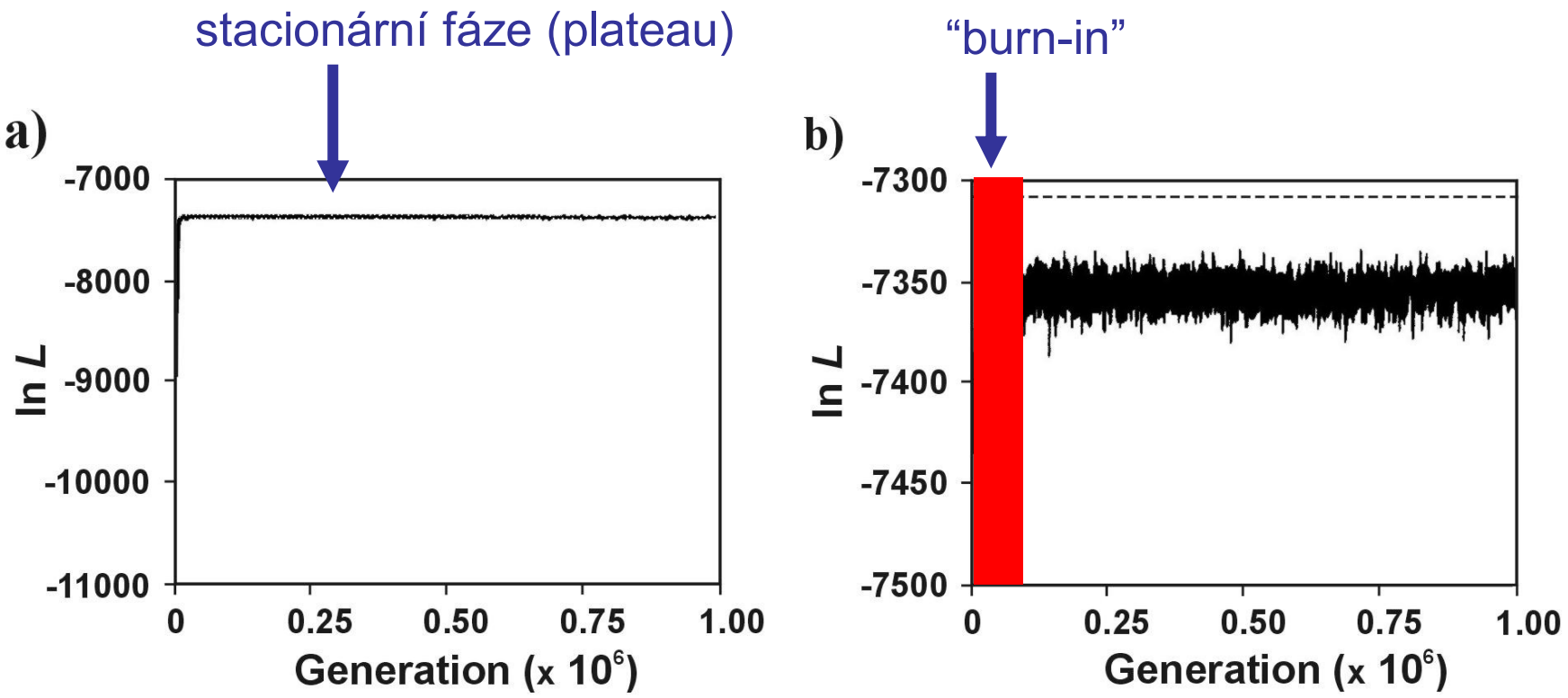
Metropolisův-Hastingsův algoritmus:

Změna parametru $x \rightarrow x'$

1. jestliže $P(x') > P(x)$, akceptuj x'
2. jestliže $P(x') \leq P(x)$, vypočti $R = P(x')/P(x)$
   protože platí, že $P(x') \leq P(x)$, musí být $R \leq 1$
3. generuj náhodné číslo U z rovnoměrného rozělení z intervalu (0, 1)
4. jestliže $R \geq U$, akceptuj x', jestli ne, ponechej x

usměrněný pohyb robota v aréně:




stacionární fáze (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)

Reverzibilní přeskokový MCMC (*reversible jump MCMC*):

umožňuje měnit počet parametrů při každém MC kroku

lze použít např. k modelování proměnlivosti evoluce mezi pozicemi v sekvenci, k výběru modelů nebo k tvorbě nehomogenních substitučních modelů (např. různé složení bází podél jednotlivých větví)

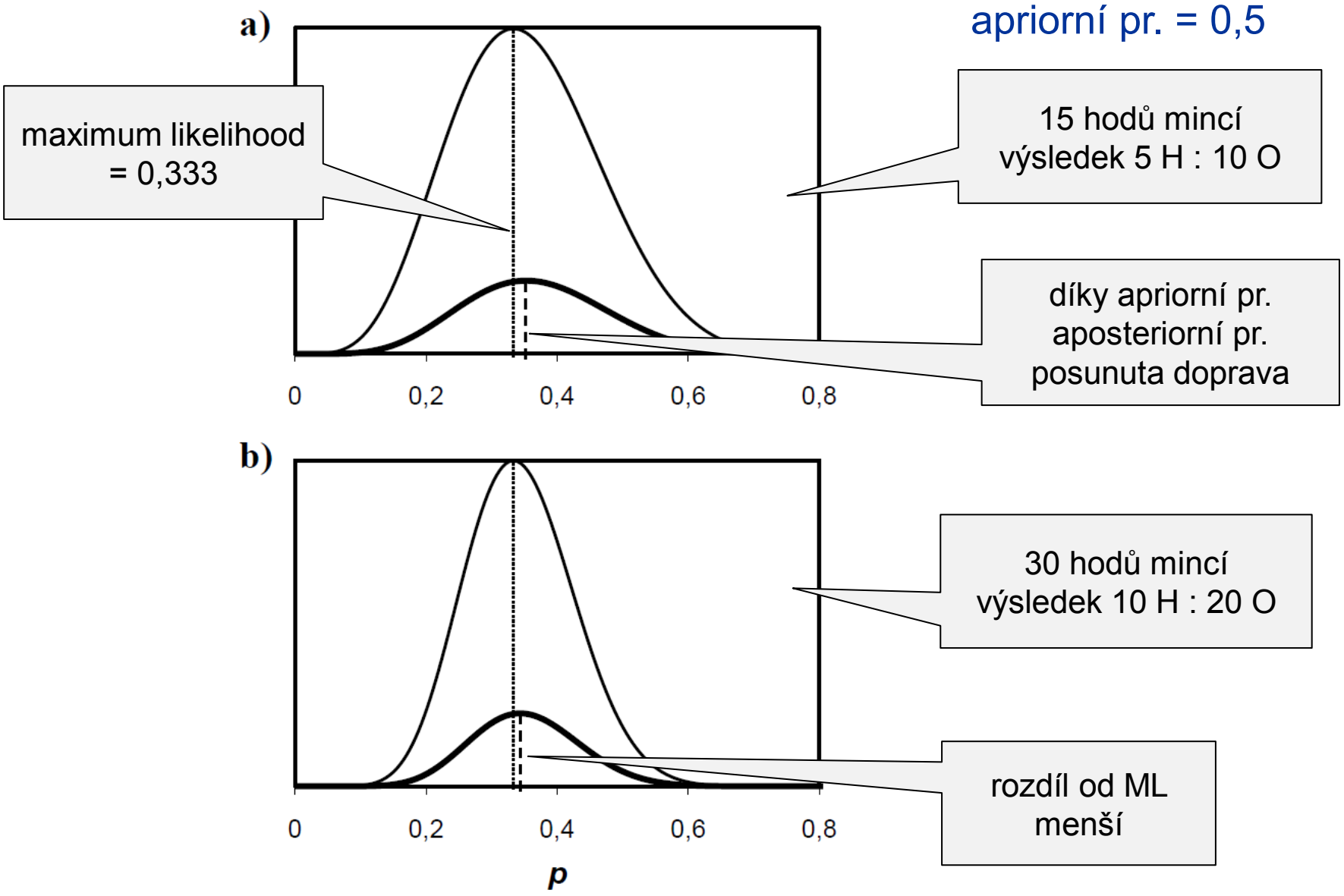
Metropolis coupled MCMC (MCMCMC, MC^3):

1 „chladný“ řetězec, 3 „zahřáté“ řetězce

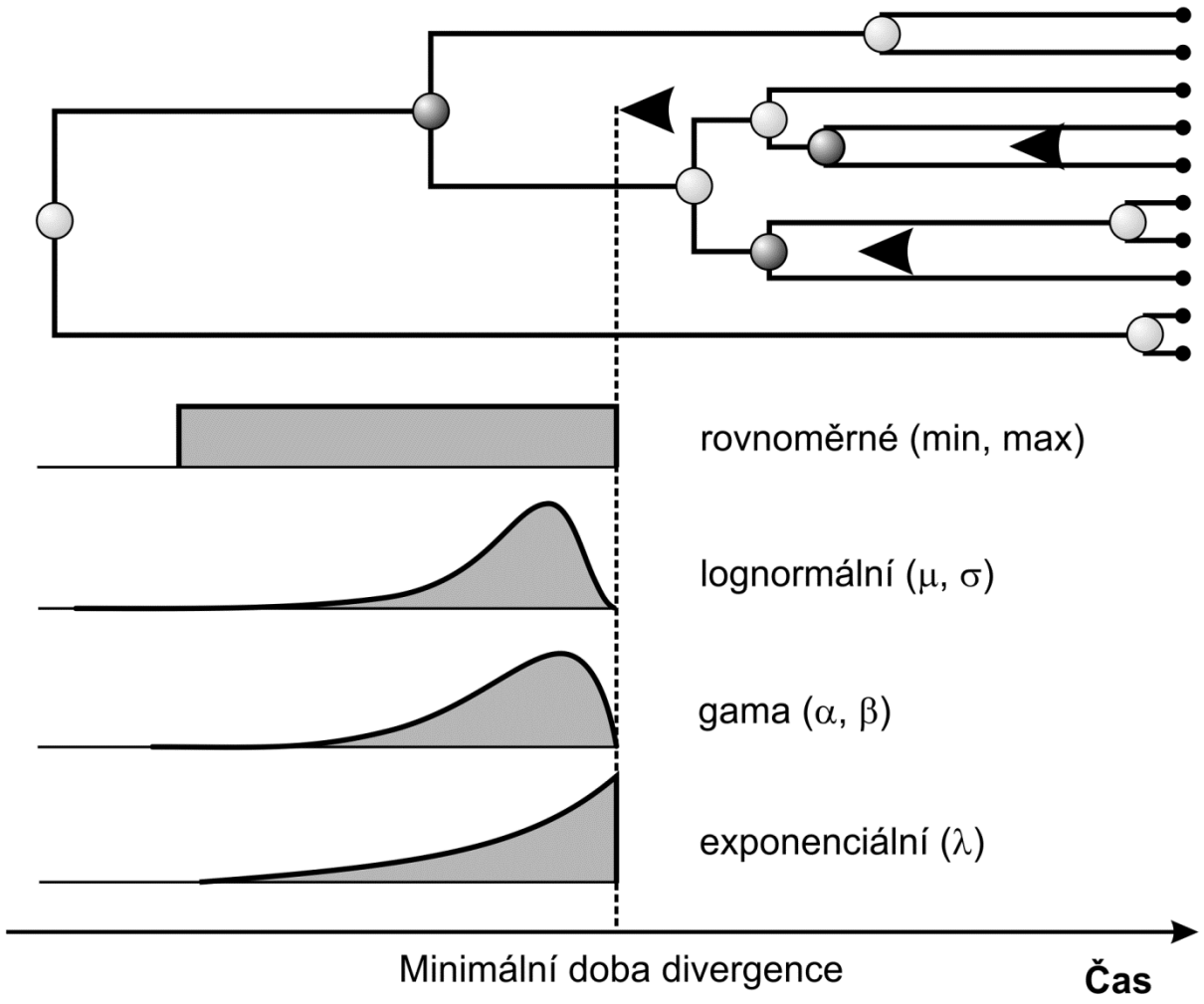
stejný výchozí bod, díky stochasticitě rychlá divergence „robotů“

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

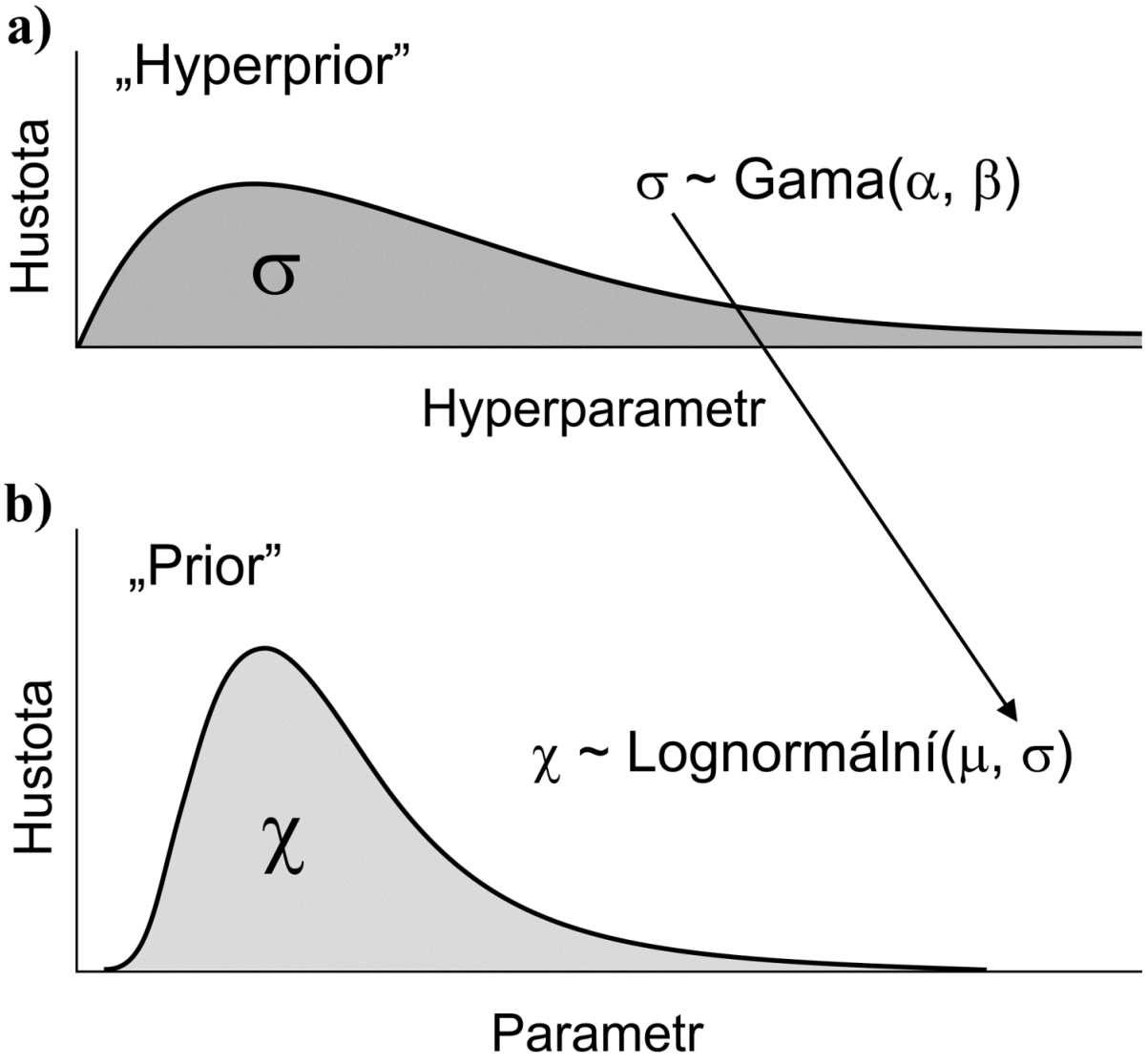
Problém apriorních pravděpodobností



Problém apriorních pravděpodobností



Stanovení apriorních pravděpodobností:

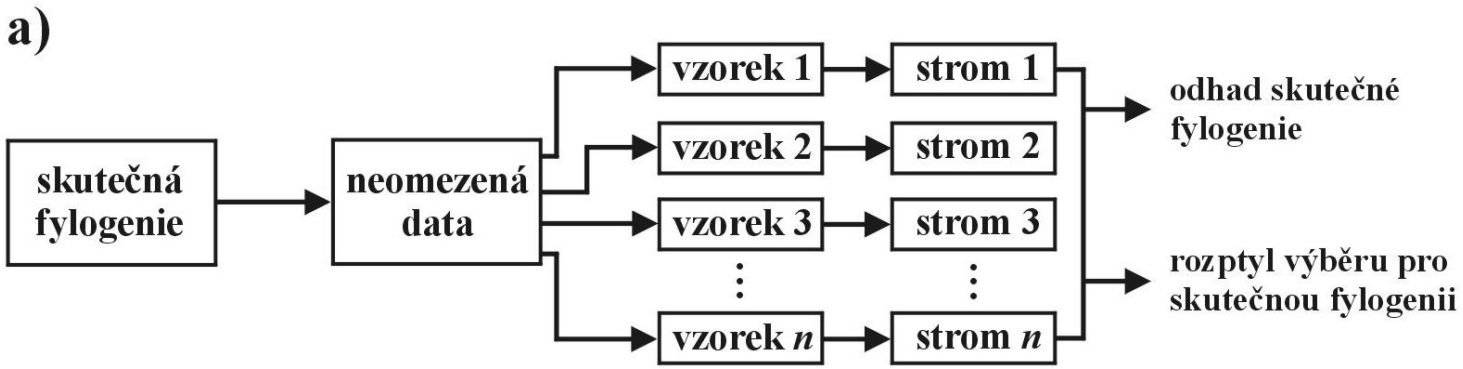


Měření spolehlivosti stromů

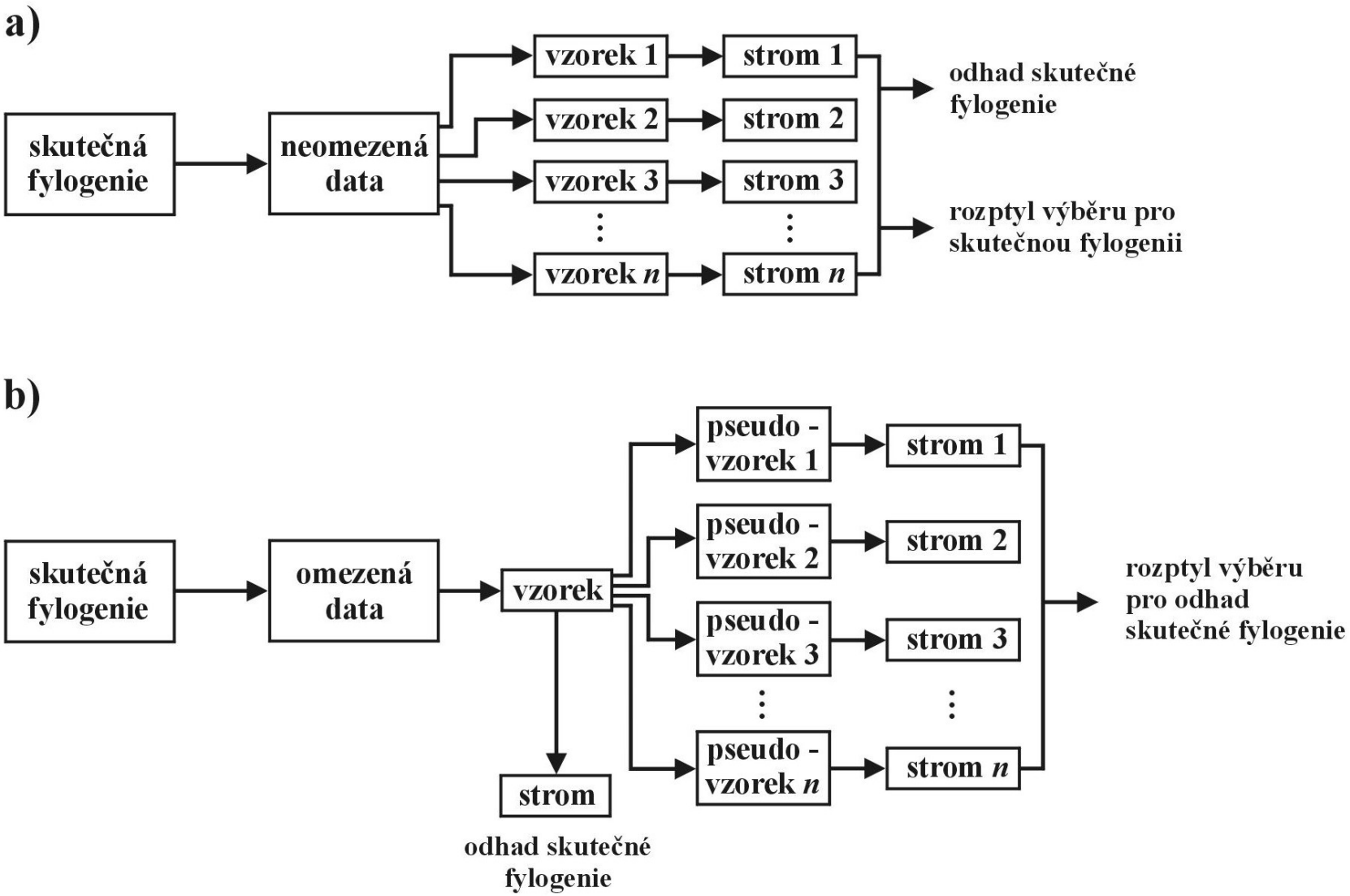
Metody opakovaného výběru

bez navrácení = jackknife
s navrácením = bootstrap

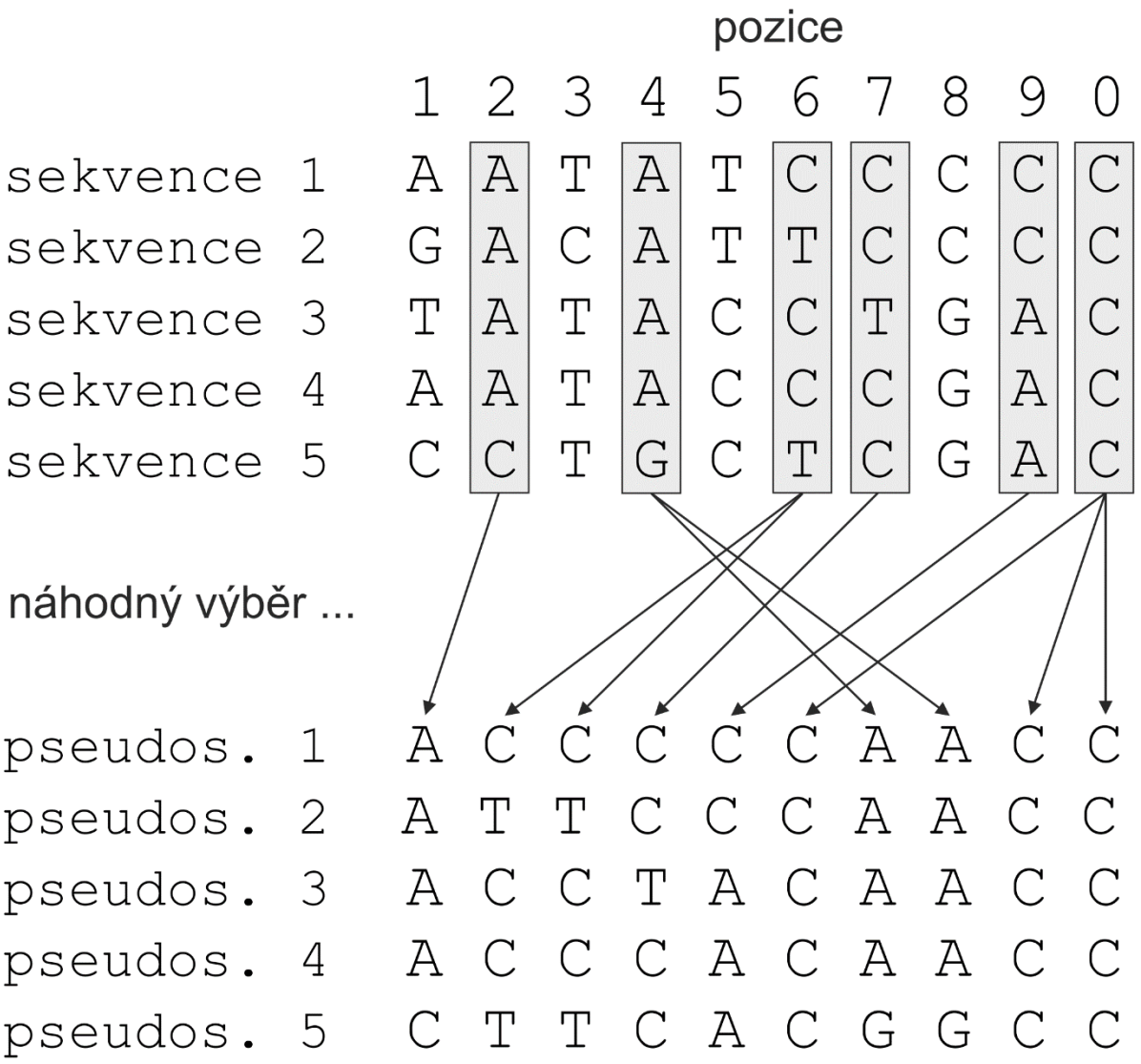
bootstrap:


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

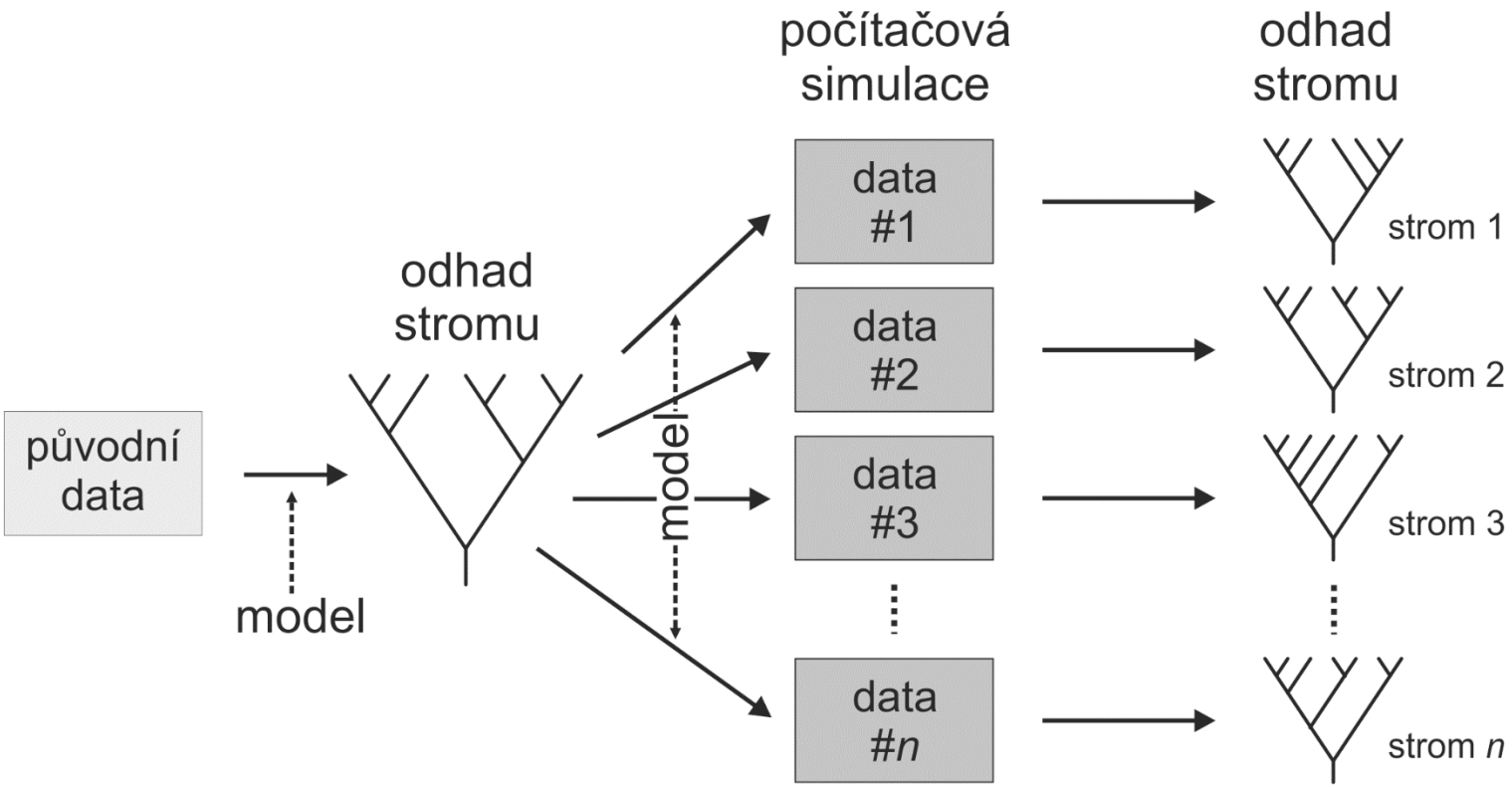
bootstrap:


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

parametrický bootstrap: evoluční model

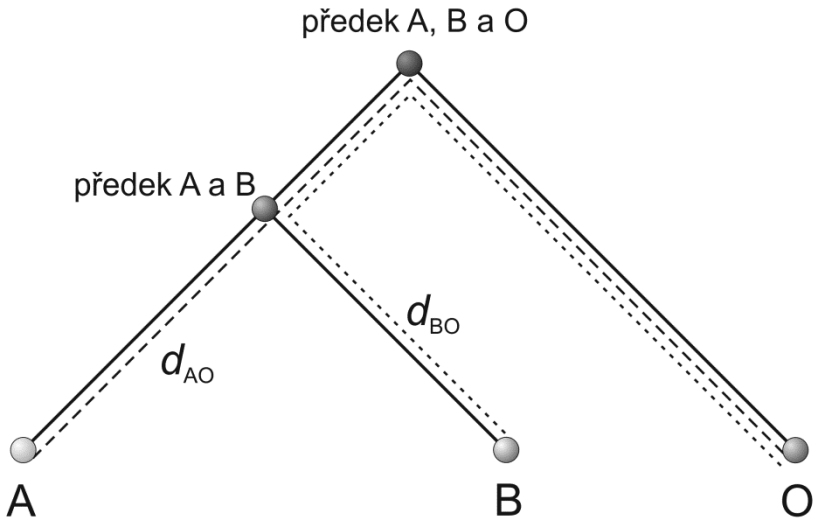


bayesovská analýza: aposteriorní pravděpodobnosti

Testování hypotéz

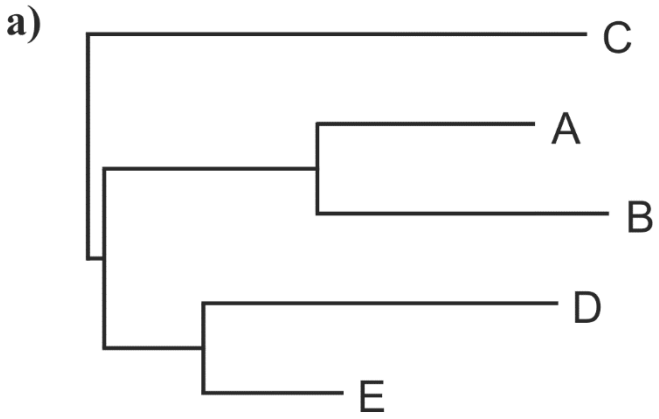
Test molekulárních hodin:

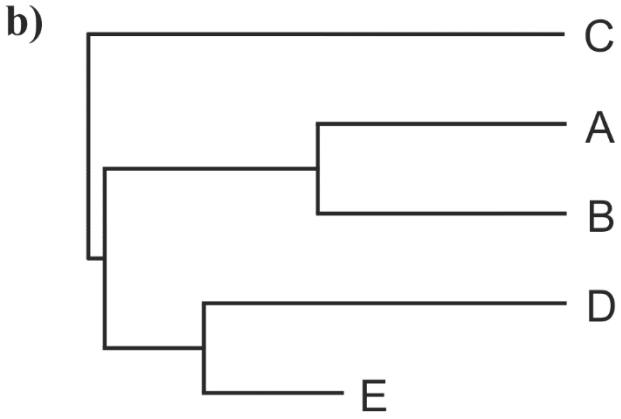
Relative rate test (RRT): AC=BC?



Linearizované stromy

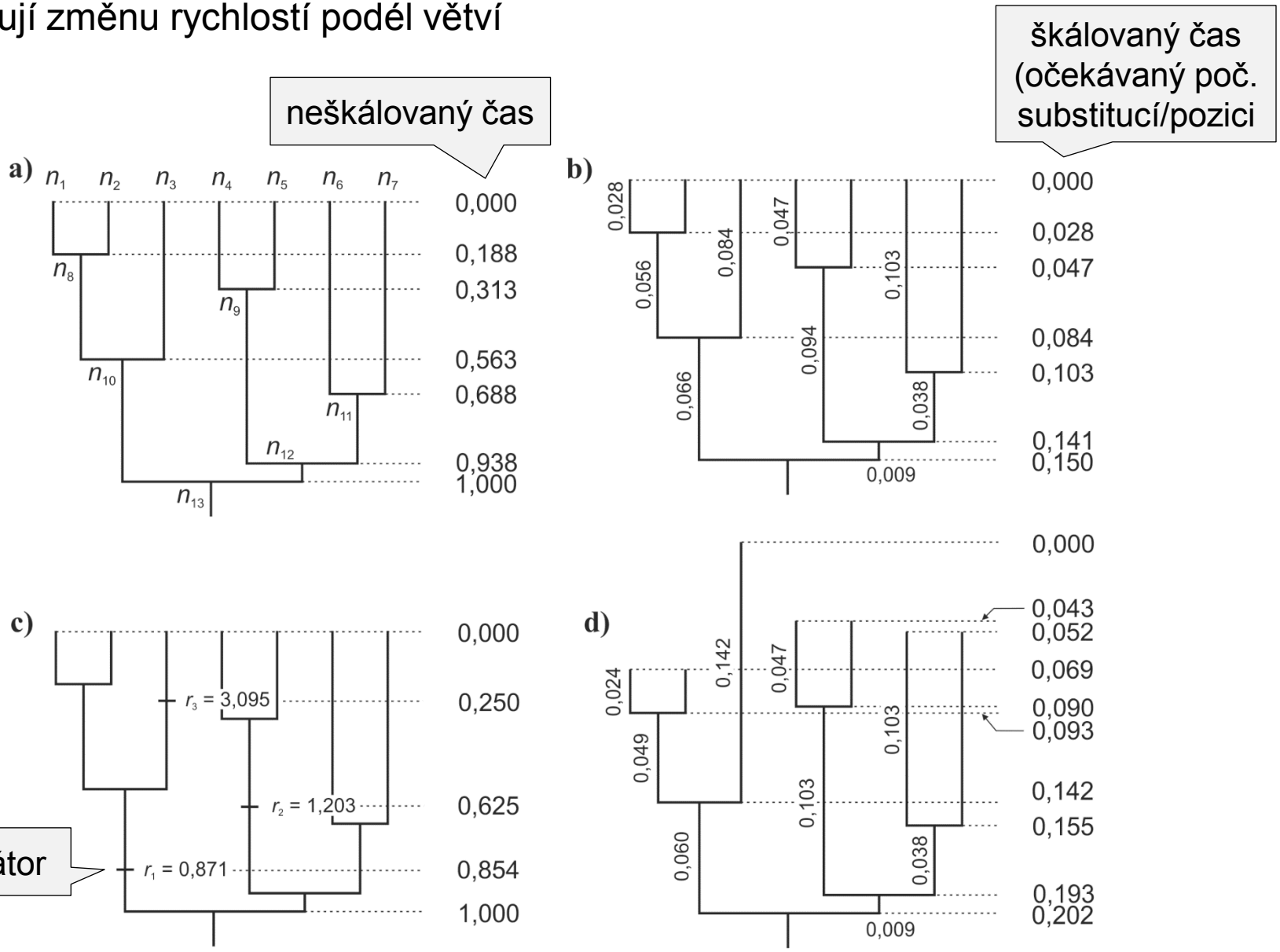
odstranění signifikantně odlišných taxonů





Relaxované molekulární hodiny

umožňují změnu rychlostí podél větví



Srovnání stromů

Jsou dva stromy signifikantně odlišné?

Testy párových pozic:

winning sites test

Felsensteinův *z* test

Templetonův test

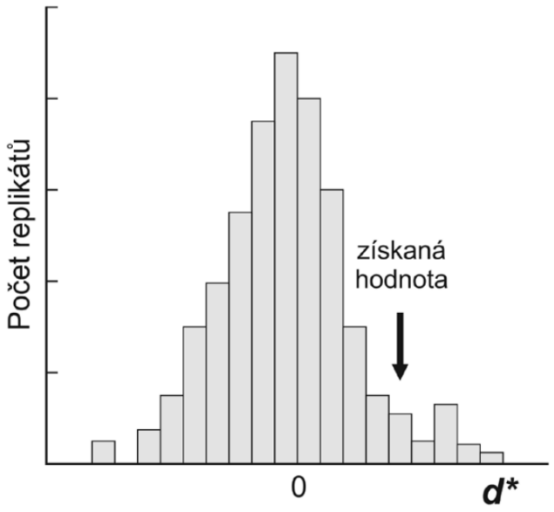
Kishinův-Hasegawův test (KHT, RELL)

a)

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

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

$d^*_n = \ln L^*_{T1} - \ln L^*_{T2}$



Pro více než dva stromy:

Shimodairův-Hasegawův (SH) test

Srovnání stromů

Do jaké míry jsou dva stromy odlišné?

Distance mezi stromy:

partition metric

quartet metric

path difference metric

metody inkorporující délky větví

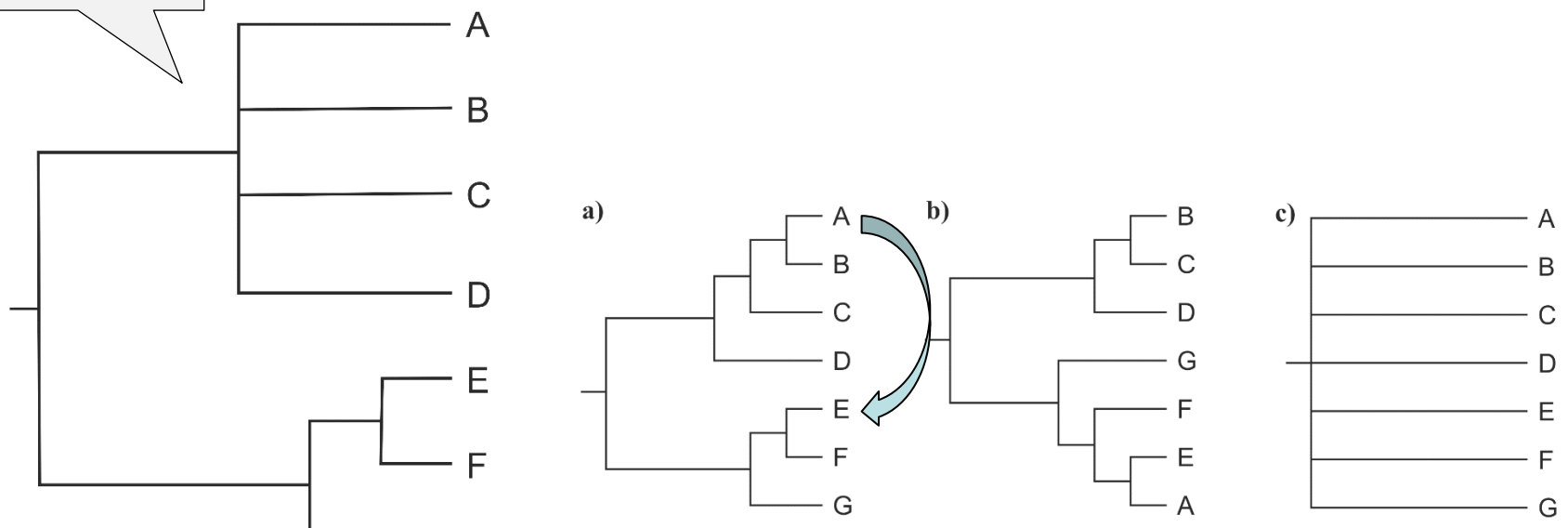
Problémy s distancemi mezi stromy

Konsensuální stromy

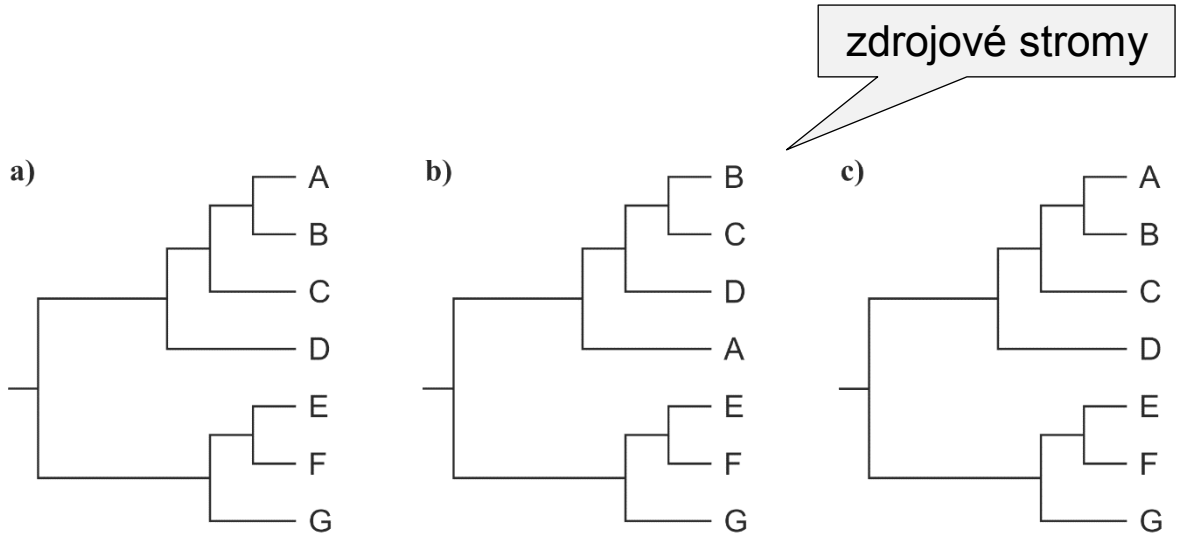
striktní konsensus

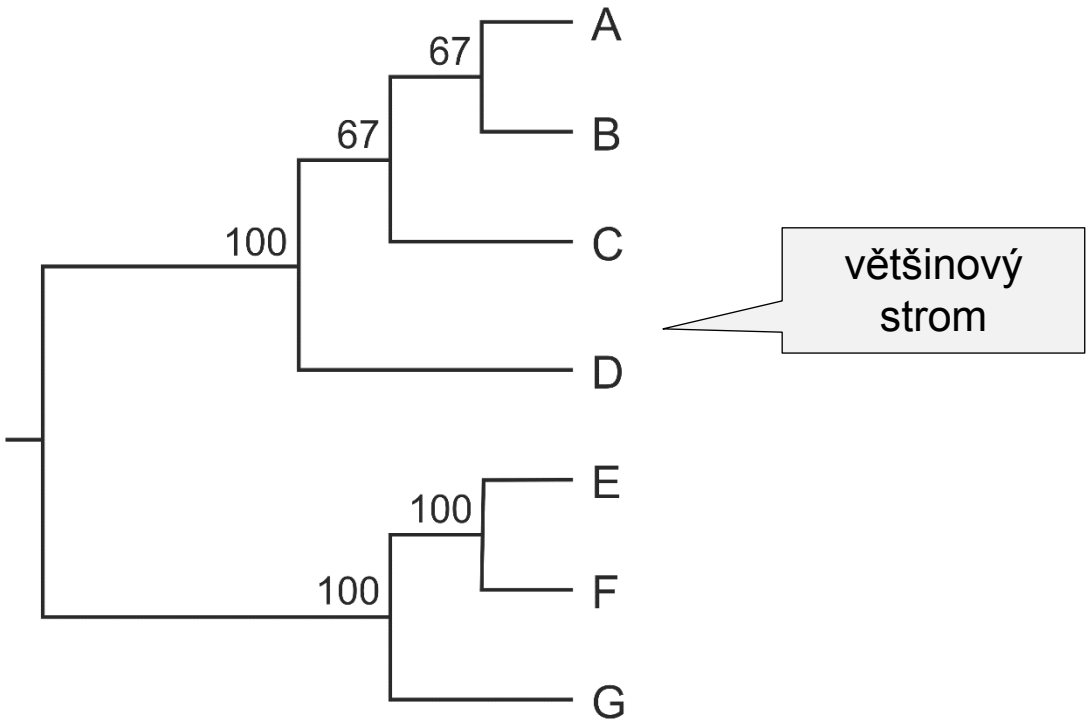
zdrojové stromy

striktně konsensuální strom



majority-rule





Konsensuální stromy

problém s konsensuálními stromy – kombinovaná vs. separátní analýza, *supermatrix vs*. *supertree*

konsensuální stromy v metodách opakovaného výběru, bayesovská analýza

Fylogenetické programy

alignment:

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

konstrukce stromů:

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

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

práce se stromy:

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