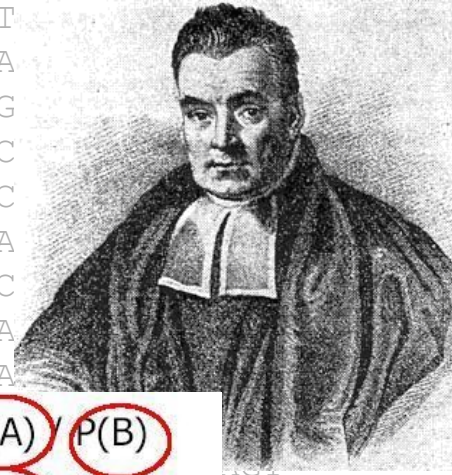
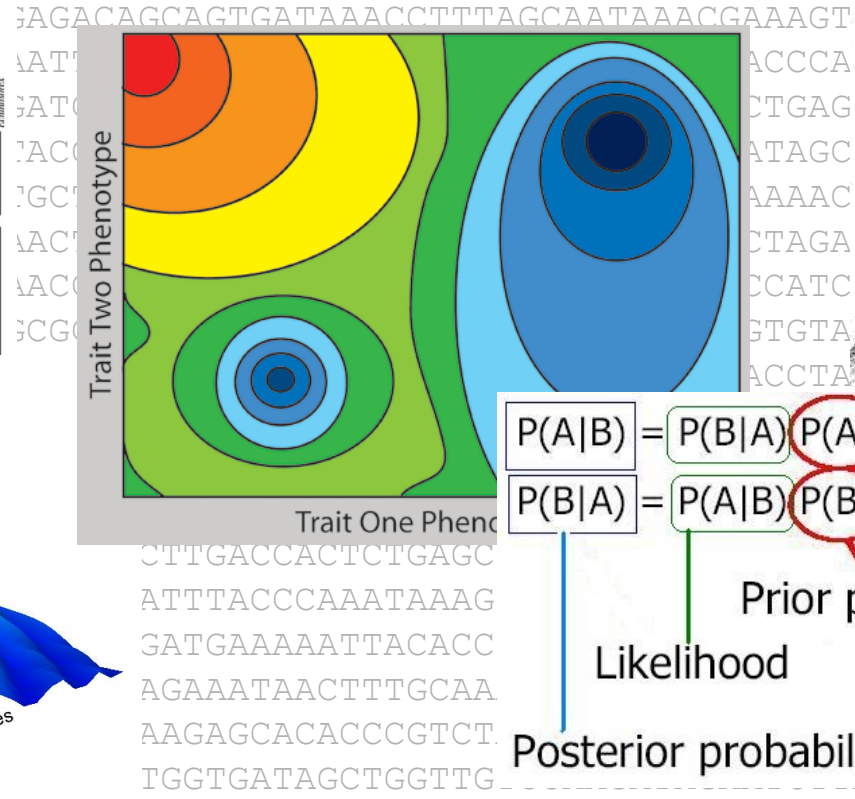
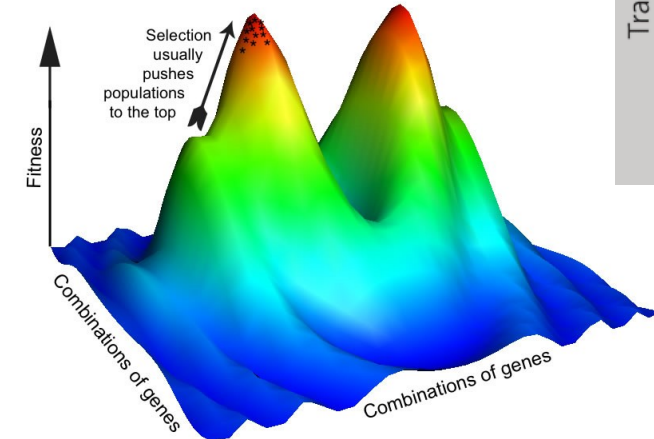
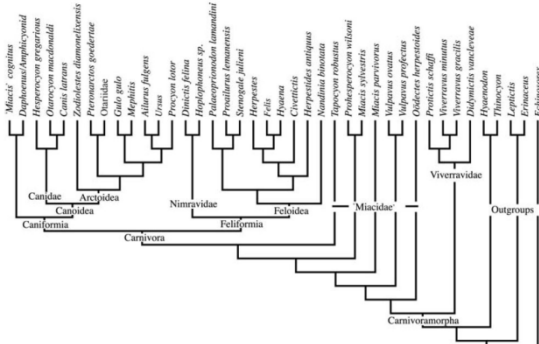


# FYLOGENETICKÁ ANALÝZA II.



15 hodů mincí:

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



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

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

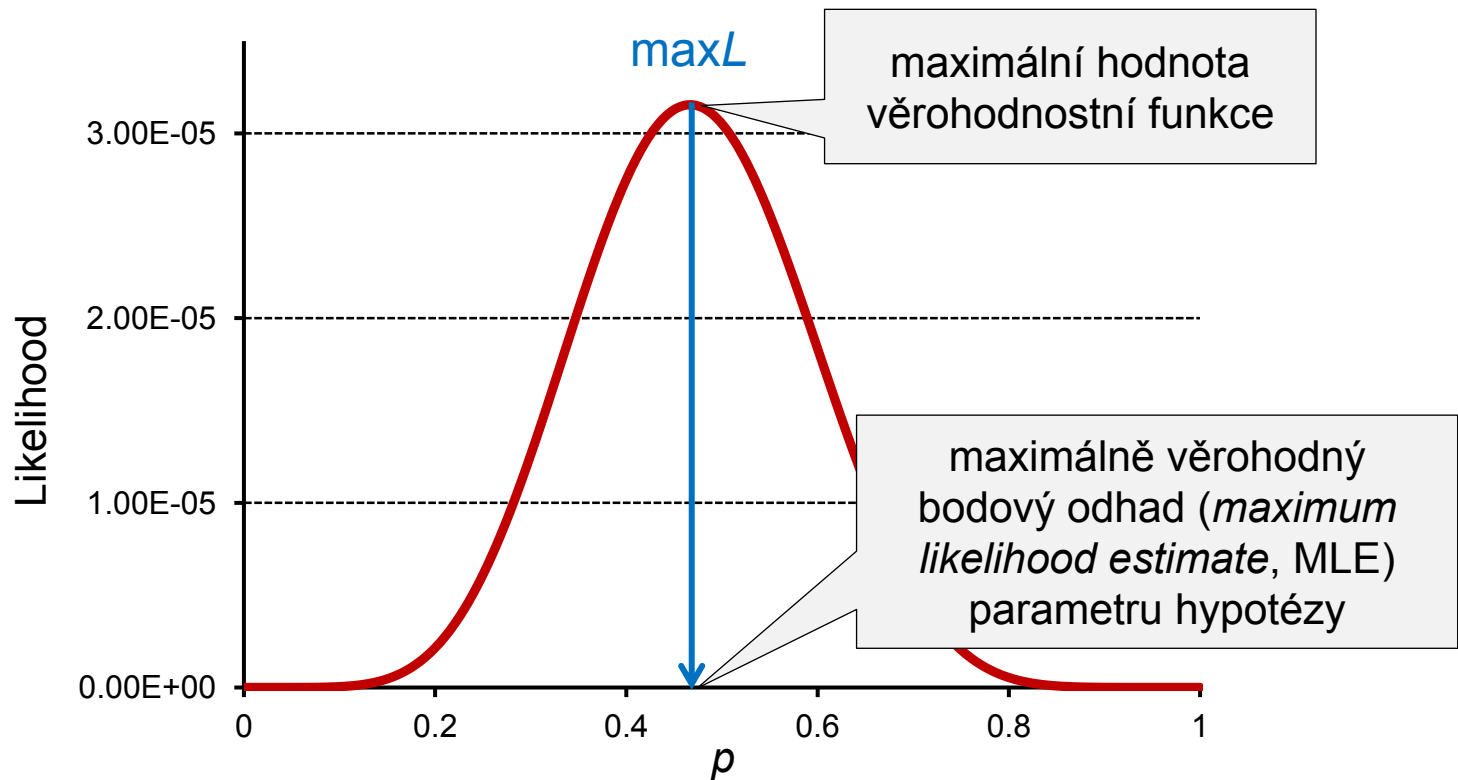
Pravděpodobnost, že padne hlava =  $p$ , orel =  $(1 - p)$

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



Protože hody nezávislé  $\Rightarrow$  pravděpodobnost výsledného skóre =  
 $(1 - p) \times (1 - p) \times p \times p \times p \times (1 - p) \times p \times (1 - p) \times (1 - p) \times (1 - p) \times p \times (1 - p) \times p \times p \times (1 - p) =$   
 $= p^7(1-p)^8$

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



## Hypotéza?

Např. H = mince není „cinknutá“, tj.  $p = 1/2 \Rightarrow L = 3,0517 \cdot 10^{-5}$

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

$$p = 1/3 \Rightarrow L = 1,7841 \cdot 10^{-5}$$

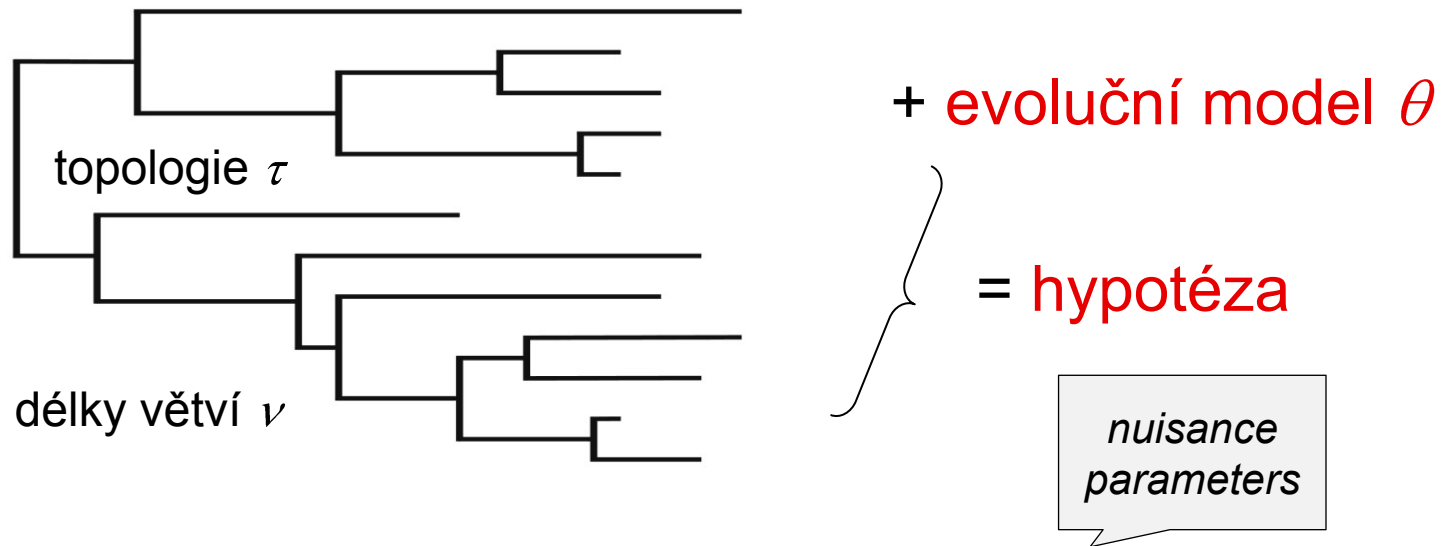
$\Rightarrow$  výsledek hodů 1,7× pravděpodobnější s pravou mincí

# Maximální věrohodnost ve fylogenetické analýze

data:

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

strom:



$L = P(D | H)$ :  $D$  = matice sekvencí (dat),  $H = \tau$  (topologie) +  $\nu$  (délky větví) +  $\theta$  (model)

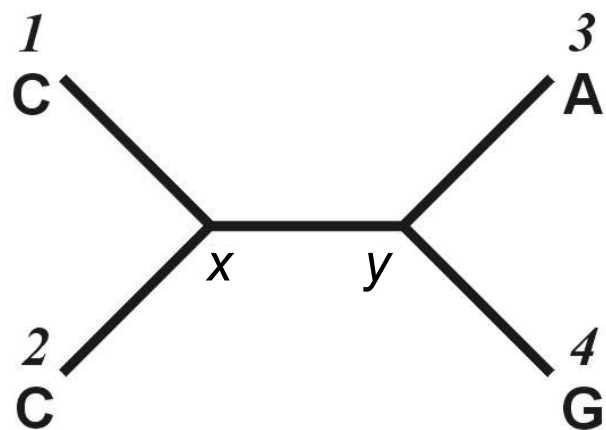
1 N

**1** TCAAAAATGGCTTTATTTC**C**TTAATGCCGTTAACCTTGCGGGGGCCATG

**2** TCCGTGATGGATTTATTT**C**GCAATGCCTGTCATCTTATTCTCAAGTATC

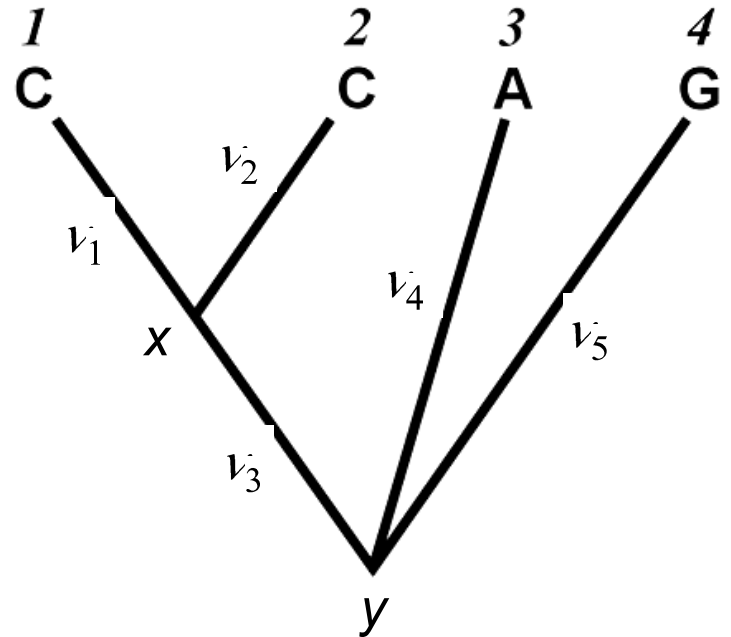
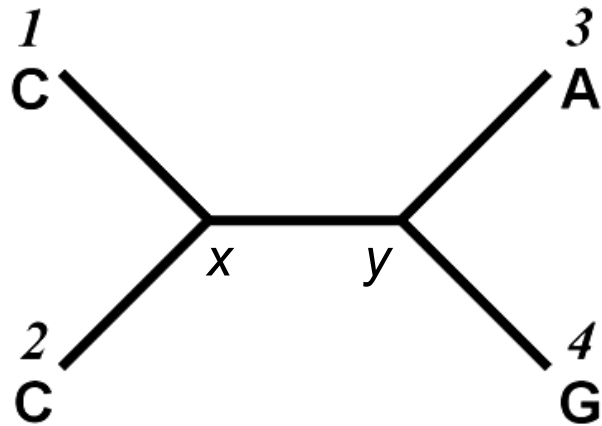
**3** TTCGTGATGGATTTATTG**A**GGTATGCCAGTCATCCTTTTCTCATCTATC

**4** TTCGTGACGGGTTTATCT**G**GCAATGCCGGTCATCCTATTTTCGAGTATT



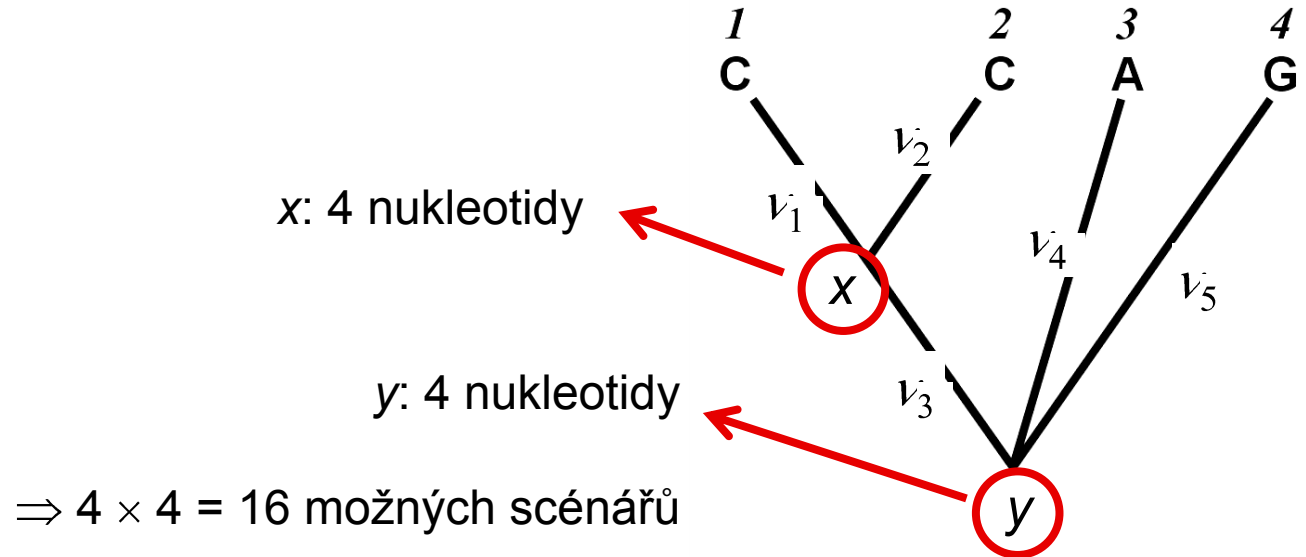
	1		$j$		$N$
<b>1</b>	TCAAAAATGGCTTTATTTC	<b>C</b>	TTAATGCCGTTAACCTT	TGCGGGGGCCATG	
<b>2</b>	TCCGTGATGGATTTATTTTC	<b>C</b>	GCAATGCCTGTCATCTT	ATTCTCAAGTATC	
<b>3</b>	TTCGTGATGGATTTATTG	<b>A</b>	GGTATGCCAGTCATCCT	TTTTCTCATCTATC	
<b>4</b>	TTCGTGACGGGTTTATCTC	<b>G</b>	GCAATGCCGGTCATCCT	ATTTTCGAGTATT	

$v_i$  = délky větví



	1	$j$	$N$
<b>1</b>	TCAAAAATGGCTTTATTTC	<b>C</b>	TTAATGCCGTTAACCTTGC
<b>2</b>	TCCGTGATGGATTTATTTTC	<b>C</b>	GCAATGCCTGTCATCTTATTCTCAAGTATC
<b>3</b>	TTCGTGATGGATTTATTG	<b>A</b>	GGTATGCCAGTCATCCTTTTTCTCATCTATC
<b>4</b>	TTCGTGACGGGTTTATCTC	<b>G</b>	GCAATGCCGGTCATCCTATTTTCGAGTATT

$v_i$  = délky větví



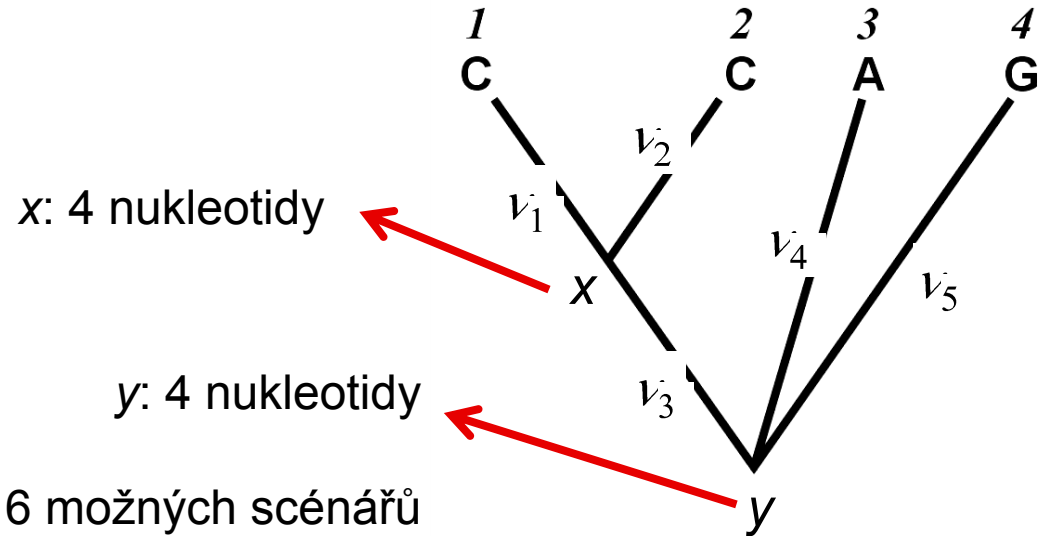
$$L(1) = P(y) \times P(y \rightarrow x) v_3 \times P(x \rightarrow C) v_1 \times P(x \rightarrow C) v_2 \times P(y \rightarrow A) v_4 \times P(y \rightarrow G) v_5$$

$$L(j) = P(\text{scénář 1}) + \dots + P(\text{scénář 16})$$



	1		$j$		$N$
<b>1</b>	TCAAAAATGGCTTTATTTC	<b>C</b>	TTAATGCCGTTAACCCCTTGC	GGGGGCCATG	
<b>2</b>	TCCGTGATGGATTTATTT	<b>C</b>	GCAATGCCTGTCATCTTATT	CTCAAGTATC	
<b>3</b>	TTCGTGATGGATTTATTG	<b>A</b>	GGTATGCCAGTCATCCTTTT	CTCATCTATC	
<b>4</b>	TTCGTGACGGGTTTATCT	<b>G</b>	GCAATGCCGGTTCATCCTATT	TTTCGAGTATT	

$v_i =$  délky větví

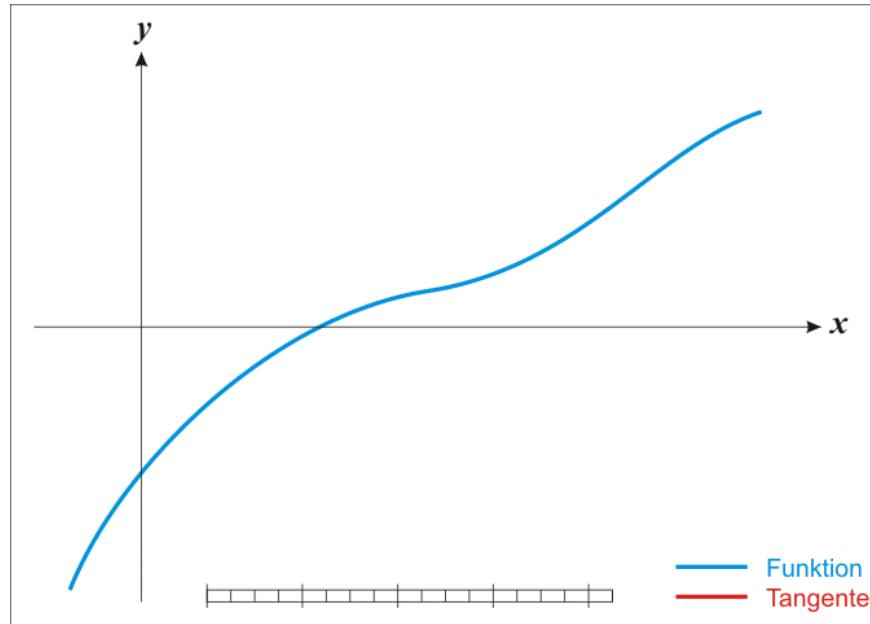


všechny pozice:  $L = L(1) \times L(2) \times \dots \times L(j) \times \dots \times L(N) = \prod_{j=1}^N L_j$

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

# Hledání maximální věrohodnosti daného stromu

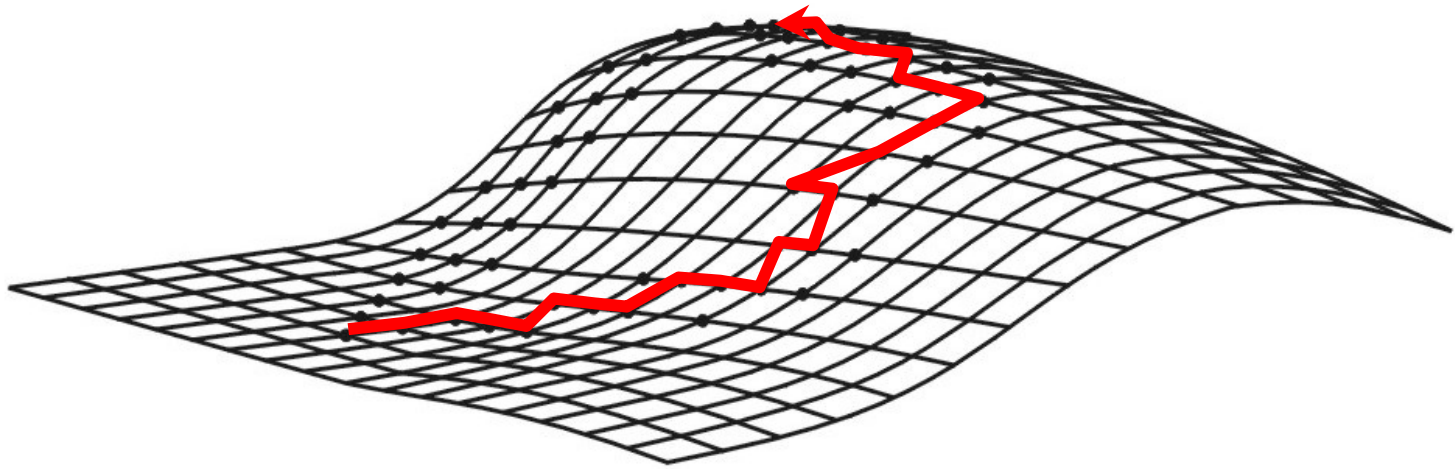
→ např. Newtonova (Newtonova-Raphsonova) metoda



[https://upload.wikimedia.org/wikipedia/commons/e/e0/NewtonIteration\\_Ani.gif](https://upload.wikimedia.org/wikipedia/commons/e/e0/NewtonIteration_Ani.gif)

Hledání nejvěrohodnějšího stromu: heuristické hledání

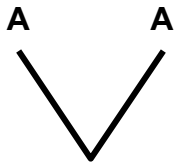
# Heuristické hledání



*stepwise addition* ... např. PHYLIP

*star decomposition* ... např. MOLPHY; *neighbor-joining tree*

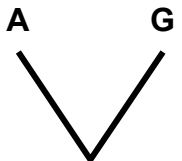
*branch swapping*



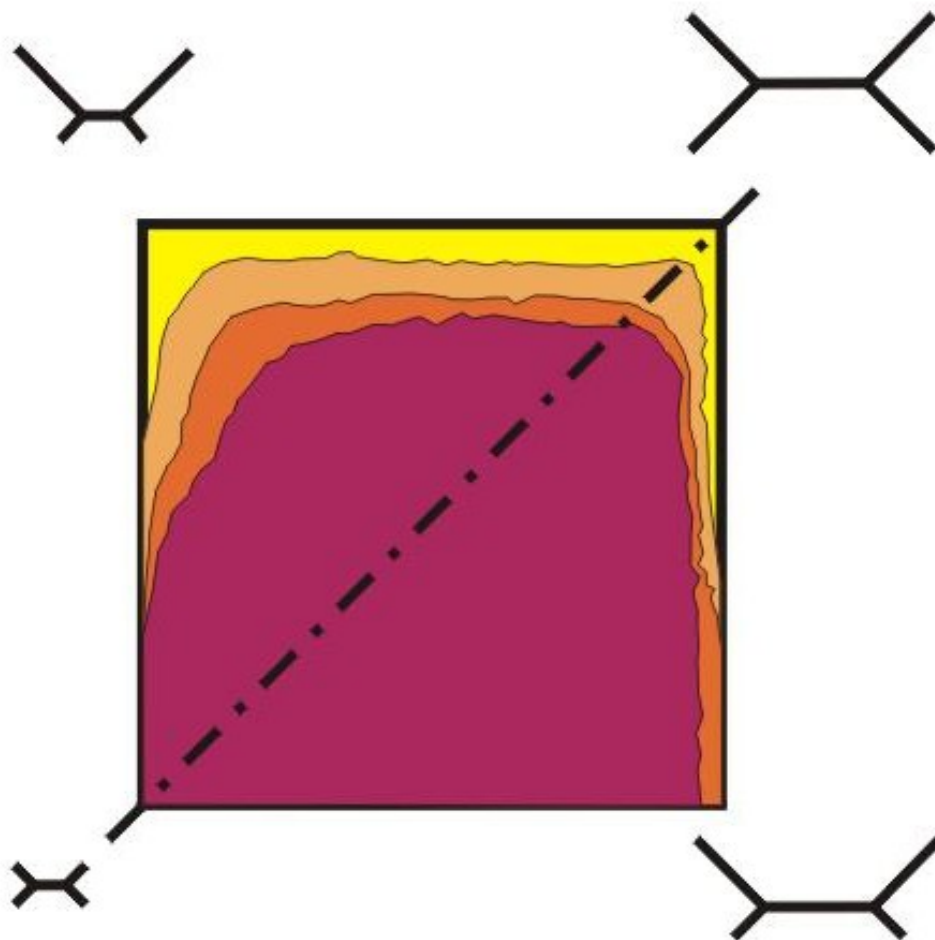
## Věrohodnost (ML) a úspornost (MP)

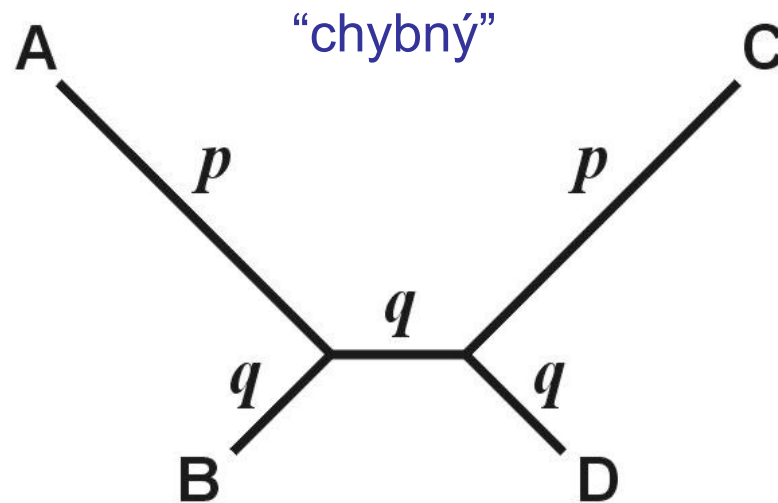
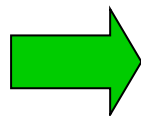
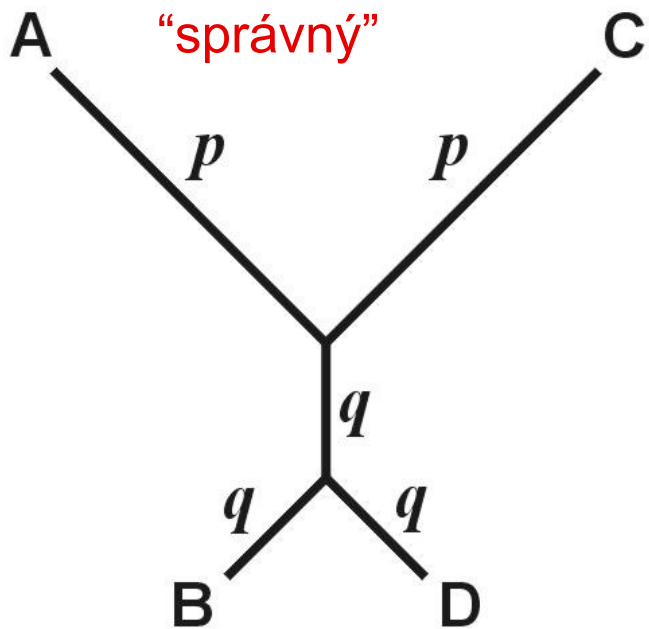
Počet změn	Parsimonie	$\nu = 0,01$	$\nu = 0,10$	$\nu = 0,20$	$\nu = 1,00$
		(0,2475)	(0,2266)	(0,20611)	(0,11192)
0	100	99,99	99,83	99,31	82,17
1	0	0,00	0,00	0,00	0,00
2	0	0,0011	0,11	0,44	9,13
3	0			0,034	3,55
4	0				0,0027

Počet změn	Parsimonie	$\nu = 0,01$	$\nu = 0,10$	$\nu = 0,20$	$\nu = 1,00$
		(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



# 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?

$$P(H | D)$$

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  $P(H | D)$  se nazývá **aposteriorní** (*posterior probability*)

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

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

**Aposterioorní pravděpodobnost, že naše kostka je falešná, je dána Bayesovou rovnicí:**

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

věrohodnost

apriorní pravděpodobnost



suma čitateľů pro všechny alternativní hypotézy





Thomas Bayes

Pro náš příklad se 2 hody kostkou:







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:

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

	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

$$\begin{aligned}
 P(\text{biased} | \text{2 dots, 6 dots}) &= \frac{P(\text{2 dots, 6 dots} | \text{biased}) \times P(\text{biased})}{P(\text{2 dots, 6 dots} | \text{biased}) \times P(\text{biased}) + P(\text{2 dots, 6 dots} | \text{fair}) \times P(\text{fair})} \\
 &= \frac{18/441 \times 2/10}{18/441 \times 2/10 + 1/36 \times 8/10} = \underline{0,269}
 \end{aligned}$$

## 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

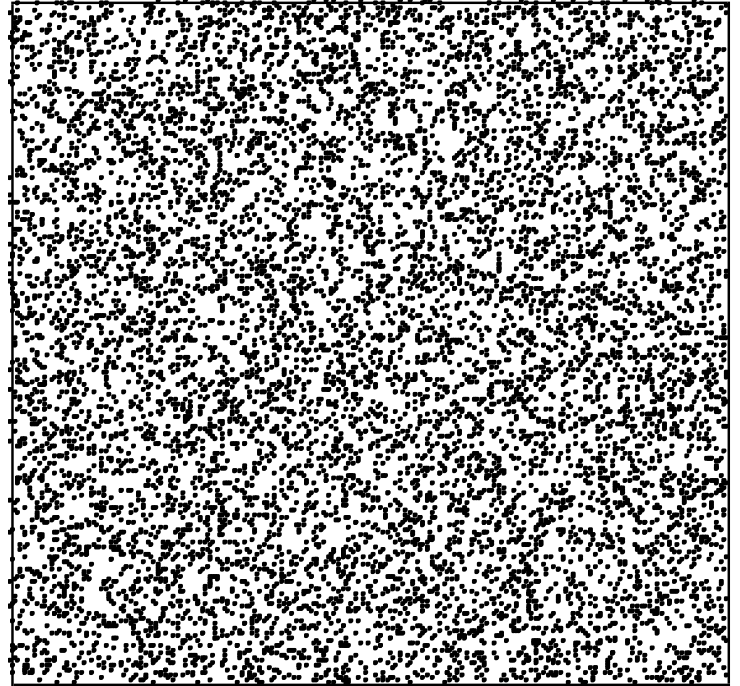
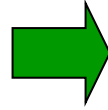
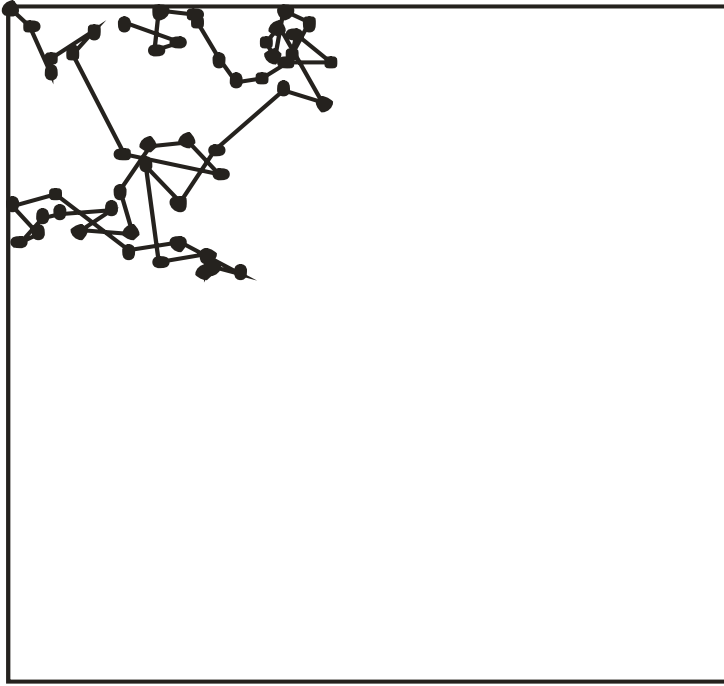
řešení: metody Monte Carlo

náhodný výběr vzorků, při velkém množství aproximace skutečnosti

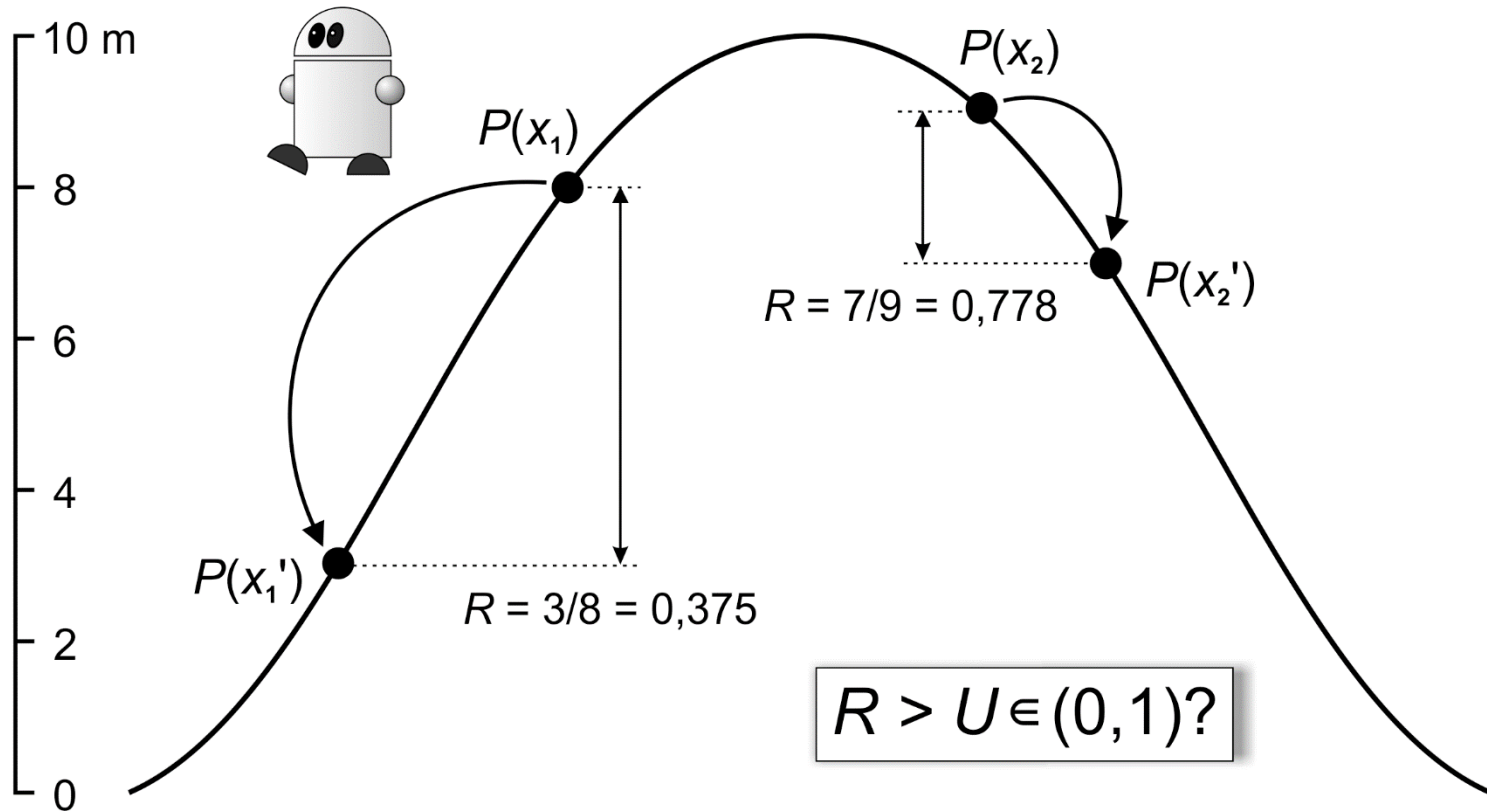
Markovovy řetězce: Markov chain Monte Carlo (MCMC)

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

...  $P$  stejná po celé fylogenii = homogenní Markovův proces



# Metropolisův-Hastingsův algoritmus:

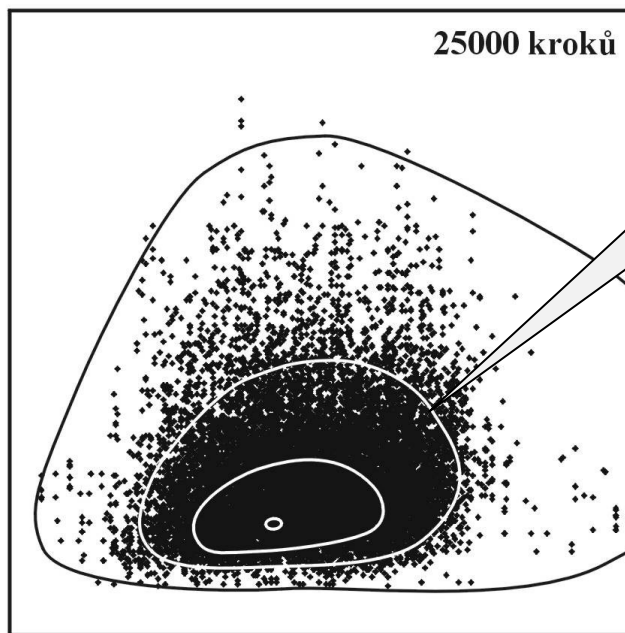
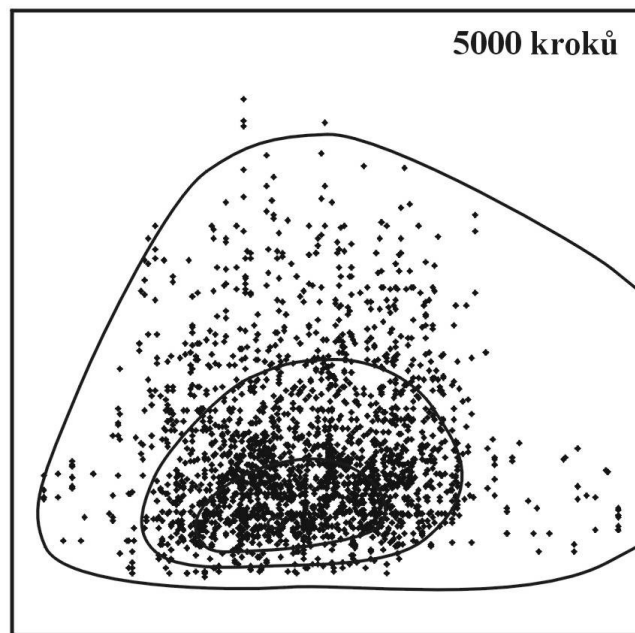


## Metropolisův-Hastingsův algoritmus:

Změna parametru  $x \rightarrow x'$

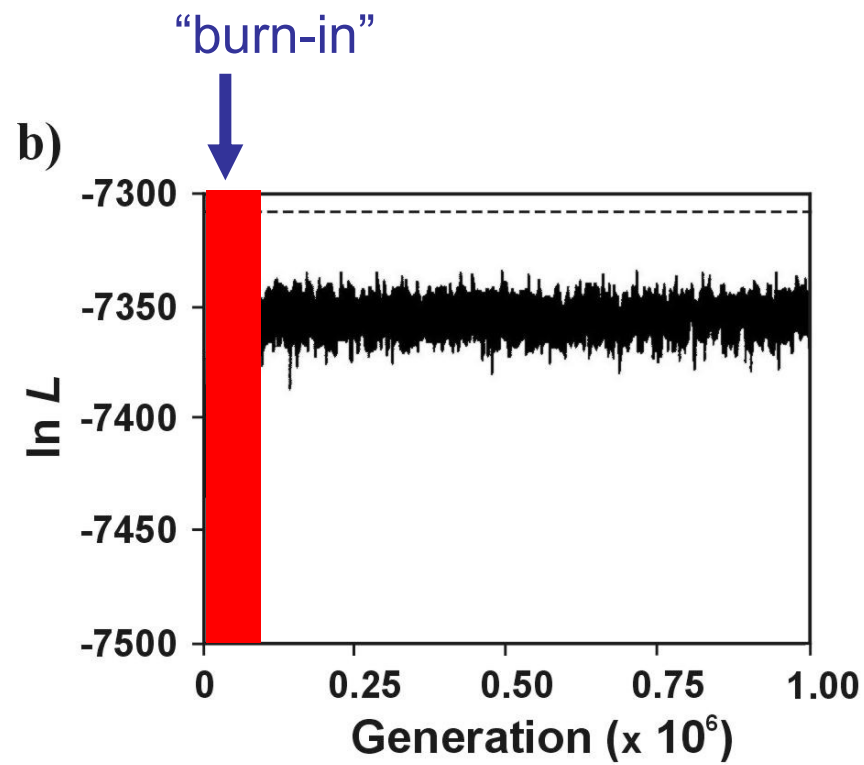
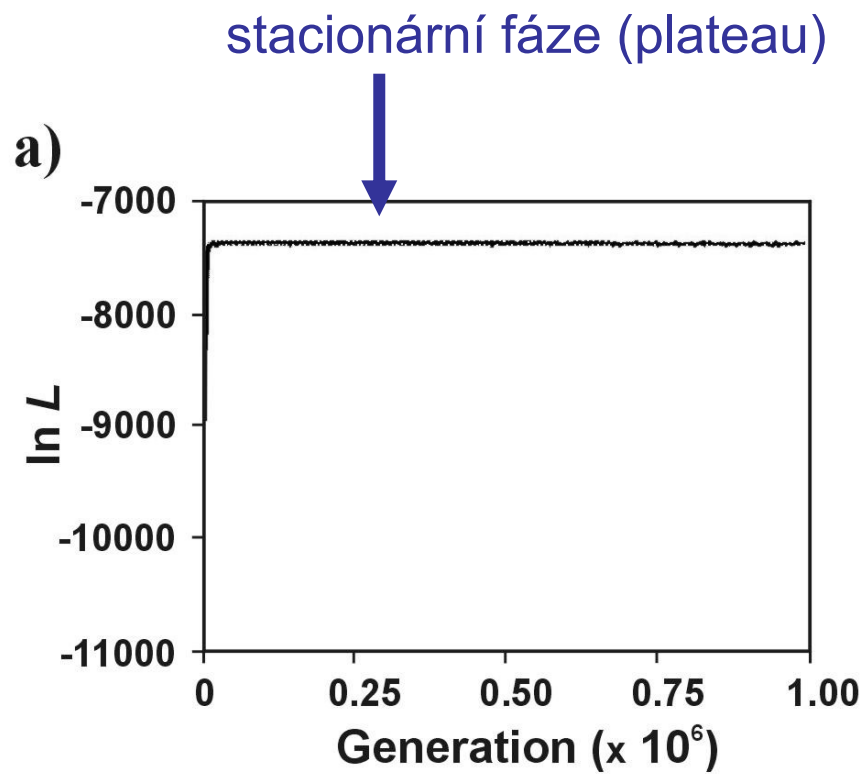
1. jestliže  $P(x') > P(x)$ , akceptuj  $x'$
2. jestliže  $P(x') \leq P(x)$ , vypočti  $R = P(x')/P(x)$   
protože platí, že  $P(x') \leq P(x)$ , musí být  $R \leq 1$
3. generuj náhodné číslo  $U$  z rovnoměrného rozdělení z intervalu  $(0, 1)$
4. jestliže  $R \geq U$ , akceptuj  $x'$ , jestli ne, ponechej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

usměrněný pohyb robota v aréně:



„vrstevnice“  
arény





## Reverzibilní přeskokový MCMC (*reversible jump MCMC*):

umožňuje měnit počet parametrů při každém MC kroku

Ize 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)

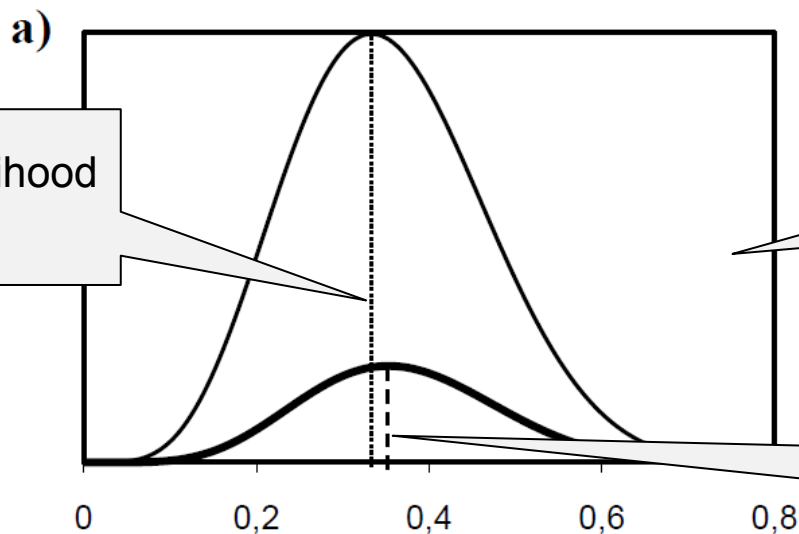
## *Metropolis coupled MCMC* (MCMCMC, MC<sup>3</sup>):

1 „chladný“ řetězec, 3 „zahřáté“ řetězce

stejný výchozí bod, díky stochasticitě rychlá divergence „robotů“

MrBayes: <http://morphbank.ebc.uu.se/mrbayes/>

# Problém apriorních pravděpodobností

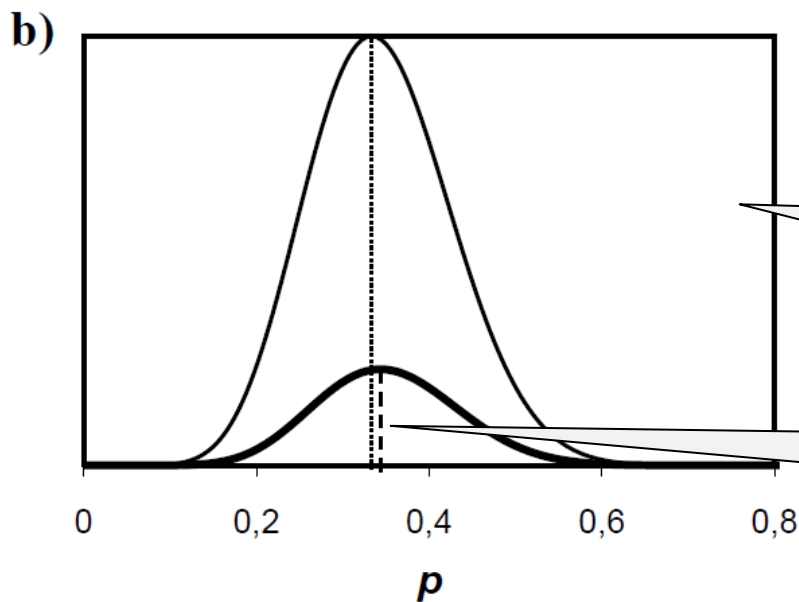


maximum likelihood  
= 0,333

apriorní pr. = 0,5

15 hodů mincí  
výsledek 5 H : 10 O

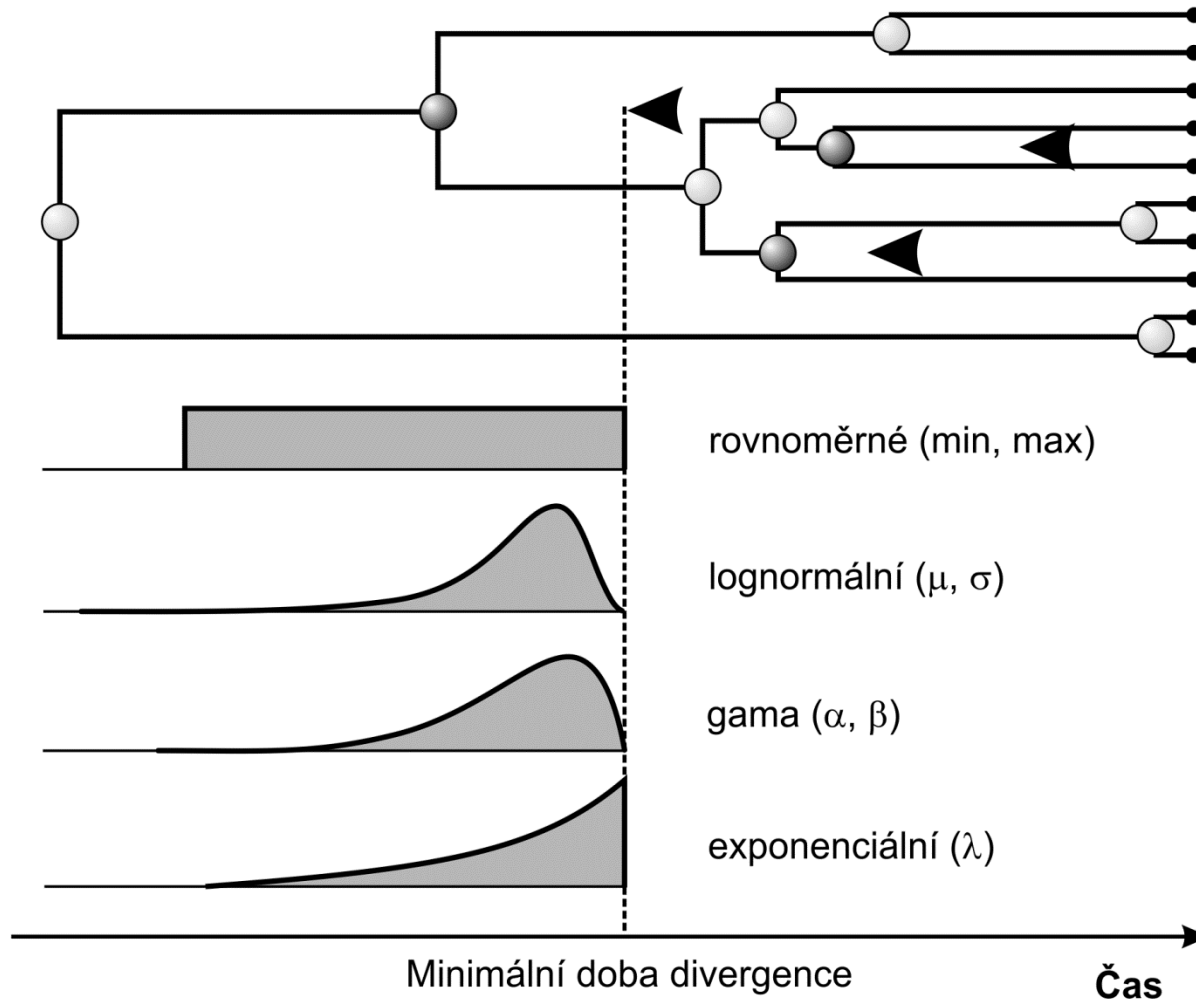
díky apriorní pr.  
aposteriorní pr.  
posunuta doprava



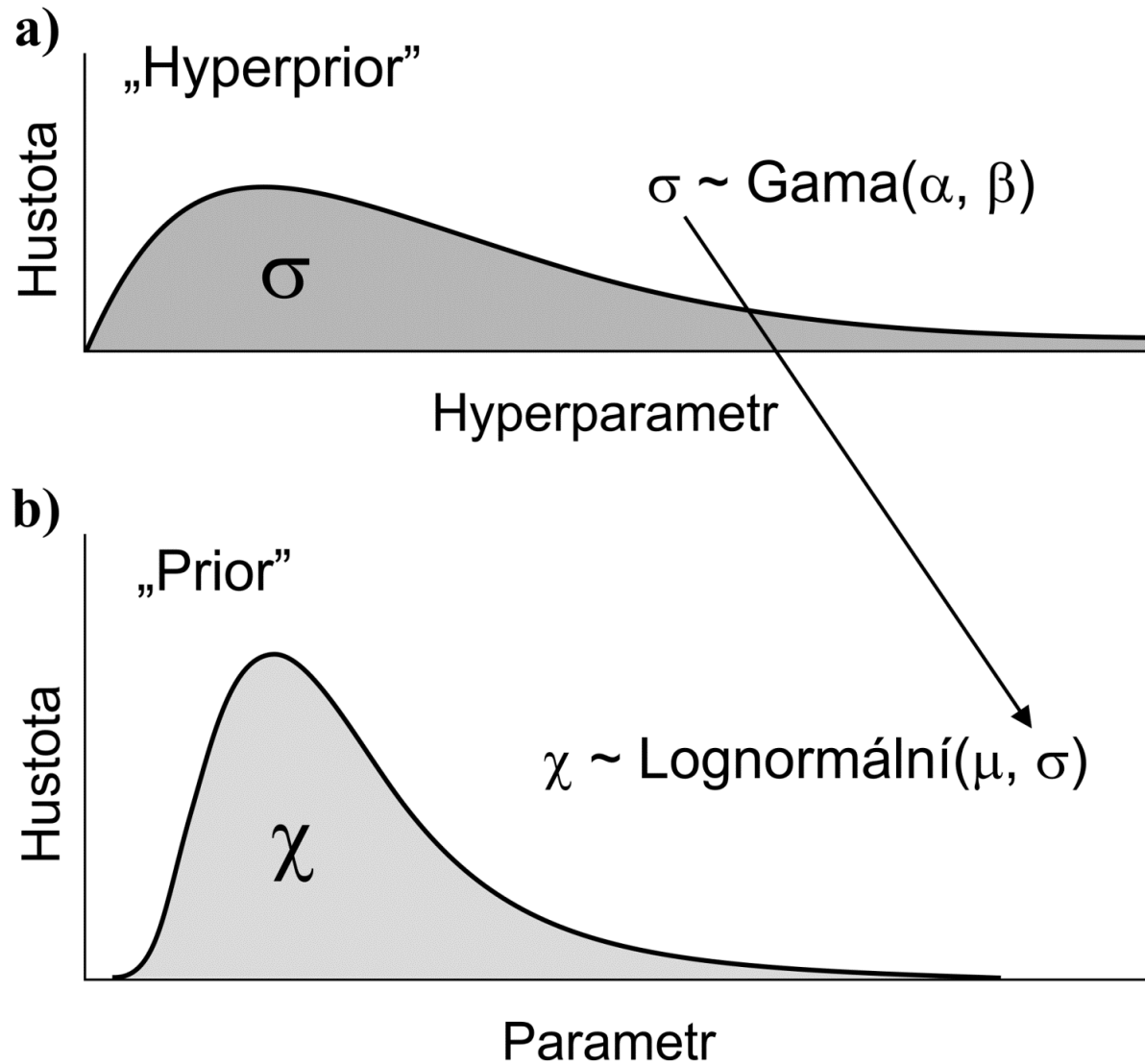
30 hodů mincí  
výsledek 10 H : 20 O

rozdíl od ML  
menší

# Problém apriorních pravděpodobností



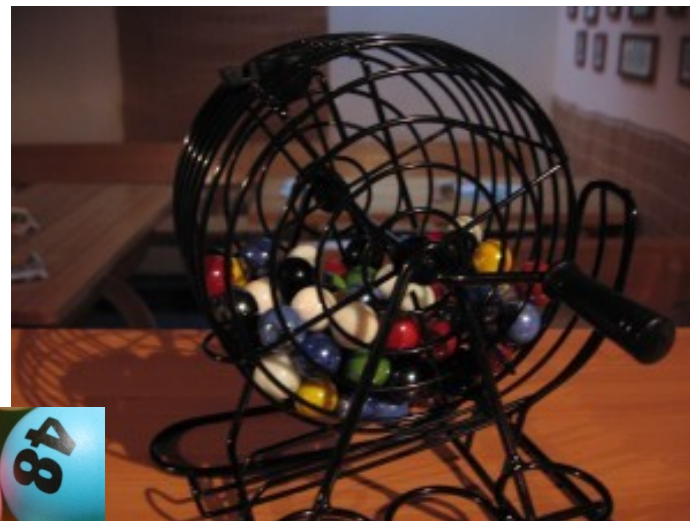
# Stanovení apriorních pravděpodobností:



# Měření spolehlivosti stromů

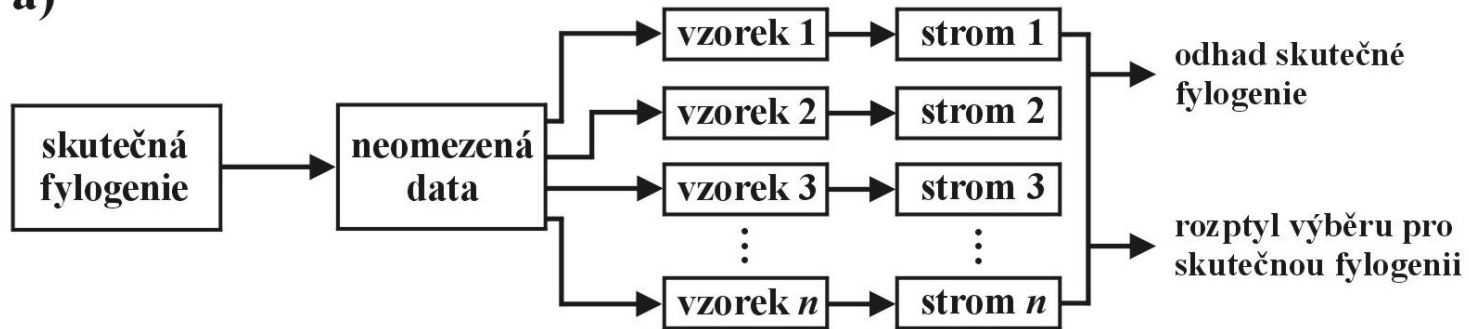
Metody opakovaného výběru

bez navrácení = **jackknife**  
s navrácením = **bootstrap**



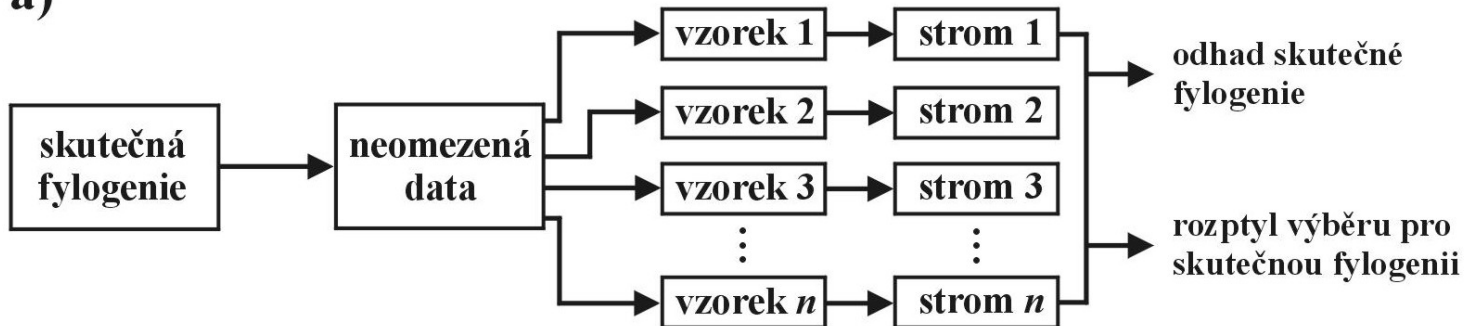
# bootstrap:

a)

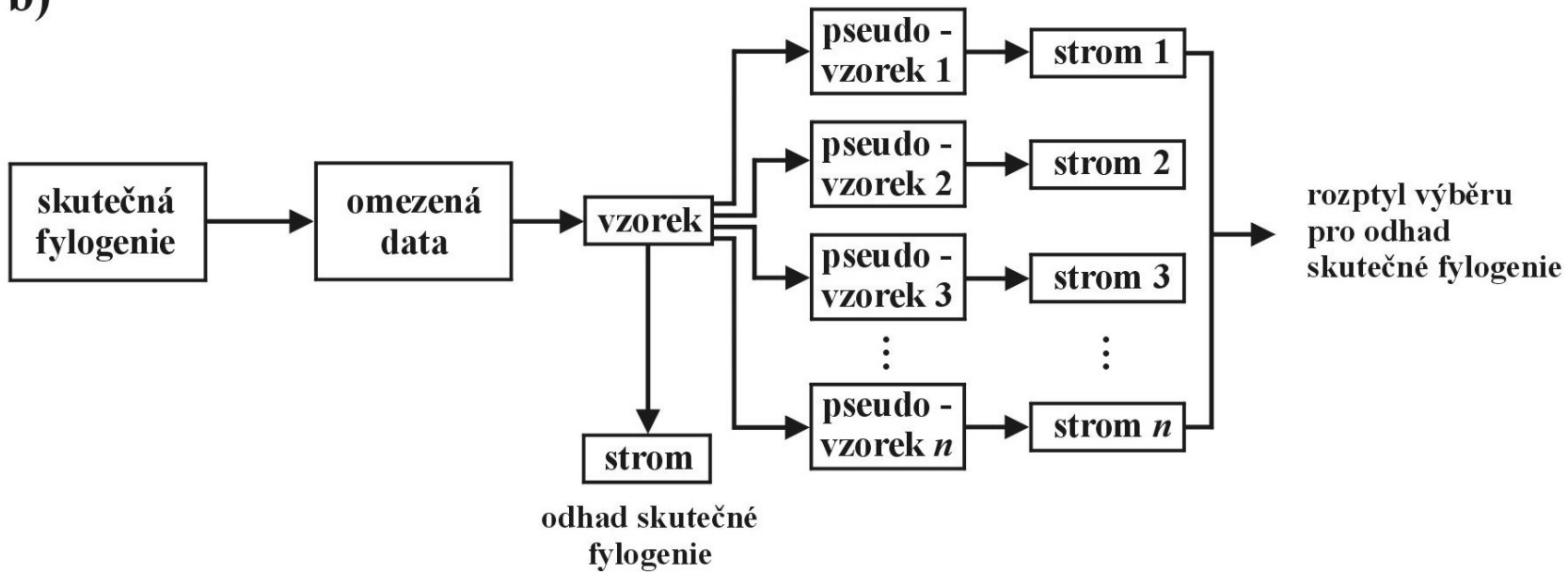


# bootstrap:

a)

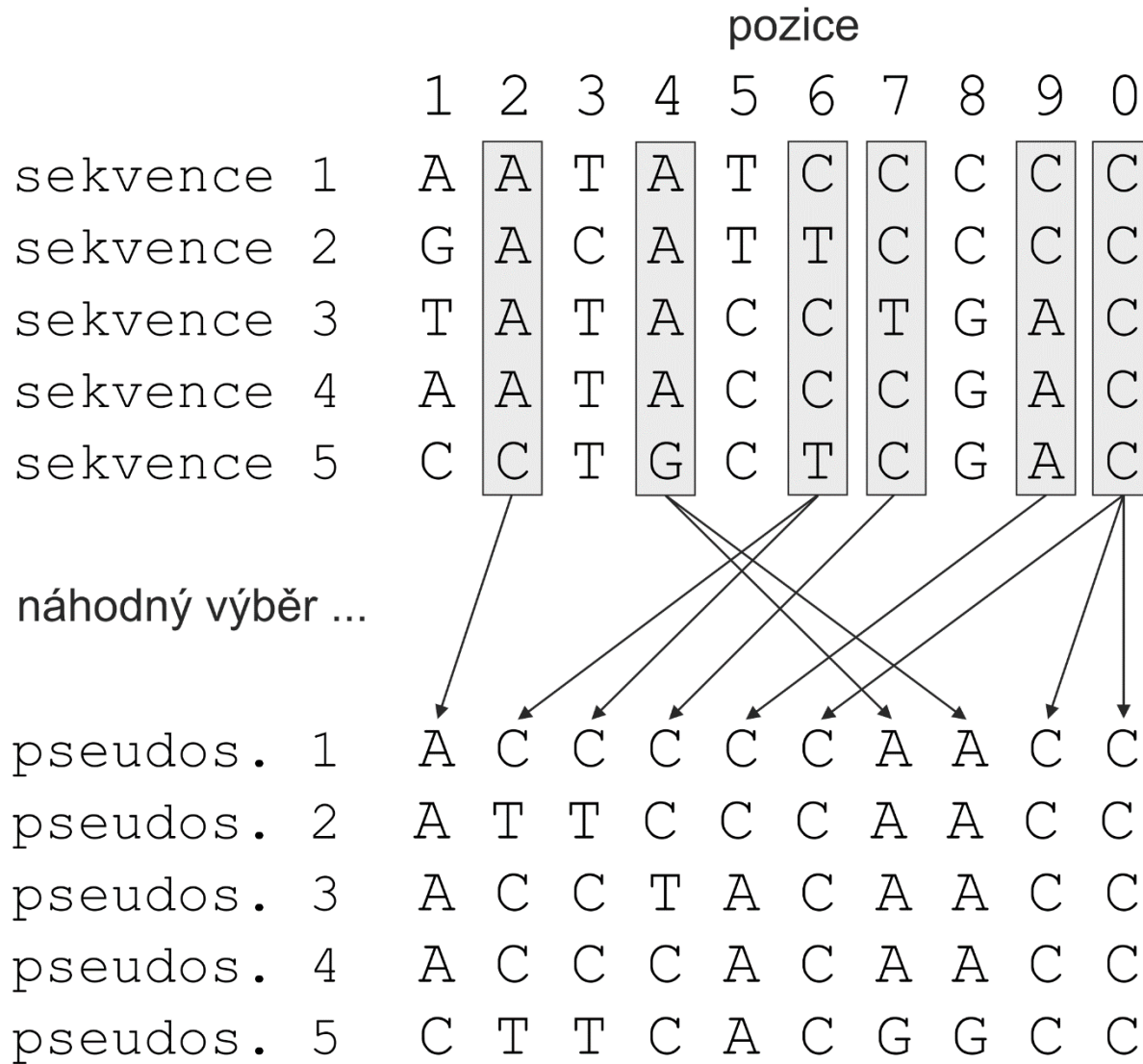


b)

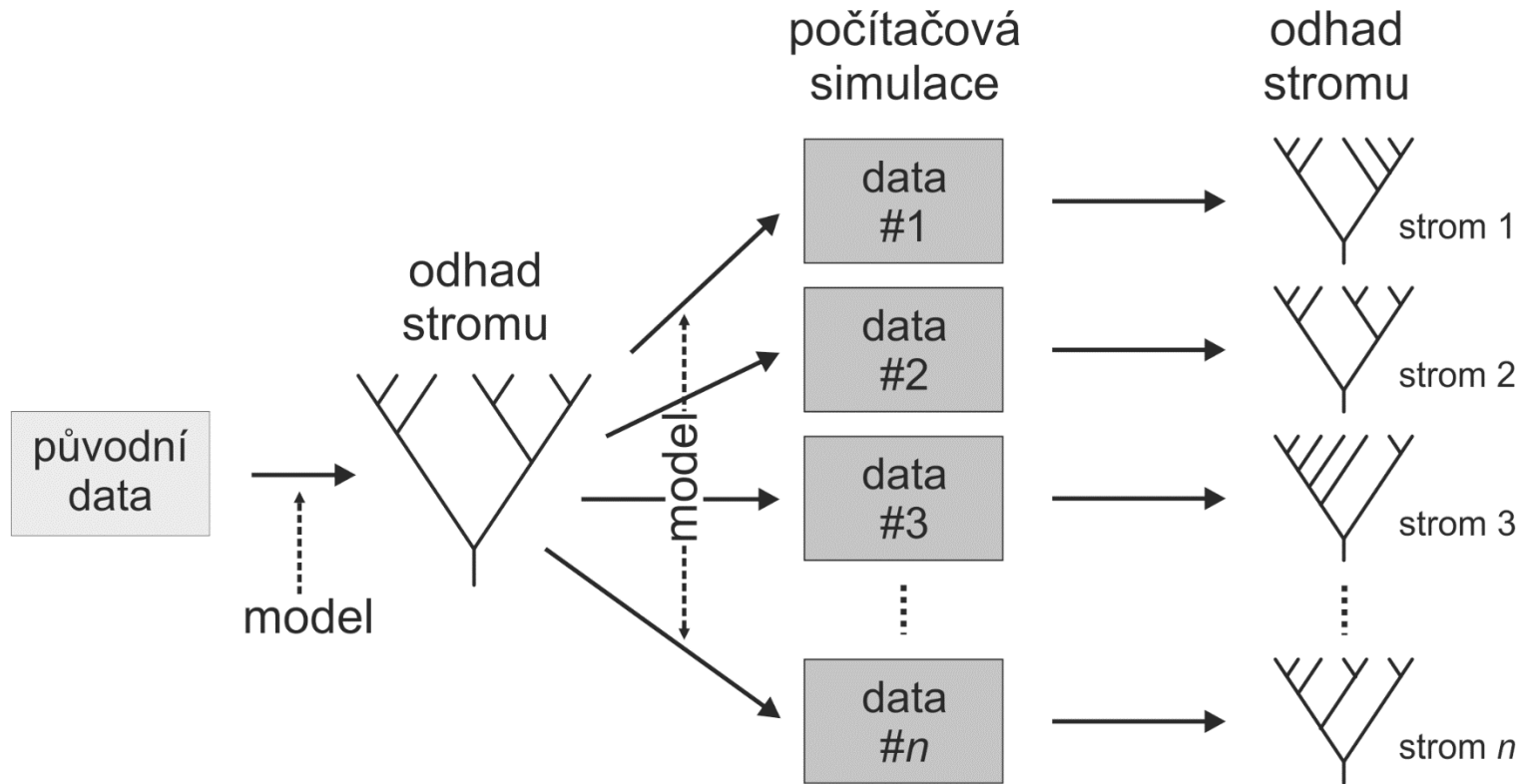




# bootstrap:



# parametrický bootstrap: evoluční model

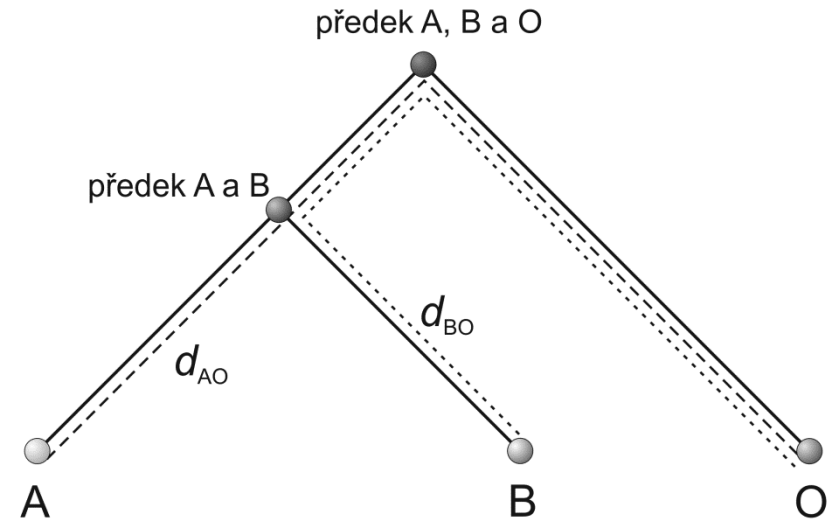


bayesovská analýza: aposteriorní pravděpodobnosti

# Testování hypotéz

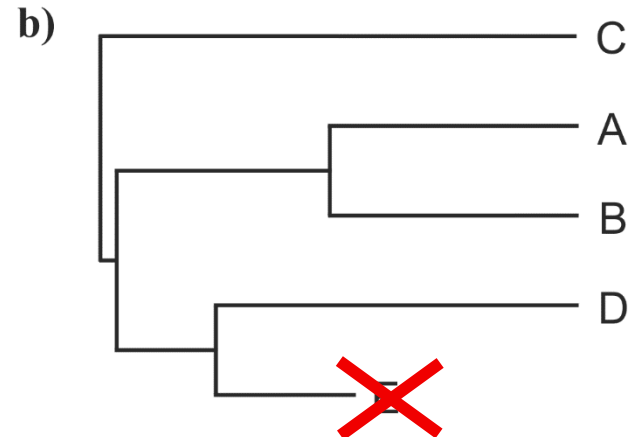
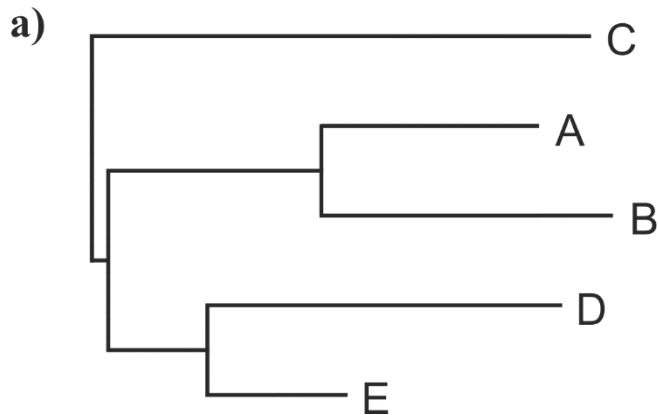
Test molekulárních hodin:

Relative rate test (RRT):  $AC=BC$ ?



Linearizované stromy

odstranění signifikantně odlišných taxonů

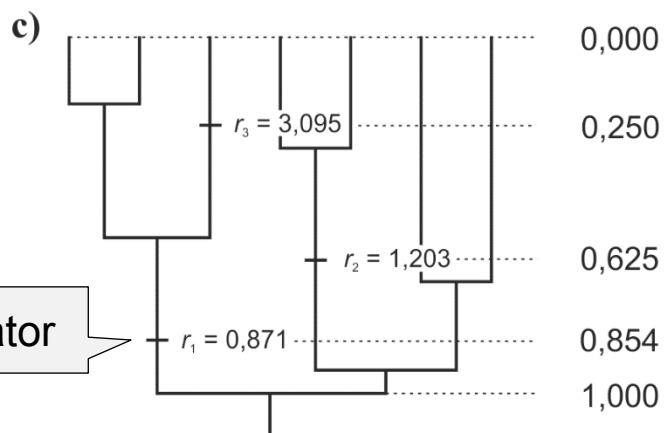
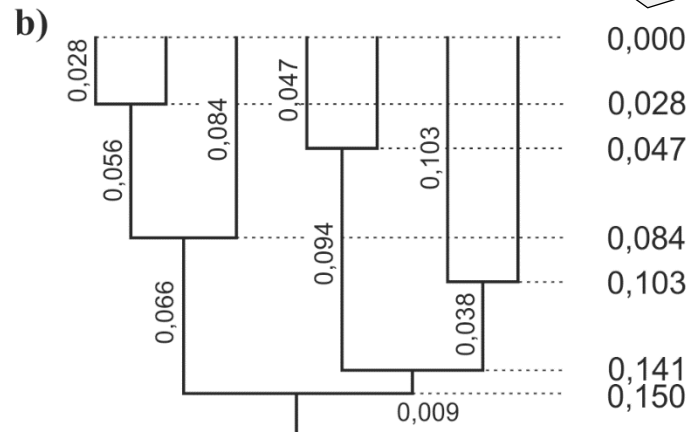
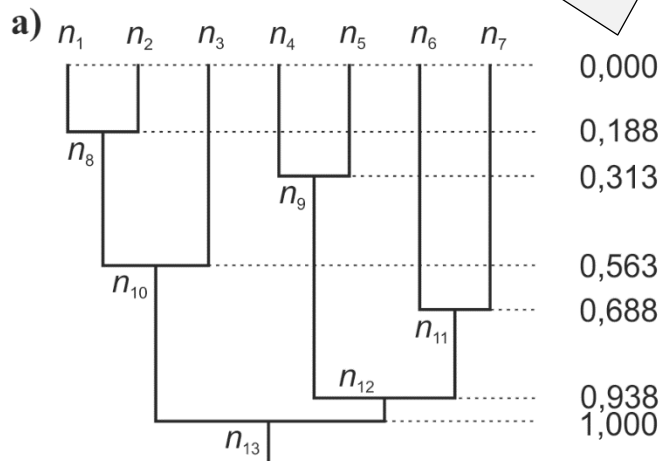


# Relaxované molekulární hodiny

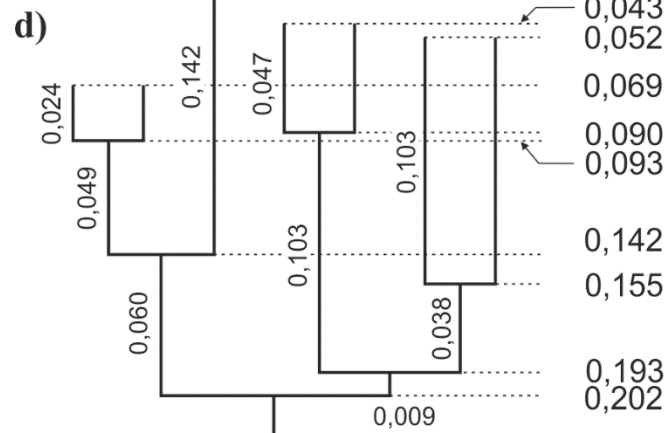
umožňují změnu rychlostí podél větví

neškálovaný čas

škálovaný čas  
(očekávaný poč.  
substitucí/pozici)



multiplikátor



# 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

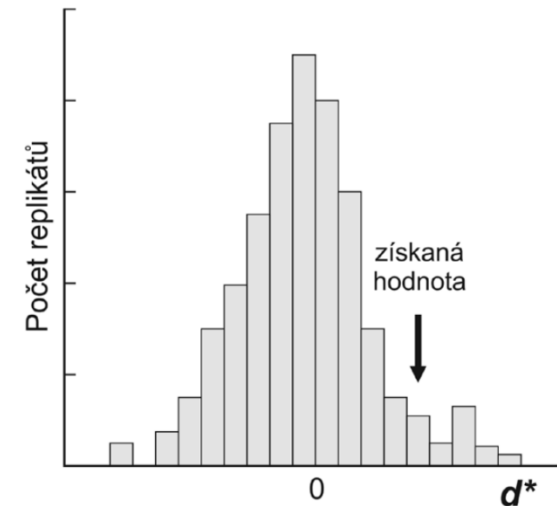
$$d_1^* = \ln L_{T1}^* - \ln L_{T2}^*$$

$$d_2^* = \ln L_{T1}^* - \ln L_{T2}^*$$

$$d_3^* = \ln L_{T1}^* - \ln L_{T2}^*$$

...

$$d_n^* = \ln L_{T1}^* - \ln L_{T2}^*$$



## 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

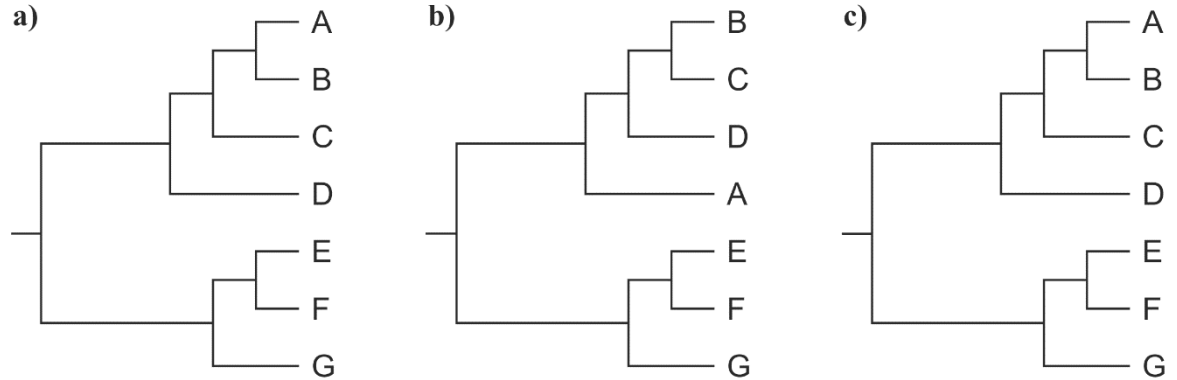
metody inkorporující délky větví

Problémy s distancemi mezi stromy

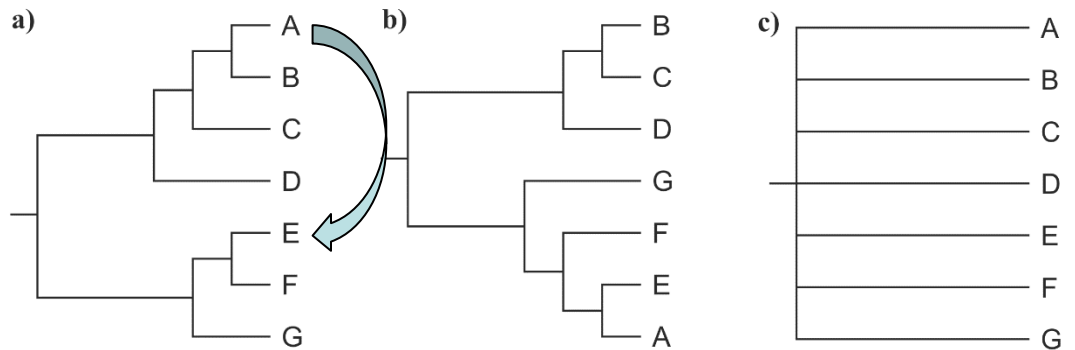
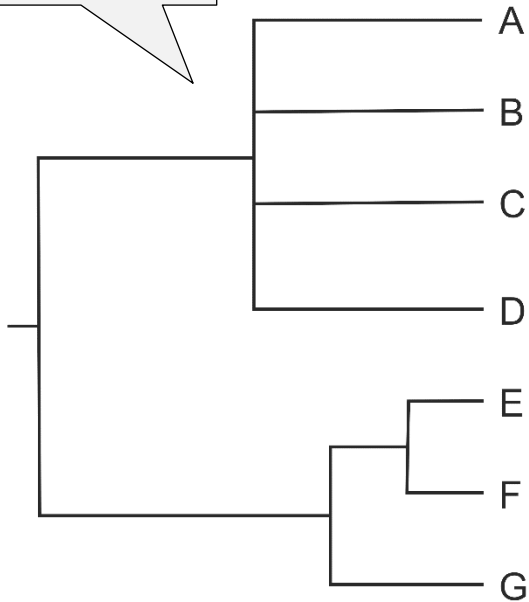
# Konsenzuální stromy

striktní konsensus

zdrojové stromy

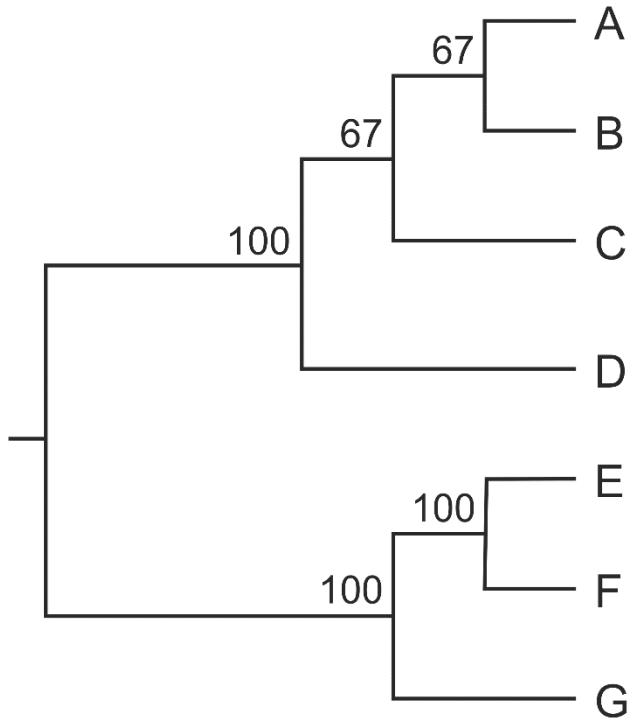
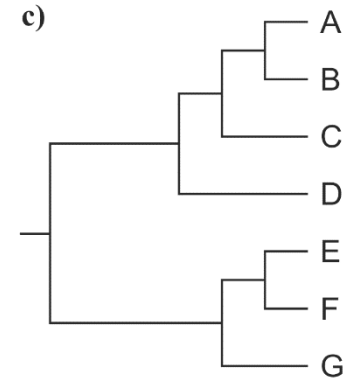
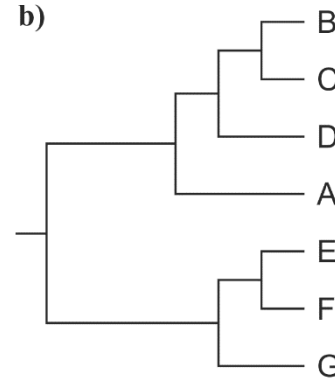
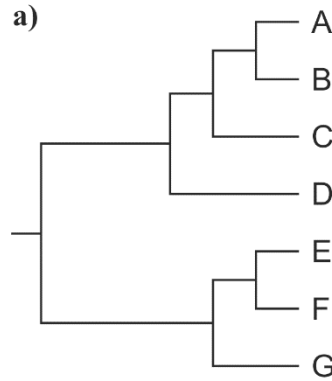


striktně konsenzuální strom



# majority-rule

zdrojové stromy



většinový strom



# Konsenzuální stromy

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

konsenzuální 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>