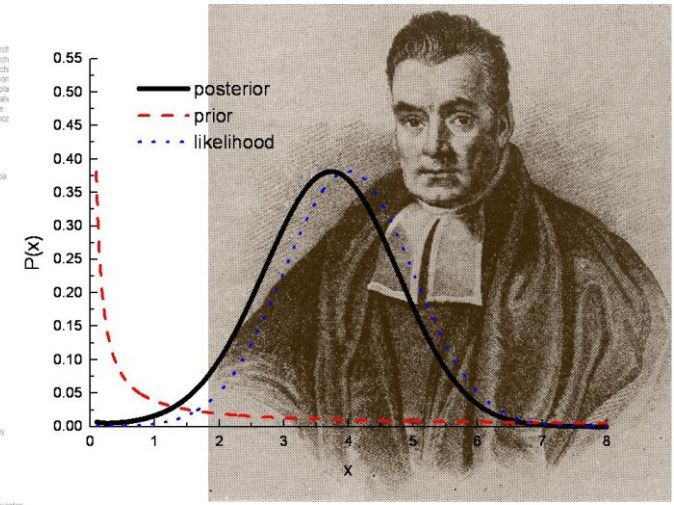
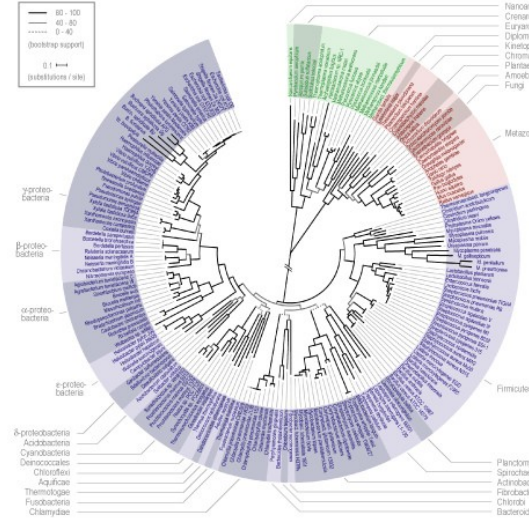
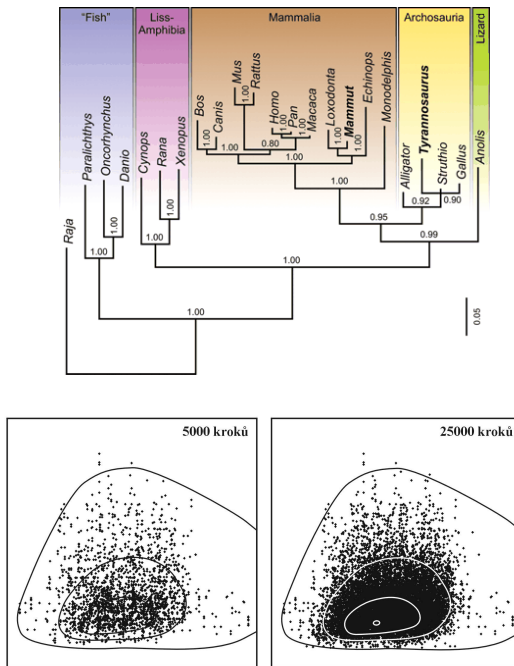


METODY ANALÝZY DAT II.



Maximální věrohodnost (Maximum likelihood, ML)
heterogenita substitučních rychlostí, ML a konzistence

Bayesovská analýza
MCMC

Měření spolehlivosti stromů
jackknife, bootstrap, parametrický bootstrap, permutační testy

Testování hypotéz
testy molekulárních hodin, srovnávání stromů, distance mezi stromy

Konsensuální stromy

Maximální věrohodnost (maximum likelihood, ML)

hod mincí 15× → skóre OOHHHOHOOOHOOHHO:

7× panna (hlava, H)8× orel (O)

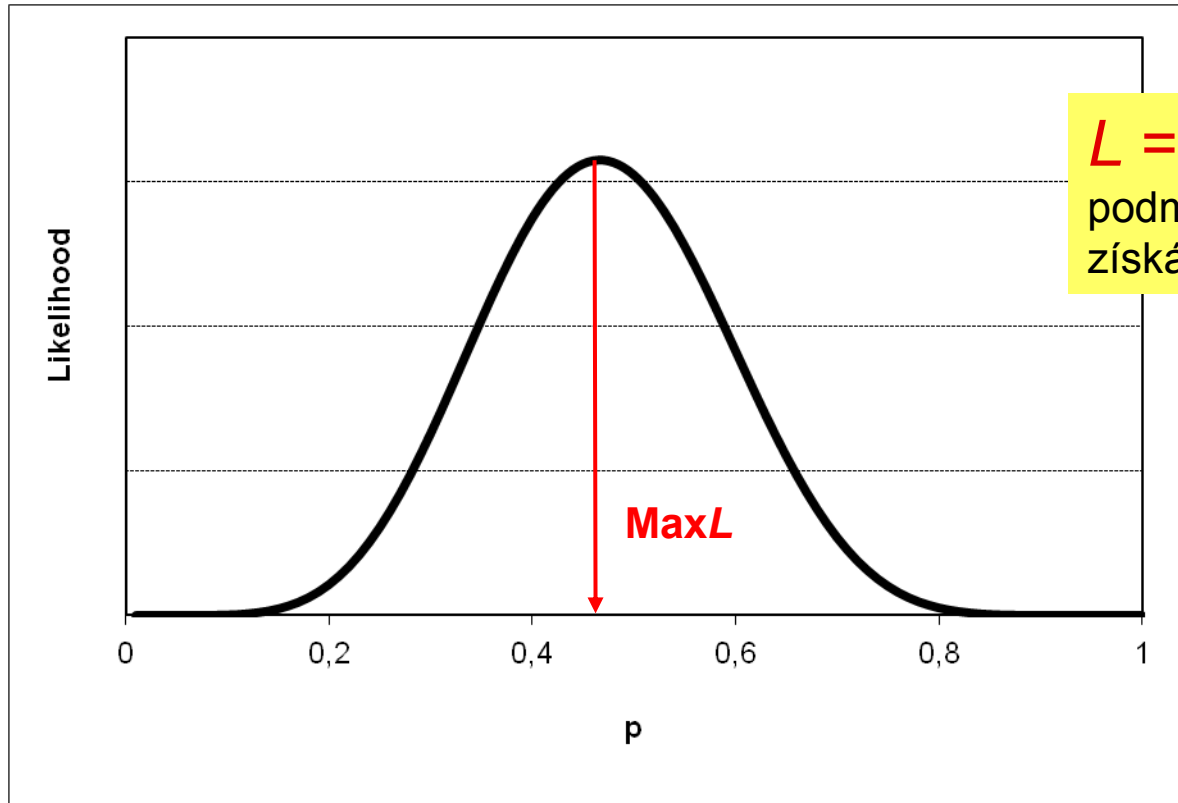
pravděpodobnost, že padne hlava = p , orel = $(1 - p)$

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 = $0,4666 \approx 7/15$

Maximální věrohodnost (maximum likelihood, ML)



$$L = (D | H)$$

podmíněná pravděpodobnost
získání dat D při hypotéze H

pravá mince: $p = 1/2 \Rightarrow L = 3,0517 \cdot 10^{-5}$

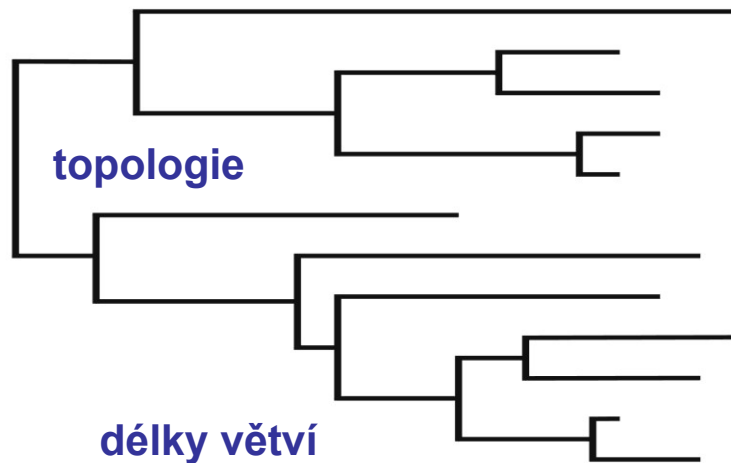
falešná mince: $p = 1/3 \Rightarrow L = 1,7841 \cdot 10^{-5} \Rightarrow$ výsledek hodů $1,7\times$
pravděpodobnější s pravou mincí

Maximální věrohodnost ve fylogenetické analýze

data:

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

strom:



evoluční model

= hypotéza

Věrohodnostní funkce: jaká je
pravděpodobnost získání daných
dat při dané hypotéze?

$$L = P(D | H),$$

kde D = matice dat
 $H = \tau$ (topologie),
 ν (délky větví),
 θ (model)

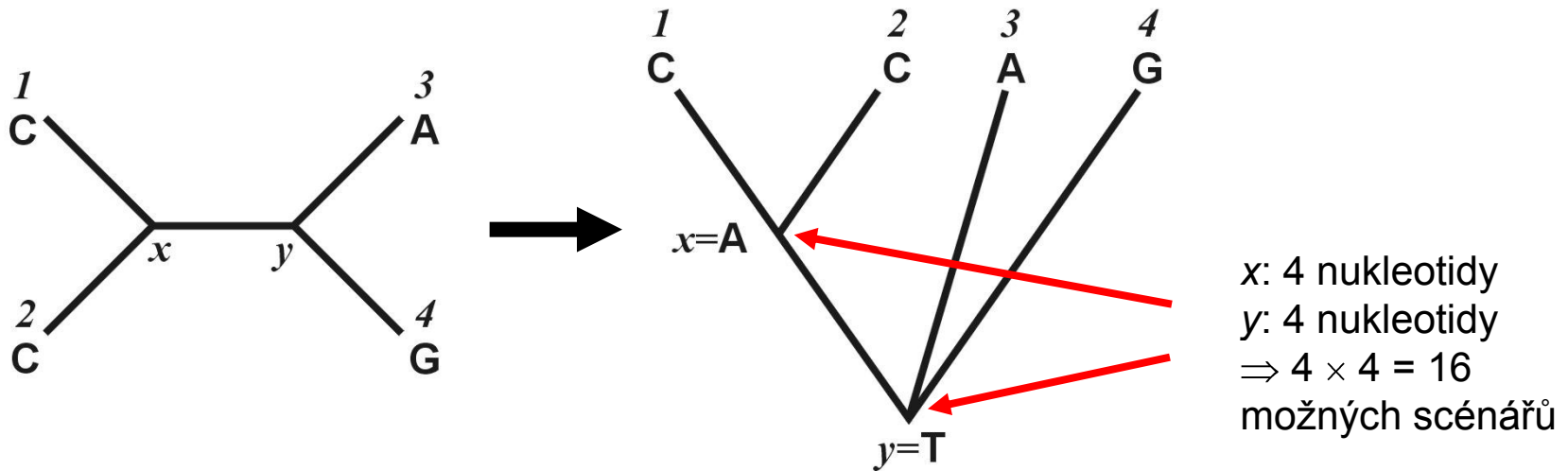
1 j N

1 TCAAAAATGGCTTTATTTCGCTTAATGCCGTTAACCCCTTGCGGGGGCCATG

2 TCCGTGATGGATTTATTTCCGCAATGCCTGTCATCTTATTCTCAAGTATC

3 TTCGTGATGGATTTATTGCAGGTATGCCAGTCATCCTTTTCTCATCTATC

4 TTCGTGACGGGTTTATCTCGGCAATGCCGGTTCATCCTATTTTCGAGTATT

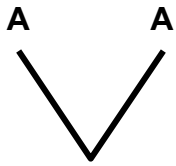


1) $L(1) = P(A) \times P(T) \times P(AC) \times P(AC) \times P(TA) \times P(TG)$

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

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

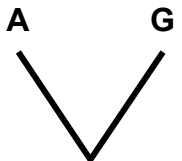
4) $\ln L = \ln L(1) + \ln L(2) + \dots + \ln L(N) = \sum_{j=1}^N \ln L_j$



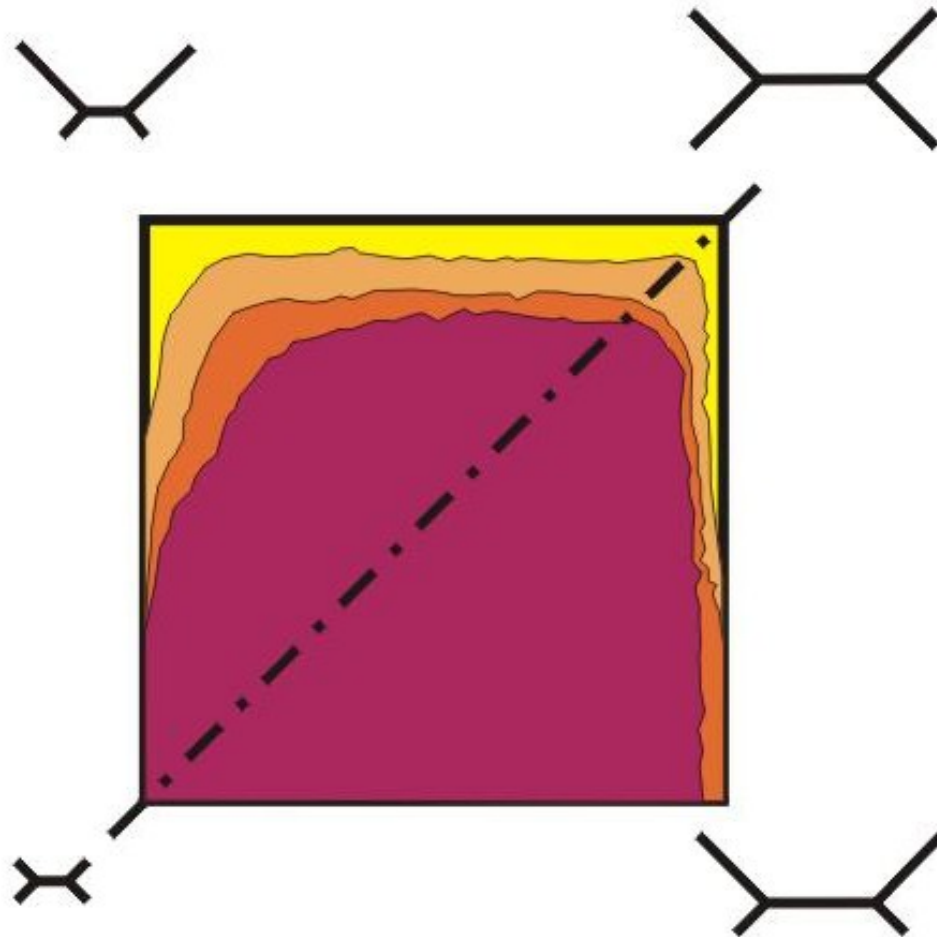
Věrohodnost (ML) a úspornost (MP)

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

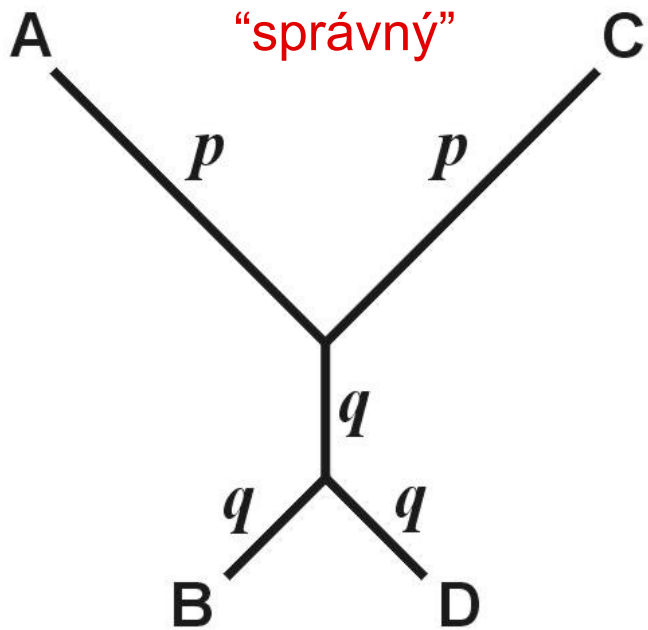
Počet změn	Parsimonie	$\nu = 0,01$	$\nu = 0,10$	$\nu = 0,20$	$\nu = 1,00$
		(0,00083)	(0,00786)	(0,01462)	(0,04602)
0	0	0,00	0,00	0,00	0,00
1	100	99,66	96,64	92,36	66,54
2	0	0,33	3,22	6,22	21,19
3	0		0,12	0,48	8,61
4	0		0,003	0,023	2,05
5	0			0,0037	0,42



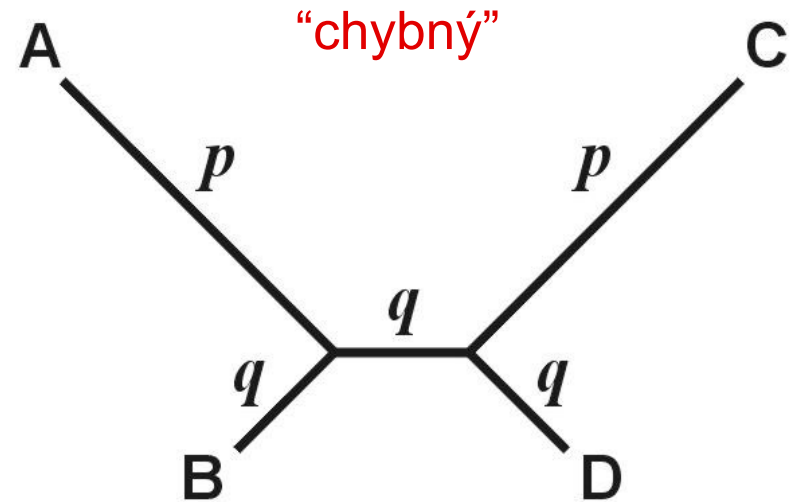
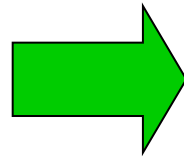
Věrohodnost a konzistence:



Věrohodnost a konzistence:



Farrisova
(anti-Felsensteinova,
inverzní Felsensteinova)
zóna



“long-branch repulsion”

Bayesovská analýza

(ML: jaká je pravděpodobnost dat při dané hypotéze?)


bayesiánský přístup - příklad:

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

pravděpodobnosti jednotlivých výsledků

u pravých kostech stejné, u falešných se liší:

házíme 2× 1. hod:  2. hod: 

→ Jaká je pravděpodobnost, že naše kostka je falešná?

	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

Bayesovská analýza

aposteriorní pravděpodobnost (posterior probability)

= pr. platnosti hypotézy při získaných datech: $P(H | D)$

a.p. je funkcí věrohodnosti $P(D | H)$ a apriorní pravděpodobnosti (prior prob.)

prior vyjadřuje náš apriorní předpoklad nebo znalost







příklad se 2 hody kostkou:



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

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



suma čitateľů pro všechny alternativní hypotézy

apriorní pravděpodobnost (falešná) = 0.2
 (20/100 falešných kostek v souboru)

	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

Pr., že dostaneme   s pravou kostkou:

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

Pr. že dostaneme   s falešnou kostkou:

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

$$P(\text{biased} | \text{1-6}) = \frac{P(\text{1-6} | \text{biased}) \times P(\text{biased})}{P(\text{1-6} | \text{biased}) \times P(\text{biased}) + P(\text{1-6} | \text{fair}) \times P(\text{fair})}$$

$$= \frac{18/441 \times 2/10}{18/441 \times 2/10 + 1/36 \times 8/10} = \underline{0.269}$$

Bayesovská metoda ve fylogenetické analýze:

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

posterior \swarrow $P(\tau, \mathbf{v}, \theta | \mathbf{X})$ \nwarrow prior
 likelihood \downarrow $P(\mathbf{X} | \tau, \mathbf{v}, \theta)$
 $\sum_{i=1}^{B(s)}$ \square summing over all possible trees

Parametry pro bayesovskou analýzu: ML odhady \rightarrow **empirická BA**
 všechny kombinace \rightarrow **hierarchická BA**

$$P(\tau, \mathbf{v}, \theta) = \int P(\tau, \mathbf{v}, \theta) dF(\mathbf{v}, \theta)$$

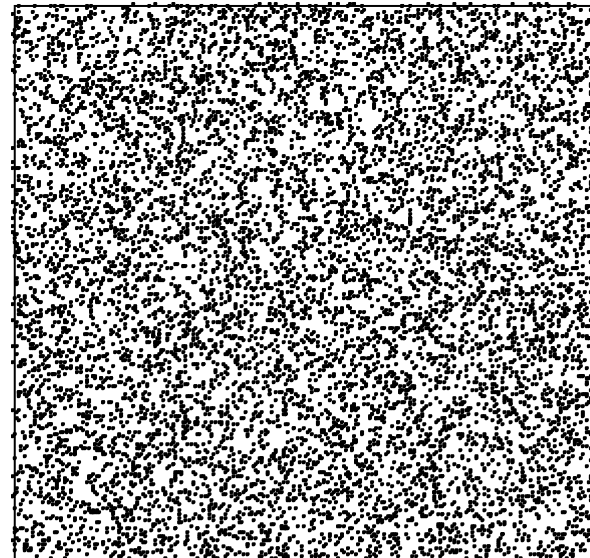
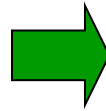
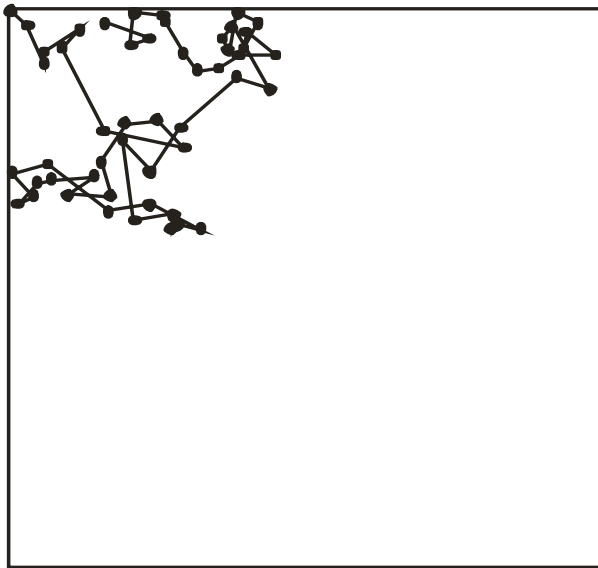
Problém: příliš složité \Rightarrow nelze řešit analyticky, pouze numericky aproximovat
řešení: **metody Monte Carlo**

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**

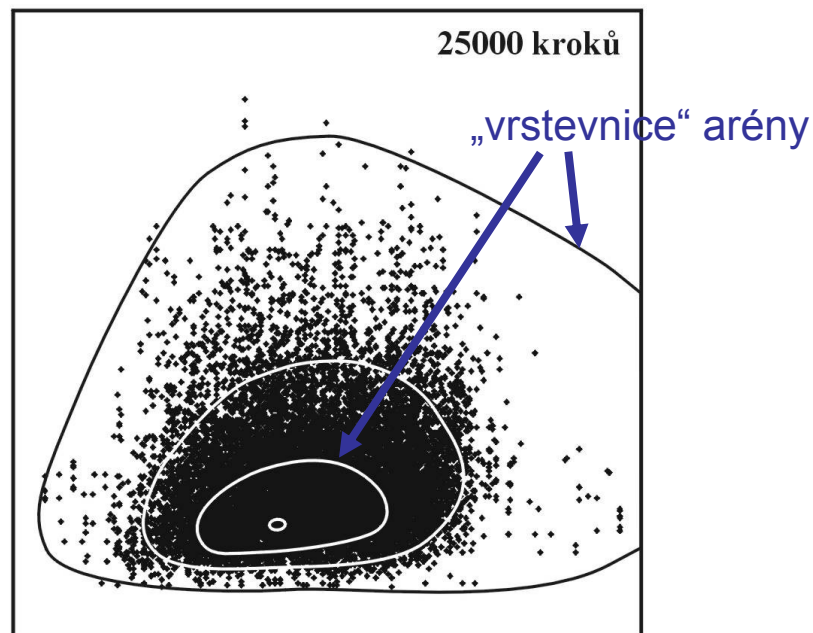
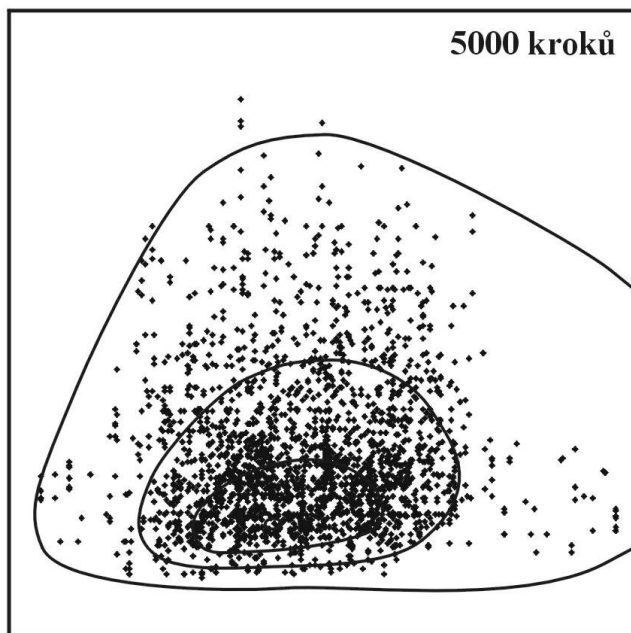


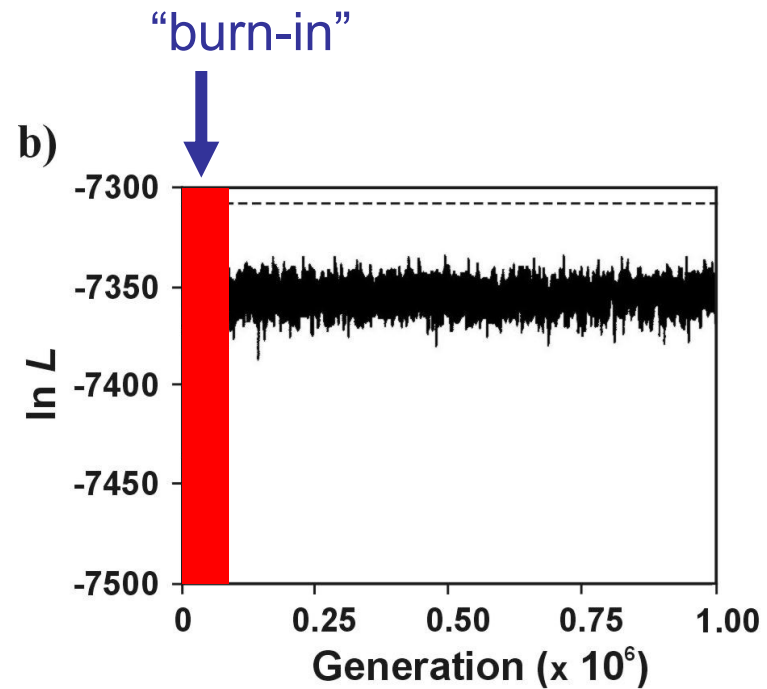
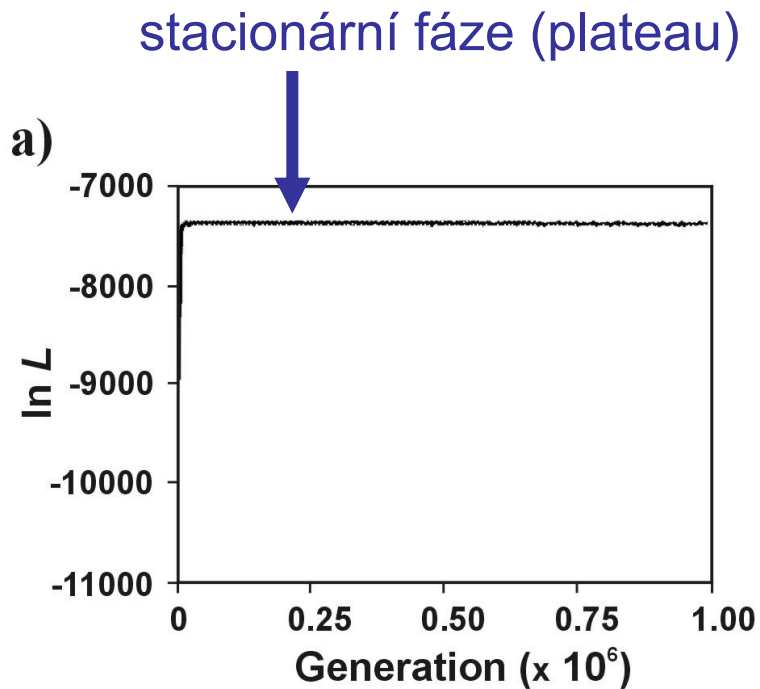
Metropolisův-Hastingsův algoritmus:

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:





MrBayes: <http://morphbank.ebc.uu.se/mrbayes/>
4 independent chains, Metropolis-coupled MCMC

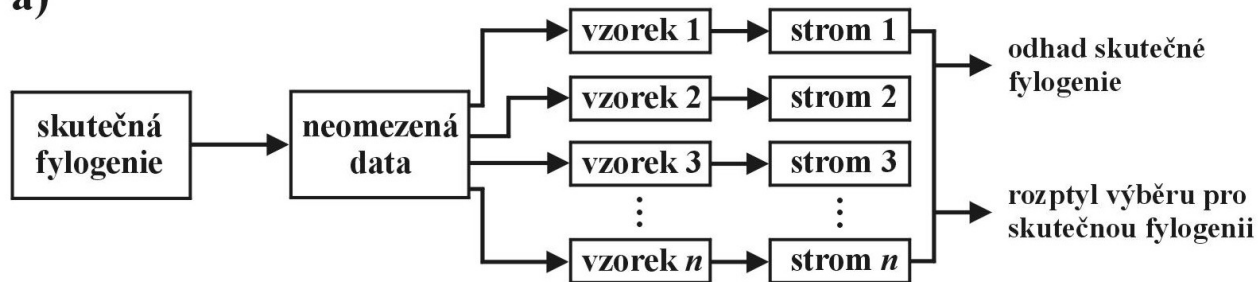
Problémy apriorních pravděpodobností!

Měření spolehlivosti stromů

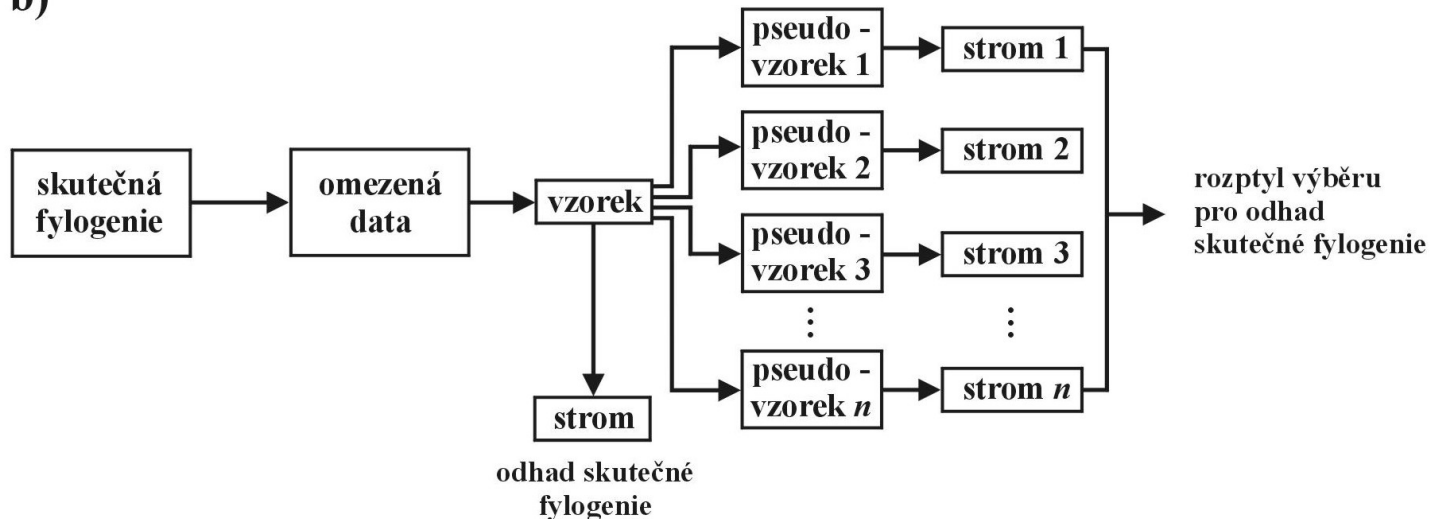
Metody opakovaného výběru

bez navrácení: jackknife
z navrácením: bootstrap

a)



b)



parametrický bootstrap: evoluční model
aposteriorní pravděpodobnosti

Je hierarchická struktura stromu reálná?

permutation tail probability test (PTP)
topology-dependent permutation tail prob. test (T-PTP)

Testování hypotéz

Testování modelů: LRT, Akaike, Bayes

Testy molekulárních hodin:

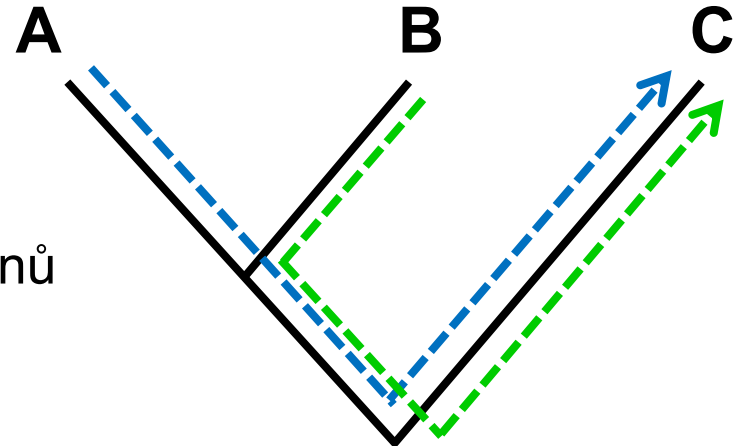
Relative rate test (RRT): $AC=BC?$

linearizované stromy

odstranění signifikantně odlišných taxonů

relaxované molekulární hodiny

umožňují změnu rychlostí podél větví



Srovnání stromů

Je jeden strom lepší než druhý?

Testy párových pozic:

winning sites test

Felsensteinův z test

Templetonův test

Kishinův-Hasegawův test (KHT, RELL)

Pro více než dva stromy:

Shimodairův-Hasegawův (SH) test

Jsou dva stromy signifikantně odlišné?

Distance mezi stromy:

partition metric

quartet metric

path difference metric

metody inkorporující délky větví

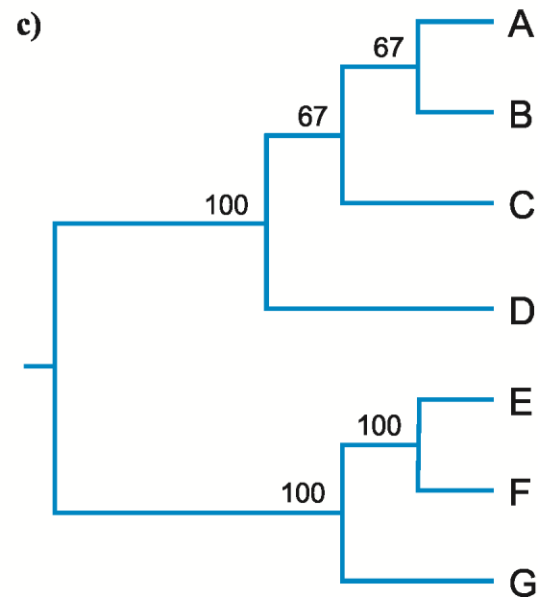
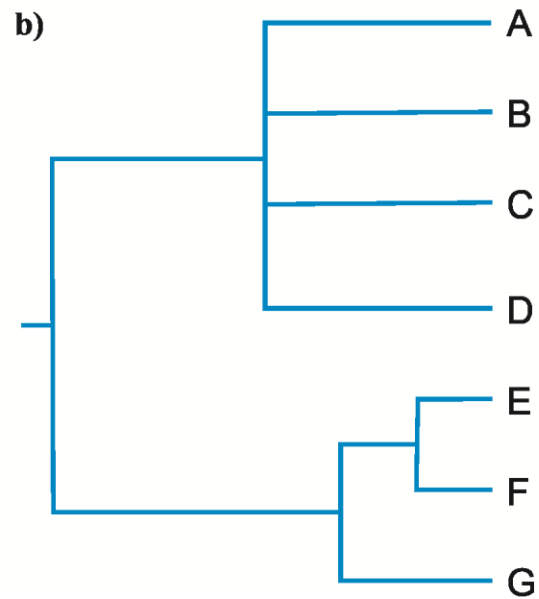
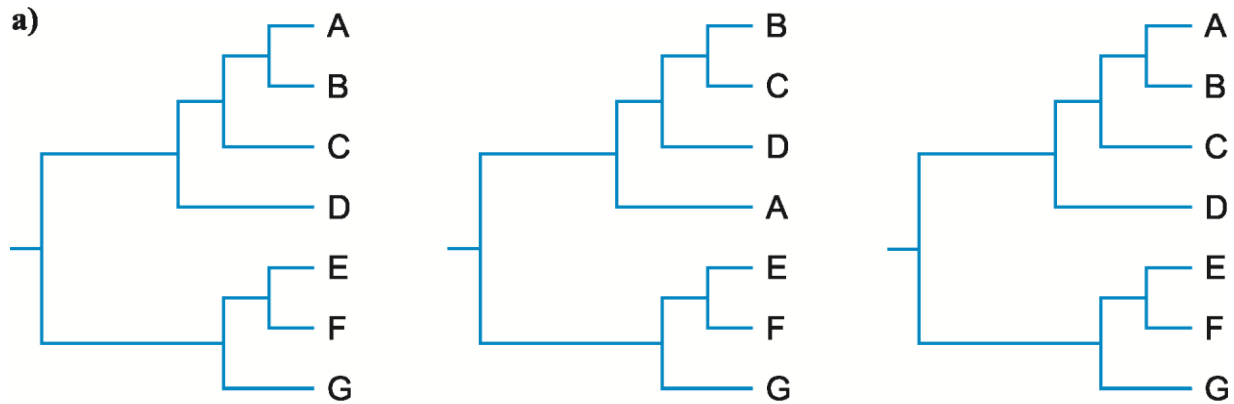
Problémy s distancemi mezi stromy!

Konsenzuální stromy

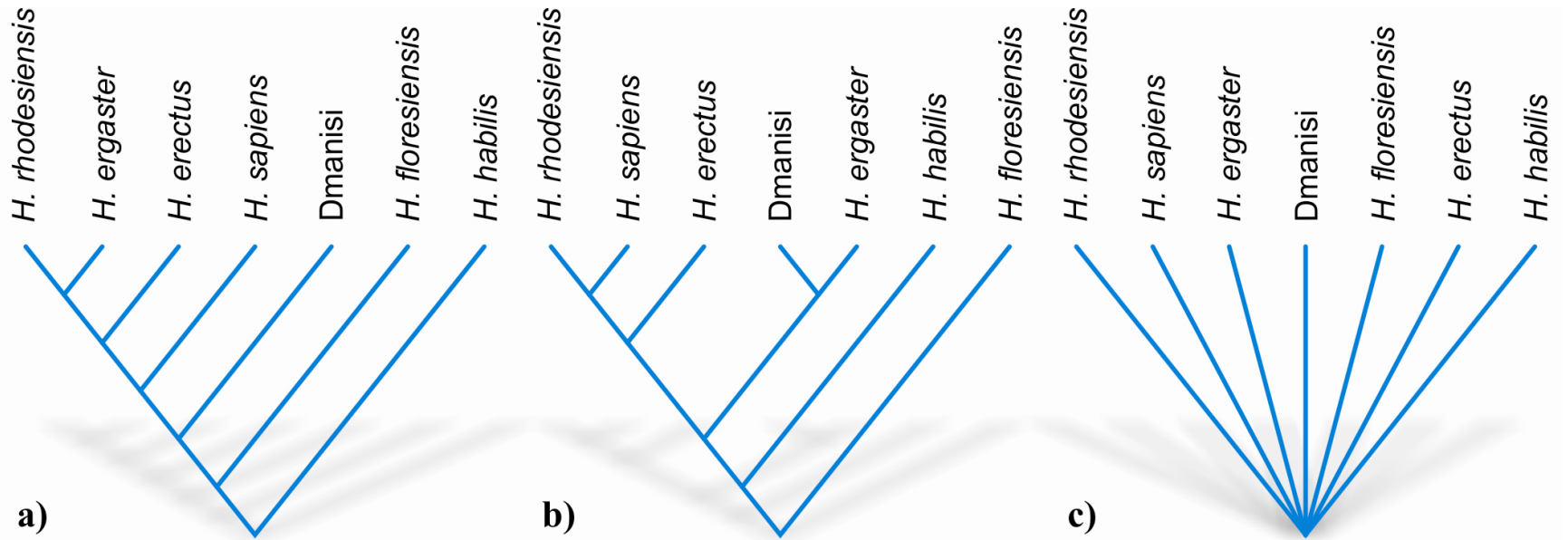
striktní konsensus
majority-rule

problém s konsenzuálními stromy – kombinovaná vs. separátní
analýza, supermatrix vs. supertree
konsenzuální stromy v metodách opakovaného výběru,
bayesovská analýza

Konsenzuální stromy



Konsenzuální stromy



Fylogenetické programy:

alignment:

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

PAUP*

PHYLIP

McClade ... MP

MOLPHY, TREE-PUZZLE ... ML

MrBayes ... BA

práce se stromy:

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