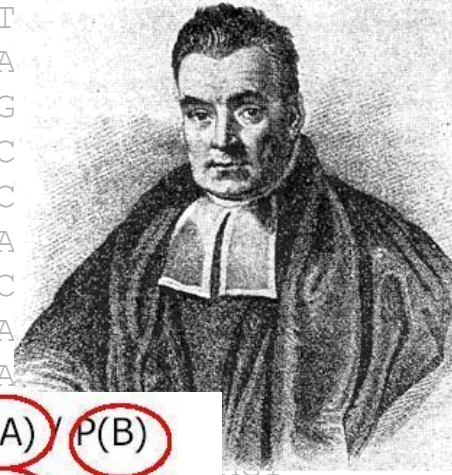
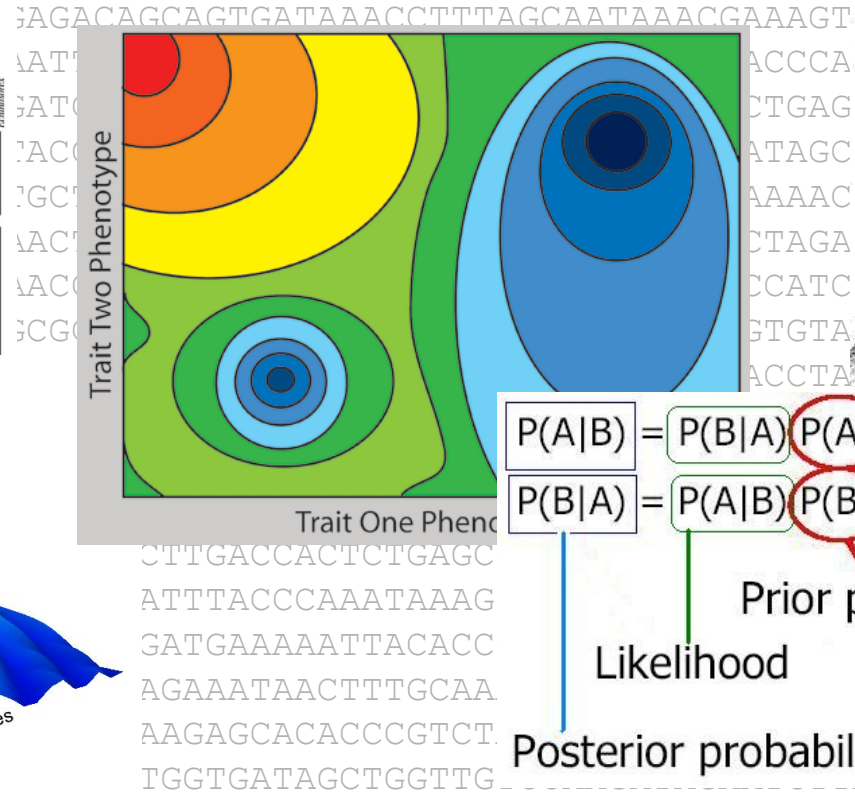
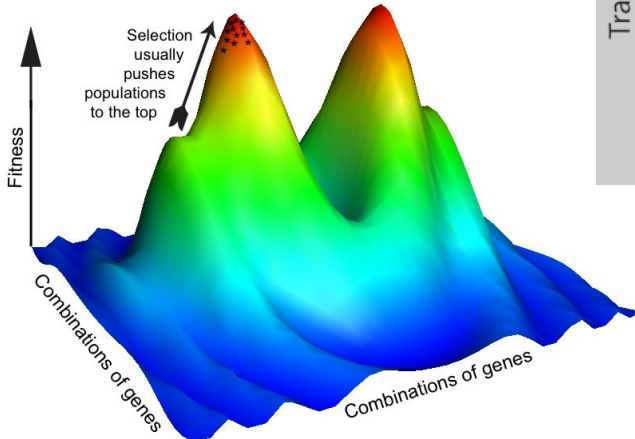
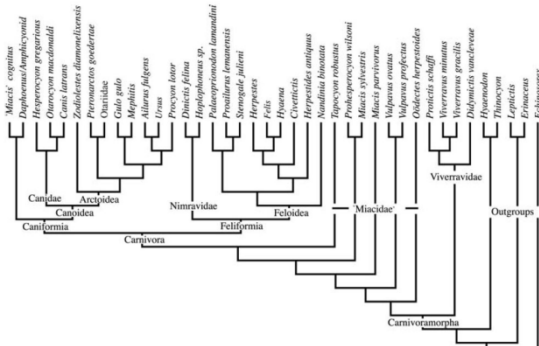


FYLOGENETICKÁ ANALÝZA II.



Maximální věrohodnost (maximum likelihood, ML)

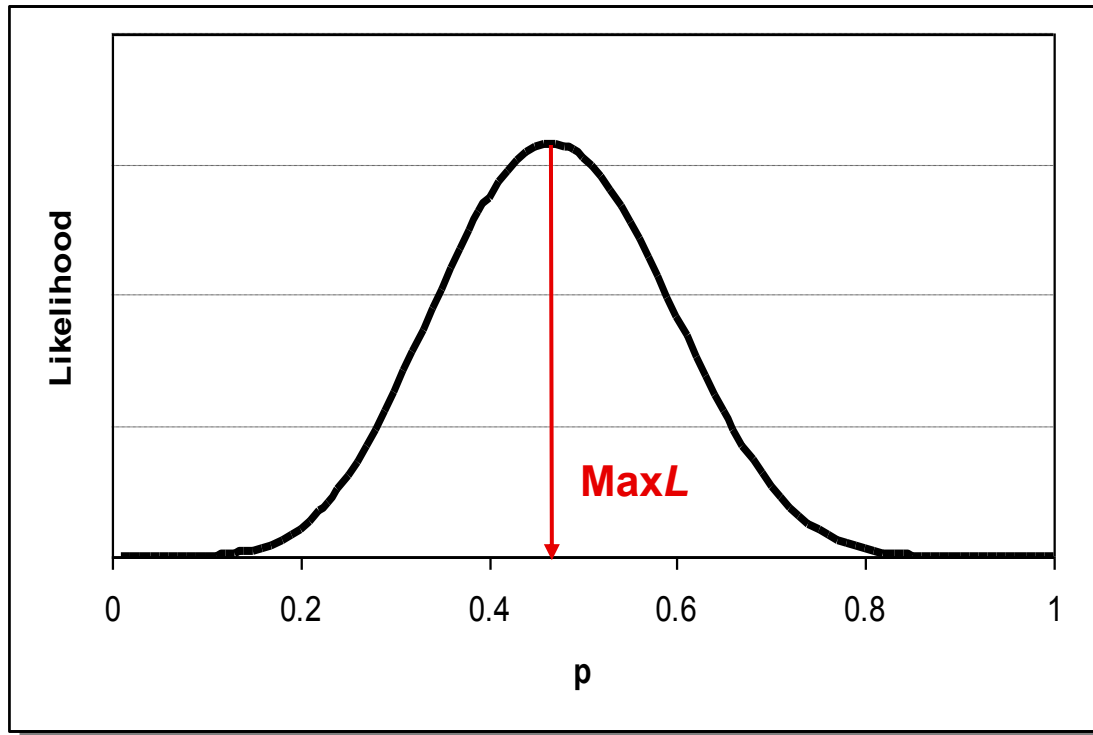
hod mincí 15× → skóre OOHHHOHOOOHOHHO: 7× panna (hlava, H), 8× orel (O)

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

hody nezávislé ⇒ 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 = $0,4666 \approx 7/15$



$$L = (D | H)$$

podmíněná pravděpodobnost
získání dat D při hypotéze H

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

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

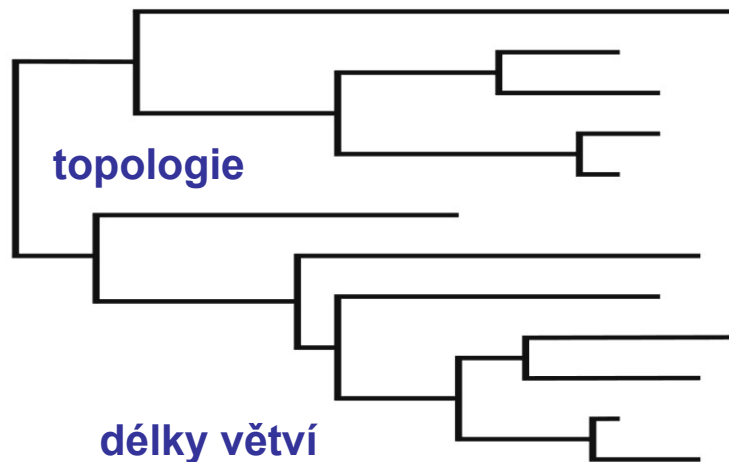
⇒ výsledek hodů 1,7×
pravděpodobnější
s pravou mincí

Maximální věrohodnost ve fylogenetické analýze

data:

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

strom:



evoluční model

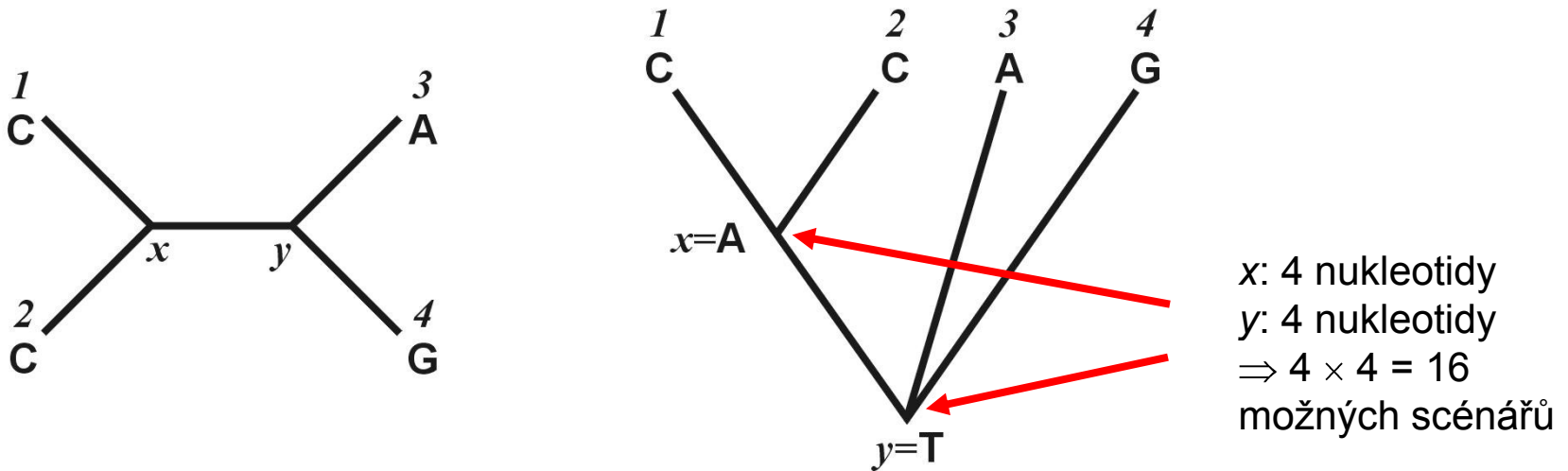
= hypotéza

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		<i>j</i>		<i>N</i>																																																
1	T	C	A	A	A	A	T	G	G	C	T	T	T	A	T	T	C	G	C	T	T	A	A	T	G	C	C	G	T	T	A	A	C	C	C	T	T	G	C	G	G	G	G	C	C	A	T	G					
2	T	C	C	G	T	G	A	T	G	G	A	T	T	T	A	T	T	T	C	C	G	C	A	A	T	G	C	C	T	G	T	C	A	T	C	T	T	A	T	T	C	T	C	A	A	G	T	A	T	C			
3	T	T	C	G	T	G	A	T	G	G	A	T	T	T	A	T	T	T	G	C	A	G	G	T	A	T	G	C	C	A	G	T	C	A	T	C	C	T	T	T	T	T	T	C	T	C	A	T	C	T	A	T	C
4	T	T	C	G	T	G	A	C	G	G	G	T	T	A	T	C	T	C	G	G	C	A	A	T	G	C	C	G	G	T	C	A	T	C	C	T	A	T	T	T	T	C	G	A	G	T	A	T	T				

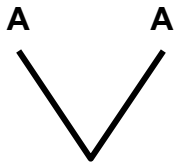


1) $L(1) = P(A) \times P(T) \times P(AC) \times P(AC) \times P(TA) \times P(TG)$

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

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

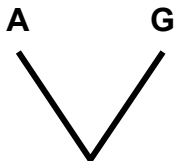
4) $\ln L = \ln L(1) + \ln L(2) + \dots + \ln L(N) = \sum_{j=1}^N \ln L_j$



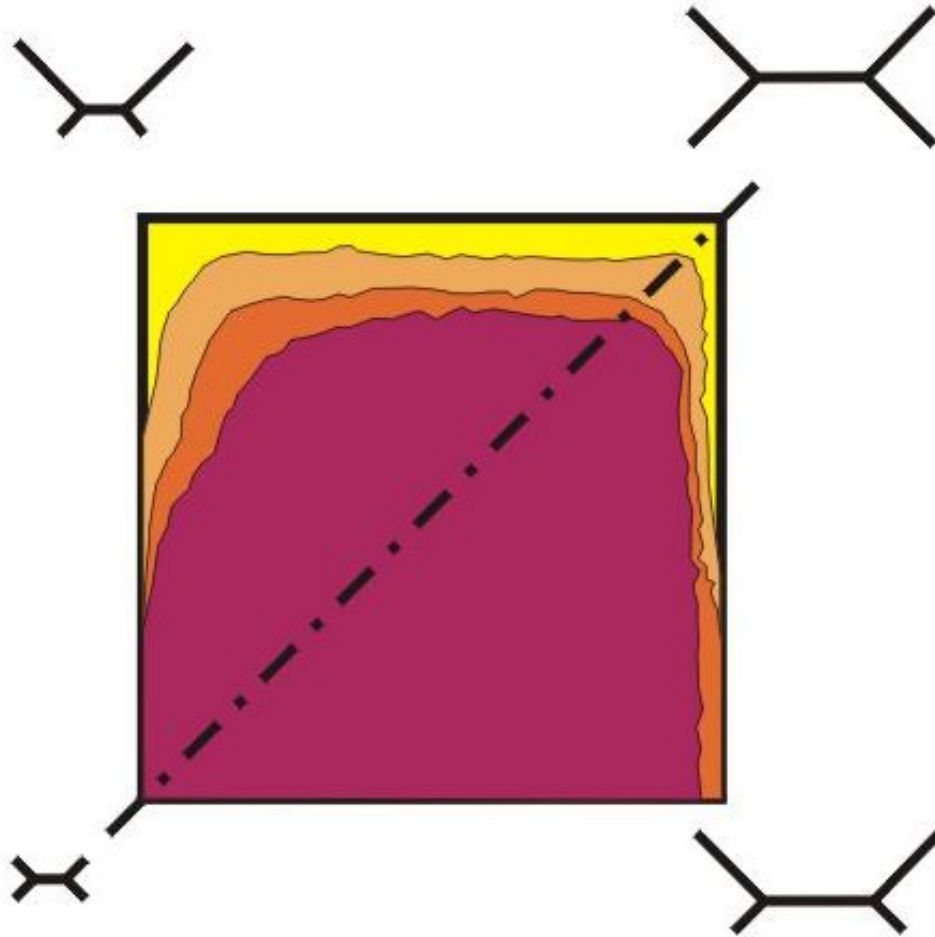
Věrohodnost (ML) a úspornost (MP)

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

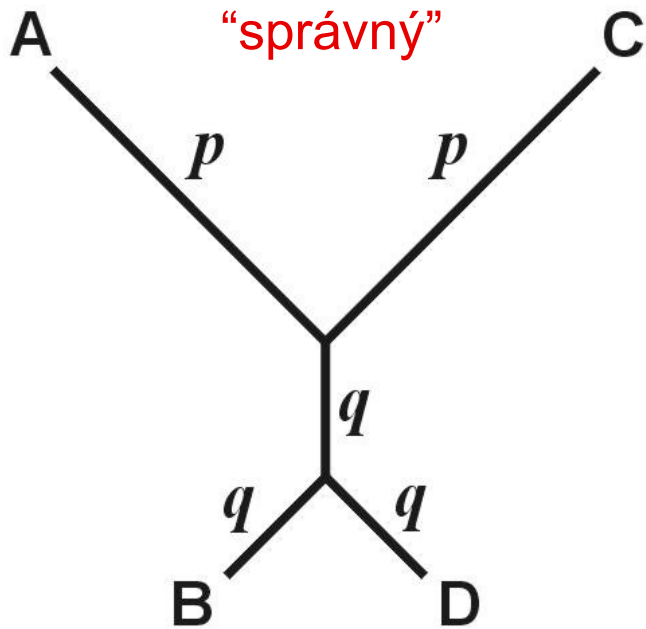
Počet změn	Parsimonie	$\nu = 0,01$	$\nu = 0,10$	$\nu = 0,20$	$\nu = 1,00$
		(0,00083)	(0,00786)	(0,01462)	(0,04602)
0	0	0,00	0,00	0,00	0,00
1	100	99,66	96,64	92,36	66,54
2	0	0,33	3,22	6,22	21,19
3	0		0,12	0,48	8,61
4	0		0,003	0,023	2,05
5	0			0,0037	0,42



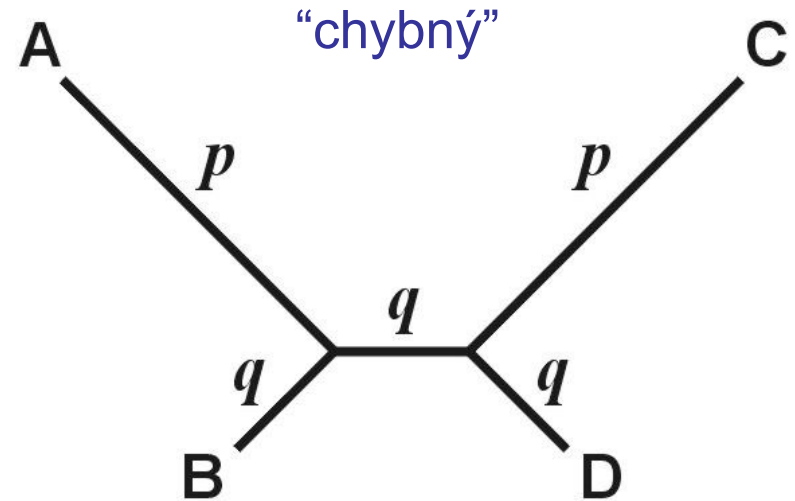
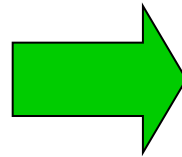
Věrohodnost a konzistence



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

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:







The diagram shows the Bayes' theorem equation with three callout boxes. The first box, labeled 'věrohodnost', points to the likelihood term $P(D | H)$ in the numerator. The second box, labeled 'apriorní pravděpodobnost', points to the prior probability term $P(H)$ in the numerator. The third box, labeled 'suma čitateľů pro všechny alternativní hypotézy', points to the denominator $\Sigma[P(D | H_i) \times P(H_i)]$.

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

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

$$P(\text{biased} | \text{1 dot, 3 dots}) = \frac{P(\text{1 dot, 3 dots} | \text{biased}) \times P(\text{biased})}{P(\text{1 dot, 3 dots} | \text{biased}) \times P(\text{biased}) + P(\text{1 dot, 3 dots} | \text{fair}) \times P(\text{fair})}$$

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

Bayesovská metoda ve fylogenetické analýze:

aposteriorní
pravděpodobnost

věrohodnost

apriorní
pravděpodobnost

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

suma přes všechny
možné stromy

Parametry pro bayesovskou analýzu: ML odhady → **empirická BA**

všechny kombinace → **hierarchická BA**

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

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

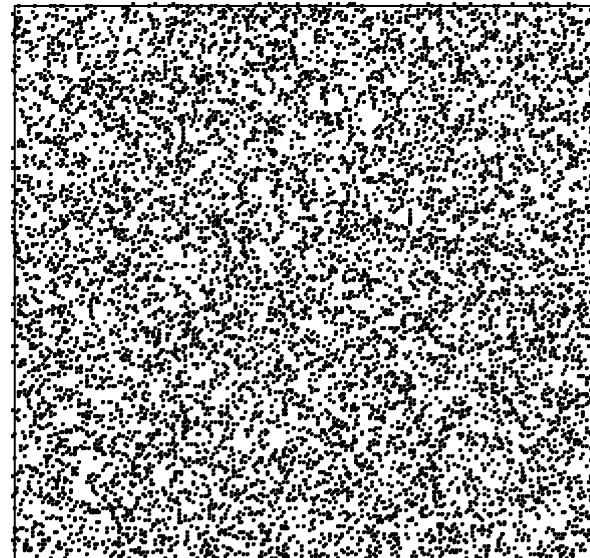
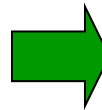
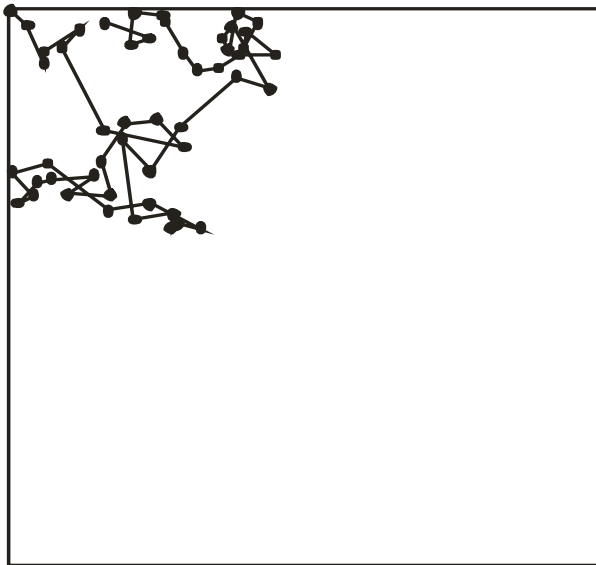
řešení: metody Monte Carlo

náhodný výběr vzorků, při velkém množství aproximace skutečnosti

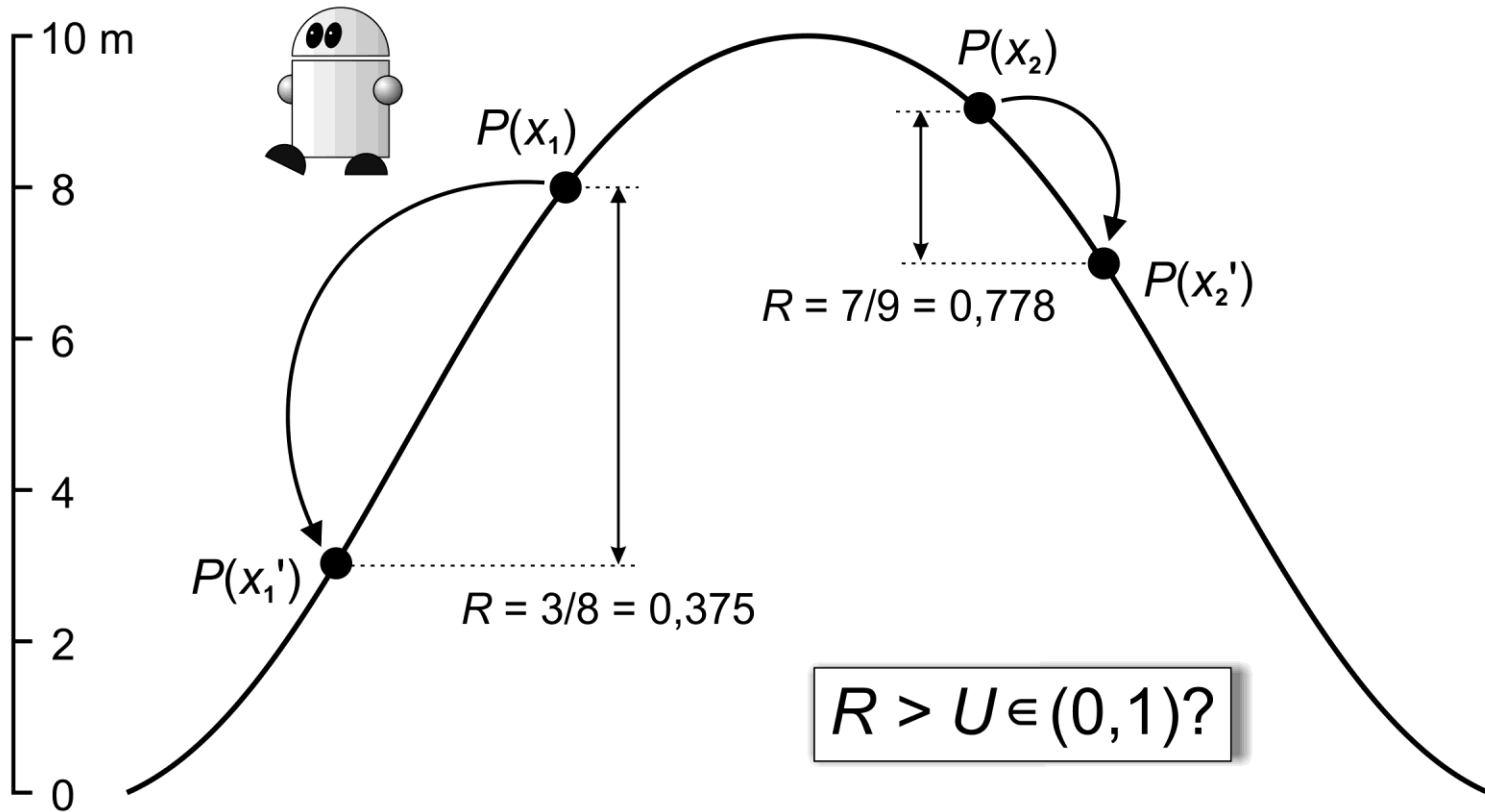
Markovovy řetězce: Markov chain Monte Carlo (MCMC)

Markovův proces: $t(-1) A \rightarrow T(0) C \rightarrow T(+1) G$

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



Metropolisův-Hastingsův algoritmus:

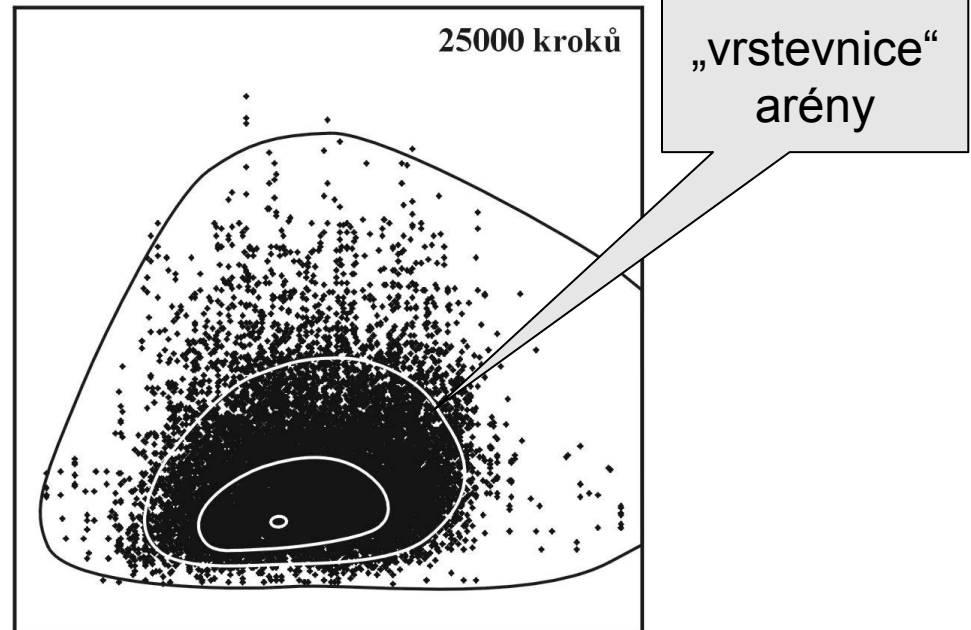
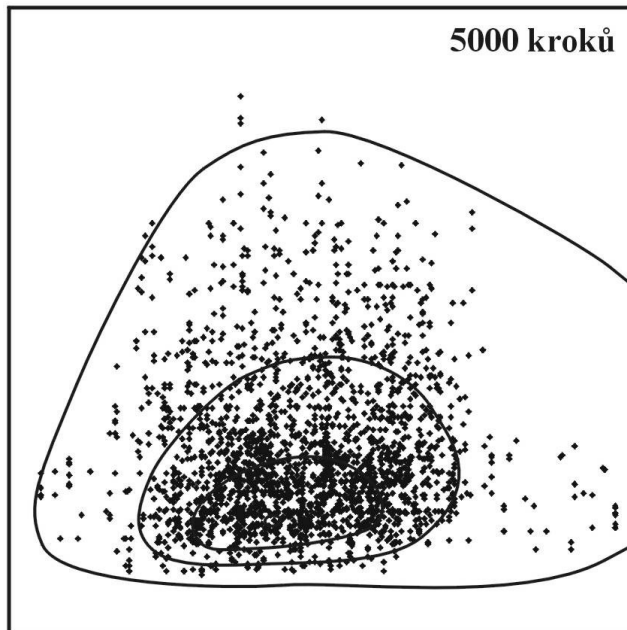


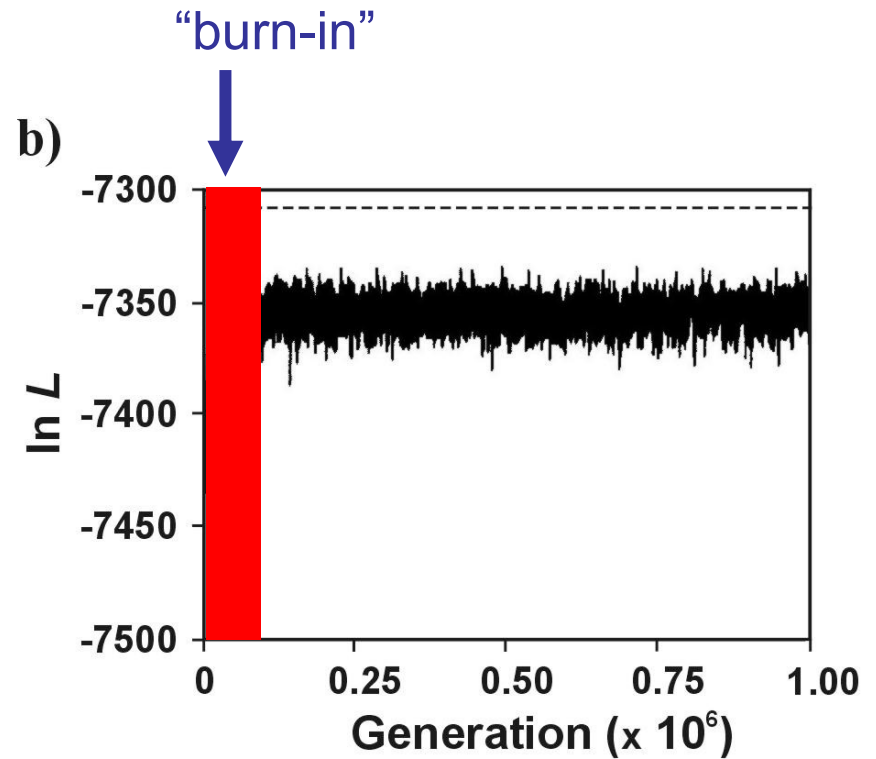
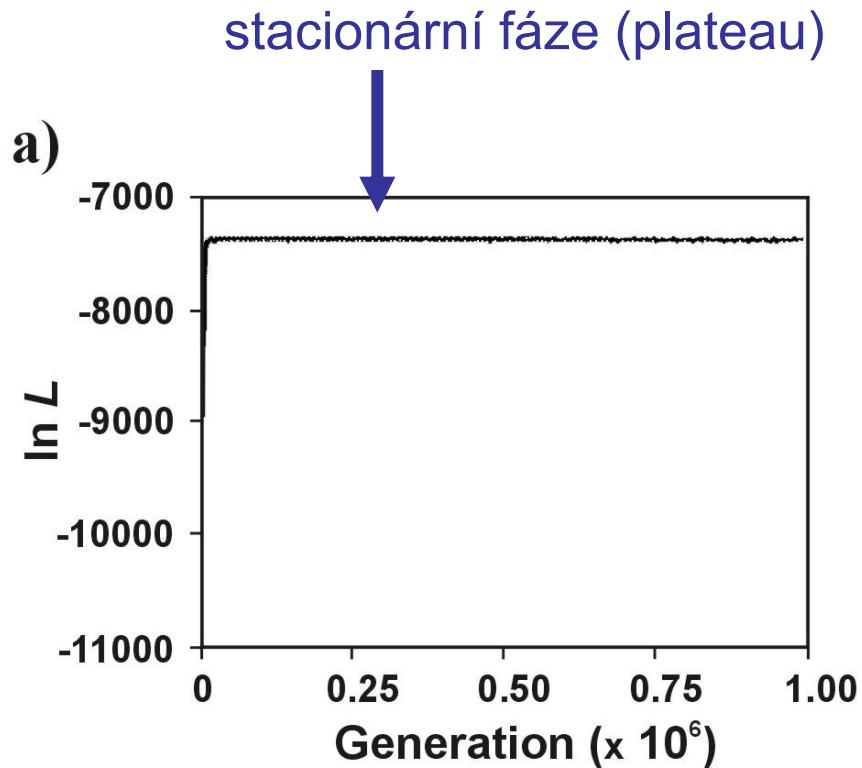
Metropolisův-Hastingsův algoritmus:

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:

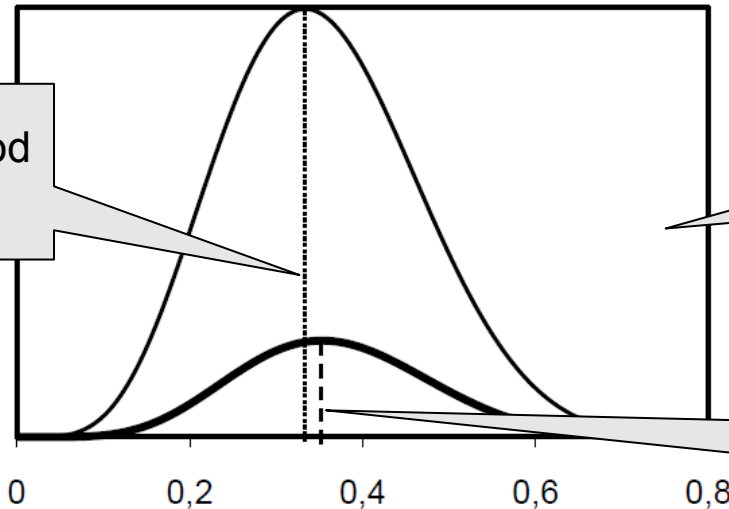




MrBayes: <http://morphbank.ebc.uu.se/mrbayes/>
4 independent chains, Metropolis-coupled MCMC

Problém apriorních pravděpodobností: subjektivnost

a)



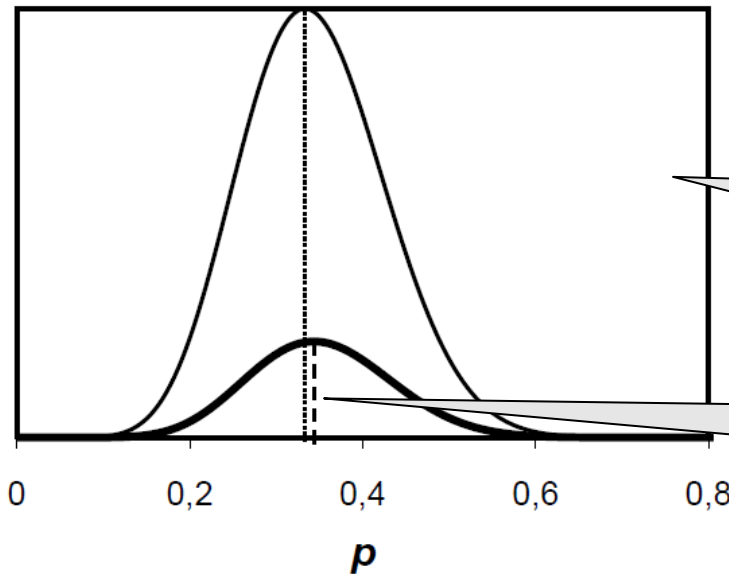
apriorní pr. = 0,5

maximum likelihood
= 0,333

15 hodů mincí
výsledek 5 H : 10 O

díky apriorní pr.
aposteriorní pr.
posunuta doprava

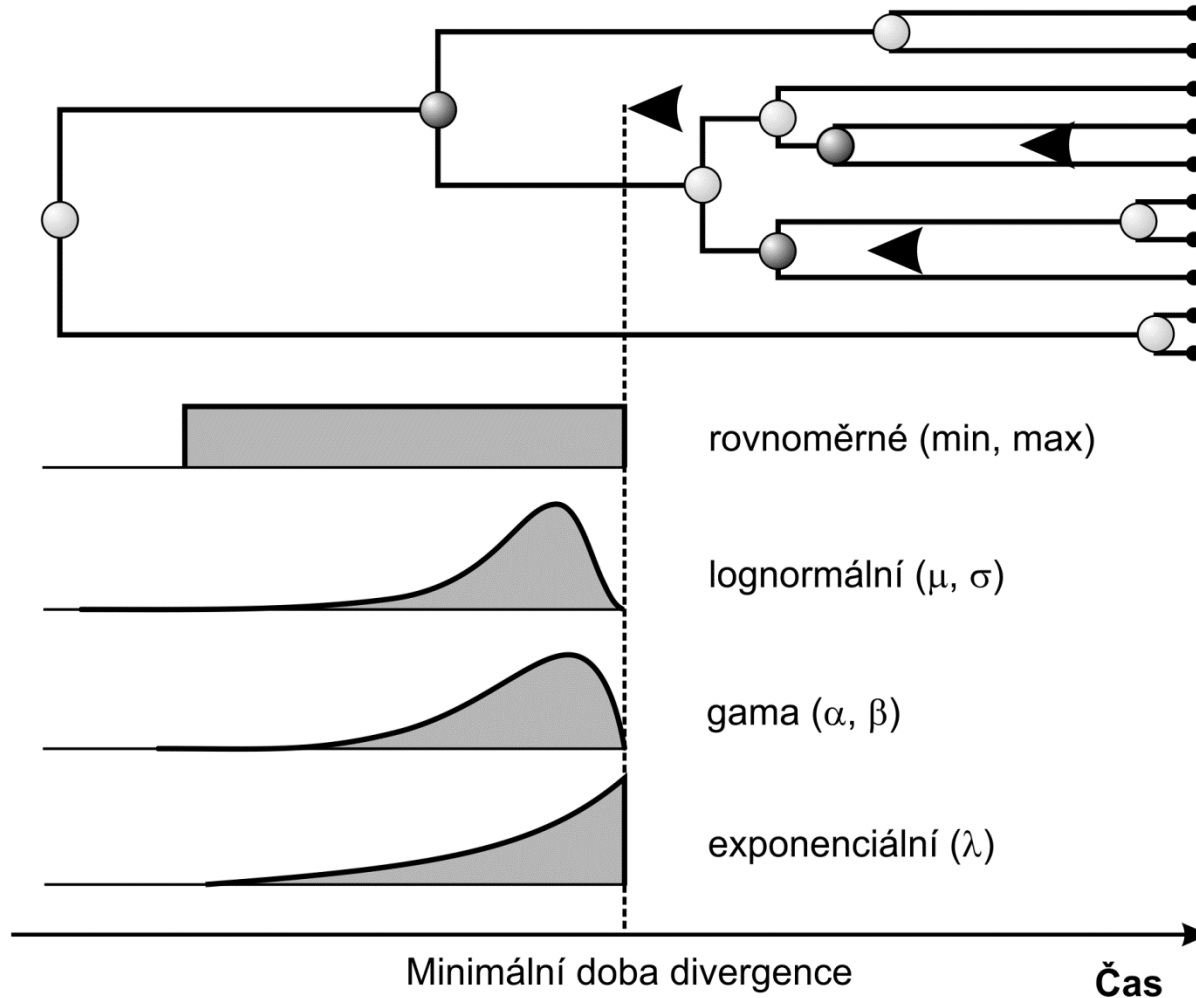
b)



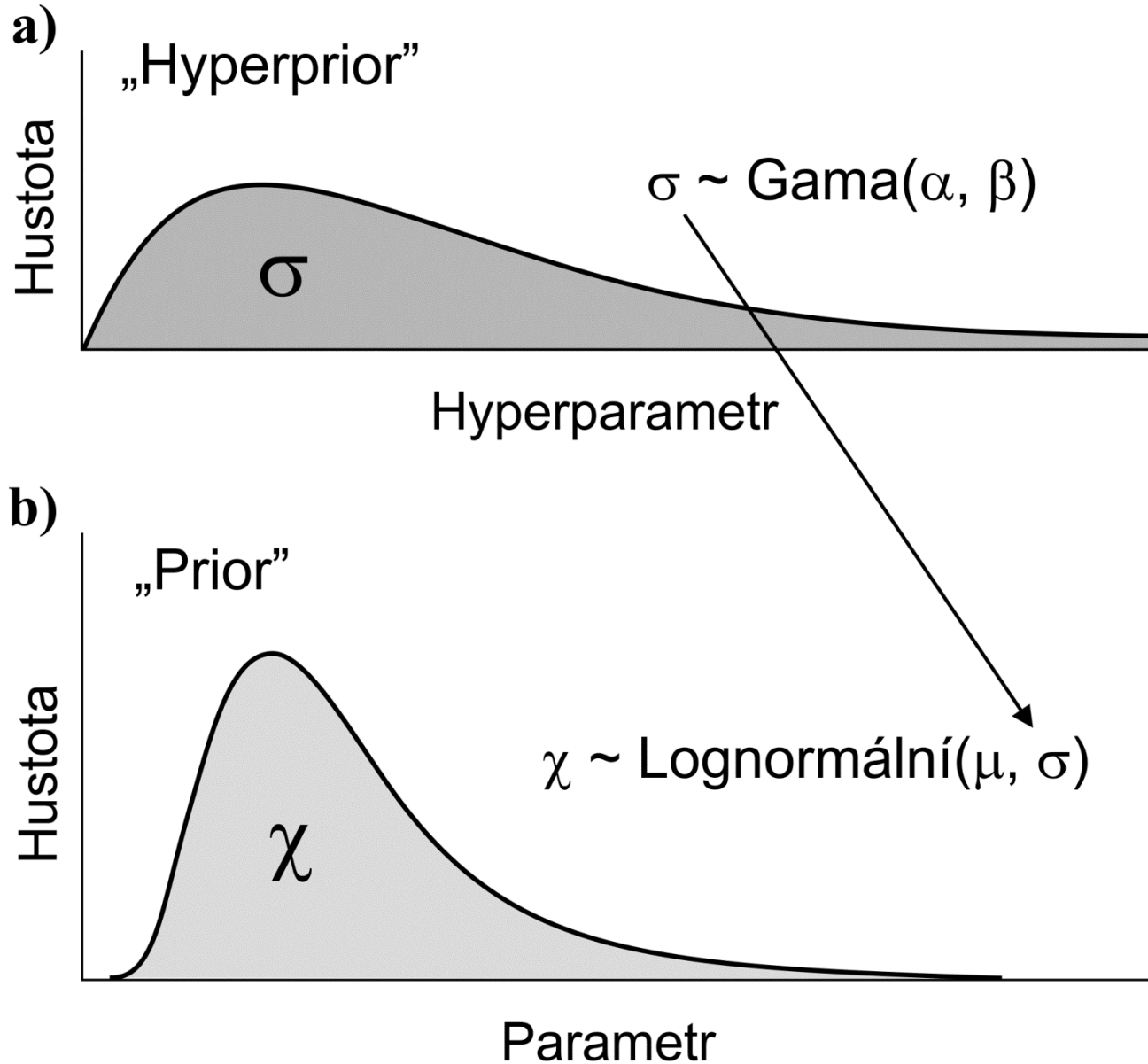
30 hodů mincí
výsledek 10 H : 20 O

rozdíl od ML
menší

Problém apriorních pravděpodobností: subjektivnost

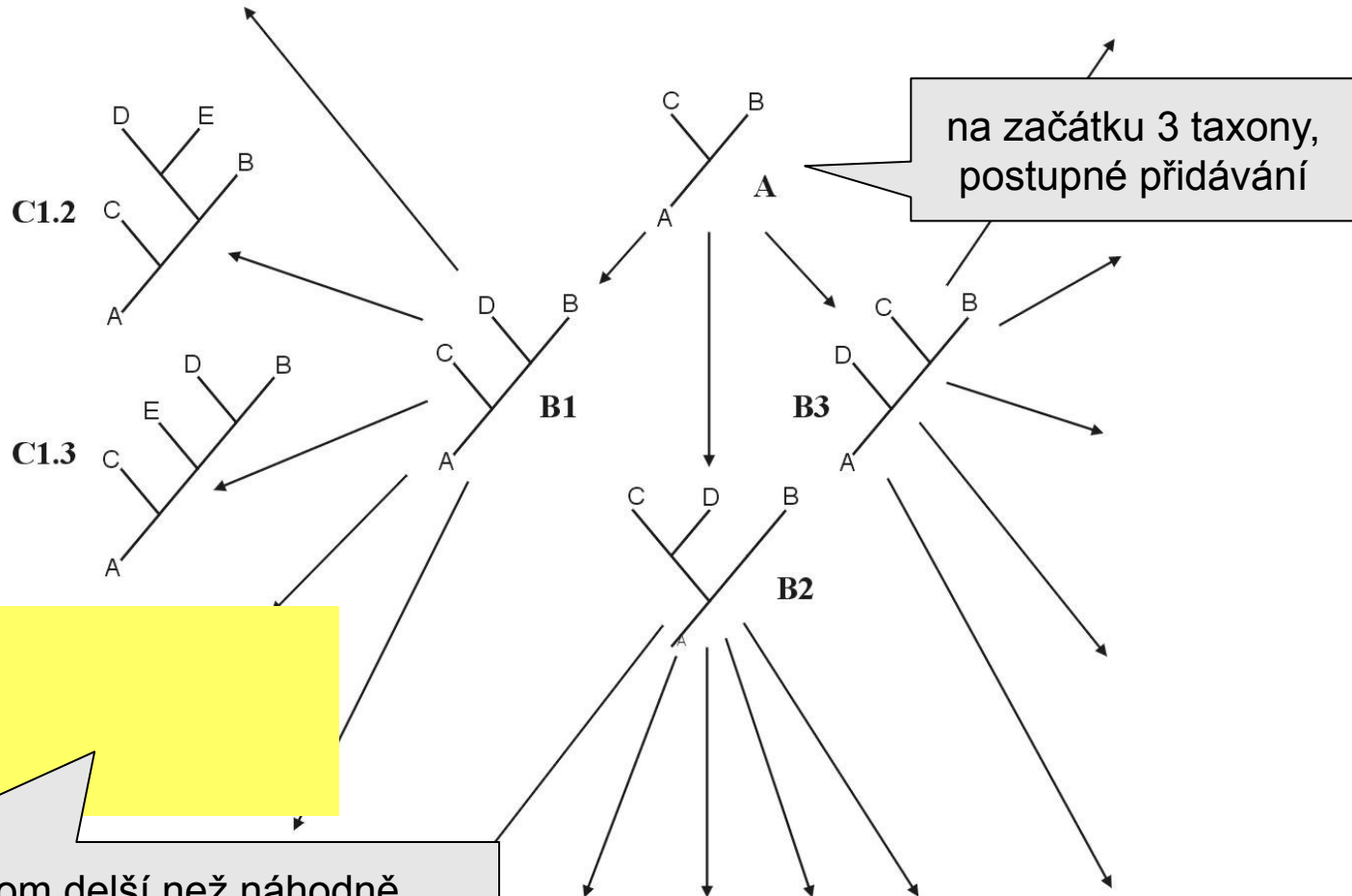


Problém apriorních pravděpodobností: subjektivnost



Hledání optimálního stromu a měření spolehlivosti

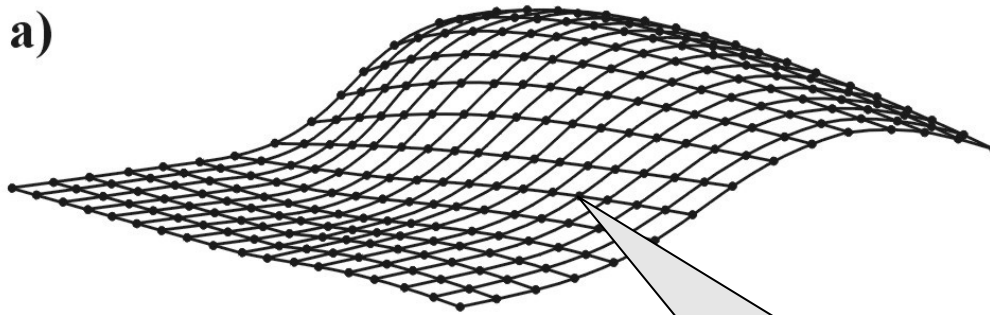
1. **Exaktní metody:** a) vyčerpávající hledání (exhaustive search)
b) branch-and-bound



je-li strom delší než náhodně vybraný, algoritmus dál nepokračuje

2. Heuristický přístup:

a)

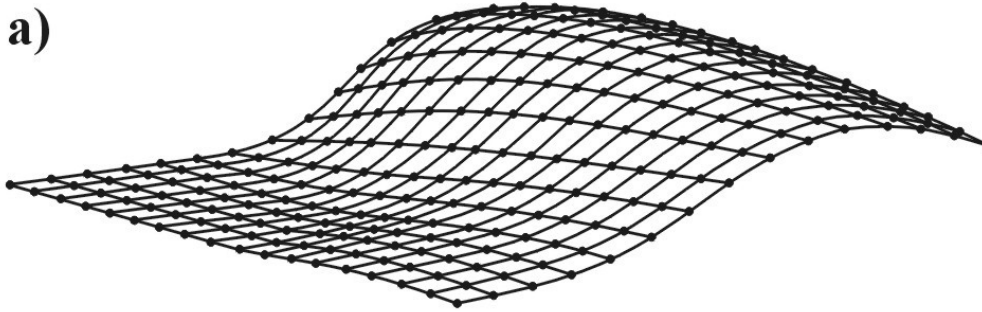


všechny možné stromy

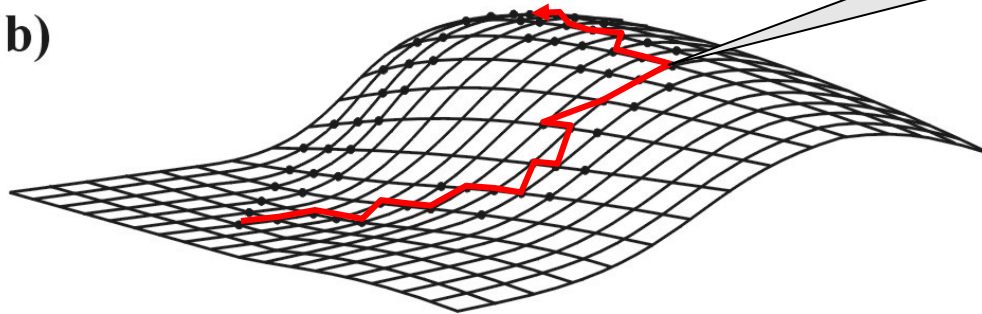
2. Heuristický přístup:

stepwise addition
star decomposition
branch swapping

a)

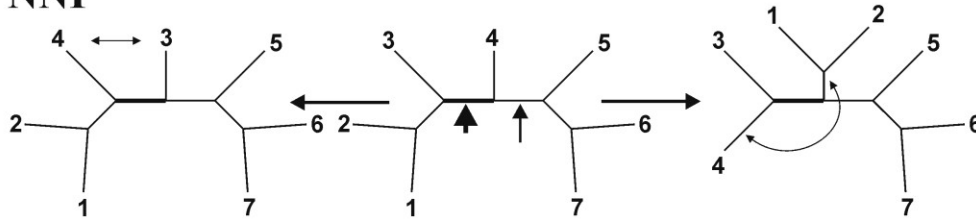


b)

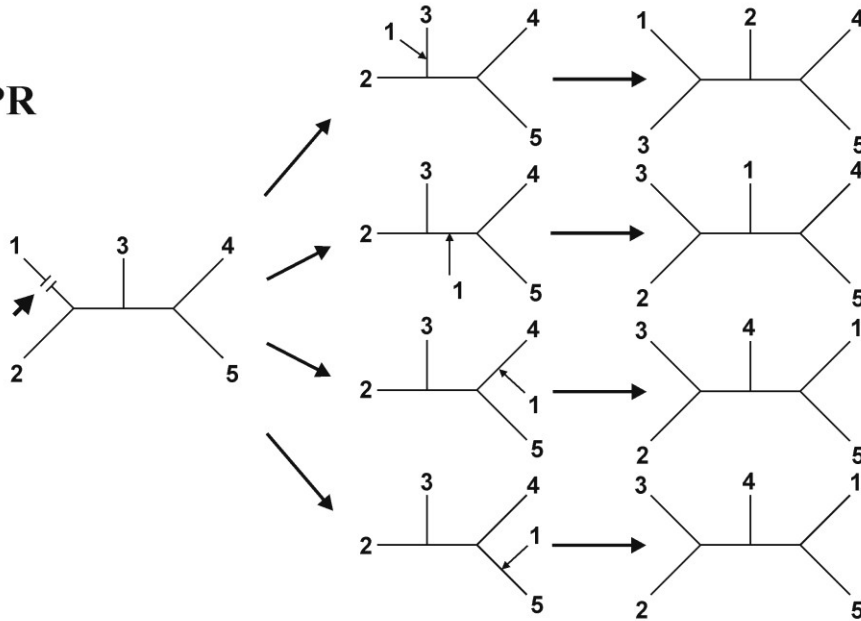


heuristické hledání

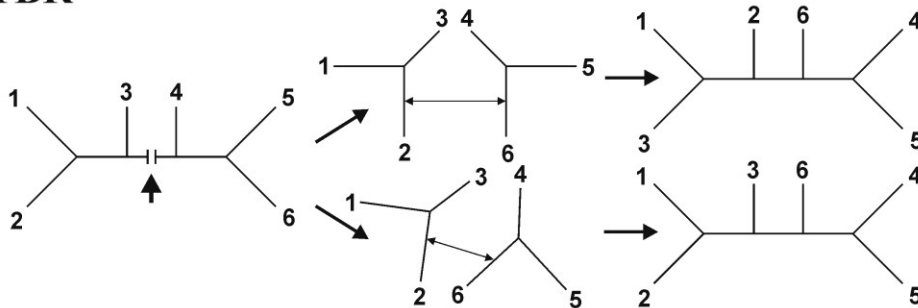
NNI



SPR



TBR



* nearest-neighbor interchanges (NNI)

* subtree pruning and regrafting (SPR)

* tree bisection and reconnection (TBR)

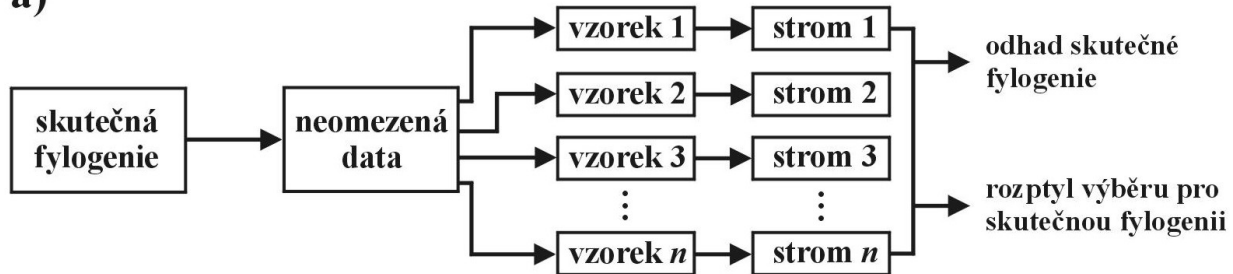
Měření spolehlivosti stromů

Metody opakovaného výběru

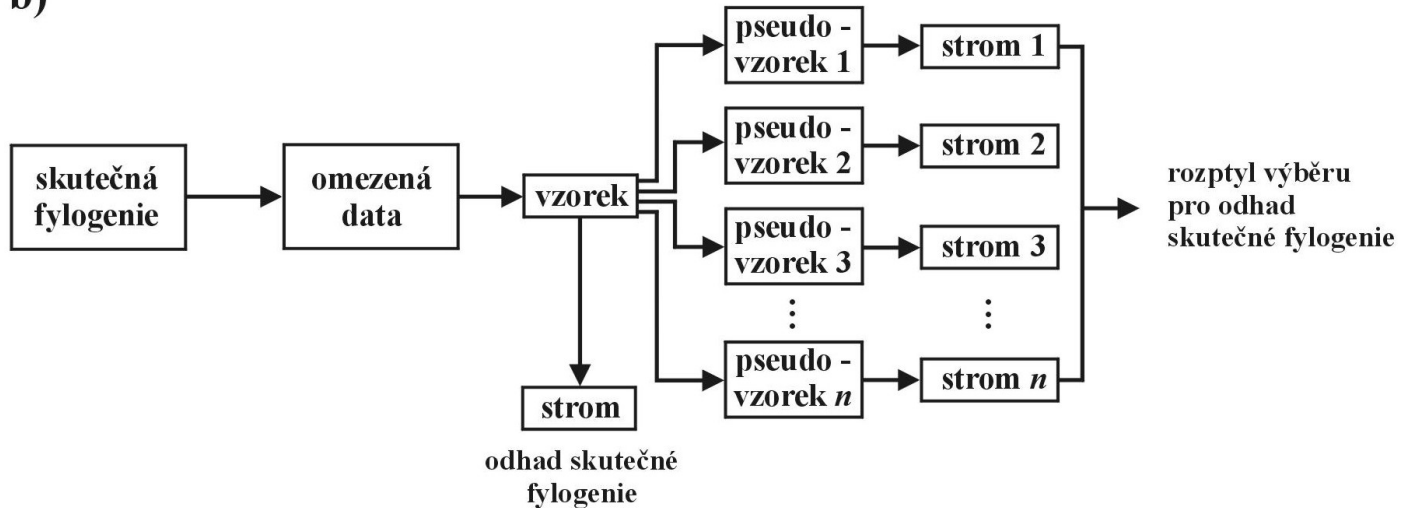
bez navrácení = jackknife

s navrácením = bootstrap

a)

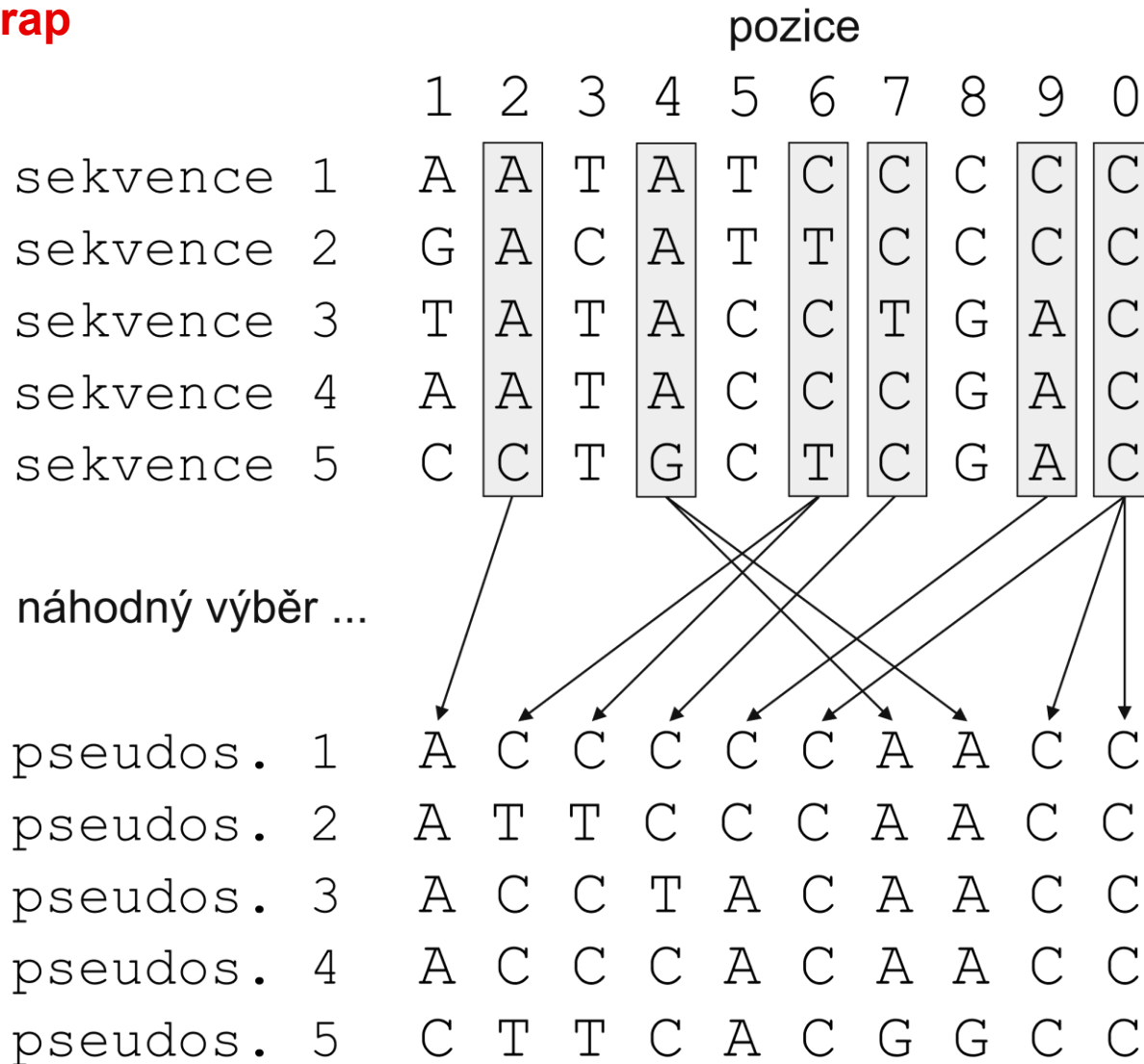


b)



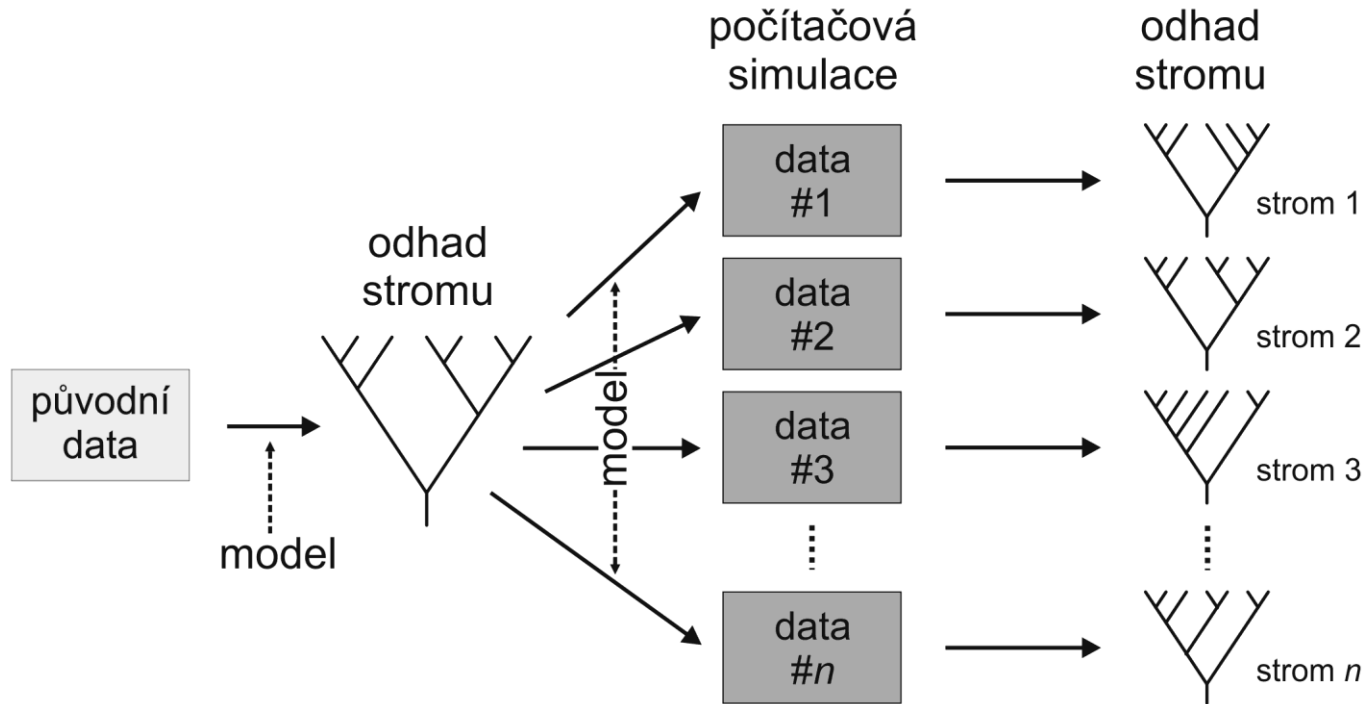
Měření spolehlivosti stromů

Bootstrap



Měření spolehlivosti stromů

parametrický bootstrap: evoluční model



bayesovská analýza: aposteriorní pravděpodobnosti

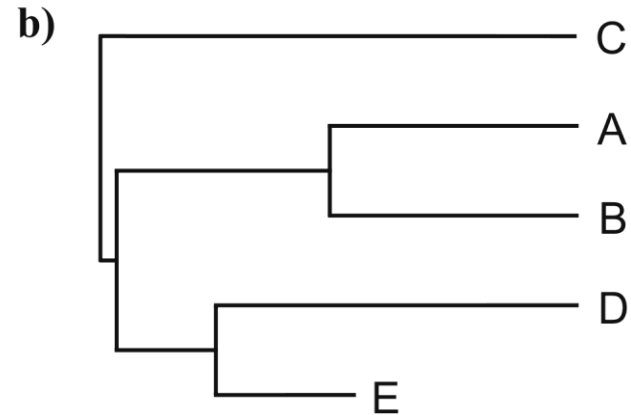
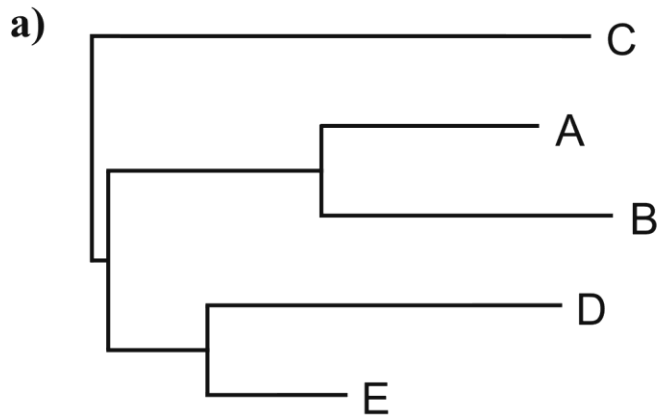
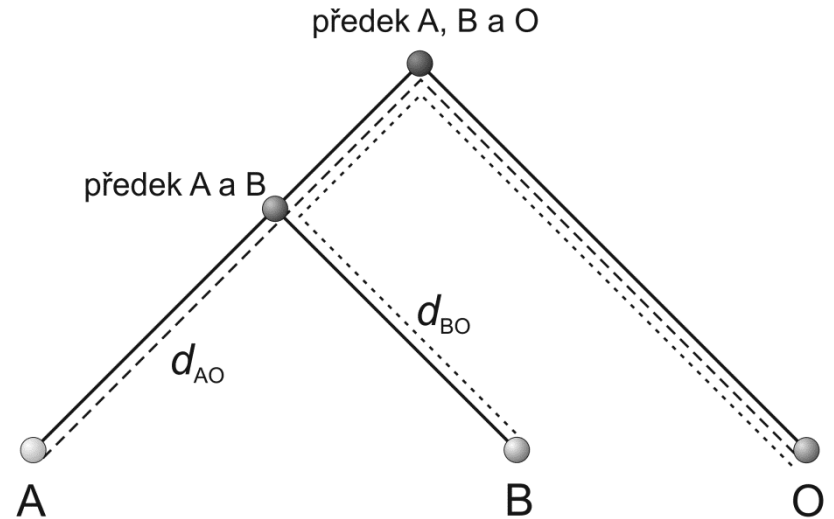
Testování hypotéz

Test molekulárních hodin

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

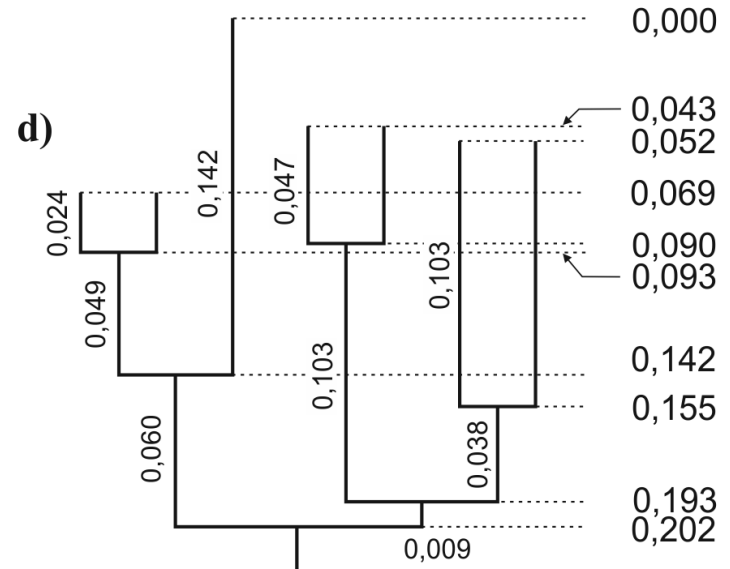
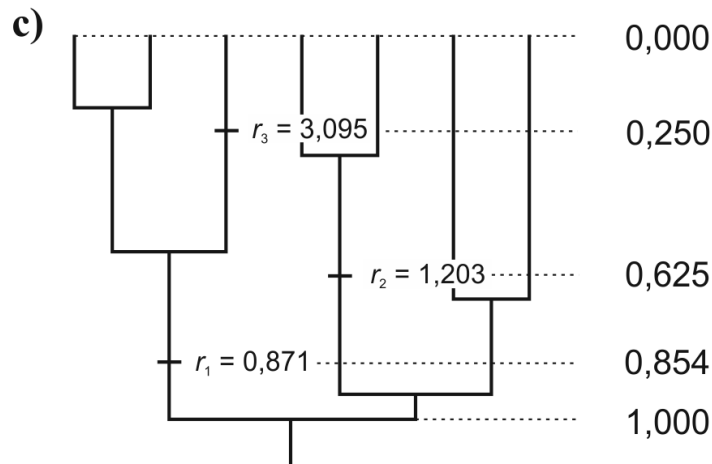
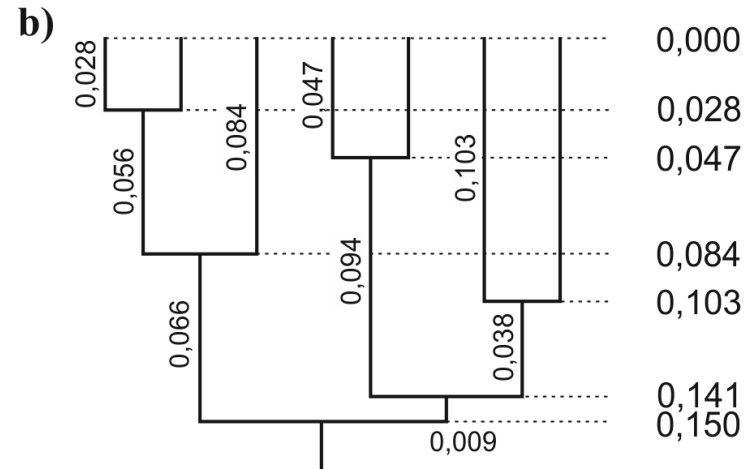
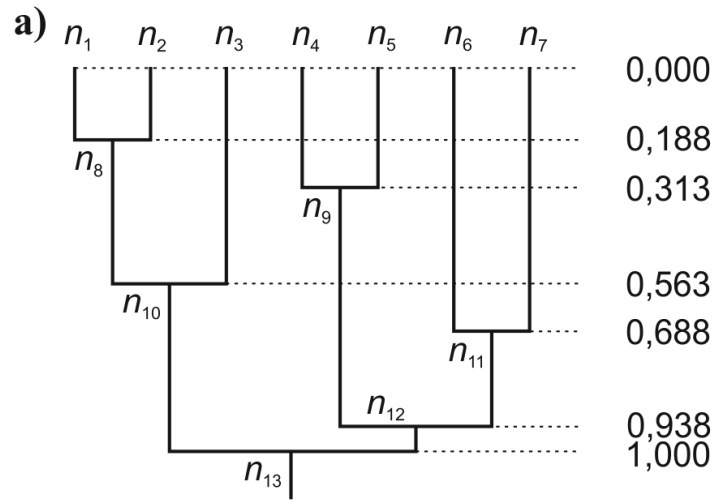
Linearizované stromy

odstranění signifikantně odlišných taxonů



Relaxované molekulární hodiny

umožňují změnu rychlostí podél větví



Srovnání stromů

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

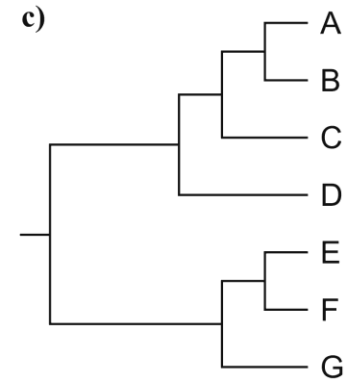
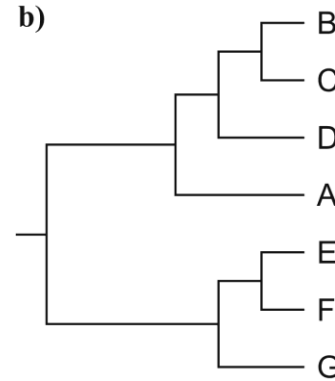
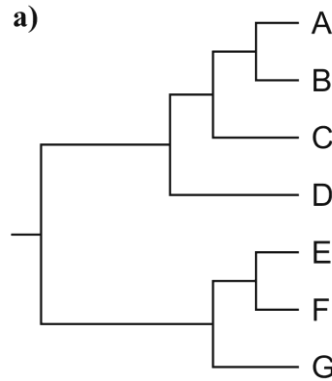
metody inkorporující délky větví

Problémy s distancemi mezi stromy

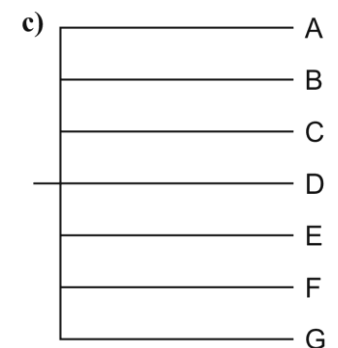
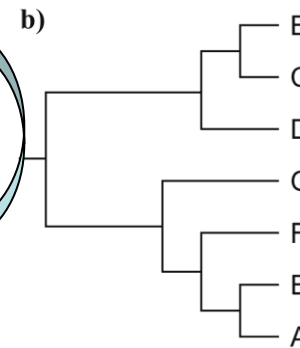
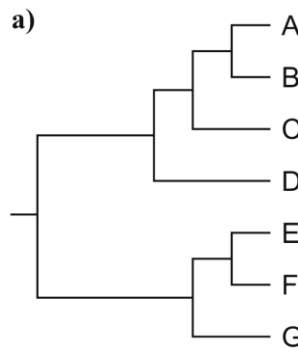
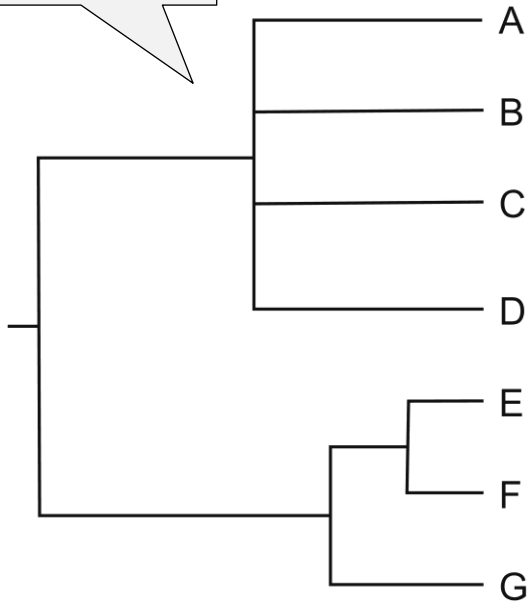
Konsenzuální stromy

zdrojové stromy

strikní konsensus



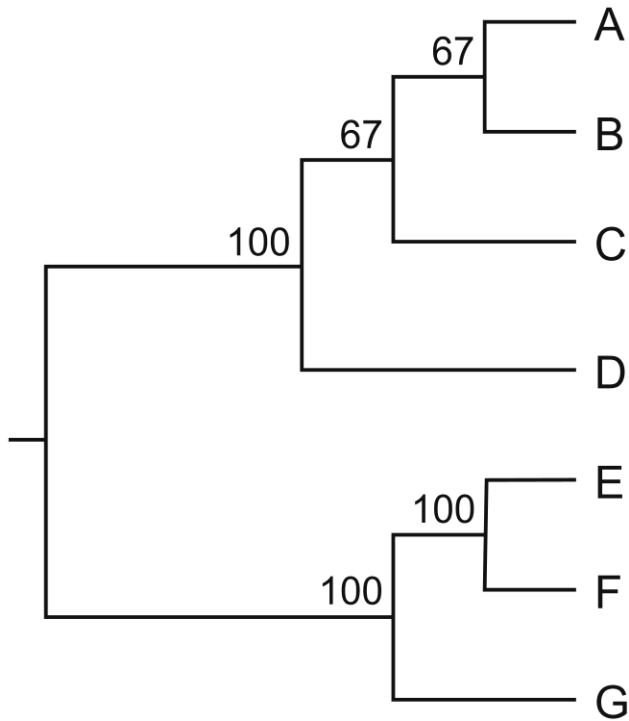
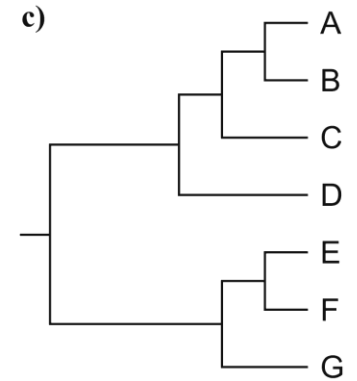
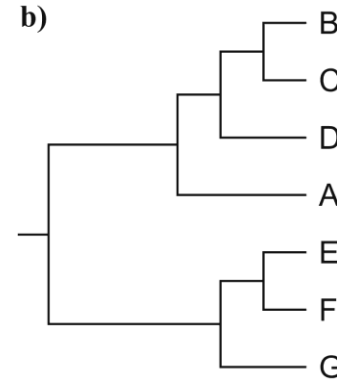
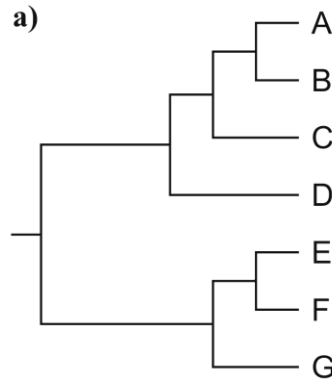
strikně konsenzuální strom



Konsensuální stromy

zdrojové stromy

majority-rule



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>