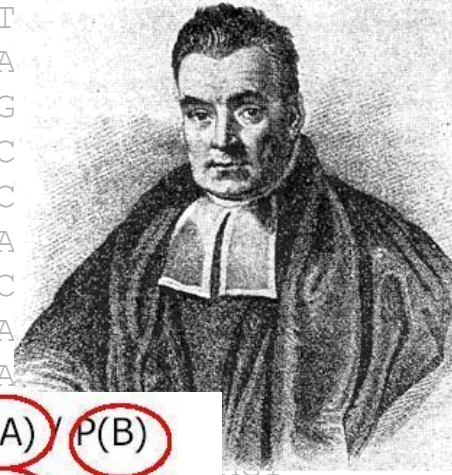
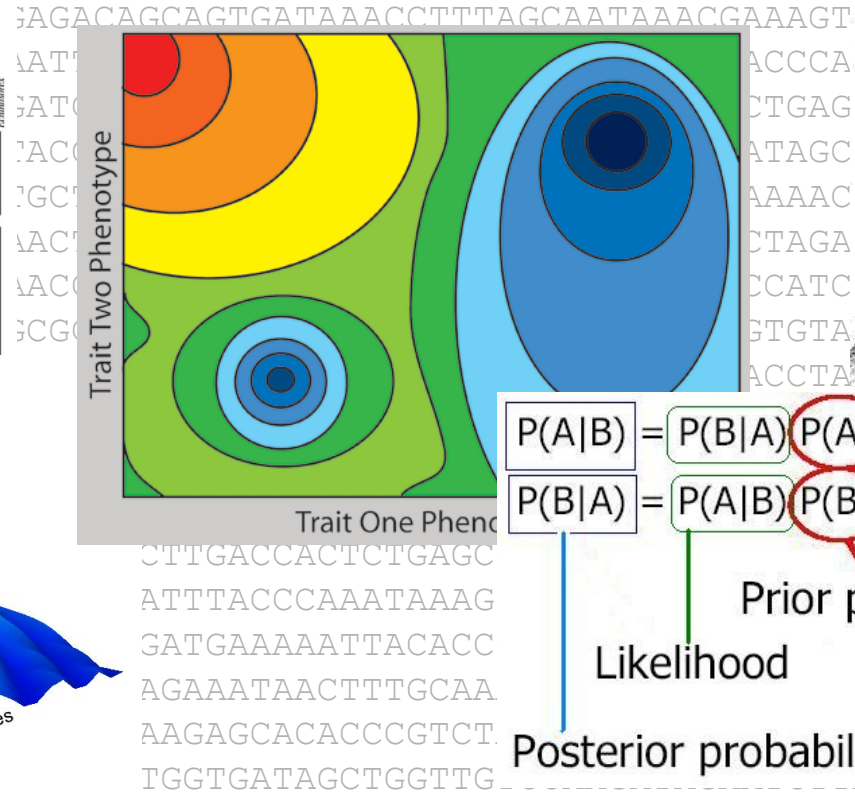
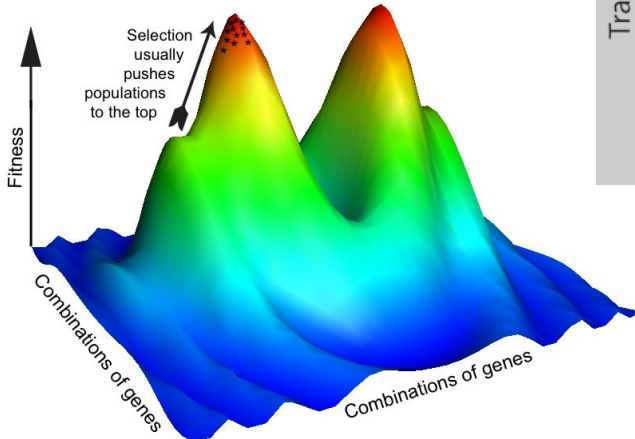
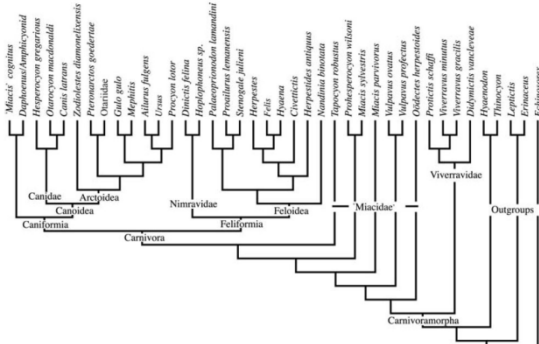


FYLOGENETICKÁ ANALÝZA II.



Maximální věrohodnost (*maximum likelihood, ML*)

15× hod mincí: skóre OOHHHOHOOOHOHHO

7× panna (hlava, H), 8× orel (O)

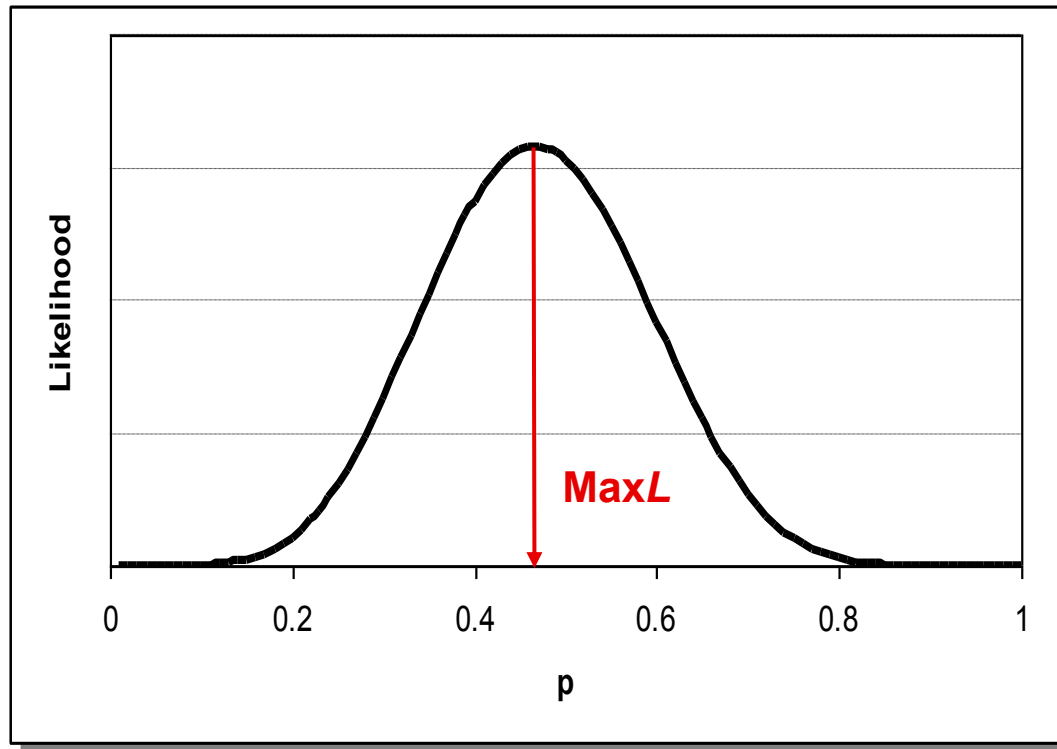


pravděpodobnost, že padne hlava = p , orel = $(1 - p)$

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 = $0,4666 \approx 7/15$



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

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

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

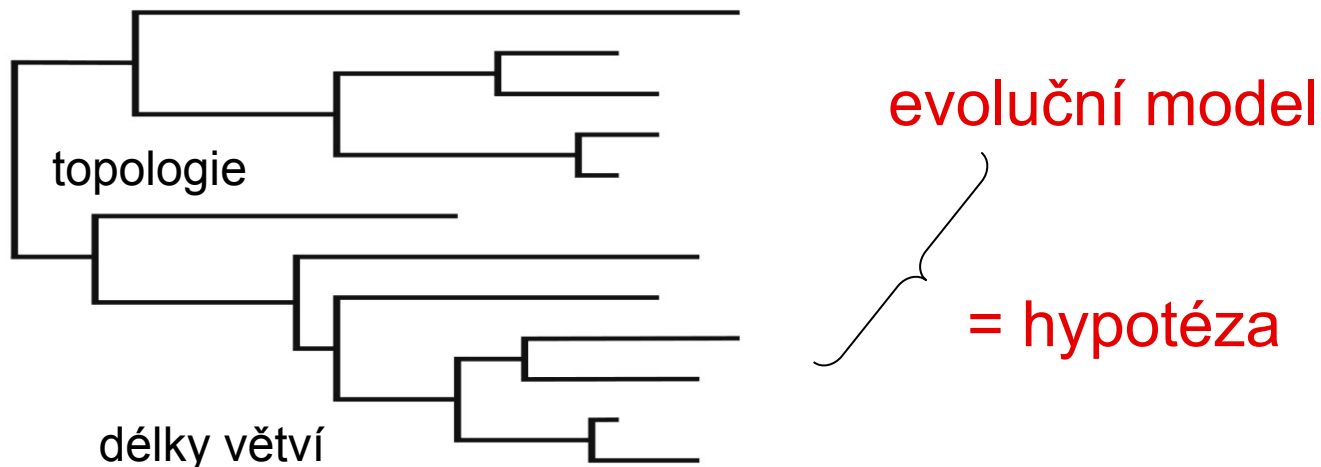
Věrohodnost (*likelihood*) = podmíněná pravděpodobnost dat (D) při platnosti dané hypotézy (H): $L = P(D | H)$

Maximální věrohodnost ve fylogenetické analýze

data:

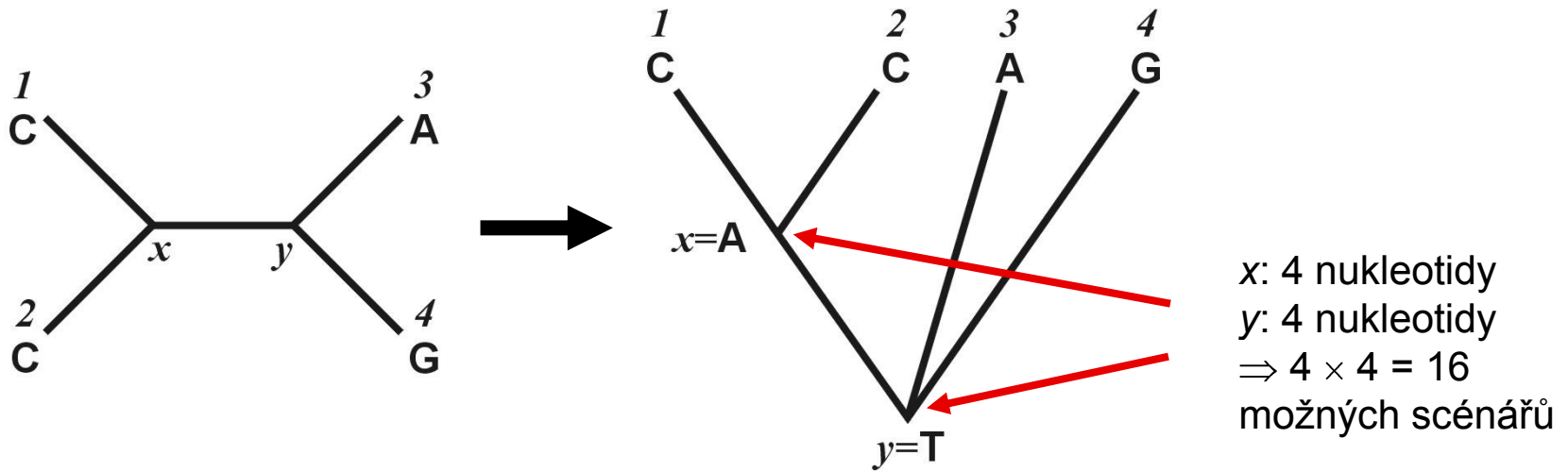
```
1   TCAAAAATGGCTTTATTTCGCTTAATGCCGTTAACCTTGCGGGGGCCATG
2   TCCGTGATGGATTTATTTCCGCAATGCCTGTCATCTTATTCTCAAGTATC
3   TTCGTGATGGATTTATTGCAGGTATGCCAGTCATCCTTTTCTCATCTATC
4   TTCGTGACGGGTTTATCTCGGCAATGCCGGTCATCCTATTTTCGAGTATT
```

strom:



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

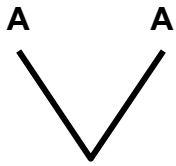
	1		j		N																																											
1	T	C	A	A	A	A	T	G	G	C	T	T	A	T	T	C	G	C	T	T	A	A	C	C	T	T	G	C	G	G	G	G	C	C	A	T	G											
2	T	C	C	G	T	G	A	T	T	T	A	T	T	T	T	C	C	G	C	A	A	T	G	C	C	T	G	T	C	A	T	C	T	T	A	T	T	C	T	C	A	A	G	T	A	T	C	
3	T	T	C	G	T	G	A	T	T	T	A	T	T	G	C	A	G	G	T	A	T	G	C	C	A	G	T	C	A	T	C	C	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



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

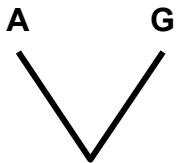
4) $\ln L = \ln L(1) + \ln L(2) + \dots + \ln L(N) = \sum_{j=1}^N \ln L_j$



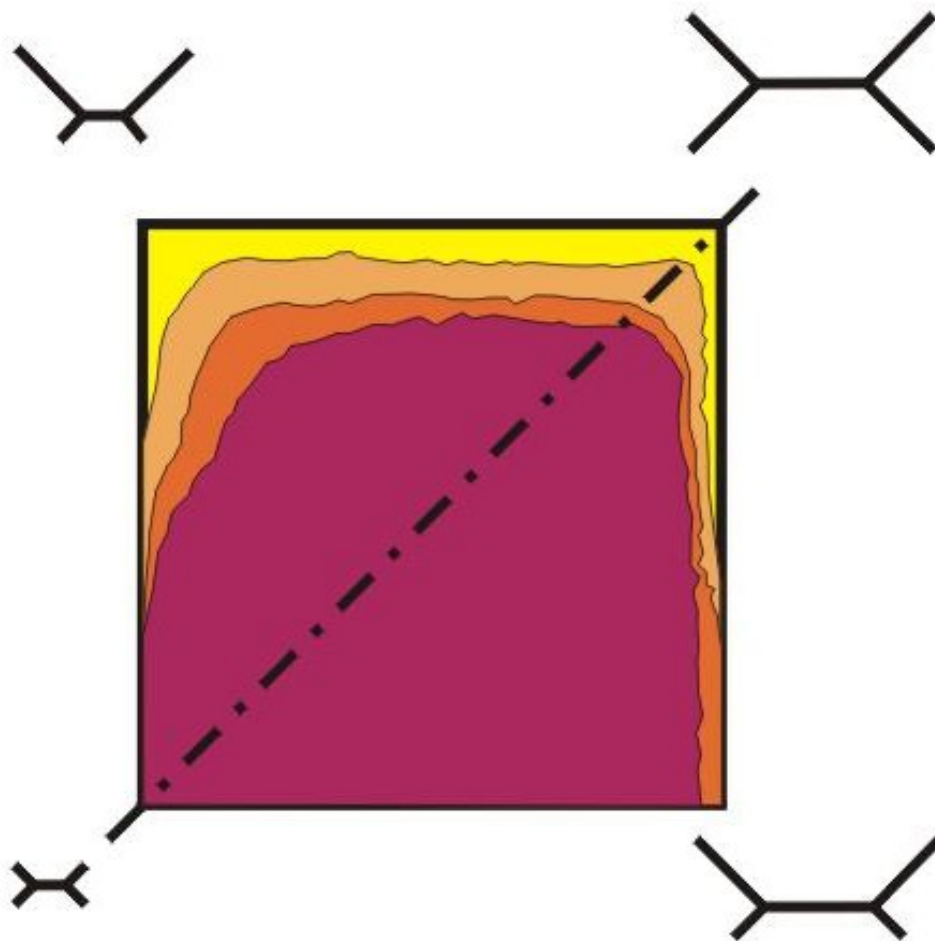
Věrohodnost (ML) a úspornost (MP)

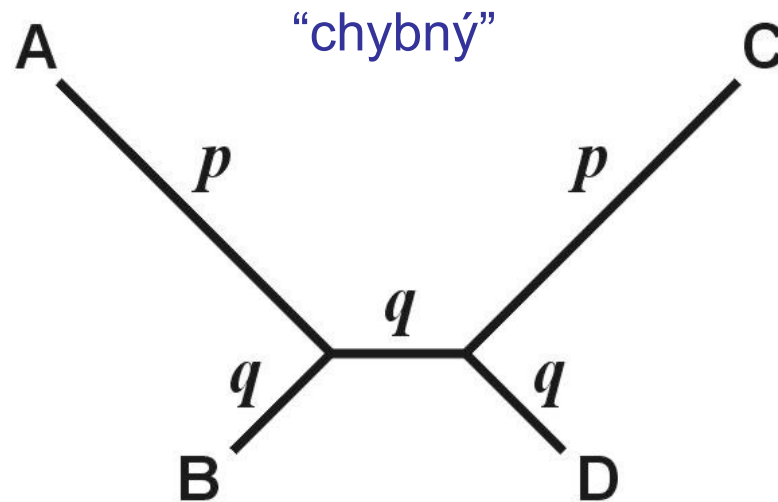
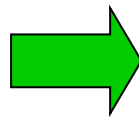
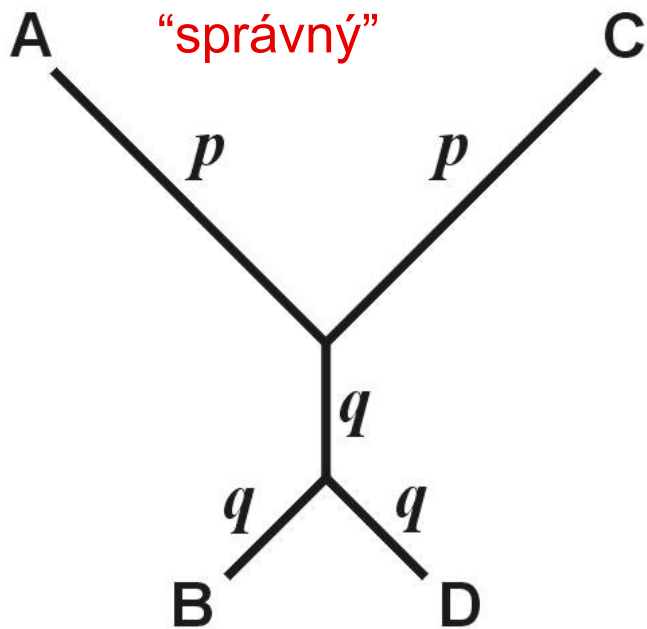
Počet změn	Parsimonie	$\nu = 0,01$	$\nu = 0,10$	$\nu = 0,20$	$\nu = 1,00$
		(0,2475)	(0,2266)	(0,20611)	(0,11192)
0	100	99,99	99,83	99,31	82,17
1	0	0,00	0,00	0,00	0,00
2	0	0,0011	0,11	0,44	9,13
3	0			0,034	3,55
4	0				0,0027

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





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

věrohodnost

apriorní
pravděpodobnost

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

suma přes všechny
možné stromy

Parametry pro bayesovskou analýzu:

ML odhady → **empirická BA**

všechny kombinace → **hierarchická BA**

$$P(\tau, \nu, \theta) = \int P(\tau, \nu, \theta) dF(\nu, \theta)$$

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

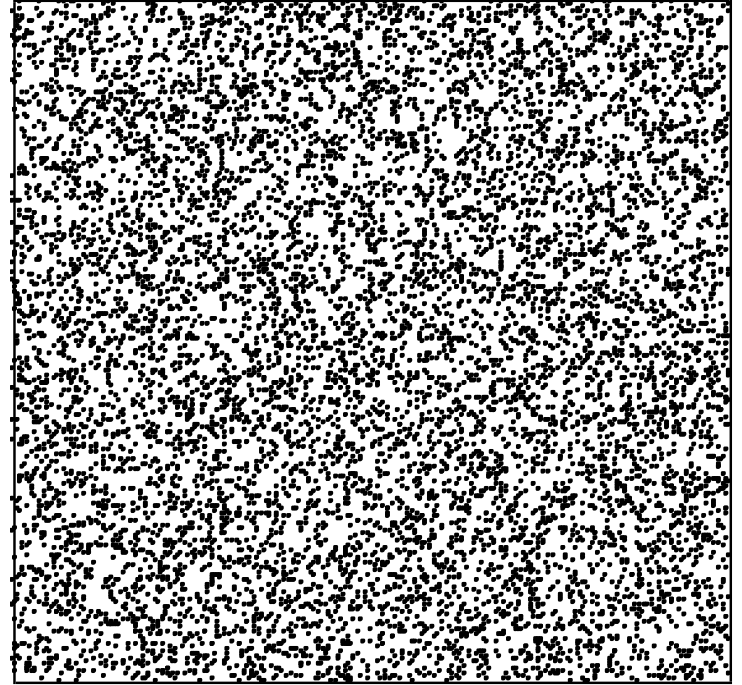
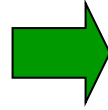
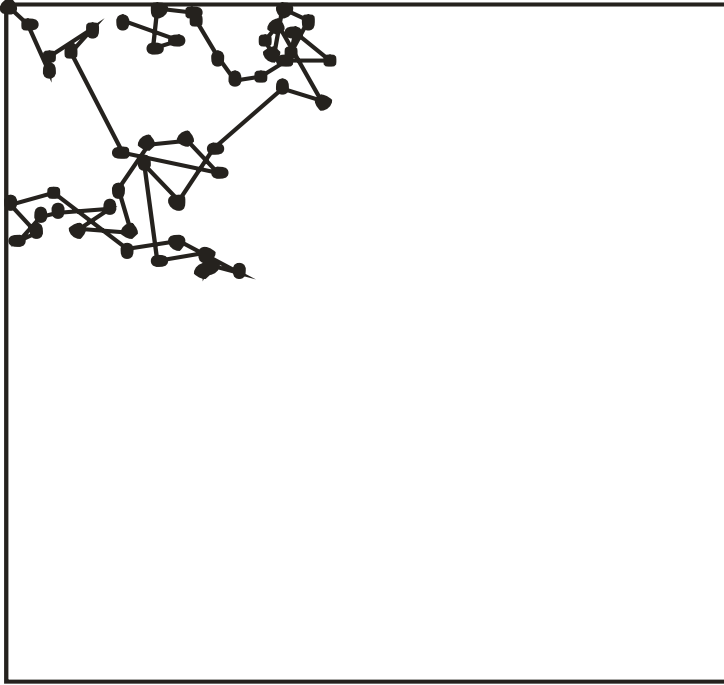
řešení: metody Monte Carlo

náhodný výběr vzorků, při velkém množství aproximace skutečnosti

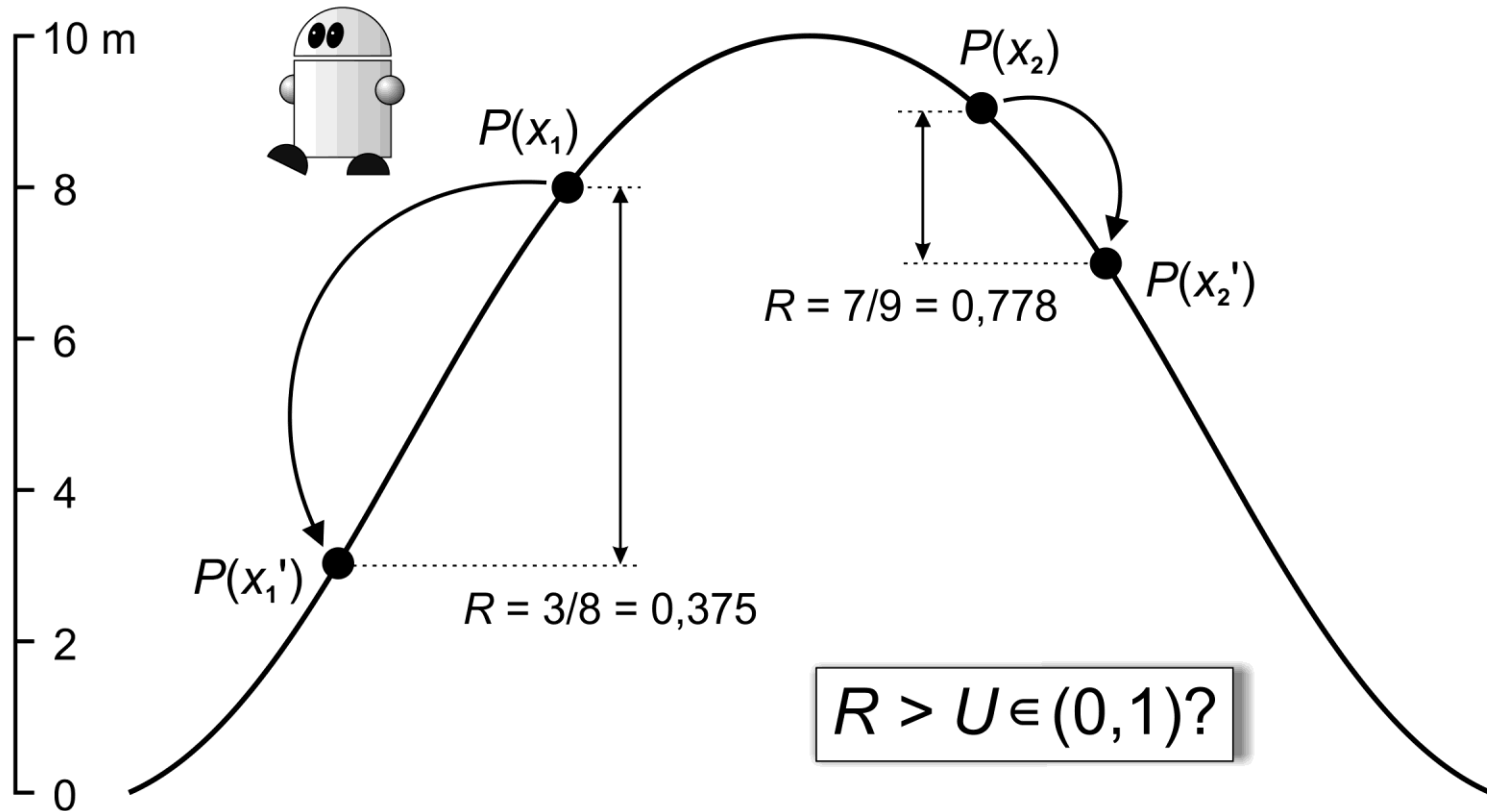
Markovovy řetězce: Markov chain Monte Carlo (MCMC)

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

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



Metropolisův-Hastingsův algoritmus:

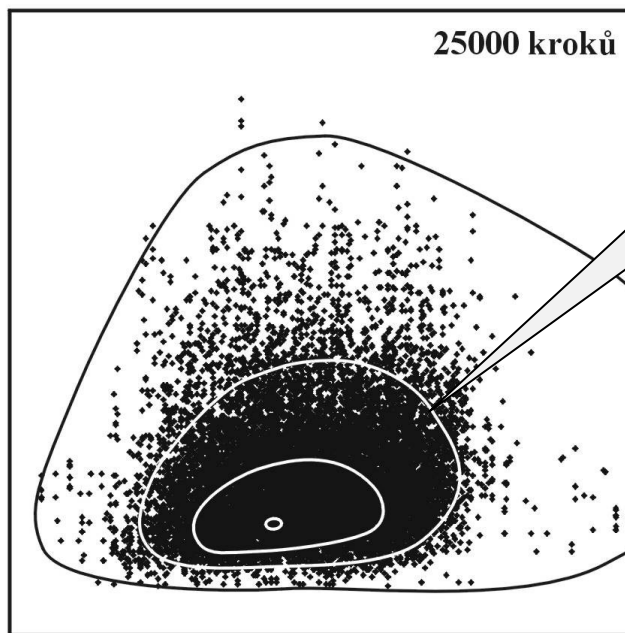
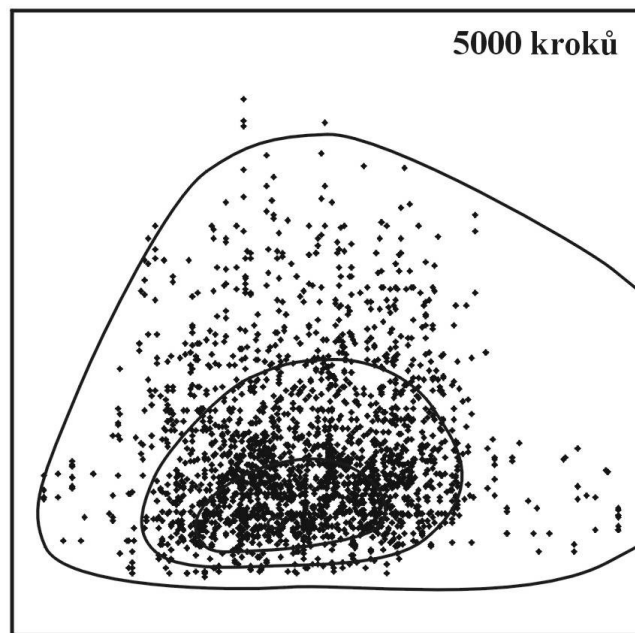


Metropolisův-Hastingsův algoritmus:

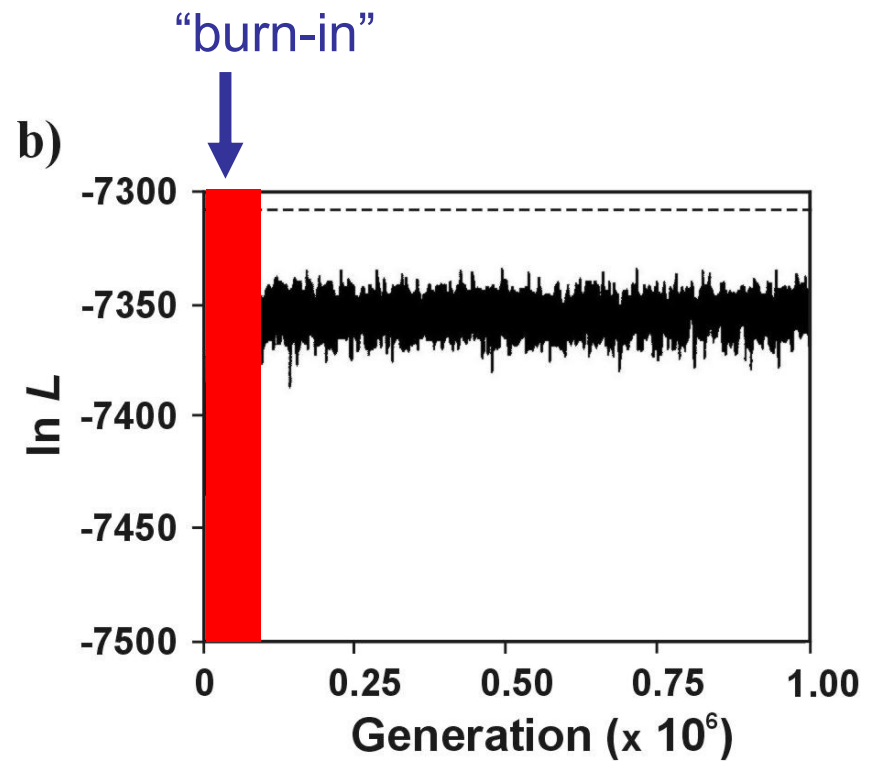
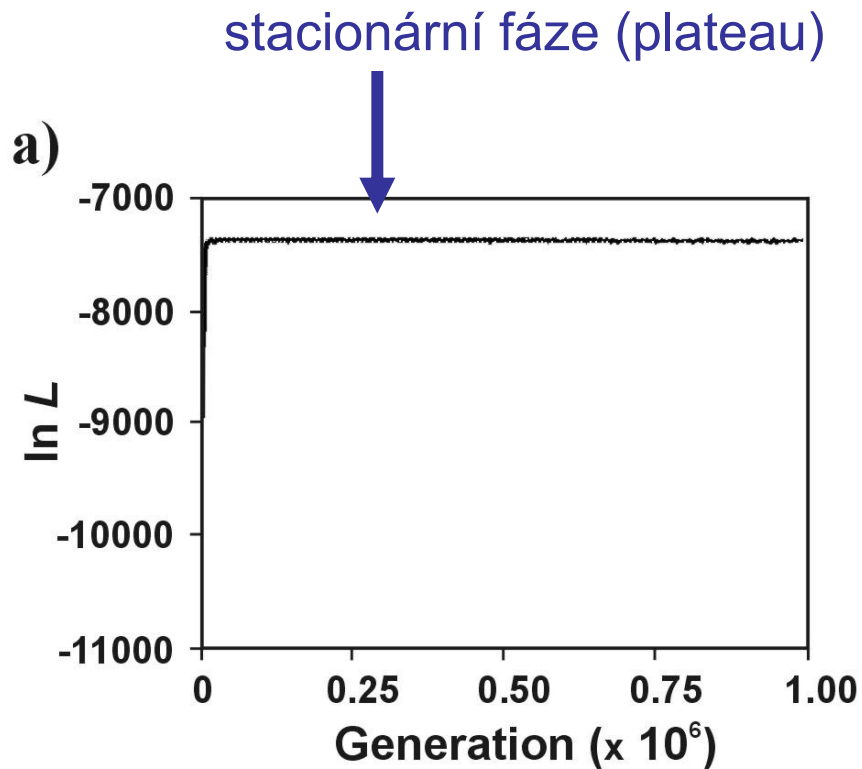
Změna parametru $x \rightarrow x'$

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:

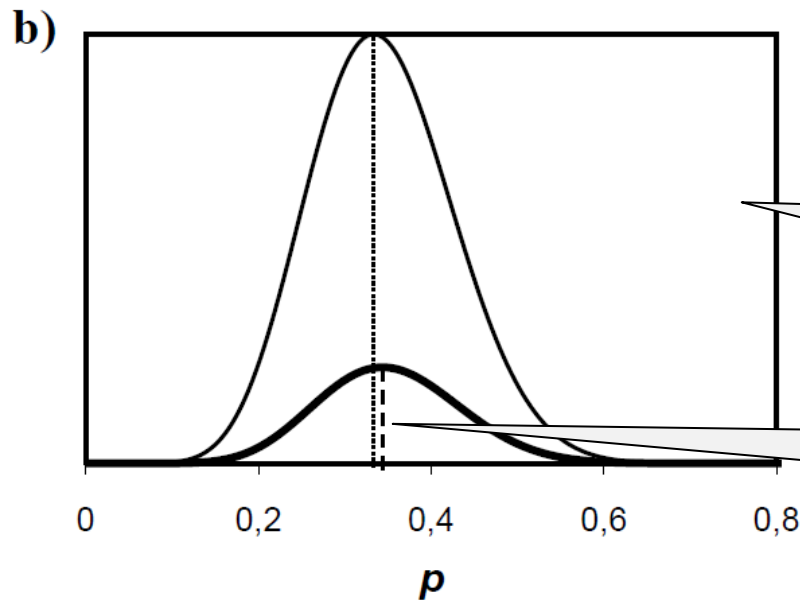
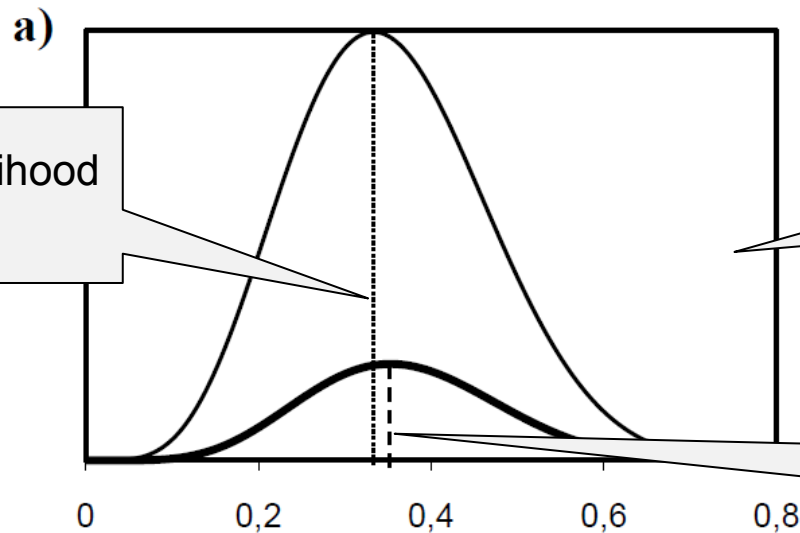


„vrstevnice“
arény

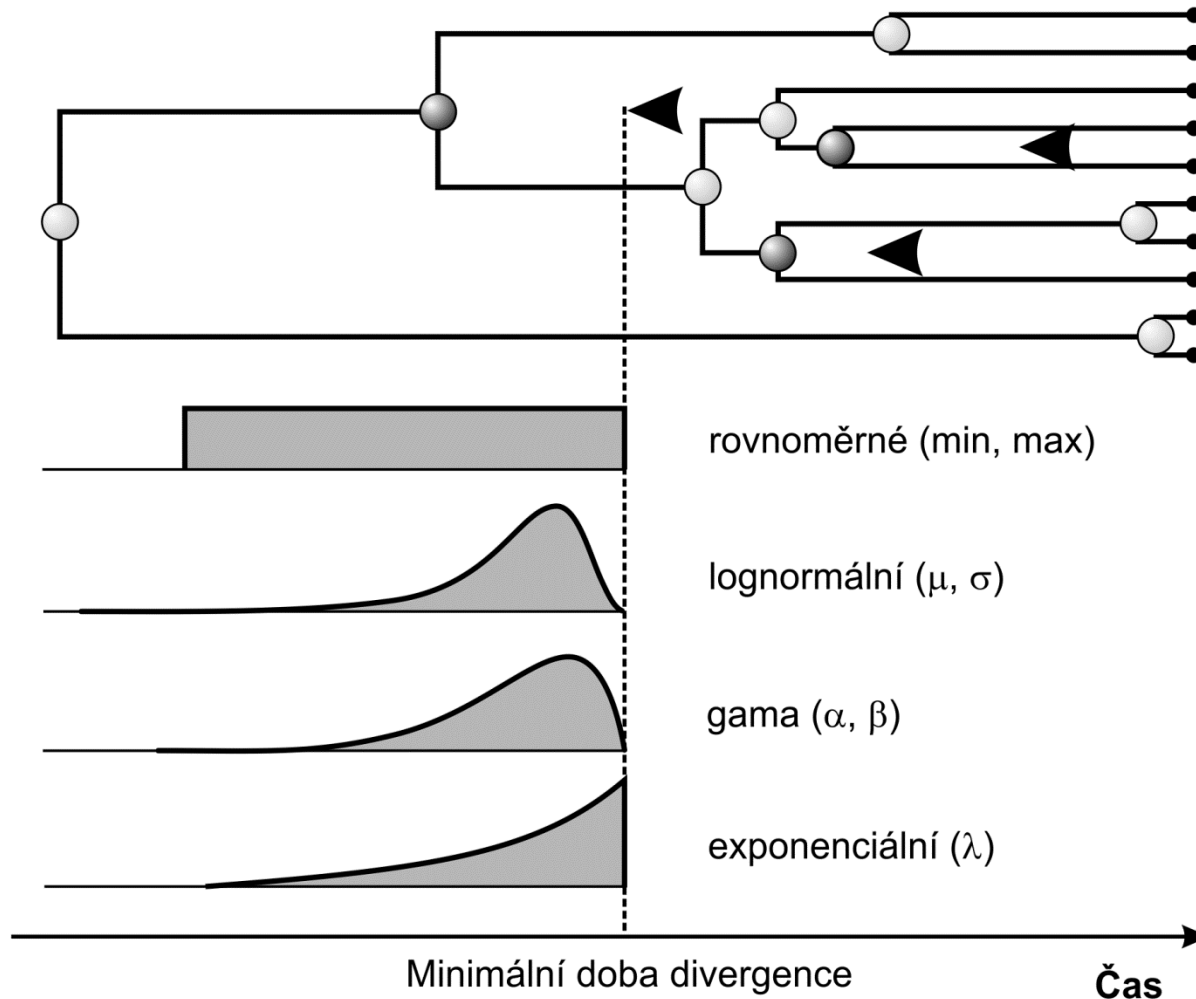


MrBayes: <http://morphbank.ebc.uu.se/mrbayes/>
4 independent chains, Metropolis-coupled MCMC

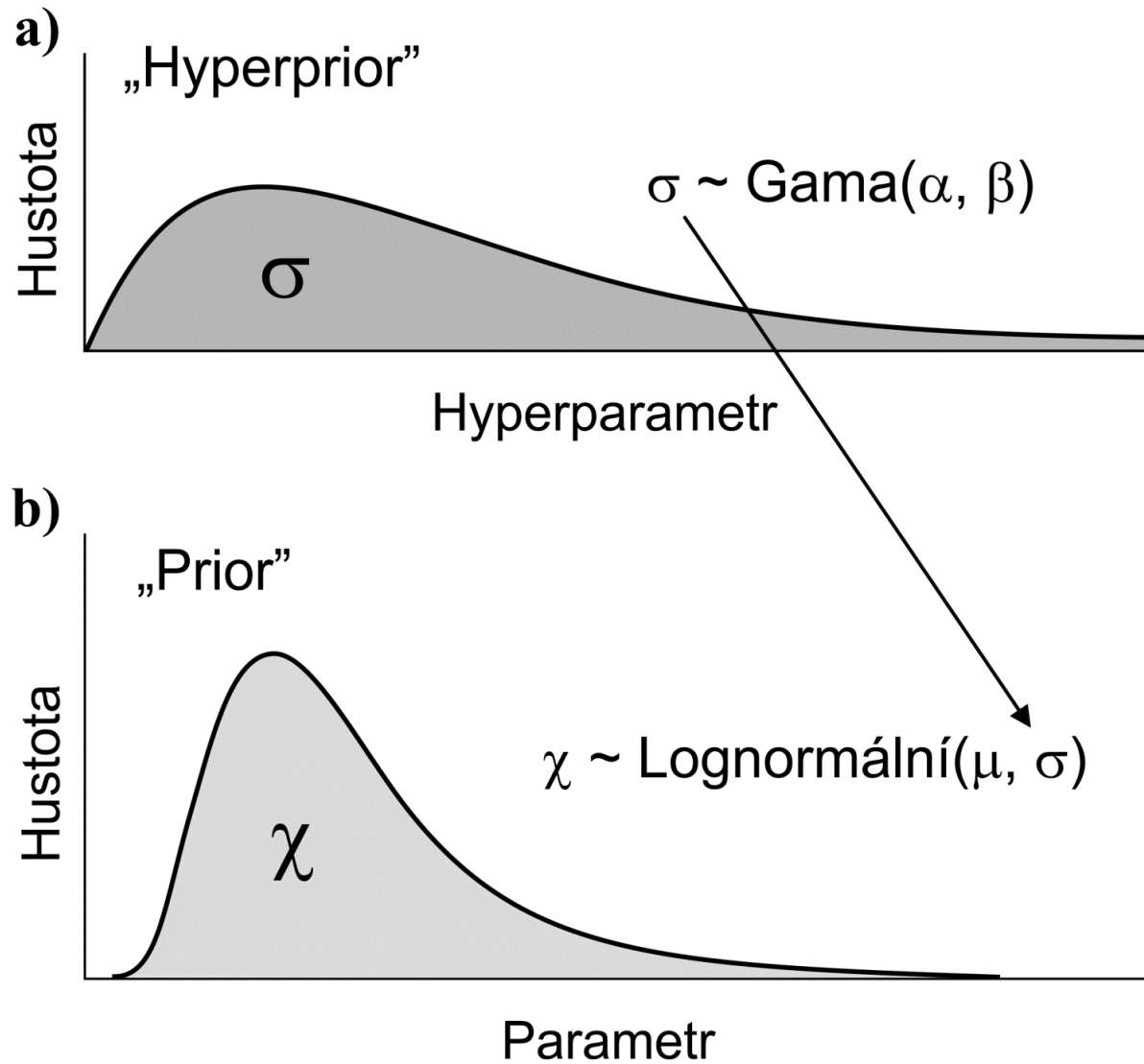
Problém apriorních pravděpodobností: subjektivnost



Stanovení apriorních pravděpodobností:



Stanovení apriorních pravděpodobností:



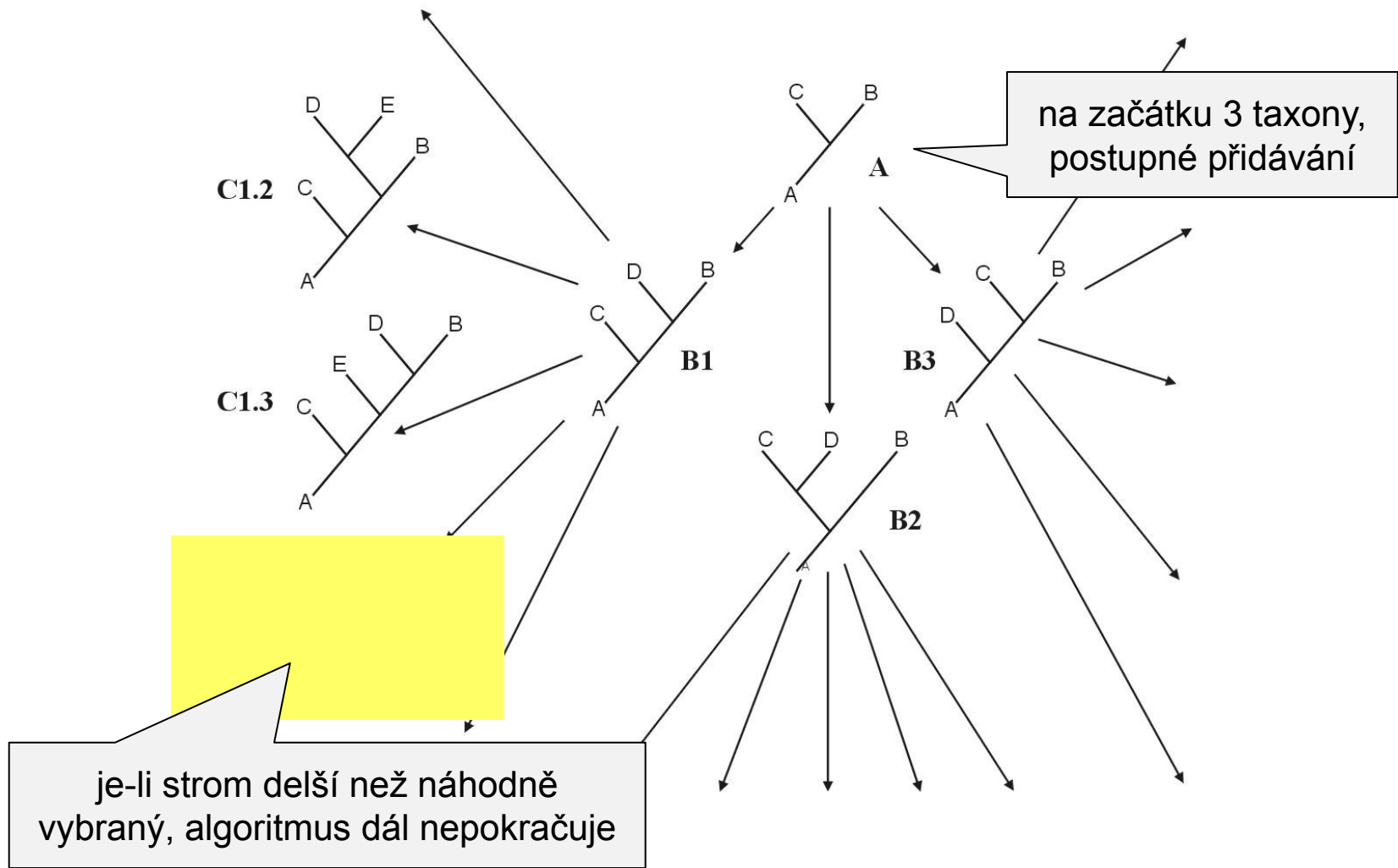
Hledání optimálního stromu a měření spolehlivosti stromů

1. Exaktní metody:

a) vyčerpávající hledání (*exhaustive search*)

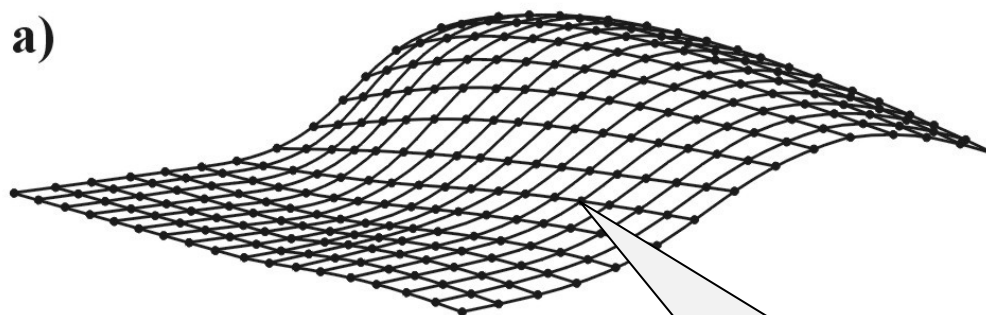
b) *branch-and-bound*

branch-and-bound



2. Heuristický přístup:

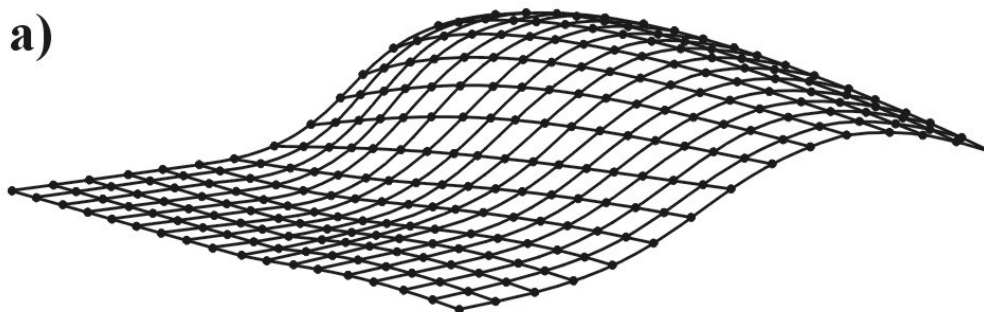
a)



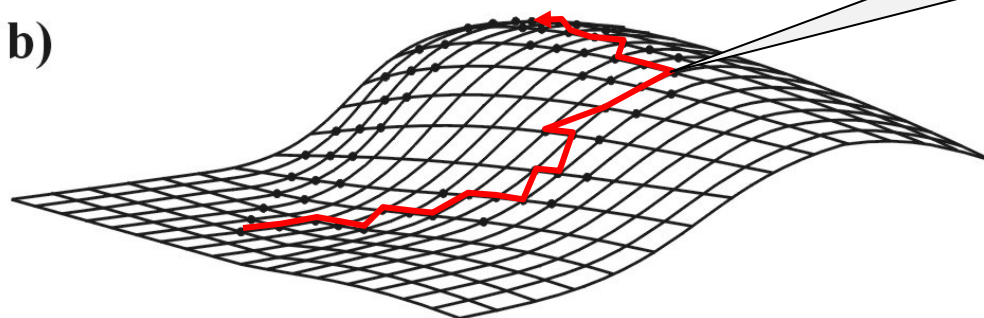
všechny možné stromy

stepwise addition
star decomposition
branch swapping

a)

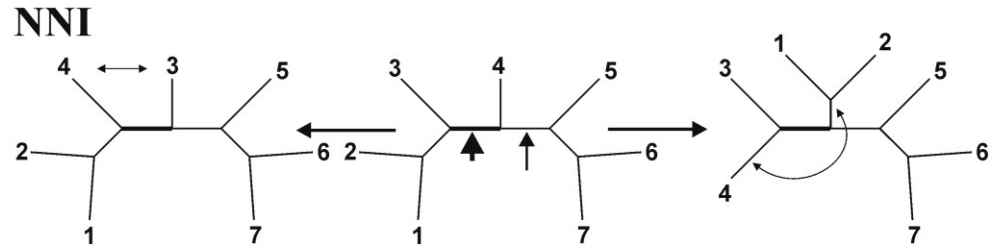


b)

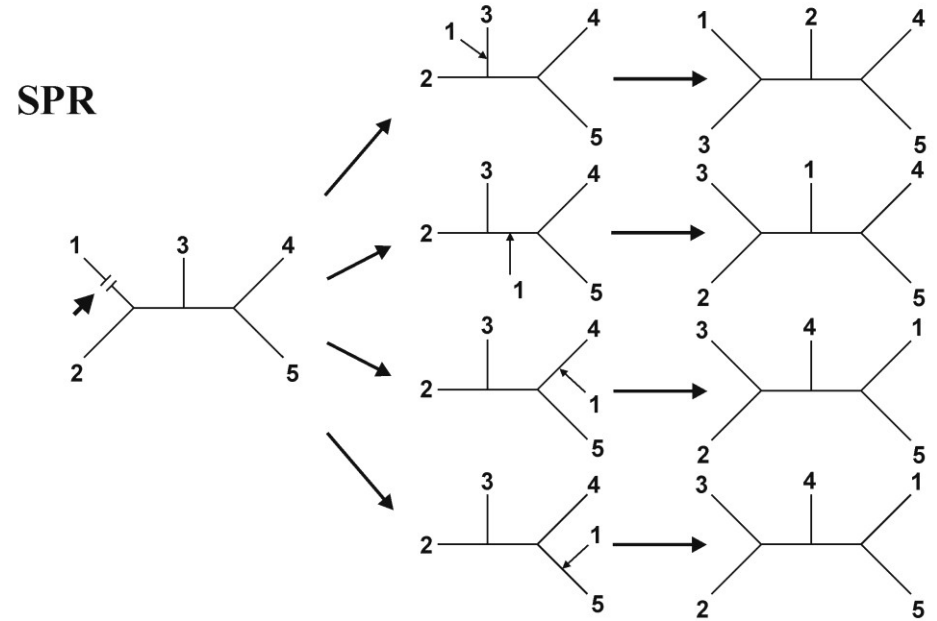


heuristické hledání

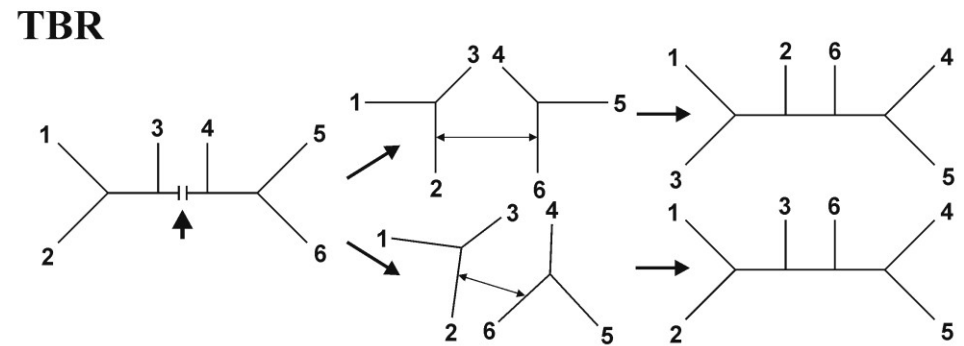
*nearest-neighbor
interchanges (NNI)*



*subtree pruning
and regrafting (SPR)*



*tree bisection and
reconnection (TBR)*



Měření spolehlivosti stromů:

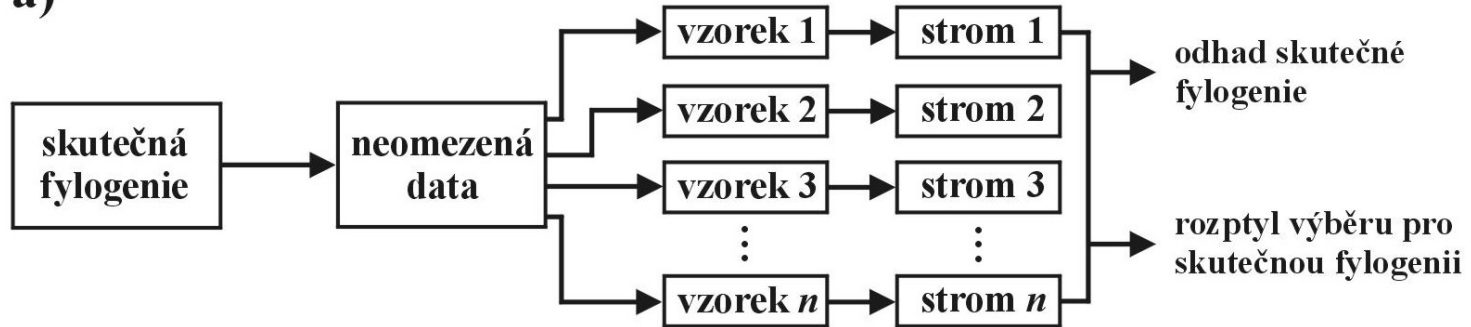
Metody opakovaného výběru

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



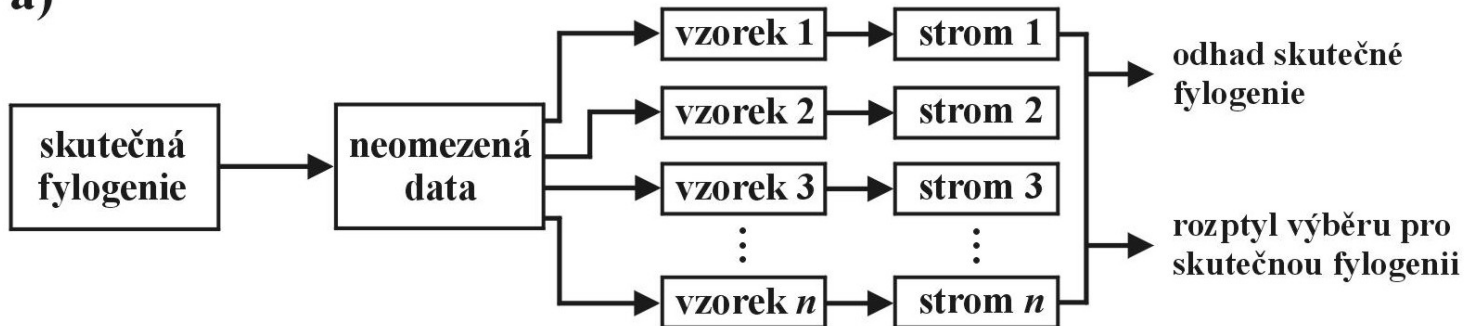
bootstrap:

a)

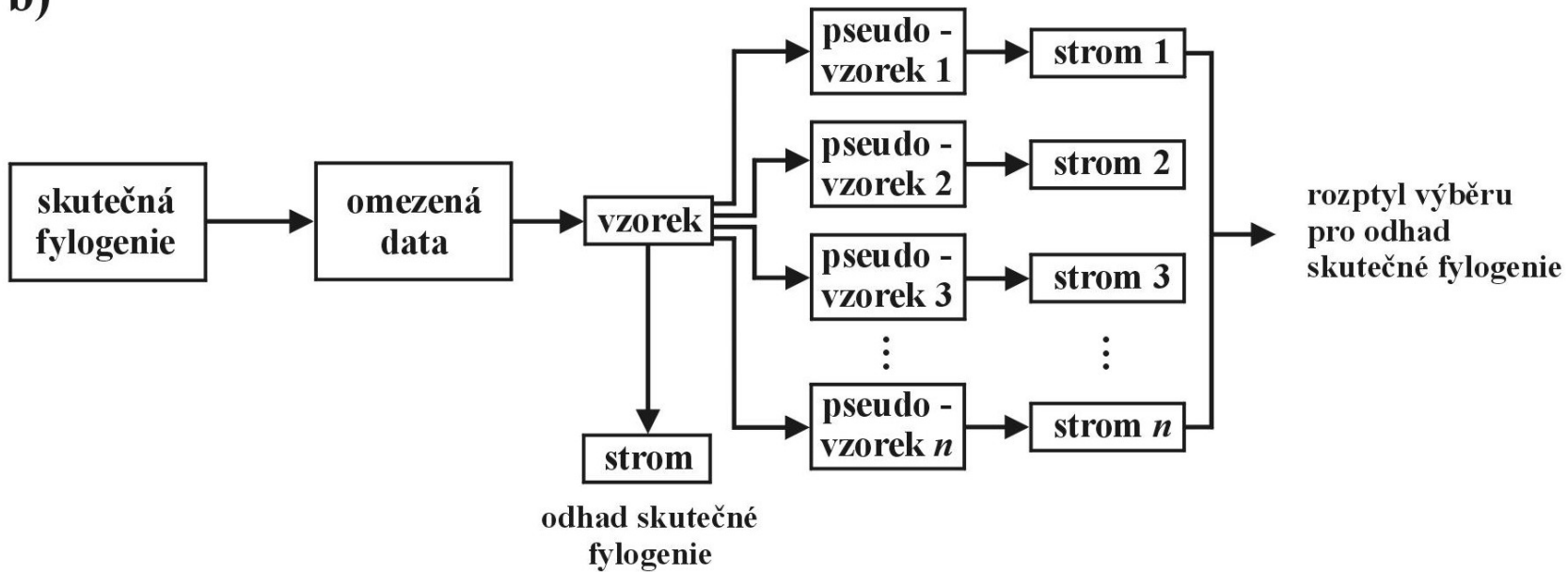


bootstrap:

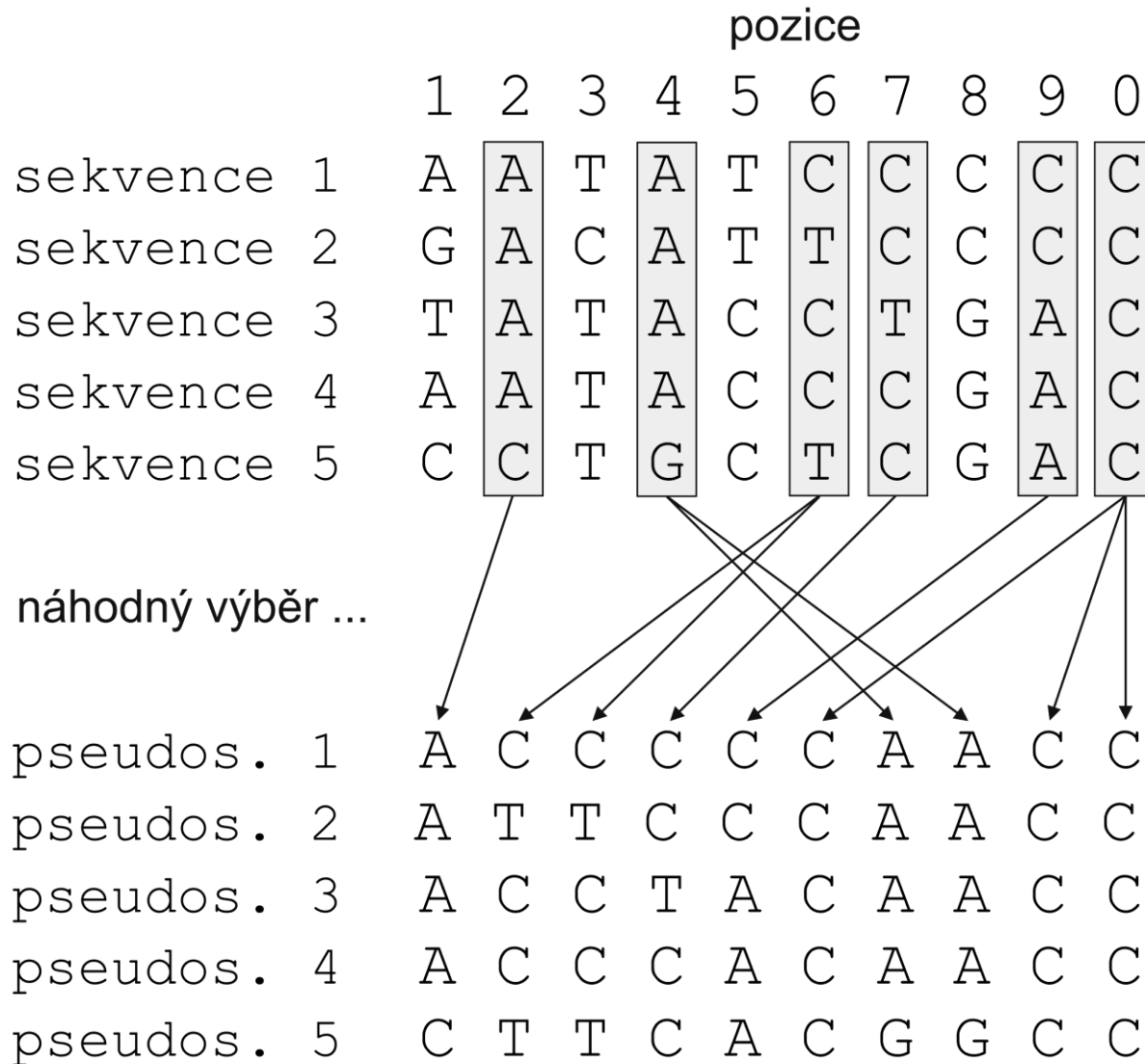
a)



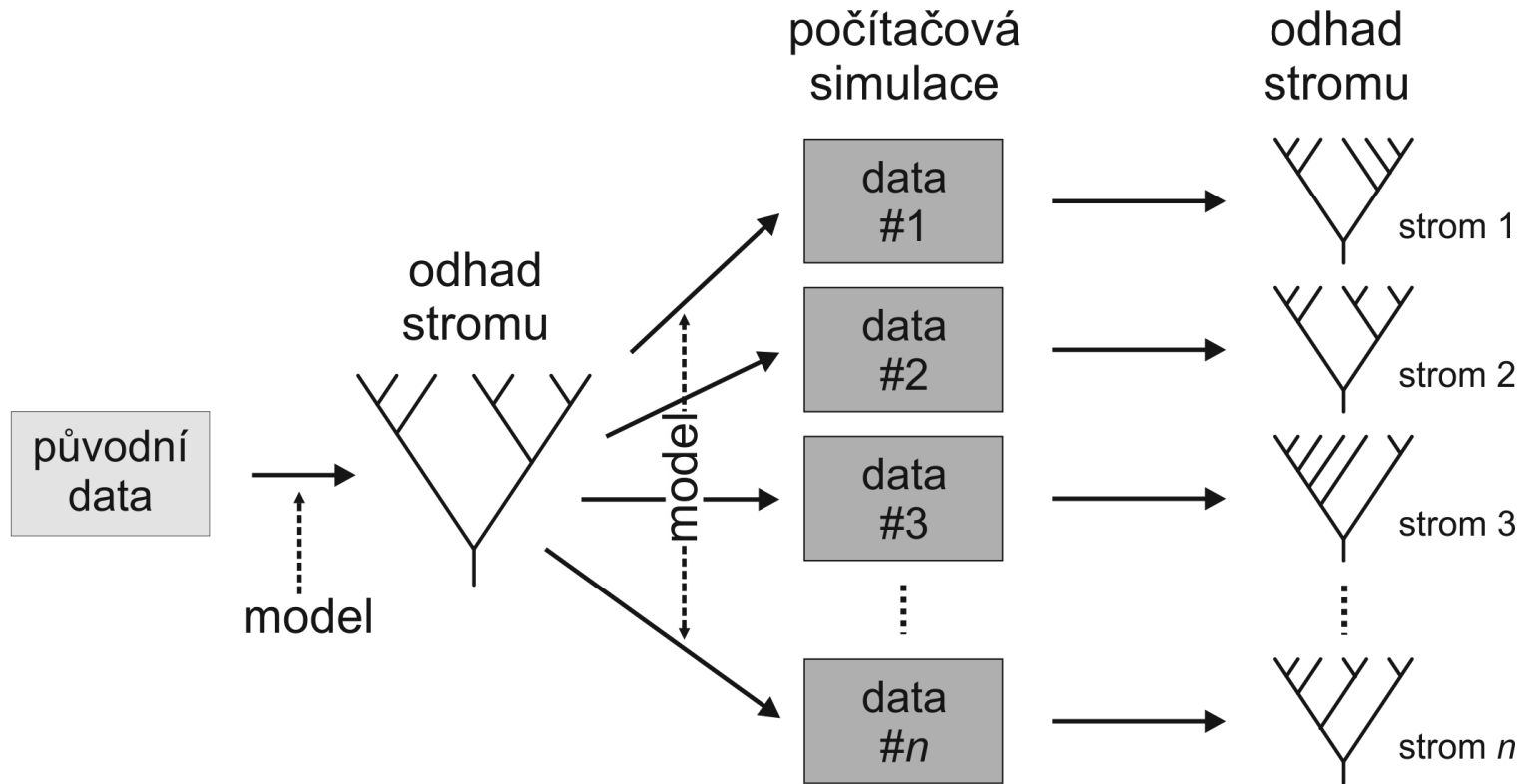
b)



bootstrap:



parametrický bootstrap: evoluční model



bayesovská analýza: aposteriorní pravděpodobnosti

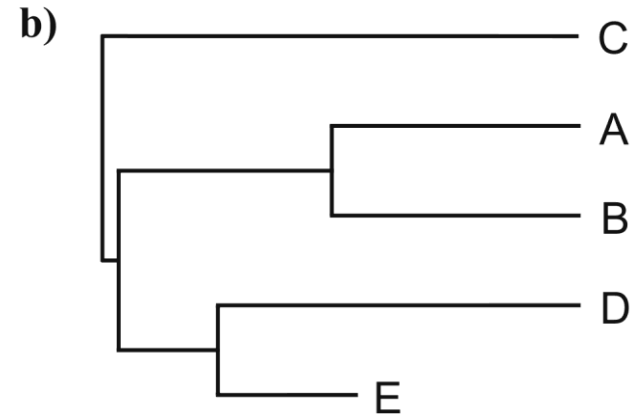
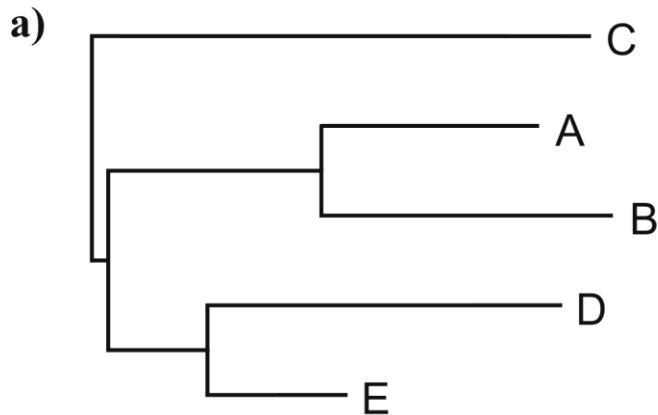
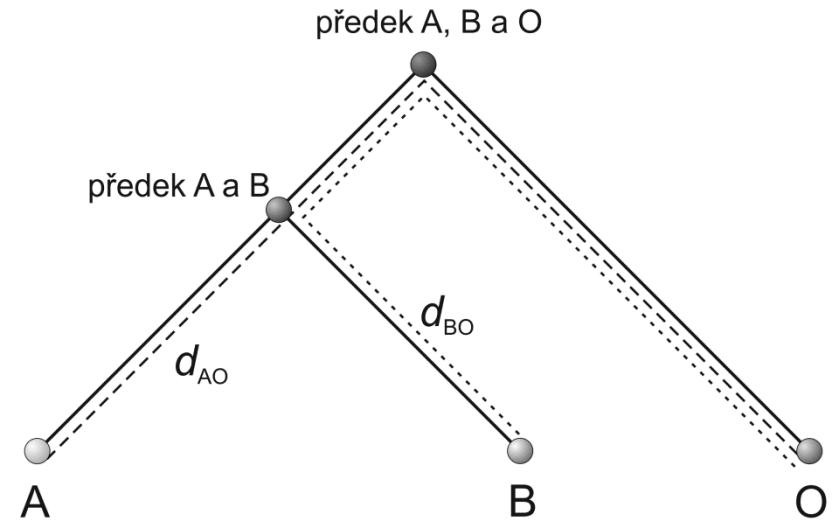
Testování hypotéz

Test molekulárních hodin:

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

Linearizované stromy

odstranění signifikantně odlišných taxonů

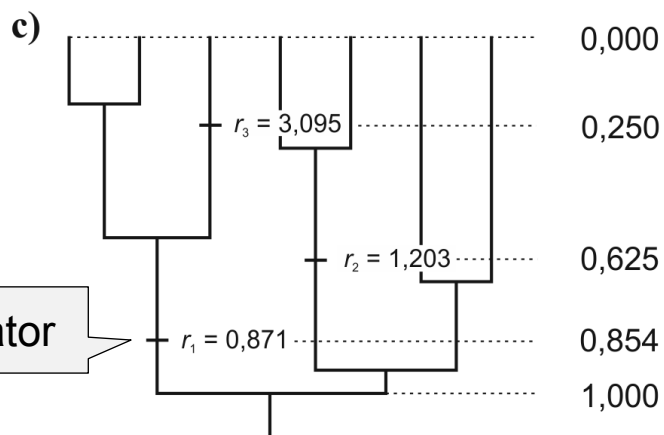
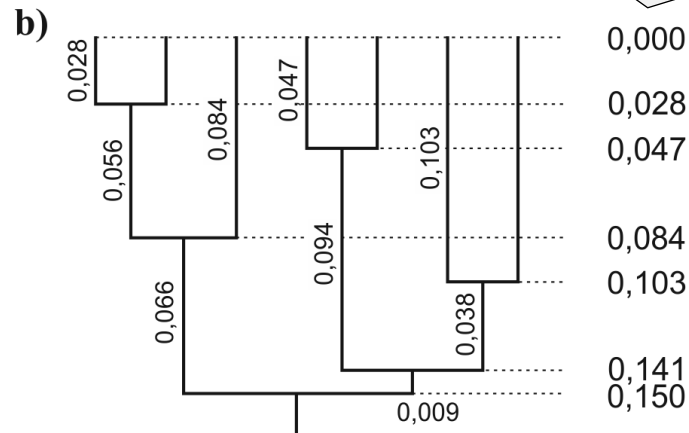
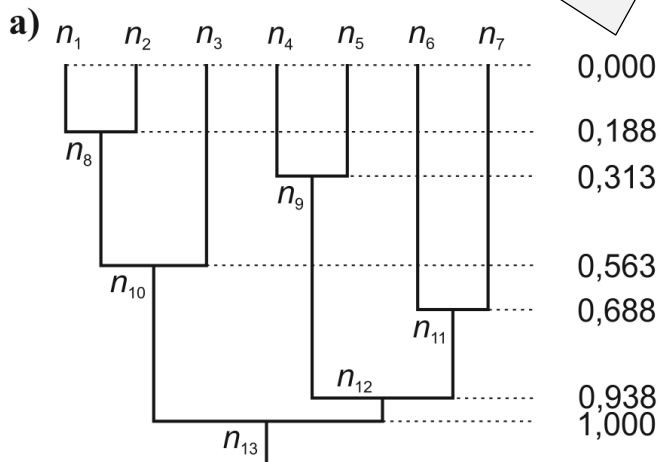


Relaxované molekulární hodiny

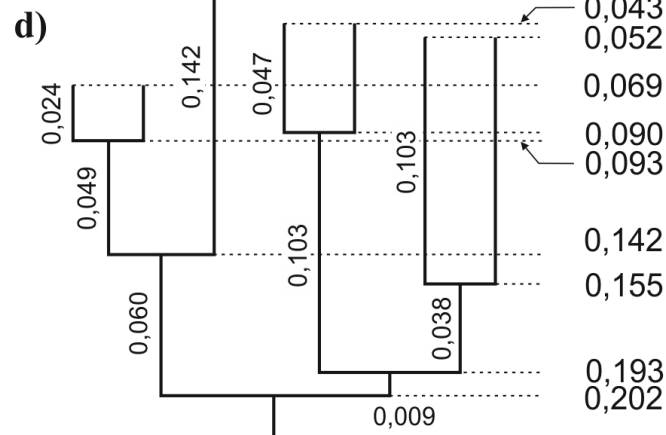
umožňují změnu rychlostí podél větví

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, RELI)

a)

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

kde i je bootstrapový replikát

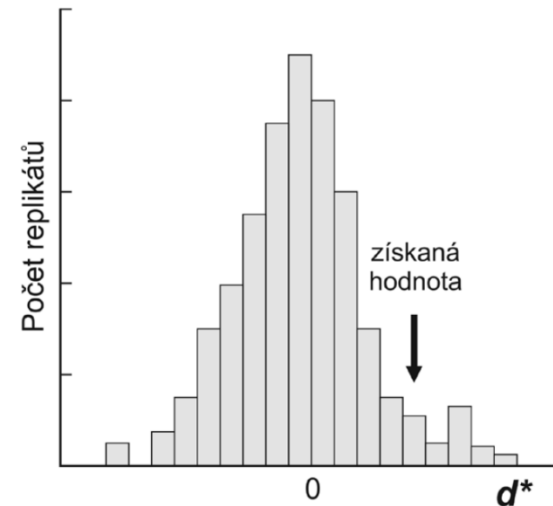
$$d_{1}^* = \ln L_{T1}^* - \ln L_{T2}^*$$

$$d_{2}^* = \ln L_{T1}^* - \ln L_{T2}^*$$

$$d_{3}^* = \ln L_{T1}^* - \ln L_{T2}^*$$

...

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



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

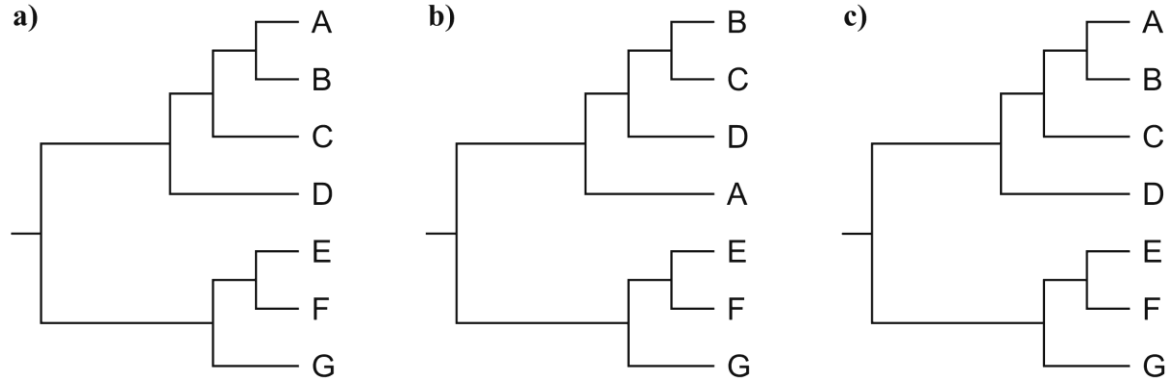
metody inkorporující délky větví

Problémy s distancemi mezi stromy

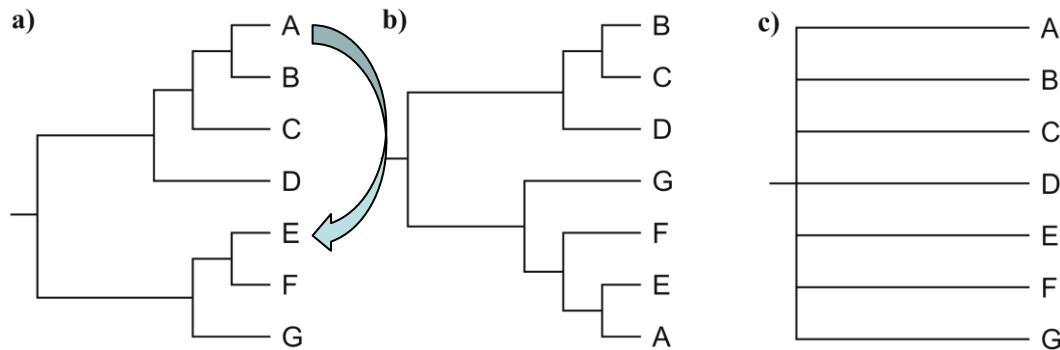
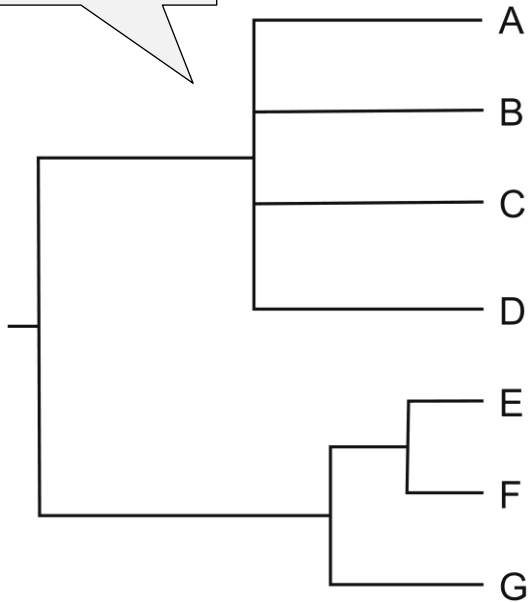
Konsenzuální stromy

striktní konsensus

zdrojové stromy



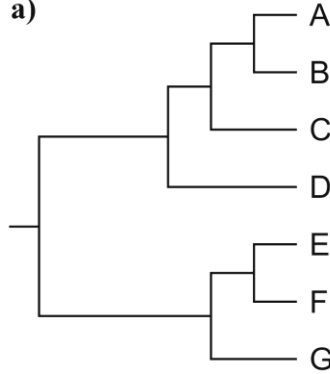
striktně konsenzuální strom



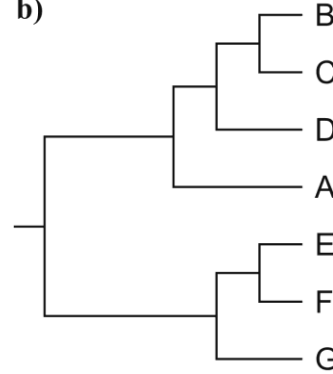
majority-rule

zdrojové stromy

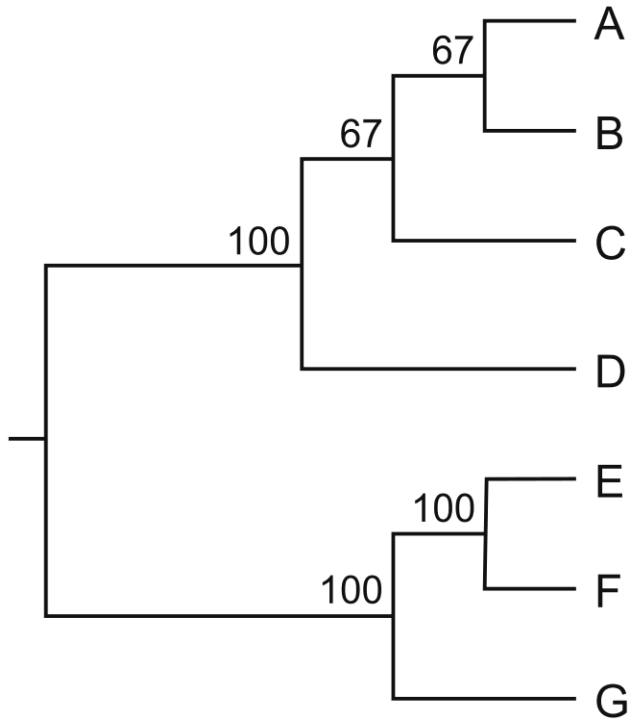
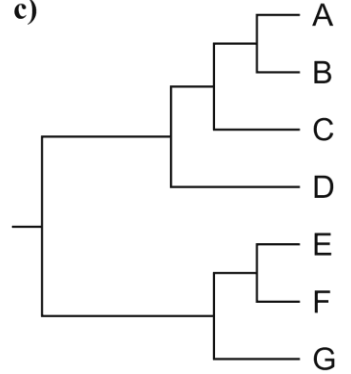
a)



b)



c)



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>