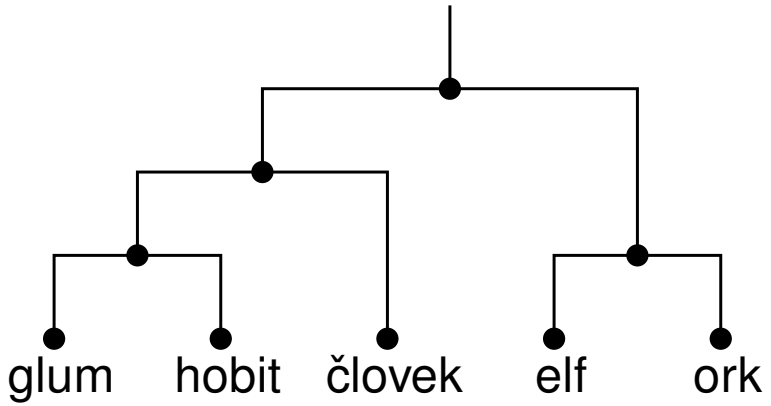
Fylogenetické stromy

Broňa Brejová

26.10.2023

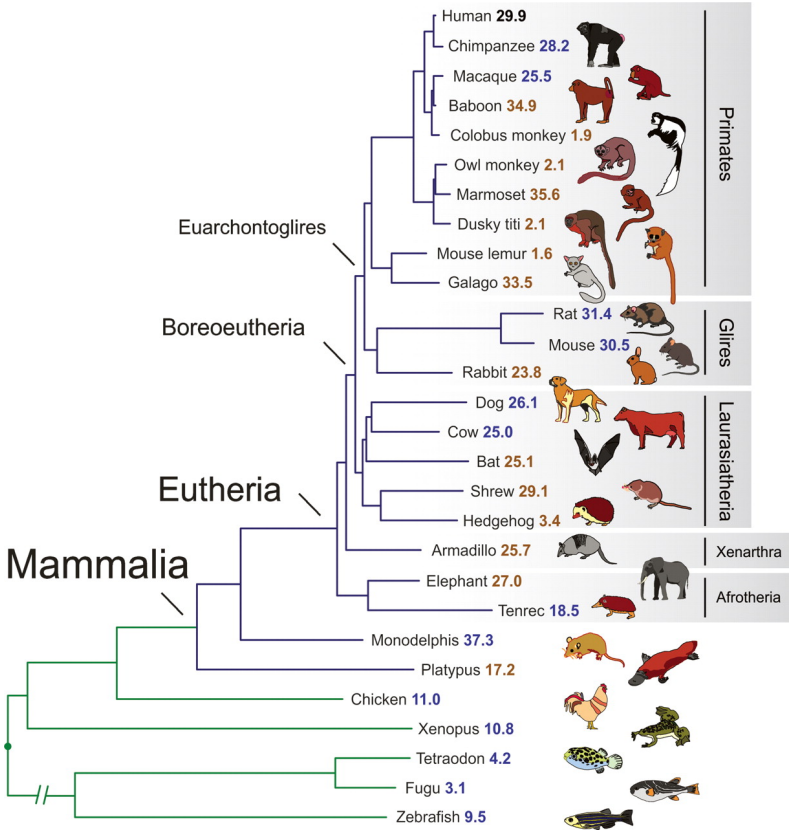
Terminológia

- zakorenený strom, rooted tree
- nezakorenený strom, unrooted tree
- hrana, vetva, edge, branch
- vrchol, uzol, vertex, node
- list, leaf, leaf node, tip, terminal node
- vnútorný vrchol, internal node
- koreň, root
- podstrom, subtree, clade



Reálne ukážky stromov z článkov (zakorenený/nezakorenený)

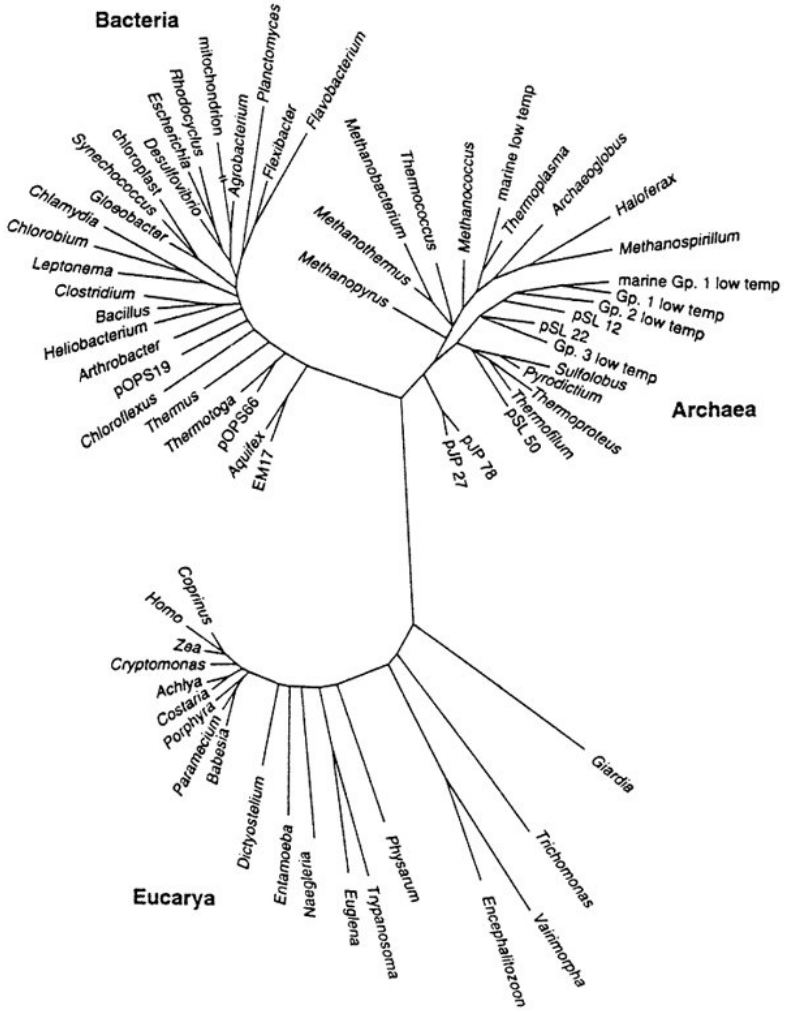
[Margulies et al. 2007]



zakorenený pomocou

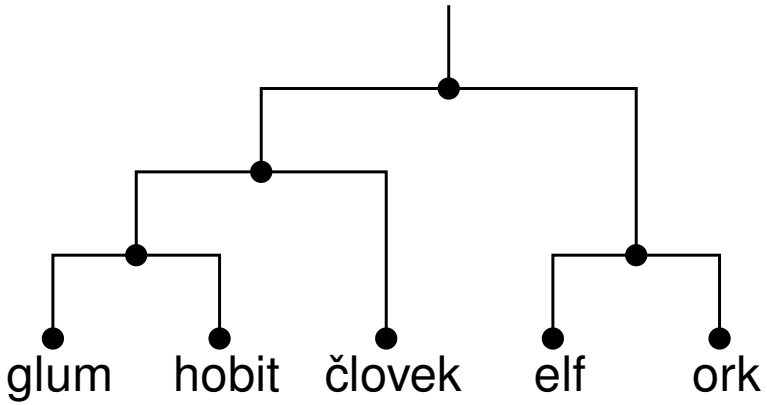
vonkajšej skupiny (outgroup)

[Pace et al 1997]



Zopár faktov o stromoch

- Majme zakorenený strom s n listami, v ktorom má každý vnútorný vrchol 2 deti. Takýto strom vždy má $n - 1$ vnútorných vrcholov a $2n - 2$ vetiev (prečo?)
- Majme nezakorenený strom s n listami, v ktorom má každý vnútorný vrchol 3 susedov. Takýto strom vždy má $n - 2$ vnútorných vrcholov a $2n - 3$ vetiev.
- Koľkými spôsobmi môžeme zakoreniť nezakorenený strom s n listami?

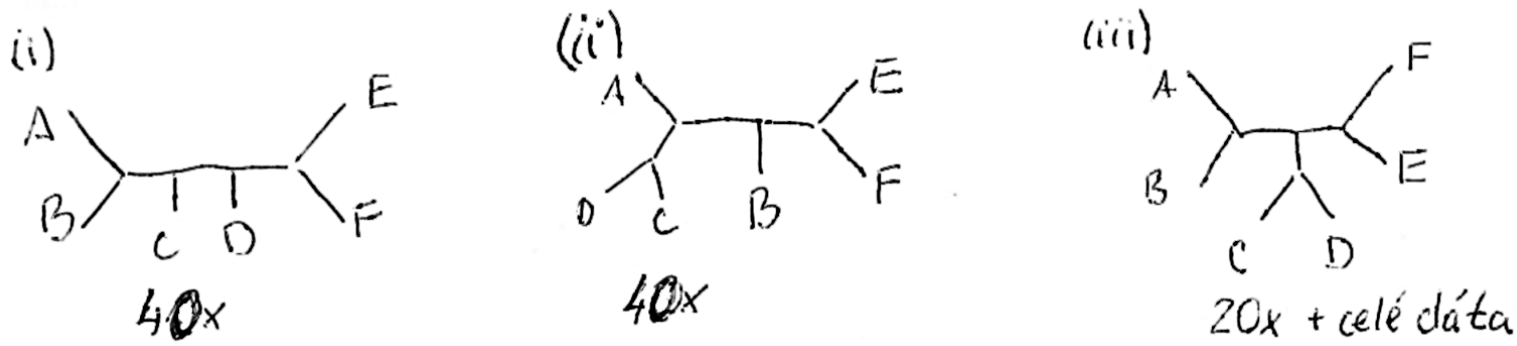


Bootstrap

- Náhodne vyberieme niektoré stĺpce zarovnania, zostrojíme strom
- Celé to opakujeme veľa krát
- Značíme si, koľkokrát sa ktorá hrana opakuje v stromoch
  (v nezakorenenom strome je hrana rozdelenie listov na dve skupiny)
- Nakoniec zostavíme strom z celých dát a pozrieme sa ako často sa ktorá jeho hrana vyskytovala
- Môžeme zostaviť aj strom z často sa vyskytujúcich hrán
- Bootstrap hodnoty sú odhadom spoľahlivosti, hlavne ak máme celkovo málo dát (krátke zarovnanie)
- Ak však dáta nezodpovedajú vybranej metóde/modelu, tak aj pre zlý strom môžeme dostať vysoký bootstrap

Bootstrap

Robili sme $100\times$ bootstrap, dostali sme tieto výsledky:



Doplňte bootstrap hodnoty hranám výsledného stromu (iii)

Ktoré ďalšie vetvy majú podporu aspoň 20%?

Aký strom by sme dostali, ak by sme chceli nechať iba vetvy s podporou aspoň 80%?

Opakovanie pravdepodobnostných modelov

Keď počítame pravdepodobnosť, rozmýšľame o **myšlienkovom experimente**, v ktorom hádžeme kockou, ťaháme gulôčky z vreca a pod.

- Dôležité je vždy si poriadne uvedomiť, ako tento experiment prebieha
- Experimenty nastavujeme tak, aby odzrkadľovali nejaké aspekty reality, napr. skutočných DNA sekvencií, ich evolúcie a pod.
- Pravdepodobnosti, ktoré spočítame v idealizovanom svete nám možno niečo povedia o reálnom svete
- Slávny citát štatistika Georga Boxa:
  All models are wrong, but some are useful.

Aké sme doteraz videli modely

- **Skórovacie matice:** porovnávame model náhodných sekvencií a model náhodných zarovnaní
- **E-value v BLASTe:** náhodne vygenerujeme databázu a dotaz (query), koľko bude v priemere medzi nimi lokálnych zarovnaní so skóre aspoň S?
- **Hľadanie génov:** model generujúci sekvenciu+anotáciu naraz (parametre nastavené na známych génoch).
  Pre danú sekvenciu, ktorá anotácia je najpravdepodobnejšia?
- **Evolúcia, Jukes-Cantorov model:** model generujúci stĺpec zarovnania.
  Neznáme parametre: strom, dĺžky hrán.
  Pre danú sadu stĺpcov zarovnania, ktoré parametre povedú k najväčšej pravdepodobnosti? $\max_{param} \Pr(data|param)$

Evolúcia, Jukes-Cantorov model

Model generujúci stĺpec zarovnania.

Neznáme parametre: strom, dĺžky hrán.

Pre danú sadu stĺpcov zarovnania, ktoré parametre povedú k najväčšej pravdepodobnosti? $\max_{param} \Pr(data|param)$

- Pravdepodobnosť zmeny/nezmeny na hrane dĺžky t:
  $Pr(A|A,t) = (1 + 3e^{-\frac{4}{3}t})/4,$
  $P(C|A,t) = (1 - e^{-\frac{4}{3}t})/4$
- Ak poznáme ancestrálne sekvencie, vieme spočítať pravdepodobnosť dát
- Ancestrálne sekvencie sú náhodné premenné, ktoré nás nezaujímajú: marginalizujeme ich (uvažujeme všetky ich možné hodnoty)