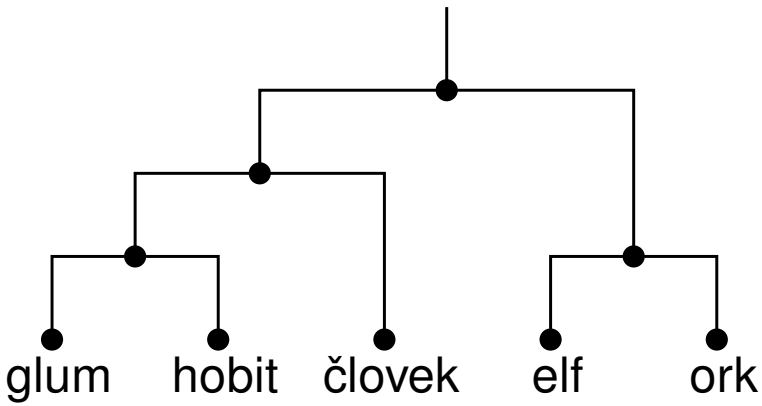
Fylogenetické stromy

Jana Černíková

30.10.2025

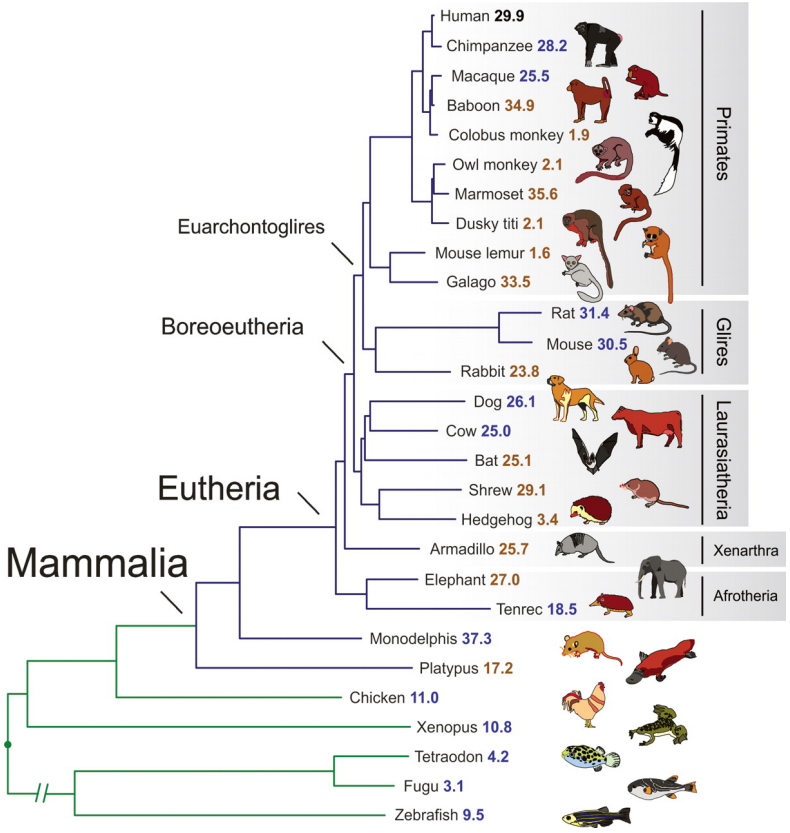
Terminológia

- 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, branch point
- koreň, root
- podstrom, subtree, clade



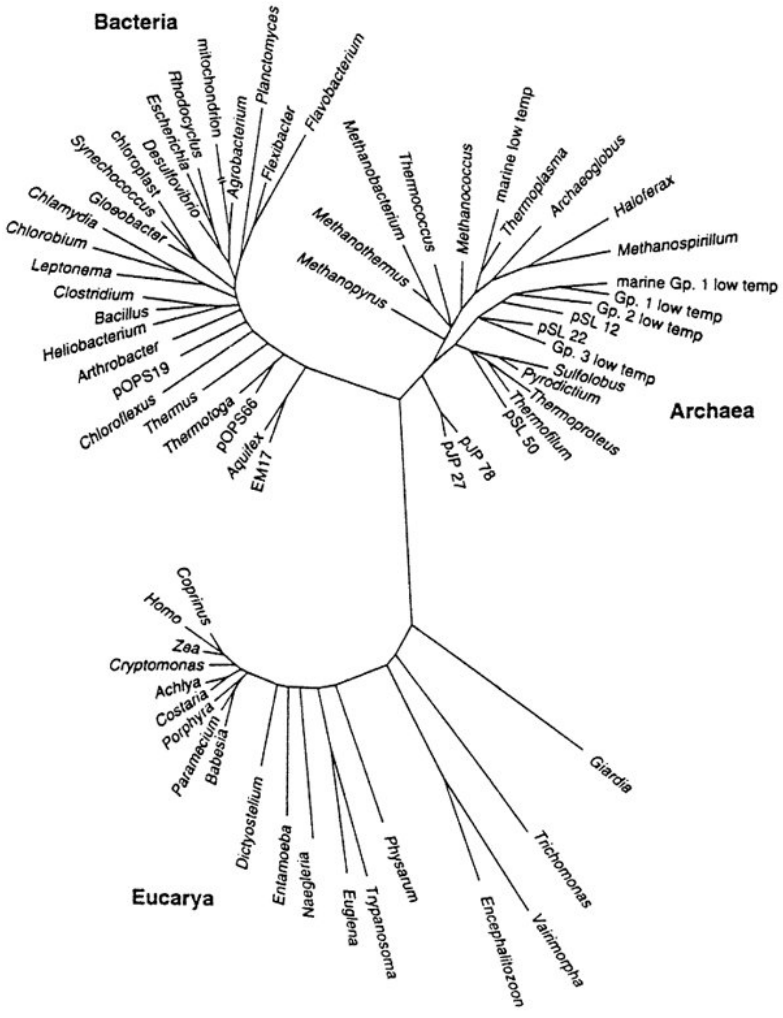
Reálne ukážky stromov z článkov (zakorenený/nezakorenený)

[Margulies et al. 2007]



zakorenený pomocou vonkajšej skupiny (outgroup)

[Pace et al 1997]



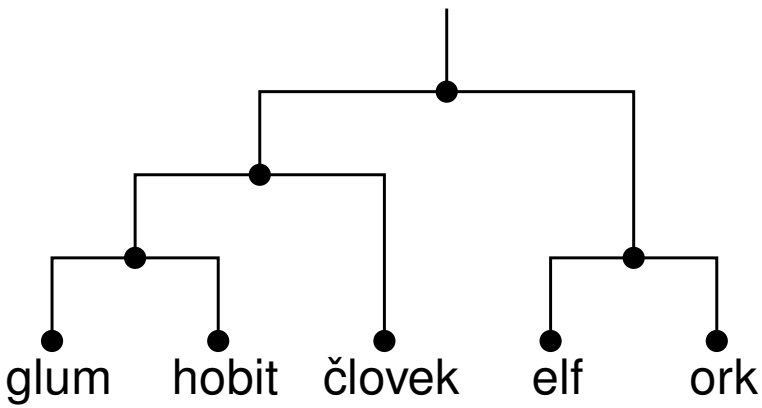
Zopár faktov o stromoch

- 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?)

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).

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).

- *hrana* v tomto prípade zodpovedá rozdeleniu vrcholov na 2 skupiny

Bootstrap

- Máme frekvenciu každej hrany (koľko krát sme ju videli v našich stromoch)
- Zostavíme strom z *celých dát* a pozrieme sa ako často sa ktorá jeho hrana vyskytovala.
- Môžeme zostaviť aj strom z často sa vyskytujúcich hrán.

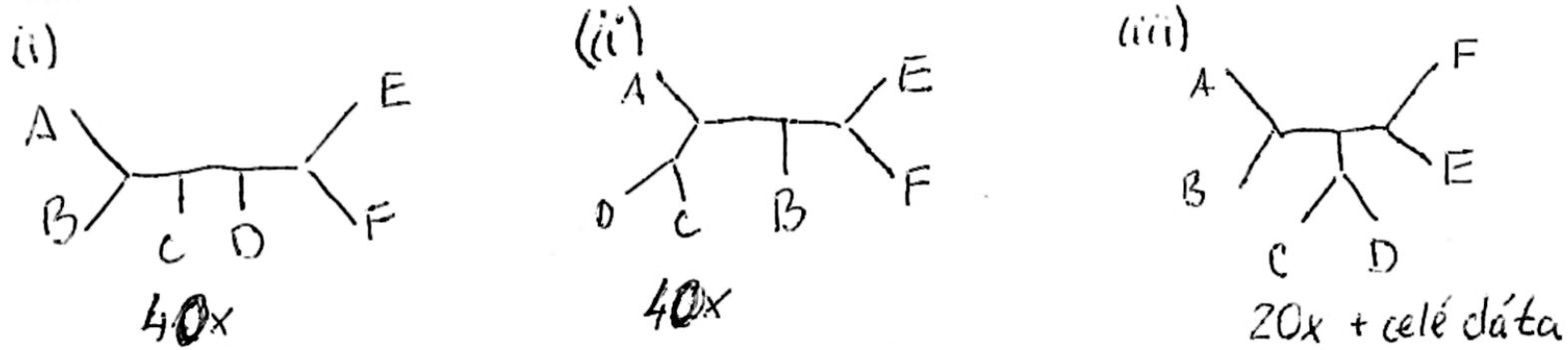
Bootstrap

- Máme frekvenciu každej hrany (koľko krát sme ju videli v našich stromoch)
- 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× 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:
  $Pr(A|A,t) = \frac{1+3e^{-\frac{4}{3}t}}{4}$,
  $P(C|A,t) = \frac{1-e^{-\frac{4}{3}t}}{4}$
- Ak poznáme ancestrálne sekvencie, vieme spočítať pravdepodobnosť dát.
- Ancestrálne sekvencie sú náhodné premenné, ktoré nás nezaujímajú: marginalizujeme ich (uvažujeme všetky ich možné hodnoty).

Zložitejšie evolučné modely

- Jukes-Cantorov model uvažuje len dĺžku hrany udanú ako priemerný počet substitúcií (vrátane tých, ktoré nevidíme)
- Jukes-Cantor predpokladá, že každá mutácia je rovnako pravdepodobná

Zložitejšie evolučné modely

- Jukes-Cantorov model uvažuje len dĺžku hrany udanú ako priemerný počet substitúcií (vrátane tých, ktoré nevidíme)
- Jukes-Cantor predpokladá, že každá mutácia je rovnako pravdepodobná

- Nie všetky substitúcie sa dejú rovnako často:
  tranzície (pyrimidín - pyrimidín, purín - purín) sú pravdepodobnejšie
  ako *tranzverzie* (purín - pyrimidín, pyrimidín - purín)
- Nie všetky nukleotidy sa v danom genóme vyskytujú rovnako často

Od Jukes-Cantorovho modelu ku všeobecnejším modelom mutácií

- Všeobecnenší model:
  zavedieme μ_{xy} *rýchlosť substitúcie* z bázy x na bázu y
- Matica rýchlostí (substitution rate matrix)

$$\begin{pmatrix} -\mu_A & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{CA} & -\mu_C & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & -\mu_G & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & -\mu_T \end{pmatrix}$$

Pre daný čas t, môžeme vypočítať pravdepodobnosť každej substitúcie z bázy x na bázu y (*transition probabilities*):
$\Pr(y = C \mid x = A, t)$

Znižovanie počtu parametrov

- rozlišujeme *tranzície* $C \Leftrightarrow T, A \Leftrightarrow G$ a *transverzie* $\{C, T\} \Leftrightarrow \{A, G\}$
- α rýchlosť tranzícií, β rýchlosť transverzií
- π_Y frekvencia bázy Y v *ekvilibriu*

$$\mu_{XY} = \begin{cases} \alpha\pi_Y & \text{ak } X \Leftrightarrow Y \text{ je tranzícia} \\ \beta\pi_Y & \text{ak } X \Leftrightarrow Y \text{ je transverzia} \end{cases}$$

$$\begin{pmatrix} -\mu_A & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{CA} & -\mu_C & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & -\mu_G & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & -\mu_T \end{pmatrix} \rightarrow \begin{pmatrix} -\mu_A & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & -\mu_C & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & -\mu_G & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & -\mu_T \end{pmatrix}$$

$-\mu_X = -(\text{súčet zvyšku riadku}) \Rightarrow$ aby súčet celého riadku bol 0

Znižovanie počtu parametrov — HKY matica

Hasegawa, Kishino a Yano

$$\begin{pmatrix} -\mu_A & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{CA} & -\mu_C & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & -\mu_G & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & -\mu_T \end{pmatrix} \rightarrow \begin{pmatrix} -\mu_A & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & -\mu_C & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & -\mu_G & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & -\mu_T \end{pmatrix}$$

$$\mu_{x,y} = \begin{cases} \kappa\pi_y & \text{ak } x \Leftrightarrow y \text{ je tranzícia} \\ \pi_y & \text{ak } x \Leftrightarrow y \text{ je transverzia} \end{cases}$$

- frekvencie $\pi_A, \pi_C, \pi_G, \pi_T$ v *ekvilibriu*
- rozlišujeme *tranzície* a *transverzie*
  tranzície sú κ krát častejšie (typicky $\kappa \approx 2$)
  - $\kappa = \alpha/\beta$: pomer rýchlostí ktorými sa dejú tranzície vs transverzie
- Máme iba štyri parametre: $\pi_A, \pi_C, \pi_G, \kappa$
  (π_T sa dopočíta do 1: $\pi_A + \pi_C + \pi_G + \pi_T = 1$)