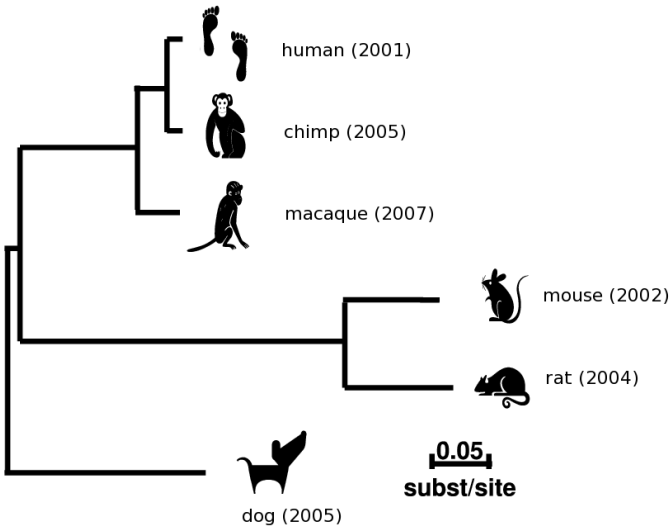
Organizačné poznámky

- Domáca úloha 2 bude zverejnená dnes, dátum odovzdania 29.11. 22:00
  Informatici budú implementovať algoritmus, ktorý bude preberaný budúci týždeň na cvičeniach
- Nezabudnite na **prvé stretnutie** ohľadom journal clubu
  (najneskôr 22.11., osobne alebo online).
  Pred stretnutím oznámte čas a miesto do diskusie skupiny na Moodli
  Po stretnutí napíšte krátku správu do diskusie
  (kto sa zúčastnil, čo sa dohodlo, či sú nejaké problémy, stačí pár viet)

Komparatívna genomika

Tomáš Vinař

9.11.2023



Komparatívna genomika

- Štúdium evolúcie genómov
  - Mutácie jednotlivých báz DNA (táto prednáška)
  - Krátke inzercie a delécie
  - Väčšie udalosti: prestavby genómu, duplikácie
- Typy mutácií:
  - Neutrálne
  - Škodlivé (deleterious)

    $\Rightarrow$ **Purifikačný výber (purifying selection)**
  - Prospešné (advantageous)

    $\Rightarrow$ **Pozitívny výber (positive selection)**
- Na základe porovnávania genómov chceme nájsť oblasti s nezvyčajnou evolučnou históriou (zachovávanie dôležitých funkcií, vývoj nových funkcií)

Komparatívna genomika

- Zostavíme viacnásobné zarovnania genómov

  (zarovnané miesta by mali pochádzať z tej istej sekvencie spoločného predka)

```
     Human AGTGGCTGCCAGGCTG---GGATGCTGAGGCCTTGTTTGCAGGGAGGT
    Rhesus AGTGGCTGCCAGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGAGGT
     Mouse GGTGGCTGCCGGGCTG---GGTGGCTGAGGCCTTGTTGGTGGGGTGGT
       Dog AGTGGCTGCCCGGCTG---GGTGGCTGAGGCCTTATTTGCAGGGAGGT
     Horse GATGGCTGCCGGGCTG---GGCTGCCGAGGCCTTGTTCGTGGGGAGGT
Armadillo AGTGGCTGCCGGGCTG---GGAGGCCAAGGCCTTGTTCGCGGGCAGGT
   Chicken AGTGGCTGCCAGTCTGCGCCGTGGCCGACGTCTTGCTCGGGGGAAGGT
   X. trop AATGGCTTCCATTTTGTGCCGCTGCTGAGGTCTTGTTCTGGGGAAGAT
```

- **Metódy:** Kombinujeme techniky na anotáciu (HMM) a pravdepodobnostné modely evolúcie

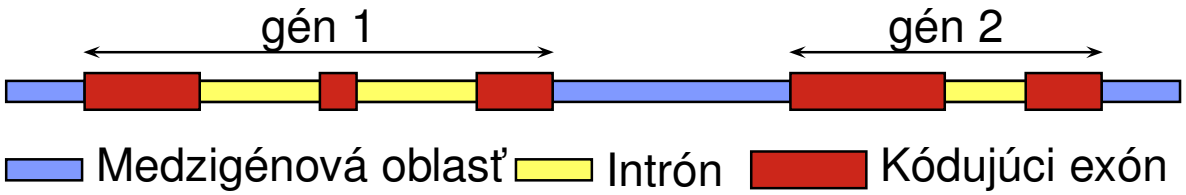
Príklad 1: Hľadanie funkčných oblastí sekvencií

Dôsledky purifikačného výberu:

- Funkčné časti sekvencie zostávajú zachované, menia sa pomalšie
- Nefunkčné sekvencie sa vyvíjajú rýchlejším tempom
- **Príklad:** gény kódujúce proteíny, porovnanie človek myš
  - kódujúce časti: 85% zhoda (zarovnanie na 98% dĺžky)
  - intróny: 69% zhoda (zarovnanie na 48% dĺžky)
- **Úloha:** Hľadáme **nadmerne dobre zachované sekvencie**
- Veľká časť bude zodpovedať známym funkčným elementom (kódujúce gény, regulačné regióny, a pod.)
- Zachované sekvencie ktoré sa neprekrývajú so známymi funkčnými elementami: zaujímavé objekty pre výskum

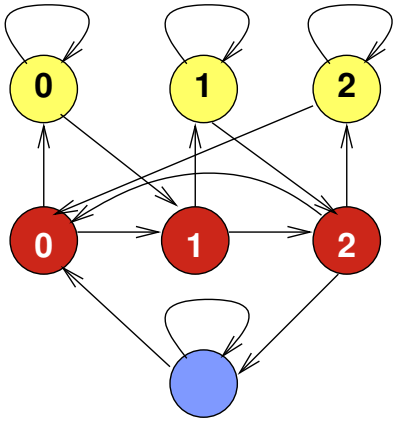
Opakovanie: hľadanie génov

Úlohou je nájsť polohu génov v genóme a ich exónovú štruktúru.



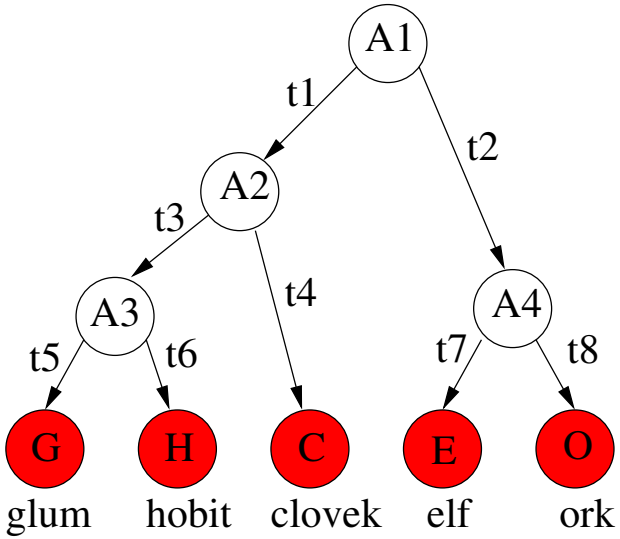
Vytvoríme skrytý Markovovský model (HMM), ktorý vie generovať sekvencie a ich anotácie podobné skutočným.

Pýtame sa, ktorá anotácia je najpravdepodobnejší pár k danej sekvencii.



Opakovanie: pravdepodobnostné modely evolúcie

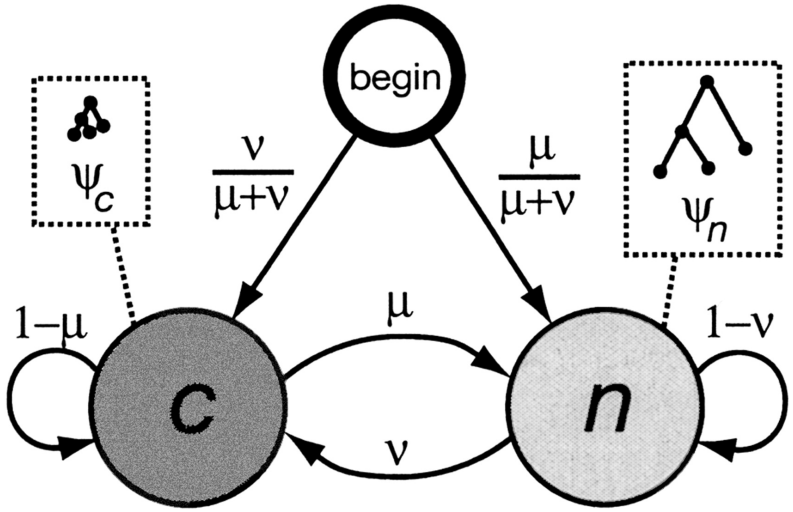
- Strom môžeme chápať ako **jednoduchý generatívny model**

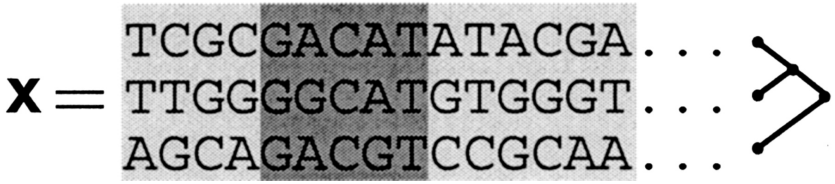


- Pre hranu z Y do X dĺžky t možno pravdepodobnosť mutácie spočítať použitím evolučného modelu, napr. Jukes-Cantor:
  $\Pr(X = C \mid Y = A, t) = \frac{1}{4}(1 - e^{-\frac{4}{3}\alpha t})$
- Pre celý strom $\Pr(G, H, C, E, O, A1, \ldots, A4) = \Pr(A1) \cdot \Pr(A2 \mid A1, t_1) \cdot \Pr(A4 \mid A1, t_2) \cdot \Pr(A3 \mid A2, t_3) \cdot \Pr(G \mid A3, t_5) \cdot \Pr(H \mid A3, t_6) \cdot \Pr(C \mid A2, t_4) \cdot \Pr(E \mid A4, t_7) \cdot \Pr(O \mid A4, t_8)$

PhastCons: detekcia dobre zachovaných sekvencií

Fylogenetické HMM: kombinácia HMM a fylogenetického stromu.

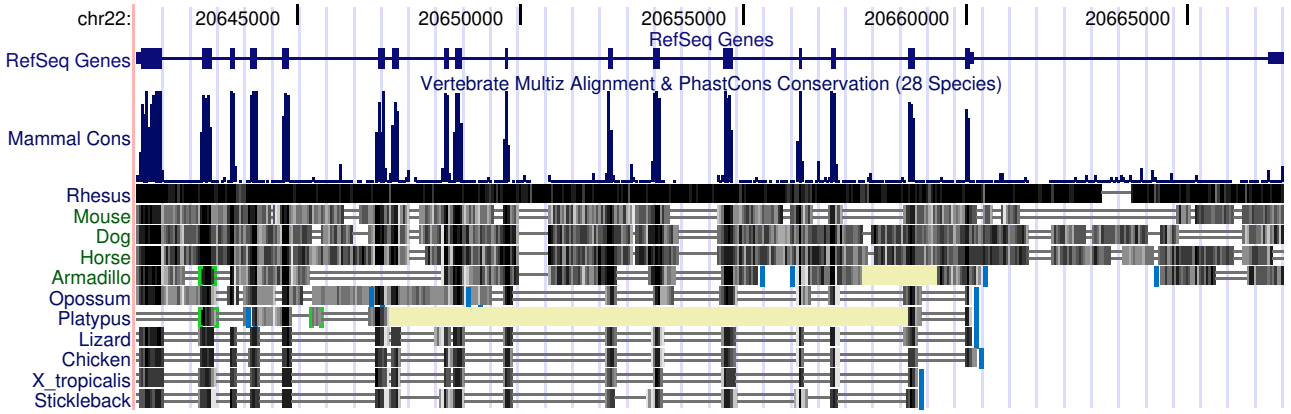




- Dva stavy: zachovaná sekv., neutrálna sekv.
- V každom stave generujeme celý stĺpec zarovnania
- Zachovaná sekvencia má kratšie hrany stromu, teda menšia divergencia sekvencií

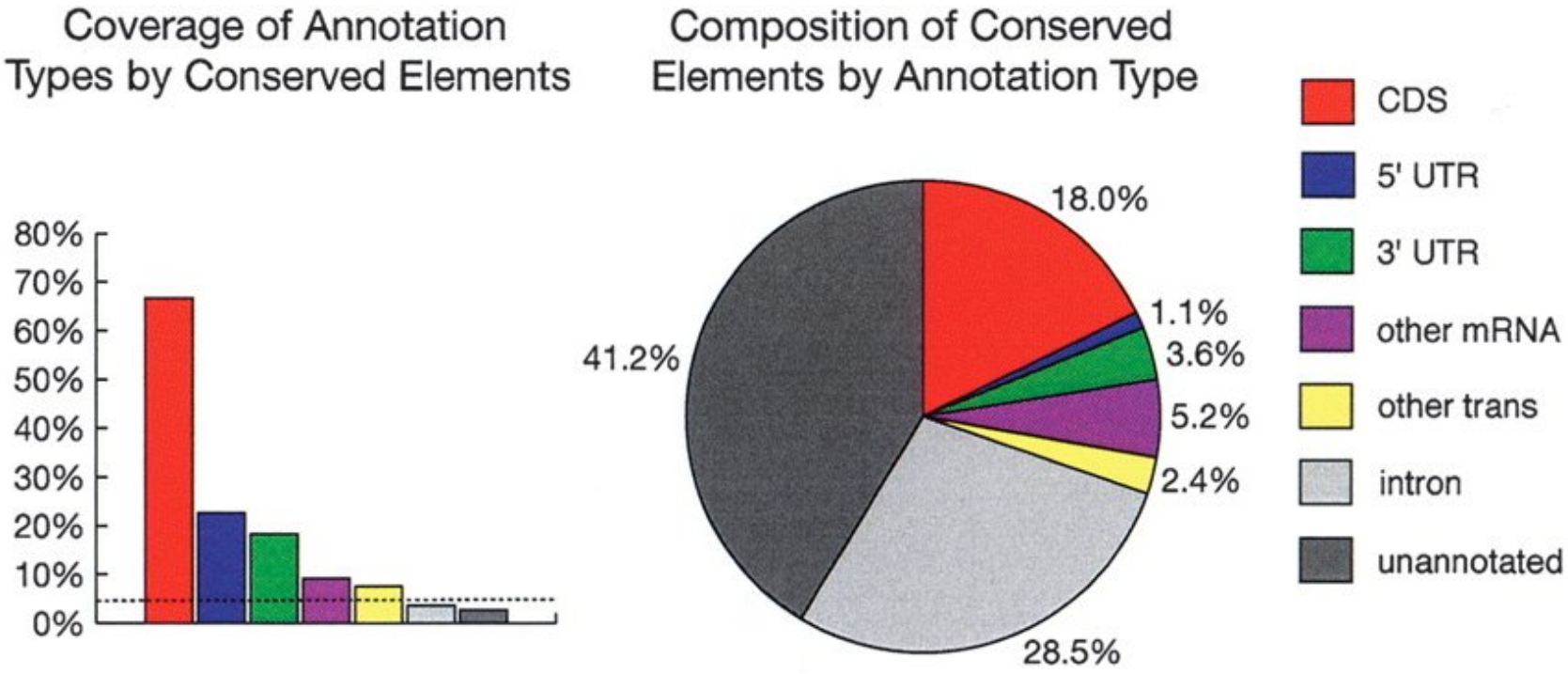
Použitie fylogenetického HMM

- Model určuje rozdelenie pravdepodobnosti cez zarovnania a anotácie (tu: anotácia = označenie zachovaných sekvencií)
- Pre dané zarovnanie hľadáme najpravdepodobnejšiu anotáciu
- Kombinácia Viterbiho a Felsensteinovho algoritmu



Výsledky celogenómovej aplikácie PhastCons-u

Zarovnania genómov človeka, myši, sliepky, fugu



Fylogenetické HMM pre hľadanie génov

- Použijeme stavy z hľadača génov
- Pre každý stav máme evolučný model
  (maticu rýchlostí, dĺžky hrán)
- Trojperiodickosť frekvencií mutácií pomáha nájsť gény

Ako veľmi pomôžu zarovnania zlepšiť presnosť

	Exóny		Gény	
Program	sn	sp	sn	sp
AUGUSTUS (1 genóm)	52%	63%	24%	17%
NSCAN (zarovnanie)	68%	82%	35%	37%

Guigo et al 2006, evaluácia na 1% ľudského genómu

Genetický kód

Ala / A	GCT, GCC, GCA, GCG	Leu / L	TTA, TTG, CTT, CTC, CTA, CTG
Arg / R	CGT, CGC, CGA, CGG, AGA, AGG	Lys / K	AAA, AAG
Asn / N	AAT, AAC	Met / M	ATG
Asp / D	GAT, GAC	Phe / F	TTT, TTC
Cys / C	TGT, TGC	Pro / P	CCT, CCC, CCA, CCG
Gln / Q	CAA, CAG	Ser / S	TCT, TCC, TCA, TCG, AGT, AGC
Glu / E	GAA, GAG	Thr / T	ACT, ACC, ACA, ACG
Gly / G	GGT, GGC, GGA, GGG	Trp / W	TGG
His / H	CAT, CAC	Tyr / Y	TAT, TAC
Ile / I	ATT, ATC, ATA	Val / V	GTT, GTC, GTA, GTG
START	ATG	STOP	TAA, TGA, TAG

Príklad 2: Hľadanie génov pod vplyvom pozitívneho výberu

- **Pozitívny výber** = proces, ktorým sa v genóme ustália **prospešné mutácie**
- Neobvykle vysoké množstvo mutácií, ktoré by mohli súvisieť so zmenou funkcie
- V rámci génov, ktoré kódujú proteíny:
  - **Synonymné mutácie** nemenia zakódovanú aminokyselinu
    napr. ACA (Thr) $\Rightarrow$ ACT (Thr)
  - **Nesynonymné mutácie** menia zakódovanú aminokyselinu
    napr. ACA (Thr) $\Rightarrow$ AAA (Lys)
- Vytvoríme pravdepodobnostný model evolúcie, ktorý bude rozlišovať synonymné a nesynonymné mutácie $\Rightarrow$ identifikácia sekvencií s neobvykle vysokým podielom nesynonymných mutácií

Od Jukes-Cantorovho modelu ku všeobecnejším modelom mutácií

- Jukes-Cantor predpokladá, že každá mutácia rovnako pravdepodobná
- 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 — HKY matica

Hasegawa, Kishino a Yano

$$\begin{pmatrix} -\mu_A & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & -\mu_C & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & -\mu_G & \kappa\pi_T \\ \pi_A & \kappa\pi_C & \pi_G & -\mu_T \end{pmatrix} \qquad \mu_{x,y} = \begin{cases} \kappa\pi_y & \text{ak } x \Leftrightarrow y \text{ je tranzícia} \\ \pi_y & \text{ak } x \Leftrightarrow y \text{ je transverzia} \end{cases}$$

- **ekvilibrium:** frekvencie $\pi_A, \pi_C, \pi_G, \pi_T$
- rozlišujeme **tranzície** $C \Leftrightarrow T, A \Leftrightarrow G$ a **transverzie** $\{C, T\} \Leftrightarrow \{A, G\}$
  tranzície sú κ krát častejšie (typicky $\kappa \approx 2$)
- Máme iba štyri parametre: $\pi_A, \pi_C, \pi_G, \kappa$
  (π_T sa dopočíta do 1)

Substitučný model pre kodóny

Namiesto jednotlivých báz uvažujeme trojice

Rýchosť zmeny z kodónu i na kodón j:

$$\mu_{i,j} = \begin{cases} 0, & \text{ak sa } i, j \text{ líšia na } > 1 \text{ pozíciách,} \\ \kappa\pi_j, & \text{synonymné tranzície,} \\ \pi_j, & \text{synonymné transverzie,} \\ \omega\kappa\pi_j, & \text{nesynonymné tranzície,} \\ \omega\pi_j, & \text{nesynonymné transverzie.} \end{cases}$$

Príklad: $\mu_{AAC,GGC} = 0$, $\mu_{CTA,CTT} = \pi_{CTT}$, $\mu_{CTA,CCA} = \omega\kappa\pi_{CCA}$

Parametre: Frekvencie kodónov π_j, ω, κ

neutrálna evolúcia $\omega = 1$, pozitívny výber $\omega > 1$,

purifikačný výber $\omega < 1$

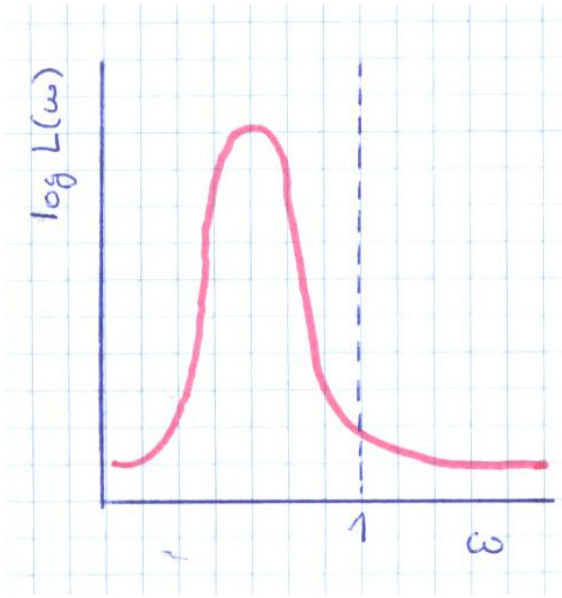
Aplikácia kodónového substitučného modelu

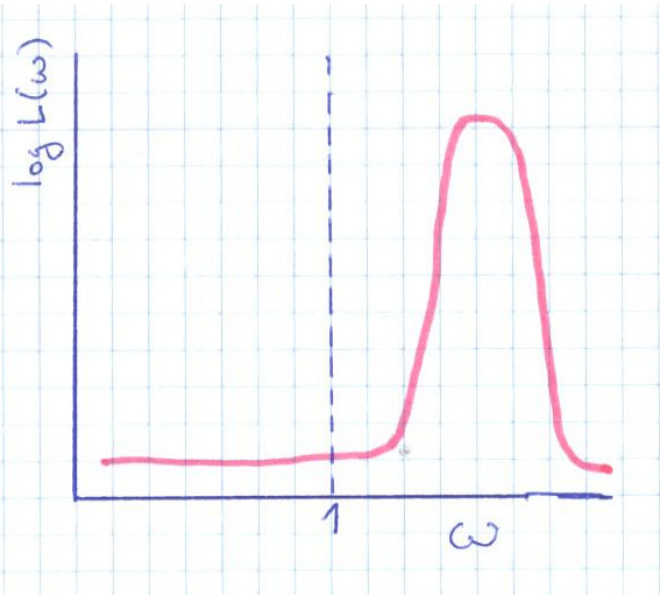
```
          F   V   I   H   D   S   E   G   D   G   E   C   M   Q   E
človek  TTT GTG ATC CAC GAC TCC GAG GGG GAC GGC GAG TGC ATG CAG GAG
kosmáč  TTT GTG ATC CAC GAG AAC AAC AAG GAC GGC GAG TGC ATG CAG GAT
          F   V   I   H   E   N   N   K   D   G   E   C   M   Q   D
```

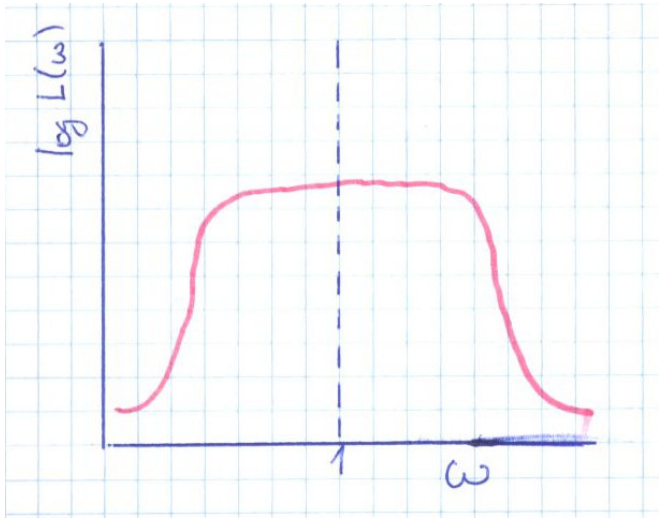
- Na základe celých genómov môžeme odhadnúť základné parametre modelu π_*, κ
- Pre dané ω a t vieme spočítať vierohodnosť

$$L(\omega, t) = \Pr(C, K \,|\, \omega, t)$$

- Sledujeme, ako sa mení $L(\omega) = \max_t L(\omega, t)$ pre rôzne hodnoty ω


log L(ω)
1
ω


log L(ω)
1
ω


log L(ω)
1
ω

Test pomerov vierohodností (Likelihood-ratio test)

- $L(\omega)$ môže byť najväčšie pre $\omega > 1$,
  ale môže to byť spôsobené len štatistickou varianciou v dátach
  $\Rightarrow$ potrebujeme štatistický test
- Spočítame vierohodnosť $L_A = \max_{\omega<1} L(\omega)$
- Spočítame vierohodnosť $L_B = \max_{\omega} L(\omega)$ (bez obmedzenia ω)
- Vždy platí $L_B \geq L_A$
- Ak skutočné $\omega < 1$, $L_A \approx L_B$ (nulová hypotéza)
  nás zaujímajú prípady $L_B >> L_A$
  $\Rightarrow$ gén pod vplyvom pozitívneho výberu (alt. hypotéza)

Za predpokladu, že $\omega < 1$, platí $2\log(L_B/L_A) \approx \chi^2_1$
$\Rightarrow$ možno priradiť P-hodnotu nulovej hypotéze $\omega < 1$

Hľadanie génov pod vplyvom pozitívneho výberu: Zhrnutie

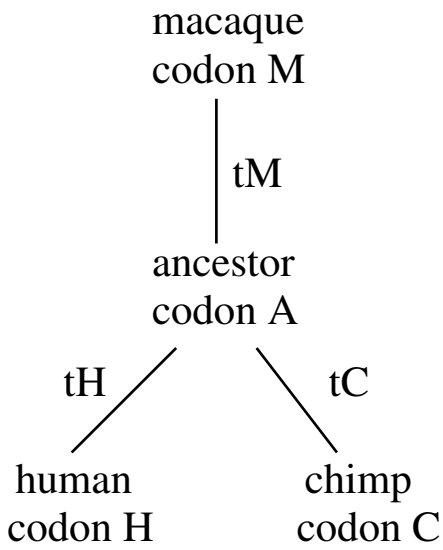
- Nájdeme zarovnanie toho istého génu z dvoch organizmov (na úrovni kodónov)
- Odhadneme základné parametre kodónového modelu na základe porovnania celých genómov
- Parameter ω modeluje selekciu
- Spočítame vierohodnosť $L_A = \max_{\omega<1} L(\omega)$
  a vierohodnosť $L_B = \max_{\omega} L(\omega)$
- Na základe štatistiky $2\log(L_B/L_A)$ priradíme P-hodnotu nulovej hypotéze $\omega < 1$
- Gény s malou P-hodnotou sú pod vplyvom pozitívneho výberu

“Jednoducho” rozšíriteľné na porovnanie viacerých organizmov

$$\Pr(A, H, C, M \mid \omega, t_H, t_C, t_M) =$$
$$\pi_A \cdot \Pr(H \mid A, t_H) \cdot \Pr(C \mid A, t_C) \cdot \Pr(M \mid A, t_M)$$

Zbavíme sa ancestrálnych sekvencií:

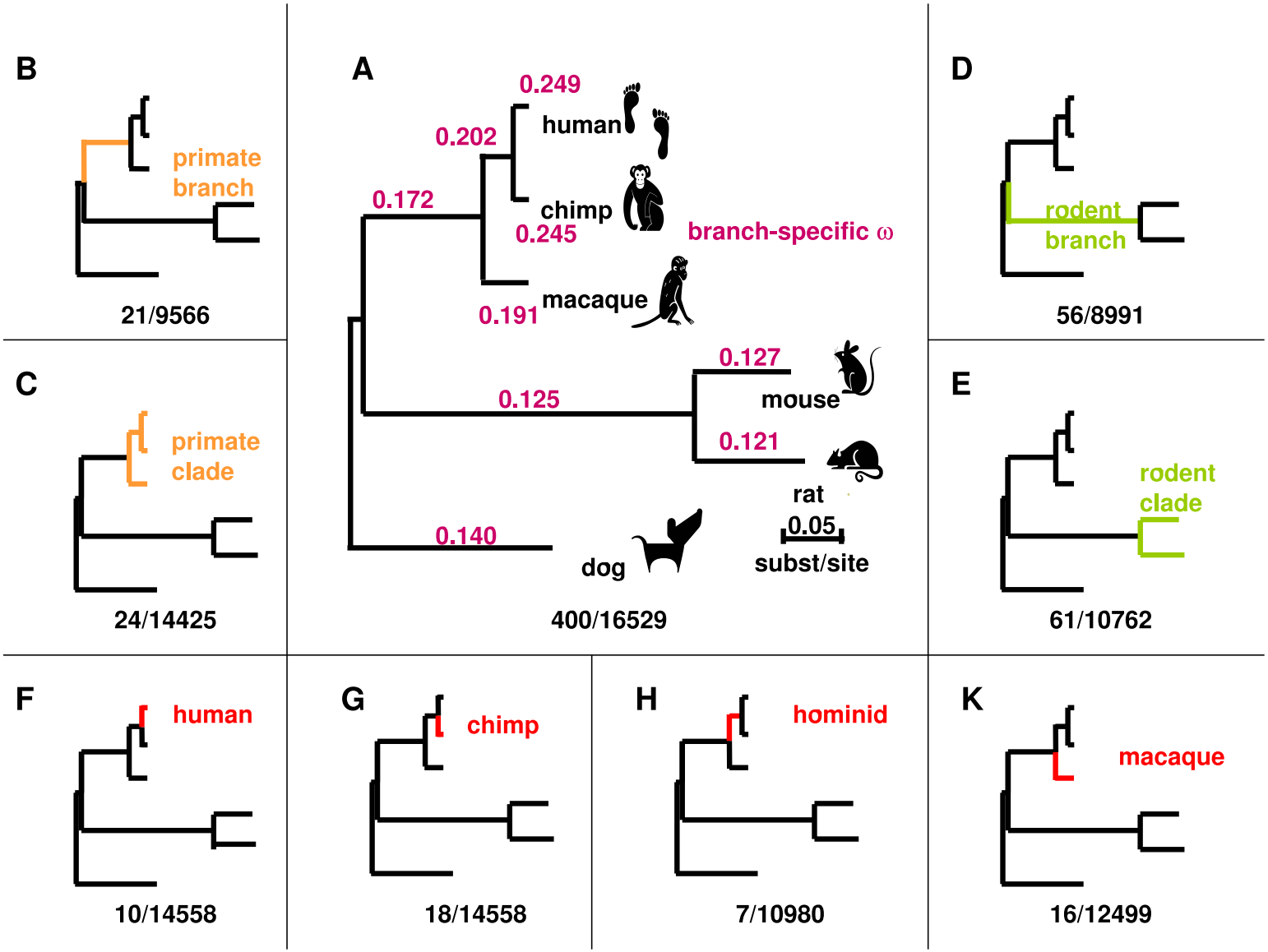
$$\Pr(H, C, M \mid \omega, t_H, t_C, t_M) =$$
$$\sum_A \Pr(A, H, C, M \mid \omega, t_H, t_C, t_M)$$



Vierohodnosť ω:

$$L(\omega) = \max_{t_H, t_C, t_M} \Pr(H, C, M \mid \omega, t_H, t_C, t_M)$$

- Existuje program PAML, ktorý takúto vierohodnosť počíta
- K dispozícii zložitejšie modely, napr. s meniacim sa ω v rámci génu


B
primate branch
21/9566
A
0.249
human
0.202
chimp
0.172
0.245
branch-specific ω
macaque
0.191
0.127
mouse
0.125
0.121
rat
0.05
subst/site
0.140
dog
400/16529
D
rodent branch
56/8991
C
primate clade
24/14425
E
rodent clade
61/10762
F
human
10/14558
G
chimp
18/14558
H
hominid
7/10980
K
macaque
16/12499

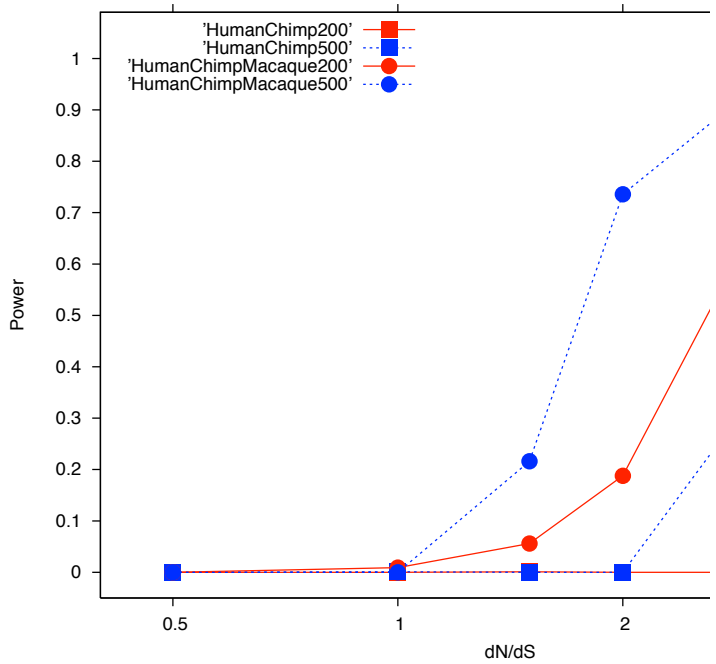
Funkčné kategórie obohatené o gény s pozitívnym výberom

Defense: cellular defense response, antigen processing and presentation, response to virus, response to bacterium

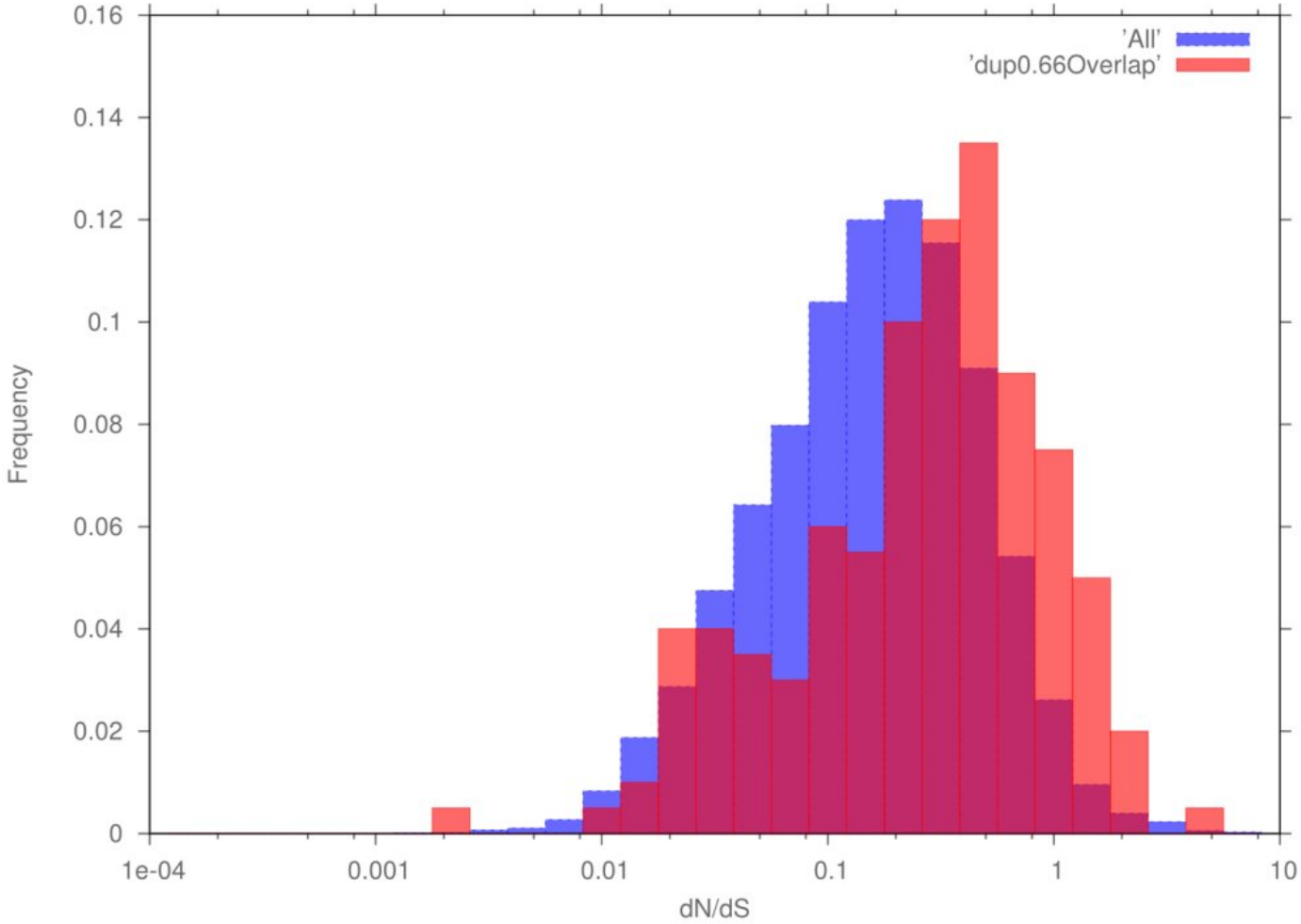
Immunity: adaptive immune response, adaptive immune response somatic recomb, lymphocyte mediated immunity, immunoglobulin mediated immune response, B cell mediated immunity, innate immune response, complement activation alternative pathway, regulation of immune system process, positive regulation of immune response, humoral immune response, complement activation classical pathway, humoral immune response circulating immunoglob, complement activation, activation of plasma proteins mute inflam resp, akute inflammatory response, response to wounding

Sensory perception: sensory perception of taste, G-protein coupled receptor protein signaling pathway, neurological process, sensory perception of chemical stimulus, sensory perception of smell

Viacej genómov pomáha vylepšiť účinnosť testov



Pozitívny výber v duplikovaných génoch



Zhrnutie

- Prirodzený výber má významnú úlohu v evolúcii
- **Purifikačný výber:**
  - Zachované regióny majú s veľkou pravdepodobnosťou nejakú funkciu
  - Pri hľadaní génov berieme do úvahy aj typické mutácie kodónov
- **Pozitívny výber:**
  - Pozitívny výber v génoch sa prejavuje veľkým pomerom nesynonymných zmien (evolúcia na proteínovej úrovni)
  - Zduplikované gény sú častejšie pod vplyvom pozitívneho výberu
  - Poľovačka pokračuje: hľadáme gény spôsobujúce charakteristické črty človeka
- **Metódy:** evolučné modely, fylogenetické HMM, test pomerov vierohodností