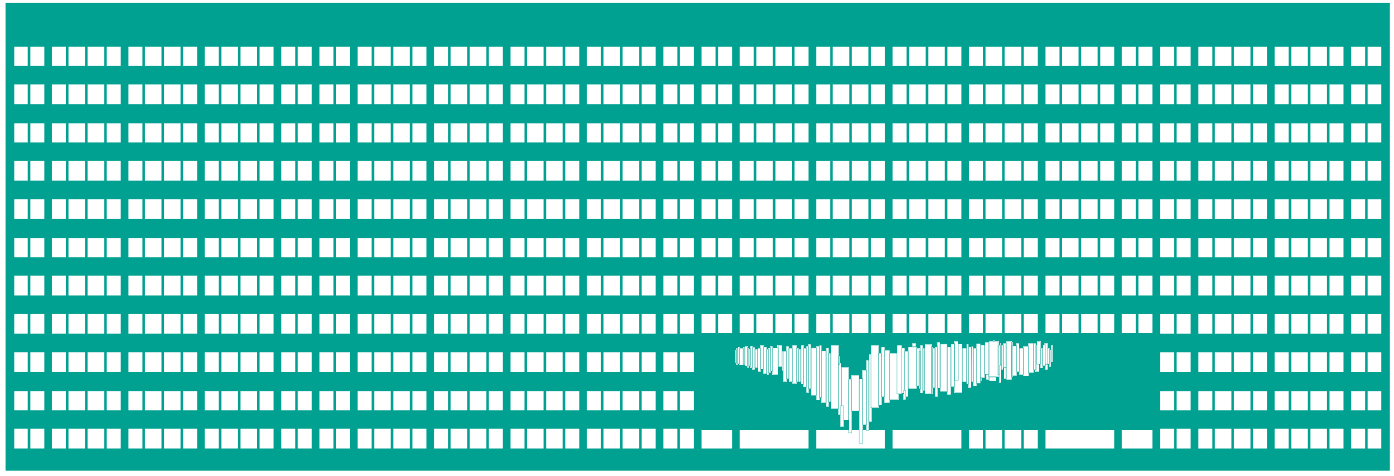
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

Algoritmy pro Bioinformatiku

Fylogenetická analýza

Michal Vašinek

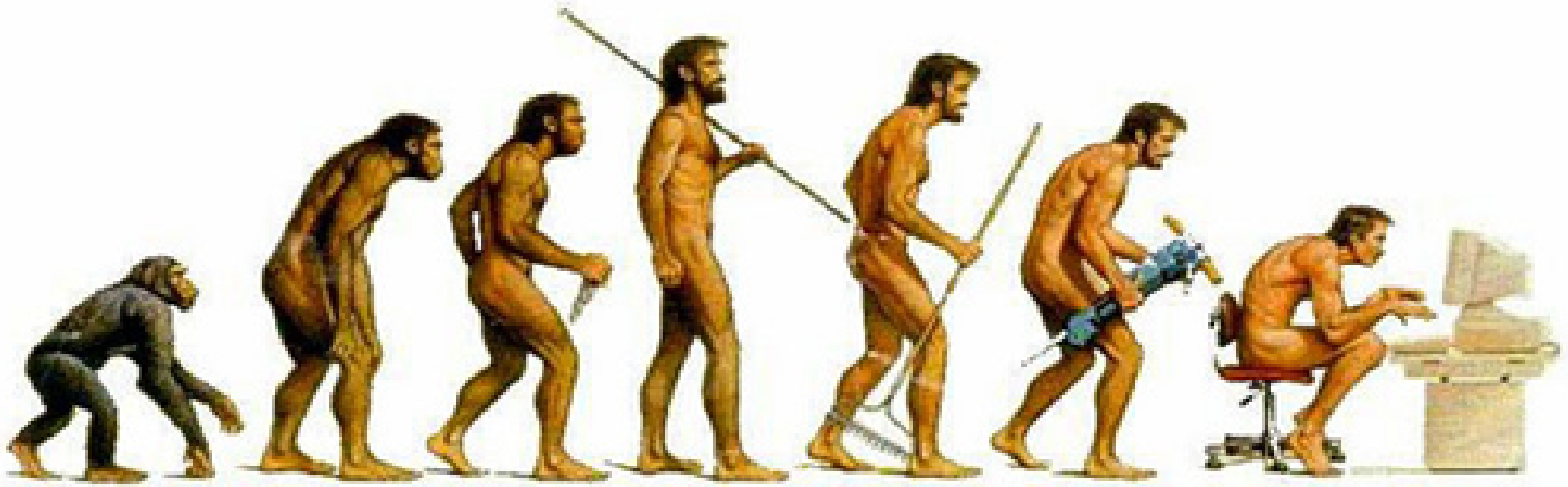
VŠB – Technická univerzita Ostrava

FEI/EA404

michal.vasinek@vsb.cz

23. listopadu 2022

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY



Přehled přednášky

- Úvod do problematiky
- Adam a Eva
- Parsimonie
- Metody založené na vzdálenosti

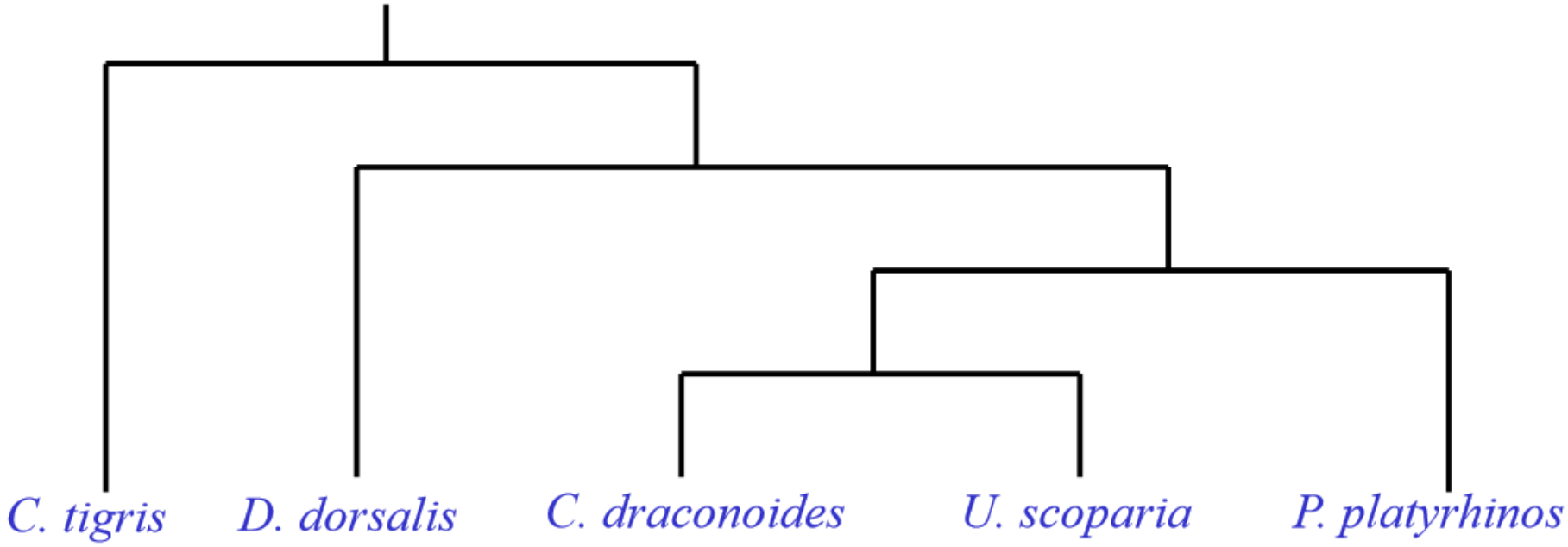
Evoluce

- DNA kóduje informace o živém organismu.
- Organismy předávají informaci v DNA svým potomkům.
- V důsledku mutací se DNA v průběhu času mění.
- V rámci milionů let se vyvinuly různé druhy ogranismů.
- Fylogenetika se zabývá studiem genetických vztahů mezi různými druhy.

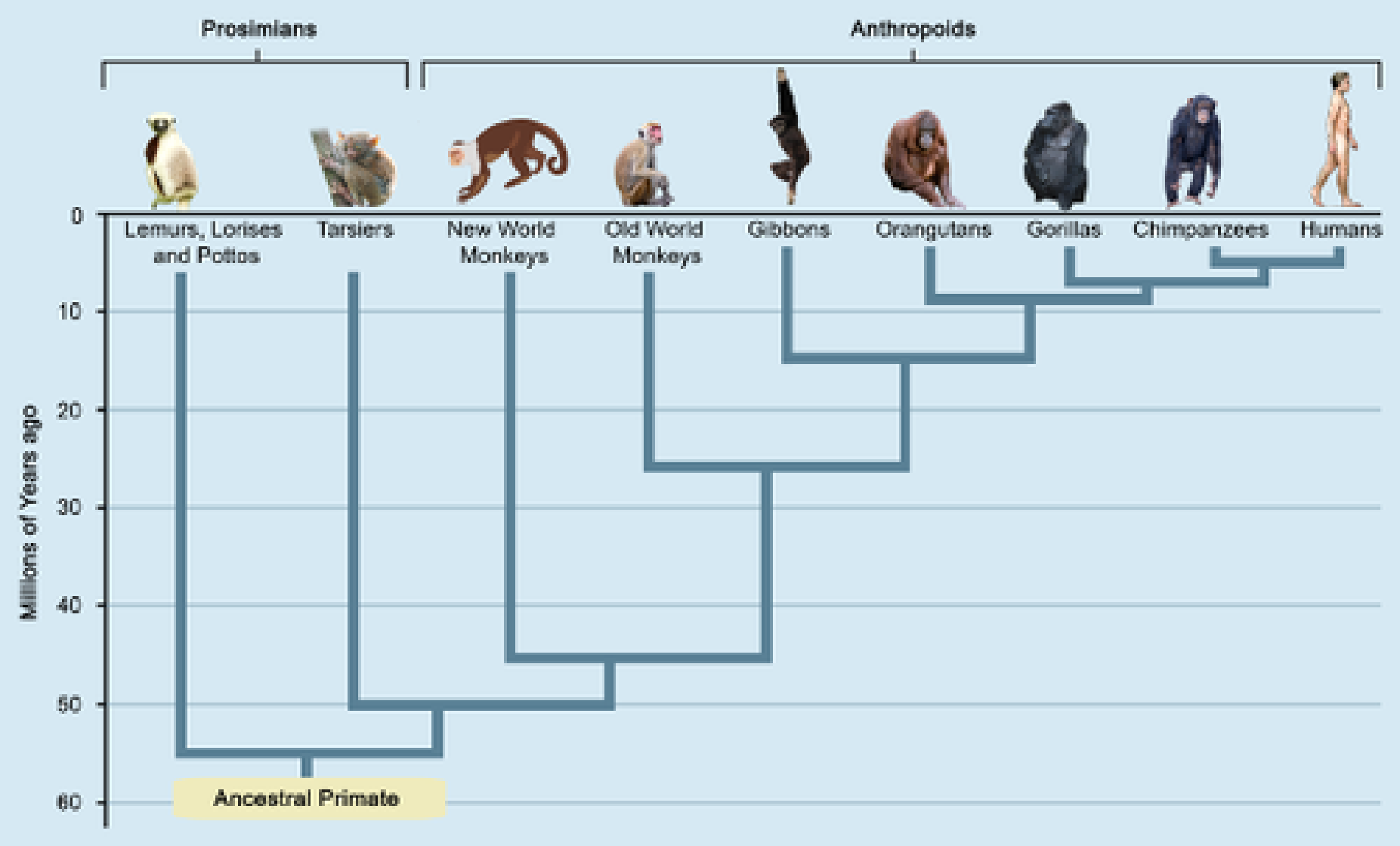
Motivace

- Porozumění evoluční historii organismů.
- Porozumění tomu, jak se vyvinuly u organismů různé funkce.
- K mapování mutací ve virech.
- K porozumění, co jsou klíčové stavební prvky organismů. Identifikovat je napříč organismy.

Příklady



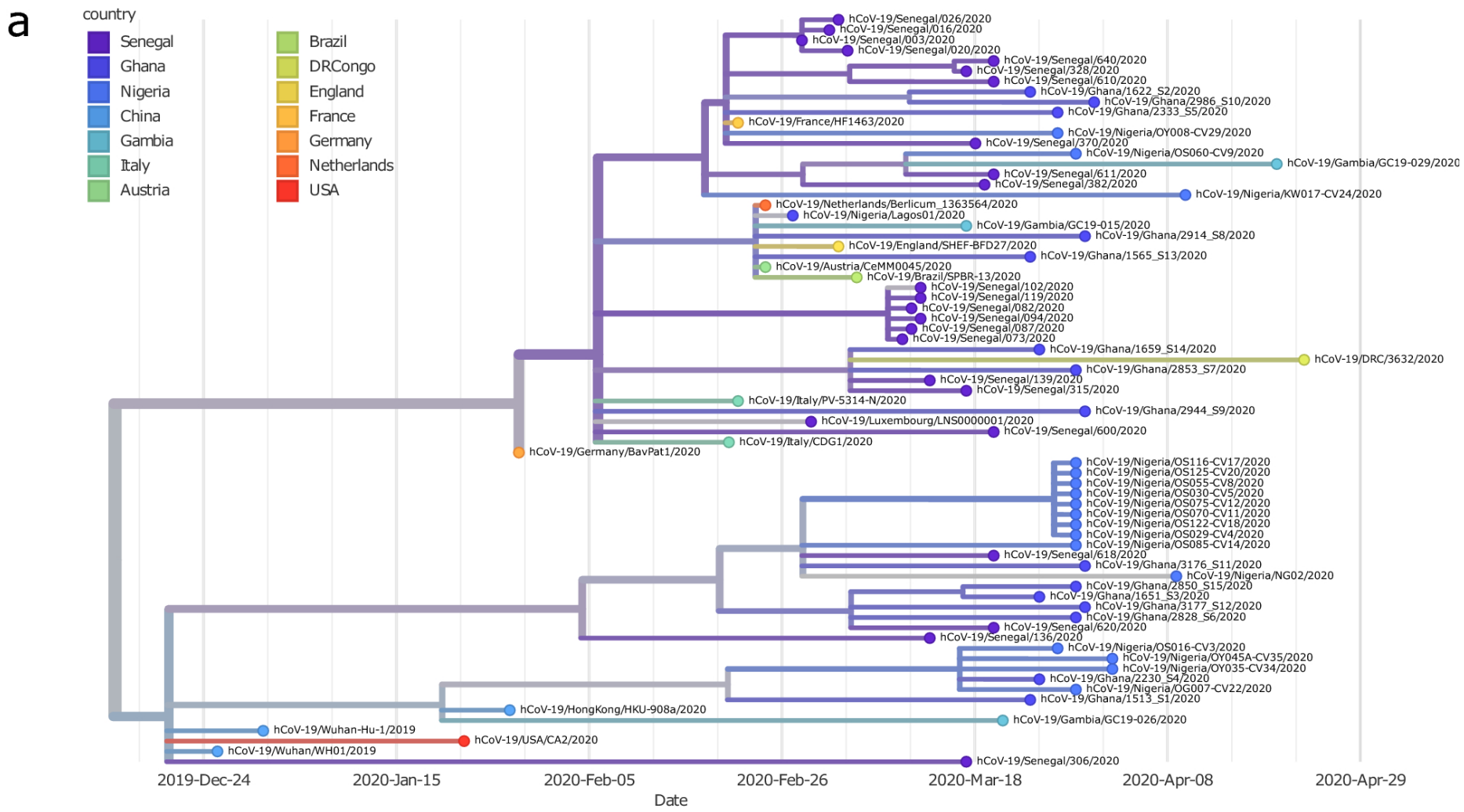
Příklady


Prosimians
Anthropoids
Lemurs, Lorises and Pottos
Tarsiers
New World Monkeys
Old World Monkeys
Gibbons
Orangutans
Gorillas
Chimpanzees
Humans
Millions of Years ago
0
10
20
30
40
50
60
Ancestral Primate

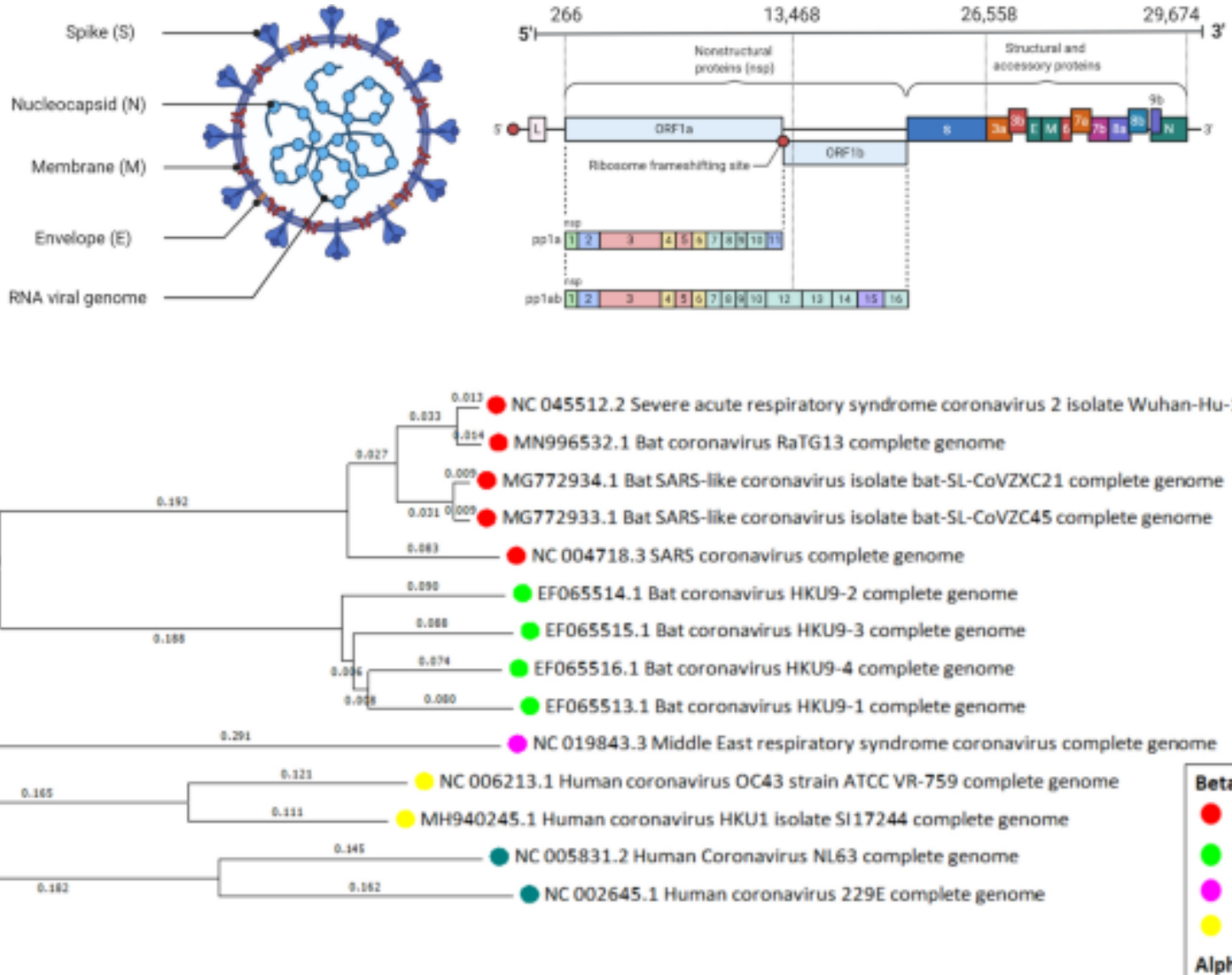
Fylogenetika a epidemiologie

- Chřipková vakcína, založena na studiu vývoje genomu chřipkového viru.
- Studium rapidního vývoje viru HIV.
- Mutace SARS-COV2.
- Fylogenetika umožňuje zjistit zdroj infekce - konkrétní organismus.

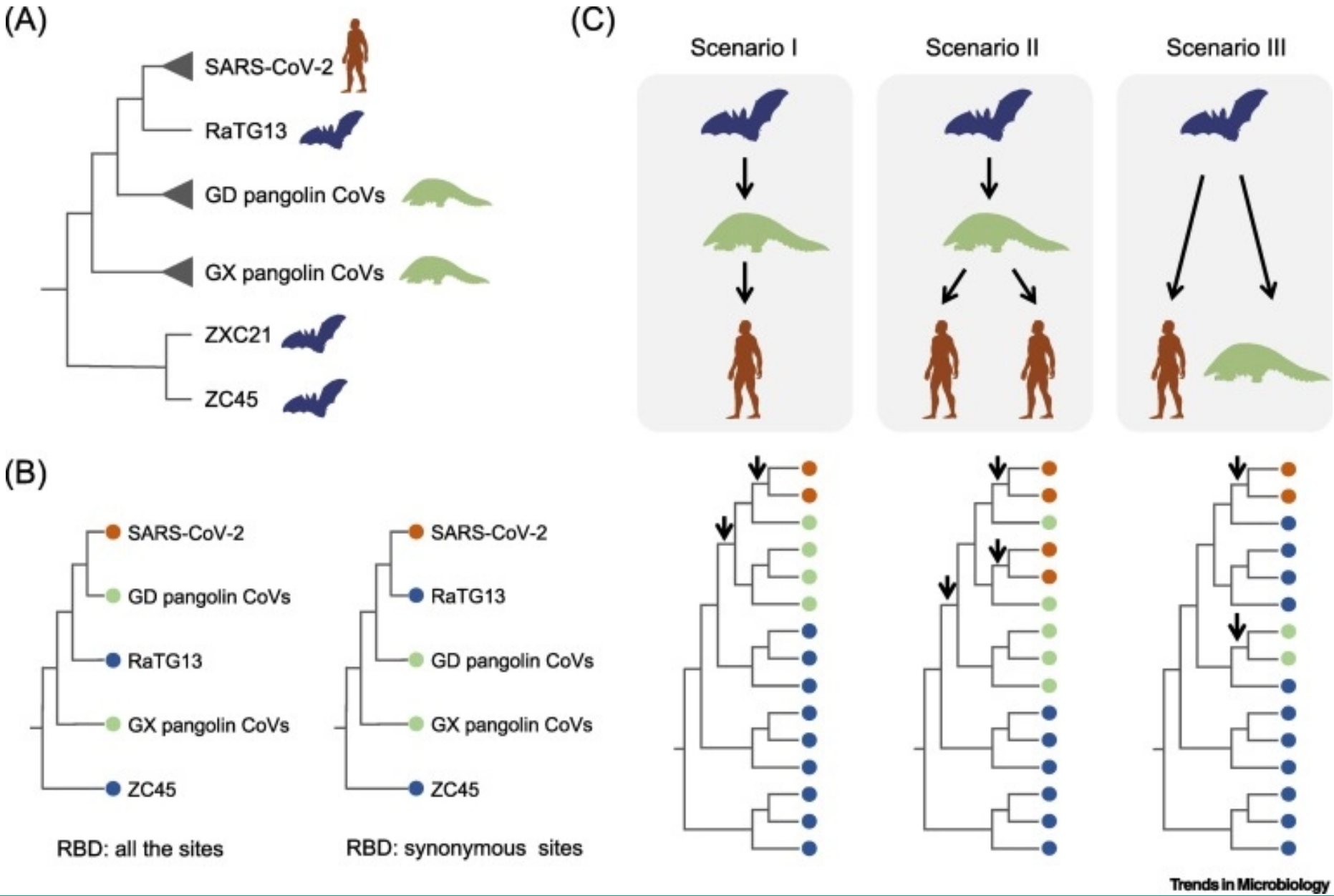
Příklady


a
country
Senegal
Ghana
Nigeria
China
Gambia
Italy
Austria
Brazil
DRCongo
England
France
Germany
Netherlands
USA
hCoV-19/Senegal/026/2020
hCoV-19/Senegal/016/2020
hCoV-19/Senegal/003/2020
hCoV-19/Senegal/020/2020
hCoV-19/Senegal/640/2020
hCoV-19/Senegal/328/2020
hCoV-19/Senegal/610/2020
hCoV-19/Ghana/1622_S2/2020
hCoV-19/Ghana/2986_S10/2020
hCoV-19/Ghana/2333_S5/2020
hCoV-19/France/HF1463/2020
hCoV-19/Nigeria/OY008-CV29/2020
hCoV-19/Senegal/370/2020
hCoV-19/Nigeria/OS060-CV9/2020
hCoV-19/Gambia/GC19-029/2020
hCoV-19/Senegal/611/2020
hCoV-19/Senegal/382/2020
hCoV-19/Nigeria/KW017-CV24/2020
hCoV-19/Netherlands/Berlicum_1363564/2020
hCoV-19/Nigeria/Lagos01/2020
hCoV-19/Gambia/GC19-015/2020
hCoV-19/Ghana/2914_S8/2020
hCoV-19/England/SHEF-BFD27/2020
hCoV-19/Ghana/1565_S13/2020
hCoV-19/Austria/CeMM0045/2020
hCoV-19/Brazil/SPBR-13/2020
hCoV-19/Senegal/102/2020
hCoV-19/Senegal/119/2020
hCoV-19/Senegal/082/2020
hCoV-19/Senegal/094/2020
hCoV-19/Senegal/087/2020
hCoV-19/Senegal/073/2020
hCoV-19/Ghana/1659_S14/2020
hCoV-19/DRC/3632/2020
hCoV-19/Ghana/2853_S7/2020
hCoV-19/Senegal/139/2020
hCoV-19/Senegal/315/2020
hCoV-19/Italy/PV-5314-N/2020
hCoV-19/Ghana/2944_S9/2020
hCoV-19/Luxembourg/LNS0000001/2020
hCoV-19/Senegal/600/2020
hCoV-19/Italy/CDG1/2020
hCoV-19/Germany/BavPat1/2020
hCoV-19/Nigeria/OS116-CV17/2020
hCoV-19/Nigeria/OS125-CV20/2020
hCoV-19/Nigeria/OS055-CV8/2020
hCoV-19/Nigeria/OS030-CV5/2020
hCoV-19/Nigeria/OS075-CV12/2020
hCoV-19/Nigeria/OS070-CV11/2020
hCoV-19/Nigeria/OS122-CV18/2020
hCoV-19/Nigeria/OS029-CV4/2020
hCoV-19/Nigeria/OS085-CV14/2020
hCoV-19/Senegal/618/2020
hCoV-19/Ghana/3176_S11/2020
hCoV-19/Nigeria/NG02/2020
hCoV-19/Ghana/2850_S15/2020
hCoV-19/Ghana/1651_S3/2020
hCoV-19/Ghana/3177_S12/2020
hCoV-19/Ghana/2828_S6/2020
hCoV-19/Senegal/620/2020
hCoV-19/Senegal/136/2020
hCoV-19/Nigeria/OS016-CV3/2020
hCoV-19/Nigeria/OY045A-CV35/2020
hCoV-19/Nigeria/OY035-CV34/2020
hCoV-19/Ghana/2230_S4/2020
hCoV-19/Nigeria/OG007-CV22/2020
hCoV-19/Ghana/1513_S1/2020
hCoV-19/HongKong/HKU-908a/2020
hCoV-19/Gambia/GC19-026/2020
hCoV-19/Wuhan-Hu-1/2019
hCoV-19/USA/CA2/2020
hCoV-19/Wuhan/WH01/2019
hCoV-19/Senegal/306/2020
2019-Dec-24
2020-Jan-15
2020-Feb-05
2020-Feb-26
2020-Mar-18
2020-Apr-08
2020-Apr-29
Date
b
0.8
0.6
0.4

Příklady


Spike (S)
Nucleocapsid (N)
Membrane (M)
Envelope (E)
RNA viral genome
266
13,468
26,558
29,674
5'
3'
Nonstructural proteins (nsp)
Structural and accessory proteins
ORF1a
ORF1b
Ribosome frameshifting site
pp1a
pp1ab
NC 045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 complete genome
MN996532.1 Bat coronavirus RaTG13 complete genome
MG772934.1 Bat SARS-like coronavirus isolate bat-SL-CoVZXC21 complete genome
MG772933.1 Bat SARS-like coronavirus isolate bat-SL-CoVZC45 complete genome
NC 004718.3 SARS coronavirus complete genome
EF065514.1 Bat coronavirus HKU9-2 complete genome
EF065515.1 Bat coronavirus HKU9-3 complete genome
EF065516.1 Bat coronavirus HKU9-4 complete genome
EF065513.1 Bat coronavirus HKU9-1 complete genome
NC 019843.3 Middle East respiratory syndrome coronavirus complete genome
NC 006213.1 Human coronavirus OC43 strain ATCC VR-759 complete genome
MH940245.1 Human coronavirus HKU1 isolate SI17244 complete genome
NC 005831.2 Human Coronavirus NL63 complete genome
NC 002645.1 Human coronavirus 229E complete genome
0.05
Beta Coronaviruses
Lineage B
Lineage D
Lineage C
Lineage A
Alpha Coronaviruses
Outgroup

Příklady


(A)
SARS-CoV-2
RaTG13
GD pangolin CoVs
GX pangolin CoVs
ZXC21
ZC45
(B)
SARS-CoV-2
GD pangolin CoVs
RaTG13
GX pangolin CoVs
ZC45
RBD: all the sites
SARS-CoV-2
RaTG13
GD pangolin CoVs
GX pangolin CoVs
ZC45
RBD: synonymous sites
(C)
Scenario I
Scenario II
Scenario III
Trends in Microbiology

Historické poznámky

- Do poloviny 50. let se fylogenetický vývoj hodnotil na základě subjektivního hodnocení expertů.
- Počátkem 60. let snaha o nalezení objektivních kritérií pro konstrukci fylogenetických vazeb.
- Fylogenetika se stala nedílnou součástí biologie:
  - Klasifikace organismů.
  - Porozumění biologickým mechanismům.

Morfologické vs. Molekulární členění

- Klasická fylogenetická analýza používala morfologické rysy organismů:
  - Počet končetin, velikost, hmotnost, receptory, ...
- Moderní metody jsou založeny na porovnávání molekul:
  - Sekvence genů.
  - Sekvence proteinů.

Výběr sekvencí

- DNA - Velmi citlivá na změny, mutace probíhá různě rychle v různých oblastech genomu.
- cDNA/RNA - Užitečná pro porovnání vzdálených organismů.
- Proteinové sekvence - Porovnání evolučně vzdálených organismů, mutace probíhají rovnoměrněji.
- Příklad: Ribosomální RNA 16S - sekvence existuje ve všech prokaryotních organismech a v určité formě také v eukaryotických organismech, je jedním ze základních stavebních kamenů ribosomů => zachovává se v evoluci.

Evoluce člověka - Adam a Eva

- Fylogenetickou analýzou můžeme zjistit jak probíhala evoluce člověka.
- Zkoumané sekvence:
  - Mitochondriální DNA.
  - Y chromozom.

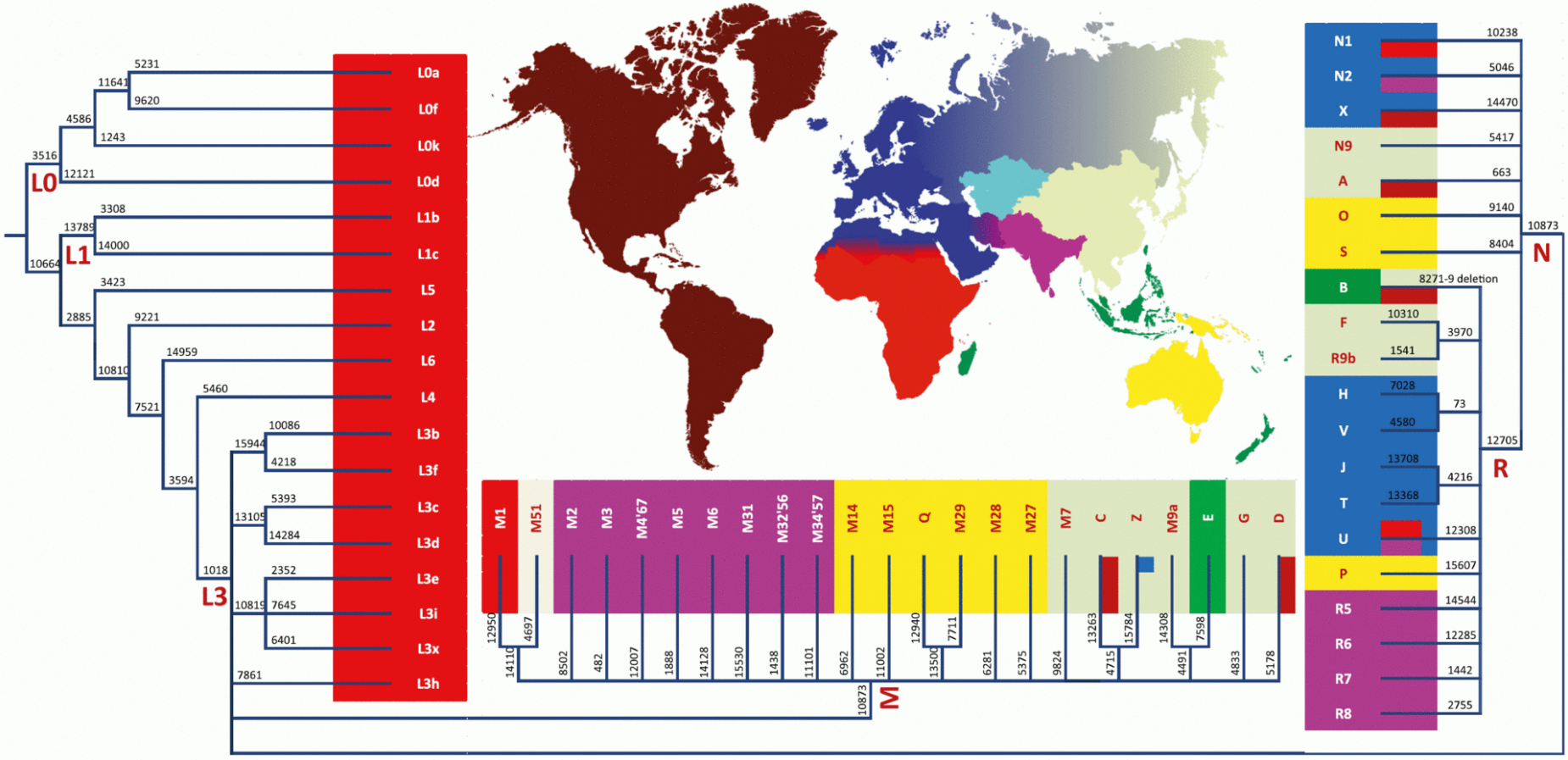
Mitochondriální DNA - Eva

- Lidská mitochondriální DNA (mtDNA)
  - Kruhová dvoušroubovicová DNA sekvence skládající se z cca 16 500 párů bazí.
  - Každý dědíme mtDNA od matky.
  - Rychlost bodových mutací je přibližně 10x větší, než li u jaderné DNA.
  - Nedochází k rekombinaci.
- Všichni jsme podědili mtDNA od jedné původní matky člověka (Evy)!.

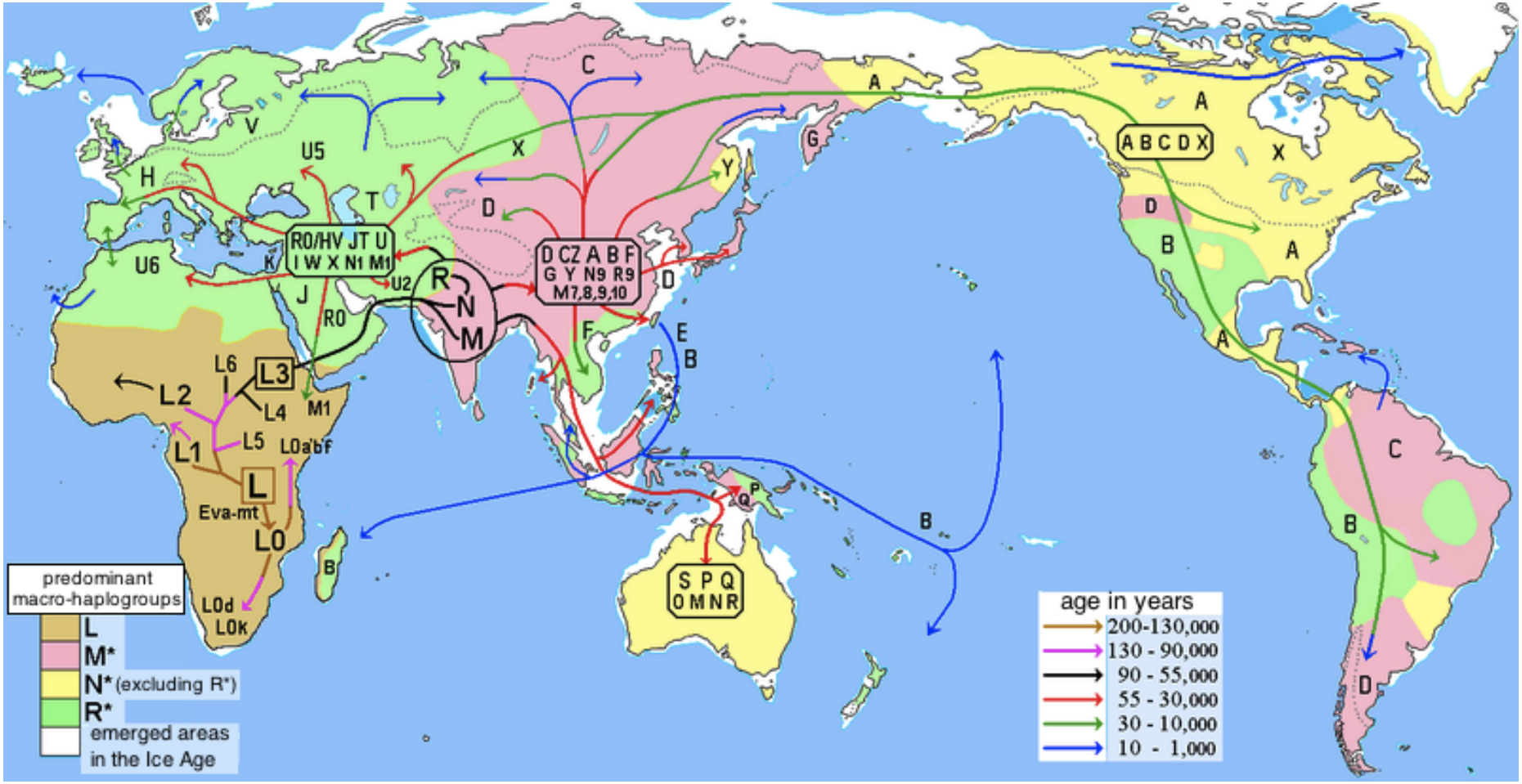
Mitochondriální DNA - Eva

- Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. Wilson, A. C. African populations and the evolution of human mitochondrial DNA. Science 253, 1503-1507 (1991).
  - Experiment: vybrali placentální tkáň u 147 žen různých ras z různých zemí.
  - Předpokládali konstantní molekulární hodiny - míra mutací je za určitý čas konstantní.
  - Jejich výzkum vedl k objevu, že Eva se objevila poprvé přibližně před 143 tisíci lety.

Mitochondriální DNA - Eva



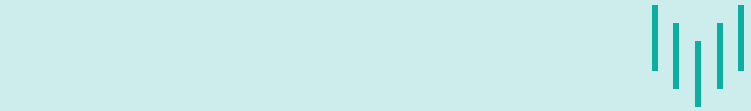
Mitochondriální DNA - Eva



Chromozome Y - Adam

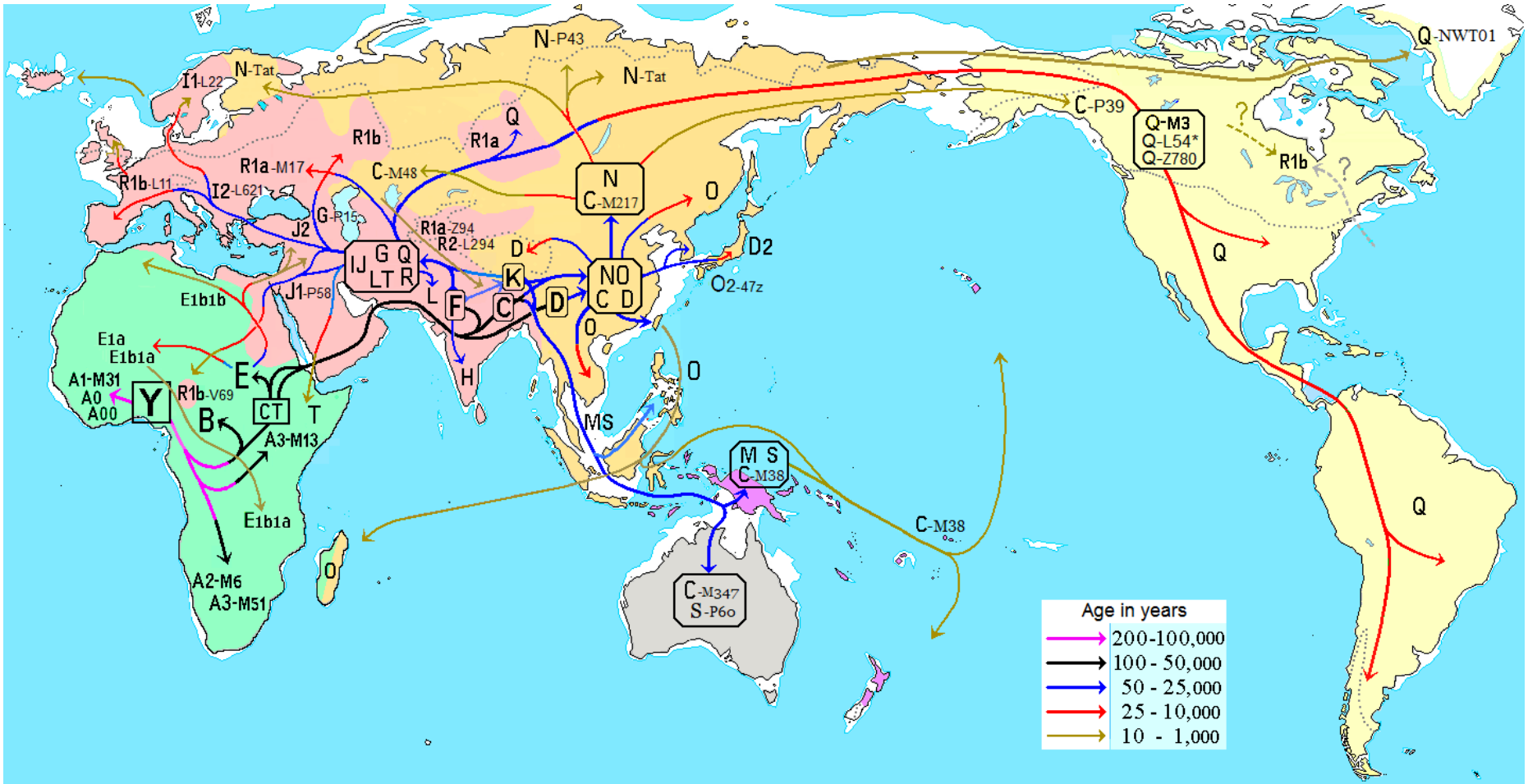
- Chromozóm Y se vyskytuje pouze u mužů => předává se od otce.
- Rychlost mutací je pomalejší, než u mtDNA => potřeba většího počtu vzorků.
- V roce 1997 studie mezi 900 muži a jejich 93 polymorfismů (bodových mutací).
  - 15% obvyvatel jihoafrických buší mají konkrétní polymorfismus A
  - 5-10% Etiopanů a Súdánců mají také A
  - Zbylá část afričanů a mimoafrických obyvatel mají T.
  - Bushmen, etiopané a sudánci jsou zřejmě geneticky nejbližšmi potomky Adama s chromozomem Y.

Chromozome Y - Adam



- Underhill et al. Y chromosome sequence variation and the history of human populations. Nature Genetic, 26:358-361, 2000.
- Studie 1062 můžů z 22 různých geografických oblastí.
- Identifikováno 167 haplotypů (polymorfismů, rozdílů).
- Stáří společného předka všech 167 haplotypů se odhaduje na 59 000 let.
- `https://en.wikipedia.org/wiki/Human_Y-chromosome_DNA_haplogroup`

Y chromozom - Adam



Výpočetní proces

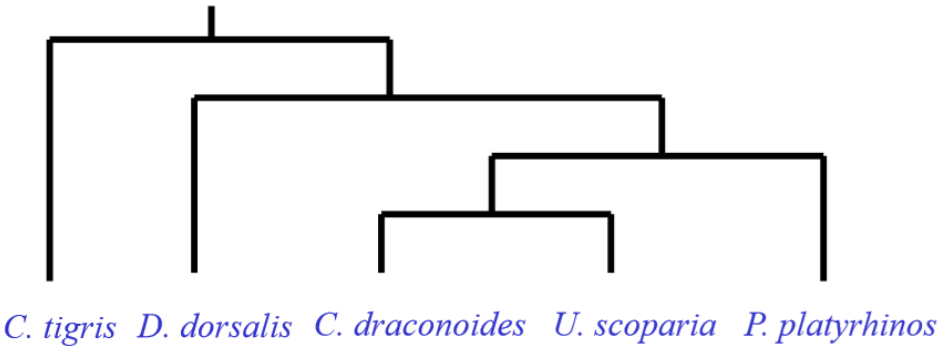
- Sekvenace DNA, RNA či proteinových sekvencí.
- Vytvoření mnohačetného alignmentů sekvencí.
- Spočtení párové vzdálenosti.
- Sestavení fylogenetického stromu.
- Odhad spolehlivosti
- Vizualizace

Fylogenetické stromy

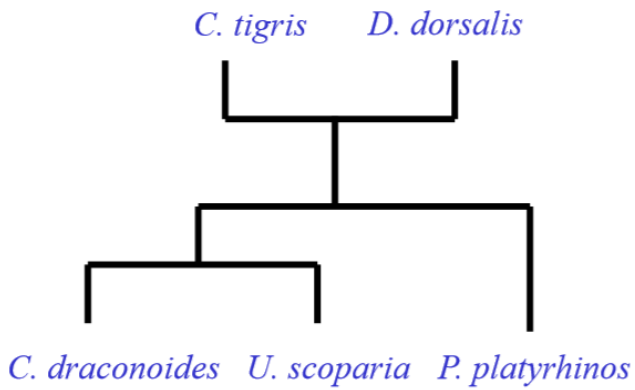
- Listy stromu reprezentují porovnávané objekty. (geny, jedince, druhy). Používá se pro ně výraz Taxon.
- Vnitřní uzly jsou hypotetičtí předci.
- V kořenovém stromu cesta od kořene k listu reprezentuje evoluční cestu, kde kořenový uzel reprezentuje společného prapředka.
- V nekořenovém stromu jsou zobrazeny vztahy mezi organismy, ale ne evoluční cesta.

Fylogenetické stromy 2

- Fylogenetické stromy jsou kořenovými stromy.
- Protože však určení společného prapředka je na vědecké úrovni velmi obtížná úloha, uvažujeme také řešení s nekořenovými stromy.



Kořenový strom



Nekořenový strom

Výběr atributů

Taxa	Selected Sequence Positions (sites) and character							
	1	2	3	4	5	6	7	8
1	A	A	G	A	G	T	G	C
2	A	G	C	C	G	T	G	C
3	A	G	A	T	A	T	C	C
4	A	G	A	G	A	T	C	C

- Pozice 1,6,8 nenesou informaci.
- Pozice 2,3,4 nepreferují žádný strom.
- Pozice 5,7 informativni.

Výběr sekvencí

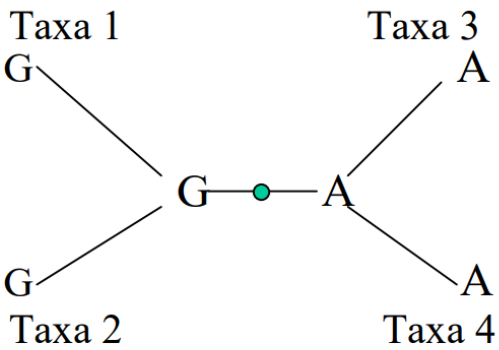
Pravidlo: Vybíráme takové pozice , kde se znaky vyskytují alespoň ve dvou taxonech a jsou zde alespoň dva různé znaky.

Výběr atributů 2

Taxa	Selected Sequence Positions (sites) and character							
	1	2	3	4	5	6	7	8
1	A	A	G	A	G	T	G	C
2	A	G	C	C	G	T	G	C
3	A	G	A	T	A	T	C	C
4	A	G	A	G	A	T	C	C

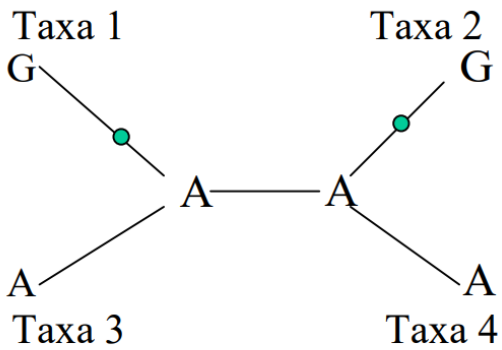
Adapted from Li and Graur 1991

Tree 1



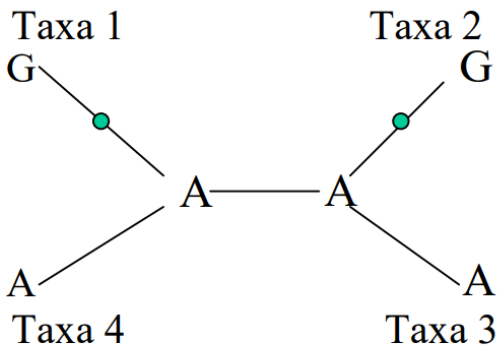
Length = 1

Tree 2



Length = 2

Tree 3



Length = 2

Volba metod

- Velmi podobné sekvence - Metody maximální parsimonie (maximum parsimony).
- Středně podobné sekvence - Metody založené na výpočtu vzdáleností.
- Velmi rozdílné sekvence - Metody maximální věrohodnosti (maximum likelihood)

Maximální parsimonie

- Nalezení stromu, který vysvětluje data s minimálním množstvím změn.
- Časově náročná metoda.
- Například programy PHYLIP a PAUP.
- Formální definice:
  - Nechť S je množina sekvencí.
  - Hammingova vzdálenost dvou sekvencí x a y je $H(x, y)$.
  - Nejúspornější (most parsimony) strom je strom T, kde listy jsou označeny sekvencemi z S a každý vnitřní uzel má přiřazenu takovou sekvenci, že platí $H(T) = \sum_{(x,y)\in E(T)} H(x, y)$ je minimální. $H(T)$ značí úspornost T (parsimony length).

Výpočetní problémy

- Small Parsimony problem - nalezení délky parsimonie (úspory) a ohodnocení hran daného stromu.
- Large Parsimony problem - nalezení nejvíce parsimonního stromu.

Small parsimony problem

Problém nalezení nejmenší parsimonie

- Vstup: Máme množinu sekvencí S a topologii kořenového stromu T s listy označenými prvky z S.
- Výstup: Délka parsimonie stromu T a označení vnitřních uzlů T.

Problém lze řešit pomocí Fitchova algoritmu.

Řešení základní úlohy - Fitchův algoritmus

Fitchův algoritmus

- Vstup: strom T, kde každý list v je označen jedním symbolem v_c.
- Výstup: nejvíce parsimonní strom T s označenými vnitřními uzly.

- Řešení ve dvou fázích:
  - Určení možné množiny pojmenování vnitřních uzlů.
  - Výběr symbolu pro vnitřní uzly.

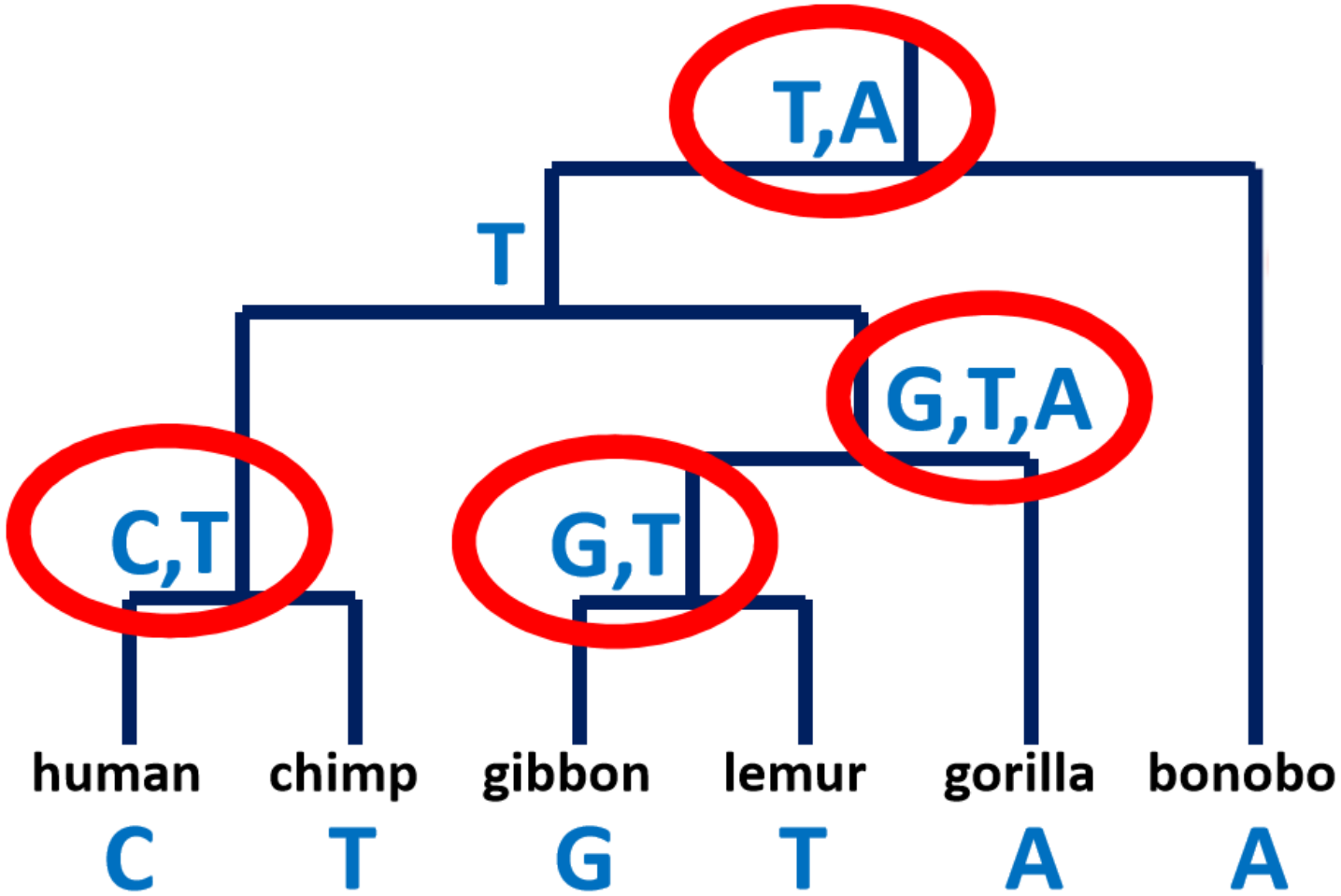
Fitchův algoritmus - potenciální označení uzlů

1. Pro každý uzel v nechť $S_v = \{v_c\}$.
2. Pro každý vnitřní uzel v s potomky u, w nechť:

$$S_v = \begin{cases} S_u \cap S_w & \text{if} S_u \cap S_w \neq \emptyset \\ S_u \cup S_w & \text{otherwise} \end{cases}$$

Délka parsimonie je rovna počtu operací sjednocení.

Fitchův algoritmus - příklad1



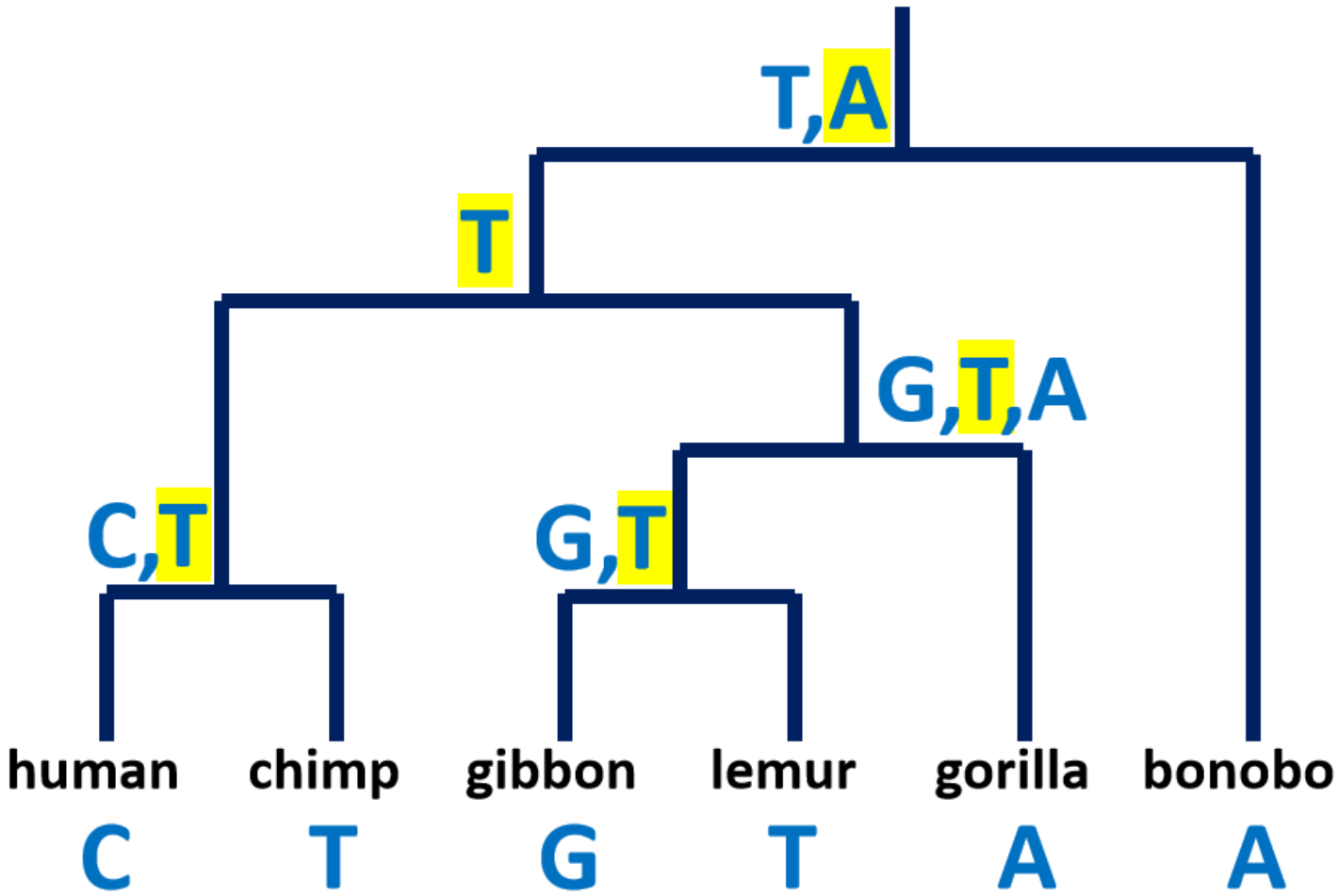
Fitchův algoritmus - volba symbolu

Varianta lineárního programování, volba symbolu pomocí backtrackingu.

1. Začneme v kořenovém uzlu, kde volíme libovolný symbol.
2. Procházíme stromem od kořene k listům.
3. Určíme v_c vnitřního uzlu v s rodičovckým uzlem u:

$$v_c = \begin{cases} \text{if} u_c \in S_v \rightarrow v_c = u_c \\ \text{otherwise} \rightarrow \text{vyberme libovolný symbol z } S_v \end{cases}$$

Fitchův algoritmus - příklad2



Fitchův algoritmus - poznámky

- Symboly na různých pozicích jsou na sobě nezávislé.
- Problém se sekvencemi libovolné délky m lze řešit pomocí m řešení základní úlohy.
- Časová složitost celého postupu je: $O(mnk)$, kde k je velikost abecedy, m je délka sekvence a n je počet sekvencí.

Large parsimony problem

Large parsinomy problem

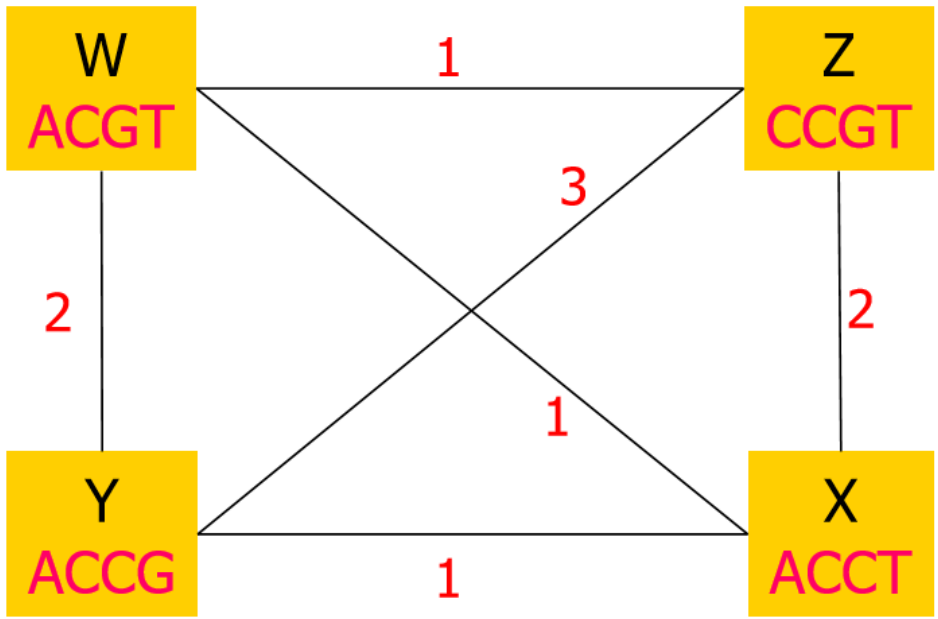
- Vstup: množina sekvencí S
- Výstup: nejvíce parsimonní strom.

- Large Parsimony je NP-těžký problém.
- Existuje však 2-aproximační polynomiální algoritmus. (Takový algoritmus, jehož řešení je garantováno nejhůře 2x horší, než je optimální řešení)

Aproximace

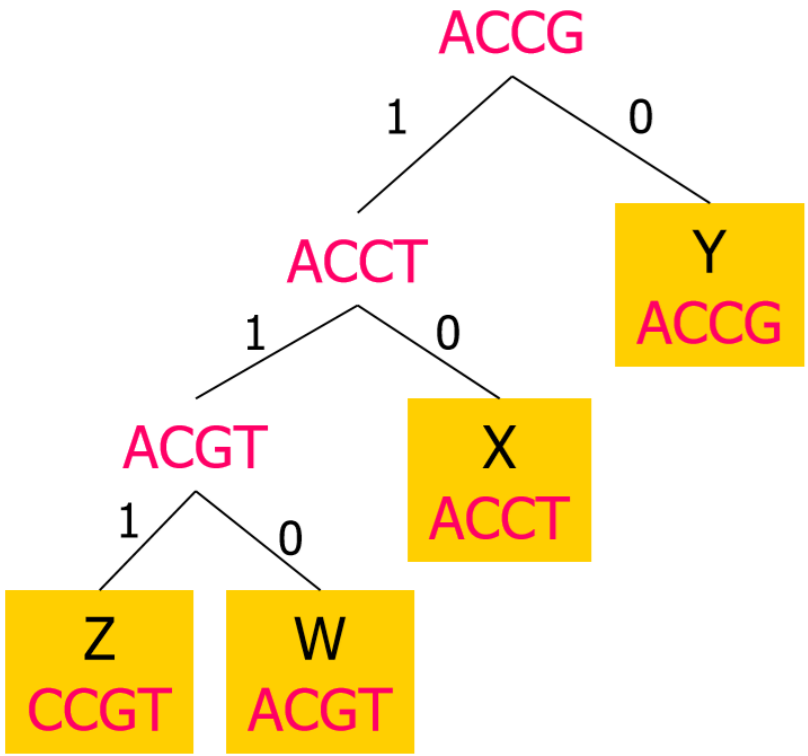
Mějme množinu sekvencí S, definujme $G(S)$ jako vážený úplný graf, jehož uzly jsou označeny sekvencemi z S a každá hrana (i, j) má váhu $H(i, j)$ (Hammingovu vzdálenost).

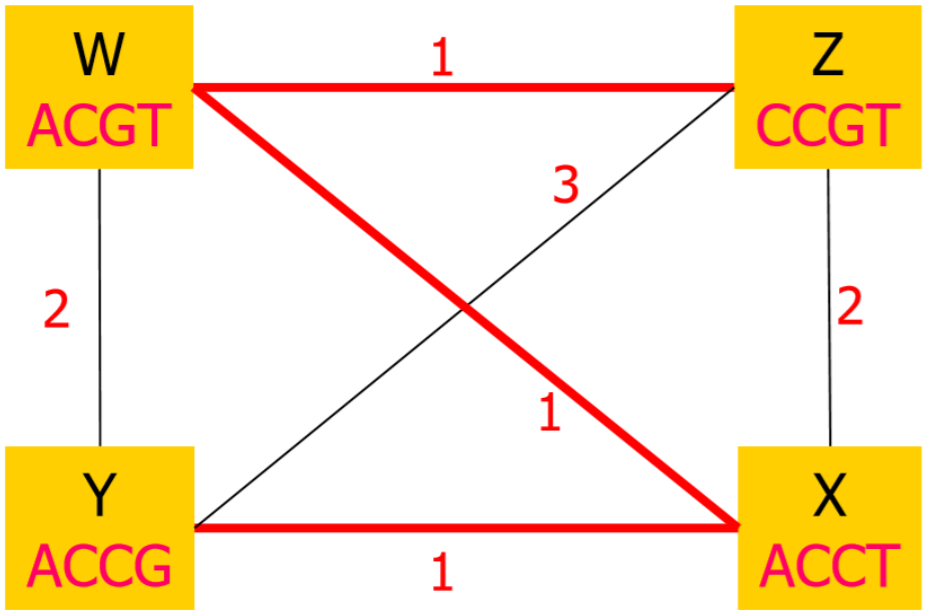
	1	2	3	4
W	A	C	G	T
X	A	C	C	T
Y	A	C	C	G
Z	C	C	G	T



Aproximační algoritmus

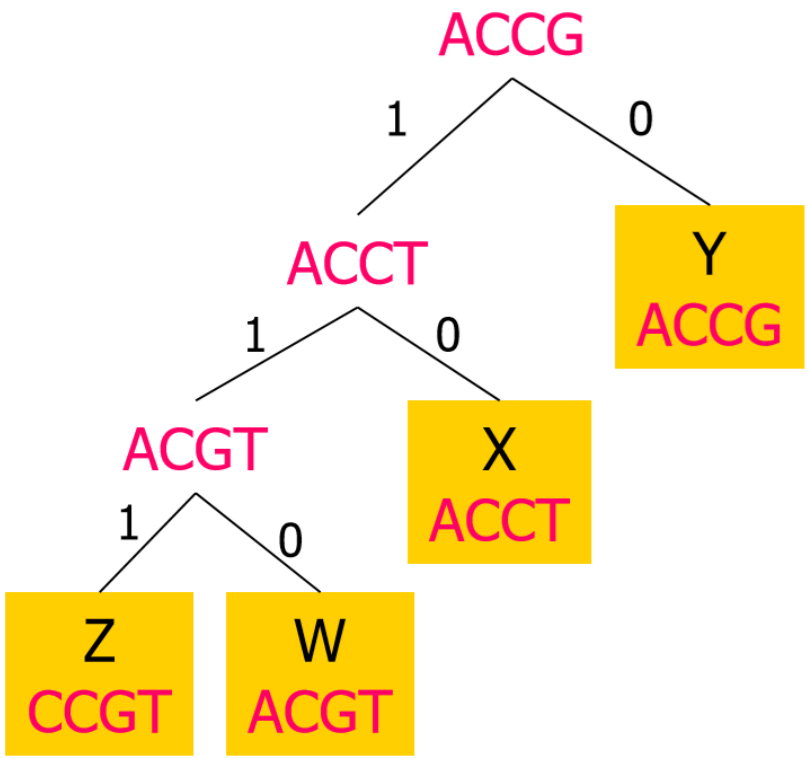
- Nechť T je minimální kostra grafu (spanning tree) $G(S)$.
- Minimální kostra grafu např. Kruskalův nebo Jarníkův algoritmus.

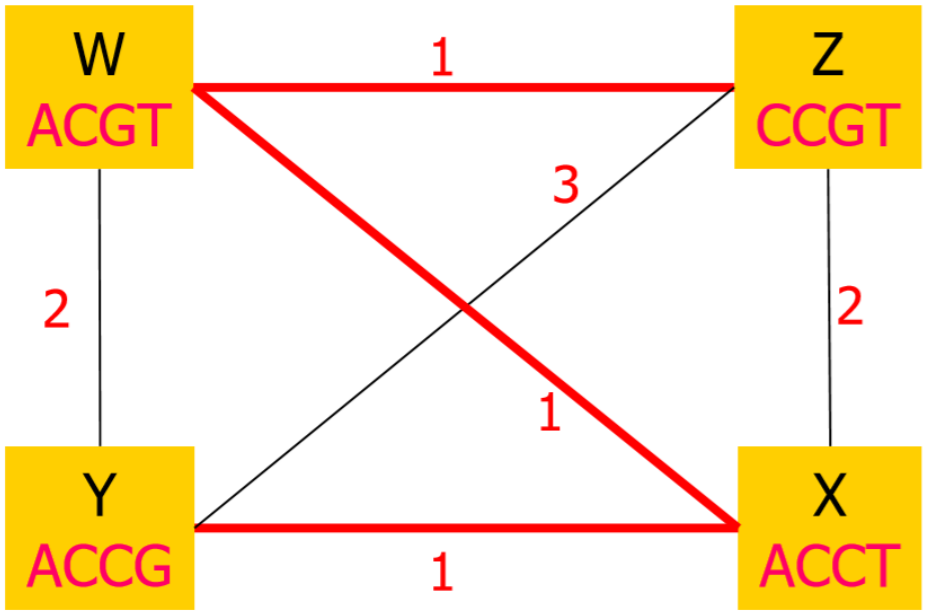




Věta: Aproximace minimální kostrou grafu

Nechť T je minimální kostra grafu $G(S)$. Potom délka parsimonie stromu T je nejvýše dvakrát tak velká, jako nejvíce parsimonní strom.





Aplikace maximální parsimonie

- Bush, R. M., C. A. Bender, et al. (1999) "Predicting the evolution of human influenza A."Science 286: 1921-1925.
- Chřipka je rapidně se vyvíjející virus.
- Bush a kolektiv ukázali, že virus lidské chřípky typu A (podtyp H3) může být použit pro predikci evolučního vývoje budoucích chřipkových kmenů.
- Tyto predikované kmeny jsou součástí chřipkové vakcíny při očkování proti chřipce.

Aplikace maximální parsimonie

- CDC - center for disease control.
- Každý rok osekvenují cca 7000 genomů chřipky.
- Určení podobnosti chřipkových genomů.
- Monitorování evoluce chřipkových virů.
- Vyhodnocení účinnosti vakcín.

Aplikace maximální parsimonie

Metody založené na vzdálenostech

- Cílem je vygenerovat strom, ve kterém podobné sekvence s krátkou vzdáleností jsou blízko a součet délek hran dvou uzlů je roven jejich vzdálenosti.
- ClustalW
- PAUP
- PHYLIP, DNADIST, PROTDIST pro vygenerování matice vzdáleností.

UPGMA

- UPGMA - unweighted pair group method with arithmetic mean.
- Předpoklad molekulárních hodin (evoluce konstaní rychlostí).
- Vytváří kořenové stromy.
- Podmínka ultrametriky: pro libovolné tři taxony a, b, c platí:

$$d_{ac} <= max(d_{ab}, d_{bc})$$

UPGMA algoritmus

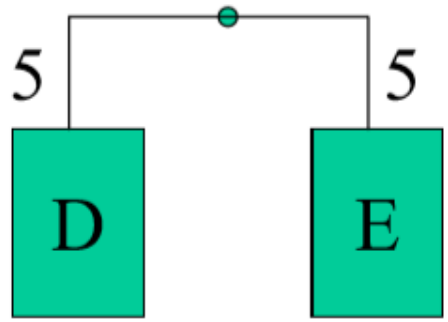
- Inicializace: Definujme množinu všech listových uzlů T, jeden pro každou sekvenci. Nechť výška každého uzlu je 0. Nechť L=T.
- Opakujte dokud je v L víc než jeden uzel.
  1. Vyberte dva nejbližší uzly (A,B) a vytvořte z nich rodičovský uzel K. Spojte A,B a K. Nastavte výsku K na $d_{ab}/2$. Nastavte délku větve mezi K a A = výška K - výška A, nastavte délku větve mezi K a B = výška K - výška B.
  2. Odeberte A,B z L a přidejte K do L. Přepočítejte vzdálenosti mezi K a dalšími uzly v L. Vzdálenost mezi K a dalšími uzly je dána průměrnou vzdáleností potomků K k ostatním uzlům.

UPGMA příklad

	A	B	C	D	E
A	-	20	26	26	26
B		-	26	26	26
C			-	16	16
D				-	10
E					-

	A	B	C	DE
A	-	20	26	26
B		-	26	26
C			-	16
DE				-

Step 1: Select D and E

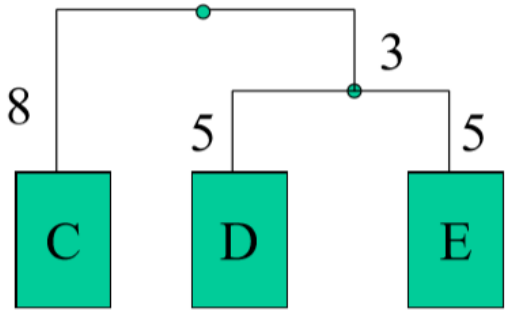


UPGMA příklad

	A	B	C	DE
A	-	20	26	26
B		-	26	26
C			-	16
DE				-

Step 2: Select (DE) and C

	A	B	DEC
A	-	20	26
B		-	26
DEC			-



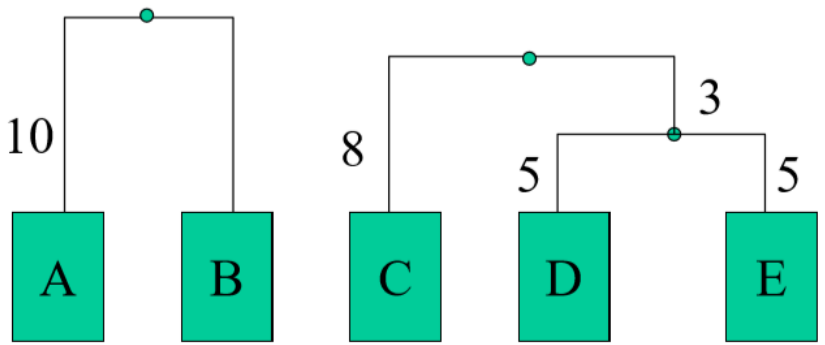
$$dist(DEC,A) = (d_{DA}+d_{EA}+d_{CA})/3 = 26$$
$$dist(DEC,B) = (d_{DB}+d_{EB}+d_{CB})/3 = 26$$

UPGMA příklad

	A	B	DEC
A	-	20	26
B		-	26
DEC			-

	AB	DEC
AB	-	26
DEC		-

Step 3: select A, B

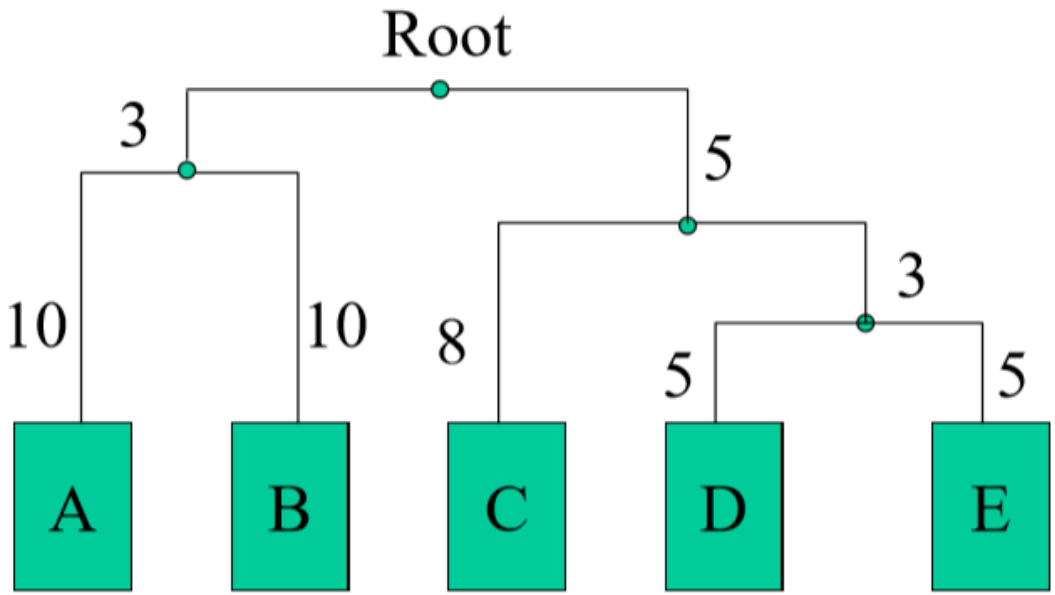


$$\text{dist(DEC,AB)} = (d_{DA}+d_{DB}+d_{EA}+d_{EB}+d_{CA}+d_{CB})/6$$
$$= 26$$

UPGMA příklad

	AB	DEC
AB	-	26
DEC		-

Step 4: select (A,B), (D,E,C)

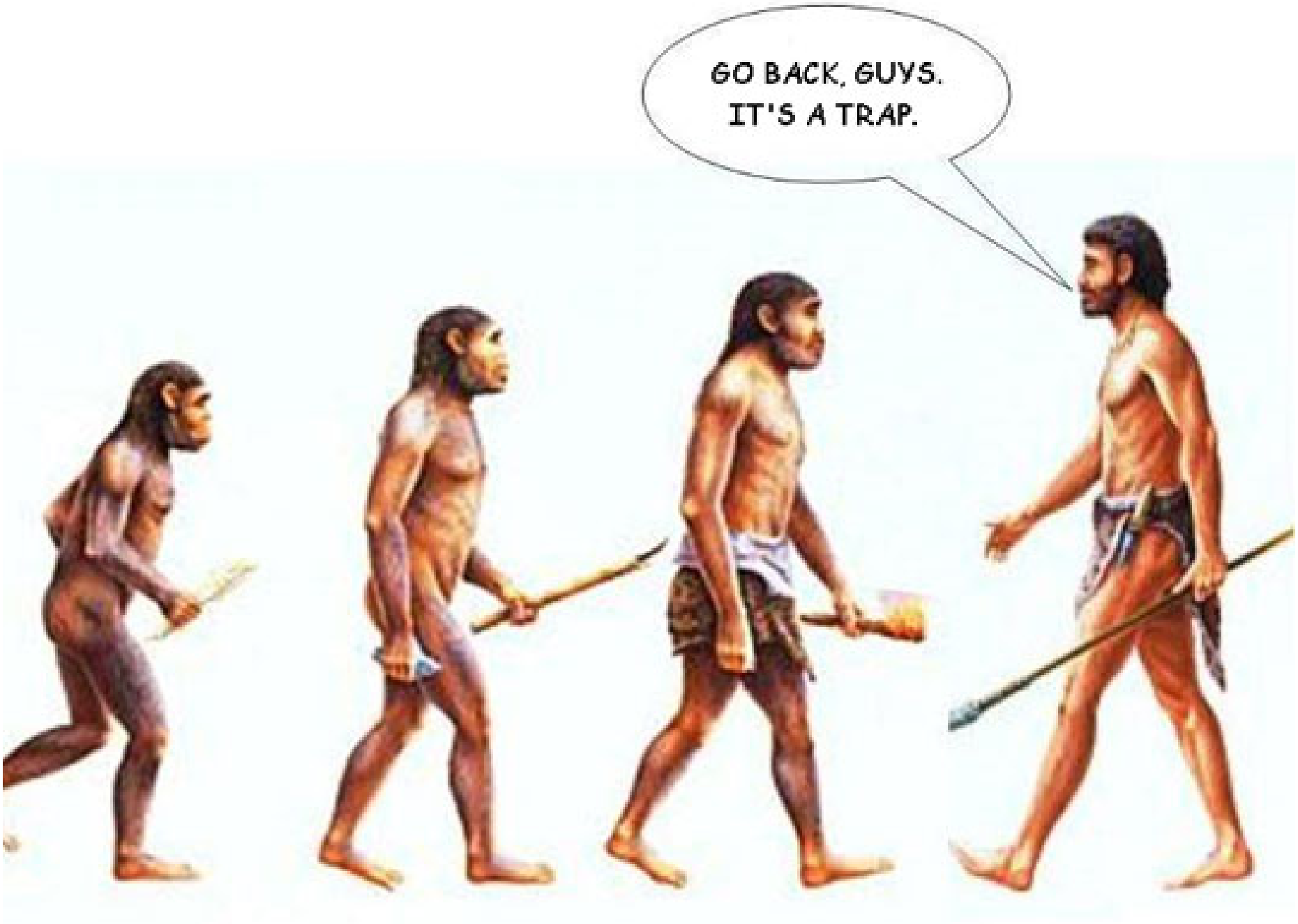


Výpočet vzdáleností

- Celkový počet bodových neshod.
- Celkový počet bodových neshod a indelů.

Ověření spolehlivosti

- Bootstrapping
- Mějme sekvence o délce m, vybíráme podmnožinu symbolů sekvencí.
- Pro každou náhodně vybranou podmnožinu vytvoříme fylogenetický strom.
- Opakujeme požadovaný počet krát.
- Ověříme vztahy mezi dvojicemi sekvencí. Pokud je jejich vazba stabilní => vyskytuje se ve většině stromů, pak jsme nejspíše obdrželi fylogenetický strom blízký skutečnému stromu.


GO BACK, GUYS.
IT'S A TRAP.

DÄŻkuji za pozornost

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA404

michal.vasinek@vsb.cz

23. listopadu 2022

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY