# Fylogenetická evoluční analýza

## Fylogeneze

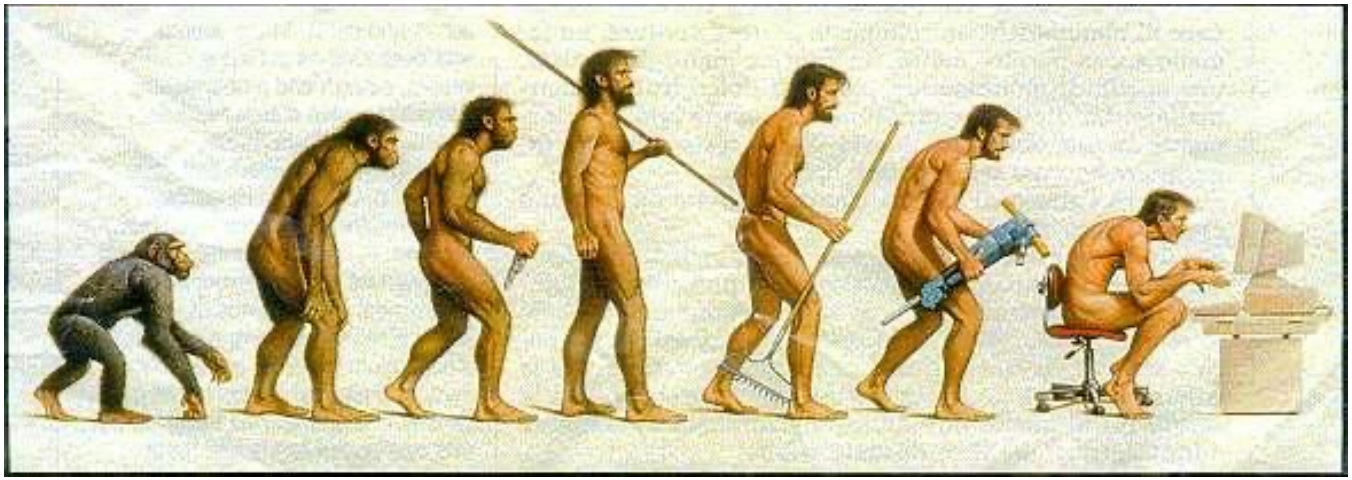
- Vývoj nových druhů procesem evoluce

## Fylogenetika

- Věda zkoumající fylogenezi, příbuzenské vztahy a vývoj organismů

## Fylogenetická analýza

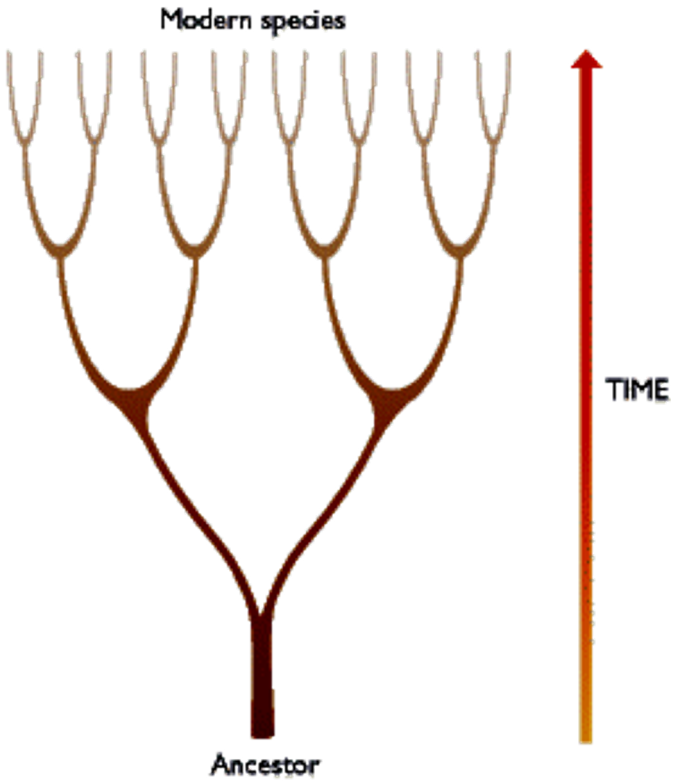
- Tvorba **fylogenetického stromu** popisujícího evoluční vztahy mezi organismy



Evoluce bioinformatika

# Fylogeneze

**Fylogeneze nezahrnuje pouze podobnosti a rozdíly mezi organismy (taxonomie)…**




**…ale také jejich evoluční vztahy.**

# Fylogenetická data

Fylogenetická data jsou získávána zkoumáním charakteristických znaků studovaných organismů

- **Morfologické znaky** (tvar)
  - Mnoho přechodových forem
  - Fosilní pozůstatky můžou být nekvalitní, neposkytují požadovná data nebo se vůbec nedochovaly potřebné informace
- **Sekvence nukleotidů nebo aminokyselin**
  - Stejná sada znaků u všech organismů → lze porovnat i velmi vzdálené životní formy

# Molekulární fylogenetická data

- **DNA** sekvence (nejčastěji)
  - Obsahuje více informací (tiché mutace, kódující vs. nekódující oblasti)
  - Pro velmi blízce příbuzné sekvence u savců je vhodná mitochondriální DNA

- Sekvence **proteinu**
  - Pro vzdáleně příbuzné organismy
  - Sekvence proteinů se zachovává během evoluce déle než sekvence DNA

# Molekulární fylogenetická data

- Jediný experiment může poskytnout informace o mnoha znacích.

```
AAGACGGCACCGACAACGACTACAACGACGCCGTCGTGGTGATCAACTGGCCGCTCGGCT
AGGATGGTACCGACATGGACTACAACGACTCCATCGTCATCCTGAACTGGCCGCTGGGCT
GGGACGGCAACGGC-TGGAC--CAAGGGCGCCTACACCGCCACGAACTGA----------
ACGACGTGCCCGGAACCTATGGCAATAACTCCGGC-TCGTTCAGTGTCAATATTGGAAAG
```

**Každá nukleotidová pozice v sekvenci může být považována za jeden ZNAK, který se vyskytuje ve ČTYŘECH rozdílných STAVECH.**

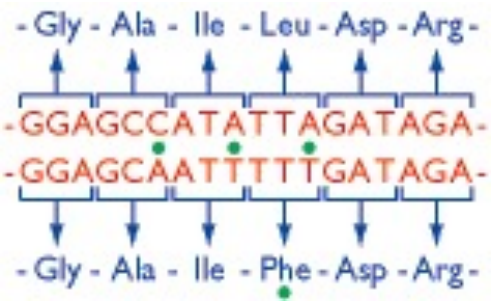
- Jednotlivé stavy jsou jednoznačné a nezaměnitelné (A x C x G x T).
  Na rozdíl od morfologických znaků (tvar), u nichž existuje mnoho přechodových forem.

- Molekulární data se dají snadno převést do „číselné“ formy.
  Vhodné pro matematické a statistické analýzy.

## Proteinové sekvence x DNA sekvence

- **Pro fylogenetickou analýzu využívány PŘEVÁŽNĚ DNA sekvence.**

**DNA poskytuje mnohem více fylogenetických informací než protein.**




**Tiché mutace**

**Variabilita uspořádání genomu (kódující x nekódují oblasti)**

**PCR, automatické sekvencování**

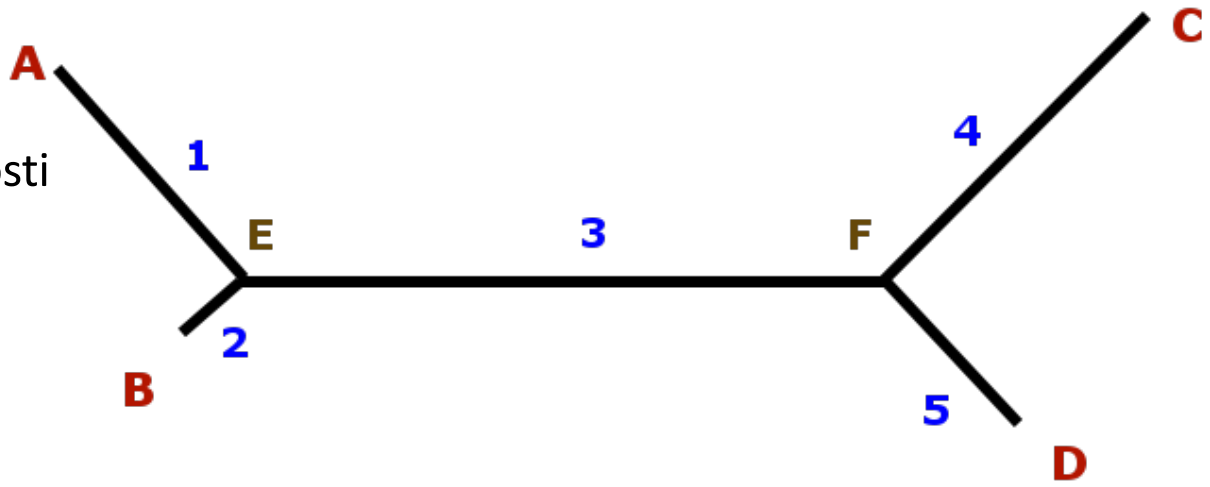
**velmi informativní pro selekci synonymních a nesynonymních mutací pro identifikaci pozitivní (adaptabilní) či negativní selekce.**

## Proteinové sekvence x DNA sekvence

- **ALE, mnoho případů, kdy je lépe pracovat s proteinovými sekvencemi:**
  - pro studium velmi odlišných (vzdálených) skupin organismů (buď pomalu se vyvíjející nt sekvence (rRNA) nebo PROTEINOVÉ sekvence (např. mezi bakteriemi a eukaryonty).
  - DNA sekvence mohou být „biased“ – preferential codon usage
  - odlišný kód mitochondriální DNA (nutnost přeložení do AA sekvencí)
  - vyšší poměr “signal-to-noise“ ve fylogenetické analýze (menší pravděpodobnost nesprávného/náhodného přiložení)

# Fylogenetický strom (fylogram)

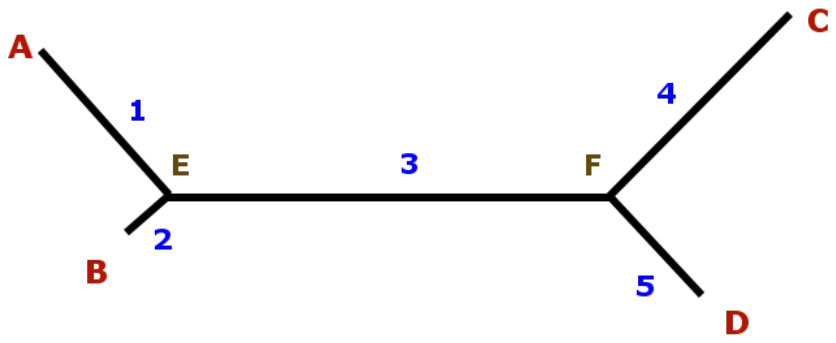
- Terminální (externí) uzly (**A, B, C, D**) – současné taxony (geny)
- Interní uzly (**E, F**) – společný předek
- Větve (**1, 2, 3, 4, 5**)
  - Vnitřní větve (**3**)
  - Periferní větve (**1, 2, 4, 5**)
  - Délky větví jsou úměrné velikosti změny v průběhu evoluce



# Typy fylogenetických stromů

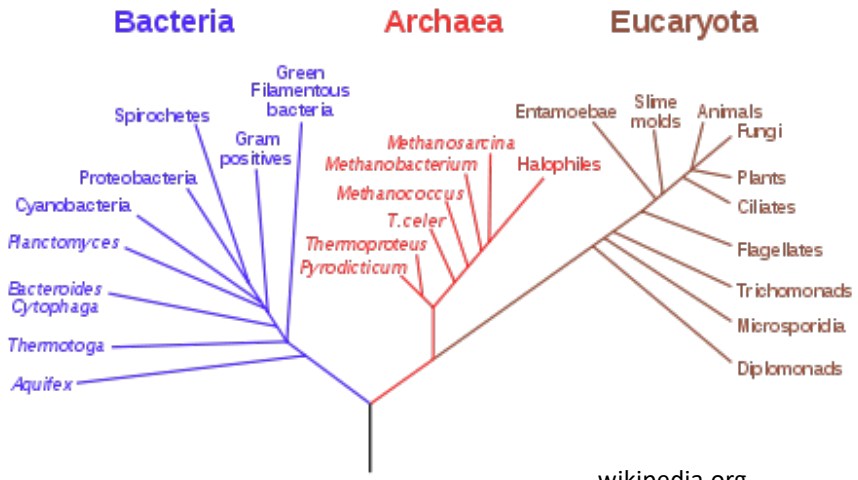
## BEZ KOŘENE

- Není známý společný předek
- Neobsahuje informace o průběhu evoluce



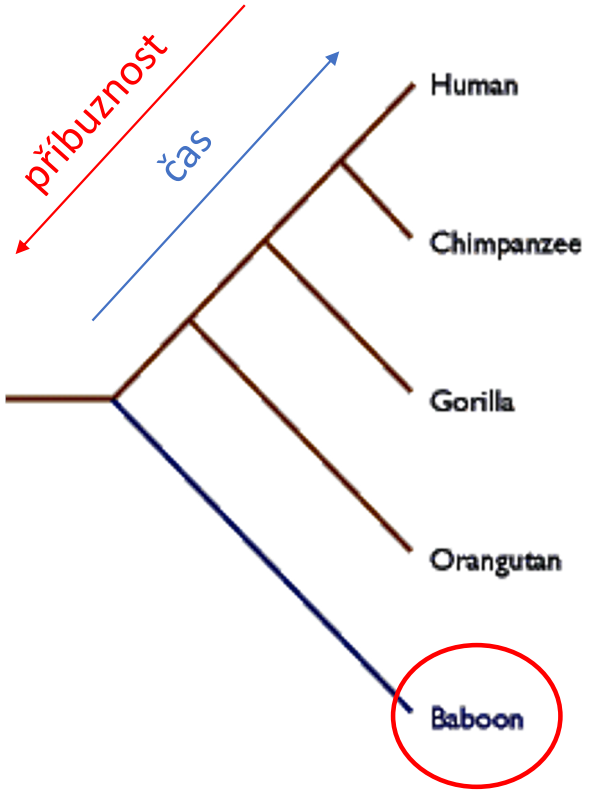

## S KOŘENEM

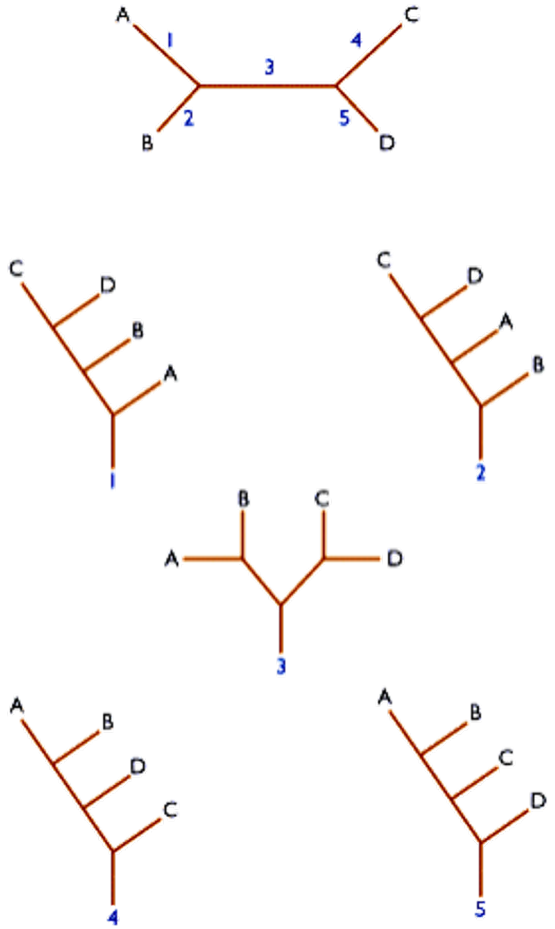
- Kořen – společný předchůdce všech taxonů
- Obsahuje informace o **průběhu evoluce**




wikipedia.org

# Fylogenetický strom (fylogram)




**Fylogenetický strom S KOŘENEM (rooted).**
**Nutný alespoň jeden gen, který je méně příbuzný s A,B,C,D, než jsou tyto geny mezi sebou navzájem = „outgroup“**

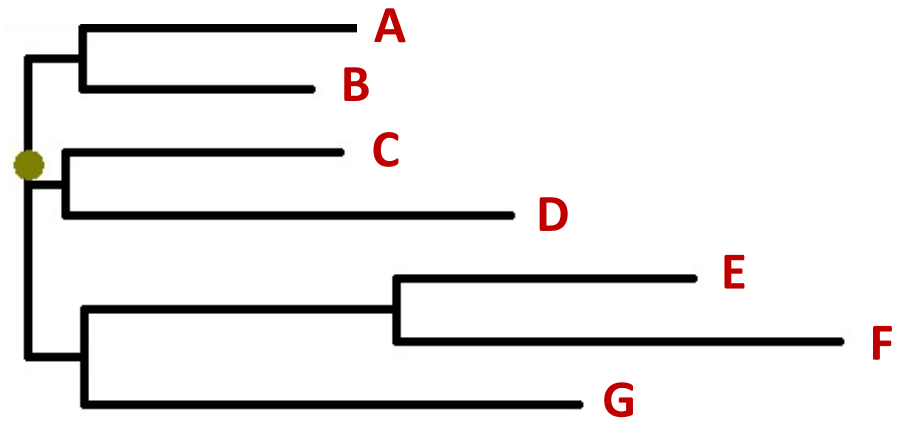
# Vlastnosti fylogenetického stromu

- Topologie – způsob, jak se strom větví, určení vztahu mezi studovanými sekvencemi (topologie je známá pro každý vytvořený strom, různé typy zápisu, např. ((A,(C,D)),B)); nebo ((A:0.5,(C:0.2,D:0.4):0.7):0.1,B:1):0.8;
- Délka větve – lineární kombinace mutacní rychlosti a casu
- Délka větve fylogenetického stromu má biologický význam (Delší větev – buď se vyvíjela nezávisle delší dobu (vnitrní vetve) nebo na ni v čase mezi divergencemi (uzly) působila vyšší mutační rychlost (např. vlivem selekce či následkem změn v populační dynamice)
- Délka stromu (tree length) – součet délek všech větví
- Délka od kořene ke špičce (root-to-tip length) – nejdelší vzdálenost od kořene k terminálnímu uzlu

# Typy fylogenetických stromů

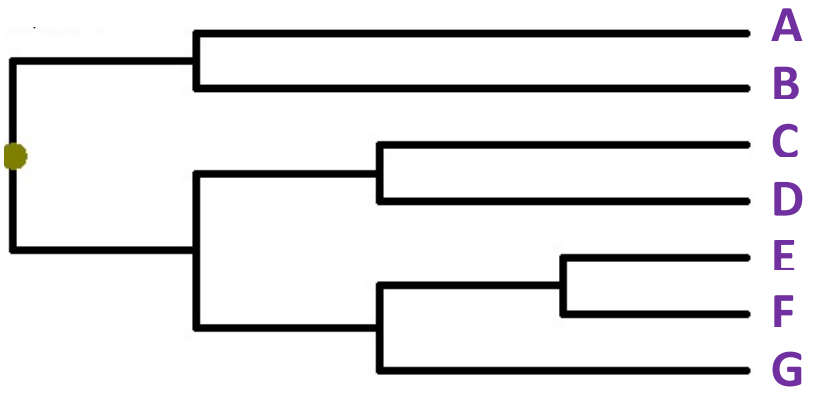
## FYLOGRAM

- rozvětvený diagram (strom), který naznačuje fylogenezi (postupný vývoj)
- délka jednotlivých větví je úměrná **velikosti změny** v průběhu evoluce

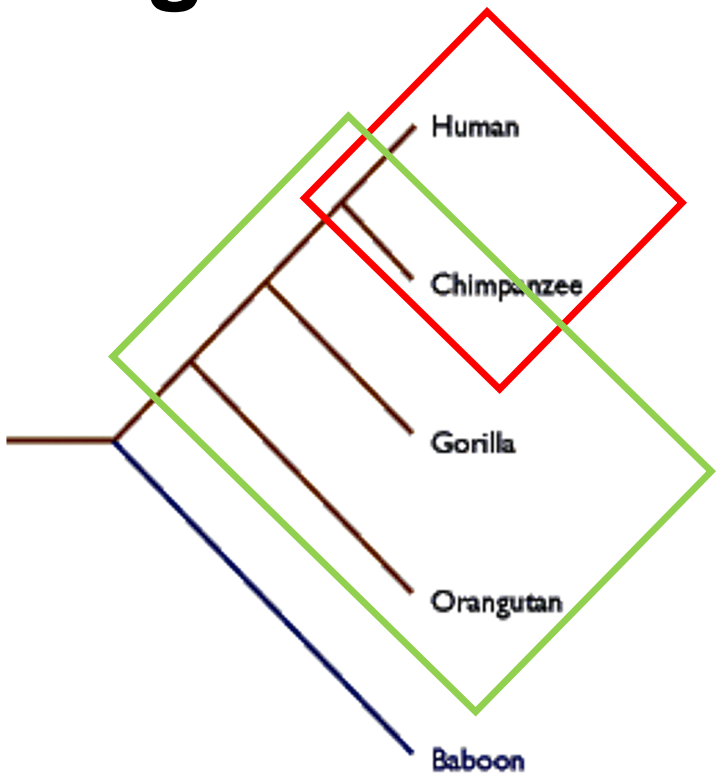



## KLADOGRAM

- všechny větve mají **stejnou délku**
- ukazuje **společné předky**, ale ne množství změn, které od té doby taxony prodělaly



# Topologie



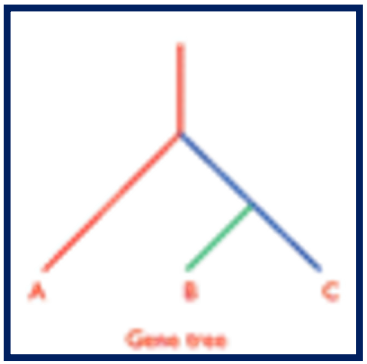

monofyletická skupina
– zahrnuje VŠECHNY větve (potomky)
a jejich nejmladšího předka

parafyletická skupina skupina
– nezahrnuje VŠECHNY větve
od společného předka

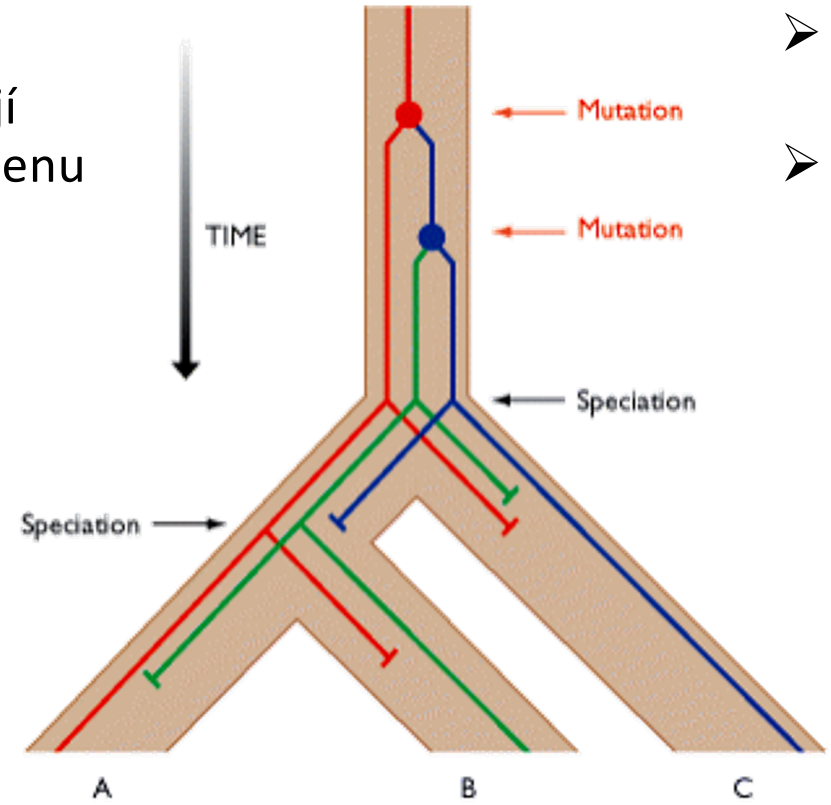
# Typy fylogenetických stromů

## GENOVÝ

- Srovnání genů
- Vnitřní uzly představují rozdělení původního genu (**mutace**)

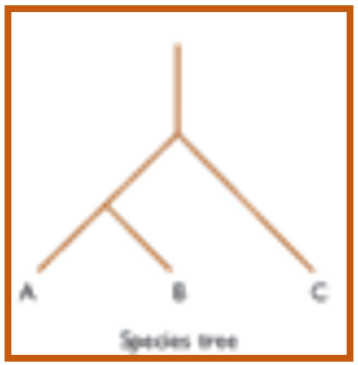



Mutace předchází separaci.




## DRUHOVÝ

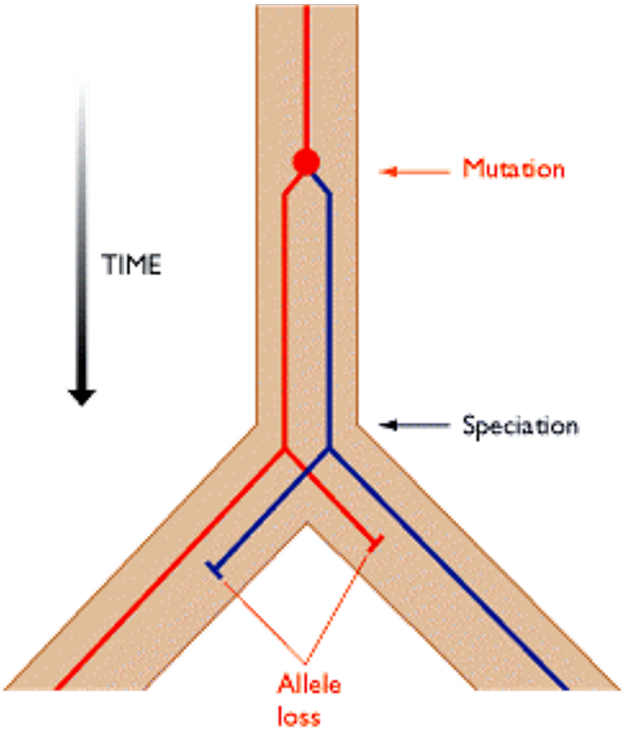
- Srovnání morfologických dat
- Vnitřní uzly představují rozdělení původního druhu (**separace**)



# „Genový“ strom x „druhový strom“

- Mutace a vznik nového druhu se s největší pravděpodobností neodehrají současně.

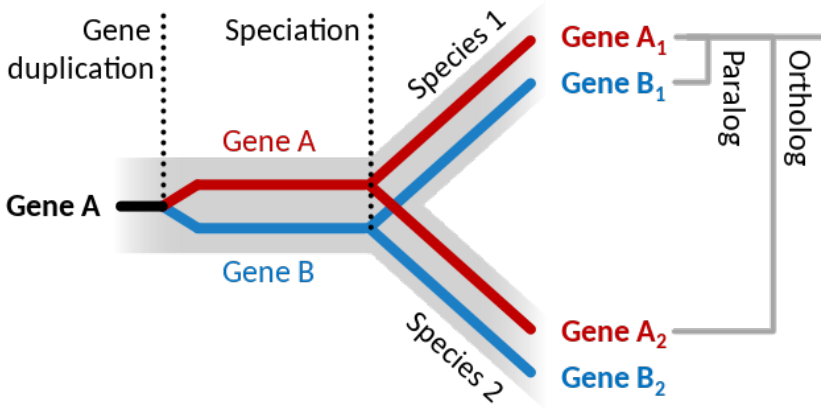
- Mutace předchází separaci – v populaci se nacházejí obě alely genu.
- Po rozdělení populací může dojít ke ztrátě jedné alely.



# Konstrukce fylogenetických stromů

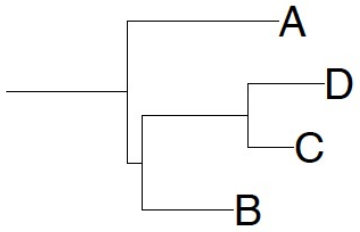
1. Alignment **ortologních** sekvencí
   - ➢ Homologní geny
     - Geny odvozené od společného předka
     - **Ortologní** x **Paralogní geny**

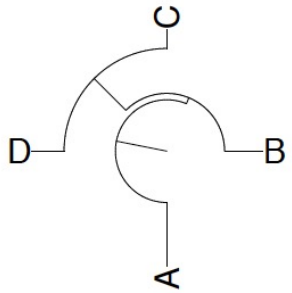
       (Nový gen vznikl speciací x duplikací)

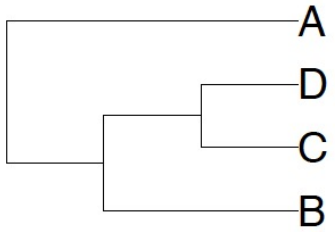



2. Určení vzdáleností mezi sekvencemi
   - ➢ Metody: Distanční metody, UPGMA (Unweighted Pair Group Method), Metoda nejmenších čtverců, Metoda minimální evoluce, Neighbor-Joining, Maximální parsimonie, Maximální věrohodnost

# Zobrazení fylogenetických stromů




- Fylogram – zobrazuje topologii a délky větví, čte se od kořene ke špičkám
- Fylogram ve vejírovém zobrazení (využívá se pro velké stromy), čte se od středu k okraji
- Kladogram – zobrazuje jenom topologii, zobrazené délky větví jsou jen uživatelská volba bez biologického významu

# Tvorba evolučních stromů

- „Alignment“ sekvencí – nezbytný pro vytvoření stromu.
  Vyhodnocení rozdílů mezi jednotlivými nukletidovými sekvencemi, většinou „multiple alignment“.

```
BclA        CGATCAACGGCAAGAAGTCGGACGGCTCGCCGTTCACGGTCAACTTCGGGATCGTCGTGT 325
BclB        CGA-CATCTTCAAGAAGAC--------------------CTACTTCGGGCTGGTCGGAT 670
BclD        CGCTGAGCGCGGGCGATACCG------------------TGTGGCTGGGCTGGCTGGGC 804
BclC        GGA-TATTTTTAAAAAATC--------------------TTATTTCGGTATTATTGGCT 754
             *   *         *  *                          **  *    *

BclA        -CGGAAGACGGCCACGACAGCGACTACAACGACGGCATCGTCGTGCTCCAGTGGCCGATC 384
BclB        -CGGAAGATGGCGGCGATGGCGACTACAACGACGGCATCGCGATCCTGAACTGGCCGCTG 729
BclD        GCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTCTGCAGTGGCCGATT 864
BclC        -CTGAAGATGGTGCGGATGATGATTATAACGATGGCATCGTGTTTCTGAACTGGCCGCTG 813
             * ***** **    **    ** ** ** ** ***** *   * **  * ****** *
```

# Guide tree vs. Phylogenetic tree

**Guide tree**

- Vypočítán na základě matice vzdáleností (distance matrix) vytvořené podle skóre pairwise alignmentů
- Výstupem je .dnd soubor
- Slouží pro vytvoření alignmentu, NEMÁ fylogenetický význam

**Phylogenetic tree**

- Vypočítán na základě vytvořeného MSA
- Vzdálenosti mezi sekvencemi jsou vypočteny a uloženy jako .ph soubor
- Následně je možno je využít pro konstrukci fylogenetického stromu (soubory .nj, .ph, .dst) pomocí zvolené metody (nj, phylip, dist)

DIST = percentage divergence (/100)
Length = number of sites used in comparison

1 vs. 2 DIST = 0.6491; length = 114
1 vs. 3 DIST = 0.6842; length = 114
1 vs. 4 DIST = 0.9298; length = 114
1 vs. 5 DIST = 0.9035; length = 114
1 vs. 6 DIST = 0.9386; length = 114
1 vs. 7 DIST = 0.9825; length = 114
2 vs. 3 DIST = 0.3772; length = 114
2 vs. 4 DIST = 0.9123; length = 114
2 vs. 5 DIST = 0.8947; length = 114
2 vs. 6 DIST = 0.9123; length = 114
2 vs. 7 DIST = 0.9386; length = 114
3 vs. 4 DIST = 0.9123; length = 114
3 vs. 5 DIST = 0.9386; length = 114
3 vs. 6 DIST = 0.9298; length = 114
3 vs. 7 DIST = 0.9474; length = 114
4 vs. 5 DIST = 0.9211; length = 114
4 vs. 6 DIST = 0.9035; length = 114
4 vs. 7 DIST = 0.9649; length = 114
5 vs. 6 DIST = 0.9561; length = 114
5 vs. 7 DIST = 0.9211; length = 114
6 vs. 7 DIST = 0.9649; length = 114

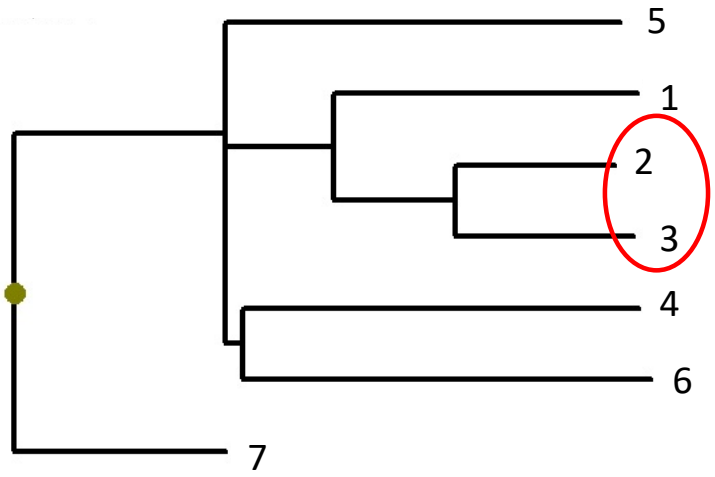
Neighbor-joining Method

Saitou, N. and Nei, M. (1987) The Neighbor-joining Method:
A New Method for Reconstructing Phylogenetic Trees.
Mol. Biol. Evol., 4(4), 406-425

This is an UNROOTED tree
Numbers in parentheses are branch lengths

Cycle 1 = SEQ: 2 ( 0.17807) joins SEQ: 3 ( 0.19912)
Cycle 2 = SEQ: 1 ( 0.34101) joins Node: 2 ( 0.13706)
Cycle 3 = SEQ: 5 ( 0.44298) joins SEQ: 7 ( 0.47807)
Cycle 4 = SEQ: 4 ( 0.44518) joins SEQ: 6 ( 0.45833)
Cycle 5 (Last cycle, trichotomy):
Node: 1 ( 0.12171) joins
Node: 4 ( 0.01864) joins
Node: 5 ( 0.02083)

.nj soubor




.dst soubor

7

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| PAIIL | 0.000 | 0.649 | 0.684 | 0.930 | 0.904 | 0.939 | 0.982 |
| RSIIL | 0.649 | 0.000 | 0.377 | 0.912 | 0.895 | 0.912 | 0.939 |
| CVIIL | 0.684 | 0.377 | 0.000 | 0.912 | 0.939 | 0.930 | 0.947 |
| BCLA | 0.930 | 0.912 | 0.912 | 0.000 | 0.921 | 0.904 | 0.965 |
| BCLB | 0.904 | 0.895 | 0.939 | 0.921 | 0.000 | 0.956 | 0.921 |
| BCLC | 0.939 | 0.912 | 0.930 | 0.904 | 0.956 | 0.000 | 0.965 |
| BCLD | 0.982 | 0.939 | 0.947 | 0.965 | 0.921 | 0.965 | 0.000 |

# Typy mutací v sekvenci

- Inzerce a delece (u genů kodujících proteiny musí mít násobky 3 (kodony)
- Mutace jsou pod selekčním tlakem (tiché mutace budou častější)
  - Obvykle snižující se pořadí: 3 - 1 – 2 (mutace ve třetím nukleotidu často nevede k záměně AA, mutace na druhé pozici se objevují nejméně často (jsou stejně pravděpodobné, ale často ”nepřežijí”, takze je nevidíme)
- Substituce
  - Tranzice – A-G, G-A, C-T, T-C
  - Transverze - A-C, A-T, C-G, T-G
- Pořadí genů (analýza genomů) - ve fylogenetice se řeší jako morfologický znak (při správné analýze (alignmentu) se genom rozdělí na jednotlivé geny a následně spojí alignmentem genů (concatenated sequence)

## Jak převést „multiple alignment“ na strom?

Fylogenetický strom – hypotéza, která vznikla co nejlepším odhadem na základě omezeného zdroje informací

- Dva přístupy
- Algoritmus – jde přímo k výsledku, co je jediný strom (odpadá srovnání vzájemně si konkurujících stromů) – metody shlukové (klastrové) analýzy (UPGMA, Neighbour-joining (NJ)) – obě využívají data vzdálenosti (distance)

- Kritérium optimálnosti – dva kroky – definování kritéria, podle kterého je hodnocen každý strom určitým skóre, které se použije k následnému srovnání všech stromů - použití specifického algoritmu pro výpočet funkce (kritérium optimálnosti) a pro získání stromu s nejlepší hodnotou této funkce

## Jak převést „multiple alignment“ na strom?

- Metody založené na distančních maticích (klastrovací metody)
- Unweighted Pair Group Method with Arithmetic Mean (UPGMA)/Weighted (WPGMA)
- Neighbor-joining (NJ)
- Metody založené na modelu/diskrétní metody
- Maximální parsimonie – MP (může kódovat indel jako pátou bázi)
- Maximální věrohodnost (maximum likelihood, ML) - skvělé, ale pomalé stromy
- Bayesiánská inference – BI (robustní vůči nastavení modelu), dnes zlatý standard

## Jak převést „multiple alignment“ na strom?

- **Distanční matice.**

  **Slouží k určení délky větví.**

Multiple alignment

```
1  AGGCCAAGCCATAGCTGTCC
2  AGGCAAAGACATACCTGACC
3  AGGCCAAGACATAGCTGTCC
4  AGGCAAAGACATACCTGTCC
```

4/20

Distance matrix

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | – | 0.20 | 0.05 | 0.15 |
| 2 | | – | 0.15 | 0.05 |
| 3 | | | – | 0.10 |
| 4 | | | | – |

## Jak převést „multiple alignment“ na strom?

- **Unweighted Pair Group Method with Arithmetic Mean (UPGMA)**
- **(and Weighted - WPGMA)**
- **– Využívá distanční matici.**

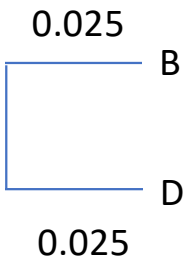
**Ultrametrická metoda, očekává, že všechny terminální konce jsou stejně vzdálené od počátku (molekulární hodiny) - všechny linie se vyvíjejí stejnou rychlostí...**

**Výsledkem je “rooted” tree**

# Jak převést „multiple alignment“ na strom?

## UPGMA

| | A | B | C | D |
|---|---|---|---|---|
| A | - | 0.20 | 0.25 | 0.15 |
| B | | - | 0.15 | 0.05 |
| C | | | - | 0.10 |
| D | | | | - |




branch – BD/2 = 0.05/2 = 0.025

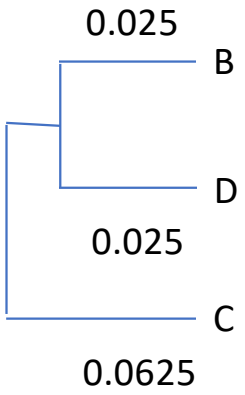
# Jak převést „multiple alignment“ na strom?

## UPGMA

| | A | B | C | D |
|---|---|---|---|---|
| A | - | 0.20 | 0.25 | 0.15 |
| B | | - | 0.15 | 0.05 |
| C | | | - | 0.10 |
| D | | | | - |



| | BD | A | C |
|---|---|---|---|
| BD | - | (0.2+0.15)/2= 0.175 | (0.15+0.1)/2= 0.125 |
| A | | - | 0.25 |
| C | | | - |

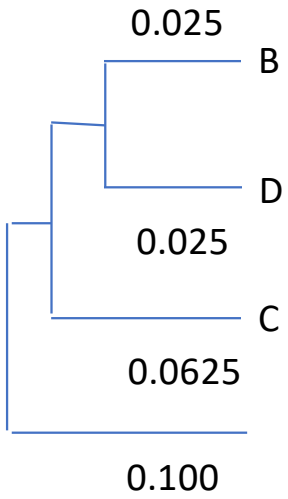



branch = 0.125/2 = 0.0625

# Jak převést „multiple alignment“ na strom?

## UPGMA

| | BDC | A |
|---|---|---|
| BDC | - | (0.2+0.25+0.15)/3= 0.2 |
| A | | - |




branch = 0.2/2 = 0.100

## Jak převést „multiple alignment“ na strom?

- **Neighbor-joining method – „spojování sousedních objektů“ (Saitou a Nei 1987) . Využívá distanční matici.**

  \+ Jednoduché = rychlé

  \+ Velmi rychlé i pro velké datasety

  \+ Vhodné pro prvotní analýzu

  \+ Mutační rychlost se na větvích může volně měnit

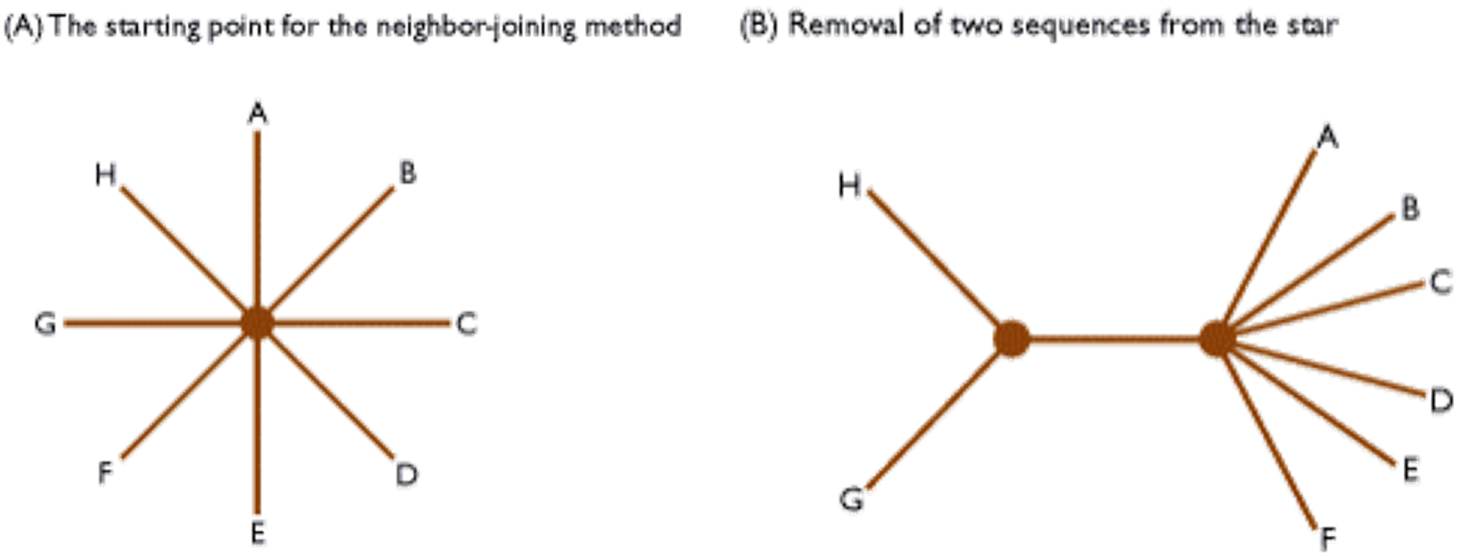
  \+ Statisticky konzistentní algoritmus při různých evolučních jevech

  \- Informace z alignmentu velmi zredukována = je převedena na distanční parametr, ztrácí se informace ze sekvence

  \- Poskytuje pouze jeden výsledný strom (netestuje jiné možnosti stromové topologie)

## Jak převést „multiple alignment“ na strom?

- **Neighbor-joining method– „spojování sousedních objektů“ (Saitou a Nei 1987) . Využívá distanční matici.**



# Jak převést „multiple alignment“ na strom? Neighbor-joining method

the NJ method does not assume the taxa to be equidistant from the root. It corrects for unequal evolutionary rates between sequences by using a conversion step. This conversion requires the calculations of “*r*-values” and “transformed *r*-values”

$$d'_{AB} = d_{AB} - 1/2 \times (r_A + r_B)$$

$d'_{AB}$… converted distance between A and B (rate corrected distance)
$d_{AB}$… actual evolutionary distance between A and B
$r_A$… sum of distances of A to all other taxa
$r_B$… sum of distances of B to all other taxa

$$r_i = \Sigma d_{ij}$$

needed for creation of a modified distance matrix

$$r'_i = r_i / n - 2$$

The transformed *r*-values ($r'_i$) are used to determine the distances of an individual taxon to the nearest node.

# Jak převést „multiple alignment“ na strom? Neighbor-joining method

| | A | B | C |
|---|---|---|---|
| B | 0.40 | | |
| C | 0.35 | 0.45 | |
| D | 0.60 | 0.70 | 0.55 |

$$r_i = \Sigma d_{ij}$$

$$r'_i = r_i / n - 2$$

calculation $r_i$ and $r'_i$ from the distance matrix

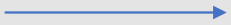
$$r_A = AB + AC + AD = 0.4 + 0.35 + 0.6 = 1.35$$
$$r'_A = r_A / (4 - 2) = 1.35/2 = 0.675$$
$$r_B = BA + BC + BD = 0.4 + 0.45 + 0.7 = 1.55$$
$$r'_B = r_B / (4 - 2) = 1.55/2 = 0.775$$
$$r_C = CA + CB + CD = 0.35 + 0.45 + 0.55 = 1.35$$
$$r'_C = r_C / (4 - 2) = 1.35/2 = 0.675$$
$$r_D = DA + DB + DC = 0.6 + 0.7 + 0.55 = 1.85$$
$$r'_D = r_D / (4 - 2) = 1.85/2 = 0.925$$

# Jak převést „multiple alignment“ na strom? Neighbor-joining method

calculation of rate corrected distances

$$d'_{AB} = d_{AB} - 1/2 * (r_A + r_B) = 0.4 - (1.35 + 1.55)/2 = -1.05$$
$$d'_{AC} = d_{AC} - 1/2 * (r_A + r_C) = 0.35 - (1.35 + 1.35)/2 = -1$$
$$d'_{AD} = d_{AD} - 1/2 * (r_A + r_D) = 0.6 - (1.35 + 1.85)/2 = -1$$
$$d'_{BC} = d_{BC} - 1/2 * (r_B + r_C) = 0.45 - (1.55 + 1.35)/2 = -1$$
$$d'_{BD} = d_{BD} - 1/2 * (r_B + r_D) = 0.7 - (1.55 + 1.85)/2 = -1$$
$$d'_{CD} = d_{CD} - 1/2 * (r_C + r_D) = 0.55 - (1.35 + 1.85)/2 = -1.05$$

| | A | B | C |
|---|---|---|---|
| B | 0.40 | | |
| C | 0.35 | 0.45 | |
| D | 0.60 | 0.70 | 0.55 |



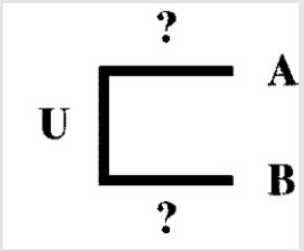
new distance matrix

| | A | B | C |
|---|---|---|---|
| B | -1.05 | | |
| C | -1 | -1 | |
| D | -1 | -1 | -1.05 |

# Jak převést „multiple alignment“ na strom? Neighbor-joining method

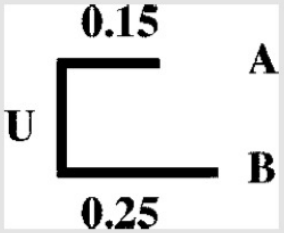
| | A | B | C |
|---|---|---|---|
| B | -1.05 | | |
| C | -1 | -1 | |
| D | -1 | -1 | -1.05 |

AB and CD are shortest --≥ we can start with AB or CD, result will be the same




$$d_{AU} = [d_{AB} + (r'_A - r'_B)]/2 = [0.4 + (0.675 - 0.775)]/2 = 0.15$$

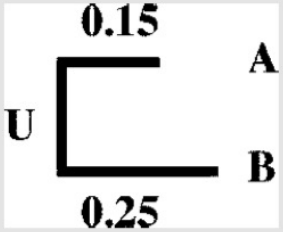
$$d_{BU} = [d_{AB} + (r'_B - r'_A)]/2 = [0.4 + (0.775 - 0.675)]/2 = 0.25$$



# Jak převést „multiple alignment“ na strom? Neighbor-joining method

original matrix

| | A | B | C |
|---|---|---|---|
| B | 0.40 | | |
| C | 0.35 | 0.45 | |
| D | 0.60 | 0.70 | 0.55 |



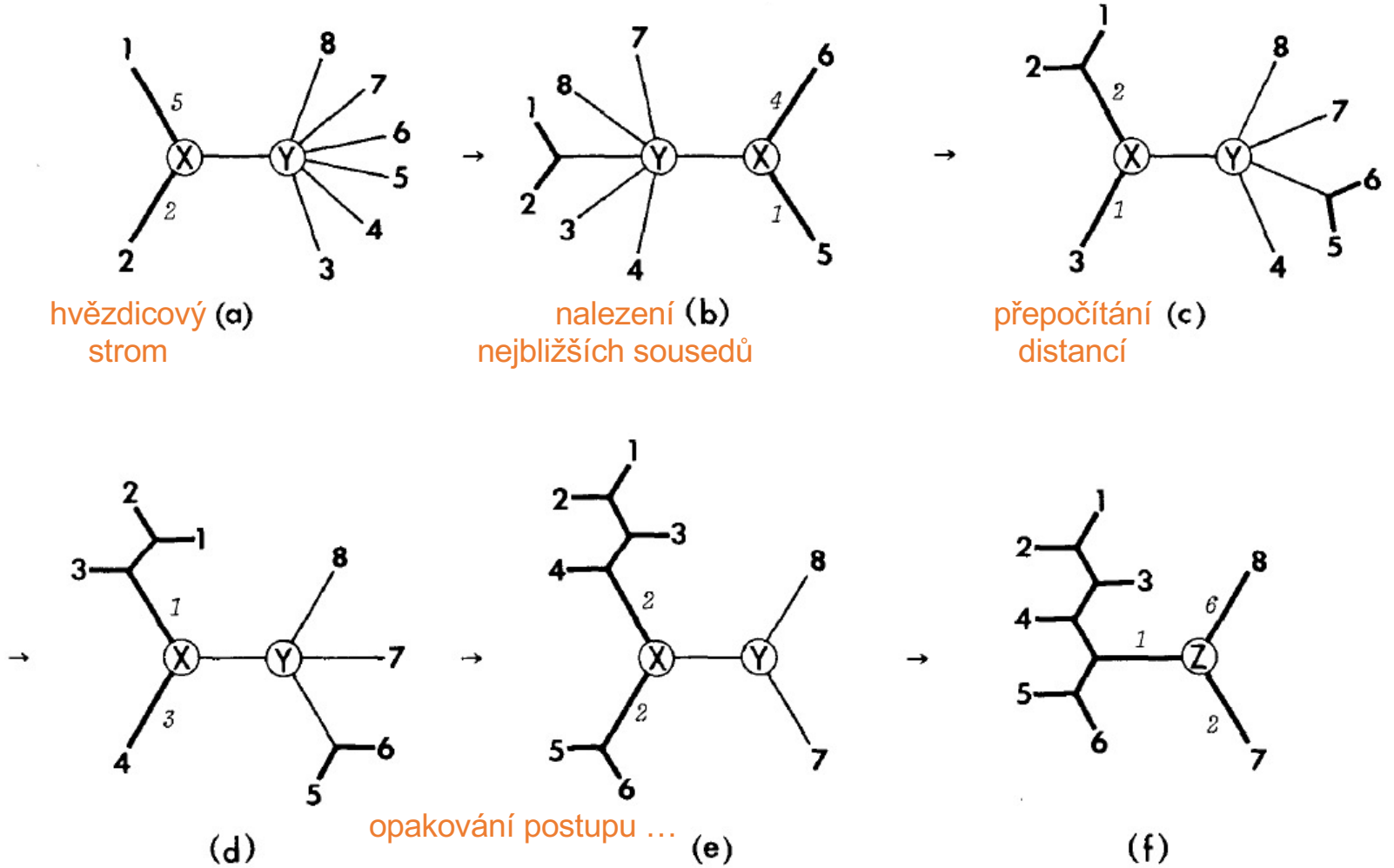

reduction of matrix:

$$d_{CU} = [(d_{AC} - d_{UA}) + (d_{BC} - d_{UB})]/2 = [(0.35 - 0.15) + (0.45 - 0.25)]/2 = 0.2$$

$$d_{DU} = [(d_{AD} - d_{UA}) + (d_{BD} - d_{UB})]/2 = [(0.6 - 0.15) + (0.7 - 0.25)]/2 = 0.45$$

| | U | C |
|---|---|---|
| C | 0.20 | |
| D | 0.45 | 0.55 |

repeating former steps (calculating $r_i$ and $r'_i$,
recalculation corrected distances,
creation rate-corrected distance matrix,
selection of the shortest distance,
creation of a node,
calculation of branch lengths, matrix reduction)




**The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees**[1]

*Naruya Saitou*[2] *and Masatoshi Nei*

# Nevýhody distančních dat

- **ztráta části informace během transformace**
  **jakmile data transformována na distance, nelze se vrátit zpět**
  (odlišné sekvence mohou dát stejné distance)
- **nelze sledovat evoluci na různých částech sekvence**
- **obtížná biologická interpretace délek větví**
- **nelze kombinovat různé distanční matice**

# Jak převést „multiple alignment“ na strom?

Character-based methods – diskrétní metody

- **Metody maximální úspornosti – maximum parsimony method.** Předpokládá (správně???), že evoluce jde nejkratší možnou cestou, tj. správný fylogenetický strom je ten, který požaduje minimum nukleotidových změn, aby bylo dosaženo daného rozdílů mezi sekvencemi (založena na výběru informativních míst v sekcencích).
- “Occamova břitva“ (sestavit sekvence tak, aby se **minimalizovalo** množství evolučních změn na stromu)
- **Neparametrická metoda** = nepoužívá substituční model (Délka větví představuje vždy počet změn, ke kterým došlo mezi dvěma uzly - u parametrických metod je délka větví evoluční změna/čas. Mutace, které se vyskytují jen u jedné sekvence datasetu, jsou neinformativní pro parsimonii, nemají význam pro určení vztahů mezi sekvencemi a pouze určují délku terminální větve)
- **Umožňuje hodnotit indely** jako pátou bázi (výhodné, pokud se indely vyskytují jednotlivě, nebo jsou delší úseky vzácné << 1% sekvence)

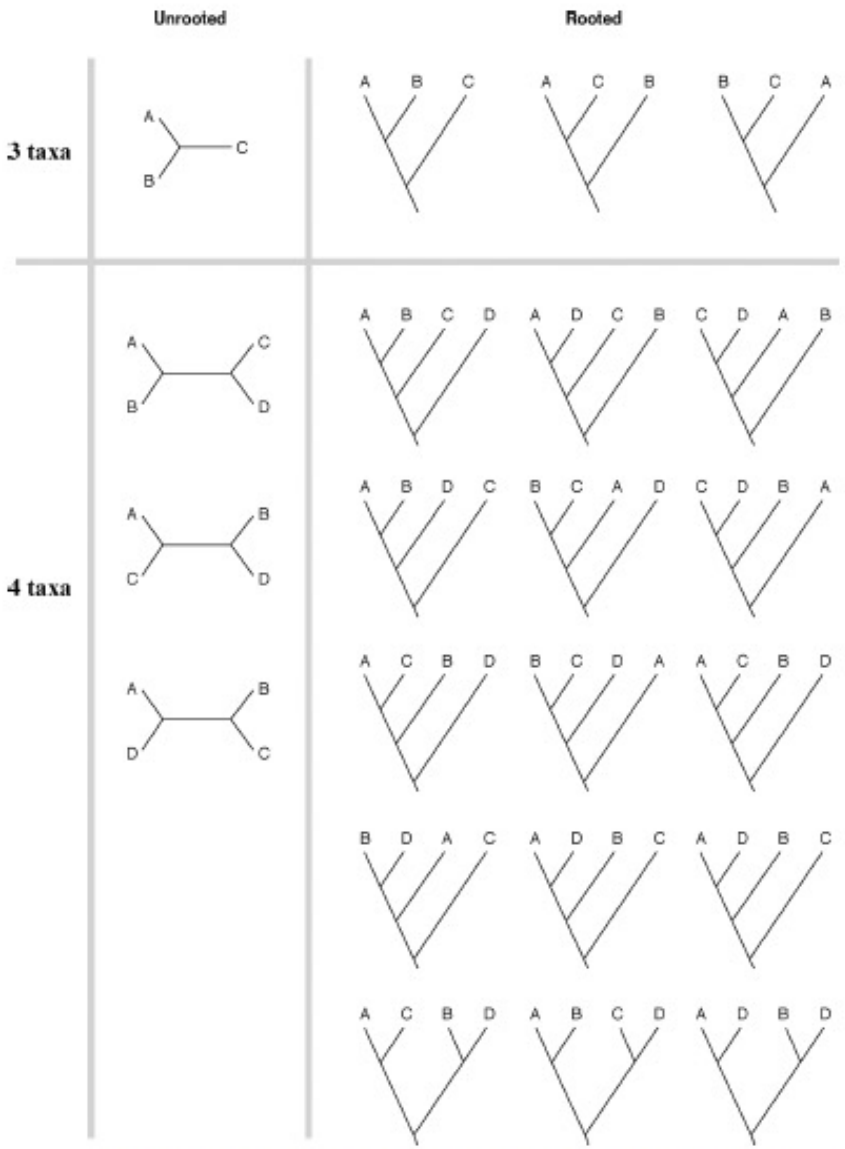
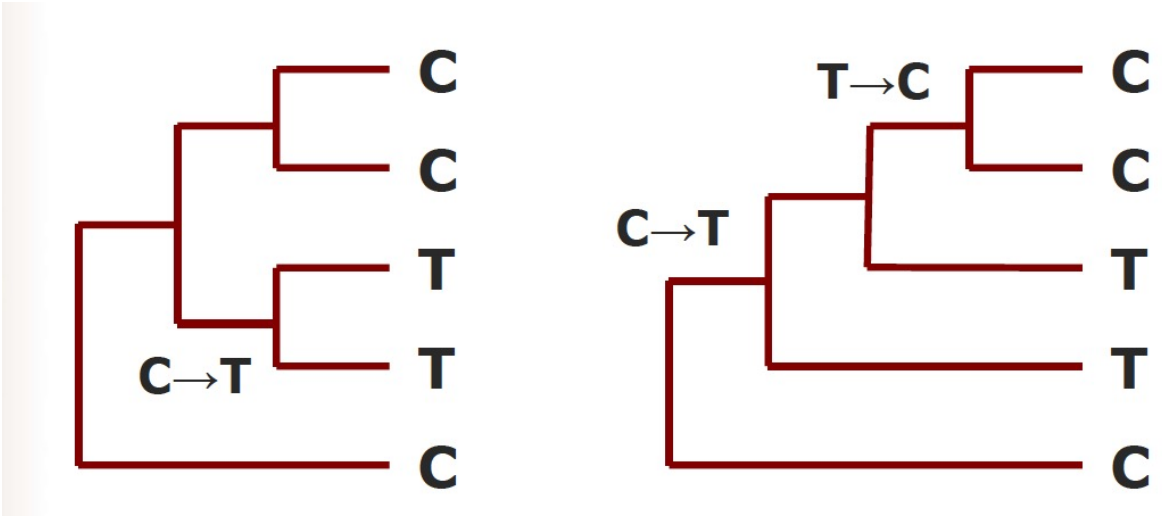



**Figure 10.7**: All possible tree topologies for three and four taxa. For three taxa, there are one unrooted and three rooted trees. For four taxa, there are three unrooted and fifteen rooted trees.

# Maximální úspornost

William of Occam (c. 1285 - c. 1349): *Occamova břitva*




1 krok

2 kroky

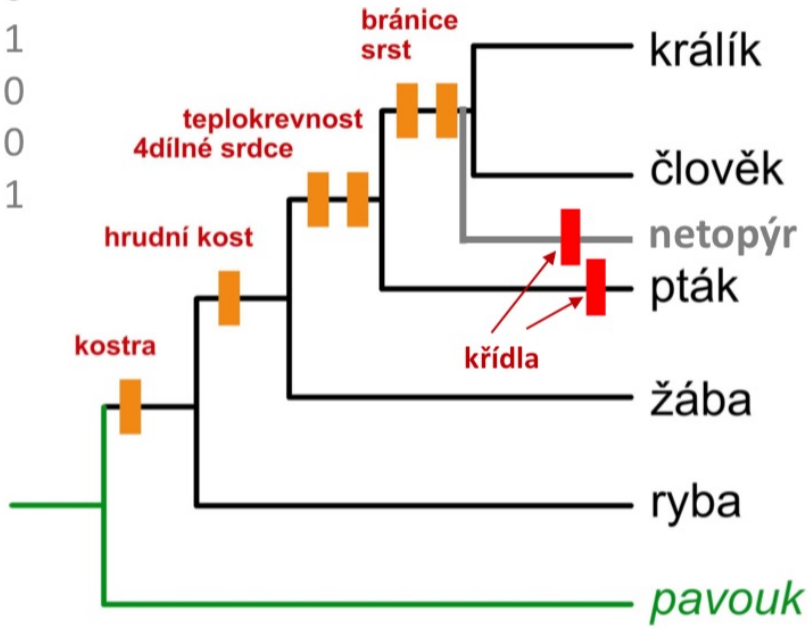
nebezpečí homoplasie

**znaky – např. morfologie, anatomie, fyziologie, atd. – může to být cokoliv**

převzato z materiálů UK – Evoluční genetika

| | kostra | teplokrevnost | hrudní kost | čtyřdílné srdce | bránice | srst | "křídla" |
|---|---|---|---|---|---|---|---|
| ryba | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| žába | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| pták | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| králík | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| člověk | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| netopýr | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

přítomnost znaků "bránice" a "srst" nám pomůže odhalit to, že křídla nevznikla jednou, ale dvakrát (a tedy že nejsou homologní, ale je to tzv. **homoplázie**)




**!!polarizace znaků – co je ancestrální víme až poté co známe fylogenezi, založenou i na jiných znacích...**

# Počet zakořeněných stromů

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

exponenciální nárůst počtu potenciálních stromů

Lze použít pro prohledávání kombinací
max 10 – 11 taxonů

| species | number of trees |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 3 |
| 4 | 15 |
| 5 | 105 |
| 6 | 945 |
| 7 | 10,395 |
| 8 | 135,135 |
| 9 | 2,027,025 |
| 10 | 34,459,425 |
| 11 | 654,729,075 |
| 12 | 13,749,310,575 |
| 13 | 316,234,143,225 |
| 14 | 7,905,853,580,625 |
| 15 | 213,458,046,676,875 |
| 16 | 6,190,283,353,629,375 |
| 17 | 191,898,783,962,510,625 |
| 18 | 6,332,659,870,762,850,625 |
| 19 | 221,643,095,476,699,771,875 |
| 20 | 8,200,794,532,637,891,559,375 |
| 30 | $4.9518 \times 10^{38}$ |
| 40 | $1.00985 \times 10^{57}$ |
| 50 | $2.75292 \times 10^{76}$ |

**+ Preciznější**

**- Větší nároky na manipulaci s daty**

**- Čím vice sekvencí, tím více topologií stromů je nutné vyzkoušet**

## Jak převést „multiple alignment“ na strom?

- **Parsimonie**:
  - * Fitchova parsimonie (X→Y, Y→ X)
  - * Wagnerova parsimonie (reverzibilita změn)
  - * Dollova parsimonie (povoluje znaku vzniknout jen jednou, paralelní a konvergentní získání znaku není povolené)
  - * Caminova-Sokalova parsimonie (změny ireverzibilní)
  - * Vážená parsimonie (ne všechny znaky jsou stejně informativní)
  - * Generalizovaná parsimonie (zobecnění uvedených typů, přirazení „costs“ všem možným typům změn)

## Jak převést „multiple alignment“ na strom?

Character-based methods – diskrétní metody

- **Metody maximální věrohodnosti – maximum likelihood method.**
- Maximalizuje věrohodnostní funkci (likelihood function)
- Nalezne hodnoty parametrů funkce, které nejlépe vysvětlují data – pravděpodobnost, že budeme pozorovat vstupní data při daném stromu a modelu s danými parametry je nejvyšší dosažená
- Ve fylogenetické rekonstrukci věrohodnostní funkce počítá pravděpodobnost konkrétní evoluční historie (strom) při konkrétním nastavení substitučního modelu, resp. posuzují se jednotlivé hypotézy o evoluční historii zkoumaných taxonů z hlediska pravděpodobnosti, že jsou v souladu se získanými daty, výsledek – maximálně pravděpodobný odhad
- **Tři součásti - vstupní data, evoluční model, fylogenetický strom s topologií i délkou větví**

# Algoritmy ML

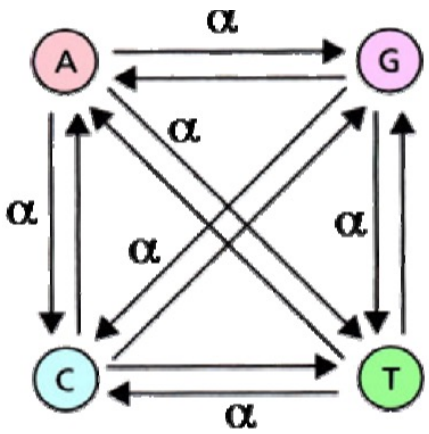
- Heuristiky – zkracují čas analýzy tím, že prohledávají jenom část vytvořených stromů
- Modifikují počáteční strom a hodnotí, nakolik změna ovlivnila věrohodnost stromu
  - Stochastický proces
  - Výměna nejbližšího souseda – nearest neighbour interchange (NNI)
  - Vyřezání a přesunutí podstromu – subtree pruning and regrafting (SPR)
  - Rozdělení a spojení stromu – tree bisection and reconnection (TBR)
- Přijímají vylepšení  původního stromu
  - Lezení do kopce – hill climbing (vylepšování věrohodnosti)
  - obdobně jako u modelování – může se dostat do lokálního optima
  - Je potřeba pustit analýzu vícekrát s různými počátečními náhodnými čísly a zkontrolovat, zda se výsledné stromy liší (obdoba Monte Carlo přístupu)

## Jak převést „multiple alignment“ na strom?

Character-based methods – diskrétní metody

- Bayesian inference (Bayesovská statistika)
- Výpočet pravděpodobnosti na základě specifikovaného modelu a na základě toho, co jsme o charakteru dat zjistili
- Základ – strom s danou topologií a délkami větví, model nukleotidových substitucí a rozložení substitučních frekvencí mezi jednotlivými nukleotidy
- Princip přístupu jako u ML
- VÝHODY – menší časová náročnost, strom zohledňující fylogenetický signál v datasetu, možnost použít i pro smíšený dataset
- Různé substituční modely (modely evoluce sekvencí)

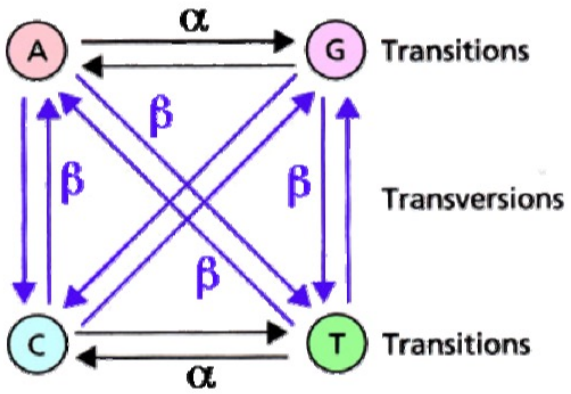
# Jukes - Cantor model (JC)




$$\mathbf{P}_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}, \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}]$$

nejjednodušší model, stejné pravděpodobnost záměny jakéhokoli nukleotidu v jiný

# Kimura´s 2-parameter model (K2P)




$$\mathbf{P}_t = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix},$$

zohledňuje pravděpodobnost záměny v rámci purinvých/anebo pyrimidinových bází:
tranzice 1, transverze 2

# Felsenstein 1981's model (F81)

Některé typy substitucí můžou být častější než jiné proto, že jsou ve zkoumaných sekvencích početnejší.
Tento model uvažuje nestejné frekvence pro všechny 4 nukleotidy

$$\mathbf{P}_t = \begin{bmatrix} . & \pi_C\alpha & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & . & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & . & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & \pi_G\alpha & . \end{bmatrix},$$

$$\mathbf{f} = [\pi_A\ \pi_C\ \pi_G\ \pi_T]$$

$\pi$ je průměrná frekvence báze i v porovnávaných sekvencích

# General time-reversible model (GTR)

Nejobecnější model, všech 6 typů substitucí má rozdílnou frekvenci

$$\mathbf{P}_t = \begin{bmatrix} . & \pi_C a & \pi_G b & \pi_T c \\ \pi_A a & . & \pi_G d & \pi_T e \\ \pi_A b & \pi_C d & . & \pi_T f \\ \pi_A c & \pi_C e & \pi_G f & . \end{bmatrix}, \qquad \mathbf{f} = [\pi_A \; \pi_C \; \pi_G \; \pi_T]$$

# Pro vysvětlení Bayesovského teoremu

- dopuručuji shlédnout pěkné video:

https://www.youtube.com/watch?v=5NMxiOGL39M

# Software pro fylogenetickou analýzu

- T-Rex (Tree and reticulogram REConstruction) - is dedicated to the reconstruction of phylogenetic trees, reticulation networks and to the inference of horizontal gene transfer (HGT) events. T-REX includes several popular bioinformatics applications such as MUSCLE, MAFFT, Neighbor Joining, NINJA, BioNJ, PhyML, RAxML, random phylogenetic tree generator and some well-known sequence-to-distance transformation models. It also comprises fast and effective methods for inferring phylogenetic trees from complete and incomplete distance matrices as well as for reconstructing reticulograms and HGT networks (Reference: Alix, C. et al. 2012. Nucl. Acids Res. **40** (W1): W573-W579).

- Phylogeny.fr - is a simple to use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences.It includes multiple alignment (MUSCLE, T-Coffee, ClustalW, ProbCons), phylogeny (PhyML, MrBayes, TNT, BioNJ), tree viewer (Drawgram, Drawtree, ATV) and utility programs (e.g. Gblocks to eliminate poorly aligned positions and divergent regions) (Reference: A. Dereeper et al,. 2008. Nucl. Acids Res. **36** (Web Server Issue):W465-9). Also available here.

- FastME provides distance algorithms to infer phylogenies. FastME is based on balanced minimum evolution, which is the very principle of NJ. FastME improves over NJ by performing topological moves using fast, sophisticated algorithms. The first version of FastME only included Nearest Neighbor Interchange (NNI). The new 2.0 version also includes Subtree Pruning and Regrafting (SPR), while remaining as fast as NJ and providing a number of facilities: distance estimation for DNA and proteins with various models and options, bootstrapping, and parallel computations. (Reference: Lefort V. et al. Molecular Biology & Evolution **32(10):** 2798-800, 2015).

- PhyML - has been widely used because of its simplicity and a fair compromise between accuracy and speed. In the meantime research on PhyML has continued, and new algorithms and methods have been implemented in the program. (Reference: V. Lefort et al. Molecular Biology and Evolution, msx149, 2017).

- RAxML (**R**andomized **Ax**elerated **M**aximum **L**ikelihood) is a program for sequential and parallel Maximum Likelihood based inference of large phylogenetic trees (Reference: Stamatakis, A. 2006. Bioinformatics 22:2688–2690).

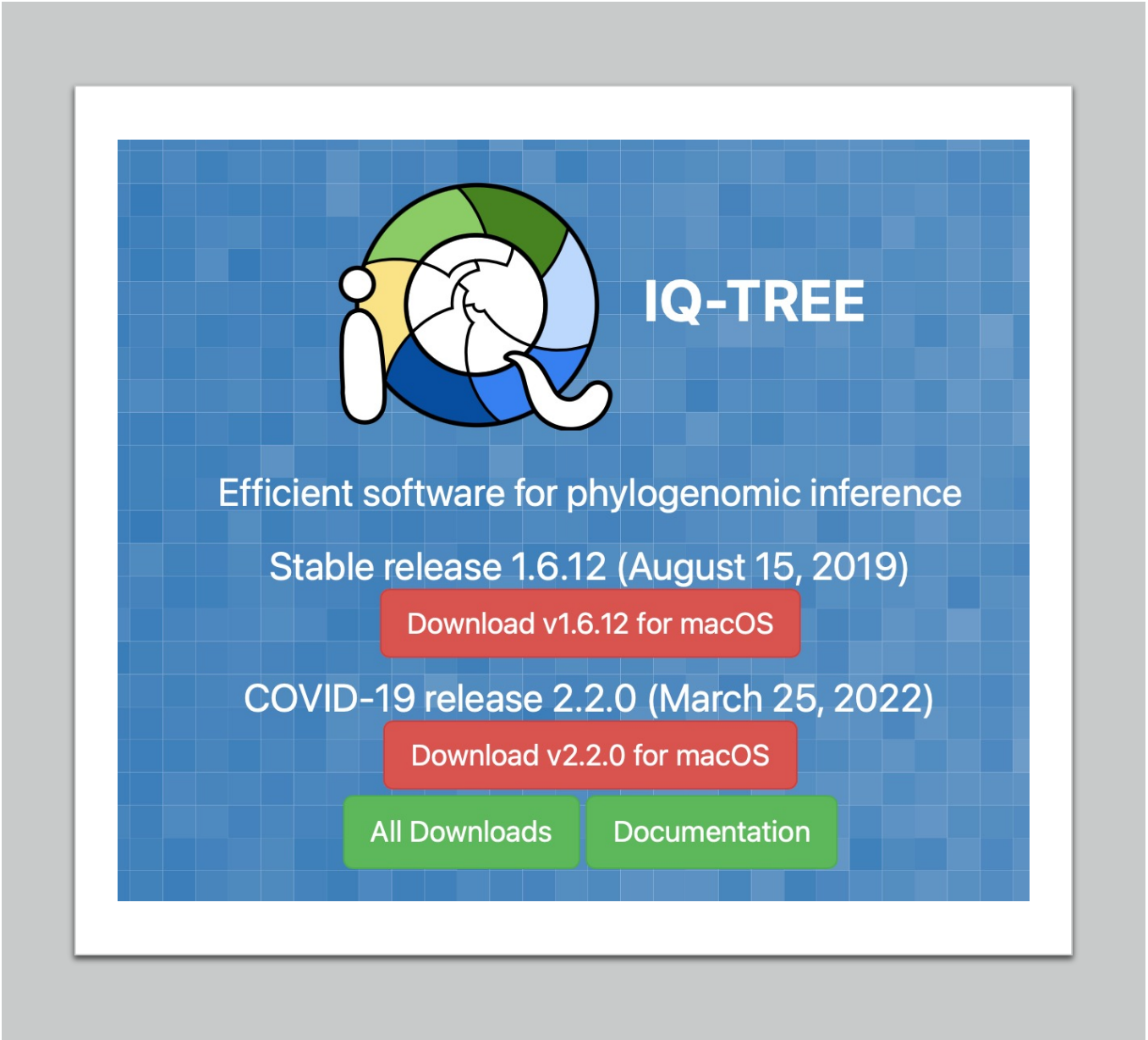
- ProtTest *(David Posada, University of Vigo, Spain)* - estimates the empirical model of aminoacid substitution that fits the data best among 64 candidate models. PROTTEST calculates AIC, AICc and BIC values, and obtain a rank of model fits, model-averaged parameter estimates, or measures of parameter importance. Mac OSX, Windows and Linux versions are available for downloading.

- Phylemon2 - a suite of web-tools for molecular evolution, phylogenetics and phylogenomics (Reference:Sánchez, R. et al. 2011.Nucl. Acids Res. 39/suppl_2/W470)

- POWER (**P**hyl**O**genetic **We**b **R**epeater) - allows users to carry out phylogenetic analysis on most programs of PHYLIP package repeatedly. POWER provide two pipelines to process the analysis. One of them includes multiple sequence alignment (MSA) at the beginning of the pipeline whereas the other begin phylogenetic analysis with aligned sequence. Very user friendly. (Reference: C.-Y. Lin. et al. 2005. Nucl. Acids Res. **33:** W553-W556).
- Phylodendron - phylogenetic tree printer *(D.G. Gilbert, Indiana Univ.)* - very useful in visualizing *.dnd file from aligments and saving the results as .GIF, .PS or .PDF files. N.B. The font style and size can be altered in the .PDF output format.
- Phylogenetic tree prediction - GeneBee service *(Belozersky Institute of Physico-chemical Biology, Moscow State University, Russia)*

# Software pro fylogenetickou analýzu

- **IQ-TREE**
**http://www.iqtree.org**
- **maximum likelihood**
- Implementované všechny substituční modely, podporu uzlu počítá nejen pomocí
- bootstrapové analýzy



# Software pro fylogenetickou analýzu

**PHYLIP**

PHYLIP (the *PHYL*ogeny *I*nference *P*ackage) is a package of programs for inferring phylogenies (evolutionary trees). It is available free over the Internet, and written to work on as many different kinds of computer systems as possible. The source code is distributed (in C), and executables are also distributed. In particular, already-compiled executables are available for Windows (95/98/NT/2000/me/xp/Vista), Mac OS X, Mac OS 8 and 9, and Linux systems. Complete documentation is available on documentation files that come with the package.

- **PHYLIP – *PHYL*ogeny *I*nference *P*ackage**

Methods that are available in the package include parsimony, distance matrix, and likelihood methods

**http://evolution.genetics.washington.edu/phylip.html**

# Software pro fylogenetickou analýzu


http://evolution.genetics.washington.edu/phylip.html

# Software pro fylogenetickou analýzu

## Phylogenetic Analysis by Maximum Likelihood (PAML)

### Introduction

PAML is a package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood. It is maintained and distributed for academic use free of charge by Ziheng Yang. ANSI C source codes are distributed for UNIX/Linux/Mac OSX, and executables are provided for MS Windows. PAML is not good for tree making. It may be used to estimate parameters and test hypotheses to study the evolutionary process, when you have reconstructed trees using other programs such as PAUP*, PHYLIP, MOLPHY, PhyML, RaxML, etc.

**http://abacus.gene.ucl.ac.uk/software/paml.html**


**http://macclade.org/index.html**