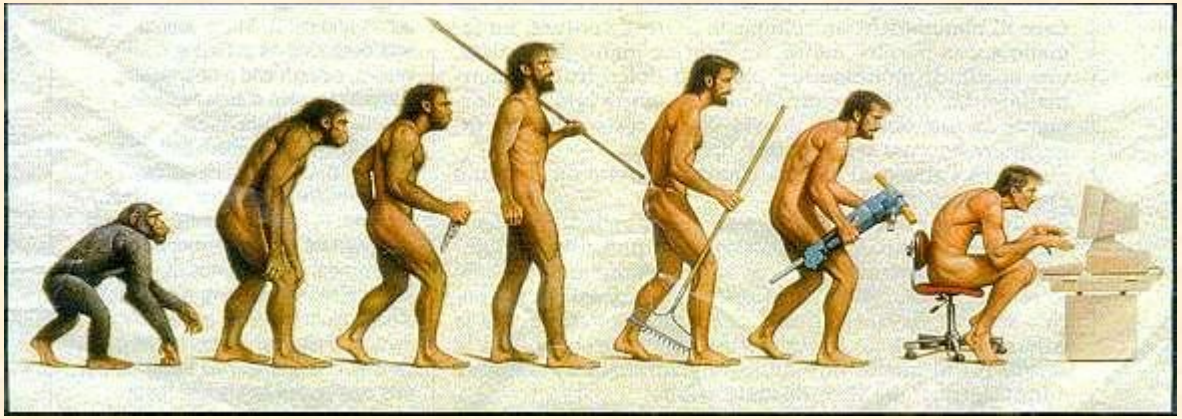
Fylogenetická evoluční analýza

Pro zajímavost…

Důležité…

Fylogeneze = vývoj druhu (vývoj nových druhů) procesem evoluce.

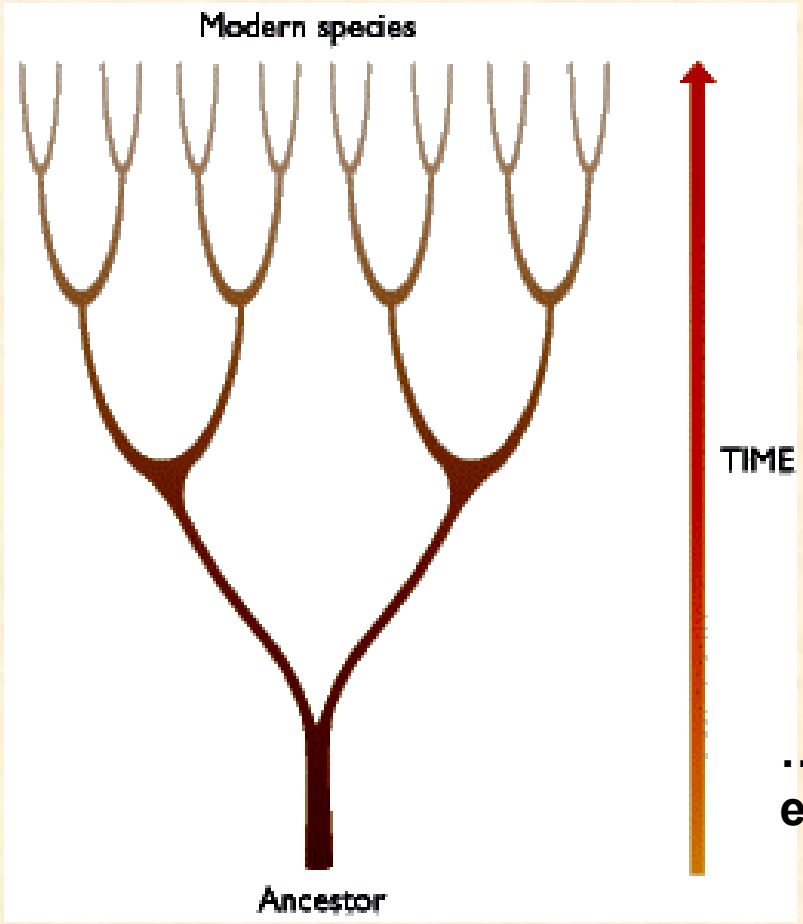
Fylogenetika = věda zkoumající fylogenezi, příbuzenské vztahy a vývoj organismů.



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.**

  Prvotně používány MORFOLOGICKÉ znaky.

  Problém – fosilní pozůstatky většinou NEKVALITNÍ, neposkytují žádané informace nebo se VŮBEC nedochovají.

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.

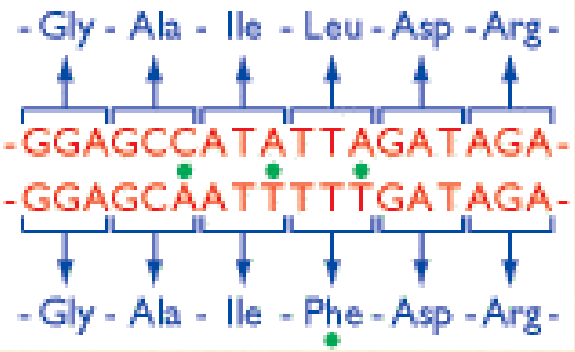
- **Jednotlivé stavy jsou jednoznačné a nezaměnitelné (A x C x G x T). Na rozdíl od morfologických znaků (tvar), u nichž existuje mnoho přechodových forem.**

- **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

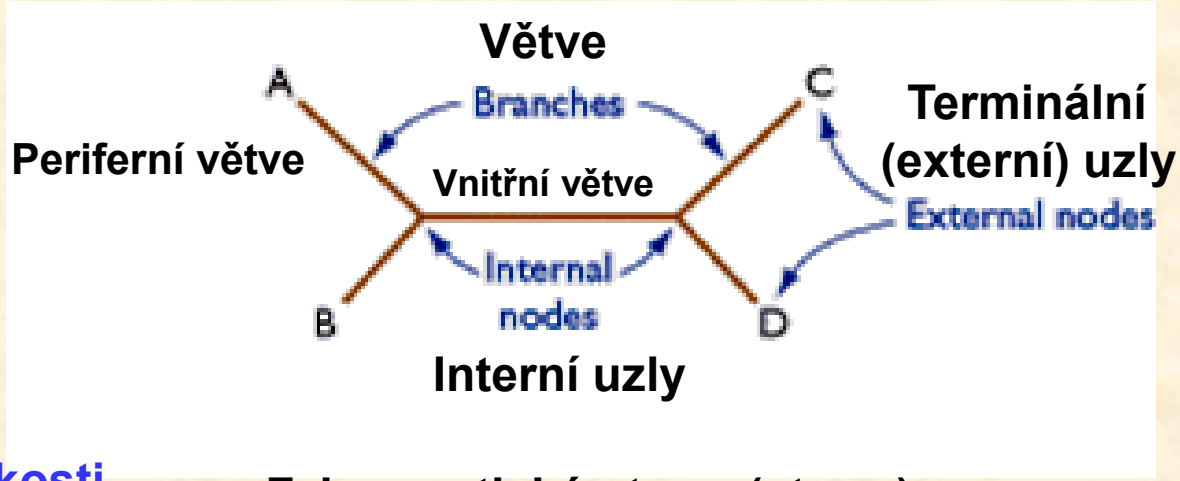
Fylogenetický strom

- **Cíl fylogenetické analýzy - fylogenetický strom popisující evoluční vztahy mezi studovanými organismy.**

Současné taxony (geny) = terminální (externí) uzly, vrcholy

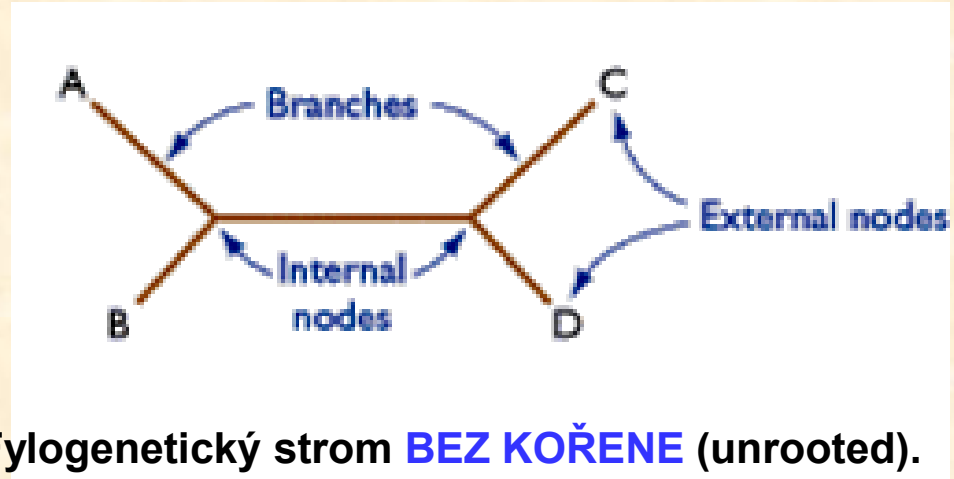
Interní uzly = rozdělení společného „předka“

Délky větví = uměrné velikosti změny v průběhu evoluce

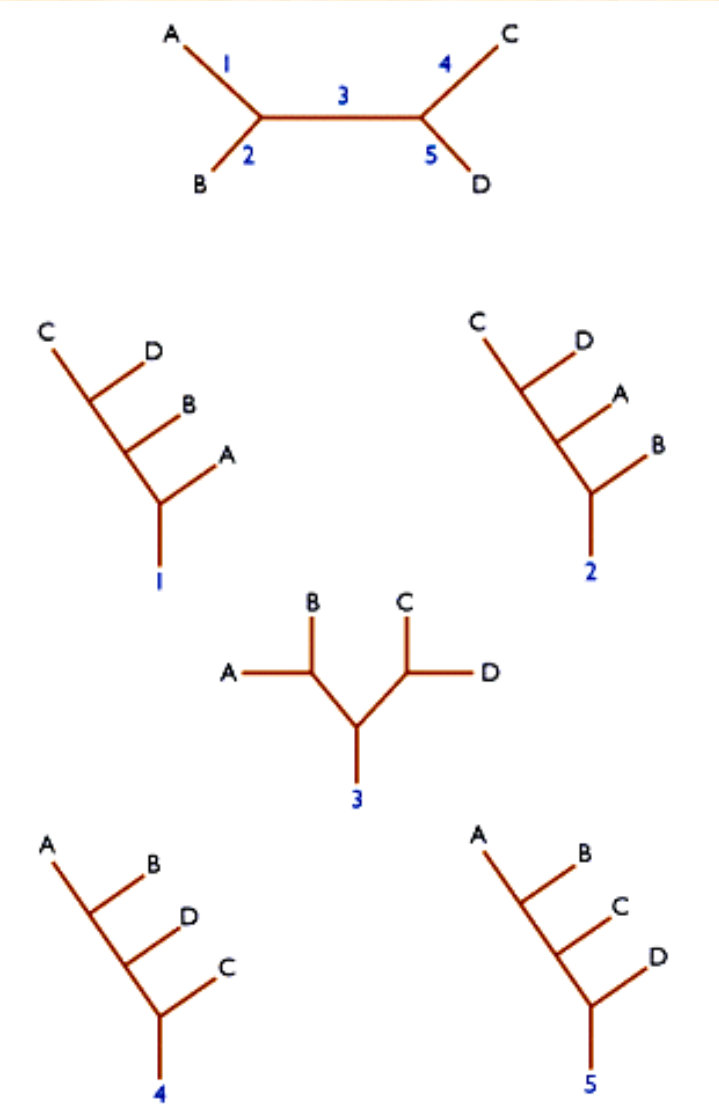


Fylogenetický strom (strom)

Fylogenetický strom

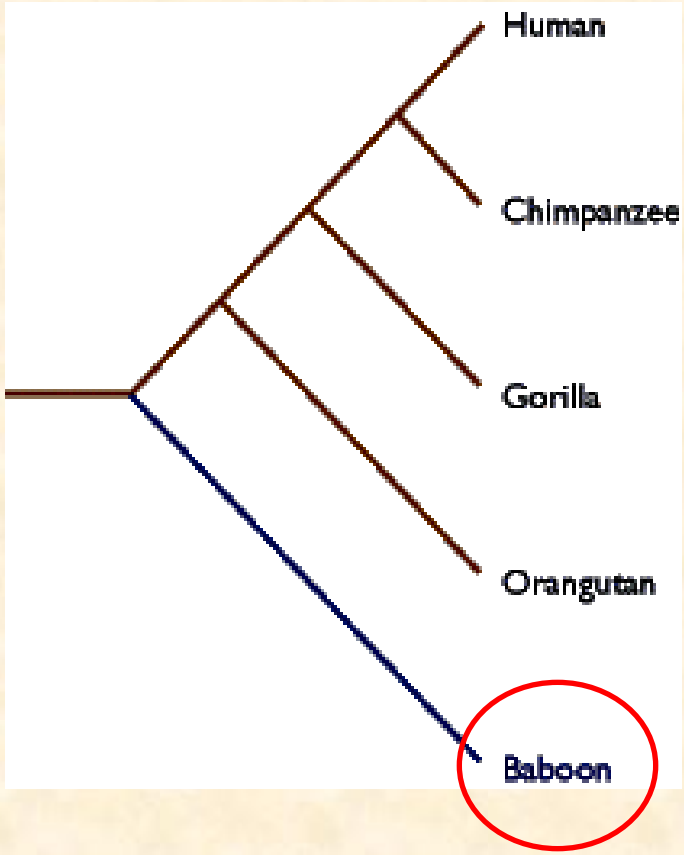


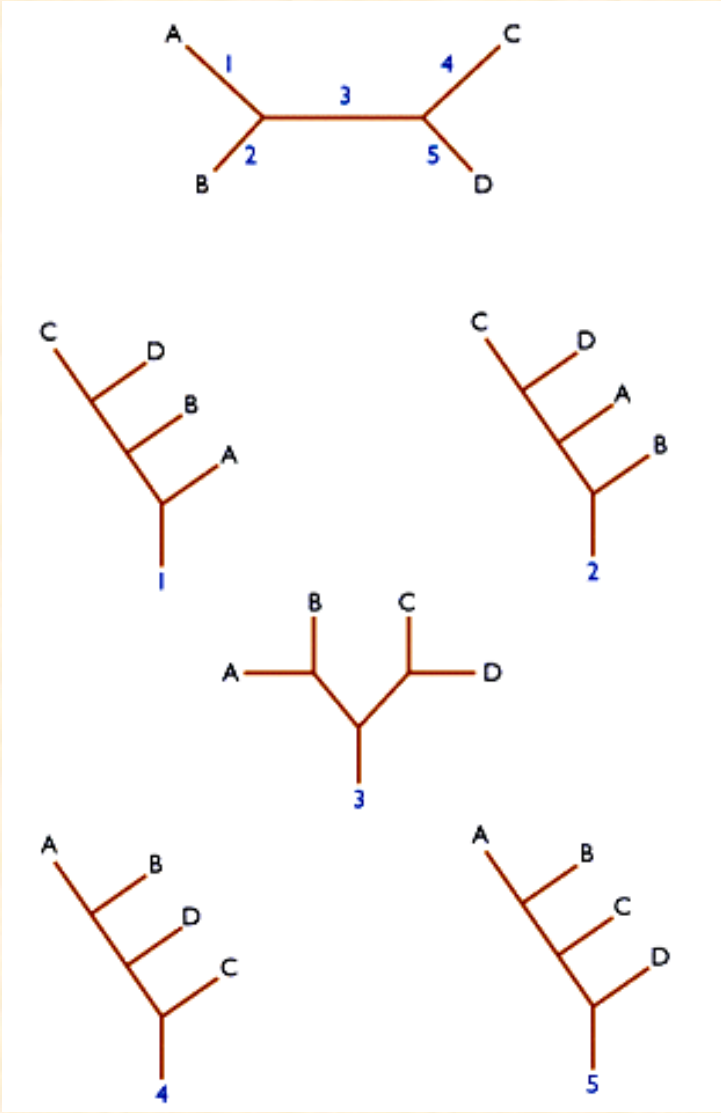
Fylogenetický strom **BEZ KOŘENE** (unrooted).
Není známý nejstarší společný předek (bod).
Vypovídá pouze o příbuzenských vztazích mezi geny, ne o „cestě“ kterou se evoluce ubírala.



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“

Fylogenetický strom





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“

„Genový“ strom x „druhový strom“

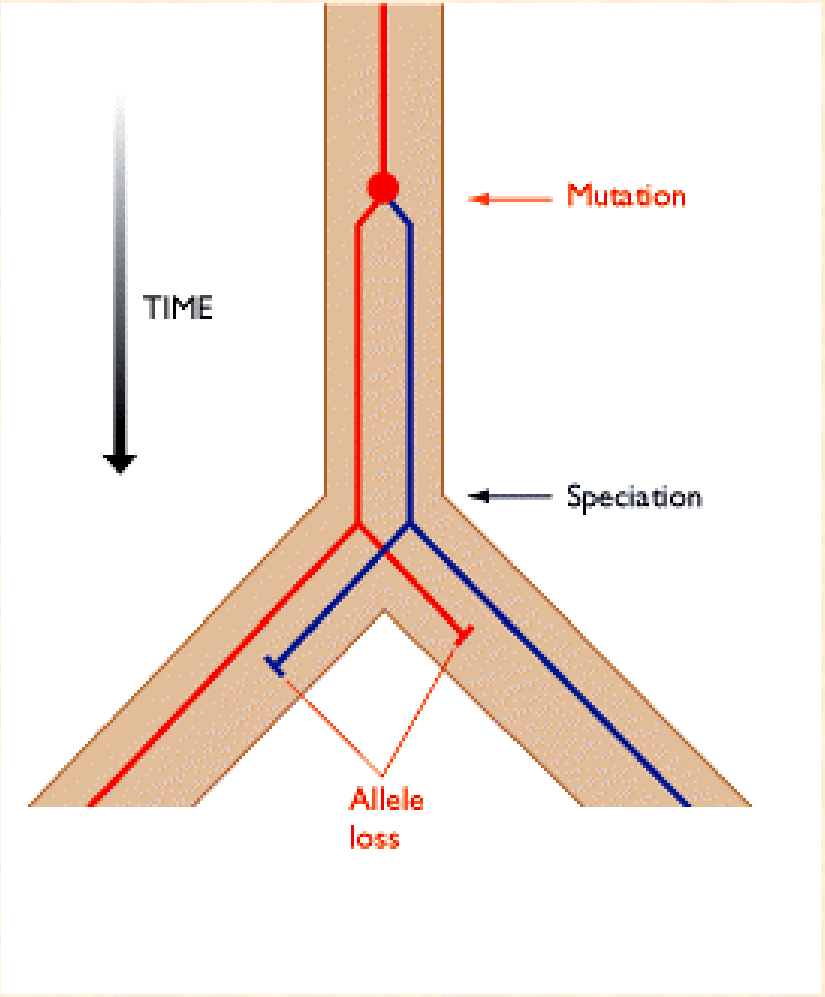
- **Genový strom – odvozen ze srovnání ortologních genů.** **Předpokládá se, že bude přesnější než strom získaný pomocí morfologických dat.**
- **Genový strom ≠ ruhový strom.**

  Genový strom – vnitřní uzly představují rozdělení původního GENU (mutace).

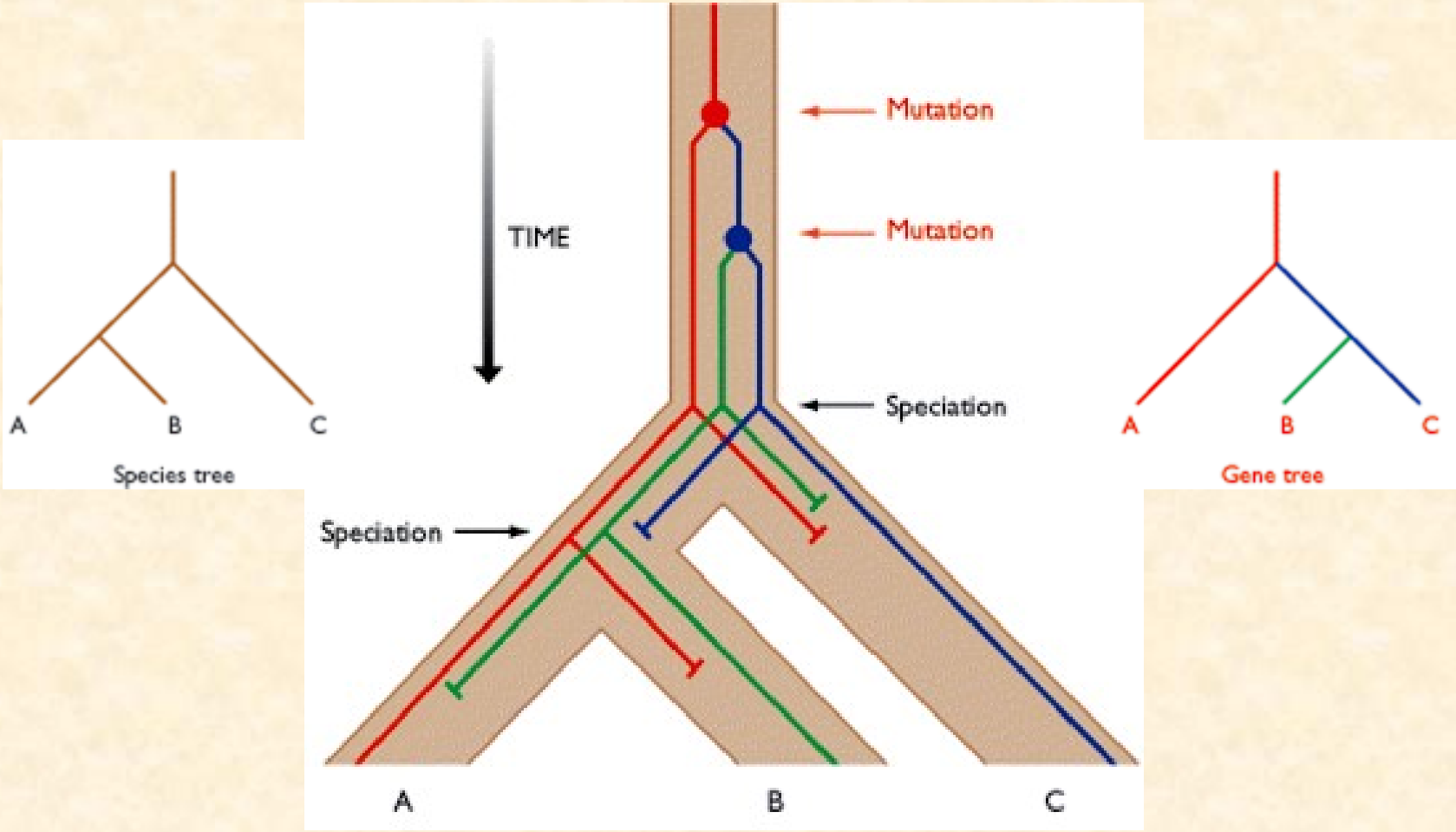
  Druhový strom – vnitřní uzly představují rozdělení populace původního DRUHU do dvou skupin (geografická izolace).

„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.**

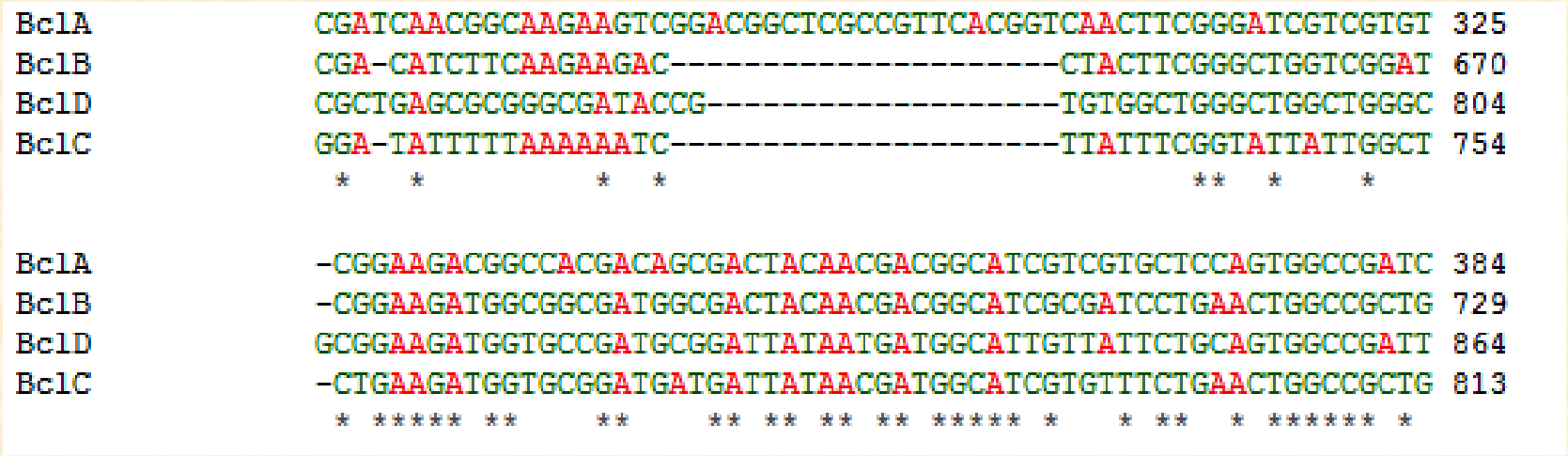


„Genový“ strom x „druhový strom“



Tvorba evolučních stromů

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



Jak převést „multiple alignment“ na strom?

- **Neexistuje „nejlepší metoda“.**
  Několik metod je používáno souběžně, žádnou nelze označit za lepší než ostatní.

- **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				–

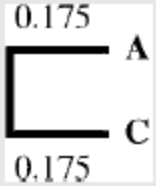
Jak převést „multiple alignment“ na strom?

- **UPGMA (Unweighted Pair Group Method with Arithmetic mean, Unweighted Pair Group Method Using Arithmetic Average). Využívá distanční matici.**

Box 11.1 An Example of Phylogenetic Tree Construction Using the UPGMA Method

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

1. Using a distance matrix involving four taxa, A, B, C, and D, the UPGMA method first joins two closest taxa together which are A and C (0.35 in grey). Because all taxa are equidistant from the node, the branch length for A to the node is AC/2 = 0.35/2 = 0.175.



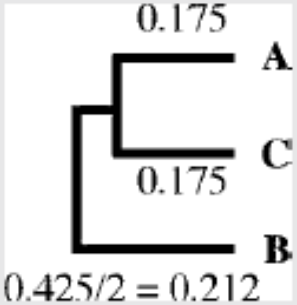
Essential Bioinformatics
JIN XIONG
Texas A&M University

Jak převést „multiple alignment“ na strom?

2. Because A and C are joined into a cluster, they are treated as one new composite taxon, which is used to create a reduced matrix. The distance of A-C cluster to every other taxa is one half of a taxon to A and C, respectively. That means that the distance of B to A-C is (AB + BC)/2; and that of D to A-C is (AD + CD)/2.

	A-C	B
B	$\frac{0.4 + 0.45}{2} = 0.425$	
D	$\frac{0.55 + 0.6}{2} = 0.575$	0.70

3. In the newly reduced-distance matrix, the smallest distance is between B and A-C (in grey), which allows the grouping of B and A-C to create a three-taxon cluster. The branch length for the B is one half of B to the A-C cluster.



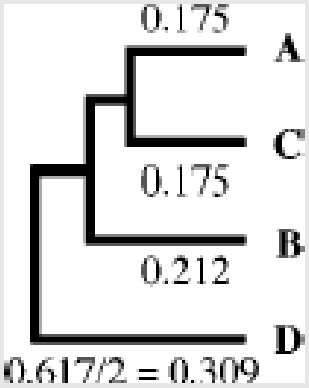
Essential Bioinformatics
JIN XIONG
Texas A&M University

Jak převést „multiple alignment“ na strom?

4. When B and A-C are grouped and treated as a single taxon, this allows the matrix to reduce further into only two taxa, D and B-A-C. The distance of D to the composite taxon is the average of D to every single component which is (BD + AD + CD)/3.

	B-A-C
D	$\frac{0.7 + 0.6 + 0.55}{3} = 0.617$

5. D is the last branch to add to the tree, whose branch length is one half of D to B-A-C.

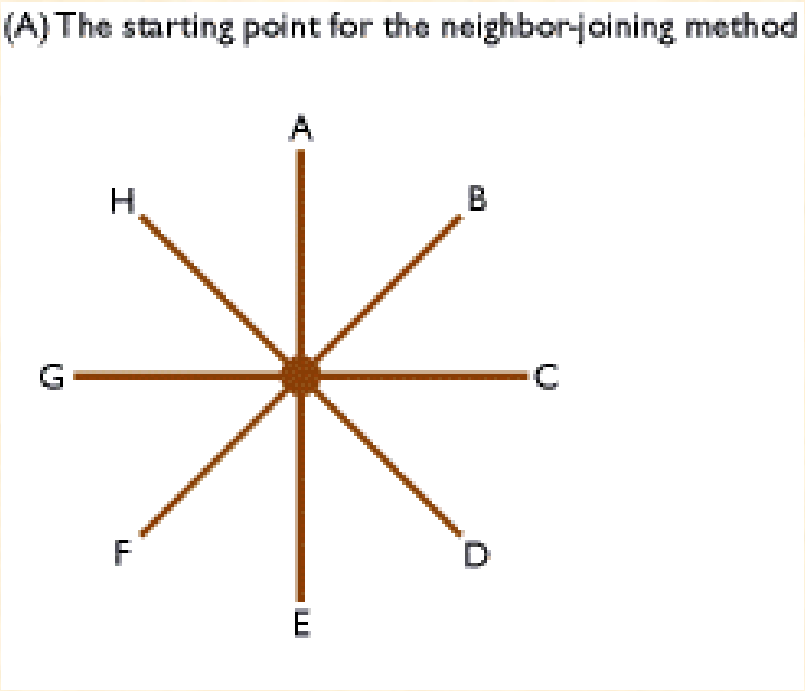


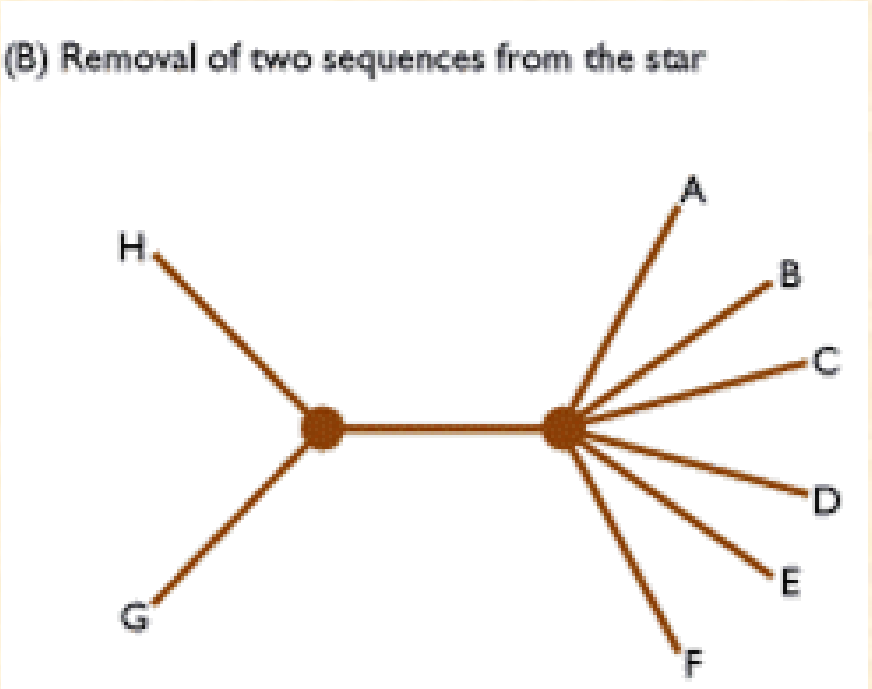
Essential Bioinformatics

JIN XIONG
Texas A&M University

Jak převést „multiple alignment“ na strom?

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





Box 11.2 Phylogenetic Tree Construction Using the Neighbor Joining Method

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

1. The NJ method is similar to UPGMA, but uses an evolutionary rate correction step before tree building. Using the same distance matrix as in the UPGMA tree building (see Box 11.1), the first step of the NJ method is r-value and r'-value calculation. According to Eq. 11.1 and 11.2, r and r' for each taxon are calculated as follows:

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

Essential Bioinformatics
JIN XIONG
Texas A&M University

2. Based on Eq. 11.4 and the above *r*-values, the corrected distances are obtained as follows:

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

3. The rate-corrected distances allow the construction of a new distance matrix.

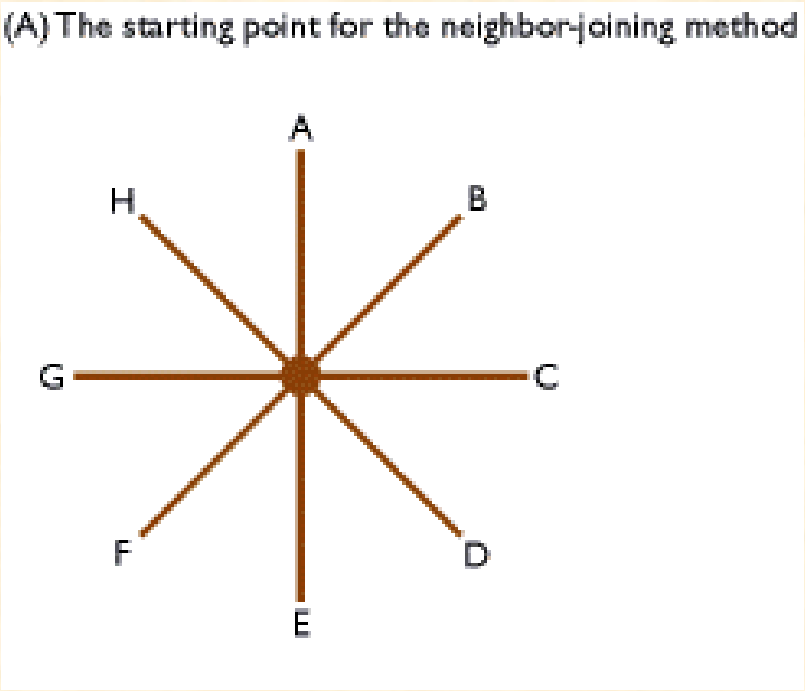
	A	B	C
B	-1.05		
C	-1	-1	
D	-1	-1	-1.05

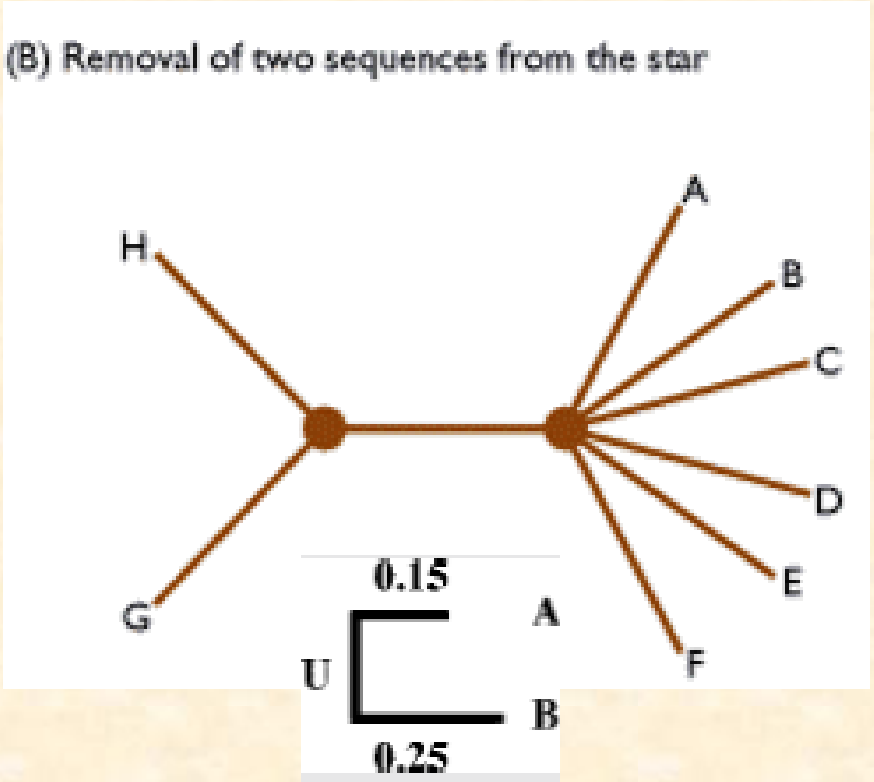
Essential Bioinformatics
JIN XIONG
Texas A&M University

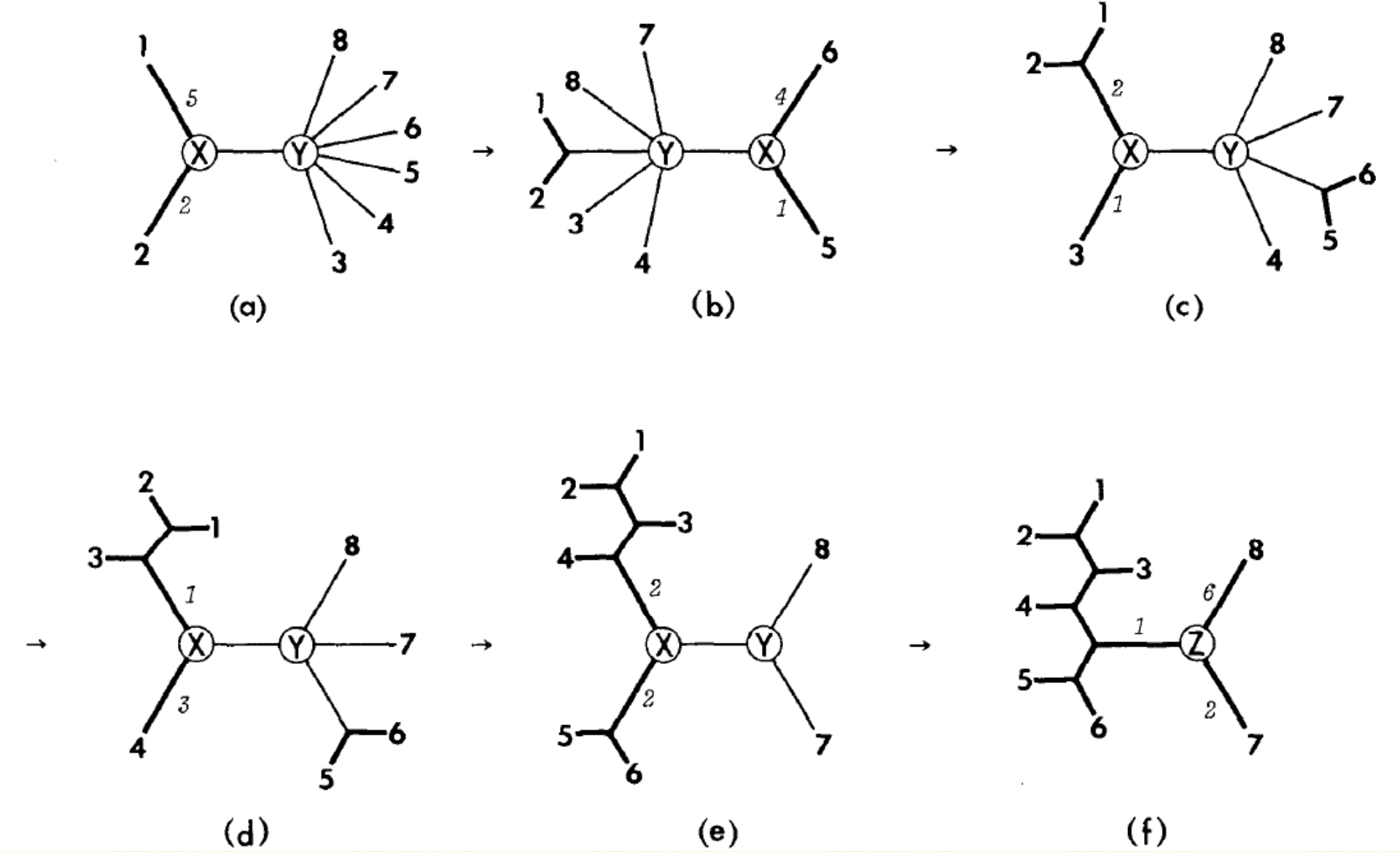
4. Before tree construction, all possible nodes are collapsed into a star tree. The pair of taxa with the shortest distances in the new

Jak převést „multiple alignment“ na strom?

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







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

Naruya Saitou[2] and Masatoshi Nei

Jak převést „multiple alignment“ na strom?

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

  \+ Jednoduché = rychlé

  \+ Vhodné pro velké soubory dat

  \+ Vhodné pro prvotní analýzu

  \- Informace z alignmentu velmi zredukována

  \- Poskytuje pouze jeden výsledný strom

Jak převést „multiple alignment“ na strom?

- **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.**

+ Preciznější

- Větší nároky na manipulaci s daty

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

- 5 sekvencí = 15 stromů, 10 sekvencí = 2 027 025 stromů

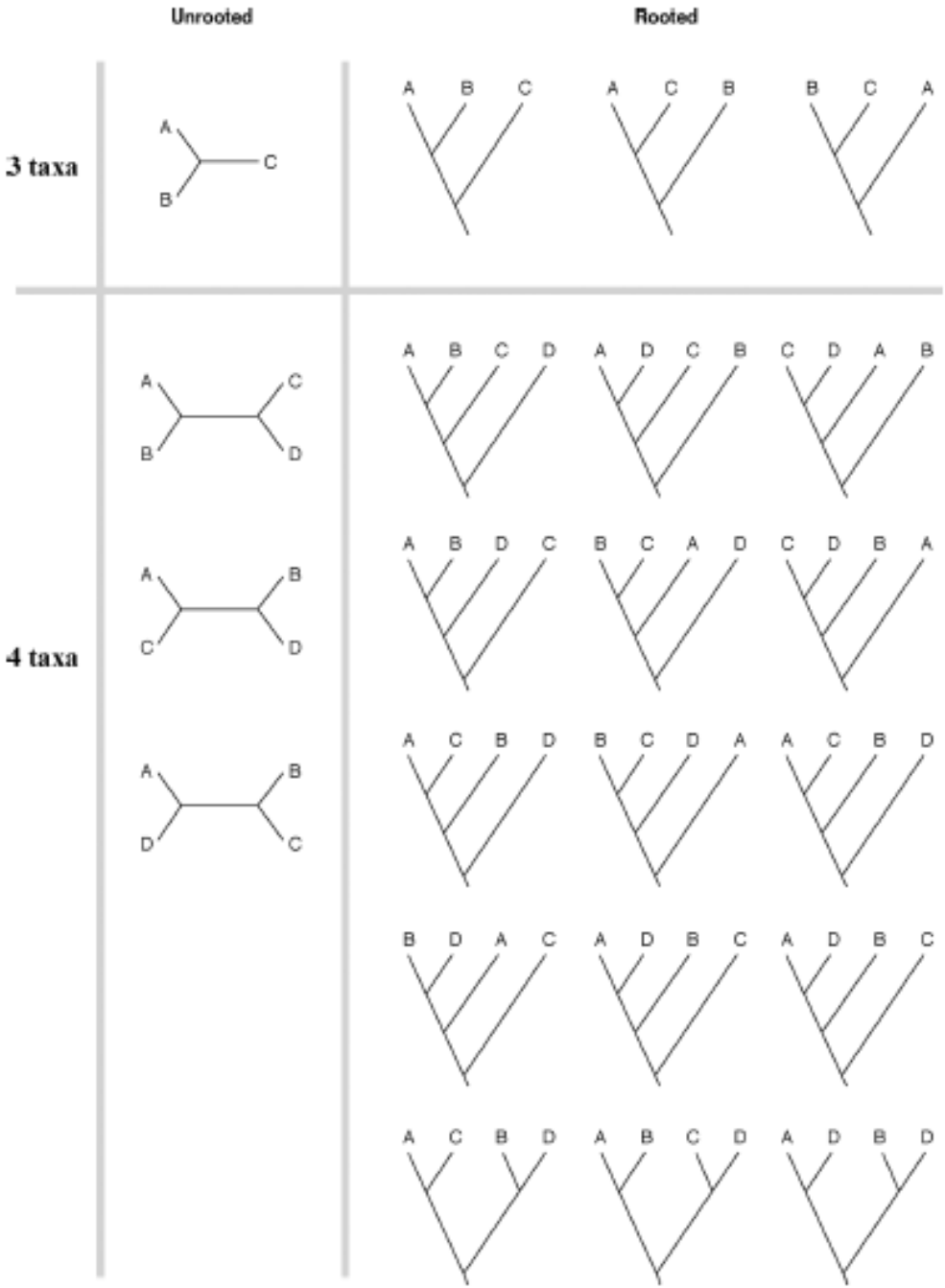


Figure 10.7: All possible tree topologies for three and four taxa. For three taxa, there are one unrooted and three rooted trees. For four taxa, there are three unrooted and fifteen rooted trees.

Jak převést „multiple alignment“ na strom?

- **Parsimonie**: **Fitchova parsimonie**

  Wagnerova parsimonie (reverzibilita změn)

  Dollova parsimonie („novinka“ může zaniknout)

  Caminova-Sokalova parsimonie (změny ireverzibilní)

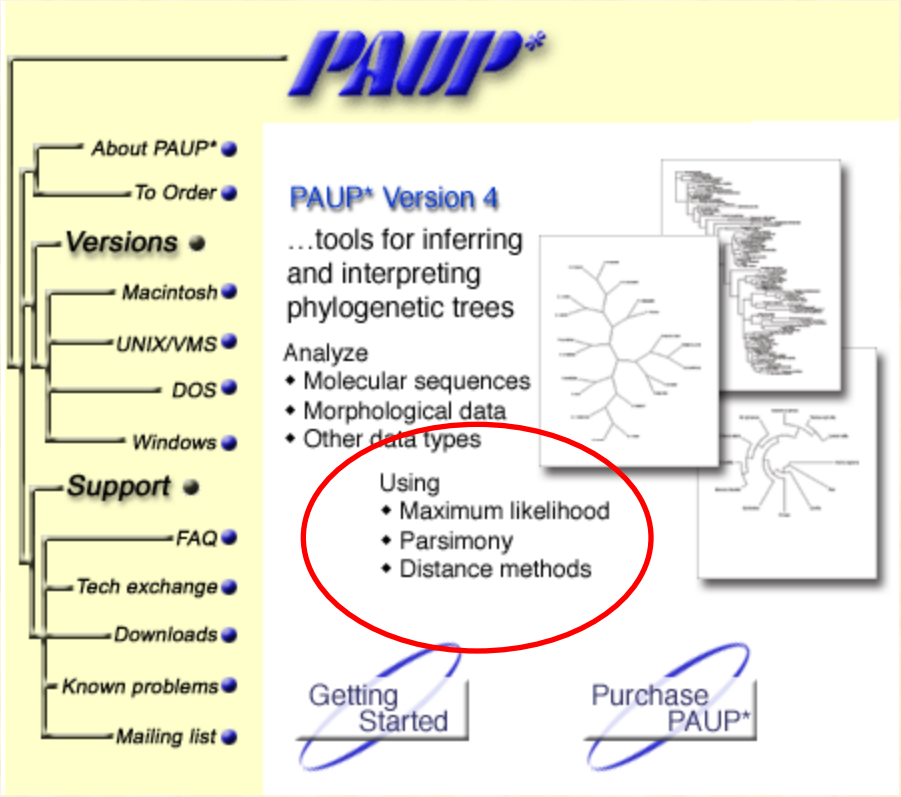
  Vážená parsimonie

  Generalizovaná parsimonie

- **Metoda maximální pravděpodobnosti**

Software pro fylogenetickou analýzu

- **BioNJ** (Neighbor-joining method)
- **PAUP** - Phylogenetic Analysis Using Parsimony



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

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

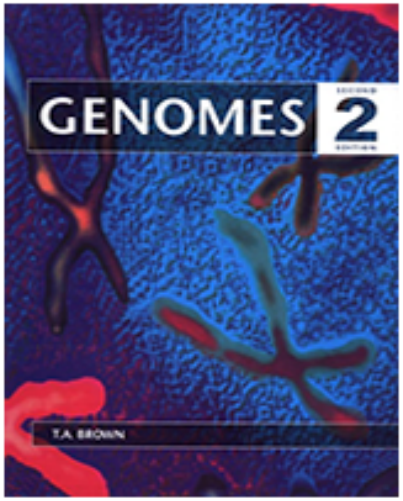
http://macclade.org/index.html

Shrnutí

- Fylogenetika = věda zkoumající fylogenezi, příbuzenské vztahy a vývoj organismů.
- Morfologická data/molekulární data (sekvence).
- Fylogenetické stromy: topologie (příbuznost + evoluce).
- Tvorba stromů: alignment + parsimonie, Neighbor-joining method, …
- BioNJ, PAUP, PHYLIP, PAML

Použitá literatura

Bookshelf ID: NBK21128 PMID: 20821850



Genomes, 2nd edition

Terence A Brown.

Department of Biomolecular Sciences, UMIST, Manchester, UK

Oxford: Wiley-Liss; 2002.
ISBN-10: 0-471-25046-5

Genomes fuses the fresh outlook of the new genomics with the traditional approach to gene expression to provide an up-to-date understanding of the role of the genome as the blueprint for life. This integrated approach focuses on the topics that are central to molecular genetics to create a teaching resource for modern molecular biology.