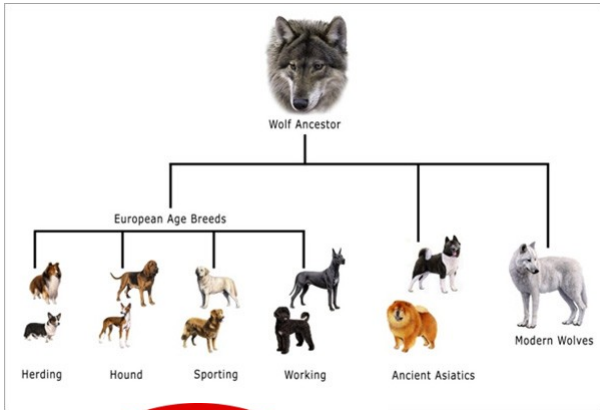
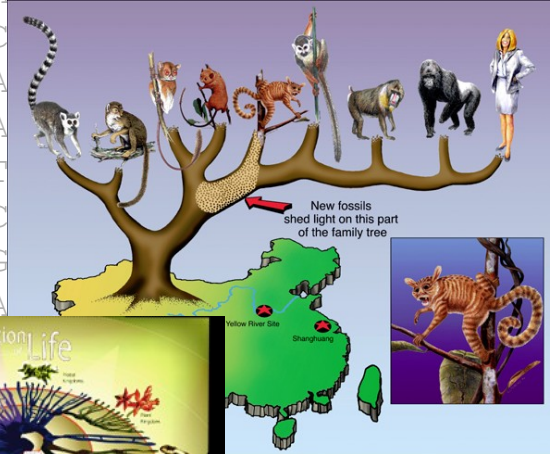


FYLOGENETICKÁ ANALÝZA I.

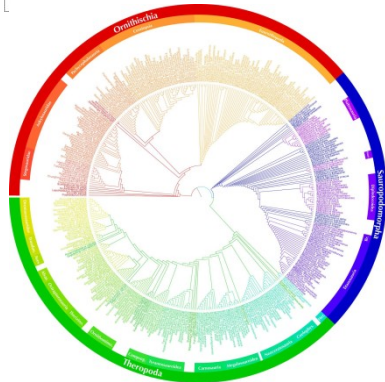
GCCTAGCCACACCCCCACGGGAGACAGCAGTGATAAACCTTTAGCAATAAACGAAAGTTTAACTAAGCCA



TCGTGCTAGCCAC
 ACCCCCCCCCCAA
 AAAGTGGCTTTAA
 TAGCCCTAAACTT
 CAAAGGACCTGGC
 TCACCGCCTCTTG
 AAGTACCACGTA



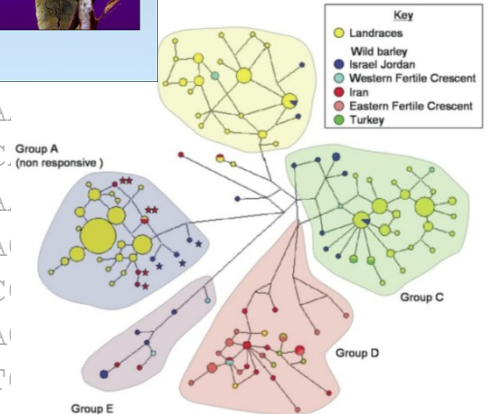
AAATAGAAA
 AAAAAAACT
 ACCCAAAC
 CGCCAGAA
 CCTGTTCT
 GCAAACCC
 ATGAGGCG
 TCGAAGCT



ATACTT
 TGTACT
 ATTTCA
 CTTAAC
 CCGCAA
 AATGAA
 AAGAAC
 ACCTAC



GTAGCTTA
 CTAGCCCC
 GCGATAGA
 ATAATACA
 ACTAAAGC
 GCAAATA
 GATAGAAT



Definice základních pojmů:

fylogenetický strom = fylogenie (*phylogeny*): s kořenem, bez kořene

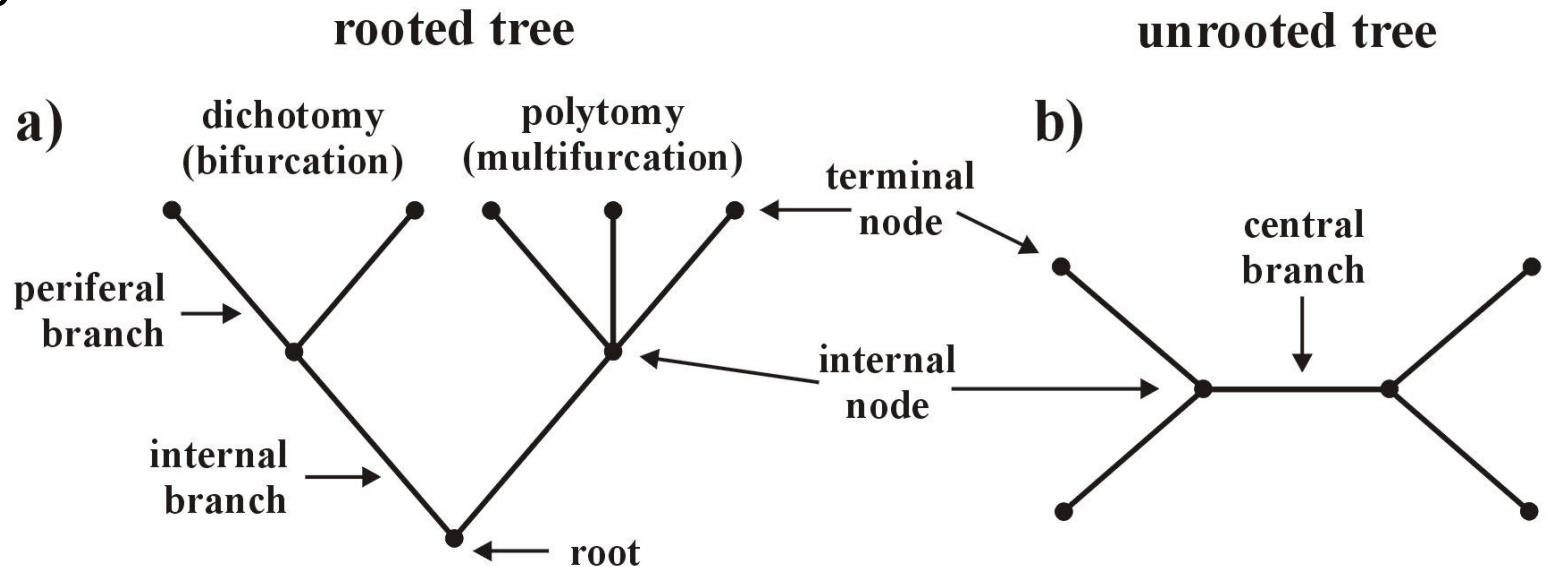
větve (*branches, edges*): vnější, vnitřní, centrální

uzly (*nodes, vertices*): vnitřní, terminální (externí)

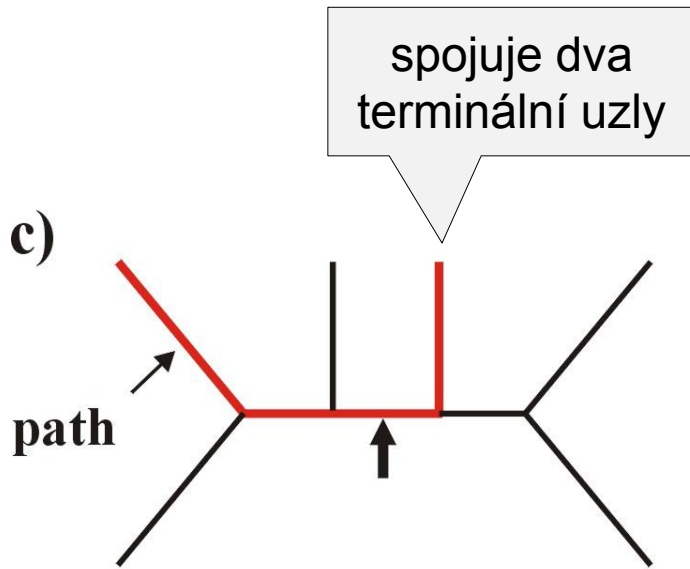
dichotomie, polytomie

OTU, HTU

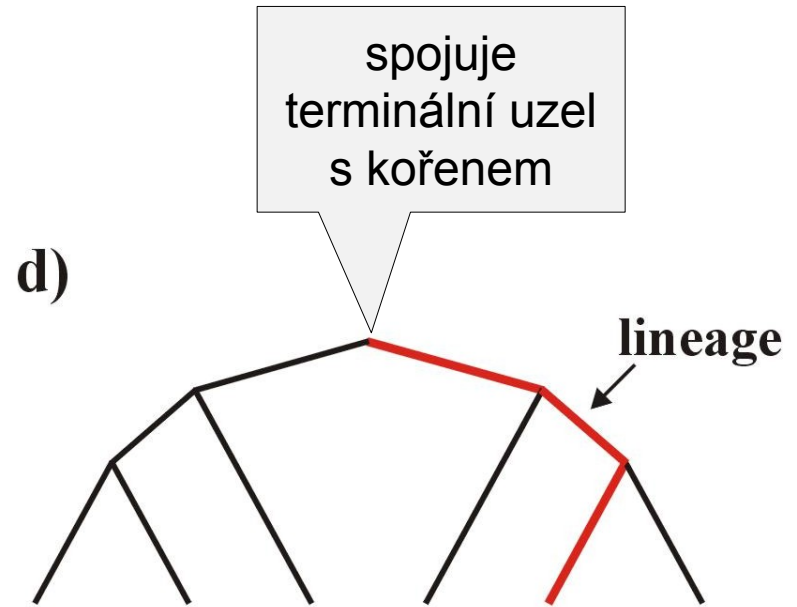
topologie



Definice základních pojmů:

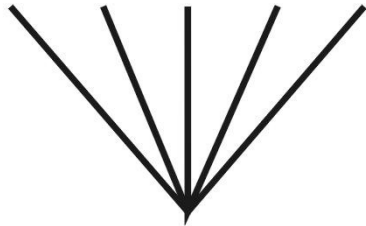


dráha

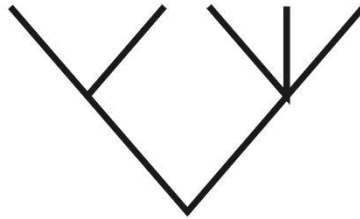


linie

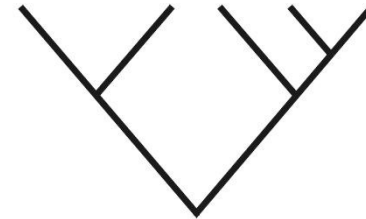
Definice základních pojmů:



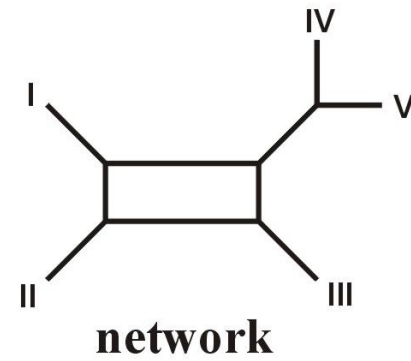
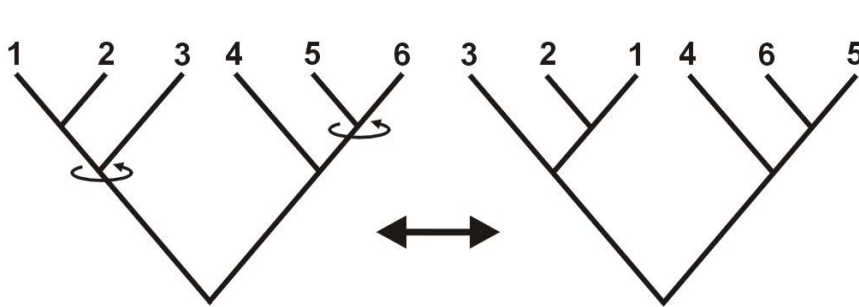
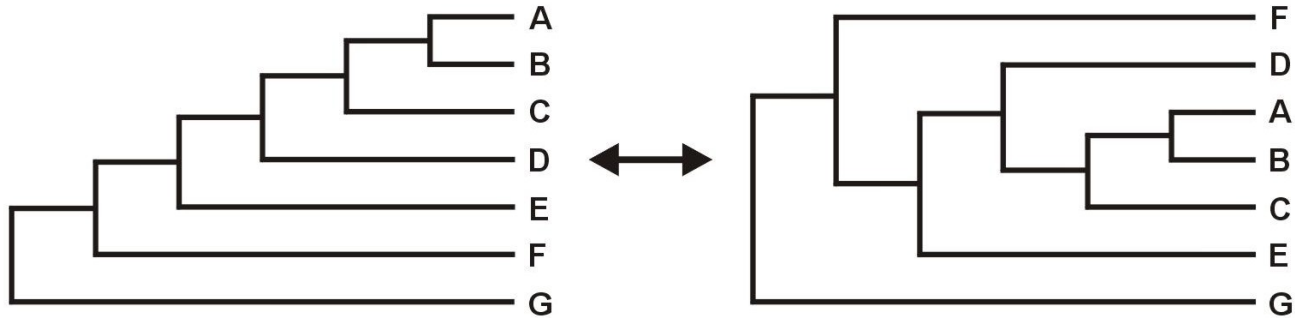
star tree



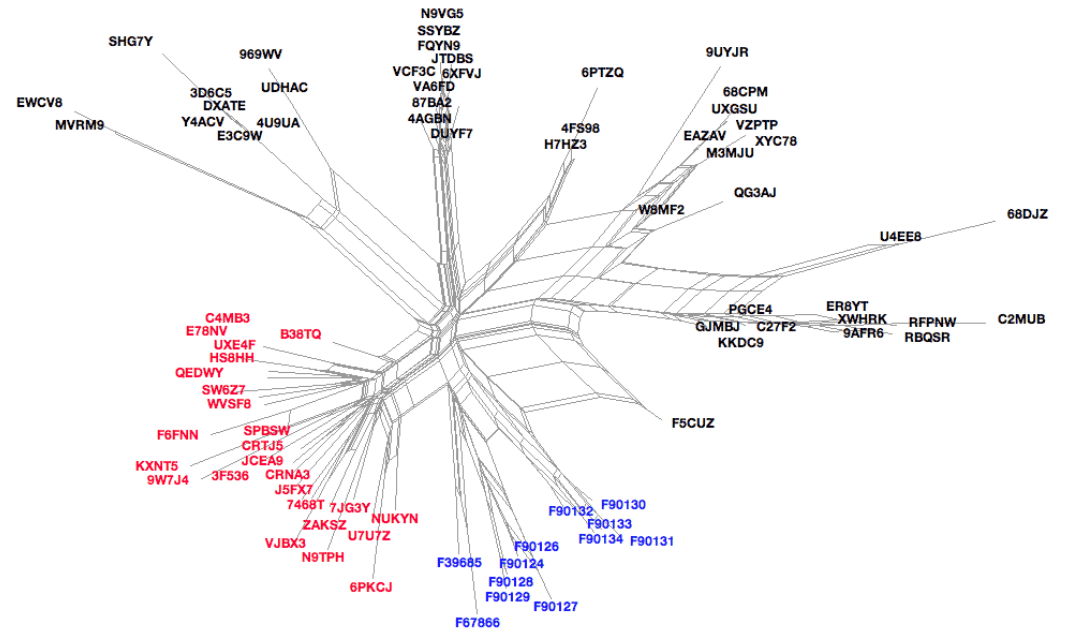
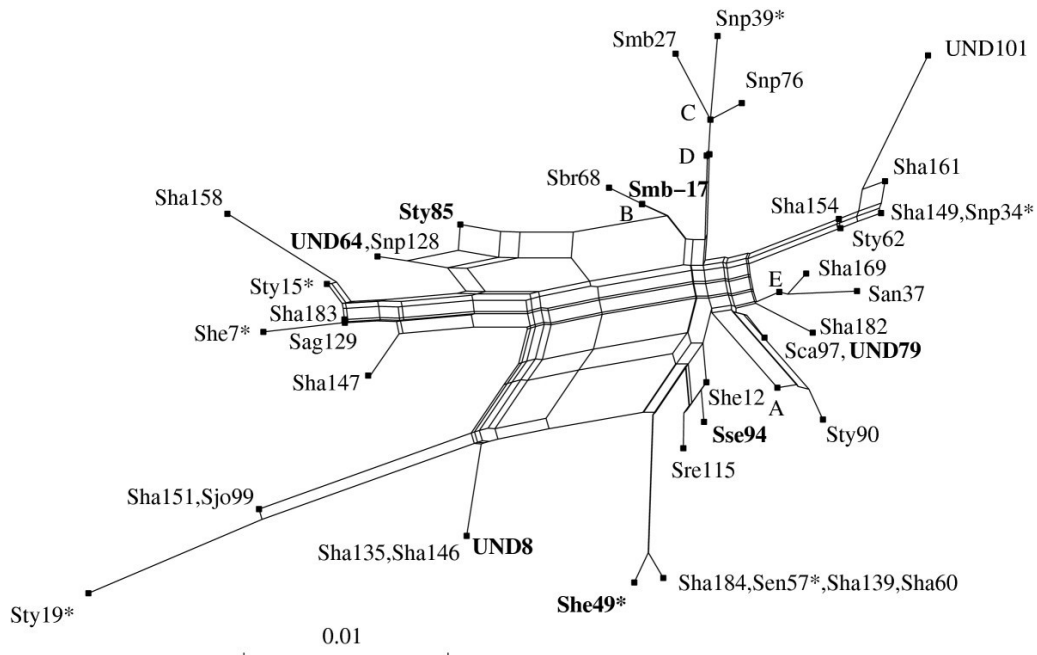
partly resolved



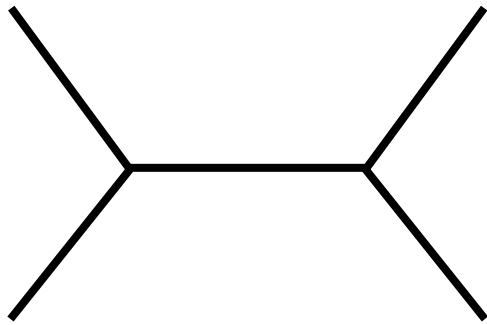
fully resolved



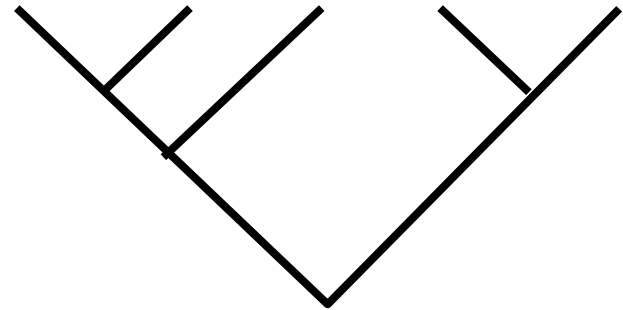
network



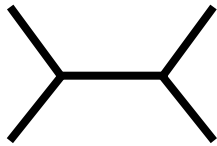
Kolik existuje stromů?



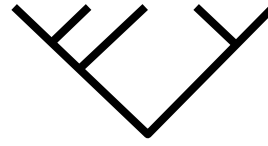
$$\frac{(2n - 1)!}{2^{n-1} (n - 1)!}$$



$$\frac{(2n - 1)!}{2^{n-1} (n - 1)!}$$



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



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

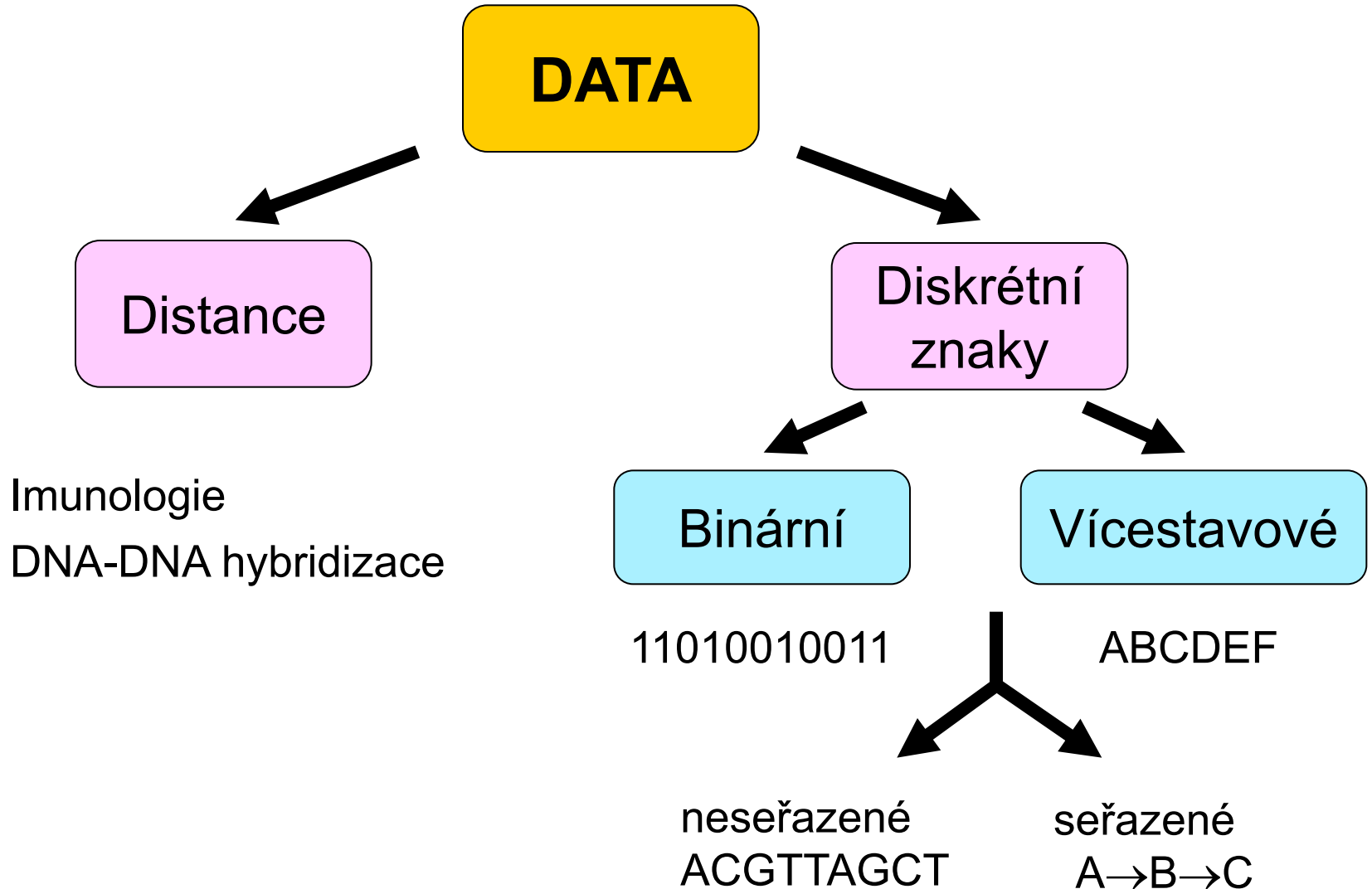
No. Taxons	Unrooted trees	Rooted trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425
11	34 459 425	654 729 075
12	654 729 075	13 749 310 575
13	13 749 310 575	316 234 143 225
14	316 234 143 225	7 905 853 580 625
15	7 905 853 580 625	213 458 046 676 875
20	213 458 046 676 875	8 200 794 532 637 891 559 375
30	8 200 794 532 637 891 559 375	4,9518×10 ³⁸
40	4,9518×10 ³⁸	1,00986×10 ⁵⁷
50	1,00986×10 ⁵⁷	4,78281×10 ⁷⁶

> Avogadrova konstanta*)

8
počet elektronů ve viditelném vesmíru (Eddingtonovo číslo)

*) 6,022 141 79×10²³ mol⁻¹

Jaké typy dat můžeme použít?



Typy dat

Nukleotidové a proteinové sekvence:

H_sapiens MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGS

báze = stav znaku

P_troglod ATGACCCCGACACGCAAAATTAACCCACTAATAAA



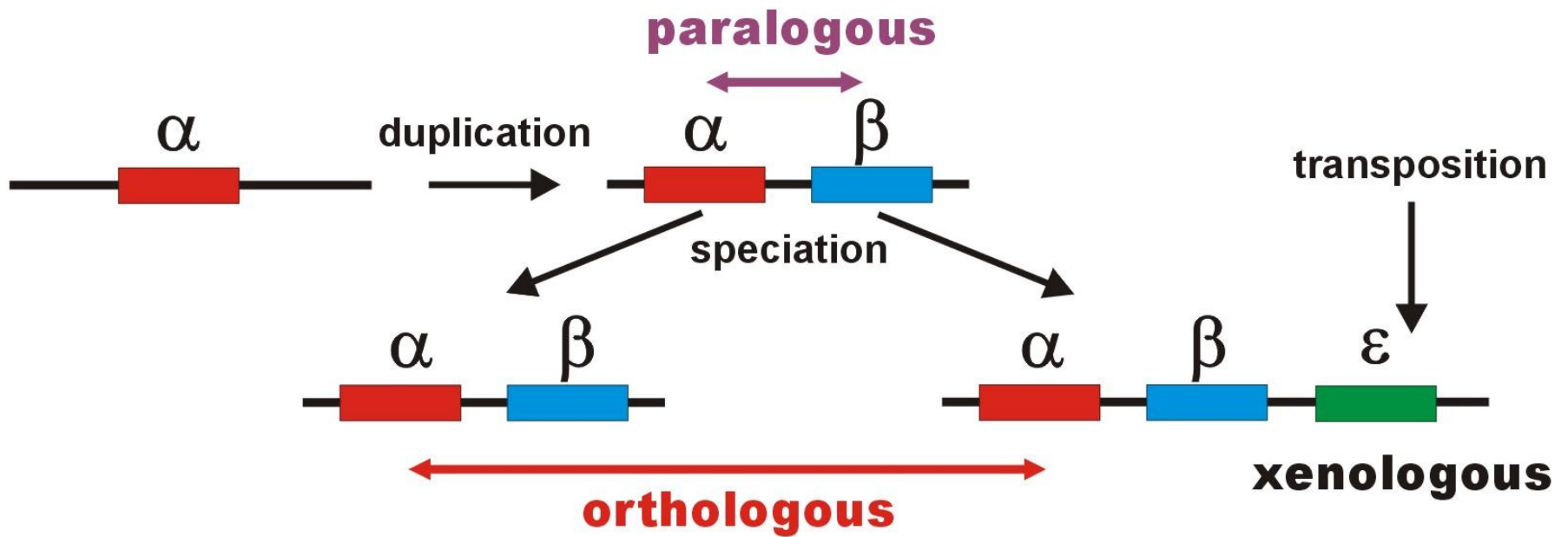
pozice (site) = znak

Typy dat

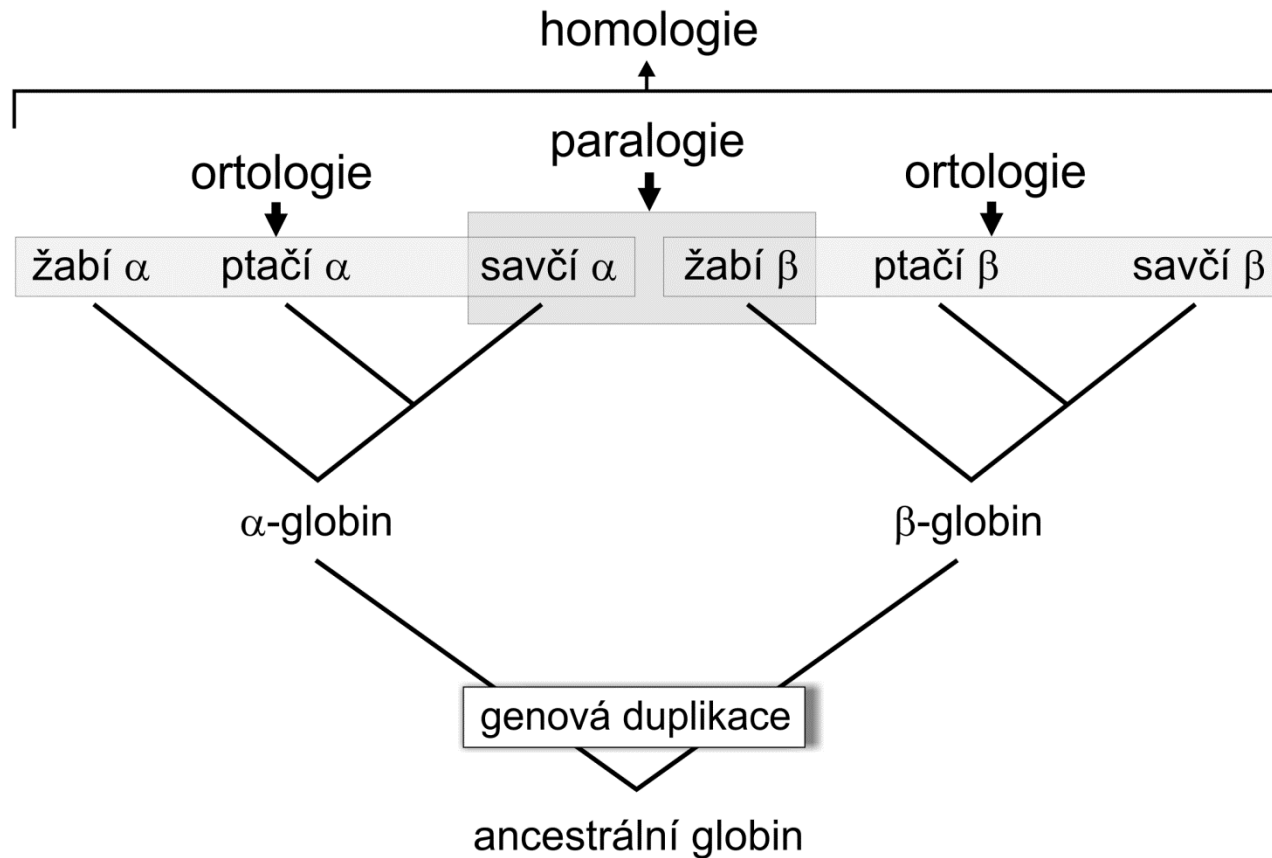
retroelementy: SINE (*Alu*, B1, B2), LINE

mikrosatelity, SNP

Problém homologie sekvencí



Problém homologie sekvencí



Pozor, ani jednotlivá místa v sekvenci DNA nejsou vzájemně zcela nezávislá!

Práce se sekvencemi

DNA databáze:

EMBL (European Molecular Biology Laboratory) – European Bioinformatics Institute, Hinxton, UK: <http://www.ebi.ac.uk/embl/>

GenBank – NCBI (National Center for Biotechnology Information), Bethesda, Maryland, USA: <http://www.ncbi.nlm.nih.gov/Genbank/>

DDBJ (DNA Data Bank of Japan) – National Institute of Genetics, Mishima, Japan: <http://www.ddbj.nig.ac.jp/>

Správa databází: většinou balíky programů Sybase nebo ORACLE

výstupy: ASCII (*American Standard Code for Information Interchange*)

Práce se sekvencemi

Proteinové databáze:

SWISS-PROT – University of Geneva & Swiss Institute of Bioinformatics:

<http://www.expasy.ch/sprot/> a <http://www.ebi.ac.uk/swissprot/>

PIR (Protein Information Resource) – NBRF (National Biomedical Research Foundation, Washington, D.C., USA) & Tokyo University & JIPID (Japanese International Protein Information Database, Tokyo) & MIPS (Martinsried Institute for Protein Sequences, Martinsried, Germany): <http://www-nbrf.georgetown.edu/>

PRF/SEQDB (Protein Resource Foundation) – Ósaka, Japan:

<http://www.prf.or.jp/en/os.htm>

PDB (Protein Data Bank) – University of New Jersey, San Diego & Super-computer Center, University of California & National Institute of Standards and Technology:

<http://www.rcsb.org/pdb/>

Formáty souborů:

FASTA:

>H_sapiens

```
ATGACCCCAATACGCAAATTAACCCCTAATAAAATTAATTAACCACTCATTTCATCGACCTCCCCACCC
CATCCAACATCTCCGCATGATGAAACTTCGGCTCACTCCTTGGCGCCTGCCTGATCCTCAAATCACCAC
AGGACTATTTCCTAGCCATACACTACTCACCAGACGCCTCAACCGCCTTTTCATCAATCGCCACATCACT
CGAGACGTAAATTATGGCTGAATCATCCGCTACCTTCACGCCAATGGCGCCTCAATATTCTTTATCTGCC
TCTTCCTACACATCGGGCGAGGCCTATATTACGGATCATTTCTCTACTCAGAAACCTGAAACATCGGCAT
...
```

>P_troglod

```
ATGACCCCGACACGCAAATTAACCCACTAATAAAATTAATTAATCACTCATTTATCGACCTCCCCACCC
CATCCAACATTTCCGCATGATGGAACTTCGGCTCACTTCTCGGCGCCTGCCTAATCCTTCAAATTACCAC
AGGATTATTTCCTAGCTATACACTACTCACCAGACGCCTCAACCGCCTTCTCGTCGATCGCCACATCACC
CGAGACGTAAACTATGGTTGGATCATCCGCTACCTCCACGCTAACGGCGCCTCAATATTTTTTATCTGCC
TCTTCCTACACATCGGCCGAGGTCTATATTACGGCTCATTTCTCTACCTAGAAACCTGAAACATTGGCAT
...
```

>P_paniscus

```
ATGACCCCAACACGCAAATCAACCCACTAATAAAATTAATTAATCACTCATTTATCGACCTCCCCACCC
CATCCAATATTTCCACATGATGAAACTTCGGCTCACTTCTCGGCGCCTGCCTAATCCTTCAAATCACCAC
AGGACTATTTCCTAGCTATACACTACTCACCAGACGCCTCAACCGCCTTCTCATCGATCGCCACATTACC
CGAGACGTAAACTATGGTTGAATCATCCGCTACCTTCACGCTAACGGCGCCTCAATACTTTTCATCTGCC
TCTTCCTACACGTCCGGTCGAGGCCTATATTACGGCTCATTTCTCTACCTAGAAACCTGAAACATTGGCAT
...
```

Formáty souborů:

GenBank:

ORIGIN

```
1  tgaaatgaag atattctctt ctcaagacat caagaagaag gaactactcc ccaccaccag
61  cacccaaagc tggcattcta attaaactac ttcttgtgta cataaattta catagtacaa
121 tagtacattt atgtatatcg tacattaaac tattttcccc aagcatataa gcaagtacat
181 ttaatcaatg atataggcca taaaacaatt atcaacataa actgatacaa accatgaata
241 ttataactaat acatcaaatt aatgctttaa agacatatct gtgttatctg acatacacca
301 tacagtcata aactcttctc ttccatatga ctatcccctt ccccatthgg tctattaatc
361 taccatcctc cgtgaaacca acaaccgccc caccaatgcc cctcttctcg ctccggggccc
421 attaaacttg ggggtagcta aactgaaact ttatcagaca tctggttctt acttcagggc
481 catcaaattg gttatcgccc atacgttccc cttaaataag acatctcgat ggtatcgggt
541 ctaatcagcc catgaccaac ataactgtgg tgtcatgcat ttgggtattht tttatthtgg
601 cctactthca tcaacatagc cgtcaaggca tgaaaggaca gcacacagtc tagacgcacc
661 tacgggtgaag aatcattagt ccgcaaaacc caatcaccta aggctaatta ttcattgctt
721 ttagacataa atgctactca ataccaaatt ttaactctcc aaacccccca accccctcct
781 cttaatgcca aacccccaaa acactaagaa cttgaaagac atatattatt aactatcaaa
841 ccctatgtcc tgatcgattc tagtagttcc caaaatatga ctcatattht agtacttgta
901 aaaatthtac aaaatcatgc tccgtgaacc aaaactctaa tcacactcta ttacgcaata
961 aatattaaca agttaatgta gcttaataac aaagcaaagc actgaaaatg cttagatgga
1021 taatthtatc cca
```

//

Formáty souborů:

PHYLIP (“interleaved” format):

6 1120

```
H_sapiens      ATGACCCCAA TACGCAAAT TAACCCCTA ATAAAATTAA TTAACCACTC
P_troglod      ATGACCCCGA CACGCAAAT TAACCCACTA ATAAAATTAA TTAATCACTC
P_paniscus     ATGACCCCAA CACGCAAAT CAACCCACTA ATAAAATTAA TTAATCACTC
G_gorilla     ATGACCCCTA TACGCAAAC TAACCCACTA GCAAACCTAA TTAACCACTC
P_pygmaeus    ATGACCCCAA TACGCAAAC CAACCCACTA ATAAAATTAA TTAACCACTC
H_lar         ATGACCCCCC TGCGCAAAC TAACCCACTA ATAAAACCTAA TCAACCACTC

ATTCATCGAC    CTCCCCACCC CATCCAACAT CTCCGCATGA TGAAACTTCG
ATTTATCGAC    CTCCCCACCC CATCCAACAT TTCCGCATGA TGGAACTTCG
ATTTATCGAC    CTCCCCACCC CATCCAATAT TTCCACATGA TGAAACTTCG
ATTCATTGAC    CTCCCTACCC CGTCCAACAT CTCCACATGA TGAAACTTCG
ACTCATCGAC    CTCCCCACCC CATCAAACAT CTCTGCATGA TGGAACTTCG
ACTTATCGAC    CTTCCAGCCC CATCCAACAT TTCTATATGA TGAAACTTTG
```

Formáty souborů:

NEXUS (PAUP*, “interleaved”):

```
#NEXUS
begin data;
dimensions ntax=6 nchar=1120;
format datatype=DNA interleave datatype=DNA missing=? gap=-;
matrix
P_troglod   ATGACCCCGACACGCAAATTAACCCACTAATAAAATTAATTAATCACTC
P_paniscus  ATGACCCCAACACGCAAATCAACCCACTAATAAAATTAATTAATCACTC
H_sapiens   ATGACCCCAATACGCAAATTAACCCCTAATAAAATTAATTAACCACTC
G_gorilla   ATGACCCCTATACGCAAACCTAACCCACTAGCAAACCTAATTAACCACTC
P_pygmaeus  ATGACCCCAATACGCAAACCAACCCACTAATAAAATTAATTAACCACTC
H_lar       ATGACCCCCCTGCGCAAACCTAACCCACTAATAAACTAATCAACCACTC

P_troglod   ATTTATCGACCTCCCCACCCCATCCAACATTTCCGCATGATGGAACTTCG
P_paniscus  ATTTATCGACCTCCCCACCCCATCCAATATTTCCACATGATGAAACTTCG
H_sapiens   ATTCATCGACCTCCCCACCCCATCCAACATCTCCGCATGATGAAACTTCG
G_gorilla   ATTCATTGACCTCCCTACCCCGTCCAACATCTCCACATGATGAAACTTCG
P_pygmaeus  ACTCATCGACCTCCCCACCCCATCAAACATCTCTGCATGATGGAACTTCG
H_lar       ACTTATCGACCTTCCAGCCCATCCAACATTTCTATATGATGAAACTTTG

end;
```

Formáty souborů:

Clustal X:

```
P_troglod ATGACCCCGACACGCAAAATTAACCCACTAATAAAAATTAATTAATCACTCATTATCGAC
P_paniscus ATGACCCCAACACGCAAAATCAACCCACTAATAAAAATTAATTAATCACTCATTATCGAC
H_sapiens ATGACCCCAATACGCAAAATTAACCCCTAATAAAAATTAATTAACCACTCATTATCGAC
G_gorilla ATGACCCCTATACGCAAAACTAACCCACTAGCAAAACTAATTAACCACTCATTATCGAC
P_pygmaeus ATGACCCCAATACGCAAAACCAACCCACTAATAAAAATTAATTAACCACTCACTCATCGAC
H_lar ATGACCCCCCTGCGCAAAACTAACCCACTAATAAAAATAATCAACCACTCACTTATCGAC
*****          *****          *****  ***          *****  *****  ** *****  * **  ***
```

```
P_troglod CTCCCACCCCATCCAACATTTCCGCATGATGAACTTCGGCTCACTTCTCGGCGCCTGC
P_paniscus CTCCCACCCCATCCAATATTTCCACATGATGAACTTCGGCTCACTTCTCGGCGCCTGC
H_sapiens CTCCCACCCCATCCAACATCTCCGCATGATGAACTTCGGCTCACTCCTTGGCGCCTGC
G_gorilla CTCCCTACCCCGTCCAACATCTCCACATGATGAACTTCGGCTCACTCCTTGGTGCCTGC
P_pygmaeus CTCCCACCCCATCAAACATCTCTGCATGATGAACTTCGGCTCACTTCTAGGCGCCTGC
H_lar CTTCAGCCCCATCCAACATTTCTATATGATGAACTTTGGTTCCTAGGCGCCTGC
** **  ****  ** ** ** **          *****  *****  ** *****  ** **  *****
```

BLAST

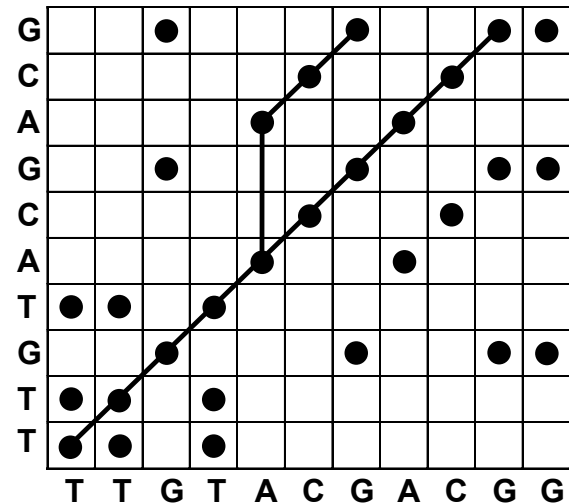
GenBank

ClustalX

Seřazení sekvencí (alignment):

Sekvence 1 **TTGTACGACGG**
 Sekvence 2 **TTGTACGACG**

TTGTACGACGG **TTGT---ACGACGG**
 | | | | | | | | | |
TTGTACGACG **TTGTACGACG**

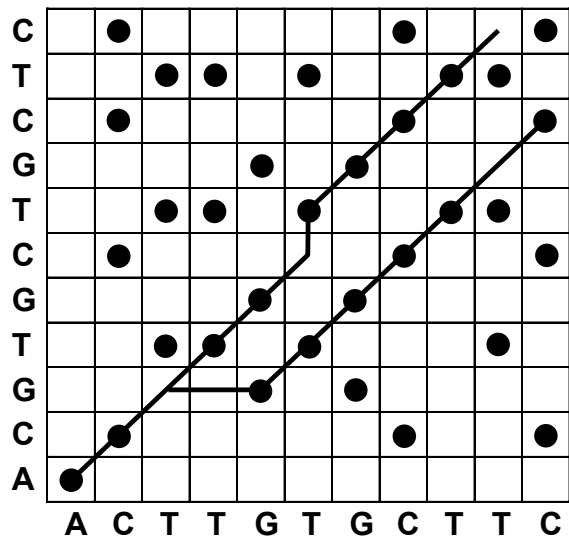


Sekvence 1 **ACTTGCTTC**
 Sekvence 2 **ACGTGCTGC**

ACTTGCTTC **ACTG-TGCTTC**
 | | | | | | | |
ACGTGCTGC

Path 1

ACTTGCTTC
 | | | | | | | |
AC--GTGCTGC



Seřazení sekvencí (alignment):

Penalizace mezer (*gap penalty*):

g = penalizace za výskyt mezery ($1 \times$)

h = extenze za každou „pomlčku“

l = délka mezery (= počet „pomlček“)

Př.: GC□ □ □ □ TTAA

$l = 5, g = x, h = 5x$

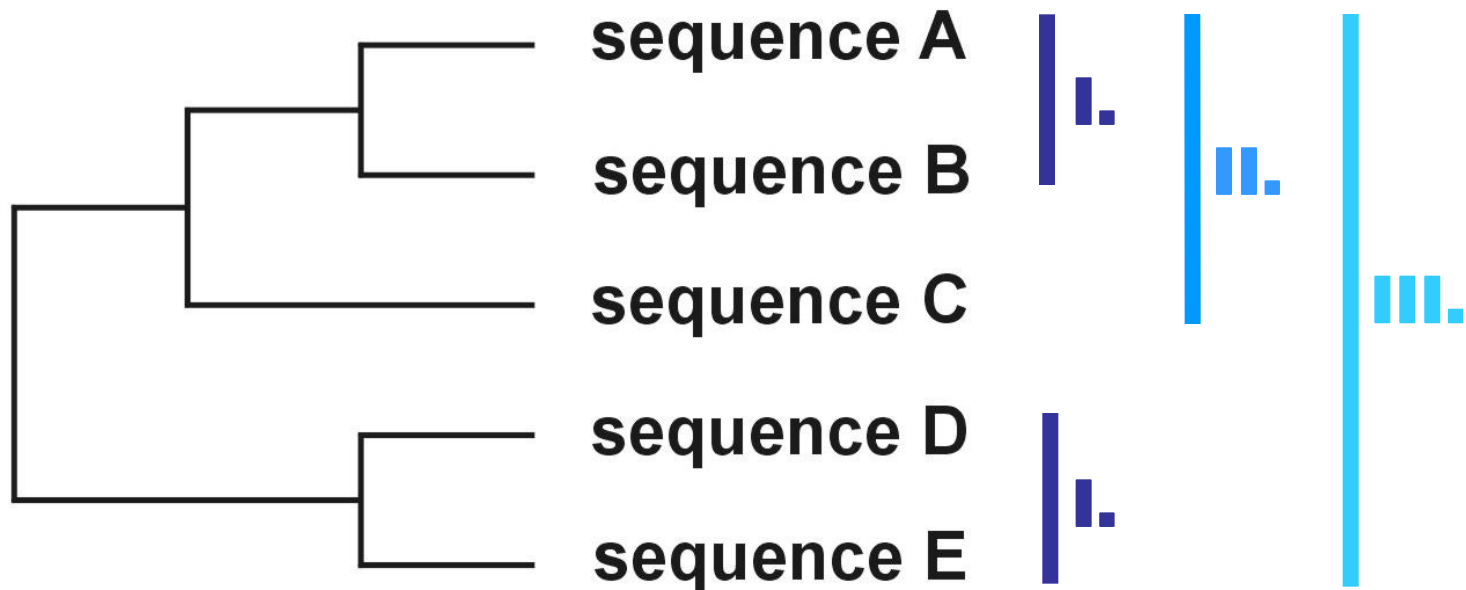
$$GP = g + hl$$

g - gap penalty
 h - gap extension
penalty
 l - gap length

Progresivní seřazení - ClustalX

3 fáze:

1. Seřazení dvojic sekvencí → párové distance
2. Konstrukce „guide tree“ (NJ)
3. Seřazení všech sekvencí podle stromu



Problém progresivního seřazení

6 druhů:

gorila
kůň
panda

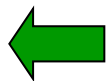
AGGTT
AG-TT
AG-TT

tučňák
kuře
pštros

A-GTT
A-GTT
AGGTT



AGGTT
AG-TT
AG-TT
AG-TT
AG-TT
AGGTT



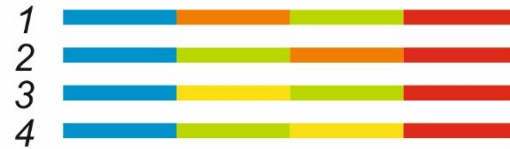
gorila
kůň
panda
tučňák
kuře
pštros

AGGTT
AG-TT
AG-TT
A-GTT
A-GTT
AGGTT

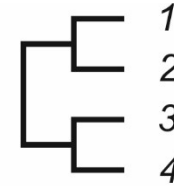


AGGTT
A-GTT
A-GTT
A-GTT
A-GTT
AGGTT

Existují i metody bez seřazení:



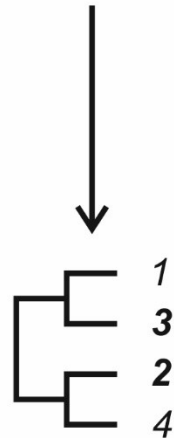
homologní sekvence



referenční strom



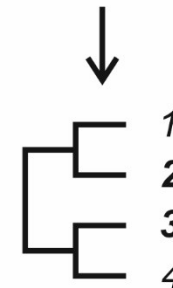
seřazení všech sekvencí



fylogenetický strom



metoda bez seřazení sekvencí



Rozdělení metod

Typy dat

distance

znaky

Metody konstrukce stromů

algorithms

kritérium
optimality

UPGMA neighbor- joining	
Fitch- Margoliash minimum evolution	maximum parsimony maximum likelihood Bayesian a.

Jak hodnotit jednotlivé metody?

výkonnost (efficiency):

jak rychlá je metoda?

síla (power):

kolik znaků je třeba?

konzistence (consistency):

vede zvyšující se počet znaků ke správnému stromu?

robustnost (robustness):

jak metoda funguje při neplatnosti předpokladů?

falzifikovatelnost (falsifiability):

umožňuje testování platnosti předpokladů?

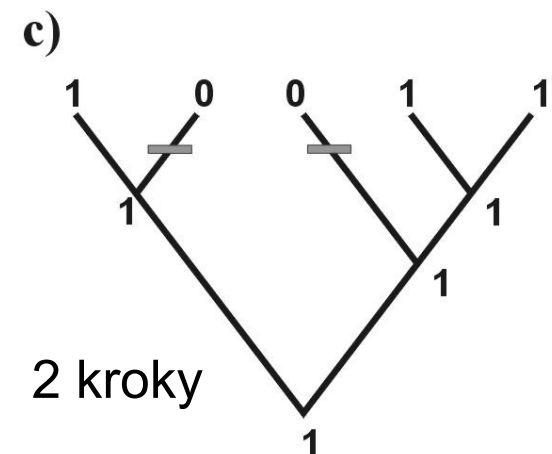
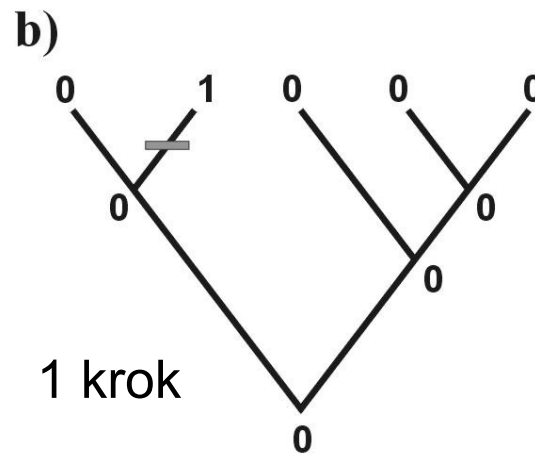
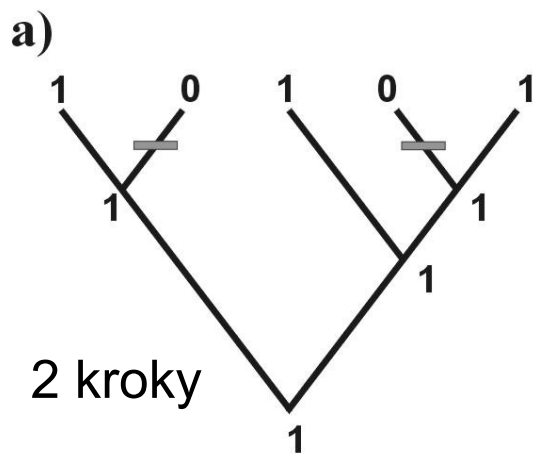
MAXIMÁLNÍ ÚSPORNOST (*maximum parsimony, MP*)

William of Ockham (c. 1287 – 1347)
Occamova břitva



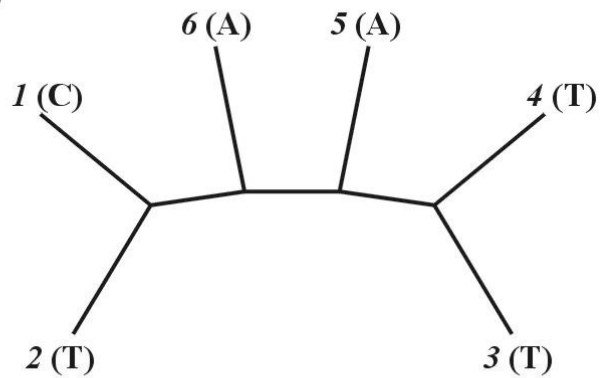
	I	II	III
A	1	0	1
B	0	0	1
C	1	0	0
D	0	1	0
E	1	0	1

minimální počet kroků = 3
skutečný počet kroků = 5
⇒ 2 extra kroky → homoplasie



Odhad počtu kroků: Fitchův algoritmus

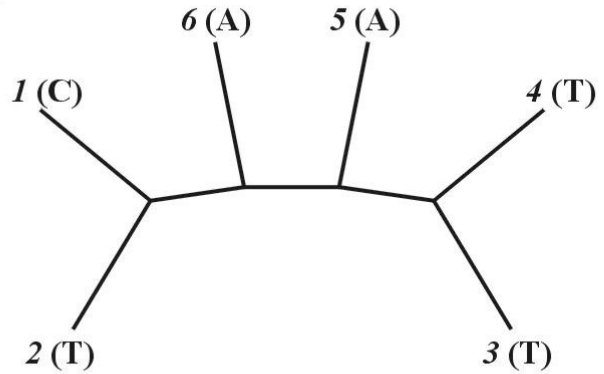
a)



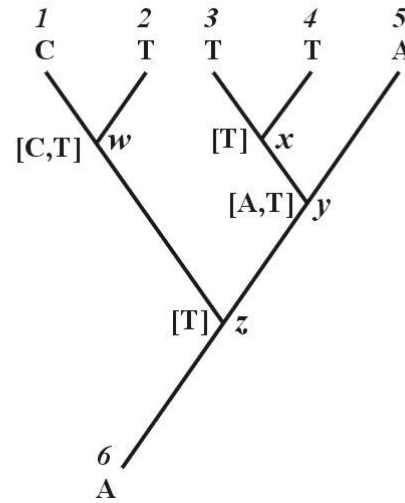
1. arbitrární kořen

Odhad počtu kroků: Fitchův algoritmus

a)



b)

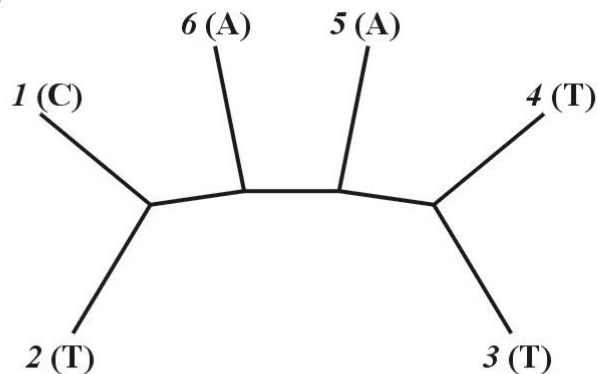


1. arbitrární kořen

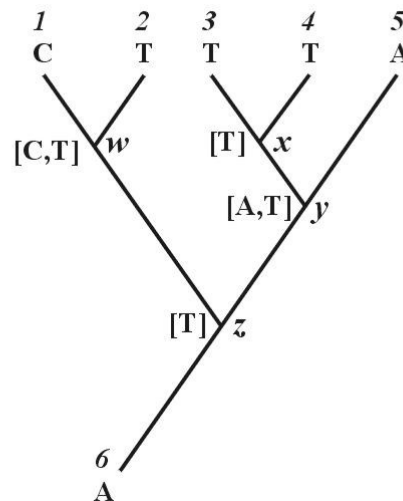
2. Shora dolů:
 $w = C$, nebo T
 $x = T$
 $y = A$, nebo T
 $z = T$

Odhad počtu kroků: Fitchův algoritmus

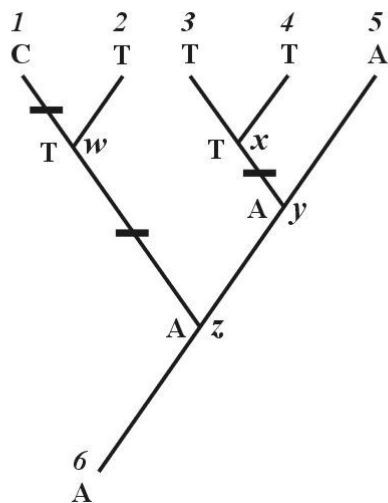
a)



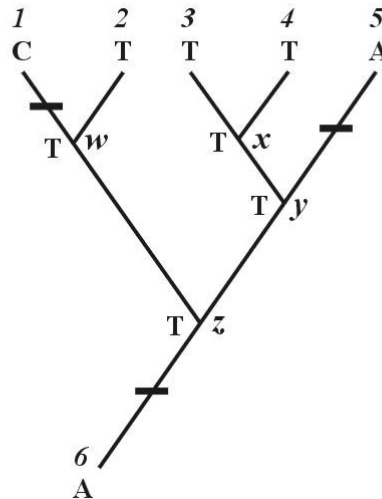
b)



c)



d)



1. arbitrární kořen

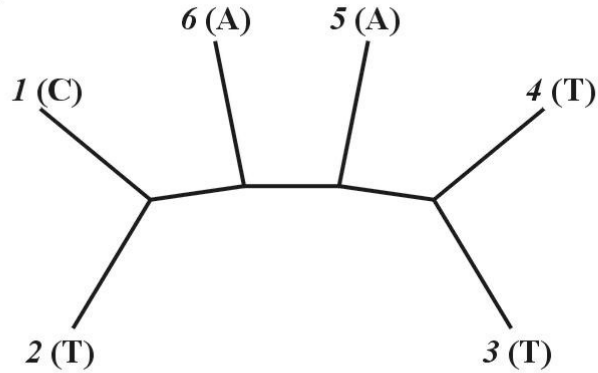
2. Shora dolů:
 $w = C$, nebo T
 $x = T$
 $y = A$, nebo T
 $z = T$

3. Zdola nahoru:
 $z = T$, nebo A

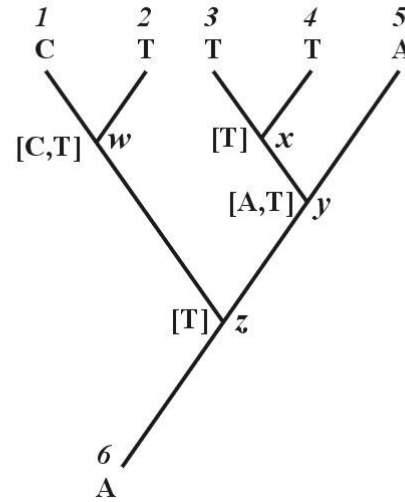
celková délka = 3

Odhad počtu kroků: Fitchův algoritmus

a)



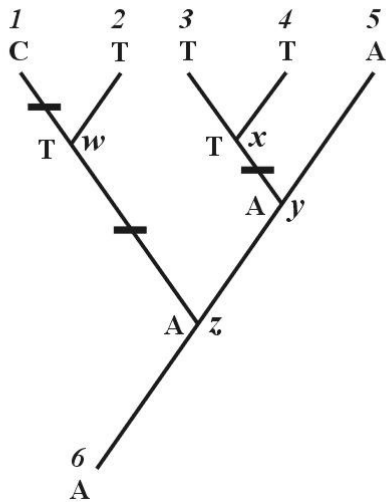
b)



1. arbitrární kořen

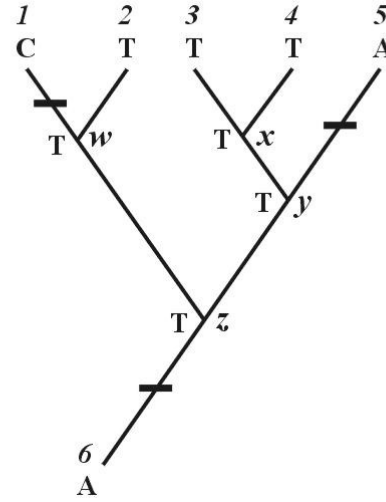
2. Shora dolů:
 $w = C$, nebo T
 $x = T$
 $y = A$, nebo T
 $z = T$

c)



DELTRAN
(DELAyed TRANsformation)

d)



ACCTRAN
(ACCelerated TRANsformation)

3. Zdola nahoru:
 $z = T$, nebo A

celková délka = 3

Problém homoplasie:

parsimony-informative and non-informative characters (sites)

- invariant sites (symplesiomorphies)
- singletons (autapomorphies)

index konzistence (consistency i., CI)

retenční index (retention i., RI)

upravený CI (rescaled CI, RC)

index homoplasie (homoplasy i., HI)

$$CI = \frac{\sum_i m_i}{\sum_i s_i} \quad RI = \frac{\sum_i g_i - \sum_i s_i}{\sum_i g_i - \sum_i m_i}$$

$$RC = CI \times RI$$

$$HI = 1 - CI$$

m = min. no. of possible steps

s = min. no. needed for explaining the tree

g = max. no. of steps for any tree

Metody parsimonie:

Fitchova: $X \rightarrow Y$ a $Y \rightarrow X$
neseřazené znaky ($A \rightarrow T$ nebo $A \rightarrow G$ etc.)

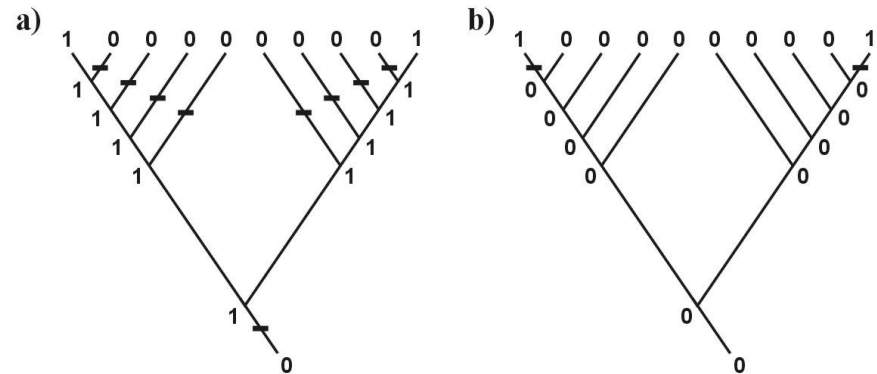
Wagnerova: $X \rightarrow Y$ a $Y \rightarrow X$
seřazené znaky ($1 \rightarrow 2 \rightarrow 3$)

Dollova: $X \rightarrow Y$ a $Y \rightarrow X$, potom nelze $X \rightarrow Y$

... *restriction-site and restriction-fragment data*

Caminova-Sokalova: $X \rightarrow Y$,
ne $Y \rightarrow X$

... SINE, LINE



vážená (weighed, transversion):

“relaxed Dollo criterion”

generalizovaná: matice nákladů (cost matrix) = kroková matice (step matrix)

Wagnerova

a)

	a	b	c	d
a	-	1	2	3
b	1	-	1	2
c	2	1	-	1
d	3	2	1	-

Fitchova

b)

	a	b	c	d
a	-	1	1	1
b	1	-	1	1
c	1	1	-	1
d	1	1	1	-

c)

	a	b	c	d
a	-	M^*)	$2M$	$3M$
b	1	-	M	$2M$
c	2	1	-	M
d	3	2	1	-

d)

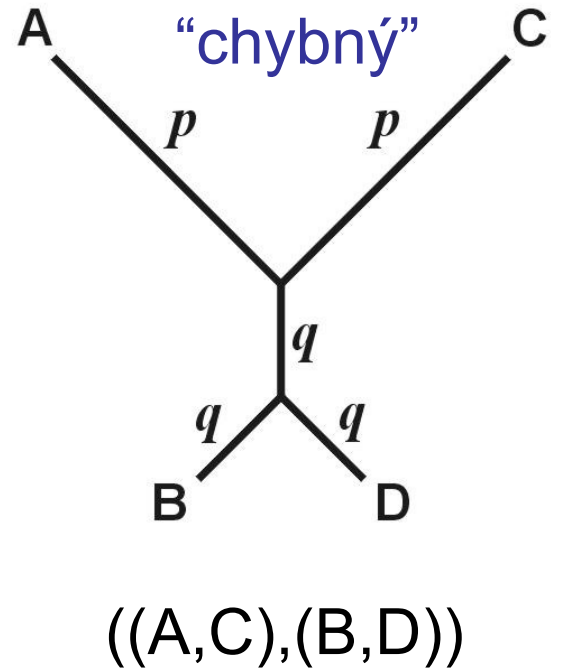
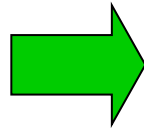
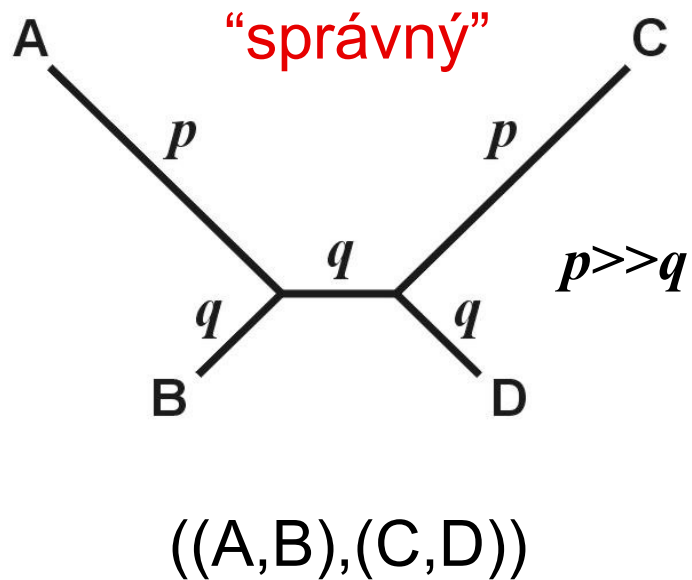
	A	C	G	T
A	-	5	1	5
C	5	-	5	1
G	1	5	-	5
T	5	1	5	-

Dollova

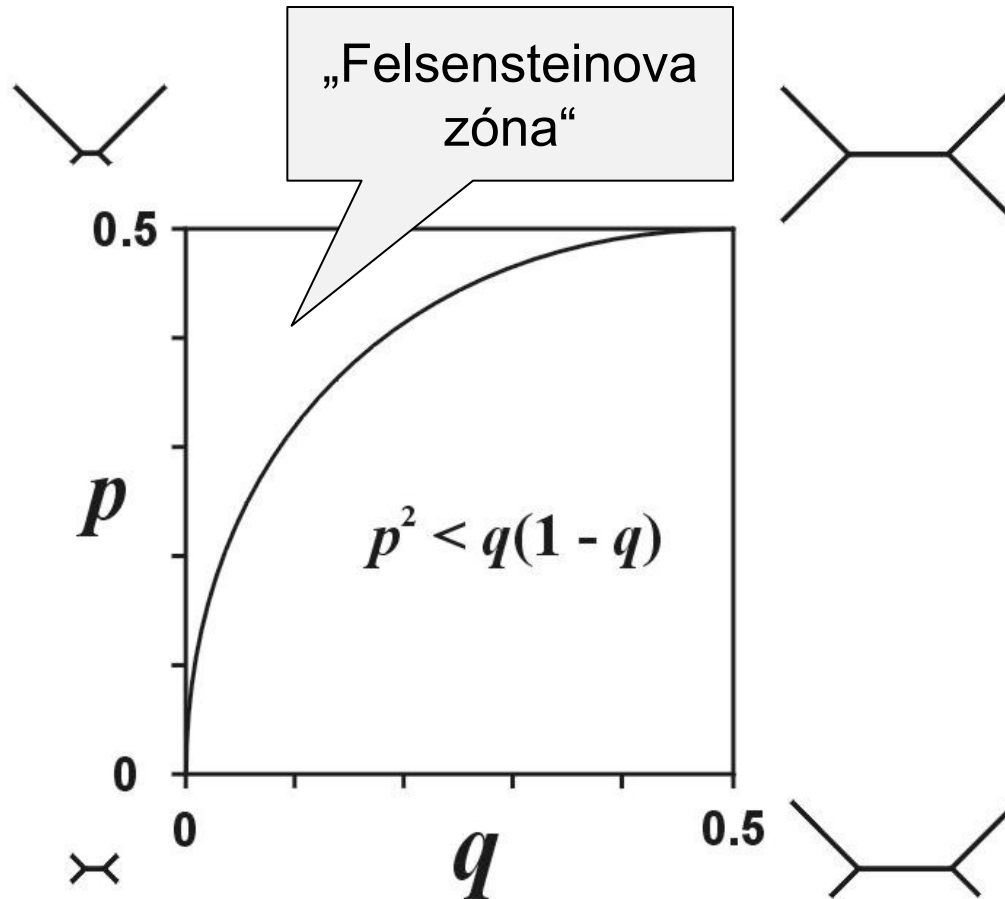
vážená
(transverzní)

*) M je libovolně velké číslo zaručující, že bude povolena pouze jedna transformace do každého odvozeného stavu.

Parsimonie a konzistence

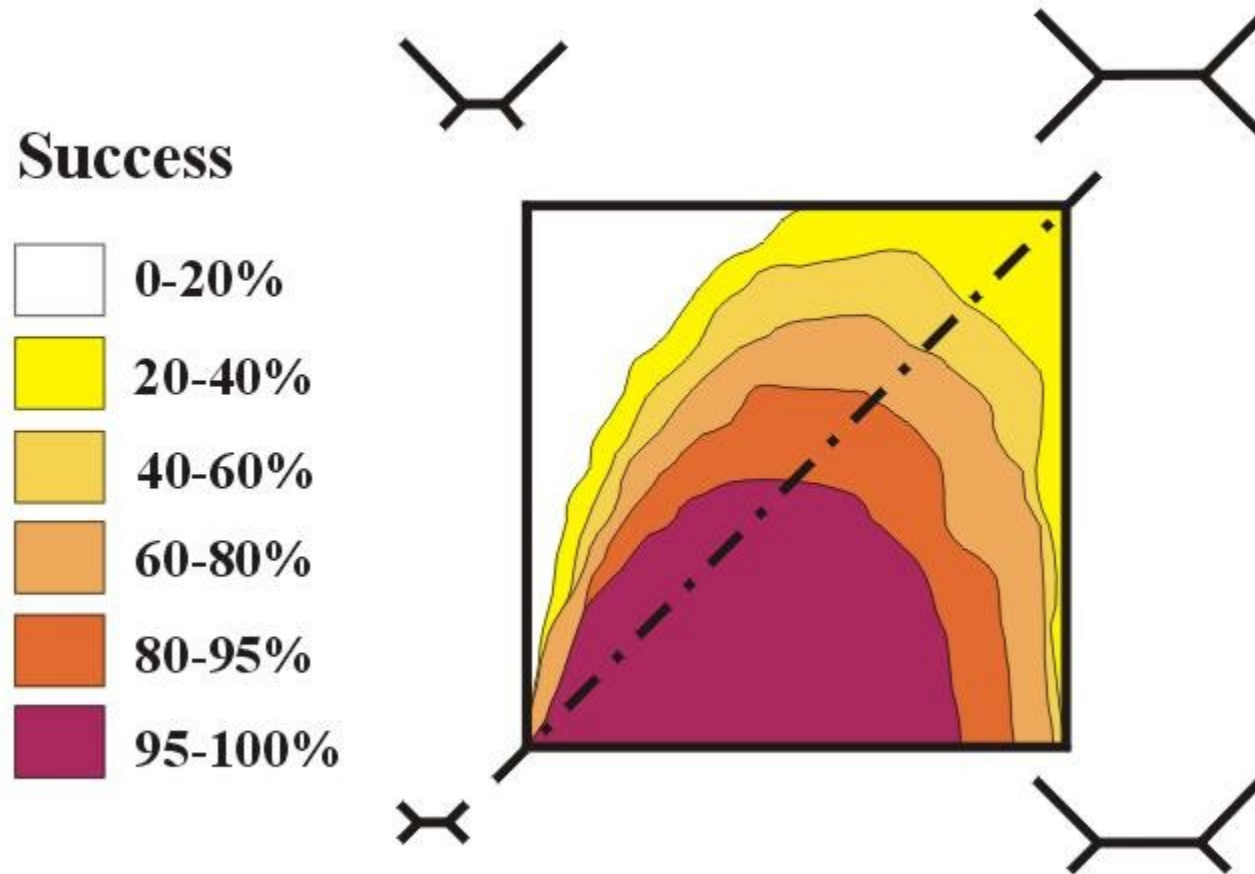


Parsimonie a konzistence

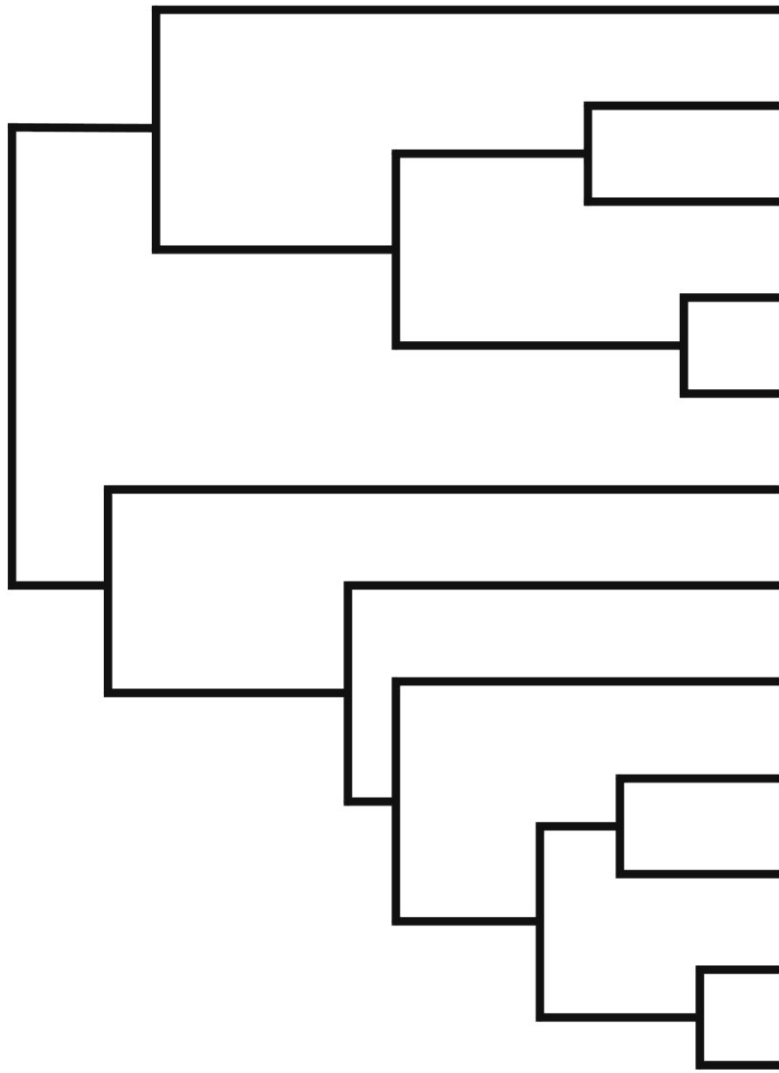


Ve Felsensteinově zóně je parsimonie nekonzistentní

Parsimonie a konzistence

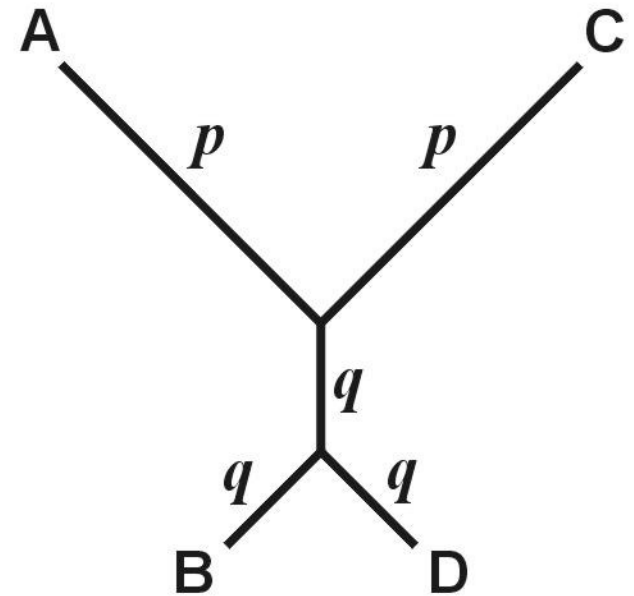
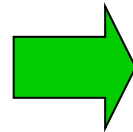


Parsimonie a konzistence



dlouhé větve

„přitažlivost dlouhých větví“
(*long-branch attraction, LBA*)



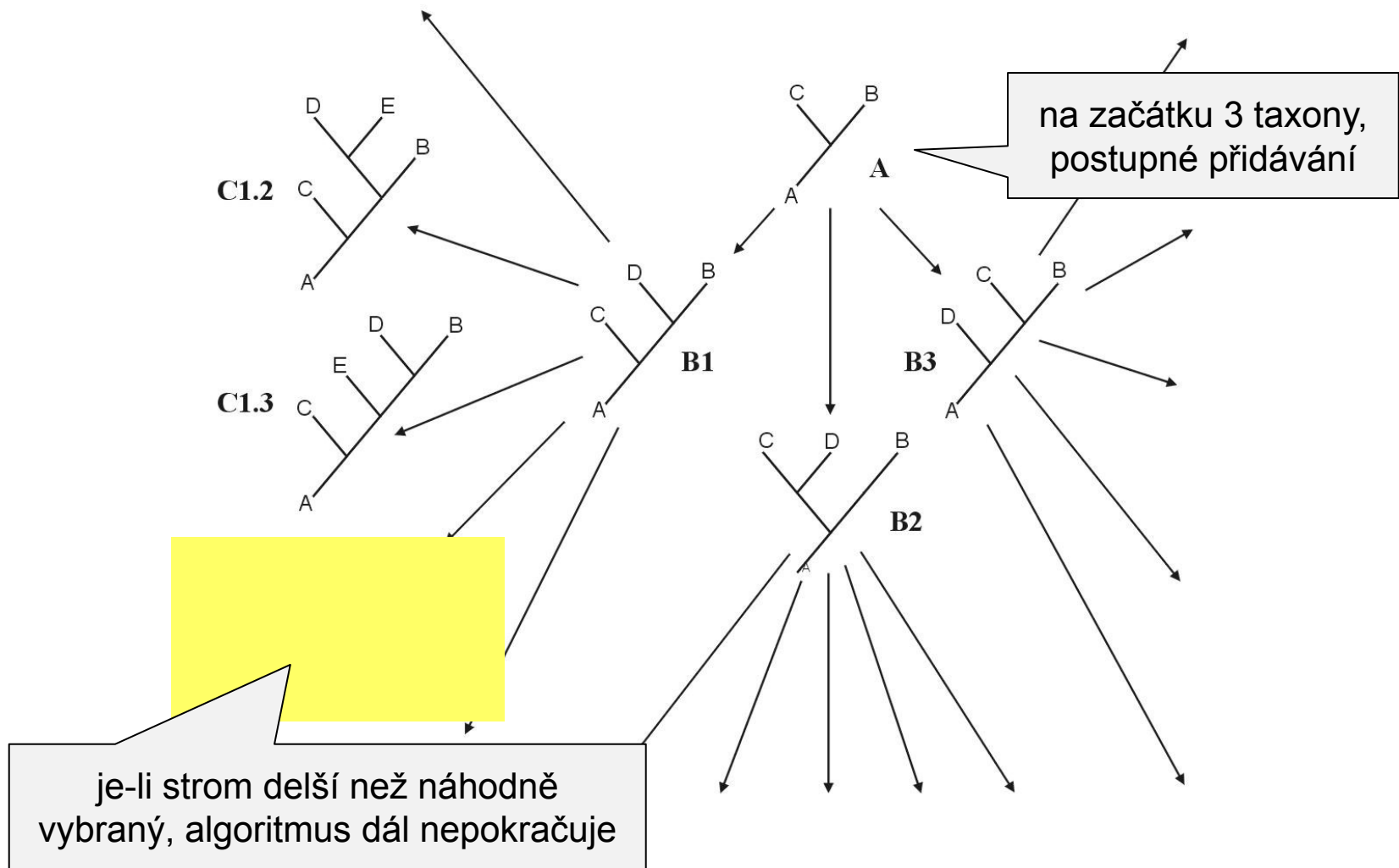
Hledání optimálního stromu

1. Exaktní metody:

a) vyčerpávající hledání (*exhaustive search*)

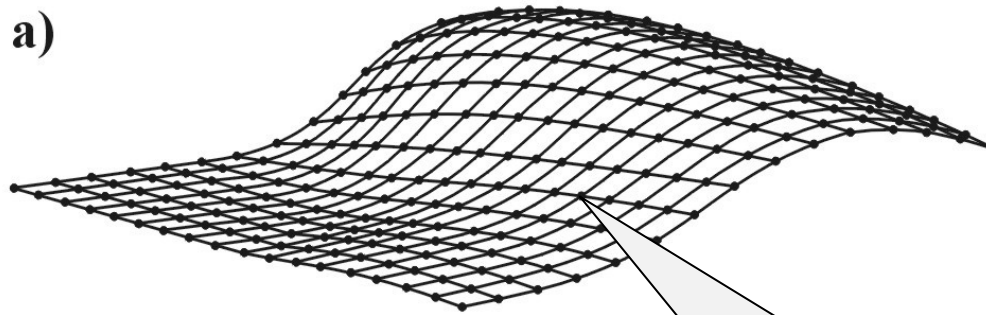
b) *branch-and-bound*

branch-and-bound



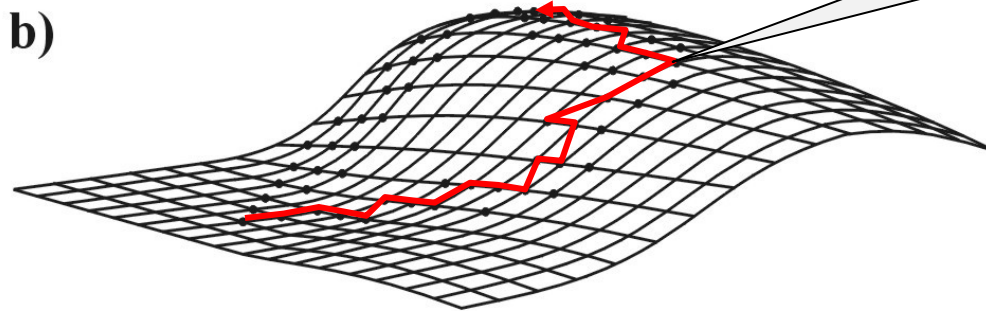
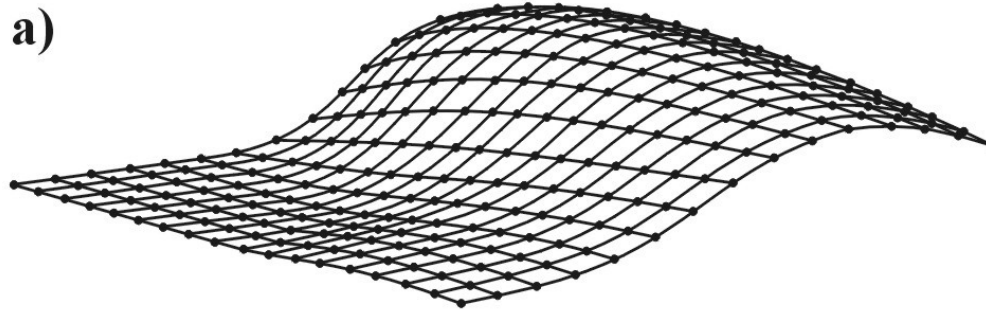
2. Heuristické hledání

a)



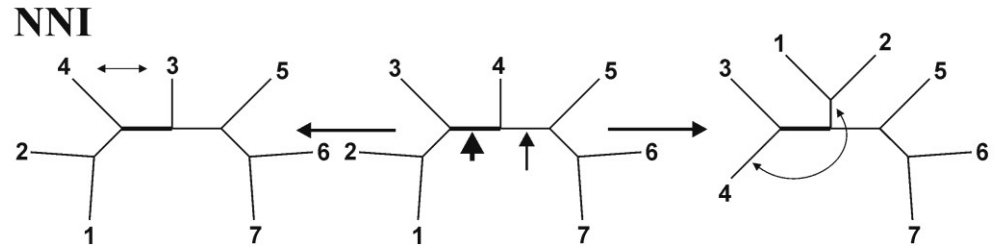
všechny možné stromy

stepwise addition
star decomposition
branch swapping

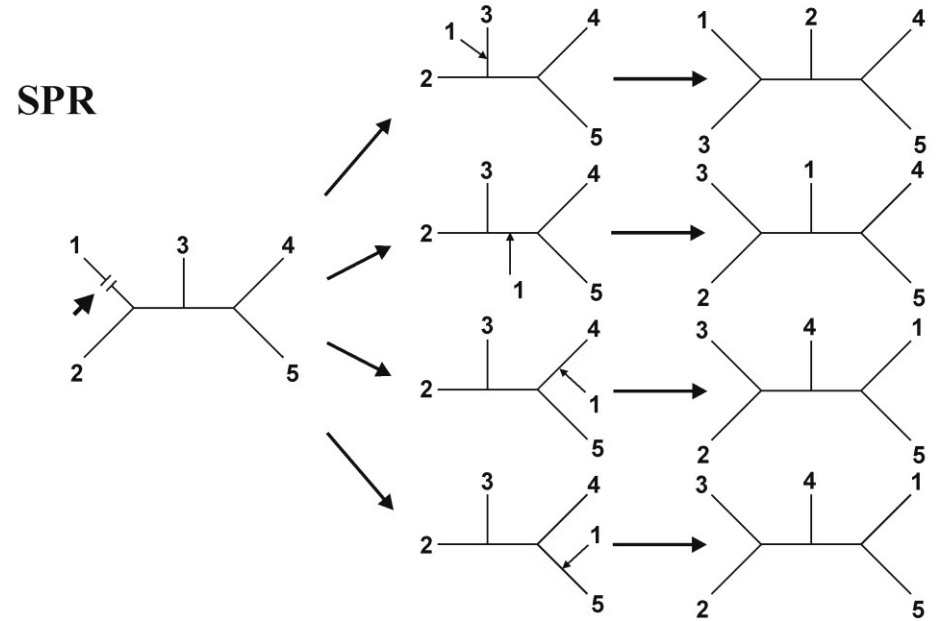


heuristické hledání

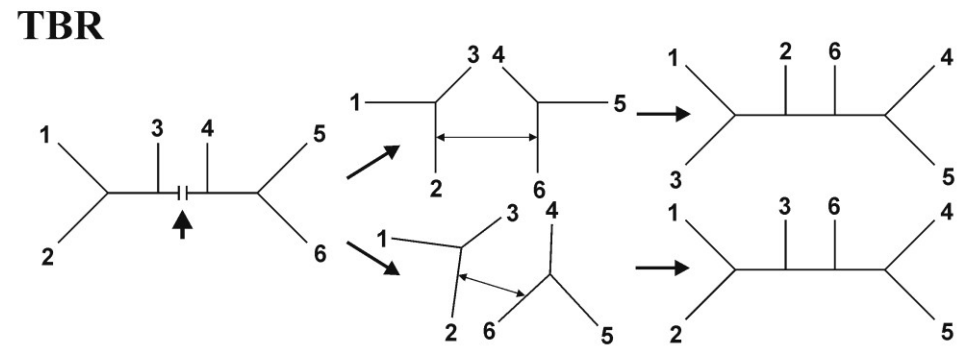
*nearest-neighbor
interchanges (NNI)*



*subtree pruning
and regrafting (SPR)*



*tree bisection and
reconnection (TBR)*



Evoluční modely a distanční metody

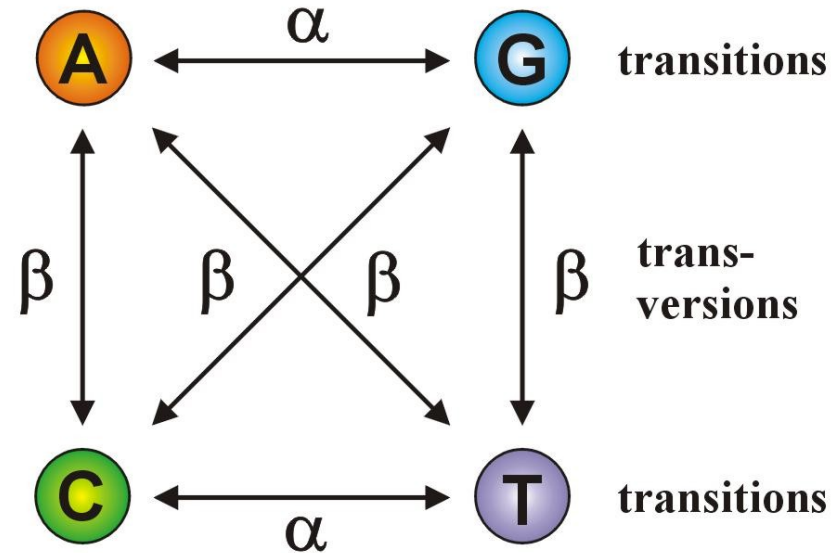
		Báze po substituci			
		A	C	G	T
Původní báze	A	$-\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	C	$\frac{1}{4}$	$-\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	G	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{3}{4}$	$\frac{1}{4}$
	T	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{3}{4}$

$$Q = \begin{pmatrix} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{pmatrix}$$

Jukes-Cantor (JC):

stejné frekvence bází
stejné frekvence substitucí

Kimura 2-parameter (K2P): transice \neq transverze



$$Q = \begin{pmatrix} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \beta \\ \beta & \alpha & \beta & - \end{pmatrix}$$

Jestliže $\alpha = \beta$, K2P = JC

Felsenstein (F81): různé frekvence bází

$$Q = \begin{pmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{pmatrix}$$

Jestliže $\pi_A = \pi_C = \pi_G = \pi_T$, F81 = JC

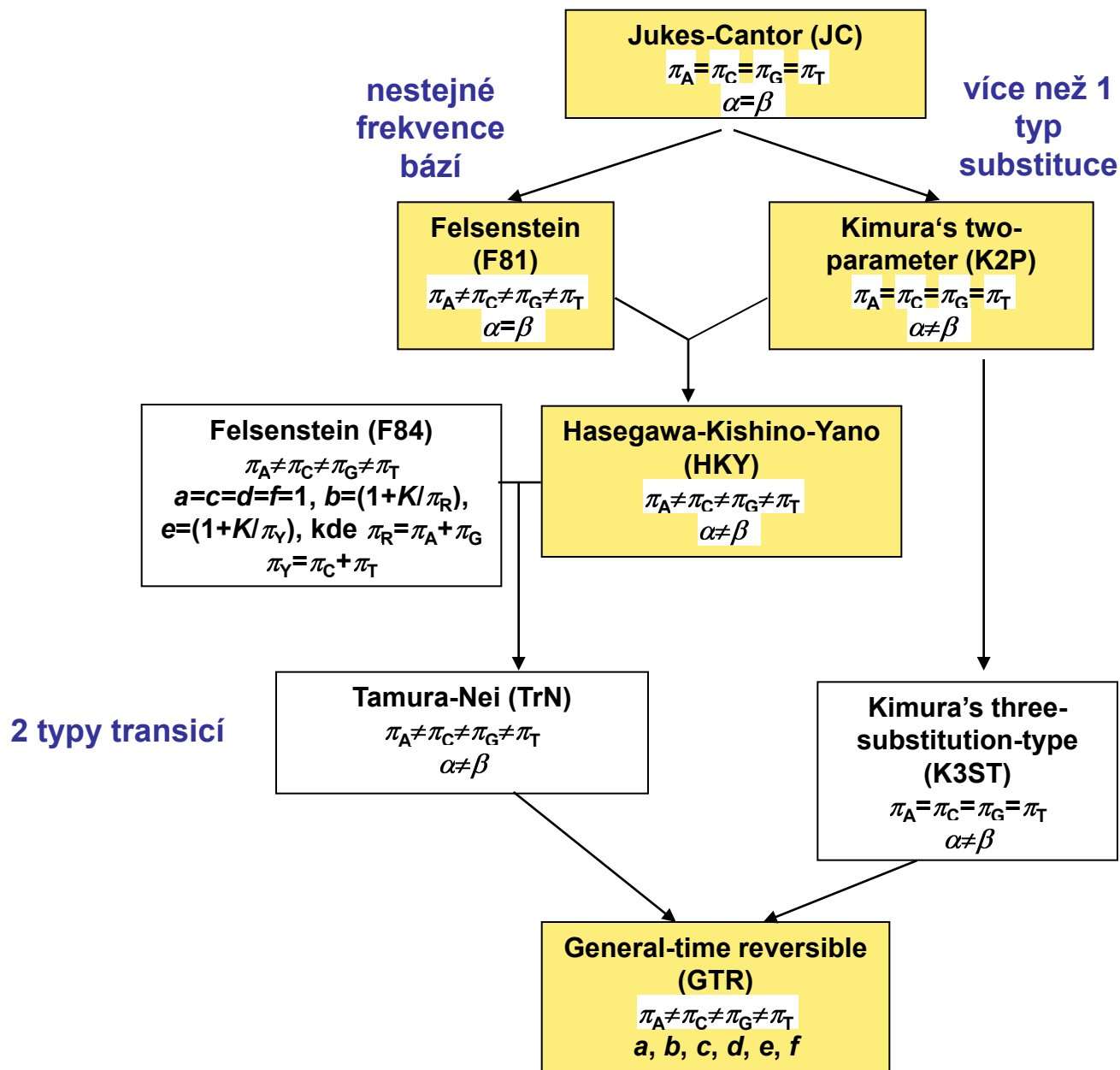
Hasegawa-Kishino-Yano (HKY):

různé frekvence bází
transice \neq transverze

$$Q = \begin{pmatrix} - & \pi_C \beta & \pi_G \alpha & \pi_T \beta \\ \pi_A \beta & - & \pi_G \beta & \pi_T \alpha \\ \pi_A \alpha & \pi_C \beta & - & \pi_T \beta \\ \pi_A \beta & \pi_C \alpha & \pi_G \beta & - \end{pmatrix}$$

General time-reversible (GTR, REV):

různé frekvence bází
různé frekvence jednotlivých typů
substitucí



Heterogenita substitučních rychlostí v různých částech sekvence

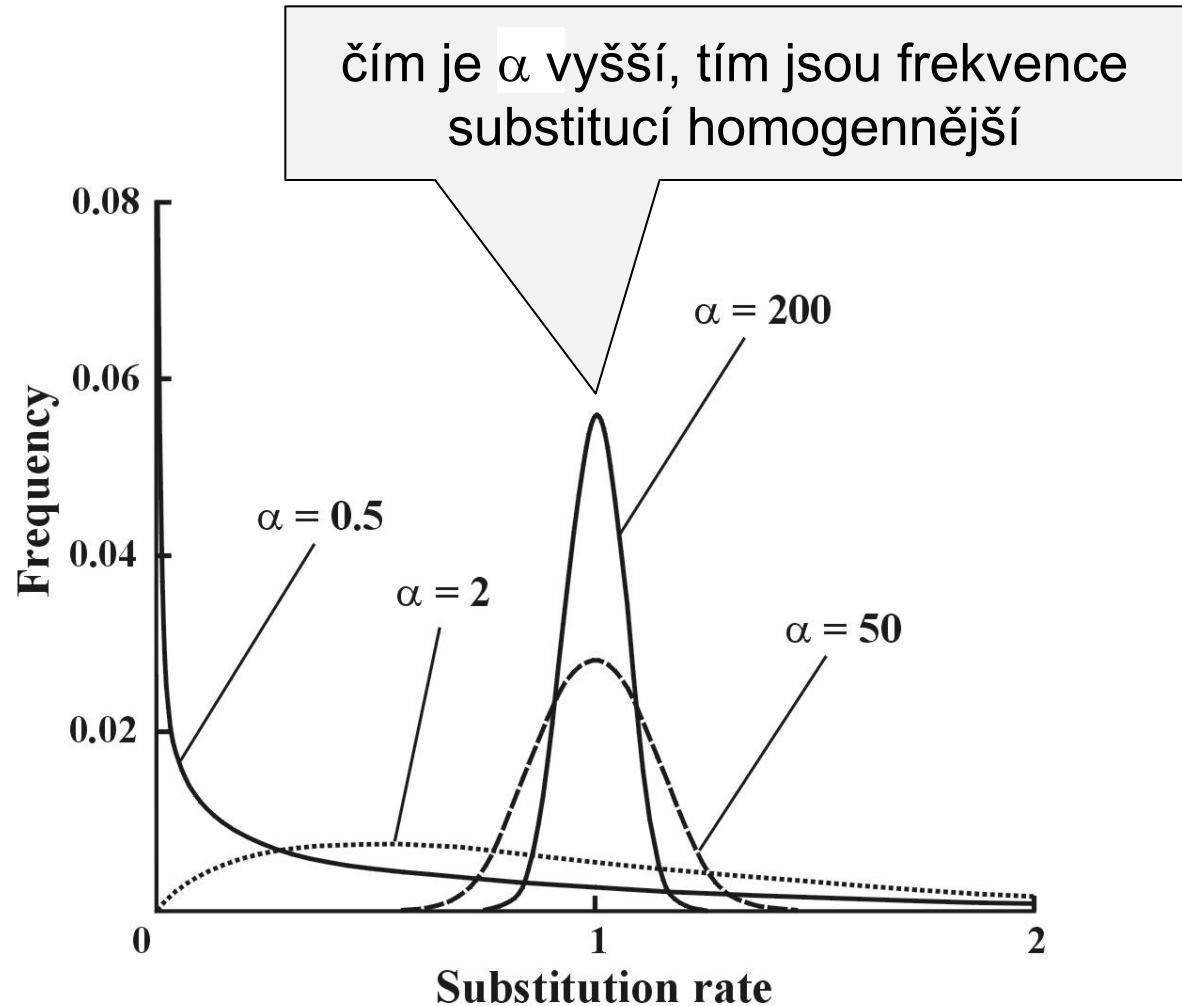
Gama (Γ) rozdělení:

parametr tvaru α

diskrétní gama model

invariantní pozice

→ GTR+ Γ +I



Porovnání modelů:

Likelihood ratio test (LRT):

zahnížděné modely (*nested models*)

$$LR = 2(\ln L_2 - \ln L_1)$$

χ^2 rozdělení, $p_2 - p_1$ stupňů volnosti

Akaike information criterion (AIC):

nonnested models

$AIC = -2\ln L + 2p$, kde p = počet volných parametrů

lepší model \rightarrow nižší AIC

Bayesian information criterion (BIC):

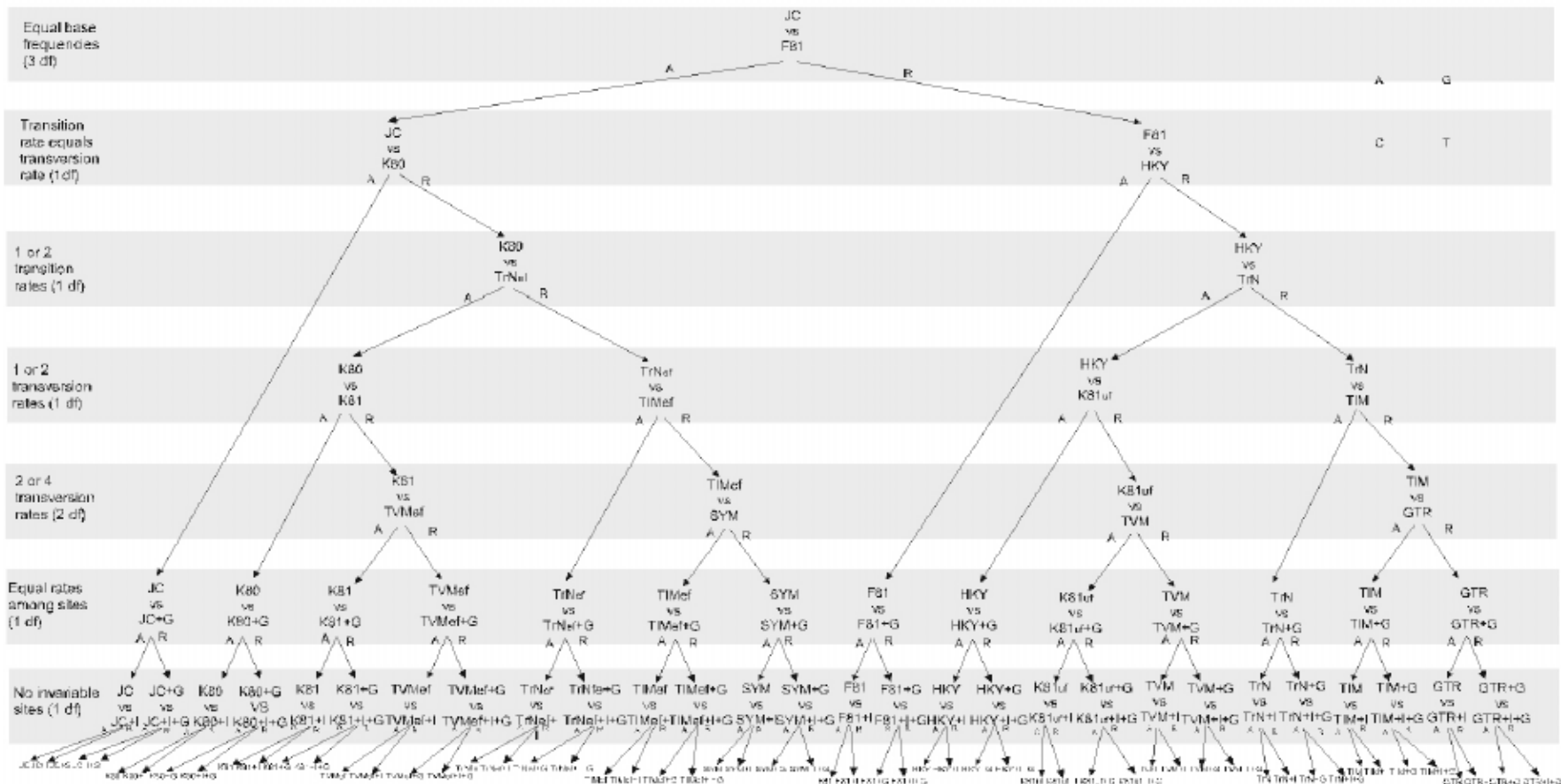
nonnested models

$BIC = -2\ln L + p\ln N$, kde N = velikost vzorku

Porovnání modelů:

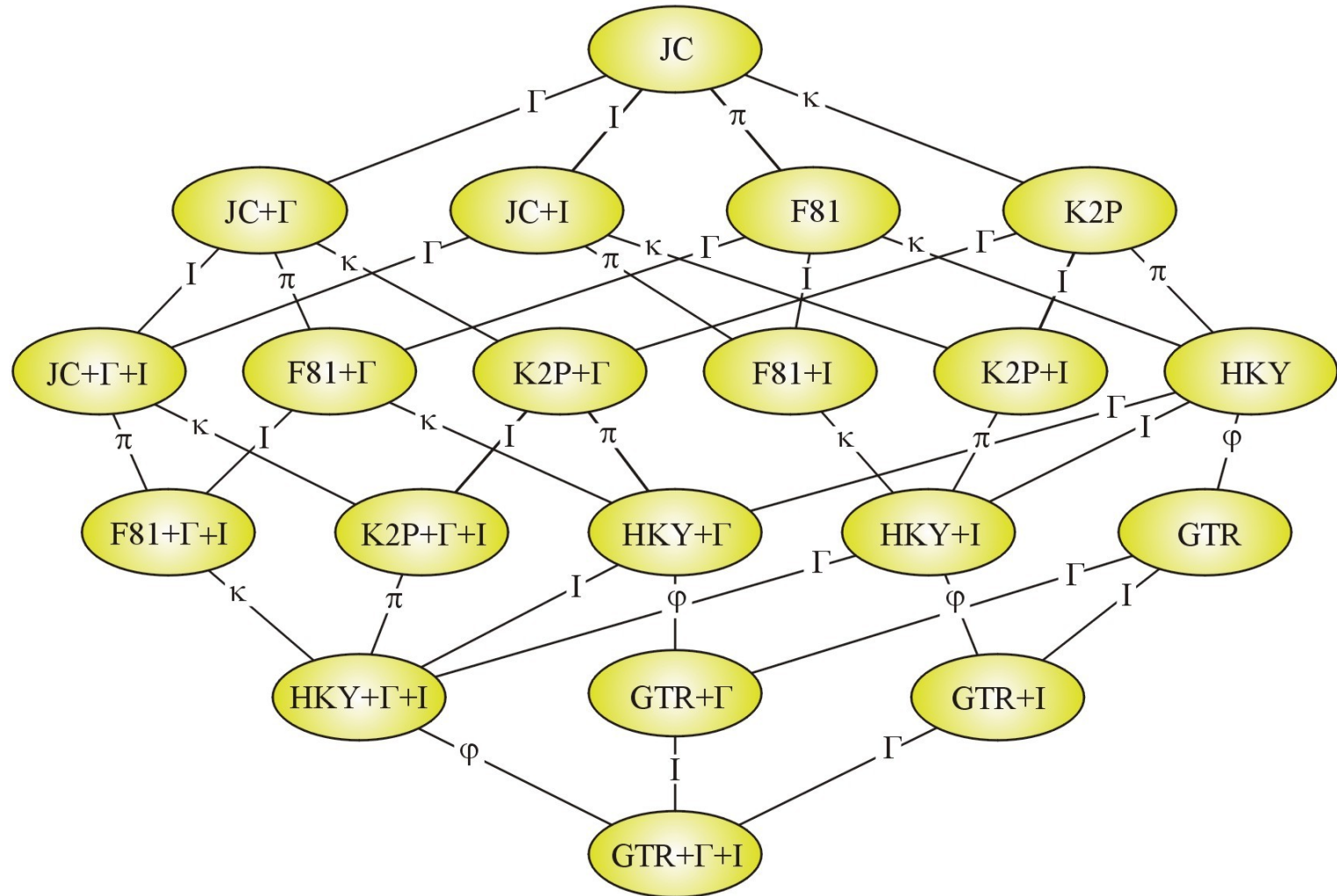
hierarchický LRT – ModelTest (Crandall and Posada)

Modeltest 3.0 hierarchy

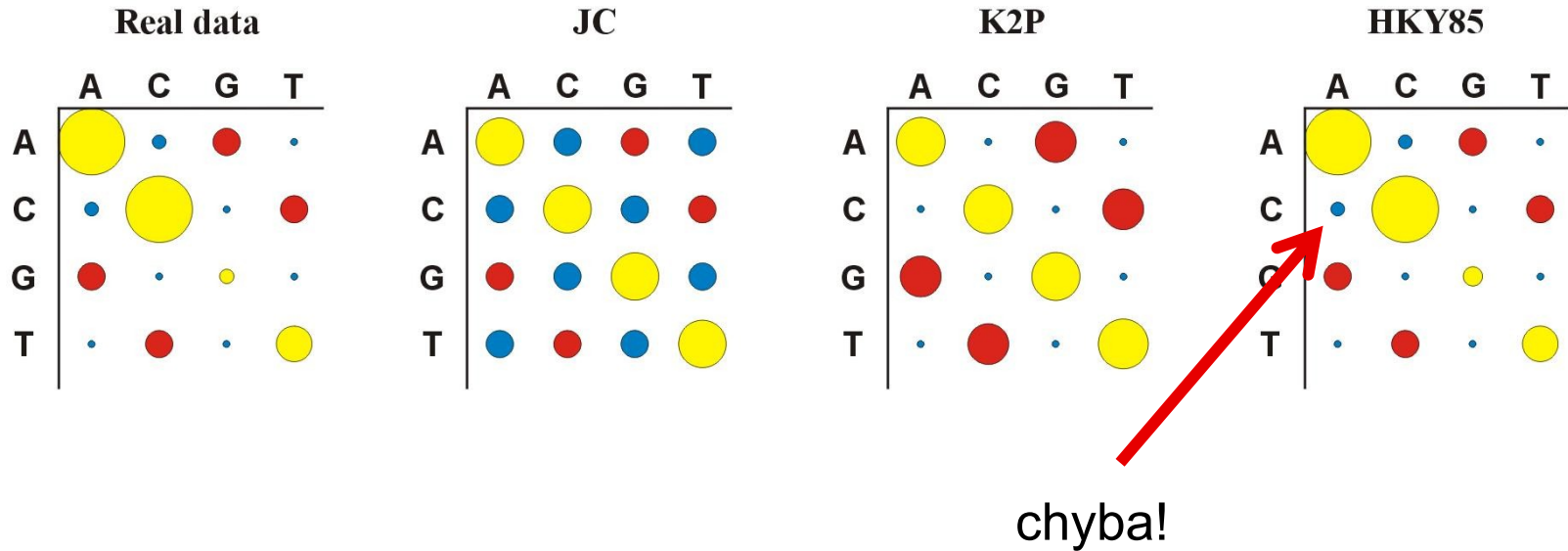


Porovnání modelů:

dynamický LRT:



Porovnání modelů:



Více parametrů \Rightarrow více realismu, ale ...

... také více neurčitosti, protože jsou odhadovány ze stejného množství dat

Distance

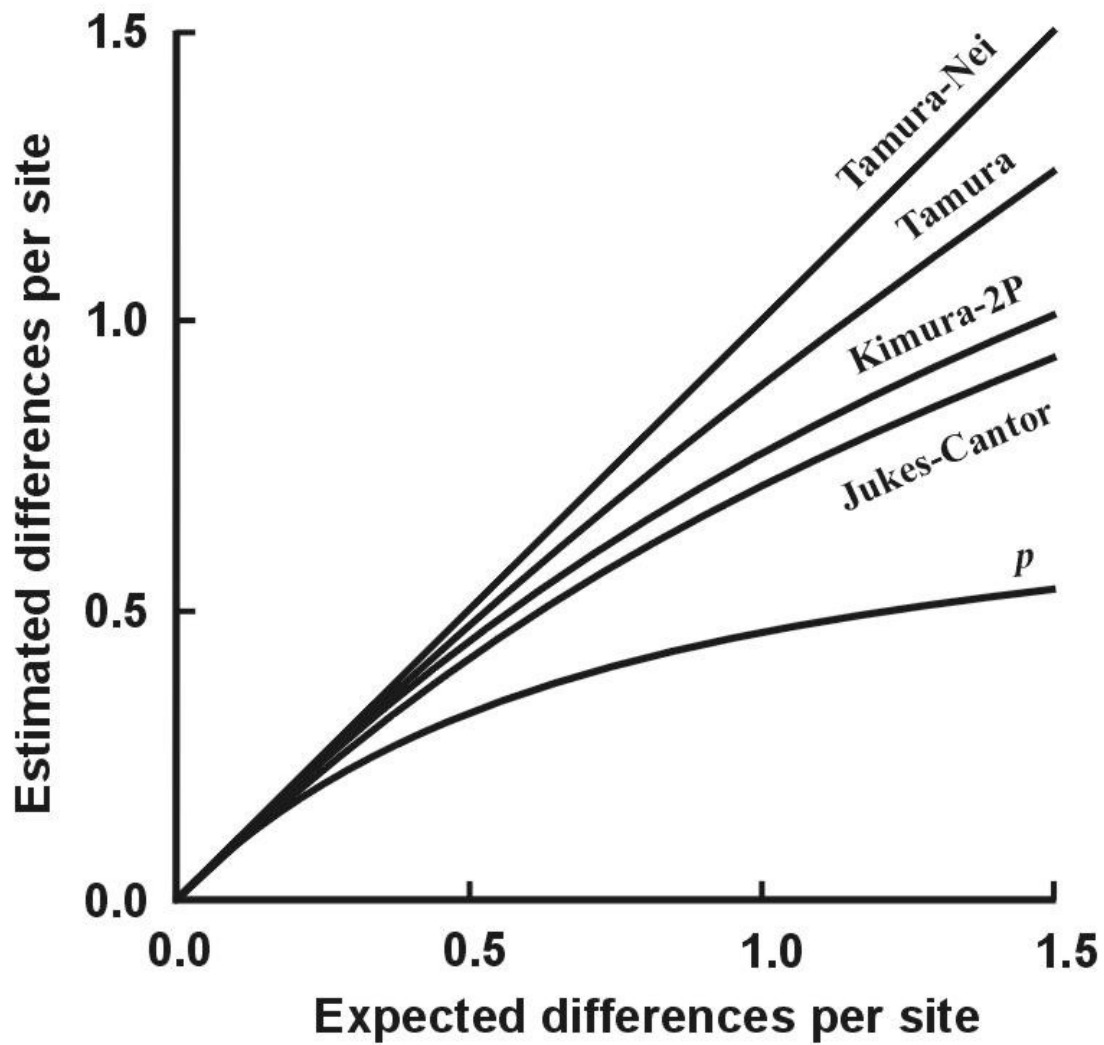
počítány pro každý pár taxonů, z matice distancí (nebo podobností)
konstruován strom

distanční metody založeny na předpokladu, že pokud bychom znali
skutečné distance mezi všemi studovanými taxony, mohli bychom
velmi jednoduše rekonstruovat správnou fylogenii

výhoda: velmi rychlé a jednoduché (lze i na kalkulačce)

Distance pro některé modely:

JC	$d_{xy} = -\frac{3}{4} \ln \left(1 - \frac{4}{3} D \right)$	$D = 1 - (a + f + k + p)$
F81	$d_{xy} = -B \ln \left(1 - \frac{D}{B} \right)$	$D = \text{jako JC}$ $B = 1 - (\pi_A^2 + \pi_C^2 + \pi_G^2 + \pi_T^2)$
K2P	$d_{xy} = \frac{1}{2} \ln \left(\frac{1}{1 - 2P - Q} \right) + \frac{1}{4} \ln \left(\frac{1}{1 - 2Q} \right)$	rozdíly typu transicí: $P = c + h + i + n$ rozdíly typu transverzí: $Q = b + d + e + g + j + l + m + o$
F84	$d_{xy} = -2A \ln \left(1 - \frac{P}{2A} - \frac{(A-B)Q}{2AC} \right) +$ $2(A-B-C) \ln \left(1 - \frac{Q}{2C} \right)$	$\pi_Y = \pi_C + \pi_T, \pi_R = \pi_A + \pi_G,$ $A = \pi_C \pi_T / \pi_Y + \pi_A \pi_G / \pi_R,$ $B = \pi_C \pi_T + \pi_A \pi_G,$ $C = \pi_R \pi_Y, P \text{ a } Q \text{ jako K2P}$
GTR	$d_{xy} = -\text{stopa} \prod \ln \prod^{-1} \mathbf{F}_{xy}$	$\Pi = \text{diagonální matice průměrných četností bází v sekvencích } X \text{ a } Y$



Shluková analýza - UPGMA

	šimp.	bonobo	gorila	člověk	orang.
šimpanz (Š)	--				
bonobo (B)	0,0118	--			
gorila (G)	0,0427	0,0416	--		
člověk (Č)	0,0382	0,0327	0,0371	--	
orangutan (O)	0,0953	0,0916	0,0965	0,0928	--

1. Najdi min $d(ij)$
2. Vypočítej novou matici: $d(\check{S}B-k) = [d(B-k)+d(\check{S}-k)]/2$
3. Opakuj 1 a 2.

	ŠB	gorila	člověk	orang.
ŠB	--			
gorila (G)	0,0422	--		
člověk (Č)	0,0355	0,0371	--	
orangutan (O)	0,0935	0,0965	0,0928	--

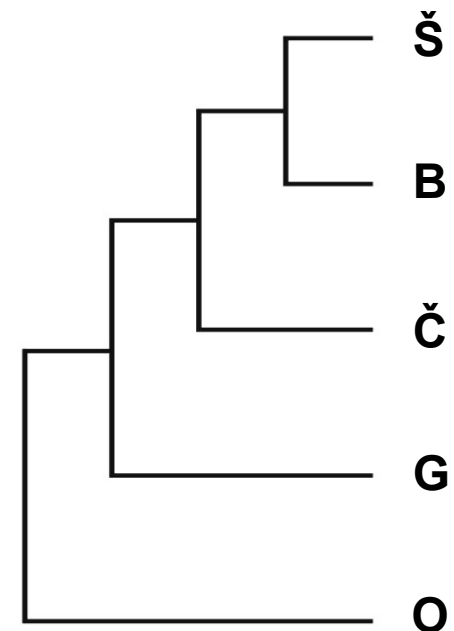
UPGMA (*unweighted pair-group method using arithmetic means*):

$$d[(BŠČ)G] = \{d(BG) + d(ŠG) + d(ČG)\} / 3$$

WPGMA: $d[(BŠČ)G] = \{d[(BŠ)G] + d(ČG)\} / 2$

single-linkage (metoda nejbližšího souseda)

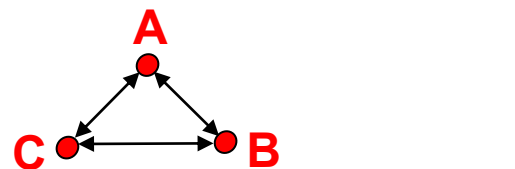
complete-linkage (m. nejvzdálenějšího souseda)



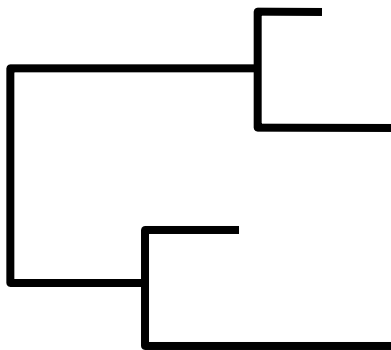
UPGMA a konzistence

aditivní distance: $d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC})$

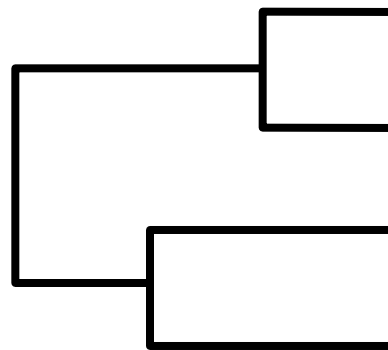
tj. vzdálenost mezi 2 taxony je rovna součtu větví,
které je spojují



ultrametrické distance: $d_{AC} \leq \max(d_{AB}, d_{BC})$

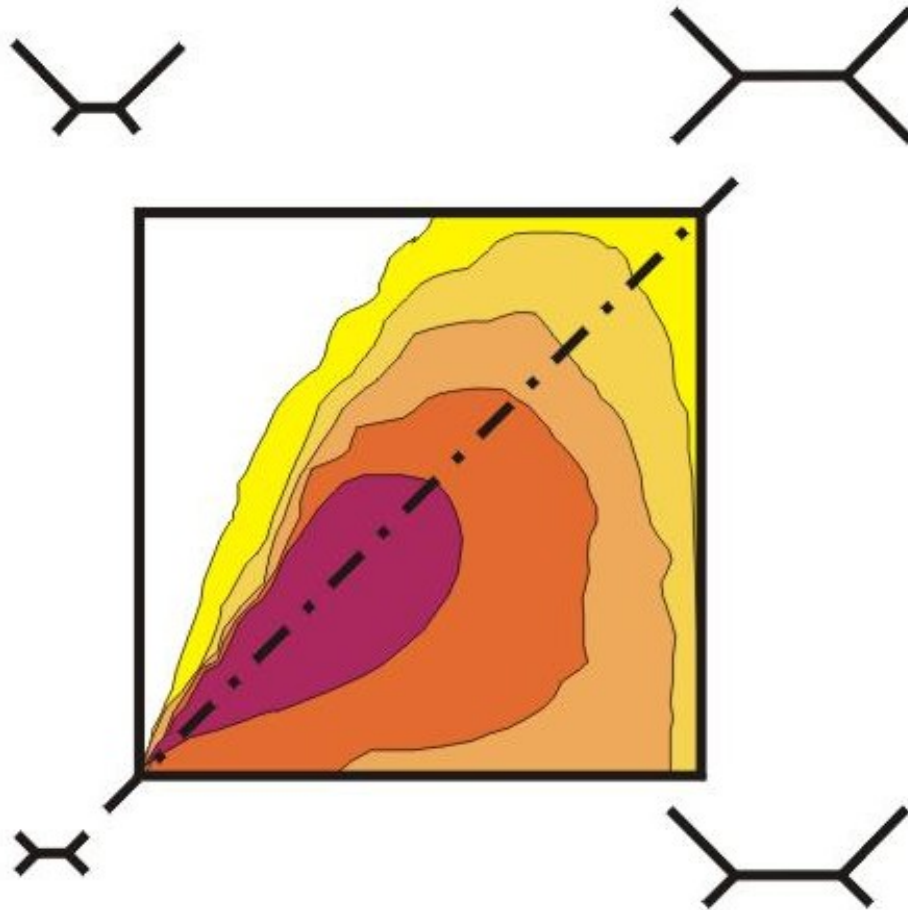


aditivní strom



ultrametrický strom

UPGMA a konzistence



Spojení sousedů (neighbor-joining, NJ)

Algoritmická metoda

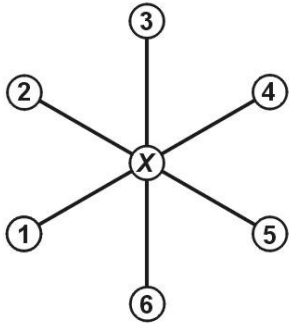
Princip minimální evoluce → minimalizuje součet délek větví S

Každý pár uzlů adjustován na základě divergence od ostatních

Konstrukce jediného aditivního stromu

hvězdicový
strom

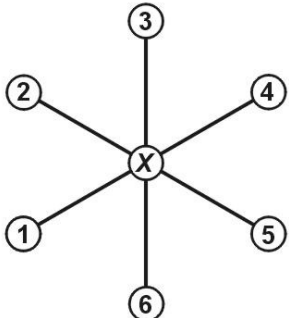
a)



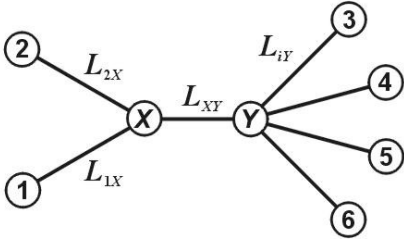
hvězdicový strom

nalezení
nejbližších
sousedů

a)

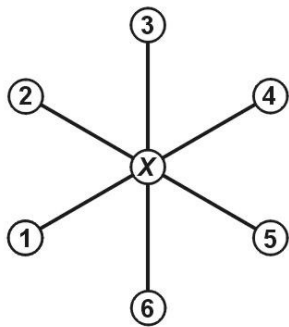


b)



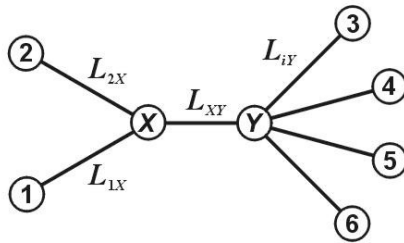
hvězdicový strom

a)



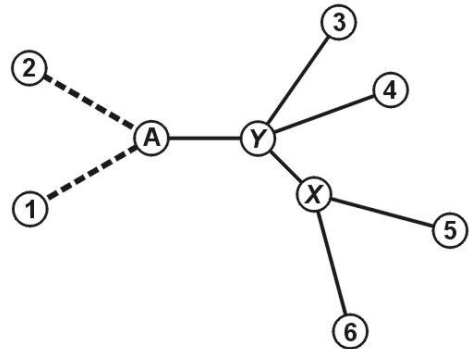
nalezení
nejbližších
sousedů

b)



přepočítání
distancí

c)

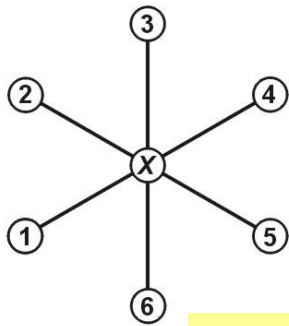


hvězdicový strom

nalezení
nejbližších
sousedů

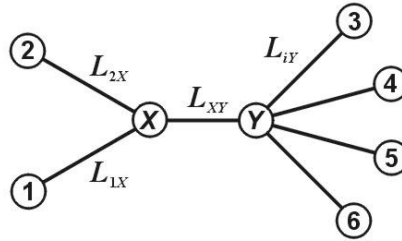
přepočítání
distancí

a)

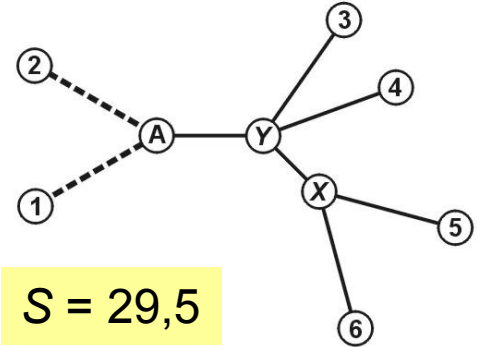


$S = 32,4$

b)

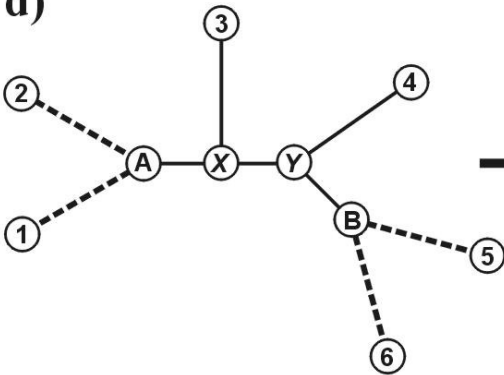


c)

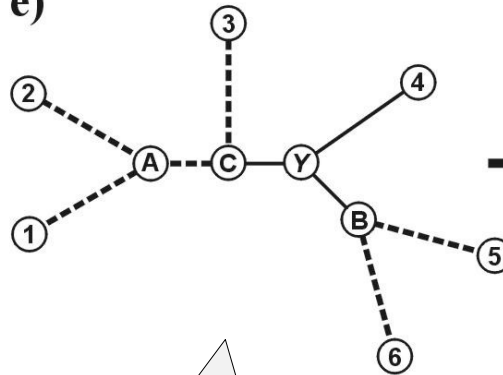


$S = 29,5$

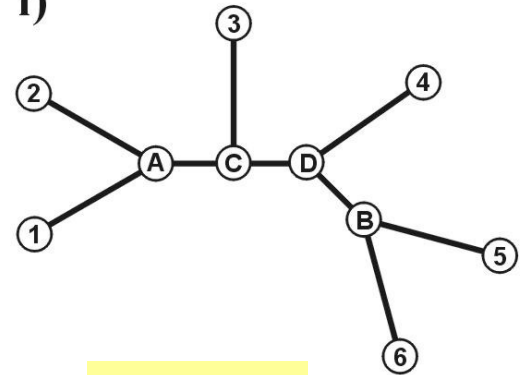
d)



e)



f)



$S = 28,0$

opakování
postupu ...

Nevýhody distančních dat:

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