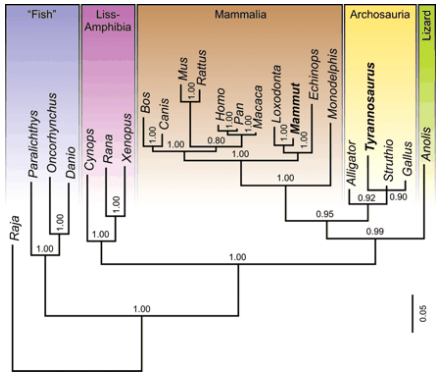
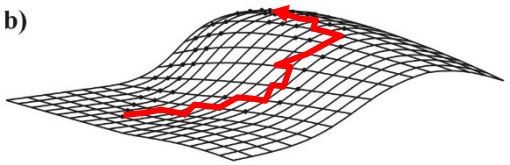
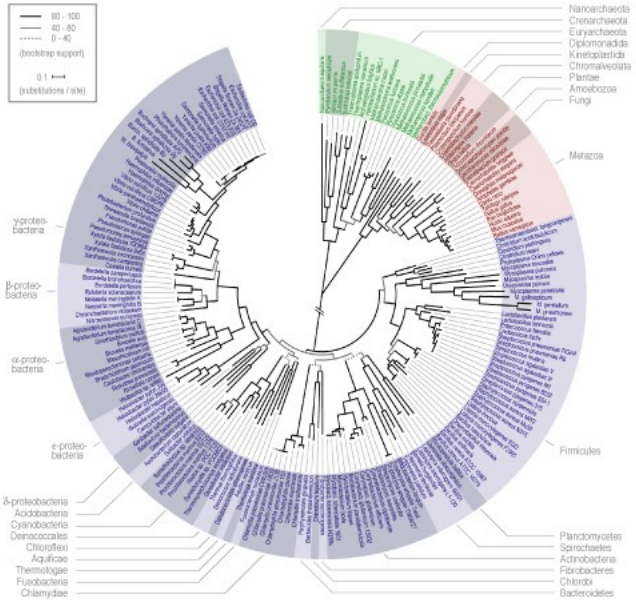
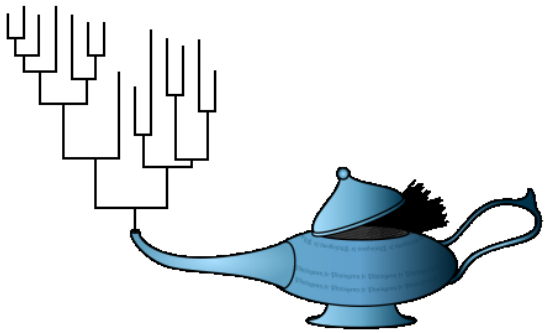
ANALÝZA DNA SEKVENCÍ





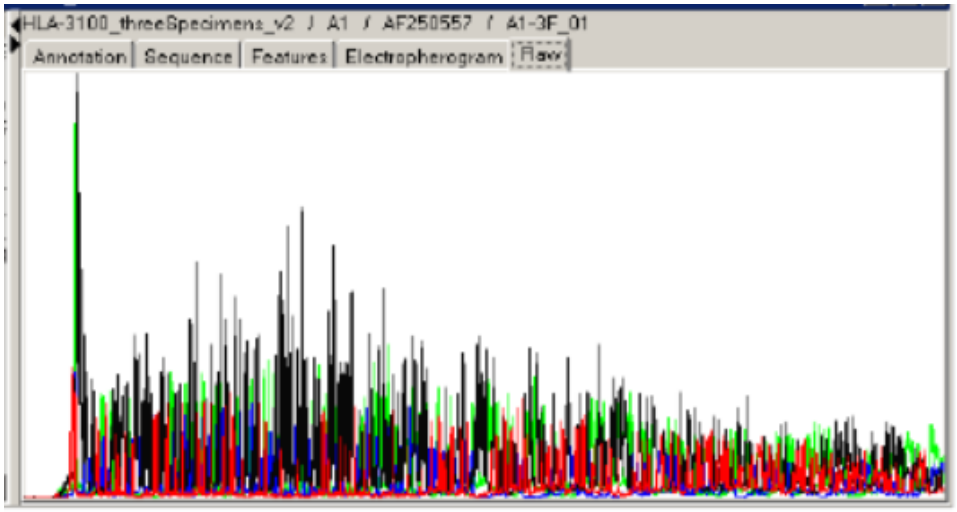




Kde získat sekvence?

- Sangerovo sekvenování – .ab1 files
- GenBank či jiná databáze
- Dryad – publikované datasety
- NGS – FASTQ (obsahuje i informaci o kvalitě sekvence) – specifická analýza dat

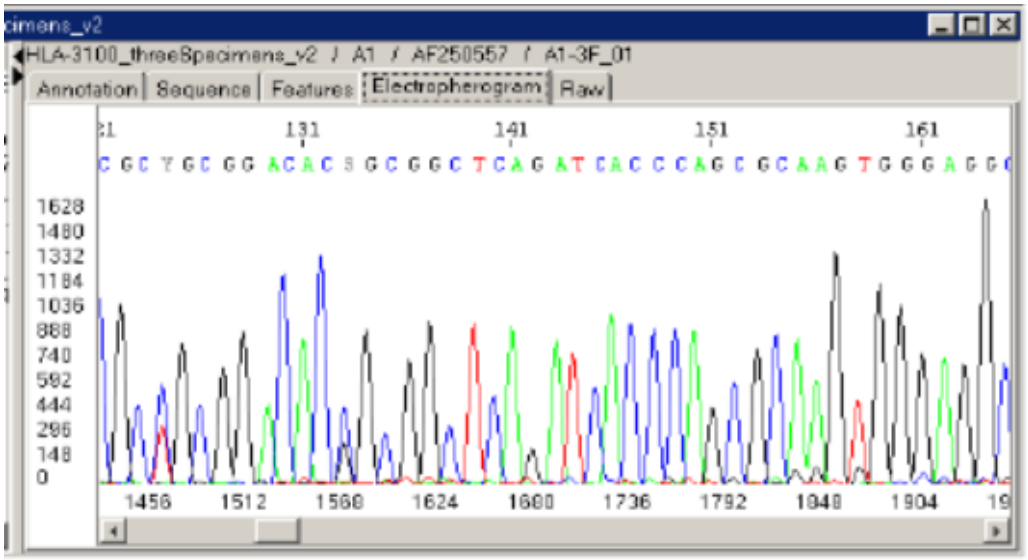
Editace sekvencí – Sangerovo sekvenování



Raw data
(.ab1 file)

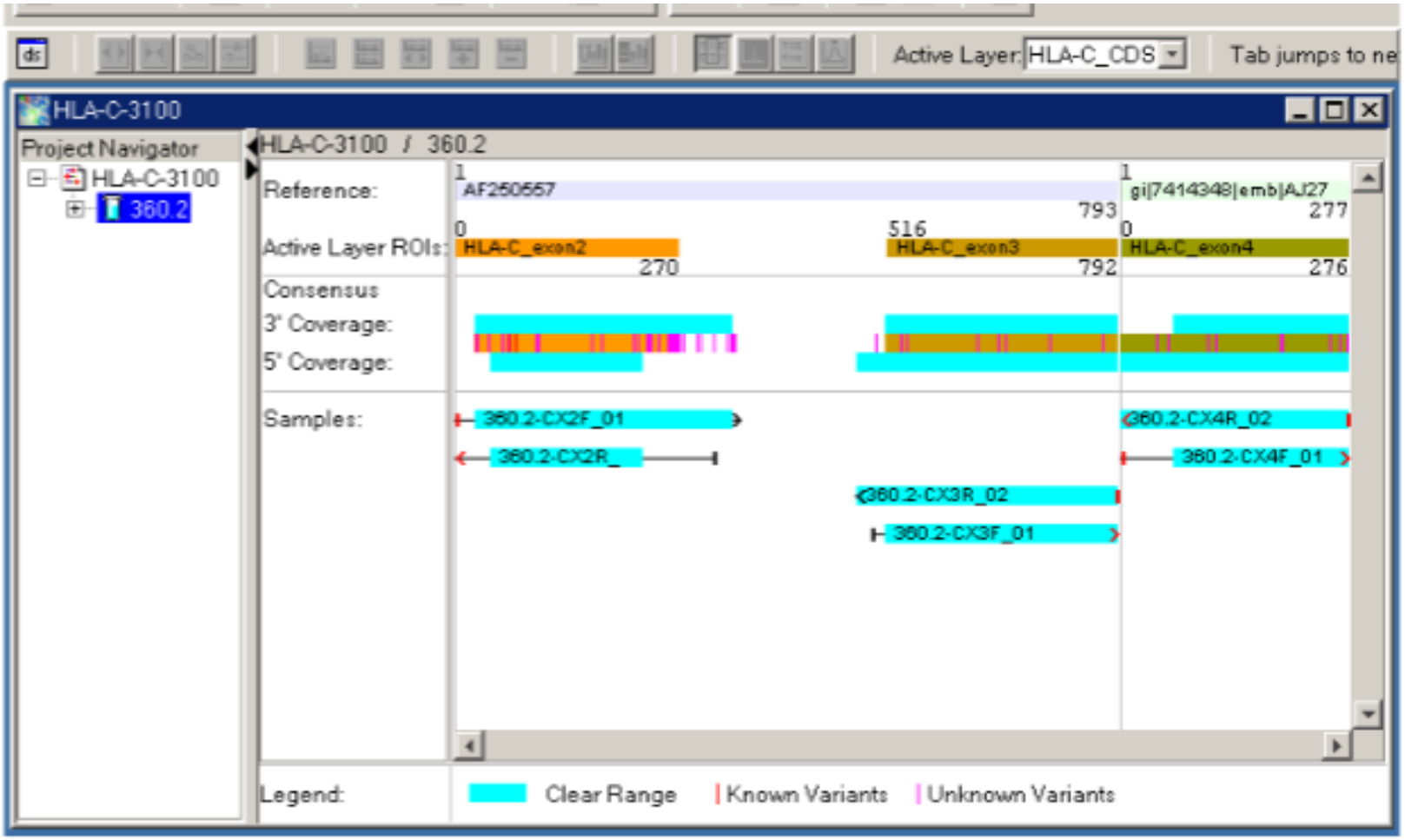
„basecalling“

SeqScape
Sequencher
Geneious

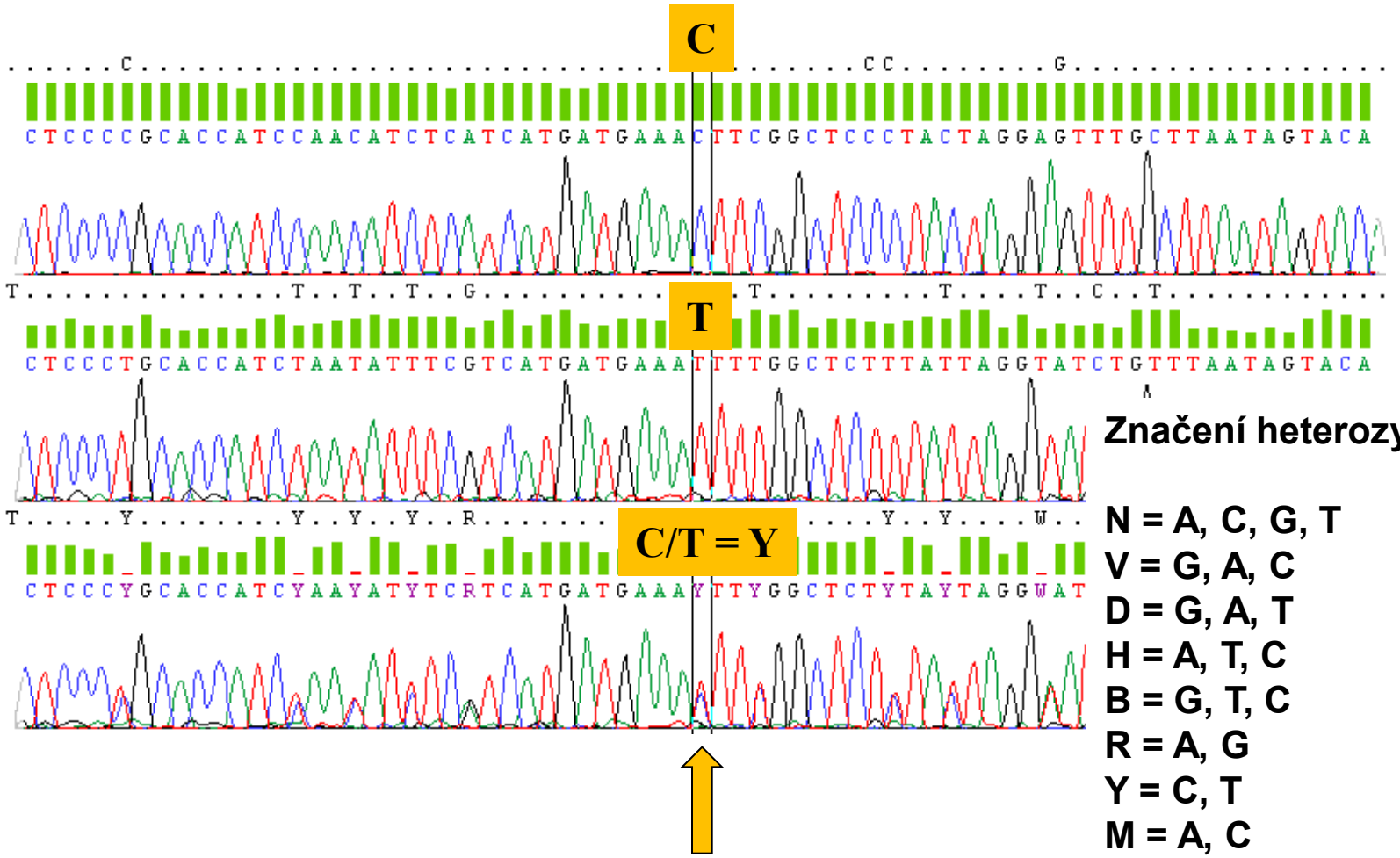


Editovaná sekvence

„Alignment“ → contig (ze stejného jedince)



Alignment sekvencí z různých jedinců – analýza polymorfismu



Značení heterozygotů

N = A, C, G, T
V = G, A, C
D = G, A, T
H = A, T, C
B = G, T, C
R = A, G
Y = C, T
M = A, C
K = G, T
S = G, C
W = A, T

Nukleotidové a proteinové sekvence:

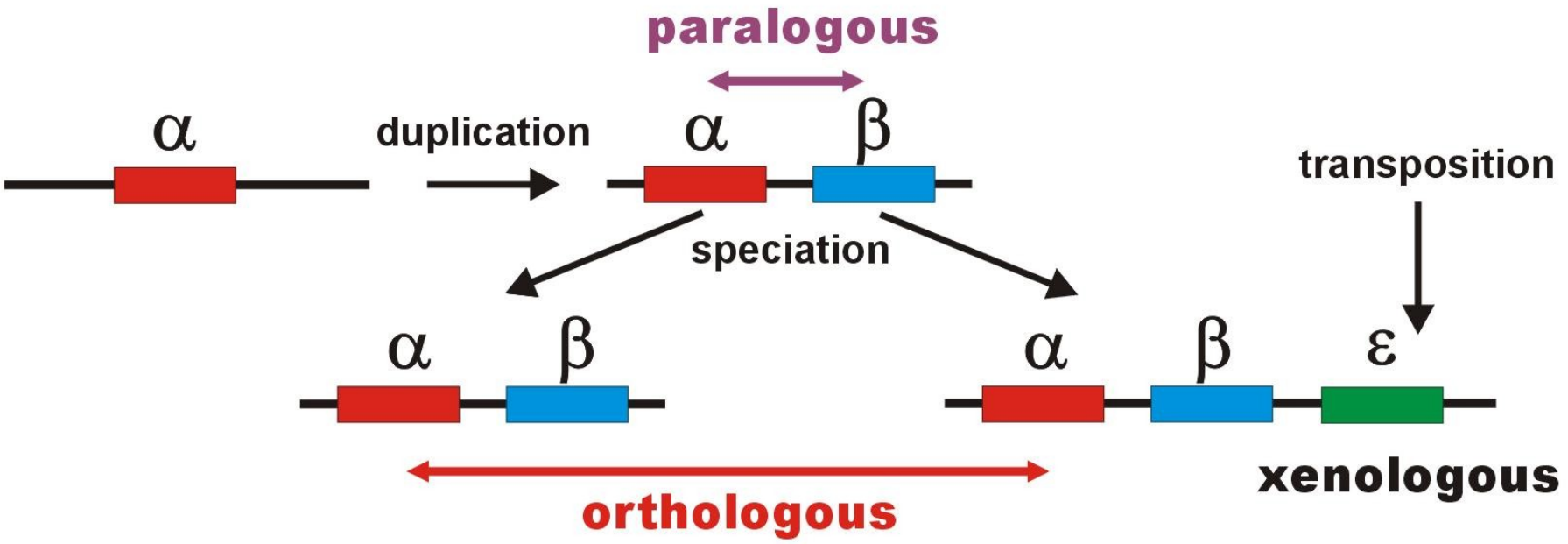
```
H_sapiens MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGS
```

báze = stav znaku

```
P_troglod ATGACCCCGACACGCAAAATTAACCCACTAATAAA
```

pozice (site) = znak

Problém homologie sekvencí



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/*

Proteinové databáze:

SWISS-PROT – University of Geneve & Swis 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
ATGACCCCAATACGCAAAATTAACCCCCTAATAAAATTAATTAACCACTCATTCATCGACCTCCCCACCC
CATCCAACATCTCCGCATGATGAAACTTCGGCTCACTCCTTGGCGCCTGCCTGATCCTCCAAATCACCAC
AGGACTATTCCTAGCCATACACTACTCACCAGACGCCTCAACCGCCTTTTCATCAATCGCCCACATCACT
CGAGACGTAAATTATGGCTGAATCATCCGCTACCTTCACGCCAATGGCGCCTCAATATTCTTTATCTGCC
TCTTCCTACACATCGGGCGAGGCCTATATTACGGATCATTTCTCTACTCAGAAACCTGAAACATCGGCAT
...
>P_troglod
ATGACCCCGACACGCAAAATTAACCCACTAATAAAATTAATTAATCACTCATTTATCGACCTCCCCACCC
CATCCAACATTTCCGCATGATGGAACTTCGGCTCACTTCTCGGCGCCTGCCTAATCCTTCAAATTACCAC
AGGATTATTCCTAGCTATACACTACTCACCAGACGCCTCAACCGCCTTCTCGTCGATCGCCCACATCACC
CGAGACGTAAACTATGGTTGGATCATCCGCTACCTCCACGCTAACGGCGCCTCAATATTTTTTATCTGCC
TCTTCCTACACATCGGCCGAGGTCTATATTACGGCTCATTTCTCTACCTAGAAACCTGAAACATTGGCAT
...
>P_paniscus
ATGACCCCAACACGCAAAATCAACCCACTAATAAAATTAATTAATCACTCATTTATCGACCTCCCCACCC
CATCCAATATTTCCACATGATGAAACTTCGGCTCACTTCTCGGCGCCTGCCTAATCCTTCAAATCACCAC
AGGACTATTCCTAGCTATACACTACTCACCAGACGCCTCAACCGCCTTCTCATCGATCGCCCACATTACC
CGAGACGTAAACTATGGTTGAATCATCCGCTACCTTCACGCTAACGGCGCCTCAATACTTTTCATCTGCC
TCTTCCTACACGTCGGTCGAGGCCTATATTACGGCTCATTTCTCTACCTAGAAACCTGAAACATTGGCAT
...
```

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 ttatactaat acatcaaatt aatgctttaa agacatatct gtgttatctg acatacacca
      301 tacagtcata aactcttctc ttccatatga ctatcccctt ccccatttgg tctattaatc
      361 taccatcctc cgtgaaacca acaacccgcc caccaatgcc cctcttctcg ctccgggccc
      421 attaaacttg ggggtagcta aactgaaact ttatcagaca tctggttctt acttcagggc
      481 catcaaatgc gttatcgccc atacgttccc cttaaataag acatctcgat ggtatcgggt
      541 ctaatcagcc catgaccaac ataactgtgg tgtcatgcat ttggtatttt tttattttgg
      601 cctactttca tcaacatagc cgtcaaggca tgaaaggaca gcacacagtc tagacgcacc
      661 tacggtgaag aatcattagt ccgcaaaacc caatcaccta aggctaatta ttcatgcttg
      721 ttagacataa atgctactca ataccaaatt ttaactctcc aaaccccccа accccctcct
      781 cttaatgcca aaccccaaaa acactaagaa cttgaaagac atatattatt aactatcaaa
      841 ccctatgtcc tgatcgattc tagtagttcc caaaatatga ctcatatttt agtacttgta
      901 aaaatttтac aaaatcatgc tccgtgaacc aaaactctaa tcacactcta ttacgcaata
      961 aatattaaca agttaatgta gcttaataac aaagcaaagc actgaaaatg cttagatgga
     1021 taattttatc cca
//
```

Formáty souborů

PHYLIP (“interleaved” format):

```
6 1120
H_sapiens     ATGACCCCAA TACGCAAAAT TAACCCCCTA ATAAAATTAA TTAACCACTC
P_troglod     ATGACCCCGA CACGCAAAAT TAACCCACTA ATAAAATTAA TTAATCACTC
P_paniscus    ATGACCCCAA CACGCAAAAT CAACCCACTA ATAAAATTAA TTAATCACTC
G_gorilla     ATGACCCCTA TACGCAAAAC TAACCCACTA GCAAAACTAA TTAACCACTC
P_pygmaeus    ATGACCCCAA TACGCAAAAC CAACCCACTA ATAAAATTAA TTAACCACTC
H_lar         ATGACCCCCC TGCGCAAAAC TAACCCACTA ATAAAACTAA 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   ATGACCCCGACACGCAAAATTAACCCACTAATAAAATTAATTAATCACTC
P_paniscus  ATGACCCCAACACGCAAAATCAACCCACTAATAAAATTAATTAATCACTC
H_sapiens   ATGACCCCAATACGCAAAATTAACCCCCTAATAAAATTAATTAACCACTC
G_gorilla   ATGACCCCTATACGCAAAACTAACCCACTAGCAAAACTAATTAACCACTC
P_pygmaeus  ATGACCCCAATACGCAAAACCAACCCACTAATAAAATTAATTAACCACTC
H_lar       ATGACCCCCCTGCGCAAAACTAACCCACTAATAAAACTAATCAACCACTC

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

end;
```

Formáty souborů

Clustal:

```
P_troglod  ATGACCCCGACACGCAAAATTAACCCACTAATAAAATTAATTAATCACTCATTTATCGAC
P_paniscus ATGACCCCAACACGCAAAATCAACCCACTAATAAAATTAATTAATCACTCATTTATCGAC
H_sapiens  ATGACCCCAATACGCAAAATTAACCCCCTAATAAAATTAATTAACCACTCATTCATCGAC
G_gorilla  ATGACCCCTATACGCAAAACTAACCCACTAGCAAAACTAATTAACCACTCATTCATTGAC
P_pygmaeus ATGACCCCAATACGCAAAACCAACCCACTAATAAAATTAATTAACCACTCACTCATCGAC
H_lar      ATGACCCCCCTGCGCAAAACTAACCCACTAATAAAACTAATCAACCACTCACTTATCGAC
           ********    *******  ***** ***  **** **** ** ****** * ** ***

P_troglod  CTCCCCACCCCATCCAACATTTCCGCATGATGGAACTTCGGCTCACTTCTCGGCGCCTGC
P_paniscus CTCCCCACCCCATCCAATATTTCCACATGATGAAACTTCGGCTCACTTCTCGGCGCCTGC
H_sapiens  CTCCCCACCCCATCCAACATCTCCGCATGATGAAACTTCGGCTCACTCCTTGGCGCCTGC
G_gorilla  CTCCCTACCCCGTCCAACATCTCCACATGATGAAACTTCGGCTCACTCCTTGGTGCCTGC
P_pygmaeus CTCCCCACCCCATCAAACATCTCTGCATGATGGAACTTCGGCTCACTTCTAGGCGCCTGC
H_lar      CTTCCAGCCCCATCCAACATTTCTATATGATGAAACTTTGGTTCACTCCTAGGCGCCTGC
           ** **  **** ** ** ** **   ****** ***** ** ***** ** ** ******
```

Seřazení sekvencí (alignment)

```
Sekvence 1    TTGTACGACGG
Sekvence 2    TTGTACGACG

TTGTACGACGG   TTGT---ACGACGG
||||||||||    ||||    |||
TTGTACGACG    TTGTACGACG
```

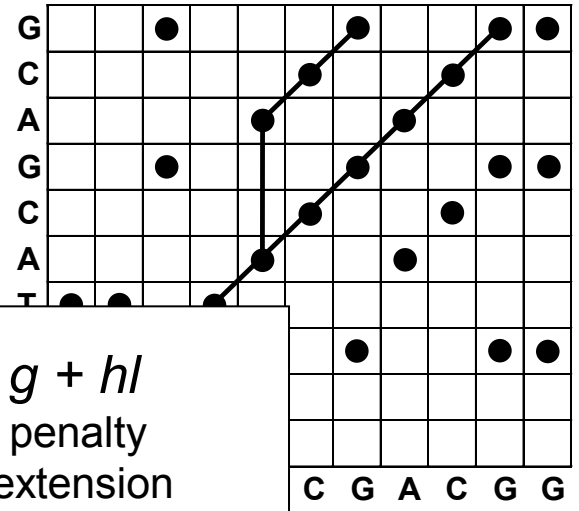
gap penalty

$$GP = g + hl$$

g - gap penalty

h – gap extension penalty

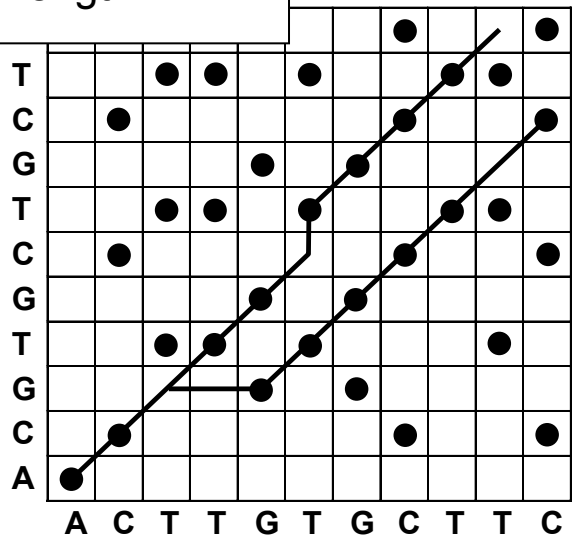
l – gap length




```
Sekvence 1    ACTTGTGCTTC
Sekvence 2    ACGTGCTGCTC

              ACTTG-TGCTTC
Path 1        ||  ||  ||||
              ACGTGCTGCTC

              ACTTGTGCTTC
Path 2        ||   ||||| |
              AC--GTGCTGCTC
```



GenBank (http://www.ncbi.nlm.nih.gov/genbank/)

BLAST
(http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome)

BioEdit nebo AliView – volně dostupné, základní editace sekvencí

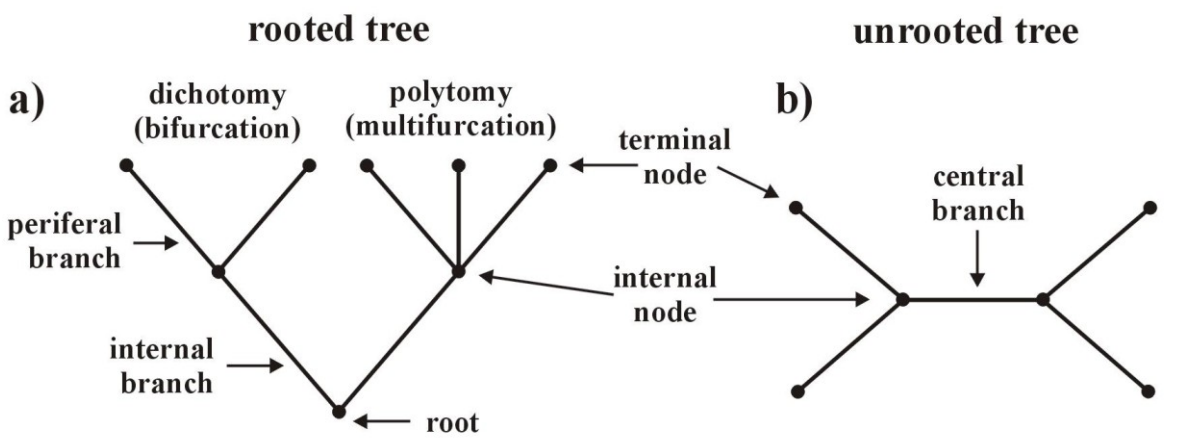
konverze formátů (zejména pro fylogenetickou analýzu):
ALTER (http://sing.ei.uvigo.es/ALTER/)

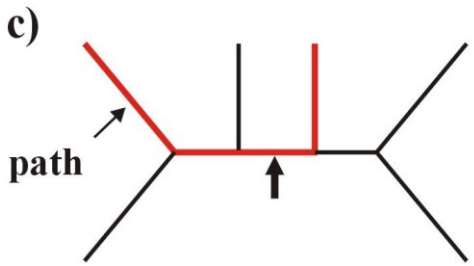
DnaSP

- např. tvorba haplotypového souboru (jednotlivé sekvence seřazeny do haplotypů – pro vytváření haplotypové sítě)

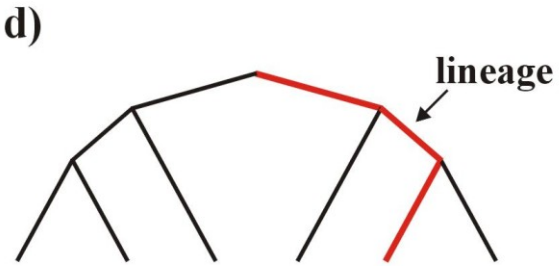
- Phase – separace heterozygotů do haplotypů (MCMC algorithm)

Fylogenetická analýza - definice základních pojmů





dráha



linie

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

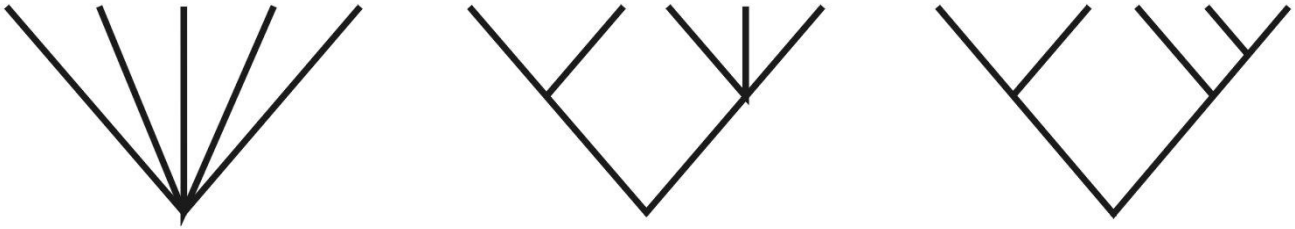
větve (branches, edges)
vnější, vnitřní, centrální

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

dichotomie, polytomie

OTU, HTU

topologie



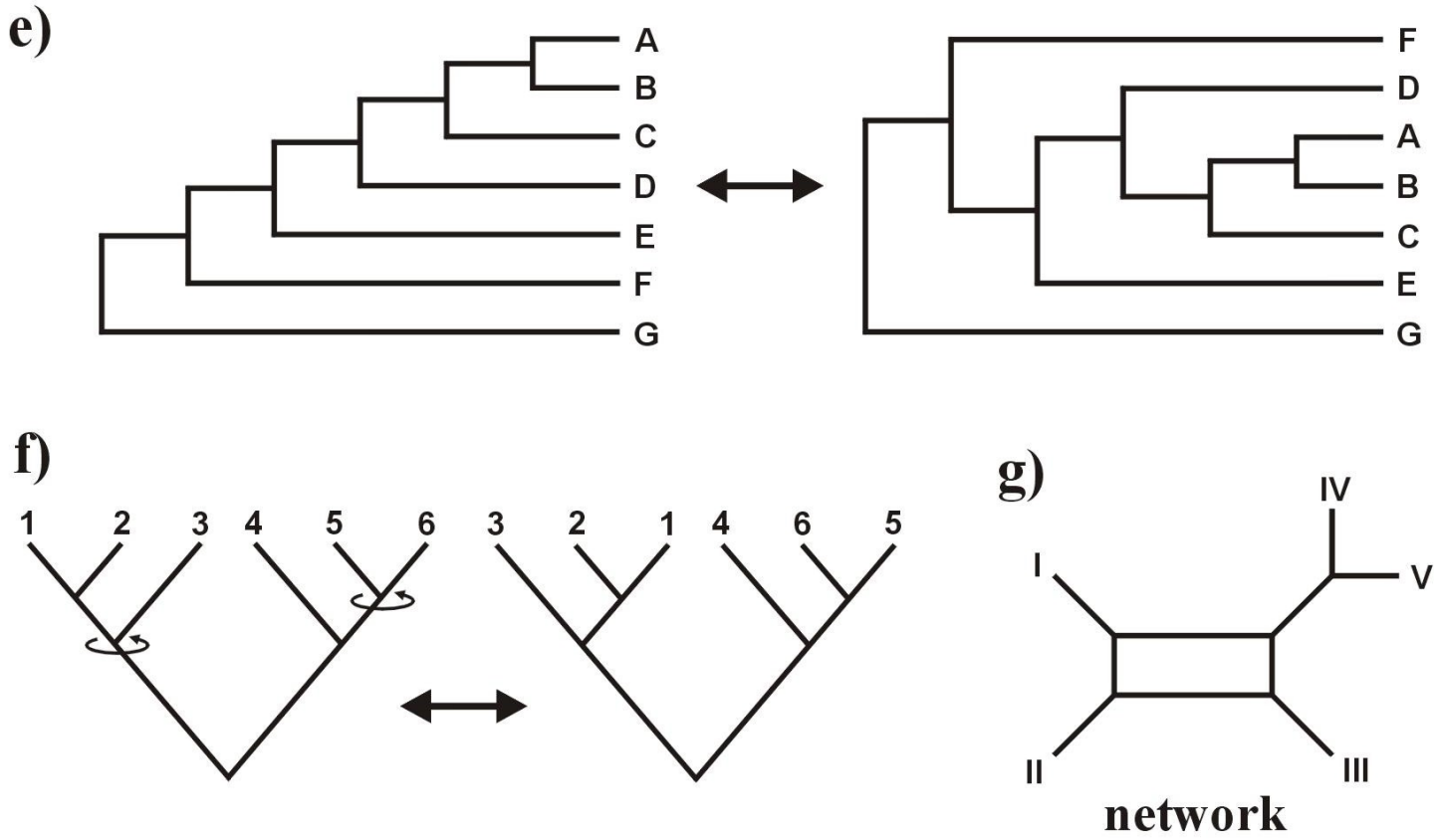
star tree

partly resolved

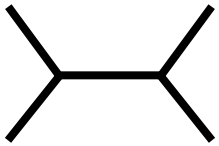
fully resolved

Definice základních pojmů

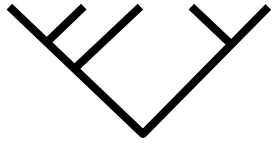
topologie:



Kolik existuje stromů?



$$\frac{(2n-5)!}{2^{n-3}(n-3)}$$

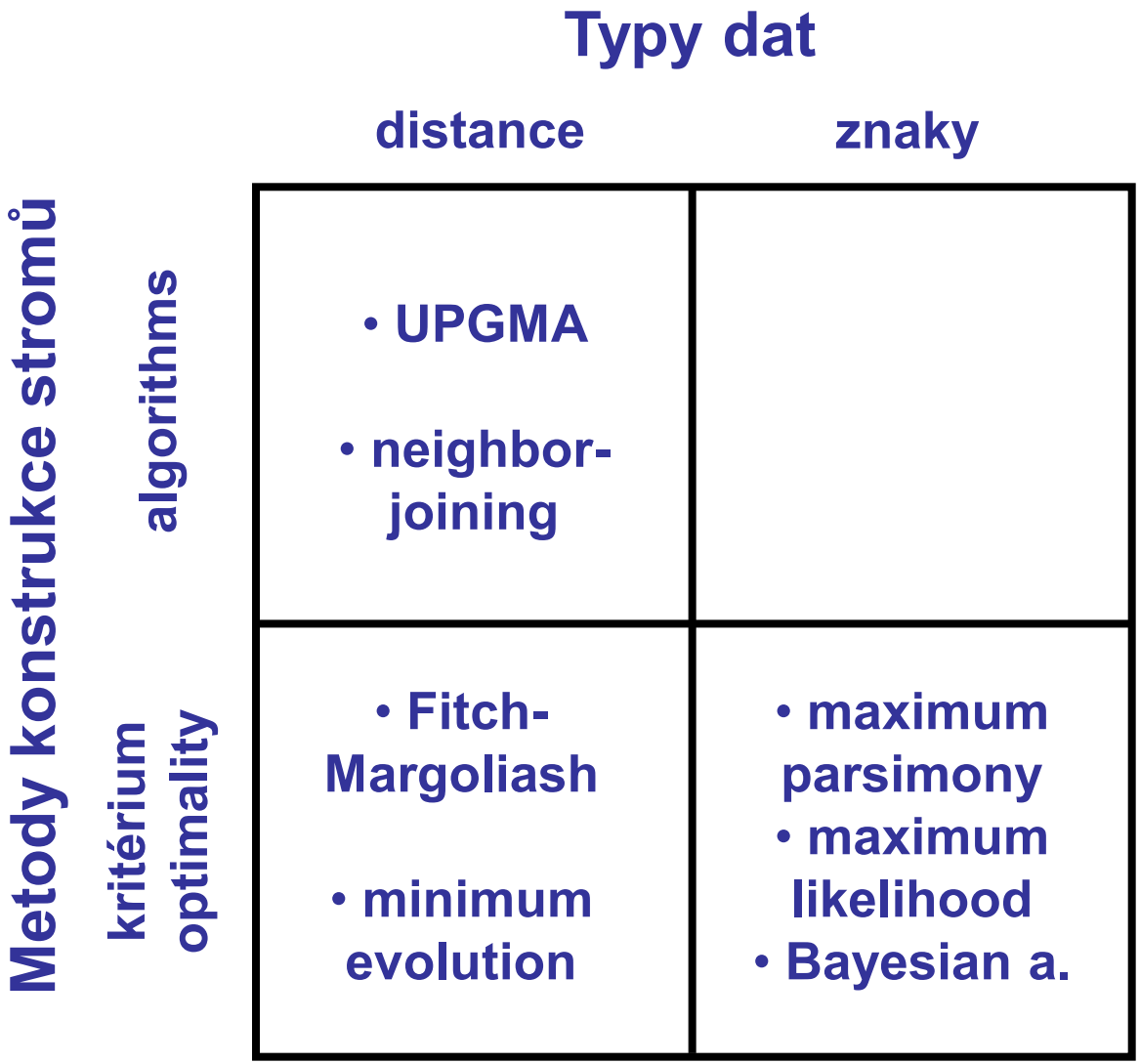


$$\frac{(2n-3)!}{2^{n-2}(n-2)}$$

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	$4{,}9518\times10^{38}$
40	$4{,}9518\times10^{38}$	$1{,}00986\times10^{57}$
50	$1{,}00986\times10^{57}$	$2{,}75292\times10^{76}$

více než elektronů ve viditelném vesmíru (Eddingtonovo číslo)

Rozdělení metod



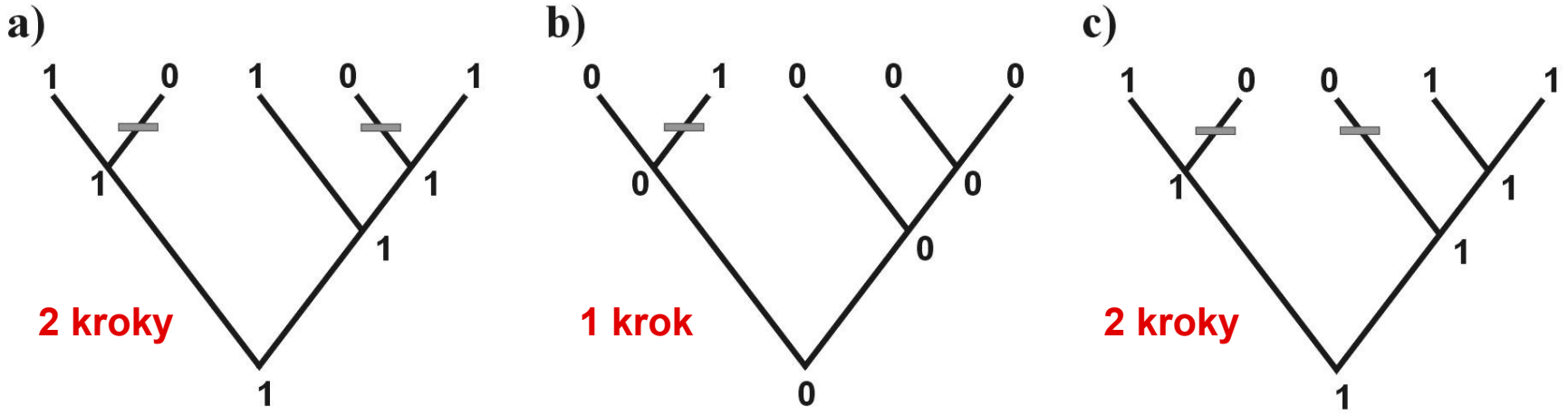
(1) Maximální úspornost (maximum parsimony, MP)

- snaha minimalizovat počet analogických stavů

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

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

minimální počet kroků = 3 (pro každý znak jedna změna)
skutečný počet kroků = 5
⇒ 2 extra kroky → analogie = homoplasie
⇒ stejný stav znaku vzniká vícekrát nezávisle



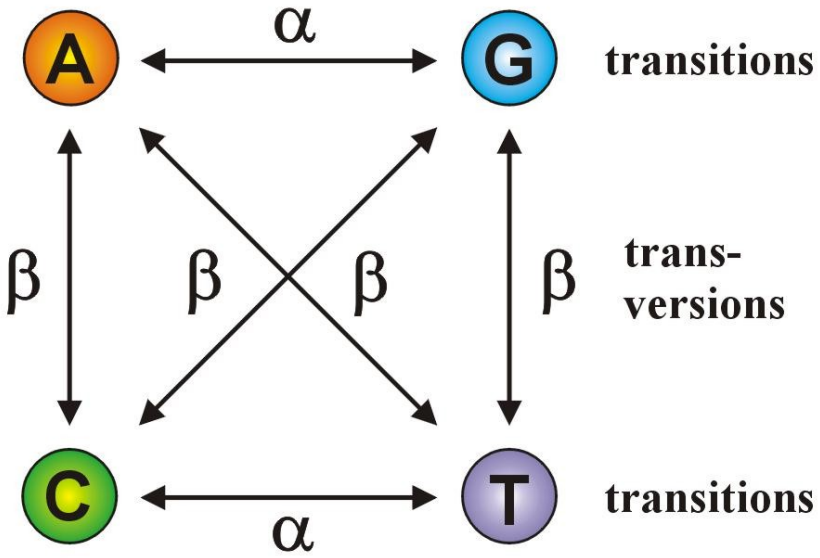
(2) Evoluční (substituční) modely a distanční metody

		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	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{3}{4}$

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

Jukes-Cantor (JC): stejné frekvence bází
stejné frekvence substitucí

Kimura 2-parameter (K2P): transice ≠ 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

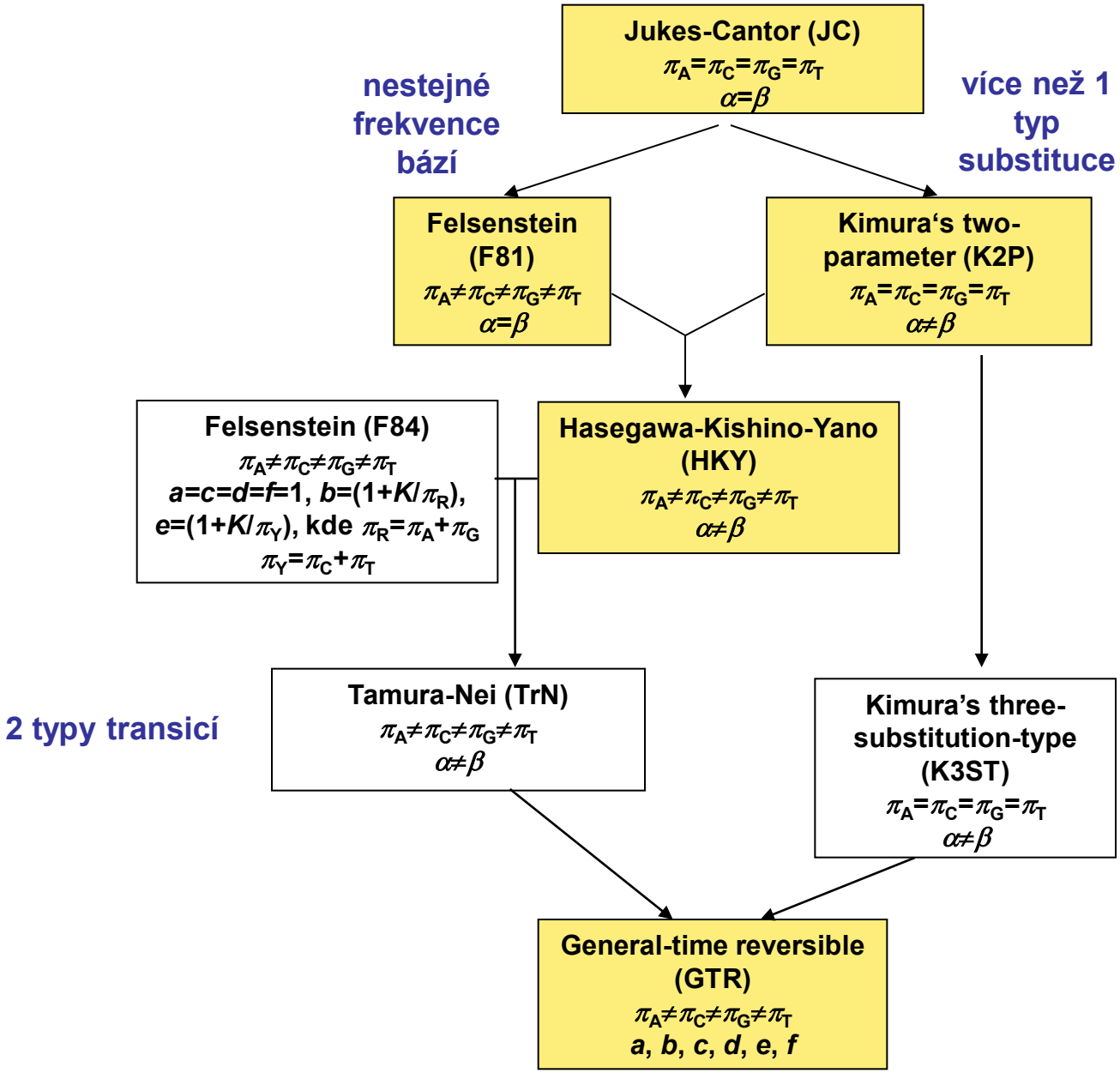
$$Q = \begin{pmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{pmatrix}$$

Jestliže $\pi A = \pi C = \pi G = \pi T$, F81 = JC

Hasegawa-Kishino-Yano (HKY): různé frekvence bází
transice ≠ transverze

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

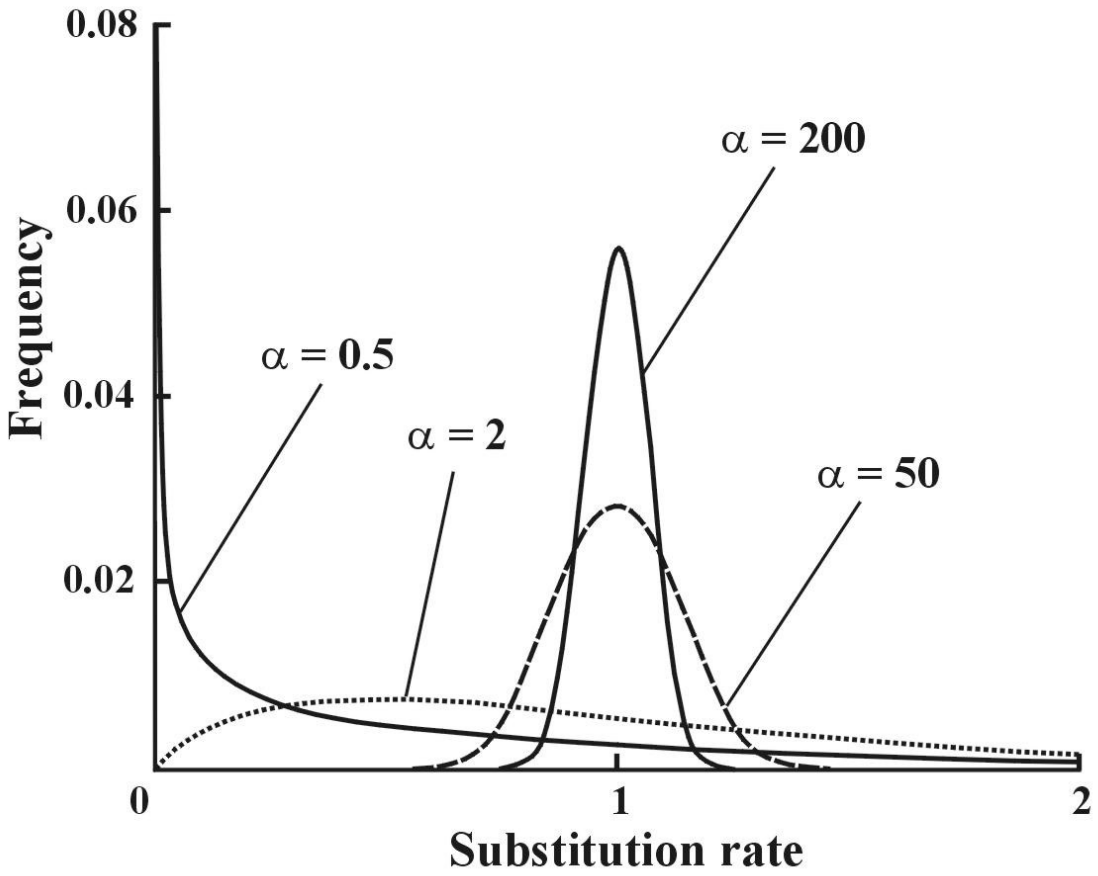
General time-reversible (GTR): různé frequence bází
různé frekvence všech substitucí


Jukes-Cantor (JC)
$\pi_A=\pi_C=\pi_G=\pi_T$
$\alpha=\beta$
nestejné frekvence bází
více než 1 typ substituce
Felsenstein (F81)
$\pi_A\neq\pi_C\neq\pi_G\neq\pi_T$
$\alpha=\beta$
Kimura's two-parameter (K2P)
$\pi_A=\pi_C=\pi_G=\pi_T$
$\alpha\neq\beta$
Felsenstein (F84)
$\pi_A\neq\pi_C\neq\pi_G\neq\pi_T$
$a=c=d=f=1$, $b=(1+K/\pi_R)$, $e=(1+K/\pi_Y)$, kde $\pi_R=\pi_A+\pi_G$
$\pi_Y=\pi_C+\pi_T$
Hasegawa-Kishino-Yano (HKY)
$\pi_A\neq\pi_C\neq\pi_G\neq\pi_T$
$\alpha\neq\beta$
2 typy transicí
Tamura-Nei (TrN)
$\pi_A\neq\pi_C\neq\pi_G\neq\pi_T$
$\alpha\neq\beta$
Kimura's three-substitution-type (K3ST)
$\pi_A=\pi_C=\pi_G=\pi_T$
$\alpha\neq\beta$
General-time reversible (GTR)
$\pi_A\neq\pi_C\neq\pi_G\neq\pi_T$
a, b, c, d, e, f

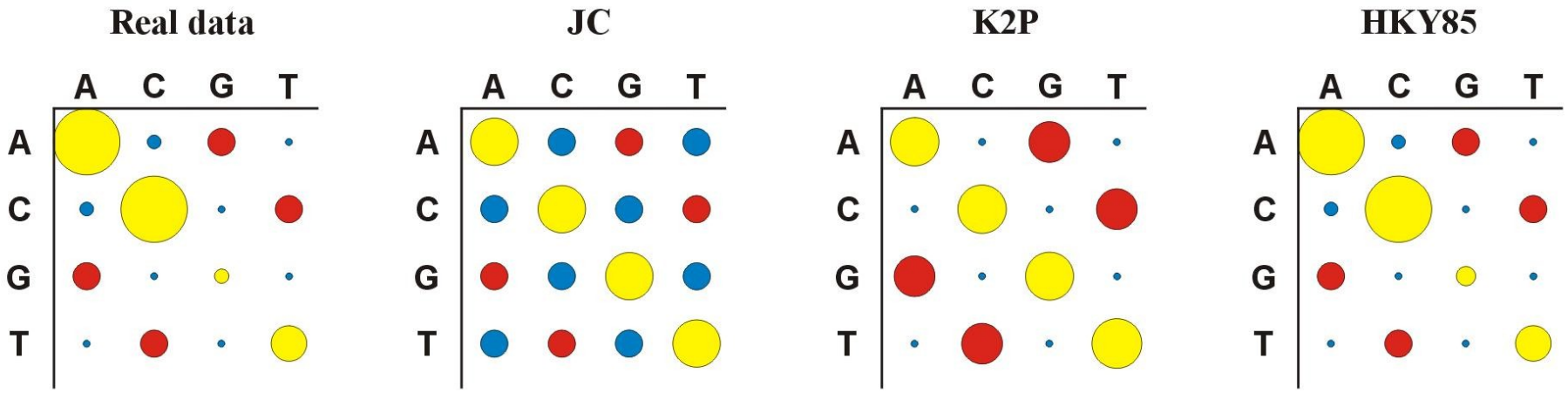
Heterogenita substitučních rychlostí v různých částech sekvence

Gama (Γ) rozdělení:

- parametr tvaru α (shape parameter)
- diskrétní gama model
- invariantní pozice
  → GTR+ Γ+I



Porovnání modelů:



Porovnání modelů:

Který model vybrat?

Likelihood ratio test (LRT):

nested models
$LR = 2(\ln L2 - \ln L1)$
Chi-square, p2 – p1 d.f.

Akaike information criterion (AIC):

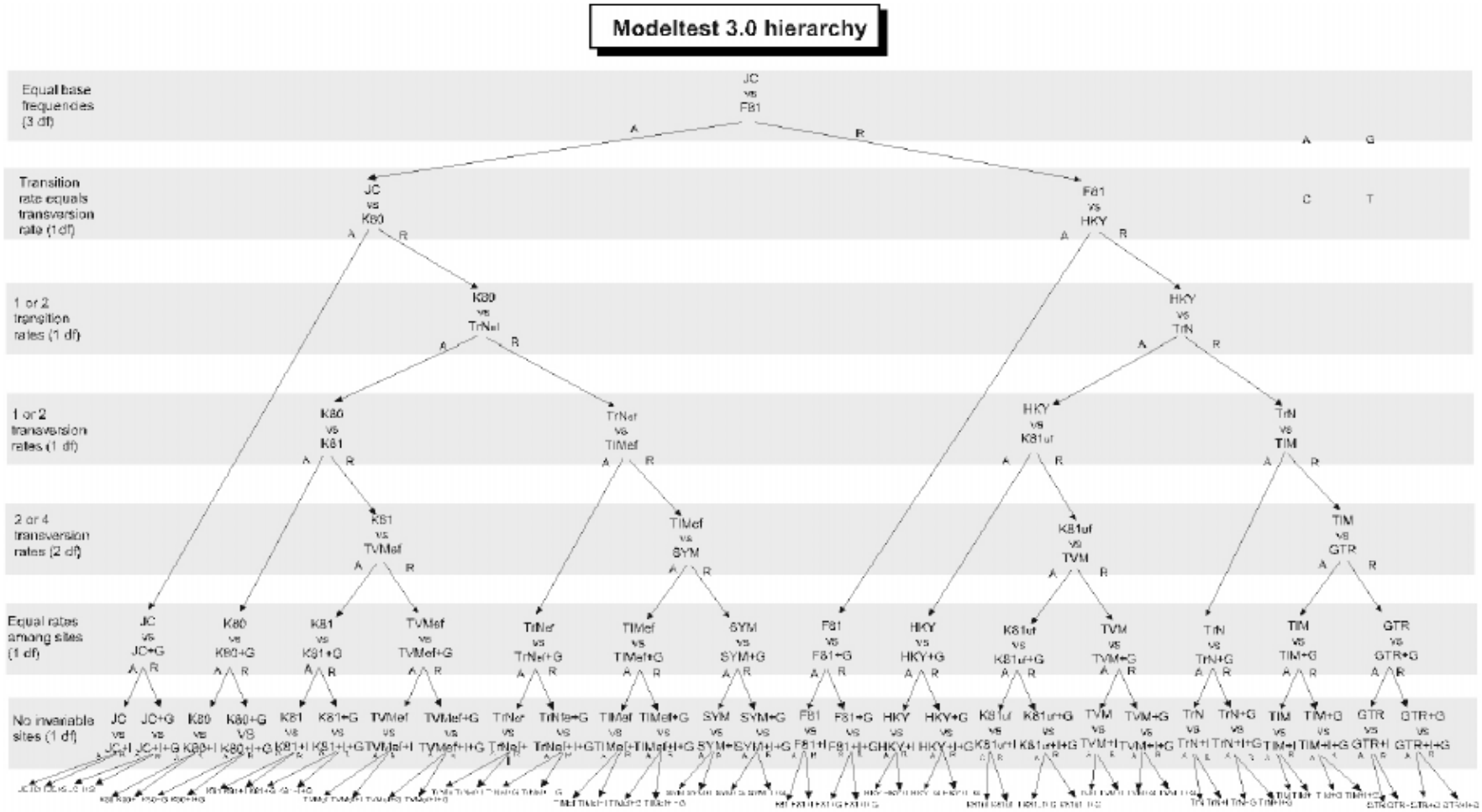
nonnested models
$AIC = -2\ln L + 2p$, where
p = number of free parameters
better model $\rightarrow$ smaller AIC

Bayesian information criterion (BIC):

nonested models
$BIC = -2\ln L + p\ln N$, where
N = sample size

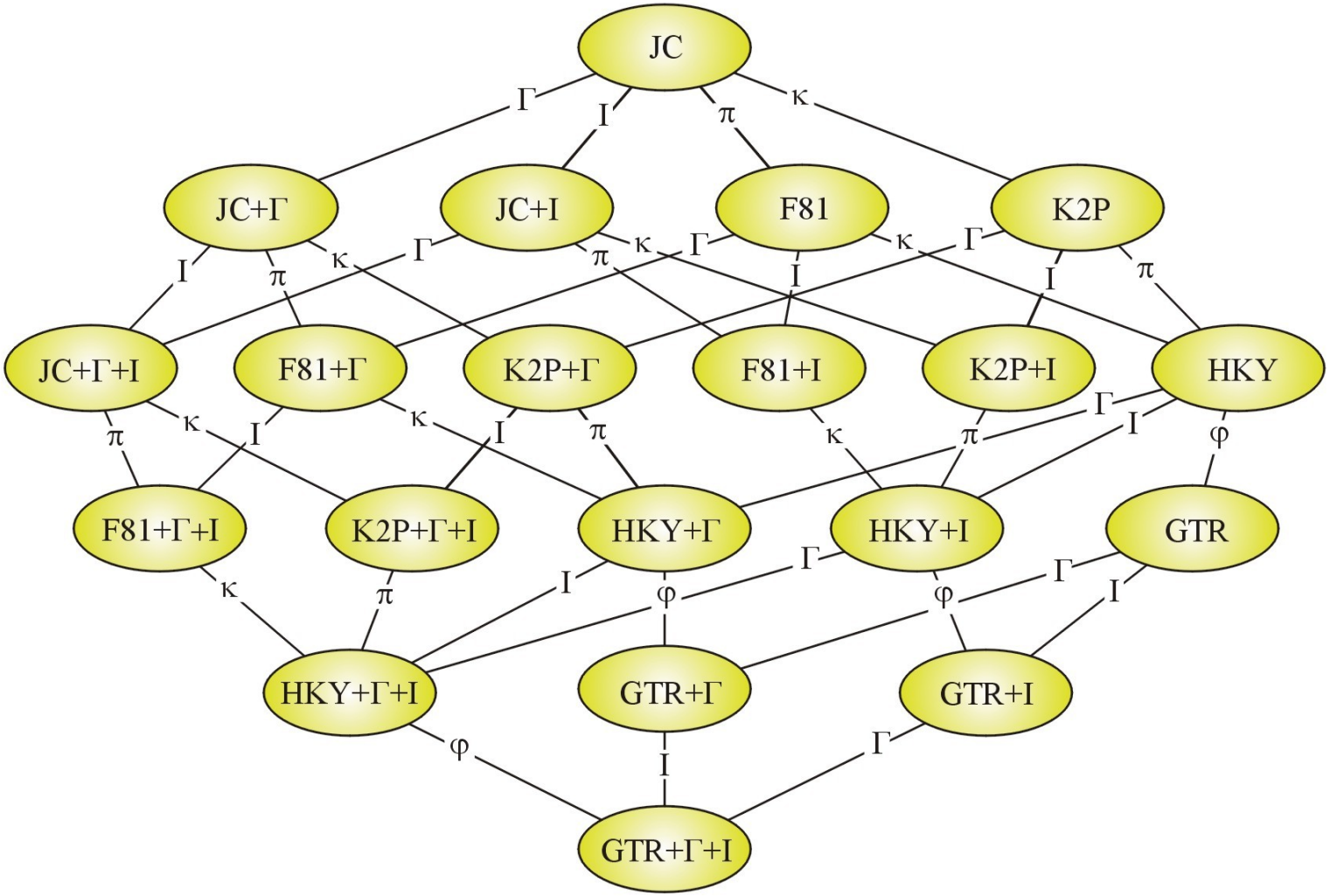
Porovnání modelů:

Hierarchický LRT – ModelTest (Crandall and Posada)

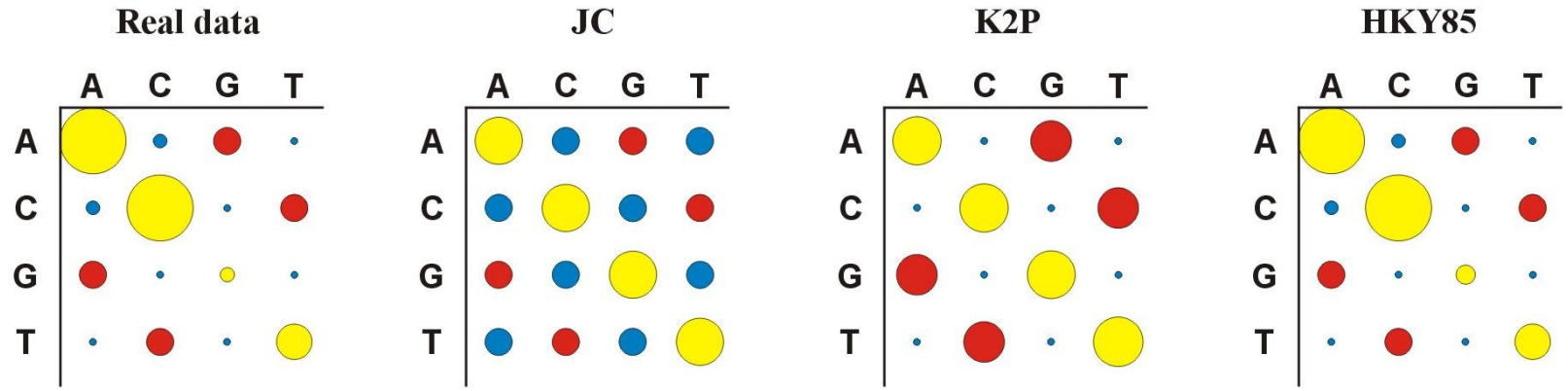


Porovnání modelů:

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

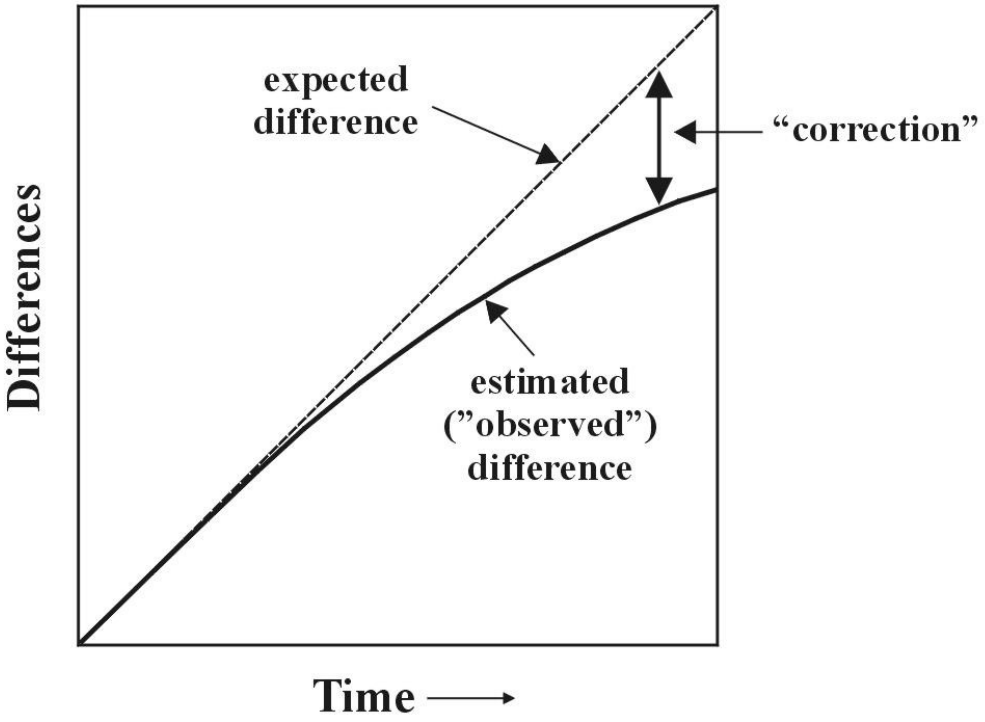
- výhoda: velmi rychlé a jednoduché (lze i na kalkulačce)

Distance

```
                1                  10                  20                  30
sekvence 1:  ACCCGTTAAGCTTAACGTACTTGGATCGAT
sekvence 2:  ACCCGTTAGGCTTAATGTACGTGGATCGAT
```

p-distance: $p = k/n = 3/30 = 0.10$

problém saturace:



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 = 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 a Q jako K2P
GTR	$d_{xy} = -\text{stopa}\ln\left\|\Pi^{-1}\mathbf{F}_{xy}\right\|$	Π = diagonální matice průměrných četností bází v sekvencích X a Y

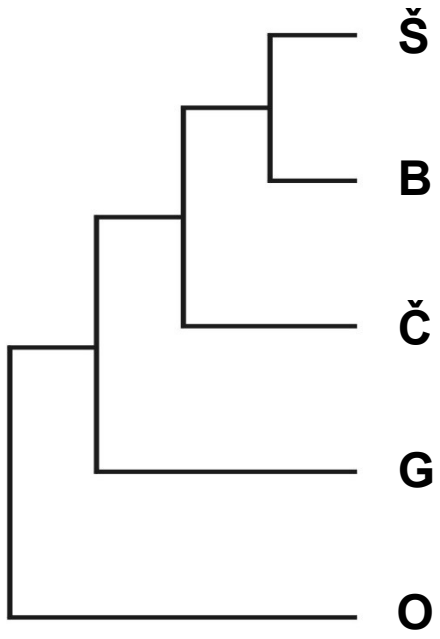
Příklad v MEGA

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(ŠB-*k*) = [*d*(B-*k*)+*d*(Š-*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	--



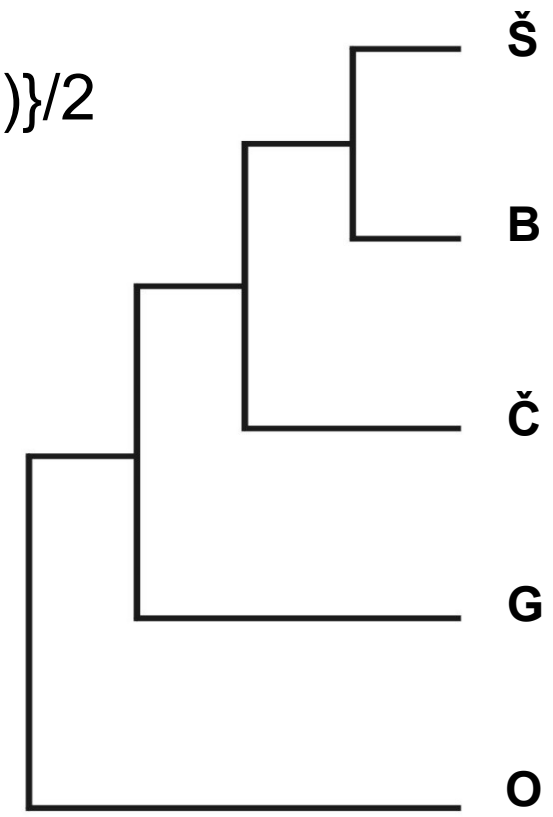
Shluková analýza - UPGMA

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

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

single-linkage

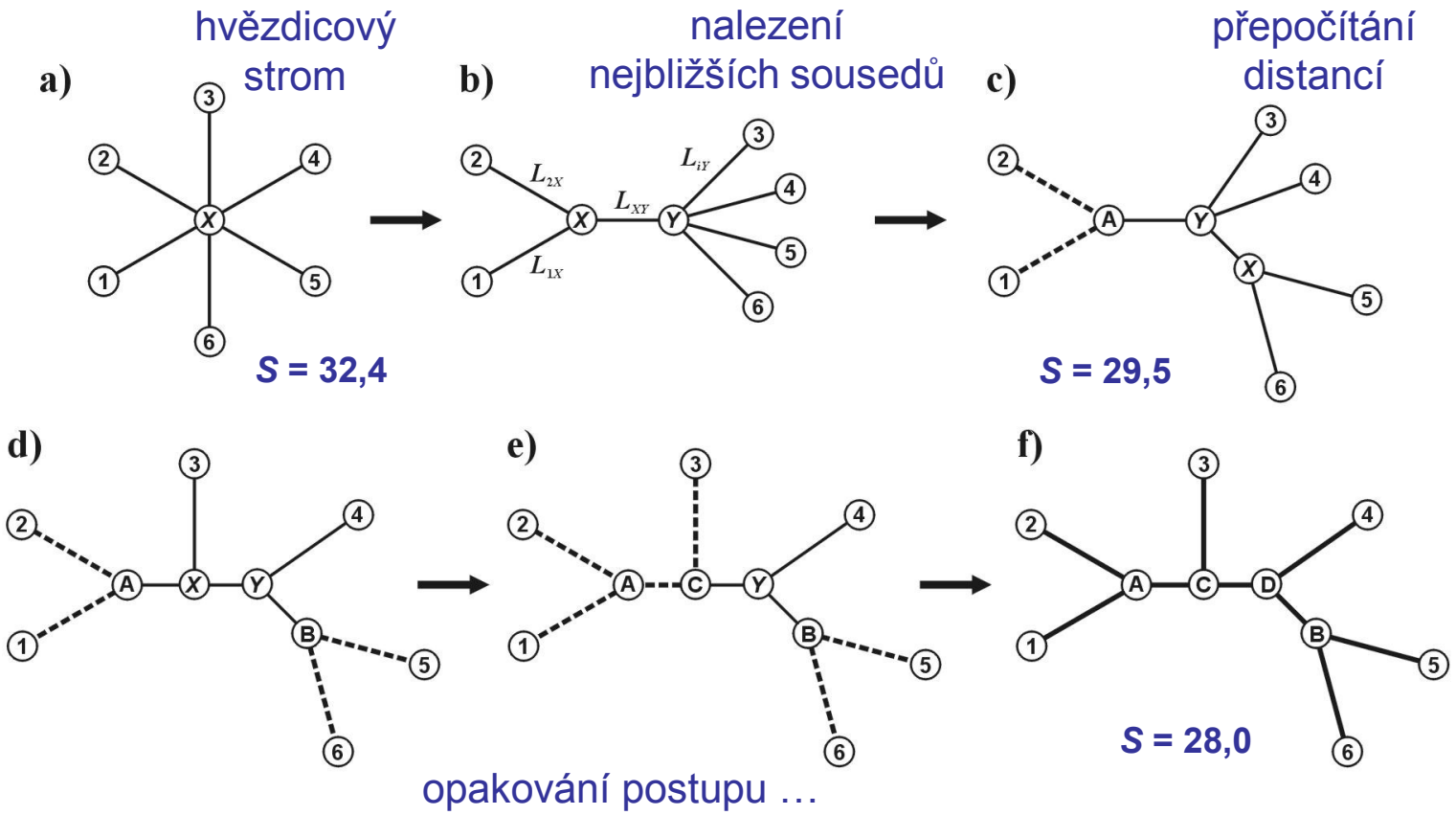
complete-linkage



Spojení sousedů (neighbor-joining, NJ)

- Algoritmická metoda
- Princip minimální evoluce $\rightarrow$ minimalizuje součet délek větví *S*
- Každý pár uzlů adjustován na základě divergence od ostatních
- Konstrukce jediného aditivního stromu

Spojení sousedů (neighbor-joining, NJ)



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

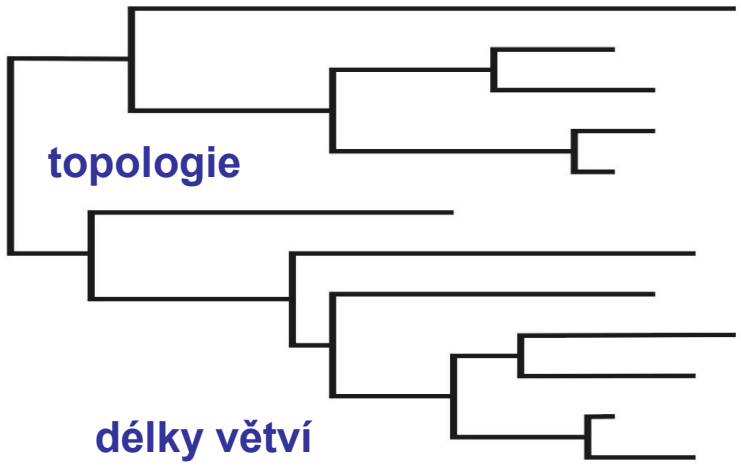
5. nelze kombinovat různé distanční matice

(3) Maximální věrohodnost (maximum likelihood)

data:

```
1    TCAAAAATGGCTTTATTCGCTTAATGCCGTTAACCCTTGCGGGGGCCATG
2    TCCGTGATGGATTTATTTCCGCAATGCCTGTCATCTTATTCTCAAGTATC
3    TTCGTGATGGATTTATTGCAGGTATGCCAGTCATCCTTTTCTCATCTATC
4    TTCGTGACGGGTTTATCTCGGCAATGCCGGTCATCCTATTTTCGAGTATT
```

strom:



evoluční model

= hypotéza

$L = P(\mathrm{D} \mid \mathrm{H})$,

kde D = matice dat

H = τ (topologie),
ν (délky větví),
θ (model)

Věrohodnostní funkce: jaká je pravděpodobnost získání daných dat při dané hypotéze?

(4) Bayesovská analýza

aposteriorní pravděpodobnost (posterior probability)
= pr. platnosti hypotézy při získaných datech: $P(H|D)$

a.p. je funkcí věrohodnosti $P(D|H)$ a apriorní pravděpodobnosti (prior prob.)

prior vyjadřuje náš apriorní předpoklad nebo znalost

Aposteriorní pravděpodobnost je dána Bayesovou rovnicí:

věrohodnost prior

$$P(H|D) = \frac{P(D|H) \times P(H)}{\Sigma[P(D|H_i) \times P(H_i)]}$$

suma čitatelů pro všechny alternativní hypotézy

Fylogenetické programy:

alignment:

ClustalX *http://inn-prot.weizmann.ac.il/software/ClustalX.html*
BioEdit
AliView

PAUP*
PHYLIP
MEGA ... MP, NJ, ML
RAxML ... ML
MrBayes ... BA

http://evolution.genetics.washington.edu/phylip/software.html#methods

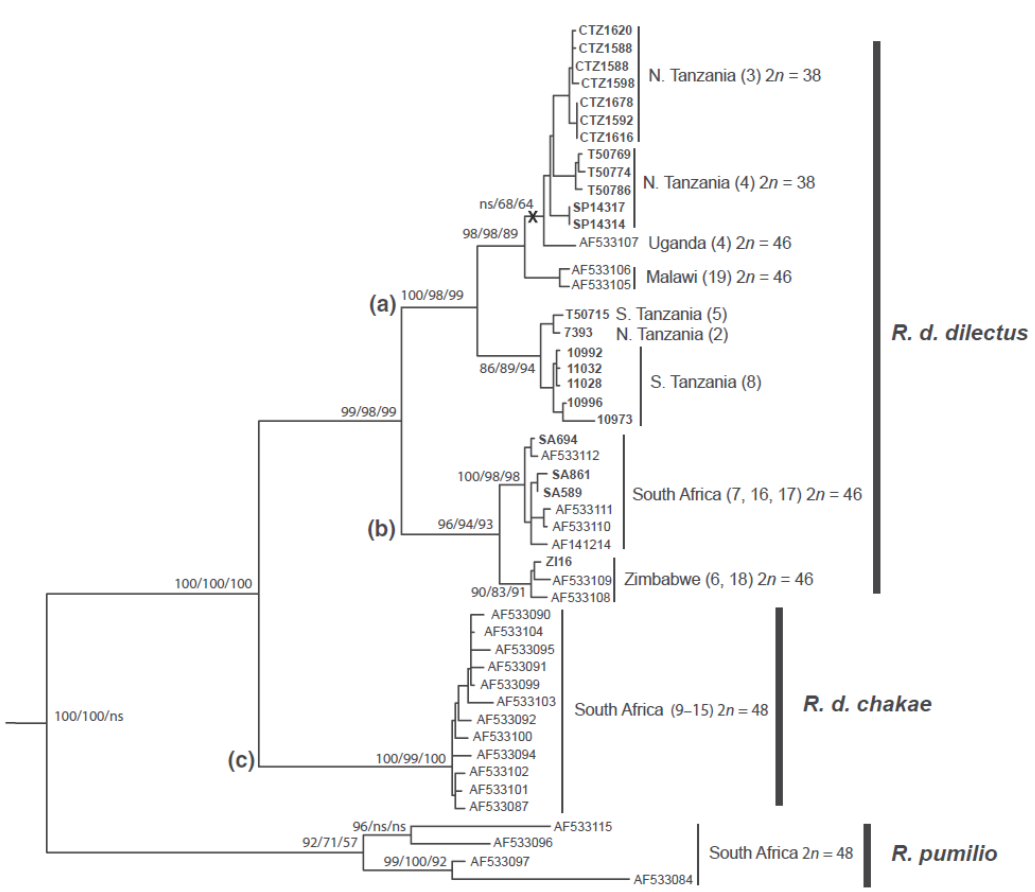
práce se stromy:

TreeView *http://taxonomy.zoology.gla.ac.uk/rod/treeview.html*
FigTree

Příklad

- určit druh zvířete, které jsme osekvenovali na mtDNA (BLAST)
- určit jeho fylogenetickou pozici v rámci rodu (alignment, NJ tree)

Rhabdomys


CTZ1620
CTZ1588
CTZ1588
CTZ1598
N. Tanzania (3) 2n = 38
CTZ1678
CTZ1592
CTZ1616
T50769
T50774
T50786
N. Tanzania (4) 2n = 38
SP14317
SP14314
ns/68/64
98/98/89
AF533107
Uganda (4) 2n = 46
AF533106
AF533105
Malawi (19) 2n = 46
(a)
100/98/99
T50715
S. Tanzania (5)
7393
N. Tanzania (2)
10992
11032
11028
S. Tanzania (8)
86/89/94
10996
10973
99/98/99
R. d. dilectus
SA694
AF533112
100/98/98
SA861
SA589
South Africa (7, 16, 17) 2n = 46
AF533111
AF533110
(b)
96/94/93
AF141214
ZI16
AF533109
Zimbabwe (6, 18) 2n = 46
AF533108
90/83/91
100/100/100
AF533090
AF533104
AF533095
AF533091
AF533099
AF533103
South Africa (9–15) 2n = 48
R. d. chakae
AF533092
AF533100
AF533094
(c)
100/99/100
AF533102
AF533101
AF533087
100/100/ns
96/ns/ns
AF533115
92/71/57
AF533096
99/100/92
AF533097
South Africa 2n = 48
R. pumilio
AF533084