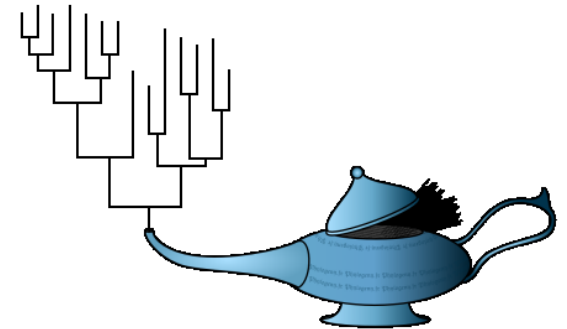
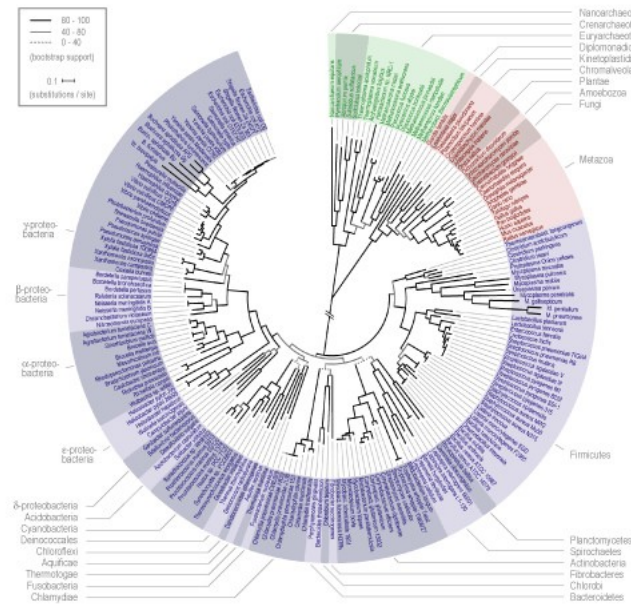
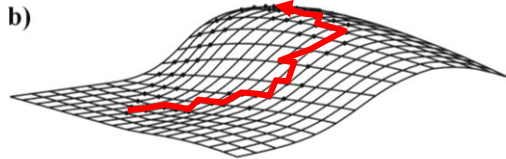
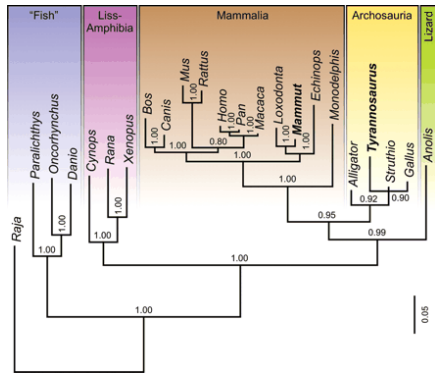


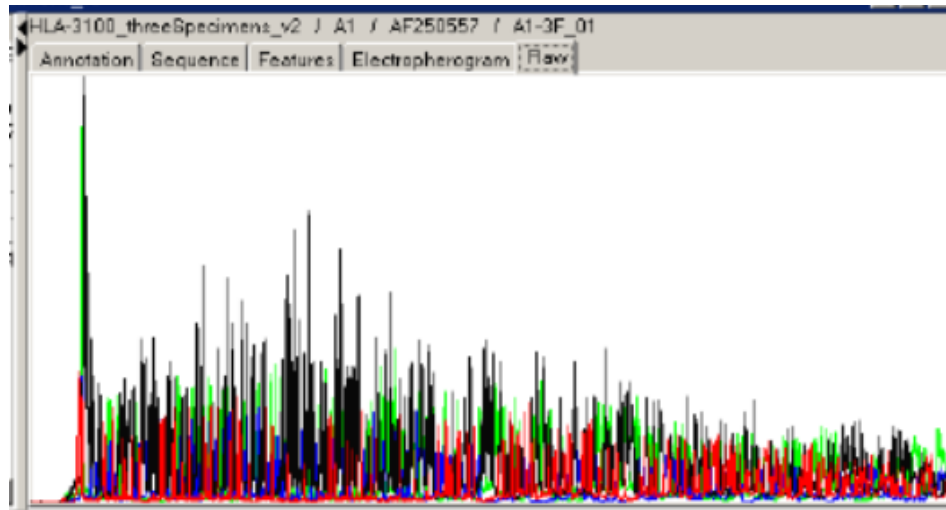
# ANALÝZA DNA SEKVENCÍ



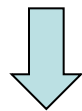
## Kde získat sekvence?

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

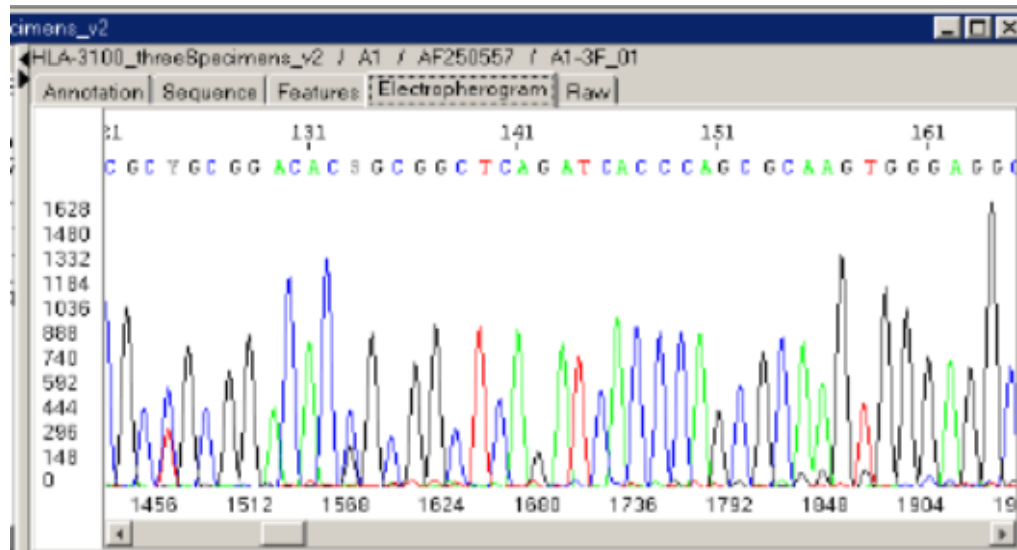
# Editace sekvencí – Sangerovo sekvenování



Raw data  
(.ab1 file)

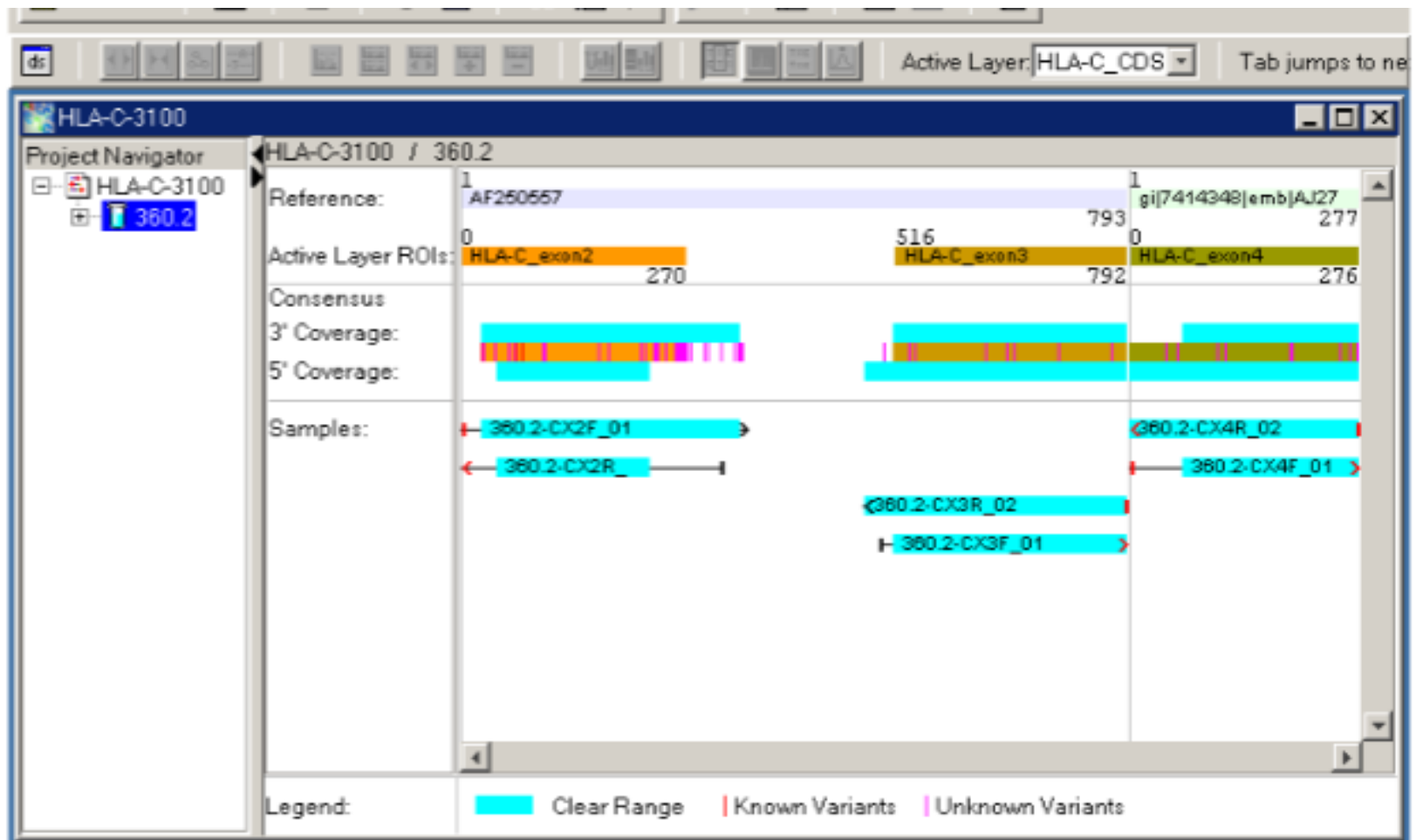


„basecalling“

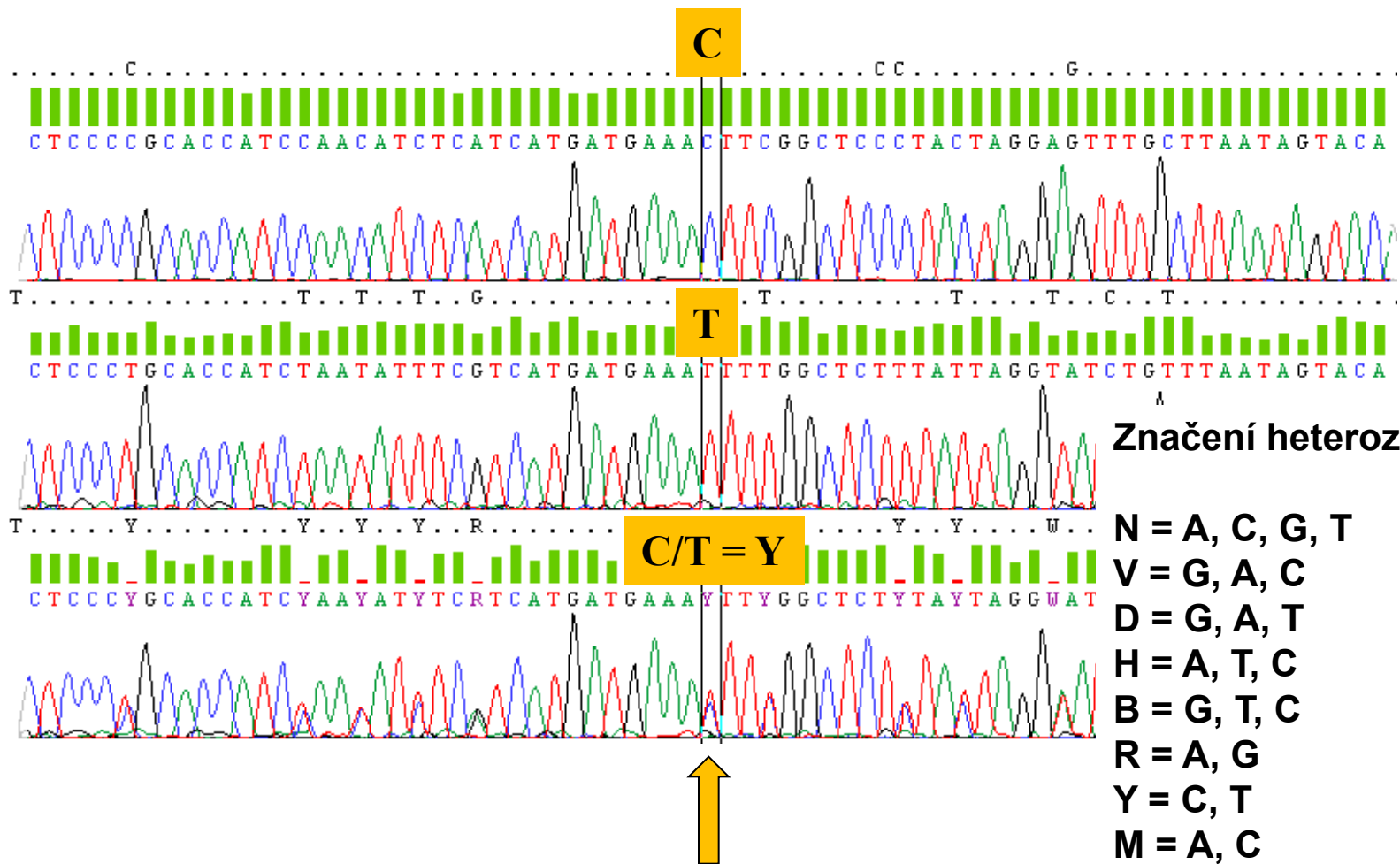


Editovaná sekvence

# „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

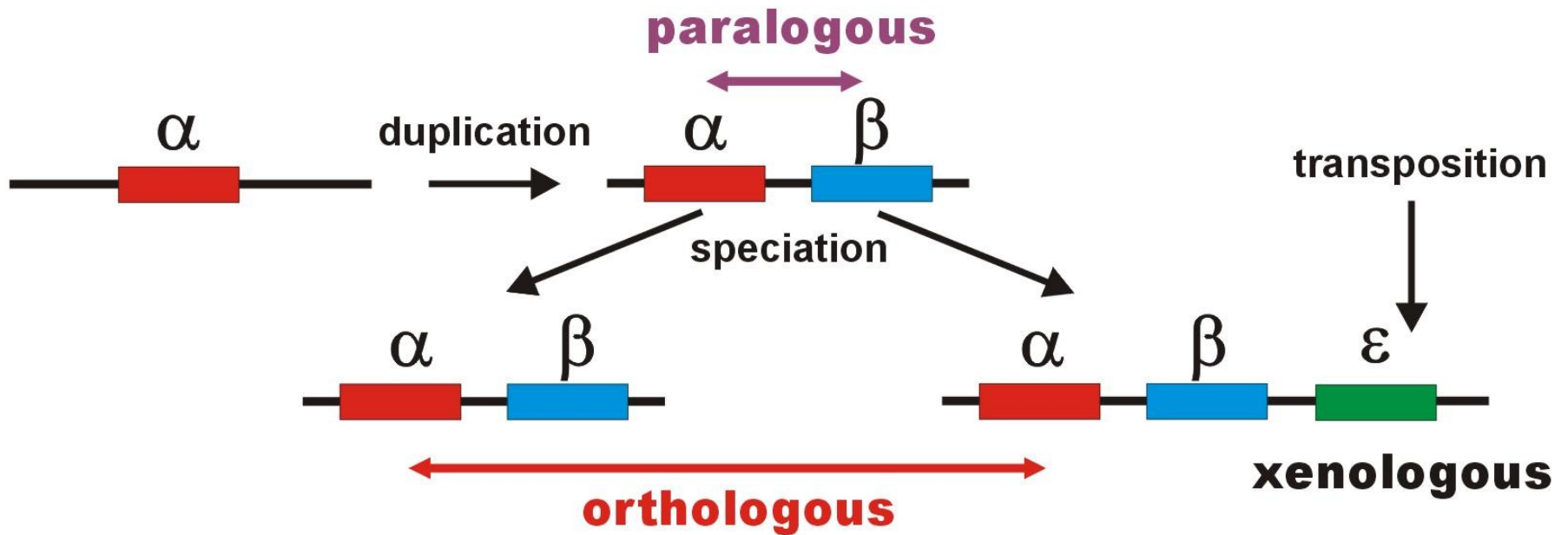
báze = stav znaku

P\_troglod ATGACCCCGA CACGCAAAATTAACCCACTAATAAA



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 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

```
ATGACCCCAATACGCAAATTAACCCCTAATAAAATTAATTAACCACTCATTCATCGACCTCCCCACCC
CATCCAACATCTCCGCATGATGAAACTTCGGCTCACTCCTTGGCGCCTGCCTGATCCTCCAAATCACCAC
AGGACTATTCCCTAGCCATACTACTCACCAGACGCCTCAACCGCCTTTTCATCAATCGCCCACATCACT
CGAGACGTAAATTATGGCTGAATCATCCGCTACCTTCACGCCAATGGCGCCTCAATATTCTTTATCTGCC
TCTTCCTACACATCGGGCGAGGCCTATATTACGGATCATTTCTCTACTCAGAAACCTGAAACATCGGCAT
```

...

>P\_troglod

```
ATGACCCCGACACGCAAATTAACCCACTAATAAAATTAATTAATCACTCATTTATCGACCTCCCCACCC
CATCCAACATTTCCGCATGATGGAACTTCGGCTCACTTCTCGGCGCCTGCCTAATCCTTCAAATTACCAC
AGGATTATTCCCTAGCTATACTACTCACCAGACGCCTCAACCGCCTTCTCGTCGATCGCCCACATCACC
CGAGACGTAAACTATGGTTGGATCATCCGCTACCTCCACGCTAACGGCGCCTCAATATTTTTTATCTGCC
TCTTCCTACACATCGGCCGAGGTCTATATTACGGCTCATTTCTCTACCTAGAAACCTGAAACATTGGCAT
```

...

>P\_paniscus

```
ATGACCCCAACACGCAAATCAACCCACTAATAAAATTAATTAATCACTCATTTATCGACCTCCCCACCC
CATCCAATATTTCCACATGATGAAACTTCGGCTCACTTCTCGGCGCCTGCCTAATCCTTCAAATCACCAC
AGGACTATTCCCTAGCTATACTACTCACCAGACGCCTCAACCGCCTTCTCATCGATCGCCCACATTACC
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 ttataactaat acatcaaatt aatgctttaa agacatatct gtgttatctg acatacacca
301 tacagtcata aactcttctc ttccatatga ctatcccctt ccccathttgg tctattaatc
361 taccatcctc cgtgaaacca acaaccgcgc caccaatgcc cctcttctcg ctccggggccc
421 attaaacttg ggggtagcta aactgaaact ttatcagaca tctggttctt acttcagggc
481 catcaaatgc gttatcgccc atacgttccc cttaaataag acatctcgat ggtatcgggt
541 ctaatcagcc catgaccaac ataactgtgg tgtcatgcat ttggtathtt tttathttgg
601 cctactttca tcaacatagc cgtcaaggca tgaaaggaca gcacacagtc tagacgcacc
661 tacgggtgaag aatcattagt ccgcaaaacc caatcaccta aggctaatta ttcatgcttg
721 ttagacataa atgctactca ataccaaatt ttaactctcc aaacccccca accccctcct
781 cttaatgcca aacccccaaa acactaagaa cttgaaagac atatattatt aactatcaaa
841 ccctatgtcc tgatcgattc tagtagttcc caaatatga ctcatathtt agtacttgta
901 aaaathttac aaaatcatgc tccgtgaacc aaaactctaa tcacactcta ttacgcaata
961 aatattaaca agttaatgta gcttaataac aaagcaaagc actgaaaatg cttagatgga
1021 taathttatc 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   ATGACCCCGACACGCAAAATTAACCCACTAATAAAAATTAATTAATCACTC
P_paniscus  ATGACCCCAACACGCAAAATCAACCCACTAATAAAAATTAATTAATCACTC
H_sapiens   ATGACCCCAATACGCAAAATTAACCCCTAATAAAAATTAATTAACCACTC
G_gorilla   ATGACCCCTATACGCAAAACTAACCCACTAGCAAAACTAATTAACCACTC
P_pygmaeus  ATGACCCCAATACGCAAAACCAACCCACTAATAAAAATTAATTAACCACTC
H_lar       ATGACCCCCCTGCGCAAAACTAACCCACTAATAAAAATAATCAACCACTC

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

end;
```

# Formáty souborů

## Clustal:

```
P_troglod ATGACCCCGACACGCAAAATTAACCCACTAATAAAAATTAATTAATCACTCATTATCGAC
P_paniscus ATGACCCCAACACGCAAAATCAACCCACTAATAAAAATTAATTAATCACTCATTATCGAC
H_sapiens ATGACCCCAATACGCAAAATTAACCCCTAATAAAAATTAATTAACCACTCATTCATCGAC
G_gorilla ATGACCCCTATACGCAAAACTAACCCACTAGCAAAACTAATTAACCACTCATTCATTGAC
P_pygmaeus ATGACCCCAATACGCAAAACCAACCCACTAATAAAAATTAATTAACCACTCACTCATCGAC
H_lar ATGACCCCCCTGCGCAAAACTAACCCACTAATAAAAATAATCAACCACTCACTTATCGAC
*****          *****  *****  ***  *****  *****  **  *****  *  **  ***
```

```
P_troglod CTCCCCACCCCATCCAACATTTCCGCATGATGGAAC TTCGGCTCACTTCTCGGCGCCTGC
P_paniscus CTCCCCACCCCATCCAATATTTCCACATGATGAAAC TTCGGCTCACTTCTCGGCGCCTGC
H_sapiens CTCCCCACCCCATCCAACATCTCCGCATGATGAAAC TTCGGCTCACTCCTTGGCGCCTGC
G_gorilla CTCCCTACCCCGTCCAACATCTCCACATGATGAAAC TTCGGCTCACTCCTTGGTGCCTGC
P_pygmaeus CTCCCCACCCCATCAAACATCTCTGCATGATGGAAC TTCGGCTCACTTCTAGGCGCCTGC
H_lar CTTCCAGCCCCATCCAACATTTCTATATGATGAAAC TTTGGTTCCTAGGCGCCTGC
** **  ****  **  **  **  **  *****  *****  **  *****  **  **  *****
```

# Seřazení sekvencí (alignment)

Sekvence 1 TTGTACGACGG  
 Sekvence 2 TTGTACGACG

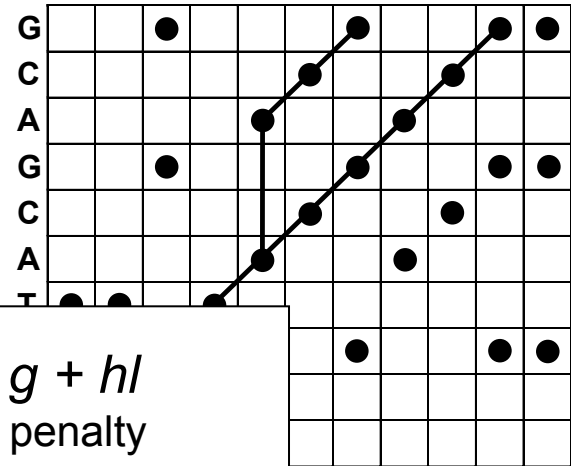
TTGTACGACGG    TTGT---ACGACGG  
 | | | | | | | |    | | |    | | |  
 TTGTACGACG    TTGTACGACG

gap penalty

Sekvence 1 ACTTGCTTTC  
 Sekvence 2 ACGTGCTGCTC

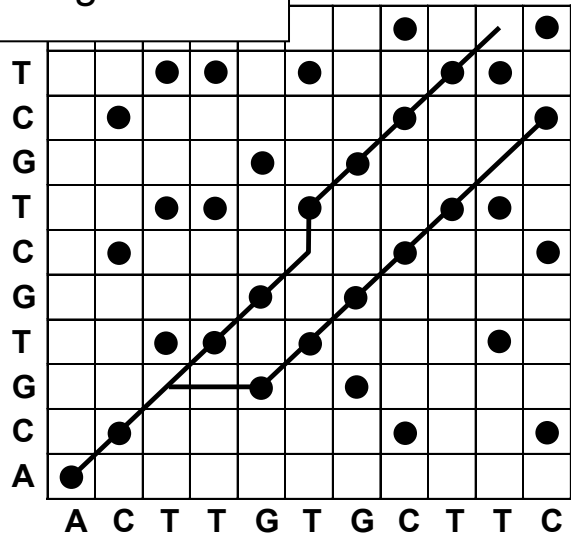
Path 1    ACTTG-TGCTTTC  
 | | | | | | | |  
 ACGTGCTGCTC

Path 2    ACTTGCTTTC  
 | |    | | | | | |  
 AC--GTGCTGCTC



$$GP = g + hl$$

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



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

**BioEdit**

**konverze formátů (zejména pro  
fylogenetickou analýzu):**

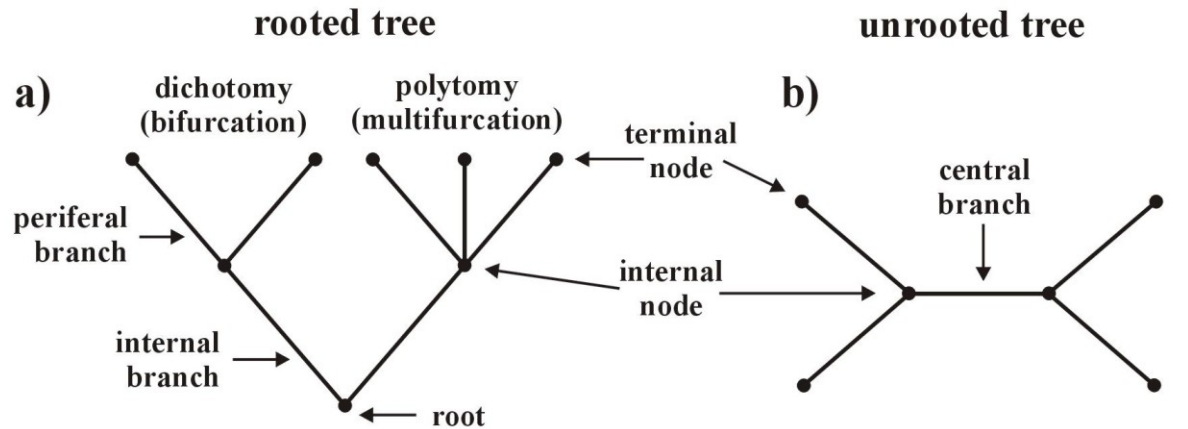
ALTER (<http://sing.ei.uvigo.es/ALTER/>)

# 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



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

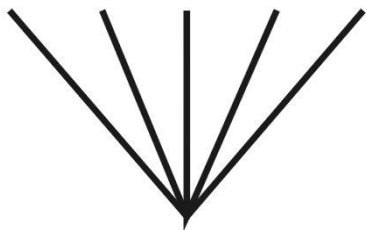
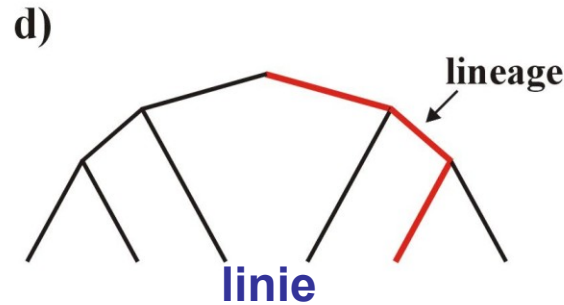
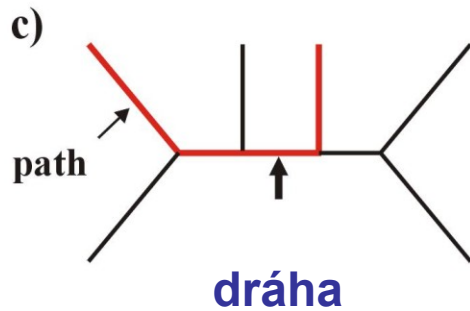
větve (branches, edges)  
vnější, vnitřní, centrální

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

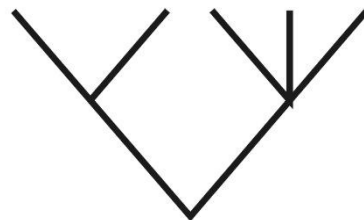
dichotomie, polytomie

OTU, HTU

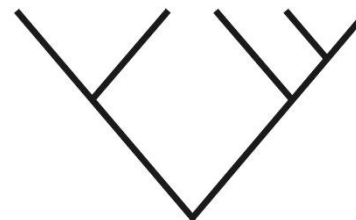
topologie



star tree



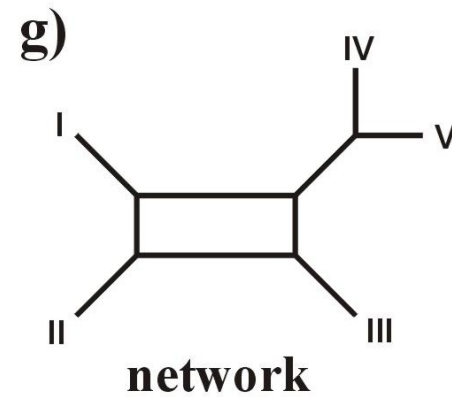
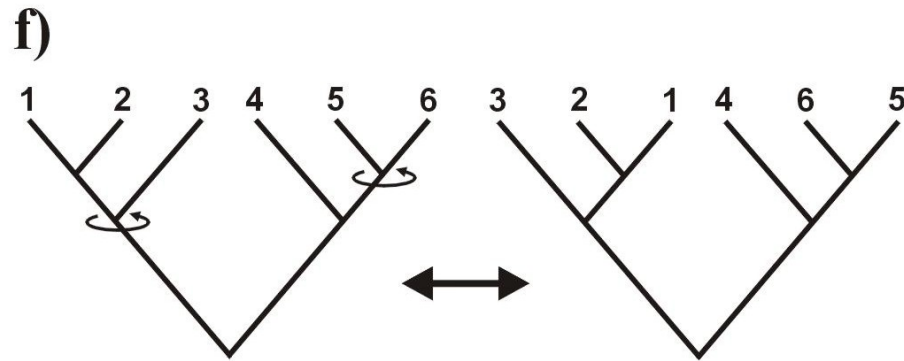
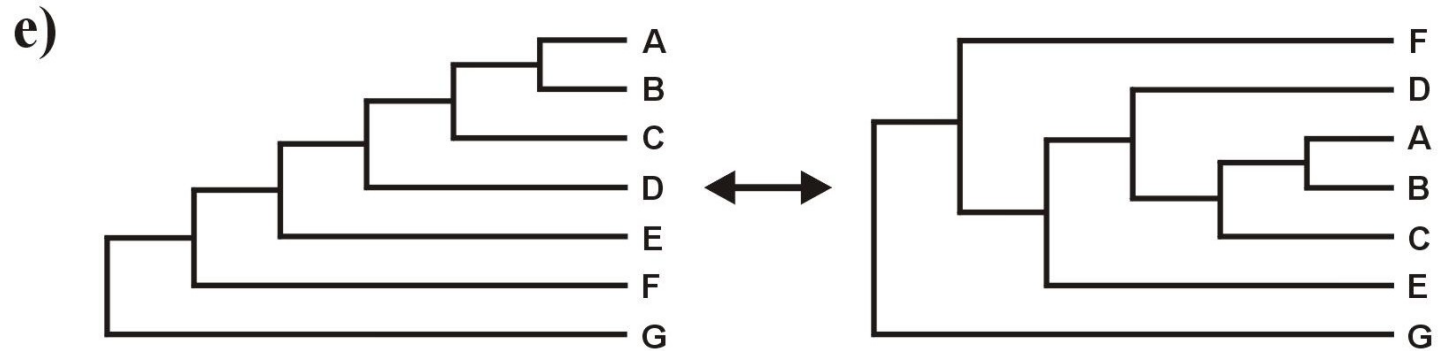
partly resolved



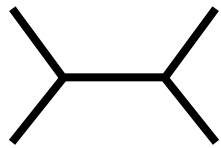
fully resolved

# Definice základních pojmů

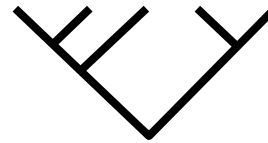
topologie:



# Kolik existuje stromů?



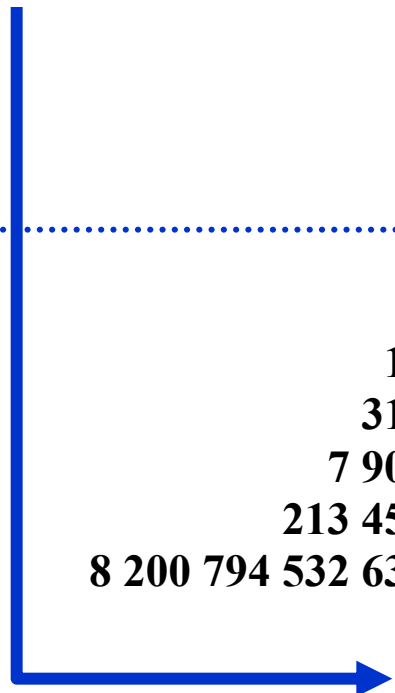
$$\frac{(2n-5)!}{2^{n-3}(n-3)!}$$



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

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	$4,9518 \times 10^{38}$
40	$4,9518 \times 10^{38}$	$1,00986 \times 10^{57}$
50	$1,00986 \times 10^{57}$	$2,75292 \times 10^{76}$

4 více než elektronů ve viditelném vesmíru (Eddingtonovo číslo)



# Rozdělení metod

## Typy dat

distance

znaky

**Metody konstrukce stromů**

algorithms

kritérium  
optimality

<ul style="list-style-type: none"><li>• UPGMA</li><li>• neighbor-joining</li></ul>	
<ul style="list-style-type: none"><li>• Fitch-Margoliash</li><li>• minimum evolution</li></ul>	<ul style="list-style-type: none"><li>• maximum parsimony</li><li>• maximum likelihood</li><li>• Bayesian a.</li></ul>

# 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?

# (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

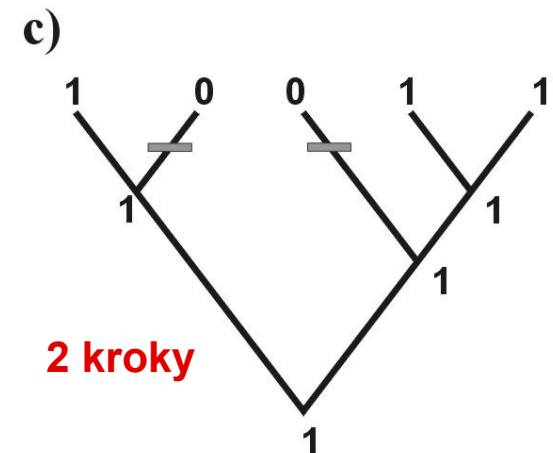
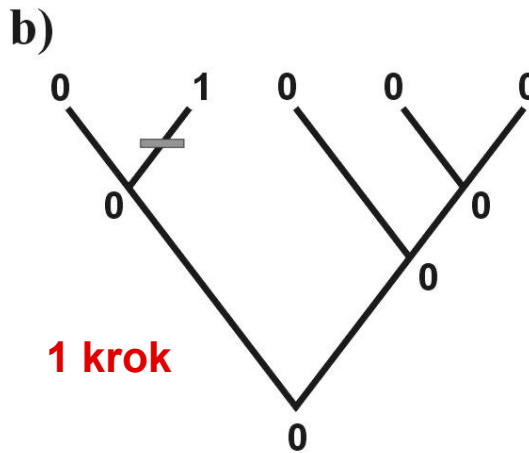
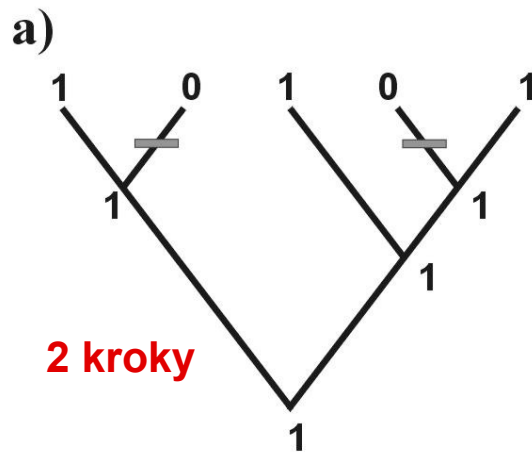
William of Occam (c. 1285 - c. 1349):  
*Occamova břitva*

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

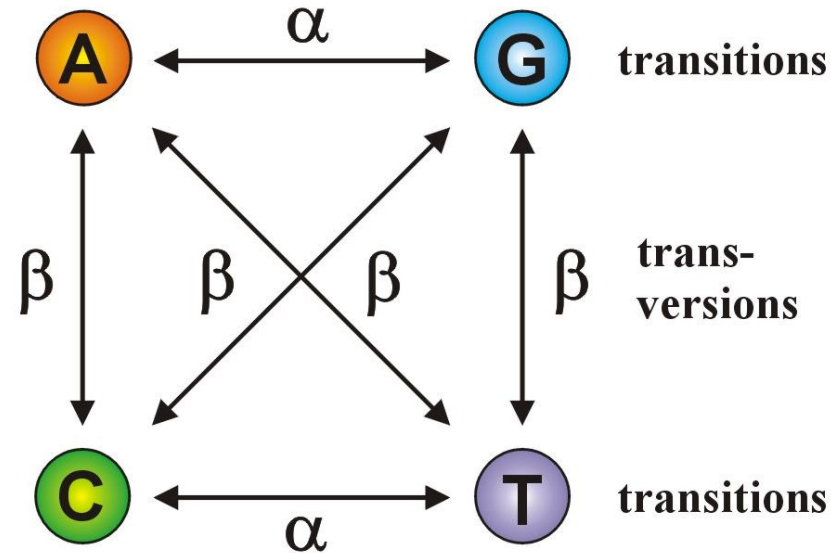
		Báze po substituci			
		A	C	G	T
Původní báze	A	$-\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	C	$\frac{1}{4}$	$-\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	G	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{3}{4}$	$\frac{1}{4}$
	T	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{3}{4}$

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

**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

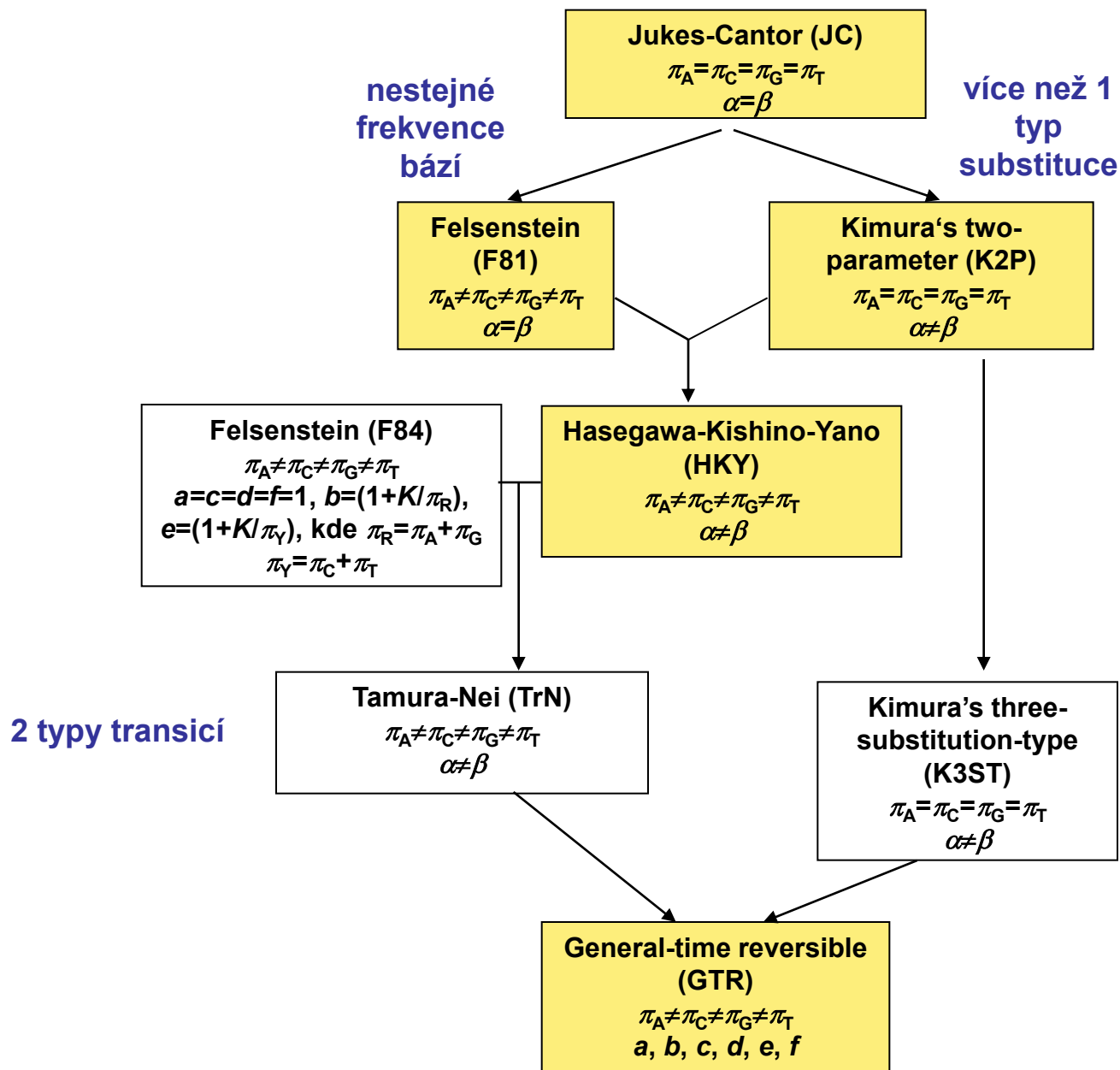
## Hasegawa-Kishino-Yano (HKY):

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):

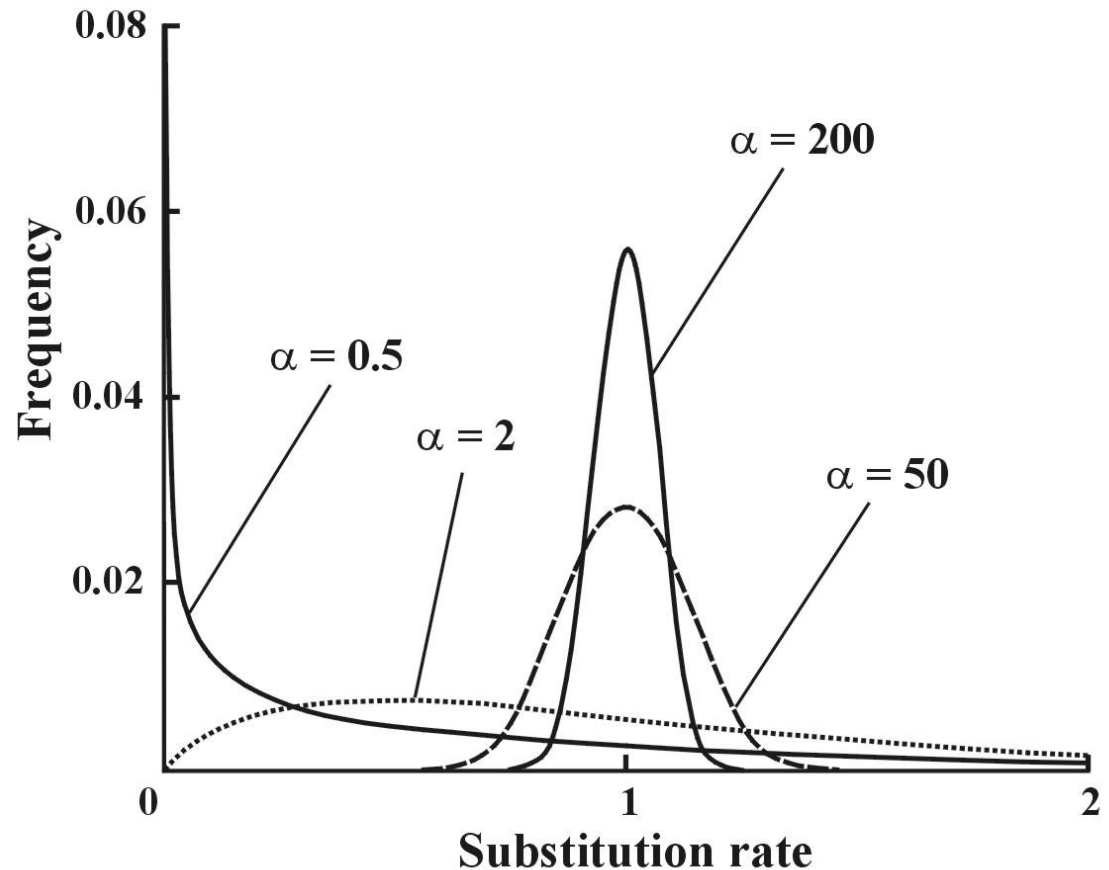
různé frekvence bází  
různé frekvence všech substitucí



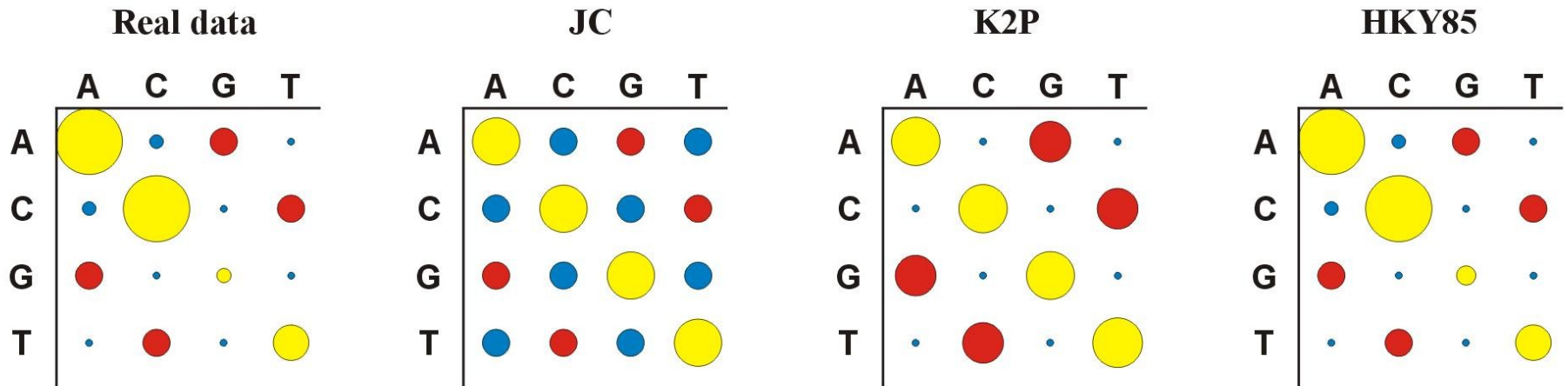
# Heterogenita substitučních rychlostí v různých částech sekvence

## Gama ( $\Gamma$ ) rozdělení:

- parametr tvaru  $\alpha$  (shape parameter)
- diskrétní gama model
- invariantní pozice  
→ GTR+ $\Gamma$ +I



# Porovnání modelů:



# Porovnání modelů:

Který model vybrat?

**Likelihood ratio test (LRT):**

nested models

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

Chi-square,  $p_2 - p_1$  d.f.

**Akaike information criterion (AIC):**

nonnested models

$$AIC = -2\ln L + 2p, \text{ where}$$

$p$  = number of free parameters

better model  $\rightarrow$  smaller  $AIC$

**Bayesian information criterion (BIC):** nonnested models

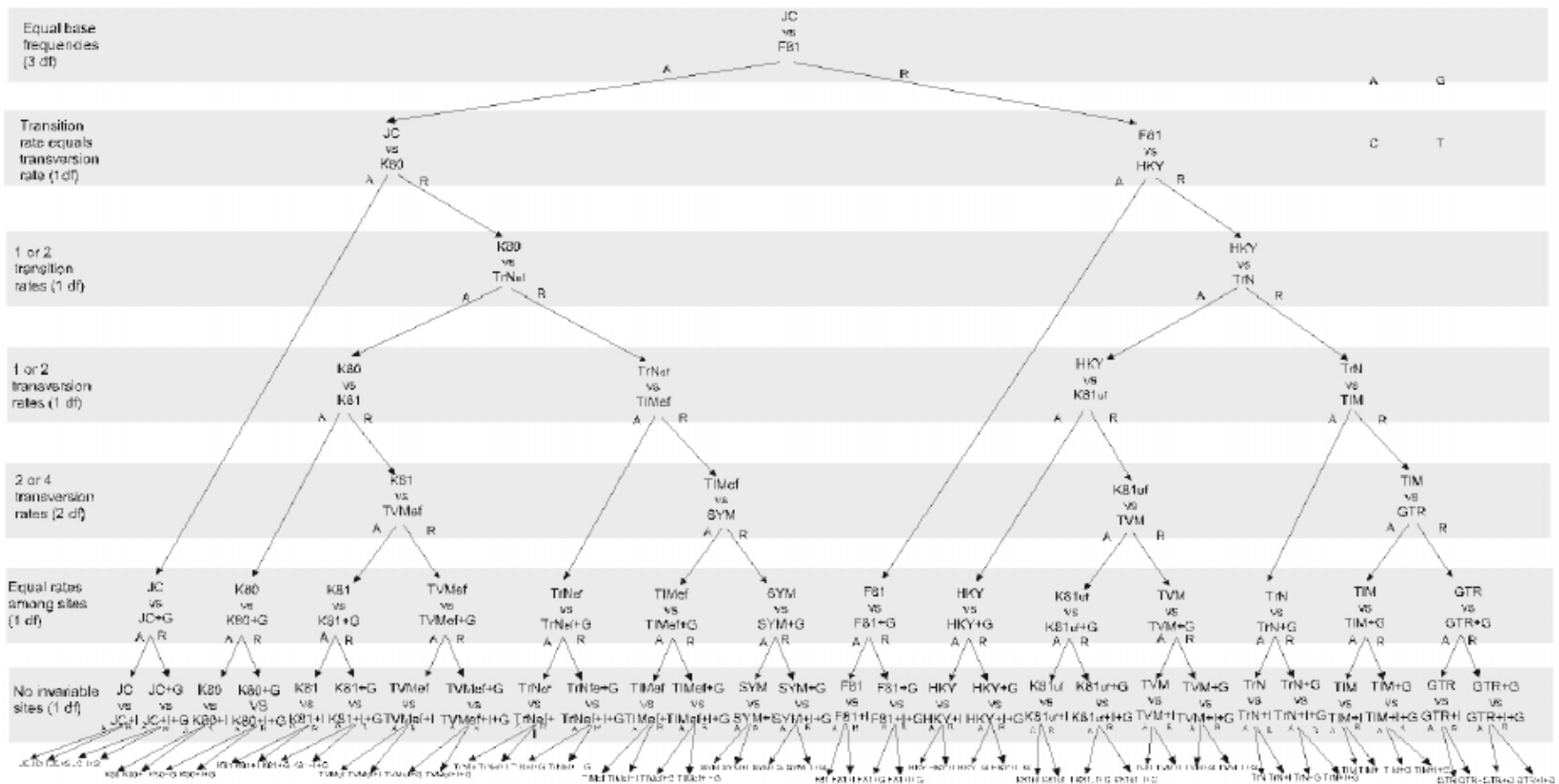
$$BIC = -2\ln L + p\ln N, \text{ where}$$

$N$  = sample size

# Porovnání modelů:

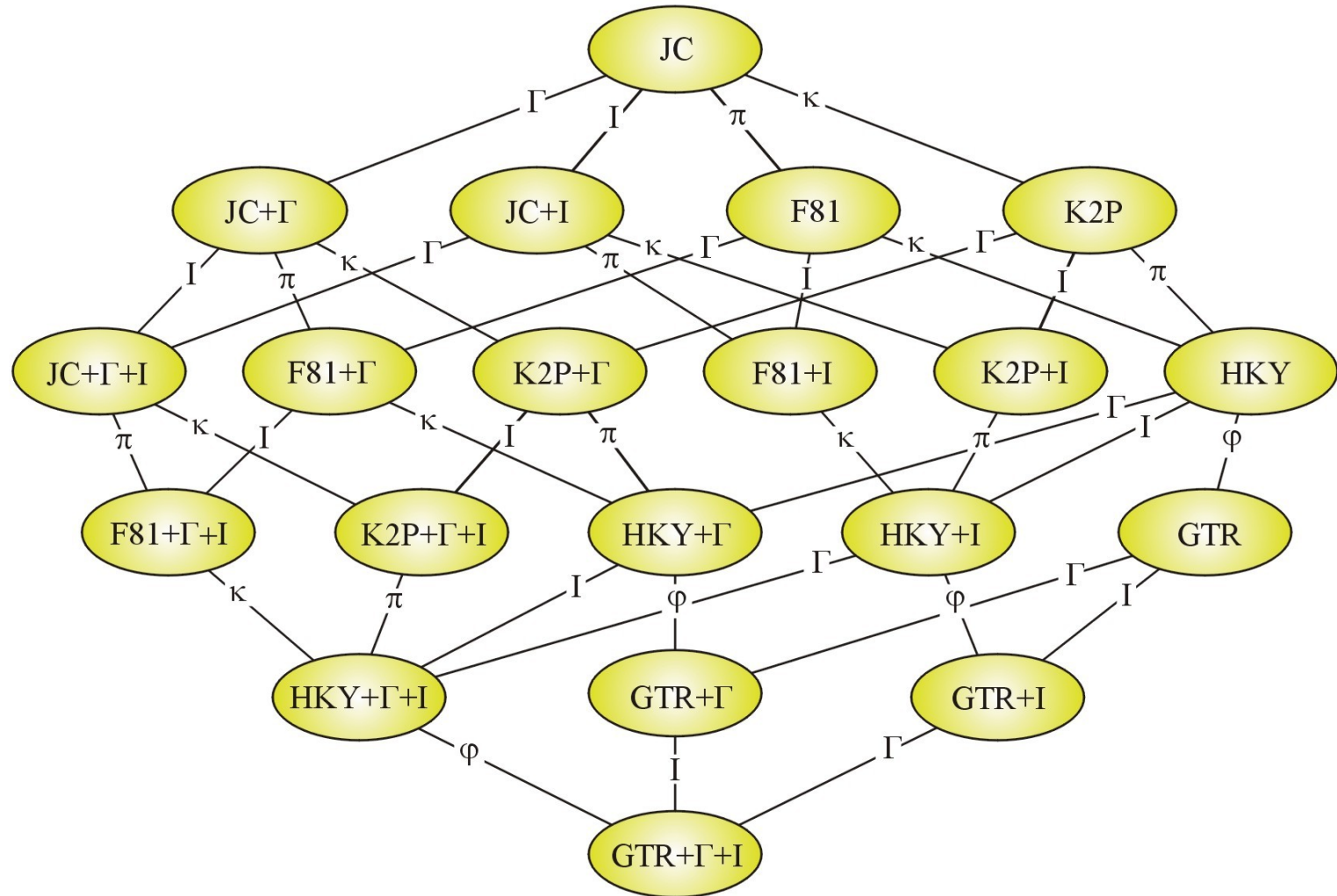
## Hierarchický LRT – ModelTest (Crandall and Posada)

Modeltest 3.0 hierarchy

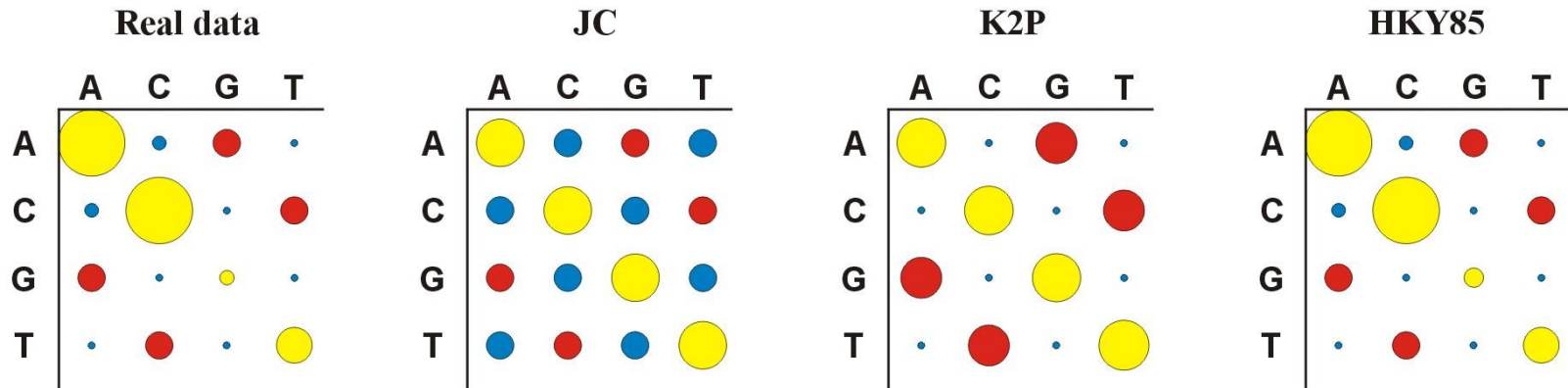


# Porovnání modelů:

## Dynamický LRT



# Porovnání modelů



Více parametrů  $\Rightarrow$  více realismu, ale ...

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



# Distance

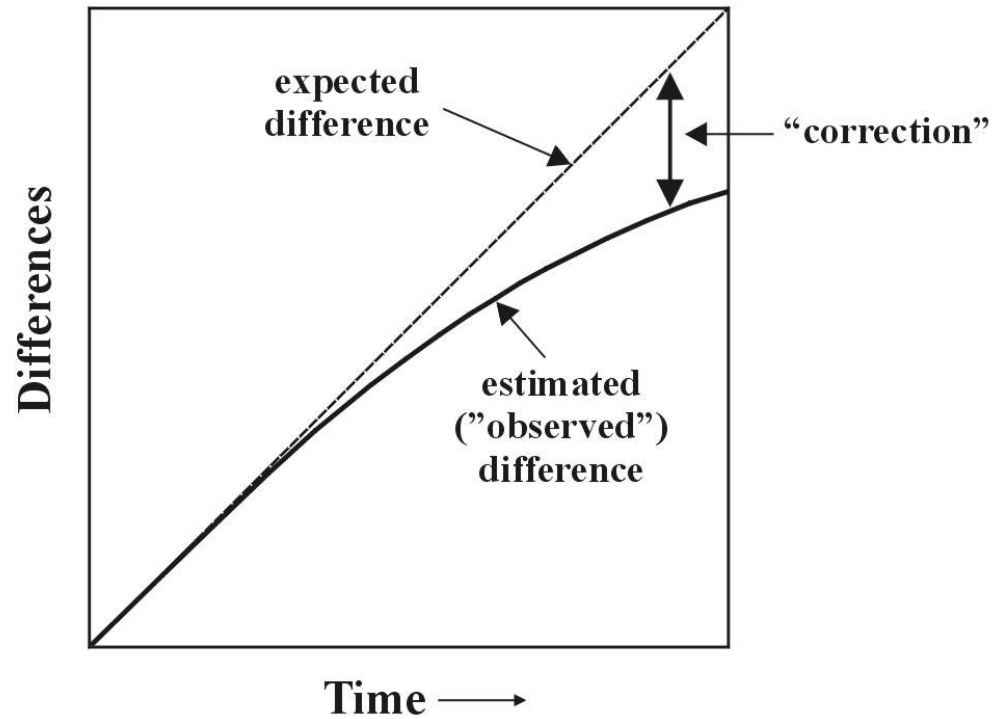
- počítány pro každý pár taxonů, z matice distancí (nebo podobností) konstruován strom
- distanční metody založeny na předpokladu, že pokud bychom znali skutečné distance mezi všemi studovanými taxony, mohli bychom velmi jednoduše rekonstruovat správnou fylogenii
- výhoda: velmi rychlé a jednoduché (lze i na kalkulačce)

# Distance

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	$d_{xy} = -B \ln\left(1 - \frac{D}{B}\right)$	$D = \text{jako JC}$ $B = 1 - (\pi_A^2 + \pi_C^2 + \pi_G^2 + \pi_T^2)$
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	$d_{xy} = -\text{stopa}\left[\prod \ln\left(\prod^{-1} \mathbf{F}_{xy}\right)\right]$	$\Pi = \text{diagonální matice průměrných četností bází v sekvencích } X \text{ a } Y$

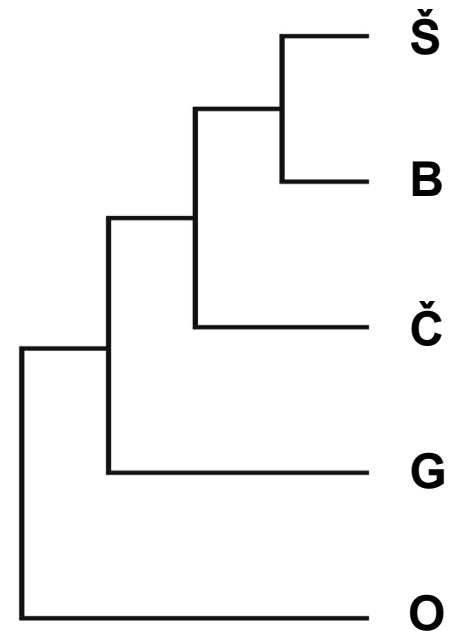
**Příklad v MEGA**

# 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  
 $d(\check{S}B-k) = [d(B-k) + d(\check{S}-k)]/2$
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	--



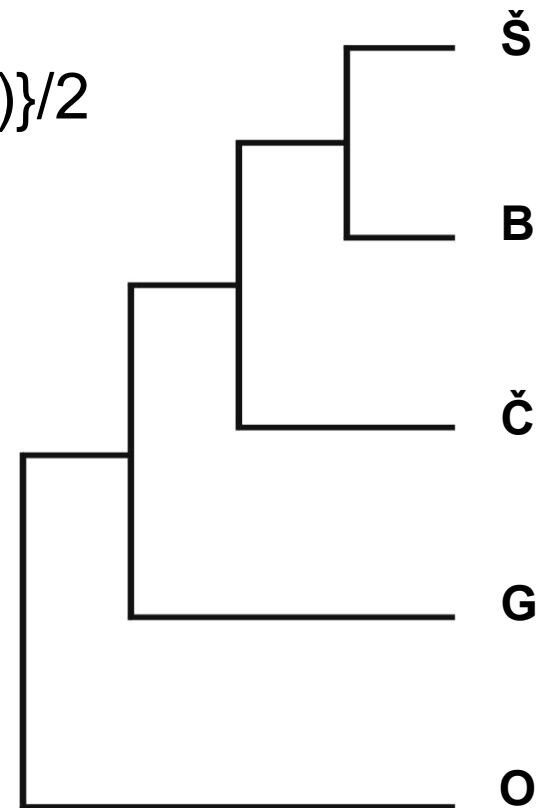
# Shluková analýza - UPGMA

UPGMA:  $d[(B\check{S}\check{C})G] = \{d(BG)+d(\check{S}G)+d(\check{C}G)\}/3$

WPGMA:  $d[(B\check{S}\check{C})G] = \{d[(B\check{S})G] + d(\check{C}G)\}/2$

single-linkage

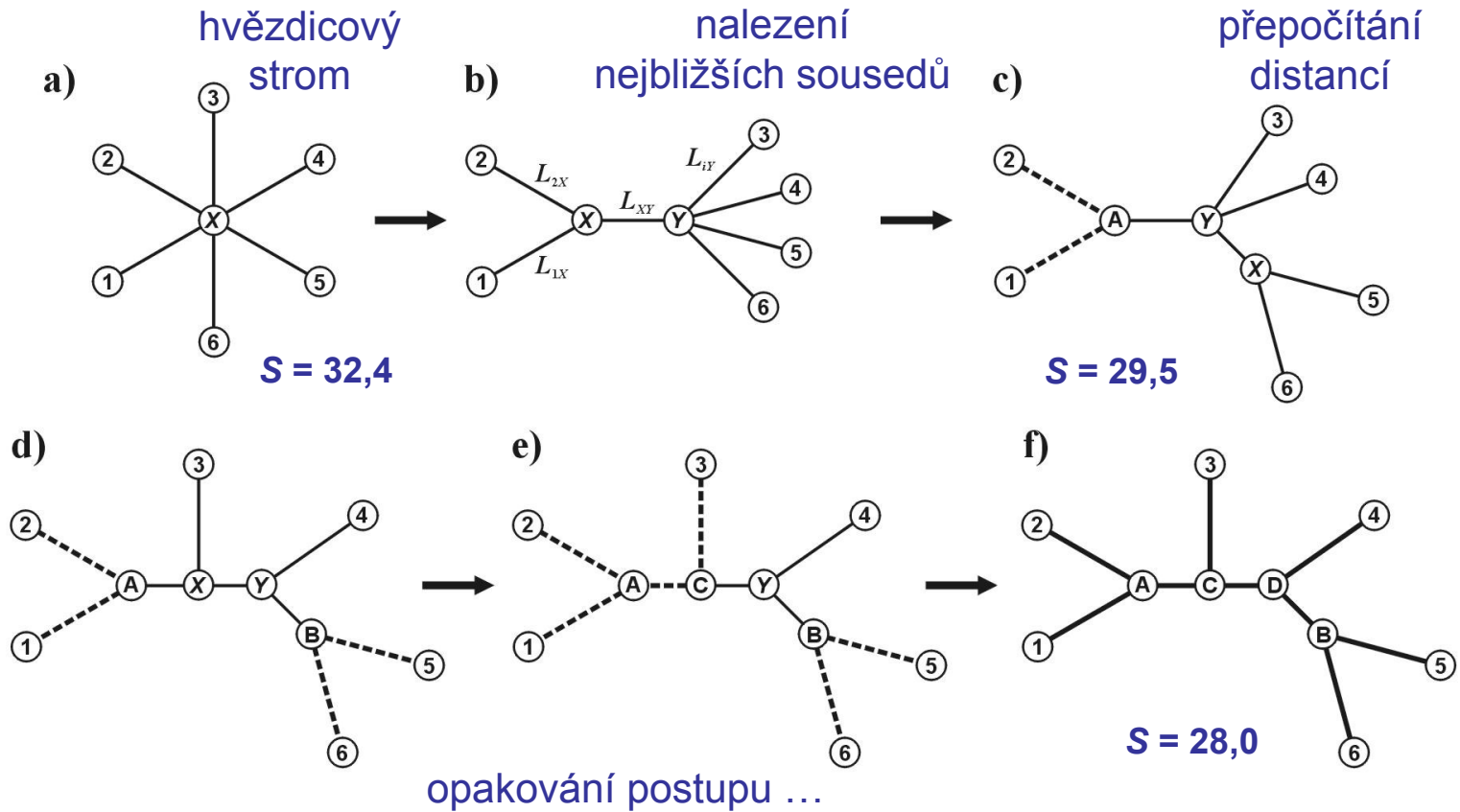
complete-linkage



# 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

# 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

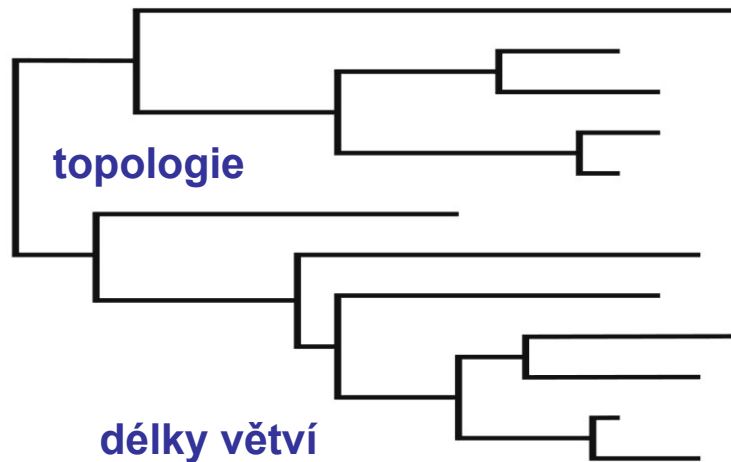


# (3) Maximální věrohodnost (maximum likelihood)

data:

```
1   TCAAAAATGGCTTTATTTCGCTTAATGCCGTTAACCTTGCGGGGGCCATG
2   TCCGTGATGGATTTATTTCCGCAATGCCTGTCATCTTATTCTCAAGTATC
3   TTCGTGATGGATTTATTGCAGGTATGCCAGTCATCCTTTTCTCATCTATC
4   TTCGTGACGGGTTTATCTCGGCAATGCCGGTTCATCCTATTTTCGAGTATT
```

strom:



evoluční model

= hypotéza

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

$$L = P(D | H),$$

kde  $D$  = matice dat  
 $H = \tau$  (topologie),  
 $\nu$  (délky větví),  
 $\theta$  (model)

## (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

Aposterioorní pravděpodobnost je dána Bayesovou rovnicí:

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

suma čitateľů pro všechny alternativní hypotézy

# Fylogenetické programy:

## alignment:

**ClustalX** *<http://inn-prot.weizmann.ac.il/software/ClustalX.html>*

**BioEdit**

**PAUP\***

**PHYLIP**

**MEGA ... MP, NJ, ML**

**McClade ... MP**

**MOLPHY, TREE-PUZZLE ... ML**

**MrBayes ... BA**

## práce se stromy:

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

**FigTree**