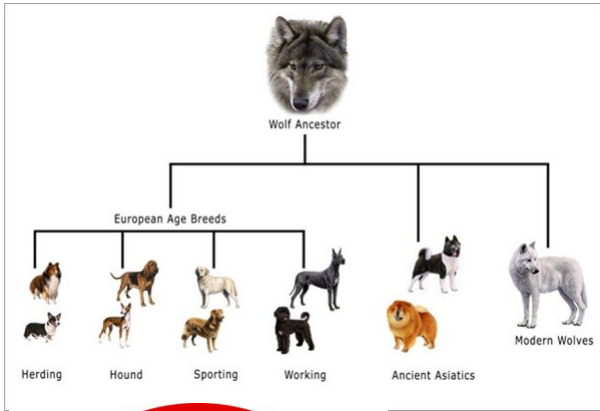
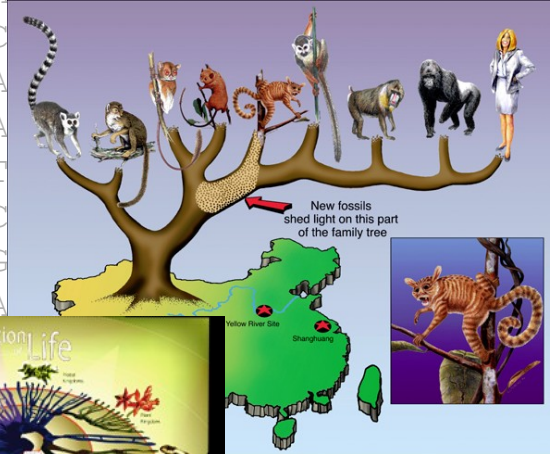


# PHYLOGENETIC ANALYSIS I.

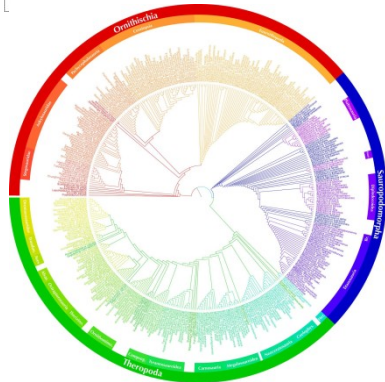
GCCTAGCCACACCCCCACGGGAGACAGCAGTGATAAACCTTTAGCAATAAACGAAAGTTTAACTAAGCCA



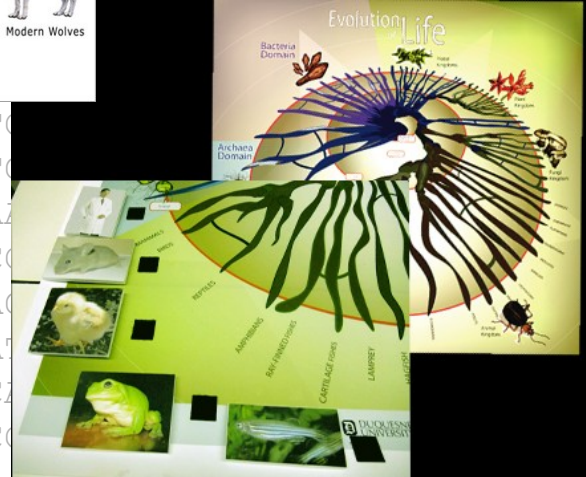
TCGTGCTAGCCAC  
 ACCCCCCCCCCAA  
 AAAGTGGCTTTAA  
 TAGCCCTAAACTT  
 CAAAGGACCTGGC  
 TCACCGCCTCTTG  
 AAGTACCACGTA



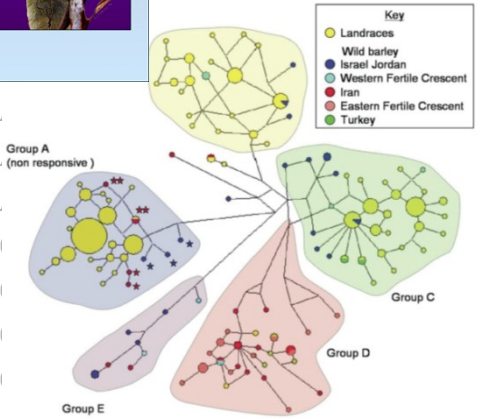
AAATAGAAA  
 AAAAAAACT  
 ACCCAAAC  
 CGCCAGAA  
 CCTGTTCT  
 GCAAACCC  
 ATGAGGCG  
 TCCGACCT



ATACTT  
 TGTACT  
 ATTTCA  
 CTTAAC  
 CCGCAA  
 AATGAA  
 AAGAAC  
 ACCTAC



GTTAGCTTA  
 CTAGCCCC  
 GCGATAGA  
 ATAATACA  
 ACTAAAGC  
 GCAAATA  
 GATAGAAT



## Definition of basic concepts:

phylogenetic tree = phylogeny (fylogenie): rooted, unrooted

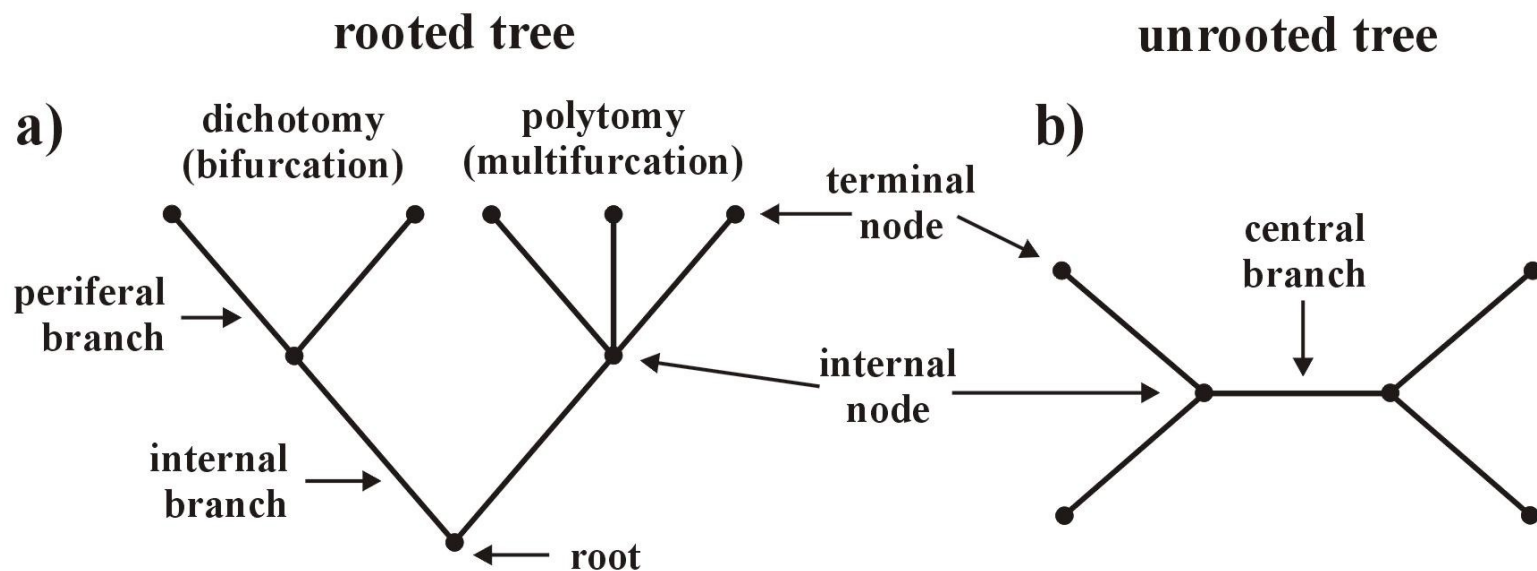
branches = edges (větvě): peripheral, internal, central

nodes = vertices (uzly): internal, terminal

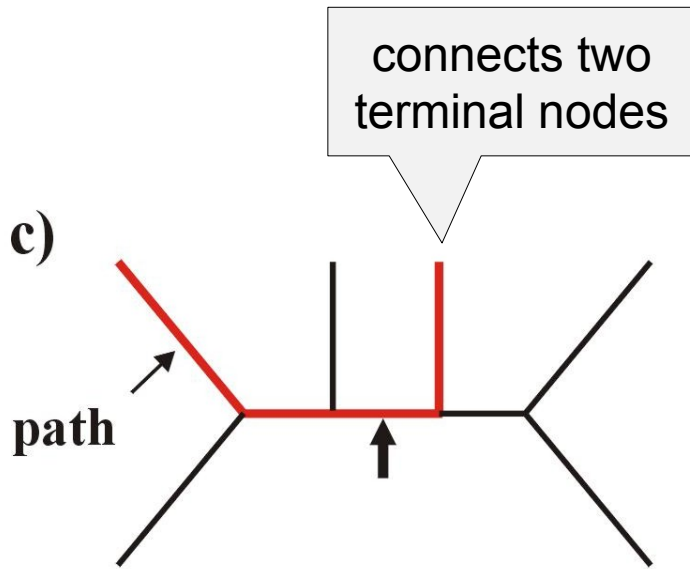
dichotomy = bifurcation, polytomy = multifurcation

OTU, HTU

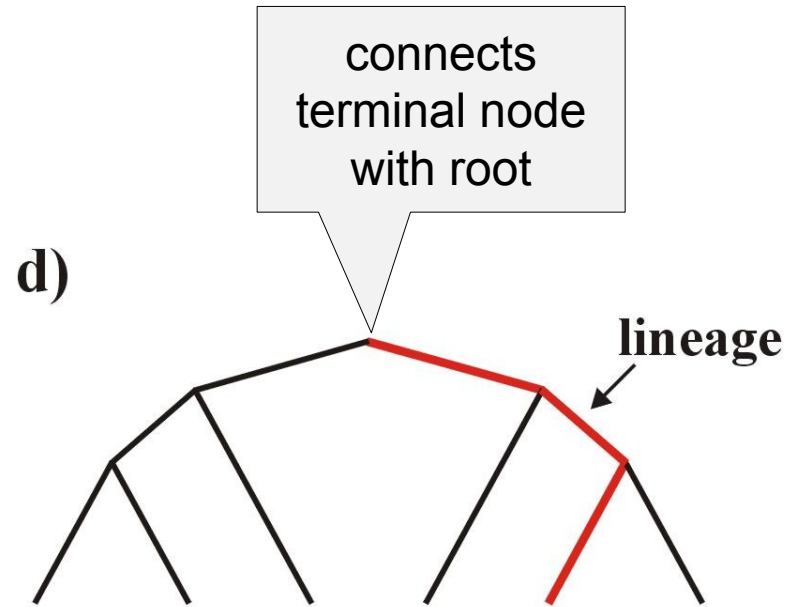
tree topology



# Definition of basic concepts:

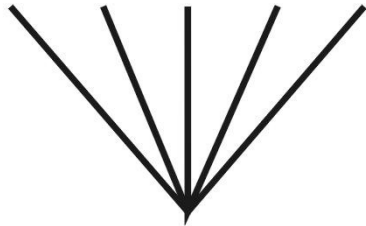


path (dráha)



lineage (linie)

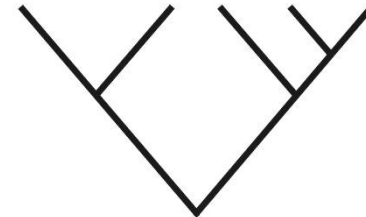
# Definition of basic concepts:



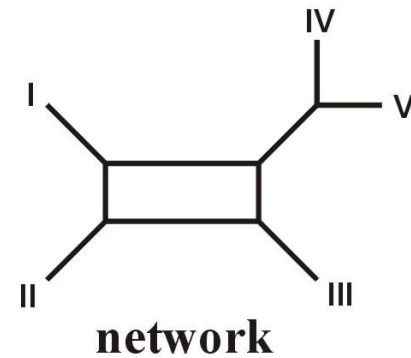
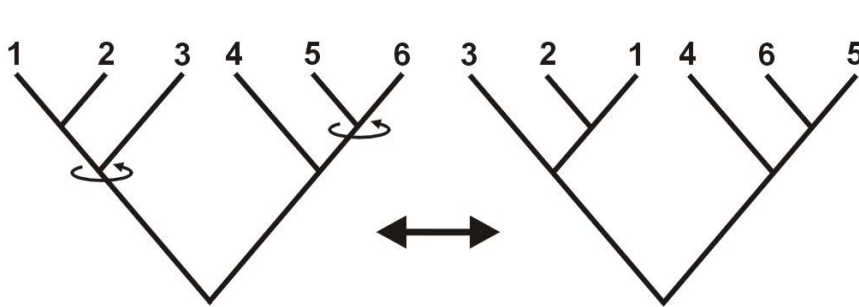
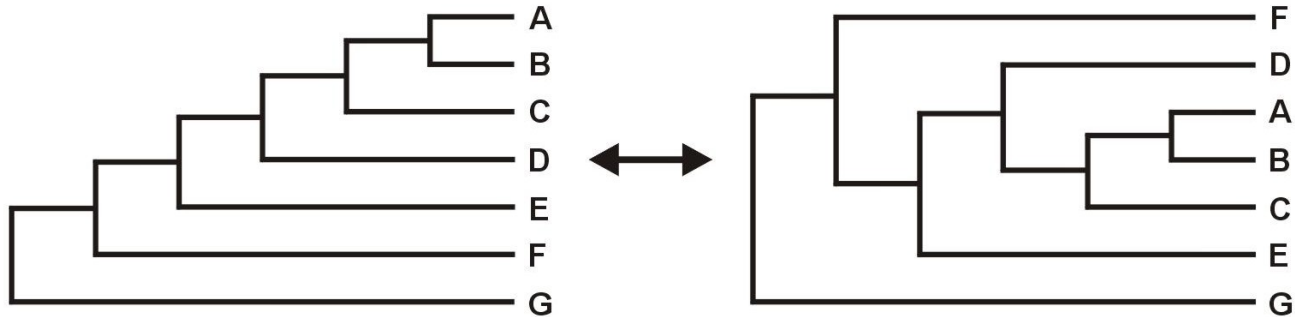
star tree

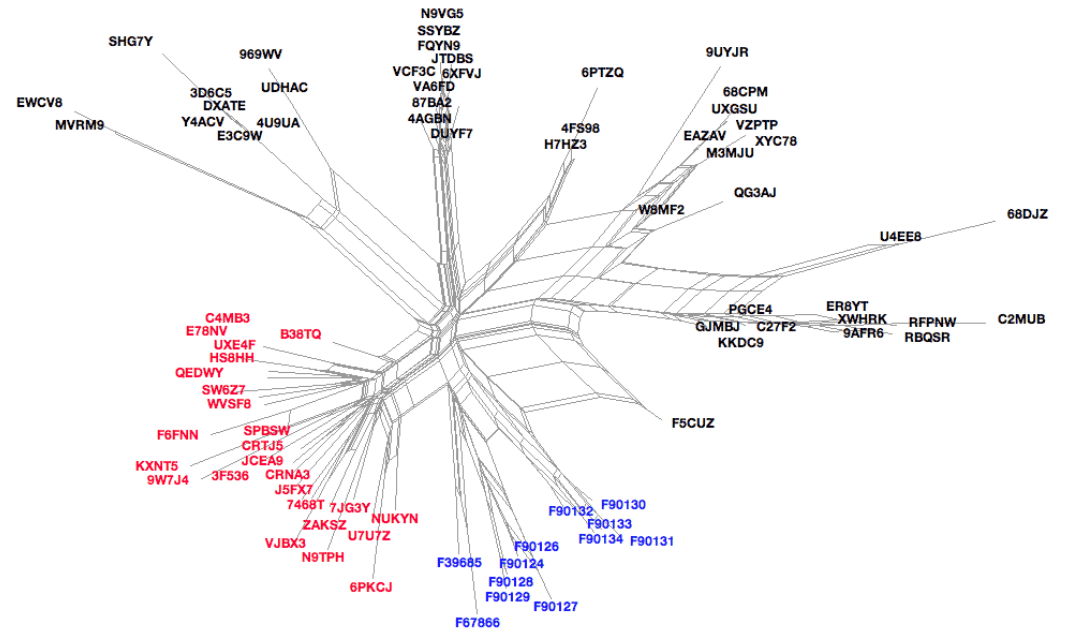
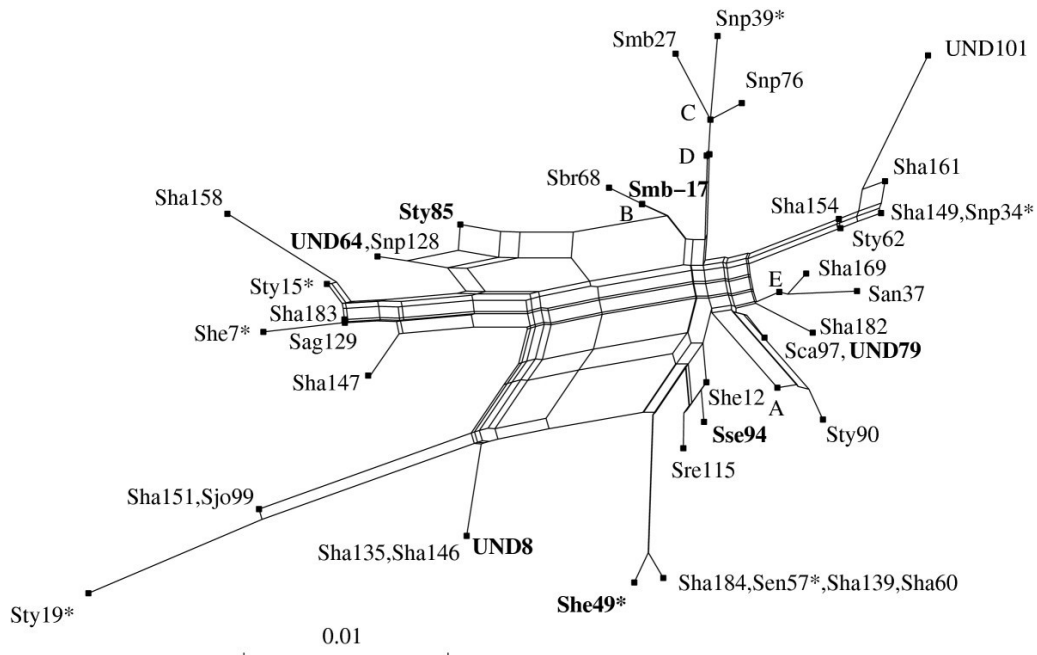


partly resolved

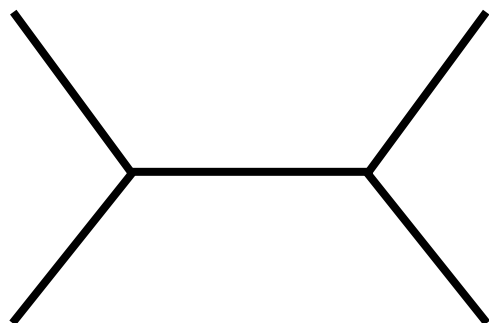


fully resolved

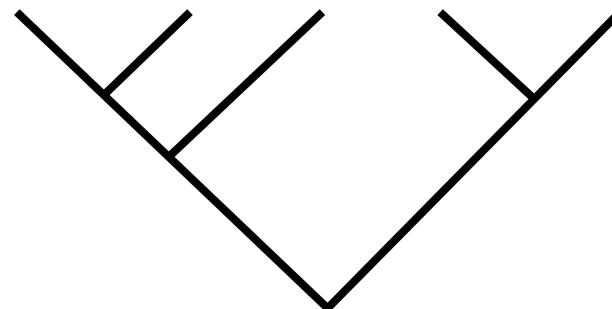




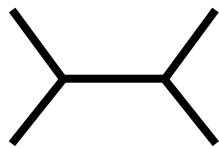
How many trees?



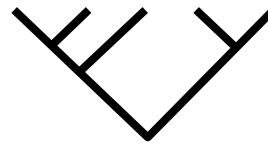
$$\frac{(2n-1)!}{2^{n-1} n}$$



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



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



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

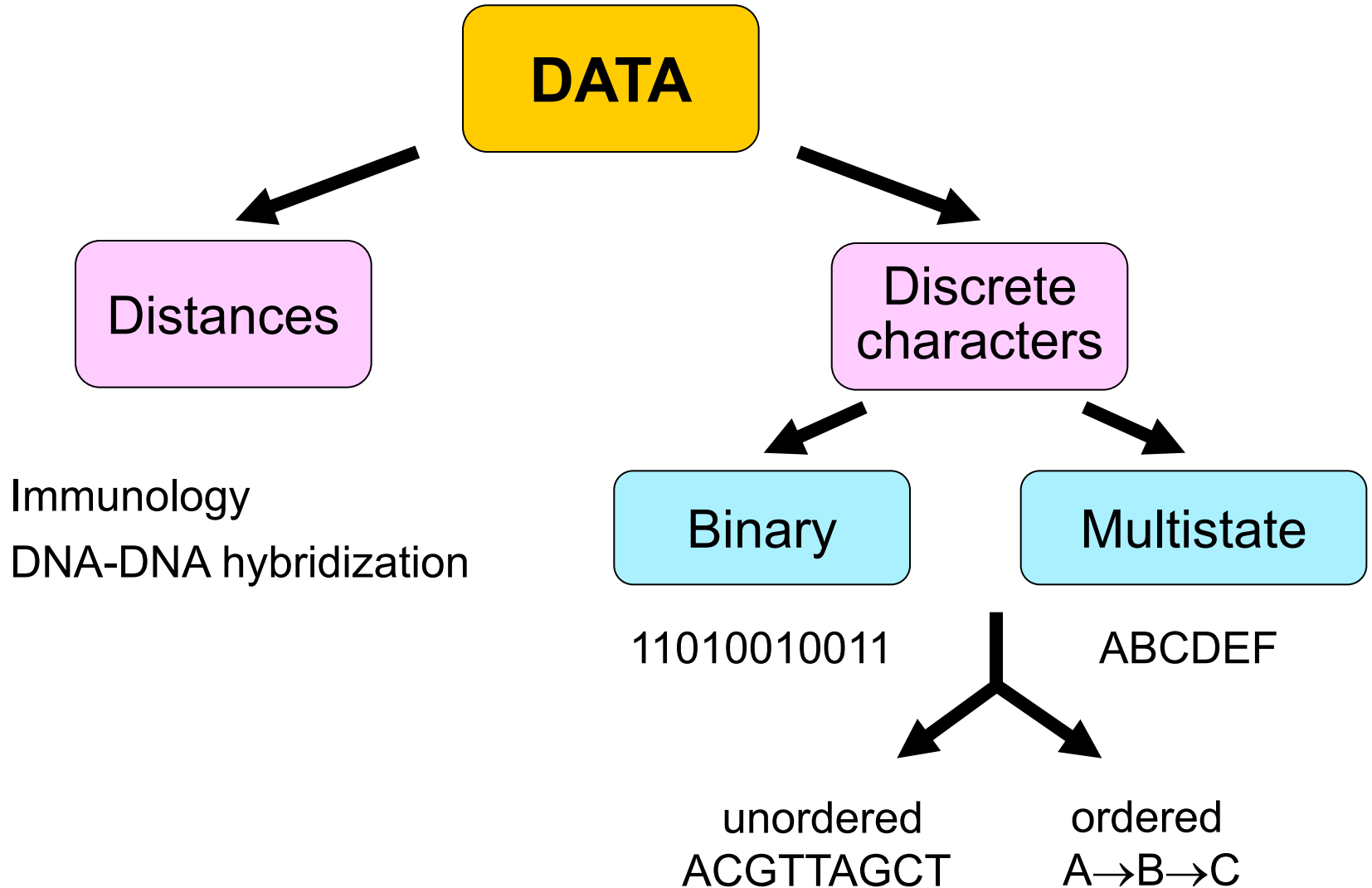
No. Taxons	Unrooted trees	Rooted trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425
11	34 459 425	654 729 075
12	654 729 075	13 749 310 575
13	13 749 310 575	316 234 143 225
14	316 234 143 225	7 905 853 580 625
15	7 905 853 580 625	213 458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

> Avogadro constant\*)

8 number of electrons in visible universe (Eddington number)

\*) 6,022 141 79×10<sup>23</sup> mol<sup>-1</sup>

# What type of data can we use?





# Types of data

Nucleotide and protein sequences:

H\_sapiens MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGS

base = character state

P\_troglod ATGACCCCGACACGCAAATAACCCACTAATAAA



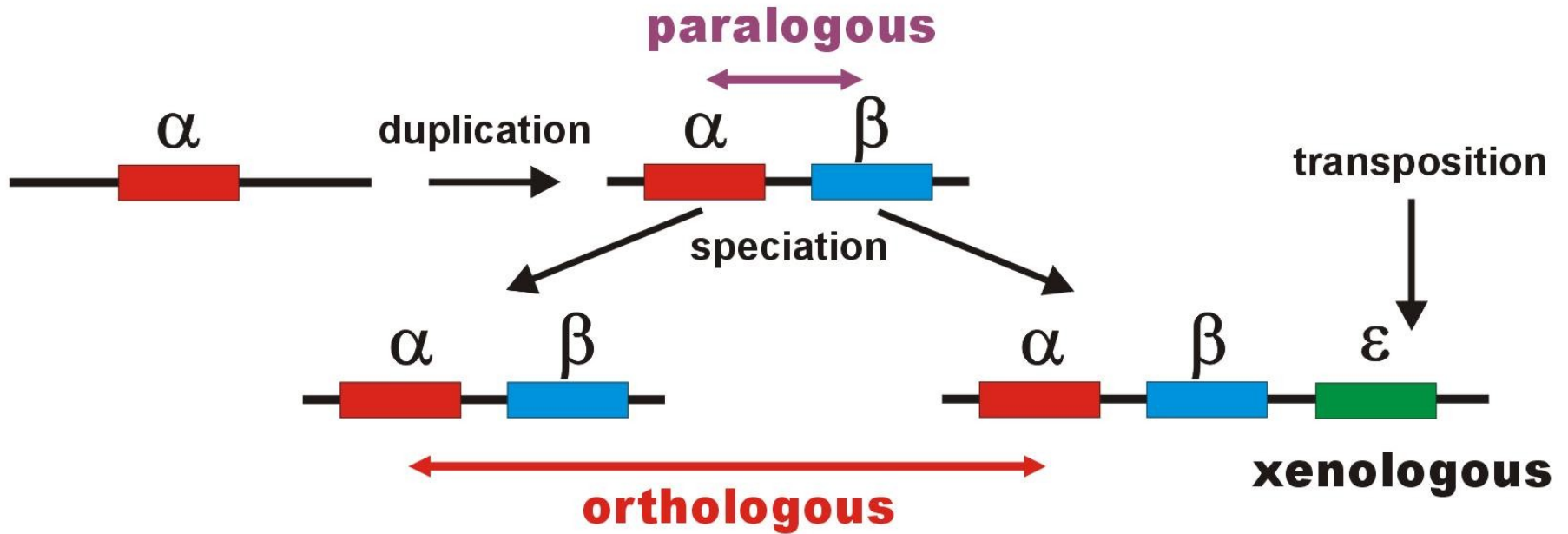
site = character

# Types of data

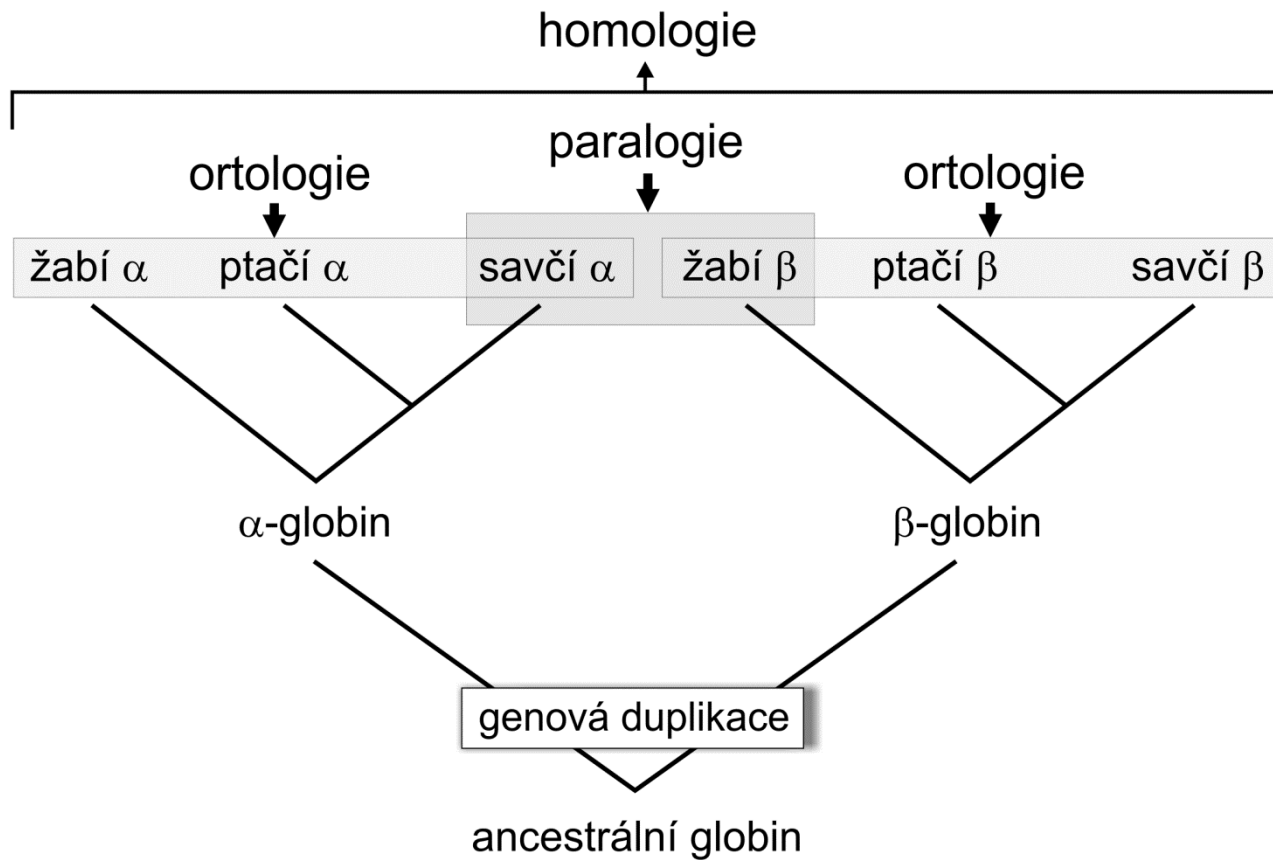
retroelements: SINE (*Alu*, B1, B2), LINE

microsatellites, SNP

# Problem with homology of sequences



# Problem with homology of sequences



Individual sites in DNA sequences may not be fully independent!

# Sequences

## DNA databases:

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/>

Database management: usually packages Sybase or ORACLE

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

# Sequences

## Protein databases:

**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/>

# File formats:

## FASTA:

>H\_sapiens

```
ATGACCCCAATACGCAAATTAACCCCTAATAAAATTAATTAACCACTCATTTCATCGACCTCCCCACCC
CATCCAACATCTCCGCATGATGAAACTTCGGCTCACTCCTTGGCGCCTGCCTGATCCTCCAAATCACCAC
AGGACTATTTCCTAGCCATACTACTCACCAGACGCCTCAACCGCCTTTTCATCAATCGCCACATCACT
CGAGACGTAAATTATGGCTGAATCATCCGCTACCTTCACGCCAATGGCGCCTCAATATTCTTTATCTGCC
TCTTCCTACACATCGGGCGAGGCCTATATTACGGATCATTTCTCTACTCAGAAACCTGAAACATCGGCAT
```

...

>P\_troglod

```
ATGACCCCGACACGCAAATTAACCCACTAATAAAATTAATTAATCACTCATTTATCGACCTCCCCACCC
CATCCAACATTTCCGCATGATGGAACTTCGGCTCACTTCTCGGCGCCTGCCTAATCCTTCAAATTACCAC
AGGATTATTTCCTAGCTATACTACTCACCAGACGCCTCAACCGCCTTCTCGTCGATCGCCACATCACC
CGAGACGTAAACTATGGTTGGATCATCCGCTACCTCCACGCTAACGGCGCCTCAATATTTTTTATCTGCC
TCTTCCTACACATCGGCCGAGGTCTATATTACGGCTCATTTCTCTACCTAGAAACCTGAAACATTGGCAT
```

...

>P\_paniscus

```
ATGACCCCAACACGCAAATCAACCCACTAATAAAATTAATTAATCACTCATTTATCGACCTCCCCACCC
CATCCAATATTTCCACATGATGAAACTTCGGCTCACTTCTCGGCGCCTGCCTAATCCTTCAAATCACCAC
AGGACTATTTCCTAGCTATACTACTCACCAGACGCCTCAACCGCCTTCTCATCGATCGCCACATTACC
CGAGACGTAAACTATGGTTGAATCATCCGCTACCTTCACGCTAACGGCGCCTCAATACTTTTCATCTGCC
TCTTCCTACACGTCCGGTCGAGGCCTATATTACGGCTCATTTCTCTACCTAGAAACCTGAAACATTGGCAT
```

...

# File formats:

## 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 ctccgggccc
421 attaaacttg ggggtagcta aactgaaact ttatcagaca tctggttctt acttcagggc
481 catcaaagtc gttatcgccc atacgttccc cttaaataag acatctcgat ggtatcgggt
541 ctaatcagcc catgaccaac ataactgtgg tgtcatgcat ttggtattht 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 caaatatga ctcatattht agtacttgta
901 aaaatthtac aaaatcatgc tccgtgaacc aaaactctaa tcacactcta ttacgcaata
961 aatattaaca agttaatgta gcttaataac aaagcaaagc actgaaaatg cttagatgga
1021 taatthtatc cca
```

//



# File formats:

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

# File formats:

## 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;
```

# File formats:

## 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
** **  ****  **  **  **  **          *****  *****  ** *****  **  **  *****
```

# File formats:

## FASTQ:

Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a FASTA title line).

Line 2 is the raw sequence letters.

Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

## Illumina sequence identifiers:

@HWUSI-EAS100R:6:73:941:1973#0/1

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 ( <i>paired-end or mate-pair reads only</i> )

# BLAST

## Basic Local Alignment Search Tool

**BLAST** finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

**NEWS**

Using BLAST+ in Docker and on the cloud: [Webinar](#) on December 9, 2020.

In this webinar, the NCBI BLAST team will demonstrate containerized BLAST+ in Docker that is ready to use locally and in the cloud.

Wed, 02 Dec 2020 12:00:00 EST [More BLAST news...](#)

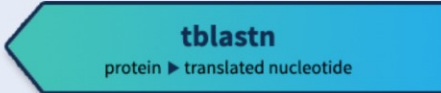
## Web BLAST



**Nucleotide BLAST**  
nucleotide ▶ nucleotide



**blastx**  
translated nucleotide ▶ protein



**tblastn**  
protein ▶ translated nucleotide



**Protein BLAST**  
protein ▶ protein

## BLAST Genomes

- [Human](#) [Mouse](#) [Rat](#) [Microbes](#)

### Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

[Reset page](#) [Bookmark](#)

#### Enter Query Sequence


Enter accession number(s), gi(s), or FASTA sequence(s)

```
CAAAAACACTAAGAACCTTGAAGACATATACCTAACTATCTAACCCCTATGTCCTGATCAATCTAGTAGTT
CAAAAAATATGACTTATATTTTAGTTCCTGTAAAAATTTTGCAAAATAATGCCCCATAAGCCAAAACCTAAAT
TATACCCCTATTACGCAATAAACAAATAGTAAAGTTAATGTAGCTTAATAAAAAGCAAAGCACTGAAAATGCTTAG
ATGGATAATTTTATCCCATAAACACAAAAGTTTGGTC
```

[Clear](#) Query subrange

From   
To

**New columns added to the Description Table**  
Click 'Select Columns' or 'Manage Columns'.



Or, upload file  Soubor nevybrán

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

#### Choose Search Set

Database  Standard databases (nr etc.):  rRNA/ITS databases  Genomic + transcript databases  Betacoronavirus

Nucleotide collection (nr/nt)

Organism   exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude  Models (XM/XP)  Uncultured/environmental sample sequences

Limit to  Sequences from type material

Entrez Query  [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search

#### Program Selection

Optimize for  Highly similar sequences (megablast)  
 More dissimilar sequences (discontiguous megablast)  
 Somewhat similar sequences (blastn)

Choose a BLAST algorithm

**BLAST** Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

[+ Algorithm parameters](#)

BLAST is a registered trademark of the National Library of Medicine

[Support center](#) [Mailing list](#) [YouTube](#)

**NCBI**  
National Center for Biotechnology Information, U.S. National Library of Medicine  
8600 Rockville Pike, Bethesda MD, 20894 USA



[Policies and Guidelines](#) | [Contact](#)

[← Edit Search](#) [Save Search](#) [Search Summary ▾](#)

**Job Title** **Nucleotide Sequence**

**RID** [WX41A1JA013](#) Search expires on 12-09 03:27 am [Download All ▾](#)

**Program** BLASTN ⓘ [Citation ▾](#)

**Database** nt [See details ▾](#)

**Query ID** lc|Query\_50051

**Description** None

**Molecule type** dna

**Query Length** 950

**Other reports** [Distance tree of results](#) [MSA viewer](#) ⓘ

ⓘ How to read this report? [▶ BLAST Help Videos](#) [↶ Back to Traditional Results Page](#)

**Filter Results**

**Organism** only top 20 will appear  exclude

Type common name, binomial, taxid or group name

+ [Add organism](#)

**Percent Identity**  to

**E value**  to

**Query Coverage**  to

[Filter](#) [Reset](#)

**Descriptions** [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

**Sequences producing significant alignments** [Download ▾](#) New [Select columns ▾](#) [Show](#)  ⓘ

select all 100 sequences selected

[GenBank](#) [Graphics](#) [Distance tree of results](#)

	Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1375 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	878	1756	100%	0.0	100.00%	1062	<a href="#">EU106210.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1373 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	878	1751	100%	0.0	100.00%	1062	<a href="#">EU106208.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1372 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	874	1747	100%	0.0	99.79%	1062	<a href="#">EU106207.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1391 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	872	1653	95%	0.0	99.79%	1015	<a href="#">EU106216.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1388 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	872	1739	100%	0.0	99.79%	1062	<a href="#">EU106214.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1387 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	872	1712	100%	0.0	99.79%	1062	<a href="#">EU106213.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1374 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	872	1717	100%	0.0	99.79%	1062	<a href="#">EU106209.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1364 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	872	1734	100%	0.0	99.79%	1062	<a href="#">EU106204.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1360 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	872	1706	100%	0.0	99.79%	1062	<a href="#">EU106200.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1371 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	872	1739	100%	0.0	99.79%	1062	<a href="#">EU106199.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1383 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	872	1712	100%	0.0	99.79%	1062	<a href="#">EU106198.1</a>
<input checked="" type="checkbox"/>	<a href="#">Mus cypricus isolate MM1358 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gen...</a>	Cyprriot mouse	872	1706	100%	0.0	99.79%	1062	<a href="#">EU106195.1</a>



100 sequences selected

[Download](#) [GenBank](#) [Graphics](#)Sort by: [Next](#) [Previous](#) [Descriptions](#)**Mus cypricus isolate MM1375 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete sequence; mitochondrial**Sequence ID: [EU106210.1](#) Length: 1062 Number of Matches: 2[See 2 more title\(s\)](#) [See all Identical Proteins\(IPG\)](#)Range 1: 1 to 475 [GenBank](#) [Graphics](#)[Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
878 bits(475)	0.0	475/475(100%)	0/475(0%)	Plus/Plus
Query 1		TGTAAACCTGAAATGAAGATATTCTTCTCAAGACATCAAGAAGAAGGAACTTATTCCC		60
Sbjct 1		TGTAAACCTGAAATGAAGATATTCTTCTCAAGACATCAAGAAGAAGGAACTTATTCCC		60
Query 61		CACCACCAACACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTAC		120
Sbjct 61		CACCACCAACACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTAC		120
Query 121		ATAGTACATTAGTACATTTATGTATATCGTACATTAATTATATTTCCCAAGCATATAAG		180
Sbjct 121		ATAGTACATTAGTACATTTATGTATATCGTACATTAATTATATTTCCCAAGCATATAAG		180
Query 181		CACGTAATTAATTAATGACATAGCACATAAAACGATATTTAACATAAAATACTACACA		240
Sbjct 181		CACGTAATTAATTAATGACATAGCACATAAAACGATATTTAACATAAAATACTACACA		240
Query 241		ACATGAATATTATATTAATACATTAAGTTAATGCTTTAAAGACATATCTGTGTTATCTG		300
Sbjct 241		ACATGAATATTATATTAATACATTAAGTTAATGCTTTAAAGACATATCTGTGTTATCTG		300
Query 301		ACATACACCATAAAGTCATAAACCCCTTCTTCCATATGACTATCCCCTTCCCATTGG		360
Sbjct 301		ACATACACCATAAAGTCATAAACCCCTTCTTCCATATGACTATCCCCTTCCCATTGG		360
Query 361		TCTATTAATCTACCATCCTCCGTGAAACCAACAACCCGCCACCTATGCCCTTCTCTCG		420
Sbjct 361		TCTATTAATCTACCATCCTCCGTGAAACCAACAACCCGCCACCTATGCCCTTCTCTCG		420
Query 421		CTCCGGGCCCATTAACCTTGGGGGTAGCTAAACTGAAACTTTATCAGACATCTGG		475
Sbjct 421		CTCCGGGCCCATTAACCTTGGGGGTAGCTAAACTGAAACTTTATCAGACATCTGG		475

Range 2: 588 to 1062 [GenBank](#) [Graphics](#)[Next Match](#) [Previous Match](#) [First Match](#)

# GenBank

ncbi.nlm.nih.gov/nuccore/?term=mus+cypriacus

Search:

**COVID-19 is an emerging, rapidly evolving situation.**  
Get the latest public health information from CDC: <https://www.coronavirus.gov>  
Get the latest research information from NIH: <https://www.nih.gov/coronavirus>  
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>

Species: Animals (29)  
Molecule types: genomic DNA/RNA (29)  
Source databases: INSDC (GenBank) (29)  
Sequence Type: Nucleotide (29)  
Genetic compartments: Mitochondrion (28)  
Sequence length: Custom range...  
Release date: Custom range...  
Revision date: Custom range...

Summary 20 per page Sort by Default order Send to: Filters: [Manage Filters](#)

Items: 1 to 20 of 29

1. [Mus cypriacus mitochondrial partial cytb gene for cytochrome b](#)  
1,140 bp linear DNA  
Accession: FR751074.1 GI: 323713991  
[Protein](#) [PubMed](#) [Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#)

2. [Mus cypriacus isolate MM1381 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete sequence; mitochondrial](#)  
1,062 bp linear DNA  
Accession: EU106281.1 GI: 157266050  
[PubMed](#) [Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#)

3. [Mus cypriacus isolate MM1377 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete sequence; mitochondrial](#)  
1,062 bp linear DNA  
Accession: EU106280.1 GI: 157266049  
[PubMed](#) [Taxonomy](#)

Find related data  
Database: Select

Search details  
"Mus cypriacus"[Organism] OR mus cypriacus[All Fields]  
 [See more...](#)

Recent activity  
[Turn Off](#) [Clear](#)  
mus cypriacus (29) Nucleotide  
mus cypriacus control region (0)

Sem zadejte hledaný výraz

CES 20:09  
CSQ 07.12.2020

# Mus cypriacus isolate MM1381 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete sequence; mitochondrial

GenBank: EU106281.1

[FASTA](#) [Graphics](#) [PopSet](#)

[Go to:](#)

LOCUS EU106281 1062 bp DNA linear ROD 16-NOV-2007

DEFINITION Mus cypriacus isolate MM1381 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete sequence; mitochondrial.

ACCESSION EU106281

VERSION EU106281.1

KEYWORDS .

SOURCE mitochondrion Mus cypriacus (Cypriot mouse)

ORGANISM [Mus cypriacus](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; Mus; Mus.

REFERENCE 1 (bases 1 to 1062)

AUTHORS Macholan,M., Vyskocilova,M., Bonhomme,F., Krystufek,B., Orth,A. and Vohralik,V.

TITLE Genetic variation and phylogeography of free-living mouse species (genus Mus) in the Balkans and the Middle East

JOURNAL Mol Ecol 16 (22), 4774-4788 (2007)

PUBMED [17908218](#)

REFERENCE 2 (bases 1 to 1062)

AUTHORS Macholan,M., Vyskocilova,M., Bonhomme,F., Krystufek,B., Orth,A. and Vohralik,V.

TITLE Direct Submission

JOURNAL Submitted (20-AUG-2007) Laboratory of Mammalian Evolutionary Genetics, Institute of Animal Physiology and Genetics, Acad. Sci. Czech Rep., Veveri 97, Brno CZ-60200, Czech Republic

FEATURES

source 1..1062  
/organism="Mus cypriacus"  
/organelle="mitochondrion"  
/mol\_type="genomic DNA"  
/isolate="MM1381"  
/db\_xref="taxon:468371"  
/country="Cyprus: Paramythia, 12 km N Limassol"  
/note="type: CY17"  
[tRNA](#) <1..37  
/product="tRNA-Thr"  
[tRNA](#) 38..105  
/product="tRNA-Pro"  
[D-loop](#) 106..982  
[gap](#) 476..587  
/estimated\_length=112  
[tRNA](#) 983..1049  
/product="tRNA-Phe"

ORIGIN

```
1 tgtaaacctg aaatgaagat attctcttct caagacatca agaagaagga acttattccc
61 caccaccaac acccaagctt ggtattctag ttaaactact tcttgagtac ataaatttac
121 atagtacatt agtacattta tgtatatcgt acattaaatt atattcccca agcatataag
181 cactgaaatt aattaatga catagacat aaaaagatata ttaacataaa atactacaca
241 acatgaatat tatattaaat acattaagtt aatgctttaa agacatatct gtgttatctg
301 acatacacca taaagtcata aacccttctc ttccatatag ctatccccct ccccatctgg
361 tctattaatc taccatcctc cgtgaaacca acaaccgcc cacctatgcc cctcttctcg
421 ctccgggccc attaaacttg ggggtagcta aactgaaact ttatcagaca tctgg
[ gap 112 bp] Expand Ns
588
601 ttatttttgg tctactttca tcaacatagc cgtcaaggca tgaaggaca gcacacagtc
661 tagacgacc tacggtgaag aatcattagt ctcataacc caatcaccca aggctaatta
721 ttcattgctt ttacacataa aattattcaa taccagattt taactctcca aacccccccc
781 accccatccc tcttaatgcc aaacccaaa aacactaaga actgaaaga catatactat
841 taactatcta accctatgct ctgatcaatt ctagtattc aaaaatatg acttatattt
901 tagttcttgt aaaaatttgg caaaataatg ccccataagc caaaactcta attatccctc
961 attacgcaat aaacaatagt aagttaatgt agcttaataa aaagcaaacg actgaaaatg
1021 cttagatgga taattttatc ccataaacac aaagtttgg tc
```

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Related information

PubMed

Taxonomy

PopSet

Recent activity

[Turn Off](#) [Clear](#)

Mus cypriacus isolate MM1381 tRNA-Thr gene, partial sequence; and tRNA-P

Nucleotide

mus cypriacus (29)

mus cypriacus control region (0)

Nucleotide

On the Probability of Fixation of Mutant Genes in a Population

The frequency of multiple paternity suggests that sperm competition is common in...

[See more...](#)

# Mus cypricus isolate MM1381 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete sequence; mitochondrial

GenBank: EU106281.1

[GenBank](#) [Graphics](#) [PopSet](#)

>EU106281.1 Mus cypricus isolate MM1381 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete sequence; mitochondrial

```
TGTAAACCTGAAATGAAGATATTCTTCTCAAGACATCAAGAAGAAGGAACCTATTCCCCACCACCAAC
ACCCAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAAATTTACATAGTACATTAGTACATTTA
TGTATATCGTACATTAATAATATATTCCCCAAGCATATAAGCACGTAATTAATTAATGACATAGCACAT
AAAACGATATTTAACATAAAAATACTACACAACATGAATATTATATAAATACATTAAAGTTAATGCTTTAA
AGACATATCTGTGTTATCTGCATACACCATAAAGTCATAAACCCCTTCTTCCATATGACTATCCCCCTT
CCCCATTTGGTCTATTAATCTACCATCCTCGTGAACCAACAACCCGCCACCTATGCCCTCTTCTCG
CTCCGGGCCATTAAACTTGGGGGTAGCTAAACTGAAACTTTATCAGACATCTGGNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CGTCAAGGCATGAAAGGACAGCACAGCTAGACGCACCTACGGTGAAGAATCATTAGTCCTCATAACC
CAATCACCAAGGCTAATATTATCATGCTGTGTAGACATAAAATATTCAATACCAGATTTAACTCTCCA
AACCCCCCAACCCCATCCTCTTAATGCCAAACCCAAAACACTAAGAATTGAAAGACATATACTAT
TAACTATCTAACCCATATGCTCTGATCAATTCTAGTAGTTCAAAAAATATGACTTATATTTAGTCTTGT
AAAAATTTGCAAATAATGCCCCATAAGCCAAAACCTAATATAACCCATTACGCAATAAACAATAGT
AAGTTAATGTAGCTTAATAAAAAGCAAAGCACTGAAAATGCTTAGATGGATAATTTATCCCATAAACAC
AAAGGTTTGGTC
```

Change region shown ▾

Customize view ▾

Analyze this sequence ▾

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Related information ▾

PubMed

Taxonomy

PopSet

Recent activity ▾

Turn Off Clear

 Mus cypricus isolate MM1381 tRNA-Thr

# Mus cypricus isolate MM1381 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete sequence; mitochondrial

GenBank: EU106281.1

[GenBank](#) [FASTA](#) [PopSet](#)
[Link To This View](#) | [Feedback](#)


Analyze this sequence ▾

Run BLAST

Pick Primers

Related information ▾


PubMed

Taxonomy

PopSet

Recent activity ▾

Turn Off Clear

 Mus cypricus isolate MM1381 tRNA-Thr gene, partial sequence; and tRNA-Pro gene,...

Nucleotide

 mus cypricus (29)

Nucleotide

 mus cypricus control region (0)

# Sequence alignment:

Sequence 1    **TTGTACGACGG**  
 Sequence 2    **TTGTACGACG**

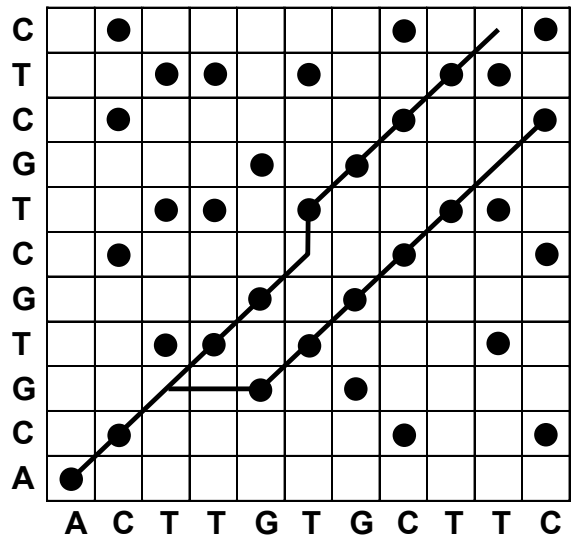
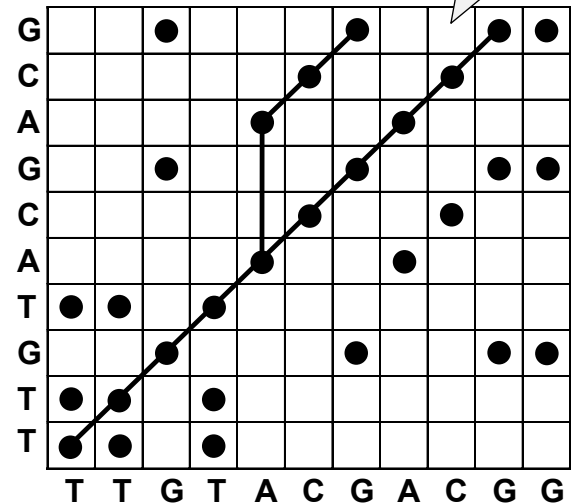
**TTGTACGACGG**    **TTGT---ACGACGG**  
 | | | | | | | | | |    | | |    | | |  
**TTGTACGACG**    **TTGTACGACG**

Sequence 1    **ACTTGCTTC**  
 Sequence 2    **ACGTGCTGC**

Path 1    **ACTTGCTTC**  
 | | | | | | | |  
**ACGTGCTGC**

Path 2    **ACTTGCTTC**  
 | |    | | | | | | |  
**AC--GTGCTGC**

dot plot



# Sequence alignment:

Gap penalty:

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

$h$  = extenze za každou „pomlčku“

$l$  = délka mezery (= počet „pomlček“)

Př.: GC□ □ □ □ TTAA

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

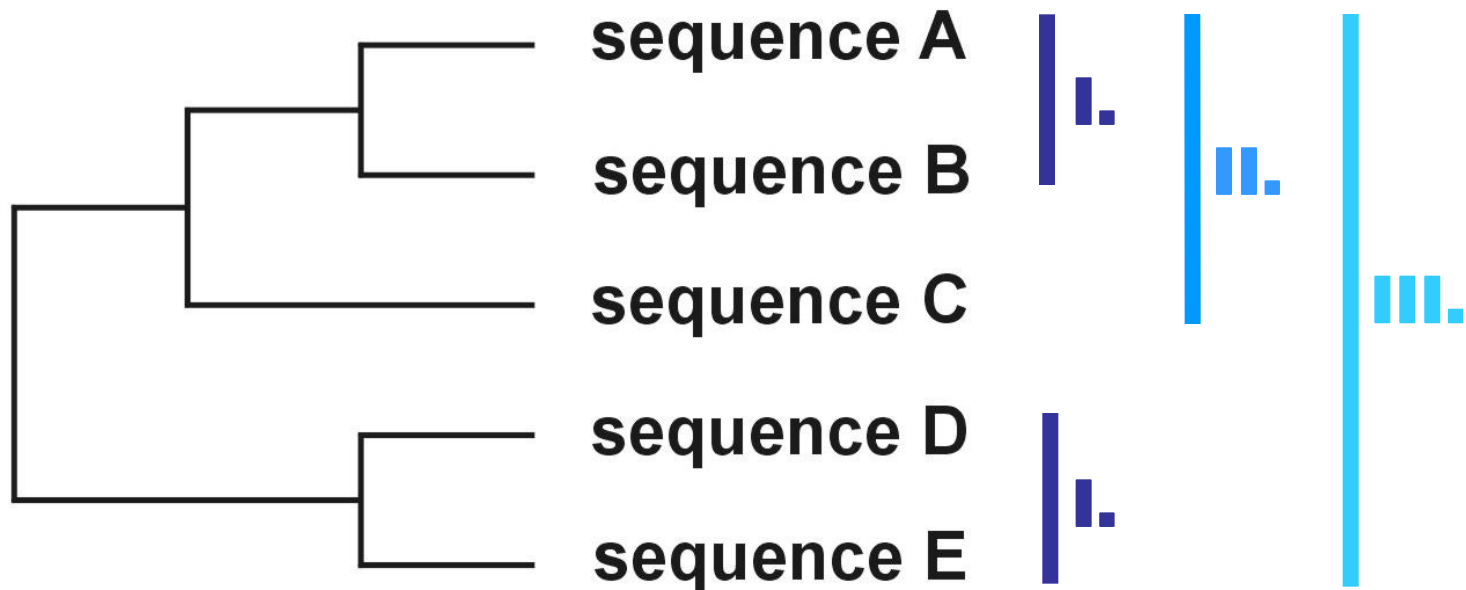
$$GP = g + hl$$

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

# Progressive alignment - ClustalX

3 phases:

1. Alignment of sequence pairs  $\rightarrow$  pairwise distances
2. Construction of guide tree (eg. Neighbor-Joining)
3. Alignment of all sequences according to guide tree



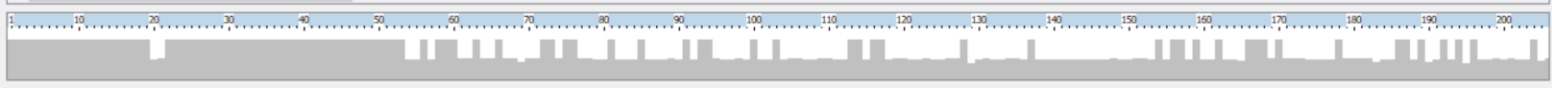
Mode: M...

- Do Complete Alignment Ctrl+L
- Do Guide Tree Only Ctrl+G
- Do Alignment from Guide Tree
- Realign Selected Sequences
- Realign Selected Residue Range
- Align Profile 2 to Profile 1
- Align Profiles from Guide Trees
- Align Sequences to Profile 1
- Align Sequences to Profile 1 from Tree
- Alignment Parameters ▶
- Iteration ▶
- Output Format Options
- Set All Parameters to default

```

*****
GAATCTCTCTCGAAGCATCAAGAGAGGGAGTTATTCDDGACCCAGACACCCAAAGCGTGTATCTAGTTAAACTACTCTTGGGTACATAAAATTACATAGTACATTAGTACATTTATGTATATCGTACATTAAATATAATTCDDGACCCATATAAGCATGTAATTAATTAAGCATAG
GAATCTCTCTCGAAGCATCAAGAGAGGGAGTTATTCDDGACCCAGACACCCAAAGCGTGTATCTAGTTAAACTACTCTTGGGTACATAAAATTACATAGTACATTAGTACATTTATGTATATCGTACATTAAATATAATTCDDGACCCATATAAGCATGTAATTAATTAAGCATAG
GAATCTCTCTCGAAGCATCAAGAGAGGGAGTTATTCDDGACCCAGACACCCAAAGCGTGTATCTAGTTAAACTACTCTTGGGTACATAAAATTACATAGTACATTAGTACATTTATGTATATCGTACATTAAATATAATTCDDGACCCATATAAGCATGTAATTAATTAAGCATAG
GAATCTCTCTCGAAGCATCAAGAGAGGGAGTTATTCDDGACCCAGACACCCAAAGCGTGTATCTAGTTAAACTACTCTTGGGTACATAAAATTACATAGTACATTAGTACATTTATGTATATCGTACATTAAATATAATTCDDGACCCATATAAGCATGTAATTAATTAAGCATAG
GAATCTCTCTCGAAGCATCAAGAGAGGGAGTTATTCDDGACCCAGACACCCAAAGCGTGTATCTAGTTAAACTACTCTTGGGTACATAAAATTACATAGTACATTAGTACATTTATGTATATCGTACATTAAATATAATTCDDGACCCATATAAGCATGTAATTAATTAAGCATAG

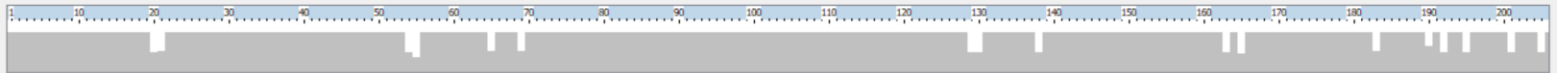
```





Mode: Multiple Alignment Mode Font: 10

```
*****  
EU106282  TGTAAAGCTGAAA TGAAGATCTTCCTCTC GAAGCATCAAGAGAGAGGAACTT TCCDDGACCCCAACACCCCAAGGCTGGTATCTAGTTAAACTACTTCTGGGTACAATAATTACATAGTACACAGTACATTTATGATATCGTACATTAAA TTATTTCCDDCAAGCATATAAGCAGTAAA TTATAAANGATATAA  
EU106283  TGTAAAGCTGAAA TGAAGATCTTCCTCTC GAAGCATCAAGAGAGAGGAACTT TCCDDGACCCCAACACCCCAAGGCTGGTATCTAGTTAAACTACTTCTGGGTACAATAATTACATAGTACACAGTACATTTATGATATCGTACATTAAA TTATTTCCDDCAAGCATATAAGCAGTAAA TTATAAANGATATAA  
EU106281  TGTAAAGCTGAAA TGAAGATCTTCCTCTC GAAGCATCAAGAGAGAGGAACTT TCCDDGACCCCAACACCCCAAGGCTGGTATCTAGTTAAACTACTTCTGGGTACAATAATTACATAGTACATTAAGTACATTTATGATATCGTACATTAAA TTATTTCCDDCAAGCATATAAGCAGTAAA TTAAATTAANGATATAG  
EU106280  TGTAAAGCTGAAA TGAAGATCTTCCTCTC GAAGCATCAAGAGAGAGGAACTT TCCDDGACCCCAACACCCCAAGGCTGGTATCTAGTTAAACTACTTCTGGGTACAATAATTACATAGTACATTAAGTACATTTATGATATCGTACATTAAA TTATTTCCDDCAAGCATATAAGCAGTAAA TTAAATTAANGATATAG  
EU106280  TGTAAAGCTGAAA TGAAGATCTTCCTCTC GAAGCATCAAGAGAGAGGAACTT TCCDDGACCCCAACACCCCAAGGCTGGTATCTAGTTAAACTACTTCTGGGTACAATAATTACATAGTACATTAAGTACATTTATGATATCGTACATTAAA TTATTTCCDDCAAGCATATAAGCAGTAAA TTAAATTAANGATATAG  
EU106283  TGTAAAGCTGAAA TGAAGATCTTCCTCTC GAAGCATCAAGAGAGAGGAACTT TCCDDGACCCCAACACCCCAAGGCTGGTATCTAGTTAAACTACTTCTGGGTACAATAATTACATAGTACATAAGTACATTTATGATATCGTACATTAAA TTATTTCCDDCAAGCATATAAGCAGTAAA TTAAATTAANGATATAG
```



CLUSTAL-Alignment file created [H:/mtDNA/Cytb/Mus.aln]

# Problem with progressive alignment

6 species:

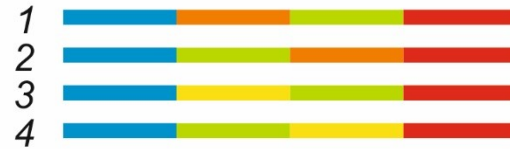
gorilla	AGGTT	penguin	A-GTT
horse	AG-TT	chicken	A-GTT
panda	AG-TT	ostrich	AGGTT



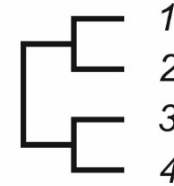
AGGTT	gorilla	AGGTT	AGGTT
AG-TT	horse	AG-TT	A-GTT
AG-TT	panda	AG-TT	A-GTT
AG-TT	penguin	A-GTT	A-GTT
AG-TT	chicken	A-GTT	A-GTT
AGGTT	ostrich	AGGTT	AGGTT



# There are also methods without alignment:



homologní sekvence



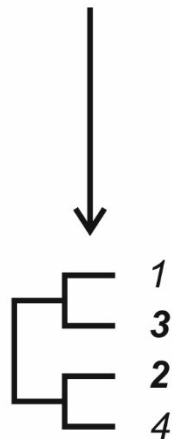
referenční strom



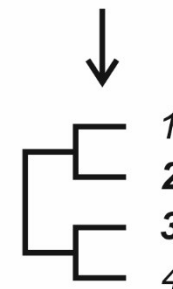
seřazení všech sekvencí



metoda bez seřazení sekvencí



fylogenetický strom



# Methods

## Data types

distances

characters

Methods of tree construction

optimality criteria algorithms

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

# How to assess the methods?

Efficiency:

how fast is the method?

Power:

how many characters we need?

Consistency:

does increasing characters result in true tree?

Robustness:

how does it work when assumptions are violated?

Falsifiability:

does it allow testing assumptions?

# MAXIMUM PARSIMONY, MP (maximální úspornost)

William of Ockham (c. 1287 – 1347)

Occam's razor

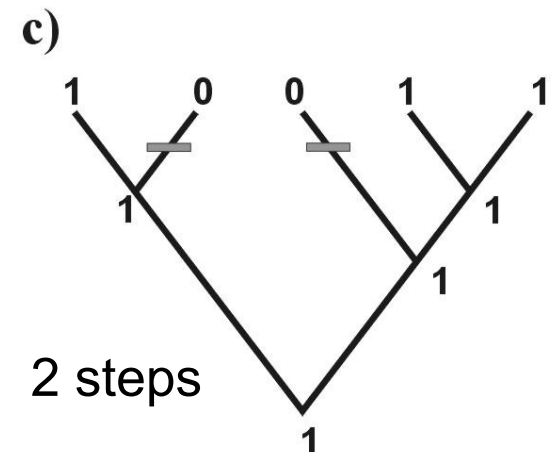
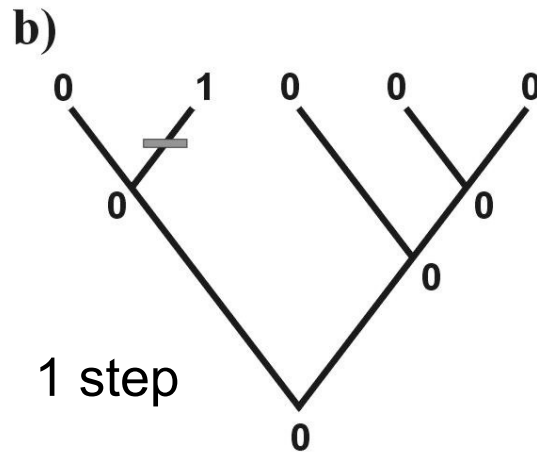
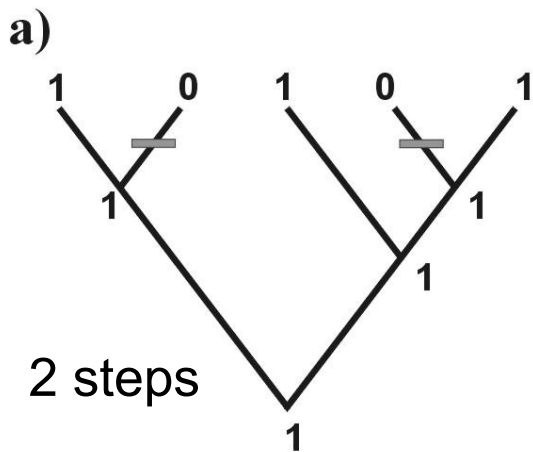


minimal number of steps = 3

real number of steps = 5

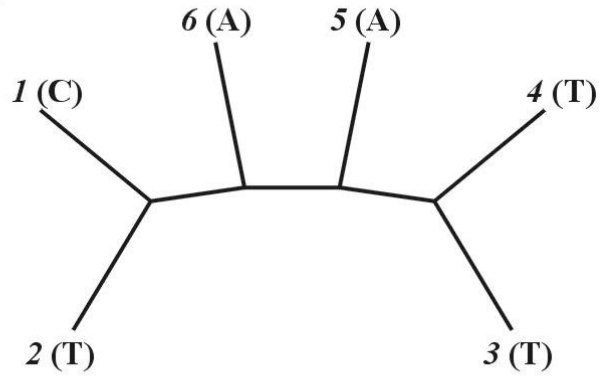
⇒ 2 extra steps → homoplasy

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



# Estimation of number of steps: Fitch algorithm

a)

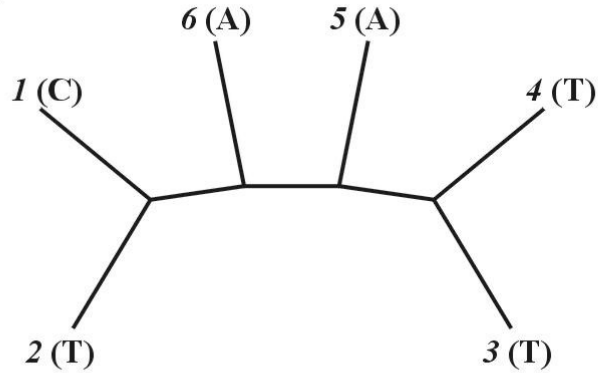


1. arbitrary root

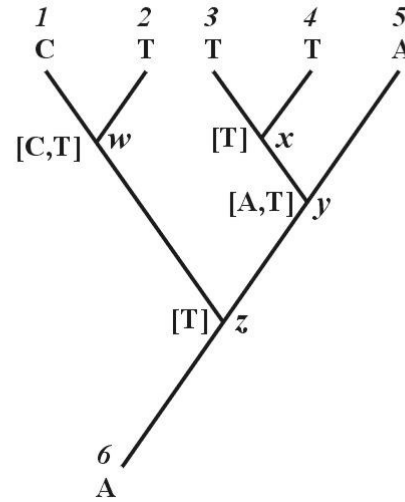


# Estimation of number of steps: Fitch algorithm

a)



b)



1. arbitrary root

2. Downward:

$w = C$  or  $T$

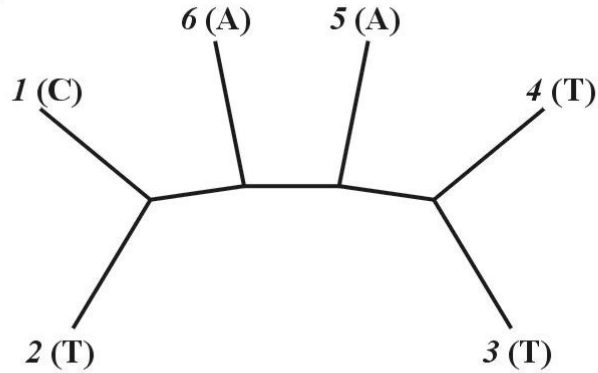
$x = T$

$y = A$  or  $T$

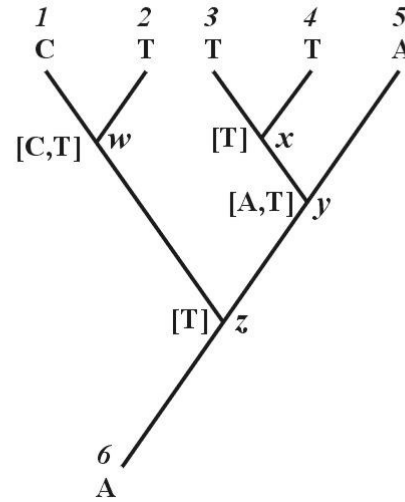
$z = T$

# Estimation of number of steps: Fitch algorithm

a)



b)



1. arbitrary root

2. Downward:

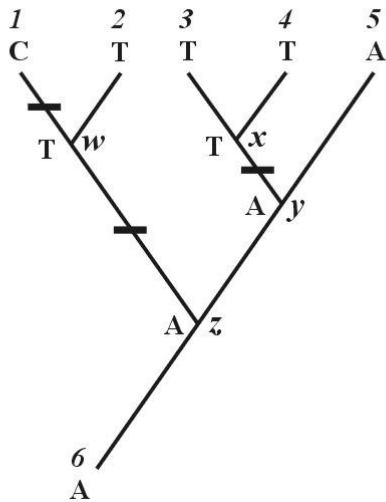
$w = C$  or  $T$

$x = T$

$y = A$  or  $T$

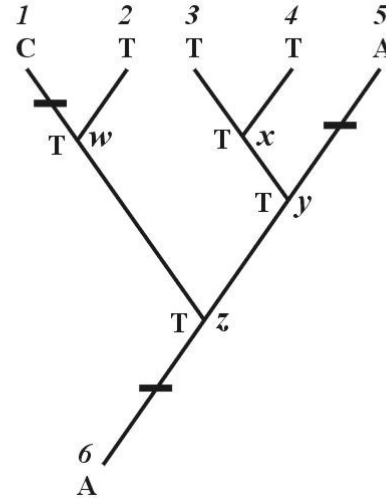
$z = T$

c)



**DELTRAN**  
(**DE**Layed **TR**ANSformation)

d)



**ACCTRAN**  
(**ACC**elerated **TR**ANSformation)

3. Upward:

$z = T$ , nebo  $A$

total length = 3

## Problem of homoplasy:

parsimony-informative and non-informative characters (*sites*)

- invariant sites (*symplesiomorphies*)
- singletons (*autapomorphies*)

index of consistency, CI

retention index, RI

rescaled consistency index, RC

homoplasy index, HI)

$$CI = \frac{s}{m} \quad RI = \frac{g - i}{g - i_{min}}$$

$$RC = CI \times RI$$

$$HI = 1 - CI$$

$m$  = min. no. of possible steps

$s$  = min. no. needed for explaining the tree

$g$  = max. no. of steps for any tree

## Methods of parsimony:

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

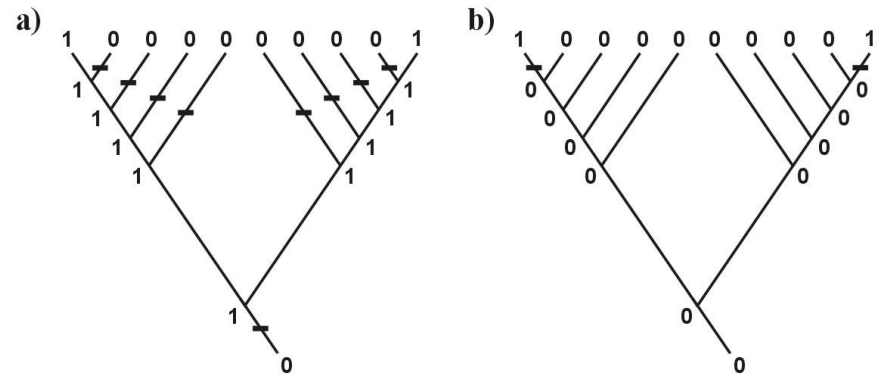
**Wagner:**  $X \rightarrow Y$  a  $Y \rightarrow X$   
seřazené znaky ( $1 \rightarrow 2 \rightarrow 3$ )

**Dollo:**  $X \rightarrow Y$  a  $Y \rightarrow X$ , potom nelze  $X \rightarrow Y$

... restriction-site and  
restriction-fragment data

**Camin-Sokal:**  $X \rightarrow Y$ ,  
not  $Y \rightarrow X$

... SINE, LINE



“relaxed Dollo criterion”

weighted = transversion p.

generalized p.: cost matrix = step matrix

Wagner

a)

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

Fitch

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)

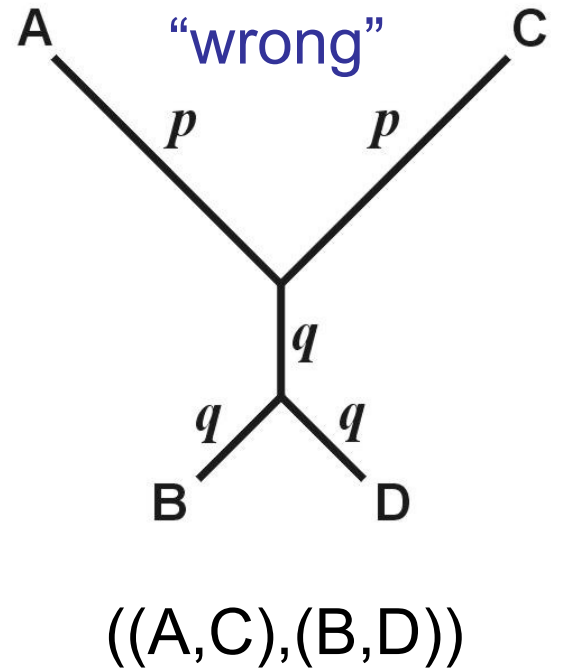
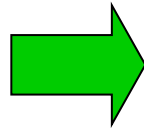
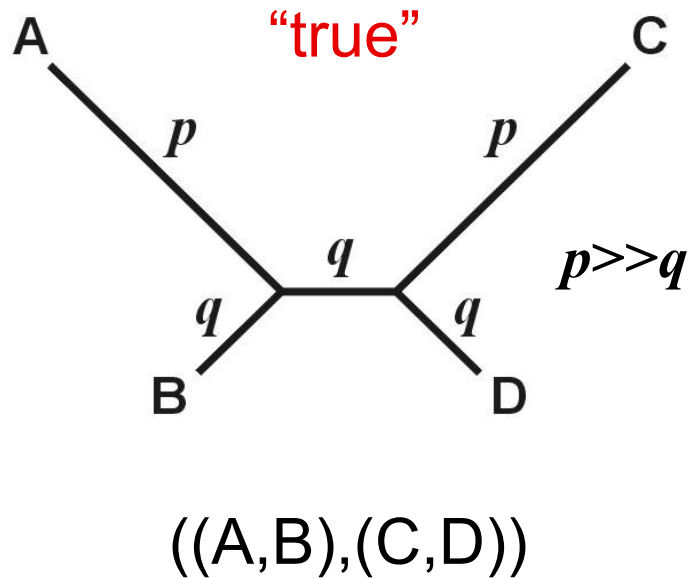
	A	C	G	T
A	-	5	1	5
C	5	-	5	1
G	1	5	-	5
T	5	1	5	-

Dollo

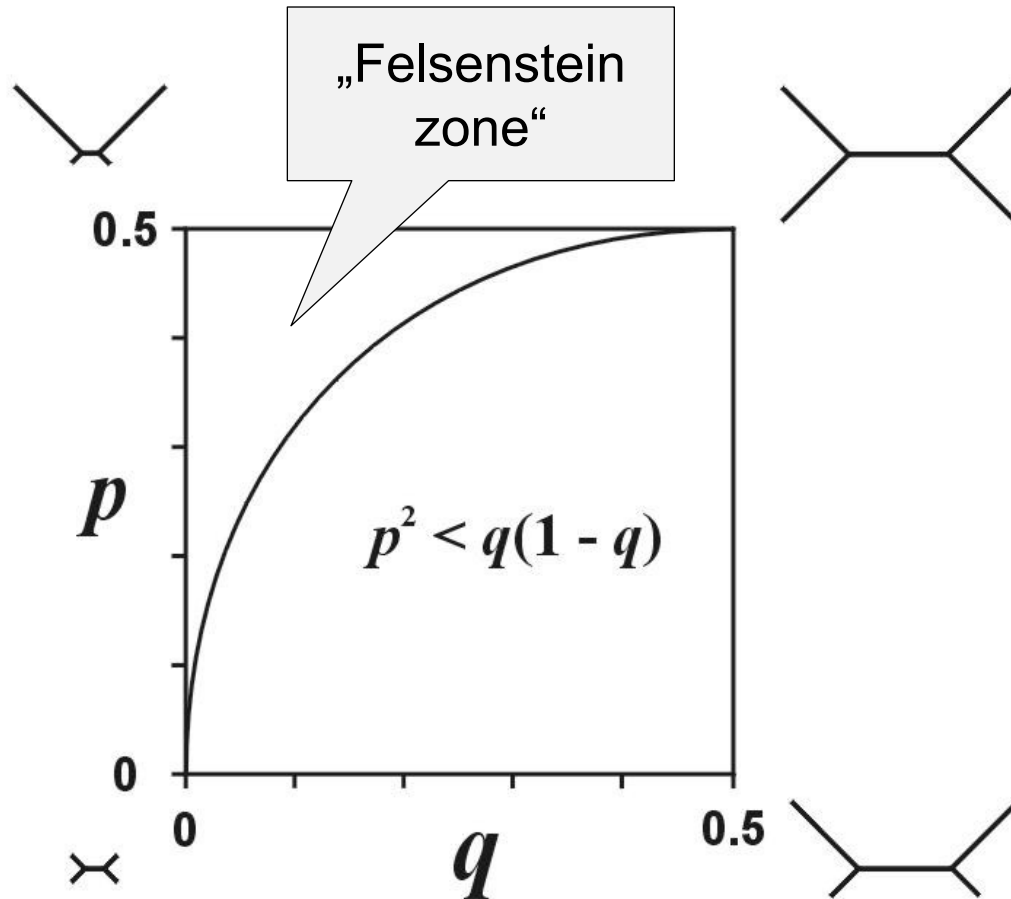
transversion

\*)  $M$  is an arbitrarily large number, guaranteeing that only one transformation to each derived state will be permitted.

# Parsimony and consistency

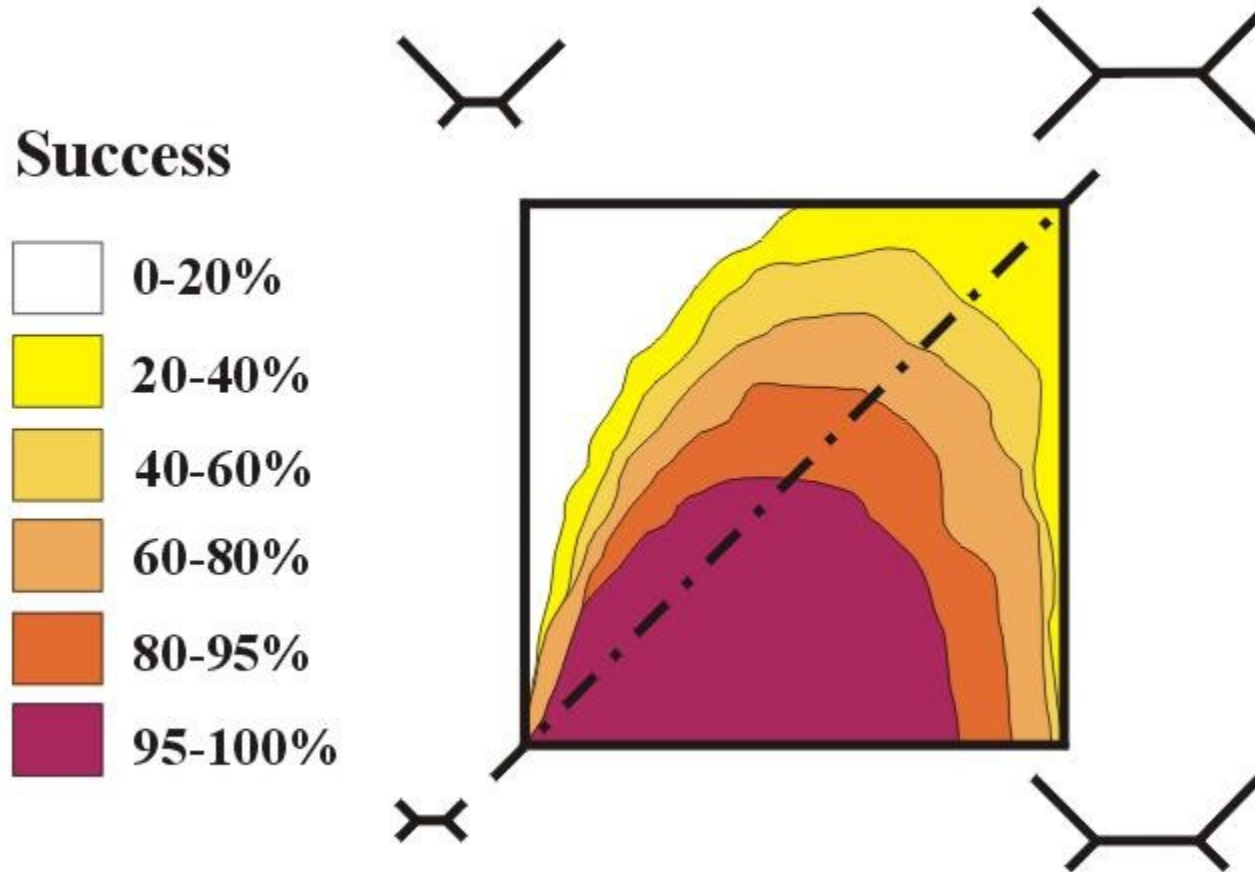


# Parsimony and consistency



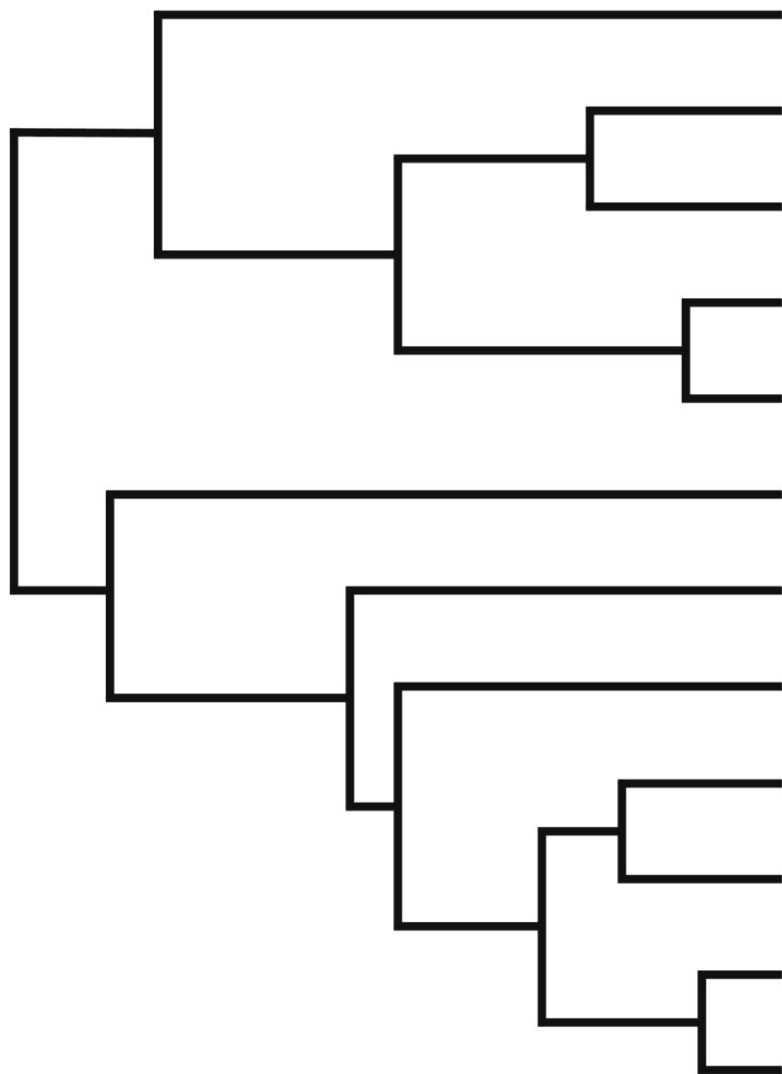
In the Felsenstein zone, parsimony is inconsistent

# Parsimony and consistency



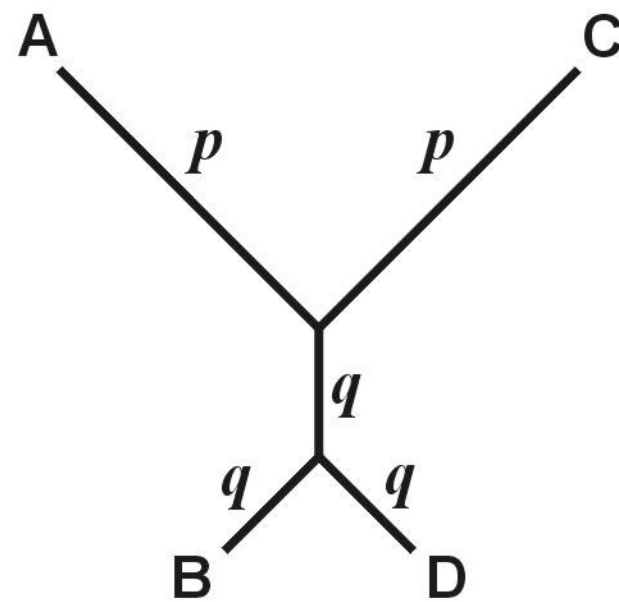
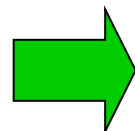


# Parsimony and consistency



long branches

long-branch attraction (LBA)

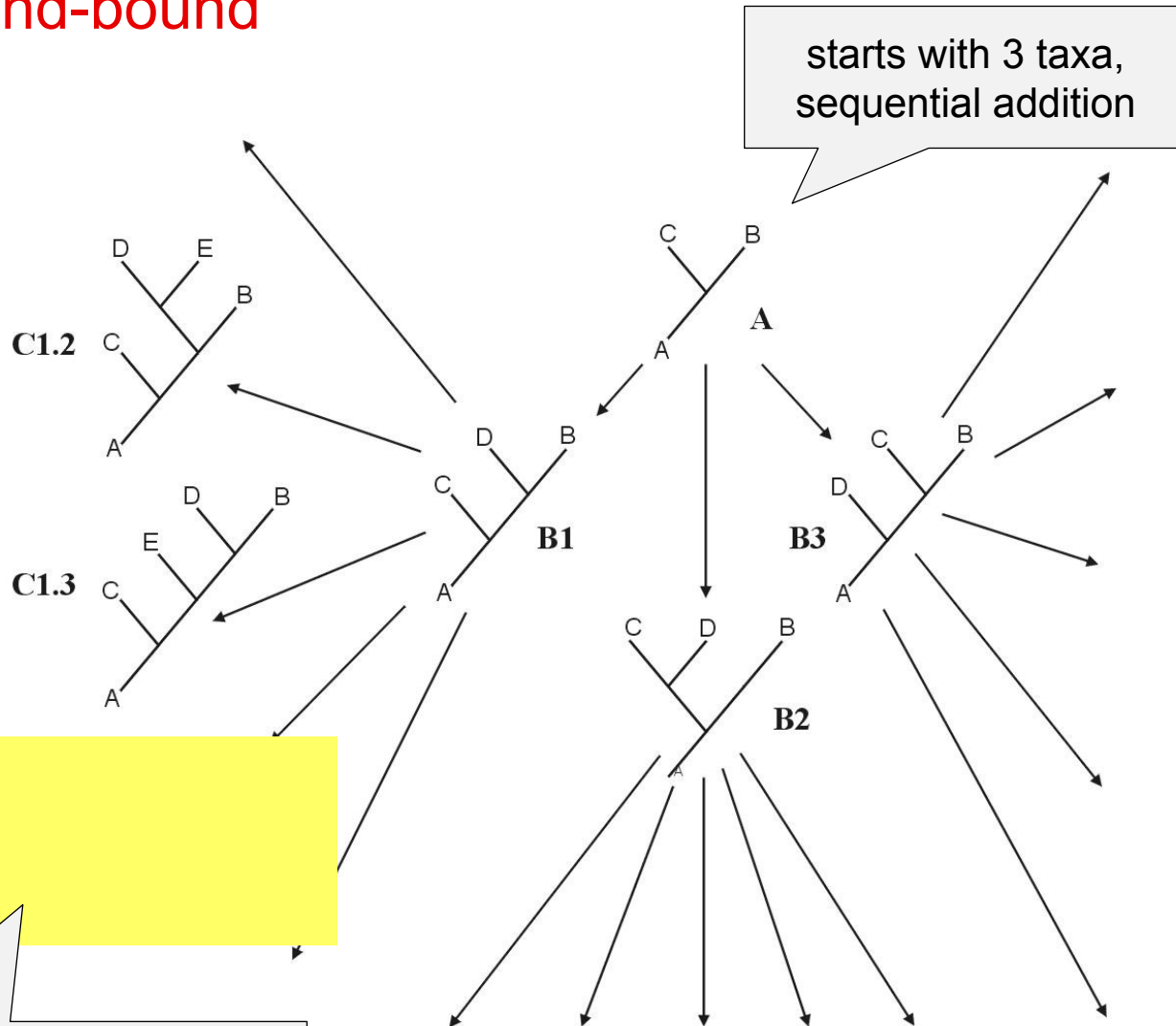


# Search for optimal tree

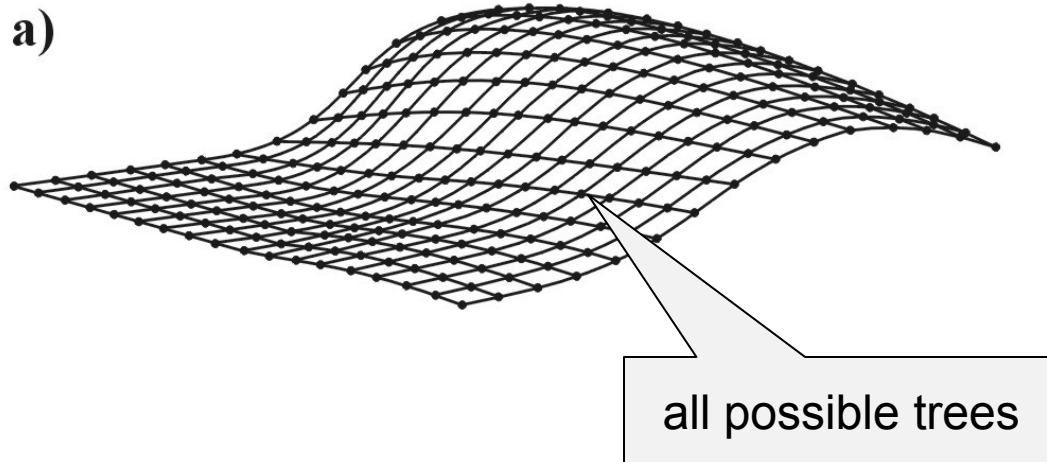
## 1. Exact methods:

- a) exhaustive search
- b) branch-and-bound

# branch-and-bound

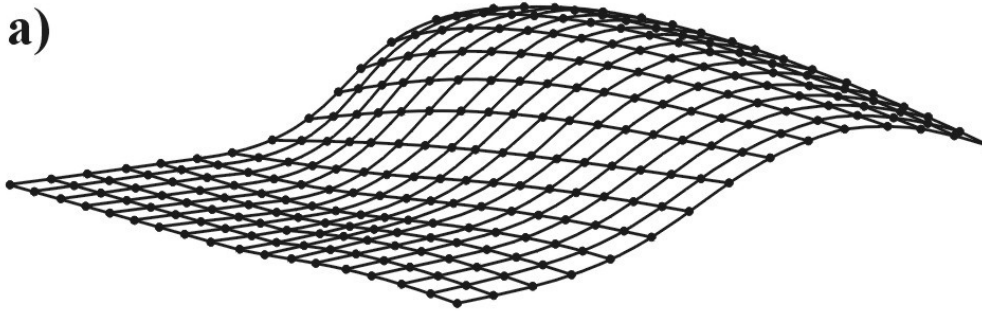


## 2. Heuristic search

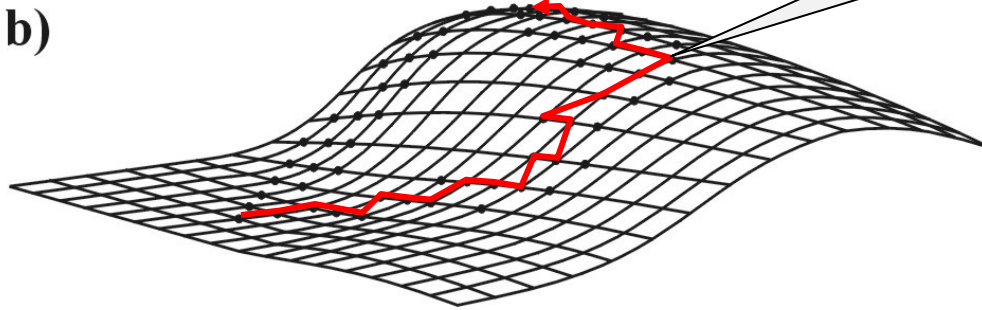


stepwise addition  
star decomposition  
branch swapping

a)

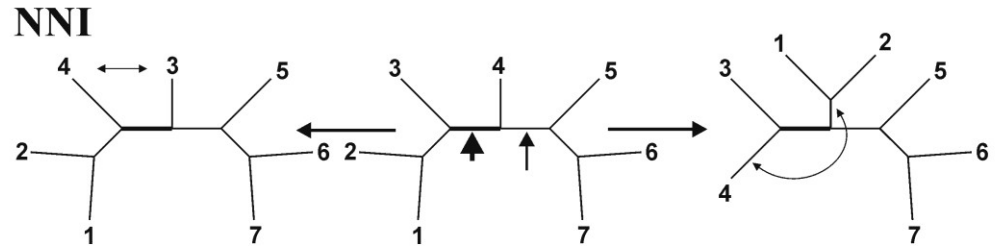


b)

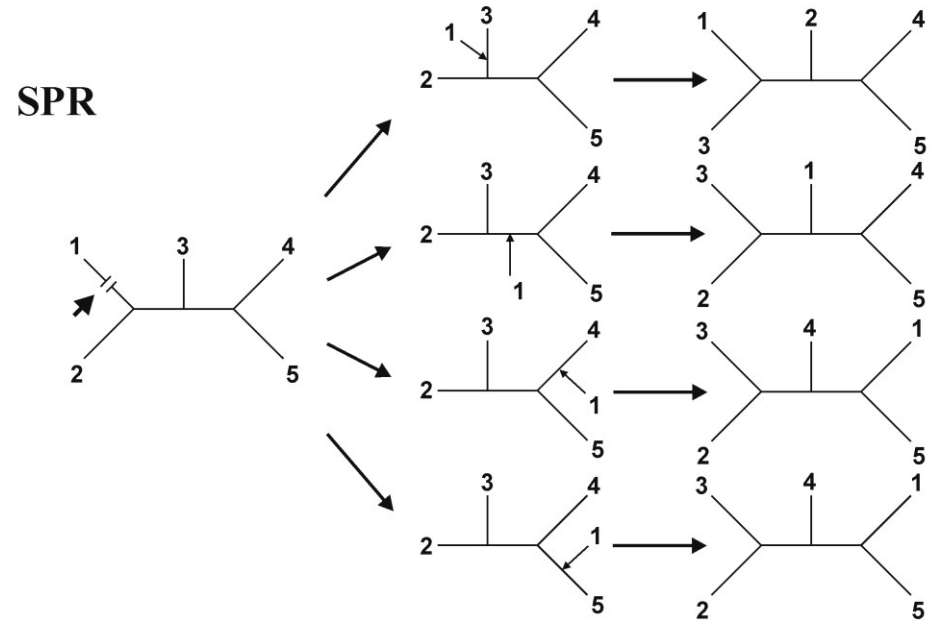


heuristic search

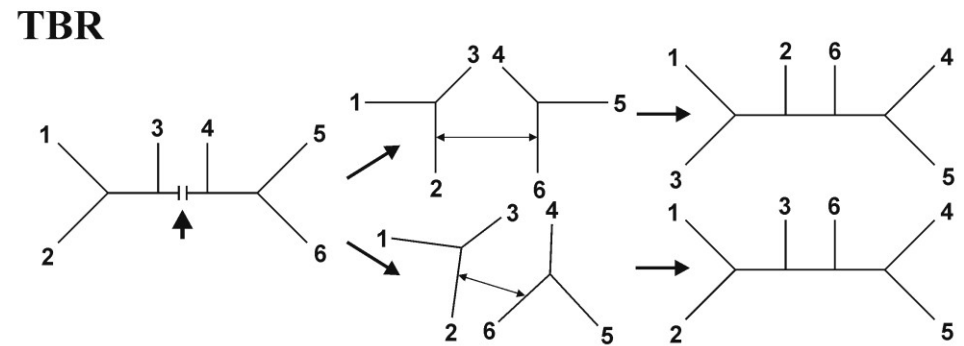
nearest-neighbor  
interchanges (NNI)



subtree pruning  
and regrafting (SPR)



tree bisection and  
reconnection (TBR)



# Evolutionary models and distance methods

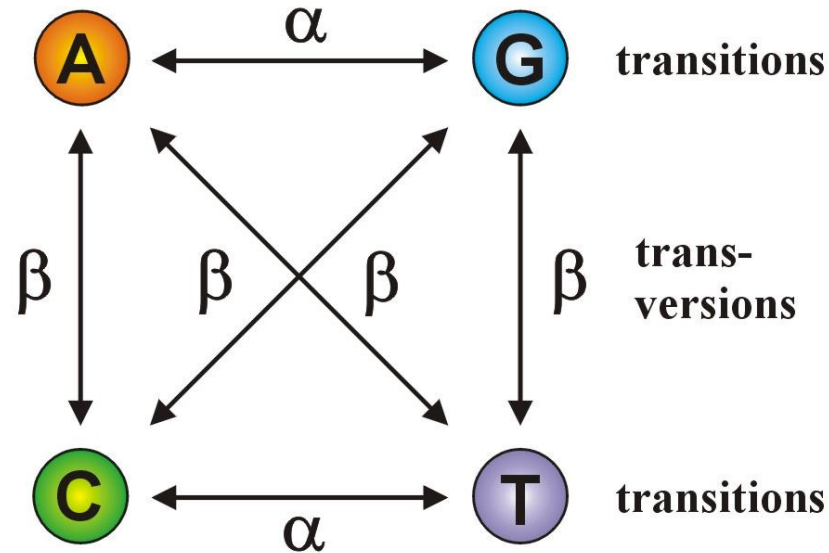
		Base after substitution			
		A	C	G	T
Original base	A	$-\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	C	$\frac{1}{4}$	$-\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
	G	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{3}{4}$	$\frac{1}{4}$
	T	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{3}{4}$

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

**Jukes-Cantor (JC):**

equal base frequencies  
equal substitution rates

## Kimura 2-parameter (K2P): transitions $\neq$ transversions



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

If  $\alpha = \beta$ , K2P = JC



**Felsenstein (F81):** different base frequencies

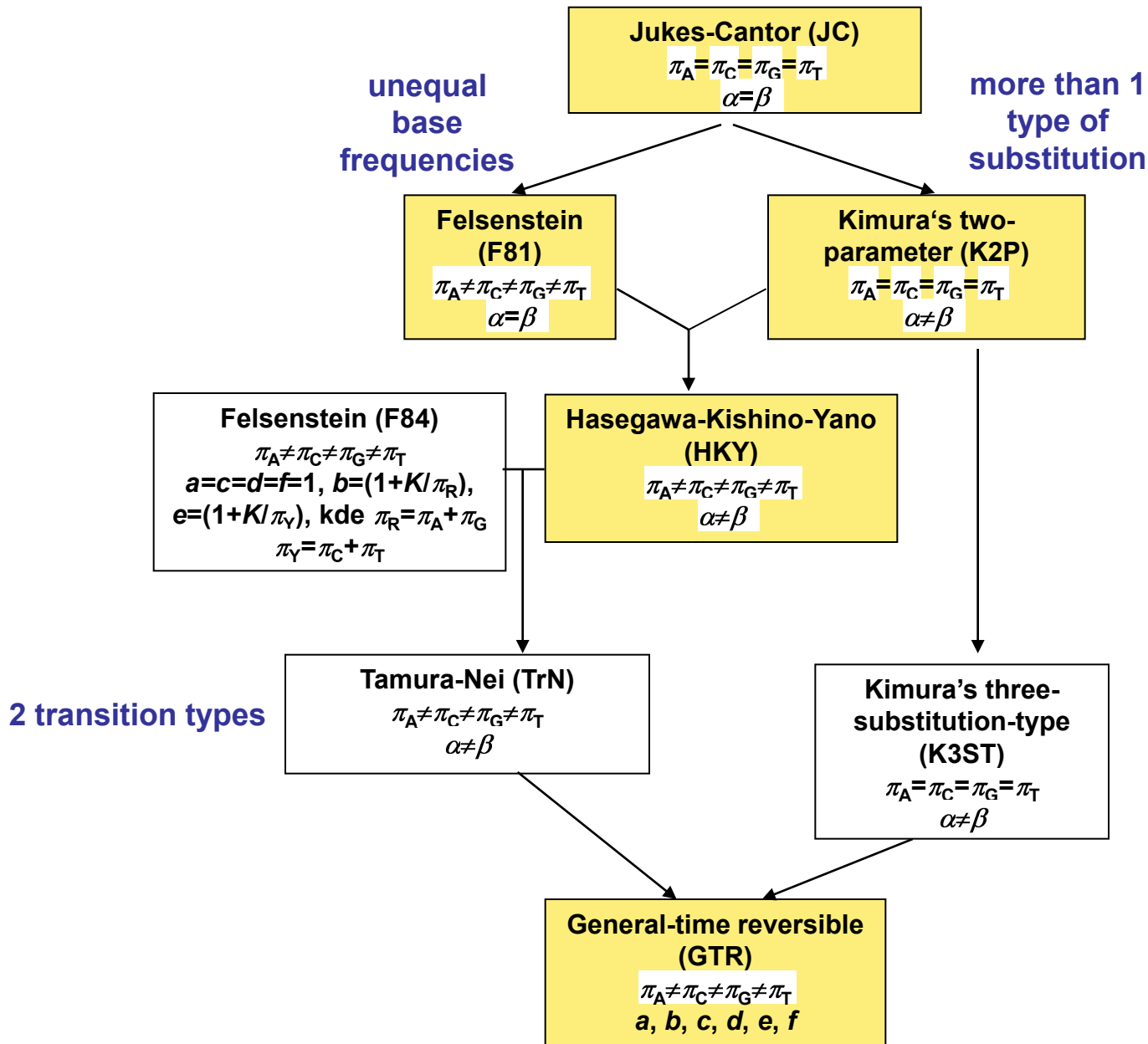
$$Q = \begin{pmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{pmatrix}$$

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

**Hasegawa-Kishino-Yano (HKY):** different base frequencies  
transitions  $\neq$  transversions

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

**General time-reversible (GTR, REV):** different base frequencies  
different substitution rates



# Heterogeneity of substitution rates in different parts of sequences

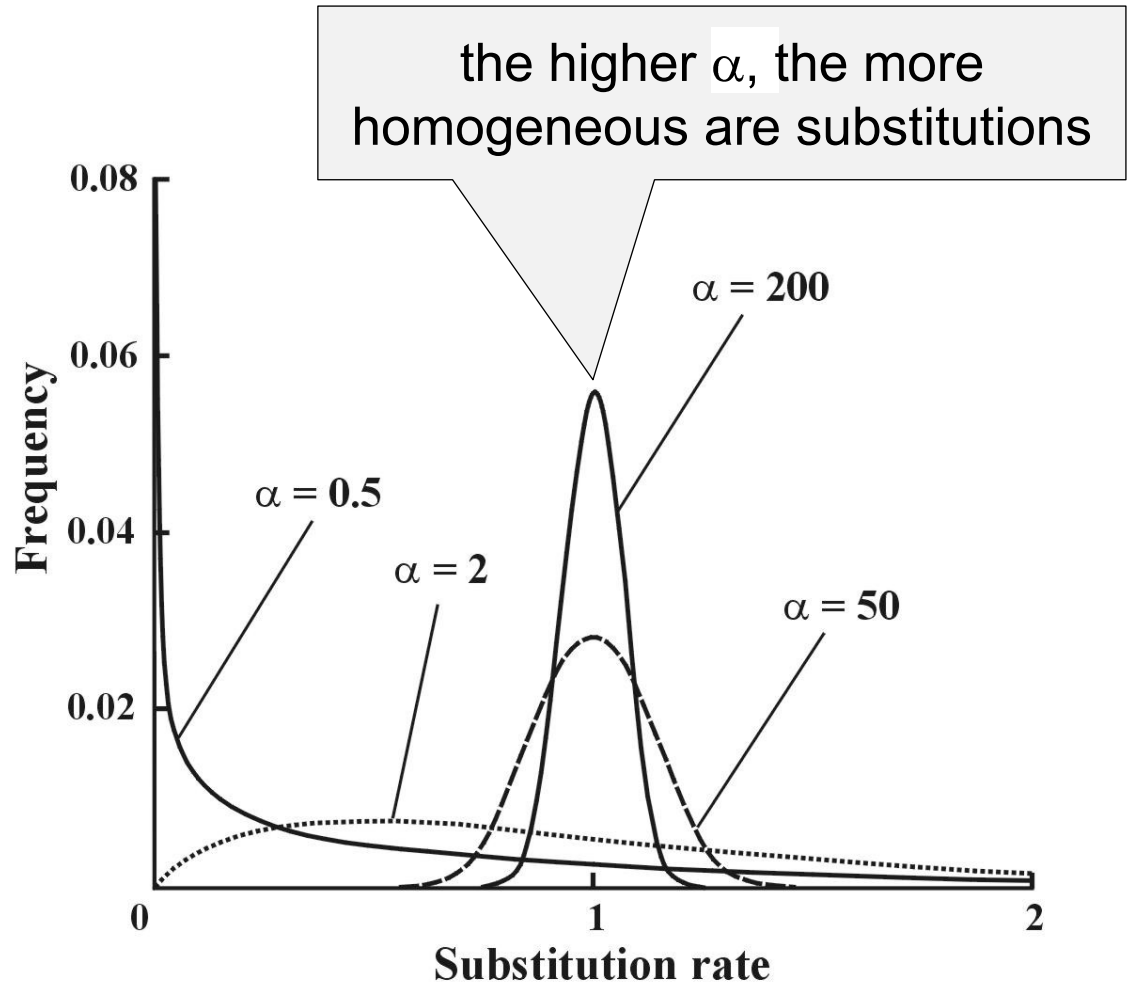
Gamma ( $\Gamma$ ) distribution:

shape parameter  $\alpha$

discrete gamma model

invariant sites

→ GTR+ $\Gamma$ +I



## Model comparison:

### Likelihood ratio test (LRT):

nested models

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

$\chi^2$  distribution,  $p_2 - p_1$  degrees of freedom

### Akaike information criterion (AIC):

nonnested models

$$AIC = -2\ln L + 2p, \text{ kde } p = \text{number of free parametres}$$

better model  $\rightarrow$  lower *AIC*

### Bayesian information criterion (BIC):

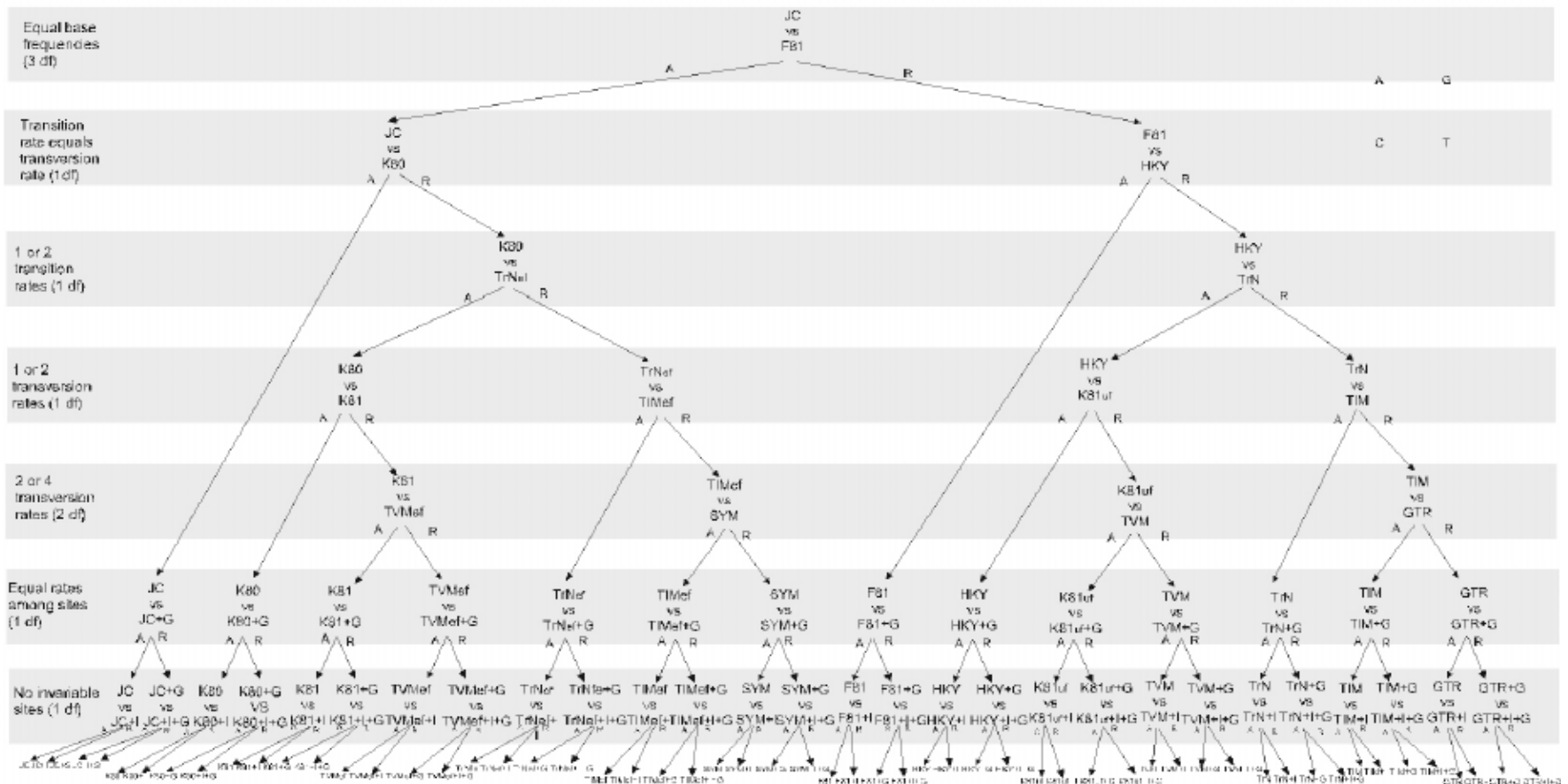
nonnested models

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

# Model comparison:

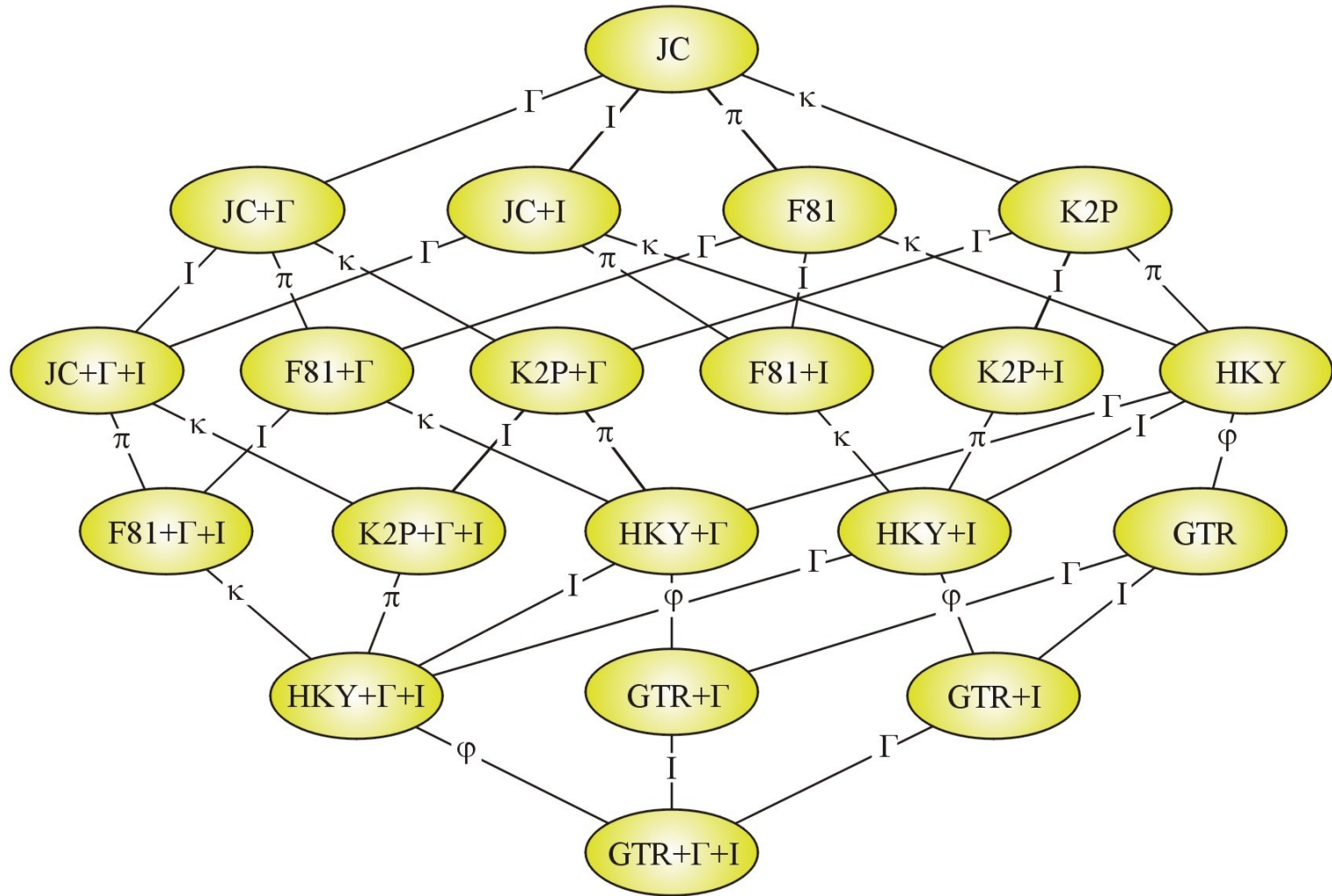
hierarchical LRT – ModelTest (Crandall and Posada), jModelTest

**Modeltest 3.0 hierarchy**

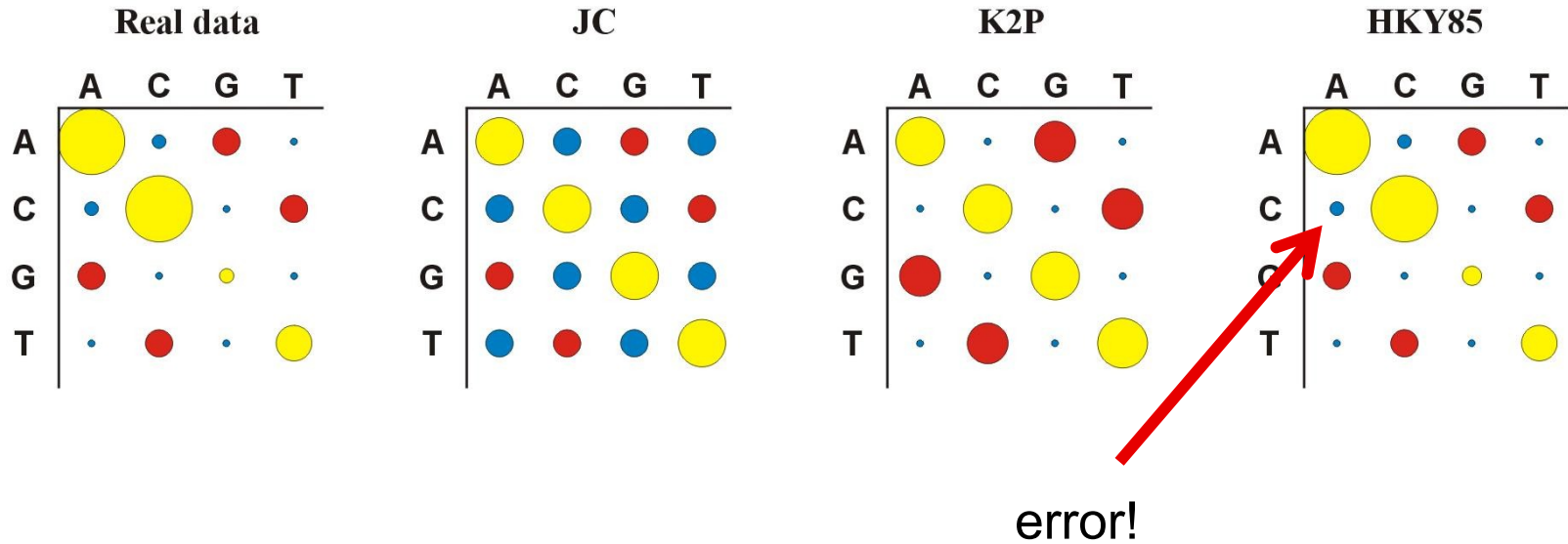


# Model comparison:

dynamic LRT:



# Model comparison:



More parameters  $\Rightarrow$  more realism, but ...

... also less confidence (estimates based on the same amount of data!)

# Distances

computed for each pair of taxa, from distance (or similarity) matrix  
– tree inference

distance methods base on assumption that if we know true distances,  
we can very easily infer the true phylogeny

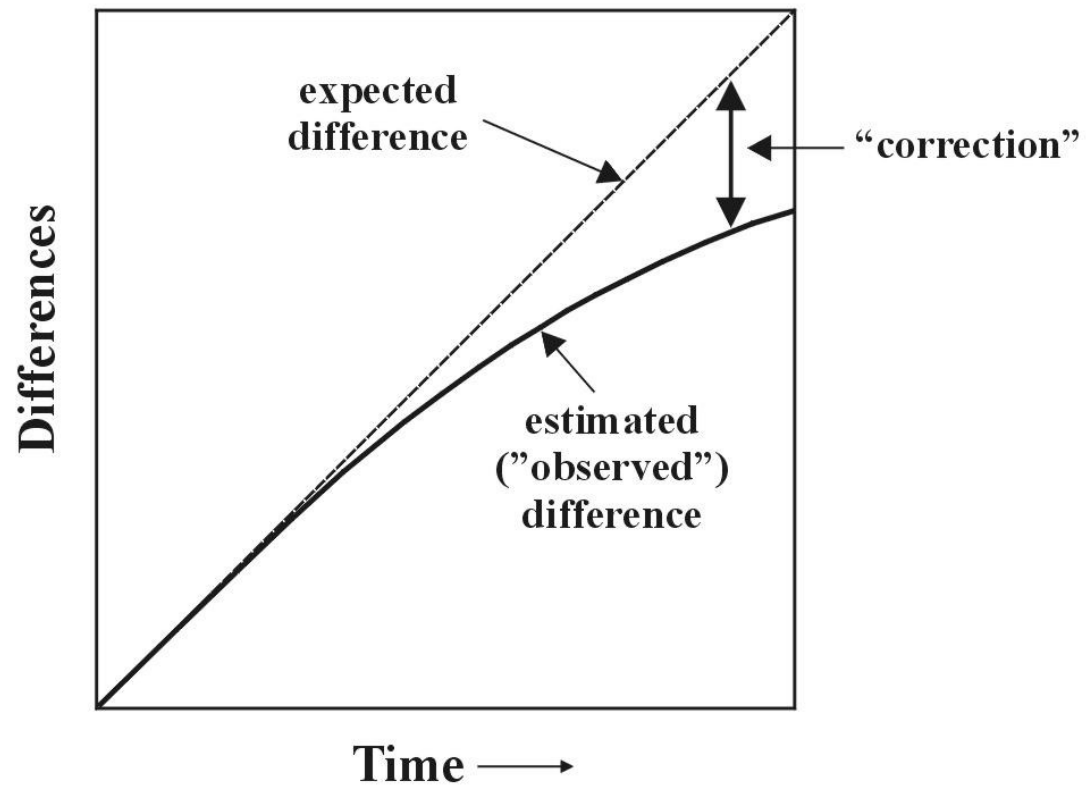
advantage: very fast and simple (also with a calculator)



1
10
20
30  
 sequence 1: ACCCGTTAAGCTTAACGTACTTGGATCGAT  
 sequence 2: ACCCGTTAGGCTTAATGTACGTGGATCGAT

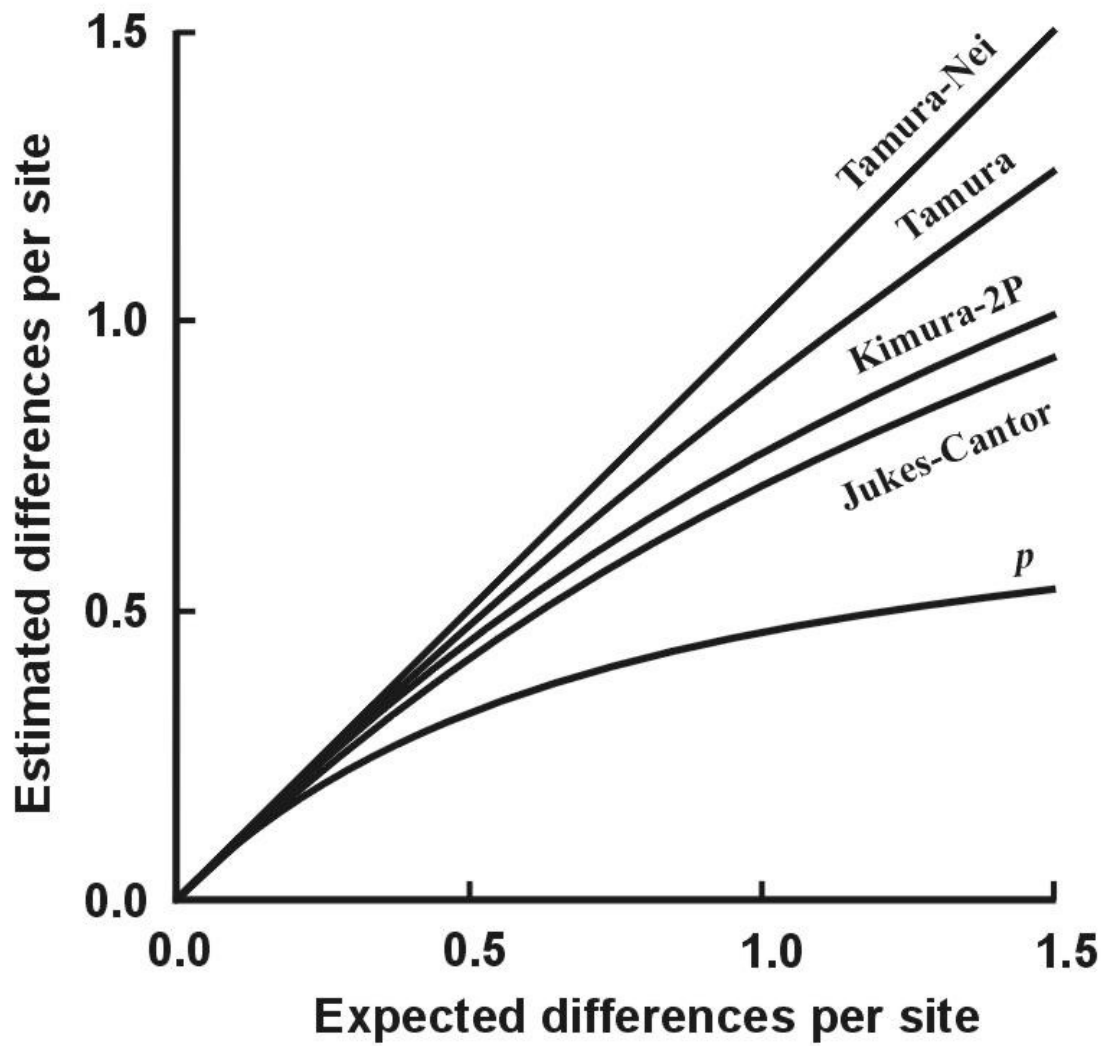
*p*-distance:  $p = k/n = 3/30 = 0,10$

problem of saturation:



## Distances for some models:

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



## Cluster analysis - UPGMA

	chimp	bonobo	gorilla	human	orang.
chimp (Š) --					
bonobo (B)	0,0118	--			
gorilla (G)	0,0427	0,0416	--		
human (Č)	0,0382	0,0327	0,0371	--	
orangutan (O)	0,0953	0,0916	0,0965	0,0928	--

1. Find min  $d(ij)$
2. Calculate new matrix  $(\check{S}B-k) = [d(B-k)+d(\check{S}-k)]/2$
3. Repeat 1 a 2.

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

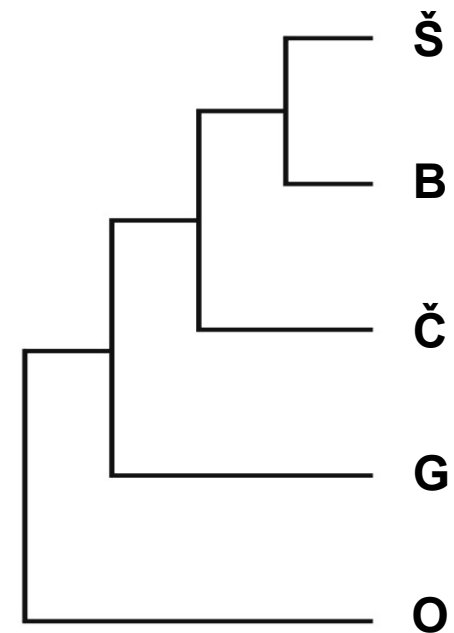
UPGMA (unweighted pair-group method using arithmetic means):

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

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

single-linkage (metoda nejbližšího souseda)

complete-linkage (m. nejvzdálenějšího souseda)

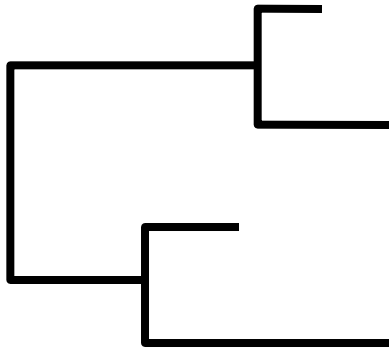
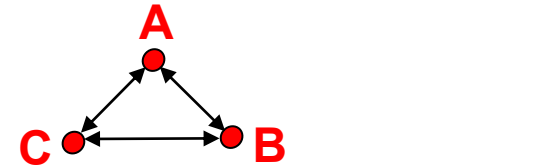


# UPGMA and consistency

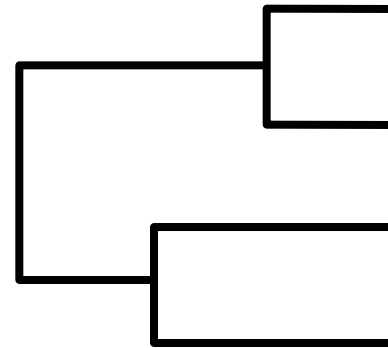
**additive distances:**  $d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC})$

tj. distance between 2 taxa equals sum of branches connecting them

**ultrametric distances:**  $d_{AC} \leq \max(d_{AB}, d_{BC})$

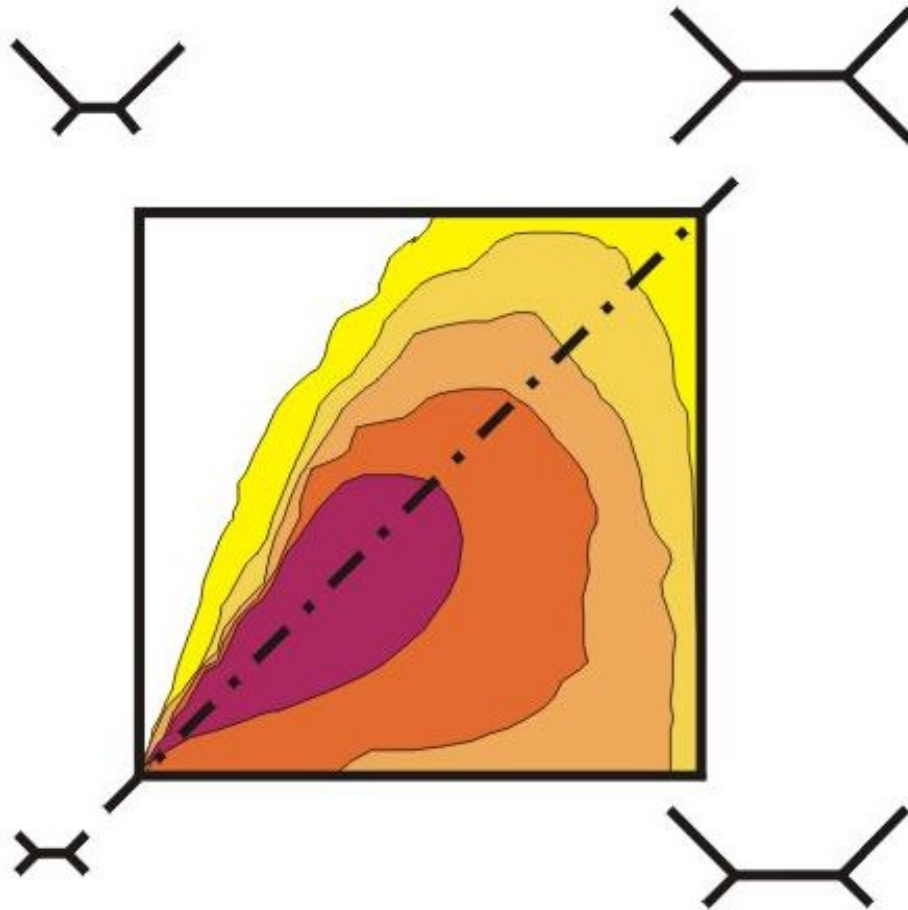


additive tree



ultrametric tree

# UPGMA and consistency



## Neighbor-Joining, NJ

Algorithmic method

Principle of minimal evolution → minimizes sum of branch lengths  $S$

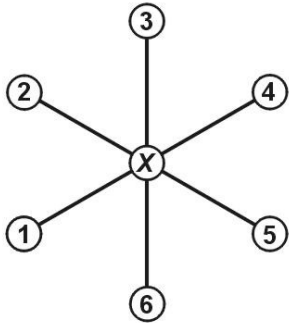
Each pair of nodes adjusted according to its divergence from others

Single additive tree



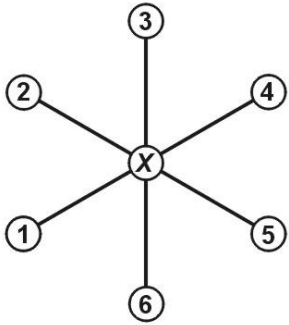
star tree

a)



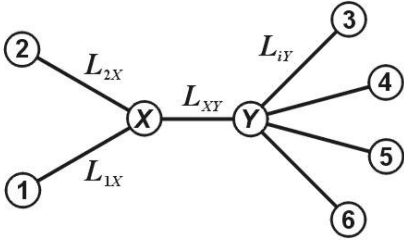
star tree

a)



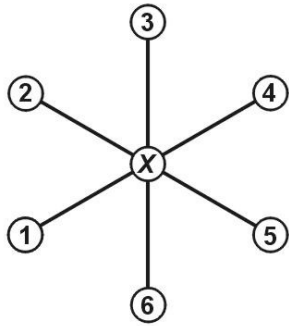
finding nearest neighbors

b)



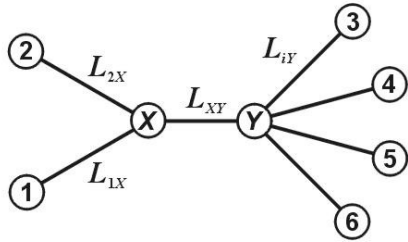
star tree

a)



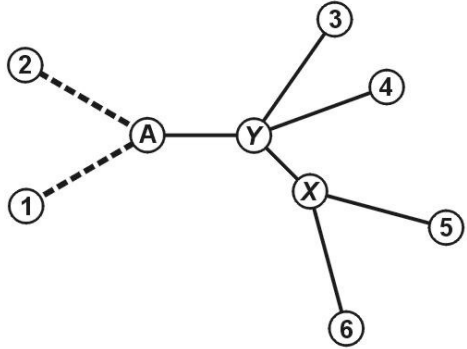
finding nearest neighbors

b)



distance recalculation

c)

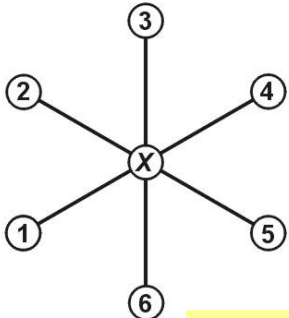


star tree

finding nearest neighbors

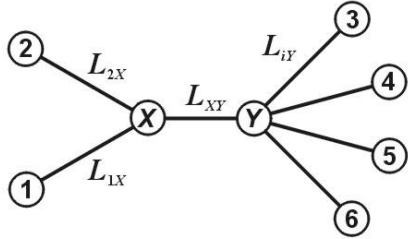
distance recalculation

a)

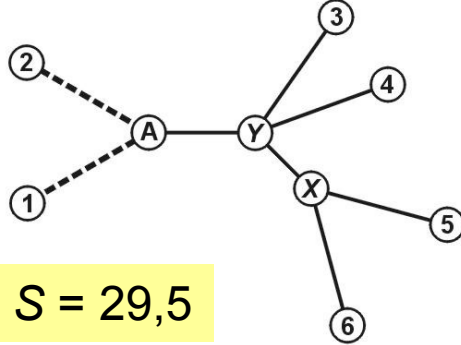


$S = 32,4$

b)

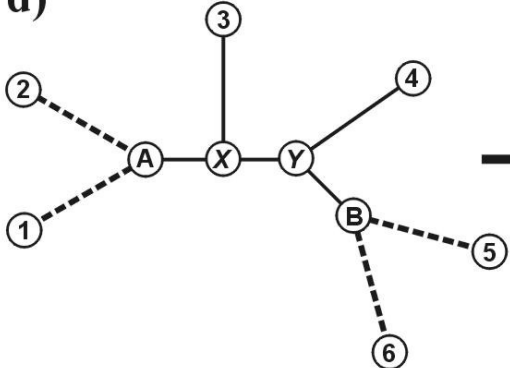


c)

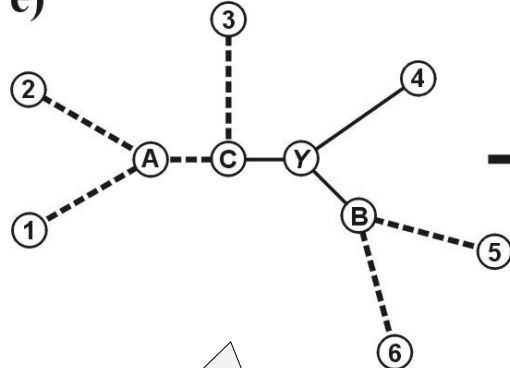


$S = 29,5$

d)

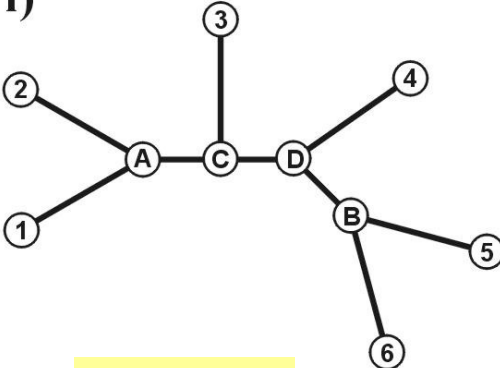


e)



repeating...

f)



$S = 28,0$

## Drawbacks of distance data:

1. loss of information during transformation
2. after transformation to distances, we cannot infer original data (different sequences may result in the same distance)
3. we cannot study the evolution in different parts of sequence
4. difficult biological interpretation of branch lengths
5. we cannot combine more distance matrices