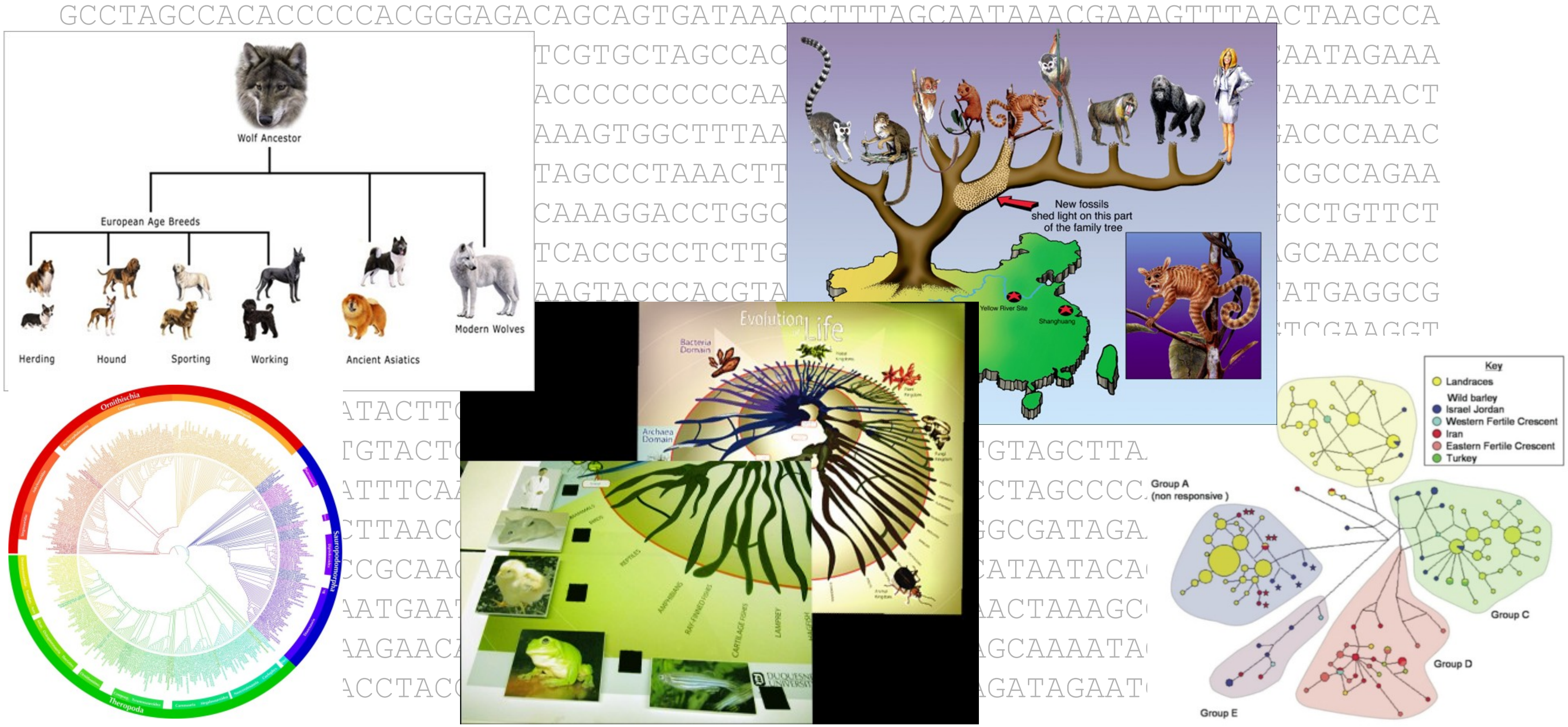
MODULARIZACE VÝUKY EVOLUČNÍ A EKOLOGICKÉ BIOLOGIE
CZ.1.07/2.2.00/15.0204

# PHYLOGENETIC ANALYSIS I.



evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

# Definition of basic concepts:

phylogenetic tree = phylogeny (fylogenie): rooted, unrooted

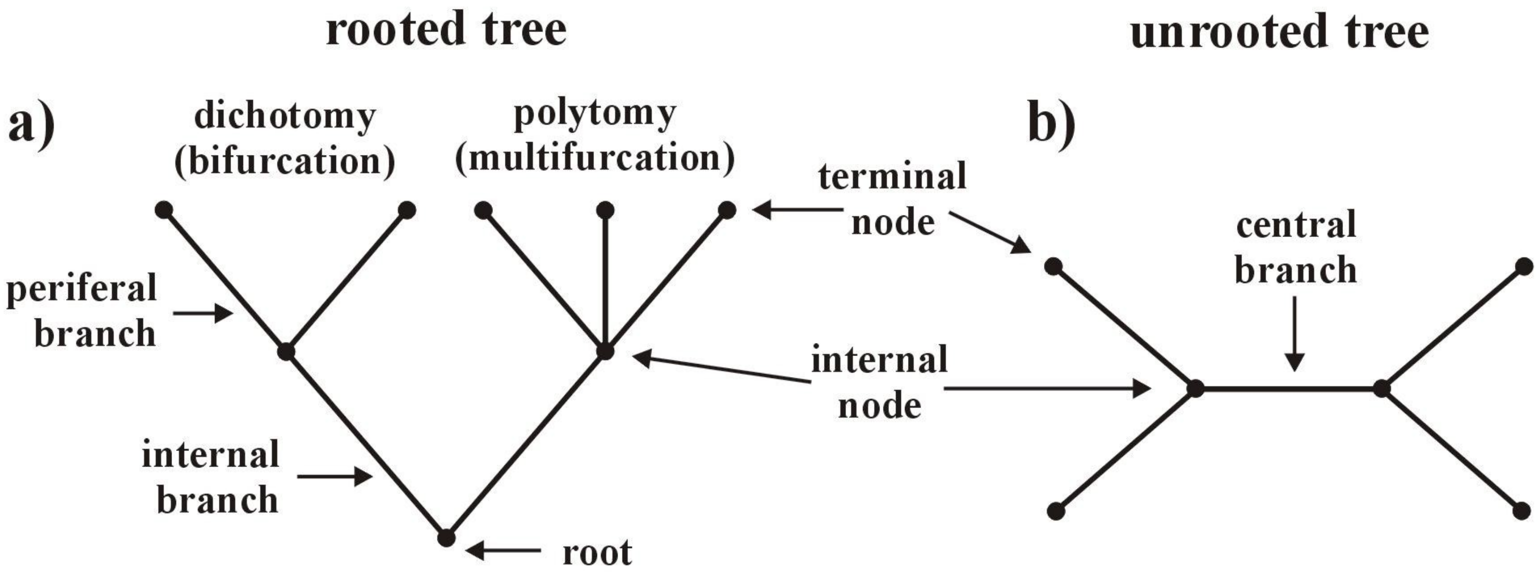
branches = edges (větve): peripheral, internal, central

nodes = vertices (uzly): internal, terminal

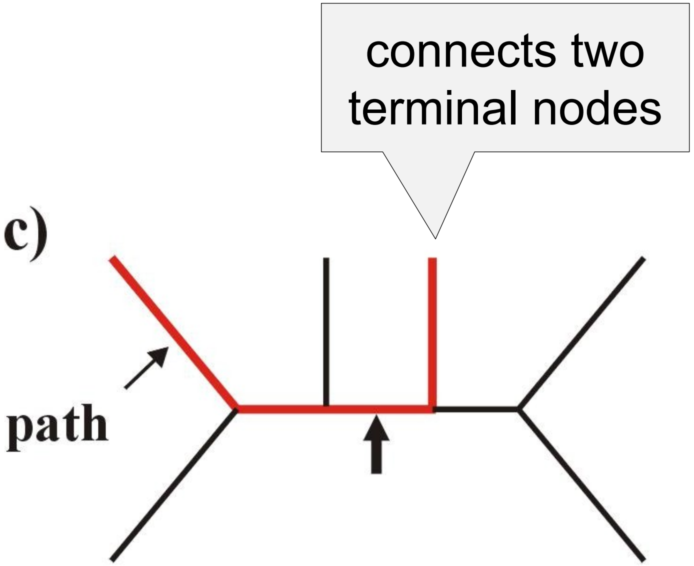
dichotomy = bifurcation, polytomy = multifurcation

OTU = operational taxonomic unit, HTU = hypothetical taxonomic unit

tree topology

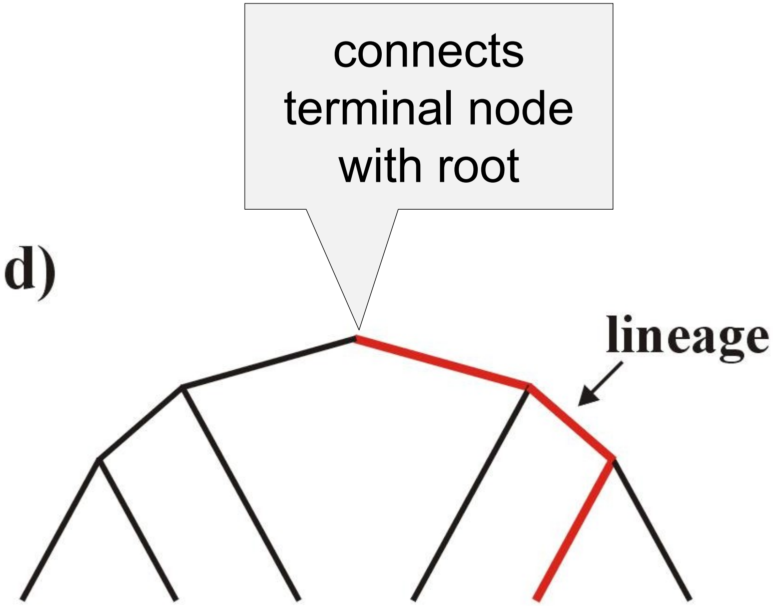


# Definition of basic concepts:



connects two terminal nodes

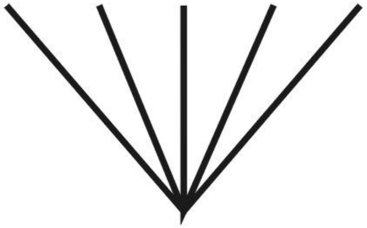
path (dráha)



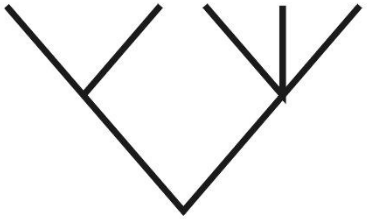
connects terminal node with root

lineage (linie)

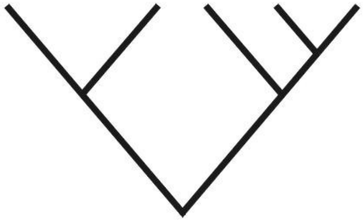
# Definition of basic concepts:



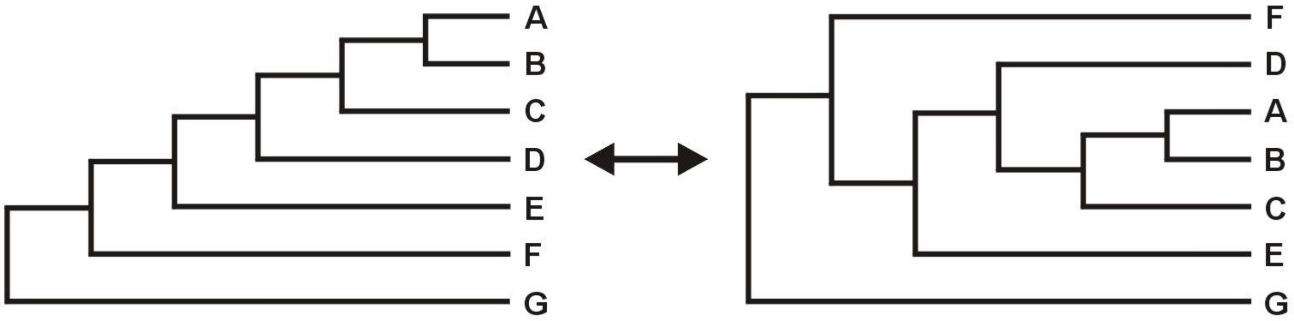
star tree

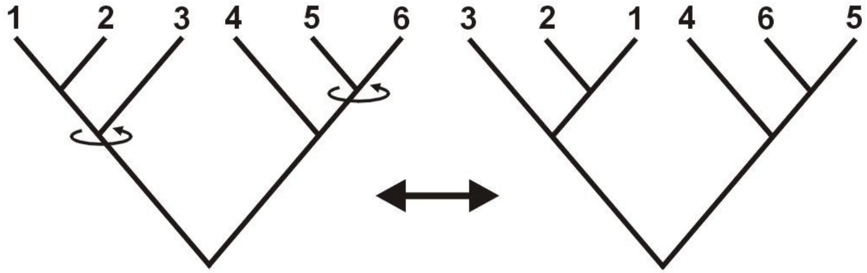


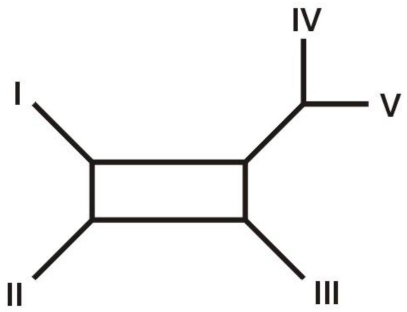
partly resolved



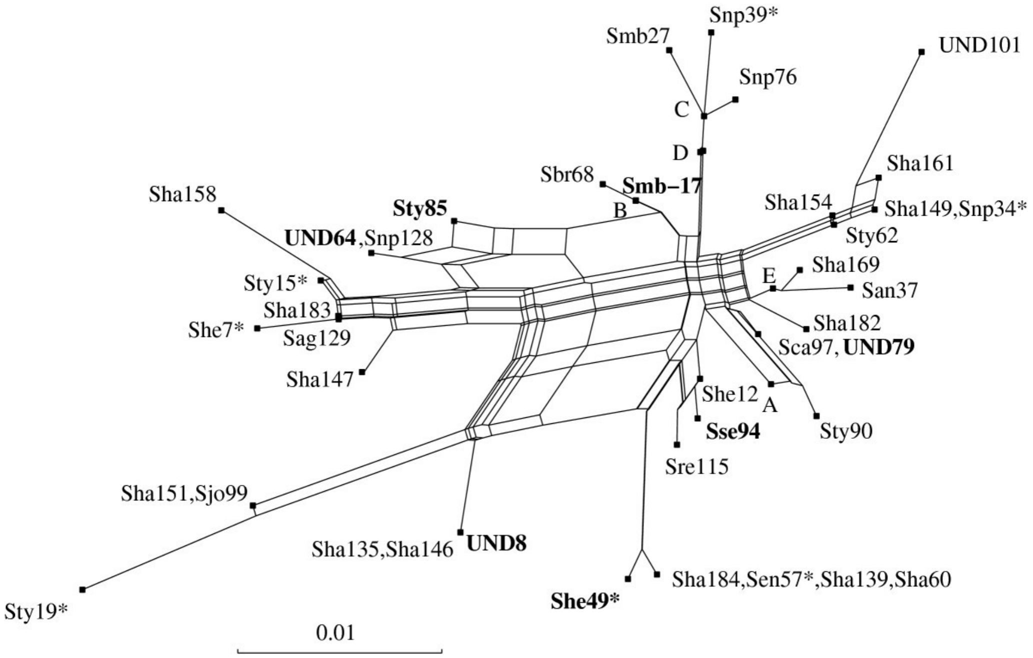
fully resolved

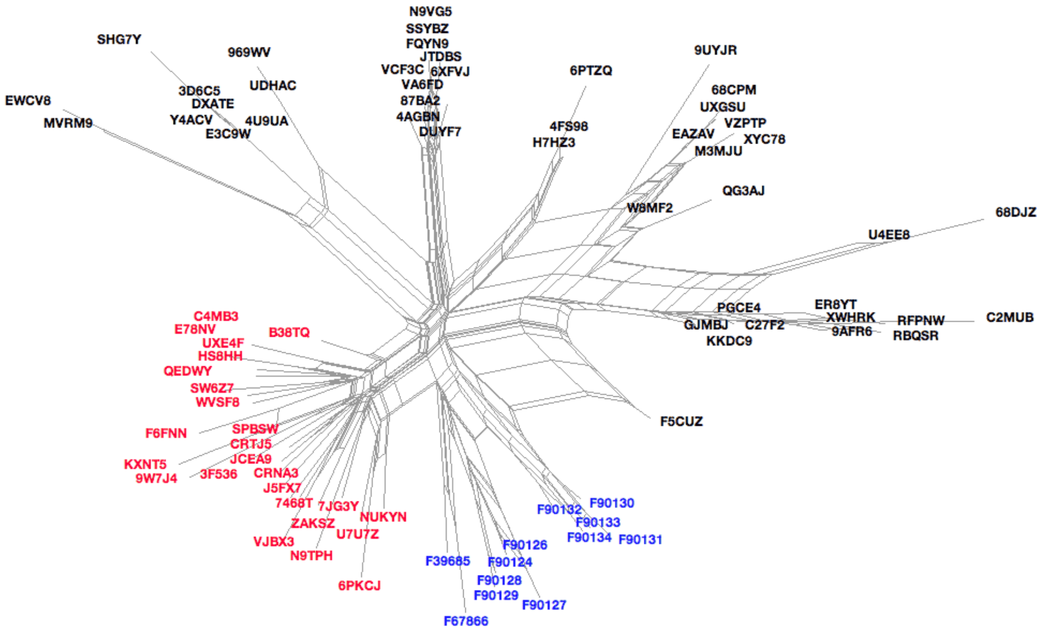




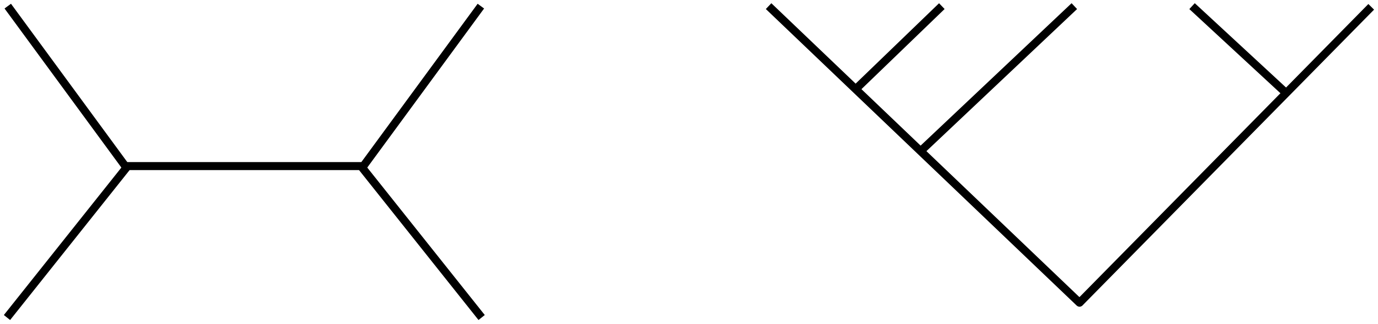


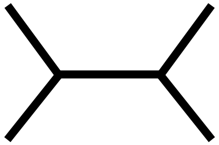
network

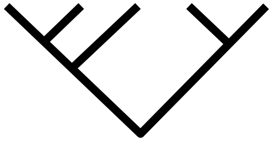

Snp39*
Smb27
UND101
Snp76
C
D
Sha161
Sbr68
Smb–17
Sha158
Sty85
Sha154
Sha149,Snp34*
B
UND64,Snp128
Sty62
Sha169
E
San37
Sty15*
Sha183
Sha182
She7*
Sag129
Sca97, UND79
Sha147
She12
A
Sse94
Sty90
Sre115
Sha151,Sjo99
Sha135,Sha146
UND8
Sha184,Sen57*,Sha139,Sha60
She49*
Sty19*
0.01


N9VG5
SSYBZ
SHG7Y
FQYN9
969WV
JTDBS
VCF3C
6XFVJ
6PTZQ
9UYJR
3D6C5
UDHAC
VA6FD
68CPM
EWCV8
DXATE
87BA2
UXGSU
MVRM9
Y4ACV
4U9UA
4AGBN
4FS98
VZPTP
E3C9W
DUYF7
H7HZ3
EAZAV
XYC78
M3MJU
QG3AJ
W8MF2
68DJZ
U4EE8
C4MB3
PGCE4
ER8YT
XWHRK
C2MUB
E78NV
B38TQ
GJMBJ
C27F2
RFPNW
UXE4F
9AFR6
HS8HH
RBQSR
KKDC9
QEDWY
SW6Z7
WVSF8
F5CUZ
F6FNN
SPBSW
CRTJ5
JCEA9
KXNT5
3F536
9W7J4
CRNA3
J5FX7
7468T
7JG3Y
F90132
F90130
NUKYN
ZAKSZ
F90133
U7U7Z
F90134
F90131
VJBX3
F90126
N9TPH
F39685
F90124
F90128
6PKCJ
F90129
F90127
F67866

# How many trees?





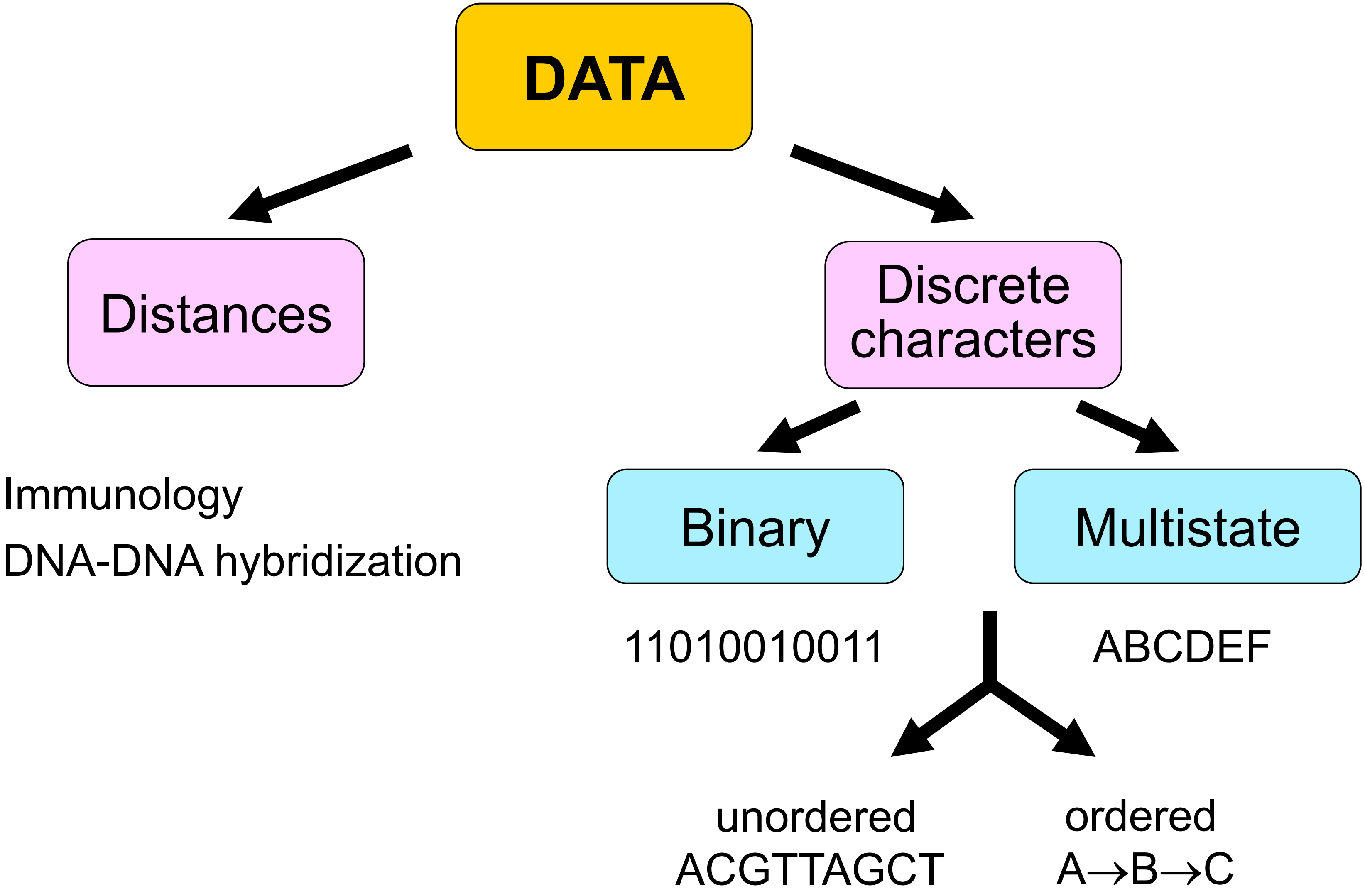


> Avogadro constant*)

number of electrons in visible universe (Eddington number)

*) $6{,}022\ 140\ 76\times10^{23}\ \mathrm{mol}^{-1}$

# What type of data can we use?



DATA

Distances

Discrete characters

Immunology

DNA-DNA hybridization

Binary

Multistate

11010010011

ABCDEF

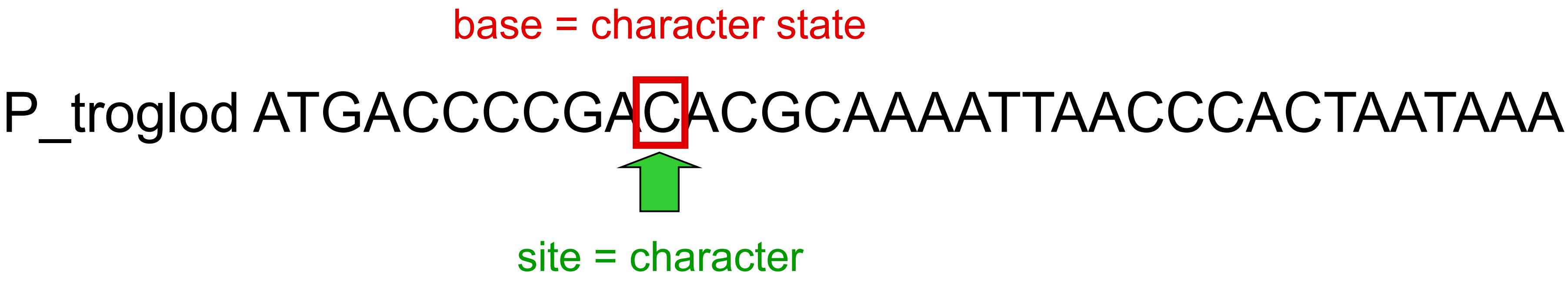
unordered
ACGTTAGCT

ordered
A→B→C

# Types of data

## Nucleotide and protein sequences:

H_sapiens MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGS



base = character state

P_troglod ATGACCCCGACACGCAAAATTAACCCACTAATAAA

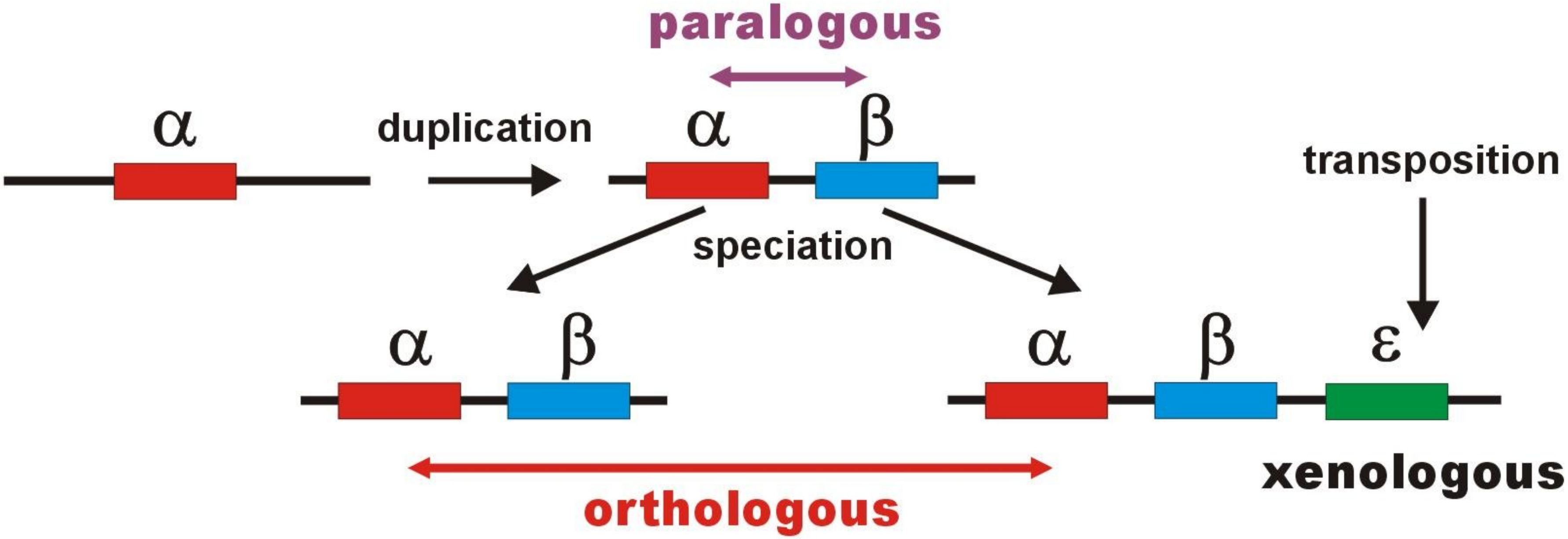
site = character

# Types of data

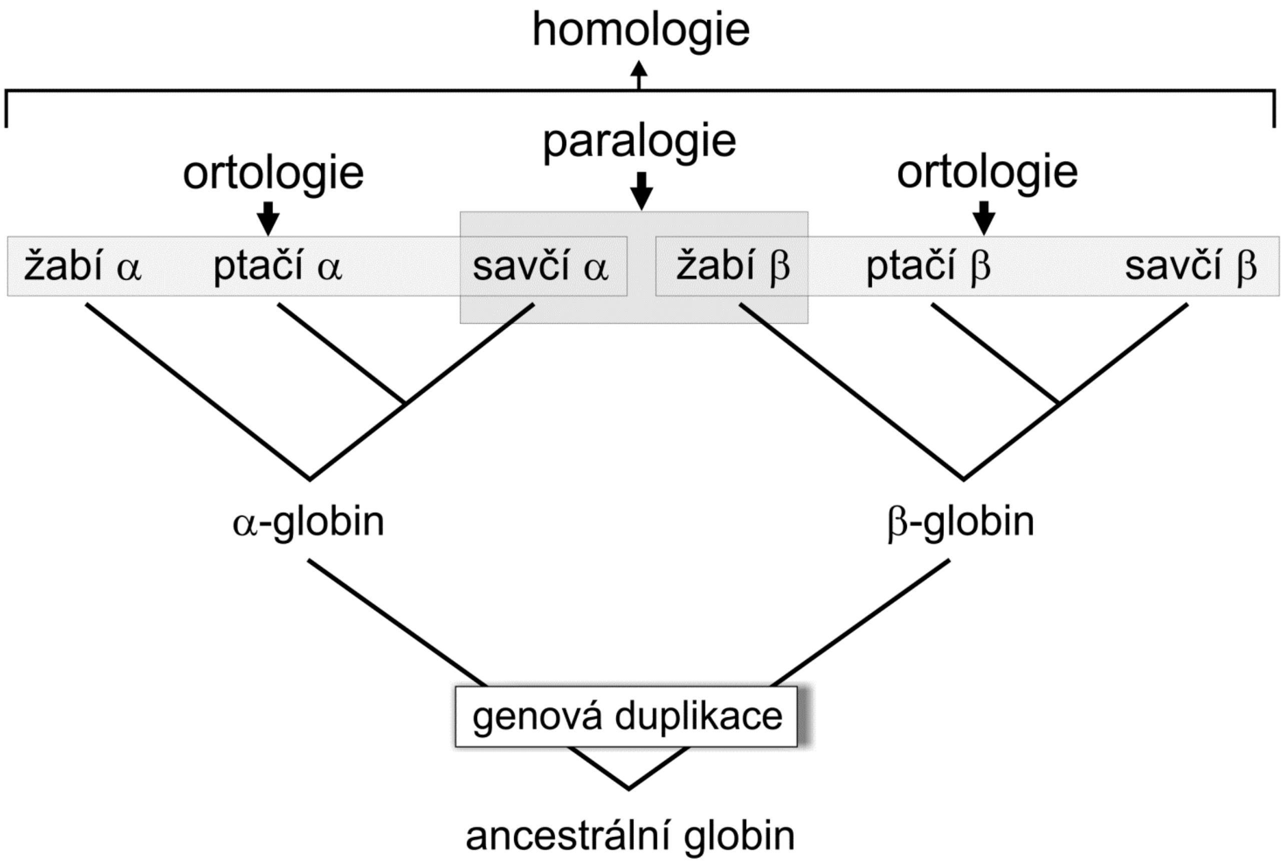
retroelements: SINE (*Alu*, B1, B2), LINE

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 managment: usually packages Sybase or ORACLE

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

# Sequences

## Protein databases:

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/*

# File formats:

## FASTA:

```
>H_sapiens
ATGACCCCAATACGCAAAATTAACCCCCTAATAAAATTAATTAACCACTCATTCATCGACCTCCCCACCC
CATCCAACATCTCCGCATGATGAAACTTCGGCTCACTCCTTGGCGCCTGCCTGATCCTCCAAATCACCAC
AGGACTATTCCTAGCCATACACTACTCACCAGACGCCTCAACCGCCTTTTCATCAATCGCCCACATCACT
CGAGACGTAAATTATGGCTGAATCATCCGCTACCTTCACGCCAATGGCGCCTCAATATTCTTTATCTGCC
TCTTCCTACACATCGGGCGAGGCCTATATTACGGATCATTTCTCTACTCAGAAACCTGAAACATCGGCAT
...
>P_troglod
ATGACCCCGACACGCAAAATTAACCCACTAATAAAATTAATTAATCACTCATTTATCGACCTCCCCACCC
CATCCAACATTTCCGCATGATGGAACTTCGGCTCACTTCTCGGCGCCTGCCTAATCCTTCAAATTACCAC
AGGATTATTCCTAGCTATACACTACTCACCAGACGCCTCAACCGCCTTCTCGTCGATCGCCCACATCACC
CGAGACGTAAACTATGGTTGGATCATCCGCTACCTCCACGCTAACGGCGCCTCAATATTTTTATCTGCC
TCTTCCTACACATCGGCCGAGGTCTATATTACGGCTCATTTCTCTACCTAGAAACCTGAAACATTGGCAT
...
>P_paniscus
ATGACCCCAACACGCAAAATCAACCCACTAATAAAATTAATTAATCACTCATTTATCGACCTCCCCACCC
CATCCAATATTTCCACATGATGAAACTTCGGCTCACTTCTCGGCGCCTGCCTAATCCTTCAAATCACCAC
AGGACTATTCCTAGCTATACACTACTCACCAGACGCCTCAACCGCCTTCTCATCGATCGCCCACATTACC
CGAGACGTAAACTATGGTTGAATCATCCGCTACCTTCACGCTAACGGCGCCTCAATACTTTTCATCTGCC
TCTTCCTACACGTCGGTCGAGGCCTATATTACGGCTCATTTCTCTACCTAGAAACCTGAAACATTGGCAT
...
```

# 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 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 aaacccccca accccctcct
      781 cttaatgcca aaccccaaaa acactaagaa cttgaaagac atatattatt aactatcaaa
      841 ccctatgtcc tgatcgattc tagtagttcc caaaatatga ctcatatttt agtacttgta
      901 aaaattttac aaaatcatgc tccgtgaacc aaaactctaa tcacactcta ttacgcaata
      961 aatattaaca agttaatgta gcttaataac aaagcaaagc actgaaaatg cttagatgga
     1021 taattttatc cca
//
```

# File formats:

## 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
```

# File formats:

## 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;
```

# File formats:

## Clustal X:

```
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
            ** **  **** ** ** ** **   ****** ***** ** ***** ** ** ******
```

# 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.

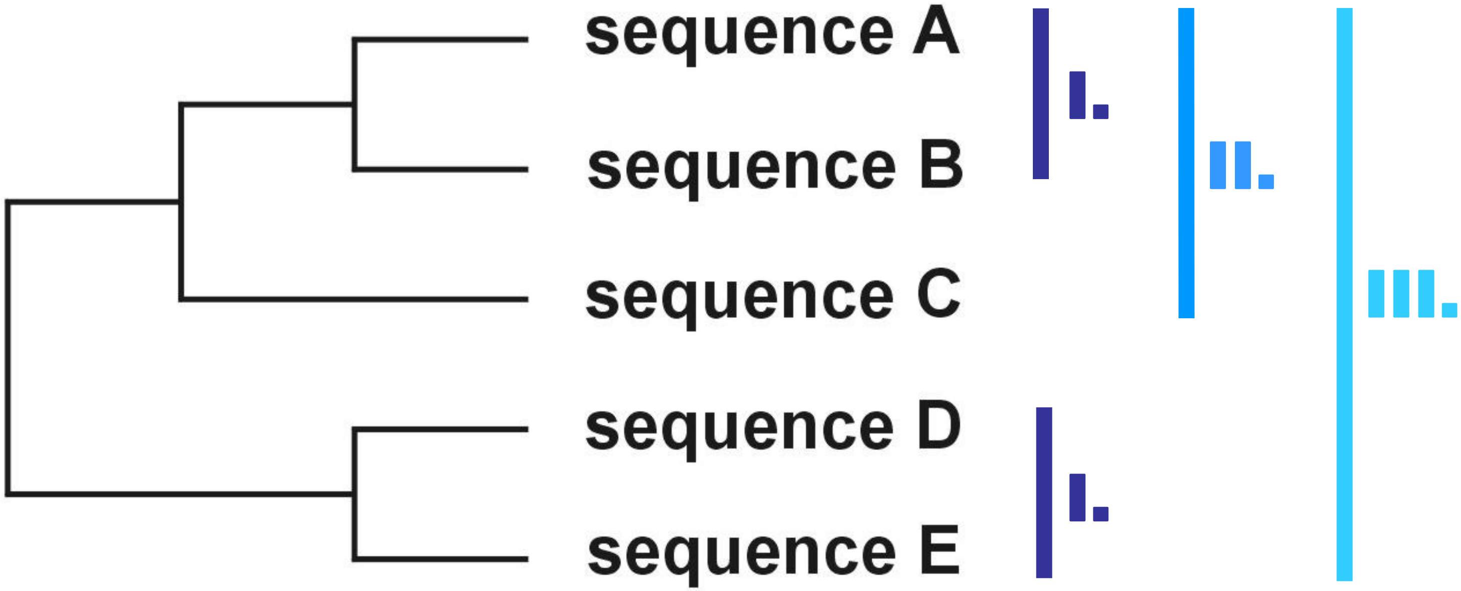
Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.

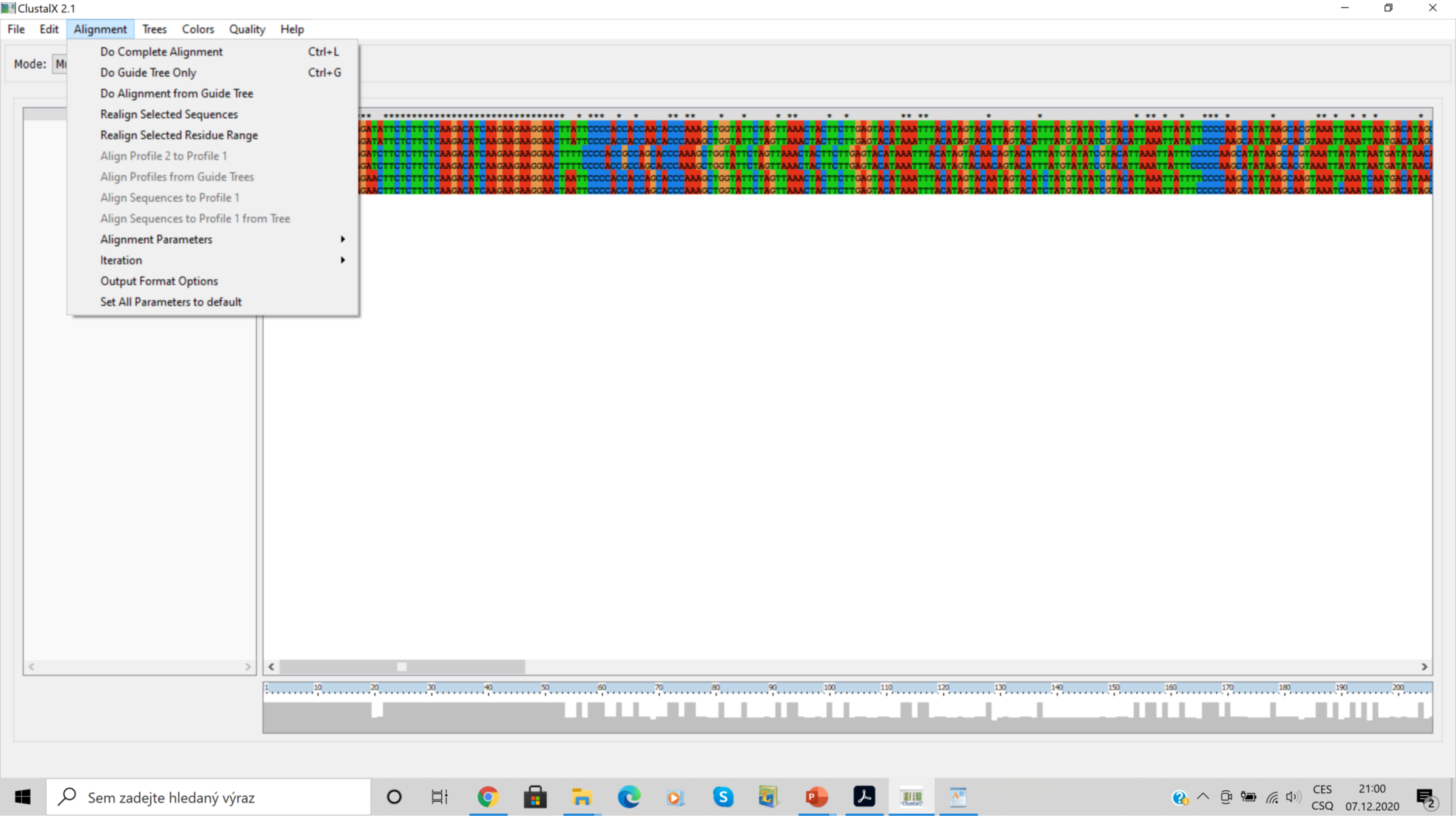
Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

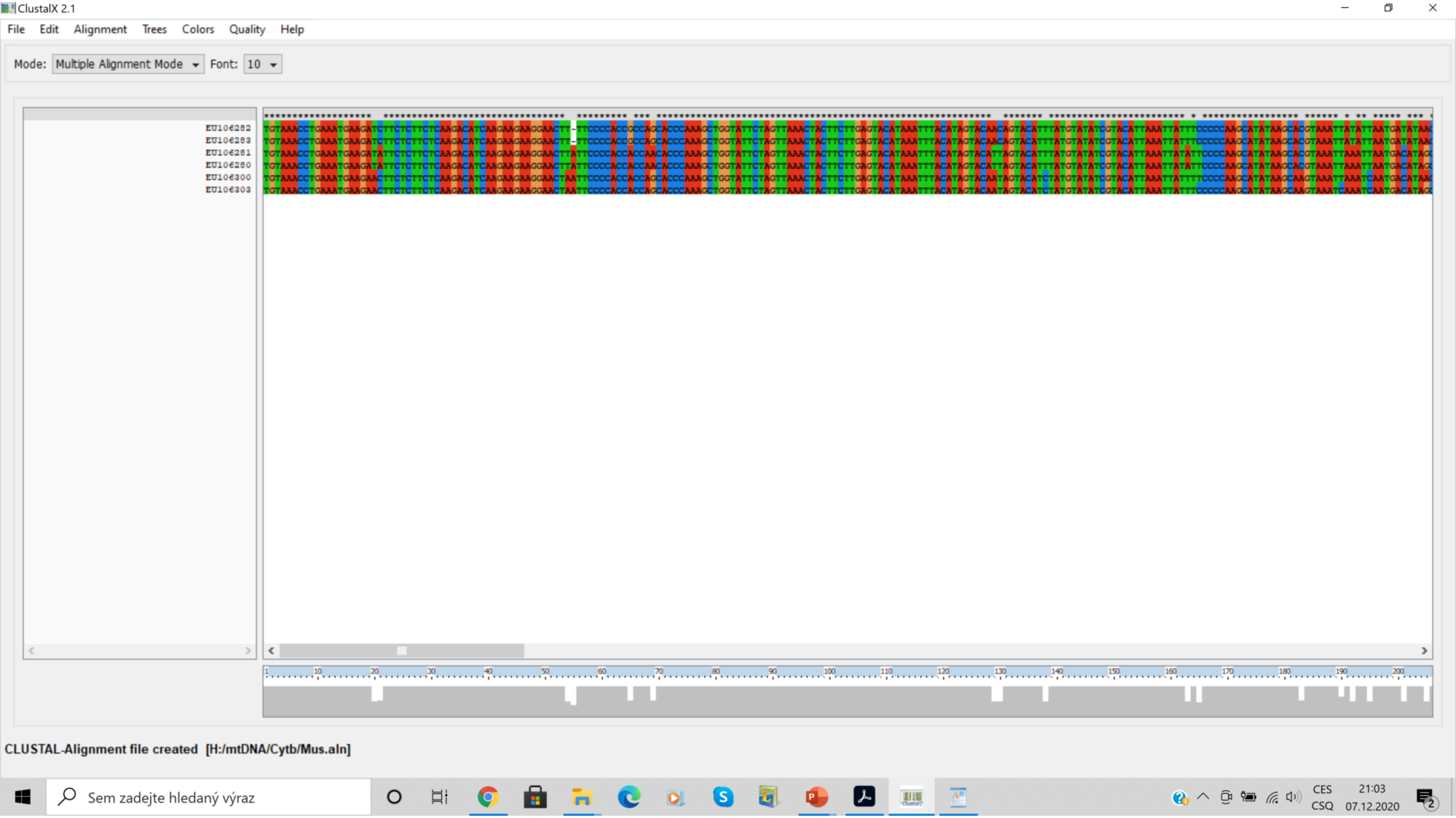
# Progressive alignment - ClustalX

3 phases:

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




ClustalX 2.1
File
Edit
Alignment
Trees
Colors
Quality
Help
Mode:
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
Sem zadejte hledaný výraz
CES
CSQ
21:00
07.12.2020



Mode: Multiple Alignment Mode Font: 10

```
          *******************  ********************************  ********* *** ***********************************************************  ******* ************************ * ***************** ****** * ** ***** *** 
EU106282  TGTAAACCTGAAATGAAGATCTTCTCTTCTCAAGACATCAAGAAGAAGGAACTT-TTCCCCACCGCCAGCACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTACATAGTACAACAGTACATTTATGTATATCGTACATTAAATTATTTCCCCCAAGCATATAAGCACGTAAATTATATTAATGATATAA
EU106283  TGTAAACCTGAAATGAAGATCTTCTCTTCTCAAGACATCAAGAAGAAGGAACTT-TTCCCCACCGCCAGCACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTACATAGTACAACAGTACATTTATGTATATCGTACATTAAATTATTTCCCCCAAGCATATAAGCACGTAAATTATATTAATGATATAA
EU106281  TGTAAACCTGAAATGAAGATATTCTCTTCTCAAGACATCAAGAAGAAGGAACTTATTCCCCACCACCAACACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTACATAGTACATTAGTACATTTATGTATATCGTACATTAAATTATATTCCCCAAGCATATAAGCACGTAAATTAAATTAATGACATAG
EU106280  TGTAAACCTGAAATGAAGATATTCTCTTCTCAAGACATCAAGAAGAAGGAACTTATTCCCCACCACCAACACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTACATAGTACATTAGTACATTTATGTATATCGTACATTAAATTATATTCCCCAAGCATATAAGCACGTAAATTAAATTAATGACATAG
EU106300  TGTAAACCTGAAATGAAGAACTTCTCTTCTCAAGACATCAAGAAGAAGGAACTAATTCCCCACCACCAGCACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTACATAGTACAATAGTACATCTATGTATATCGTACATTAAATTATTTTCCCCAAGCATATAAGCAAGTAAATTAAATCAATGACATAA
EU106303  TGTAAACCTGAAATGAAGAACTTCTCTTCTCAAGACATCAAGAAGAAGGAACTAATTCCCCACCACCAGCACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTACATAGTACAATAGTACATCTATGTATATCGTACATTAAATTATTTCCCCCAAGCATATAAGCAAGTAAATCAAATCAATGACATAG
```

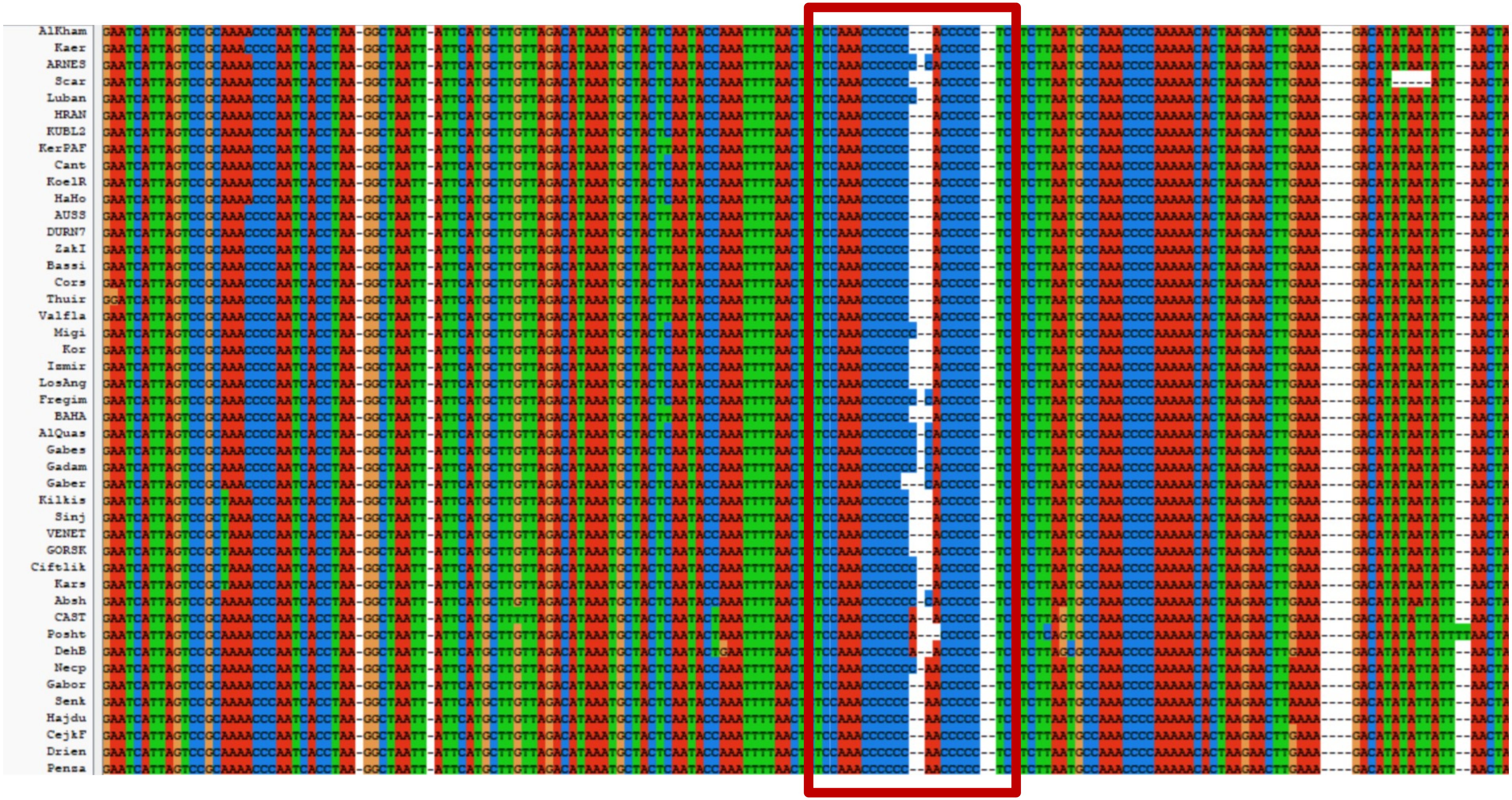
CLUSTAL-Alignment file created [H:/mtDNA/Cytb/Mus.aln]

# Problem with progressive alignment

6 species:

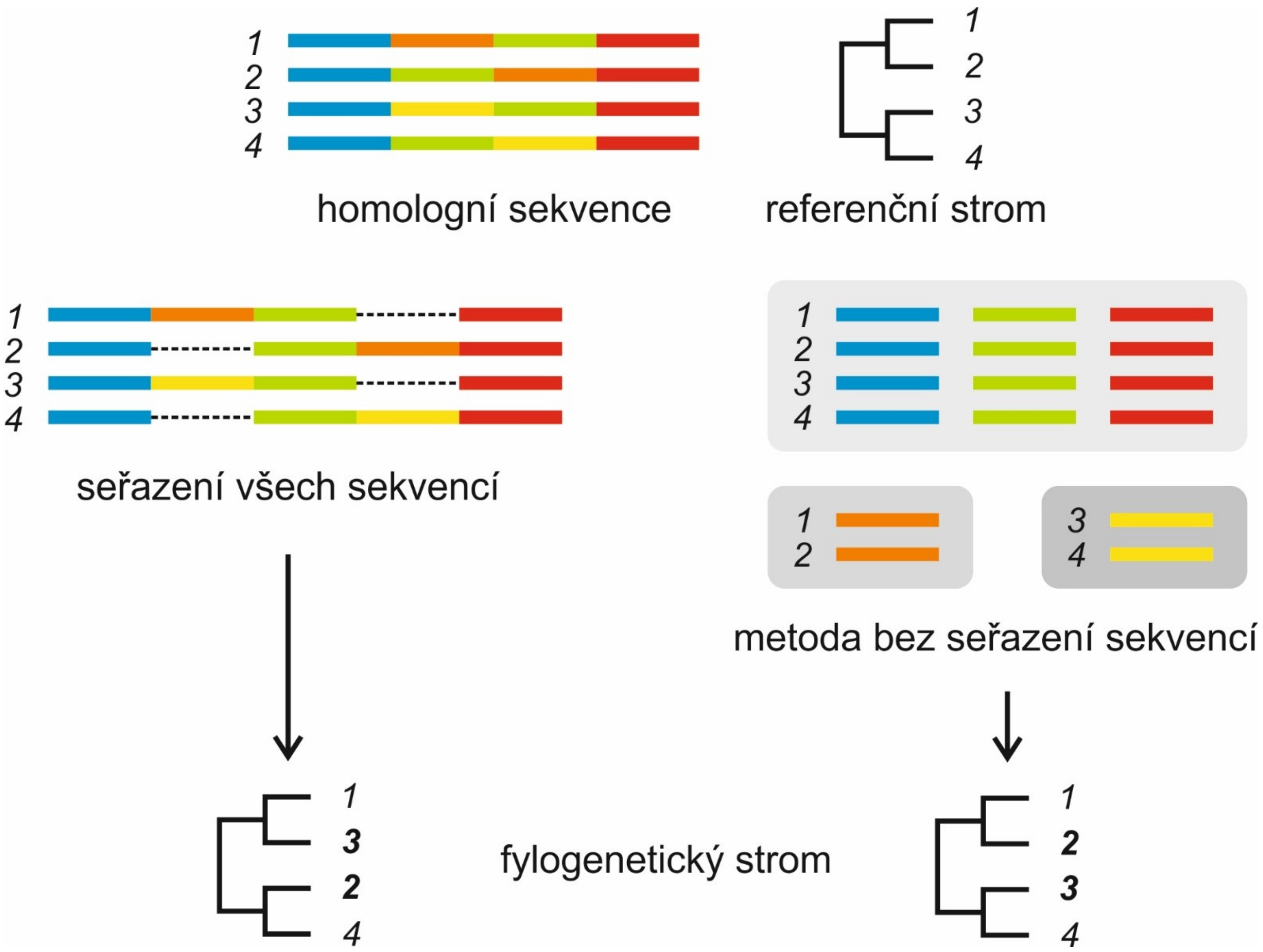
| | | | |
|---|---|---|---|
| gorilla | AGGTT | penguin | A-GTT |
| horse | AG-TT | chicken | A-GTT |
| panda | AG-TT | ostrich | AGGTT |

| | | | |
|---|---|---|---|
| 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 |



Many other alignment programs: e.g. MAFFT, MUSCLE, Geneious...

# There are also methods without alignment:



# Methods

| | | Data types | |
|---|---|---|---|
| | | distances | characters |
| Methods of tree construction | algorithms | UPGMA<br>neighbor-joining | |
| | optimality criteria | Fitch-Margoliash<br>minimum evolution | maximum parsimony<br>maximum likelihood<br>Bayesian a. |

# How to assess the methods?

Efficiency: how fast is the method?

Power: how many characters we need?

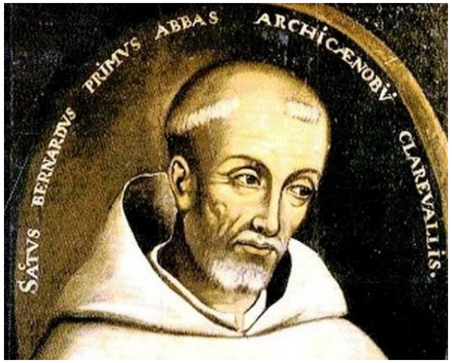
Consistency: does increasing characters result in true tree?

Robustness: how does it work when assumptions are violated?

Falsifiability: does it allow testing assumptions?

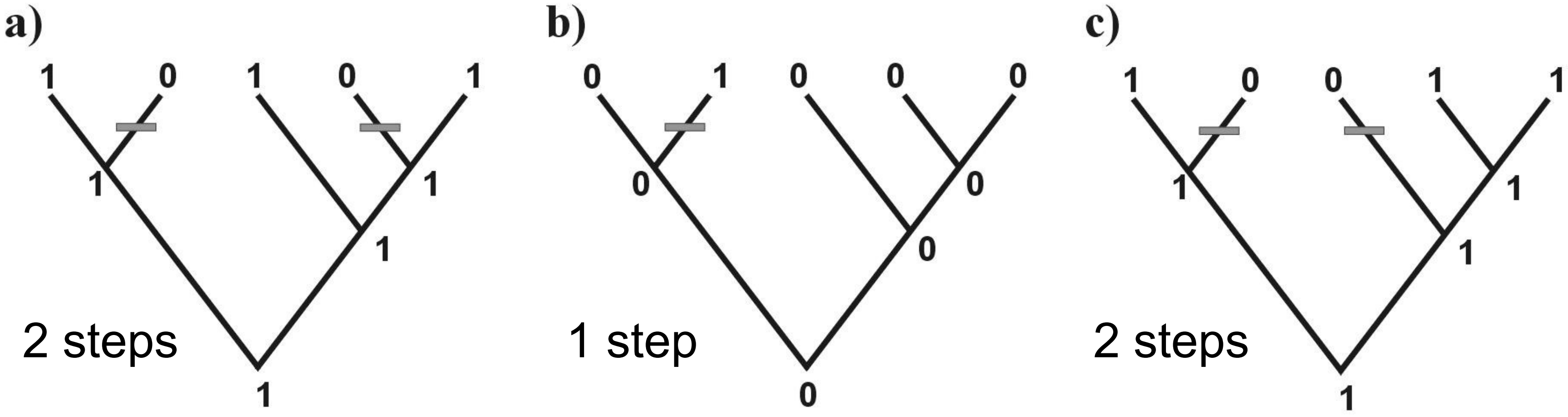
# MAXIMUM PARSIMONY, MP (maximální úspornost)

William of Ockham (c. 1287 – 1347)
Occam's razor

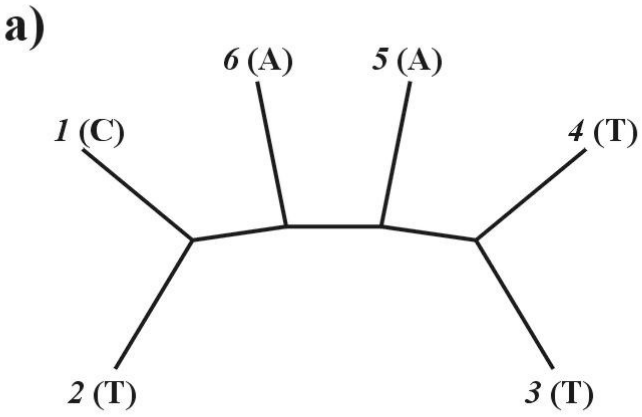


|   | I | II | III |
|---|---|---|---|
| A | 1 | 0 | 1 |
| B | 0 | 0 | 1 |
| C | 1 | 0 | 0 |
| D | 0 | 1 | 0 |
| E | 1 | 0 | 1 |

minimal number of steps = 3
real number of steps = 5
⇒ 2 extra steps → homoplasy

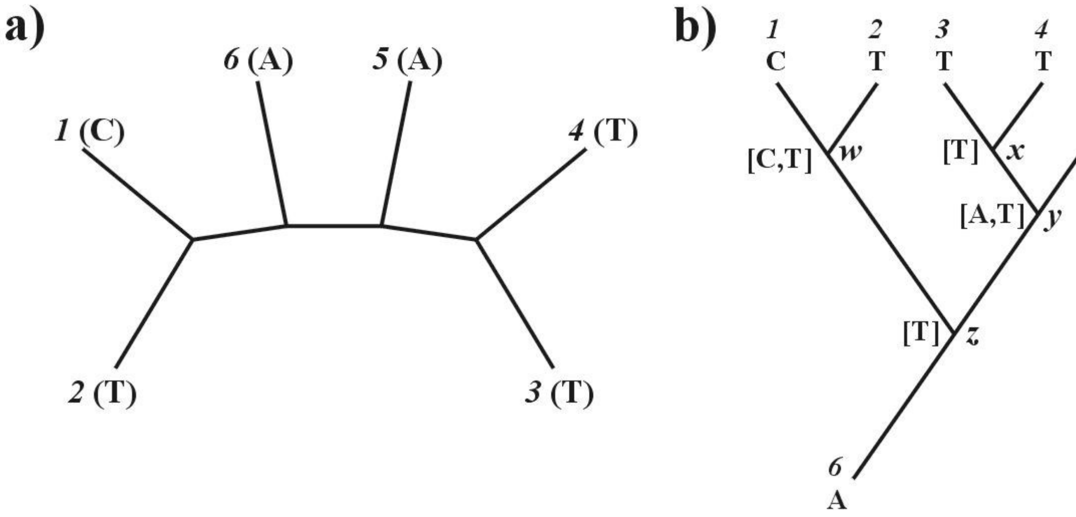


# Estimation of number of steps: Fitch algorithm



1. arbitrary root

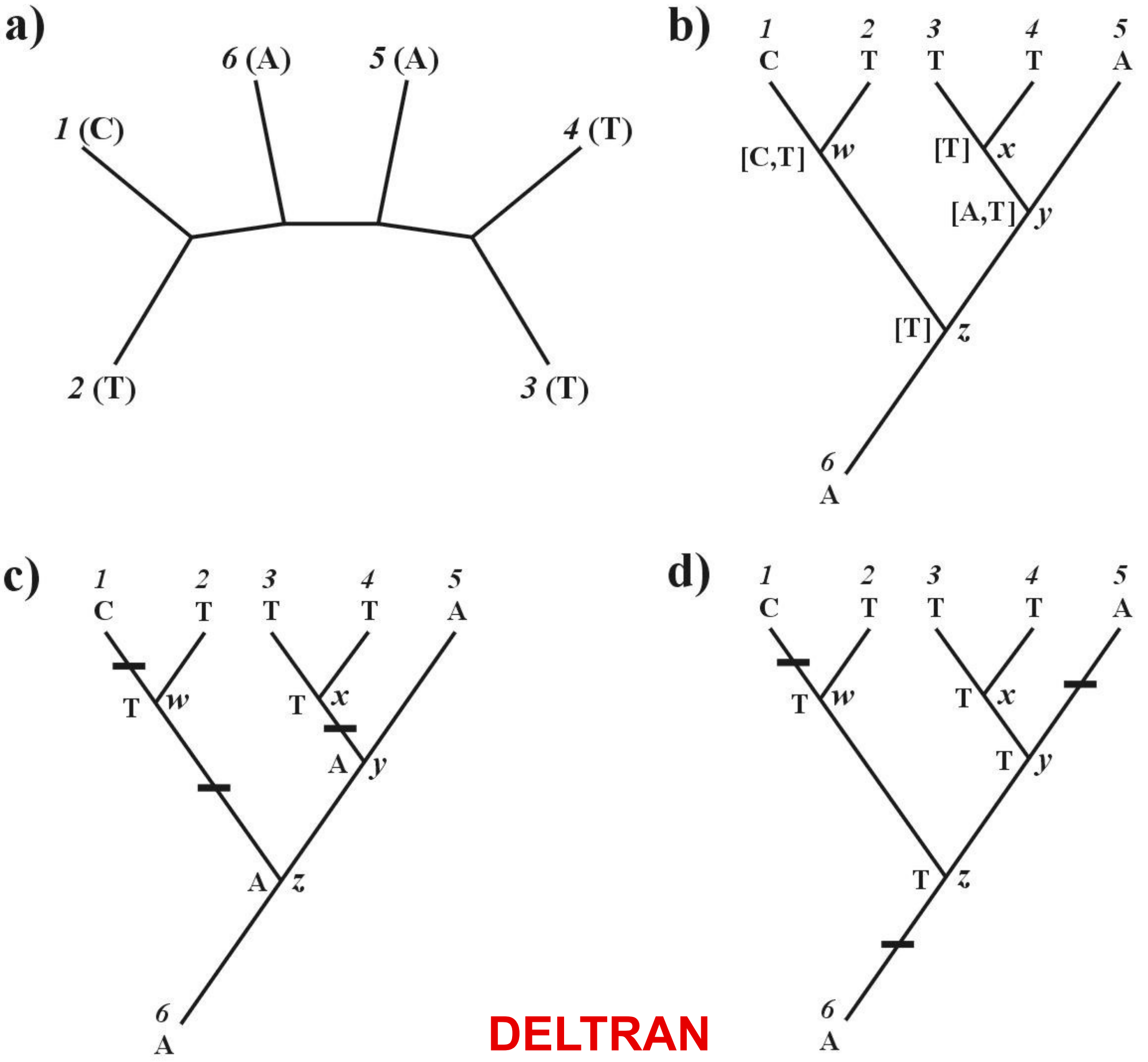
# Estimation of number of steps: Fitch algorithm



1. arbitrary root

2. Downward:
   $w$ = C or T
   $x$ = T
   $y$ = A or T
   $z$ = T

# Estimation of number of steps: Fitch algorithm



1. arbitrary root

2. Downward:
   *w* = C or T
   *x* = T
   *y* = A or T
   *z* = T

3. Upward:
   *z* = T, nebo A

total length = 3

**DELTRAN**
**(DELayed TRANsformation)**

**ACCTRAN**
**(ACCelerated TRANsformation)**

# 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

**RC = CI × 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

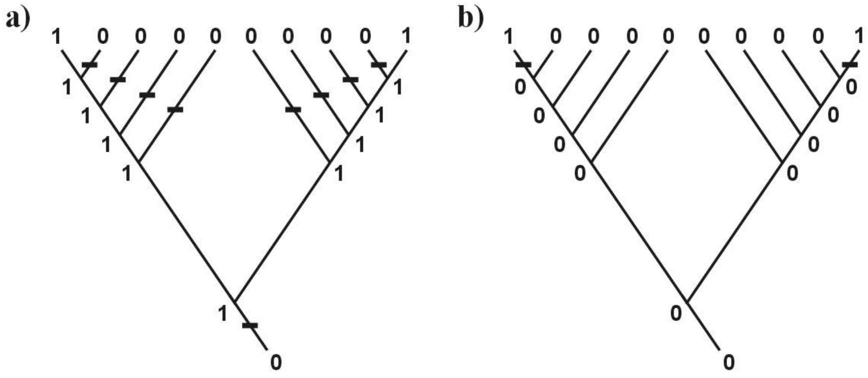
# Metods of parsimony:

Fitch: X → Y a Y → X
neseřazené znaky (A → T nebo A → G etc.)

Wagner: X → Y a Y → X
seřazené znaky (1 → 2 → 3)

Dollo: X → Y a Y → X, potom nelze X → Y

… *restriction-site and restriction-fragment data*



“*relaxed Dollo criterion*”

Camin-Sokal: X → Y,
not Y → X

… SINE, LINE

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 | - |

Dollo

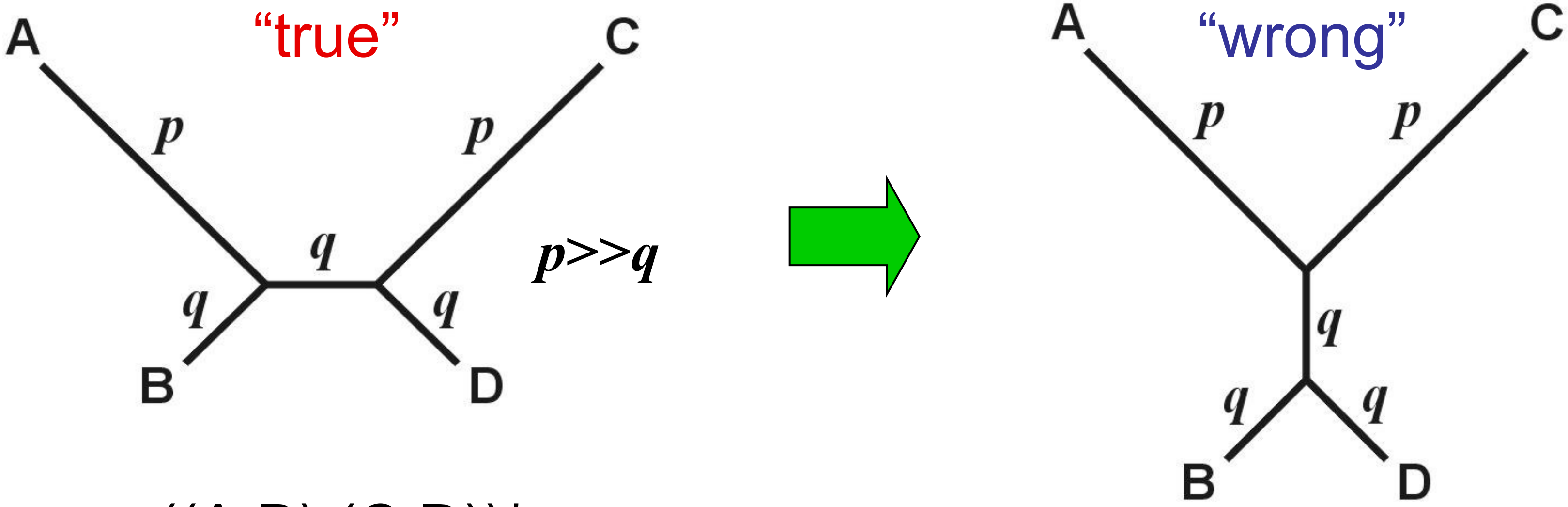
**d)**

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

transversion

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

# Parsimony and consistency



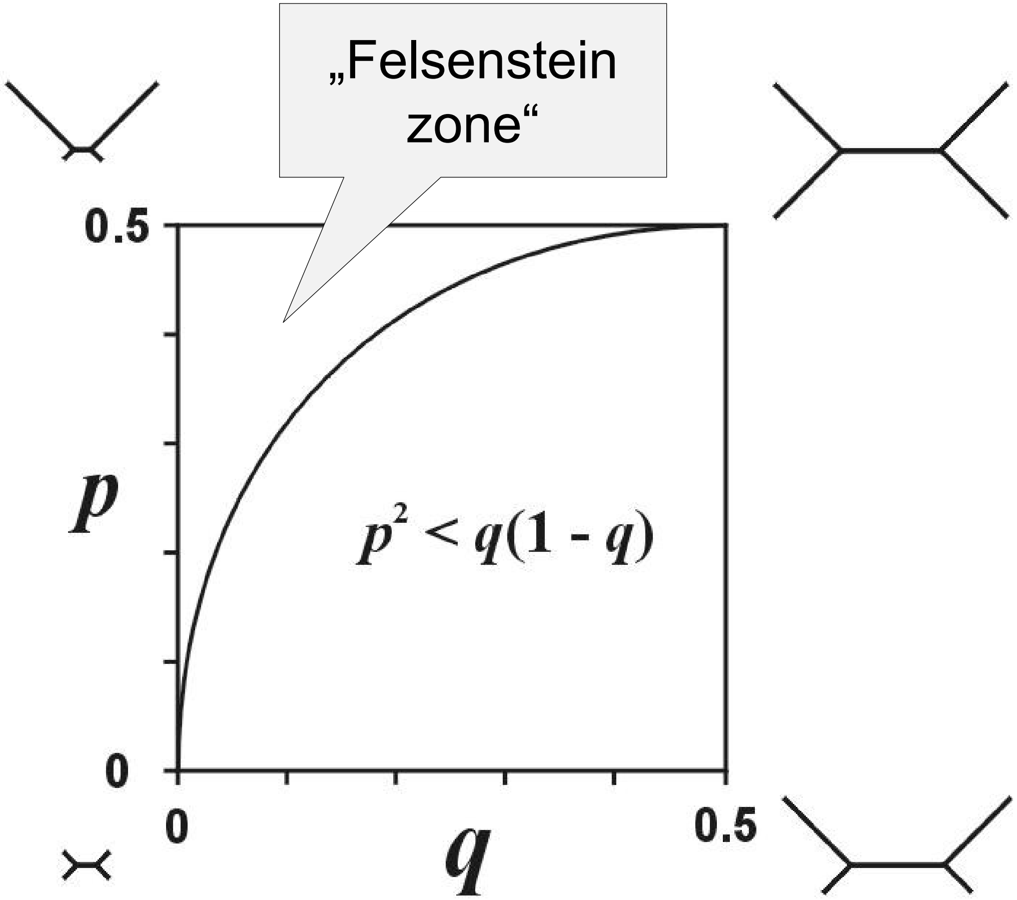
$p>>q$

((A,B),(C,D))*

((A,C),(B,D))*

* tree written in Newick format

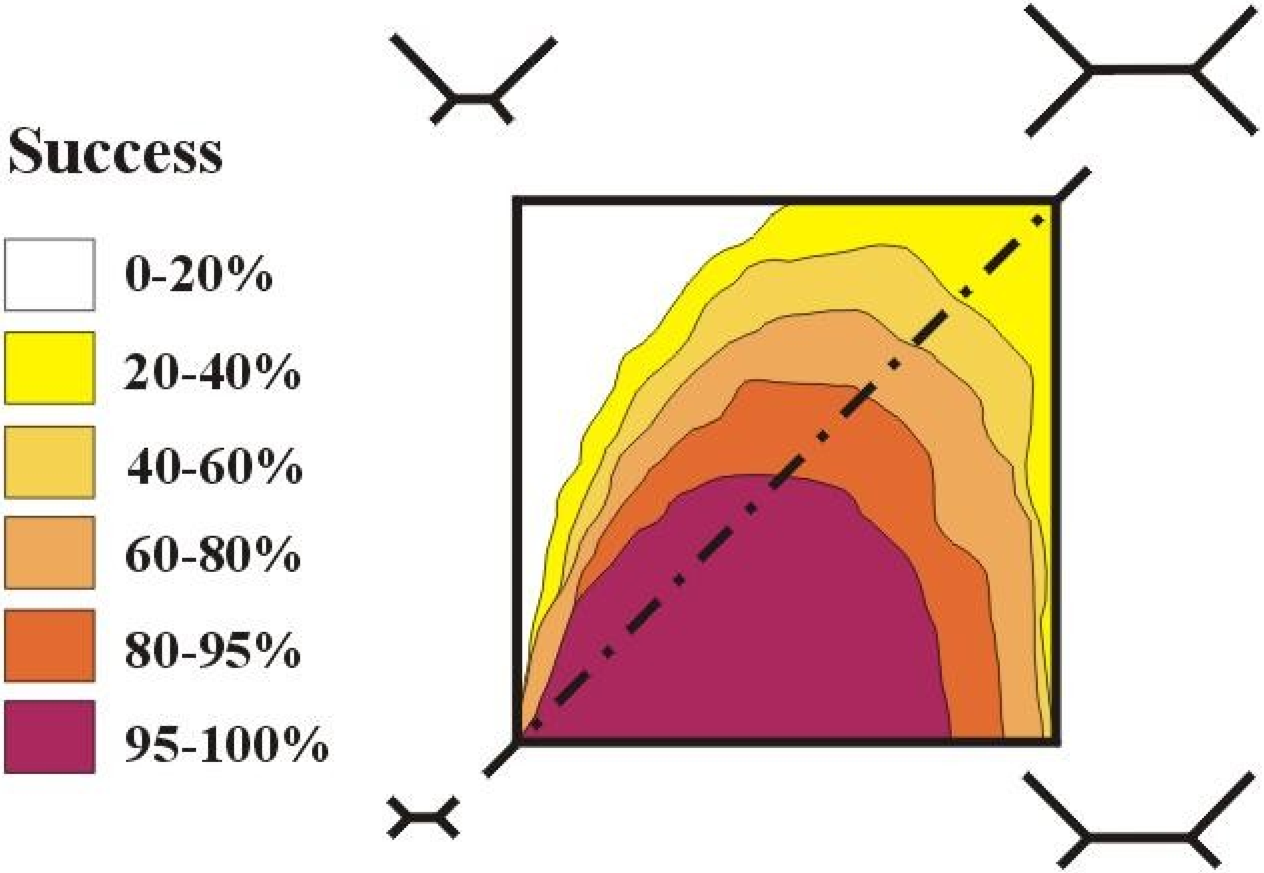
# Parsimony and consistency



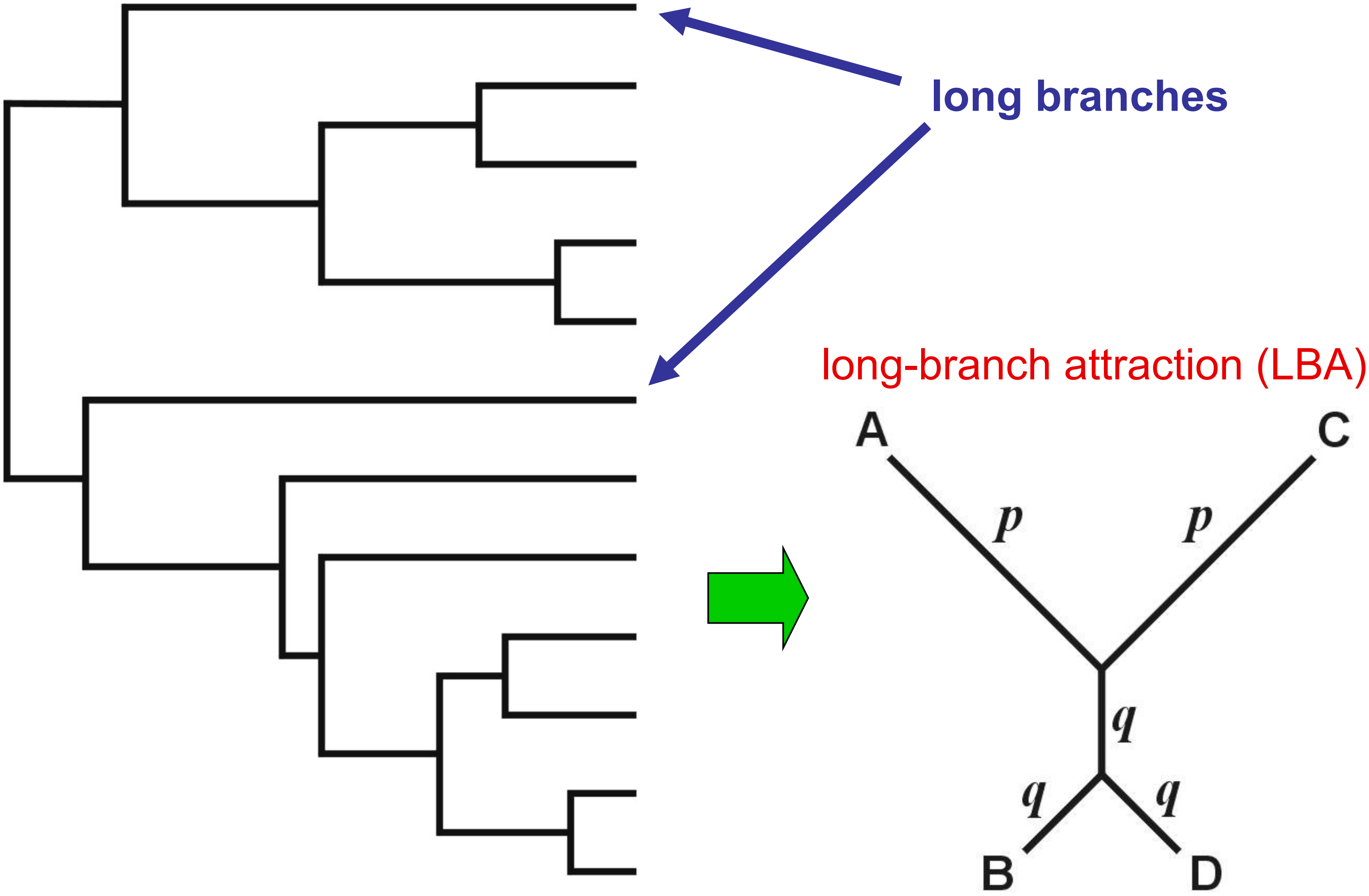
„Felsenstein zone“

In the Felsenstein zone, parsimony is inconsistent

# Parsimony and consistency



# Parsimony and consistency



long branches

long-branch attraction (LBA)

# Search for optimal tree

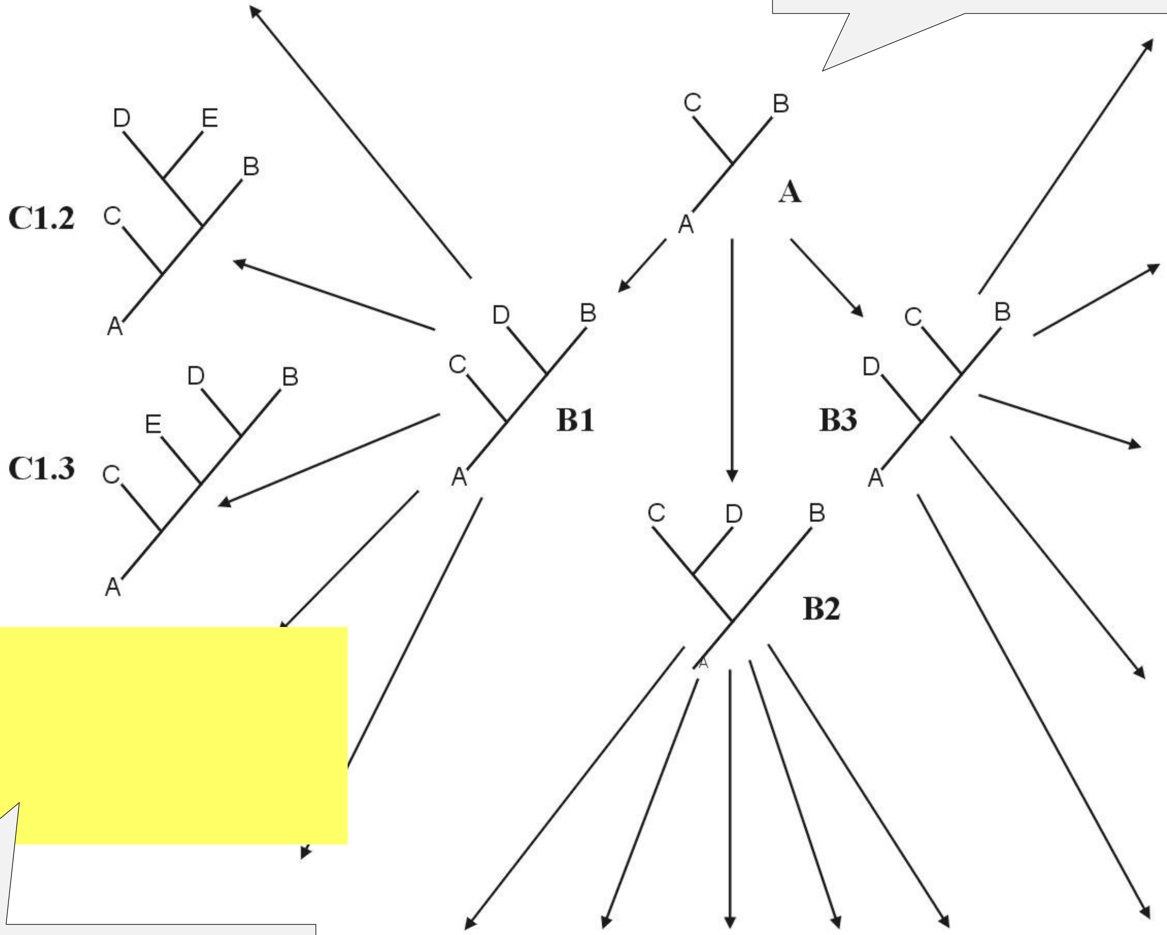
1. Exact methods:

    a) exhaustive search

    b) branch-and-bound

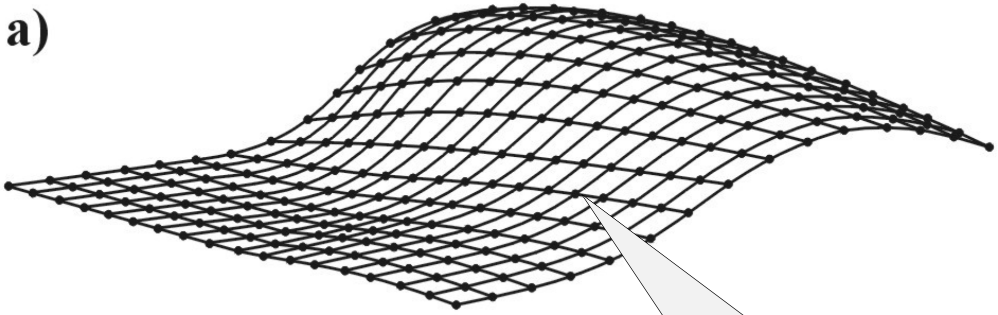
# branch-and-bound

starts with 3 taxa, sequential addition



if the tree is longer than a randomly chosen tree, the process is terminated

# 2. Heuristic search

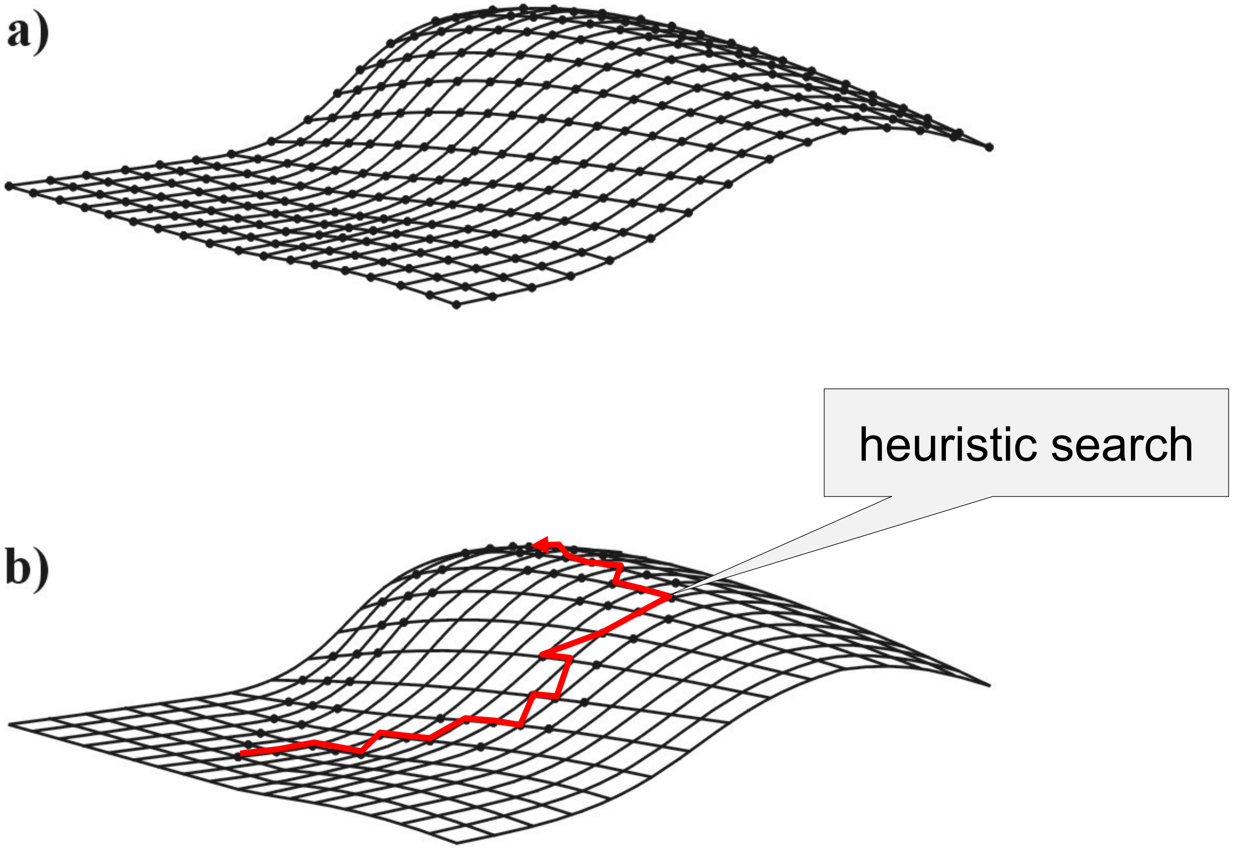


all possible trees

stepwise addition

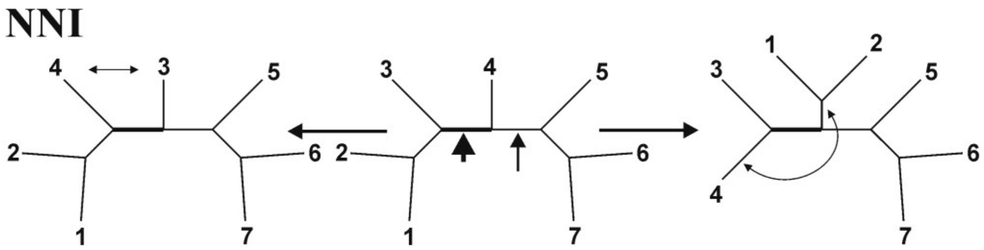
star decomposition

branch swapping

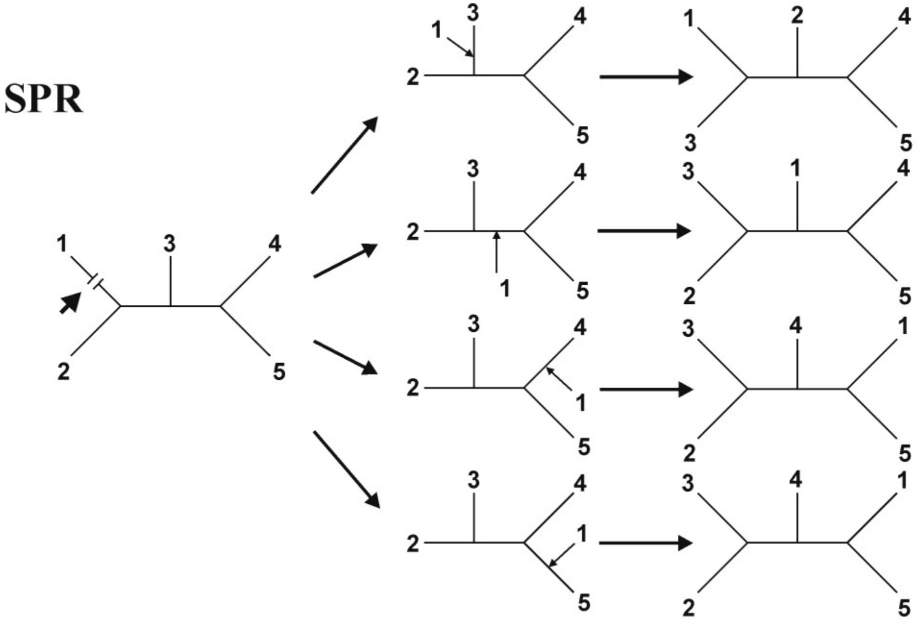


heuristic search

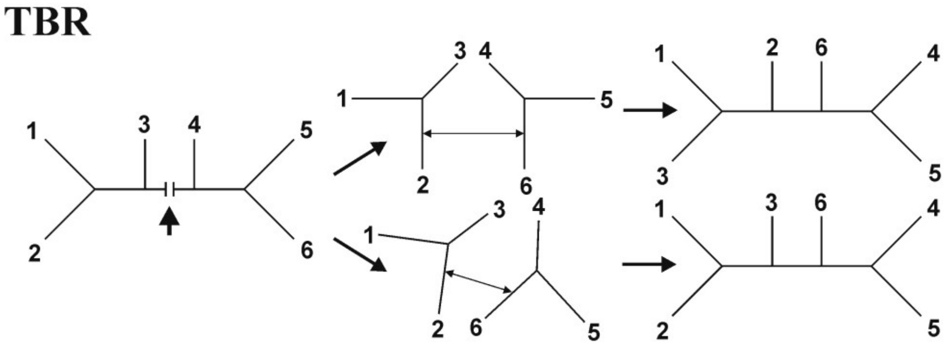
nearest-neighbor interchanges (NNI)



subtree prunning and regrafting (SPR)



tree bisection and reconnection (TBR)



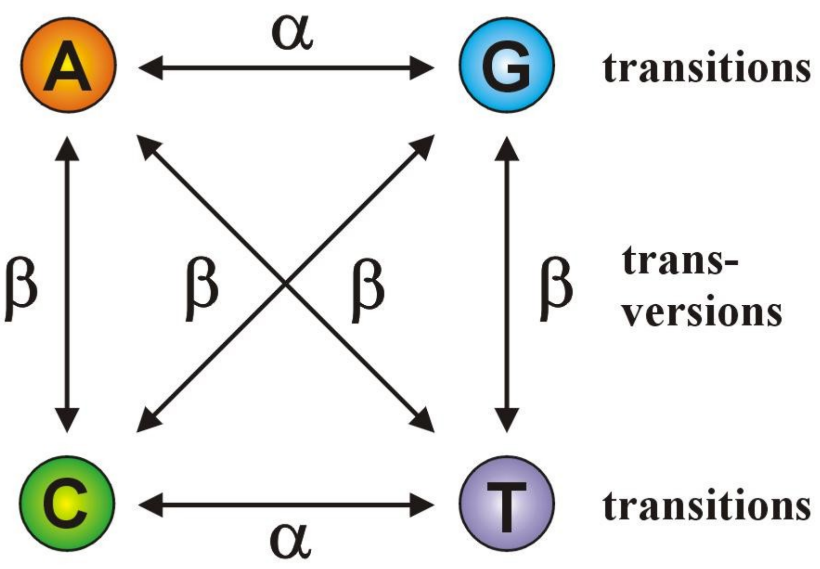
# Evolutionary models and distance methods

| | | Base after substitution | | | |
|---|---|---|---|---|---|
| | | A | C | G | T |
| | A | -¾ | ¼ | ¼ | ¼ |
| Original base | C | ¼ | -¾ | ¼ | ¼ |
| | G | ¼ | ¼ | -¾ | ¼ |
| | T | ¼ | ¼ | ¼ | -¾ |

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

**Jukes-Cantor (JC):** equal base frequencies
equal substitution rates

## Kimura 2-parameter (K2P): transitions ≠ 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

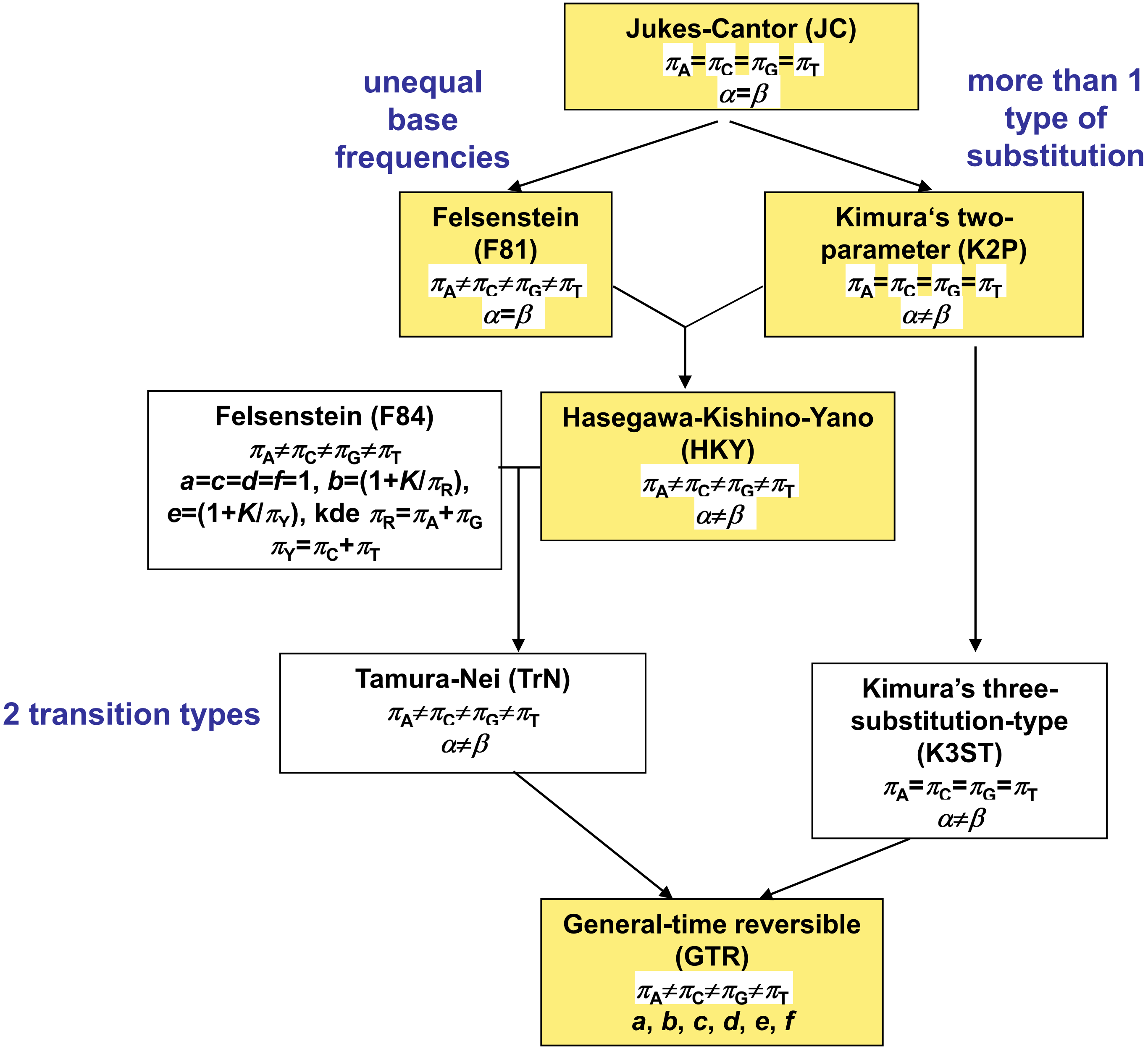
$$Q = \begin{pmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{pmatrix}$$

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

**Hasegawa-Kishino-Yano (HKY):** different base frequencies
transitions ≠ 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



**Jukes-Cantor (JC)**
$\pi_A=\pi_C=\pi_G=\pi_T$
$\alpha=\beta$

unequal base frequencies

more than 1 type of substitution

**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 transition types

**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*

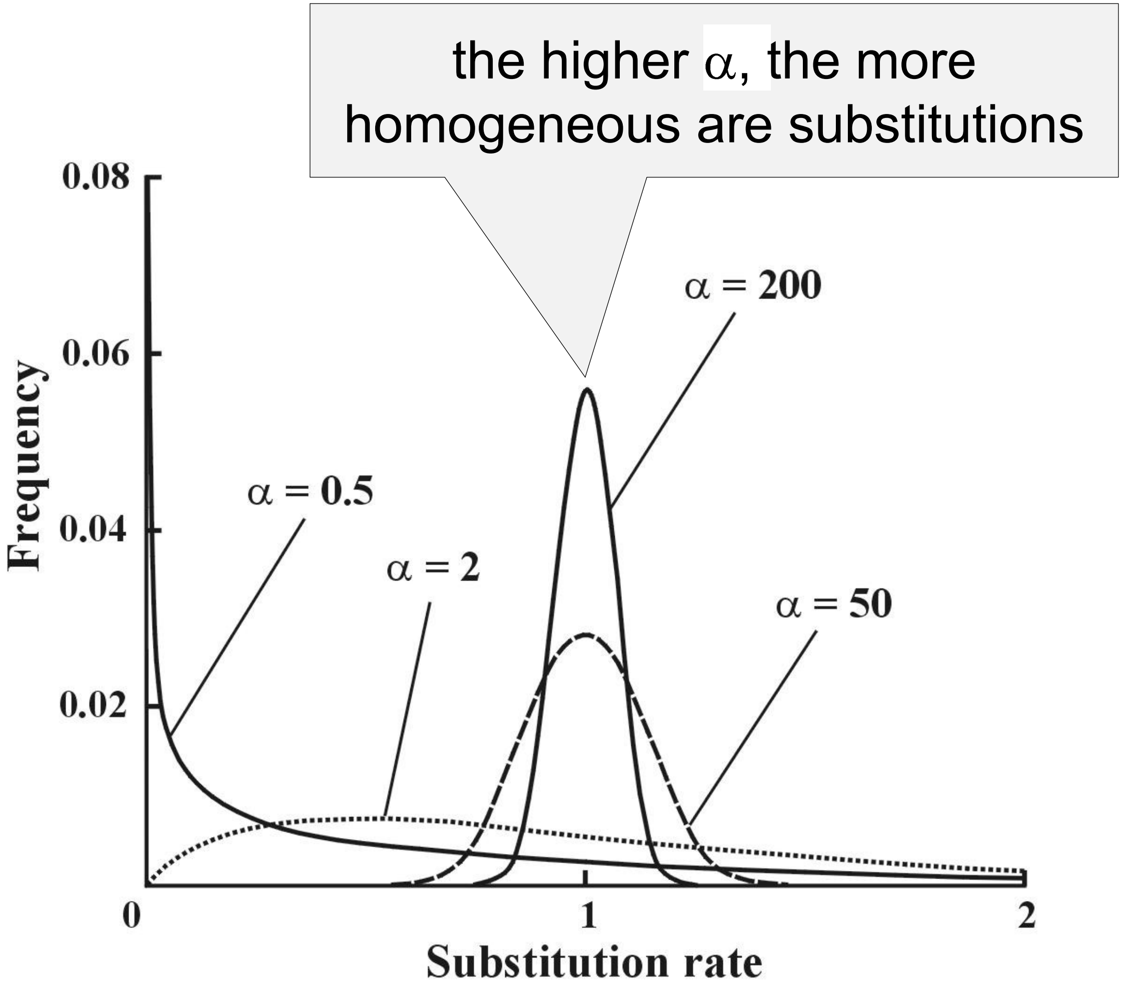
# Heterogenity of substitution rates in different parts of sequences

Gamma distribution:

shape parameter α

discrete gamma model

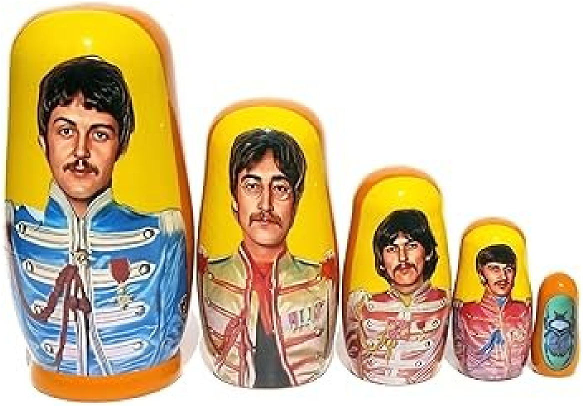
invariant sites
→ GTR+Γ+I
nebo GTR+G+I



the higher α, the more homogeneous are substitutions

# Model comparison:

## Likelihood ratio test (LRT):



nested models

$LR = 2(\ln L2 - \ln L1)$

$\chi^2$ distribution, p2 – p1 degrees of freedom

## Akaike information criterion (AIC):

nonnested models

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

better model → lower $AIC$

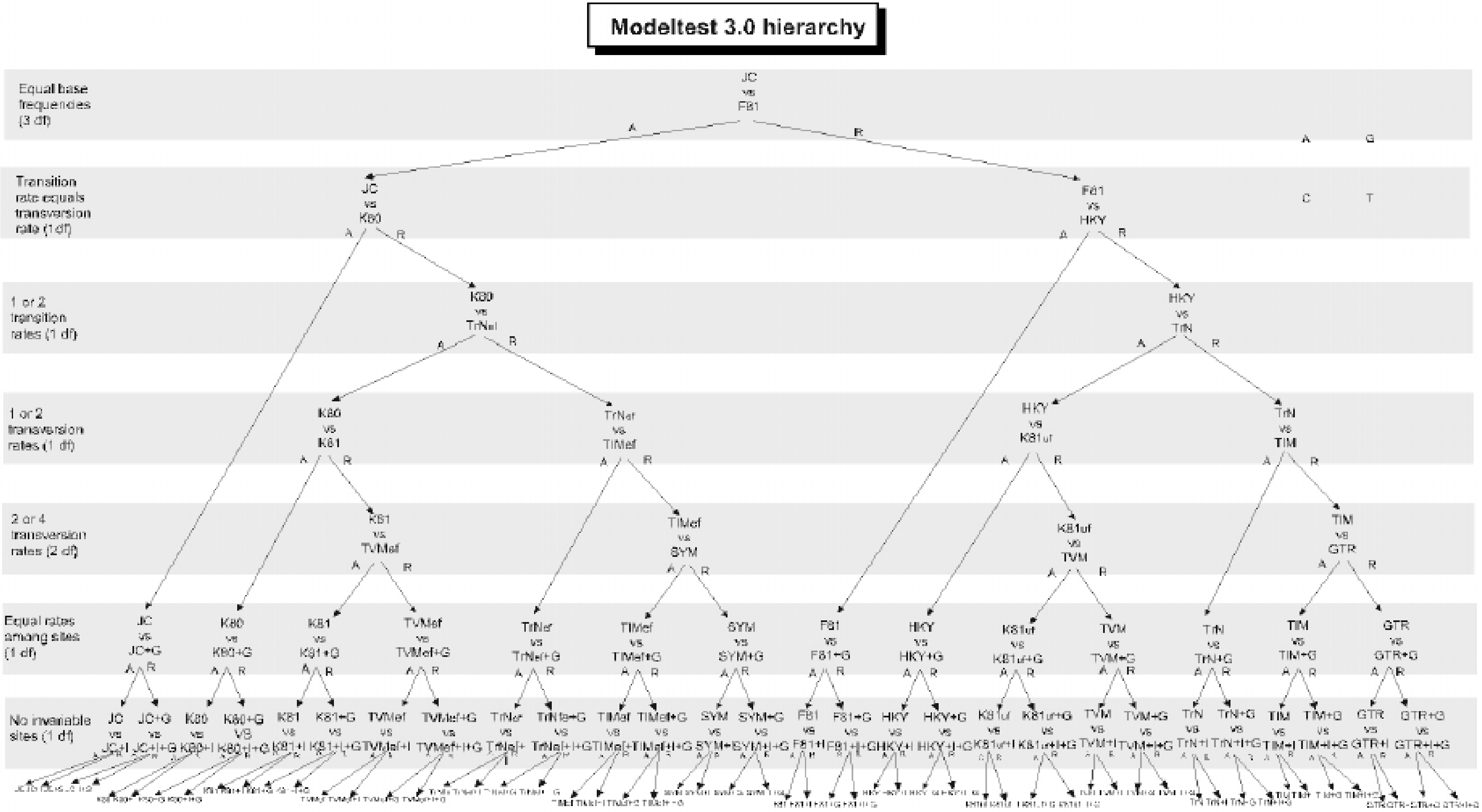
## Bayesian information criterion (BIC):

nonnested models

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

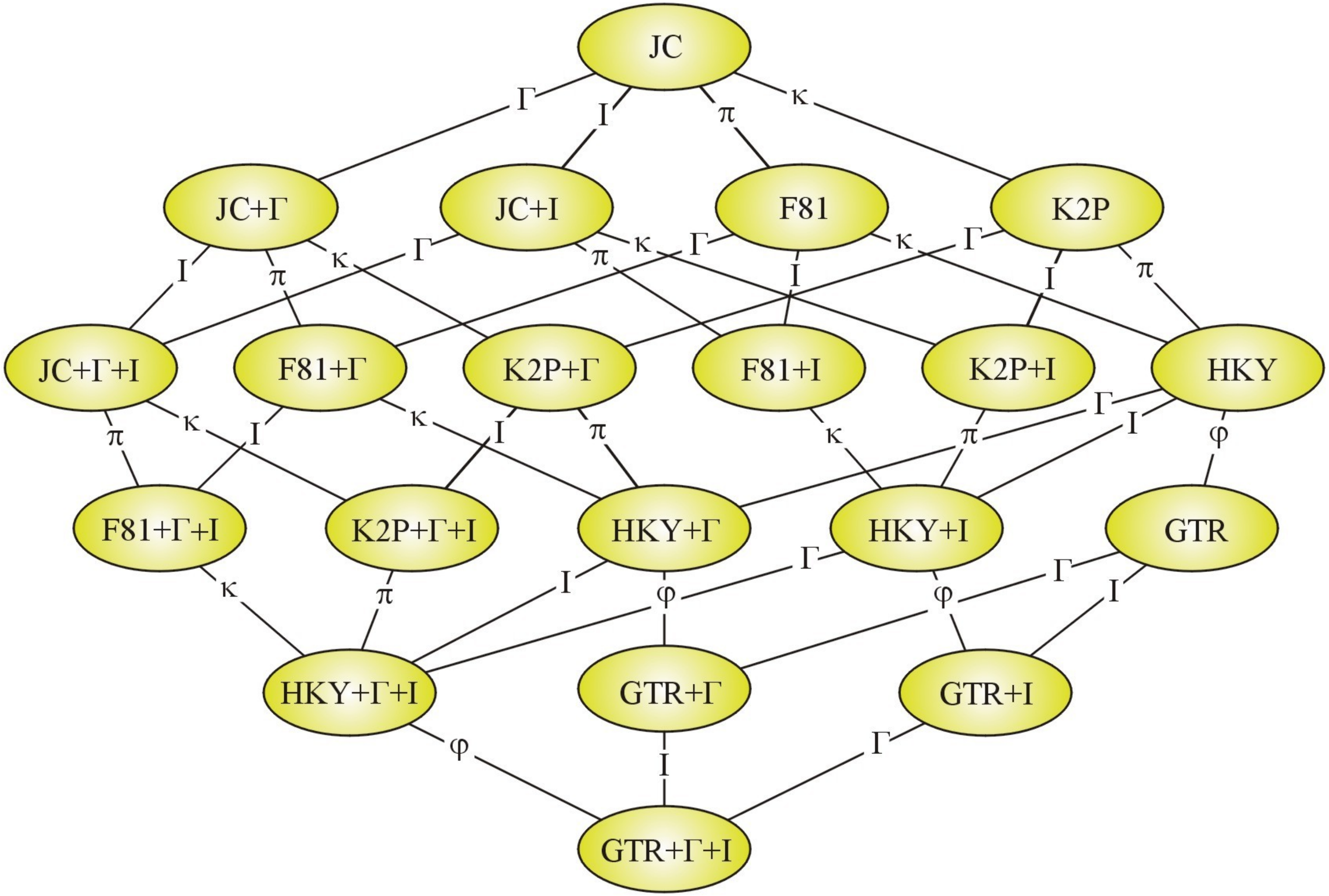
# Model comparison:

hierarchical LRT – ModelTest (Crandall and Posada), jModelTest

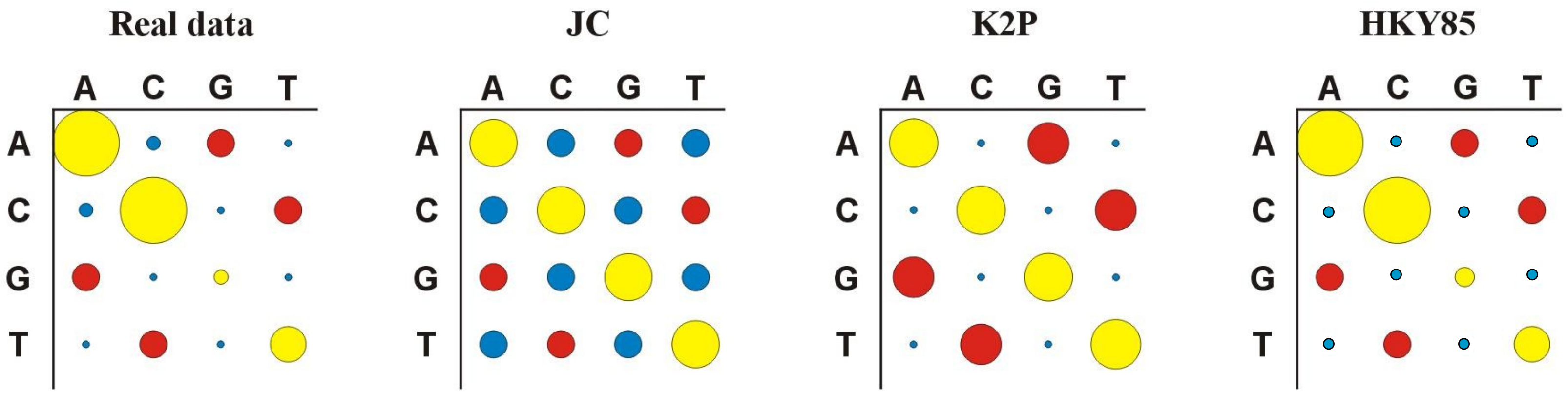


# Model comparison:

dynamic LRT:



# Model comparison:



More parametres ⇒ 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

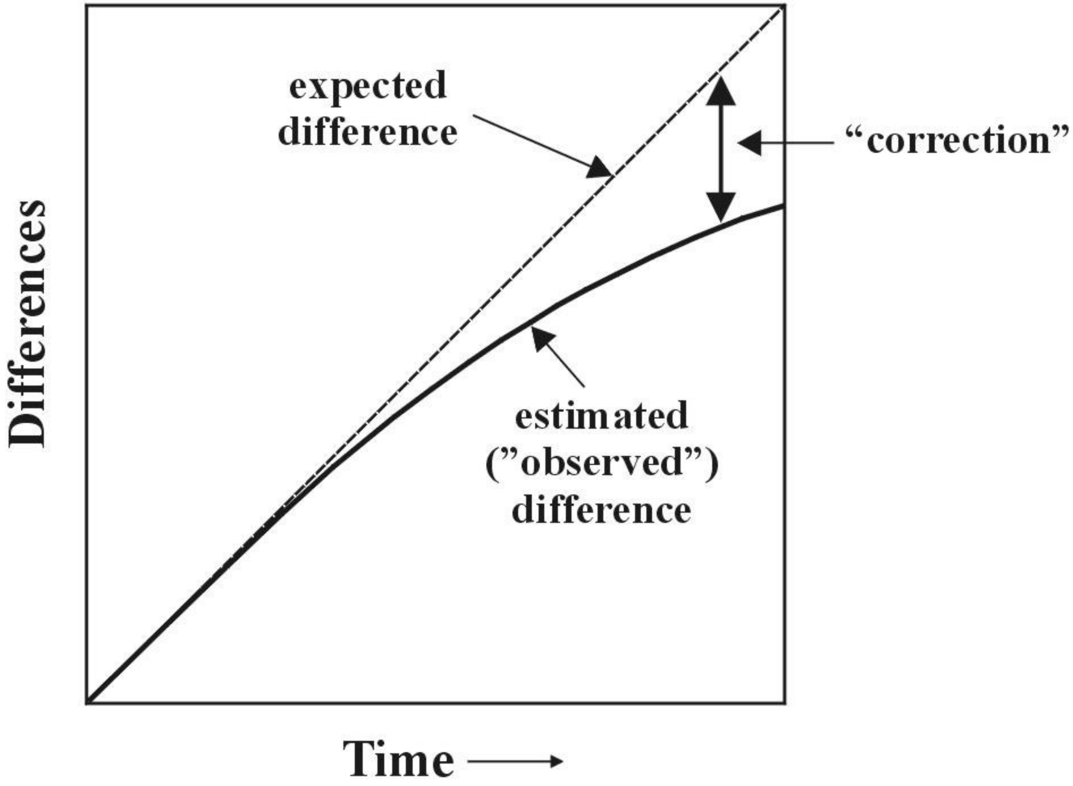
distance methods base on assumption that if we know true distances, we can very easily infer the true phylogeny

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

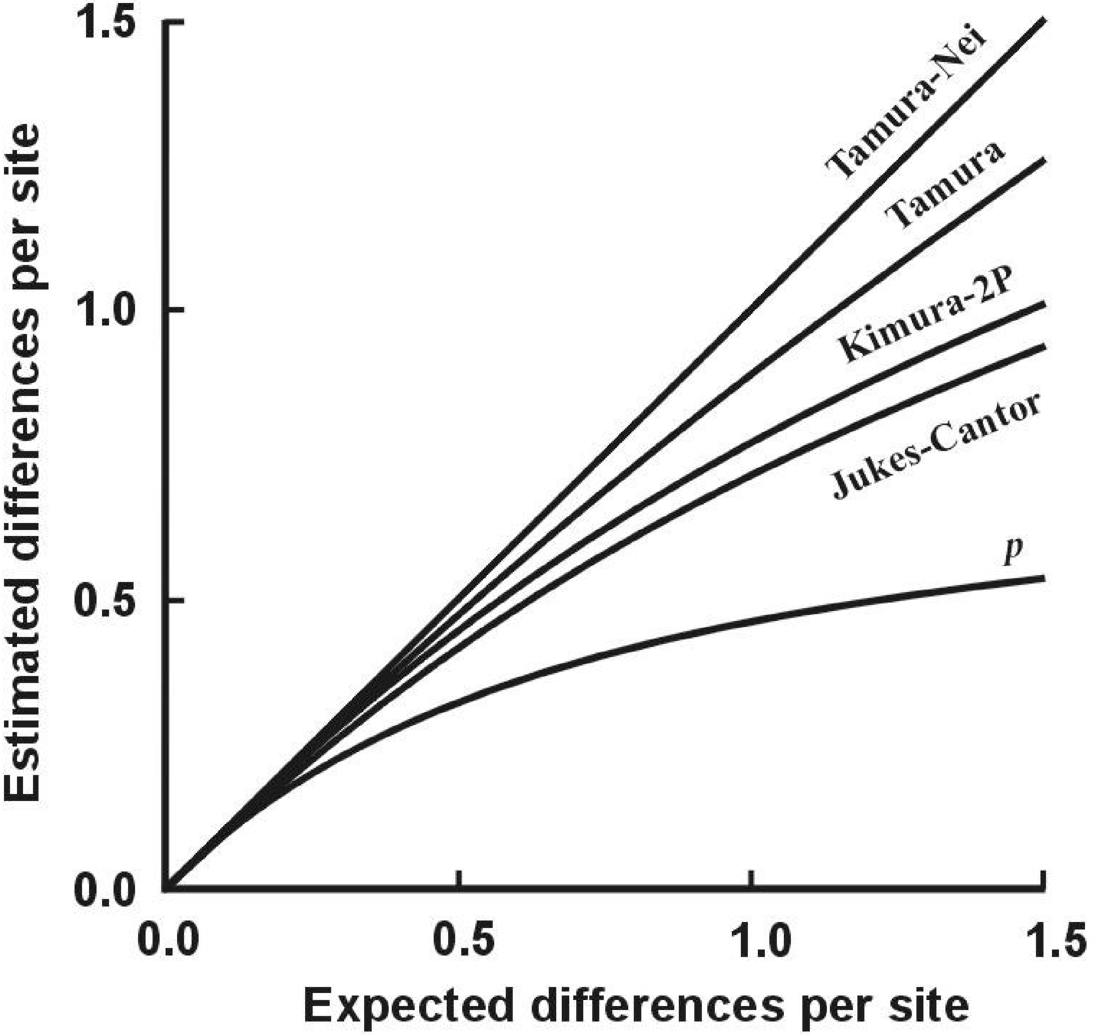
```
              1        10        20        30
sequence 1:   ACCCGTTAAGCTTAACGTACTTGGATCGAT
sequence 2:   ACCCGTTAGGCTTAATGTACGTGGATCGAT
```

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

problem of saturation:



# Distances for some models:


Estimated differences per site
Expected differences per site
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
Tamura-Nei
Tamura
Kimura-2P
Jukes-Cantor
p

# 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 (ŠB-*k*) = [*d*(B-*k*)+*d*(Š-*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 | -- |

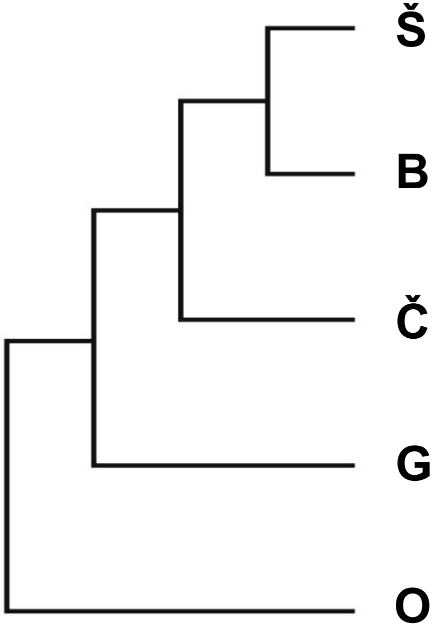
UPGMA (unweighted pair-group method using arithmetic means):

d[(BŠČ)G] = {d(BG)+d(ŠG)+d(ČG)}/3

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

single-linkage (metoda nejbližšího souseda)

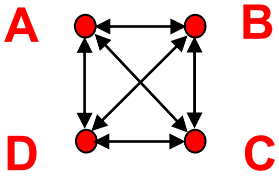
complete-linkage (m. nejvzdálenějšího souseda)



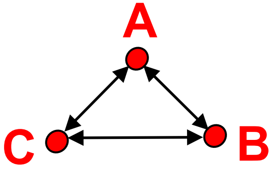
# UPGMA and consistency

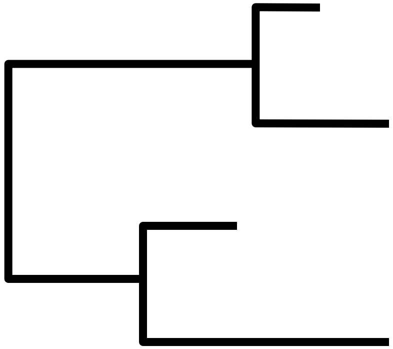
additive distances: $d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC})$

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

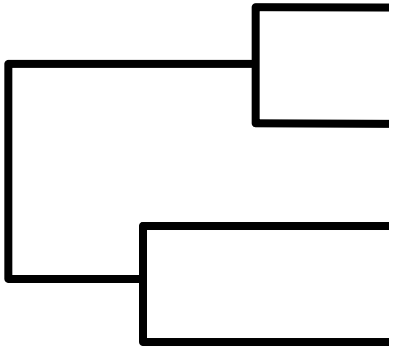


ultrametric distances: $d_{AC} \leq \max(d_{AB}, d_{BC})$



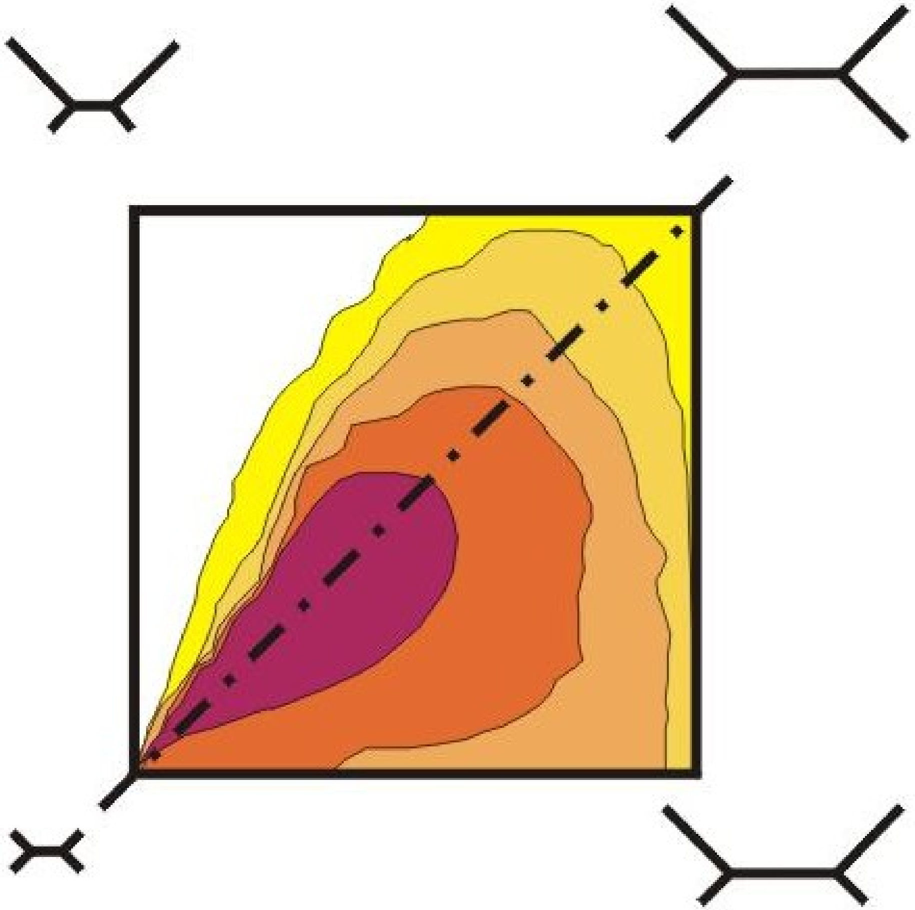


additive tree



ultrametric tree

# UPGMA and consistency



# Neighbor-Joining, NJ

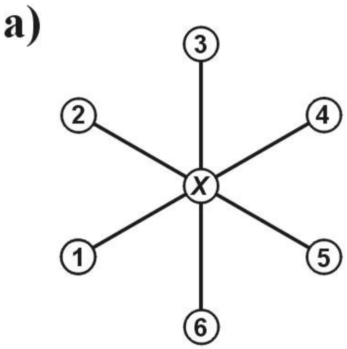
Algorithmic method

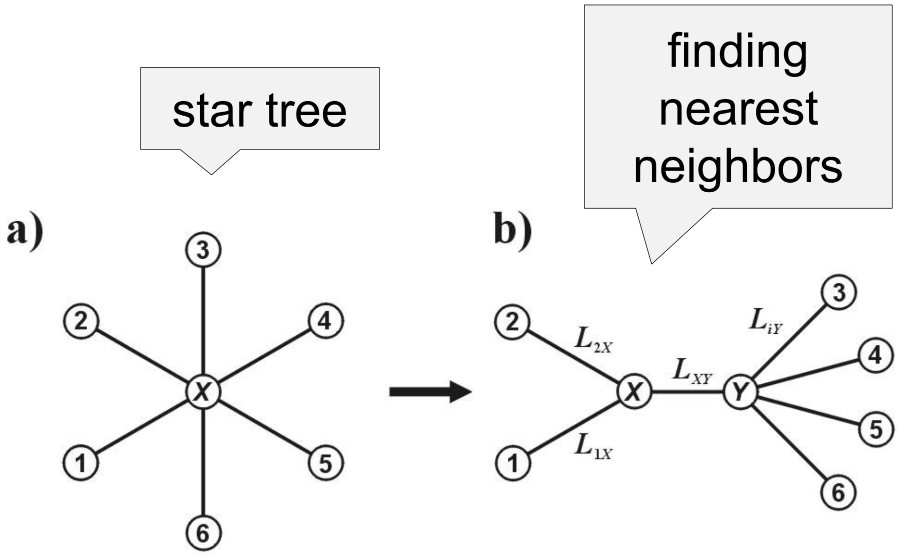
Principle of minimal evolution → minimizes sum of branch lenghts *S*

Each pair of nodes adjusted according to its divergence from others

Single additive tree

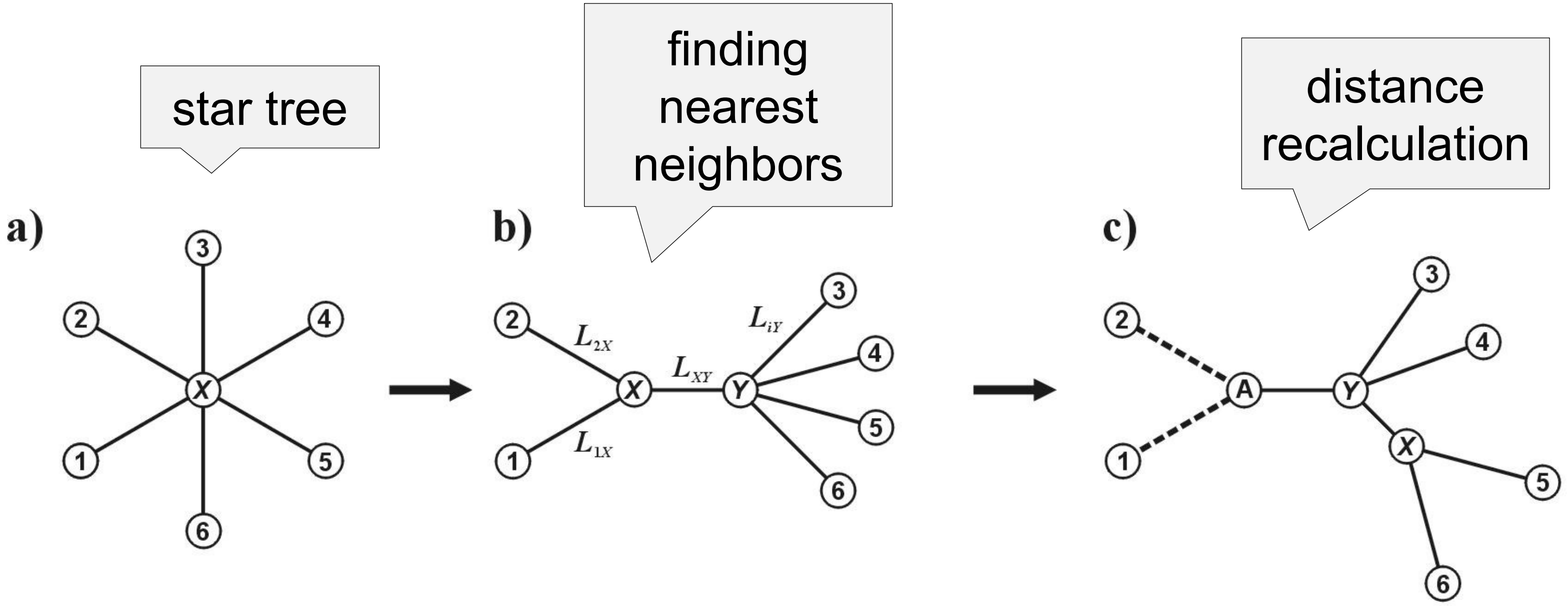
star tree





star tree

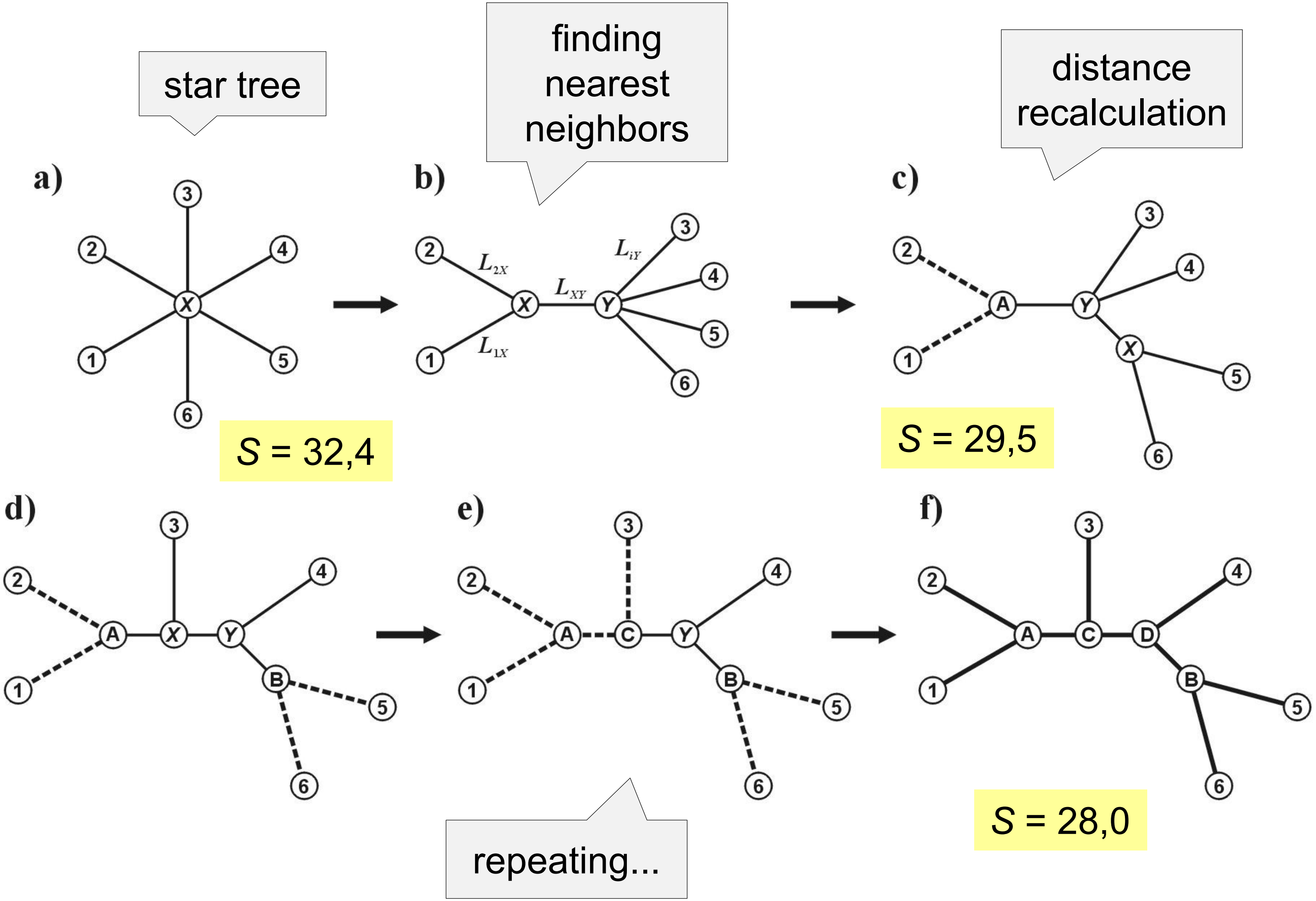
finding nearest neighbors



star tree

finding nearest neighbors

distance recalculation



star tree

finding nearest neighbors

distance recalculation

repeating...

*S* = 32,4

*S* = 29,5

*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