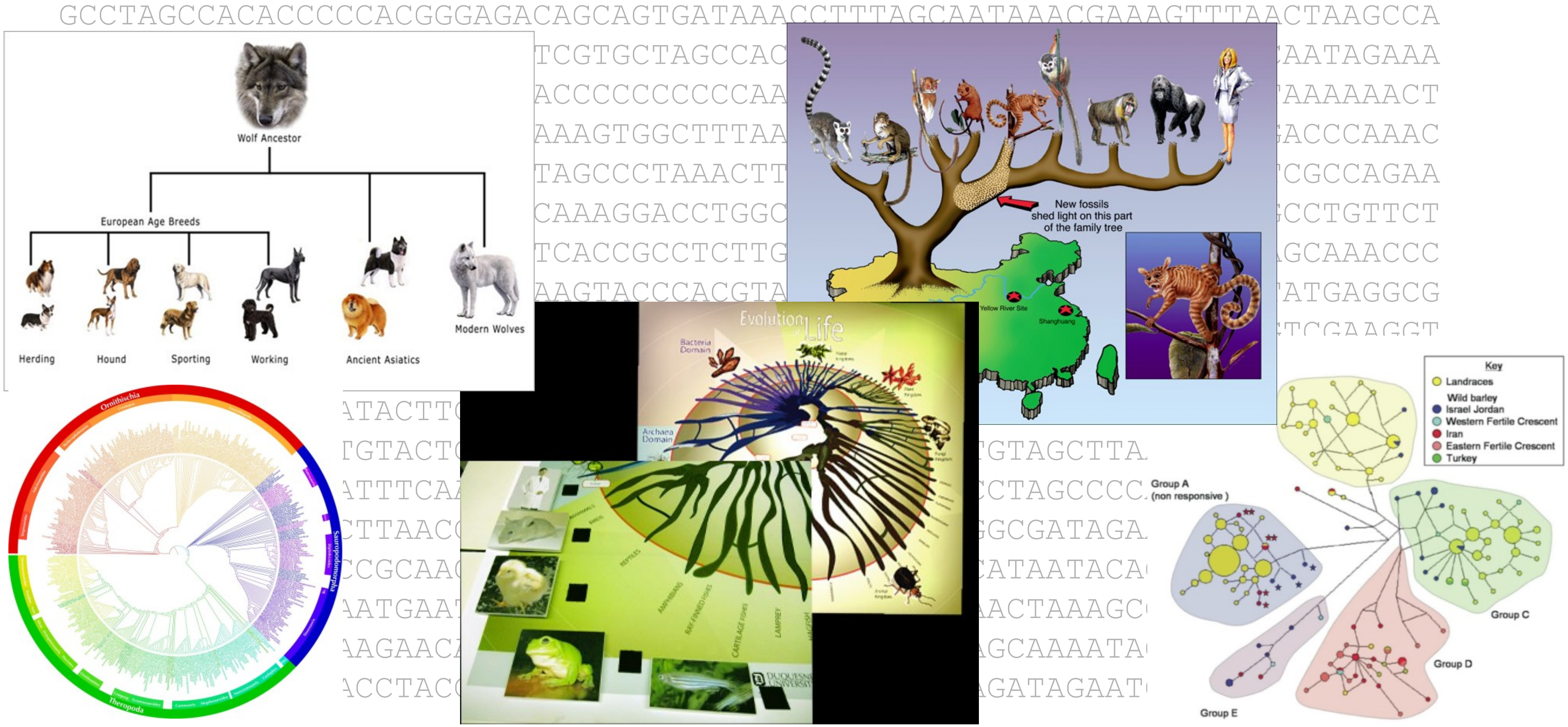
MODULARIZACE VÝUKY EVOLUČNÍ A EKOLOGICKÉ BIOLOGIE
CZ.1.07/2.2.00/15.0204

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

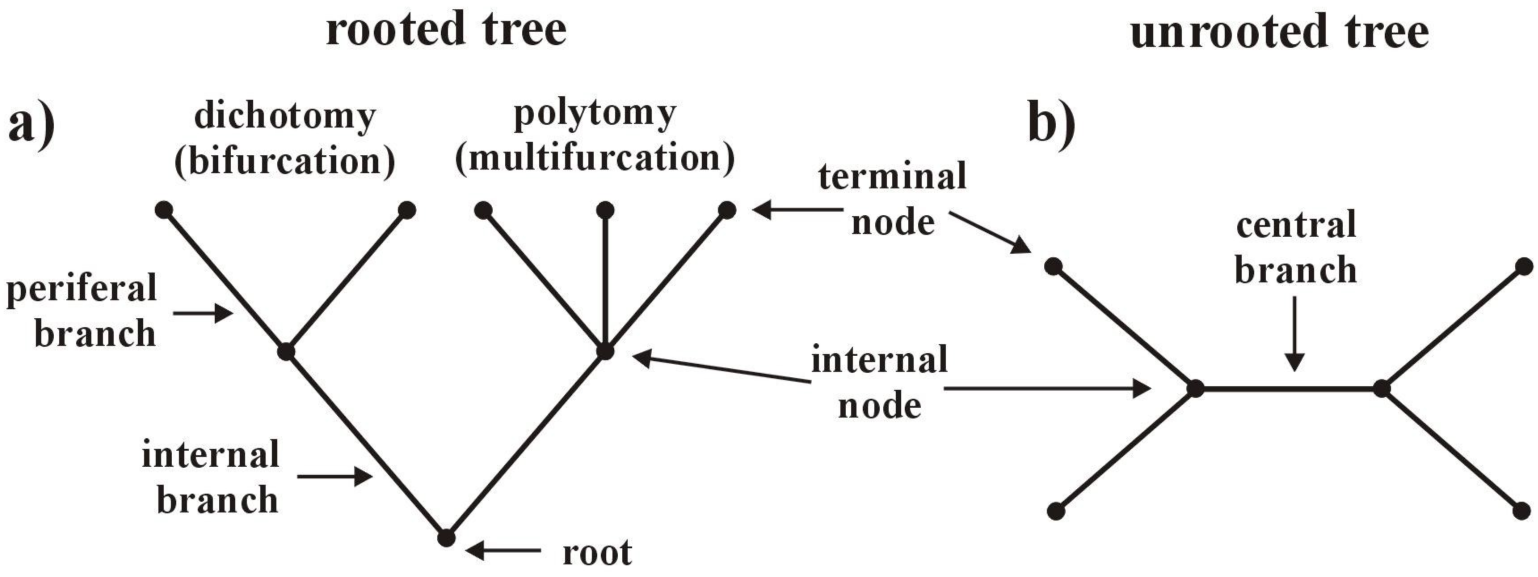
branches = edges (větve): peripheral, internal, central

nodes = vertices (uzly): internal, terminal

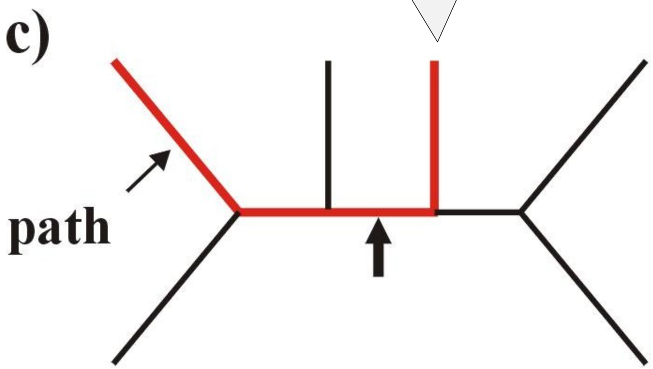
dichotomy = bifurcation, polytomy = multifurcation

OTU, HTU

tree topology

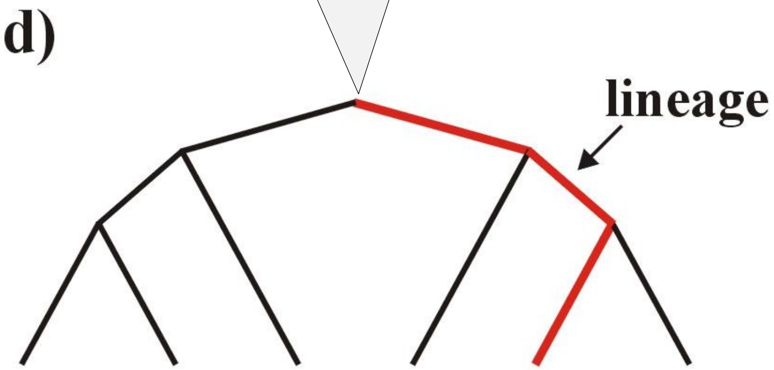


Definition of basic concepts:



connects two terminal nodes

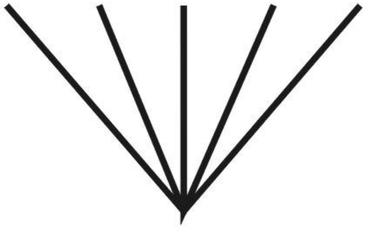
path (dráha)



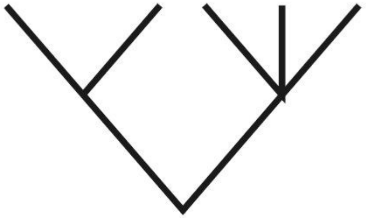
connects terminal node with root

lineage (linie)

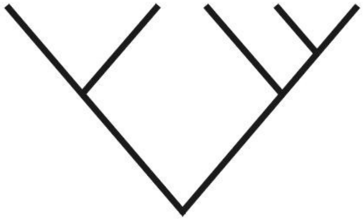
Definition of basic concepts:



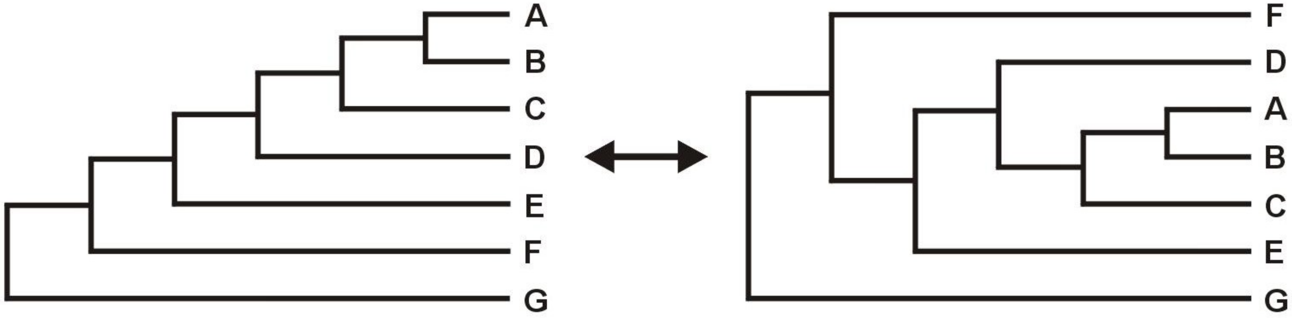
star tree

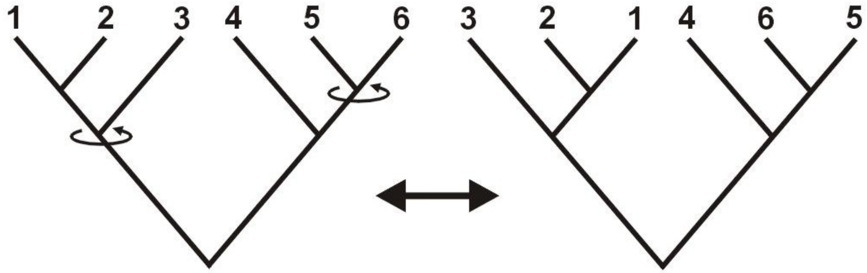


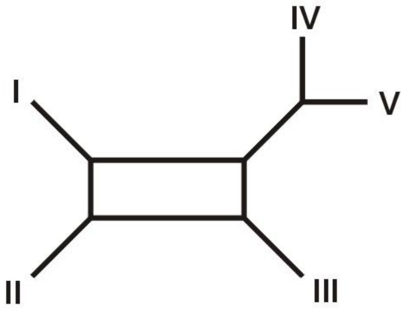
partly resolved



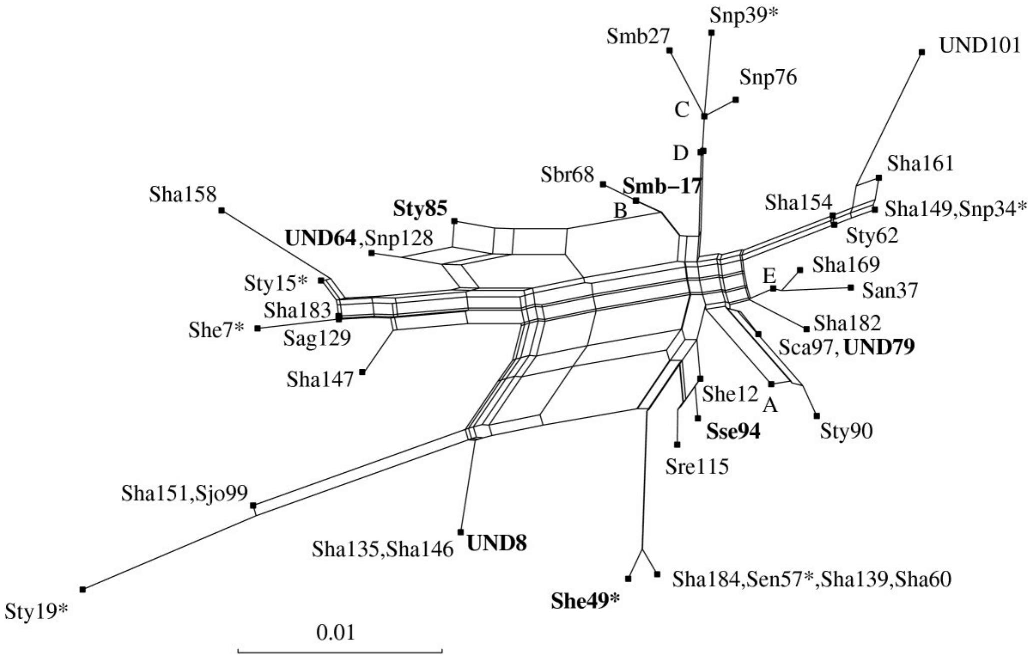
fully resolved

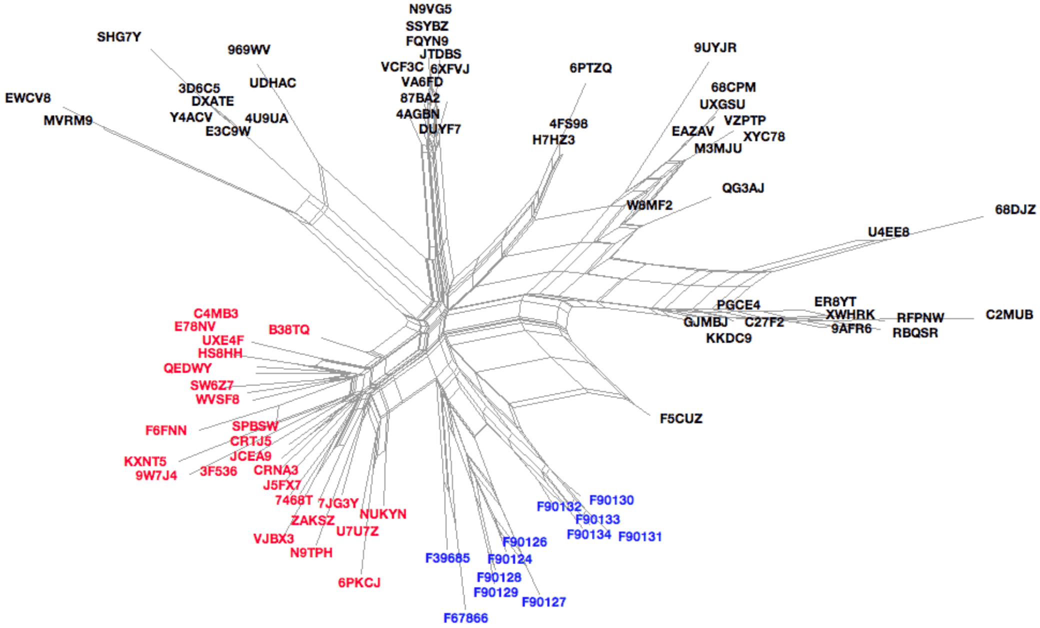




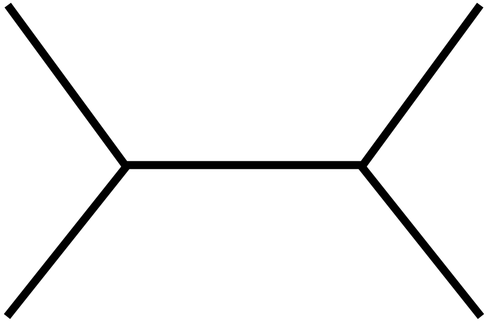


network

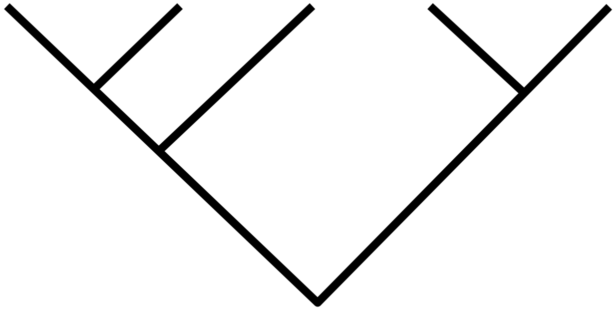

Snp39*
Smb27
UND101
Snp76
C
D
Sha161
Sbr68
Smb−17
Sha158
Sty85
B
Sha154
Sha149,Snp34*
UND64,Snp128
Sty62
E
Sha169
Sty15*
San37
Sha183
Sha182
She7*
Sag129
Sca97, UND79
Sha147
She12
A
Sse94
Sty90
Sre115
Sha151,Sjo99
Sha135,Sha146
UND8
Sha184,Sen57*,Sha139,Sha60
Sty19*
She49*
0.01


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

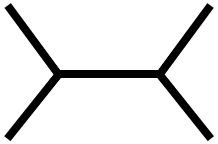
How many trees?



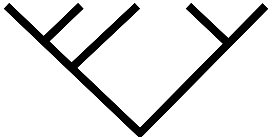
$$\frac{(2n-\;)!}{2^{n-}(n-\;)}$$



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



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



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

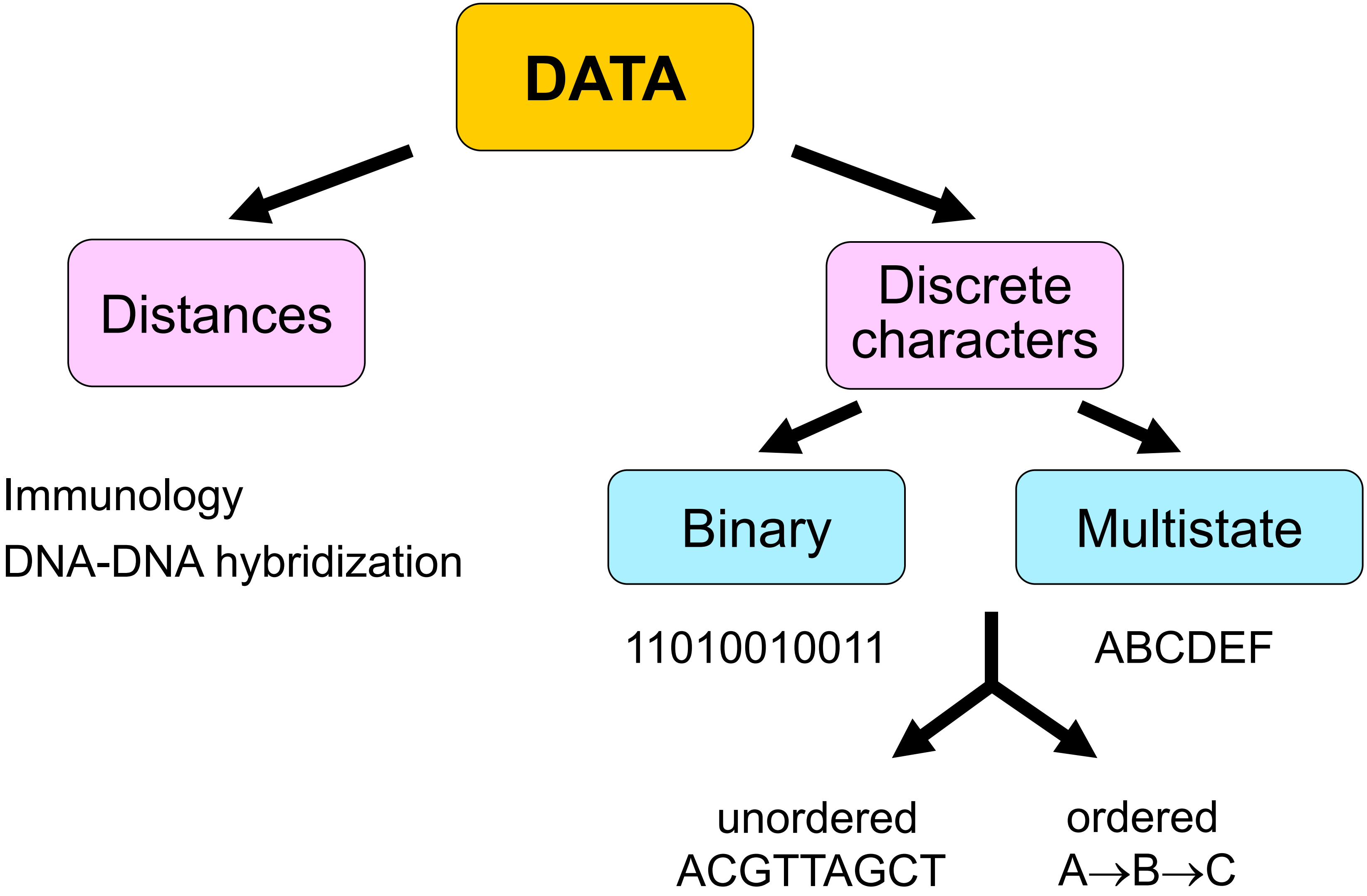
No. Taxons	Unrooted trees	Rooted trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425
11	34 459 425	654 729 075
12	654 729 075	13 749 310 575
13	13 749 310 575	316 234 143 225
14	316 234 143 225	7 905 853 580 625
15	7 905 853 580 625	213 458 046 676 875
20	213 458 046 676 875	8 375
30	8 200 794 532 637 891 559 375	10^{38}
40	$4{,}9518\times10^{38}$	10^{57}
50	$1{,}00986\times10^{57}$	$2{,}75292\times10^{76}$

> Avogadro constant*)

number of electrons in visible universe (Eddington number)

*) 6,022 141 79×10^{23} mol^{-1}

What type of data can we use?



DATA

Distances

Immunology

DNA-DNA hybridization

Discrete characters

Binary

11010010011

Multistate

ABCDEF

unordered

ACGTTAGCT

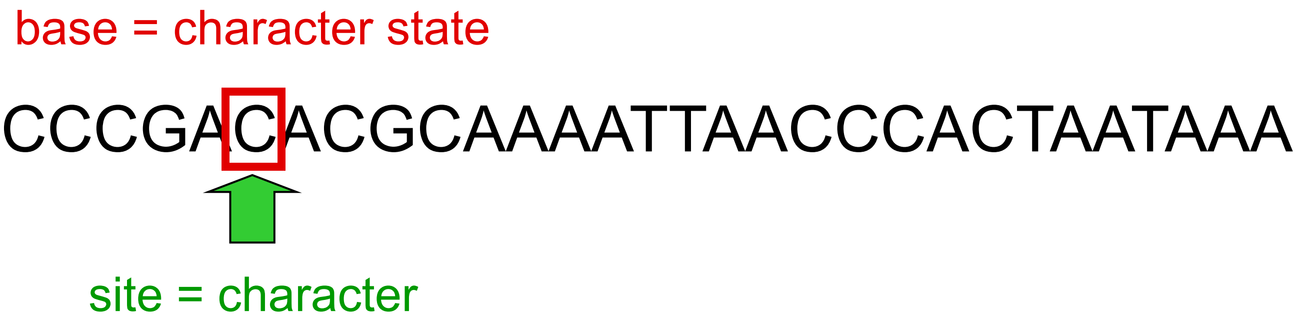
ordered

A→B→C

Types of data

Nucleotide and protein sequences:

H_sapiens MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGS



base = character state

P_troglod ATGACCCCGACACGCAAAATTAACCCACTAATAAA

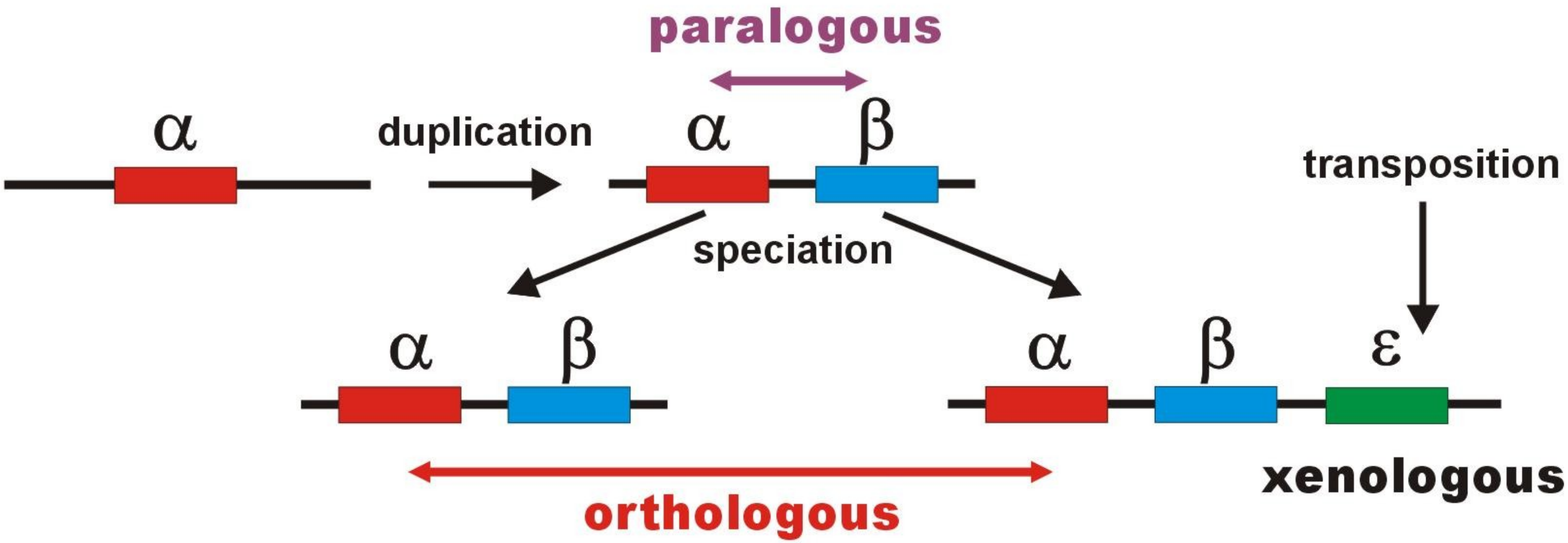
site = character

Types of data

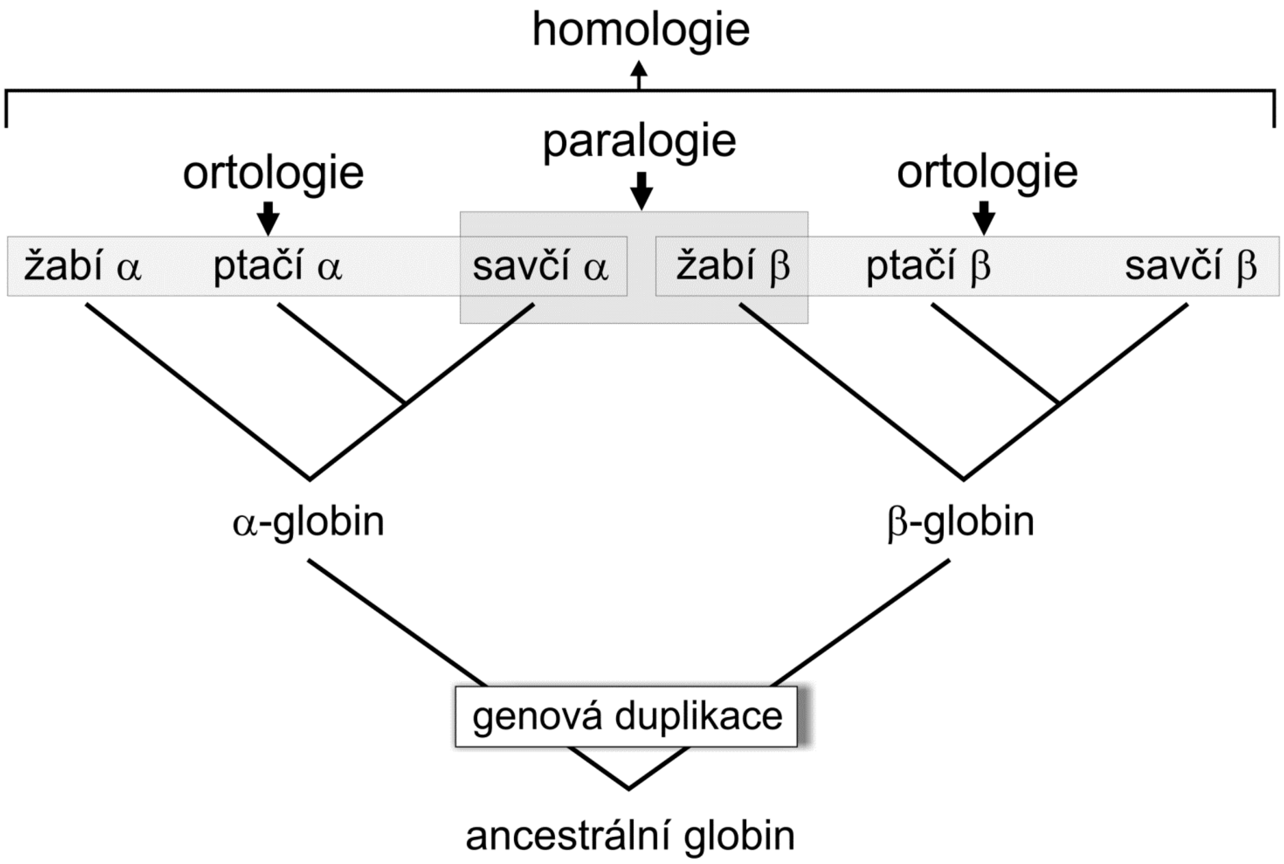
retroelements: SINE (*Alu*, B1, B2), LINE

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 gggtagcta 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.

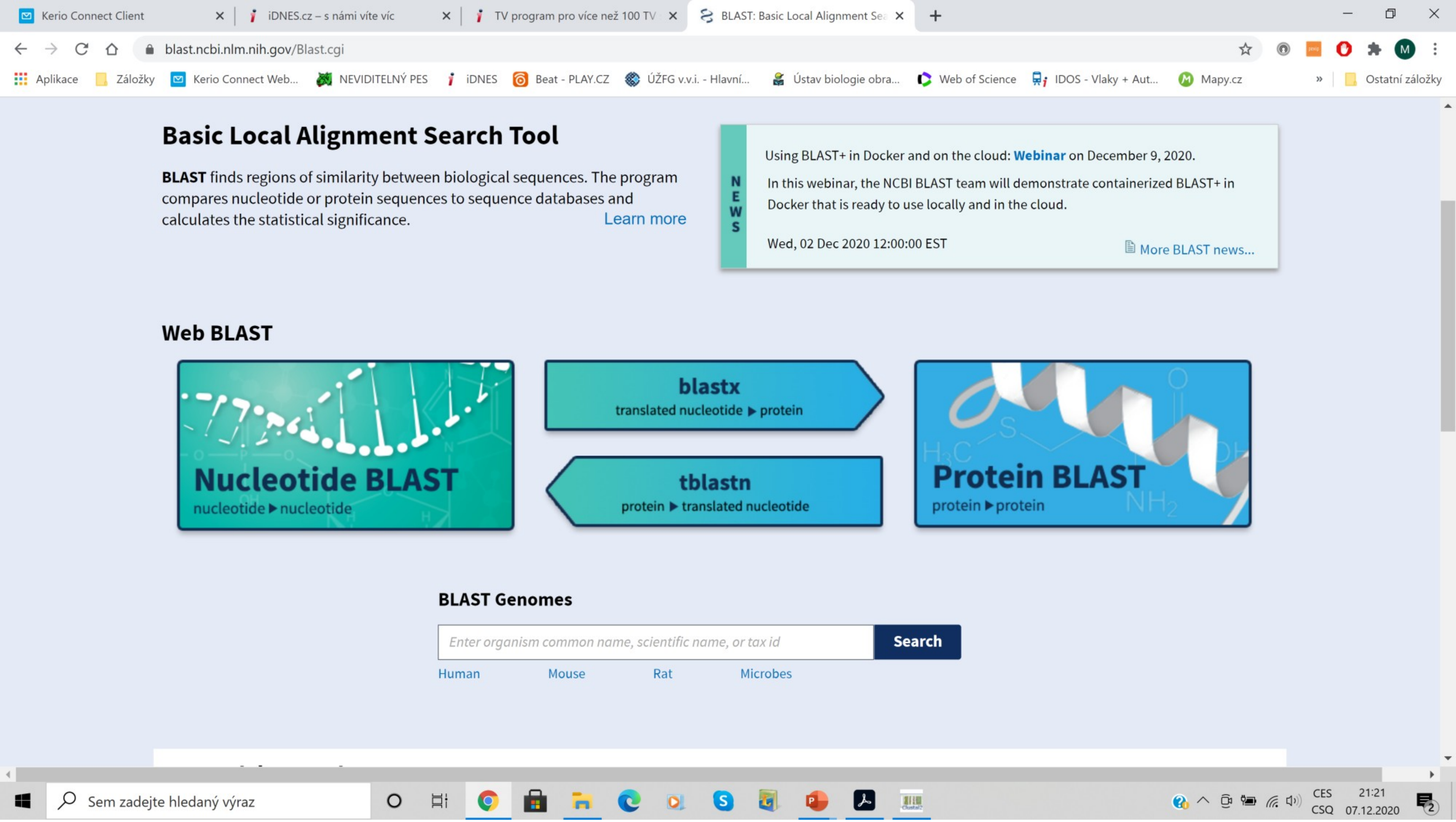
Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

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 *(paired-end or mate-pair reads only)*

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.

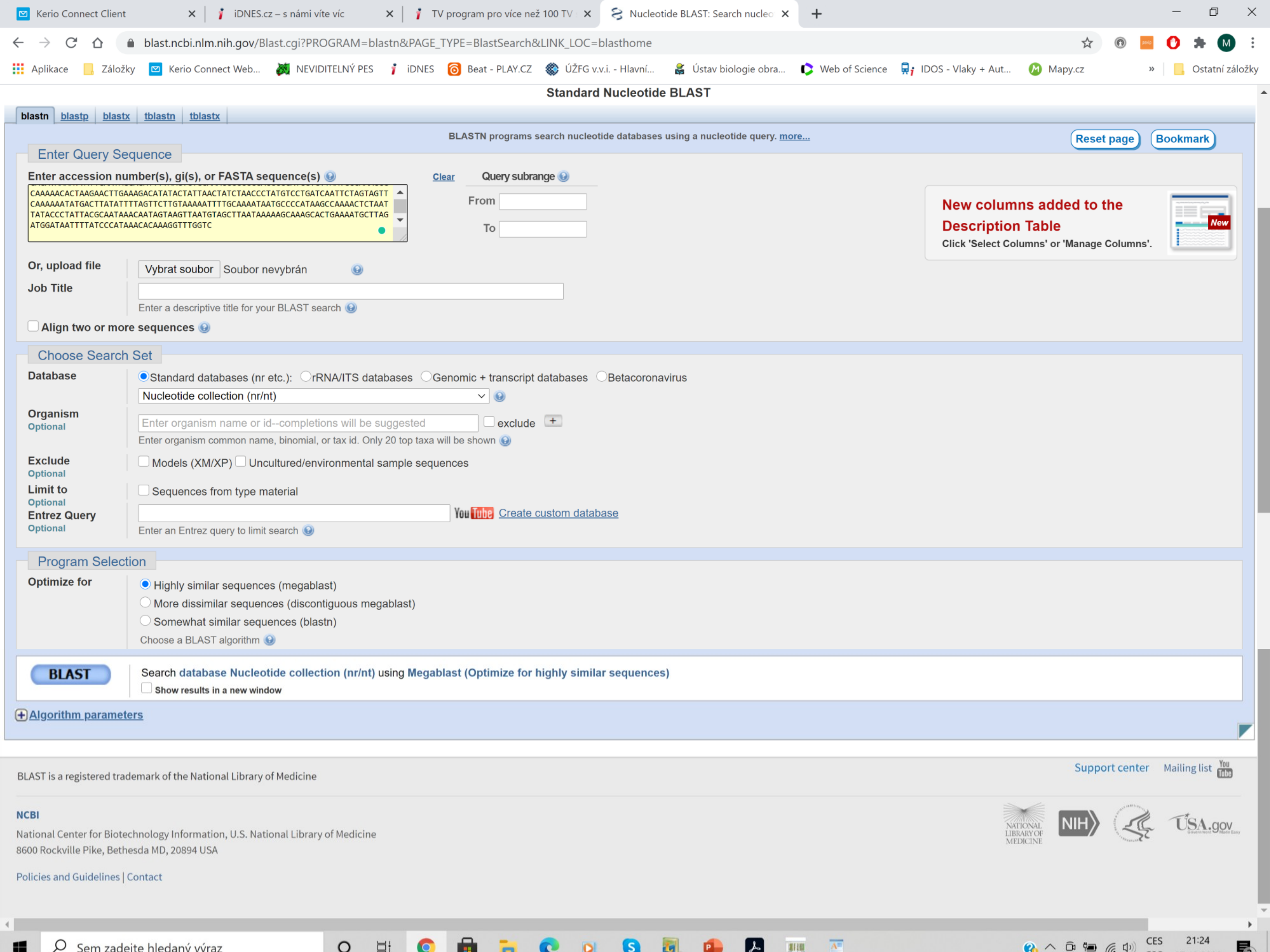
Wed, 02 Dec 2020 12:00:00 EST

More BLAST news...

Web BLAST

BLAST Genomes

Enter organism common name, scientific name, or tax id



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

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

```
CAAAAACACTAAGAACTTGAAAGACATATACTATTAACTATCTAACCCTATGTCCTGATCAATTCTAGTAGTT
CAAAAAATATGACTTATATTTTAGTTCTTGTAAAAATTTTGCAAAATAATGCCCCATAAGCCAAAACTCTAAT
TATACCCTATTACGCAATAAACAATAGTAAGTTAATGTAGCTTAATAAAAAGCAAAGCACTGAAAATGCTTAG
ATGGATAATTTTATCCCATAAACACAAAGGTTTGGTC
```

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 Optional

Enter organism name or id--completions will be suggested ☐ exclude

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

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

Limit to Optional ☐ Sequences from type material

Entrez Query Optional 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

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

☐ Show results in a new window

Algorithm parameters

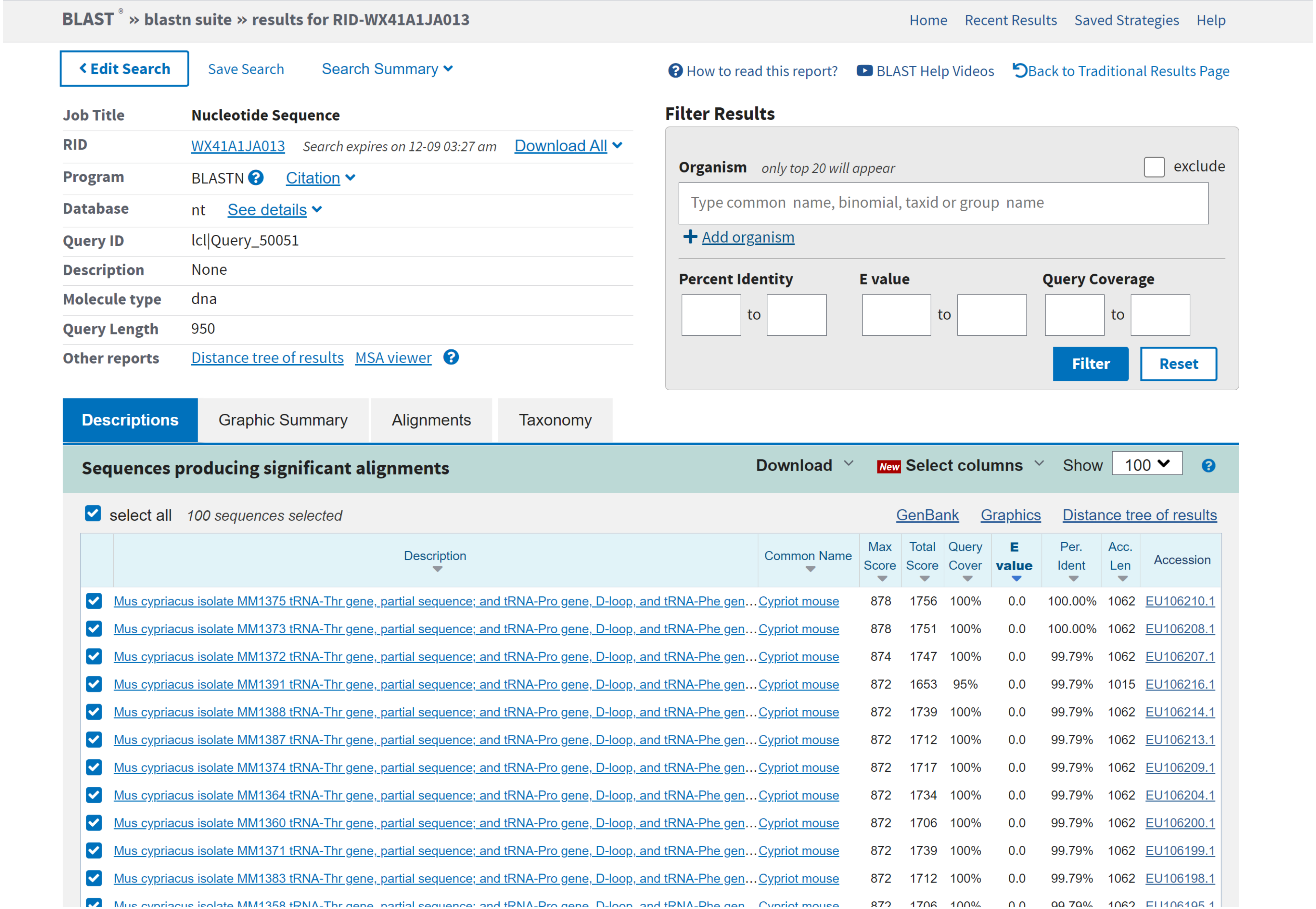
BLAST is a registered trademark of the National Library of Medicine

Support center Mailing list

NCBI

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

Policies and Guidelines | Contact



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	lcl|Query_50051
Description	None
Molecule type	dna
Query Length	950
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* ☐ exclude

Type common name, binomial, taxid or group name

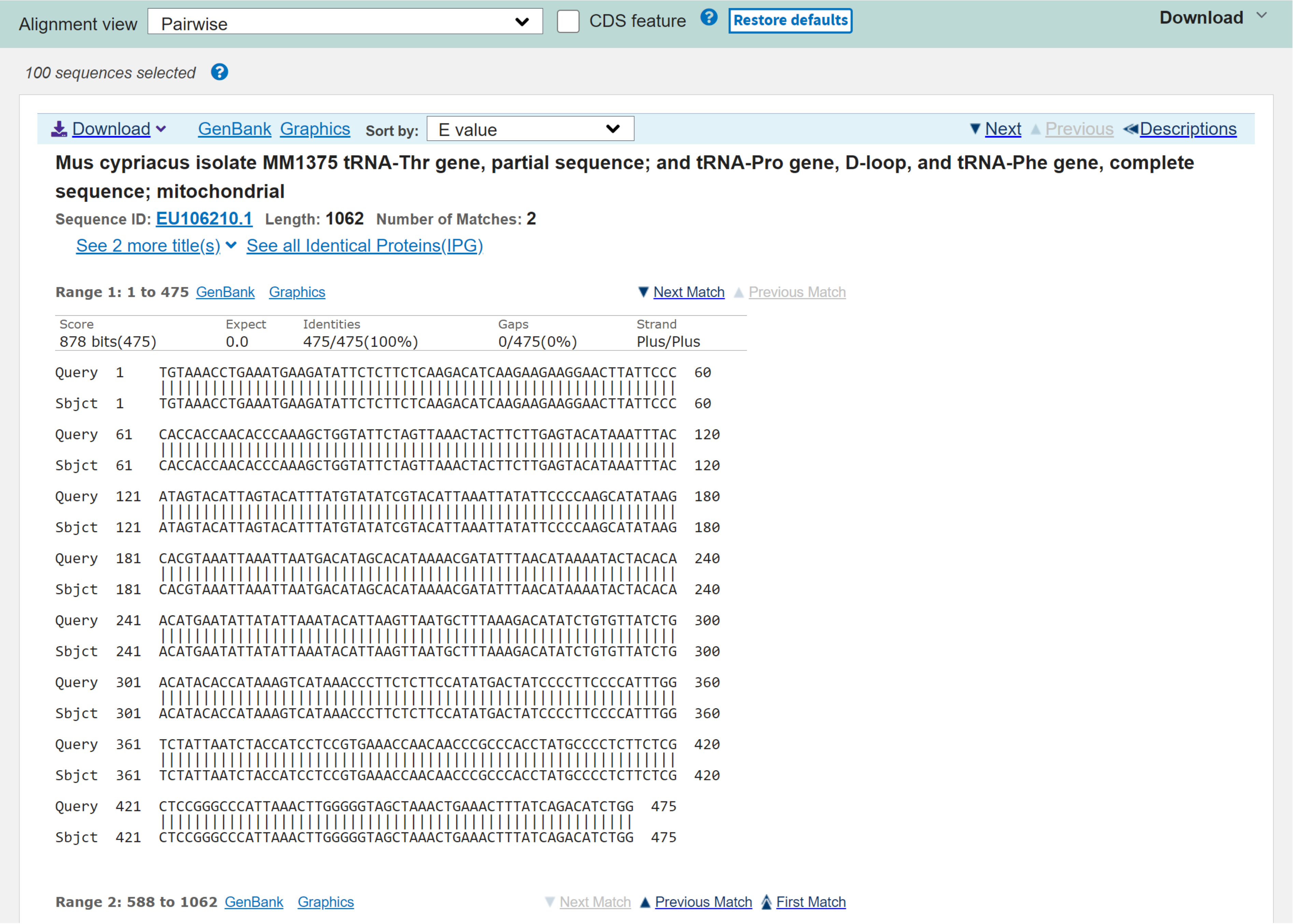
Add organism

Percent Identity ☐ to ☐ **E value** ☐ to ☐ **Query Coverage** ☐ to ☐

Sequences producing significant alignments

☑ select all *100 sequences selected*

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



100 sequences selected

Mus cypriacus 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)

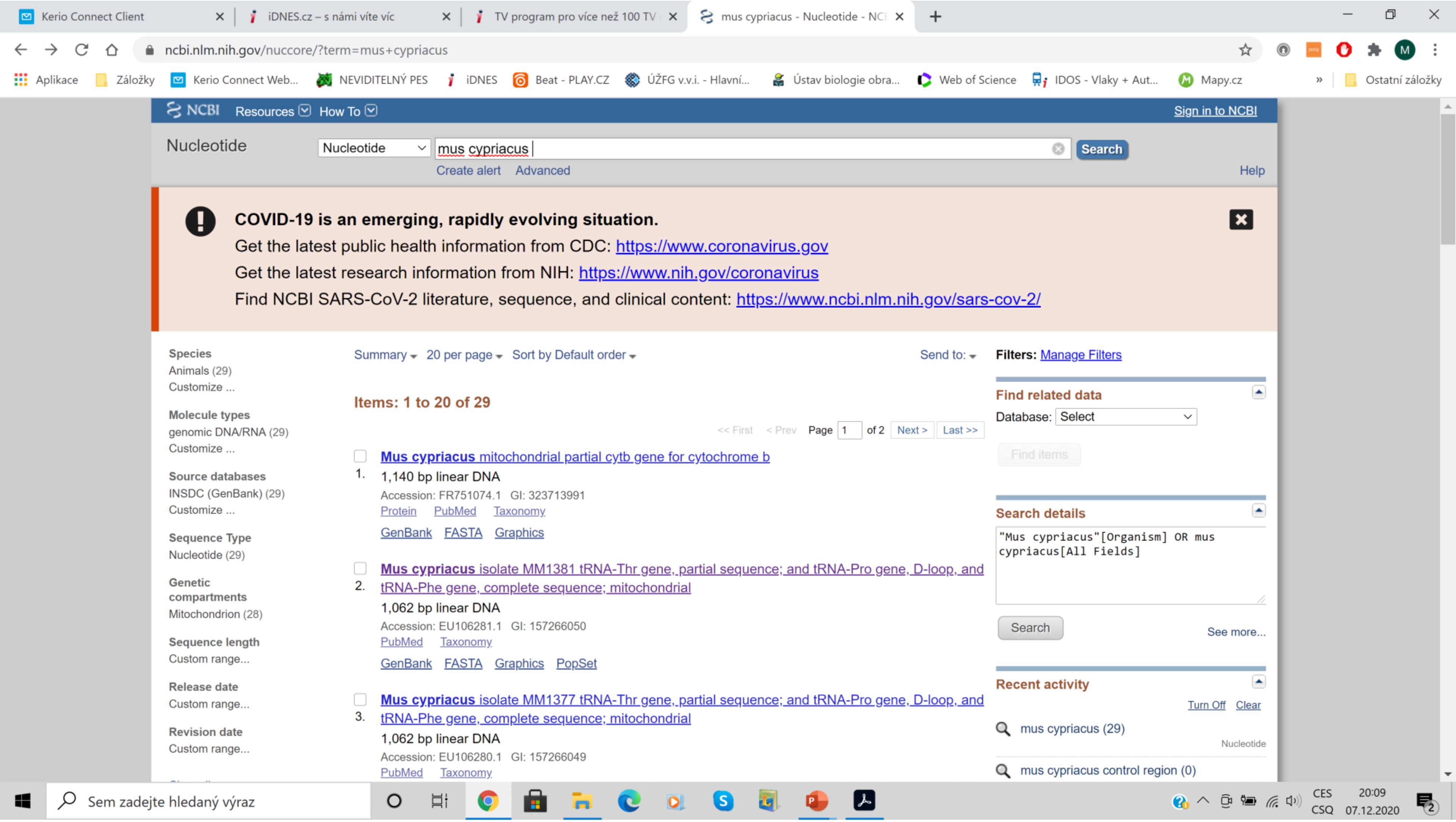
Range 1: 1 to 475 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
878 bits(475)	0.0	475/475(100%)	0/475(0%)	Plus/Plus

```
Query  1    TGTAAACCTGAAATGAAGATATTCTCTTCTCAAGACATCAAGAAGAAGGAACTTATTCCC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    TGTAAACCTGAAATGAAGATATTCTCTTCTCAAGACATCAAGAAGAAGGAACTTATTCCC  60

Query  61   CACCACCAACACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTAC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   CACCACCAACACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTAC  120

Query  121  ATAGTACATTAGTACATTTATGTATATCGTACATTAAATTATATTCCCCAAGCATATAAG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  ATAGTACATTAGTACATTTATGTATATCGTACATTAAATTATATTCCCCAAGCATATAAG  180

Query  181  CACGTAAATTAAATTAATGACATAGCACATAAAACGATATTTAACATAAAATACTACACA  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  CACGTAAATTAAATTAATGACATAGCACATAAAACGATATTTAACATAAAATACTACACA  240

Query  241  ACATGAATATTATATTAAATACATTAAGTTAATGCTTTAAAGACATATCTGTGTTATCTG  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  ACATGAATATTATATTAAATACATTAAGTTAATGCTTTAAAGACATATCTGTGTTATCTG  300

Query  301  ACATACACCATAAAGTCATAAACCCTTCTCTTCCATATGACTATCCCCTTCCCCATTTGG  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  ACATACACCATAAAGTCATAAACCCTTCTCTTCCATATGACTATCCCCTTCCCCATTTGG  360

Query  361  TCTATTAATCTACCATCCTCCGTGAAACCAACAACCCGCCCACCTATGCCCCTCTTCTCG  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  TCTATTAATCTACCATCCTCCGTGAAACCAACAACCCGCCCACCTATGCCCCTCTTCTCG  420

Query  421  CTCCGGGCCCATTAAACTTGGGGGTAGCTAAACTGAAACTTTATCAGACATCTGG  475
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  CTCCGGGCCCATTAAACTTGGGGGTAGCTAAACTGAAACTTTATCAGACATCTGG  475
```

Range 2: 588 to 1062 GenBank Graphics

GenBank



Nucleotide

Nucleotide | mus cypriacus | Search

Create alert Advanced Help

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)
Customize ...

Molecule types
genomic DNA/RNA (29)
Customize ...

Source databases
INSDC (GenBank) (29)
Customize ...

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: ▾

Items: 1 to 20 of 29

<< First < Prev Page 1 of 2 Next > Last >>

☐ 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

Filters: Manage Filters

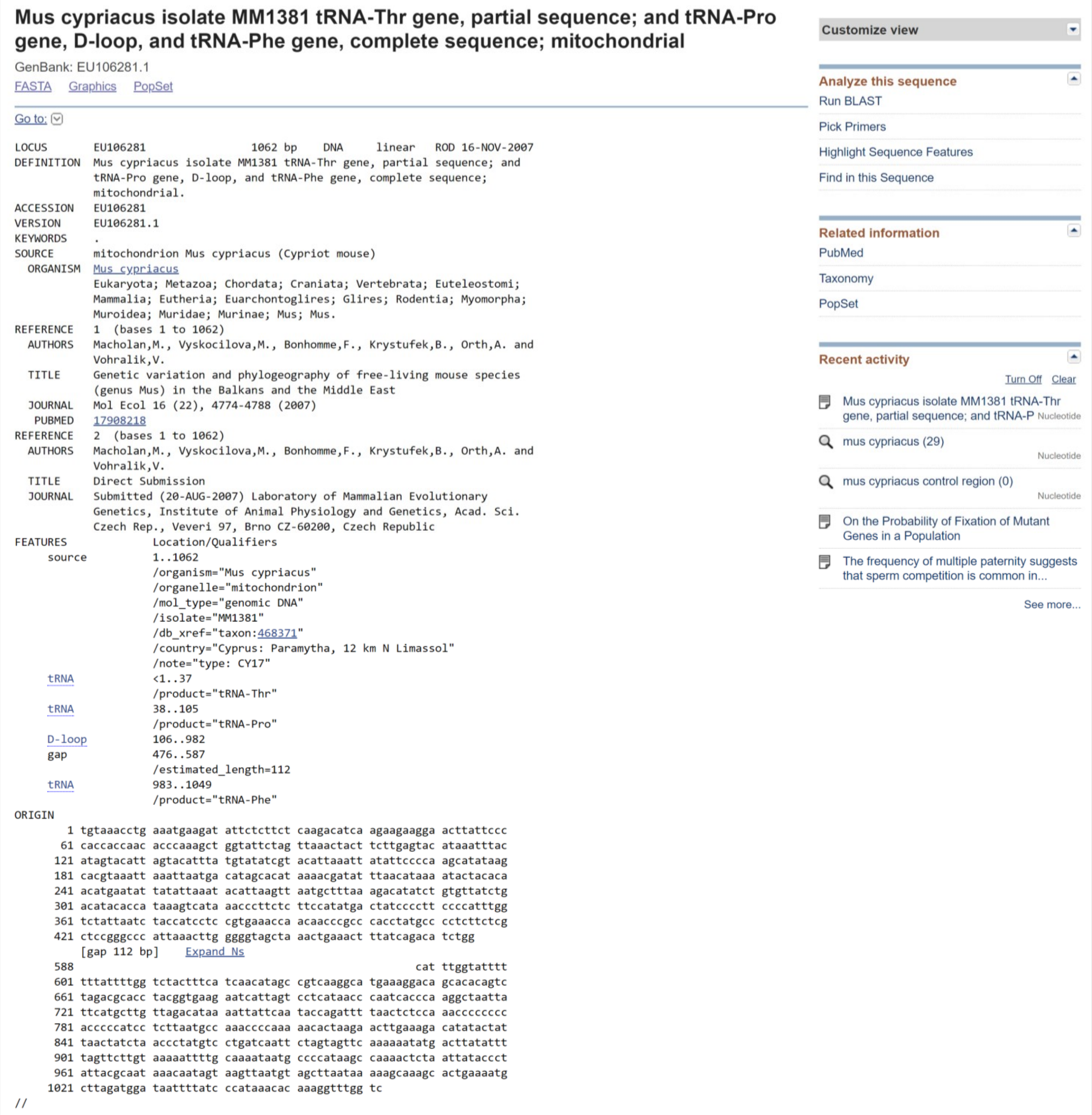
Find related data
Database: Select
Find items

Search details

```
"Mus cypriacus"[Organism] OR mus
cypriacus[All Fields]
```

Search See more...

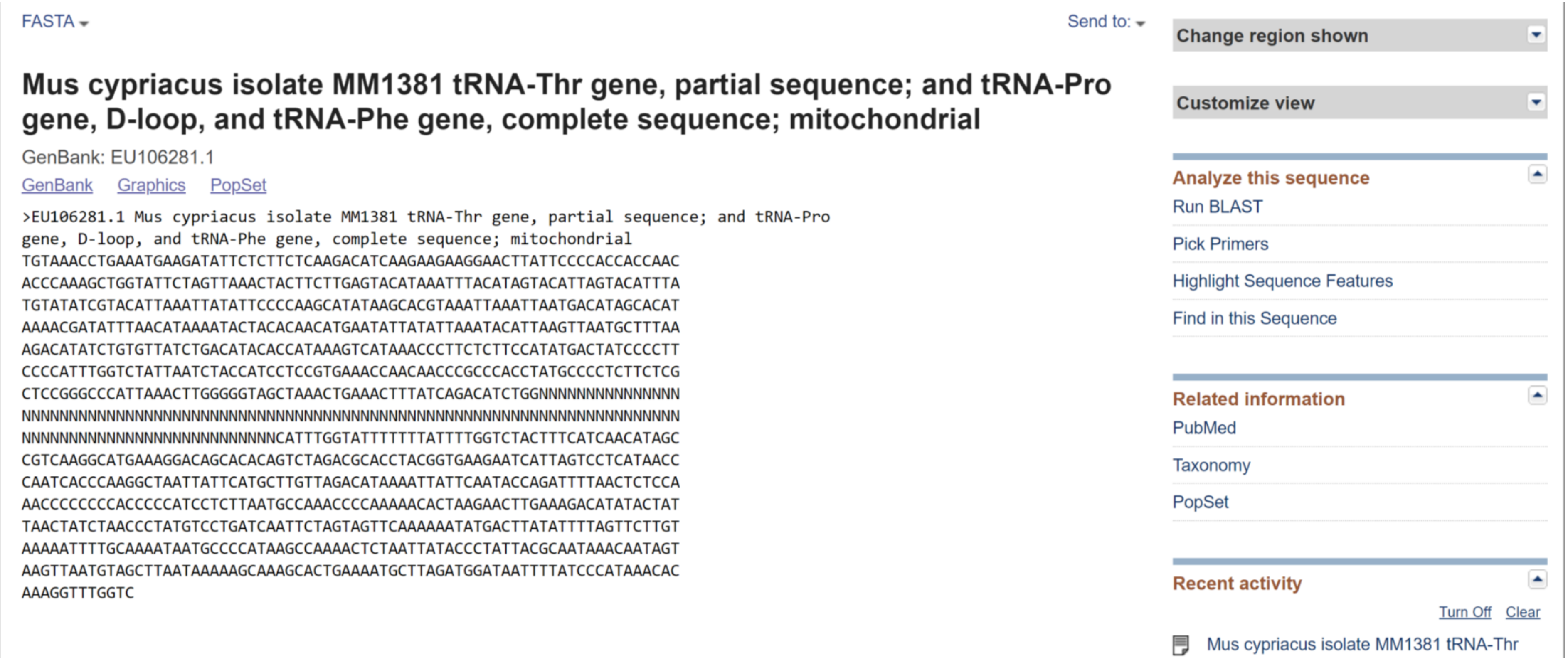
Recent activity
Turn Off Clear
mus cypriacus (29) Nucleotide
mus cypriacus control region (0)



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

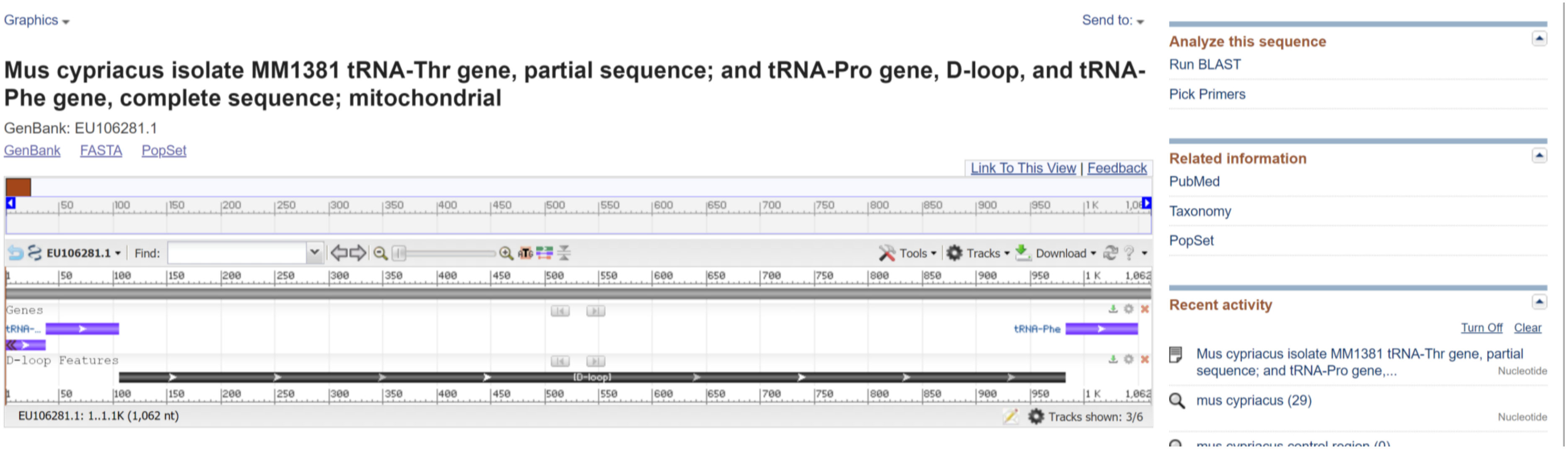
```
LOCUS       EU106281                1062 bp    DNA     linear   ROD 16-NOV-2007
DEFINITION  Mus cypriacus isolate MM1381 tRNA-Thr gene, partial sequence; and
            tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete sequence;
            mitochondrial.
ACCESSION   EU106281
VERSION     EU106281.1
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             Location/Qualifiers
     source          1..1062
                     /organism="Mus cypriacus"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /isolate="MM1381"
                     /db_xref="taxon:468371"
                     /country="Cyprus: Paramytha, 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 acccaaagct ggtattctag ttaaactact tcttgagtac ataaatttac
      121 atagtacatt agtacattta tgtatatcgt acattaaatt atattcccca agcatataag
      181 cacgtaaatt aaattaatga catagcacat aaaacgatat ttaacataaa atactacaca
      241 acatgaatat tatattaaat acattaagtt aatgctttaa agacatatct gtgttatctg
      301 acatacacca taaagtcata aacccttctc ttccatatga ctatcccctt ccccatttgg
      361 tctattaatc taccatcctc cgtgaaacca acaacccgcc cacctatgcc cctcttctcg
      421 ctccgggccc attaaacttg ggggtagcta aactgaaact ttatcagaca tctgg
          [gap 112 bp]    Expand Ns
      588                                                    cat ttggtatttt
      601 tttattttgg tctactttca tcaacatagc cgtcaaggca tgaaaggaca gcacacagtc
      661 tagacgcacc tacggtgaag aatcattagt cctcataacc caatcaccca aggctaatta
      721 ttcatgcttg ttagacataa aattattcaa taccagattt taactctcca aacccccccc
      781 acccccatcc tcttaatgcc aaaccccaaa aacactaaga acttgaaaga catatactat
      841 taactatcta accctatgtc ctgatcaatt ctagtagttc aaaaaatatg acttatattt
      901 tagttcttgt aaaaattttg caaaataatg ccccataagc caaaactcta attataccct
      961 attacgcaat aaacaatagt aagttaatgt agcttaataa aaagcaaagc actgaaaatg
     1021 cttagatgga taattttatc ccataaacac aaaggtttgg tc
//
```



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

```
>EU106281.1 Mus cypriacus isolate MM1381 tRNA-Thr gene, partial sequence; and tRNA-Pro
gene, D-loop, and tRNA-Phe gene, complete sequence; mitochondrial
TGTAAACCTGAAATGAAGATATTCTCTTCTCAAGACATCAAGAAGAAGGAACTTATTCCCCACCACCAAC
ACCCAAAGCTGGTATTCTAGTTAAACTACTTCTTGAGTACATAAATTTACATAGTACATTAGTACATTTA
TGTATATCGTACATTAAATTATATTCCCCAAGCATATAAGCACGTAAATTAAATTAATGACATAGCACAT
AAAACGATATTTAACATAAAATACTACACAACATGAATATTATATTAAATACATTAAGTTAATGCTTTAA
AGACATATCTGTGTTATCTGACATACACCATAAAGTCATAAACCCTTCTCTTCCATATGACTATCCCCTT
CCCCATTTGGTCTATTAATCTACCATCCTCCGTGAAACCAACAACCCGCCCACCTATGCCCCTCTTCTCG
CTCCGGGCCCATTAAACTTGGGGGTAGCTAAACTGAAACTTTATCAGACATCTGGNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNCATTTGGTATTTTTTTATTTTGGTCTACTTTCATCAACATAGC
CGTCAAGGCATGAAAGGACAGCACACAGTCTAGACGCACCTACGGTGAAGAATCATTAGTCCTCATAACC
CAATCACCCAAGGCTAATTATTCATGCTTGTTAGACATAAAATTATTCAATACCAGATTTTAACTCTCCA
AACCCCCCCCACCCCCATCCTCTTAATGCCAAACCCCAAAAACACTAAGAACTTGAAAGACATATACTAT
TAACTATCTAACCCTATGTCCTGATCAATTCTAGTAGTTCAAAAAATATGACTTATATTTTAGTTCTTGT
AAAAATTTTGCAAAATAATGCCCCATAAGCCAAAACTCTAATTATACCCTATTACGCAATAAACAATAGT
AAGTTAATGTAGCTTAATAAAAAGCAAAGCACTGAAAATGCTTAGATGGATAATTTTATCCCATAAACAC
AAAGGTTTGGTC
```



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

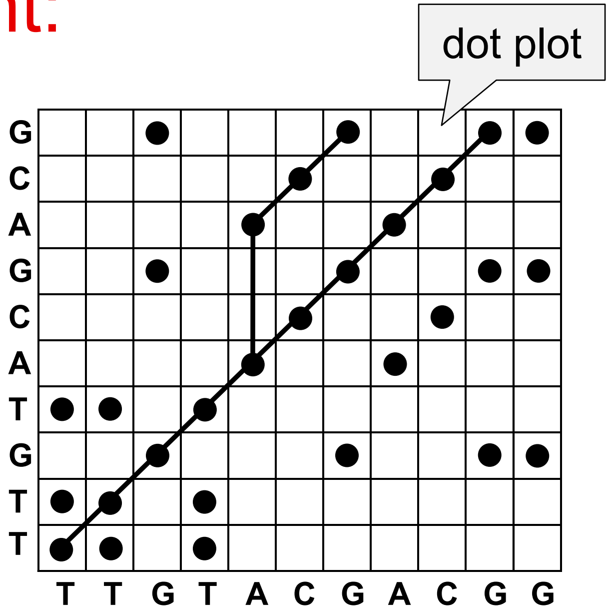
EU106281.1: 1..1.1K (1,062 nt)

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

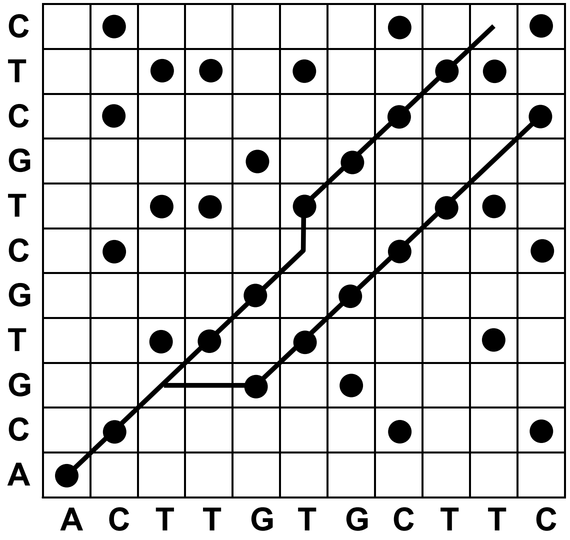
mus cypriacus (29)

Sequence alignment:

```
Sequence 1    TTGTACGACGG
Sequence 2    TTGTACGACG

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



```
Sequence 1    ACTTGTGCTTC
Sequence 2    ACGTGCTGCTC

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

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



Sequence alignment:

$GP = g + hl$
g - gap penalty
h – gap extension penalty
l – gap length

Gap penalty:

g = penalizace za výskyt mezery (1×)

h = extenze za každou „pomlčku“

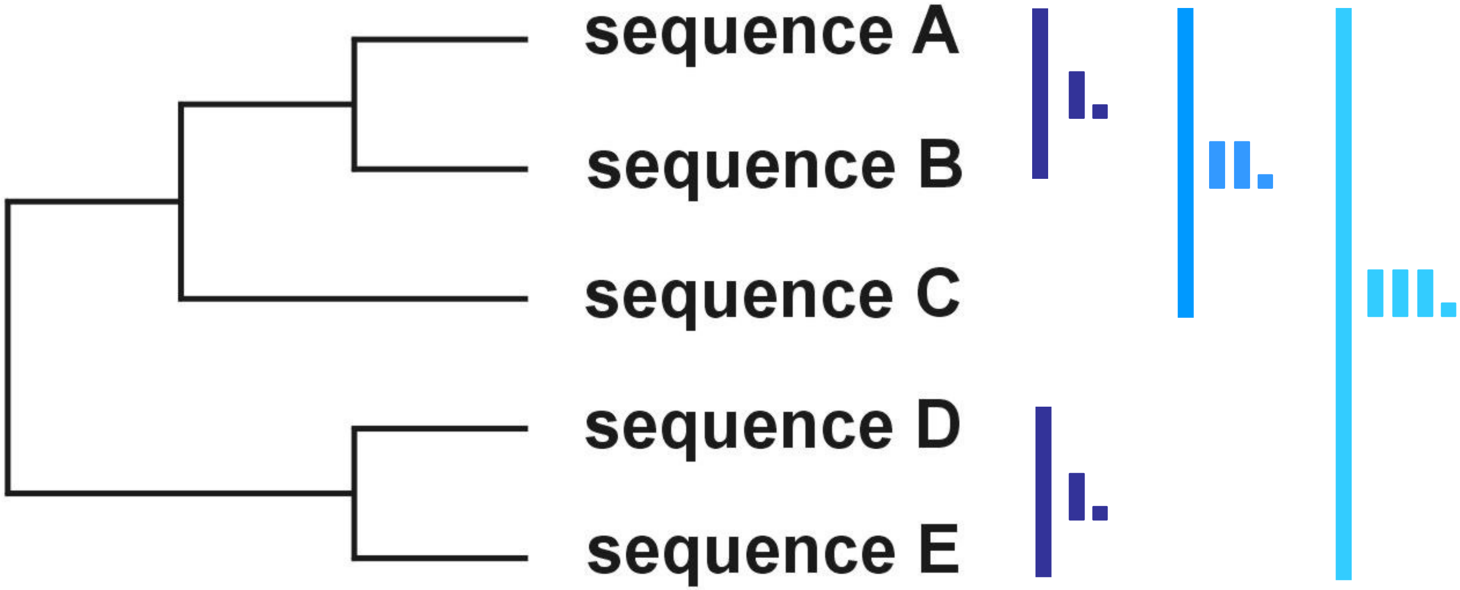
l = délka mezery (= počet „pomlček“)

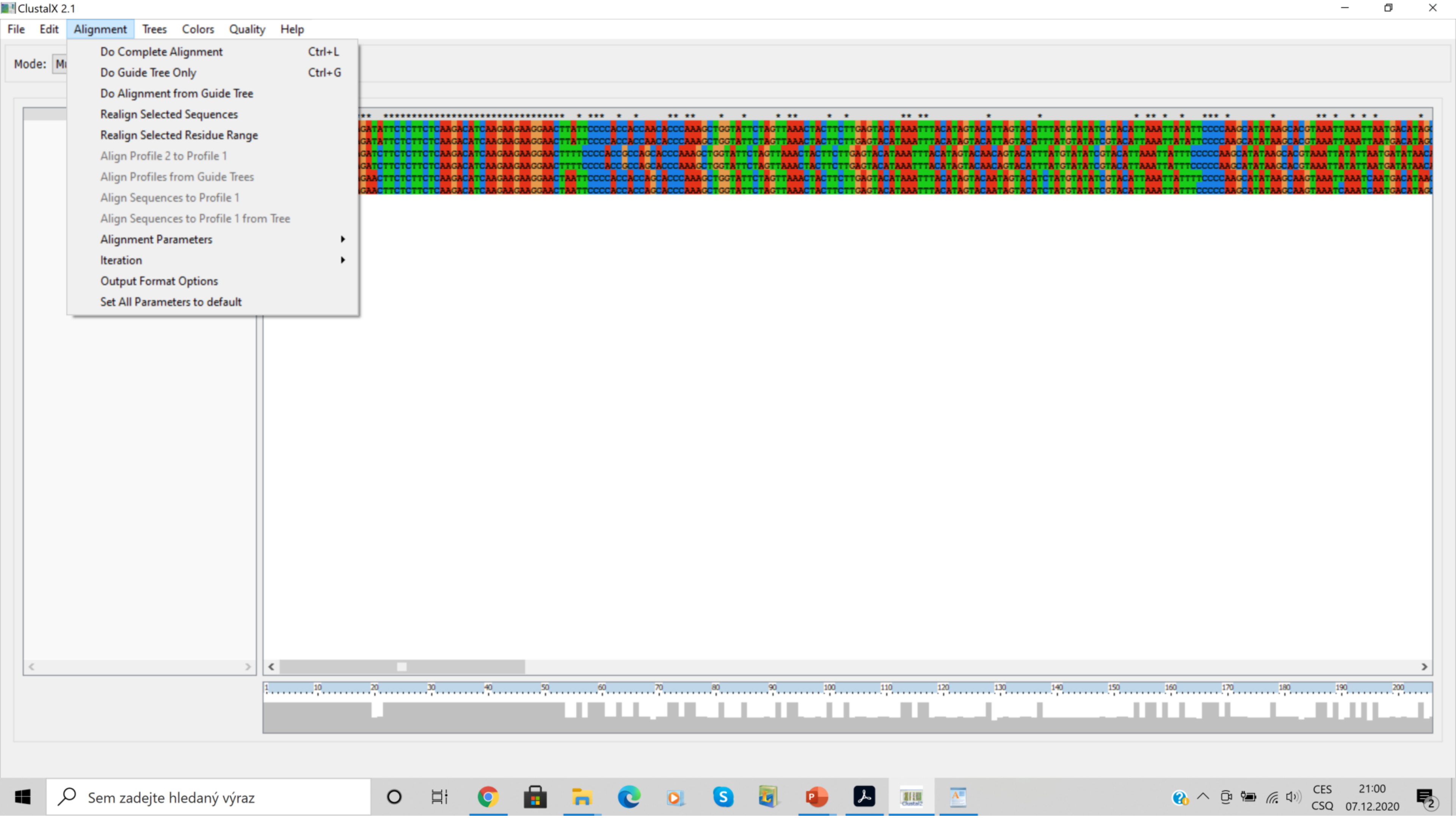
Př.: GC□ □ □ □ □ TTAA
$l = 5$, $g = x$, $h = 5x$

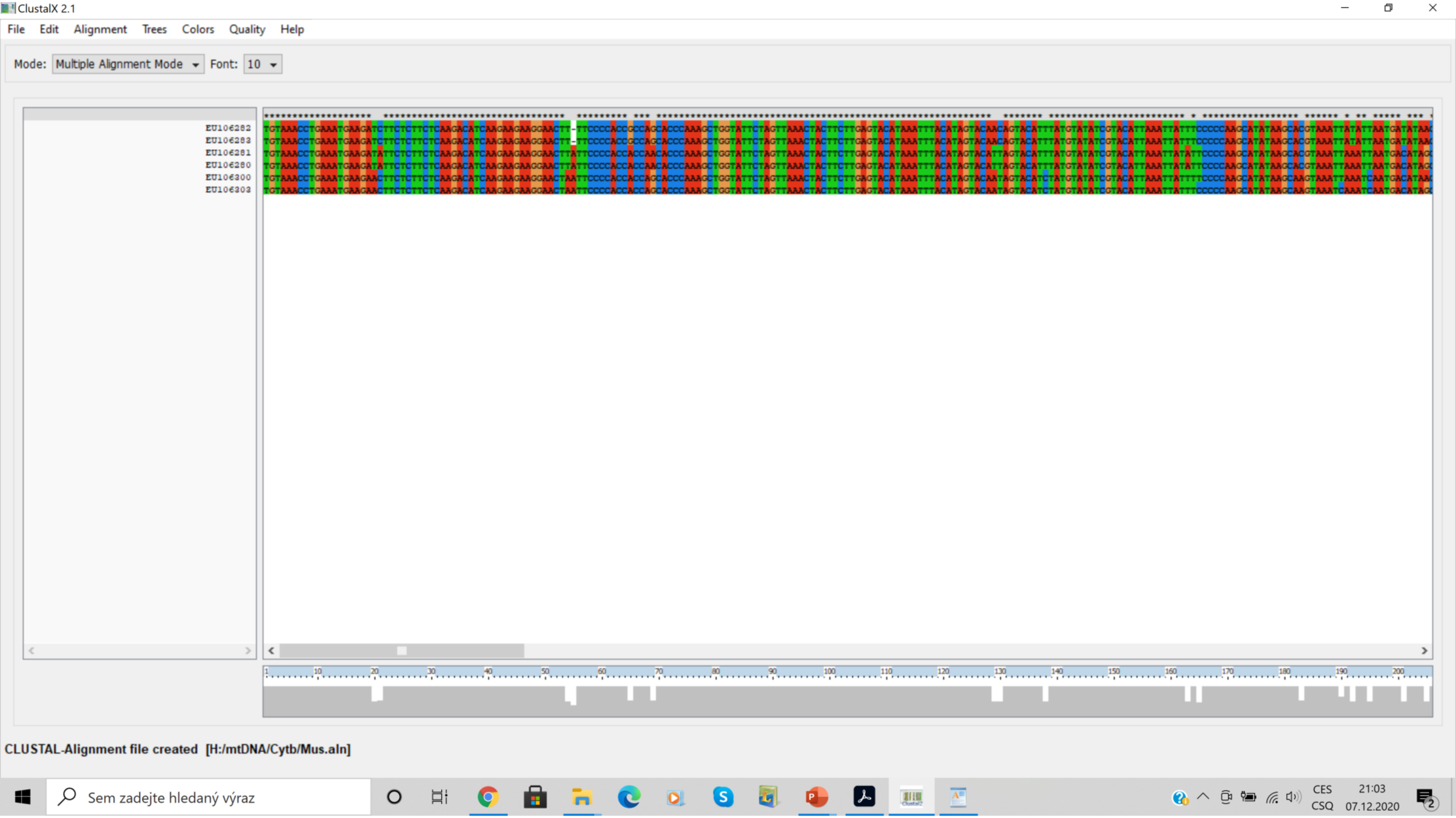
Progressive alignment - ClustalX

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
EU106283
EU106281
EU106280
EU106300
EU106303

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

Problem with progressive alignment

6 species:

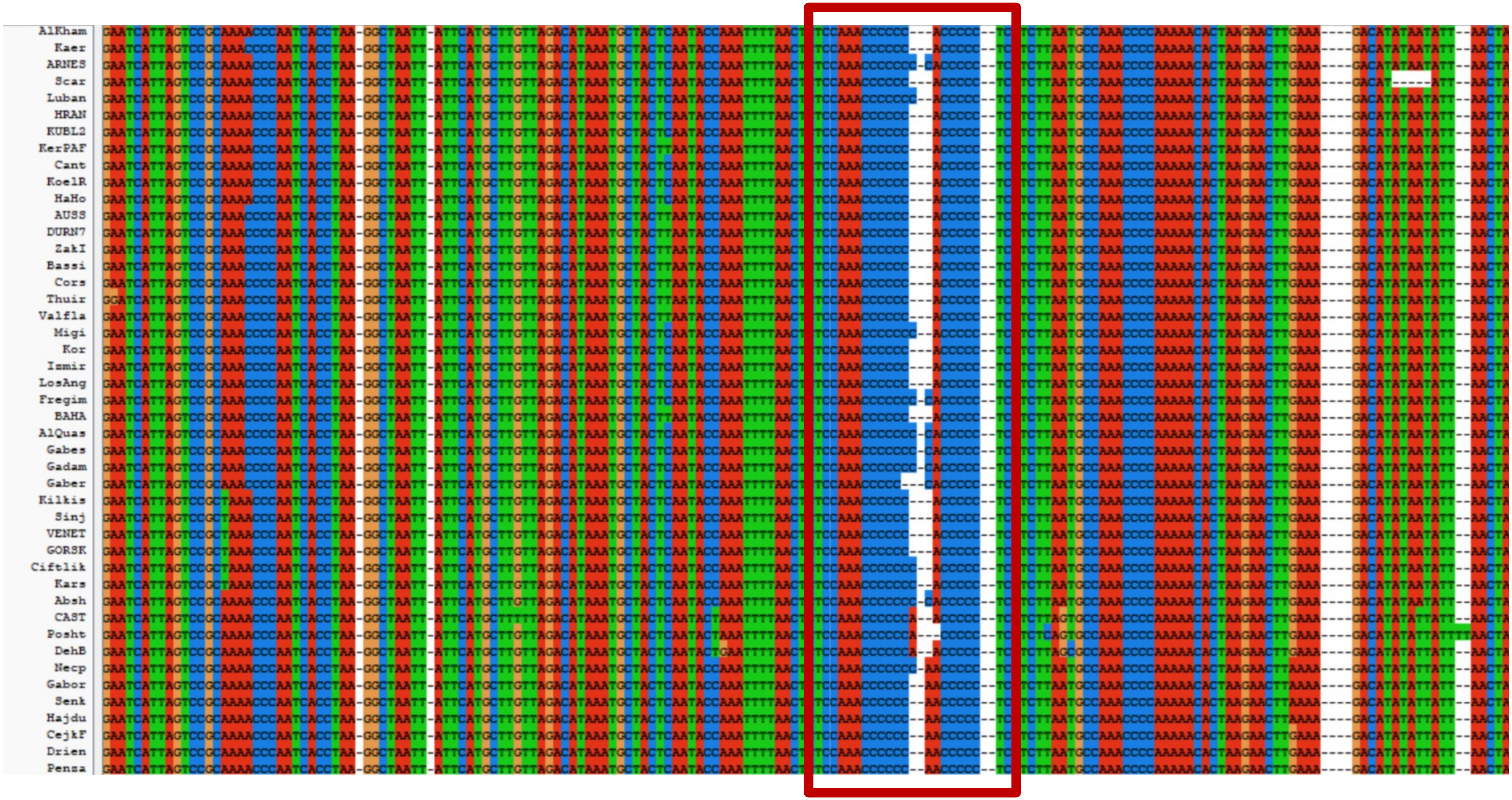
gorilla	**AGGTT**
horse	**AG-TT**
panda	**AG-TT**

penguin	**A-GTT**
chicken	**A-GTT**
ostrich	**AGGTT**

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

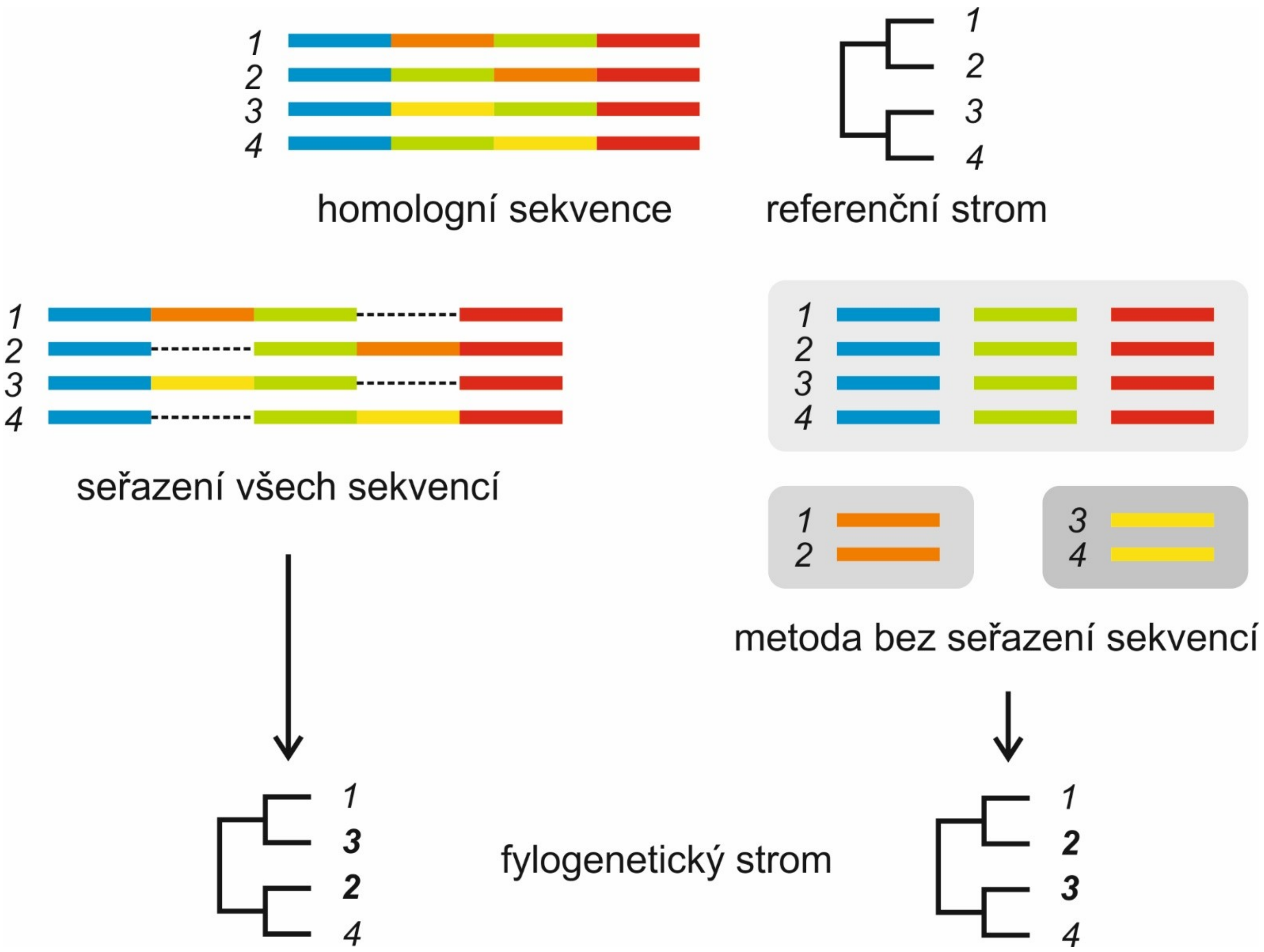
gorilla	**AGGTT**
horse	**AG-TT**
panda	**AG-TT**
penguin	**A-GTT**
chicken	**A-GTT**
ostrich	**AGGTT**

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



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

There are also methods without alignment:



Methods

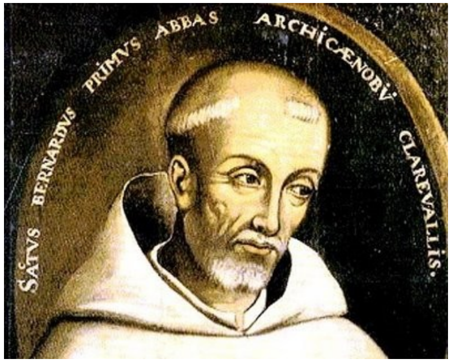
		Data types	
		distances	characters
Methods of tree construction	algorithms	UPGMA neighbor-joining	
	optimality criteria	Fitch-Margoliash minimum evolution	maximum parsimony maximum likelihood Bayesian a.

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?

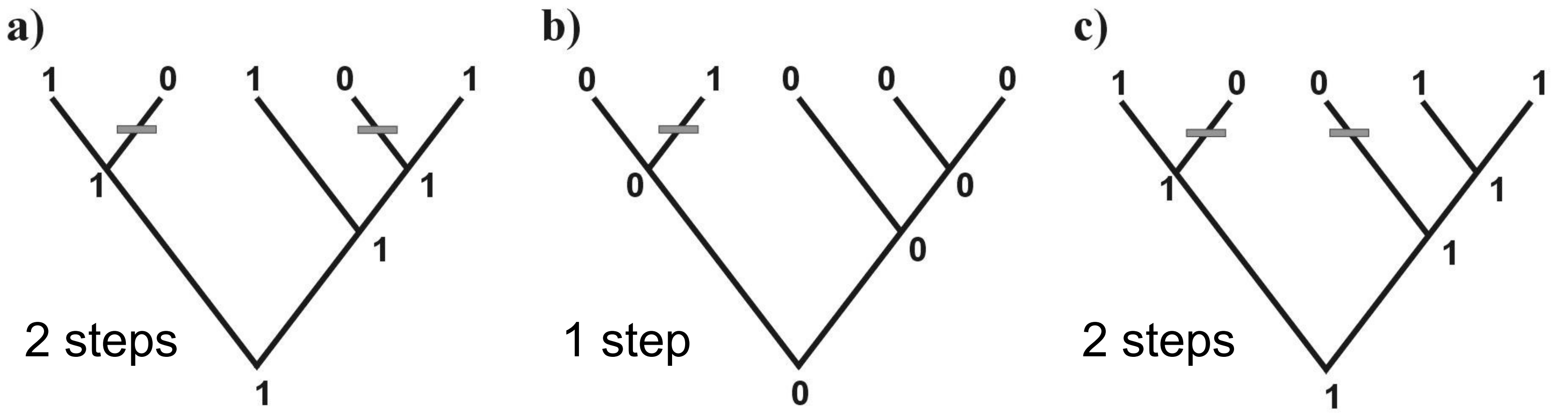
MAXIMUM PARSIMONY, MP (maximální úspornost)

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

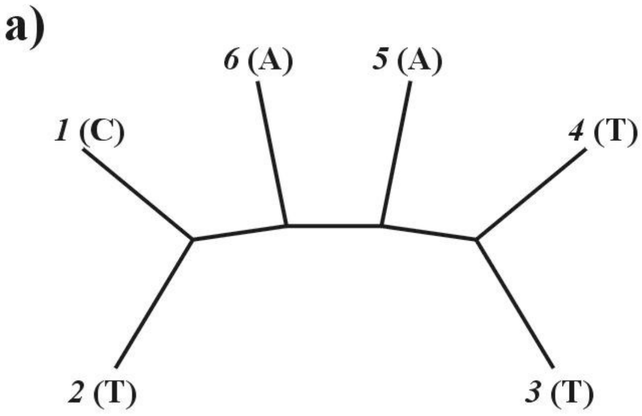


minimal number of steps = 3
real number of steps = 5
⇒ 2 extra steps → homoplasy

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

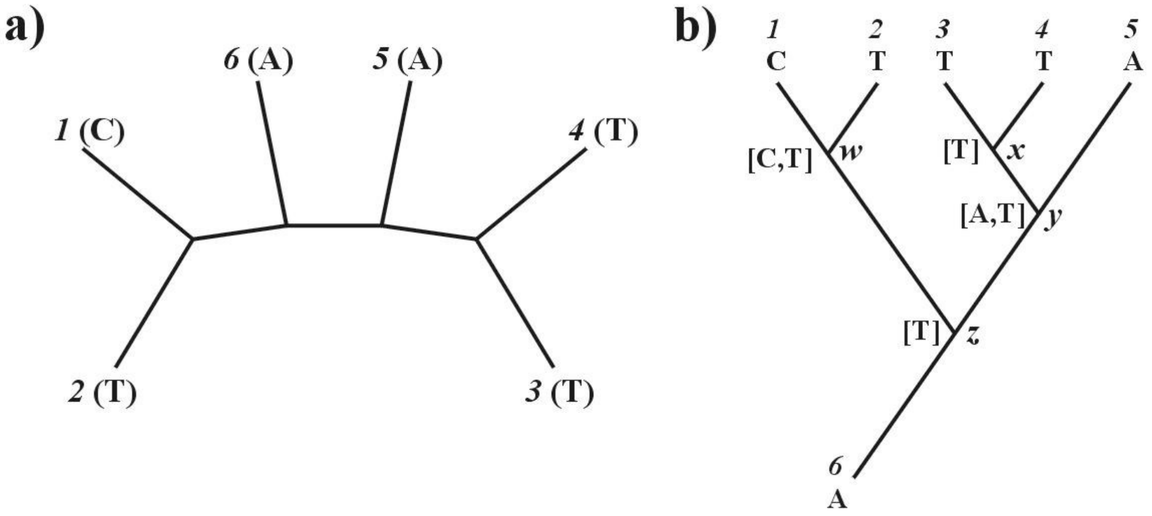


Estimation of number of steps: Fitch algorithm



1. arbitrary root

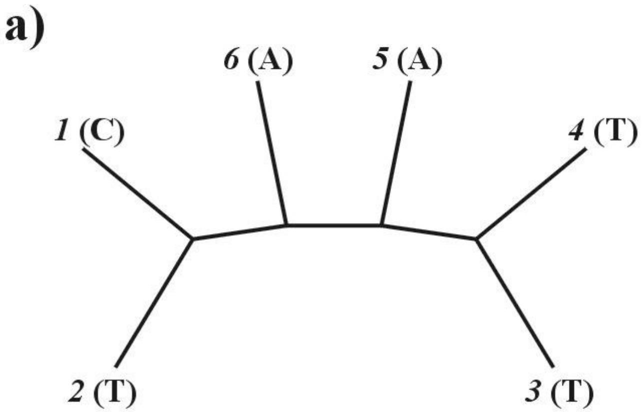
Estimation of number of steps: Fitch algorithm

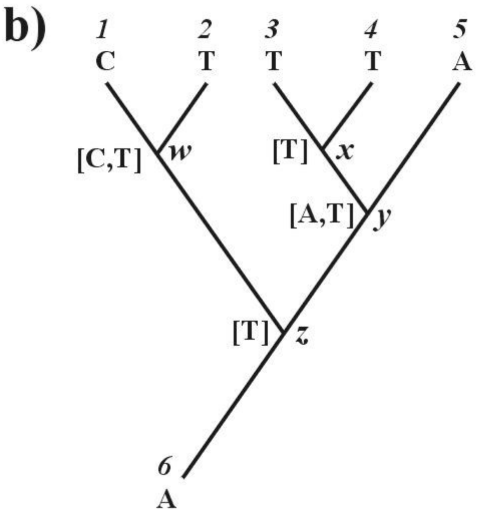


1. arbitrary root

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

Estimation of number of steps: Fitch algorithm



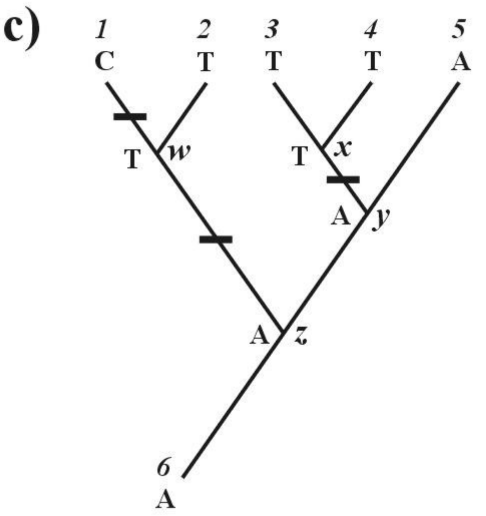


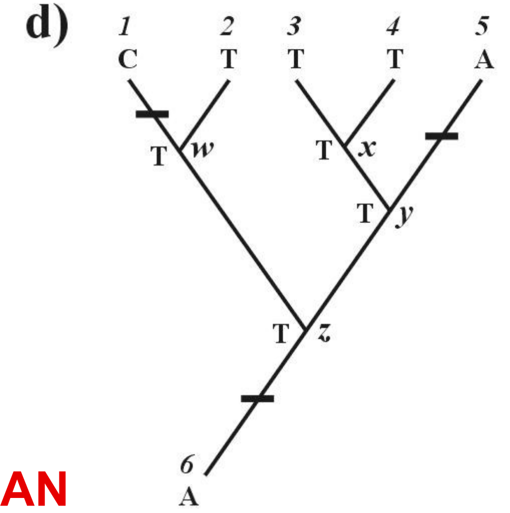
1. arbitrary root

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)

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

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

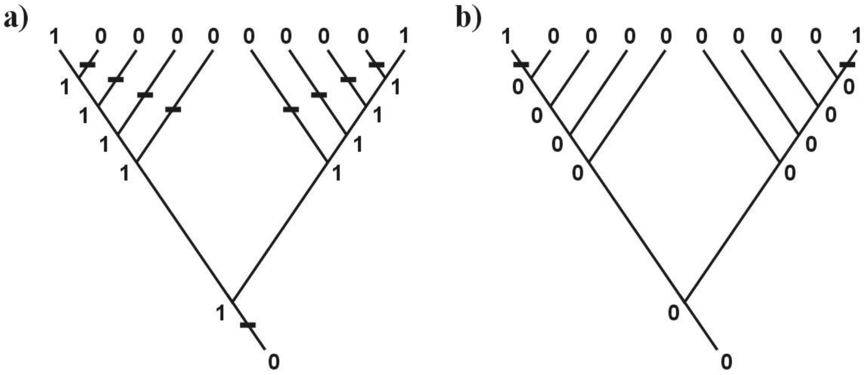
Metods of parsimony:

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

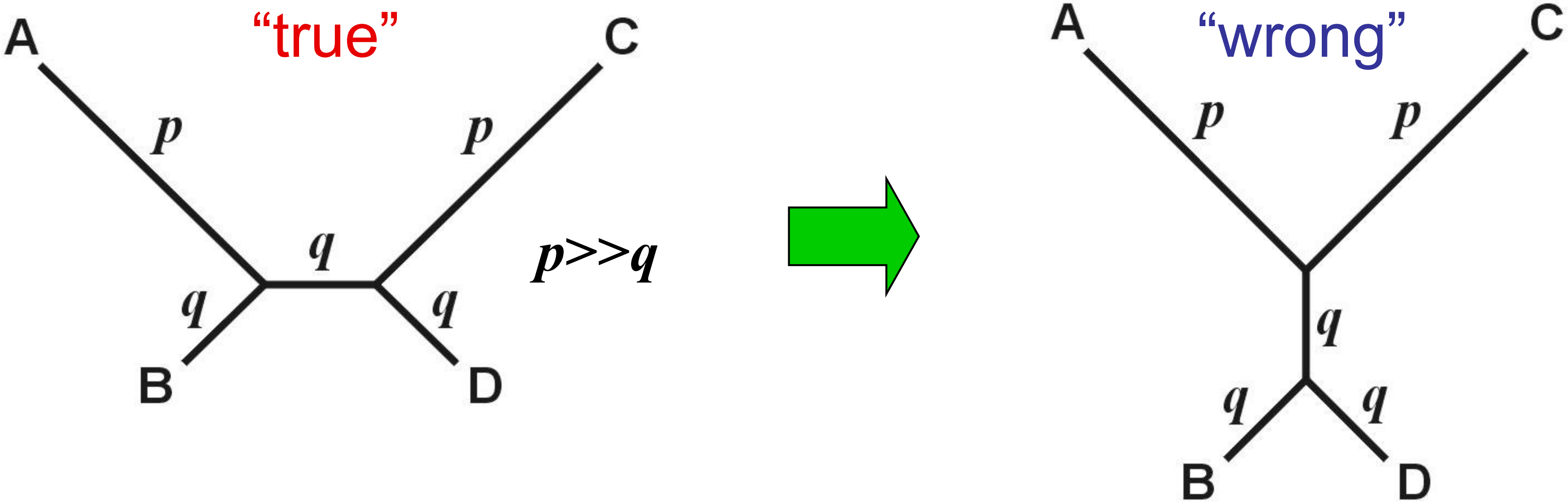
d)

	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.

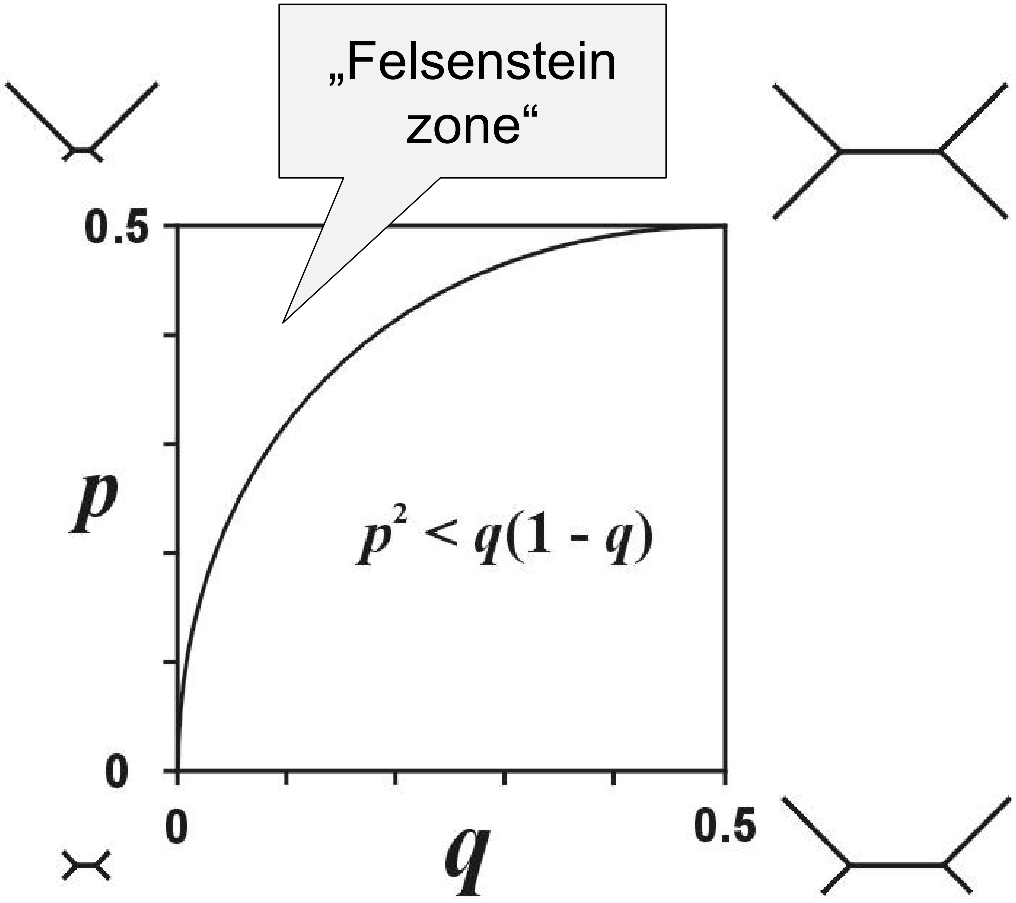
Parsimony and consistency



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

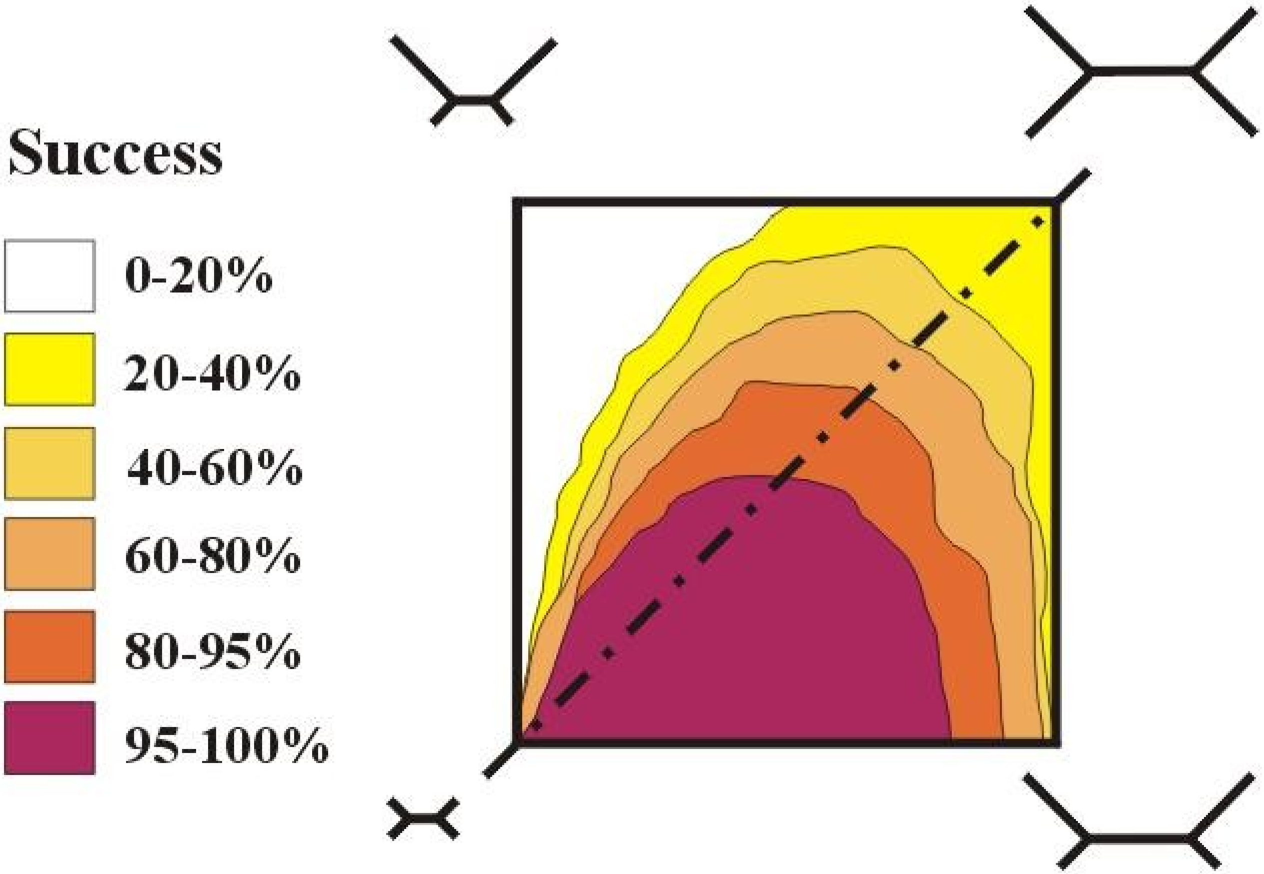
((A,C),(B,D))

Parsimony and consistency

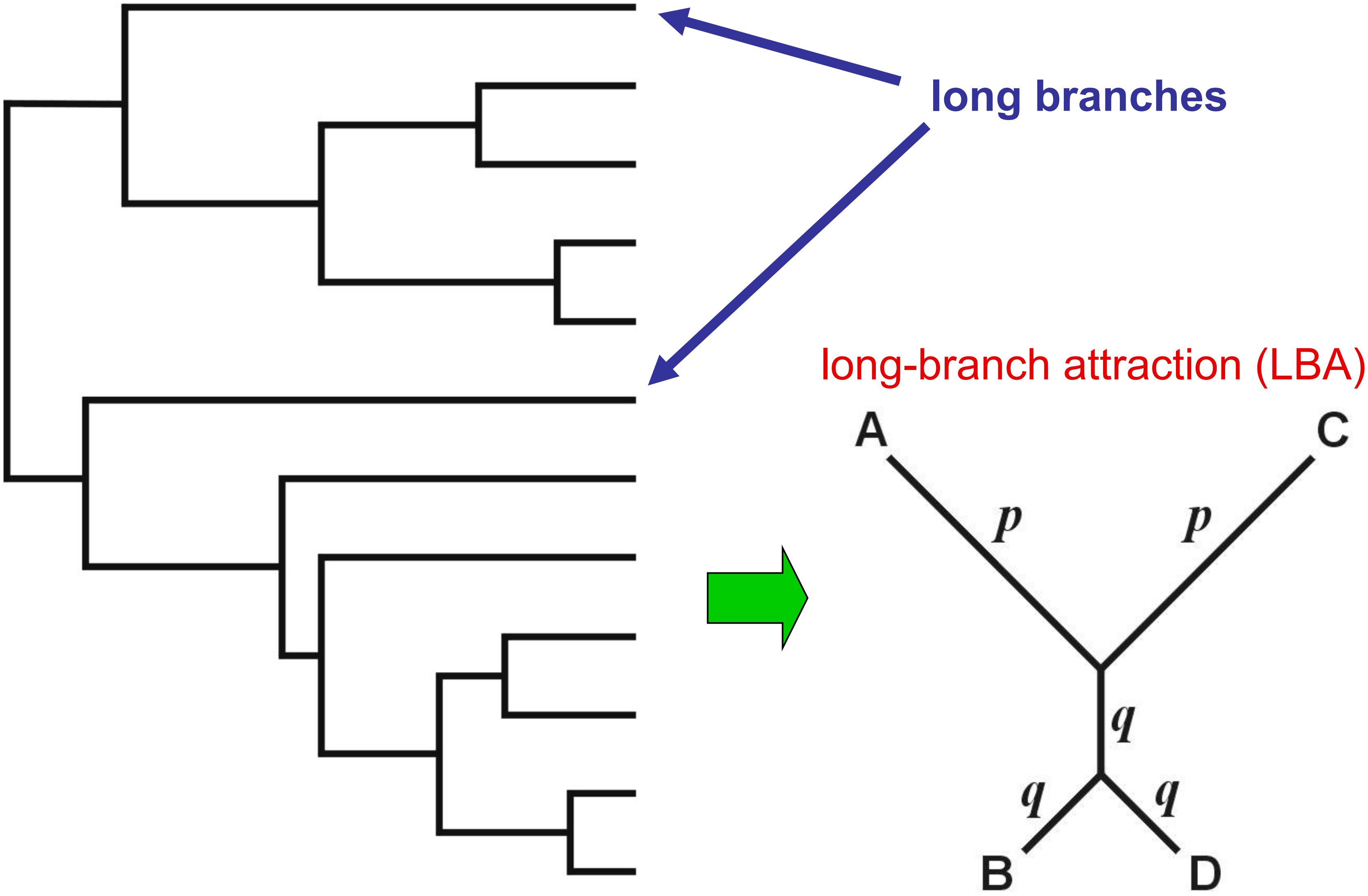

„Felsenstein zone“

In the Felsenstein zone, parsimony is inconsistent

Parsimony and consistency



Parsimony and consistency



long branches

long-branch attraction (LBA)

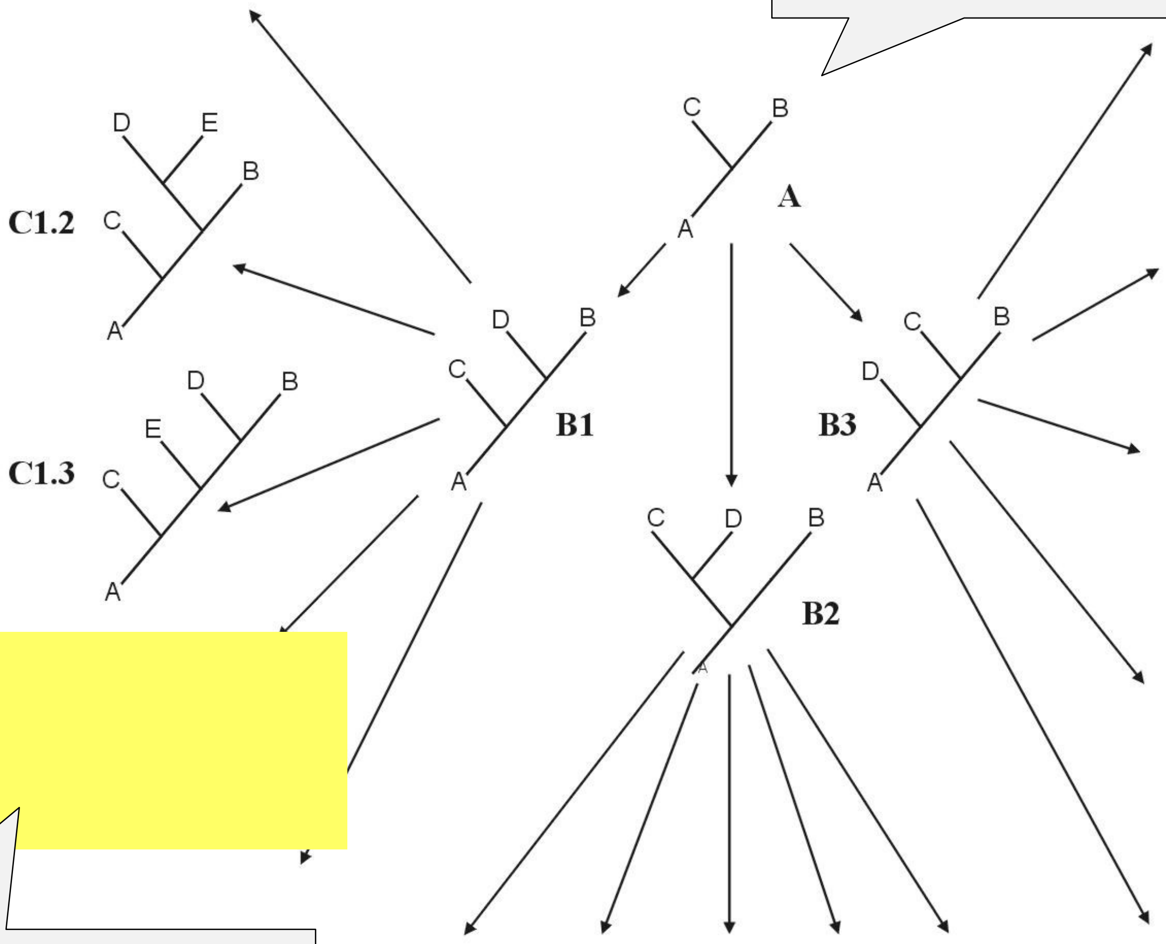
Search for optimal tree

1. Exact methods:

   a) exhaustive search

   b) branch-and-bound

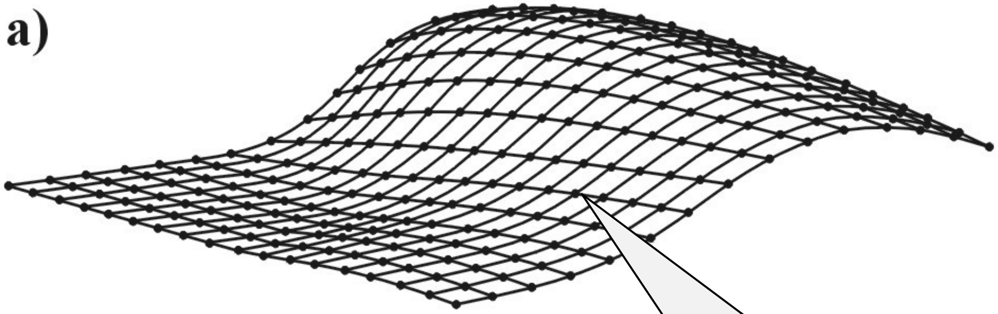
branch-and-bound



starts with 3 taxa, sequential addition

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

2. Heuristic search

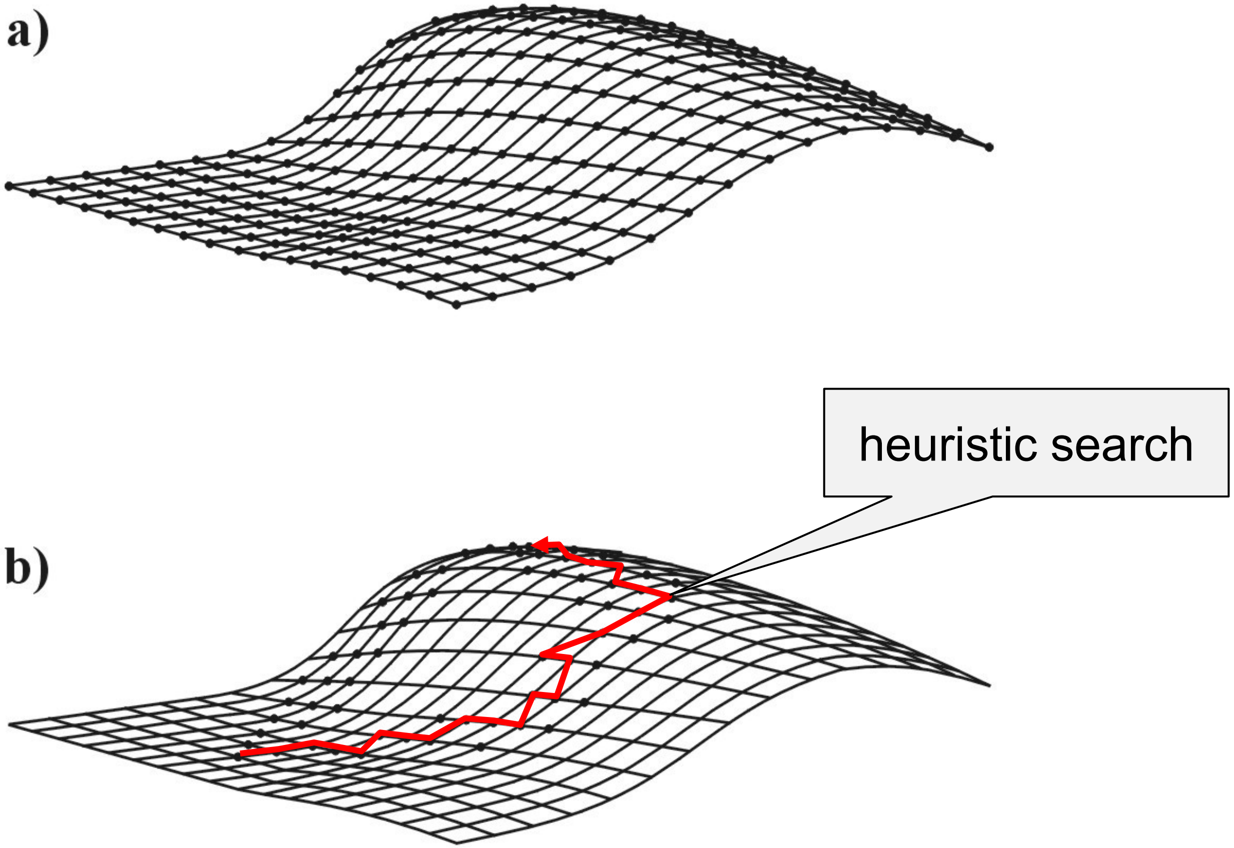


all possible trees

stepwise addition

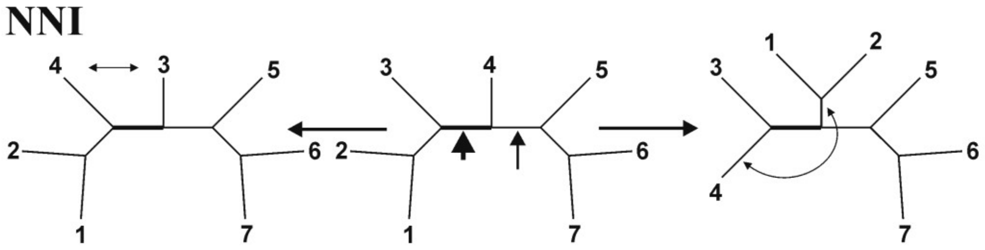
star decomposition

branch swapping

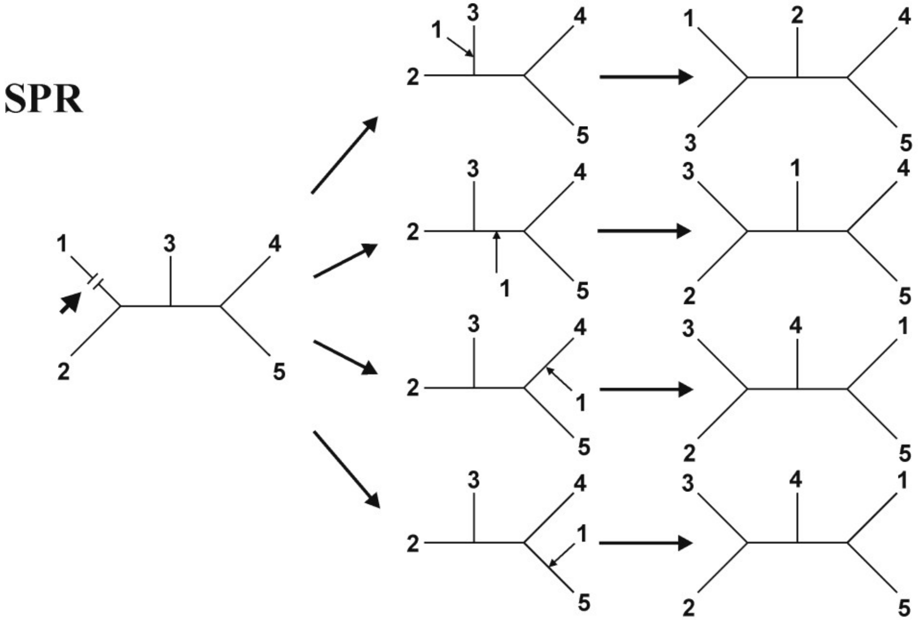


heuristic search

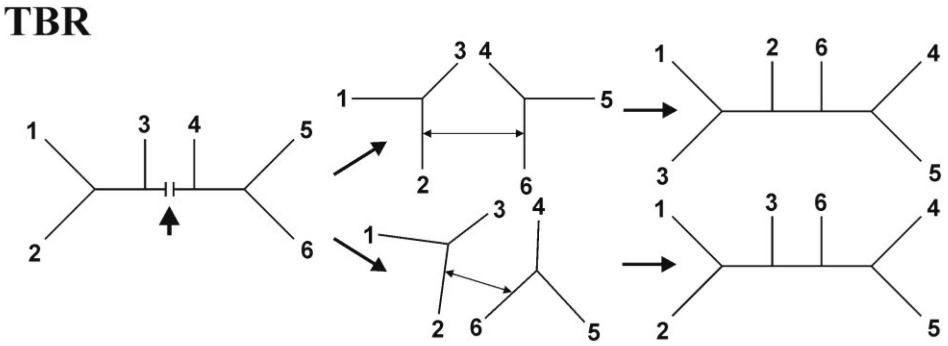
nearest-neighbor interchanges (NNI)



subtree prunning and regrafting (SPR)



tree bisection and reconnection (TBR)



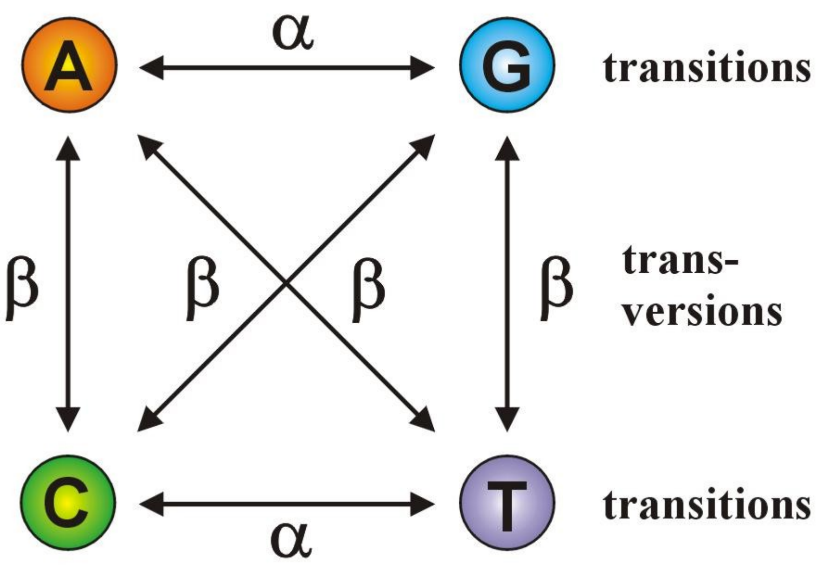
Evolutionary models and distance methods

		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

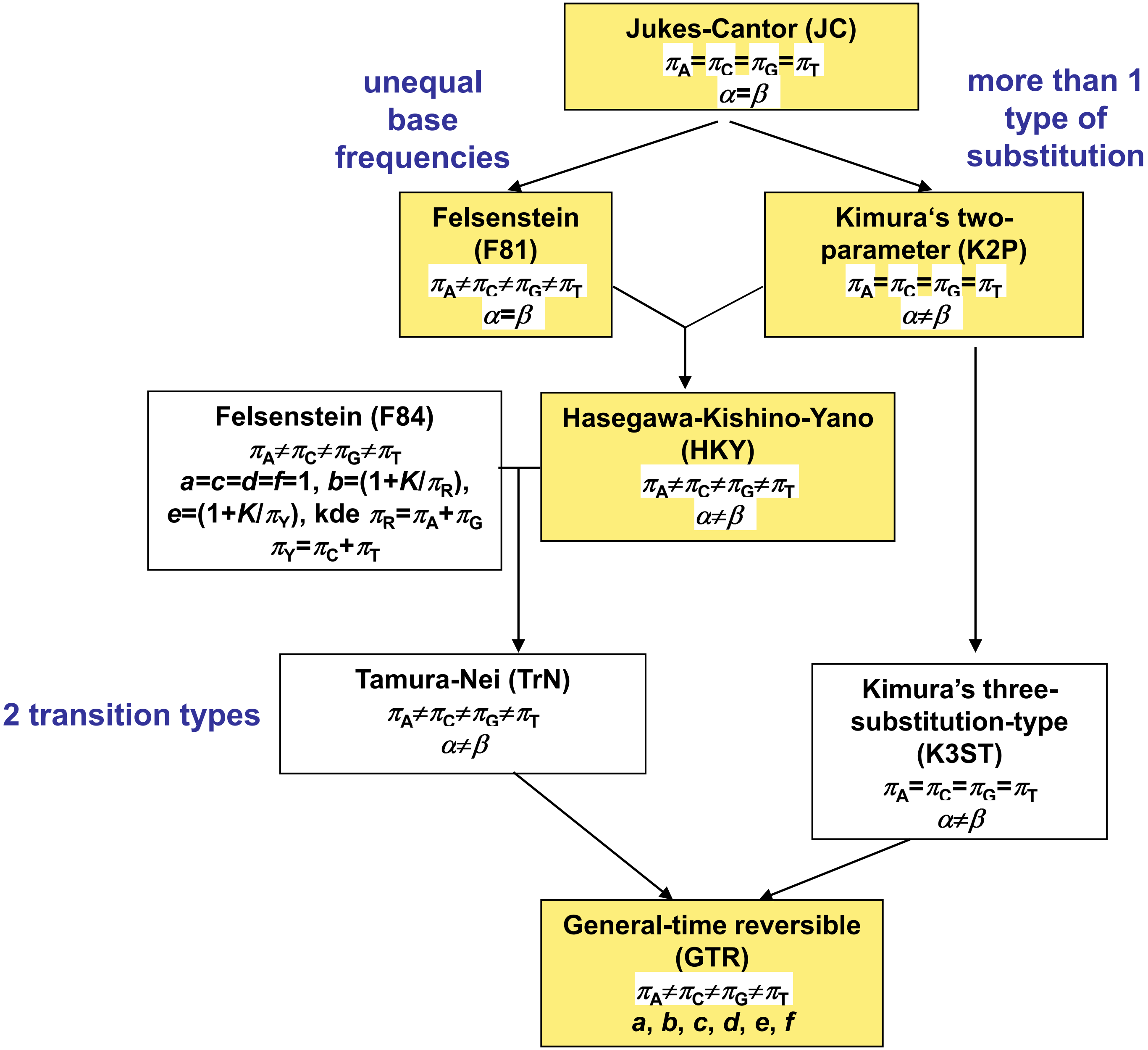
$$\mathbf{Q} = \begin{pmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{pmatrix}$$

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

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

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

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

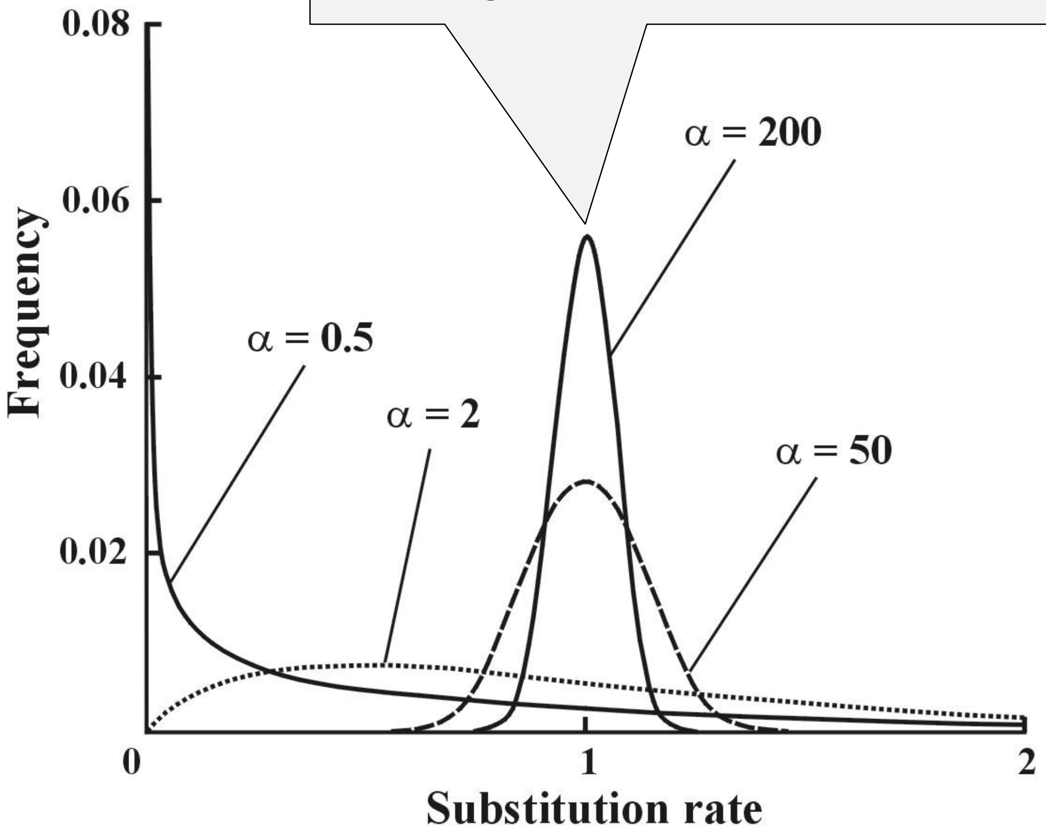
Heterogenity of substitution rates in different parts of sequences

Gamma (Γ) distribution:

shape parameter α

discrete gamma model

invariant sites
→ GTR+ Γ+I



the higher α, the more homogeneous are substitutions

Model comparison:

Likelihood ratio test (LRT):

nested models

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

χ^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

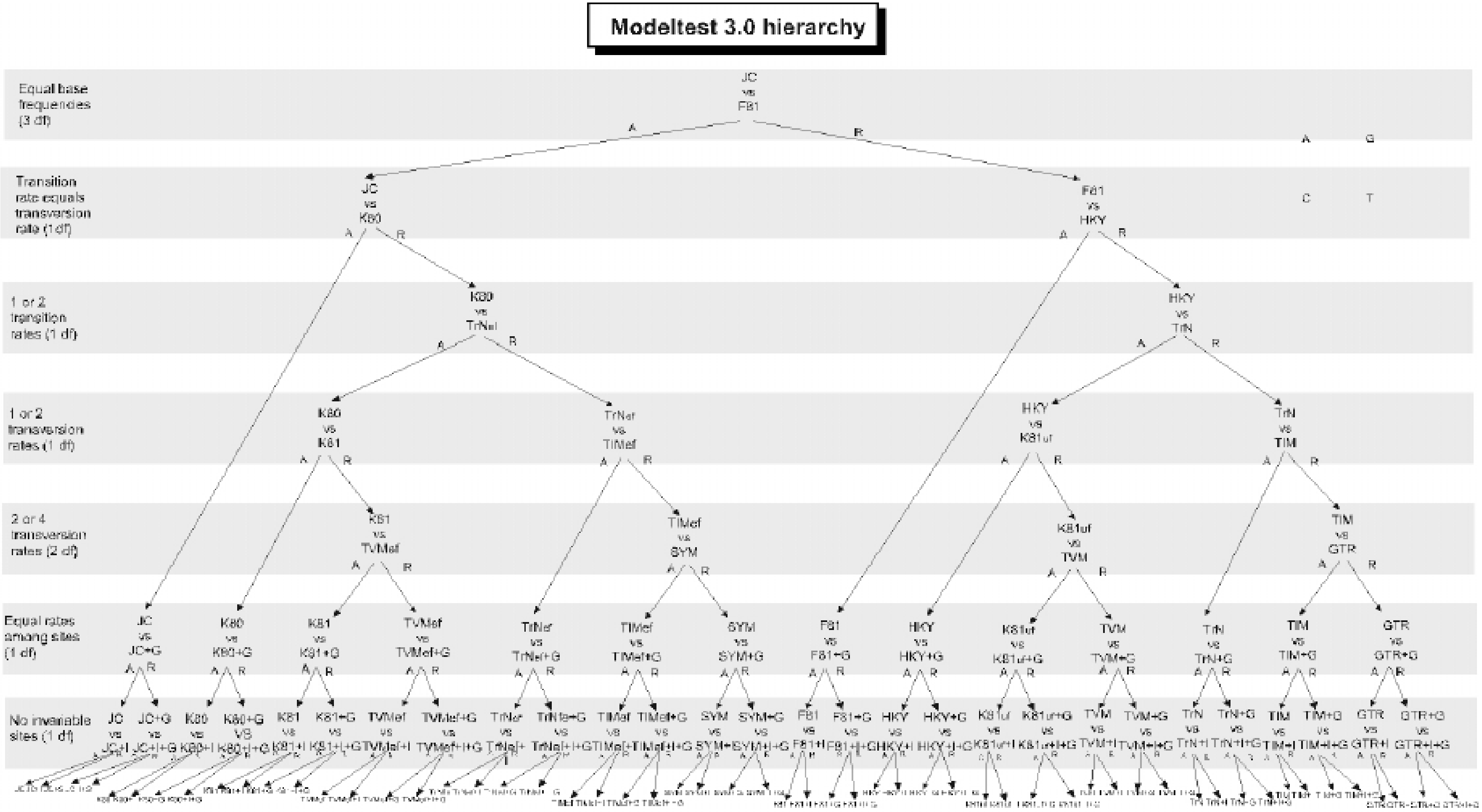
Bayesian information criterion (BIC):

nonested models

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

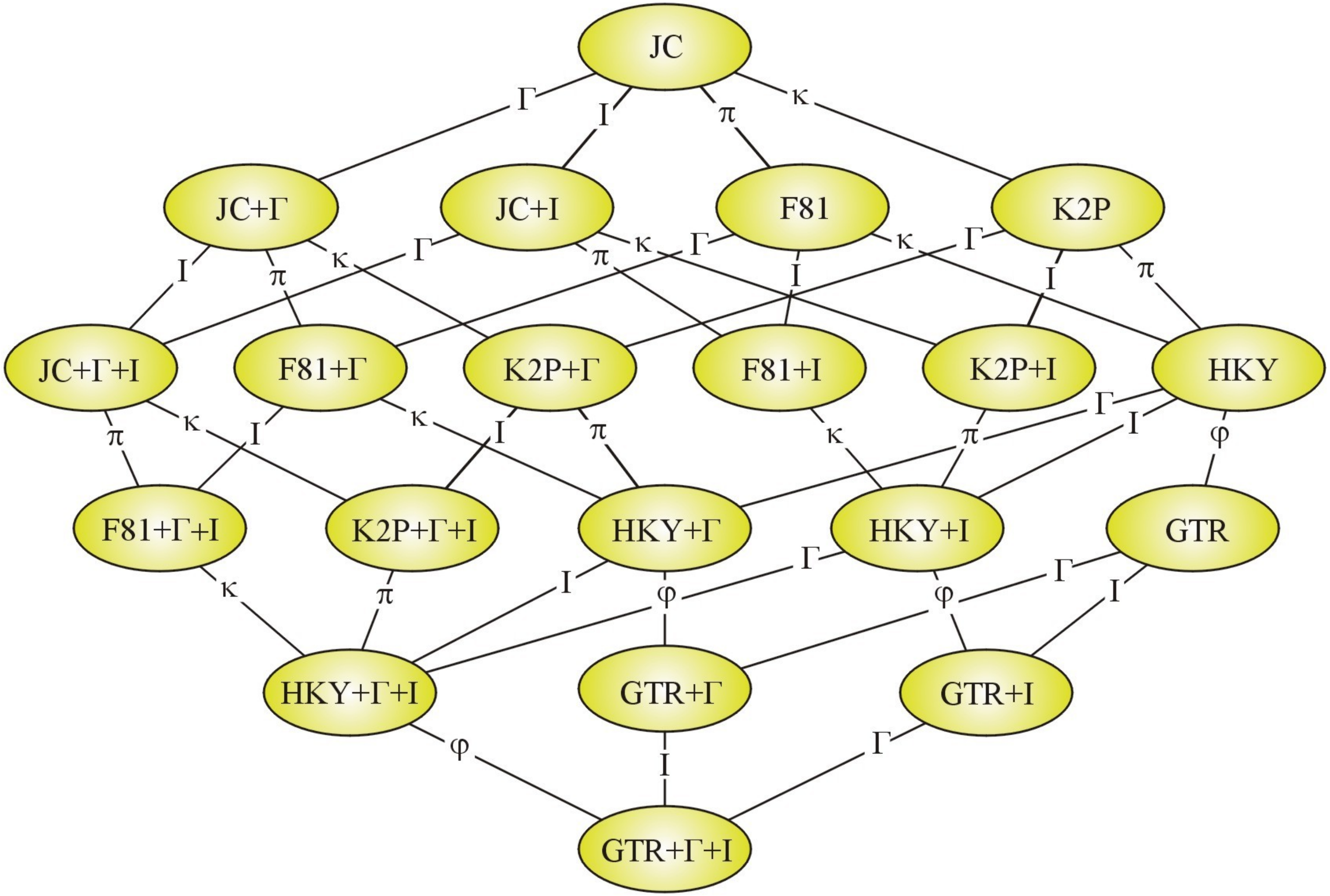
Model comparison:

hierarchical LRT – ModelTest (Crandall and Posada), jModelTest

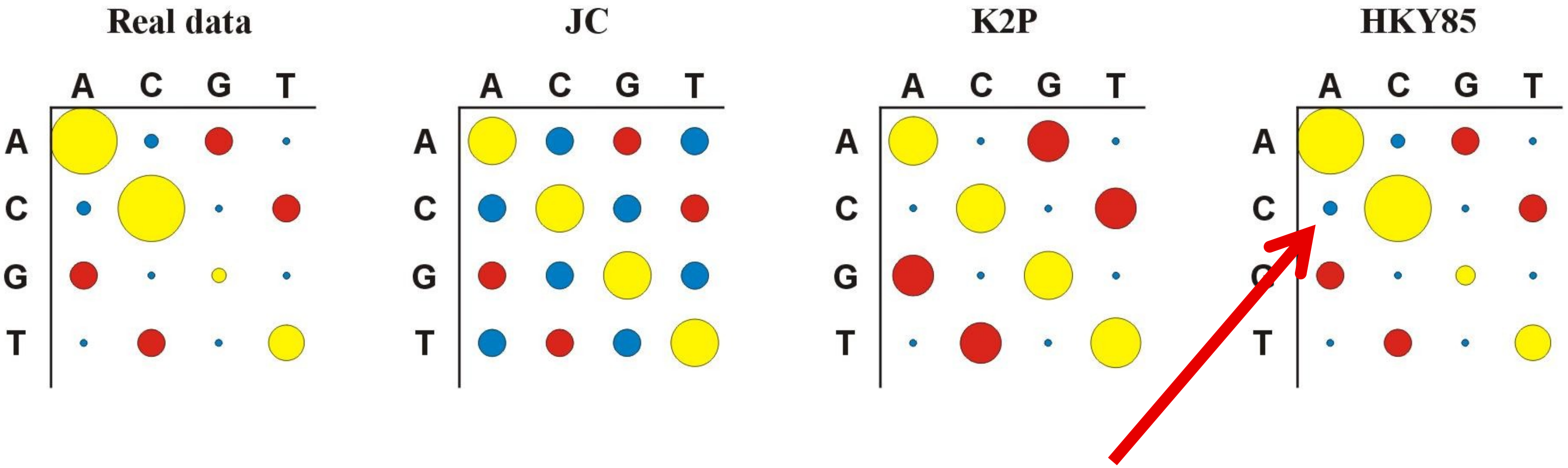


Model comparison:

dynamic LRT:



Model comparison:



error!

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

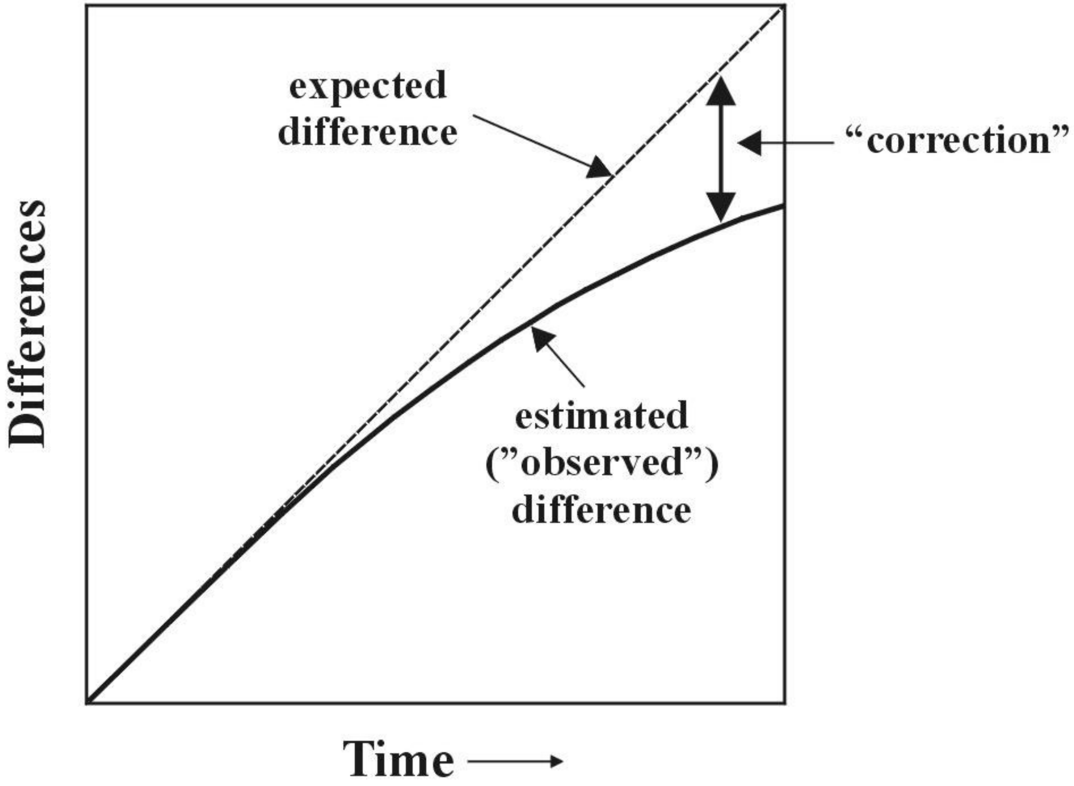
distance methods base on assumption that if we know true distances, we can very easily infer the true phylogeny

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

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

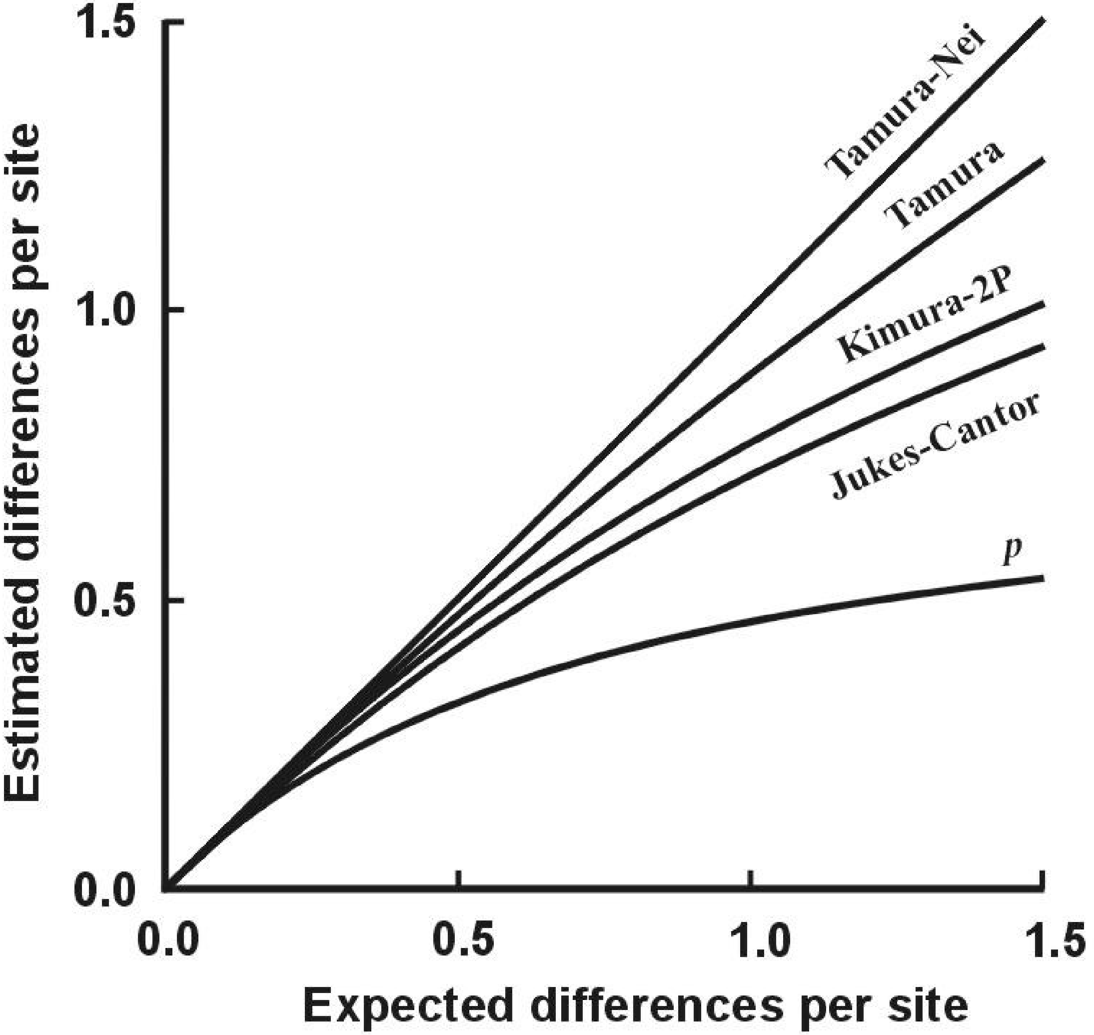
p-distance: $p = k/n = 3/30 = 0{,}10$

problem of saturation:



Distances for some models:

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


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

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	--

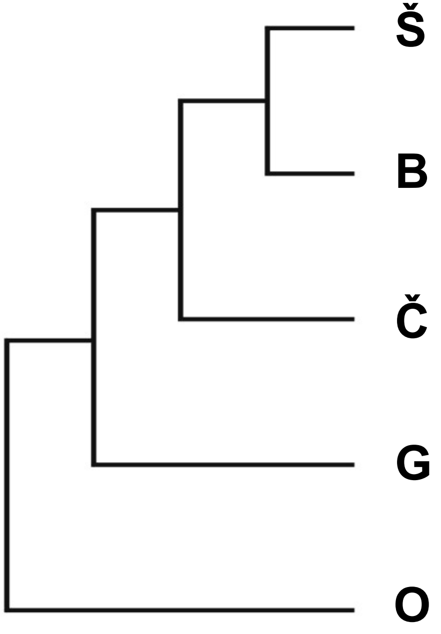
UPGMA (unweighted pair-group method using arithmetic means):

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)

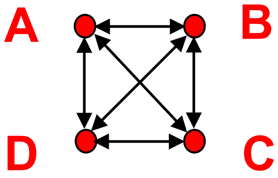
complete-linkage (m. nejvzdálenějšího souseda)

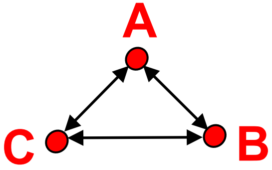


UPGMA and consistency

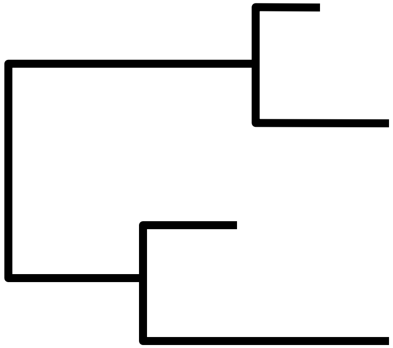
additive distances: $d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC})$

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

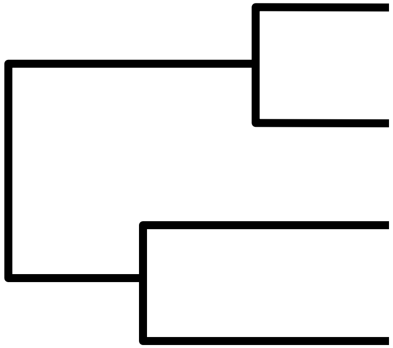




ultrametric distances: $d_{AC} \leq \max(d_{AB}, d_{BC})$

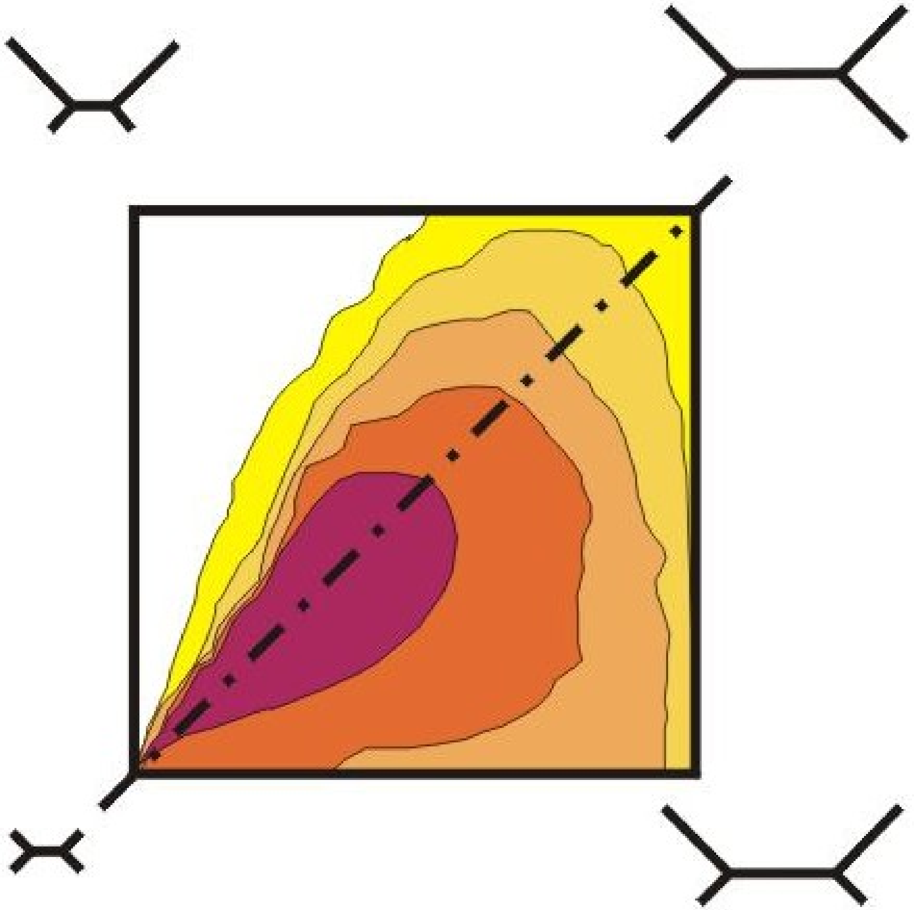


additive tree



ultrametric tree

UPGMA and consistency



Neighbor-Joining, NJ

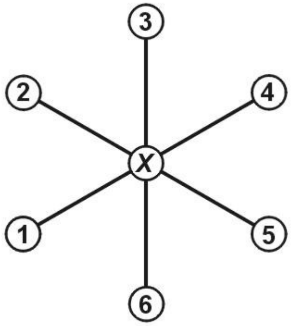
Algorithmic method

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

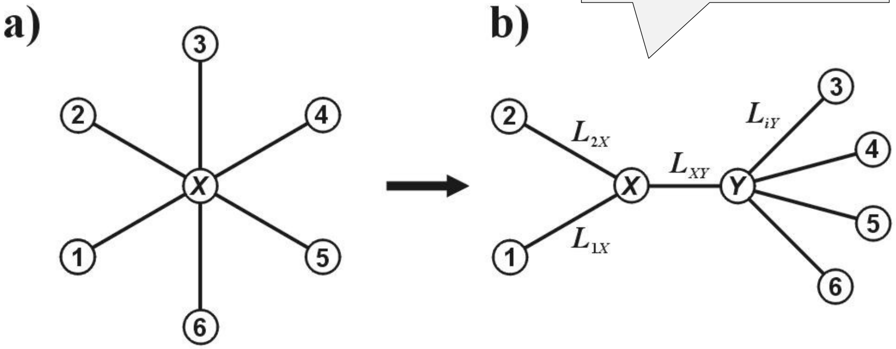
Each pair of nodes adjusted according to its divergence from others

Single additive tree

a)

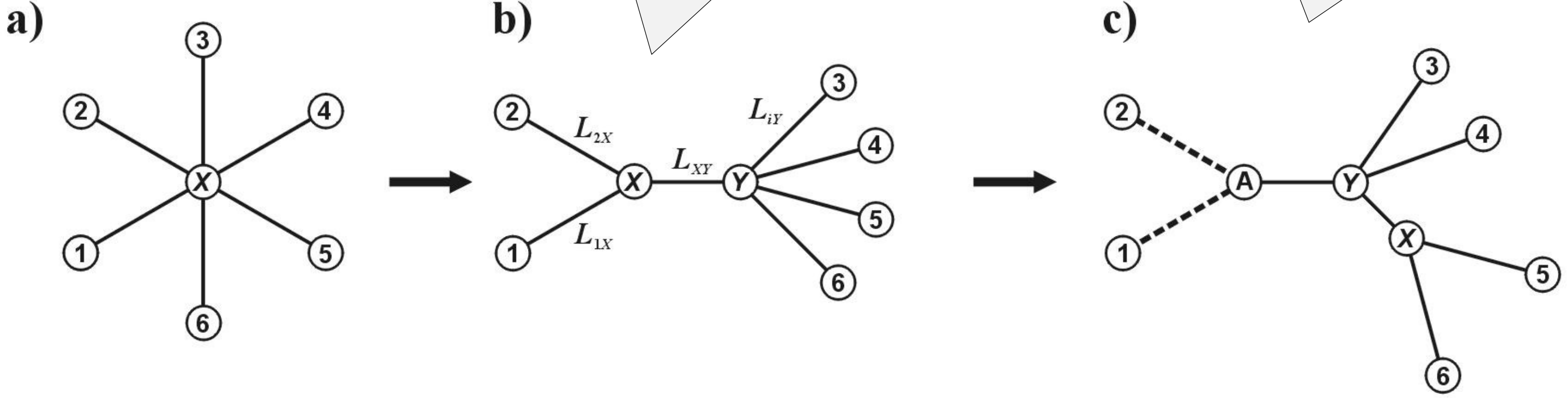


star tree



star tree

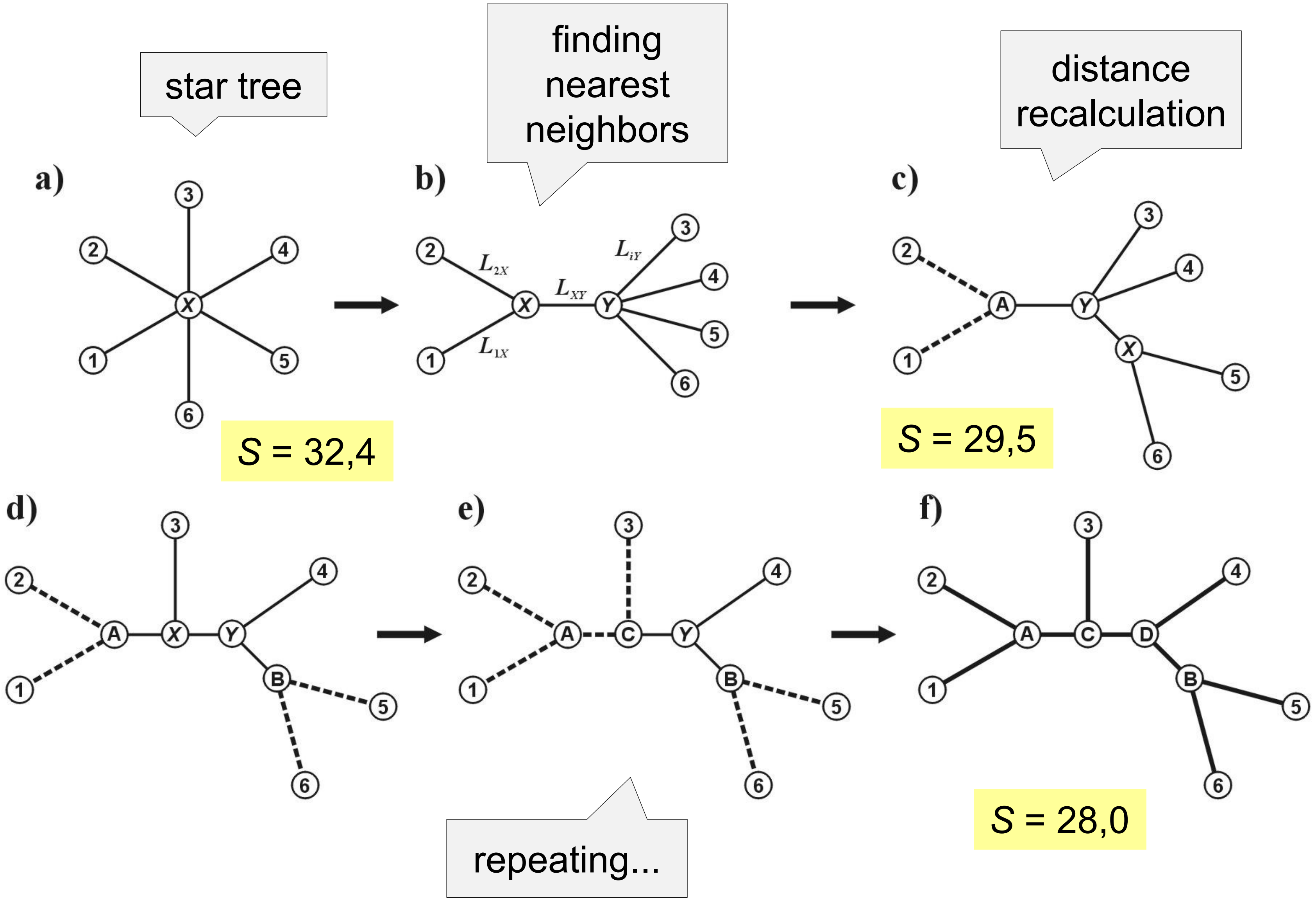
finding nearest neighbors



star tree

finding nearest neighbors

distance recalculation



star tree

finding nearest neighbors

distance recalculation

S = 32,4

S = 29,5

repeating...

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