

Genetika a taxonomie virů

MVDr. Jana Kvičerová, Ph.D.

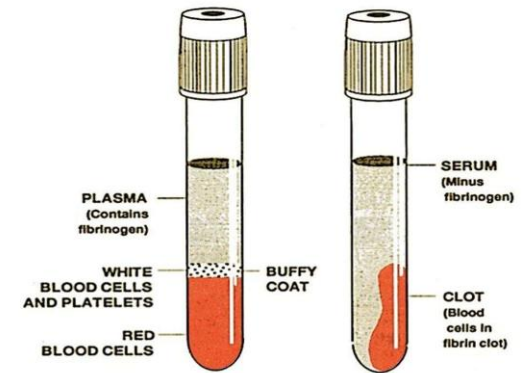
Katedra parazitologie
Přírodovědecká fakulta JU
České Budějovice

2.-3.12. 2024

Zásady odběru vzorků pro izolaci nukleových kyselin virů

Druhy vzorků (tkání):

- krev
- trus / stolice
- moč
- sliny
- stěry
- orgány (slezina, ledviny, játra, plíce, srdce, mozek)

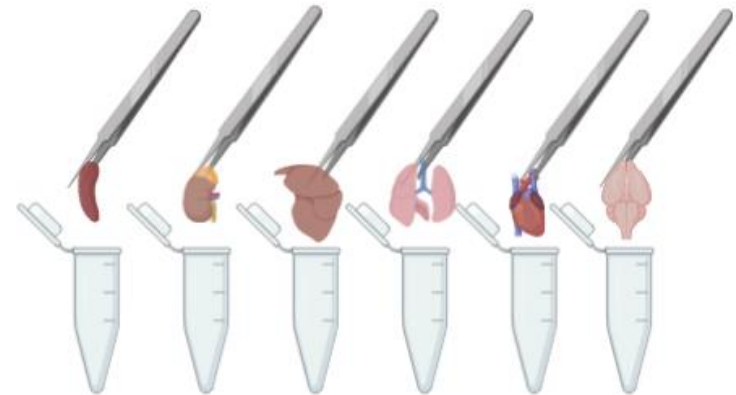


V humánní medicíně:

- **krev**
- stolice
- stěry

U zvířat:

- krev
- trus
- **orgány (postmortálně)**



Zásady odběru vzorků pro izolaci nukleových kyselin virů

Obecné zásady odběru vzorků pro izolaci NK virů:

- vzorky **co nejčerstvější**

in vivo – není problém

post mortem – co nejdříve

- **včasné zpracování** (krev na sérum – co nejdříve) či **fixace/konzervace**

- **sterilní odběr** (sterilní nástroje, sterilní zkumavky,
případně prostředí flowboxu)

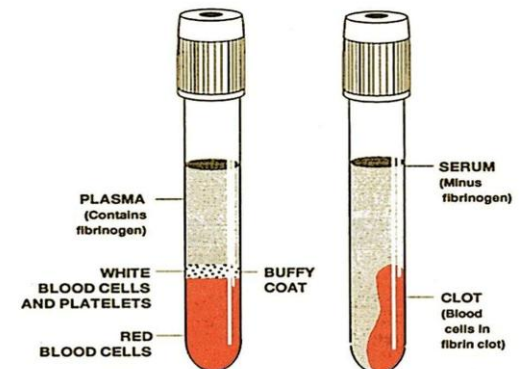
- **použití rukavic** (vždy), případně respirátoru

- **(biohazard boxy)**

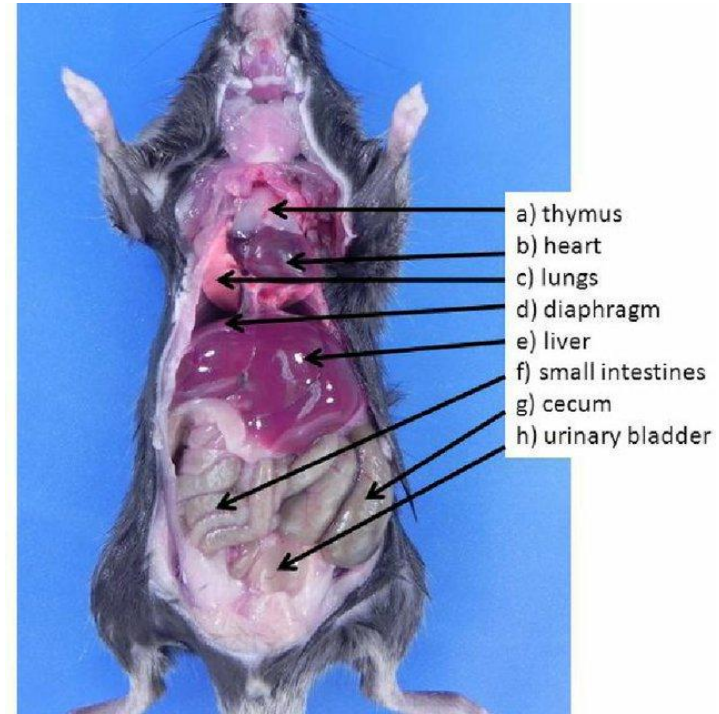
Zásady odběru vzorků pro izolaci nukleových kyselin virů

Krevní sérum:

- odběr plné žilní krve, její vysrážení a centrifugace
 - **centrifugace 5-10 min při 1500 otáčkách/min**
 - **odebrat sérum** (tj. žlutavou tekutinu vyloučenou na povrch)
 - umístění séra do sterilní mikrozkušavky a **zamrazení**
- odstranění krevních elementů, fibrinogenu, a většiny koagulačních faktorů



Orgány a zásady jejich odběru (*Mus musculus*)



Sterilně vyjmout: 1. slezina, 2. ledviny, 3. játra, 4. srdce, 5. plíce, 6. mozek

Média pro fixaci vzorků tkání pro virovou diagnostiku

- **bez fixačního média** (tzv. nasucho) – u čerstvých odběrů

výhody: finanční, časová úspora při odběru, bez nutnosti eliminace média před izolací nukleových kyselin

nevýhody: nutnost okamžitého zamrazení

- **RNAlater**

výhody: bez nutnosti okamžitého zamrazení (např. při odběrech vzorků v terénu v delší vzdálenosti od laboratoře)

nevýhody: finanční, časové, nutnost eliminace před izolací NK

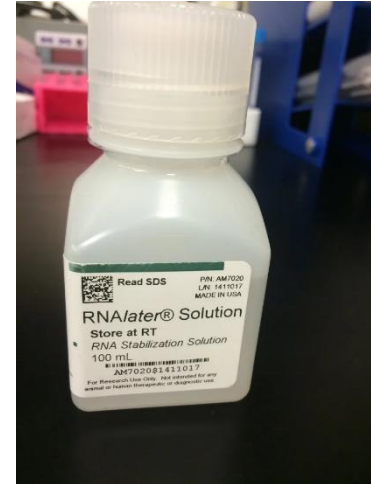
- **DNA/RNA Shield**

výhody: bez nutnosti okamžitého zamrazení (např. při odběrech vzorků v terénu v delší vzdálenosti od laboratoře),

lze uchovávat až 30 dní při pokojové teplotě (!)

nevýhody: finanční, časové, nutnost eliminace před izolací NK

RNAlater



DNA/RNA Shield



Hydraflock Swab



Screwcap Scoop



Ultra High-Density Beads



Blood



Saliva



Urine

Reagent

Swab

Stool

Tissue

DNA/RNA Shield

DNA/RNA Shield™

Catalog Nos:	R1100-50	(50 ml)
	R1100-250	(250 ml)
	R1200-25	(25 ml) 2X concentrate
	R1200-125	(125 ml) 2X concentrate
Storage:	Reagent stable -80°C to 70°C.	



Features

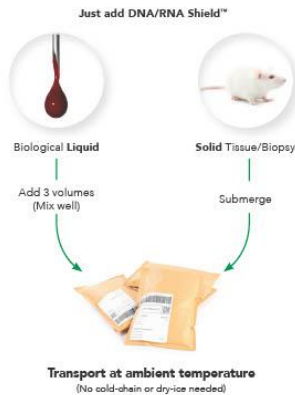
- Eliminates cold-chain. Ensures DNA/RNA stability during sample transport/storage at ambient temperatures.
- Inactivates infectious agents (*viral, bacterial, fungal and parasitic*).
- Easy sample processing. DNA/RNA can be isolated directly without sample precipitation or reagent removal (*compatible with most DNA and RNA purification kits/ high-throughput workflows*).

Sample Stability

Samples in DNA/RNA Shield™ are stable for long periods of time before subsequent purification of high-quality DNA/RNA.

Temperature	Time
-20°C and below	Indefinite
4°C – 25°C (ambient)	Minimum 30 days
35-40°C	Up to 7 days

Instructions for Sample Storage/Transport



Unsure of sample type? Just add 9 volumes of DNA/RNA Shield™. If sample is viscous, add more DNA/RNA Shield™.

Suggested Volumes

DNA/RNA Shield™	300 µl	600 µl
Cell pellets	10 ⁶	10 ⁷
Tissue	30 mg	60 mg
Biological liquid	100 µl	200 µl

DNA/RNA Shield™ (2X concentrate):
(1) Use only for immediate sample processing/purification.
(2) For sample storage, dilute with nuclease-free water (1:1) prior to use.

DNA/RNA Purification

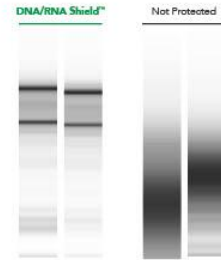
- Samples in DNA/RNA Shield™ are directly compatible with:
- All Zymo Research Purification Kits.
 - Most kits and workflows from Qiagen, Roche, Thermo-Fisher, Macherey Nagel, etc.
 - High-throughput automated workflows from Hamilton, Tecan, bioMérieux, PerkinElmer, Eppendorf, Promega, etc.

Recommended Purification Kits for DNA/RNA Shield™:

Product / Format	Size	Cat. No.
Quick-DNA/RNA™ Miniprep Plus Spin Column	80 preps	D7003
Quick-DNA/RNA™ MagBead Magnetic Bead	96 preps 384 preps	R2130 R2131

Protection from freeze-thaw damage

Are your samples protected from the stress caused by freeze-thaw cycling? DNA/RNA Shield™ is not only beneficial for sample transport, but also for the long-term storage of biological samples. DNA/RNA Shield™ protects DNA/RNA from multiple freeze-thaw cycles, even in the most complex of samples (i.e., whole blood).



High quality RNA from blood stored in DNA/RNA Shield™ that was freeze-thawed from -50°C to room temperature.

Pathogen Inactivation

DNA/RNA Shield abides by Center for Disease Control's (CDC) guidelines for pathogen inactivation.

Validated organisms by various research groups:

Bacteria	Viruses	Yeast & Eukaryotes
<i>B. subtilis</i>	Parvovirus	<i>C. albicans</i>
<i>E. faecalis</i>	Chikungunya Virus	<i>C. neoformans</i>
<i>E. coli</i>	Dengue Virus	<i>S. cerevisiae</i>
<i>L. fermentum</i>	Ebola virus	<i>P. malariae</i>
<i>L. monocytogenes</i>	Herpes Simplex Virus-1	
<i>M. tuberculosis</i>	Herpes Simplex Virus-2	
<i>P. aeruginosa</i>	Influenza A	
<i>S. enterica</i>	Rhinovirus	
<i>S. aureus</i>	MERS-coronavirus	
<i>S. pneumoniae</i>	West Nile Virus	
<i>X. fastidiosus</i>		

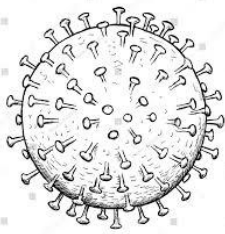


DNA/RNA Shield™ Collection Devices

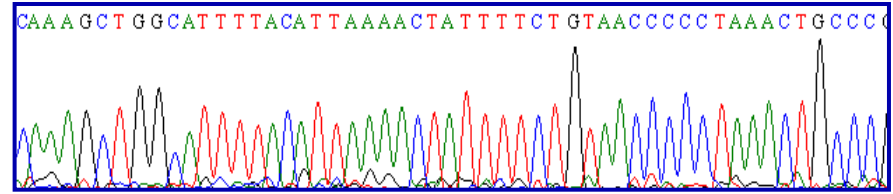


Product	Description	Recommended processing
Fecal collection tube R1101	20 x 76 mm container pre-filled with 9 ml DNA/RNA Shield™ for the direct collection of up to 1 gram or 1 ml stool. Collection spoon is included in the container screwcap.	ZymoBIOICS™ DNA/RNA D4301 (spin-column); D4302 (magbead)
Swab & collection tube R1106, R1107, R1108, R1109	12 x 80 mm self-centering screwcap container filled with DNA/RNA Shield™ (1 ml, 2 ml) and a sterile swab for specimen collection.	
Saliva collection kit R1210	Saliva collection tube with funnel and tube filled with 2 ml of DNA/RNA Shield™ for the direct collection of 2 ml saliva.	Quick-DNA/RNA™ Plus D7001 (spin-column); R2130 (magbead)
Collection & lysis tube R1102, R1103, R1104, R1105	2 ml screw-cap tube with DNA/RNA Shield™ and Bashing Beads (lysis tubes: microbe, tissue, pathogen) for the collection and homogenization of tough-to-lyse samples.	
Blood collection tube R1150	16 x 100 mm evacuated blood tube filled with 6 ml of DNA/RNA Shield™ for the direct collection of 3 ml whole-blood (human or animal).	Quick-DNA/RNA™ Blood Tube R1151 (spin-column); R2130 (magbead)

For bulk reagent, large volume, and custom device requests, please contact Zymo Research directly.

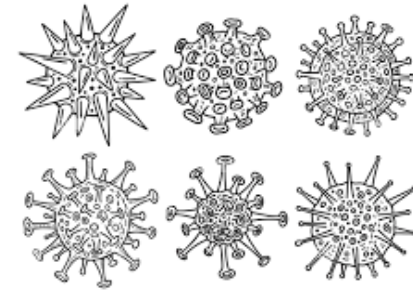
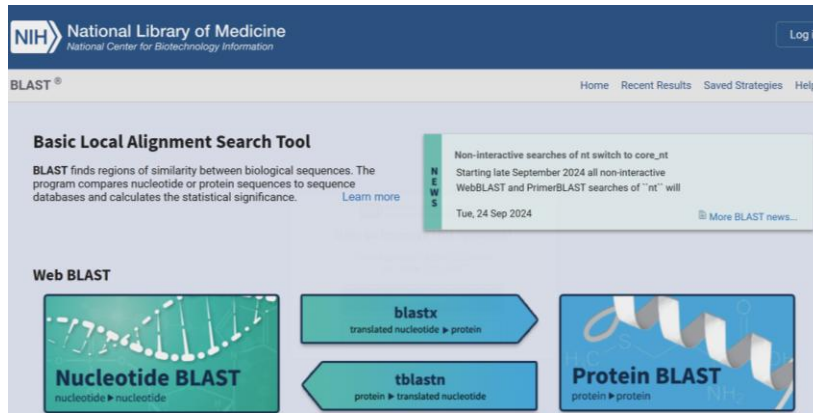


Izolace DNA / RNA, PCR, sekvenování



Organismus, který chci fylogeneticky charakterizovat (měl bych vědět proč)

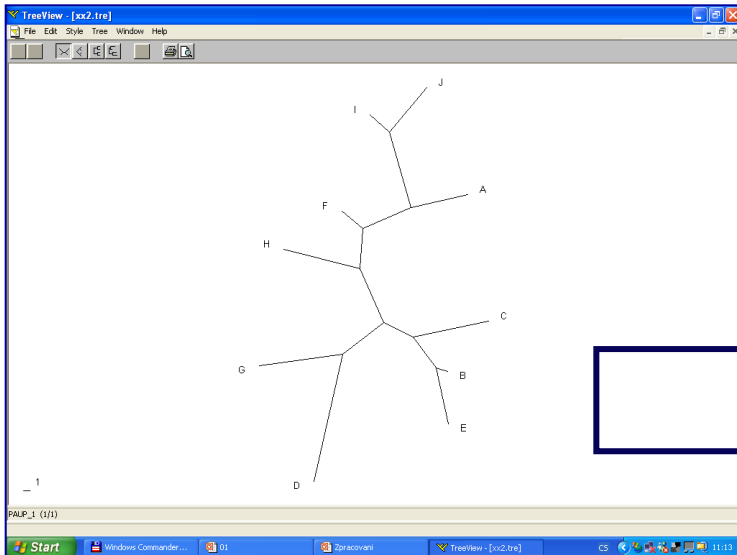
Vyhledání homologických sekvencí pro další taxony (měl bych vědět, pro které)



Vytvoření alignmentu

```
GTGGGAAGGAAA- - - CCTGGTGGTTAATA- - - CCCA
GTCGTGAGGAAGG- - - TGGTGTGTTAATAGCAGCA
BTAGCGAGGAAGG- - - CATTAGTTTAATAGACTAG
GCGGGGAGGAAGG- - - CGTGAGAGCGAATACCTTTC
GTAGGGAGGAAGGC- - - AA- TATCCTTAATACGGTTA
```

Fylogenetická analýza zvolenou metodou



To hlavní a podstatné:
interpretace fylogenetických vztahů

Fylogenetická analýza

vlastní sekvence 1
vlastní sekvence 2
vlastní sekvence 3

outgroup 1
outgroup 2
vlastní sekvence 1
vlastní sekvence 2
vlastní sekvence 3
ingroup sekvence A
ingroup sekvence B
ingroup sekvence C
ingroup sekvence D

fylogenetická analýza matice

1. nalezení
vhodných
sekvencí
(databáze)
a doplnění
matice

2. příprava
matice:
alignment


prohledávání databáze

multiple alignment

identifikace příbuzenských
vztahů mezi sekvencemi

Databáze GenBank

www.ncbi.nlm.nih.gov



National Library of Medicine
National Center for Biotechnology Information

Log in

All Databases Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation


Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)


Submit

Deposit data or manuscripts into NCBI databases




Download

Transfer NCBI data to your computer




Learn

Find help documents, attend a class or watch a tutorial




Develop

Use NCBI APIs and code libraries to build applications




Analyze

Identify an NCBI tool for your data analysis task



Research

Explore NCBI research and collaborative projects



Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Explore 3D Molecular Structures with iCn3D

21 Nov 2024

Do you want to analyze three-dimensional structures and highlight

Try Out a Development Version of NCBI's Publicly Available Annotation Tool, EGAPx

Databáze GenBank – PubMed

→ vyhledávání publikací

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

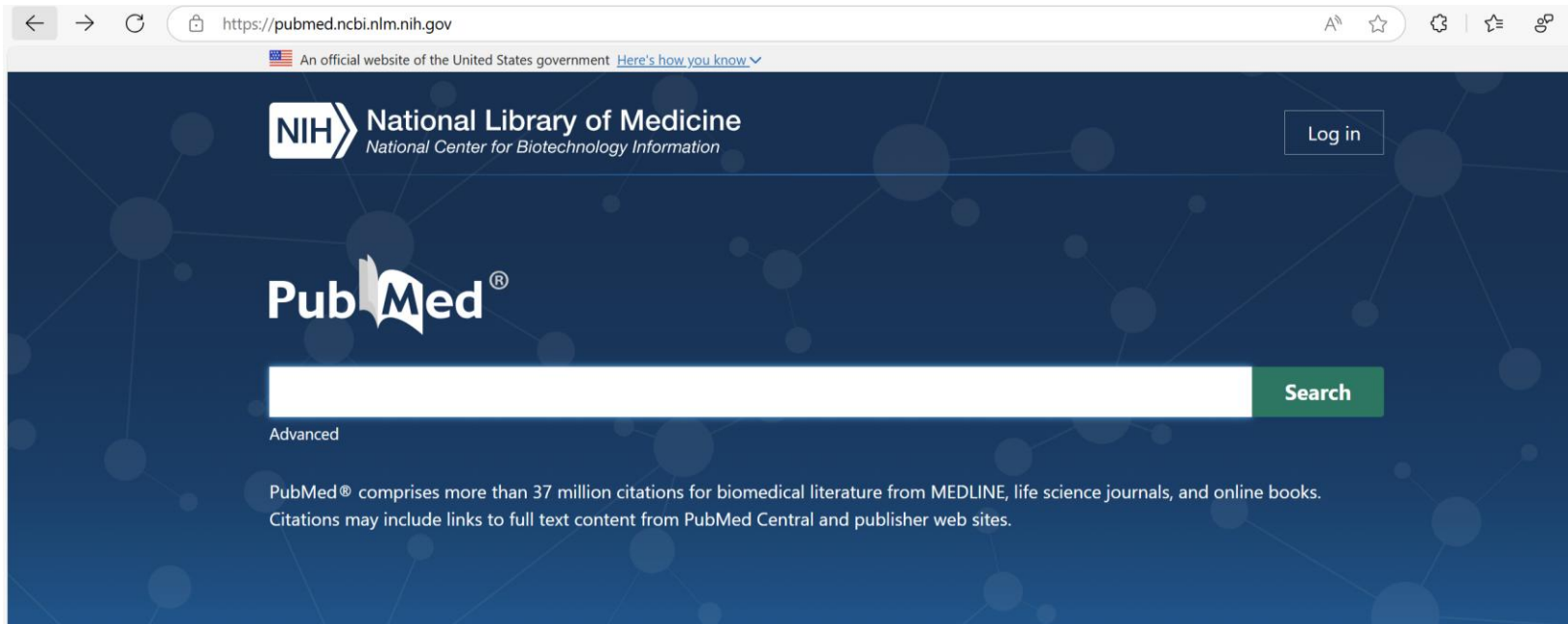
Research
Explore NCBI research and collaborative projects

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog
Explore 3D Molecular Structures with iCn3D
21 Nov 2024
Do you want to analyze three-dimensional structures and highlight
Try Out a Development Version of NCBI's Publicly Available Annotation Tool, EGAPx

Databáze GenBank – PubMed

→ vyhledávání publikací



The image shows a screenshot of the PubMed website homepage. The browser address bar displays the URL <https://pubmed.ncbi.nlm.nih.gov>. The page features the NIH logo and the text "National Library of Medicine National Center for Biotechnology Information" in the top left. A "Log in" button is located in the top right. The main heading is "PubMed®". Below it is a search bar with a green "Search" button. The word "Advanced" is positioned below the search bar. A descriptive paragraph states: "PubMed® comprises more than 37 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites." The background is dark blue with a network diagram pattern.

Databáze GenBank – Submit (BankIt, Sequin, SRA...)

→ vkládání sekvencí (po přihlášení heslem a uživatelským jménem)

The screenshot shows the NCBI National Library of Medicine homepage. At the top left is the NIH logo and the text 'National Library of Medicine National Center for Biotechnology Information'. A 'Log in' button is in the top right. Below the header is a search bar with a dropdown menu set to 'All Databases' and a 'Search' button. On the left is a navigation menu with items like 'NCBI Home', 'Resource List (A-Z)', 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The main content area is titled 'Welcome to NCBI' and contains a paragraph about the center's mission. Below this are six main sections: 'Submit' (with a red circle and arrow), 'Download', 'Learn', 'Develop', 'Analyze', and 'Research'. Each section has an icon and a brief description. On the right side, there are 'Popular Resources' (PubMed, Bookshelf, PubMed Central, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem) and 'NCBI News & Blog' with a recent article about 3D molecular structures.

Databáze GenBank – Submit (BankIt, Sequin, SRA...)

→ vkládání sekvencí (po přihlášení heslem a uživatelským jménem)

Submission Portal

Submit to the world's largest public repository of biological and scientific information

Type a few words about the sequence data you are submitting and select an option to learn more. You can also browse submission information below.

What do you want to submit?

Enter a few words about your sequence data.

16S rRNA



Suggest tool

Suggested tools

[GenBank >](#)

Submit ribosomal RNA (rRNA), rRNA-ITS, SARS-CoV-2, Influenza, Norovirus, Dengue, metazoan COX1 or eukaryotic nuclear mRNA

[SRA >](#)

SRA accepts unassembled reads from high throughput sequencing platforms. Submitted data files should generally be minimally processed and include per-base quality scores.

Databáze GenBank – Taxonomy

→ vyhledávání sekvencí podle názvu taxonu (viru)

The screenshot shows the NCBI website interface. At the top, there is a blue header with the NIH logo and the text "National Library of Medicine National Center for Biotechnology Information". A "Log in" button is located in the top right corner. Below the header, there is a search bar with a dropdown menu set to "All Databases" and a "Search" button. On the left side, there is a vertical sidebar menu with various categories. The "Taxonomy" link is highlighted with a red circle and a red arrow. The main content area is titled "Welcome to NCBI" and contains several sections: "Submit" (Deposit data or manuscripts into NCBI databases), "Download" (Transfer NCBI data to your computer), "Learn" (Find help documents, attend a class or watch a tutorial), "Develop" (Use NCBI APIs and code libraries to build applications), "Analyze" (Identify an NCBI tool for your data analysis task), and "Research" (Explore NCBI research and collaborative projects). On the right side, there are sections for "Popular Resources" (PubMed, Bookshelf, PubMed Central, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem) and "NCBI News & Blog" (Explore 3D Molecular Structures with iCn3D, 21 Nov 2024, Do you want to analyze three-dimensional structures and highlight, Try Out a Development Version of NCBI's Publicly Available Annotation Tool, EGAPx).

Databáze GenBank – Taxonomy

→ vyhledávání sekvencí podle názvu taxonu (viru)

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Taxonomy

All Databases Downloads Submissions Tools How To

Databases
Taxonomy
Contains the names and phylogenetic lineages of more than 160,000 organisms that have molecular data in the NCBI databases. New taxa are added to the Taxonomy database as data are deposited for them.

Downloads
[FTP: NCBI Taxonomy](#)
This site contains the full taxonomy database along with files associating nucleotide and protein sequence records with their taxonomy IDs. See the taxdump_readme.txt and gi_taxid.readme files for more information.

Submissions
[GenBank: Barcode](#)
Tool for submission to the GenBank database of Barcode short nucleotide sequences from a standard genetic locus for use in species identification.

Quick Links
[Taxonomy](#)
[Taxonomy Browser](#)
[Taxonomy Common Tree](#)

Databáze GenBank – Taxonomy

→ vyhledávání sekvencí podle názvu taxonu (viru)

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

Taxonomy Taxonomy Orthohantavirus Search

Limits Advanced Help

Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

Using Taxonomy

- [Quick Start Guide](#)
- [FAQ](#)
- [Handbook](#)
- [Taxonomy FTP](#)
- [Important Update: Phyla Changing](#)
- [Important Update: New Flu species Names](#)

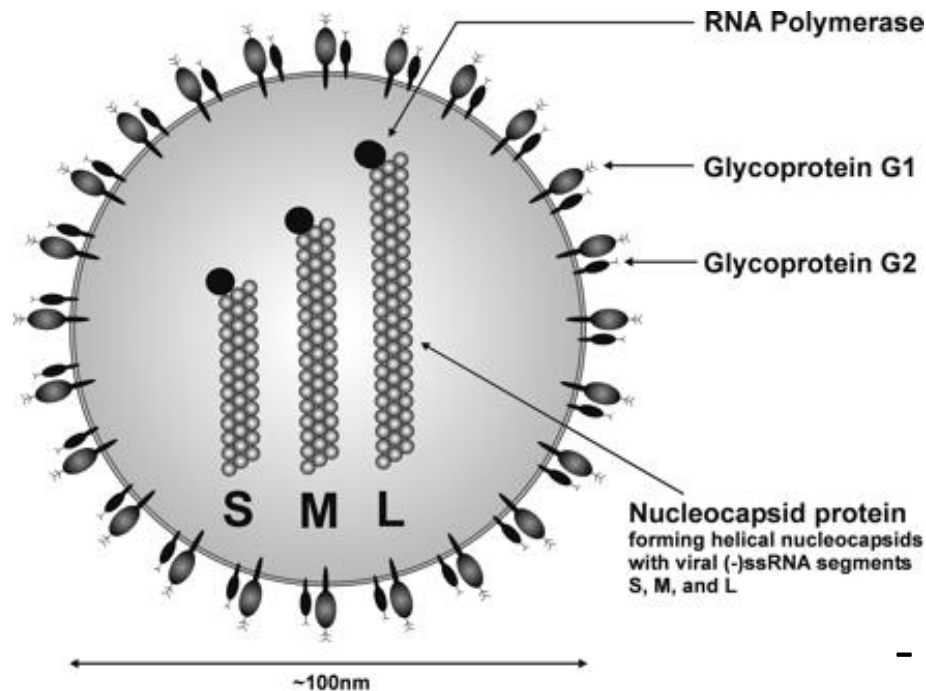
Taxonomy Tools

- [Browser](#)
- [Common Tree](#)
- [Statistics](#)
- [Name/ID Status](#)
- [Genetic Codes](#)
- [Linking to Taxonomy](#)
- [Extinct Organisms](#)

Other Resources

- [GenBank](#)
- [LinkOut](#)
- [E-Utilities](#)
- [Batch Entrez](#)
- [INSDC](#)

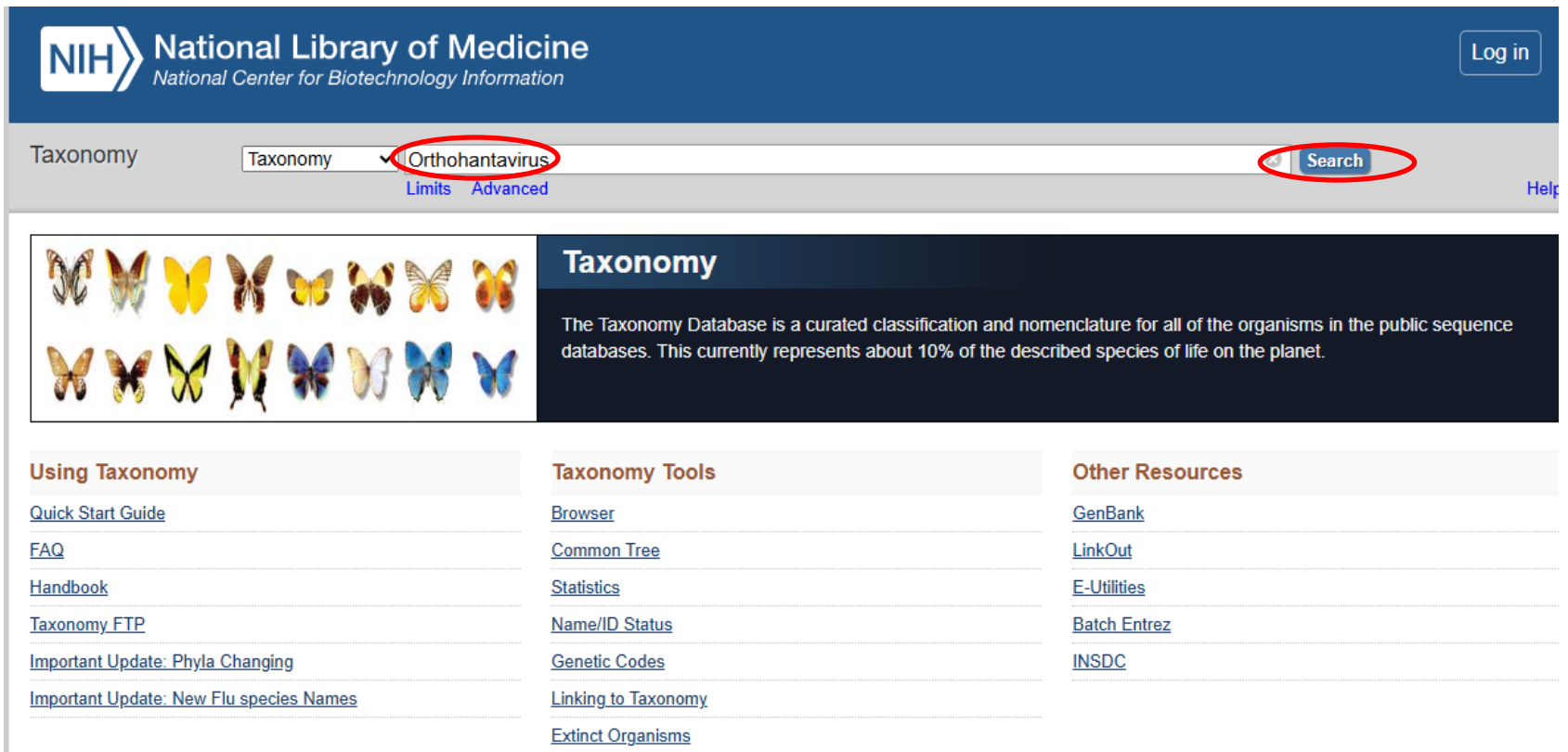
rod *Orthohantavirus* (*Elliiovirales*: *Hantaviridae*)



- ssRNA
- negativní polarita
- segmentovaný genom (S, M a L segment)
- M segment → reassortment
- zoonotický potenciál

Databáze GenBank – Taxonomy

→ vyhledávání sekvencí podle názvu taxonu (viru)



NIH National Library of Medicine
National Center for Biotechnology Information

Taxonomy Taxonomy **Orthohantavirus** Search

Limits Advanced Help

Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

Using Taxonomy

- [Quick Start Guide](#)
- [FAQ](#)
- [Handbook](#)
- [Taxonomy FTP](#)
- [Important Update: Phyla Changing](#)
- [Important Update: New Flu species Names](#)

Taxonomy Tools

- [Browser](#)
- [Common Tree](#)
- [Statistics](#)
- [Name/ID Status](#)
- [Genetic Codes](#)
- [Linking to Taxonomy](#)
- [Extinct Organisms](#)

Other Resources

- [GenBank](#)
- [LinkOut](#)
- [E-Utilities](#)
- [Batch Entrez](#)
- [INSDC](#)

Databáze GenBank – Taxonomy

→ vyhledávání sekvencí podle názvu taxonu (viru)

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy BioCollections

Search for [] as complete name [x] lock [Go] [Clear]

Display 3 levels using filter: none

Nucleotide Protein Structure Genome Popset SNP Conserved Domains GEO Datasets PubMed Central
 Gene HomoloGene SRA Experiments LinkOut BLAST GEO Profiles Protein Clusters Identical Protein Groups BioProject
 BioSample Assembly dbVar Genetic Testing Registry Host Viral Host PubChem BioAssay

Lineage (full): [Viruses](#); [Riboviria](#); [Orthornavirae](#); [Negarnaviricota](#); [Polyploviricotina](#); [Ellioviricetes](#); [Bunyavirales](#); [Hantaviridae](#); [Mammantavirinae](#)

o [Orthohantavirus](#) *Click on organism name to get more information.*

o [Amga orthohantavirus](#)

o [Amga virus](#)

o [Orthohantavirus andesense](#)

o [Araraquara virus](#)

o [Araraquara-like virus strain P5/Cajuru](#)

o [Bermejo virus](#)

o [Castelo dos Sonhos virus](#)

o [Central Plata virus](#)

o [Hantavirus Ac1Hu](#)

o [Hantavirus Bu3Hu](#)

o [Hantavirus Ce20](#)

o [Hantavirus Ce22](#)

o [Hantavirus Pe4Hu](#)

o [Hantavirus Sa63](#)

o [Hantavirus U89](#)

o [Hu39694 virus](#)

o [Jabora hantavirus](#)

o [Juititaba-like virus strain On576](#)

o [Juititaba-like virus strain PB 1002](#)

o [Juititaba-like virus strain PB 1033](#)

o [Juititaba-like virus strain PB 981](#)

o [Lechiguana virus](#)

o [Maciel virus](#)

o [Neembucu hantavirus](#)

o [Oran virus](#)

o [Pergamino virus](#)

o [Tunari virus](#)

o [Orthohantavirus asamaense](#)

o [Asama virus](#)

o [Orthohantavirus asikkalaense](#)

o [Orthohantavirus bayoui](#)

o [Catacamas virus](#)

Databáze GenBank – Taxonomy

→ vyhledávání sekvencí podle názvu taxonu (viru)

The screenshot displays the NCBI Taxonomy Browser interface. At the top, there are navigation tabs for Entrez, PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and BioCollections. The search bar contains the text 'Amga virus' and is set to search 'as complete name'. Below the search bar, the results for 'Amga virus' are shown, including its Taxonomy ID (1511732), current name, and various links for further information. On the right side, there is a table titled 'Entrez records' which lists the number of records in different databases for this taxon. The 'Nucleotide' record count is highlighted with a red circle.

Amga virus

Taxonomy ID: 1511732 (for references in articles please use NCBI:txid1511732)

current name: **Amga virus**

NCBI BLAST name: **viruses**

Rank: **no rank**

Genetic code: [Translation table 1 \(Standard\)](#)

Host: vertebrates

[Lineage \(full\)](#)
[Viruses](#); [Riboviria](#); [Orthornavirae](#); [Negarnaviricota](#); [Polyploviricotina](#); [Ellioviricetes](#); [Bunyavirales](#); [Hantaviridae](#); [Mammantavirinae](#); [Orthohantavirus](#); [Amga orthohantavirus](#)

[View and Analyze sequences in NCBI Virus](#)
[ICTV homepage](#)

Information from sequence entries

[Show organism modifiers](#)

Entrez records	
Database name	Direct links
Nucleotide	14
Protein	14
Genome	1
PubMed Central	9
Identical Protein Groups	14
Assembly	1
Taxonomy	1

Databáze GenBank – Taxonomy

→ získávání a stahování sekvencí (data mining)

Items: 14

- [Amga virus strain MSB148347 nucleocapsid gene, partial cds](#)
 1. 582 bp linear cRNA
Accession: KM201419.1 GI: 808178698
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

- [Amga virus strain MSB148558 nucleocapsid gene, complete cds](#)
 2. 1,627 bp linear cRNA
Accession: KM201411.1 GI: 808178696
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

- [Amga virus strain MSB148347 glycoprotein precursor, gene, partial cds](#)
 3. 729 bp linear cRNA
Accession: KM201420.1 GI: 808178694
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

- [Amga virus strain MSB148457 glycoprotein precursor, gene, partial cds](#)
 4. 683 bp linear cRNA
Accession: KM201417.1 GI: 808178692
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

- [Amga virus strain MSB148436 glycoprotein precursor, gene, partial cds](#)
 5. 727 bp linear cRNA
Accession: KM201415.1 GI: 808178690
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

Databáze GenBank – Taxonomy

→ získávání a stahování sekvencí

Amga virus strain MSB148347 RNA-dependent RNA polymerase gene, partial cds

GenBank: KM201421.1

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

[Go to:](#) ☺

LOCUS KM201421 1302 bp cRNA linear VRL 31-JAN-2016
DEFINITION Amga virus strain MSB148347 RNA-dependent RNA polymerase gene,
partial cds.
ACCESSION KM201421
VERSION KM201421.1
KEYWORDS .
SOURCE Amga virus
ORGANISM [Amga virus](#)
Viruses; Riboviria; Orthornavirae; Negarnaviricota;
Polyploviricotina; Ellioviricetes; Bunyavirales; Hantaviridae;
Mammantavirinae; Orthohantavirus; Amga orthohantavirus.
REFERENCE 1 (bases 1 to 1302)
AUTHORS Kang,H.J., Gu,S.H., Cook,J.A. and Yanagihara,R.
TITLE Amga virus, a newly identified hantavirus in the Laxmann's shrew
(Sorex caecutiens)
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 1302)
AUTHORS Kang,H.J., Gu,S.H., Cook,J.A. and Yanagihara,R.
TITLE Direct Submission
JOURNAL Submitted (15-JUL-2014) Department of Tropical Medicine and Medical
Microbiology and Pediatrics, John A. Burns School of Medicine,
University of Hawaii at Manoa, 651 Ilalo Street, Honolulu, HI
96813, USA
FEATURES Location/Qualifiers
source 1..1302
/organism="Amga virus"
/mol_type="viral cRNA"
/strain="MSB148347"
/isolation_source="lung"
/host="Sorex caecutiens"
/db_xref="taxon:1511732"
/geo_loc_name="Russia"
/collection_date="14-Aug-2006"

Co lze zjistit:

- druh viru
- sekvenovaný lokus
- Accession number
- (hostitel)
- (tkáň hostitele)
- (lokalita sběru)
- (datum sběru)

Databáze GenBank – Taxonomy

→ získávání a stahování sekvencí (data mining)

```
CDS
<1..>1302
/codon_start=1
/product="RNA-dependent RNA polymerase"
/protein_id="AKD00010.1"
/translation="SFLSRVIYKHYKSLISEVTTCCFFLFEKGLHGNVNEEAKIHLETV
EWALKFRAKEEKYGSHLVEHGYRISDLYLNPSLVEQQMYCQDVVELGAFELNNMLLSK
TQVVGNSIINKHWNLPYFSQTRNISLKGMSGSIQEDGHLASVTLIEAIRYLQNSRHN
PSLLQLYEETRTAKAQARIVRKYQRTEADRGFFITTLPTRCRLEIIEDYYDAIAKNVP
EEYISYGGERKILNIQQALEKALRWASGESHLELSTGKLIIPMKRKLMYVSADATKWSP
GDNSAKFRRTAVLHNGLRDDKLRNCVIDALINIYKTDFFMSRKLKKYIGNMESLDEH
VKAFLSFFPDGNSGEVHGNWLQGNLNKCSSLFAVGMSLLFKRNVKELFPELECFEFA
HHSDDGLFIYGYLEPVDDGTDWFMVYVTTQQIQAGNHHWYSVNT"
```

ORIGIN

```
1 tcattcttat ctagagtgat atataagcat tataaaagct tgatttcaga agtgactaca
61 tgtttttttc tatttgagaa gggattacat gggaaatgtaa atgaggaggc aaaaatacat
121 ttggaaactg tagaatgggc actaaagtcc cgggcaaagg aagagaaata tggttccacc
181 ttggtagaac atggatagat gatcagtgat ttgtatttga acccatctct ggttgagcaa
241 caaatgtatt gtcaagatgt agtggaaacta ggagcatttg aactaaataa tatgctgcta
301 tcgaagacac aagtgtgttg aaattctatt attaataagc attggaactt accatatttt
361 agtcaaacgc ggaatattag tttaaaggga atgtctggtt caatccaaga agatggacat
421 ctatctgctt cagtaactct tatagaagca atccgggtatt tacaanaattc acgtcacaac
481 cctagtctcc tccagttata tgaagagacc agaacagcaa aagcacaggc tagaattggt
541 aggaaatatac agagaactga ggcagataga ggattcttca taacgactct ccctaccaga
601 tgtagactag aaatcattga ggactattat gatgctatag caaagaatgt cccagaagaa
661 tatatttcat atgggtgtga gcgaaagata ttaaatattc agcaagcact tgaaaaagct
721 ttaagatggg cttctggtga gagtcatcta gagctctcaa cgggtaagct gataccaatg
781 aaacggaaat tgatgtatgt aagcgcagat gctacaaaat gggtcacctgg tgataattca
841 gctaaattcc gtcgattcac tgctgtttta cataatggat tgagggatga taaattgcca
901 aattgtgtga tagatgcact aatcaatatt tataagactg atttttttat gtctagaaaa
961 ctcaaaaagt atataggaaa tatggagagt ttagatgagc atgtaaaggc cttctgtcgc
1021 ttttttccag atggaaattc tggatgaagta catggtaatt gggtacaggg gaatctaacc
1081 aaatgctctt cctgtgttgc tgtaggatag tctttacttt ttaaaagagt ctggaaagaa
1141 ttatttccgg aactggaatg cttttttgaa tttgcacatc attcagatga tgggcttttc
1201 atatatggtt atcttgaacc agttgatgat ggaacagact ggtttatgta tgtcacacag
1261 caaattcagg caggtaatca tcattggtac agtgtaaata ca
```

Lze zjistit:

- nukleotidovou sekvenci
- aminokyselinovou sekvenci

Databáze GenBank – Taxonomy

→ stahování sekvencí (**formát FASTA**)

FASTA ▾

Send to: ▾

Amga virus strain MSB148347 RNA-dependent RNA polymerase gene, partial cds

GenBank: KM201421.1

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

```
>KM201421.1 Amga virus strain MSB148347 RNA-dependent RNA polymerase gene, partial cds
TCATTCTTATCTAGAGTGATATATAAGCATTATAAAAAGCTTGATTTTCAGAAGTGACTACATGTTTTTTTC
TATTTGAGAAGGGATTACATGGGAATGTAATGAGGAGGCCAAAAATACATTTGGAAACTGTAGAATGGGC
ACTAAAGTTCCGGGCAAAGGAAGAGAAAATATGGTTCACACTTGGTAGAACATGGATATAGGATCAGTGAT
TTGTATTTGAACCCATCTCTGGTTGAGCAACAAATGTATTGTCAAGATGTAGTGGAAGTGGGAGCATTG
AACTAAATAATATGCTGCTATCGAAGACACAAGTTGTTGGAAATCTATTATTAATAAGCATTGGAACTT
ACCATATTTTAGTCAAACGCGGAATATTAGTTTAAAGGGAATGTCTGGTTCATCCAAGAAGATGGACAT
CTATCTGCTTCAGTAACTCTTATAGAAGCAATCCGGTATTTACAAAATTACGTCACAACCCTAGTCTCC
TCCAGTTATATGAAGAGACCAGAACGACAAAAGCACAGGCTAGAATTGTTAGGAAATATCAGAGAACTGA
GGCAGATAGAGGATTCTTCATAACGACTCTCCCTACCAGATGTAGACTAGAAATCATTGAGGACTATTAT
GATGCTATAGCAAAGAATGTCCCAGAAGAATATATTTTCATATGGTGGTGAGCGAAAAGATATTAATATTC
AGCAAGCACTTGAAAAAGCTTTAAGATGGGCTTCTGGTGAGAGTCATCTAGAGCTCTCAACGGGTAAGCT
GATACCAATGAAACGGAAATTGATGTATGTAAGCGCAGATGCTACAAAATGGTCACCTGGTGATAATTCA
GCTAAATCCGTCGATCTACTGCTGTTTTACATAATGGATTGAGGGATGATAAATTGCGAAATGTGTGA
TAGATGCACTAATCAATATTTATAAGACTGATTTTTTATGTCTAGAAAACCAAAAAGTATATAGGAAA
TATGGAGAGTTTAGATGAGCATGTAAAGGCCTCTTGTGCTTTTTCCAGATGGAAATCTGGTGAAGTA
CATGGTAATTGGTTACAGGGGAATCTAAACAAATGCTCTTCCCTGTTTGTCTGATGGTATGCTTTACTTT
TTAAAAGAGTCTGGAAAAGAATTTTCCGGAAGTGAATGCTTTTTTGAATTTGCACATCATTTCAGATGA
TGGGCTTTTCATATATGGTTATCTTGAACCAAGTTGATGATGGAACAGACTGGTTTATGTATGTCACACAG
CAAATTCAGGCAGGTAATCATCATTGGTACAGTGTAAATACA
```

Taxonomie virů

Jak zjistím, které taxony potřebuji v datasetu?

Jak zjistím druhy virů, které se nacházejí v dané čeledi / řádu?

→ **ICTV (International Committee on Taxonomy of Viruses)**



Home About Taxonomy Report Information Forums Help

International Committee on Taxonomy of Viruses: ICTV

Official Taxonomic Resources

- ICTV Taxonomy Browser**
Search and browse the virus taxonomy
- Master Species List (MSL)**
MSL: Spreadsheet of all current species
- Virus Metadata Resource (VMR)**
VMR: Virus exemplars for every species

This web site collects information on user preferences and usage statistics so we can provide you with a more personalized experience. [More info](#)

Taxonomie virů

Jak zjistím, které taxony potřebuji v datasetu?

Jak zjistím druhy virů, které se nacházejí v dané čeledi / řádu?

→ ICTV (International Committee on Taxonomy of Viruses)

The screenshot shows the ICTV Taxonomy Browser interface. At the top, there is a navigation bar with links for Home, About, Taxonomy, Report, Information, Forums, and Help. Below this is a blue header with the text "Current ICTV Taxonomy Release". The main heading is "Taxonomy Browser". A paragraph explains that the browser provides access to the current virus taxonomy and is updated when a new release is approved. There are three buttons: "Taxonomy Search", "Taxonomy Browser", and "Download Current Taxonomy Spreadsheet (MSL)". A search box contains the text "Orthohantavirus" and has "Search" and "Reset" buttons. Below the search box is a checkbox for "Select to search across all ICTV releases" and a "Show 10 entries" dropdown. A table displays search results with columns for Release, Rank, and Name. The first result is for the year 2023, Genus, and the name "Riboviria · Orthornavirae · Negarnaviricota · Polyploviricotina · Bunyaviricetes · Elliovirales · Hantaviridae · Mammantavirinae · Orthohantavirus". At the bottom, there is a blue footer with a privacy notice: "This web site collects information on user preferences and usage statistics so we can provide you with a more personalized experience." and buttons for "Don't collect" and "Accept".

Home > Current ICTV Taxonomy Release

Current ICTV Taxonomy Release

Taxonomy Browser

This taxonomy browser provides access to the current virus taxonomy. This page will be updated whenever a new taxonomy release has been approved by the ICTV.

[Taxonomy Search](#) [Taxonomy Browser](#) [Download Current Taxonomy Spreadsheet \(MSL\)](#)

Unless the "Select to search across all ICTV releases" button is checked below, your search will be against the taxonomy release indicated below the search box (or below the search result set, if present). To search against the current release, refresh the page.

Orthohantavirus

Select to search across all ICTV releases

Show 10 entries

Release	Rank	Name
2023	Genus	Riboviria · Orthornavirae · Negarnaviricota · Polyploviricotina · Bunyaviricetes · Elliovirales · Hantaviridae · Mammantavirinae · Orthohantavirus

This web site collects information on user preferences and usage statistics so we can provide you with a more personalized experience.

[More info](#)

Taxonomie virů

Jak zjistím, které taxony potřebuji v datasetu?

Jak zjistím druhy virů, které se nacházejí v dané čeledi / řádu?

→ ICTV (International Committee on Taxonomy of Viruses)

Orthohantavirus

Select to search across all ICTV releases

Show 10 entries

	Release	Rank	Name
<input type="button" value="View"/> <input type="button" value="History"/>	2023	Genus	<i>Riboviria</i> › <i>Orthornavirae</i> › <i>Negarnaviricota</i> › <i>Polyploviricotina</i> › <i>Bunyaviricetes</i> › <i>Elliovirales</i> › <i>Hantaviridae</i> › <i>Mammantavirinae</i> › Orthohantavirus
<input type="button" value="View"/> <input type="button" value="History"/>	2023	Species	<i>Riboviria</i> › <i>Orthornavirae</i> › <i>Negarnaviricota</i> › <i>Polyploviricotina</i> › <i>Bunyaviricetes</i> › <i>Elliovirales</i> › <i>Hantaviridae</i> › <i>Mammantavirinae</i> › Orthohantavirus › Orthohantavirus <i>andesense</i>
<input type="button" value="View"/> <input type="button" value="History"/>	2023	Species	<i>Riboviria</i> › <i>Orthornavirae</i> › <i>Negarnaviricota</i> › <i>Polyploviricotina</i> › <i>Bunyaviricetes</i> › <i>Elliovirales</i> › <i>Hantaviridae</i> › <i>Mammantavirinae</i> › Orthohantavirus › Orthohantavirus <i>artybashense</i>
<input type="button" value="View"/> <input type="button" value="History"/>	2023	Species	<i>Riboviria</i> › <i>Orthornavirae</i> › <i>Negarnaviricota</i> › <i>Polyploviricotina</i> › <i>Bunyaviricetes</i> › <i>Elliovirales</i> › <i>Hantaviridae</i> › <i>Mammantavirinae</i> › Orthohantavirus › Orthohantavirus <i>asamaense</i>
<input type="button" value="View"/> <input type="button" value="History"/>	2023	Species	<i>Riboviria</i> › <i>Orthornavirae</i> › <i>Negarnaviricota</i> › <i>Polyploviricotina</i> › <i>Bunyaviricetes</i> › <i>Elliovirales</i> › <i>Hantaviridae</i> › <i>Mammantavirinae</i> › Orthohantavirus › Orthohantavirus <i>asikkalaense</i>
<input type="button" value="View"/> <input type="button" value="History"/>	2023	Species	<i>Riboviria</i> › <i>Orthornavirae</i> › <i>Negarnaviricota</i> › <i>Polyploviricotina</i> › <i>Bunyaviricetes</i> › <i>Elliovirales</i> › <i>Hantaviridae</i> › <i>Mammantavirinae</i> › Orthohantavirus › Orthohantavirus <i>bayoui</i>
<input type="button" value="View"/> <input type="button" value="History"/>	2023	Species	<i>Riboviria</i> › <i>Orthornavirae</i> › <i>Negarnaviricota</i> › <i>Polyploviricotina</i> › <i>Bunyaviricetes</i> › <i>Elliovirales</i> › <i>Hantaviridae</i> › <i>Mammantavirinae</i> › Orthohantavirus › Orthohantavirus <i>boweense</i>

This web site collects information on user preferences and usage statistics so we can provide you with a more personalized experience.

Don't collect

[More info](#)

Databáze GenBank – Taxonomy

→ stahování sekvencí (**formát FASTA**)

FASTA ▾

Send to: ▾

Amga virus strain MSB148347 RNA-dependent RNA polymerase gene, partial cds

GenBank: KM201421.1

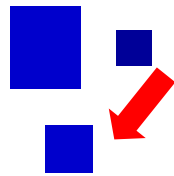
[GenBank](#) [Graphics](#)

```
>KM201421.1 Amga virus strain MSB148347 RNA-dependent RNA polymerase gene, partial cds
TCATTCTTATCTAGAGTGATATATAAGCATTATAAAAAGCTTGATTTTCAGAAGTGACTACATGTTTTTTTC
TATTTGAGAAGGGATTACATGGGAATGTAATGAGGAGGCCAAAAATACATTTGGAAACTGTAGAATGGGC
ACTAAAGTTCCGGGCAAAGGAAGAGAAAATATGGTTCACACTTGGTAGAACATGGATATAGGATCAGTGAT
TTGTATTTGAACCCATCTCTGGTTGAGCAACAAATGTATTGTCAAGATGTAGTGGAAGTACTAGGAGCATTTG
AACTAAATAATATGCTGCTATCGAAGACACAAGTTGTTGGAAATCTATTATTAATAAGCATTGGAACTT
ACCATATTTTAGTCAAACGCGGAATATTAGTTTAAAGGGAATGTCTGGTTCATCCAAGAAGATGGACAT
CTATCTGCTTCAGTAACTCTTATAGAAGCAATCCGGTATTTACAAAATTACGTCACAACCCTAGTCTCC
TCCAGTTATATGAAGAGACCAGAACGACAAAAGCACAGGCTAGAATTGTTAGGAAATATCAGAGAACTGA
GGCAGATAGAGGATTCTTCATAACGACTCTCCCTACCAGATGTAGACTAGAAATCATTGAGGACTATTAT
GATGCTATAGCAAAGAATGTCCCAGAAGAATATATTTTATATGGTGGTGAGCGAAAAGATATTAATATTC
AGCAAGCACTTGAAAAAGCTTTAAGATGGGCTTCTGGTGAGAGTCATCTAGAGCTCTCAACGGGTAAGCT
GATACCAATGAAACGGAAATTGATGTATGTAAGCGCAGATGCTACAAAATGGTCACCTGGTGATAATTCA
GCTAAATCCGTCGATTCAGTCTGTTTTACATAATGGATTGAGGGATGATAAATTGCGAAATGTGTGA
TAGATGCACTAATCAATATTTATAAGACTGATTTTTTTATGTCTAGAAAACCAAAAAGTATATAGGAAA
TATGGAGAGTTTAGATGAGCATGTAAGGCCCTCTTGTCTGTTTTTCCAGATGGAAATCTGGTGAAGTA
CATGGTAATTGGTTACAGGGGAATCTAAACAAATGCTCTTCCCTGTTTGTCTGATGGTATGCTTTACTTT
TTAAAAGAGTCTGGAAAAGAAATTTCCGGAAGTGAATGCTTTTTTGAATTTGCACATCATTTCAGATGA
TGGGCTTTTTCATATATGGTTATCTTGAACAGTTGATGATGGAACAGACTGGTTTATGTATGTCACACAG
CAAATTCAGGCAGGTAATCATCATTGGTACAGTGTAAATACA
```

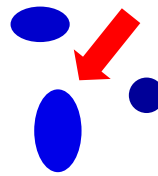
Dataset pro fylogenetické analýzy

reprezentativní výběr taxonů do datasetu, **tentýž gen**

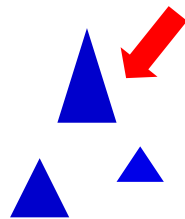
čeleď čtverečků



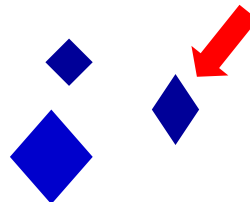
čeleď koleček



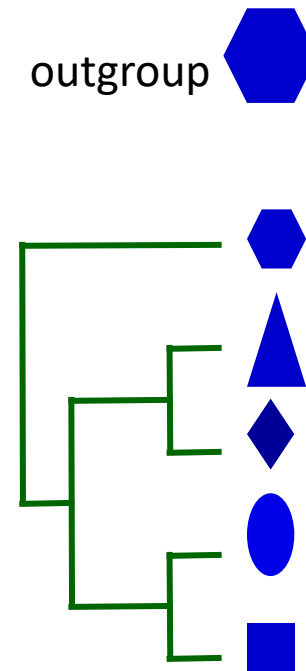
čeleď trojúhelníků



čeleď kosočtverců



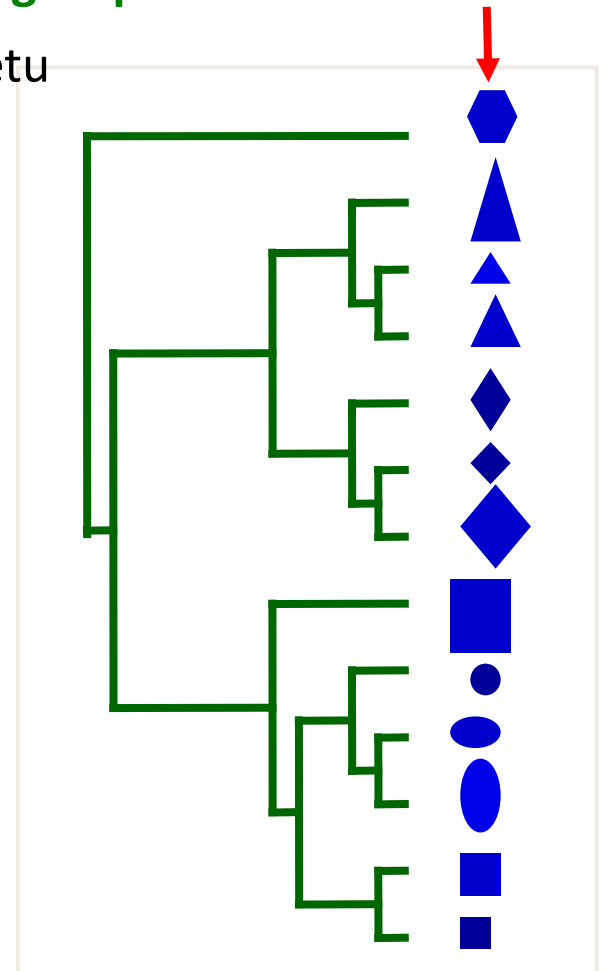
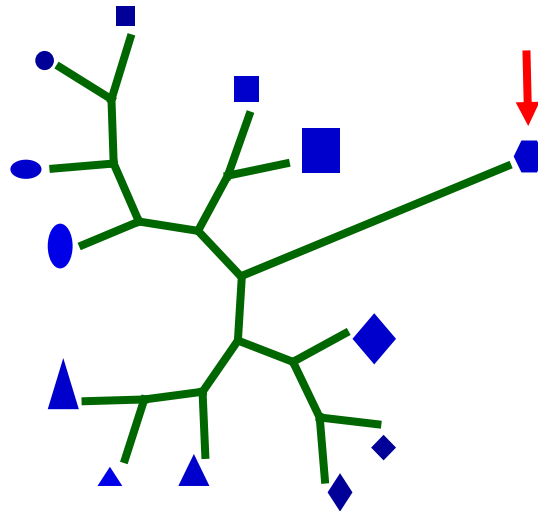
outgroup



Výběr outgroupu pro fylogenetické analýzy

Výběr outgroupu (tj. kořene) - požadavky na outgroup:

- sekvence **stejného genu** jako u taxonů v datasetu
- **vzdálenější, ale příbuzný organismus**
(outgroup musí být out-group)
(neměl by ale být příliš vzdálený)



1. Praktická úloha – tvorba datasetu

Postup:

1. Vyhledat sekvence TÉHOŽ GENU (tj. například L segment či jeho synonymum, RNA dependent RNA polymerase) virů rodu *Orthohantavirus*
dataset musí být VŽDY složen ze sekvencí stejného genu
2. Vyhledané a vybrané sekvence ve formátu FASTA zkopírovat pod sebe (Ctrl+C, Ctrl+V) do Poznámkového bloku
3. Pokud je v GenBanku k danému druhu viru více sekvencí téhož genu, vybíráme:
 - nejdelší sekvenci / sekvence
 - sekvence pocházející z různých hostitelů
 - sekvence pocházející z různých geografických oblastí
4. Vybrat vhodný outgroup a vložit ho do datasetu

1. Praktická úloha – tvorba datasetu

Hantavirový genom:

L segment = RNA dependentní RNA polymeráza (RdRp)

GenBank: *RdRp*, RNA-dependent RNA polymerase,
segment L, L gene

M segment = glykoproteiny G1 a G2 (Gn a Gc)

GenBank: *Gc protein*, *M polyprotein*, *glycoprotein gene*,
glycoprotein precursor

S segment = nukleokapsidový protein N

GenBank: *nucleocapsid protein*, *nucleoprotein S*,
nucleoprotein gene

Tvorba datasetu pro fylogenetické analýzy

Soubor Úpravy Formát Zobrazení Nápověda

```
>FJ495100.1 Tula virus strain TULV/Sestrze/Mag98_02 segment L RNA dependent RNA polymerase gene, partial cds
CAGGAGACAATTCTGCAAAATTCAGAAGGTTTACTCAAACATTGCATGATGGTTTACGAGATGATAAACT
GAAAAGGTGTGTTAGATGCTTTACGGAACATATATGAAACTGATTTTTTTATGCTAGAAAAGCTACAT
AGGTATATTGATGGTATGGACGATTTGTCTGAATTTGTAGAGGATTTTTTGTCAATTTTCCCAAACAAGG
TGCTGCTGCCATAAAAAGGAAATTTGGCTTCAGGGAAATCTAAACAATGTTTCATCATTATTTGGTGCAGC
AGTTTCCCTGTTATTTAGAAAAGTCTGGAACCTCTTTATCCAGAACTGGATTGCTTTTTTGAGTTTGCA
CATCATTCT
```

```
>MK605685.1 Orthohantavirus tulaense isolate TULV/Magr1100/Beskydy/CZ/2014 segment L RNA-dependent RNA polymerase gene, partial cds
CCAGGGGATAACTCAGCAAAATTCGAAGATTTACCCAGACATTACATGATGGTTTCCGAGATGATAAAC
TGAAAAGGTGTGAGTTGATGCATTGCGGAATATATATGAAACTGATTTTTTTATGCTCGAAAAGTTACA
TAGGTATATTGATGGTATGGATGATTTATCTGAATTTGTAGAAGACTTTTTATCATNCTTCCAAATAAG
GTATCTGCTGCTATTAAGGGAATTTGGCTTCAGGGAAACCTAAACAAGTGCATCATTATTTGGCGCAG
CAGTTTCCCTGCTATTTAGAAAATTTGGAGCTCCTTTATCCAGAATTAGACTGCTTTTTTGAG
```

```
>KC522413.1 Tula virus isolate JiTr/Opava/12 RNA-dependent RNA polymerase gene, partial cds
ACGAAGTGGTCTCCAGGAGATAATTCAGCAAAATTTCAAGATTTACTCAGGCATTACATGATGGTCTGC
GAGATGATAAATTAAGGTTGTGAGTGGATGCATTGCGGAATATATATGAAACTGACTTTTTTTATGTC
TAGAAAAGTTACATAGGTATATAGATGGTATGGATGATTTGTCTGAATTTGTAGAAGACTTTTTATCATT
TTCCGAAATAAGGTATCTGCTGCTATTAAGGAAATTTGGCTTCAAGGAAATCTAAACAAGTGCATCAT
TATTTGGTGCAGCAGTTTCTTGTCTATTTAGAAAAGATTTGGAGCTCCTTTATCCAGAATTAGACTGCTT
TTTTGAATTTGCCACCACCTCGAACGACGCA
```

```
>FJ495102.1 Tula virus strain TULV/Sred ob Dravi/Ms51_97 segment L RNA dependent RNA polymerase gene, partial cds
CAGGAGATAATTCAGCAAAATTCGAAGATTCAGGCATTGCATGATGGTTTACGAGATGATAAAT
GAAGAGATGTGTAGTGGATGCTTTGCGGAATATATATGAAACTGATTTTTTTATGCTCAAGAAAAGCTACAT
AGATATATTGATGGTATGGATGATTTGTCTGAATTTGTAGAAGACTTTTTTGTCAATTTTCCCAAATAAGG
TATCTGCTGCTATTAAGGAAATTTGGCTTCAAGGAAACTTAAATAAGTGTTCATCGTTGTTTGGTGCAGC
AGTTTCCCTGTTATTTAGAAAAGATTTGGAGTCTCCTTTATCCGGAATTGGATTGCTTTTTTGAGTTTGCA
CACCATTCT
```

```
>ON243802.1 Tula orthohantavirus isolate 56AA RNA-dependent RNA polymerase gene, partial cds
ACTCAGACATTACACGATGTTTACGGGATGATAAATTAAGGTTGCGTAGTAGATGCATTACGGAATA
TATATGAGACTGATTTTTTTATGCTAGAAAATACATAGATACATTGATGGTATGGATGATTTATCTGA
GTTTGTAGAAAGATTTTTATCATTCTTTCCAAATAAGGTATCCGCTGCTATTAAGGAAACTGGCTTCAA
GGAAACCTAAATAAATGCTCATCATTATTTGGTGCAGCAGTTTCCCTGCTATTTAGGAAAATTTGGAGTC
TGCTTT
```

```
>HQ728459.1 Tula virus isolate GER/152/Arv segment L RNA dependent RNA polymerase gene, partial cds
GATGCTACTAAATGGTCGCCAGGTGATAACTCCGCAAAATTTAGGAGTTTACTCAGGCACTACATGATG
GTTTAAAGAGATGATAAATTAAGAGATGTGTGGTTGATGCCTAGAGAAACATATATGAGACTGACTTTTT
TATGCTAGAAAATACATAGATACATTGATGGTATGGATGACTTGTCCGAATTTGTGGAGGACTTCTT
TCATTTTTTCCAAATAAAGTATCAGCTGCTATTAAGGGAATTTGGCTACAAGGCAATTTGAATAAGTGTCT
CTTCATTATTTGGTGCAGCTGTTTCCCTGTTATTCAGAAAATATGGAATCTTCTTACCAGAATTAGA
TTGTTTCTTTGAATTTGCACACCACCTCTGATGATGCATTATTTATTTGTTATTTAGAACCTACAGAT
GACGGA
```

Úprava datasetu pro fylogenetické analýzy

>FJ495100

```
CAGGAGACAATTCTGCAAAATTCAGAAGTTTACTCAAACATTGCATGATGGTTTACGAGATGATAAACT
GAAAAGGTGTGTTGATAGTCTTTACGGAAATATATGAAACTGATTTTTTTATGCTAGAAAAGCTACAT
AGGTATATTGATGGTATGGACGATTGTCTGAATTTGTAGAGGATTTTTTGTCAATTTTCCCAAACAAGG
TGCTGTCTGCCATAAAAGGAAATTTGGCTCAGGGAAATCTAAACAAATGTTTCATCATTATTTGGTGCAGC
AGTTTCCCTGTTATTTAGAAAAGTCTGGAACTTCTTTATCCAGAATGGATTGCTTTTTTGGAGTTTGCA
CATCATTCT
```

>MK605685

```
CCAGGGGATAACTCAGCAAAATTCGAAGATTTACCCAGACATTACATGATGGTTTGGCAGATGATAAAC
TGAAAAGGTGTGATGTTGATGCATTGCGGAATATATGAAACTGATTTTTTTATGCTCGAAAAGTTACA
TAGGTATATTGATGGTATGGATGATTTATCTGAATTTGTAGAAGACTTTTTATCATNCTTCCAAATAAG
GTATCTGCTGCTATTAAGGGAATTTGGCTCAGGGAAACCTAAACAAGTGTCTATCATTATTTGGCGCAG
CAGTTTCCCTGCTATTTAGAAAATTTGGAGCTCCTTTATCCAGAATTAGACTGCTTTTTTGGAG
```

>KC522413

```
ACGAAGTGGTCTCCAGGAGATAATTCAGCAAAATTTCAAGATTTACTCAGGCATTACATGATGGTCTGC
GAGATGATAAATTAAGGAGGTGTAGTGGATGCTTGCAGGAATATATGAAACTGACTTTTTTATGCT
TAGAAAAGTTACATAGGTATATAGATGGTATGGATGATTTGTCTGAATTTGTAGAAGACTTTTTATCATT
TTCCC GAATAAGGTATCTGCTCTATTAAAGGAAATTTGGCTCAAGGAAATCTAAACAAGTGTCTATCAT
TATTTGGTGCAGCAGTTTCTTGTCTATTTAGAAAAGATTTGGAGCTCCTTTATCCAGAATTAGACTGCTT
TTTTGAATTTGCCACCACCTCGAACGACGCA
```

>FJ495102

```
CAGGAGATAATTCAGCAAAATTCGAAGATTCAGGCATTGCATGATGGTTTACGAGATGATAAAT
GAAGAGATGTGTAGTGGATGCTTTGCGGAATATATGAAACTGATTTTTTTATGCTCAAGAAAGCTACAT
AGATATATTGATGGTATGGATGATTTGTCTGAATTTGTAGAAGACTTTTTTGTCAATTTTCCCAAATAAGG
TATCTGCTGCTATTAAGGAAATTTGGCTCAGGAAACTTAAATAAGTGTTCATCGTTGTTGGTGCAGC
AGTTTCCCTGTTATTTAGAAAAGATTTGGAGTCTCCTTTATCCGAATTTGGATTGCTTTTTTGGAGTTTGCA
CACCATTCT
```

>ON243802

```
ACTCAGACATTACACGATGGTTACGGGATGATAAATTAAGGAGTGCCTAGTAGATGCATTACGGAATA
TATATGAGACTGATTTTTTTATGCTAGAAAATACATAGATACATTGATGGTATGGATGATTTATCTGA
GTTTGTAGAAAGATTTTTTATCATTCTTTCCAAATAAGGTATCCGCTGCTATTAAGGAAACTGGCTTCAA
GGAAACCTAAATAAATGCTCATCATTATTTGGTGCAGCAGTTCCCTGCTATTTAGGAAAATTTGGAGTC
TGCTTT
```

>HQ728459

```
GATGCTACTAAATGGTCGCCAGGTGATAACTCCGCAAAATTTAGGAGGTTTACTCAGGCACTACATGATG
GTTAAGAGATGATAAATTAAGAGATGTGTGGTTGATGCACTGAGAAACATATGAGACTGACTTTTTT
TATGCTAGAAAATACATAGATACATTGATGGTATGGATGACTTTGTCGGAATTTGTGGAGGACTTTCTT
TCATTTTTCCCAAATAAGTATCAGCTGCTATTAAGGGAATTTGGCTACAAGCAATTTGAATAAGTGTCT
CTTCATATTTGGTGCAGCTGTTTCCCTGTTATTCAGAAAATATGGAATCTTCTTTACCCAGAATTAGA
TTGTTTCTTTGAATTTGCACACCACCTCTGATGATGATATTTATTTATGGTTATTTAGAACCCTACAGAT
GACGGA
```


Alignment

→ různé programy, různé algoritmy

BioEdit

Clustal W

T-Coffee

MUSCLE

MegAlign – součást balíčku DNASTAR

MAFFT – online

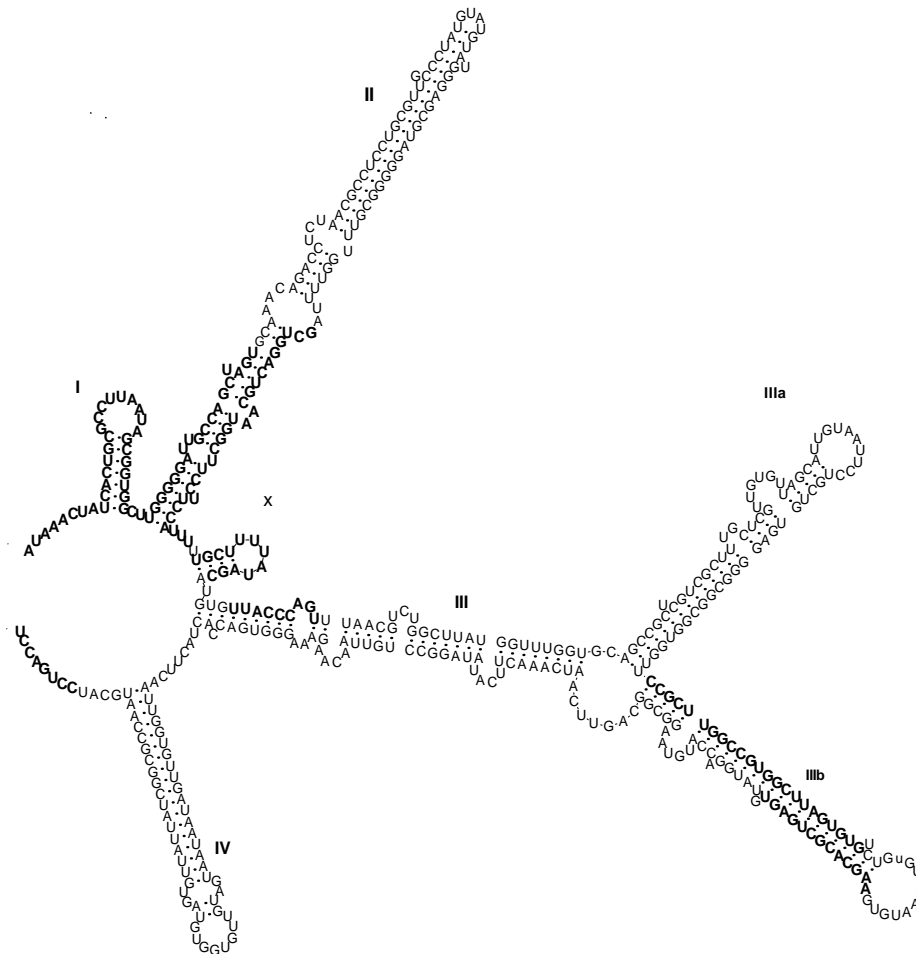
SeaView – možnost konkatenace

konkatenovaný alignment = spojený alignment 2 a více různých genů

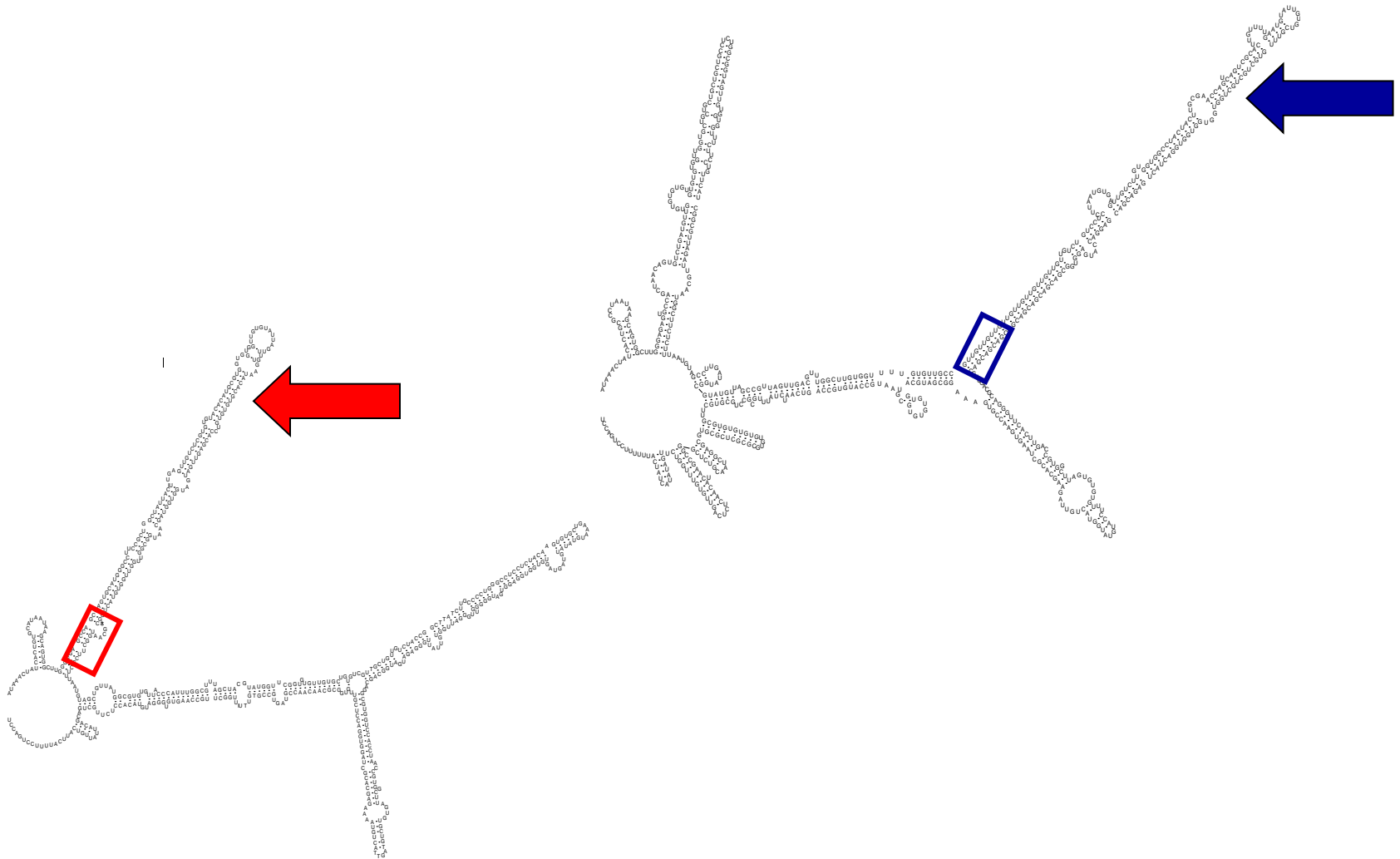
→ **další možnosti alignování:**

- využití „předalignovaných“ souborů z databází
- alignování zohledňující sekundární strukturu

Alignování za pomoci sekundární struktury



Alignování za pomoci sekundární struktury



MAFFT

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences



[Download version](#)

[Mac OS X](#)

[Windows](#)

[Linux](#)

[Source](#)

[Online version](#)

[Alignment](#)

[mafft --add](#)

[Merge](#)

[Phylogeny](#)

[Rough tree](#)

[Merits / limitations](#)

[Algorithms](#)

[Tips](#)

[Benchmarks](#)

[Feedback](#)

Strategy:

- Auto (FFT-NS-1, FFT-NS-2, FFT-NS-i or L-INS-i; depends on data size) [Updated](#)

Progressive methods

- FFT-NS-1 (Very fast; recommended for >2,000 sequences; progressive method)
- FFT-NS-2 (Fast; progressive method)
- G-INS-1 (Slow; progressive method with an accurate guide tree)

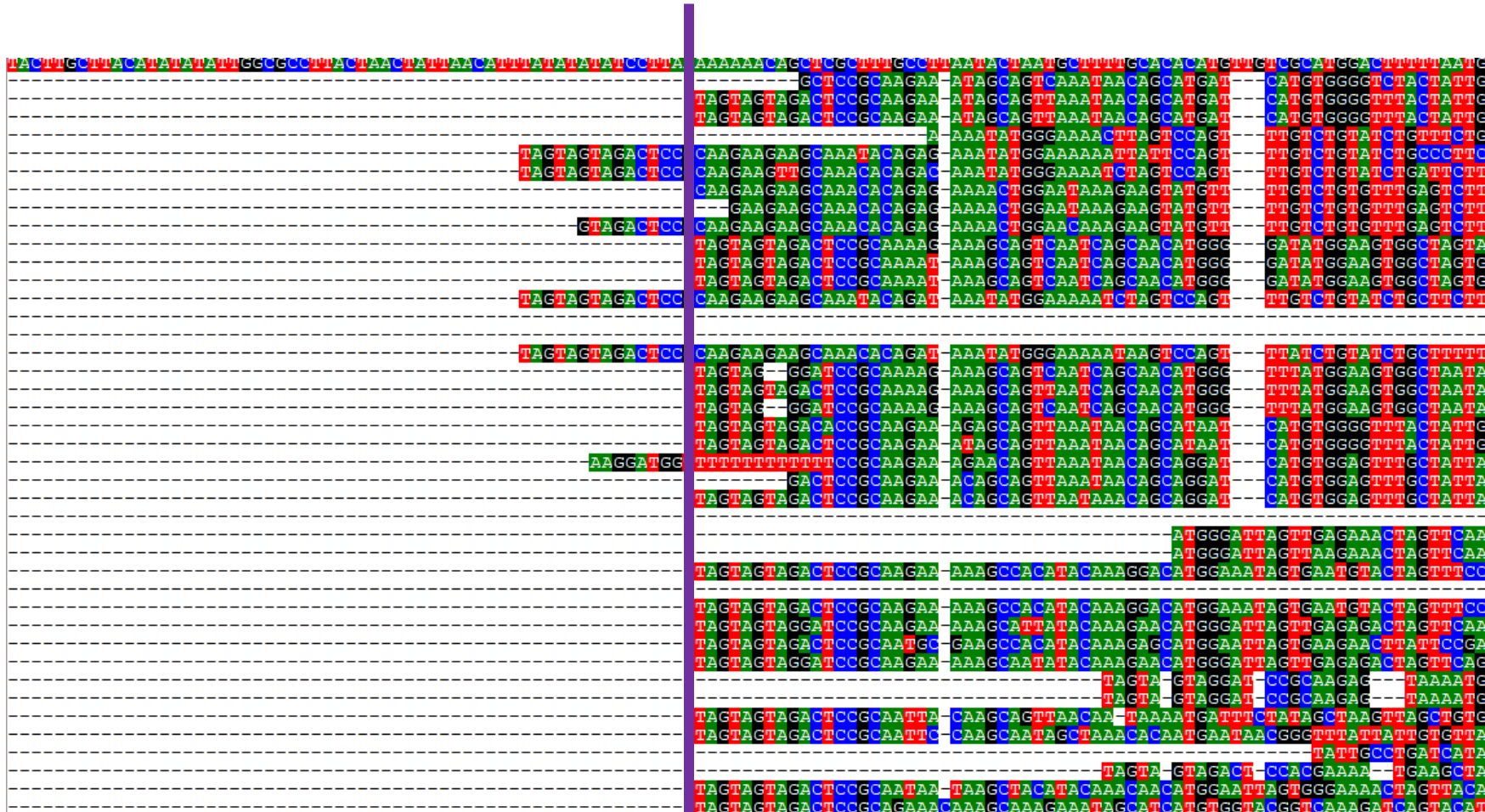
Iterative refinement methods

- FFT-NS-i (Slow; iterative refinement method)
- E-INS-i (Very slow; recommended for <200 sequences with multiple conserved domains and long gaps; **2 iterative cycles only**) [Help](#) [Updated](#) (2015/Jun)
- L-INS-i (Very slow; recommended for <200 sequences with one conserved domain and long gaps; **2 iterative cycles only**) [Help](#)
- G-INS-i (Very slow; recommended for <200 sequences with global homology; **2 iterative cycles only**) [Help](#)
- Q-INS-i (Extremely slow; secondary structure of RNA is considered; recommended for a global alignment of highly divergent ncRNAs with <200 sequences \times <1,000 nucleotides; the number of iterative cycles is restricted to two, 2016/May) [Help](#)

<https://mafft.cbrc.jp/alignment/server/index.html>

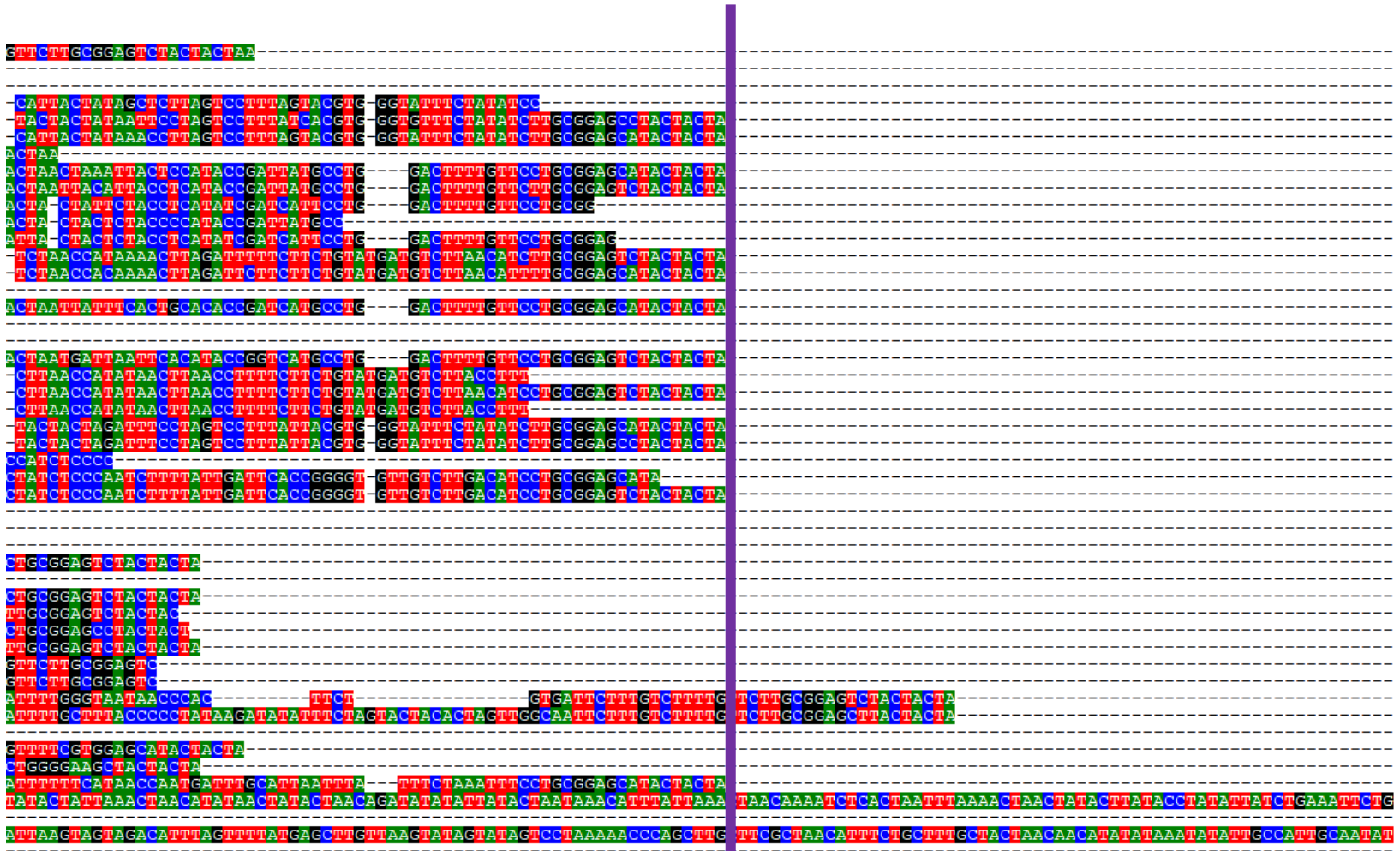
Alignment

→ ořezání (trimování) alignmentu na jednotnou délku (začátek i konec)



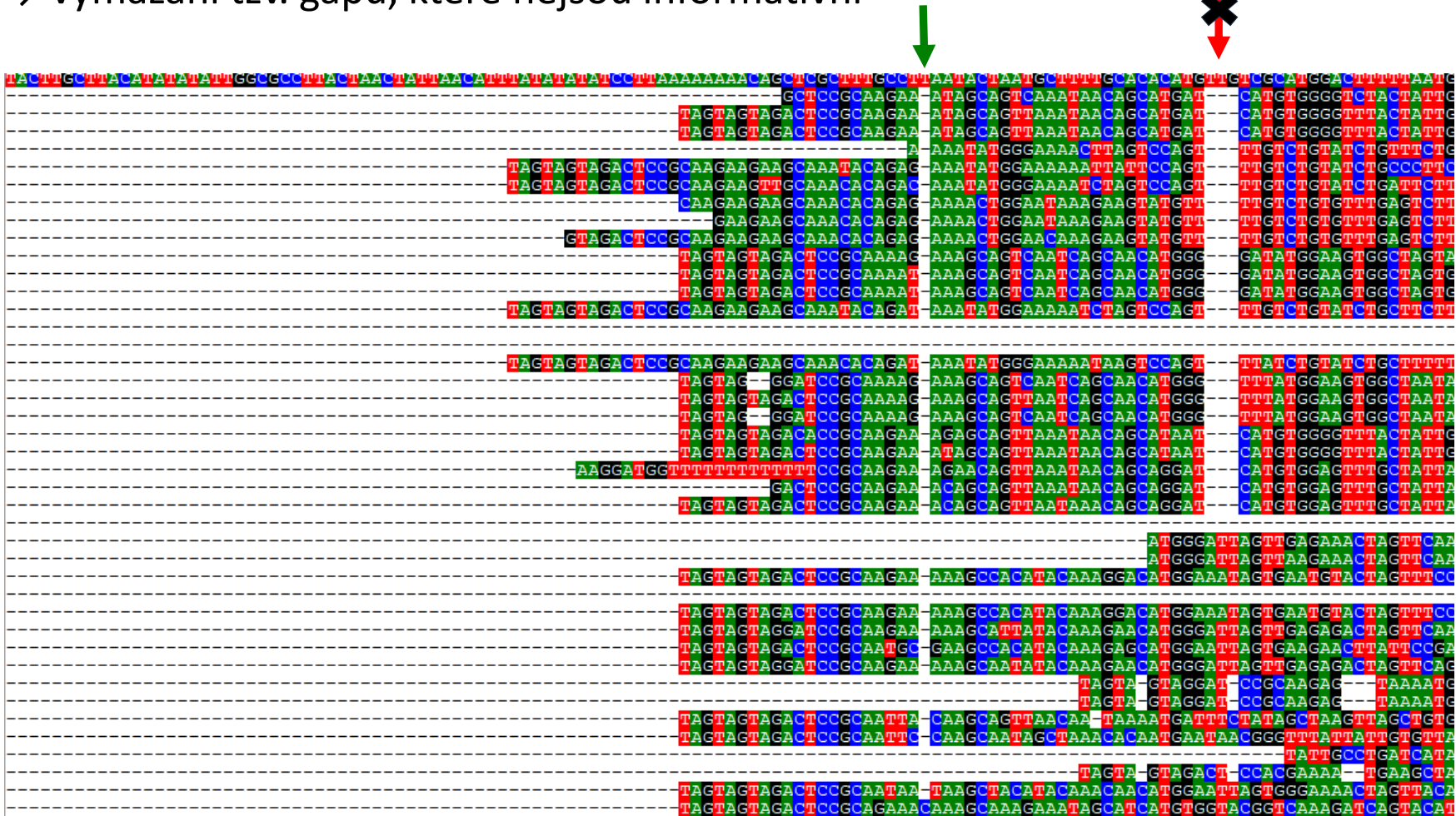
Alignment

→ ořezání (trimování) alignmentu na jednotnou délku (začátek i konec)



Alignment

→ vymazání tzv. gapů, které nejsou informativní



Alignment

Ořezávání alignmentu:

Geny nekódující protein – ořezáváme v nukleotidovém formátu

Geny kódující protein – ořezáváme v aminokyselinovém formátu
(aby nedošlo k porušení čtecího rámce)

```
ATAAACAAGGATTAGAAAGAACTATAGGAAAGCTTTATTGTTTGCTGGGGTACCTAA
ACAAACAAGACTAGAAAGAACTATCGGAAAGGCTCTGTTATTGCTGGAGTACCGAA
ACAATAAAGATAAAGAGCAGCCCAATTGGTCTGGCTTAATGGCTGGTGTCCCAA
ACAATAAAGATAGGGAAAGCCCAATAGGTTAGGCTCTTTAATGGCTGGGTACCCAA
ACAATAAAGATAAAGAAACAACAATTGGTCTGGCTTATTAATGGCTGGTGTCCCAA
ACAACAAGACTAGAAAGCAGCCCTATAGGCCATGCTTGGCTGATGGCAGGAGTACCGAA
ACAACAAGATAAGGAGCAGCCCAATAGGCCATGCTTGGCTGATGGCTGGGGTACCAA
ACAACAAGATAAGGAGCAGCCCAATAGGCAATGCTTTGCTGATGGCTGGGGTACCAA
ATAATAAAGATTAGAAAGCAATAGGAAAGTGCCTTATTATTGCAAGGGGTCCCTTC
ATAATAAAGATTAGAAAGCAATAGGAAAGTCTATTATTGCAAGGGGTCCCTTC
ACAATAAAGACAGGGAAAGCCCTATGGCCATGCTGGCTGTTGATGGCTGGGGTACCAA
ATAATAAAGATAGAGAAACAACCTATTGGTCAATGATATTATAATGGCAGGTGTTCAGAA
ATAATAAAGACAAGGAAAGCCCAATGGTCAATGCTTATTGATGGCTGGAGTCCCAA
ACAATAAAGACAAGAAAGCAGCCCTATGGTCAATGCTGTTAATGGCTGGGGTCCCAA
ATAACAAGATAAGAAAGCAATAGGCAAGTATTATTATTGCAAGGTGTCCCTTC
ATAACAAGATTAGAGAAACAATAGGAAAGTGCCTTATTATTGCAAGGATTCCTTC
ATAACAAGACTAGAAAGCAATAGGAAAGTCTATTATTGTTACAGGTTATCCCTTC
ATAACAAGATTAGAAAGAACTATAGGAAAGCTTATTATTATTGCGCGAGTCCCAA
ATAACAAGATTAGAAAGAACTATAGGAAAGCTTATTATTATTGCGCGAGTCCCAA
ACAATAAAGACAGTAGGAGCCCAATAGGTAAGGTTGCTATTGCTGGAAATACCTAA
ACAATAAAGACAGTAGGAGCCCAATAGGTAAGGTTGCTATTGCTGGAAATACCTAA
ACAATAAAGACAGTAGGAGCCCAATAGGTAAGGTTGCTATTGCTGGAAATACCTAA
```



```
GLMETIRQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSEVIKT
GLMEETRRVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
GLMEETRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
GLMEETRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
ELVQTRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
ELVQTRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
ELVQTRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
ELVQTRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
EMVQTRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
EMVQTRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
KLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
KLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
KLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
KLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
GLMEETRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
GLMEETRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
GLMEETRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
GLMEETRRKVQARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLICIQALEKALRWASGSESHEL
GLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
GLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
GLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
GLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
GLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
GLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
GLMEETRRQKAMARIVRKQRTEADRGFFITLPTRVLEIEDVDAISKNVEENISYGGERKLIAIQALEKALRWASGSESHEL
```

Alignment

The screenshot shows a sequence alignment software window titled "C:\Users\Kvicej00\Desktop\Lseg_FIN.phy". The window displays 83 total sequences. The main area shows a multiple sequence alignment of protein sequences. The sequences are color-coded by amino acid type: Cysteine (yellow), Aspartic acid (green), Glutamic acid (light green), Asparagine (cyan), Glutamine (light blue), Lysine (blue), Arginine (dark blue), Histidine (purple), Proline (dark purple), Glycine (grey), Alanine (light grey), Valine (orange), Leucine (light orange), Isoleucine (red-orange), Methionine (red), Phenylalanine (pink), Tyrosine (light pink), Serine (light red), Threonine (red), and Tryptophan (dark red). The alignment is shown against a ruler at the top, with positions ranging from 810 to 940. The sequences are aligned to a reference sequence, and the alignment is shown with gaps (dashes) where the sequences do not match. The interface includes a menu bar with options like "Select / Slide", "Mode", "Selection: 0", "Position: 53: JQ425313", "Sequence Mask: None", "Numbering Mask: None", and "Start ruler at: 1". There is also a toolbar with various icons for editing and viewing the alignment, and a scroll bar at the bottom.

Alignment

Ukládání alignmentu:

- v různých formátech, v závislosti na účelu dalšího použití

FASTA (.fas) – možnost otevření jak v programu alignmentu, tak v textové podobě (Poznámkový blok)
– možnost mazání či přidávání nových sekvencí do již zalignovaného souboru (→ usnadnění práce)

PHYLIP (.phy) – pro maximum likelihood analýzy (PhyML, RAxML apod.)

NEXUS (.nex) – pro Bayesovskou analýzu (MrBayes) a analýzy maximum parsimony (PAUP)

2. Praktická úloha – alignment

Úloha – tvorba a úprava alignmentu v programu BioEdit:

- načíst dataset, který mám uložený jako .txt, do programu BioEdit
- Accessory Application – ClustalW Multiple alignment
- Run ClustalW – OK
 - načte sekvence a porovnává každou s každým, řadí dle podobnosti
- převod do aminokyselin pomocí Ctrl+T, zpět do nukleotidů také Ctrl+T
- ořezat začátek a konec alignmentu na vhodnou délku
- zkontrolovat alignment, vymazat gapy
- uložit alignment ve formátech FASTA a PHYLIP (File – Save As)

Fylogenetické analýzy

Fylogenetická analýza = klasifikace organismů založená na evoluční historii

Účel:

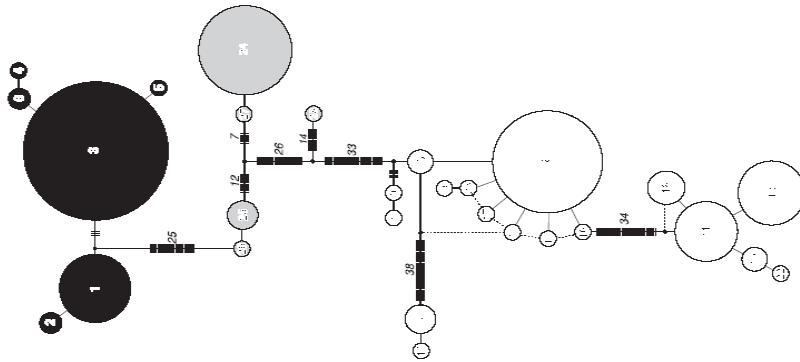
- zjišťování příbuznosti a vztahů mezi organismy / taxony
- evoluční interpretace fylogenetických vztahů
- taxonomie
- koevoluce

→ pro pozorovaná data hledáme adekvátní vysvětlení

Výstup:

grafické zobrazení pomocí **fylogenetického stromu (fylogramu)**

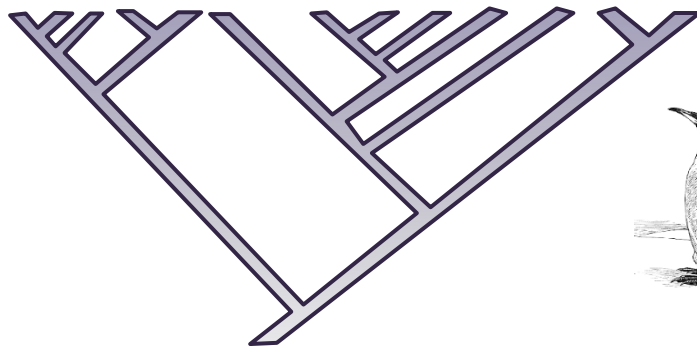
Fylogenetické analýzy



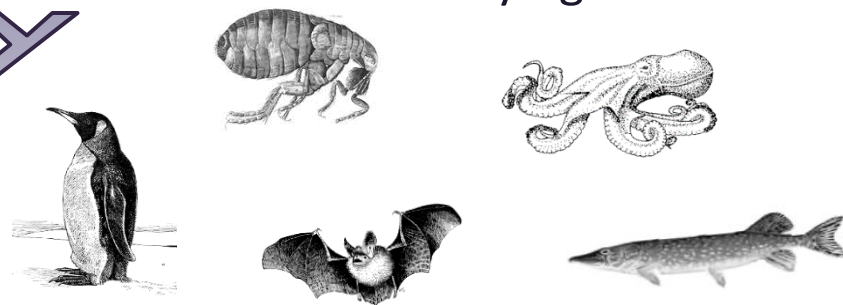
vnitrodruhové vztahy
(populační genetika)



mezidruhové vztahy

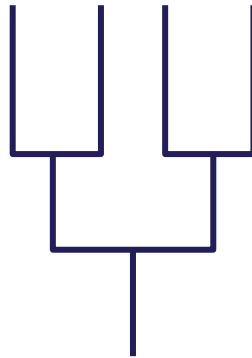


fylogenetické vztahy



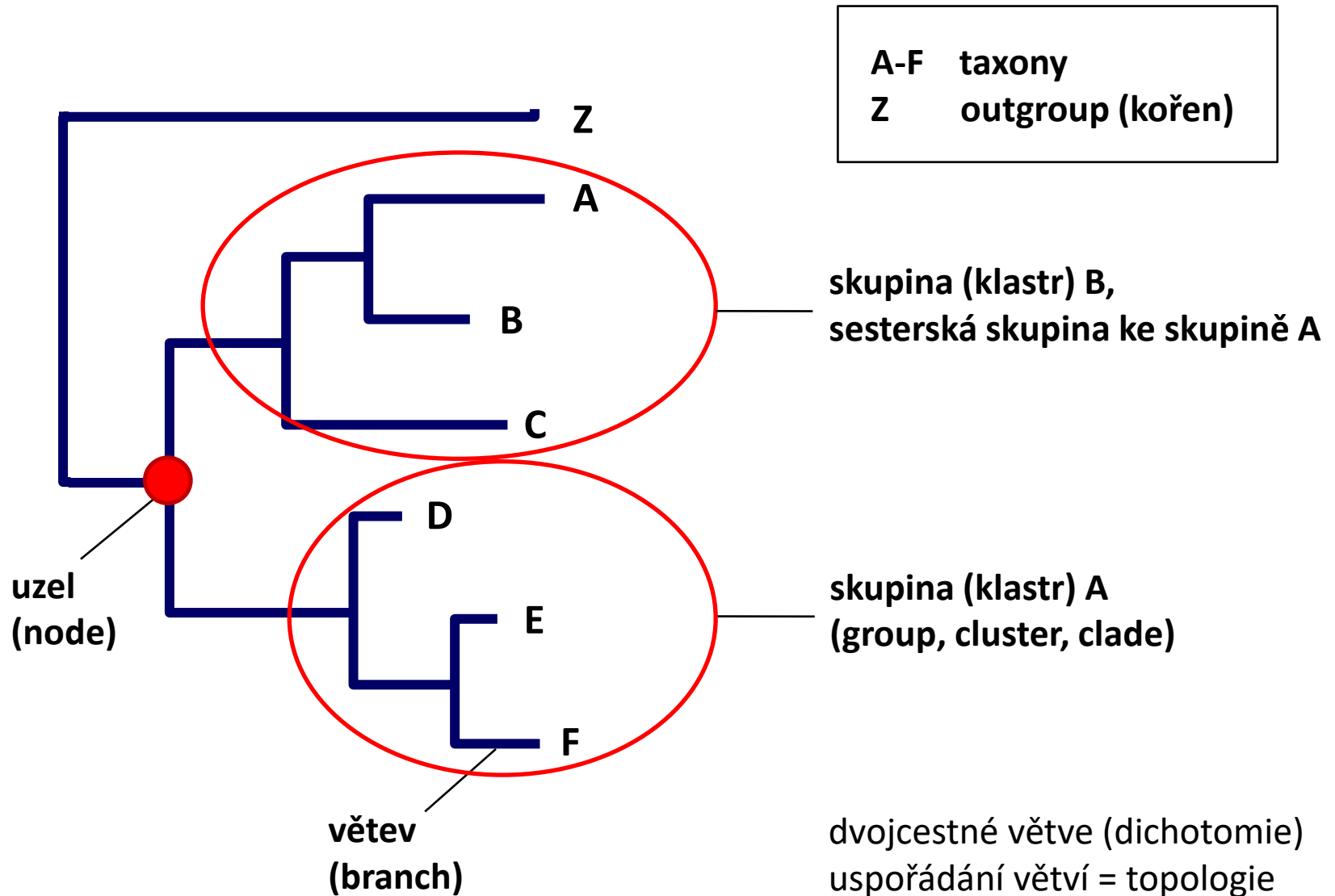
Fylogenetické analýzy

GAATCATCCC
GACCAAACCTA
GAATCATCCC
GACCAAACCTA



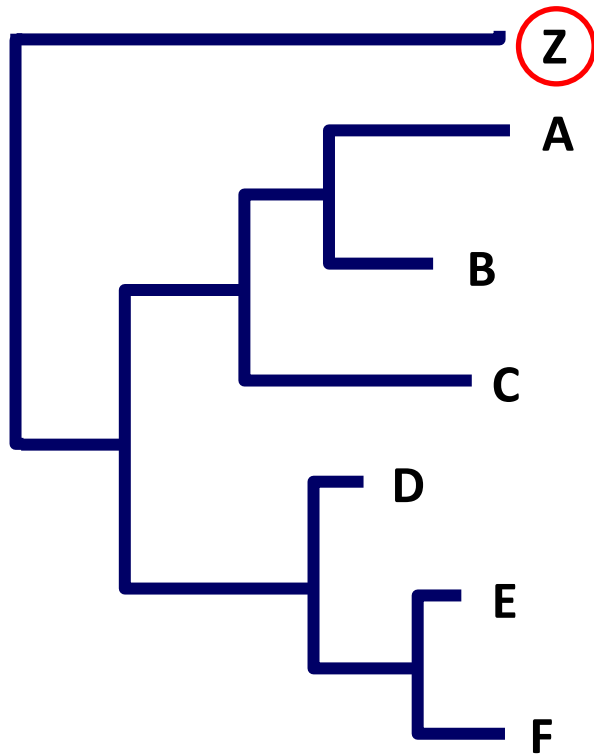
?

Fylogenetický strom

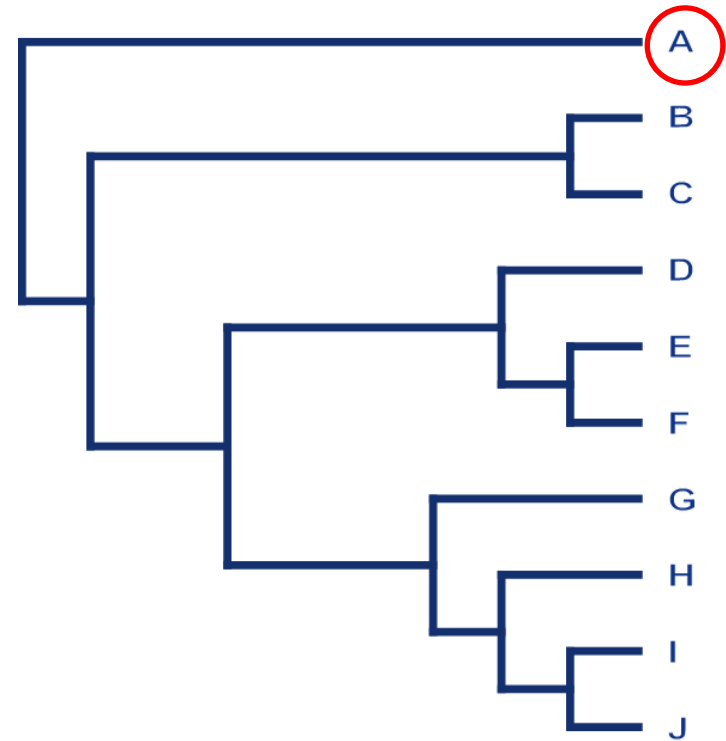


Zobrazení fylogenetických stromů

ZAKOŘENĚNÝ,
S DÉLKOU VĚTVÍ

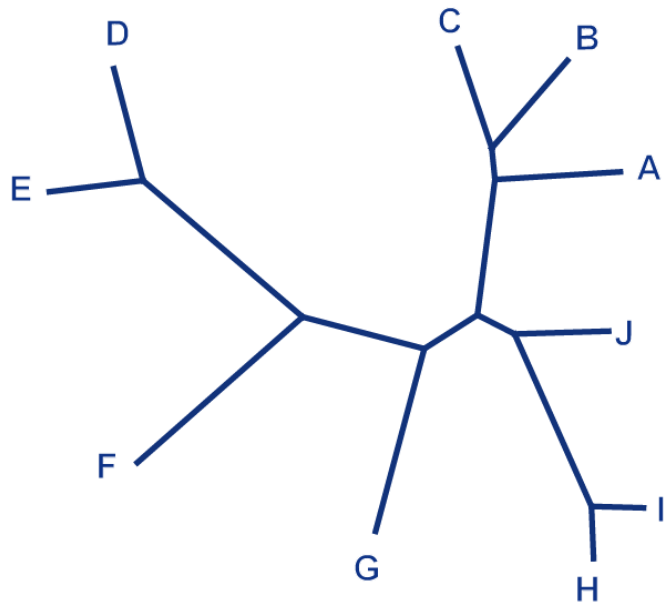


ZAKOŘENĚNÝ,
BEZ DÉLKY VĚTVÍ



Zobrazení fylogenetických stromů

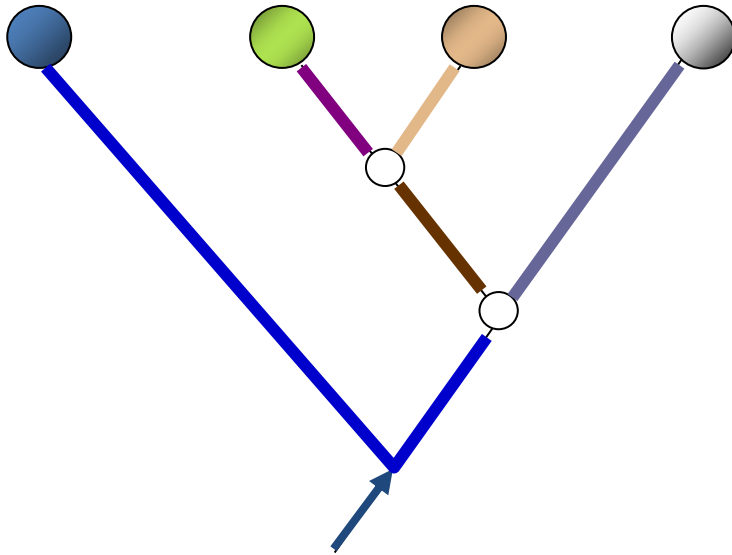
NEZAKOŘENĚNÝ



NEWICK FORMÁT

```
(A ((B C) ((D (E F))(G (H (I J))))))
```

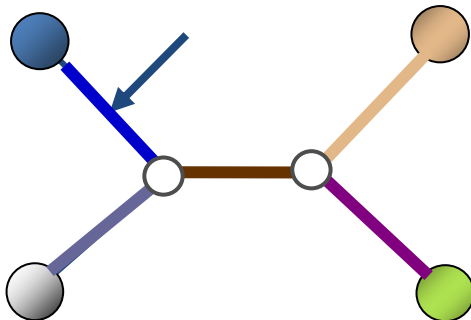
Fylogenetický strom



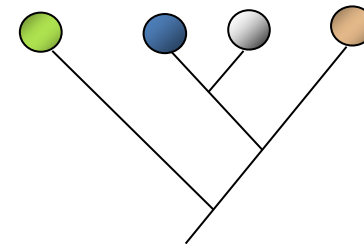
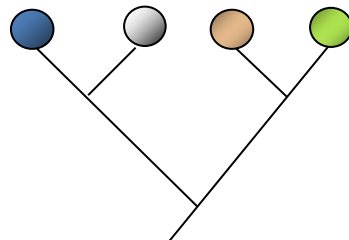
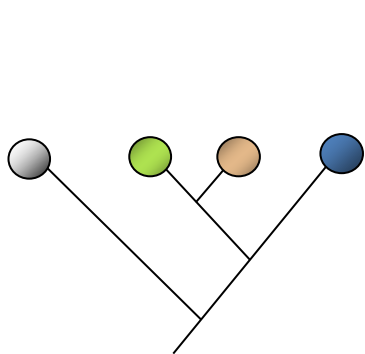
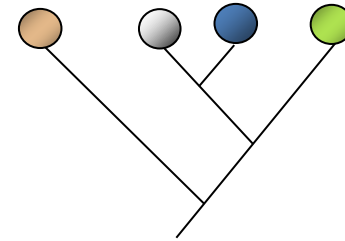
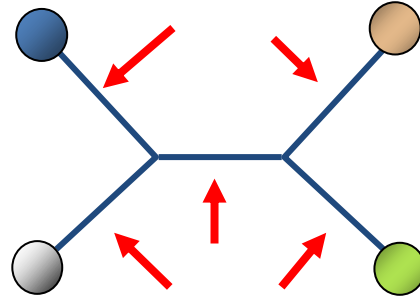
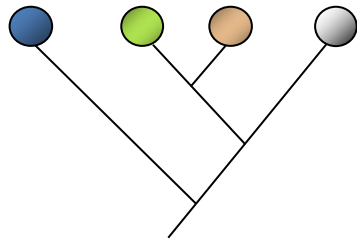
n počet taxonů ve stromu

$n-2$ uzlů

$2n-3$ větví

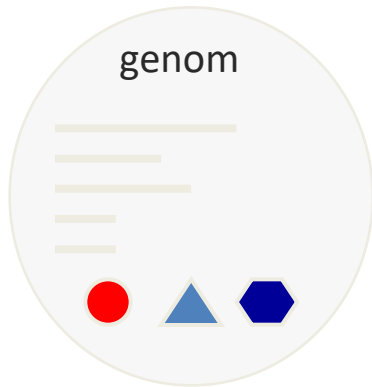


Fylogenetický strom



zakořeněných stromů
je více než nezakořeněných

Zdroje fylogenetické informace



gen A

gen B

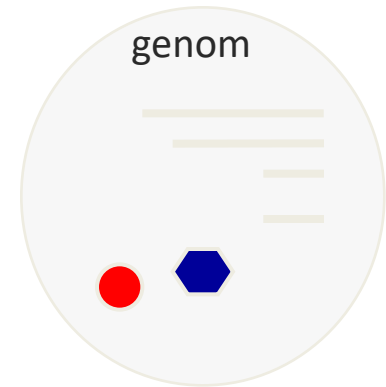
gen C

amplikonová data z NGS

„multilokusové“ informace
(RFLP, RAPD, AFLP,
mikrosatelity)

pořadí genů

sekvence nukleotidů



gen A

gen C

gen B

GAATCATCCGGACCAAACTTA

GAATCATCCCGACCAAACTTA



Fylogenetická data

Požadavky na fylogenetická data:

- informativnost (tj. přiměřená variabilita mezi taxony)
- nezávislost
- homologie
- dostatečné množství

```
GACACGGTCCAGACTCTTACGGGAGGCA GCA GTGGGGAATATTGCACAATGGGC GAAA
GACACGGTCCAGACTCCTACGGGAGGCA GCA GTGGGGAATATTGCACAATGGGC GAAA
GATACGGTCCAGACTCCTACGGGAGGCA GCA GTGGGGAATATTGCACAATGGGC GAAA
GACACGGTCCAGACTCCTACGGGAGGCA GCA GTGGGGAATATTGCACAATGGGC TAAA
GACACGGTCCAGACTCCTACGGGAGGCA GCA GTGGGGAATATTGCACAATGGGC GAAA
GATACGGTCCAGACTCCTACGGGAGGCA GCA GTGGGGAATCTTGCACAATGGGC GAAA
GATACGGTCCAGACTCCTACGGGAGGCA GCA GTGGGGAATATTGCACAATGGGGGAAA
GACACGGTCCAGACTCCTACGGGAGGCA GCA GTGGGGAATATTGCACAATGGGGGAAA
GACACGGTCCAGACTCCTACGGGAGGCA GCA GTGGGGAATATTGCACAATGGGGGAAA
```

```
GACGGGTGAGTAAAGCGTGGGAATCTGCCTTGCA GTGGGGATAACCCGGGGAAACT
GACGGGTGAGTAAAGCCTAGGAAATTGCCCTGAT GTGGGGATAACCAATTGAAACG
GACGGGTGAGTAAAGCTTGGGAATCTAGCTTATGGA GGGGGATAACTACGGGAACT
GACGGGTGAGTAAAGTCTGGGATCTGCCCGATGGA GGGGGATAACTACTGAAACG
AACGGGTGAGTAAAGTCTGGGATCTGCCCGATGGA GGGGGATAACTATTGAAACG
AACGGGTGAGTAAATCTAGGGATCTACCTAATGGA GGGGATAACTATTGAAACG
GACGGGTGAGTAA GATATGGGAATCTACCTAAAGAT GGGGGATAACTATCGGAAACG
```

```
CA-----AAGG-----TC-----TTCGGA-TT GAGTAGCGT
CAGACA-----GAGGAACTTGTTCTTGGTGG-----CGAGCG-
TAAC-----ATGAAGA-----AGCTTGCT-----TCTTTG-ATGACGAGTG-
CA-----CGGAAA-----GAAGCTTG-----CTTCTTTG-CCGGCGAGCG-
-----TTTATG-----CAGCTCTG-CTGGCAAGCG-
TA-----ACAAAAA-----TA-----TTTTTTG-TT-----AAGCG-
AATG-----ATGAAAA-----TATTAGCG-
-----TTTAGCG-
```


Fylogenetická analýza

Několik přístupů (algoritmů):

→ pro pozorovaná data hledáme adekvátní vysvětlení

- **parsimonie** → co nejúspornější uspořádání dat

MP (Maximum parsimony) – počet substitucí

- **pravděpodobnostní přístup** → co nejpravděpodobnější uspořádání dat na základě daných předpokladů

ML (Maximum likelihood) – počet substitucí „na pozici“

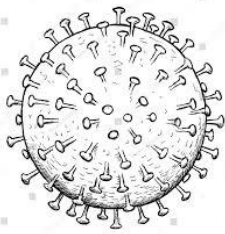
BI (Bayesian inference)

je nutné „znát“ (stanovit) model, jak probíhá evoluce

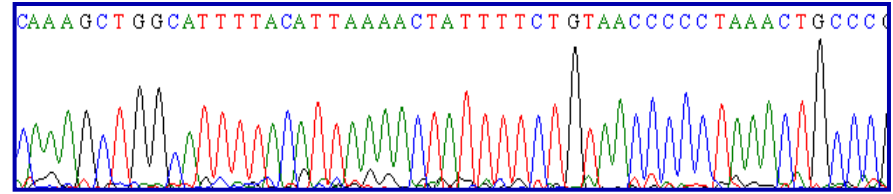
→ výběr vhodného modelu molekulární evoluce

Evoluční modely:

→ korekce na opakované substituce

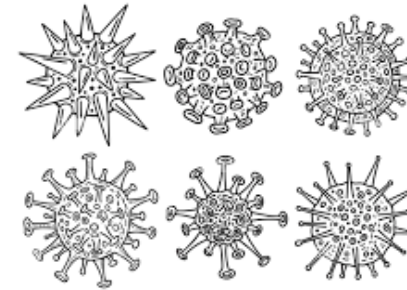
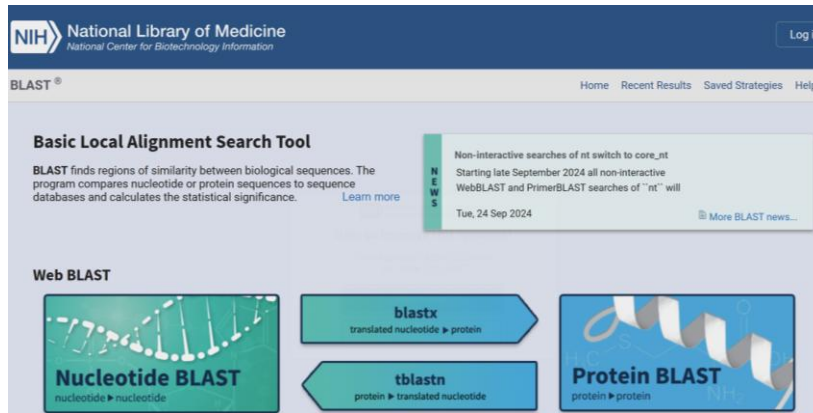


Izolace DNA / RNA, PCR, sekvenování



Organismus, který chci fylogeneticky charakterizovat (měl bych vědět proč)

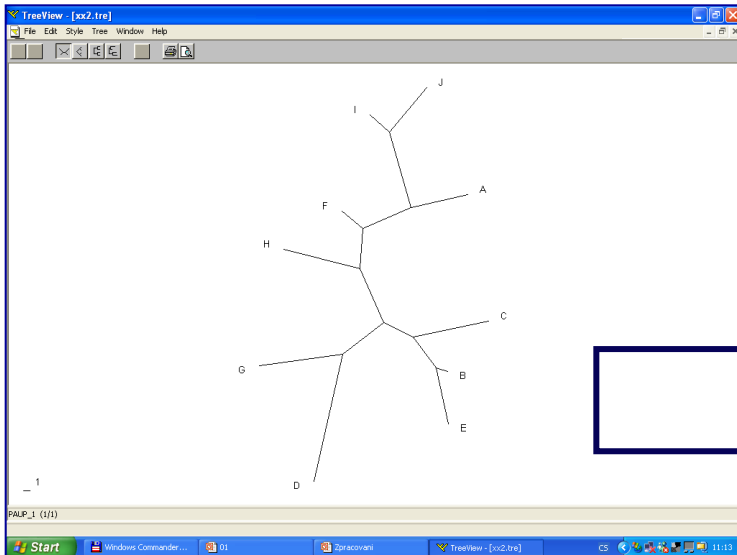
Vyhledání homologických sekvencí pro další taxony (měl bych vědět, pro které)



Vytvoření alignmentu

```
GTGGGAAGGAAA- - - CCTGGTGGTTAATA- - - CCCA
GTCGTGAGGAAGG- - - TGGTGTGTTAATAGCAGCA
BTAGCGAGGAAGG- - - CATTAGTTTAATAGACTAG
GCGGGGAGGAAGG- - - CGTGAGAGCGAATACCTTTC
GTAGGGAGGAAGGC- - - AA- TATCCTTAATACGGTTA
```

Fylogenetická analýza zvolenou metodou



To hlavní a podstatné:
interpretace fylogenetických vztahů

Maximální parsimonie (MP)

Maximální parsimonie (Maximum parsimony, MP)

- pracuje s délkou stromu a hledá nejkratší kladogram (tj. co nejúspornější uspořádání dat)

Znaky (pozice alignmentu):

- neinformativní (konstantní)
- variabilní neinformativní
- **variabilní informativní**

→ délka stromu je množstvím všech změn na všech pozicích alignmentu

→ sestavení všech stromů

Délka stromu

Počet pozic

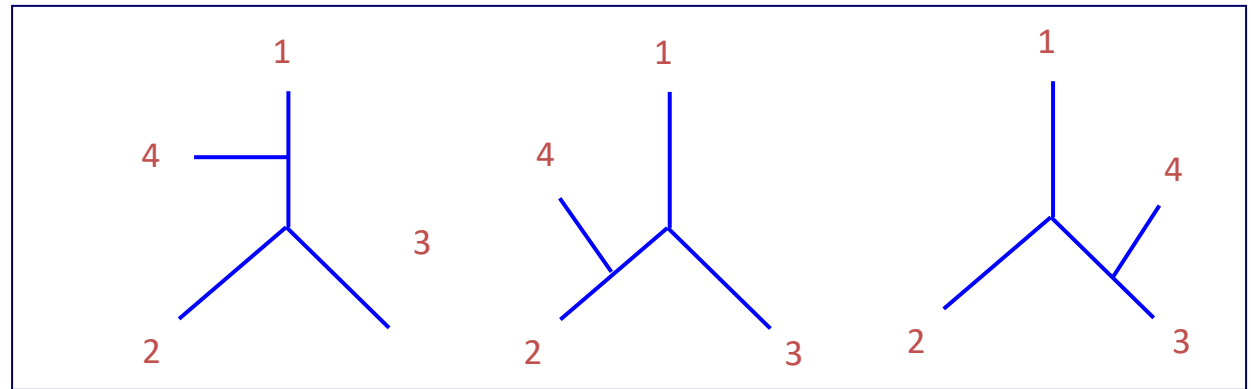
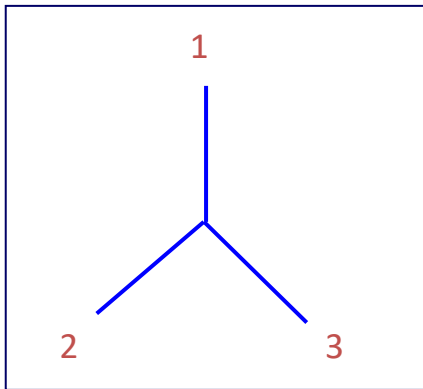
$$L = \sum_{i=1}^k l_i$$

Délka stromu pro danou pozici

Maximální parsimonie (MP)

Například:

- Výchozí strom z libovolných 3 taxonů
- Sestavení tří možných stromů s přidáním 4. taxonu
- Z každého stromu vzniklého v předešlém kroku se postaví pět dalších stromů přidáním 5. taxonu
- Pokračujeme až do přidání posledního taxonu



3 taxony → 1 nezakořeněný strom

4 taxony → 3 nezakořeněné stromy

→ a pak to prudce roste...

počet možných kombinací roste jako faktoriál n

Maximální parsimonie (MP)

No. taxa	Number of unrooted trees	Number of rooted trees
	1	1
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

Maximální parsimonie (MP)

Maximální parsimonie (Maximum parsimony, MP)

- sestavení co nejkratšího stromu (výchozího)
- jeho následné zlepšování pomocí přeskupování větví (branch swapping)
- heuristické metody (různé algoritmy)

Výhody MP:

- jednoduchá, pochopitelná, rychlá
- minimální množství předpokladů o evoluci
- dobře prostudována matematicky

Problémy MP:

- předpoklad parsimonie je zcela jistě nesprávný pro sekvence s rychlou evolucí

Pravděpodobnostní metody

- pojem pravděpodobnost nelze definovat, je značně subjektivní

R.A. Fischer (evoluční biolog) → likelihood pro biologická data

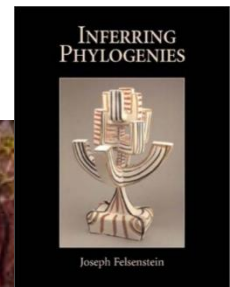
Thomas Bayes (matematik) → pravděpodobnost

pravděpodobnost (probability) ≠ věrohodnost (likelihood)

Joseph Felsenstein – přišel na to, že Fischerův přístup by šel napasovat na fylogenetiku (kniha „Inferring Phylogenies“)

máme data (tj. pozorovaný jev) → chceme ho adekvátně vysvětlit (hypotéza)

→ pravděpodobnost mezi daty a fylogenezí



Pravděpodobnostní metody

**pravděpodobnost
(probability)**

vs.

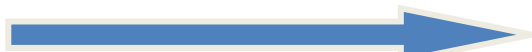
**věrohodnost
(likelihood)**

vysvětlení (*hypotéza*)

pozorovaný jev (*data*)

- někdo skládal mince aby to bylo hezké, a takhle to dopadlo.
- roztrhl se pytlík s mincemi a takhle to padlo náhodu
- spodní strana mince je těžší a pod stolem jsou magnety

probability($H | D$)



že tato hypotéza způsobila tato data

probability($D | H$)



že se objeví takováto data, pokud nastane předpokládaná hypotéza

likelihood



Pravděpodobnostní metody

Fylogenetika → hledáme hypotézu s maximální věrohodností

v likelihoodu jsou délky větví velmi důležité

Hypotéza = topologie vč. délky větví a modelu

→ vezmu data, prohledávám topologie, délky větví v každé topologii
na každé úrovni najdu nejvěrohodnější a z nich vyberu ten nejvěrohodnější
(na nejvyšší úrovni)

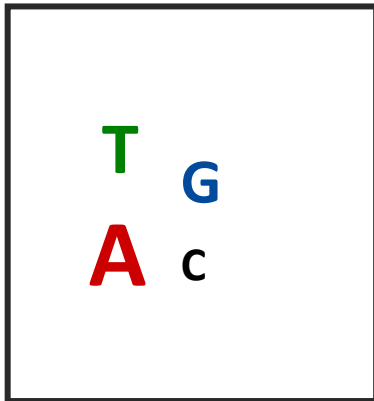
Model:

- evoluce na jednotlivých pozicích je nezávislá
- evoluce v jednotlivých liniích je nezávislá

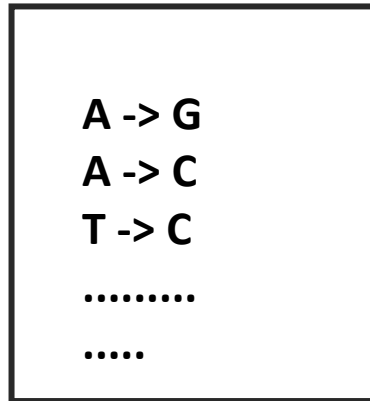
G	A	C	T	C	A	T	C	C . . . m
G	C	A	T	C	A	T	T	C
G	C	A	T	C	A	T	G	C
G	C	A	T	T	A	T	T	C
G	G	A	T	C	A	T	T	C

Evoluční modely

frekvence nukleotidů
(base composition)



transition
probabilities



rychlost substitucí
na různých pozicích matice
(distribution of rates)



Výběr vhodného evolučního modelu:

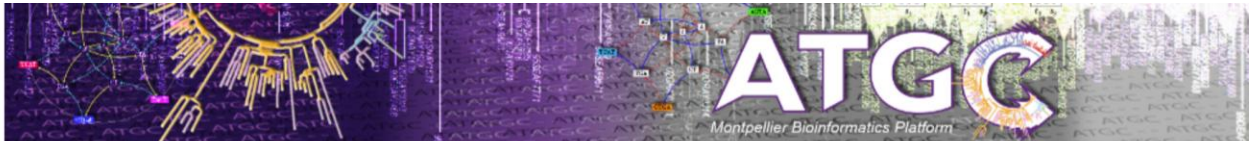
různé volně stažitelné nebo online běžící programy, např.

SMS: Smart Model Selection

www.atgc-montpellier.fr/sms/

Evoluční modely

SMS: Smart Model Selection



SMS: Smart Model Selection in PhyML

Vincent Lefort, Jean-Emmanuel Longueville, Olivier Gascuel
Molecular Biology and Evolution, 34(9):2422-2424, 2017.
[Access the recommendation on F1000Prime](#)

SMS online execution

Input Data

Input data (PHYLIP format) Nevybrán žádný soubor

Data type **Protein** **DNA**






Selection criterion

Name of your analysis

Your email

Please confirm your email

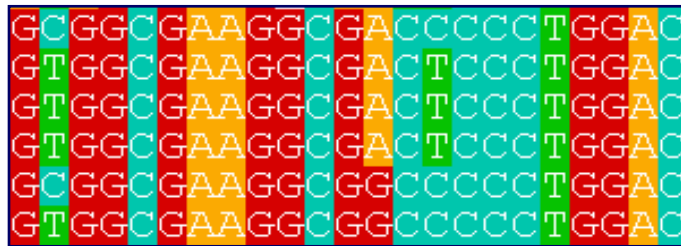
Contact: [Webmaster](#), LIRMM.



Maximální věrohodnost (ML)

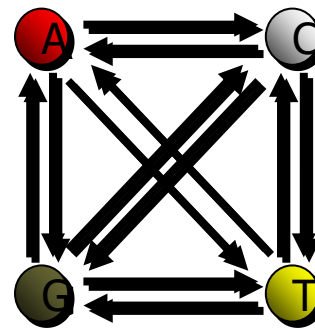
Maximální věrohodnost (Maximum likelihood, ML)

→ hledání hypotézy s maximální věrohodností



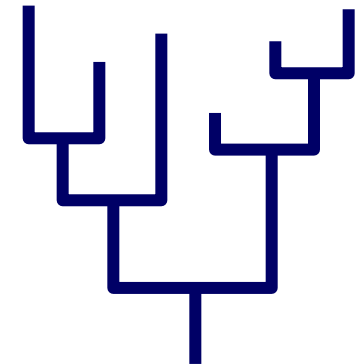
matice

data



model

hypotéza



topologie
délky větví)

(včetně

Maximální věrohodnost (ML)

Maximální věrohodnost (Maximum likelihood, ML)

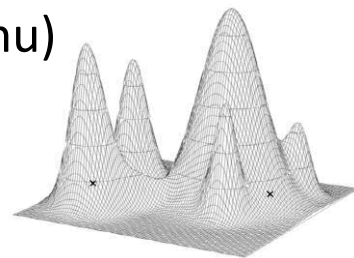
- hledání hypotézy s maximální věrohodností
- evoluce na jednotlivých pozicích je nezávislá
- evoluce v jednotlivých liniích je nezávislá

Výhody ML:

- jednoznačně lepší, pravděpodobnější stromy než MP

Nevýhody ML:

- hledání ML stromu je „optimality criterion“ proces – stejný problém jako u MP
- nastavení parametrů analýzy, výběr vhodného modelu
- výpočetně náročná (hledání nejlepší kombinace modelu a stromu)
- časově náročná



Maximální věrohodnost (ML)

→ různé programy

- volně stažitelné, online běžící, součást balíčků (např. Geneious)

PhyML – zdarma, běží poměrně rychle a efektivně

RAxML

→ načítá soubory ve formátu PHYLIP (.phy)

Maximální věrohodnost (ML)

PhyML

```
C:\Users\kvic00\Desktop\pt x + v
- PHYML v2.4.3 -

Settings for this run:

D          Data type (DNA/AA)      DNA
I      Input sequences interleaved (or sequential)  interleaved
S          Analyze multiple data sets      no
B      Non parametric bootstrap analysis      no
M      Model of nucleotide substitution      HKY
E      Base frequency estimates (empirical/ML)  empirical
T          Ts/tv ratio (fixed/estimated)  fixed (ts/tv = 4.00)
V      Proportion of invariable sites (fixed/estimated)  fixed (p-invar = 0.00)
R      One category of substitution rate (yes/no)  yes
U          Input tree (BIONJ/user tree)      BIONJ
O          Optimise tree topology      yes

Are these settings correct? (type Y or letter for one to change) |
```

Maximální věrohodnost (ML)

PhyML

```
C:\Users\kvicaj00\Desktop\pt x + v
- PHYML v2.4.3 -

Settings for this run:

D          Data type (DNA/AA)      DNA
I      Input sequences interleaved (or sequential)  interleaved
S          Analyze multiple data sets      no
B      Non parametric bootstrap analysis  yes (1000 replicates)
M          Model of nucleotide substitution      GTR
E      Base frequency estimates (empirical/ML)  empirical
V      Proportion of invariable sites (fixed/estimated)  estimated
R          One category of substitution rate (yes/no)  no
C          Number of substitution rate categories      4
A      Gamma distribution parameter (fixed/estimated)  estimated
U          Input tree (BIONJ/user tree)      BIONJ
O          Optimise tree topology      yes

Are these settings correct? (type Y or letter for one to change) |
```

Maximální věrohodnost (ML)

PhyML

```
C:\Users\kvícej00\Desktop\pl x + v
. Optimisation of the proportion of invariable sites...
. Optimisation of the gamma shape parameter...
. Log(lik) : -7267.916468 -> -7267.402566 0 swap done
. Log(lik) : -7267.402566 -> -7267.177048 0 swap done
. Log(lik) : -7267.177048 -> -7267.040841 0 swap done
. Log(lik) : -7267.040841 -> -7266.952651 0 swap done
. Optimisation of the GTR parameters...
. Optimisation of the proportion of invariable sites...
. Optimisation of the gamma shape parameter...
. Log(lik) : -7263.886298 -> -7263.436597 1 swap done
. Log(lik) : -7263.436597 -> -7262.615926 1 swap done
. Log(lik) : -7262.615926 -> -7262.415365 0 swap done
. Log(lik) : -7262.415365 -> -7262.294920 0 swap done
. Optimisation of the GTR parameters...
. Optimisation of the proportion of invariable sites...
. Optimisation of the gamma shape parameter...
. Log(lik) : -7261.013053 -> -7260.651473 0 swap done
. Log(lik) : -7260.651473 -> -7260.418495 0 swap done
. Log(lik) : -7260.418495 -> -7260.264279 0 swap done
```

Maximální věrohodnost (ML)

PhyML

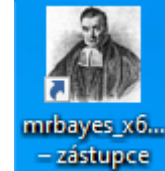
```
C:\Users\kvickej00\Desktop\ph... x + v - □ ×
. Log(lk) : -7252.678381 -> -7252.678380
. Optimisation of the GTR parameters...
. Optimisation of the proportion of invariable sites...
. Optimisation of the gamma shape parameter...
. Log(lk) : -7252.678380 -> -7252.678378
. Optimisation of the GTR parameters...
. Optimisation of the proportion of invariable sites...
. Optimisation of the gamma shape parameter...
. Log(lk) : -7252.678378 -> -7252.678378
. Optimisation of the GTR parameters...
. Optimisation of the proportion of invariable sites...
. Optimisation of the gamma shape parameter...

. Non parametric bootstrap analysis
[.....] 10/1000
[.....] 20/1000
[.....]
```

Bayesovská analýza (BI)

Bayesovská analýza (Bayesian Inference, BI)

→ načítá soubory ve formátu NEXUS (.nex)

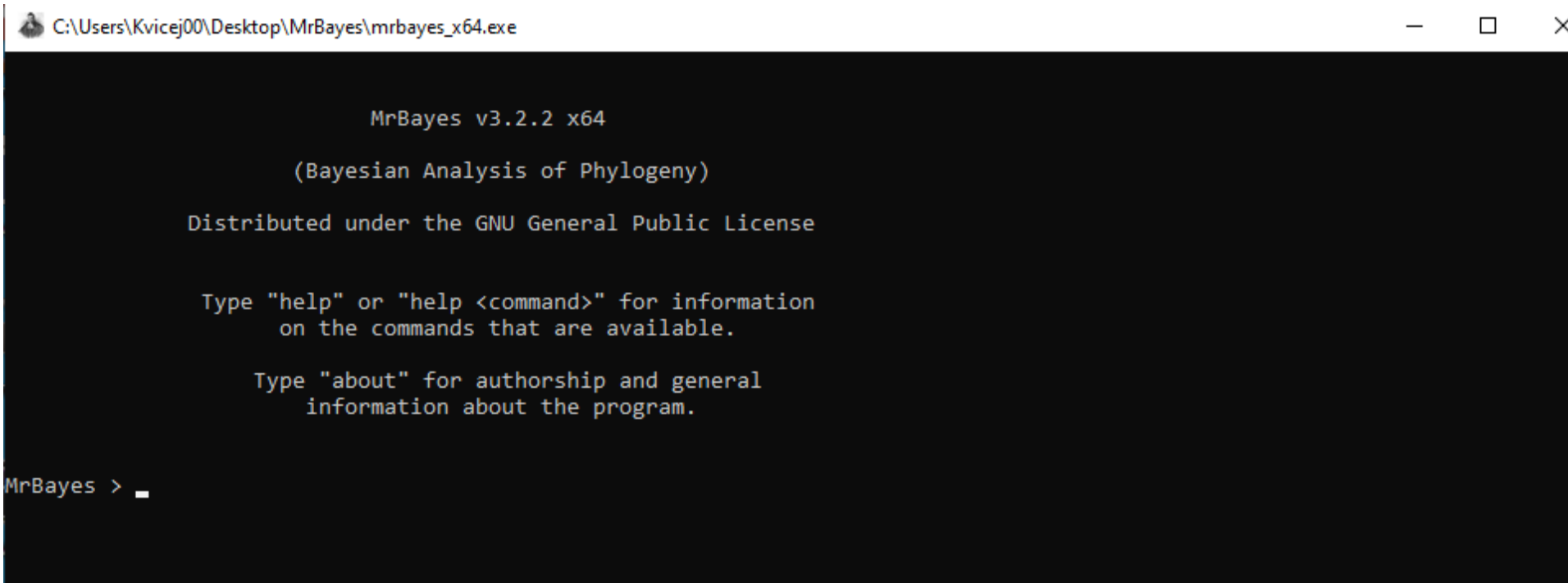


Thomas Bayes



Bayesovská analýza (BI)

Mr. Bayes



```
C:\Users\Kvicej00\Desktop\MrBayes\mrbayes_x64.exe
```

```
MrBayes v3.2.2 x64  
(Bayesian Analysis of Phylogeny)  
Distributed under the GNU General Public License  
  
Type "help" or "help <command>" for information  
on the commands that are available.  
  
Type "about" for authorship and general  
information about the program.  
  
MrBayes > _
```


Bayesovská analýza (BI)

Mr. Bayes

```
(Bayesian Analysis of Phylogeny)

Distributed under the GNU General Public License

Type "help" or "help <command>" for information
on the commands that are available.

Type "about" for authorship and general
information about the program.

MrBayes > execute L_zkus.nex

Executing file "L_zkus.nex"
DOS line termination
Longest line length = 117
Parsing file
Expecting NEXUS formatted file
Reading data block
  Allocated taxon set
  Allocated matrix
  Defining new matrix with 31 taxa and 346 characters
  Data is Dna
  Missing data coded as -
  Taxon 1 -> EU837270H
  Taxon 2 -> JX119008A
  Taxon 3 -> KX064268S
  Taxon 4 -> MK605659D
  Taxon 5 -> AJ410618S
  Taxon 6 -> MK605680K
  Taxon 7 -> MK605661K
  Taxon 8 -> KM192209S
  Taxon 9 -> KU529945S
  Taxon 10 -> JX316008T
  Taxon 11 -> KC522413T
  Taxon 12 -> MK386156T
  Taxon 13 -> KF974361S
  Taxon 14 -> NC043068A
  Taxon 15 -> JX028271M
  Taxon 16 -> KX815401P
  Taxon 17 -> AB574184P
  Taxon 18 -> MF077572P
  Taxon 19 -> MF077567P
  Taxon 20 -> MF077564P
  Taxon 21 -> MN832779P
  Taxon 22 -> KF776606P
  Taxon 23 -> 801F
  Taxon 24 -> 783F
  Taxon 25 -> MN026176P
```

Bayesovská analýza (BI)

Mr. Bayes

```
Taxon 2 -> JX119008A
Taxon 3 -> KX064268S
Taxon 4 -> MK605659D
Taxon 5 -> AJ410618S
Taxon 6 -> MK605680K
Taxon 7 -> MK605661K
Taxon 8 -> KM192209S
Taxon 9 -> KU529945S
Taxon 10 -> JX316008T
Taxon 11 -> KC522413T
Taxon 12 -> MK386156T
Taxon 13 -> KF974361S
Taxon 14 -> NC043068A
Taxon 15 -> JX028271M
Taxon 16 -> KX815401P
Taxon 17 -> AB574184P
Taxon 18 -> MF077572P
Taxon 19 -> MF077567P
Taxon 20 -> MF077564P
Taxon 21 -> MN832779P
Taxon 22 -> KF776606P
Taxon 23 -> 801F
Taxon 24 -> 783F
Taxon 25 -> MN026176P
Taxon 26 -> MF077566P
Taxon 27 -> MN026178P
Taxon 28 -> MN026173P
Taxon 29 -> KF425497D
Taxon 30 -> KF177176D
Taxon 31 -> MK605679D
Successfully read matrix
Setting default partition (does not divide up characters)
Setting model defaults
Seed (for generating default start values) = 1733088771
Setting output file names to "L_zkus.nex.run<i>.<p|t>"
Exiting data block
Reached end of file

MrBayes > lset nst=6 rates=invgamma ngammacat=4

Setting Nst to 6
Setting Rates to Invgamma
Setting Ngammacat to 4
Successfully set likelihood model parameters

MrBayes > mcmc ngen= 1000000
```

Bayesovská analýza (BI)

Mr. Bayes

Chain results (1000000 generations requested):

```
0 -- [-7001.037] (-7423.560) (-7310.681) (-7471.296) * [-7262.191] (-7463.806) (-7017.065) (-7446.010)
500 -- [-4520.078] (-4568.591) (-4567.438) (-4851.601) * (-4710.812) (-4842.278) [-4640.954] (-4755.677) -- 0:00:00
1000 -- [-4191.650] (-4311.517) (-4349.056) (-4403.381) * (-4296.386) (-4245.397) [-4264.047] (-4304.791) -- 2:46:39
1500 -- [-4062.136] (-4172.653) (-4192.854) (-4137.050) * (-4207.548) [-4071.215] (-4170.401) (-4113.387) -- 1:51:05
2000 -- [-3999.329] (-4055.607) (-4131.093) (-4062.699) * (-4147.245) [-3995.171] (-4062.825) (-4011.408) -- 2:46:38
2500 -- (-4004.056) [-3957.566] (-4076.370) (-3994.656) * (-4120.848) [-3953.068] (-3957.554) (-4004.623) -- 2:13:18
3000 -- (-3980.809) [-3932.391] (-4011.878) (-3978.856) * (-4088.130) [-3933.328] (-3931.256) (-3972.958) -- 1:51:04
3500 -- (-3957.155) [-3910.655] (-3971.395) (-3947.220) * (-4044.541) [-3926.653] (-3918.208) (-3925.643) -- 2:22:48
4000 -- [-3921.520] (-3921.043) (-3963.030) (-3919.616) * (-4012.472) (-3911.313) (-3911.143) [-3912.661] -- 2:04:57
4500 -- (-3925.067) (-3903.955) [-3915.394] (-3909.239) * (-3964.335) (-3915.977) [-3906.873] (-3895.661) -- 2:28:04
5000 -- (-3929.571) (-3906.815) [-3898.958] (-3897.281) * (-3966.163) (-3916.646) (-3911.149) [-3888.063] -- 2:13:16
```

Average standard deviation of split frequencies: 0.111959

```
5500 -- (-3930.309) (-3910.432) [-3909.824] (-3911.371) * (-3937.099) (-3904.676) (-3915.333) [-3890.150] -- 2:31:25
6000 -- (-3910.707) (-3916.071) (-3916.499) [-3887.104] * (-3919.871) (-3893.397) (-3908.263) [-3899.397] -- 2:18:48
6500 -- [-3893.685] (-3911.698) (-3903.295) (-3898.103) * (-3922.663) (-3881.209) [-3903.879] (-3891.859) -- 2:33:44
7000 -- (-3895.051) (-3901.029) [-3899.044] (-3897.073) * (-3909.663) [-3886.384] (-3910.816) (-3887.637) -- 2:22:45
7500 -- (-3904.609) (-3903.049) (-3898.748) [-3895.102] * (-3914.970) (-3886.157) (-3917.214) [-3886.752] -- 2:35:26
8000 -- (-3892.106) (-3893.815) [-3904.499] (-3899.277) * (-3911.833) [-3887.747] (-3910.232) (-3903.989) -- 2:25:43
8500 -- [-3902.334] (-3894.355) (-3902.532) (-3888.610) * [-3905.619] (-3888.486) (-3903.637) (-3904.968) -- 2:17:08
9000 -- (-3902.992) [-3889.185] (-3898.782) (-3880.044) * [-3891.998] (-3891.271) (-3888.769) (-3910.201) -- 2:28:00
9500 -- (-3902.893) (-3902.981) (-3897.013) [-3899.451] * (-3889.090) [-3900.555] (-3900.639) (-3906.465) -- 2:20:13
10000 -- (-3895.596) [-3888.290] (-3900.646) (-3906.445) * (-3895.943) [-3892.994] (-3903.430) (-3908.482) -- 2:29:51
```

Average standard deviation of split frequencies: 0.106066

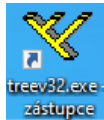
```
10500 -- (-3895.464) (-3896.796) [-3899.754] (-3916.701) * (-3905.663) [-3886.282] (-3925.605) (-3898.088) -- 2:22:42
11000 -- (-3892.916) [-3892.540] (-3898.288) (-3907.304) * [-3895.142] (-3889.422) (-3901.840) (-3896.691) -- 2:31:20
11500 -- (-3891.138) [-3887.290] (-3897.505) (-3898.465) * [-3891.534] (-3906.512) (-3896.863) (-3896.493) -- 2:24:45
12000 -- (-3886.158) (-3895.316) (-3899.951) [-3895.952] * (-3901.181) (-3909.347) (-3896.408) [-3883.249] -- 2:32:35
12500 -- [-3887.538] (-3902.739) (-3902.760) (-3912.968) * [-3892.912] (-3897.436) (-3906.132) (-3899.721) -- 2:26:29
13000 -- (-3895.558) (-3902.703) [-3903.714] (-3914.507) * (-3890.671) [-3909.579] (-3903.474) (-3897.025) -- 2:20:50
13500 -- [-3891.386] (-3903.404) (-3898.820) (-3912.319) * (-3901.203) [-3909.100] (-3904.095) (-3894.503) -- 2:27:56
14000 -- [-3893.204] (-3889.766) (-3901.741) (-3916.616) * (-3892.042) (-3906.833) (-3900.489) [-3882.750] -- 2:22:39
14500 -- (-3907.001) [-3889.357] (-3882.366) (-3921.408) * (-3892.004) (-3916.012) [-3896.471] (-3885.342) -- 2:29:12
15000 -- (-3903.163) [-3885.333] (-3885.999) (-3920.101) * [-3888.005] (-3912.557) (-3907.089) (-3893.594) -- 2:24:13
```

Average standard deviation of split frequencies: 0.079817

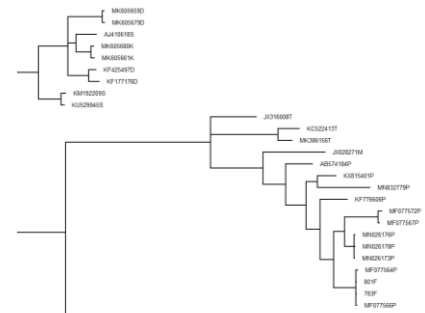
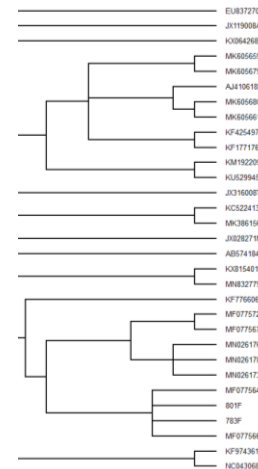
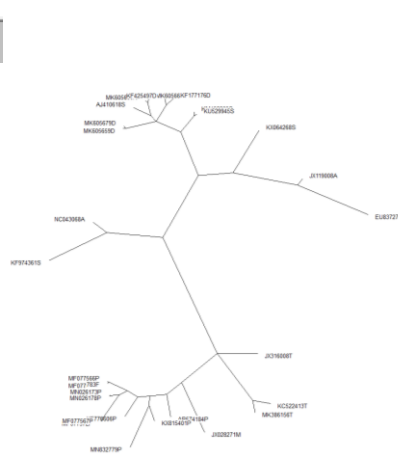
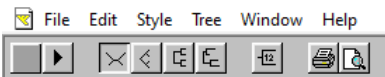
```
15500 -- (-3911.243) [-3890.943] (-3901.495) (-3904.671) * [-3880.738] (-3896.484) (-3902.559) (-3896.746) -- 2:30:18
16000 -- (-3908.134) (-3896.962) [-3891.923] (-3910.491) * [-3890.882] (-3909.738) (-3893.035) (-3904.252) -- 2:25:36
16500 -- (-3901.485) (-3896.219) [-3900.147] (-3921.860) * [-3885.614] (-3912.660) (-3889.887) (-3907.672) -- 2:31:15
17000 -- (-3894.608) [-3892.851] (-3902.664) (-3905.756) * (-3890.610) (-3896.670) (-3895.539) [-3904.515] -- 2:26:48
17500 -- (-3904.587) [-3894.477] (-3886.298) (-3920.261) * [-3902.131] (-3900.916) (-3908.223) (-3910.062) -- 2:32:06
18000 -- (-3893.863) (-3892.201) [-3880.665] (-3910.568) * (-3913.601) [-3890.487] (-3903.432) (-3908.082) -- 2:27:52
18500 -- (-3893.014) [-3884.708] (-3890.133) (-3892.454) * (-3919.170) (-3898.925) (-3901.211) [-3909.553] -- 2:23:52
19000 -- (-3899.257) [-3893.612] (-3899.712) (-3909.788) * (-3914.812) (-3897.487) (-3910.528) [-3898.014] -- 2:28:50
```

Vizualizace a úprava fylogenetických stromů

TreeView



- vizualizace stromu ve formátu **.tre** vytvořeného fylogenetickým programem
- možnost vizualizace různých typů stromů (s délkou větví, bez délky větví)
- možnost zobrazení statistických podpor uzlů
- možnost nastavení outgroupu / outgroupů a zakořenění stromu
- možnost uložení ve formátu **.emf** umožňujícím grafické úpravy v dalších programech
- možnost tisku hrubého, graficky neupraveného stromu



Vizualizace a úprava fylogenetických stromů

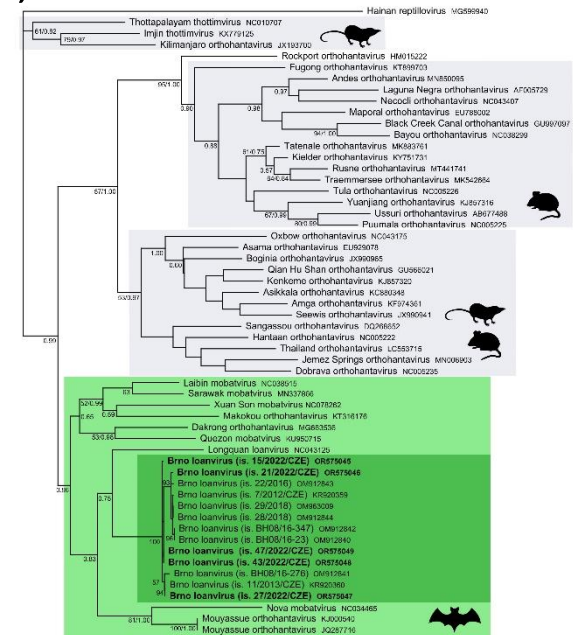
Finální grafické úpravy fylogenetických stromů:

FigTree

Adobe Illustrator



- vizualizace hrubého, graficky neupraveného stromu, ve formátech **.emf**, **.ai** nebo **.pdf**
- možnost textových úprav (názvy taxonů, velikost a typ písma)
- možnost barevných úprav a vkládání obrázků



Odkazy

- <https://www.ncbi.nlm.nih.gov>
- <https://ictv.global>
- <https://mafft.cbrc.jp/alignment/server/index.html>
- <https://phylipweb.github.io/phylip/software.html>
- <https://molbiol-tools.ca/Phylogeny.htm>

Methods By computer Cross-referenced Data types Web servers New programs Submitting

Phylogeny Programs

Changes Waiting list Other lists Old programs Not listed News

Here are 392 phylogeny packages and 54 [free web servers](#), (almost) all that I know about. It is an attempt to be completely comprehensive. I have not made any attempt to exclude prog