Instructions for Exercises

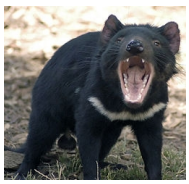
Prerequisites:

For the practical, you need the MEGA software. Go to the website https://www.megasoftware.net/ and download “MEGA 11”. To download, you have to select your operating system and the software version. In this practical, we will use the graphical (GUI) version. 64-bit should be fine with most modern PCs (younger than 7 years). Also note the installation instructions:https://www.megasoftware.net/web_help_10/index.htm#t=Part_I_Getting_Started%2FInstalling_MEGA%2FInstalling_MEGA.htm
The software should run fine on Windows, MacOS and Ubuntu.

These are just the instructions to the exercises. Please remember to answer the questions/perform the tasks on the questions sheet!

Exercise 1: Build phylogenetic trees

You will build phylogenetic trees based on the amino acid sequences of the protein alpha-globin from several species. First, you will download the sequences, then you will use the software “MEGA 11” to generate the phylogenetic trees. The species to use are:



Sarcophilus harrisii (tasmanian devil/Tasmanischer Teufel)

Papio anubis (baboon/Pavian)

Bos taurus (cow/Kuh)

Gallus gallus (chicken/Huhn)

Dasyurus viverrinus (quoll/Beutelmarder)

Rhincodon typus (whale shark/Walhai)

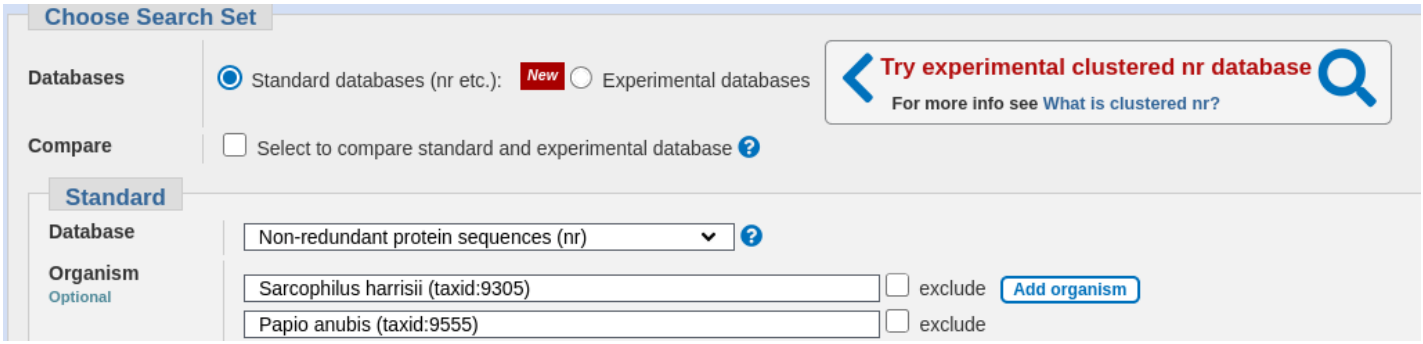
Alligator sinensis (alligator/Krokodil)

Xenopus laevis (frog/Krallenfrosch)

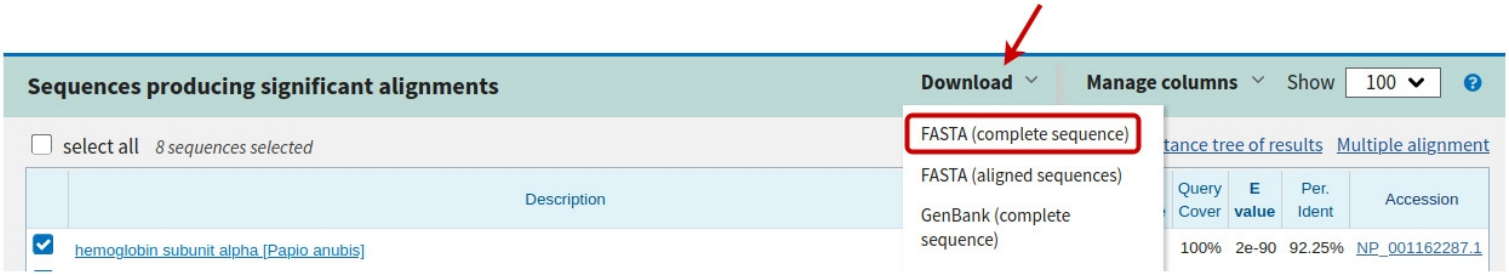
Cyprinus carpio (carp/Karpfen)

1.1 Download sequences in “fasta” format

1) Open https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome and enter the accession number “ABD95911.1” in the query field (this is the accession number for the alpha-globin gene in human).
2) Under “Choose Search Set” → ”Organism” enter the name of the species you want to run BLAST against (the nine species mentioned above).
   Start typing the Latin name and select the correct entry from the appearing menu. If everything went well, you will see the name and a taxid in the “Organism” field (e.g.: “*Papio anubis* (taxid:9555)”).



3) You can add more taxa by clicking on the “Add organism” field next to the “Organism” field. Also, you need to change the algorithm parameters and set the “Max target sequences” to 250.
4) Then, click on “BLAST”. Now, the human alpha-globin amino acid sequence will be compared to all other amino acid sequences in the taxa you specified.
5) In the results, choose the best hit for each taxon and tick the respective check boxes on the left. Once you selected all the sequences, you can download the complete multi-FASTA file for them. Hint: Find some information about the E-value and its importance for scoring alignments. Please remember the accession numbers of the sequences you downloaded.



How to download “FASTA” sequences from the BLAST output.

1.2 Checking the multi-FASTA file

Check your multi-FASTA file from the previous step in “MEGA 11”. Start the software “MEGA 11” (in Windows, write “MEGA” in the search window in the bottom left, and select “MEGA 11”) and the “main window” of MEGA opens.

1. Click on File → Edit a Text File, then open the *.fasta file you have downloaded from the BLAST output.
2. Each sequence must at least have two lines (not less!): A “header line” starting with a “>” sign followed by lines containing the amino acid sequence. Remove any empty lines, if necessary.
3. Edit the “header line” of each sequence into a short, meaningful name. For example, change “>NP_001162287.1 hemoglobin subunit alpha [Papio anubis]” to “>Baboon”, or “>Pavian”. Just remember, that there still needs to be a “>” sign at the beginning. MEGA will use the header names as labels in the tree. Thus, shorter names look better in the output.
4. Save for FASTA file under a meaningful name, i.e. under “hemoglobin.fasta”. It is highly recommended that you file name ends with “.fasta”.

```
>Papio anubis
MVLSPDDKKHVKAAWGKVGEHAGEYGAEALERMFLSFPTTI
LSKLSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP/
>Bos taurus
MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTI
LSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDFTP/
```

A valid “multi-fasta” file.

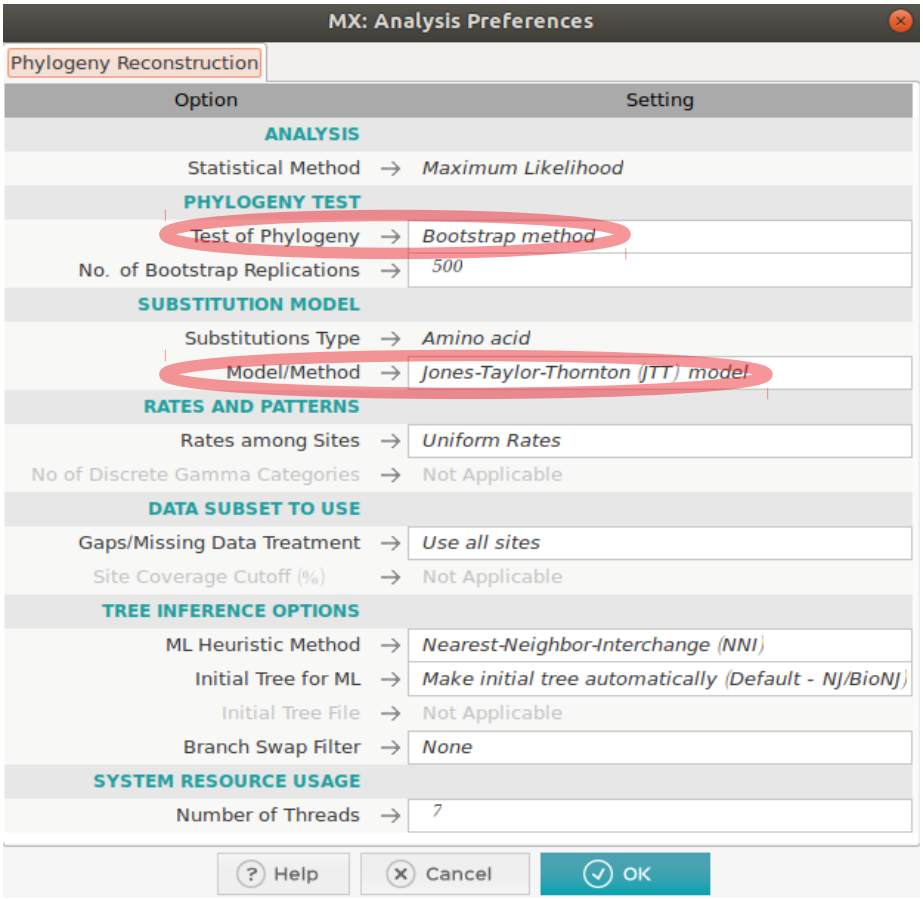
```
Papio anubis
MVLSPDDKKHVKAAWGKVGEHAGEYGAEALERMFLSFPTTKTY
LSKLSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVH

Bos taurus
MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTY
LSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDFTPAVH
```

An **invalid** “multi-fasta” file
(no “>” in header line and empty lines!).

1.3 Build phylogenetic trees in MEGA

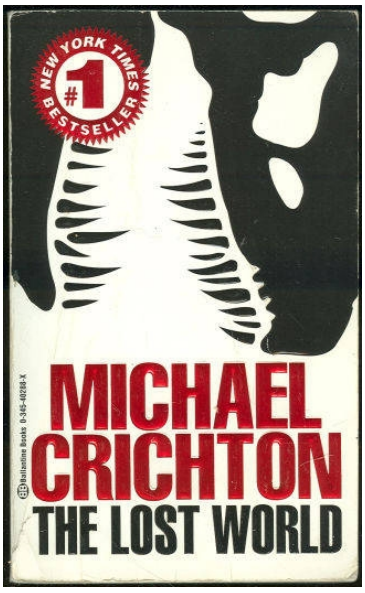
Start “MEGA 11”, if necessary. The “main window” of MEGA opens. Now follow these steps to first generate a sequence alignment and then build phylogenetic trees based on the alignment:

1. Generate sequence alignment of all nine sequences:
    a. Click on Align → Edit/Build alignment → Retrieve Sequences from a file → OK → choose your “multi-FASTA” file from step 1.2 → Open
    b. The “Alignment Explorer” window has opened (the “main window” is now in the background)
    c. In the “Alignment Explorer” window, on the top menu bar, click on Alignment → Align By MUSCLE (use the default parameters. If you are asked, if you want to “Select all”, click OK. Leave the alignment settings at default.
    d. Save the output file in MEGA format: Data → Export alignment → MEGA format.
2. Build phylogenetic trees **(with changed parameters!)**:
    a. Please remember in which MEGA window you generate which tree, as in the final tree the name of the phylogenetic method is not given!
    b. In the “main window” click on Phylogeny. Here you see five different methods. Use each of them, so that at the end you will have five trees. Choose your saved alignment file as the input data (maybe you have to change the file type to “MEGA file”) . **Set the parameters as follows**:
        - “Test of Phylogeny”=”Bootstrap method”
        - “No. of Bootstrap replications”=500 (usually researchers use 1000 bootstraps, but 500 is faster)
        - “Substitution Model” → Model/Method = ”Jones-Taylor-Thornton (JTT) model” (not possible for Maximum Parsimony).



c. You can run all the calculations in parallel, at the same time.
d. Save the resulting trees (**original, not bootstrap consensus**) as PDF files by clicking on “Image” → “Save as PDF file” *(UPGMA.pdf, NJ.pdf, ML.pdf, …)* → PDF in the top menu.

Exercise 2: BLAST



Michael Crichton's science fiction book about cloning dinosaurs, “The Lost World” (sequel of Jurassic Park), contains a putative dinosaur DNA sequence. You can download this sequence from the Teaching website, or copy it from this document.
Use nucleotide-nucleotide BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) against the default nucleotide database (nr) to identify the real source of the DNA sequence. Remember about the E-value from 1.1.

```
>LostWorld DinoDNA from the book The Lost World
gaattccggaagcgagcaagagataagtcctggcatcagatacagttggagataaggacg
gacgtgtggcagctcccgcagaggattcactggaagtgcattacctatcccatgggagcc
atggagttcgtggcgctgggggggccggatgcgggctcccccactccgttccctgatgaa
gccggagccttcctggggctgggggggggcgagaggacggaggcgggggggctgctggcc
tcctaccccccctcaggccgcgtgtccctggtgccgtgggcagacacgggtactttgggg
accccccagtgggtgccgcccgccacccaaatggagccccccccactacctggagctgctg
caacccccccggggcagccccccccatccctcctccgggcccctactgccactcagcagc
gggcccccacccctgcgaggcccgtgagtgcgtcatggccaggaagaactgcggagcgacg
gcaacgccgctgtggcgccgggacggcaccgggcattacctgtgcaactgggcctcagcc
tgcgggctctaccaccgcctcaacggccagaaccgcccgctcatccgccccaaaaagcgc
ctgcgggtgagtaagcgcgcaggcacagtgtgcagccacgagcgtgaaaactgccagaca
tccaccaccactctgtggcgtcgcagcccccatgggggaccccgtctgcaacaacattcac
gcctgcggcctctactacaaactgcaccaagtgaaccgcccccctcacgatgcgcaaagac
ggaatccaaacccgaaaccgcaaagtttcctccaagggtaaaaagcggcgccccccgggg
gggggaaacccctccgccaccgcgggaggggggcgctcctatgggggaggggggggacccc
tctatgccccccccgccgcccccccccggccgccgcccccccctcaaagcgacgctctgtac
gctctcggccccgtggtcctttcgggccattttctgccctttggaaactccggagggttt
tttgggggggggcgggggggttacacggcccccccggggctgagcccgcagatttaaata
ataactctgacgtgggcaagtgggccttgctgagaagacagtgtaacataataatttgca
cctcggcaattgcagagggtcgatctccactttggacacaacagggctactcggtaggac
cagataagcactttgctccctggactgaaaaagaaaggatttatctgtttgcttcttgct
gacaaatccctgtgaaaggtaaaagtcggacacagcaatcgattatttctcgcctgtgtg
aaattactgtgaatattgtaaatatatatatatatatatatatatctgtatagaacagcc
tcggaggcggcatggacccagcgtagatcatgctggatttgtactgccggaattc
```

Exercise 3: Substitution Matrices

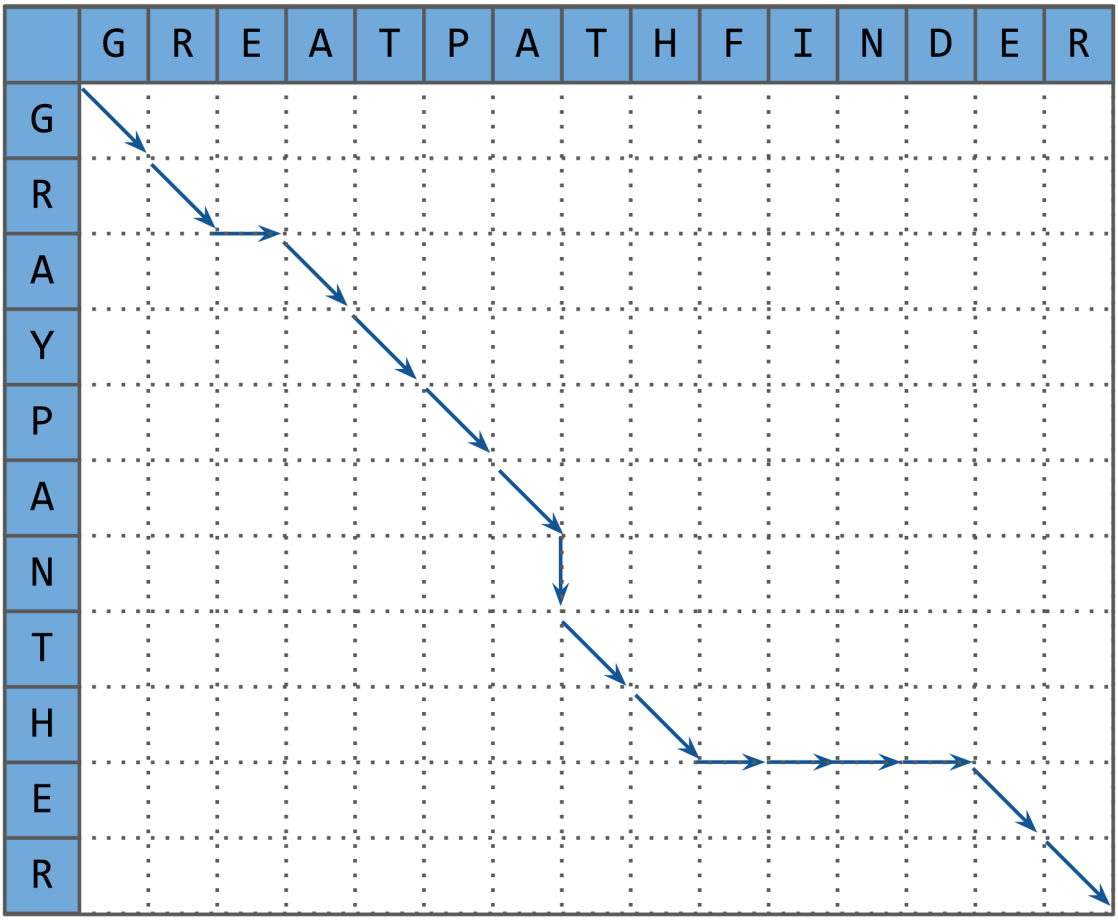
Use the BLOSUM62 and PAM250 matrices from our teaching web site, to answer the questions on the question sheet.

Exercise 4: Dotplot

On your question sheet, plot the dotplot for the word ABRACADABRACADABRA against itself, using these parameters: window size = 3, step = 1, threshold = 2.
Put a dot at the middle position of each triplet that meets the threshold criteria.

Exercise 5: Dynamic Programming

The following figure depicts a path for aligning two sequences. This is the highest scoring global alignment of the two sequences, according to the Needleman Wunsch algorithm. Answer the questions about it on the question sheet.



Exercise 6: Use BLASTx to find a hidden message

Mark Boguski of the NCBI supplied the author Michael Crichton with the “dinosaur” sequence (from the BLAST exercise). He embedded a hidden message in the sequence he provided. To see Mark's message use the translating BLAST (blastx) page with the provided sequence to translate the nucleotide sequence into an amino acid sequence. See his original publication here: https://web.archive.org/web/20210331020723/http://markboguski.net/docs/publications/BioTechniques-1992.pdf

Hint: to see the message click on the top hit and focus on the gapped part of the alignment.

Exercise 7: Add the human sequence ABD95911.1 to the tree

Download the human sequence of alpha-globin ABD95911.1 (see exercise 1) from https://www.ncbi.nlm.nih.gov/protein/ABD95911.1/ (Send to → File → FASTA). Add it to your first “multi-fasta” file (make sure the format is OK, see exercise 1.2). Next, build a Neighbor Joining tree for the new, extended data.

Good luck and have fun :)

Image sources:

https://upload.wikimedia.org/wikipedia/commons/thumb/a/ac/Tasdevil_large.jpg/252px-Tasdevil_large.jpg
https://upload.wikimedia.org/wikipedia/commons/8/81/Papio_phylogeny_%28ita%29.png
https://www.flickr.com/photos/usdagov/16802162424
https://commons.wikimedia.org/wiki/File:Dasyurus_viverrinus_Gould.jpg
https://commons.wikimedia.org/wiki/File:Gallus_gallus_bankiva_1876.jpg
http://pngimg.com/download/13168
https://commons.wikimedia.org/wiki/Category:Xenopus_laevis#/media/File:Xenopus_laevis_02.jpg
https://de.wikipedia.org/wiki/Datei:Cyprinus_carpio_GLERL_1.jpg
https://commons.wikimedia.org/wiki/File:Rhincodon_typus.png