Lab01

Introduction

During the labs, we will use different bioinformatics tools to construct sequence alignments, distance matrices, and phylogenetic trees from the initial data obtained.

Some software has a GUI available, but some don't. Some tools can be implemented online (on the Internet). But we will use the **command line environment** (not in Lab01) if your computer has options (e.g., Phylip, MEGA-CC). Therefore, it would be crucial for you to be familiar with this environment.
Assignments highlighted by a number in the Labs must be submitted to the myCourses Assignments tab. The instructor will check your progress and provide feedback to correct common mistakes. We will occasionally have **Literature Discussions** (mostly about bioinformatic protocols) that are part of **Assignments** (grading) during the Labs.

If you are installing or running a tool (in the terminal or online) and need help, please get in touch with me. Proper documentation of all the steps taken before encountering an issue will help troubleshoot the problem.

Gene scavenger hunt: Asynchronous Activity for Tuesday, January 14th

For phylogenetic analysis, first, we need a set of homologous genes (DNA or protein).
To obtain gene sequences, we are going to use the NCBI (GenBank):
https://www.ncbi.nlm.nih.gov/
You can create a user account with your **RIT credentials**. This will save your recent activity in your dashboard, such as the genes you looked for, which database you picked, the terms and parameters used, etc.
If you have done a Blast search, NCBI does not automatically save it in your dashboard. You need to click on "Save search" additionally.

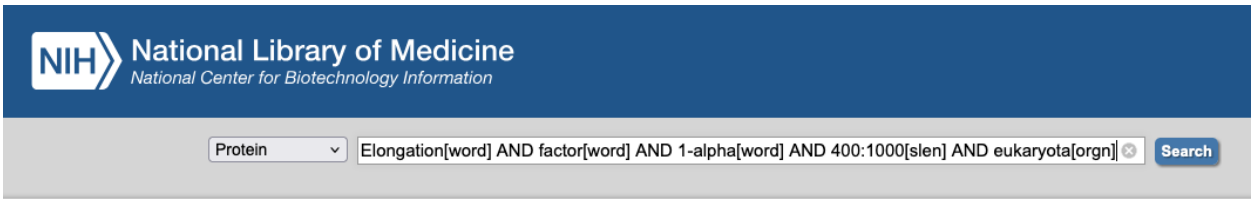
First, we are going to find a DNA sequence that interests you. You can search for genes through several databases. If you are interested in humans, one is Online Mendelian Inheritance in Man (https://www.omim.org/). Once you have your gene of interest…

Example_1:

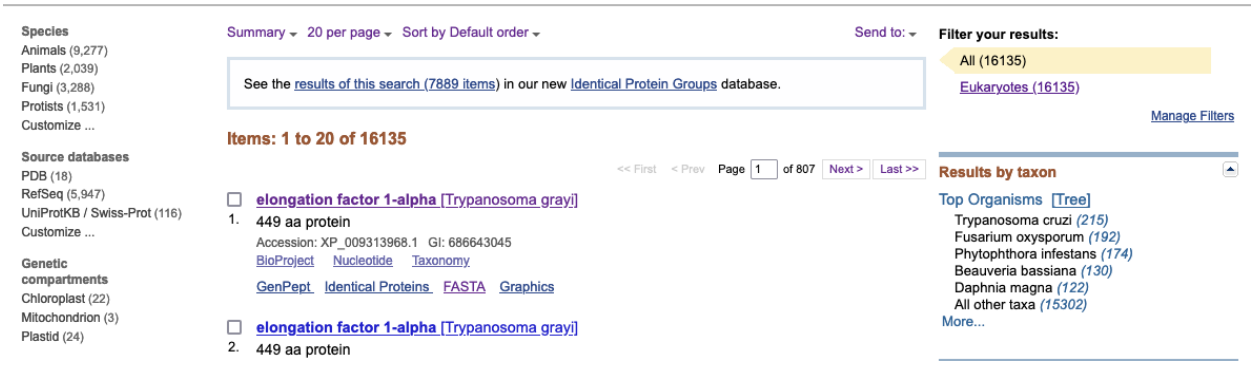
Feel free to choose any other protein and taxon to run this activity.

At the NCBI search bar, select **protein as database** and search only in eukaryotes:

Elongation[word] AND factor[word] AND 1-alpha[word] AND 400:1000[slen] AND eukaryota[orgn]



Let's see the output screen.
-How many entries?
-How is the distribution within Eukaryotes? Check results by taxon.

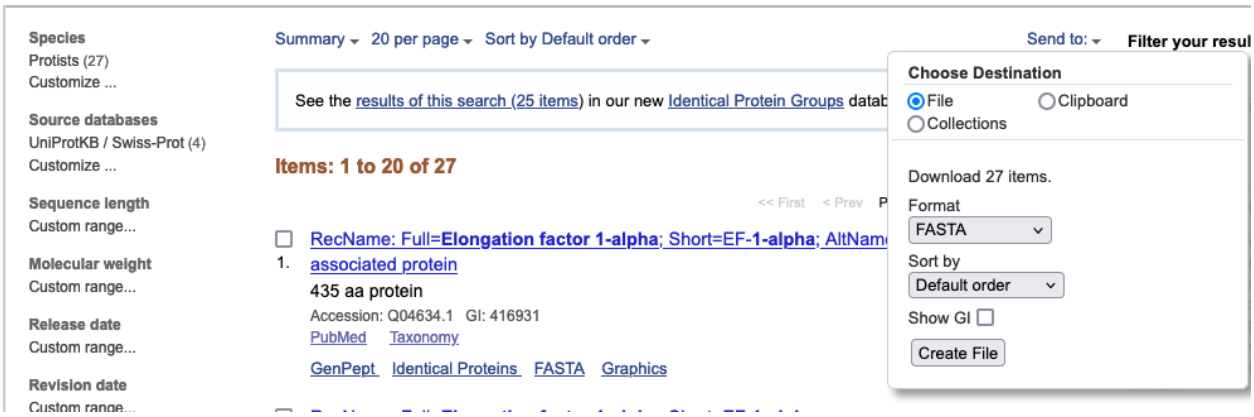


Multiple sequence fasta file
-We need a small dataset to start working with (16K is a lot!). The more extensive (more sequences), the more computation time it will take.
-Search again in the NCBI bar, under protein DB, for:

"Elongation[word] AND factor[word] AND 1-alpha[word] AND 400:1000[slen] AND ciliates[orgn]"

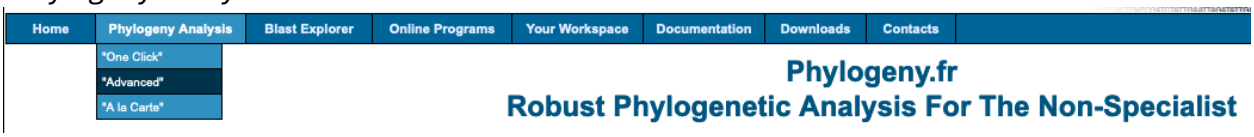
-Within ciliates, there should be only 27 entries. **DO NOT PICK MORE THAN 30 FOR THE NEXT STEPS!**

-Save the entries for future sequence analysis. ***Send to***.... *File...FASTA*
Name it something like "EF1a_ciliates.fasta"
-Good! We now have a multiple Fasta file with our gene of interest and the taxon clade we want to focus on.



Multiple sequence alignment

Let's do an online alignment of your sequences in the fasta file.
-Go to (eg.) http://www.phylogeny.fr/index.cgi
-*Phylogeny Analysis…Advance*



-Select ONLY Multiple alignments with MUSCLE and "*Create Workflow*"
-Next, Input the data (fasta file)
-Check (only check, no need to change anything) *Advance Settings*
-Submit the job. *Including your email, especially for big datasets, which can take some time (not in this case), is always good.

-Check your output and pattern. What are the different colors/shades for?
Tip: look at the bottom of the page.
-Save the output: There are three options: Fasta, Phylip, and crustal. At this point, saving multiple formats is always suitable for future analysis: Fasta and Phylip.

We can recheck the alignment in one of the viewers we have installed locally (e.g., AliView, MEGA).

Discussion 1.1
Once you finish, submit the report to myCourses. The final alignment should be submitted as a text file in Fasta (*.fasta), Phylip (*.phy), or Mega (*.meg) format.

Phylogenetic tree
Don't trust 100% what a program gives you as a result. Always double-check with your own eyes, in this case, the alignment, and see if something is off. Check for big gaps or a region

without proper alignment (no similarity blocks highlighted in a specific area). This can be shown later in the phylogenetic tree (odd long branch separated from the rest).
Let's make a quick tree with our alignment.

-Phylogeny Analysis…Advance…Check only **PhyML** and upload your file (*aln.phy).
-Go over the different settings. What happens if you change between Protein/DNA-RNA DataType?
- Click *Submit*
-Now we wait…

"PhyML has been launched! Maximum likelihood computation in progress, please wait…"

-You will get a phylogenetic tree representation in your browser. It is better to open the tree in a tree viewer (FigTree, MEGA).
-Save the tree as Newick format (*.nwk).
-Check the tree in newick format as text.
-Open the tree with MEGA or FigTree, and save it in PDF of JPG.

Discussion 1.2
Submit your Newick format tree *.nwk (text file), and a tree picture (*jpg or *.pdf).

Here ends the Asynchronous Activity for Tuesday, January 14th
We'll continue with the second part on Thursday

Homolog sequences

We are going back starting with our gene of interest, and we are going to identify homologs to it by using the BLAST server at NCBI

-Search for the BLAST portal at NCBI
-Which BLAST program do we need for your sequence?

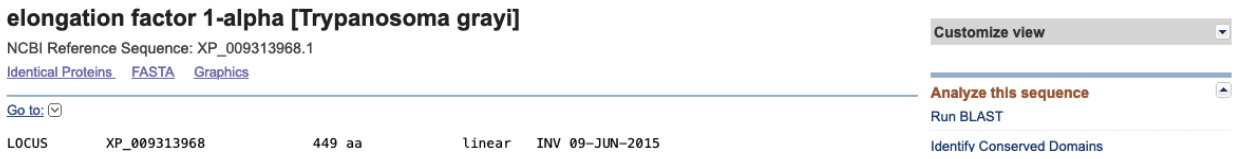
Example_2:

-At NCBI search bar, select **protein as database** and search only in eukaryotes:

Elongation[word] AND factor[word] AND 1-alpha[word] AND 400:1000[slen] AND eukaryota[orgn]

-Pick the first entry and format it as fasta. You can copy the fasta sequence in a file, and paste it at the NCBI BLAST server https://blast.ncbi.nlm.nih.gov/Blast.cgi

Or simply click on Analyze this sequence > Run BLAST



-Now, what databases are available? Which one should be used. Let's use the Model Organisms (landmark) DB.

-Algorithm/Parameters Selection (discussed in class).
-Looking good? Hit "*BLAST*" and check the results.

Activity 1.1
Download and submit (myCourses) your description table (as a text table) with a selected group of entries (no more than ten sequences). For selecting, sort by Percent identity %. Then, go to *Download*...Hit table (text). You will save a text file which looks like this:

```
# blastp
# Iteration: 0
# Query: XP_009313968.
# RID: CTWRM00T013
# Database: SMARTBLAST/landmark
# Fields: query acc.ver, subject acc.ver, % identity, alignment
length, mismatches, gap opens, q. start, q. end, s. start, s. end,
evalue, bit score, % positives
# 4 hits found
XP_009313968.1   XP_003859873.1   92.273     440   34    0     1     440
      1     440   0.0   847   95.45
```

<u>Assignment – Activity 1 (myCourses)</u>

We are going to repeat the phylogenetic analysis steps, from sequence alignment to phylogenetic tree representation, using the selected group (7-10 sequences total) from the BLAST output (landmark DB).

*Avoid repeated species.
*Sort by percentage identity. Use % > 70%

While you continue working in the blast results window, you can save the entries as a multisequence fasta file, and proceed to their phylogenetic analysis.

*Click on **GenPept** (see GenBank report for selected sequences)

*Send to…FILE…Format…FASTA…and create (save) file

*This is a good point to rename the FASTA headers: make them sorter (Gene-name plus species name).

Now you have your multisequence Fasta file. As we discussed in the previous section, you can run a sequence alignment analysis and construct a phylogenetic tree.

Activity 1.2

Submit your Newick format tree as a text file and a tree figure as PDF or JPG.
In myCourses, I have uploaded a zip file (**Lab01_data.zip**) with multiple sequences in fasta. You can build your own fasta file from the Blast step before or use the file **EF1a_Blast.fasta** for phylogenetic analysis.