Lab03

Introduction

We will run FASTA and BLAST searches at different levels.

Document all work during the dry lab for each exercise, defining all your tools and parameters, data output, and interpretation.

From Lab03, there are three Assignments to be submitted to myCourses:
Discussion 3.1
Discussion 3.2
Activity 3.1

Search databases with FASTA (Discussion 3.1)

https://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi

K-tup parameter
https://fasta.bioch.virginia.edu/mol_evol/ktup_info.html

Use U. of Virginia FASTA server to search with the **protein::protein algorithm**. You can try any of the sequences we have used in other analyses (eg. Elongation factor 1a, protein sequence).

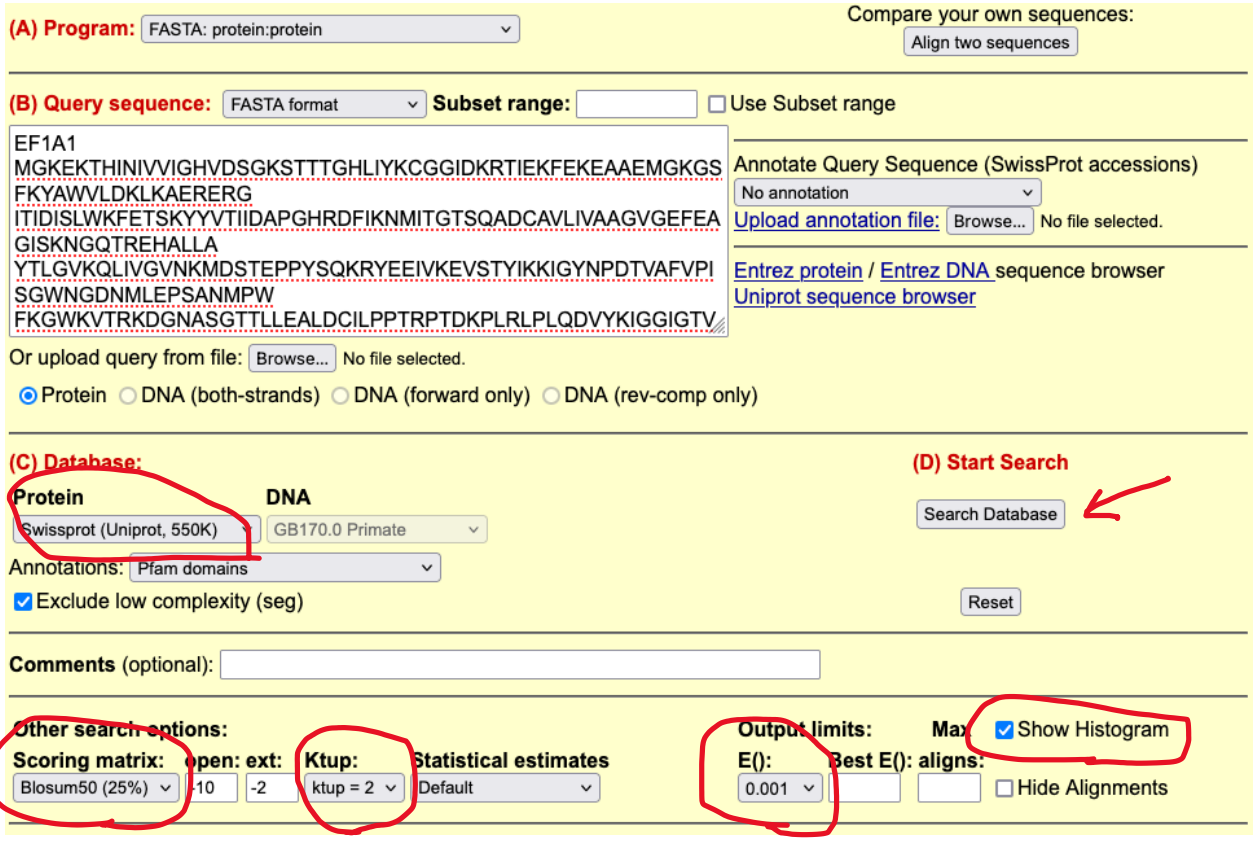
Pick a combination of parameters:
-Database (eg. Uniprot, human/Uniprot, mouse/Uniprot, Celegans/Uniprot)
-Scoring Matrix (Blosum50, Blosum80, ...)
-ktup (1 or 2)
Mark the "Show Histogram" option to get the distribution of similarity scores.

With the same input (query) sequence, modify one parameter at a time and check what has changed in the output.

Annotate your input parameters and output values and your reflection on why they have changed.
Hint: you can pick "far apart" parameters (databases from a non-related taxon, scoring matrices, etc) to better view any changes.
Hint: to organize your data with all your searches, you can **build a table** with parameters selected and output values.

Example
Using the human elongation factor 1 alpha (462 aa)

https://www.ncbi.nlm.nih.gov/protein/P68104.1



Check the **size of the database**/s (residues/sequences)
Look the **distribution of scores** (normalized similarity scores, number of sequences, E – number of expected sequences.
Explain the values you obtain from the **best score** in the hit list table.
When does the time of computation (**Scan time**) increase/decrease?

Discussion 3.1
Once you finish, submit the report with explanations to myCourses (in Assignments). You do not need to copy/paste (or screenshot) the full histograms or hit tables.

NCBI BLAST (Discussion 3.2)

Let's go back to NCBI search with blast algorithms
https://www.ncbi.nlm.nih.gov/

Again, we need some query sequences we would like to check. We are going to use the sequence of the netrin receptor from *Homo sapiens* (NP_005206.2).
Search for "NP_005206.2" and find the fasta format. You can copy and paste it in the NCBI blastp window query. Or go to the right panel to Analyze this sequence > Run BLAST.

Let's go over the options/parameters one more time. Modify the following:
Database: Reference proteins (Refseq)
Expect threshold: 0.001
Word size: 3
Filter: low complexity regions

Why are we modifying these?

Go over the result section one by one:

Descriptions//Graphic Summary//Alignments//Taxonomy

In the Graphic Summary section, what new information are we getting about the protein? What's the distribution in Taxonomy? What taxon is being represented?

THIS PART WAS DONE IN CLASS
Now, let's move on to a different gene: BAJ99511.1
https://www.ncbi.nlm.nih.gov/protein/BAJ99511.1/

What can you tell about that specific gene? How it was sequenced. From what? See all the metadata associated with the gene. Could you run a similar blastp (like the one above)? Use the "Non-redundant protein sequences" Database for this search. Again, go over all the result sections…do you find something odd?

Collect all your BLAST search discussion points into Discussion 3.2.
Discussion 3.2
Once you finish, submit the report to myCourses (in Assignments).

Standalone BLAST (Activity 3.1)

Download the simple blast-like implementations (for nucleotide sequences) into your computer.
https://git.rc.rit.edu/biol-530-630/Lab01/Lab03/blastn

Check if the Perl scripts are working in your environment. Read the documentation (README file). Download a reference sequence (aka Database) and query sequence locally. You can use the chloroplast genome we used in the Dotplot assignment (NC_014267 in gitlab as database.fasta) and a candidate gene from within to check the script's functionality.
Eg: Chlp_gene_nt.fa

```
>NC_014267.1:c1890-1774 Kryptoperidinium foliaceum chloroplast,
complete genome
```

ATGACAGGTCCAAATCCAAATAAACAAGCAGTCGAATTAAATAGAACGTCACTTTACTGGGGACTTTTAT
TAATTTTTGTTTTAGCAGTATTATTTTCAAGTTACTTTTTCAATTAA

Inspect the scripts. First, without database creation (*blastn.pl*). And then use build_blastdb.pl to create a (binary) Database file to work with *simple_blastn.pl* and *simple_blastn_hits.pl*

Save all your input/outputs from your terminal and discuss the results with the steps (modules) inside each script. Can you perform any of the following:

- Modify output format
- Change parameters within the script. Different outcome?
- Add any extra step (threshold?)

It is OK if you are not successful in any of these tasks. But let's try! ☺

Activity 3.1
Once you finish, submit the report to myCourses (in Assignments).

Edit: I know many prefer Python rather than Perl. In other Labs, we will be using Jupiter notebooks (Python). But for a rudimentary approximation to the BLAST steps, Perl was easy for me to implement. There is an attempt in *blastn_simulation_output.ipynb*, but it is far from perfect (not like the Perl scripts are way better). If you would like to try and make a script in your preferred language, feel free to do so!