Sequence bioinformatics, computer exercise IX
University of Potsdam, WS 2023/2024, Stefanie Hartmann

Give it all you've got! And since the mere presence of smartphones reduces available cognitive capacity (`https://www.journals.uchicago.edu/doi/full/10.1086/691462`), I suggest you put your cell phone out of sight and into your bag for the duration of this computer lab.

Fast database searches using BLAST

Objective

In today's exercise you will use BLAST to search hundreds or thousands of sequences for those that are similar to a query sequence. This will allow you to get familiar with the command-line BLAST program, its parameters, results, and some of its output formats. You'll also learn how to efficiently parse the output of BLAST in a manner that is suitable for large-scale analyses.

1 Logging into the cluster and preparing for the analyses

☞ Log into the university's computing cluster and change to your directory for this class on the /work partition. Create a new directory for today's exercise (e.g., lab09) and change into it.

☞ You'll use the BLAST software, which is globally installed. Load the corresponding module by typing `module load bio/BLAST/2.11.0-Linux_x86_64`

☞ Start an interactive session: `srun --cpus-per-task=1 --mem=5G -p teaching --pty bash`
Please coordinate with your neighbors so not everyone starts the interactive session at the same time.

2 The data

You will need for this exercise:

- the file TC753_ARATH.fasta from last week.
- a file with all *Arabidopsis thaliana* (predicted) protein sequences, which will be used as a database for the first set of today's exercises: "athaProteins.fasta"
- for section 6 you will need a file "moreProteins.fasta" with all the (predicted) protein sequences of nine different genomes. It contains data from the following nine plant, algal, and bacterial species; the corresponding prefixes for the sequence IDS are given in parentheses):
  - *Arabidopsis thaliana* (ath)
  - *Chlamydomonas reinhardtii* (cre)
  - *Ostreococcus lucimarinus* (olu)
  - *Chondrus crispus* (ccr)
  - *Galdieria sulphuraria* (gsu)
  - *Vitis vinifera* (vvi)
  - *Prochlorococcus marinus* (pma)
  - *Synechocystis PCC6803* (ssp)
  - *Zea mays* (zma)

☞ The two database files are available as lab09.tgz in /home/group/seqBioinf/ . The query sequence is available from your own and the shared folder, from lab08. Copy all data to a folder for today's exercise, extract files that need to be extracted, and verify that you've got all three files for today.

☞ How many sequences do the files 'athaProteins.fasta' and 'moreProteins.fasta' contain? Find the answer to this question using Linux system utilities.

☞ Can you also come up with a linux command to count how many sequences per species are in the file 'moreProteins.fasta'? grep "^>" moreProteins.fasta | uniq -c --check-chars=3 | less

20112 >ath_0
9807 >ccr_0
19528 >cre_0
7174 >gsu_0
7603 >olu_0
1844 >pma_0
3519 >ssp_0
26346 >vvi_0
46530 >zma_0

3 Formatting sequences for use as a BLAST database

The program `makeblastdb` can be used to format sequences as a BLAST database: Sequences are indexed and made readable for BLAST, and for this, the fasta file is converted into (at least) three files in binary format.

The command `makeblastdb -in athaProteins.fasta -dbtype prot` would make all sequences in the file "athaProteins.fasta" available as a database that has the same name as the fasta file. This command creates additional files that the BLAST program needs to efficiently access sequences and their IDs.

☞ Format the *Arabidopsis* sequence file as a BLAST database.

> **NOTE: For all of today's BLAST-commands (i.e., makeblastdb and the different blastp runs), you'll need a bit of patience: after issuing a command, you will have to wait for about two minutes or so before the prompt returns. Do not enter any additional commands before the process is done and the prompt is back!**

4 Using the command-line BLAST

In the terminal, similar to the online BLAST at NCBI, different BLAST programs are available for different types of searches: blastp, blastx, etc. To get to the help pages of these programs, use the option `-help`.

The minimum set of parameters that blastp requires are the following:

- the name of the database to be searched: **-db**
- the name of the file containing the query sequence **-query**
- the name of the output file **-out**

Your program calls will thus look like this; you will have to fill in the blanks with the appropriate information:

`blastp -query` _____ `-db` _____ `-out` _____

5 Searching the Toc75-3 sequence against all Arabidopsis proteins

☞ Search the Arabidopsis protein sequence against the *Arabidopsis thaliana* protein sequences in the file "athaProteins.fasta": Write down and then execute the BLAST call based on the format given above. Note that the database name is the .fasta file; the other generated files will then be recognized automatically.

☞ Look at the resulting file with the linux tool `less` blastp -query TC753_ARATH.fasta -db athaProteins.fasta -out test1

☞ Examine the best hits and the corresponding E-value, raw score, and bit score. How many identical and similar amino acids does the query sequence with the second and third best hits share? Do these hits cover the entire query sequence? (And what does the first hit correspond to?) 얘가 뭐냐?

☞ Can you identify all values in the output that are required to compute the E-value? (m, n, S, S', k, λ)

☞ Provide the definition of an E-value, once in general terms and once for a specific search result.

☞ Ignoring the E-values for now, inspect the alignments: Assess the length and quality of each alignment. Also take into consideration how much of each sequence the alignment covers.

seq	bits	E-value
ath_5611	1681	0.0
ath_6949	519	5e-179
ath_19446	488	5e-167

☞ Based only on your evaluation of the alignments: which sequences listed in the BLAST output are homologous to the query sequence, and which are not?

☞ Which E-value is at the threshold of your alignment-based decision?

6 Searching the Toc75-3 sequence against proteins from nine genomes

Next you'll collect (e.g., for further analysis such as a multiple sequence alignment) members of the toc75 gene family from more than one genome. This will again involve looking at the results using the default parameters, possibly adjusting some parameters, and deciding on an appropriate cutoff for your specific purpose.

☞ Format the sequences in "moreProteins" for use in a BLAST search, then use blastp as before to search the TC753_ARATH sequence against this file.

☞ Inspect the result using `less`: again, ignore the E-values and just evaluate the pairwise alignments. Assess the length and quality of each alignment. And how much of each sequence does the alignment cover? Also, you'll need to take into consideration the evolutionary distance of the genomes/organisms.

☞ After evaluating the alignments, determine a cutoff: which of the sequences listed in the BLAST output do you consider to be homologous to the query sequence? What is the E-value for the alignment between these two sequences?

6.1 Optimizing the generation of the output for easier parsing

Going through the alignments and evaluating them is necessary for (almost) all BLAST searches you'll do. This will allow you to determine a cutoff. For extracting sequence IDs and/or doing further analyses based on this cutoff, however, there is a better way to run the BLAST search: generate the output in a table.

☞ Again, search the TC753_ARATH sequence against the "moreProteins" database, but this time use an additional parameter, `-outfmt 7` . Using this option will output the BLAST result in tabular format. After a few explanatory lines, the entire result will be given without the actual alignments but with all the important information about the alignments.

☞ Open the result with `less` and get familiar with this output format and the different columns. You may have to increase the width of your terminal window for optimal display of the output table. The line starting with "# Fields:" tells you which information is in which column. Make sure you understand what's in each of the columns before proceeding! If necessary, compare with the result file that contains the alignments.

☞ Are all database sequence IDs in the second column unique? If not, how do you explain this?

다 각기 다른 종의 각기 다른 시퀀스를 나타냄.

☞ In the table, identify the last database sequence you decided to be homologous to the query sequence. Look at the alignment lengths, bit scores, and E-values above and below this cutoff.

☞ Now, for the third time, search the TC753_ARATH sequence against the "moreProteins" database, but this time use the parameter, `-outfmt 6` (instead of 7) . This will also result in the a tabular output, but the comment lines are omitted. Look at the result with `less`.

☞ You can exclude some alignments from the output altogether by using additional parameters. If you use blastp with the parameter `-evalue 1e-20`, for example, the output only includes matches below this threshold. In this case, a simple extraction of all values in the second column (e.g,. using cut) will give you the corresponding IDs. Try it out!

☞ Regardless of the cutoff you decided on, and regardless of how you obtained the final list of sequences that you consider homologous to the query sequence: For which of the nine species did you find homologs? Which species have multiple homologs of the toc75 genes?

☞ Optional: Play around with some of the other BLAST parameters to learn about their effects on the results.

7 Exiting the terminal

Log out of the interactive session with `exit`.

Then log out of the current session to the cluster, type `exit` again.

Finally, to cleanly close the terminal window, type `exit` yet again.

Do not close a terminal window with the mouse, especially not if an ssh session is still active.

8 Summary

☞ Summarize what you have learned about the biology/sequences/evolution of the toc75 gene family from today's exercise. No additional information needs to be looked up, and speculative answers are again encouraged.

9 OPTIONAL: online BLAST searches

If you have not yet done a BLAST search online, do this now. Direct your browser to `www.ncbi.nlm.nih.gov/BLAST/` and do an online search of the the query sequence to all or to a subset (e.g., only *Arabidopsis* or plant sequences) of the data in GenBank.

Compare the output to the results you obtained using the locally installed BLAST version. Pay attention to different results in the header region and the graphical summaries, which are only part of the online search result. Also, note that you can select BLAST hits for retrieval in the online version.

➜ Don't forget to take your flash drive with you when you leave!