Protein Domain Exercises

Elias Dohmen

05.12.2017

Exercise 1

Go to ensembl.org and download the proteome of Danio rerio in fasta format. (Try first yourself and then ask if you have problems.
Tip: You should finally get the file: 'Danio_rerio.GRCz10.pep.all.fa.gz' that you can unzip for example with gzip in the console. The ftp server of ensembl is a good location to search for this.)

Exercise 2

From the file above extract the Sequence of the protein 'ENSDARP00000018318.9'

Exercise 3

Go to
(Website: http://pfam.xfam.org and scan the Sequence from Exercise 2 for protein domains. See what can you find out about the domains and where to get different information on the webpage.

Exercise 4

Go to the InterPro webpage
(Website: https://www.ebi.ac.uk/interpro/) and scan the Sequence from Exercise 2 for protein domains. See what kind of information you get here and which databases were used.
Compare results from Exercise 3 and 4.

Exercise 5

Run pfam_scan.pl on your machine to scan the proteome of Danio rerio (Exercise 1) for protein domains. To try first if your command works properly just save the single sequence from exercise 2 in a file and use that as the input file. (The scan will take some time, just start the scan and continue with the exercises)
Tip: type

```
pfam\_scan.pl --help
```

to get an overview of the available parameters. Please choose parameters:

- '-cpu 4' to run the scan with 4 cores
- '-dir /global/databases/pfam/v31/' to scan against all pfam domains in the database version31 (we put the database already in that directory)

Check the output and see what kind of information you get with a PfamScan.

Exercise 6

Run interproscan.sh on your machine to scan the proteome of Danio rerio (Exercise 1) for protein domains. To try first if your command works properly just save the single sequence from exercise 2 in a file and use that as the input file. (The scan will take some time, just start it and continue with the exercises)
Tip: type

```
interproscan.sh
```

to get an overview of the available parameters (If you get an error/warning doing that, please ask for help)
Check the output and see what kind of information you get with an InterProScan.

Exercise 7

Check the quality of your annotated proteome (Exercise 5) with DOGMA.
Download DOGMA here:
(Website: http://domainworld.uni-muenster.de/programs/dogma/) and download and extract the matching database for pfam domains in version 31 into the DOGMA directory:
(Download: http://domainworld.uni-muenster.de/public/data/dogma/core-sets/DOGMA_v3/pfam31.tbz).
Run DOGMA to assess the quality of the annotated proteome. Tip: type

```
dogma.py --help'
```

to get an overview of the available parameters.

Extra Exercises

You can do these additional Exercises if you have time and want to practice the learned topics from the first part of the course:
Exercise 1 - extra: Using bash commands or a python script, can you download automatically all proteomes of species given in a text file? Hint: check out the 'wget' command.
Exercise 2 - extra: Using bash commands or a python script, can you extract the first n sequences from a fasta file and save it to a new one? Can you extract certain sequences specified by their identifiers in a text file and save them in a new fasta file?