Function Prediction \2017

Practical 7 - Comp Gen 2017
Assistants: Daniel Morgan and Mateusz Kaduk

PURPOSE

To analyse the proteins in your genome sequences for functional properties such as subcellular localization.

OBJECTIVES

1. Domain annotation of 100 proteins
2. A script for performing whole-genome analysis with Phobius
3. The results on one genome of the Phobius analysis
4. An xy plot of the fraction of TM (predicted transmembrane segments) proteins vs average nr of TM segments for your genomes
5. A script for performing whole-genome analysis with targetP
6. The results on your eukaryotic genome of the TargetP analysis

ACTIVITY

Perform the following steps in this order

Domain annotation

Find the Pfam domain organization for the first 100 proteins encoded in your genomes.

1. The way to do this is to use the hmmscan program.
2. Easiest is to run it as

```
# Normally you would download library from Pfam
wget ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz

# Each Pfam file is described by release notes
# ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/relnotes.txt
# Than you decompress is and prepare it for hmmscan tool
gzip -d Pfam-A.hmm.gz
hmmpress Pfam-A.hmm

# We have done that for you already, and the Pfam/ files are provided in
course directory

# Run the hmmscan
hmmscan --cut_ga --acc <hmm database file> <query protein file>
# Where hmm database file is Pfam-A.hmm and query protein file is your
proteome in multi-fasta format
```

    a. What do the options mean?
    b. As database file, use **Pfam/Pfam-A.hmm.** This may take a while.
    c. To parse the results, you can use the command

```
cat hmmscan-output.txt | perl hmmscan_parser.pl
```

3. What kind of output does this give?
4. Using the **pfam2go** map (http://geneontology.org/external2go/pfam2go), assign gene ontology terms to each of the genes.
    a. Best is if you write a script to do this. If you do not, you can use
    b. **pfam2goTransfer.py** which takes as arguments the pfam2go file and the output from `hmmscan_parser`**.pl**. If you do use this, you must answer the following questions:
        i. How many dictionaries are used in the program? What is gained by using them in the places where they are used?
        ii. What are the purposes of the third and fourth split commands?
        iii. What type of variable is arch? What is its biological meaning?
5. Do the results from what you ran above differ from simple BLAST (hypothetically domain sequence vs proteome file)? How, why and to what extent?

Simple one-gene analysis using Phobius

6. We will use Phobius, a fast and accurate predictor of TM topology

Before you start the genome analysis, make sure that Phobius works from the command line. Make a test file called Q8TCT8 with the sequence of Q8TCT8 (see **software documentation** http://phobius.sbc.su.se/instructions.html).

```
/afs/pdc.kth.se/home/e/erison/Public/bin/phobius/1.01/phobius 8QTCT8.fa
```

Then run it on the web server. The results should be the same.

Whole-proteome analysis with Phobius

7. Now we want to run Phobius for all proteins in a genome. This assumes you have a fasta file with all proteins in each proteome from previous practicals. We need a script that launches Phobius for each sequence, and parses the output of Phobius. This is a very typical script in bioinformatics so it is a very general exercise. All we want to collect for now is the number of predicted signal peptides and TM (predicted transmembrane segments) segments for each protein, and find out for one proteome:
    a. The fraction of proteins with 0 TM segments.
    b. The fraction of proteins with > 0 TM segments.
    c. The average number of TM segments for those with >0 segments.
    d. The fraction of proteins with > 0 signal peptide.
    e. The fraction of those (with > 0 signal peptide) with > 0 TM segment.

    Tips:

    You can call Phobius for each of sequence, or simply run

```
phobius -short <proteome file>
```

to get a one-line summary for each protein that is easily parsed without BioPython.

Comparative proteome analysis with Phobius

8. Now run the previous analysis on all your (real) genomes. Make an xy scatter plot showing the fraction of TM proteins on one axis and the average nr of TM segments on the other axis. Is there a trend?

More protein localization analysis with targetP

TargetP can predict more subcellular localizations, namely mitochondrial (only eukaryotes) and chloroplast (only for plants). Test it on your yeast chromosome via: http://www.cbs.dtu.dk/services/TargetP/

9. What does Plant/Non-Plant parameter do ?
    a. What fraction are predicted mitochondrial proteins ?
    b. How many are both predicted mitochondrial and to have a signal peptide? (Comment on such predictions, are they biologically sound?)
10. Would your run targetP for all of your genomes ? Why ?