Comparative Genomics 2018

Practical 7: Function Prediction

Assistants: Stefanie Friedrich, Miguel Castresana, Deniz Secilmis

All forms of plagiarism are forbidden, and if detected it will result in a lower grade.

PURPOSE

In this practical you will investigate the proteins in your genome sequences for functional properties such as identification of transmembrane (TM) segments, signal peptides (SP), mitochondrial proteins, and sub-cellular localization by using different tools and databases.

KEY QUESTIONS

Summarise shortly this practical

Domain annotation of the first 100 proteins

1. What are Pfam domains?
2. What do the options of the offered hmmscan command mean?
3. What kind of output gives hmmscan_parser.pl on the hmmscan output?
4. What are gene ontology terms? Why are they needed? What kind of information do they offer concerning molecular functions? (the GO documentation is very helpful: http://www.geneontology.org/)
5. Write your own Python script to assign gene ontology terms to each of the genes; or if you prefer to use the given script pfam2goTransfer.py answer the following questions:
    a. How many dictionaries are used in the program? What is gained by using them in the places where they are used?
    b. What are the purposes of the third and fourth split commands?
    c. What type of variable is arch and what is its biological meaning?
6. Choose the GO identity numbers of three genes (i.e, one or more domains per gene); which molecular functions are linked to these genes/domains?
    a. To understand your genomes even more thoroughly, one could further analyse the number of proteins with either one or more than one domain, and the domains with either one or more than one GO:term. However, you will go in more detail when analysing TM segments and SP for your genes in the following task.

Whole-proteome analysis with Phobius

1. To analyse TM segments and SPs a simple analysis can be performed. For each proteome investigate and present:
    a. The number of proteins for both, with 0 and at least one TM segment, and the share of both categories (should sum up to 1)
    b. The average number of TM segments for those with >0 segments
    c. The number of proteins for both, with 0 and at least one SP, and the share of both categories (should sum up to 1)
    d. The share of those with both, SP segments > 0 and TM segment > 0
2. Make two scatter plots with the results obtained for your five genomes
    a. One is showing the share of TM proteins on the x-axis versus the number of proteins with at least one TM segment on the y-axis
    b. The other one shows the share of TM proteins on the x-axis and the average number of TM segments on the other axis.
    c. Is there a trend? Is the data sufficient for the five genomes to generalise your observation?

More protein localization analysis with TargetP

1. What does the Plant/Non-Plant parameter mean and affect?
2. What number and share of predicted mitochondrial proteins do you observe?
3. How many proteins are both, predicted mitochondrial and have a signal peptide?
    a. Comment on such predictions, are they biologically sound?
4. Would you run targetP for all of your genomes? Motivate your choice!

MATERIALS & TOOLS

1. HMMER version 3.1b2 (hmmscan) to scan for gene families in Pfam
2. Gene Ontology consortium to study gene ontologies http://www.geneontology.org/
3. Phobius version 1.01 to investigate transmembrane segments and signal peptides; program `module load phobius` & webtool http://phobius.sbc.su.se
4. TargetP to predict subcellular location http://www.cbs.dtu.dk/services/TargetP/

ACTIVITIES

Domain annotation

1. Find the Pfam domain organization for the first 100 proteins encoded in your genomes (HMMscan takes a while to run, therefore, you will only analyse the first 100 proteins).
    a. First, extract the first 100 genes from your proteomes with a python script you write
    b. Second, find the domains in these 100 genes with hmmscan; easiest is to run it with:

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

# Each Pfam file is described by release notes
ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/relnotes.txt

# Decompress and prepare for the hmmscan tool
gzip -d Pfam-A.hmm.gz
hmmpress Pfam-A.hmm

# We have done the steps above already and the files are provided
in the course directory data/Pfam/
# The files are large, no need to download them; you can insert the
path to the hmm_database_file into the command

# Go through the available options with
hmmscan --help /hmmscan -h

# Run hmmscan where hmm_database_file is Pfam-A.hmm and
query_protein_file is your proteome in multi-fasta format (this may
take a while ...)

hmmscan --cut_ga --acc <path/hmm_database_file>
<query_protein_file>
```

2. Parse the results with this command pipeline:

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

3. Use the pfam2go map (http://geneontology.org/external2go/pfam2go) to assign gene ontology terms to each of the 100 genes:
   a. Try to write a script for this
   b. If you do not, you can use pfam2goTransfer.py which takes as arguments the pfam2go file and the output from hmmscan_parser.pl.

Simple one-gene analysis using Phobius

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

1. Before you start the genome analysis, make sure that Phobius works from the command line.

```
module load phobius
#help and available options with
```

```
phobius --help
```

2. Make a test fasta file called Q8TCT8 with the sequence of Q8TCT8 (see software documentation http://phobius.sbc.su.se/instructions.html).

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

3. Run it on the web server (see Materials & Tools) to test if the results are the same.

Whole-proteome analysis with Phobius

1. Run Phobius for all proteins in each of our genomes.
    a. We need a command pipeline or a script that launches Phobius for each genome and parses the output of Phobius (the latter should be a python script). All we want to collect for now is the number of both, predicted transmembrane (TM) segments and predicted signal peptides (SP), for each protein.

       Tip: To get a one-line summary of each protein that is easily parsed without BioPython, you can call Phobius for each sequence simply with:

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

More protein localization analysis with targetP

TargetP can predict more sub-cellular localisations, namely mitochondrial (only eukaryotes) and chloroplast (only for plants).

1. Test targetP on your eukaryotic genome at http://www.cbs.dtu.dk/services/TargetP/