Course KB8007 Comparative Genomics

Practical 7: Function prediction

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

The report should be formatted in one .doc file and sent to oliver.frings@sbc.su.se before the end of the week and contain the following results. Missing or failed items will result in a reduced grade for this practical.

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

Procedures:

Simple one-gene analysis using Phobius

We will use Phobius, a fast and accurate predictor of TM topology, see http://phobius.sbc.su.se/. It is installed locally as a module at SBC, type “module add phobius” to activate it.

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 http://phobius.sbc.su.se/instructions.html). Run like so:

```
phobius Q8TCT8
```

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

Whole-proteome analysis with Phobius

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 segments for each protein, and find out for one proteome:

- The fraction of proteins with 0 TM segments.
- The fraction of proteins with > 0 TM segments.
- The average number of TM segments for those with >0 segments.
- The fraction of proteins with > 0 signal peptide.
- The fraction of those (with > 0 signal peptide) with > 0 TM segment.

Tips:

- You can either read in the protein sequences with the SeqIO.parse function in BioPython and call Phobius for each of sequence, or simply run ‘phobius –short genome0.pep’ to get a one-line summary for each protein that is easily parsed without BioPython.

<u>Comparative proteome analysis with Phobius</u>

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?

<u>More protein localization analysis with targetP</u>

TargetP can predict more subcellular localizations, namely mitochondrial (only eukaryotes) and chloroplast (only for plants). Test it on a single sequence:

```
module add targetp
targetp –N Q8TCT8
```

What does the -N specify? Now modify your previous script such that it will run targetP on each protein. Modify the parser so that it parses out the mitochondrial proteins. Questions:

- What fraction are predicted mitochondrial?
- How many are both predicted mitochondrial and to have a signal peptide? (Comment on such predictions, are they biologically sound?)