Practical 2: Gene Prediction

Outline
In this practical you will use genscan and glimmer to predict the genes of your genomes and to get the protein sequences for the genes.

If anything does not work or is unclear, contact Kristoffer (krifo@sbc.su.se) as soon as possible.

Aim
1. To get familiar with the programs and their output.
2. Script a pipeline to run the programs and parse their output properly.
3. **Get protein sequences for the genes of your genomes (**you will need these for later practicals!**)**

Details
To start with
- Read the details thoroughly, some parts might take more time than others and some may be possible to perform simultaneously.
- Browse http://genes.mit.edu/GENSCANinfo.html and http://www.cbcb.umd.edu/software/glimmer/.
  - You might also want to look at /afs/pdc.kth.se/home/k/krifo/Public/bin/glimmer3.02/glim302notes.pdf as well as at /afs/pdc.kth.se/home/k/krifo/Public/bin/genscan/README

Glimmer
- Run glimmer for a genome using g3-from-scratch.csh or g3-iterated.csh. This script is located in /afs/pdc.kth.se/home/k/krifo/Public/bin/glimmer3.02/scripts so you will have to add this directory to your path first.
  - Write a script to repeat this for all your genomes. (looping over all genome files is suitable)
  - Is glimmer suitable for all your genomes?
  - Why?
- Examine the raw output from glimmer.
  - Can you see some difference between eukaryote and prokaryote glimmer output?
  - If so, why do you think that is?
- Write a script to parse the glimmer output and extract the protein sequences corresponding to the predicted gene coordinates. The end result should be proteome multifasta files, that is, a file on the format:

  >[some protein ID 1]
  [AMINO ACID SEQUENCE 1, single line format]
  >[some protein ID 2]
  [AMINO ACID SEQUENCE 2, single line format]
  ...

The glimmer <tag>.predict files will contain the predicted gene coordinates for the query genome. This means that, in order to generate the final proteome, you might have to extract the corresponding region from the original genome fasta file for each gene, then translate it codon-by-codon into amino acids. Generate such proteome files for each prokaryote or archaeal genome you have.

Genscan

- Run genscan for the eukaryote genomes. The binary is in /afs/pdc.kth.se/home/k/krifo/Public/bin/genscan so you might want to add that directory to your path. Look at the README to find the syntax. You need to supply a parameter file, this should be /afs/pdc.kth.se/home/k/krifo/Public/bin/genscan/HumanIso.smat as we approximate the characteristics of the other eukaryotes with those of human.
  - Genscan requires a hefty amount of memory, what implications might this have?
  - Is there ways to reduce/get around potential problems stemming from the memory requirement?
  - Where does the output from genscan end up, do you need to redirect it?
    - It might be good to try with a short test file e.g. the first few lines of one of your genomes.
- Write a script to parse the output from genscan to create a proteome multifasta file as for glimmer above.
- Generate proteome multifasta files for all your eukaryote genomes.

Report:

Your written report for this practical should include:

- A short summary of what you have done as well as specification and motivation of what methods and parameters you have chosen.
- Discuss the questions in the lab specification as well as any questions and thoughts of your own.
- Discussion of results.
- Well commented code for any scripts you have written.