Comparative Genomics 2018

Practical 2: Gene prediction

Assistants: Miguel Castresana, Deniz Seçilmiş, Stefanie Friedrich

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

In this practical you will learn how to use **Glimmer** and **GENSCAN** to predict the genes of your genomes and to obtain the protein sequences for those genes.

Suggestion: Start by familiarizing yourself with **GENSCAN** and **Glimmer**.

Glimmer

1. To understand the program and their commands go to : https://ccb.jhu.edu/software/glimmer/glim302notes.pdf
2. A.L. Delcher, K.A. Bratke, E.C. Powers, and S.L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer, Bioinformatics 23:6 (2007), 673-679.

GENSCAN

1. Burge C1, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997 Apr 25;268(1):78-94.
2. http://genes.mit.edu/README

Both programs are pre-installed

```
/afs/pdc.kth.se/projects/sbc/vol/software/genscan/1.0/install/i386_ubuntu8.10/genscan
/usr/bin/tigr-glimmer
```

If they are not running use:

```
module add genscan
module add glimmer
```

Exercise 1 - Glimmer

Steps for the first part of the exercise. Please explain the parameters you are using to run Glimmer. Check those in the PDF provided.

1. Find long ORF from genome

```
tigr-glimmer long-orfs -n -t 1.15 01.fa 01.long-orf-coords
```

2. Extract long ORF

```
tigr-glimmer extract -t 01.fa 01.long-orf-coords > 01.longorf
```

3. Prepare training set

```
tigr-glimmer build-icm -r 01.icm < 01.longorf
```

4. Start glimmer

```
tigr-glimmer glimmer3 -o50 -g110 -t30 01.fa 01.icm 01.glimmer
```

5. Long ORFs are provided to construct the training set, what other two sources of sequences can be used instead of or in addition to long ORFs ?
6. Is Glimmer suitable for all genomes ? Why ?
7. Make a histogram of predicted gene lengths for each genome in R

```
install.packages('ggplot2')
library(ggplot2)

plotGlimmer = function(file='01.glimmer.predict') {
  t = read.table(file, header = F, skip = 1)
  c = data.frame(size=abs(t[,2]-t[,3]))
  ggplot(c, aes(size))+geom_histogram(binwidth=1000)+ggtitle(file)
}
plotGlimmer()
```

8. Do all gene sizes follow the same distribution in all genomes ?
9. Extract the protein sequences from the predicted genes obtained. Use the script **parseGlimmer.py.2** available in the script directory.

Exercise 2 - GENSCAN

Steps for the second part of the exercise.

Run **GENSCAN** for the eukaryote provided in Practical 1. Run it, using **HumanIso.smat**.

1. From GENSCAN output, extract the amino acid and nucleotide sequences and make separate files for each.
2. Create the PostScript (graphical) output, which is a diagram of the locations and DNA strand of all predicted exons/genes.
3. Using BLAST and the nucleotide sequences extracted from GENSCAN output, tell me the protein names of the first two nucleotide sequences.