Gene Prediction \2017

Practical 2 - Comp Gen 2017
Assistants: Daniel Morgan and Mateusz Kaduk

PURPOSE **1**

MATERIALS NEEDED **1**

Glimmer 2

GeneScan 2

PURPOSE

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

MATERIALS NEEDED

Start by familiarizing yourself with GENESCAN and Glimmer

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

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.

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
```

Glimmer

Steps for the first part of the exercise

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 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 ?

GeneScan

Steps for the second part of the exercise

1. Run Genscan for the eukaryota (identified with blast in practical 1) genomes using HumanIso.smat (genvertseq.fa)
2. Genscan requires a hefty amount of memory, what implications might this have?
3. It might be good to try with a short test file e.g. The first few lines of one of your genomes.

Extract protein sequences (proteomes) and gene sequences (genomes) from glimmer and genescan for relevant genomes. You can validate your dna->protein translation script by using blast on resulting sequences and inspecting fasta file. Save fasta files with protein and nucleotide sequences.