

Exercises for course 27104,
The Scientific Communication of Comparative Genomics

Teacher: Prof. David W. Ussery
Technical support and exercises: Matloob Qureshi
Assistant teacher and exercises: Tammi Vesth

Week 3 tasks:

Day 1, Wednesday 21 September:

1. Genefinding using prodigalRunner
2. Number of genes/proteins, compare the new count with the previous one
3. Amino acid and codon usage

Day 2, Thursday 22 September:

1. BLAST matrix

Output of the week:

- Protein coding genes in FASTA format
- List of number of annotated genes compared to the published number of genes
- Amino and codon usage plots for selected genomes
- BLAST matrix

1 Day 1, exercises:

1.1 Run genefinding algorithm on extracted DNA (hypothetical genes/proteins).

Up until now, you have worked with the genes/proteins published along with the genome sequence. Now we will run our own genefinding algorithm on the DNA sequence of the genome. This is very often the same that the publishers of the genome has done. The good thing about running one algorithm on all the genomes is that the results will be standardized which is not the case with published annotations. First, create a folder for the files generated during genefinding:

```
mkdir ProdigalFiles
```

Somewhere in your folder-structure, you have saved the DNA FASTA files for your genomes (probably called `<name>.fna`). Find these files in your directories and copy these files into the *ProdigalFiles* directory:

```
cp <path to <name>.fna files>/*fna ProdigalFiles/
```

Enter the *ProdigalFiles* directory and run genefinding on the DNA FASTA files:

```
cd ProdigalFiles
prodigalrunner <name>.fna
```

This can be wrapped in a *for* loop, BUT make sure that it works with a single genome first!

```
for x in <name1> <name2> <name3> <name4>
> do
> prodigalrunner $x.fna
> done
```

Or with the wild-card for all files called `<something>.fna`

```
for x in *.fna
> do
> prodigalrunner $x
> done
```

Output files include:

```
<name>.gff =>raw prodigal output, you will not use this file
<name>_prodigal.orf.fsa => protein file in FASTA format
<name>_prodigal.orf.fna => gene file in FASTA format
<name>_prodigal.gbk =>"fake" GenBank file, you will not use this file
```

1.2 Count number of hypothetical genes/proteins, from FASTA files (amino acids).

This procedure is something you have done before. Be aware that you are now working on protein files from your local gene finding. The number of genes/proteins might be different than the numbers you obtained from the published genome data. Compare the numbers you obtained here with the previous numbers for the same genome.

```
grep -c ">" <name>_prodigal.orf.fsa
```

```
for x in *prodigal.orf.fsa
> do
> echo $x
> grep -c ">" $x
> done
```

1.3 Calculate amino acid and codon usage for hypothetical genes/proteins.

These basic statistics for each genome are visualized in a PDF summary with three plots and a text summary. The calculations have been implemented in a program called *basicgenomeanalysis* and takes gene FASTA files as input. In this setup the gene FASTA files have the suffix **.orf.fna*. Run the analysis as follows, note that the name is WITHOUT the suffix:

```
basicgenomeanalysis <name>_prodigal /usr/bin/gnuplot
```

2 Day 2, exercises:

2.1 Construct BLAST matrix from hypothetical genes/proteins.

A BLAST matrix is a comparison of proteomes (proteins from a genome) used to estimate how many proteins is found in common between two genomes. We will construct a matrix from the **.orf.fsa* files created by *prodigal*. The BLAST matrix algorithm has been implemented in a program called *blastmatrix*. First we will construct an input file for this program. The input file must be of the format *XML* which is a nice computer-reading format but not very friendly to human eyes. A small program called *makebmdest* construct a *XML* file from all the **orf.fsa* files (protein files in FASTA format) in a directory. Make sure that you have the right files in the current working directory. In class you will run this procedure on 3 genomes (< *genome1* >, < *genome2* >, < *genome3* >). Select 3 genomes and create a folder for them:

```
mkdir TestMatrix
```

Copy the needed files into the test directory:

```
cp <genome1>_prodigal.orf.fsa TestMatrix/  
cp <genome2>_prodigal.orf.fsa TestMatrix/  
cp <genome3>_prodigal.orf.fsa TestMatrix/
```

Enter the *TestMatrix* directory:

```
cd TestMatrix
```

Create BLAST matrix XML file from the protein FASTA files within the *TestMatrix* directory (current working directory):

```
makebmdest . > bmdest.xml
```

The "." indicates that the path to the files is the path to the current working directory. If this dot is not included the code will create a *XML* file with no references to the actual protein files.

the BLAST matrix is then generated using the *blastmatrix* program with the file *bmdest.xml* as input.

```
blastmatrix bmdest.xml > blastmatrix.ps
```

In class you should do this for a small set of your genomes. Homework will be to start this plot for all of your organisms and let it run overnight.