Genome Sequence Analysis Assignment

Issued: Friday 16 November 2018
Due: Monday 26 November 2018

Please submit an electronic version via Turnitin (instructions on how to upload your submission are in the Coursework Dropbox pages of Blackboard.)

A hard copy of your report is not required.

AIM: This practical introduces some of the basic concepts used during the analysis of eukaryotic genomic sequences. A number of widely used Unix-based analysis programs will be run on a segment of genomic sequence from a fungus. We will explore how the results of these analyses are used to build a composite picture of the features present within the sequences, each contributing its own independent evidence to our production of putative gene models.

You will be required to:

- Discuss the individual results generated and how they lead to the identification of features in the sequence.
- Use this information to suggest a putative gene model or gene models, including the number of exons.

Write-up:

- Please produce a report of no more than 1,000 words, with a maximum of 8 figures/tables. Please provide relevant key references. The components of the report should be a logical representation of the work and critical assessment of the results.
- When discussing the output (results) of particular programs, please illustrate your comments with relevant excerpts or summaries from the output files generated. Please do not include unfiltered, un-annotated output. It is up to you to decide what is relevant but don't forget to take into account the statistical validity of the results you discuss.
- Basic standards for write-ups apply: please make sure that all figures and tables have descriptive titles and are correctly referenced in the text. Please do fully cite papers, sites and other materials you refer to.

Instructions

First collect a copy of the mystery fungal genomic sequence you will need to analyse (**fungal_18.fa**).

This is available on Blackboard.

1. Repeats

Use the program RepeatMasker to investigate the repeats present in the sequence.

In order to use RepeatMasker you need to first load the appropriate module by typing:

```
module load repeatmasker
```

You can view the RepeatMasker documentation by typing

```
less /project/soft/linux64/RepeatMasker/repeatmasker.help
```

A file of suitable repeats to search has been provided in a file called MScRepeats.lib. RepeatMasker can search specific repeat libraries by used the -lib flag:

```
RepeatMasker –X –lib /project/data/blastdb/MScRepeats.lib  fungal_18.tfa
```

The “-X” flag tells RepeatMasker to replace repeat sequences with ”Xs” rather than the default “N”. The reason for this is that the sequence already contains a few Ns so using X will enable you to see the repeat locations.

RepeatMasker produces 5 separate output files, each named after the input file, with the following endings - .tbl, .log, . out, .masked and .cat. Have a look at each and briefly list what it contains. What repeats are found?

Many of the repeats in the MScRepeats.lib are predictions made by the REPET package (https://urgi.versailles.inra.fr/Tools/REPET), particularly transposable elements (TEs). These repeats are named as Grouper, Piler or Recon, reflecting the clustering method used in their prediction. The repeat class/family is unknown.

2. Homology Methods:

Now use BLAST to look for matches to known genes or other features in the genomic sequence.

In order to use BLAST you need to first load the appropriate module by typing:

```
module load blast
```

A selection of suitable blast databases are stored in the directory /project/data/blastdb. Each one is made up from multiple constituent files. To obtain a list of available nucleotide databases you can type

ls –l /project/data/blastdb /*.nal

(when you specify the name of the database to search, omit the .nal ending).

Similarly, a list of protein blast database names can be produced by replacing the **.nal** with **.pal.** *Please note that due to the complex way the blast databases are stored and partitioned, files that appear to pertain to the same database may not all have the same file creation and modification dates. This is normal. For the purposes of this practical, you will use static local versions of databases. These may not be up to date.*

You will need to search both DNA and protein blast databases. Results will be saved in new files specified by the name after the –out flag as below, e.g.:

blastn -query fungal_18.tfa -db ena_std_fun -out fungal.tfa.blastn.ena_std_fun

Please justify your choice of databases, discuss the results generated and their significance, and explain how this adds to your knowledge of features in your genomic sequence.

Include key parts of the blast output in your write-up (<u>not</u> the entire output files please) and refer to them when you discuss what you have learned from them.

<u>3</u>. ***Ab initio*** **<u>Methods - Genscan:</u>**

In order to use Genscan you need to first load the appropriate module by typing:

module load genscan

Run Genscan using the parameter set for human sequences.

genscan /project/soft/linux64/src/genscan/HumanIso.smat fungal_18.tfa > genscan_human.out

Comment on the differences between the resulting gene predictions generated here and those generated by FgenesH in the next section. How do these differences arise and Which program is more likely to be accurate in this instance?

4 Ab initio gene prediction - fgenesh:

Use the web-site for fgenesh at http://www.softberry.com to run fgenesh on your sequence. You will find a great deal of help information on the web-site and also links to other gene prediction methods that may be useful later on in the practical. Remember to consider the origin of your sequence and chose appropriate training set(s). Save your results, discuss their meaning briefly and justify the choice of training set(s) used. **If you experience problems contacting this external web-site on several attempts, please email d.huntley@imperial.ac.uk.**

5 Protein-based annotations

Several of the *ab initio* gene predictors can produce conceptual translations of putative proteins. Very briefly compare and contrast the protein predictions made, referring back to the predicted exons/genes where necessary.

Now choose the protein sequence(s) that you think best fits the evidence (compare *ab initio* results with your BLAST results) for each putative gene and use this sequence(s) to perform further analyses to gain more information about the gene structure(s) you have identified, and the encoded protein(s). Include a brief description of the methods you used, the results generated and a discussion of how the information can be used to improve our knowledge of this genome segment and the encoded protein.

Can you identify the organism that the original genomic sequence most probably was derived from?

Additional information to help your write-up

A simple way to browse for entries in specific DNA and protein databases is to use the databases at the NCBI via Entrez http://www.ncbi.nlm.nih.gov/gquery. Individual databases can be selected, or multiple ones.. If you already know the accession number you are interested in (e.g. as the result of a BLAST search), you can search with the accession number. Once you have found the required entries they can be browsed or saved to a file.