November 2016

Assignment 2, Genome Informatics Module, Computational Biology MPhil

This is a group assignment. For groups see Moodle.

Presentations will be on **Tuesday 15th November**, please submit files by **11.45 pm Monday 14th November**. This is a fixed deadline and there will be no extensions as per the MPhil handbook.

Working as a group: You may split the assignment between group members however you wish, though all members must contribute. If someone is not taking part please let me know as soon as possible. Consider that some sections need to be done before others. In your presentation make clear who did what sections of the assignment.

Presentations: These will be 20 minutes long with 5-10 minutes for questions at the end. All members of the group must give part of the presentation.

Marking: All members of the group will receive roughly the same mark, weighted by your portion of the presentation, your answers to questions and the parts of the assignment that you contributed to.

Submission: Each member of the group should submit a copy of the group presentation via Moodle. Each group must supply a printed version of your slides on the Tuesday for reference during your presentations.

If you have any issues with the assignment please email Dr Alastair Crisp (eadc2@cam.ac.uk).

Assignment Description

Each group will be given the names of a number of *Drosophila* species equal to the number of members of the group (names on Moodle). They must first produce annotation for these species (as described below) then compare the annotations across species and to the model organism *Drosophila melanogaster*.
In your presentation describe what you did, what you found and any conclusions you drew from this. Do not feel limited to the example methods below, feel free to try different or multiple methods, but remember this assignment is assessed only on the presentation and you will be asked to justify your choices. The questions in each section are intended as a starting point, not an exclusive list of things to consider in the comparisons.

Three groups will do the gene sections and three groups will do the protein sections. This is shown on Moodle.

1 Per species annotation

This section is to be done for each species your group is assigned. Each group will do **either** the gene sections or the protein sections.

1.1 Genes

1.1.1 Annotation by alignment

Align the *D. melanogaster* transcripts and proteins to your species genome, e.g. using blastn/x.

1.1.2 Annotation using probabilistic model

Annote the genome using a probabilistic model e.g. genscan.

1.1.3 Compare annotation

1. Compare the annotation from the two different methods. Which is better/more complete/more accurate?
2. How do your annotations compare to the avaliable genes? e.g. Number, position, etc.

1.2 Proteins

1.2.1 Annotation by alignment

1. Annotate the available proteins by aligning them to a protein database, for instance to the Swissprot database (/local/data/public/genome_informatics/assignment_2/swissprot_database/) using blastp.
2. Assign GO IDs from this alignment. Details of the GO IDs assigned to each protein in Swissprot are in the same folder as the database.

1.2.2 Annotation from domains

1. Annotate protein domains using profile HMMs, for instance using HMMer and the PFAM database (/local/-data/public/genome_informatics/assignment_2/pfam_database/).
2. Assign GO IDs from this annotations. Details of the GO IDs assigned to each PFAM entry are in the same folder as the database.

1.2.3 Compare annotation

1. Compare the annotation from the two different methods. Which is better/more complete/more accurate?
2. Compare your annotation to the annotation on e.g. FlyMine. Do they differ? e.g. Number of genes/ number of GO IDs/range of GO IDs annotated.

2 Cross-species comparison

Compare all species to each other and to *D. melanogaster*. Each group will do **either** the gene sections or the protein sections.

2.1 Genes

Compare your annotations and the available gene annotations across all species and with *D. melanogaster*. Consider things like gene number, length etc.

2.2 Proteins

Compare your annotations and the avaliable annotations across all species and with *D. melanogaster*. Consider things like number of genes/ number of GO IDs/range of GO IDs annotated. Are the different species enriched for different GO terms? If so can you thing of a reason why?

Help and Guidance

Connecting to the server

You can connect to the server using secure shell (SSH)
From the linux command line you connect with the following command.
ssh <username>@subliminal.maths.cam.ac.uk
Windows users can use PUTTY `http://www.chiark.greenend.org.uk/~sgtatham/putty/`
There are plenty of HOW-TO guides online.

Using (or being) nice

To promote the sharing of resources when multiple users are trying to use the same server you can add:
nice -n x
where x is a number from 1 to 19 (1 highest priority) and the server will assign CPU time based on the priorities. This means you can run your programs using all 64 cores of subliminal (where possible) without worrying about preventing other people using the server and your programs will automatically fill the avaliable CPUs if other programs finish. This only works well if EVERYONE uses it. I would suggest you all use the same priority, say 5.

SFTP - SSH file transfer protocol

SFTP can be used to move files to and from the server.

GNU Screen

Screen is a terminal multiplexer which allow processes to continue running even after the client disconnects from the server. You can then reconnect to the screen session when you next login. After SSHing into the server, type 'screen' to start a new screen session. There are numerous screen tutorials online to help you use its full functionality.

Obtaining Data

Sequence data and annotation (where avaliable) for each species may be obtained from the list of websites found at the end of Lecture 5.

Your may use any databases you choose, but for convenience Pfam and Swissprot databases are found at: /local/data/public/genome_informatics/assignment_2

Bioinformatics Software

You may use any programs you like. The following programs are avaliable in
/local/data/public/genome_informatics/programs/ :
BLAST
genscan
HMMer

n.b. genscan does not take input fasta larger than 2Mb or use multiple sequences per file so you will need to run each scaffold separately. In some cases you may need to split scaffolds.