CS31810 Assignment 2020-2021

Deadline: Weds 6th January at 1pm (Blackboard submission)

1 Submission and Assessment

This assignment is worth **100%** of the marks for CS31810. You should submit a **single zip file** containing your article and any supplementary files. The article must be in PDF format, with contents as described below. The zip file is to be submitted via Blackboard before 1pm on 6th January 2021. Do **not** submit this to PeerJ!

By submitting via Blackboard, you are implicitly declaring the work to be your own.

The assignment must be in your own words; it is not acceptable to construct your assignment by copying and pasting chunks of text from the web. Please take note of the information on Unacceptable Academic Practice which can be found at `https://www.aber.ac.uk/en/aqro/handbook/regulations/uap/`. It is important to indicate clearly in your own work where you have included the work of others. In Computer Science this could include reuse of designs and code as well as copying or quoting text.

Marking will be according to the assessment criteria for essays (appendix AC at `https://impacs-inter.dcs.aber.ac.uk/images/editor-content/Documentation/Handbooks/Appendices/AppendixAC.pdf`).

Feedback will be returned on or before 27th January 2021.

In case of circumstances affecting this coursework, please provide a special circumstances form `https://www.aber.ac.uk/en/aqro/exams/special-circumstances/` to fbrstaff@aber.ac.uk.

If you have specific questions relating to the assignment itself, please contact Amanda Clare, afc@aber.ac.uk.

2 Summary of the assignment

The assignment involves the analysis of some genomes that have been constructed from sequence data from metagenomics. You are given four genomes, by email. You are to analyse your genomes using your choice of analysis tools and methods, and to write a short paper about your findings.

3 The data

You are given four of the genomes that were produced from the following study:

Stewart *et al.* (2018) Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nature Communications 9:870[1].

The genomes are incomplete, and consist of multiple contigs/scaffolds. The set of contigs/scaffolds for each genome is provided as a single FASTA file and will be emailed to each of you individually.

4 The paper

You will write a short paper about your findings. The paper will be written using the formatting standards of the PeerJ journal (`https://peerj.com/about/author-instructions/`), as discussed in class, with a strict page limit of four pages (this is likely to be less than 3000 words). Please use the PeerJ template (Latex or Word), and convert the final result to PDF before submission.

[1] The paper is available at `https://www.nature.com/articles/s41467-018-03317-6`, and the genomes data was downloaded from `https://datashare.is.ed.ac.uk/handle/10283/3009`

Your paper should include an abstract of no more than 500 words (fewer is okay, but it should summarise your paper). It should then contain the following sections: Introduction, Materials & Methods, Results, Discussion, Conclusions. The marks will be awarded as follows: Abstract (10%), Introduction (10%), Materials and Methods (20%), Results (30%), Discussion and Conclusion (20%). The final 10% will be awarded for clear writing, suitable presentation and overall coherence.

You may include supplementary data in addition to the page limit in extra files. Supplementary data may include tables, figures, code or data (for example, GFF files or CSV files). Each supplementary data file should be clearly named, and clearly referenced in the paper (for example "See Supplemental Table S1 for full details regarding ORF length distribution", with the corresponding table in a file called Supplemental_Table_S1.pdf).

5 The analysis

The analysis must contain at a minimum a statistical summary describing the number of contigs, length of contigs and GC content for each of your four genomes. After this, the choice of analysis is left to you, and could concentrate in detail on one genome only, or could compare two, or all four. Some examples are given below. Mixing more than one of the list below is allowed and, indeed, this list should not limit your explorations:

- Give a good k-mer composition analysis. Are the four genomes distinguishable by their k-mer composition? Are the contigs distinguishable? Are they more similar to *E. coli* or to *Bacillus subtilis* or a selection of other genomes?
- Annotate your genome(s) for ORFs and/or other interesting features. What kinds of ORF are present? What properties do they have? Is this expected?
- (Possibly harder) Compare the behaviour of two or more existing bioinformatics tools on your genomes, in terms of speed, memory usage, ease of use, documentation, features and results output. What could be improved, if anything for these tools?
- (Harder) Find out what genus/species are your genomes. Can you determine the species with any confidence? What literature is available that describes these species and how is it applicable to your genome?
- (Harder) Provide a detailed comparison of your genomes with the closest well-annotated reference genomes available (does it contain the same genes? which parts are missing?).
- (Harder) Test out some newly available bioinformatics software or methodology on your genomes - take a recent bioRxiv paper (`https://www.biorxiv.org/`) or Bioinformatics journal paper (`https://academic.oup.com/bioinformatics`) and try out what the authors recommend. Was it good or not so good (and why)?

The difficulty of the analysis that you have attempted, and the thoroughness of your work will both be taken into account when marking. A simple analysis can score highly, if completed very well. A more challenging analysis that is incomplete but has a good discussion of the limitations can also score well. The assignment is expected to take you around 40 hours work.