MSc ASSIGNMENT

BIO720P: AI & DATA SCIENCE IN BIOLOGY 2023

What you need to do

The aim this exercise is to explore and understand the effects of dengue fever on human gene expression. Our starting point for the assignment is the transcriptomics dataset we used in the exploratory data analysis practical, i.e. GEO dataset GDS5093, which contains gene expression data from four populations: (a) patients infected with dengue fever; (b) patients with dengue haemorrhagic fever; (c) patients recovering from dengue fever and (d) healthy controls.

Specifically, you need to use this dataset to answer the following research questions:

- To what extent does the gene expression profile of a patient differ between the four populations? You should use exploratory data analysis methods for this.
- Which, if any, individual genes differ significantly in expression between the four disease states? Volcano plots would be useful here. If there are differences, what do they mean biologically?
- From a clinical treatment perspective, it would be useful if transcriptomics data could be used to distinguish between patients with dengue fever and dengue haemorrhagic fever. Is it possible to do this using machine learning?

To be clear, we want you to apply exploratory, statistical and machine learning methods to the dengue dataset in order to answer these questions. The practical sessions covered all the data analysis skills necessary to complete this assignment. The assignment gives you an opportunity to cement your understanding of the topics covered, finish parts of the practicals that you didn't previously have time for and expand on the work you have already done during the module.

> ⓘ If you have any questions about the assignment, please post these on the discussion forum for this module on QMplus so that everyone gets the benefit of any answers that are posted.

What you need to submit

We want you to produce a scientific report that describes the analysis you have done, and discusses what you have discovered. In future this report should act as an *aide memoire* of skills learned in this module and be something you can show potential employers or PhD supervisors to demonstrate your proven and wide-ranging expertise[1].

Your report should have the following sections:

Abstract: A short (no more than half a page) summary of the work, including the aim, summary of methods, results and your conclusion.

[1] Note that we do not keep your work (e.g. on JupyterHub or Posit Cloud) indefinitely so you should keep your own copies of anything you may want to use in future. To avoid future cheating, please don't make your assignment work available online.

Introduction: Explain the aim of the work, and what your starting point was (in this case exercise explain the nature of the samples and the analytical method(s) used to collect the data). Keep this brief: no more than 1-2 paragraphs.

Methods: Information about the data and how you analysed it. This should have as its core a textual explanation, with diagrams and/or equations if necessary. You do not need to include extensive background explanations about common methods such a PCA, random forests and performance metrics if they are already explained well elsewhere (e.g. in papers, books or online documentation).

Results and discussion: This is where you put your results (e.g. tables, figures) and explain the key scientific findings from these.

Conclusion: A single paragraph describing your key findings and their significance.

> ⓘ A data science report has the same structure as any other scientific report so you can use your previous scientific writing experience to decide what goes where.

Submission process

You must submit two assignment files in PDF format via the QMPlus page for this module by the date and time specified there. The files you need to submit are:

- Your report as a PDF file. Use *your* surname as the name of the file, e.g. `sanchez.pdf` if you're called Rick Sanchez.
- Submit a PDF of the code (e.g. Jupyter Notebook) you wrote to produce your results – this should be in separate file named like `sanchez_code.pdf`.

Marking scheme

You will be awarded a single mark for this assignment, based on the BIO720P report marking guidance, which is based on the SBBS MSc dissertation marking guidance. Some advice for getting high marks:

- In the Methods section, we will be looking for the use of appropriate methods and parameters, and justifications for these, explained accurately, clearly and concisely.
- Results should be professionally presented (e.g. the right type of plot or table with appropriate labelling, readable font size, legend, etc.) and explained clearly and concisely.
- Discussion should focus on the biological findings of your analysis, within the context of existing literature, but should also note any caveats or limitations of your analysis.

There is no target length for these reports, so you need to use your judgement to decide how to get the necessary information across clearly and concisely. This applies to both the text and figures – think carefully about which figures are worth including. Marks will not be given for your Python code *per se*, but we may use this to understand how you arrived at your results and it may be referred to if academic misconduct is suspected.

Last updated 7 December 2023 by Conrad Bessant
`c.bessant@qmul.ac.uk`