MSc BIOINFORMATICS ASSIGNMENT

BIO725P: POST GENOMIC BIOINFORMATICS

Background

The aim of this assignment is to produce a portfolio of work based on three of the hands-on practical exercises that you did during the module. This 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 that you have already done. In future this portfolio should act as an *aide memoire* of skills learned in this module, and be something to show potential employers or PhD supervisors to demonstrate your proven and wide-ranging expertise in post genomic bioinformatics.

For each exercise, you should write a report with the following sections:

Introduction: Explain the aim of the work, and what your starting point was (e.g. in a data analysis 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 – otherwise you will mostly be repeating what was written in the instructions.

Methods: How you analysed the data. This should have as its core a textual explanation, with diagrams and equations if necessary.

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 findings and their significance.

The exercises that you need to write up are described below, together with the percentage of total marks that they carry. In each case we have picked out specific aspects of the associated practical that we would like you to cover and, in some cases, added a bonus task.

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.

Marking scheme

In terms of the methods section, high marks will be awarded to reports in which appropriate methods and parameters have been applied, and the justifications for using them have been explained accurately, clearly and concisely. Results should be presented appropriately (e.g. the right type of plot or table with appropriate labelling, readable font size, etc.) and explained clearly and concisely. High marks will be awarded to those reports that meet these requirements and also demonstrate a full understanding of the problem tackled and provide critical insight into the results obtained. To achieve the very highest marks you will need to have completed the bonus tasks.

Submission process

You must submit your assignment files via the QMPlus page for this module.

Please submit all three reports in a single PDF file, each starting on a new page. Use *your* surname as the name of the file, e.g. `farnsworth.pdf`.

Exercise 1: Exploratory analysis of transcriptomics data (35%)

For this exercise you worked on GEO dataset GDS5093, which contained 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. Use R to visualise the relationships between the gene expression profiles of the four populations and identify the genes that show the most significant difference in expression between those populations. (You must include the core R code used to produce each visualisation – if any individual piece of code exceeds half a page you should put it in an appendix.)

Bonus Task: Perform functional enrichment analysis to explain the biological significance of the virus-induced differences in gene expression. If you complete this task, be sure to include your methods and results in the relevant parts of your report.

Exercise 2: Machine learning for disease diagnosis (40%)

In this exercise you were given GC-MS data collected from patients diagnosed with one of three different gut diseases (IBS, ulcerative colitis or Crohn's disease) and a group of healthy controls. For each patient you have data from breath, blood, urine and faeces. Using the classyfire R package, you need to rank these sample types according to their usefulness to diagnose Crohn's disease. Think carefully about what you use as input to classyfire, and how to choose the optimum parameters (e.g. number of ensembles and bootstraps to use) – you will need to justify these choices. You do not need to include any code for this exercise, unless you've done something particularly innovative that you'd like to highlight.

Based on your results, which sample type would be most useful for diagnosing Crohn's disease and why? Among the samples of this type, are there any individual samples that are particularly difficult to classify correctly?

You have more time to work on this than you had in our practical session, so can run more complex models, and more of them. However, be aware that the RStudio Cloud project has limited resources (4 CPU cores) and a maximum background execution time of six hours. You may get better performance running RStudio on your own computer – as part of the setup for this you would need to download and install the classyfire package from the Day 4 section of the module page.

Bonus Task: Random forests are a popular alternative to SVMs for building classification models. For the sample type that you found most diagnostic, use the R package *caret* (https://topepo.github.io/caret/) to see whether random forests can outperform classyfire's SVM models in terms of classification accuracy. If you complete this task, be sure to include your methods and results in the relevant parts of your report.

Exercise 3: Investigating the human kinase network (25%)

In this exercise, you were asked to generate a kinase signalling network from information held in the SIGNOR database. Within the context of the overall report structure described earlier you should briefly (a) explain the main steps used to build and visualise the network, (b) show the kinase network produced as clearly as possible and (c) answer the biological questions from Task 4 of the practical instructions and explain the reasoning behind your answers. You do not need to include any Python code.

In addition, you should plot the degree distribution of the network – how does this compare with the degree distribution of networks found in other areas of biology and elsewhere? Are any particular network motifs overrepresented in your kinase interaction network?