15 January 2021

MED3007 «Statistical Principles in Genomics: an Introduction with Rstudio»

Take-home exam

This take-home exam is in the form of a data analysis project for the leukaemia gene expression data set by Golub *et al.* (1999).

Background and description of the exam data set (MED3007_exam_data.Rdata):

This dataset comes from a proof-of-concept study published in 1999 by Golub *et al.* It was one of the first studies to show that it is possible to classify new cases of cancer based on their gene expression data (via DNA microarray) and thereby provided a general approach for identifying new cancer classes and assigning tumours to known classes.

The data set contains 47 patients with acute lymphoblastic leukaemia (ALL) and 25 patients with acute myeloid leukaemia (AML). The samples were assayed using Affymetrix Hgu6800 chips and data on the expression of 7129 genes (Affymetrix probes) are available in the matrix called “expr”. The data frame “clin” contains additional information about the patients.

Reference:
Golub *et al.* (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 531-537.

Exam task:

The task of this exam is quite open: to analyse the data set with any of the statistical approaches discussed in class using R and RStudio and summarise the results as plots and / or tables. Use of the gene expression matrix (“expr”) is mandatory; the use of the “clin” data frame is optional.

In addition to the analysis results, please provide all the R code and an appropriate documentation of analysis plans and motivations as well as a brief discussion/ interpretation of the results. This can be done either as a reproducible source code file (.R, .Rmd, ...) which is appropriately commented, or it can be done as a document file (.doc, .pdf, ...) which combines all the analysis details, analysis results and discussion, and the source code.

The requirement for passing the exam is that the analysis is correctly motivated and implemented and that results and R code are shown. Please note that there is no right or wrong answer here, since genomics data will typically be analysed in various ways depending on the research question.

Language:

The language for this exam is English.

Time plan:

Friday, 15 Jan 2021 (13:00)
Distribution of the homework exam assignment.

Tuesday, 19 Jan 2021 (13:00)
Deadline for handing in of homework exam in Inspera.