Practical – Microarray analysis – Data loading and exploration

(STAT2)

Jacques van Helden

2020-03-06

Contents

Introduction 1
Study case 1
Data pre-processing 1
Availability of the pre-processed data 1
Data download 2
Data loading 2
Marginal statistics 2
Empirical distributions 2
Sample classes 3

Introduction

In this practical, we will load a dataset that will be used as study case to apply different approaches of multivariate analysis:

- data exploration
- multidimensional scaling
- differential analysis
- clustering (unsupervised classification)
- supervised classification

Study case

Reference: Den Boer ML *et al.* (2009). A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol. 2009 10:125-34.

- **DOI**: [doi: 10.1016/S1470-2045(08)70339-5]
- **Pubmed**: [PMID 19138562].
- **Raw data** available at Gene Expression Omnibus, series [GSE13425]
- **Preprocessed data**: https://github.com/jvanheld/stat1/tree/master/data/DenBoer_2009.

Data pre-processing

The raw microarray data has been pre-processed in order to dispose of a ready-to-use dataset. pre-processing included

- filtering of barely detected or poorly expressed genes,
- log2 transformation to normalise the raw measurements
- between-sample standardisation to enable comparison between the different samples.

Availability of the pre-processed data

The preprocessed data is available here: https://github.com/jvanheld/stat1/tree/master/data/DenBoer_2009.

It contains the following files.

File	Contents	Structure
GSE13425_group_descriptions.tsv.gz	Description of the patient groups	Tab-delimited file with one row per group and one column per type of description (group name, label)
phenoData_GSE13425.tsv.gz	Metadata (sample descriptions)	Tab-delimited file with one row per biological sample (one per patient) and one column per type of information about the biological sample
GSE13425_Norm_Whole.tsv.gz	Normalised microarray data	Tab-delimited file with one row per gene and one column per patient
GSE13425_AMP_Whole.tsv.gz	Detection status of each gene in each sample (Absent, Marginal, Present)	Tab-delimited file with one row per gene and one column per patient

Data download

Write an R script that perform the following operations

1. Creates a directory to store a local copy for this practical on your computer. (`~/STAT2_CMB_practicals/den-boer-2009/`
2. Creates a sub-directory for the data (`~/STAT2_CMB_practicals/den-boer-2009/data/`).
3. Lists the files available on the data Web site (https://github.com/jvanheld/stat1/tree/master/data/DenBoer_2009).
4. For each of these files, checkS if it is already present in your local local data directory, and downloads it if it is not the case.

Try to make your code re-usable, in the perspective to apply it soon in order to download data sets from other web sites.

Data loading

Write an R script that loads the data tables from your local data directory.

Marginal statistics

Write an R script that computes marginal statistics (mean, sd, min, percentiles 5, 10, 25, 50, 75, 90, 95, max, IQR) on each row of the normalised data table (one row corresponds to one gene).

Do the same for each column (patient).

Empirical distributions

Draw a plot that displays the empirical distribution of normalised expression values in the whole data table.

Draw another plot that displays the empirical distribution of normalised expression values in the different samples (one polygon frequency per sample).

Compute a table with the mean expression profile per cancer type (one row per gene, one column per cancer type) and draw them with box plots.

Sample classes

Load the pheno table. We will use the three following columns:

- `Sample.GEO.ID`: the identifier of the sample in the Gene Expression Omnibus (GEO) database.
- `Sample.title`: cancer type of each sample.
- `sample.labels`: A short label (1 or 2 letters) for each cancer type

Cound the number of samples per group and draw a barplot with the result.