Ciências ULisboa | Estatística e Investigação Operacional

MÉTODOS ESTATÍSTICOS EM BIOINFORMÁTICA(1)
MÉTODOS ESTATÍSTICOS EM GENÉTICA(2)

2.° semestre – 2023/2024

Trabalho de Avaliação - Parte 2
Prazo: 30 de junho de 2024

(1) individual (85% da nota final do projeto)
(2) grupos de 2 elementos (85% da nota final do projeto)

Choose one question from the following three.

Always justify your answers and interpret the graphs and tables you present. Provide a brief description of the statistical methods applied in each *package* you use.

1. **Microarray Data**

   In an experiment conducted at a prestigious Molecular Cardiology Laboratory of a Brazilian Institute, mRNA samples were obtained from saphenous vein tissues (veins from the inner leg) of patients operated on for cardiac disease. These tissues were maintained in an *ex vivo* culture and subjected to two experimental conditions: arterial and venous regimes. For each patient, the saphenous vein tissue sample was subjected to the two aforementioned experimental conditions. The two resulting samples, for each patient, were hybridized on the same *cDNA microarray* (two channels). A total of 2994 genes were analyzed.

   The file `chip1.txt` refers to the first patient and contains the expression levels of the genes from the tissue subjected to the arterial regime (Art), the genes from the tissue subjected to the venous regime (Ven), and their respective background

values ($BgArt$ and $BgVen$). The Art sample corresponds to the red channel and the Ven sample to the green channel. Subsequently, two more patients were operated on, and two *arrays* were obtained under identical conditions to the first. The intensities are recorded in `chip2.txt` and `chip3.txt`. The intensities in the 3 files are not normalized!

Compile the data from each file into a single data matrix and answer the following questions:

- Normalize the data using the *package* `limma`. Perform this task as completely as possible, making use of the graphical representations available for this purpose.
- Apply the Bayesian method of Lönnstedt and Speed (*package* `limma`) to the normalized data. Establish some command lines to identify the genes with differential expression and justify the cut-off point considered.
- Still using the *package* `limma`, with the normalized data, re-identify the genes with differential expression, this time using the moderated t-statistic. Set a significance level that seems appropriate to you.
- Now consider the *package* `RankProd` to identify differentially expressed genes. The analysis should be carried out on the same normalized data, setting an appropriate FDR.
- Compare the results obtained by the three methods and comment.

2. **RNA-Seq Data**

In this exercise, the student must describe and demonstrate how to handle an RNA-Seq database using the indicated *packages*. The following points should be considered:

Data

In addition to read counts, the database should contain information about gene length and/or GC content. The options for the database to be used are:

- Swirl Zebrafish, through the *package* `zebrafishRNASeq`;
- Any other dataset, provided it is not included in the databases of the *packages* `limma`, `degSeq`, `edgeR`, `edaseq`.

Methodology 1

- Normalization: use the *package* `edaseq`;
- Identify differentially expressed genes using the *package* `edgeR`.

Methodology 2

- Normalization: use the `voom` function from the *package* `limma`;
- Identify differentially expressed genes using the *package* `limma`.

Methodology 3

- Normalization: use the `voom` function from the *package* `limma`;
- Identify differentially expressed genes using the *package* `RankProd`.

Compare the results obtained by the three methods and comment.

3. **Microarray, RNA-Seq or other Omics Data**

The student must choose a *package* from `Bioconductor`, which fits within the scope of the subject taught. This choice requires prior authorization from the professor responsible for the course.

- Prepare a report containing a description of the statistical methodology presented in the article that underpins the selected *package*;
- Apply the selected *package* to a dataset, which may be (1) a dataset available on the web or in other *packages*, own data, etc., or (2) simulated data."

Bom Trabalho!