**BS9001 Research Experience Laboratory Report**

**Meta-analysis of schizophrenia using data science approaches**

**Done by: Austin Chia Cheng En (U1740366A)**

**Abstract**

In this project, the first few steps in the meta-analysis of post-mortem prefrontal cortex microarray data are covered. The datasets are extracted from NCBI dataset browser through GEOquery using the R programming language. The datasets are cleaned and compressed for preparation for statistical analysis. The genefilter package is used for the rowttests function in data analysis. A confusion matrix is made to better determine the feasibility of the statistical t-tests. Jaccard coefficients are calculated for the different genes. Further research and analysis of the Jaccard coefficients have to be done before any substantial findings can be confirmed.

**Introduction**

**Meta-analysis**

The statistical analysis of a large collection of analysis results from different separate studies to integrate their conclusion. It is used to overcome size limit or scope in individual studies for reliable information about treatment effects. It combines the results of the different findings into a new conclusion. Meta-analysis is typically used in small trials to provide essential information on treatment effects to avoid the delay and expense of a large scale randomized clinical trial. It can also be useful for clinical trials that would be deemed impractical or unethical.

A meta-analysis usually begins with selecting appropriate data, a dataset that has lesser “NA” or missing values. This data is then compressed with expression data. Once compressed, the dataset is ready for statistical analysis.

In summary, a meta-analysis is an important and valuable tool for summarizing data from multiple studies. However, it is not an easy task and requires careful thought and planning to provide accurate and useful information.

**Bioconductor**

Bioconductor is an open-source package that contains tools for the analysis and comprehension of microarray data. It is only available on the R programming language. Bioconductor contains the “GEOquery” library that contains useful functions that allow the extraction of GDS datasets from NCBI, such as the “getGEO” function.

**Method**

**Sourcing for suitable datasets using the NCBI browser**

The terms “schizophrenia” and “homo sapiens” were entered into the search query function on the NCBI dataset browser. Datasets that involved studies of schizophrenia in homo sapiens were chosen. The datasets are further narrowed down after each data is explored. Incomplete and messy datasets were removed from inclusion in the study.

**Compression and binding**

The "GEOquery" library was imported first. The getGEO function from the GEOquery package was used to store the GDS dataset into a vector The GDS vector is then converted into an expression set through the use of the "GDS2eSet" function. The expression data is stored into a new vector. The third column names were renamed "Gene\_Symbol" and its feature data was stored into a new vector The expression data was stored into a dataframe. The expression data was bound to the respective genes.

The bound dataframe was cleaned from NA values and empty fields using the `is.na` function and unique gene symbols are subsetted using the "unique" function. An empty list is initialized for the loop output. A for loop is used to select out each row in the expression matrix and add into the initialized list. The `colMeans` function was applied onto the list and stored as a dataframe. The dataframe is saved by writing into a csv file.

**Bootstrapping**

All the necessary libraries are first imported. `readr` for writing dataframes into csv files and `genefilter` for the `rowttests` function. The metadata of the GDS was accessed and columns 1 to 12 are for the control group and columns 38 to 49 are for the disease class (schizophrenia). The expression data of both groups were subsetted. A progress bar was first added through the "progress" package. A factor with the corresponding "normal" and "disease" classes was defined. Working input dataframe was coerced into a matrix for the rowttest function. An empty list was initialized to store output data from the loop. The nested list output derived from the `for` loop was coerced into a dataframe.

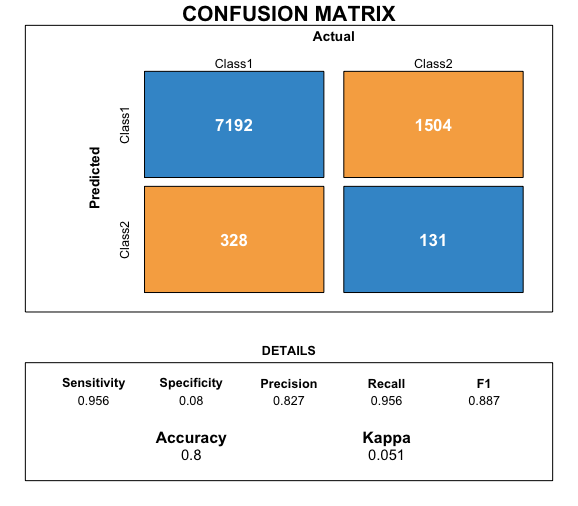
**Jaccard coefficient test**

A Jaccard coefficient function was defined. A progress bar was included to view run progress. The Jaccard coefficients were assigned into a table through the use of a nested for loop. The bottom half of the table is not filled up, as the values will be a mirror image of those at the top half.

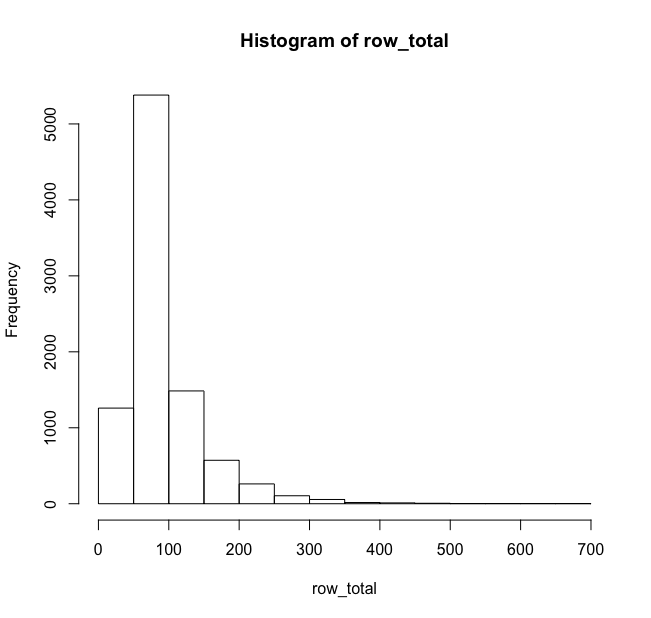
**Confusion matrix**

Based on the bootstrap, each gene was then determined if it was significant. `caret` package was used and a confusion matrix was made using the `confusionMatrix` function. The matrix was drawn using a user-defined function.

**Results**

****

****

****

****

**Discussion**

**Data processing and cleaning**

Other methods for cleaning and processing data can be used instead of the method used. Data imputation methods should be used if any further analysis of data is required.

**Histogram of distribution of gene significance**

This histogram was used as a benchmark to determine the significance of a gene. As a cut-off point of 200 out of 1000 was chosen, a confusion matrix might show a falsely positive result. This might potentially lead to errors in any conclusive findings.

**Confusion Matrix**

Based on the confusion matrix plotted, the t-tests that were done across the 2 data groups were largely high in precision and recall. However, these findings may not be representative of true gene significance. More confusion matrices should be made to better determine a more suitable cut-off point for gene significance.

**Further recommendations**

Further analysis and visualization of the Jaccard test has to be done to better represent the significance of the data chosen.

**References**

Berman NG, Parker RA. Meta-analysis: Neither quick nor easy. BMC Medical Research Methodology. 2002;2(1). doi:10.1186/1471-2288-2-10

Nakagawa S, Noble DWA, Senior AM, Lagisz M. Meta-evaluation of meta-analysis: ten appraisal questions for biologists. BMC Biology. 2017;15(1). doi:10.1186/s12915-017-0357-7

Tutorial 9 (Biological Databases: Ensembl). 2019