Introduction into machine learning and analyzes of Breast Cancer Proteomes

true

June 15th, 2022

Contents

Dataset

information about the data set and the three give files :

About Dataset Context: This data set contains published iTRAQ proteome profiling of 77 breast cancer samples generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). It contains expression values for ~12.000 proteins for each sample, with missing values present when a given protein could not be quantified in a given sample.

Content:

- **File:** 77cancerproteomesCPTACitraq.csv
  - **RefSeqaccessionnumber:** RefSeq protein ID (each protein has a unique ID in a RefSeq database)
  - **gene_symbol:** a symbol unique to each gene (every protein is encoded by some gene)
  - **gene_name:** a full name of that gene
  - **Remaining columns:** log2 iTRAQ ratios for each sample (protein expression data, most important), three last columns are from healthy individuals
- **File:** clinicaldatabreast_cancer.csv
  - **First column** "Complete TCGA ID" is used to match the sample IDs in the main cancer proteomes file (see example script).
  - **All other columns** have self-explanatory names, contain data about the cancer classification of a given sample using different methods. 'PAM50 mRNA' classification is being used in the example script.
- **File:** PAM50_proteins.csv
  - **Contains** the list of genes and proteins used by the PAM50 classification system. The column RefSeqProteinID contains the protein IDs that can be matched with the IDs in the main protein expression data set.

Past Research: The original study: *http://www.nature.com/nature/journal/v534/n7605/full/nature18003.html* (paywall warning)

In brief: the data were used to assess how the mutations in the DNA are affecting the protein expression landscape in breast cancer. Genes in our DNA are first transcribed into RNA molecules which then are translated into proteins. Changing the information content of DNA has impact on the behavior of the proteome, which is the main functional unit of cells, taking care of cell division, DNA repair, enzymatic reactions and signaling etc. They performed K-means clustering on the protein data to divide the breast cancer patients into sub-types, each having unique protein expression signature. They found that the best clustering was achieved using 3 clusters (original PAM50 gene set yields four different subtypes using RNA data).

```
proteomes_data <- read.csv(file = "Data//77_cancer_proteomes_CPTAC_itraq.csv")
clinical_data <- read.csv(file = "Data/clinical_data_breast_cancer.csv")
pam50_protein_data <- read.csv(file = "Data/PAM50_proteins.csv")

head(proteomes_data[1:5], n = 5)
```

```
##   RefSeq_accession_number gene_symbol          gene_name A0.A12D.01TCGA
## 1               NP_958782        PLEC  plectin isoform 1       1.096131
## 2               NP_958785        <NA> plectin isoform 1g       1.111370
## 3               NP_958786        PLEC plectin isoform 1a       1.111370
## 4               NP_000436        <NA> plectin isoform 1c       1.107561
## 5               NP_958781        <NA> plectin isoform 1e       1.115180
##   C8.A131.01TCGA
## 1       2.609943
## 2       2.650422
## 3       2.650422
## 4       2.646374
## 5       2.646374
```

position the caret at any line or the code chunk, then click "+".

The code chunk appears:

Type any R code in the chunk, for example: