

mrr_project_bio

2025-11-10

Classification of Cancer outcome using Genetic and Clinical data

Introduction

This project investigates the effect of genetic and clinical variables on the survival outcome of breast cancer patients. The data contains records of more than 1000 breast cancer patients from several research institutions. Clinical data contains patient-related and tumor-related information. Additionally mRNA gene expression data is available for each patient. The gene expression data has been processed to include only the top 5000 most variable genes on the transformed scale ($\log_2(\text{counts} + 1)$).

We consider that the main outcome variable of interest is **vital_status**, which is defined in clinical data.

Load data

```
load("mrr_bio.Rdata")
```

Genes data

```
dim(GeneX)
```

```
## [1] 1231 5000
```

```
str(GeneX)
```

```
## num [1:1231, 1:5000] 0 0 10.92 3.17 5 ...
## - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:1231] "EW-A2FS-01A-11R-A17B" "OL-A6VR-01A-32R-A33J" "E9-A226-01A-21R-A157" "A8-A08H-0...
##   ..$ : chr [1:5000] "CLEC3A" "SCGB2A2" "CPB1" "GSTM1" ...
```

```
colnames(GeneX)[1:10]
```

```
## [1] "CLEC3A"      "SCGB2A2"      "CPB1"        "GSTM1"       "TFF1"        "SCGB1D2"
## [7] "KCNJ3"       "MUCL1"       "LINC00993"   "ANKRD30A"
```

```
rownames(GeneX)[1:10]
```

```

## [1] "EW-A2FS-01A-11R-A17B" "OL-A6VR-01A-32R-A33J" "E9-A226-01A-21R-A157"
## [4] "A8-A08H-01A-21R-A00Z" "D8-A27H-01A-11R-A16F" "D8-A3Z6-01A-11R-A239"
## [7] "B6-A1KN-01A-11R-A13Q" "BH-A0DL-01A-11R-A115" "A8-A09X-01A-11R-A00Z"
## [10] "BH-A2L8-01A-11R-A18M"

```

GeneX est un tableau de 1231 lignes par 5000 colonnes :

- Les lignes correspondent aux patients avec le nom de chaque ligne l'identifiant du patient comme EW-A2FS-01A-11R-A17B.
- Les colonnes correspondent aux gènes avec chaque nom de colonne un gène comme CLEC3A, qui sont les 5000 gènes les plus variables d'une personne à l'autre.
- Chaque case correspond à la valeur d'expression (un entier) d'un certain gène (colonne) pour un certain patient (ligne)

Clinical data

```

load("mrr_bio.Rdata")
dim(clinical_data)

```

```
## Loading required namespace: S4Vectors
```

```
## NULL
```

```
head(clinical_data[, 0:5])
```

```

## DataFrame with 6 rows and 5 columns
##           initial_weight tissue_type laterality
##           <numeric> <character> <character>
## EW-A2FS-01A-11R-A17B      110     Tumor      Left
## OL-A6VR-01A-32R-A33J      380     Tumor     Right
## E9-A226-01A-21R-A157      510     Tumor      Left
## A8-A08H-01A-21R-A00Z      250     Tumor      Left
## D8-A27H-01A-11R-A16F      170     Tumor     Right
## D8-A3Z6-01A-11R-A239       60     Tumor      Left
##           tissue_or_organ_of_origin age_at_diagnosis
##           <character> <integer>
## EW-A2FS-01A-11R-A17B      Breast, NOS        15302
## OL-A6VR-01A-32R-A33J      Breast, NOS        17702
## E9-A226-01A-21R-A157      Breast, NOS        16511
## A8-A08H-01A-21R-A00Z      Breast, NOS        24137
## D8-A27H-01A-11R-A16F      Breast, NOS        26588
## D8-A3Z6-01A-11R-A239      Breast, NOS        20629

```

```
colnames(clinical_data)
```

```

## [1] "initial_weight"          "tissue_type"
## [3] "laterality"             "tissue_or_organ_of_origin"
## [5] "age_at_diagnosis"       "primary_diagnosis"
## [7] "prior_treatment"        "diagnosis_is_primary_disease"

```

```

## [9] "ajcc_pathologic_t"           "morphology"
## [11] "classification_of_tumor"    "sites_of_involvement"
## [13] "days_to_last_follow_up"     "follow_ups_disease_response"
## [15] "race"                      "gender"
## [17] "ethnicity"                 "vital_status"
## [19] "age_at_index"              "days_to_birth"
## [21] "age_is_obfuscated"         "bcr_patient_barcode"
## [23] "primary_site"              "disease_type"

```

`clinical_data` est un tableau de 1231 lignes par 24 colonnes :

- Les lignes correspondent aux noms des patients comme pour `GeneX`.
- Les colonnes sont les noms des données cliniques, comme `initial_weight`, `sites_of_involvement` et `vital_status`.
- Chaque case correspond à la valeur de la donnée clinique pour chaque patient, donc ce peut être un entier ou une chaîne de caractères.

Outcome variable

The outcome variable is the vital status of the patient (dead or alive) :

```

y <- clinical_data$vital_status
table(y)

```

```

## y
## Alive  Dead
## 1029   201

```

Exploratory data analysis: example

```

# summary statistics of first 10 genes
apply(GeneX[,1:10], 2, summary)

```

```

##          CLEC3A  SCGB2A2      CPB1      GSTM1      TFF1  SCGB1D2      KCNJ3
## Min.    0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
## 1st Qu. 1.000000 7.48779 4.523562 0.000000 6.679463 5.614710 2.807355
## Median  4.523562 11.70434 7.686501 5.491853 10.796040 9.273796 5.930737
## Mean    5.925758 11.12788 8.313829 5.357693 9.712977 9.026789 6.687515
## 3rd Qu. 10.175539 15.06285 11.151939 10.344845 13.207240 12.274950 10.321917
## Max.   19.455835 22.36921 22.745581 15.874141 18.927850 21.479643 16.672715
##          MUCL1 LINC00993 ANKRD30A
## Min.    0.000000 0.000000 0.000000
## 1st Qu. 5.894767 5.643856 5.554589
## Median  8.945444 10.512740 10.338736
## Mean    9.151482 8.864887 8.650306
## 3rd Qu. 12.404866 12.230770 11.990273
## Max.   20.547814 15.914922 16.758184

```

```
# clinical variables: gender
unique(clinical_data$gender) # unique values

## [1] "female" "male"   NA

table(clinical_data$gender) # frequency

##
##   female    male
##   1217      13
```