Introduction into machine learning and analyzes of Breast Cancer Proteomes

true

September 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: Original research paper: https://www.researchgate.net/publication/303509927_Proteogenomics_connects_somatic_mutations_to_signaling_in_breast_cancer

Summary: 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). my question is are there different ways to categorize subtypes of breast cancer other than the PAM50 method, and define them as benign or malignant?

```r
# packages
library('pander')
```

```r
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")

# showing succeseful loading of data
head(proteomes_data[1:5], n = 5)
```

```
##   RefSeq_accession_number gene_symbol          gene_name AO.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
```

```r
head(clinical_data, n=5)
```

```
##   Complete.TCGA.ID Gender Age.at.Initial.Pathologic.Diagnosis ER.Status
## 1     TCGA-A2-A0T2 FEMALE                                  66  Negative
## 2     TCGA-A2-A0CM FEMALE                                  40  Negative
## 3     TCGA-BH-A18V FEMALE                                  48  Negative
## 4     TCGA-BH-A18Q FEMALE                                  56  Negative
## 5     TCGA-BH-A0E0 FEMALE                                  38  Negative
##   PR.Status HER2.Final.Status Tumor Tumor..T1.Coded Node Node.Coded Metastasis
## 1  Negative          Negative    T3         T_Other   N3   Positive         M1
## 2  Negative          Negative    T2         T_Other   N0   Negative         M0
## 3  Negative          Negative    T2         T_Other   N1   Positive         M0
## 4  Negative          Negative    T2         T_Other   N1   Positive         M0
## 5  Negative          Negative    T3         T_Other   N3   Positive         M0
##   Metastasis.Coded AJCC.Stage Converted.Stage Survival.Data.Form Vital.Status
## 1         Positive   Stage IV   No_Conversion           followup     DECEASED
## 2         Negative  Stage IIA       Stage IIA           followup     DECEASED
## 3         Negative  Stage IIB   No_Conversion         enrollment     DECEASED
## 4         Negative  Stage IIB   No_Conversion         enrollment     DECEASED
## 5         Negative Stage IIIC   No_Conversion           followup       LIVING
##   Days.to.Date.of.Last.Contact Days.to.date.of.Death OS.event OS.Time
## 1                          240                   240        1     240
## 2                          754                   754        1     754
## 3                         1555                  1555        1    1555
## 4                         1692                  1692        1    1692
## 5                          133                    NA        0     133
##   PAM50.mRNA SigClust.Unsupervised.mRNA SigClust.Intrinsic.mRNA miRNA.Clusters
## 1 Basal-like                          0                     -13              3
## 2 Basal-like                        -12                     -13              4
## 3 Basal-like                        -12                     -13              5
## 4 Basal-like                        -12                     -13              5
## 5 Basal-like                          0                     -13              5
##   methylation.Clusters RPPA.Clusters CN.Clusters
## 1                    5         Basal           3
## 2                    4         Basal           4
## 3                    5         Basal           1
## 4                    5         Basal           1
## 5                    5         Basal           1
##   Integrated.Clusters..with.PAM50. Integrated.Clusters..no.exp.
## 1                                2                            2
## 2                                2                            1
## 3                                2                            2
## 4                                2                            2
## 5                                2                            2
##   Integrated.Clusters..unsup.exp.
## 1                               2
## 2                               1
## 3                               2
## 4                               2
## 5                               2
```

```
head(pam50_protein_data, n=5)
```

```
##   GeneSymbol RefSeqProteinID      Species                           Gene.Name
## 1        MIA       NP_006524 Homo sapiens        melanoma inhibitory activity
## 2      FGFR4       NP_002002 Homo sapiens fibroblast growth factor receptor 4
## 3      FGFR4       NP_998812 Homo sapiens fibroblast growth factor receptor 4
## 4      FGFR4       NP_075252 Homo sapiens fibroblast growth factor receptor 4
## 5     GPR160       NP_055188 Homo sapiens      G protein-coupled receptor 160
```

```
# showing structure of dataframe
str(proteomes_data)
```

```
## ’data.frame':    12553 obs. of  86 variables:
##  $ RefSeq_accession_number: chr  "NP_958782" "NP_958785" "NP_958786" "NP_000436" ...
##  $ gene_symbol            : chr  "PLEC" NA "PLEC" NA ...
##  $ gene_name              : chr  "plectin isoform 1" "plectin isoform 1g" "plectin isoform 1a" "plec
##  $ A0.A12D.01TCGA         : num  1.1 1.11 1.11 1.11 1.12 ...
##  $ C8.A131.01TCGA         : num  2.61 2.65 2.65 2.65 2.65 ...
##  $ A0.A12B.01TCGA         : num  -0.66 -0.649 -0.654 -0.632 -0.64 ...
##  $ BH.A18Q.02TCGA         : num  0.195 0.215 0.215 0.205 0.215 ...
##  $ C8.A130.02TCGA         : num  -0.494 -0.504 -0.501 -0.51 -0.504 ...
##  $ C8.A138.03TCGA         : num  2.77 2.78 2.78 2.8 2.79 ...
##  $ E2.A154.03TCGA         : num  0.863 0.87 0.87 0.866 0.87 ...
##  $ C8.A12L.04TCGA         : num  1.41 1.41 1.41 1.41 1.41 ...
##  $ A2.A0EX.04TCGA         : num  1.19 1.19 1.19 1.19 1.2 ...
##  $ AO.A12D.05TCGA         : num  1.1 1.1 1.1 1.1 1.09 ...
##  $ AN.A04A.05TCGA         : num  0.385 0.371 0.371 0.378 0.375 ...
##  $ BH.A0AV.05TCGA         : num  0.351 0.367 0.367 0.361 0.371 ...
##  $ C8.A12T.06TCGA         : num  -0.205 -0.162 -0.167 -0.184 -0.167 ...
##  $ A8.A06Z.07TCGA         : num  -0.496 -0.499 -0.496 -0.492 -0.488 ...
##  $ A2.A0CM.07TCGA         : num  0.683 0.694 0.698 0.687 0.687 ...
##  $ BH.A18U.08TCGA         : num  -0.265 -0.252 -0.252 -0.252 -0.252 ...
##  $ A2.A0EQ.08TCGA         : num  -0.913 -0.928 -0.928 -0.932 -0.928 ...
##  $ AR.A0U4.09TCGA         : num  -0.0332 -0.0302 -0.0272 -0.0302 -0.0302 ...
##  $ AO.A0J9.10TCGA         : num  0.02 0.012 0.012 0.0039 0.012 ...
##  $ AR.A1AP.11TCGA         : num  0.461 0.461 0.461 0.461 0.461 ...
##  $ AN.A0FK.11TCGA         : num  0.974 0.977 0.977 0.97 0.985 ...
##  $ AO.A0J6.11TCGA         : num  0.831 0.857 0.857 0.837 0.865 ...
##  $ A7.A13F.12TCGA         : num  1.28 1.28 1.28 1.28 1.28 ...
##  $ BH.A0E1.12TCGA         : num  0.762 0.762 0.766 0.758 0.766 ...
##  $ A7.A0CE.13TCGA         : num  -1.12 -1.12 -1.12 -1.13 -1.13 ...
##  $ A2.A0YC.13TCGA         : num  0.819 0.815 0.815 0.799 0.819 ...
##  $ AO.A0JC.14TCGA         : num  -0.307 -0.307 -0.307 -0.307 -0.301 ...
##  $ A8.A08Z.14TCGA         : num  0.569 0.569 0.569 0.569 0.569 ...
##  $ AR.A0TX.14TCGA         : num  -0.583 -0.573 -0.567 -0.583 -0.573 ...
##  $ A8.A076.15TCGA         : num  1.87 1.87 1.87 1.86 1.87 ...
##  $ AO.A126.15TCGA         : num  0.196 0.196 0.196 0.219 0.2 ...
##  $ BH.A0C1.16TCGA         : num  -0.518 -0.51 -0.507 -0.518 -0.513 ...
##  $ A2.A0EY.16TCGA         : num  1.17 1.18 1.18 1.17 1.18 ...
##  $ AR.A1AW.17TCGA         : num  0.578 0.582 0.578 0.59 0.586 ...
##  $ AR.A1AV.17TCGA         : num  -0.76 -0.76 -0.749 -0.736 -0.749 ...
##  $ C8.A135.17TCGA         : num  1.12 1.14 1.14 1.14 1.12 ...
##  $ A2.A0EV.18TCGA         : num  0.453 0.473 0.473 0.459 0.473 ...
##  $ AN.A0AM.18TCGA         : num  1.5 1.51 1.5 1.5 1.5 ...
##  $ D8.A142.18TCGA         : num  0.539 0.542 0.542 0.535 0.542 ...
##  $ AN.A0FL.19TCGA         : num  2.46 2.48 2.48 2.46 2.48 ...
##  $ BH.A0DG.19TCGA         : num  -0.206 -0.206 -0.206 -0.215 -0.206 ...
##  $ AR.A0TV.20TCGA         : num  -1.51 -1.53 -1.53 -1.53 -1.51 ...
##  $ C8.A12Z.20TCGA         : num  -0.787 -0.756 -0.756 -0.775 -0.772 ...
##  $ AO.A0JJ.20TCGA         : num  0.757 0.781 0.774 0.764 0.771 ...
##  $ AO.A0JE.21TCGA         : num  0.56 0.563 0.56 0.542 0.56 ...
##  $ AN.A0AJ.21TCGA         : num  -0.428 -0.406 -0.406 -0.406 -0.406 ...
##  $ A7.A0CJ.22TCGA         : num  -1.001 -1.005 -1.005 -0.998 -1.001 ...
##  $ AO.A12F.22TCGA         : num  -1.95 -1.95 -1.96 -1.95 -1.96 ...
##  $ A8.A079.23TCGA         : num  1.05 1.05 1.05 1.06 1.05 ...
##  $ A2.A0T3.24TCGA         : num  0.584 0.581 0.581 0.587 0.587 ...
##  $ A2.A0YD.24TCGA         : num  0.0638 0.0933 0.0845 0.0667 0.0845 ...
##  $ AR.A0TR.25TCGA         : num  -1.1 -1.11 -1.11 -1.1 -1.11 ...
##  $ AO.A03O.25TCGA         : num  1.05 1.06 1.06 1.06 1.06 ...
##  $ AO.A12E.26TCGA         : num  0.265 0.276 0.276 0.278 0.278 ...
##  $ A8.A06N.26TCGA         : num  0.239 0.25 0.244 0.25 0.25 ...
##  $ A2.A0YG.27TCGA         : num  -0.0782 -0.0681 -0.0714 -0.0579 -0.0647 ...
##  $ BH.A18N.27TCGA         : num  1.1 1.1 1.1 1.09 1.11 ...
##  $ AN.A0AL.28TCGA         : num  0.324 0.327 0.327 0.33 0.327 ...
##  $ A2.A0T6.29TCGA         : num  0.794 0.818 0.815 0.801 0.818 ...
##  $ E2.A158.29TCGA         : num  -1.09 -1.1 -1.1 -1.1 -1.1 ...
##  $ E2.A15A.29TCGA         : num  2.18 2.18 2.18 2.18 2.18 ...
##  $ AO.A0JM.30TCGA         : num  1.4 1.41 1.41 1.41 1.41 ...
##  $ C8.A12V.30TCGA         : num  0.674 0.689 0.689 0.678 0.689 ...
##  $ A2.A0D2.31TCGA         : num  0.1075 0.1042 0.1075 0.0975 0.1042 ...
##  $ C8.A12U.31TCGA         : num  -0.482 -0.478 -0.482 -0.471 -0.482 ...
##  $ AR.A1AS.31TCGA         : num  1.22 1.22 1.22 1.2 1.22 ...
##  $ A8.A09G.32TCGA         : num  -1.52 -1.51 -1.51 -1.52 -1.51 ...
##  $ C8.A131.32TCGA         : num  2.71 2.73 2.74 2.73 2.75 ...
##  $ C8.A134.32TCGA         : num  0.14 0.126 0.133 0.112 0.126 ...
##  $ A2.A0YF.33TCGA         : num  0.311 0.296 0.296 0.296 0.296 ...
##  $ BH.A0DD.33TCGA         : num  -0.692 -0.659 -0.664 -0.657 -0.662 ...
##  $ BH.A0E9.33TCGA         : num  1.47 1.48 1.47 1.46 1.47 ...
##  $ AR.A0TT.34TCGA         : num  -0.511 -0.526 -0.526 -0.533 -0.53 ...
##  $ AO.A12B.34TCGA         : num  -0.964 -0.938 -0.944 -0.935 -0.935 ...
##  $ A2.A0SW.35TCGA         : num  -0.488 -0.488 -0.488 -0.488 -0.504 ...
##  $ AO.A0JL.35TCGA         : num  -0.107 -0.107 -0.107 -0.107 -0.107 ...
##  $ BH.A0BV.35TCGA         : num  -0.0658 -0.0559 -0.0658 -0.0559 -0.0625 ...
##  $ A2.A0YM.36TCGA         : num  0.656 0.658 0.656 0.656 0.651 ...
##  $ BH.A0C7.36TCGA         : num  -0.552 -0.548 -0.552 -0.552 -0.557 ...
##  $ A2.A0SX.36TCGA         : num  -0.399 -0.393 -0.393 -0.393 -0.396 ...
##  $ X263d3f.I.CPTAC        : num  0.599 0.607 0.604 0.604 0.604 ...
##  $ blcdb9.I.CPTAC         : num  -0.191 -0.184 -0.186 -0.186 -0.167 ...
##  $ c4155b.C.CPTAC         : num  0.567 0.579 0.577 0.577 0.577 ...
```

```
str(clinical_data)
```

```
## ’data.frame’:    105 obs. of  30 variables:
##  $ Complete.TCGA.ID                   : chr  "TCGA-A2-A0T2" "TCGA-A2-A0CM" "TCGA-BH-A18V" "TCGA-BH-A
##  $ Gender                             : chr  "FEMALE" "FEMALE" "FEMALE" "FEMALE" ...
##  $ Age.at.Initial.Pathologic.Diagnosis: int  66 40 48 56 38 57 74 60 61 67 ...
##  $ ER.Status                          : chr  "Negative" "Negative" "Negative" "Negative" ...
##  $ PR.Status                          : chr  "Negative" "Negative" "Negative" "Negative" ...
##  $ HER2.Final.Status                  : chr  "Negative" "Negative" "Negative" "Negative" ...
##  $ Tumor                              : chr  "T3" "T2" "T2" "T2" ...
##  $ Tumor..T1.Coded                    : chr  "T_Other" "T_Other" "T_Other" "T_Other" ...
##  $ Node                               : chr  "N3" "N0" "N1" "N1" ...
##  $ Node.Coded                         : chr  "Positive" "Negative" "Positive" "Positive" ...
##  $ Metastasis                         : chr  "M1" "M0" "M0" "M0" ...
##  $ Metastasis.Coded                   : chr  "Positive" "Negative" "Negative" "Negative" ...
##  $ AJCC.Stage                         : chr  "Stage IV" "Stage IIA" "Stage IIB" "Stage IIB" ...
##  $ Converted.Stage                    : chr  "No_Conversion" "Stage IIA" "No_Conversion" "No_Convers
##  $ Survival.Data.Form                 : chr  "followup" "followup" "enrollment" "enrollment" ...
##  $ Vital.Status                       : chr  "DECEASED" "DECEASED" "DECEASED" "DECEASED" ...
##  $ Days.to.Date.of.Last.Contact       : int  240 754 1555 1692 133 309 425 643 775 964 ...
##  $ Days.to.date.of.Death              : int  240 754 1555 1692 NA NA NA NA NA NA ...
##  $ OS.event                           : int  1 1 1 1 0 0 0 0 0 0 ...
##  $ OS.Time                            : int  240 754 1555 1692 133 309 425 643 775 964 ...
##  $ PAM50.mRNA                         : chr  "Basal-like" "Basal-like" "Basal-like" "Basal-like" ...
##  $ SigClust.Unsupervised.mRNA         : int  0 -12 -12 -12 0 0 0 -12 -12 -12 ...
##  $ SigClust.Intrinsic.mRNA            : int  -13 -13 -13 -13 -13 -13 -13 -13 -13 -13 ...
##  $ miRNA.Clusters                     : int  3 4 5 5 5 5 3 5 2 5 ...
##  $ methylation.Clusters                 : int  5 4 5 5 5 5 5 5 5 5 ...
##  $ RPPA.Clusters                        : chr  "Basal" "Basal" "Basal" "Basal" ...
##  $ CN.Clusters                          : int  3 4 1 1 1 1 1 1 1 3 ...
##  $ Integrated.Clusters..with.PAM50.     : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ Integrated.Clusters..no.exp.         : int  2 1 2 2 2 2 2 2 2 2 ...
##  $ Integrated.Clusters..unsup.exp.      : int  2 1 2 2 2 2 2 2 2 2 ...
```

```
str(pam50_protein_data)
```

```
## 'data.frame':    100 obs. of  4 variables:
##  $ GeneSymbol     : chr  "MIA" "FGFR4" "FGFR4" "FGFR4" ...
##  $ RefSeqProteinID: chr  "NP_006524" "NP_002002" "NP_998812" "NP_075252" ...
##  $ Species        : chr  "Homo sapiens" "Homo sapiens" "Homo sapiens" "Homo sapiens" ...
##  $ Gene.Name      : chr  "melanoma inhibitory activity" "fibroblast growth factor receptor 4" "fibrob
```

codebook

```
codebook1 <- read.csv2("Data/77_cancer_proteomes_CPTAC_codebook.txt")
codebook2 <- read.csv2("Data/clinical_data_breast_cancer_codebook.txt")
```

Type any R code in the chunk, for example: