

Introduction into machine learning and analyzes of Breast Cancer Proteomes

true

September 15th, 2022

Contents

Dataset	1
About Dataset	1
Exploratory Data Analysis	2
Data observation	9
Data cleaning and altering	11

Dataset

About Dataset

information about the data set and the three give files :

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:** clinicalatabreast_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?

```
# packages
library(pander)
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##   combine

library(stringr)
```

Exploratory Data Analysis

loading of the dataframes and showing the successful loading and its dimensions. note only the first 5 columns of “77_cancer_proteomes_CPTAC_itraq.csv” are shown since after column 4 they are the same type.

```

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

```

```

# only showing first 5 columns of proteomes
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

```

```

head(clinical_data, n=5)

```

```

##   Complete.TCGA.ID Gender Age.at.Initial.Pathologic.Diagnosis ER.Status
## 1   TCGA-A2-AOT2 FEMALE                                66 Negative
## 2   TCGA-A2-AOCM FEMALE                                40 Negative
## 3   TCGA-BH-A18V FEMALE                                48 Negative
## 4   TCGA-BH-A18Q FEMALE                                56 Negative
## 5   TCGA-BH-AOEO 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 the structure/dimensions of dataframe
```

```
cat("77_cancer_proteomes_CPTAC_itraq [ number of rows:", nrow(proteomes_data), "number of columns:", ncol(proteomes_data), "\n")
```

```
## 77_cancer_proteomes_CPTAC_itraq [ number of rows: 12553 number of columns: 86
```

```
cat("clinical_data [ number of rows:", nrow(clinical_data), "number of columns:", ncol(clinical_data), "\n")
```

```
## clinical_data [ number of rows: 105 number of columns: 30
```

```
cat("pam50_protein_data [ number of rows:", nrow(pam50_protein_data), "number of columns:", ncol(pam50_protein_data), "\n")
```

```
## pam50_protein_data [ number of rows: 100 number of columns: 4
```

Checking if the Proteomes data has been correctly read.

```
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.AOEX.04TCGA	: num	1.19 1.19 1.19 1.19 1.2 ...
## \$ A0.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.AOCM.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.AOEq.08TCGA	: num	-0.913 -0.928 -0.928 -0.932 -0.928 ...
## \$ AR.AOU4.09TCGA	: num	-0.0332 -0.0302 -0.0272 -0.0302 -0.0302 ...
## \$ A0.AOJ9.10TCGA	: 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.AOFK.11TCGA	: num	0.974 0.977 0.977 0.97 0.985 ...
## \$ A0.AOJ6.11TCGA	: 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.AOE1.12TCGA	: num	0.762 0.762 0.766 0.758 0.766 ...
## \$ A7.AOCE.13TCGA	: num	-1.12 -1.12 -1.12 -1.13 -1.13 ...
## \$ A2.AOYC.13TCGA	: num	0.819 0.815 0.815 0.799 0.819 ...
## \$ A0.AOJC.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.AOTX.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 ...
## \$ A0.A126.15TCGA	: num	0.196 0.196 0.196 0.219 0.2 ...
## \$ BH.AOC1.16TCGA	: num	-0.518 -0.51 -0.507 -0.518 -0.513 ...
## \$ A2.AOEY.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.AOEV.18TCGA	: num	0.453 0.473 0.473 0.459 0.473 ...
## \$ AN.AOAM.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.AOFL.19TCGA	: num	2.46 2.48 2.48 2.46 2.48 ...
## \$ BH.AODG.19TCGA	: num	-0.206 -0.206 -0.206 -0.215 -0.206 ...
## \$ AR.AOTV.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 ...
## \$ A0.AOJJ.20TCGA	: num	0.757 0.781 0.774 0.764 0.771 ...
## \$ A0.AOJE.21TCGA	: num	0.56 0.563 0.56 0.542 0.56 ...
## \$ AN.AOAJ.21TCGA	: num	-0.428 -0.406 -0.406 -0.406 -0.406 ...
## \$ A7.AOCJ.22TCGA	: num	-1.001 -1.005 -1.005 -0.998 -1.001 ...
## \$ A0.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.AOT3.24TCGA	: num	0.584 0.581 0.581 0.587 0.587 ...
## \$ A2.AOYD.24TCGA	: num	0.0638 0.0933 0.0845 0.0667 0.0845 ...
## \$ AR.AOTR.25TCGA	: num	-1.1 -1.11 -1.11 -1.1 -1.11 ...
## \$ A0.A030.25TCGA	: num	1.05 1.06 1.06 1.06 1.06 ...

```
## $ A0.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 ...
## $ A0.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 ...
## $ A0.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 ...
## $ A0.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 ...
```

Nothing strange about the Proteomes dat everything seems to be read correct.

Checking if the clinical data has been correctly read.

```
str(clinical_data)
```

```
## 'data.frame':   105 obs. of  30 variables:
## $ Complete.TCGA.ID      : chr  "TCGA-A2-AOT2" "TCGA-A2-AOCM" "TCGA-BH-A18V" "TCGA-BH-A18V" ...
## $ 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_Conversion" ...
## $ 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 ...
```

Nothing strange about the clinical data everything seems to be read correct.

Checking if the pam50 protein data has been correctly read.

```
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" "fibroblast growth factor receptor 4" ...
```

Nothing strange about the pam50 protein data everything seems to be read correct.

codebook

loading of the created codebooks for the three dataframes. showing also its contents and successful loading

```
cancer_proteomes_CPTAC_codebook <- read.csv2("Data/77_cancer_proteomes_CPTAC_codebook.txt")
clinical_data_codebook <- read.csv2("Data/clinical_data_breast_cancer_codebook.txt")
PAM50_protein_codebook <- read.csv2("Data/PAM50_protein_codebook.txt", sep = ";")
```

```
cancer_proteomes_CPTAC_codebook
```

```
##           Column           Description data.type unit
## 1 RefSeq_accession_number RefSeq protein ID   string  NA
## 2           gene_symbol Gene abbreviation code string  NA
## 3           gene_name     Name of the gene    string  NA
## 4      Remaining columns      log2 iTRAQ ratios float  NA
```

```
clinical_data_codebook
```

```
##           Column           Description
## 1 Complete_TCGA_ID         TCGA ID
## 2           Gender         Gender
```

## 3	Age_at_Initial_Pathologic_Diagnosis	Age at Initial Pathologic Diagnosis
## 4	ER Status	Estrogen receptor Status
## 5	PR Status	Progesterone receptor Status
## 6	HER2 Final Status	Human Epidermal growth factor Receptor 2
## 7	Tumor	Tumor
## 8	Tumor--T1 Coded	Tumor--T1 Coded
## 9	Node	Node
## 10	Node-Coded	Node-Coded
## 11	Metastasis	Metastasis
## 12	Metastasis-Coded	Metastasis-Coded
## 13	AJCC Stage	American Joint Committee on Cancer Stage
## 14	Converted Stage	Converted Stage
## 15	Survival Data Form	Survival Data Form
## 16	Vital Status	Vital Status
## 17	Days to Date of Last Contact	Days to Date of Last Contact
## 18	Days to date of Death	Days to date of Death
## 19	OS event	OS event 0= NO, 1= YES
## 20	OS Time	OS Time
## 21	PAM50 mRNA	PAM50 mRNA
## 22	SigClust Unsupervised mRNA	SigClust Unsupervised mRNA
## 23	SigClust Intrinsic mRNA	SigClust Intrinsic mRNA
## 24	miRNA Clusters	miRNA Clusters
## 25	methylation Clusters	methylation Clusters
## 26	RPPA Clusters	RPPA Clusters
## 27	CN Clusters	CN Clusters
## 28	Integrated Clusters (with PAM50)	Integrated Clusters (with PAM50)
## 29	Integrated Clusters (no exp)	Integrated Clusters (no exp)
## 30	Integrated Clusters (unsup exp)	Integrated Clusters (unsup exp)
##	type data.type unit	
## 1	name chr <NA>	
## 2	name chr <NA>	
## 3	Descriptive chr <NA>	
## 4	Descriptive chr <NA>	
## 5	Descriptive chr <NA>	
## 6	Descriptive chr <NA>	
## 7	Descriptive chr <NA>	
## 8	Descriptive chr <NA>	
## 9	Descriptive chr <NA>	
## 10	Descriptive chr <NA>	
## 11	Descriptive chr <NA>	
## 12	Descriptive chr <NA>	
## 13	Descriptive chr <NA>	
## 14	Descriptive chr <NA>	
## 15	Descriptive chr <NA>	
## 16	Descriptive chr <NA>	
## 17	Time int Days	
## 18	Time int Days	
## 19	Descriptive int <NA>	
## 20	Time int Hours	
## 21	Descriptive chr <NA>	
## 22	Count int <NA>	
## 23	Count int <NA>	
## 24	Count int <NA>	
## 25	Count int <NA>	


```
## 26 Descriptive      chr <NA>
## 27      Count      int <NA>
## 28      count      int <NA>
## 29      count      int <NA>
## 30      count      int <NA>
```

```
PAM50_protein_codebook
```

```
##      Column      Description type      unit
## 1      GeneSymbol      Gene abbreviation chr      <NA>
## 2 RefSeqProteinID Unique reference identifier chr      <NA>
## 3      Species      Species chr latin name
## 4      Gene.Name      Name of the gene chr      <NA>
```

Data observation

there are 12553 rows in the data, these are proteins identifiable with a RefSeq ID number and have 86 columns of which the last 83 are samples (with named with their identifiers and the last three from healthy individuals). to further use the data i shall reshape it to make the rows samples and each column a protein

```
# first making a data frame with only the numerical data, samples start at column number 4 til the end
proteomes_data_numerical <- proteomes_data[4:86]
```

```
# checking data
head(proteomes_data_numerical,n=5)
```

```
##      AO.A12D.01TCGA C8.A131.01TCGA AO.A12B.01TCGA BH.A18Q.02TCGA C8.A130.02TCGA
## 1      1.096131      2.609943      -0.6598280      0.1953407      -0.4940596
## 2      1.111370      2.650422      -0.6487422      0.2154129      -0.5038992
## 3      1.111370      2.650422      -0.6542851      0.2154129      -0.5006193
## 4      1.107561      2.646374      -0.6321133      0.2053768      -0.5104589
## 5      1.115180      2.646374      -0.6404277      0.2154129      -0.5038992
##      C8.A138.03TCGA E2.A154.03TCGA C8.A12L.04TCGA A2.A0EX.04TCGA AO.A12D.05TCGA
## 1      2.765081      0.8626593      1.407570      1.185108      1.100688
## 2      2.779709      0.8701860      1.407570      1.192612      1.100688
## 3      2.779709      0.8701860      1.410312      1.188860      1.100688
## 4      2.797995      0.8664226      1.407570      1.185108      1.100688
## 5      2.787023      0.8701860      1.413053      1.200116      1.093358
##      AN.A04A.05TCGA BH.A0AV.05TCGA C8.A12T.06TCGA A8.A06Z.07TCGA A2.A0CM.07TCGA
## 1      0.3845877      0.3505357      -0.2049179      -0.4964091      0.6834035
## 2      0.3713928      0.3674053      -0.1624185      -0.4985089      0.6944241
## 3      0.3713928      0.3674053      -0.1666684      -0.4964091      0.6980976
## 4      0.3779903      0.3606575      -0.1836682      -0.4922095      0.6870771
## 5      0.3746916      0.3707793      -0.1666684      -0.4880099      0.6870771
##      BH.A18U.08TCGA A2.A0EQ.08TCGA AR.A0U4.09TCGA AO.A0J9.10TCGA AR.A1AP.11TCGA
## 1      -0.2650304      -0.9126703      -0.03322133      0.020007050      0.4610875
## 2      -0.2516423      -0.9279787      -0.03021642      0.011955318      0.4610875
## 3      -0.2516423      -0.9279787      -0.02721152      0.011955318      0.4610875
## 4      -0.2516423      -0.9318057      -0.03021642      0.003903587      0.4610875
## 5      -0.2516423      -0.9279787      -0.03021642      0.011955318      0.4610875
##      AN.A0FK.11TCGA AO.A0J6.11TCGA A7.A13F.12TCGA BH.A0E1.12TCGA A7.A0CE.13TCGA
## 1      0.9735642      0.8311317      1.279185      0.7620444      -1.123173
```

## 2	0.9774761	0.8565398	1.275167	0.7620444	-1.123173
## 3	0.9774761	0.8565398	1.275167	0.7663844	-1.116861
## 4	0.9696523	0.8367780	1.279185	0.7577045	-1.129486
## 5	0.9852998	0.8650092	1.279185	0.7663844	-1.129486
##	A2.A0YC.13TCGA	A0.A0JC.14TCGA	A8.A08Z.14TCGA	AR.A0TX.14TCGA	A8.A076.15TCGA
## 1	0.8188241	-0.3072668	0.5688946	-0.5834286	1.873982
## 2	0.8148772	-0.3072668	0.5688946	-0.5725489	1.870383
## 3	0.8148772	-0.3072668	0.5688946	-0.5671090	1.870383
## 4	0.7990900	-0.3072668	0.5688946	-0.5834286	1.859587
## 5	0.8188241	-0.3010327	0.5688946	-0.5725489	1.870383
##	A0.A126.15TCGA	BH.A0C1.16TCGA	A2.A0EY.16TCGA	AR.A1AW.17TCGA	AR.A1AV.17TCGA
## 1	0.1958767	-0.5183665	1.174881	0.5783087	-0.7598231
## 2	0.1958767	-0.5100020	1.183209	0.5822129	-0.7598231
## 3	0.1958767	-0.5072138	1.183209	0.5783087	-0.7491137
## 4	0.2189346	-0.5183665	1.174881	0.5900212	-0.7357270
## 5	0.1997197	-0.5127902	1.179045	0.5861170	-0.7491137
##	C8.A135.17TCGA	A2.A0EV.18TCGA	AN.A0AM.18TCGA	D8.A142.18TCGA	AN.A0FL.19TCGA
## 1	1.120502	0.4529859	1.501967	0.5385958	2.455138
## 2	1.137618	0.4725901	1.510348	0.5422105	2.480137
## 3	1.137618	0.4725901	1.501967	0.5422105	2.480137
## 4	1.137618	0.4585871	1.501967	0.5349810	2.461956
## 5	1.120502	0.4725901	1.501967	0.5422105	2.477864
##	BH.A0DG.19TCGA	AR.A0TV.20TCGA	C8.A12Z.20TCGA	A0.A0JJ.20TCGA	A0.A0JE.21TCGA
## 1	-0.2056375	-1.514278	-0.7871950	0.7571881	0.5597770
## 2	-0.2056375	-1.528285	-0.7559406	0.7808707	0.5634069
## 3	-0.2056375	-1.528285	-0.7559406	0.7741042	0.5597770
## 4	-0.2150062	-1.531087	-0.7746932	0.7639546	0.5416274
## 5	-0.2056375	-1.514278	-0.7715678	0.7707210	0.5597770
##	AN.A0AJ.21TCGA	A7.A0CJ.22TCGA	A0.A12F.22TCGA	A8.A079.23TCGA	A2.A0T3.24TCGA
## 1	-0.4281815	-1.0012398	-1.947792	1.048959	0.5837133
## 2	-0.4063780	-1.0046198	-1.952718	1.052257	0.5806231
## 3	-0.4063780	-1.0046198	-1.955180	1.052257	0.5806231
## 4	-0.4063780	-0.9978599	-1.947792	1.058852	0.5868034
## 5	-0.4063780	-1.0012398	-1.957643	1.052257	0.5868034
##	A2.A0YD.24TCGA	AR.A0TR.25TCGA	A0.A030.25TCGA	A0.A12E.26TCGA	A8.A06N.26TCGA
## 1	0.06377853	-1.101675	1.053225	0.2648591	0.2385471
## 2	0.09333637	-1.108783	1.055948	0.2757113	0.2498182
## 3	0.08446902	-1.108783	1.055948	0.2757113	0.2441826
## 4	0.06673431	-1.096937	1.058671	0.2784244	0.2498182
## 5	0.08446902	-1.111152	1.058671	0.2784244	0.2498182
##	A2.A0YG.27TCGA	BH.A18N.27TCGA	AN.A0AL.28TCGA	A2.A0T6.29TCGA	E2.A158.29TCGA
## 1	-0.07820182	1.101261	0.3236627	0.7939756	-1.086529
## 2	-0.06805814	1.101261	0.3269726	0.8181815	-1.095492
## 3	-0.07143937	1.097767	0.3269726	0.8147235	-1.095492
## 4	-0.05791445	1.090779	0.3302826	0.8008915	-1.095492
## 5	-0.06467691	1.108248	0.3269726	0.8181815	-1.095492
##	E2.A15A.29TCGA	A0.A0JM.30TCGA	C8.A12V.30TCGA	A2.A0D2.31TCGA	C8.A12U.31TCGA
## 1	2.180123	1.395247	0.6739047	0.10749090	-0.4815502
## 2	2.180123	1.408922	0.6887176	0.10416449	-0.4778898
## 3	2.180123	1.412341	0.6887176	0.10749090	-0.4815502
## 4	2.180123	1.408922	0.6776079	0.09751166	-0.4705692
## 5	2.180123	1.408922	0.6887176	0.10416449	-0.4815502
##	AR.A1AS.31TCGA	A8.A09G.32TCGA	C8.A131.32TCGA	C8.A134.32TCGA	A2.A0YF.33TCGA
## 1	1.222507	-1.523343	2.707250	0.1401818	0.3113192

```
## 2      1.218974      -1.512646      2.733832      0.1260538      0.2961771
## 3      1.222507      -1.509972      2.737629      0.1331178      0.2961771
## 4      1.204839      -1.517995      2.733832      0.1119257      0.2961771
## 5      1.222507      -1.509972      2.752819      0.1260538      0.2961771
## BH.A0DD.33TCGA BH.A0E9.33TCGA AR.A0TT.34TCGA AO.A12B.34TCGA A2.A0SW.35TCGA
## 1      -0.6923158      1.466665      -0.5114212      -0.9639039      -0.4877725
## 2      -0.6594687      1.482283      -0.5260667      -0.9382095      -0.4877725
## 3      -0.6641611      1.474474      -0.5260667      -0.9439194      -0.4877725
## 4      -0.6571224      1.458856      -0.5333894      -0.9353546      -0.4877725
## 5      -0.6618149      1.474474      -0.5297281      -0.9353546      -0.5038532
## AO.A0JL.35TCGA BH.A0BV.35TCGA A2.A0YM.36TCGA BH.A0C7.36TCGA A2.A0SX.36TCGA
## 1      -0.10668      -0.06583842      0.6558497      -0.5522120      -0.3985598
## 2      -0.10668      -0.05589267      0.6581426      -0.5477494      -0.3926014
## 3      -0.10668      -0.06583842      0.6558497      -0.5522120      -0.3926014
## 4      -0.10668      -0.05589267      0.6558497      -0.5522120      -0.3926014
## 5      -0.10668      -0.06252317      0.6512639      -0.5566746      -0.3955806
## X263d3f.I.CPTAC blcdb9.I.CPTAC c4155b.C.CPTAC
## 1      0.5985845      -0.1912845      0.5669753
## 2      0.6066975      -0.1839177      0.5787017
## 3      0.6039931      -0.1860225      0.5767473
## 4      0.6039931      -0.1860225      0.5767473
## 5      0.6039931      -0.1670792      0.5767473
```

Data cleaning and altering

Transposing

transposing the created data frame “proteomes_data_numerical”, and adding the refseq ID as column name

```
proteomes_data_numerical_transposed <- as.data.frame(t(proteomes_data_numerical))
colnames(proteomes_data_numerical_transposed) <- proteomes_data$RefSeq_accession_number

# checking is succesfull
cat("proteomes_data_numerical_transposed[ number of rows:", nrow(proteomes_data_numerical_transposed),
    "number of columns:", ncol(proteomes_data_numerical_transposed), '\n')
```

```
## proteomes_data_numerical_transposed[ number of rows: 83 number of columns: 12553
```

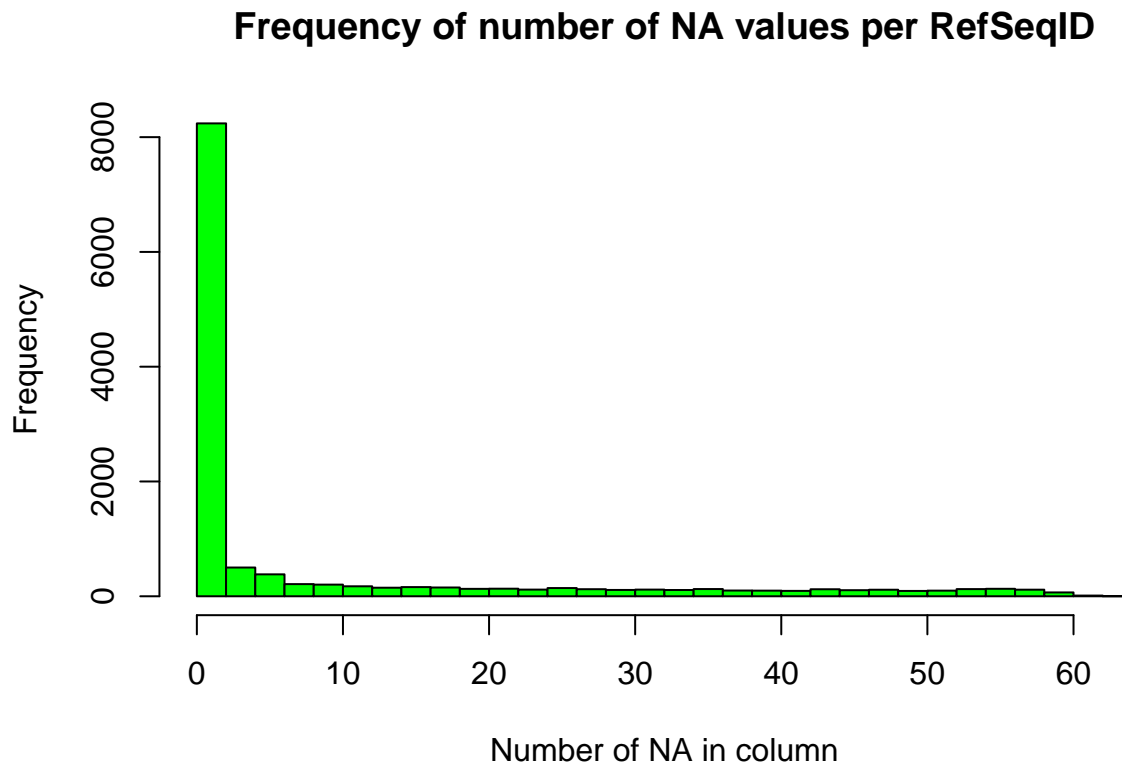
```
# lets add a healthy group and cancer, the last 3 columns where from healthy poeple
# *see content section for the reference
proteomes_data_numerical_transposed$sample_type <- "Cancer"
proteomes_data_numerical_transposed$sample_type[81:83] <- "Healthy"
```

cleaning

since there are NA values in the data lets see how much

```
count_na_func <- function(x) sum(is.na(x))
# getting NA values per RefSeqID(column)
Na_per_col <- sapply(proteomes_data_numerical_transposed, count_na_func)
```

```
hist(Na_per_col, breaks = 40,
     xlab = "Number of NA in column",
     ylab = "Frequency",
     main = "Frequency of number of NA values per RefSeqID",
     col = "green")
```



```
cat("number of proteins with NA values in them:", sum(Na_per_col > 0), '\n' )
```

```
## number of proteins with NA values in them: 4559
```

```
# TODO decide how many NA values are permitted per protein
proteomes_filtered_data <- proteomes_data_numerical_transposed[Na_per_col < 8]
cat("number of proteins with 8 or more NA values in them and deleted from data:",
    sum(Na_per_col > 8), '\n')
```

```
## number of proteins with 8 or more NA values in them and deleted from data: 3219
```

```
cat("proteomes_filtered_data[number of rows:",
    nrow(proteomes_filtered_data),
    "number of columns:",
    ncol(proteomes_filtered_data), '\n')
```

```
## proteomes_filtered_data[number of rows: 83 number of columns: 9200
```