Introduction into machine learning and analysis of Breast Cancer Proteomes

Theme09 - Introduction to Machine Learning

Mats Slik
344216
BFV3
November 12, 2022
Dave Langers (LADR)
Bart Barnard (BABA)

Introduction into machine learning and analysis of Breast Cancer Proteomes

Theme09 - Introduction to Machine Learning

Mats Slik
344216
Bioinformatics
Institute for Life Science & Technology
Hanze University of Applied Sciences
Dave Langers (LADR)
Bart Barnard (BABA)
November 12, 2022

Abstract

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. my question is: Are there different ways to categorize breast cancer based on protein expression data, with machine learning being used to classify them without using the pam50 proteins?

Table of Contents

List of Abbreviations

EDA	Exploratory Data Analysis
TCGA	The cancer Genome Atlas Program
CPTAC	Clinical Proteomic Tumor Analysis Consortium
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid

List of Figures

List of Tables

1 Introduction

??

2 Methods

3 Results

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.

my question is: Are there different ways to categorize breast cancer based on protein expression data, with machine learning being used to classify them without using the pam50 proteins?

to answer that question we must first look at the data.

The 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. this data was sampled from 105 originally from the TCGA (The Cancer Genome Atlas Program - NCI), which was further filterd to 77 samples containing high quality protein expression data.

When looking at the dimensions of the data set we can see there are a lot of proteins see table 1

Row.names	Tumor	NP_958782	NP_958785	NP_958786	NP_000436
blcdb9.I.CPTAC	Healthy	-0.1913	-0.1839	-0.186	-0.186
c4155b.C.CPTAC	Healthy	0.567	0.5787	0.5767	0.5767
TCGA-A2-A0CM	T2	0.6834	0.6944	0.6981	0.6871
TCGA-A2-A0D2	T2	0.1075	0.1042	0.1075	0.09751
TCGA-A2-A0EQ	T2	-0.9127	-0.928	-0.928	-0.9318
TCGA-A2-A0EV	T1	0.453	0.4726	0.4726	0.4586

```
## number of rows: 80 number of columns: 9201
```

After this first assessment of the data we started looking at the number of missing values as seen in the figures' fig 1 and 2 below

As we can see in these figures 1 and 2 the distribution is very much to the left where a lot of proteins have one or only two missing values, further more there are still a couple of proteins that have a high number of missing values these are to be filtered out because this can create a false set of results when we are using them in our machine learning algoritme for clustering them o there cancer stage. so to further see how the data is we took the distribution of a couple of proteins in a multi boxplot as seen in figure 3

In this figure 3 we can clearly see that for the first 70 protein that most have a distribution of their log2 itraq expression between 5 and -5 but there are some that have higher numbers. To further make sense of all the 12 to 9 thousand proteins in the data we calculated the standard deviation of them see figure 4

In this figure 4 we compared the normal data set and the one filtered that has had protein with more tha 10% of their values missing removed. In it, we can clearly see that a lot of proteins with high deviation are removed from the data. To make a further analyse of these samples

In this figure 5 we can see how the different samples are spread according to there cancer stages.

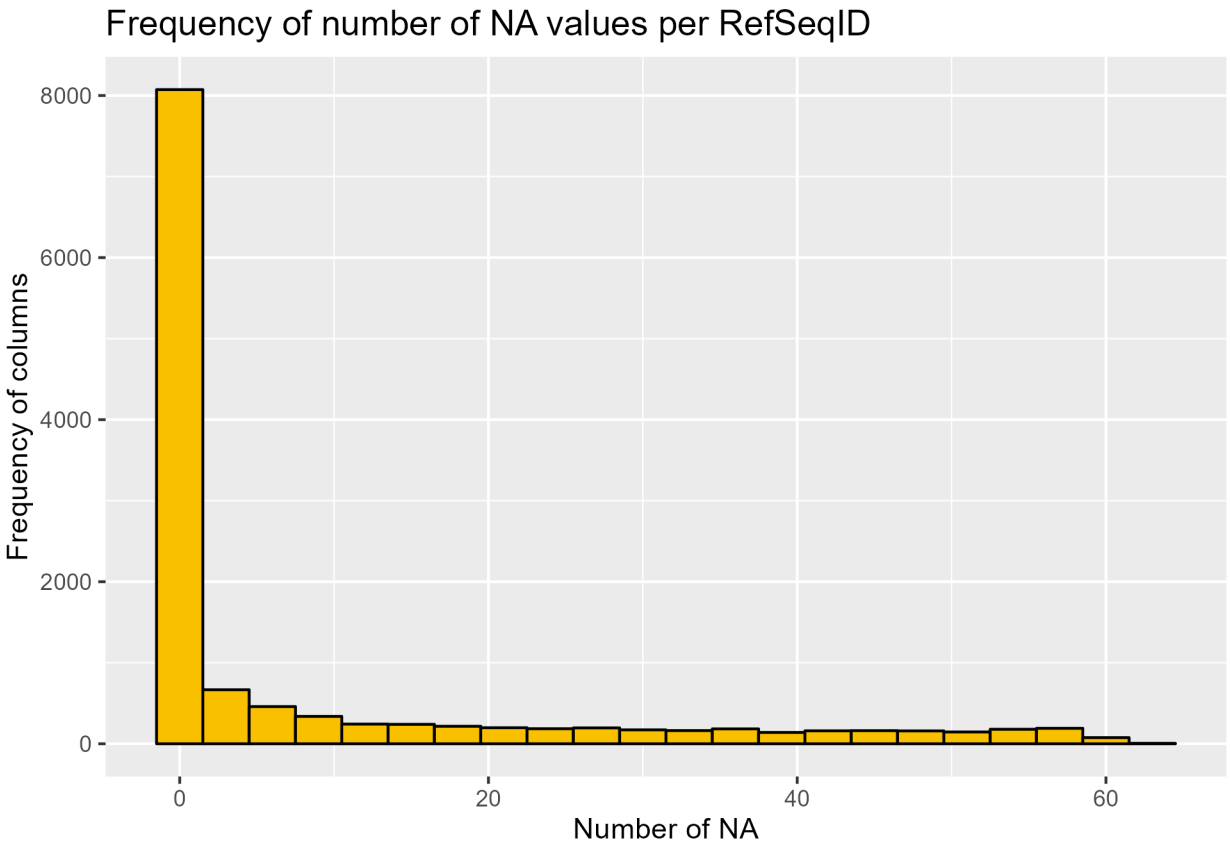

Figure 1: figure 1

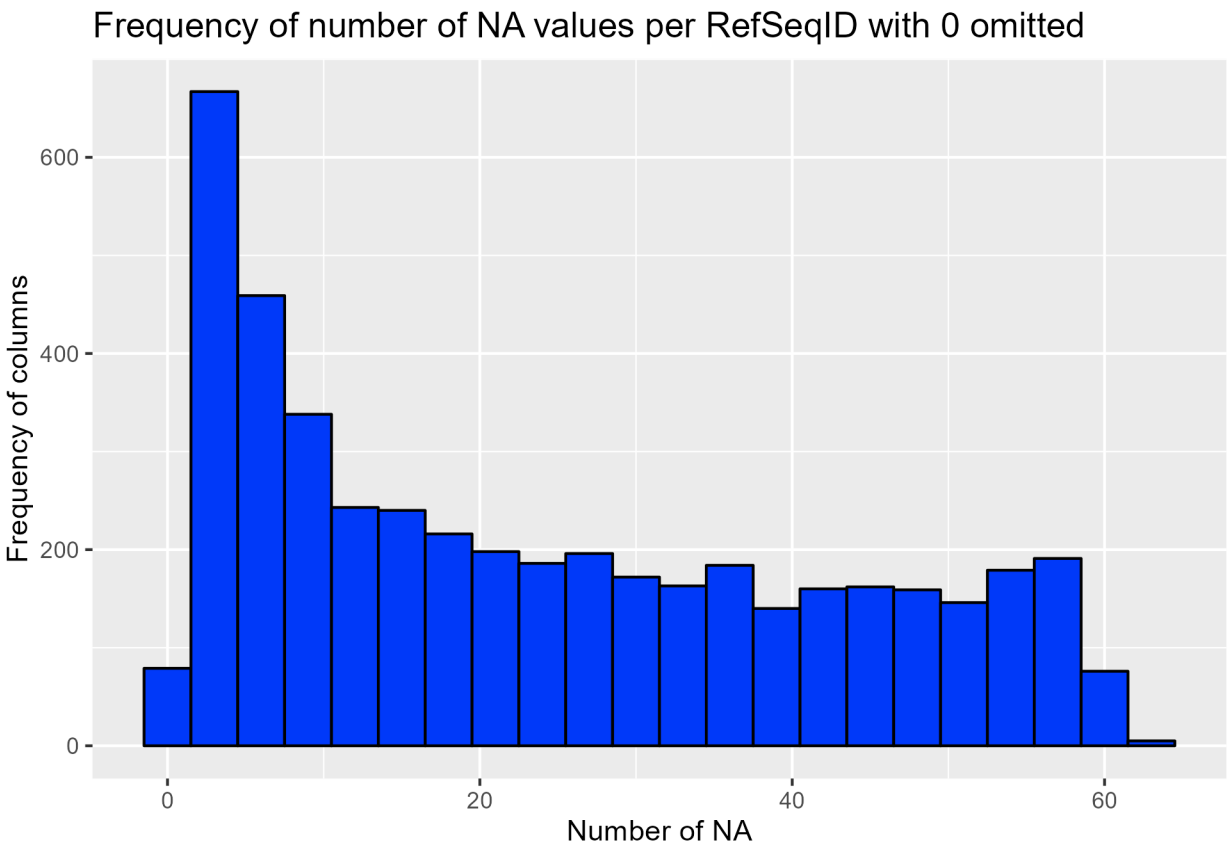

Figure 2: figure 2

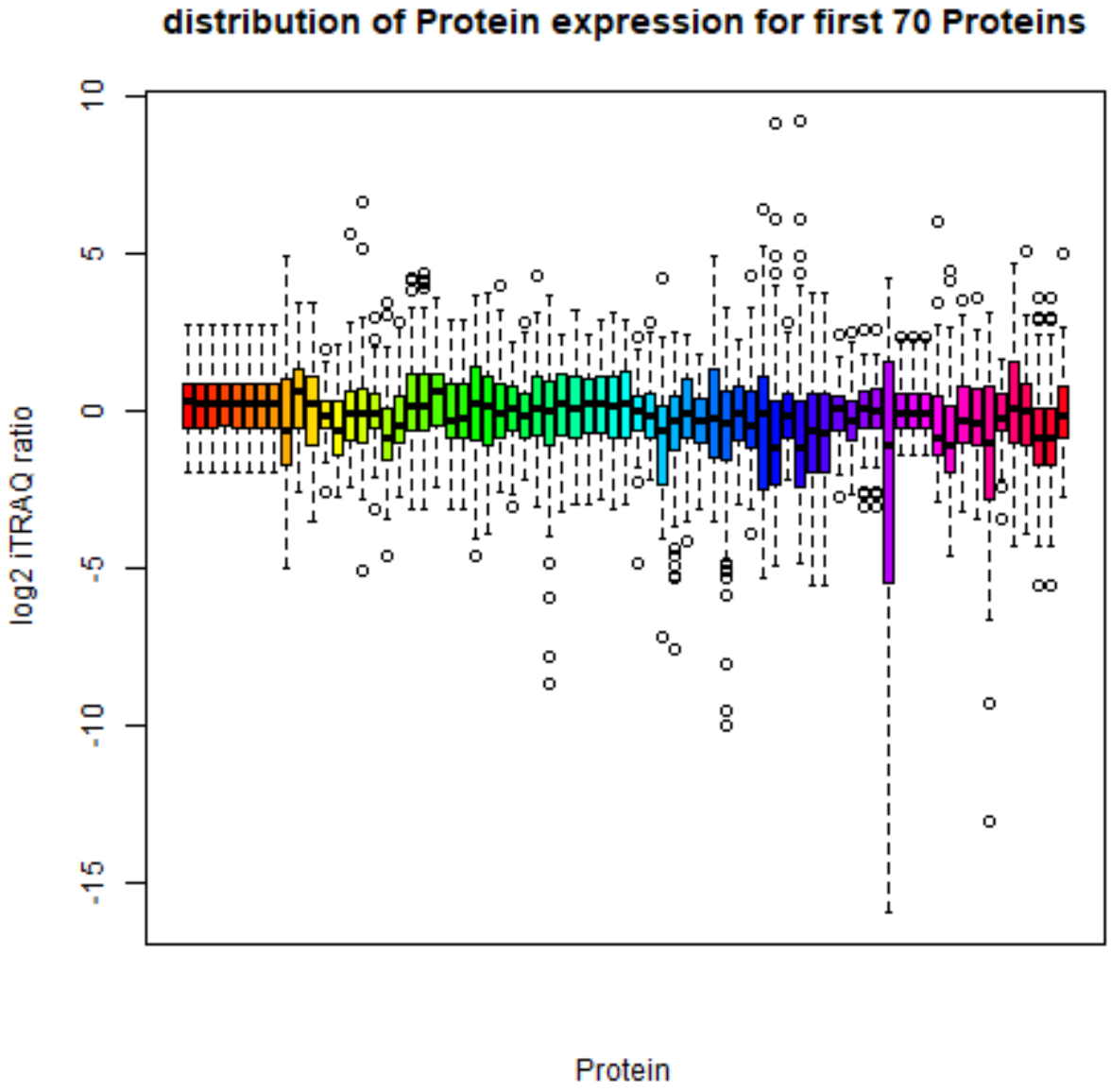

Figure 3: figure 3

Figure 4: figure 4

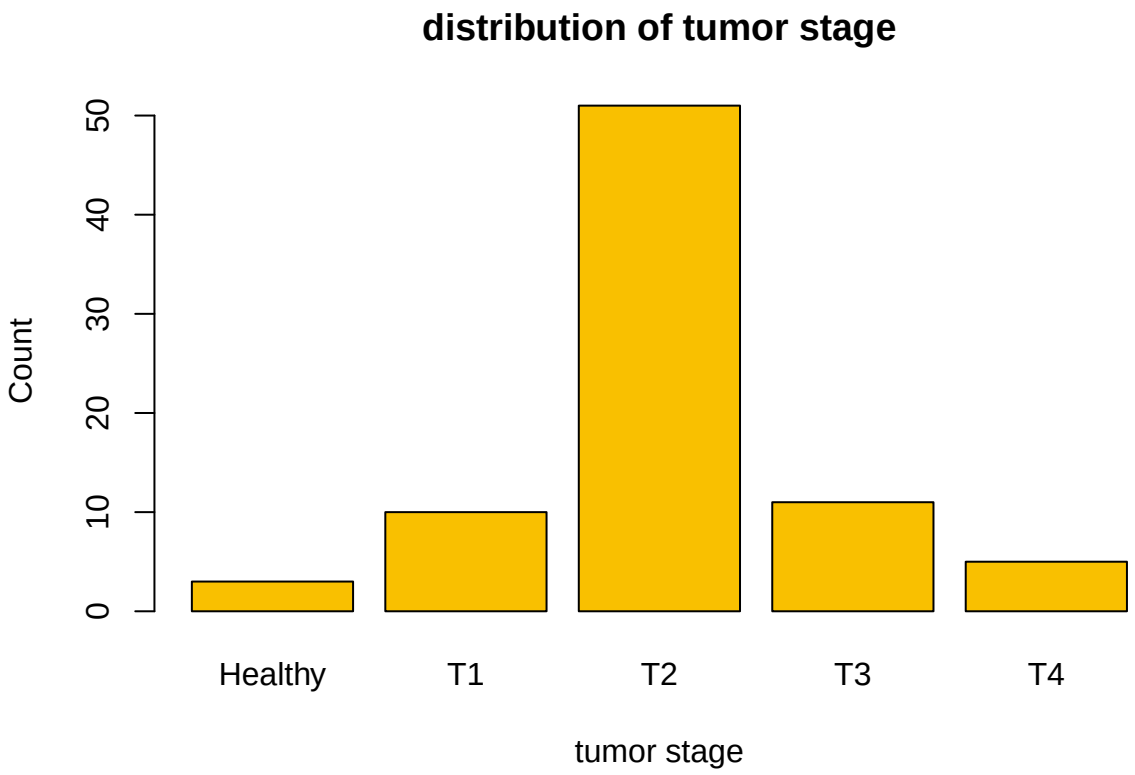

Figure 5: distribution of amount of samples per tumor stage

4 Discussion and Conclusion

In the results' section we can see from the figures 1 and two that although the data set was supplied with the label as high quality there are still proteins in the data with more tha 10% of there expression values missing, this combined with the need for using the expressing data with the clinical categorical data, the sample names needed to be changed to be compared. all this wasn't something to be expecting of high quality data. also in figure 5 it is clearly visible that the categorisation of the tumor stage there are a lot of T2 stages in the samples than any other. furthermore the sheer amount of proteins recorded in this data is very useful for my purpose of trying to use another classification as the PAM50 protein list

5 References

Mertins, Philipp, D R Mani, Kelly Ruggles, Michael Gillette, Karl Clauser, Pei Wang, Xianlong Wang, et al. 2016. "Proteogenomics Connects Somatic Mutations to Signaling in Breast Cancer." *Nature* 534 (May). https://doi.org/10.1038/nature18003.

6 Appendices