Introduction into machine learning and analysis of Breast Cancer Proteomes

Theme09 - Introduction to Machine Learning

Mats Slik
344216
BFV3
October 4, 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)
October 4, 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

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

Row.names	Tumor	Tumor..T1.Coded	AJCC.Stage	Vital.Status	NP_958782
TCGA-A2-A0CM	T2	T_Other	Stage IIA	DECEASED	0.6834
TCGA-A2-A0D2	T2	T_Other	Stage IIB	LIVING	0.1075
TCGA-A2-A0EQ	T2	T_Other	Stage IIA	LIVING	-0.9127
TCGA-A2-A0EV	T1	T1	Stage IA	LIVING	0.453
TCGA-A2-A0EX	T3	T_Other	Stage IIB	LIVING	1.185
TCGA-A2-A0EY	T2	T_Other	Stage IIB	LIVING	1.175

```
## number of rows: 77 number of columns: 9204
```

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

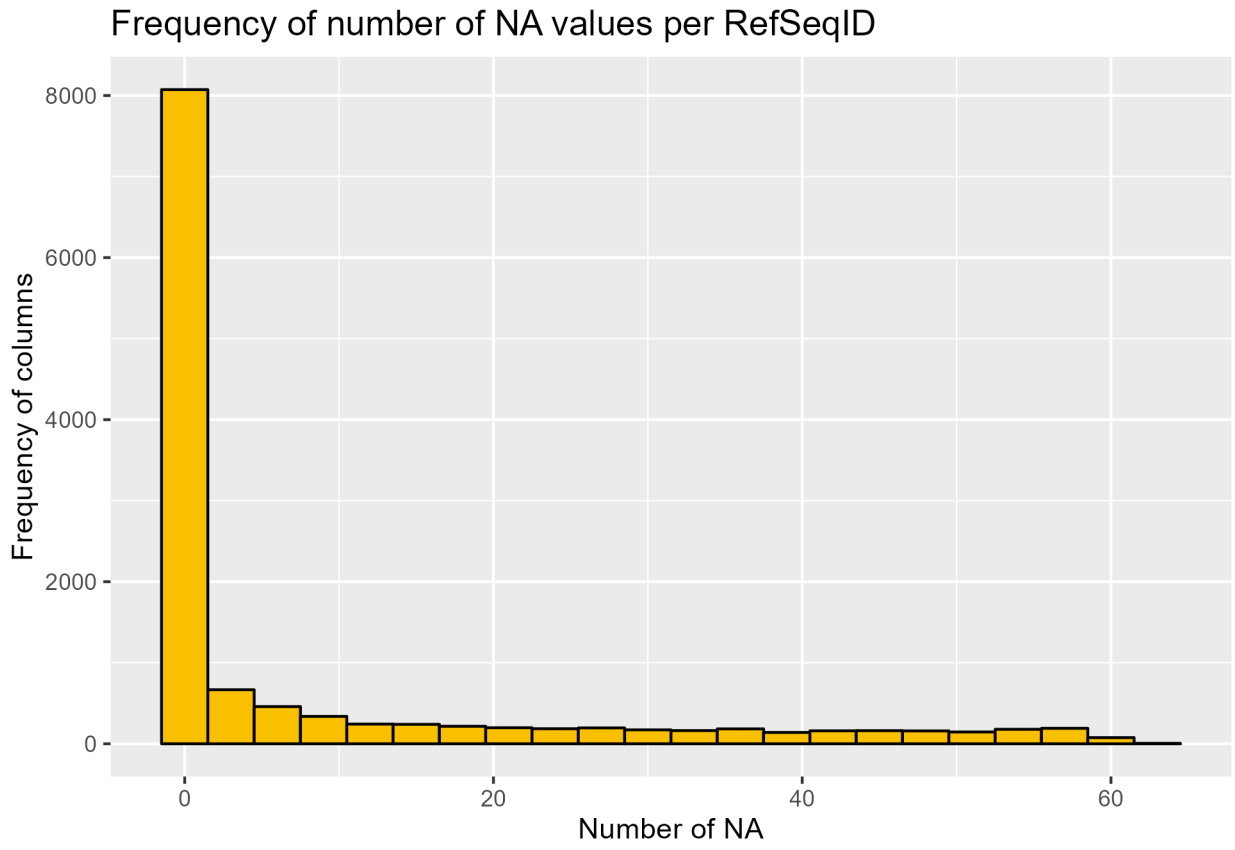

Figure 1: figure 1

as we can see in thes figures 1 and 2 the distribution is very much to the left where allot of proetis 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 becaus this can create a false set of results when we are using them in oure machine learning algoritme for clustering them o there cancer stage.

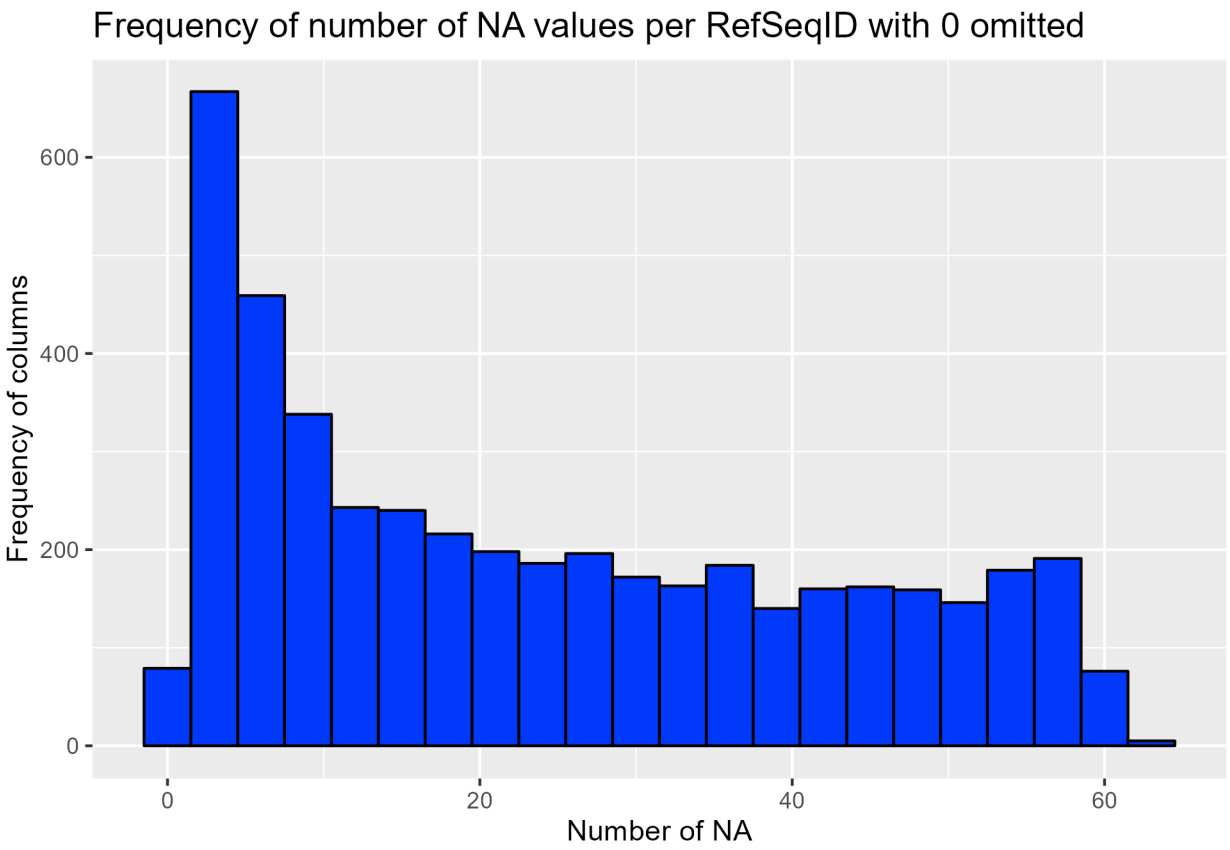

Figure 2: figure 2

so to furthere see how the data is we took the distribution of a couple of proteins in a mullti boxplot as seen in figure 3

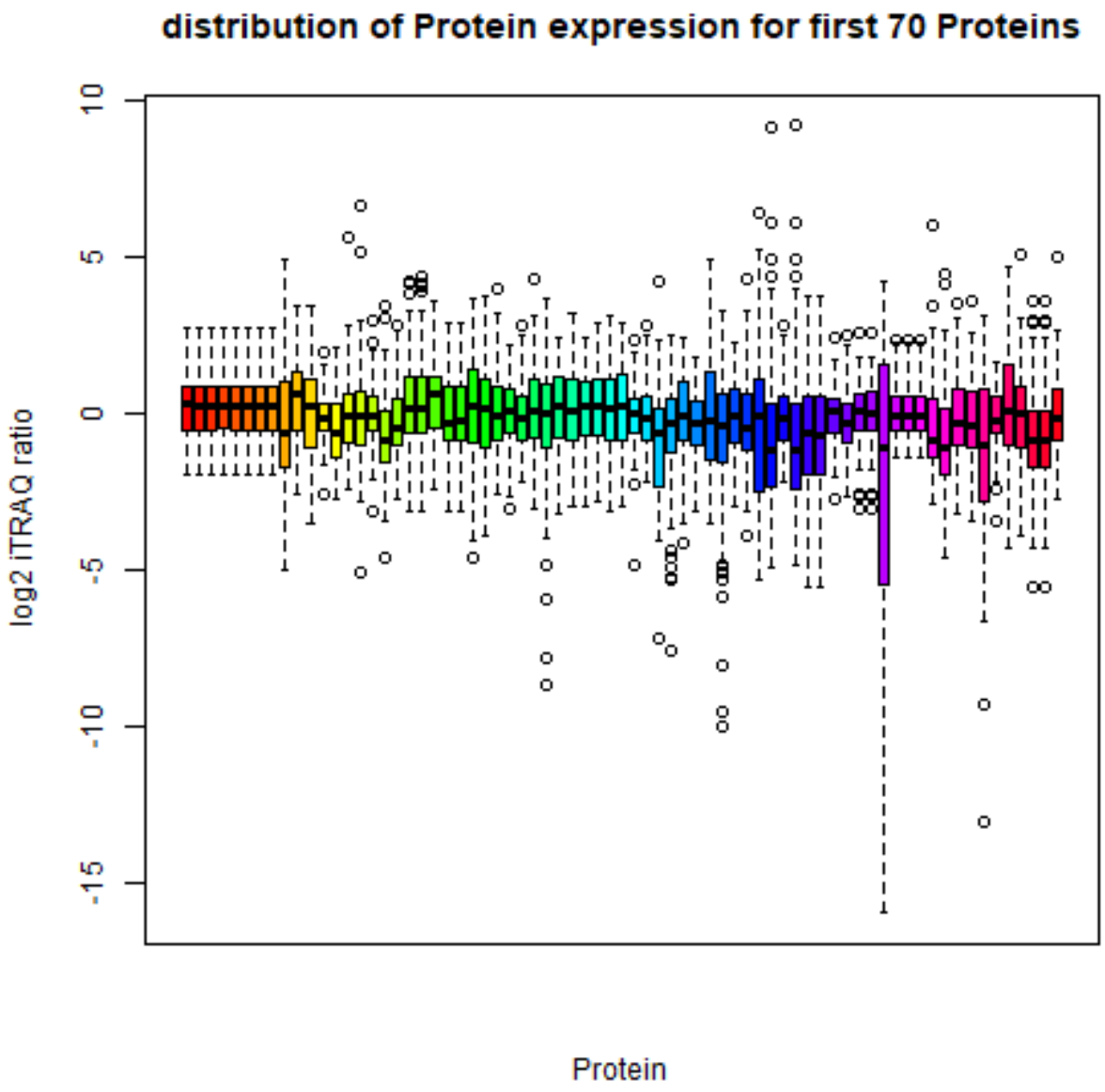

Figure 3: figure 3

in this figure 3 we can clearly see that for the first 70 protein thaty most have a distribution of ther log2 itraq expression between 5 and -5 but there are some that have higher numbers. to further make sense of all of 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 filterd that has had protein with more tha 10% of there values missing removed. in it we can clearly see that allot of proteins with high deviation are removed from the data. to make a further analyse of thease samples

Figure 4: figure 4

Discussion and Conclusion

4 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.

5 Appendices