res_conc

Rob Meulenkamp

10/5/2021

Introduction

For this research the breast cancer remains the primary focus. Somatic mutations play an import role in the development of breast cancer. The effect of the mutations remains poorly understood. The data provide 105 annotated breast cancer, which 77 provides high-quality data. Machine learning provides calculated patterns, which could help to interpret the effects of somatic mutations with breast cancer.

Exploratory Data Analysis

Loading the data in R.

```
p_data <- read.csv("data/77_cancer_proteomes_CPTAC_itraq.csv")

clinic <- read.csv("data/clinical_data_breast_cancer.csv")
```

```
n <- p_data$RefSeq_accession_number

#Get all but first 3 columns
proteomes <- as.data.frame(t(p_data[,4:83]))
colnames(proteomes) <- n

proteomes <- cbind(rownames(proteomes), data.frame(proteomes, row.names=NULL))
colnames(proteomes)[1] <- "Complete.TCGA.ID"

#Function string manipulation
get.clinical.id <- function(proteome.id) {
  x = substr(proteome.id, 4, 7)
  y = substr(proteome.id, 0, 2)
  paste("TCGA",y,x,sep="-")
}

proteomes$Complete.TCGA.ID <- sapply(proteomes$Complete.TCGA.ID, get.clinical.id)
proteomes_all <- proteomes
```

Cleaning data

```
#Remove the duplicates
proteomes_all <- proteomes_all[!duplicated(proteomes_all$Complete.TCGA.ID),]


#Merge the Tumor column from clinic data
```

```r
merged_file <- merge(clinic, proteomes_all, BY = "Complete.TCGA.ID")
merged_file <- merged_file[-c(2:6, 8:30)]


# remove variable with >20% missing data
proteomes_all <- merged_file[ , colSums(is.na(merged_file))  / nrow(proteomes_all) < 0.20]


# replace remaining NA-values with the mean
for (i in which(sapply(proteomes_all, is.numeric))) {
    proteomes_all[is.na(proteomes_all[, i]), i] <- mean(proteomes_all[, i], na.rm = TRUE)
}

# change the format from wide to long
long <- melt(proteomes_all)
```

```
## Using Complete.TCGA.ID, Tumor as id variables
```

Results

The results contain graphs and tables which help to visualize the data. The aim is to explore the data and to make the data ready for machine learning experiments.

The histogram gives a clear view about the data distribution. This could help investigate the difference in expression values between the groups. But this doesn't imply when the variation is high within the groups.

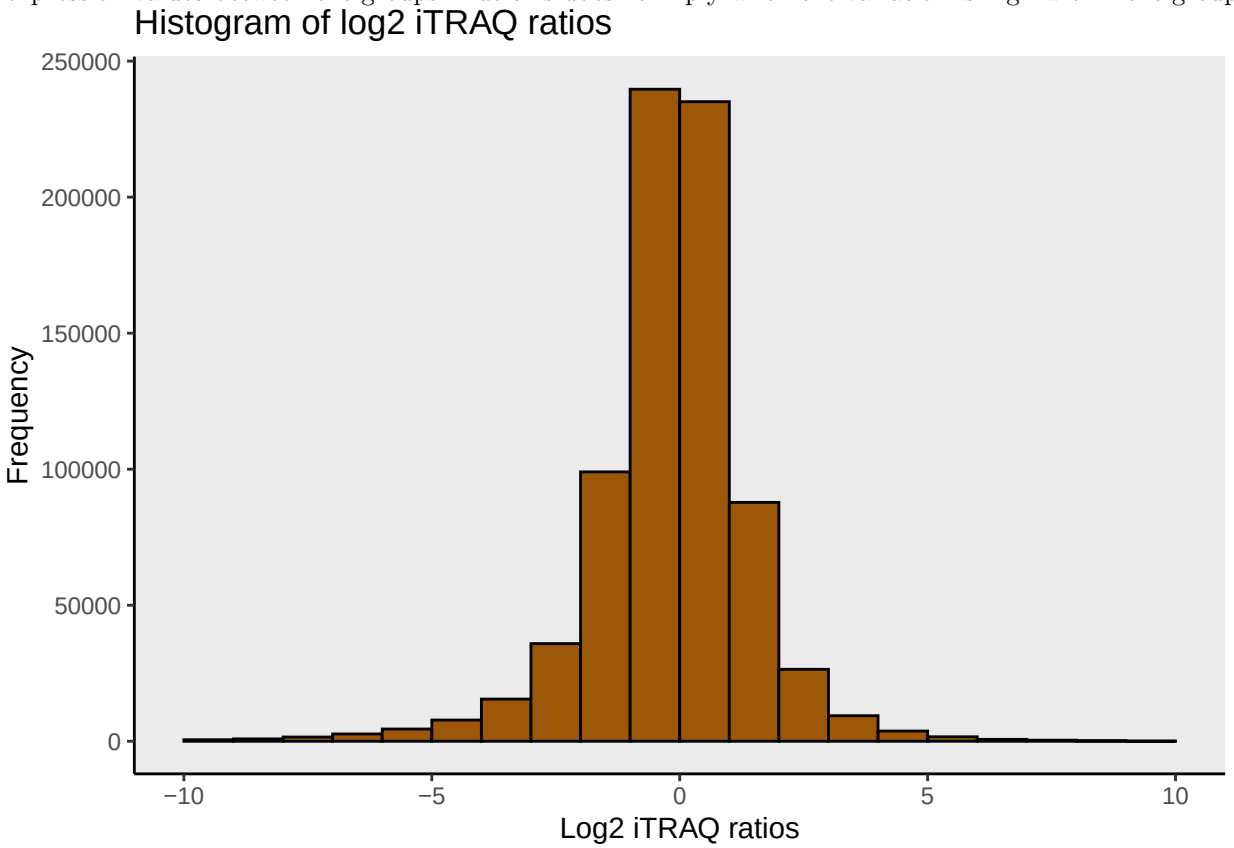

Figure 1. *Normalized distribution log2 iTRAQ ratios for every protein.* The bins between 0 and -10 are higher in comparison with the bins between 0 and 10. So there are probably more downregulated proteins

instead of upregulated proteins.

The aim of the table is to give a quick view of the amount of records from the dataset before and after the filtering. This is a way to keep the transparency of the research.

status	records	values
Before filter	86	12553
After filter	77	10056

Table 1: The number of columns and rows before and after the filtering

The dataset started with 86 records. The first columns existed out RefSeq_accession_number for every protein. The RefSeq accession number was being stored in a variable. The goal was to make the observations in rows instead of columns. The second column contained gene symbol and the third column gene name. Both the columns were left out of the dataset for the research. After filtering out the unused columns the dataset exists of 83 records. The dataset included three replicates and three healthy patients. The last six records were filtered out to ready the dataset for machine learning experiments.

The proteins were selected based on the number of missing values. Every protein having more than twenty percent of missing values get selected out. All proteins with less than twenty percent of missing values, still has NA-values. The remaining NA-values are replaced with the mean. Roughly 2500 proteins didn't pass the filter.

The pie chart shows the data ratio between the four tumor groups. The proportional representation of patients in every tumor group is displayed in the frame.

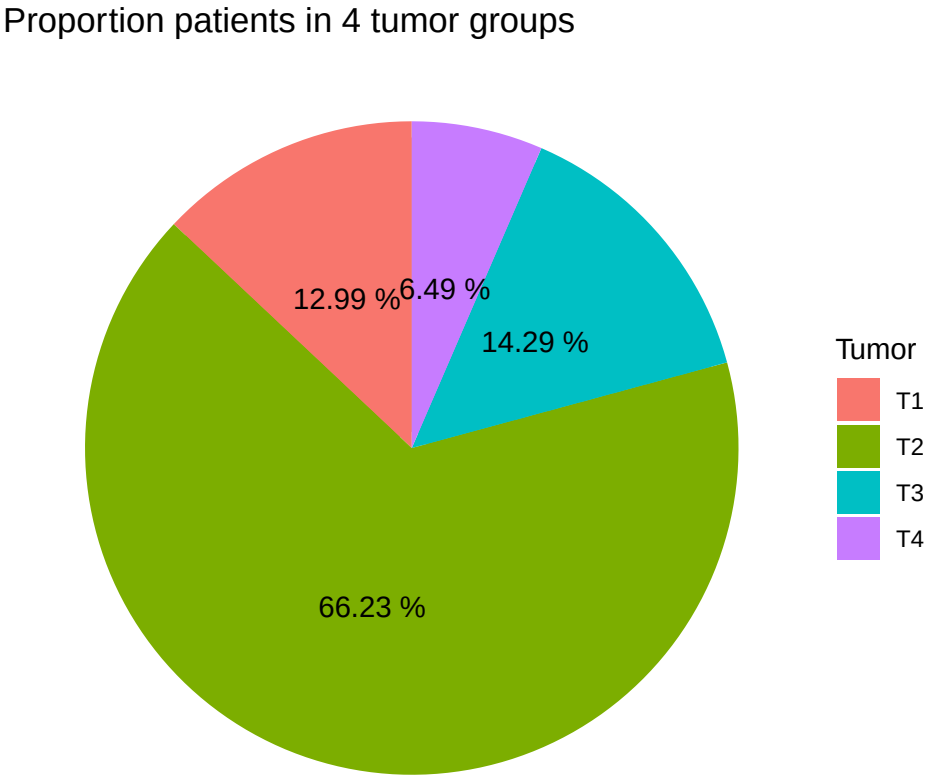

Figure 2. *Representation patients in percentages for every tumor group.*

The group T2 exists of 66.23 % of the patients. The group T1, T3 and T4 have combined less patients in comparison with group T2. This could have an effect on the accuracy of predicting the tumor group for the less represented tumor groups.

Discussion & Conclusion

The exploratory data analyses help to get a better view of the quality of the dataset. The research provides a high-quality dataset. The delivered dataset "77_cancer_proteomes_CPTAC_itraq.csv" contained 83 patients. The file included three healthy patients at the end of the file and three duplicates. Both were filtered out of the file. To predict the stage of breast cancer based on the protein expression values there's no need to keep the three healthy patients. The duplicates were removed because they give more weight in the training. This could affect the accuracy to predict the cancer stage. If the quantity of duplicates is equal to the quantity of patients there would be no major difference between the tumor groups.

The last step to improve the quality is to look for missing values for every protein. If a single protein contains more than twenty percent missing values, the protein gets filtered out. After this filter, all the remaining missing values were replaced with the mean value. The dataset doesn't have many outliers so the mean is chosen to replace the missing values.

Figure 1 indicates there are more downregulated proteins instead of upregulated proteins.

Figure 2 shows the representation for every tumor stage. All the 77 patients are divided based on the tumor stage. Tumor stage 1, stage 3 and tumor stage 4 have combined less patients in comparison with tumor stage 2. This could lead to an algorithm with a lower accuracy to predict the tumor stage 1, stage 3 and stage 4 because you have less data.

Overall the quality of the data was great. The only thing holding accountable to is the representation for every tumor group. Tumor group 2 has a lot more data available.

Future research

For the future research, filter twenty five or thirty percent of the missing value instead of twenty percent. This could improve the performance as long as there is enough data available.

Second, It's really important that every tumor group is evenly represented. Getting more data for the tumor group 1, group 3 and group 4 would impact the accuracy of prediction.

Last, instead of leaving the three healthy patients out of the dataset add more healthy patients. So the model could recognize the difference in expression value between the healthy person and the person with breast cancer.