TCGA Project

# Abstract

In this paper an attempt was made to differentiate cancer data using RNA-Seq data from TCGNA. The results showed a good separation between namely ovary, skin and testis cancer types, but it proved difficult to separate colon and rectum data. An attempt to separate the two cancer types using machine learning methods proved difficult with an FPR of x%. Feature selection was then used

# Introduction

The cancer genome project is …

Wanting to differentiate between rectum, colon, skin and testis/ovarian cancer

Whether it is rectum or colon cancer plays a big difference treatment wise (<https://healthblog.uofmhealth.org/cancer-care/how-colon-and-rectal-cancer-differ>)

As per the ***Law of Parsimony*** of ‘*Occam’s Razor’*, the best explanation to a problem is that which involves the fewest possible assumptions. Thus, feature selection becomes an indispensable part of building machine learning models. From <https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/>

# Methods

<https://www.nature.com/articles/s41597-019-0207-2>

Data was extracted from TCGA using code in Appendix 1. The data contained RNA-Seq data from Skin, Ovary, Testis, Colon and Rectum cancer patients. The criteria for the datasets was that the sample had to have at least 20 million reads and genes had to have over 1024 mapped reads in at least 10 samples. From each cancer project, 100 samples was randomly selected.

Principal component analysis (PCA) was performed using prcomp in the stats package. The two principal components was then visualized using the built in plot from base.

Differential expression (DE) analysis was performed using LmFit and eBayes functions from the limma package (version 3.12). All genes were

First, we used PCA to get an initial overview of the data.

Then used a feature selection or DEG analysis to reduce dimensionality of the data.

After that we used two different machine learning methods, namely KNN and SVM with and without cross validation to compare. Afterwards we initialized a differential co-expression network for each condition and used hierarchical clustering to find altered clusters in the different conditions differentia matrix. We only compared testis vs ovary and rectum vs colon.

Supervised learning:

(a) PCA/hierarchal clustering/heatmap to get an initial overview of the data,

(b) differential expression analysis to reduce the dimensionality,

(c) machine learning to predict outcome,

(d) extract some biological insight from the machine learning model

Implement at least two different machine learning methods and use feature selection (differential expression) and cross validation to evaluate them. Søk etter

# Results

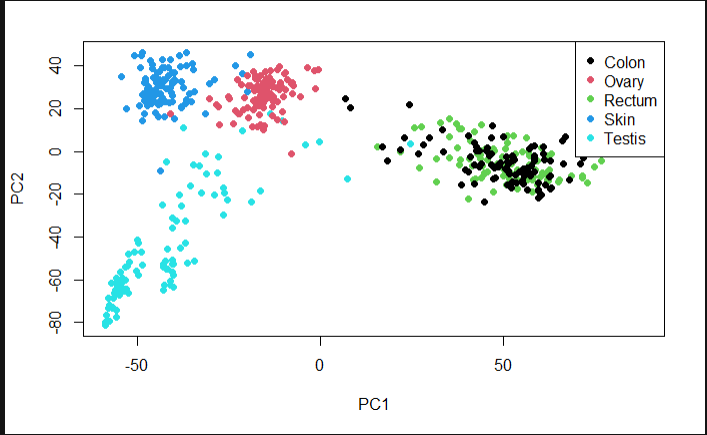


Figure 1 PCA of cancer data. Principal components only account for x% of the deviation

The PCA plot in figure 1 shows a good separation on Ovarian, testicular and skin cancer data, but struggles to separate rectum and colon as expected.

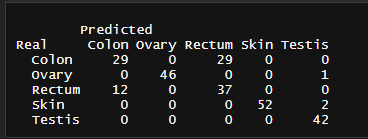


Figure 2 KNN table without cross validation

After running KNN without cross validation the initial observation from PCA plot is again observed. It appears to be difficult to separate rectum and colon data, whereas the other data is easier to separate. Without cross validation the accuracy was calculated to be 0.824, whereas with it was at 0.822.

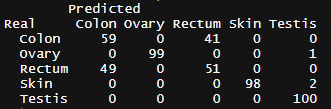


Figure 3 KNN table with cross validation showing predicted vs real

Continuing with rectum vs colon cancer in comparisment, after cross validation the accuracy has decreased somewhat (, but it is still difficult to separate rectum and colon.

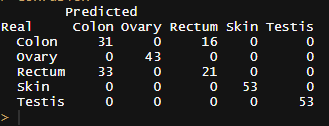


Figure 4 SVM without cross validation

SVM without cross validation shows and accuracy of 0.82 where the error lies mostly in the rectum and colon data. After running a cross validation with 10 indices the error rate was calculated to be 0.838 (see appendix x).

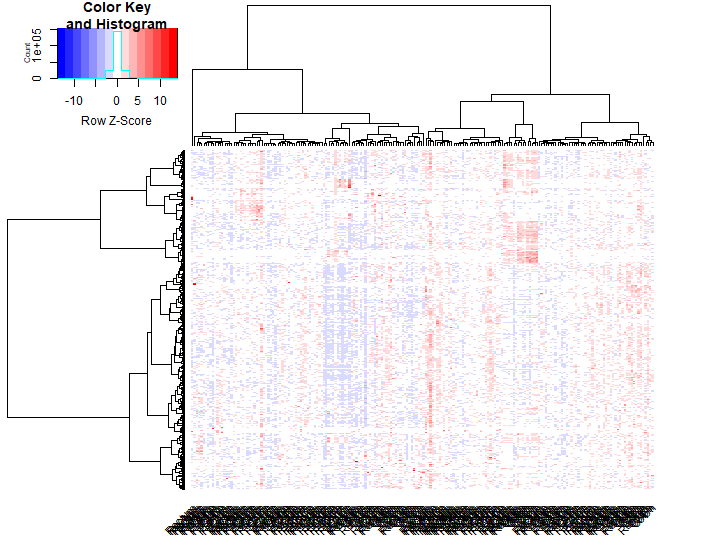
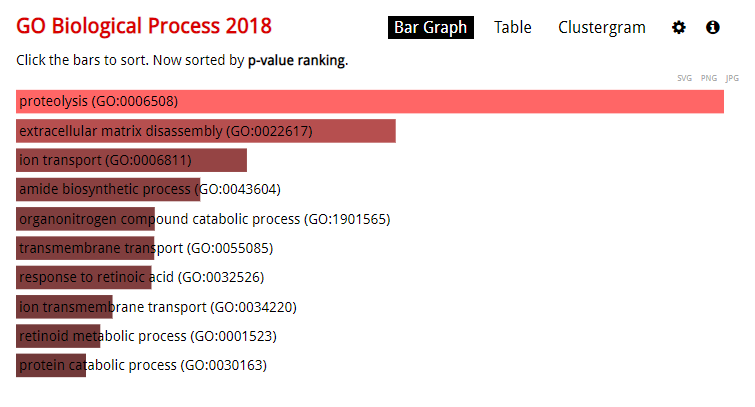


Figure 5 Heatmap with correlation for colon and rectum comparisement. Shows a high occurence of no correlation, meaning same expression in most cases.



# Discussion

The two machine learning methods has a similar accuracy at around 0.82.

# References