Biological network and data analysis

Giuliani Riccardo

July 2020

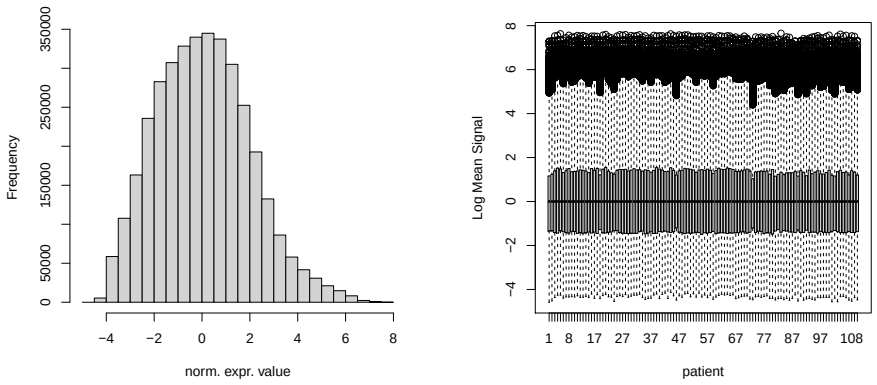
1 Introduction

The aim of this project is to analyze a gene expression profile dataset of 80 patients affected by lung cancer (divided in 40 smokers and 40 non smokers) and compare it with the set of genes belonging to 30 healthy individuals. With this work will maybe possible to infer possible biological variations behind the malignancy in these individuals, hoping they could be used to better understand the causes of this specific tumor and help the development of a cure.

2 Analysis and results

2.1 Data Normalization

Data present this dataset were already normalized and log2 transformed. In this step it was only applied a median normalization to increase the robustness of data to the presence of outliers in the further steps.

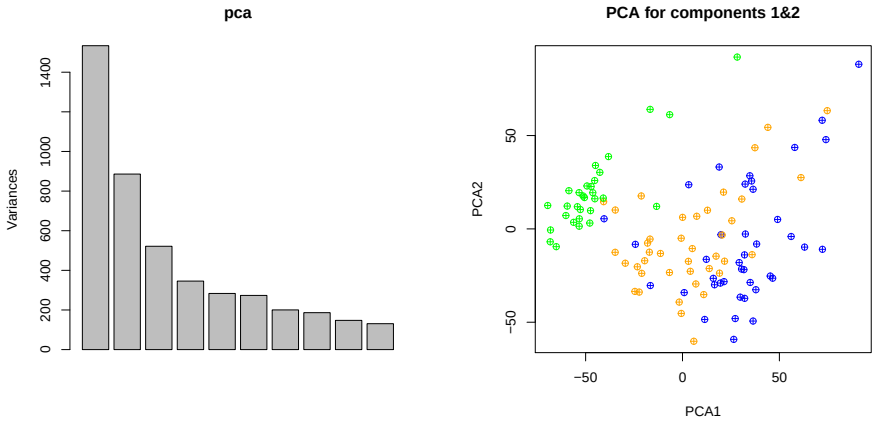


(a) *Overall data histogram* (b) *Dataset boxplot*

Figure 1: From the histogram(a) we can assess that data tends to follow a normal distribution, meaning the normalization was successfull, while in the boxplot(b) we can notice the presence of some outliers in the top part and that the median normalization has shifted all the medians to 0 value

2.2 PCA

To perform a quick overview of the dataset a PCA analysis have been performed by labelling with different colors the three categories analyzed.

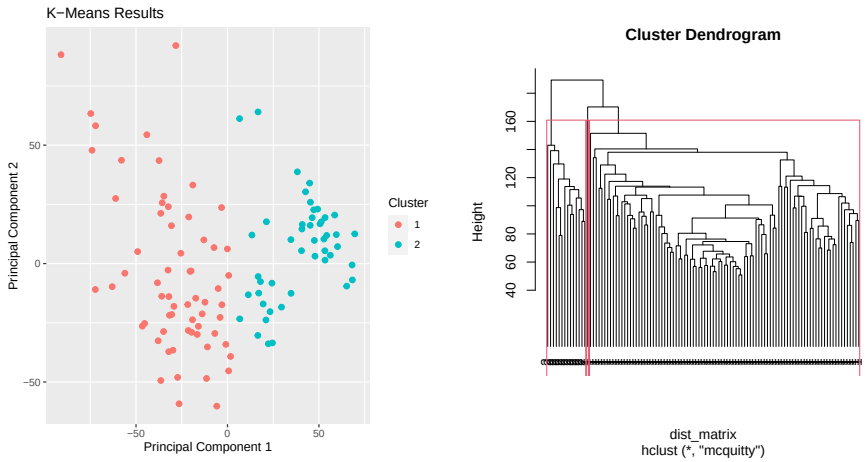


(a) *Screeplot* (b) *PCA*

Figure 2: The screeplot(a) suggests the presence of few components containing most of the information about data. In the PCA(b) the first two components were used to plot the data, it is clearly visible a distinction between normal (green) and tumors (orange and blue), moreover it is possible to see a different pattern between affected smokers (blue) and non-smoker (orange)

2.3 Unsupervised algorithms

Unsupervised learning is a type of machine learning with the aim to find patterns in the dataset able to divide into subgroups its elements, without pre-existing labels given by the user and human supervision. In the further step of the analysis I employed two relatively simple unsupervised method, k-means and hierarchical clustering.



(a) *K-means* (b) *Hierarchical clustering*

Figure 3: K-means(a) was able to divide data into two groups of 47 and 63 points respectively, whilst Hierarchical clustering(b) two groups of 96 and 14 points respectively

k-means was able to to identify 2 clusters (figure 3.a) in which the dimensionality, with a certain degree of error, can be comparable with the original proportion between affected and control patients. On the other hand hierarchical clustering (figure 2.b) has shown poor results, resulting unable to distinguish the two data types. In the last step of this analysis data were divided in three subcategories, to check if they were able to subgroup the tumoral set of patients into smoker and non-smoker. Unfortunately none of them has proved successfull in this specific case.

2.4 Supervised algorithms

The next step of the analysis consist on using a set of supervised algorithms to perform an analysis on the database. Supervised learning is the algorithm task of learning a function that maps an input to an output based on example input-output pairs. They infer a function from labeled training data consisting of a set of training examples, and are usually validated by applying this function to a set of testing data.
Random forest: The algorithm was successfully able to divide the tumoral population from the normal, moreover in a further step was able to distinguish between the two tumoral subgroups (smoker and non smoker) either.

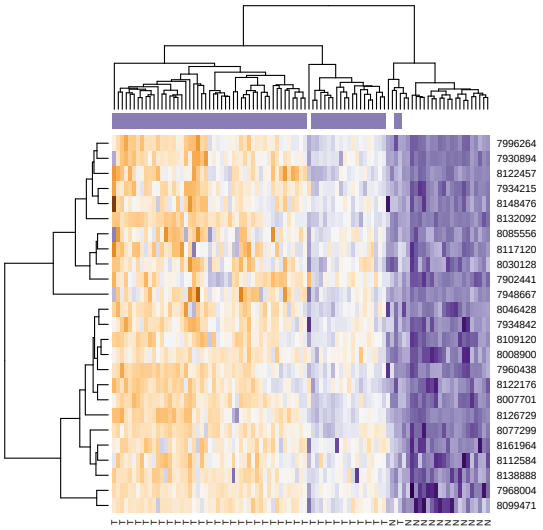


Figure 4: Heathmap of the best 200 genes able to divide the tumoral population from the normal, notice the presence of some borderline results

LDA: Prior to LDA analysis it was necessary apply a feature selections to reduce the dimension of the dataset in order to fit for the LDA algortithm.
Then LDA was performed on a set of tumoral and normal samples of the reduced dataset and on a dataset of only tumoral samples. In both cases it has proved successfull displaying a AUC of 1. LDA and random forest model were better evaluated performing a 10 times repeated 10 fold cross validation to reduce biases derived from the analysis. Both method have displayed very good results with an accuracy close to 1 (figure 5).

Lasso regression: The technique was applied in the reduced dataset of tumoral samples, it was able to fully classify the data provided in the test set, resulting in a AUC of 1.

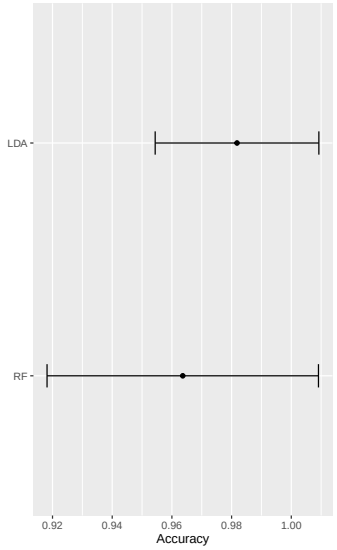


Figure 5: Accuracy derive from 10 times repeated 10 fold cross validation. LDA on the top, random forest on the bottom

SCUDO: Scudo analysis was performed on the whole dataset, labelling normal patients and the two tumoral sub populations.

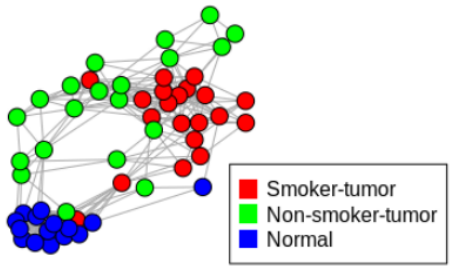


Figure 6: Scudo plot, distinctions among the three populations are clearly visible

2.5 Functional enrichment analysis

This step on the analysis consist on trying to infer the biological implications that can be the cause for the disease. Specific tools allow to get the biological implications of a set of genes given in input by querying specific databases and identify the GO categories with significantly more density of genes than expected by chance. This method is quite useful because analyzing the most relevant genes in distinguishing the tumoral patients from the normal, it is possible to infer the biological causes behind the disease.
Before proceding with the analysis was necessary to select the most informative genes present in the dataset, to this scope the 200 most informative genes were selected from random forest.

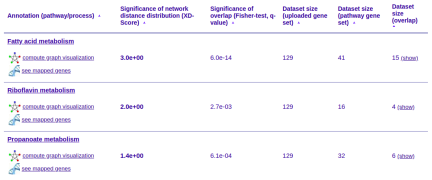
DAVID: 164 of the 200 provided genes were found by DAVID and used for the analysis.
g:Profiler: By comparing the two analysis done with the two different software, discordant results about the genes functionalities were obtained, probably because the set of genes provided in input is not unambiguous, the two software query different annotated databases and they use different algorithms for the analysis.

2.6 Network analysis

In the final part of the project was performed a network analysis, that like the functional enrichment one has the scope to infer the possible biological variations that are the causes of the malignancy in patients. This analysis differs from first because with these starting list of genes it builds a network, sometimes by adding new nodes formed of closely related genes stored in specific databases. An enrichment analysis is finally performed on the network.

Cytoscape Only 106 of our total 200 genes were found against kegg patways db. From this analysis was found out there is a correlation between the genes given in input and the fatty acid related pathway, retinol metabolism and the production of riboflavin and tyrosine. Moreover it was also possible to look which specific tissues were the most related with the set of genes, interesting to notice the low correlation with tumor and very high correlation with fetal lung tissue.

Enrichnet: This tool provided similar results to Cytoscape, giving a further suggestion about the involvement of input genes in the fatty acid,riboflavin and tyrosine metabolisms, but it didn't seem to notice any implication in retinol metabolism as suggested by cytoscape.



(a) *Enrichnet*

Pathway or Process	XD-score
Fatty acid metabolism	1.74552
Retinol metabolism	1.61090
Drug metabolism - cytochrome P450	1.09620
Tyrosine metabolism	1.09620
Riboflavin metabolism	0.95181
Propanoate metabolism	0.91354

(b) *Cytoscape*

ID	Term_Description	Fold_Enrichment	occurrence	lowest_p	highest_p
R-HSA-392154	Nitric oxide stimulates guanylate cyclase	14.650633	10	5.2e-05	5.2e-05
R-HSA-177929	Signaling by EGFR	4.156208	10	2.1e-03	2.1e-03
R-HSA-418346	Platelet homeostasis	4.650995	10	4.3e-03	4.3e-03
R-HSA-203615	eNOS activation	8.879171	10	6.5e-03	6.5e-03
R-HSA-419812	Calcitonin-like ligand receptors	9.767089	10	8.0e-03	8.0e-03

(c) *PathfindR*

Figure 7: Enrichnet(a), Cytoscape(b) and PathfindR (c) most relevant pathways

PathfindR:
This analysis found different results from the previous (figure 7.c) two, prbably due to the different approach used by this tool from the others.

3 Materials and Methods

- PCA: unsupervised technique that relies on eigenvectors and eigenvalues to find the main direction responsible for the partition of the data
- k-means: K = 2 and K = 3 in two separated analysis
- Hierarchical: K = 2 and K = 3 in two separated analysis
- Random forest: algorithm that builds a set of trees where each node is a chosen attribute able to divide into two subgroups the data. The selection of these attributes is based on the entropy of the information they give, nodes on the top are the best in reducing the overall entropy of the data partition. Bagging usually is applied to choose the best attributes to avoid possible biases. nof trees = 800, seeds = 1357, in each analysis 5 samples for each subset are left out as test set.
- LDA: supervised technique able to apply a dimensionality reduction to keep the best attributes for dividing the dataset into the labelled subgroups. In order to do so it employs an objective function to maximize the

distance of the projection of the centroids and minimize the scattering of each cluster. The feature selection was performed by filtering. Row t-test adjusted with Benjamini Hochberg was performed and the genes with p-value ¿ 0.001 were maintained.

- Random forest and LDA validation: performed with caret R package, applying a 10 times repeated 10 folds cross validation.
- Lasso regression: technique derived from linear regression, that uses data (predictor) to build a linear model. In order to choose the best model often relies on minimizing the least squares function. The least squares function is adjusted by the squared sum of the beta terms in order to perform a dimensionality reduction ad selecting the most informative attributes. Performed by glmnet R library, applying a 5 fold cross validation.
- SCUDO: peculiar algortithm designed to overcome the problem of batch effect. It consists on sorting the list of genes for each patient by decreasing expression level. Then it selects the top and bottom elements of these lists and perform a gene set enrichment analysis to build a sort of distance matrix among samples to build a distance related graph. Performed from R library rScudo, nTop = 25, nBottom = 0.25, N = 0.3
- DAVID and g:Profiler: online enrichment tools able to query different databases and to obtain erichment results about possible pathways implication, tissue specificity etc. of the set of genes given in input. selected the 200 most informative genes from Random Forest. Gene IDs were converted in Entrez with DAVID ad hock tool (196 over 200 correctly converted), In the selection of unambiguous genes only the first term was kept. These 196 genes are used in both analyses. Analysis done on pathway databases only, relevant genes selected with a p-value ¿ 0.001
- Cytoscape: the same set of 196 genes used in DAVID and g:profiler were reused for this analysis. Analysis done querying Kegg database.
- Enrichnet: tool based on finding closely related genes with the seed genes and rank all the possible pathways basing on the distance between the seed genes and the related ones by performing a random walk for the evaluation giving in output a XD-score.To avoid selection by chance Fisher ttest or hypergeometric test are employed. Same condition used in cytoscape reused for this analysis
- PathfindR: tool that look in which pathways sub network our genes are. Then basing on the overall number of genes present on each subnetwork and the p.value provided by the user for each gene, it filters the most significant pathways related with our genes. 196 relevant genes converted from Entrez gene ID to official gene symbols. Only 113 over 196 were correctly converted in this step.

4 Conclusion

With this project was possible to experiment different analysis strategies on a cancer vs control dataset. In particular the PCA analysis gave good results in distinguishing the set of patients, in contrary the other unsupervised learning method provides not sufficiently good results.
Supervised learning method provided better result in distinguishing the subsets of patients, that is not necessary a good result because it can be determined by overfitting, despite a quick cross validation had still given good results. Only by using these models on different datasets in further analysis would be possible to determine the true effectiveness of these methods.
Enrichment analysis with DAVID and gene profiler had given different results, about in which pathways are affected by our genes. Different results from the previous were obtained also with network analysis, despite the fact that Enrichnet and Cytoscape came to similar conclusions.