SWIM analysis on gene regulatory networks

Davide Toma

October 1, 2020

Contents

1 Abstract

We are going to analyse, through SWIM procedure, one of the most common cancer, the Lung Squamous Cell Carcinoma. SWItchMiner (SWIM) is a software implementation of a procedure, able to extract information contained in complex networks. Specifically, SWIM allows unearthing the existence of a new class of hubs, called “fight-club hubs”, characterized by a marked negative correlation with their first nearest neighbors. Among them, a special subset of genes, called “switch genes”, appears to be characterized by an unusual pattern of intra - and inter - module connections that confers them a crucial topological role. During the analysis we will have as dataset, a set of genes monitored in two different conditions: normal and tumour. First of all, will be found, properly applying SWIM software, the set of ”switch genes” related to drastic changes in many biological settings. Once this set of genes is found, we will apply an enrichment analysis with the web tool 'Enrich r' that will find out if our genes have a statistical significance with some processes and interactions in 4 different categories: Ontologies, Diseases/Drugs, Transcription and Pathways.

2 A brief overview of the tumor

Squamous-cell carcinoma (SCC) of the lung is a histologic type of non-small-cell lung carcinoma (NSCLC). It is the second most prevalent type of lung cancer after lung adenocarcinoma and usually occur in the central part of the lung or in one of the main airways (left or right bronchus). Its tumor cells are characterized by a squamous appearance, similar to the one observed in epidermal cells. Squamous-cell carcinoma of the lung is strongly associated with tobacco smoking and this is very important because LUSC is one of the tumor types with the highest number of mutations since smoking, the main driver of the disease, is a strong mutagenic factor. [1]

3 SWIM

In this part of the report, we will discuss about the decisions we took in order to get the set of switch genes with the help of SWIM software. SWIM allows to find new hubs, called “fight-club hubs”, with the characteristic of a negative correlation with their first nearest neighbors. Then we will try to identify a subset of genes, called “switch genes”, characterized by an unusual patterns that give them a crucial role.

In other words, the purpose of our analysis with SWIM is finding switch genes that could be associated with changes in the physiological state of cells or tissues induced by the cancer development.

3.1 Preprocessing phase

First of all we started the preprocessing phase where, after the charging of the data we know we have:

- Number of columns = 76
- Number of rows = 20531
- Number of microRNAs = 1046

Our first decision is to log2-transform the data. Then, we had to fix two thresholds in order to filter out genes that had expressions equal to zero more than the chosen treshold or that change very little according to the IQR percentile measure. The choice of these two threshold was made looking at the following plots, and after some trials and errors we found good values for the parameters we were looking for:

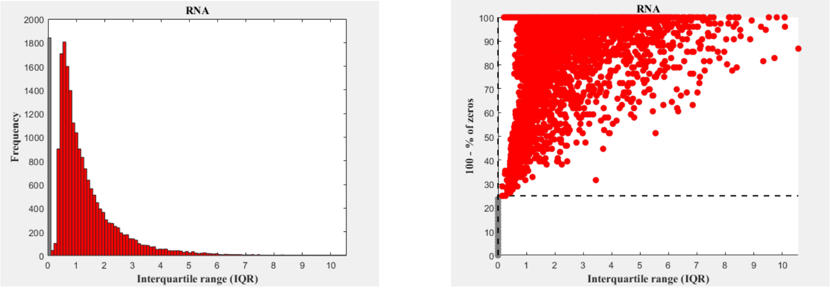

Figure 1: RNA discarded genes

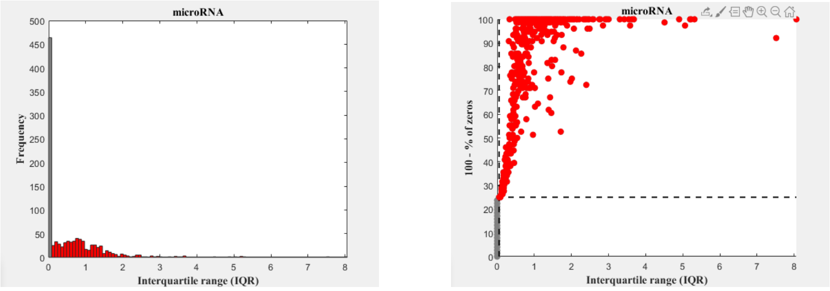

Figure 2: microRNAs discarded genes

From the two plots it's clear that we reached the final goal of this phase. In fact, all the samples with a lot of zeros were discarded both for RNA and microRNAs.

These are the final choices for the thresholds:

- Treshold for allowed zeros = 75
- Treshold for IQR = 8
- Treshold for allowed zeros for microRNAs = 75

- Treshold for IQR for microRNAs = 44

These parameters set the number of allowed zeros for RNA and microRNAs equal to 57 over 76. In this way the number of rows kept is 18691 and the for the microRNAs is 586.

3.2 Differentially expressed genes - Statistical significance filters

After preprocessing, we had to fix two other parameters in order to keep only differentially expressed genes between the two states (tumor and normal). In other words, we want only genes that have a statistical significance in their expressions between the tumorous and the normal conditions.

The first parameter has influence on the log fold-change, that allowed us to identify genes whose expression in the two groups of samples considered varies by a certain proportion, while the second is referred to the False Discovery Rate (FDR) value and is a correction method to avoid the false positive problem that affects the log fold-change method.[2]

Also for the choice of these parameters we used the approach of try and evaluate the results, and the final choice for the parameters is the following:

- Linear fold change = 4.74
- FDR = 0.05

We can see from the plot below how these two parameters work on the dataset:

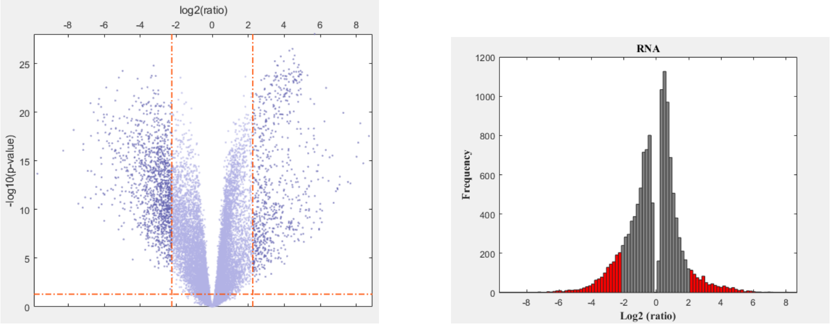

Figure 3: RNA discarded genes

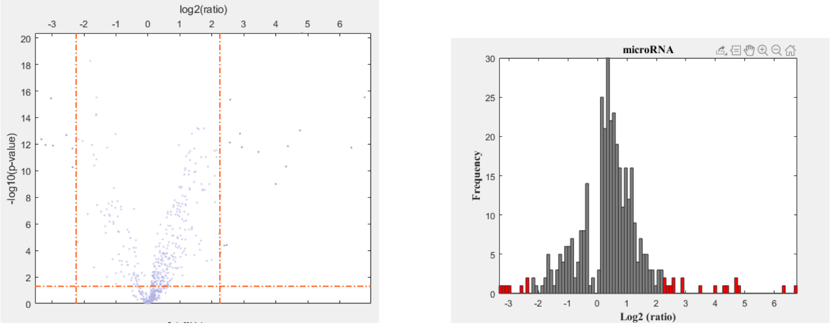

Figure 4: RNA discarded genes

In this way we kept 2059 rows and 24 microRNAs. The number of rows is exactly the 10% of the starting ones (20531), and this is good because one way to be sure the parameters chosen were correct was to check if the ratio between the rows we kept and the initial rows is around 10%, so we can proceed in the next phase of the analysis.

3.3 Correlation network

Now we had to calculate the correlation between a pair of genes thanks to a given threshold for the Pearson Correlation. This value had to be chosen adding the lowest number of edges that kept the number of connected components as low as possible. [2] We chose the parameter for the correlation percentile equal to 0.85 and we got the following results:

- Correlation threshold = 0.74
- Number of selected correlated pairs = 502727
- Number of nodes in the correlation network = 2001

The image below shows which were, according to the correlation, the grey genes (the ones we discarded) and the red ones, the ones we kept in our network.

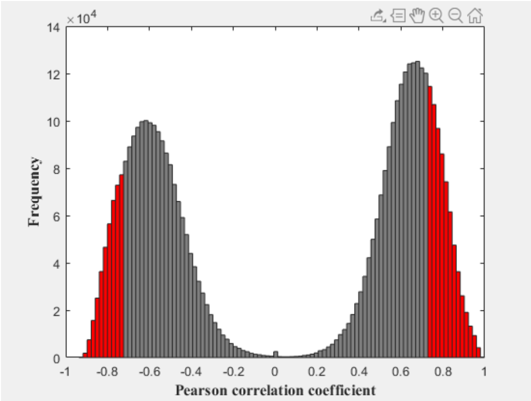

Figure 5: Genes distribution according to the correlation

3.4 Local communities

We now have a gene co-expression network (GCN), an undirected graph, where each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them or, in other words, if and only if they have a correlation higher then the one fixed in the previous point.[2] SWIM then tried to make clustering inside this network running the k-means algorithm. We decided run K-means 10 times for each dierent values of K from 1 to 10, and the K that best divided our data was found with the Elbow method.

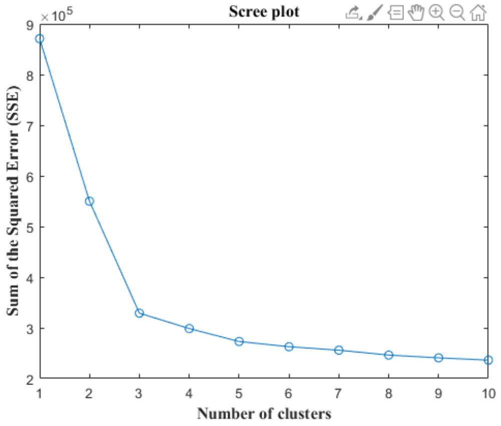

Figure 6: Elbow method

From the picture above we can say that the best value of K was 3, according to the Elbow method, and the sum of the squared error with K equal to 3 is 329017.

3.5 The set of switch genes

Built the network and found the local communities, now we want to assign a class to each community. The Swim software assigned classes based on the Heat Cartography map for our correlation network. The following plot shows seven regions identified with R1-R7. The nodes are coloured with respect to their average Pearson correlation coecient and the blue ones are the fight-club hubs, nodes with an average negative correlation in expression with their interaction partners (the first order neighbors). The fight-club hubs that fall into the R4 region are the switch genes, that are the genes we wanted to isolate with this first part of the analysis. [2]

There were found 340 switch genes out of 1973 nodes.

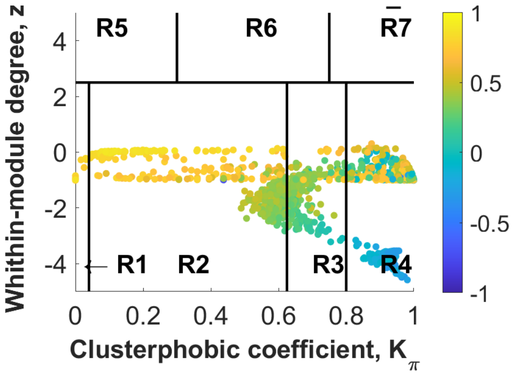

Figure 7: Genes clusters

Now we show another plot that represents the average Pearson correlation coefficient distribution of our network. We can see that the distribution has three peaks, which respectively represents three types of hubs: fight-club hubs (on the left), date hubs (center) and party hubs (right).

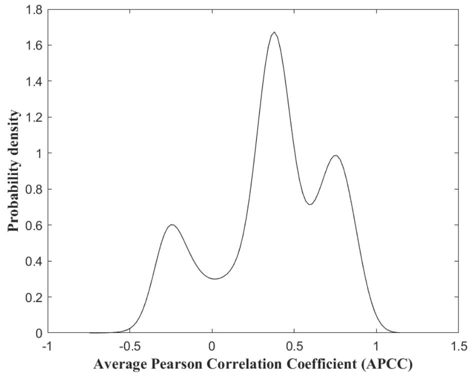

Figure 8: Average Pearson correlation distribution

3.6 Connectivity considerations of the network

In this last part of the SWIM analysis we wanted to check the connectivity of our correlation network.

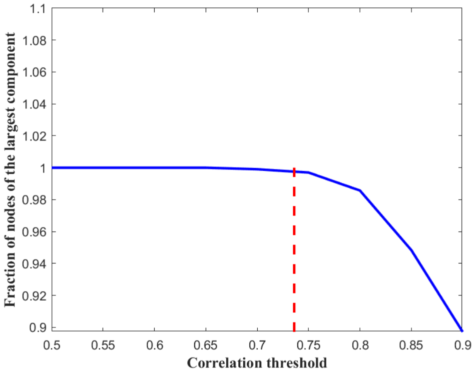

Figure 9: Network connectivity

The plot above has a range of values that vary from 0.5 to 0.9 (from 0.1 to 0.5 it was, obviously, fully connected) and shows that the network is fully connected until 0.75, then it decreases really fast. As we said before during our analysis, our pick was 0.74 that kept the network fully connected and this has a reply in the plot.

The last result of the analysis with SWIM is an overview on the integrity of the network. In fact, the software tried to eliminate each time a node from one of the classes (hubs, fight-club, date, party) and evaluate the average shortest path every time. This part was really slow and computational expensive, but the resulting plot is really interesting and we can see that even if we remove a really small fraction of switch genes, the average shortest path increases. There is also another interesting fact we can see from the plot: in fact the random removal of nodes has a collapse in the trend that however can be explained by the randomness in the choice of the fraction of the removed nodes (and it's also a small change in the order of the second decimal place).

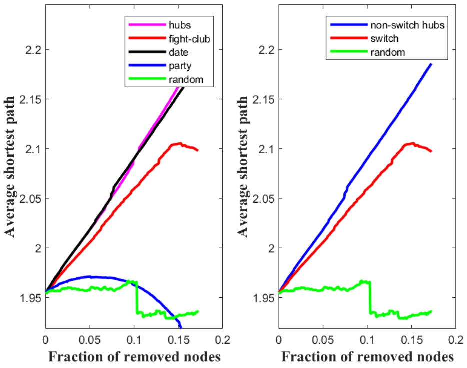

Figure 10: Network connectivity

4 Enrichr

Now we will use a web tool to continue our analysis starting from the set of switch genes we found in the previous analysis with SWIM. For our Enrichr analysis, we considered 4 different categories with a subset of their datasets:

- Ontologies: GO Biological Process 2018, GO Molecular Function 2018, GO Cellular Component 2018
- Diseases/Drugs: DisGeNET, GWAS Catalog 2019
- Transcription: TRANSFAC and JASPAR PWMs
- Pathways: KEGG 2019 Human

4.1 Ontologies

We started our analysis with Enrichr by looking at the ontology and we considered these three different types of datasets in the fields of Biological process, molecular function and cellular component. We found out that the most enriched terms are:

- DNA metabolic process (GO:0006259) with p-value of 9.82×10^{-19}
- DNA helicase activity (GO:0003678) with p-value of 4.34×10^{-11}
- spindle (GO:0005819) with p-value of 6.47×10^{-22}

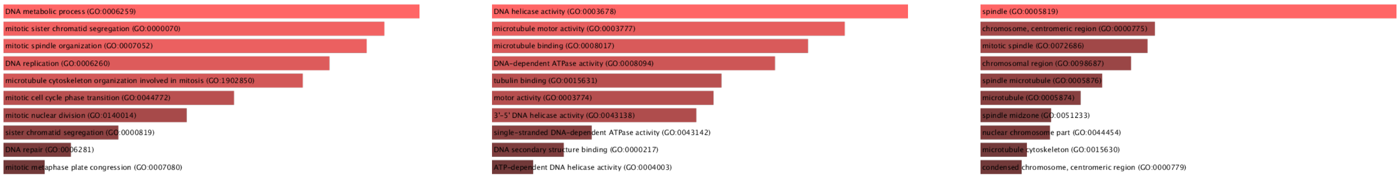

Figure 11: Enrich r output for Ontologies

Now we will focus on the MCM4 gene that, as we can see thanks to the following cluster-gram (that underline the single association between an input gene and an enriched term), has the lowest p-value associated with the DNA metabolic process.

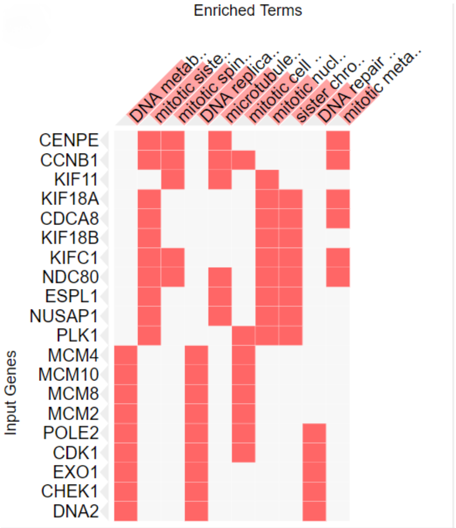

Figure 12: Clustergram ontologies - DNA metabolic process

MCM4 is a gene that encodes a really important protein; the protein encoded by this gene is one of the highly conserved mini-chromosome maintenance proteins (MCM) that are essential for the initiation of eukaryotic genome replication. [4]

4.2 Disease/Drugs

In this part we worked with two of the biggest datasets for gene disease association and gene drugs association, DisGeNet and GWAS Catalog 2019.

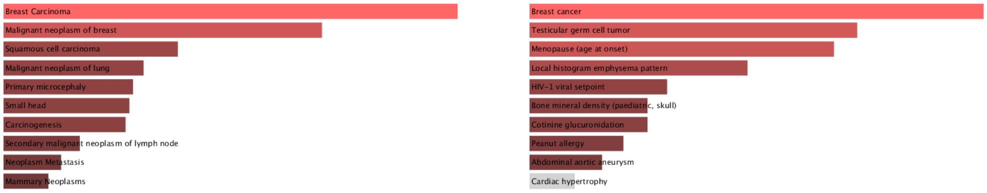

Figure 13: Enrich r output for Diseases/Drugs

We found out that the most enriched diseases are breast carcinoma (for DisGeNET) and breast cancer (for GWAS Catalog 2019); so, from two different datasets we found the same enriched term. Breast cancer is the most common cancer diagnosed in women. Survival rates in the developed world are high, with between 80% and 90% of those in England and the United States alive for at least 5 years while in developing countries, survival rates are poorer.

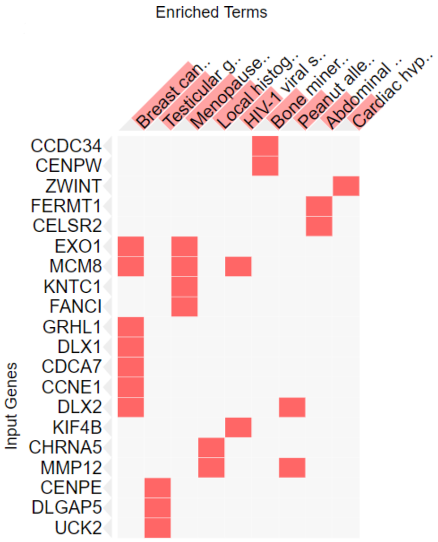

Figure 14: Clustergram disease/drugs: Breast cancer

EXO1 is the gene with the highest p-value. This gene encodes a protein with 5' to 3' exonuclease activity as well as an RNase H activity (so in the transcription framework). It is similar to the Saccharomyces cerevisiae protein which interacts with Msh2 and which is involved in mismatch repair and recombination. Alternative splicing of this gene results in three transcript variants encoding two different isoforms. [5]

4.3 Transcription

Now we wanted to consider the transcription framework of Enrichr, and as usual we evaluated the bar plot given from the software for the dataset we are considering in this framework:

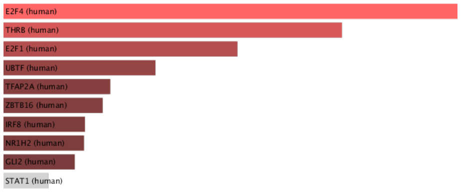

Figure 15: Enrichr output for Transcription

We can see that the most enriched factor in this case is E2F4 (human) with p-value of 2.303×10^{-4}. The protein encoded by this gene is a member of the E2F family of transcription factors. The E2F family plays a crucial role in the control of cell cycle and action of tumor suppressor proteins and is also a target of the transforming proteins of small DNA tumor viruses. The E2F proteins contain several evolutionally conserved domains found in most members of the family. These domains include a DNA binding domain, a dimerization domain which determines interaction with the differentiation regulated transcription factor proteins (DP), a transactivation domain enriched in acidic amino acids, and a tumor suppressor protein association domain which is embedded within the transactivation domain. It plays an important role in the suppression of proliferation-associated genes, and its gene mutation and increased expression may be associated with human cancer. [3][6]

4.4 Pathways

The last framework we considered is Pathways and we focused on the KEGG 2019 human dataset. From the analysis of the bar plot in the figure it turned out that the most significant category, according to the p-value, is the cell cycle.

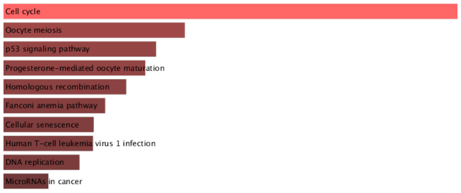

Figure 16: Enrichr output for Pathways

The next step was analyse the associated cluster-gram to find out which one is the gene with the lowest p-value with the cell cycle category that was the most significant:

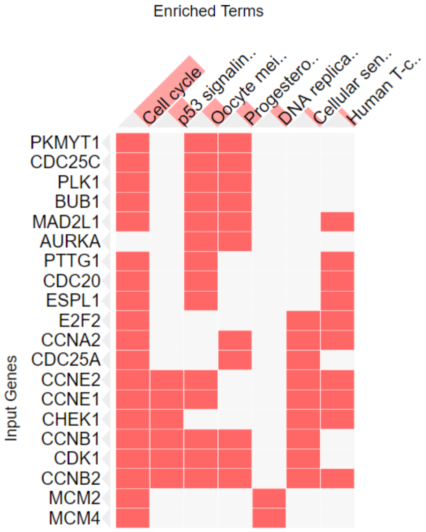

Figure 17: Enrichr clustegram for Pathways

We found out that the gene most correlated with the cell cycle category that was the most enriched one is PKMYT1. This gene encodes a member of the serine/threonine protein kinase family. The encoded protein is a membrane-associated kinase that negatively regulates the G2/M transition of the cell cycle by phosphorylating and inactivating cyclin-dependent kinase 1. The activity of the encoded protein is regulated by polo-like kinase 1. Alternatively spliced transcript variants encoding multiple isoforms have been observed for this gene. [7]

5 Conclusions

This project had without doubts some interesting results. In the first part with SWIM we found a set of 340 switch genes that with the help of doctors and biologists, and with more advanced researches, could be an interesting starting point to fight the squamous cell carcinoma of the lung. In the second part of the project, thanks to Enrichr, we showed how the set of switch genes found has implication with some fundamental processes in the human organism, such as the cell cycle (with particular attention to the transcription process) and the DNA replication, and the study of their interactions can make possible, in future times, really significant discoveries in the field of the medicine and biology.

6 References

Here some documents we found useful to perform the analysis:

[1] ”Squamous cell carcinoma of the lung” - Harvard health publishing - https://www.health.harvard.ed/
[2] ”SWIM software” - Teaching material
[3] ”Cell cycle” - TeachMe Physiology - https://teachmephysiology.com/basics/cell-growth-death/cell-cycle/
[4] ”MCM4” - GeneCards - https://www.genecards.org/cgi-bin/carddisp.pl?gene=MCM4
[5] ”Exo1” - GeneCards - https://www.genecards.org/cgi-bin/carddisp.pl?gene=Exo1
[6] “E2f family” - https://www.genecards.org/cgi-bin/carddisp.pl?gene=e2f4
[7] ”PKMYT1” - GeneCards - https://www.genecards.org/cgi-bin/carddisp.pl?gene=PKMYT1