Universitat Oberta de Catalunya (UOC)

Master's Degree in Data Science (*Data Science*)

MASTER'S THESIS

Area: Data Mining and Machine Learning

Study of the Transcriptional Function of Cyclin D1 in Leukemia

Comparison between the gene expression regulated by Cyclin D1 in lymphomas and the gene expression in DNA damage.

Author: Antonio Milán Otero

Tutor: Carles Barceló

Professor: Jordi Casas Roma

Barcelona, June 12, 2019

License

This work is licensed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License.

Creative Commons 4.0 International License.

THESIS INDEX CARD

Title:	Study of the transcriptional function of Cyclin D1 in leukemia
Author:	Antonio Milán Otero
Teacher collaborator:	Carles Barceló
Teacher responsible for the subject:	Jordi Casas Roma
Date of delivery (mm/aaaa):	06/2019
Degree or program:	MSc in Data Science
Thesis area:	Data Mining and Machine Learning
Language:	English
Keywords	Leukemia, Cyclin-D1, Machine-Learning

Dedication

This thesis and the conclusion of the MSc in Data Science that it represents is dedicated to my wife Cristina. Without her support and love it would be impossible to accomplish any of this. Thanks for all the support during the long weekends that we had to spend at home because I had work to finish, you never complained, and even better, you always had a smile and a word of encouragement to cheer me up and help me to continue at difficult times.

Acknowledgment

I would like to thank my advisor, Carles Barceló for all the guidance, support and patience offered during the execution of this project.

In addition, I would like to thank Jordi Casas Roma and all the professors involved in the MSc in Data Science for all the teachings provided along this MSc degree.

Finally, thanks to Universitat Oberta de Catalunya for providing the necessary resources to make the MSc in Data Science a reality.

Abstract

Leukemia is a type of cancer that starts in blood-forming tissue, such as the bone marrow. It causes the production of large numbers of abnormal blood cells that end up entering into the bloodstream. [1]

Within the different types of leukemia, Mantle Cell Lymphoma (MCL) is the one with the worst prognosis due to the short survival average of a patient, which is close to three years. This tumor is characterized by the over-expression of Cyclin D1, a protein that helps control cell division. MCL is also characterized by the binding of this protein to certain regions of DNA involved in the regulation of DNA-damage response (DDR).

The presented study aims to identify similarities between the gene expression regulated by Cyclin D1 in lymphomas and the gene expression in DNA damage. That knowledge will allow the exploration of essential mechanisms of carcinogenesis and help in the identification of genes that could be an interesting therapeutic target in the process of tumor progression. Additionally, new biomarkers that could be used in early diagnosis can be found.

With the addition of Machine Learning algorithms to the biology analysis pipeline, this project explores new ways to improve the traditional methodologies and boost the identification of significantly enriched genes that will serve the purposes mentioned above.

The result of such a pipeline is the accurate selection of genes correlated with Cyclin D1, involved in MCL and DDR, and its posterior analysis and identification of significantly enriched gene sets.

As a conclusion, the results obtained in this study suggested that targeting of Notch pathway and studying potential common mechanisms of hypoxia and apoptosis resistance would be of great interest for possible future studies on treatments of MCL.

Keywords: Leukemia, Cyclin-D1, Machine-Learning.

[1] https://www.cancer.gov/publications/dictionaries/cancer-terms/def/leukemia

Resumen

La leucemia es un tipo de cáncer que empieza en los tejidos generadores de sangre, tales como la médula ósea. Esta enfermedad causa la producción de un gran número de células sanguíneas anormales que van a parar al flujo sanguíneo. [2]

Dentro de la leucemia encontramos diferentes tipos, siendo el Linfoma de las células del manto (en adelante, Mantle Cell Lymphoma o MCL) el que peor pronóstico tiene, debido a la corta media de supervivencia del paciente, cercana a los tres años. Este tumor se caracteriza por la sobre-expresión de la Cyclina D1, proteína que ayuda a controlar la división celular, y también por la unión de ésta proteína a ciertas regiones de ADN involucradas en la regulación del proceso de reparación del daño al ADN (en adelante, DNA-Damage Response o DDR).

El presente estudio tiene como objetivo identificar las similitudes entre la expresión génica derivada de la regulación por la Cyclina D1 en Linfoma con la que se da en el caso del daño en el ADN. Este conocimiento permitirá la exploración de los mecanismos esenciales de carcinogénesis y ayudará en la identificación de genes que pueden ser un objetivo terapéutico interesante en el proceso de progresión tumoral. Adicionalmente, se podrían hallar nuevos biomarcadores para el diagnóstico precoz de la enfermedad.

Con la adición de Machine Learning al proceso de análisis biológico, este proyecto explora nuevas formas de mejorar las metodologías tradicionales e impulsar la identificación de genes significativamente enriquecidos que servirán a los propósitos mencionados anteriormente.

El resultado de dicho proceso analítico es la precisa selección de genes correlacionados con la Cyclina D1, involucrados en MCL y DDR, y su posterior análisis.

Como conclusión, los resultados obtenidos en este estudio sugirieron que el tratamiento del *Notch pathway* y el potencial estudio de los mecanismos comunes de resistencia a la hipoxia y apoptosis serían de gran interés para posibles estudios futuros sobre tratamientos de MCL.

Keywords: Leucemia, Cyclin-D1, Machine-Learning.

[2] https://www.cancer.gov/publications/dictionaries/cancer-terms/def/leukemia

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General description of the problem

Leukemia is a set of tumor processes that causes an uncontrolled increase in leukocytes (white blood cells) in the blood or lymphatic organs.

Cyclin D1 is an oncogene frequently overexpressed in cancer, especially in leukemia. It is known that Cyclin D1 is one of the main regulators of the cell cycle, but its role as a regulator of transcription (the process that generates the proteins needed to control all cellular processes) remains unknown.

It is also well-known that Cyclin D1 binds to the promoter regions of many genes, although the result of its transcriptional activity remains unknown as well. That transcriptional activity is believed to be fundamental in the development of leukemia.

After the creation of the Gene Expression Omnibus repository[1][2] an enormous amount of genomics data is publicly available for its study. Among others, data related to leukemia is available and susceptible to be analyzed by Data Mining and Machine Learning techniques, being its interpretation fundamental to know the basic mechanisms of the cells that can lead to leukemia. Obviously, the generation of new drugs will depend on knowing these processes in detail.

In human cells, both metabolic activities and environmental factors, such as UV rays or radioactivity, can cause DNA damage. Many of these lesions produce potentially harmful mutations in the genome of the cell, which affects the survival of their descendant cells at the time of mitosis or induces malignant processes that end up leading to a tumor.

Several human cancers have been linked to DNA abnormalities such as dislocations, deletions and mutations. The clarification of the mechanisms that initiate the process of repairing DNA damage (DNA-damage response or DDR) will lead to improve the prediction of cancer risk and the treatment in the early stages. More extensive studies of the damage and DNA repair

pathways could lead to the development of new therapies aimed at strengthening the natural defense systems of the cells that prevent a tumor from being developed.

Within leukemia, Mantle Cell Lymphoma (MCL) has the worst prognosis due to the fact that the survival average of patients is close to three years. Identified in the 1990s, it is a difficult disease to diagnose and rarely considered cured. The research to find biomarkers to improve its diagnosis is actively pursued all over the world. This tumor is characterized by the overexpression of Cyclin D1 and the binding of this protein to certain regions of DNA involved in the regulation of DDR.

The project presented here aims to analyze the similarity of the gene expression regulated by Cyclin D1 of MCL with respect to gene expression of DDR. This would allow exploring possible essential mechanisms of carcinogenesis and focus on the genes that could be interesting therapeutic targets in the process of tumor progression. In addition, new biomarkers could be used to help in an early diagnosis, often linked to a better survival rate.

The data sets published for MCL (GSE21452 [3]) and DDR (GSE25848 [4]) will be used to generate a gene signature in which the significantly enriched genes will be identified in order to study their possible role as a therapeutic target and as a biomarker.

The *R* environment will be used to align the reads, generate quality controls and finally generate the gene signature through Gene Set Enrichment Analysis (GSEA). As stated in its documentation: "Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes)."[5]

All this process will be boosted with the addition of Machine Learning methodologies.

1.2 Motivation

1.2.1 Why this project?

The amount of data collected in scientific researches has increased exponentially during the last decades, making the usage of Data Science methodologies a good fit for improving the final analysis and results. This project is a clear example of how the advances in Data Science can trigger new ways of doing science, expanding the existing tools in order to achieve better results.

1.2.2 What can I add?

During my MSc in Data Science, I have been learning about all the different aspects of a Data Science project, starting from a project management point of view and continuing with all the different phases of acquisition, storage, hypothesis and modeling, visualization and deployment. For this specific project, though, my focus will be in the area of Data Mining and Machine Learning, and that is what I think I can add to the project, my accumulated experience in the commented area.

1.2.3 Personal interest

My personal interest in this project comes from the fact, or bad luck, of having close family and friends affected for leukemia, therefore, as soon as I saw the proposal of this project I felt emotionally connected to it.

Apart from that first reason, I also consider that one of the best usages of the advances of Data Mining and Machine Learning is to help in the creation of a better society, being one of its strongest foundations the improvement of the quality of the health of each individual. Therefore, I feel responsible for using my new acquired knowledge in areas that can lead to that goal.

1.2.4 How can this project improve my CV?

At the time of writing this, I am working as a research engineer in the control software group in MAX IV Laboratory, a synchrotron located in the south of Sweden that has the purpose of improving the scientific researches in a global encompass. Until now, my main activities has been related to the build of software for all the different aspects of the control system, from the very low level control, writing drivers for equipment, up to the high level software like graphical user interfaces that enable the scientist to perform their jobs, passing through all the layers in

between, like software libraries, servers, etc. In other words, I have been always close to the control, synchronization and data acquisition, with this project, I can expand my coverage and help also in the next phase of a scientific research, the analysis of the generated data.

1.3 Project Objectives

1.3.1 General

Analysis of the similarities of the gene expression regulated by Cyclin D1 in Mantle Cell Lymphoma (MCL) and the gene expression of the DNA-damage response (DDR) using Machine Learning and Gene Set Enrichment Analysis (GSEA).

1.3.2 Specific

- Identify significantly enriched genes that can act as a therapeutic target and as a biomarker.
- Create Machine Learning models to boost the process of identification of the significantly enriched genes.

1.4 Description of the Methodology Used

This project has been carried out using a quantitative methodology. This type of methodology is based on the quantification of the results, being the main objective the generation of mathematical models, theories and hypothesis to extract information from an observable phenomena.

This quantitative methodology was applied in all the different steps in the pipeline implemented, including the final step, GSEA, where the interpretation of the final results are based on the outcome of the statistical tests applied.

Along with the mentioned methodology, a CRISP-DM methodology has been adopted. Cross-industry standard process for Data Mining, also known as CRISP-DM, was born in 1996 with the goal of provide a specific methodology suited for the needs of a Data Mining project. Although this methodology was born with a clear business orientation, it is easy to adapt to the purposes of the work presented here.

The CRISP-DM approach divide the process of Data Mining in six well differentiated phases shown in the figure 1.1.

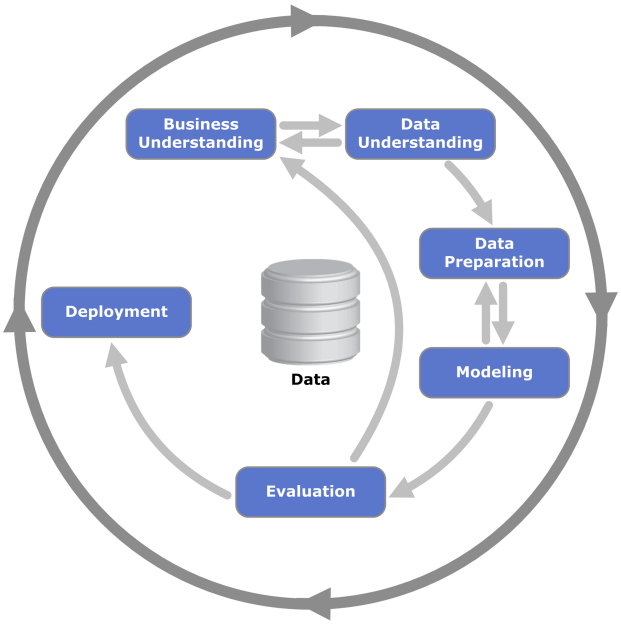

Figure 1.1: CRISP-DM Process diagram by Kenneth Jensen (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons.

In the following paragraph, a short description of the different phases in the context of this work is provided.

- **Business Understanding:** the main objective in this phase is to understand the problem that is intended to be investigated or solved, set objectives to be accomplished and

create a project plan.

- **Data Understanding:** in this phase, the data is collected and explored in order to get familiar with it, understand it or identify possible interesting subsets.
- **Data Preparation:** during this phase, the data is cleaned and transformed if needed in order to produce the final dataset that will be used in the next step.
- **Modeling:** in this phase, modeling techniques are applied in order to obtain a model that allows to give an answer to the initial objectives.
- **Evaluation:** after the generation of the model, an evaluation must be done. During this phase, the model or models obtained in the previous phase are evaluated in order to assess that the results matches the acceptance criteria. After this evaluation is done, it will be possible to analyze the results and extract knowledge from them.
- **Deployment/Publication:** once the results are evaluated and analyzed, the model can be deployed. In the case of being the project a scientific study as in the case of this project, this phase will consist in the publication of the results.

An important point to mention here is the iterative nature of the processes. It means that the order of the phases are not fixed, allowing the return to any previous phase in case of necessity. This point is important in any Data Mining or Machine Learning project, but specially in the work presented here, because one key point is the knowledge discovery, and that may require data or specific domain knowledge not contemplated at the early phases of the project.

1.5 Project Research Plan

The figure 1.2 show a Gantt Diagram with the initial planning of the project. It is divided in the following milestones needed to the accomplishment of the project:

- **Definition and planning:** this milestone consists on the definition of the project and its planning.
- **State of the art:** to accomplish this milestone, a deep study on the recent activities in the field will be done.
- **Design and implementation:** during this stage, the implementation of the study will be carried on.
- **Write the report:** once the previous step is finish, a report explaining the results must be written.
- **Thesis defense:** a presentation of the current work must be done to accomplish this milestone.
- **Public presentation:** as a last step, the work must be publicly presented to an academic trial.

An extra column has been added to the Gantt diagram to illustrate the relation between the milestones and the corresponding phase or phases of a CRISP-DM project.

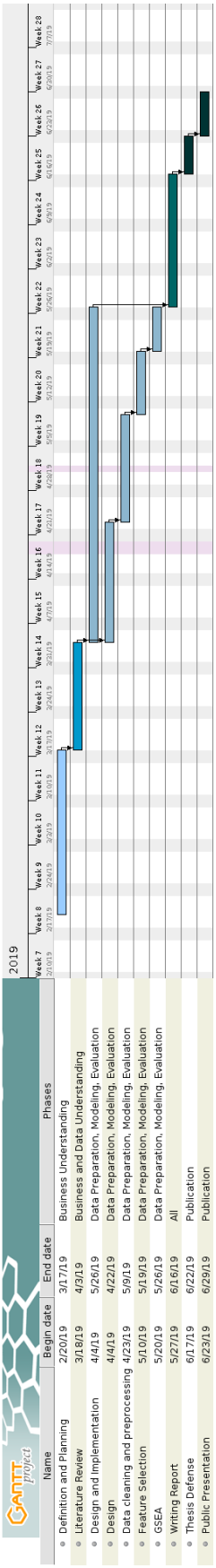

Figure 1.2: Gantt Project

Chapter 2

State of the Art

2.1 Foundations

The overexpression of Cyclin D1 in human cancer is well-known[6] and has been reported in several studies[7]. An interesting recent work conducted by Albero et al.[8], focuses on the study of this overexpression and how it produces a global trancriptional donwmodulation in lymphoid neoplasms. In their own words:

"This finding of global transcriptional dysregulation expands the known functions of oncogenic Cyclin D1 and suggests the therapeutic potential of targeting the transcriptional machinery in Cyclin D1–overexpressing tumors." [8]

Studies like the one performed by Mohanty et al.[9] show the importance of Cyclin D1 (CCND1) in the maintenance of MCL tumor cell lines, but leave unclear the protective role of this gene in preventing DNA damage during replication in MCL. This mentioned study point out some conflicts with another study on CCND1 performed by Klier et al.[10], which reports that silencing CCND1 in MCL for up to seven days cause growth arrest but not cell death in MCL.[9].

Another interesting study performed by Tiemann et al.[11] demonstrate how targeting Cyclin D1 and Cyclin D2 in chemotherapy can lead to enhance the efficacy of chemotherapy agents.

The above cited studies are a small set of examples on how important it is to increase the knowledge of the transcriptional function of Cyclin D1 in order to improve the prevention and treatment of MCL.

In parallel, other studies [12] [13], have shown the role of Cyclin D1 in the cell cycle and its influence in the DNA-damage repair process.

In addition to the already commented works, the field of Artificial Intelligence and in particular the Machine Learning discipline inside of it, has been winning attention in many fields,

being medicine one of them. Machine Learning has been widely used to study different types of cancer, and some examples of it will be provided in the following section. As a result, Machine Learning has been added to the pipeline of biological studies and, among other important consequences, it has produced a big impact improving the identification of discriminant pathways, as shown in the study done by Barla et a.[14].

Another foundation for this project is the Gene Set Enrichment Analysis method (GSEA) [15] which was presented in 2005. GSEA has had a big impact in the statistical analysis of gene sets, prove of that is its more than 10000 citations. GSEA is an analytical method that allows the researchers to focus on gene sets instead of individual genes, as it was done before. Thanks to that, it enables the detection of biological processes like metabolic pathways, transcriptional programs or stress responses. Apart from being a statistical analysis method, it also provides a software package and a database composed by more than 1000 gene sets that facilitates its usage and experimentation.

Although the intention of this text is not to provide an exhaustive explanation of how GSEA works, it is important to offer a minimum introduction to the method, to be able to understand better the way how Machine Learning can improve it.

A basic schema about how GSEA works is described in the following figure:

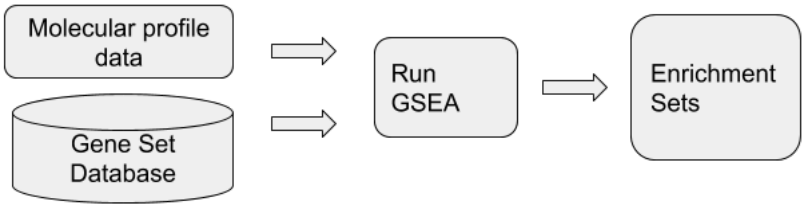

Figure 2.1: GSEA Schema.

GSEA receives two inputs, a molecular profile data and a Gene Set Database. Using this two inputs, GSEA will calculate an Enrichment Score (ES) between phenotypes for each gene contained in the molecular profile. Thanks to this ES, GSEA will be able to identify which set of genes offers statistical significance and will make possible the identification of biological processes.

Due to the big amount of data, in terms of genes, that this kind of analysis can face, it is really important to make a good selection of them beforehand. This is one of the situations where Machine Learning is able to help, providing Feature Selection algorithms that can optimize the genes selected to be passed as an input to the GSEA method.

In relation with that methodology, it is also interesting to remark the importance of choosing a proper metric for the ranking of genes, as shown in the work carried out by Zyla et al.[16].

All the above concepts have acted as a foundation to trigger the main ideas behind the

objectives of the project presented here.

2.2 Similar Work

As commented before, it is easy to find examples of the usage of Machine Learning in the field of cancer study. It has been widely used for different purposes such as classification[17] and prediction of tumors, treatment prediction, and also to boost the performance of the biological analysis pipelines using techniques like feature selection[18][19].

An interesting example is found in the study conducted by Ten et al.[20] where Machine Learning techniques were introduced in their pipeline in order to improve the analysis of multiple gene expression profiles in cervical cancer. A particular important fact extracted from that article, is that previous studies were focused either in statistical analysis methods or Machine Learning methods, but that one integrates both methodologies for the meta-analysis, which is also one of the objectives pursued by this work.

Another similar interesting work is found in the study elaborated by Park et al.[21]. In there, the identification of disease-related genes and disease mechanism is investigated using Machine Learning techniques. The study presents a novel method for gene-gene interaction (GGI) based on the usage of the Random Forest algorithm. This method is suitable for the discovery of significant GGI from heterogeneous gene expression datasets, and has the potential to be used in the research of different disease groups.

2.3 Ongoing and Future Projects

Several studies[22][23][24][25] agree on the necessity of the development of more personalized (patient-centric) treatments. Such treatments will be possible through an evaluation of each patient unique set of genomic complications and will result in more accurate treatments that will be highly effective and will not over-treat the patient. Is also important to comment that together with personalized treatments, more reliable predictive tools to improve the prognostic of each patient need to be developed, but before this point will be reached, new biomarkers and pathways that will enhance the understanding of MCL need to be discovered. This is an active area of study and it is also one of the purposes of the work presented here.

Apart from the ongoing studies on MCL, it is worth to mention that the field of Artificial Intelligence continues its expansion. Every day, new studies, methods and developments are performed. As a consequence, more fields are adopting Artificial Intelligence approaches to improve their results. Of course, medical research is also profiting from all the consequent research and development. This Artificial Intelligence explosion has the potential to change

drastically the way a scientific research will be done in the future, as a small example of it, new ways of knowledge discovery (cognitive discovery) are studied and developed, combining different areas of Artificial Intelligence like Knowledge Graphs, Natural Language Processing, Semantic Search, etc. The products obtained from that work will help the future researchers to find remarkable literature during the literature review that takes place at the starting phase of a scientific research.[26]

Chapter 3

Methods

3.1 Introduction

As described in the project objectives, this study aims to find similarities between the gene expression regulated by Cyclin D1 in MCL and the gene expression of the DDR, using Data Mining and Machine Learning techniques combined with Statistical Tests and biological analysis, to identify the significant enriched genes that can act as therapeutic target and biomarkers.

In order to fulfill these goals, this study has been developed following two different approaches that share a common part on the data cleaning and the final GSEA process, where the identification of enriched gene sets will be performed. Both approaches differ on the data normalization and posterior gene selection processes. The following sections will explain in details the two scenarios, but before going into details, it is worth to mention that two public data sets from Gene Expression Omnibus repository (GEO) were selected for this study. These data sets are:

- GSE25848 [4]: which contains data about DDR.
- GSE21452 [3]: which contains data from MCL tumors.

3.2 Scenario 1

The first developed pipeline is described in the following figure.

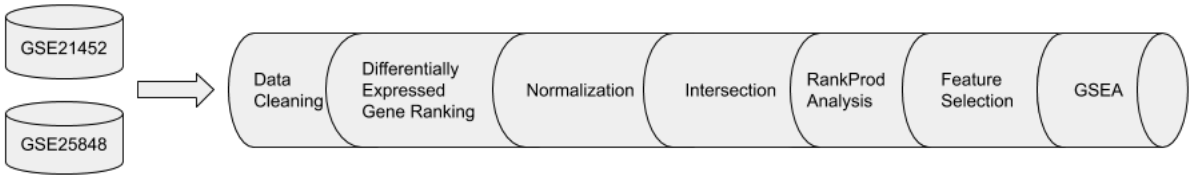

Figure 3.1: Scenario 1 pipeline

As seen in the figure, the pipeline is composed by seven steps or sub-processes:

- **Data cleaning:** the data from the two original data sets are cleaned and prepared for the posterior steps.
- **Differentially expressed gene ranking:** a first ranking of genes is generated. This ranking is done per each individual data set.
- **Normalization:** process to allow the posterior merge of both data sets.
- **Intersection:** common genes from both data sets are discovered and merged into a new data set.
- **RankProd analysis:** common genes are analyzed in order to find up and down regulated genes.
- **Feature selection:** Machine Learning algorithms are applied to find the most important features (genes).
- **GSEA:** analysis to find significantly enriched genes.

3.2.1 Data Cleaning

The first step in the pipeline consist on a general inspection of the data sets and the posterior data cleaning process. This data cleaning process consisted mainly in removing the entries containing empty or *NA* values.

3.2.2 Differentially Expressed Gene Ranking

A first selection of genes was done in this step. The purpose of this selection is to get the top ten thousand differentially expressed genes from each of both data sets. This process was done using the multiClust[27] package in R/Bioconductor.

As a result of this step, two new data sets were created. Each of these data sets were composed by ten thousand of the most differentially expressed genes of its parent data set.

3.2.3 Normalization

Once the previous steps were performed, a normalization process was applied to the new created (and reduced) data sets. In this case, a log2 transformation was applied to the data coming from the data set GSE25848. The data contained in GSE21452 was kept as it was, due to the fact that it was already log2 transformed.

Thanks to this transformation, the data coming from both data sets were in a similar scale, and able to be merged.

3.2.4 Intersection

During this process, a match between the two new reduced data sets was performed, and as a result, a new data set containing only the matched genes was created. This new data set was the one used for the posterior analysis, but before passing to the next step, another data cleaning process was performed.

In this case, some empty values (*NA*) were generated after the log2 transformation, and they needed to be treated. Two different treatments were applied. First, the Cyclin D1 gene (CCND1) was identified as one containing a few *NA* values. As this is the main gene for our study, the missing values were imputed using the mean of the *non NA* values for that gene. As a second treatment, the genes containing *NA* values were discarded.

Apart from those treatments, the name of the genes were reviewed in order to ensure that they were valid for the upcoming steps.

3.2.5 RankProd Analysis

The following step in the pipeline is the identification of up-regulated and down-regulated genes in our data set. For that purpose, a RankProd analysis was performed using the Bioconductor package RankProd.

As a result of this process, 262 up-regulated genes and 268 down-regulated genes were selected. This selection was done using a cut-off value of 0.05 on the p-values.

3.2.6 Feature Selection

Once the up-regulated and down-regulated genes were identified, a Feature Selection process was carried out. A Random Forest method was applied as an Unsupervised Feature Selection method. As a consequence, a list of features sorted by importance was generated. From that list, the top 20 features identified by the algorithm were collected. This identification was done for both, the up and down regulated genes.

3.2.7 GSEA

As a final step, two different ways of executing a GSEA were carried out. First, the Bioconductor package FGSEA[28] was used, but the outcome of this process was not satisfactory due to the lack of valid results obtained. After that first attempt, the original GSEA method[15] was performed, but again, with unsatisfactory results, where no significantly enriched genes were obtained.

This negative results forced the re-design of the experiment, and the second scenario was designed.

3.3 Scenario 2

For this second attempt, the focus was placed in the correlation between gene expressions and the Cyclin D1 expression.

The complete pipeline is represented in the next figure.

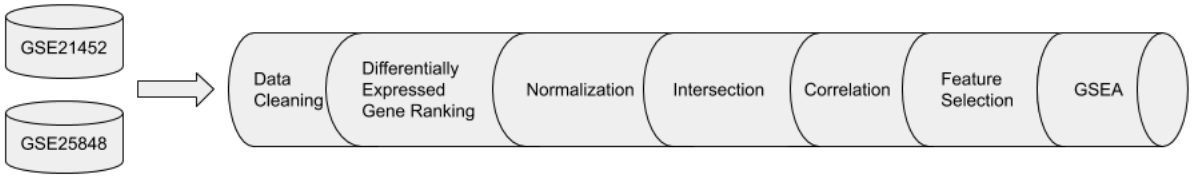

Figure 3.2: Scenario 2 pipeline

As seen in the figure, the pipeline is composed by similar sub-processes to the previous pipeline from scenario 1. The main difference in this one is the different normalization applied and the substitution of the RankProd analysis by the correlation analysis.

The first two steps from the previous scenario (data cleaning and differentially expressed gene ranking) were shared with this one, therefore, this scenario starts with an already existing ranked list of ten thousand genes per data set, which was obtained at the end of the Differentially Expressed Gene Ranking process. Due to that fact, the detailed explanation of those two sub-processes are going to be skipped in the following paragraphs.

3.3.1 Normalization

The third step in the pipeline is a normalization process. The main goal was to transform the data from both data sets to the same scale. Such transformation allowed the integration and correlation of data from both data sets.

The normalization applied in this case was a z-scored normalization.

3.3.2 Intersection

The intersection of the two data sets was carried out in a similar way than the one executed in the scenario 1. As a result, a new data set with the common genes is obtained.

The difference between this one and the one generated in the first scenario is that in this one, the data has suffered a different normalization, and therefore, the *NA* values that were generated in the log2 transformation are not present, which means that fewer data had to be removed and no data needed to be imputed.

3.3.3 Correlation with Cyclin D1

Continuing with this second pipeline, the next step was the calculation of the correlation between the expression of the available genes and the expression of Cyclin D1.

Once this correlation was calculated, a K-Means algorithm was applied. The purpose of running this method was to create three clusters and classify the data into three different types of correlation: positive correlation, no significant correlation and negative correlation.

At the end of this correlation process, two new columns were added to the data set, the first one containing the correlation value of each gene and the second one with the cluster where each gene belongs.

Because of time constrains, only the genes in the positive correlation cluster were studied in the following steps, keeping for future studies the possibility to run GSEA with the negative correlation cluster.

As a result, a 316 genes were selected for the next step.

3.3.4 Feature Selection

Similar to the Feature Selection process in the previous pipeline, a Random Forest algorithm was executed. This time, the method was executed in an unsupervised and supervised way, giving as a result a list of features (genes) ranked by importance. From the ranked list of genes obtained from the Unsupervised Random Forest, 205 genes were selected to be passed to the final GSEA process. From the ranked list of genes obtained from the Supervised Random Forest, 100 genes were selected to be passed to the final GSEA process. Both numbers were chosen to ensure that the Cyclin D1 (CCND1) was present in the selection.

3.3.5 GSEA

As a final step, a GSEA process was carried out. GSEA takes as input a molecular profile data set and a gene set database. The selection of genes obtained in the Feature Selection process was passed as first input. The gene set databases used in this process were obtained from MSigDB. Those databases are:

- **Hallman gene set (H):** coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
- **Oncogenic gene set (C6):** defined directly from microarray gene expression data from cancer gene perturbations.
- **GO gene sets (C5):** genes annotated by the same GO terms.

- **Curated gene sets (C2):** curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.
- **Immunologic gene sets (C7):** defined from microarray gene expression data from immunologic studies.

Is important to mention the phenotype argument used for the GSEA process. As this study focuses on Cyclin D1, its corresponding gene was selected to be used as phenotype, in that way, a correlation with this gene was used.

The result of running GSEA on those inputs is presented in the next chapter.

Chapter 4

Results

4.1 Introduction

This chapter summarizes the results achieved by the developed pipeline in the scenario 2, starting with a brief compilation of results from the whole pipeline and excluding the GSEA process which is kept for detailed explanation in the third section of this chapter.

4.2 Across the Pipeline

As commented in the chapter 3, the pipeline starts with the cleaning of the two selected data sets. It was specially important to clean GSE25848 as it contained 32443 out of 48803 genes without any data. Those genes were removed from the data set, resulting in a new one with 16360 genes with an expression value.

The second step was to make a first selection of genes. This selection was done individually per each data set. It consisted in a differentially expressed gene ranking where the top 10000 genes from each data set were selected.

After that process, a normalization and intersection processes took place, giving as a result a new data set composed by common genes. Such data set contained 1305 genes and 76 samples.

The next step was the computation of the correlation between each gene and CCND1. Once these values were computed, they were used to run an unsupervised clustering method, K-Means. The result of this method can be seen in the figure 4.1.

The output of that method was the classification of genes in three clusters, one for the negative correlations (cluster 3), another for the positive correlation (cluster 2) and a final one containing the non significant correlation genes (cluster 1). Only the second cluster was studied, and it was composed by 316 genes.

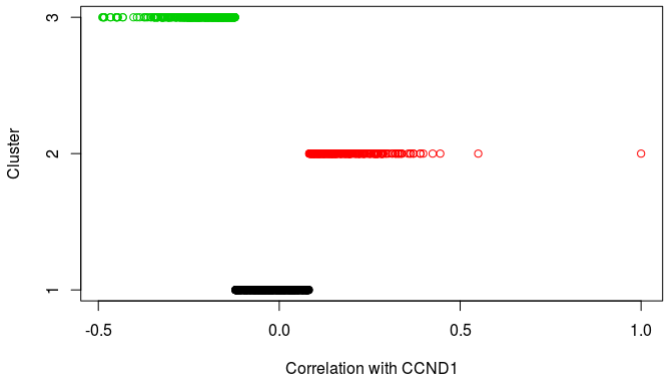

Figure 4.1: Clustering by correlation with CCND1.

As a penultimate stage in the pipeline, a Feature Selection process was carried out. A Random Forest algorithm was run in an unsupervised and supervised way, showing better results in the posterior GSEA process the supervised one. Because of that, the following explanations will only consider the supervised Random Forest. The outcome of this stage was the identification of the most significant features.

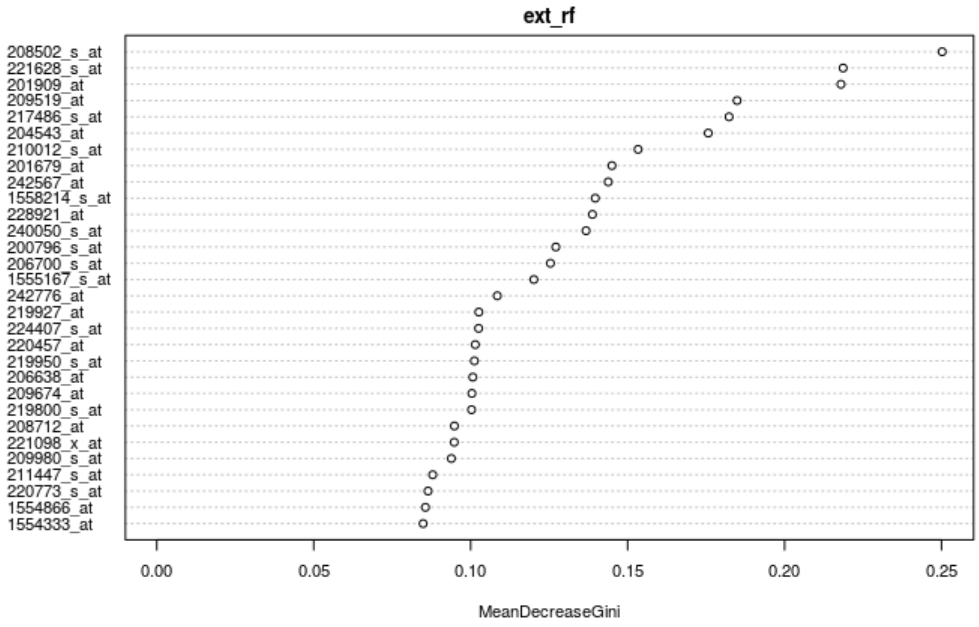

Figure 4.2: Variance importance plot obtained from Random Forest.

The model obtained had an out-of-bag of around 0.16, which was considered good enough due to the lack of interest in running predictions to classify any data. Also, as the main interest was placed in obtaining a ranked list of feature importance, no effort was spent in optimizing

the model.

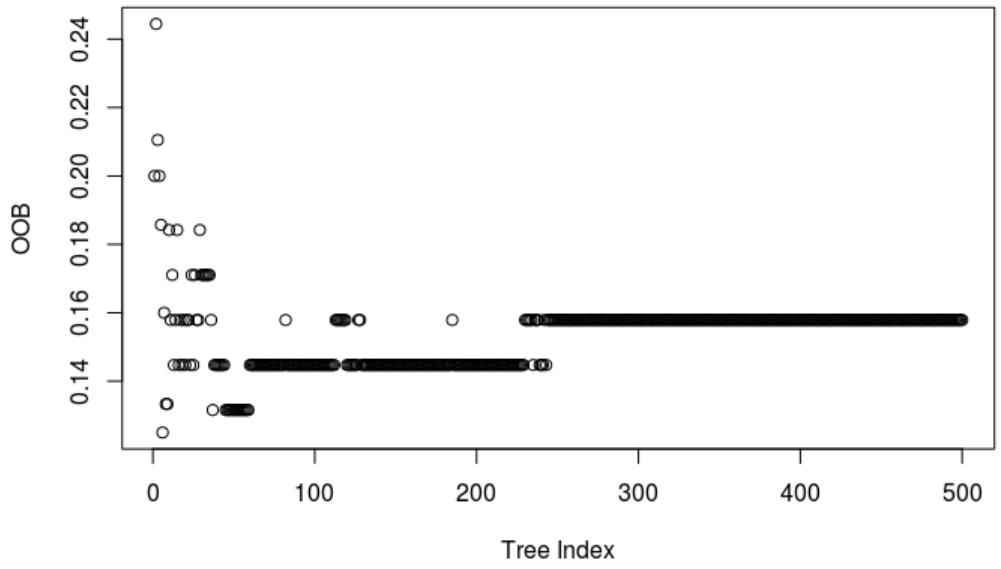

Figure 4.3: OOB from Random Forest

4.3 GSEA

The last step in the pipeline is a Gene Set Enrichment Analysis, which offers the final results of this study and deserves the dedication of an individual section in this chapter.

In order to give a better understanding of the data obtained from this study, this section starts with a short explanation of the main statistics that GSEA computes. Its main purpose is to expose the basic knowledge needed to interpret the final results.

For a deeper explanation on the statistics enumerated here and its interpretation, please refer to the GSEA documentation page [5].

4.3.1 GSEA Statistics

The following information has been obtained from the official GSEA documentation page [5].

There are four key statistics obtained from a gene set enrichment analysis:

- **Enrichment Score (ES):** the degree to which a gene set is over-represented at the top or bottom of the ranked list of genes in the expression dataset.
- **Normalized Enrichment Score (NES):** the enrichment score for a gene set after it has been normalized across analyzed gene sets. This value can be used to compare analysis results across gene sets.
- **False Discovery Rate (FDR):** the estimated probability that a normalized enrichment score represents a false positive finding.
- **Nominal P Value:** the statistical significance of the enrichment score. The nominal p value is not adjusted for gene set size or multiple hypothesis testing; therefore, it is of limited use in comparing gene sets.

Having this four statistics defined, the procedure to analyse the results is the following. First, the identified gene sets are ranked using the NES value. Then, a cut-off on FDR needs to be applied. The generalized cut-off on FDR is 25%, which indicates that the result is likely to be valid 3 out of 4 times. The gene sets that passes the FDR cut-off are the most interesting ones to generate hypothesis for further research.

Finally, the nominal p value is consulted. If a gene set has a small nominal p value and a high FDR value, it means that it is not as significant when compared with other gene sets in the empirical null distribution. The reason behind that could be that there are not enough samples, the biological signal is subtle, or the gene sets do not represent the biology in question. In case of a high nominal p value and a low FDR value, the result is considered negative, representing

that the gene set is not significant and there are other sets that are weaker. There are two cut-off defined for the nominal p value, 1% and 5%.

4.3.2 GSEA Results

The selection of genes which achieve the best results is the one obtained using a Supervised Random Forest as a Feature Selection method. These genes were used as an input of GSEA together with the MSigDB collections explained in the previous chapter: H, C6, C5, C2 and C7.

The following tables summarize the results received after running GSEA using the commented gene selection in combination with the different MSigDB collections.

The table 4.1 shows that several gene sets were identified as enriched for positive correlation with CCND1. On the other hand, as it can be seen in the table 4.2, there is only one gene set that passes the FDR cut-off for the negative correlation in the different collections. That is a normal result as there was a filtering process on positive correlated genes with CCND1 applied in early stages of the pipeline. Therefore, in the following sections, the focus will be placed in the positive correlation results showed in the table 4.1.

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Hallmark, H	13/28	7	4	5
Oncogenic, C6	45/107	6	3	5
GO, C5	919/1824	36	60	95
Curated, C2	908/1598	65	141	175
Immunologic, C7	1707/3175	0	41	106

Table 4.1: Enrichment in phenotype for positive correlations with CCND1. Each row shows the results obtained from each of the MSigDB collections.

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Hallmark, H	15/28	0	0	2
Oncogenic, C6	62/107	0	1	7
GO, C5	905/1824	0	10	41
Curated, C2	690/1598	1	37	86
Immunologic, C7	1468/3175	0	37	111

Table 4.2: Enrichment in phenotype for negative correlations with CCND1. Each row shows the results obtained from each of the MSigDB collections.

4.3.2.1 Using H collection: Hallmark gene sets

As shown in the table 4.3, using the Hallmark gene sets collection, the enrichment in phenotype for positive correlations shows that 13 from 28 gene sets are up-regulated. Seven of those passes the cut-off of FDR smaller than 25%. In addition to that, 4 gene sets have a nominal p-value less than 1%.

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Hallmark, H	13/28	7	4	5

Table 4.3: Enrichment in Phenotype for positive correlation using H.

GS	SIZE	NES	NOM p-val	FDR q-val	LEADING EDGE
HALLMARK_ESTROGEN_RESPONSE_EARLY	3	1.55	0.010	0.071	tags=33%, list=0%, signal=32%
HALLMARK_HYPOXIA	2	1.52	0.006	0.051	tags=50%, list=3%, signal=51%
HALLMARK_ESTROGEN_RESPONSE_LATE	2	1.41	0.043	0.145	tags=50%, list=0%, signal=49%
HALLMARK_APOPTOSIS	3	1.39	0.075	0.126	tags=100%, list=22%, signal=124%
HALLMARK_NOTCH_SIGNALING	1	1.33	0.000	0.186	tags=100%, list=0%, signal=99%
HALLMARK_ANDROGEN_RESPONSE	1	1.33	0.000	0.155	tags=100%, list=0%, signal=99%
HALLMARK_TNFA_SIGNALING_VIA_NFKB	4	1.32	0.132	0.137	tags=75%, list=22%, signal=92%

Table 4.4: Up-regulated gene sets for the H collection.

As seen in the table 4.4, the common genes from MCL and DDR resulted in an up-regulated identification of the following biological states or processes:

- Early and late response to estrogen.
- Hypoxia. Genes up-regulated in response of low oxygen levels.
- Apoptosis. Genes mediating programmed cell death by activation of caspases.
- Genes up-regulated by activation of Notch signaling.
- Androgen response.
- TNFA signaling response via NFKB.

In all of them, except in the Hypoxia state, CCND1 was identified as up-regulated.

4.3.2.2 Using C6 collection: Oncogenic signatures

The tables 4.5 and 4.6 show an enumeration of the results obtained from running GSEA with the C6 collection. In this case, the detected signatures of cellular pathways were CCND1 is involved are:

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Oncogenic, C6	45/107	6	3	5

Table 4.5: Enrichment in Phenotype for positive correlation using C6.

Gene Set	Size	ES	NES	NOM p-val	FDR q-val	Leading Edge
PRC2_EED_UP.V1_DN	3	0.96	1.54	0.004	0.145	tags=100%, list=6%, signal=103%
BMI1_DN.V1_UP	4	0.83	1.54	0.037	0.077	tags=50%, list=3%, signal=49%
BMI1_DN_MEL18_DN.V1_UP	4	0.75	1.43	0.079	0.212	tags=50%, list=3%, signal=49%
MEL18_DN.V1_UP	4	0.75	1.43	0.079	0.159	tags=50%, list=3%, signal=49%
RAF_UP.V1_DN	3	0.79	1.43	0.070	0.129	tags=33%, list=0%, signal=32%
IL2_UP.V1_UP	2	0.93	1.39	0.035	0.161	tags=100%, list=8%, signal=107%

Table 4.6: Up-regulated gene sets for the C6 collection.

- BMI1_DN.V1_UP. Genes up-regulated in DAOY cells (medulloblastoma) upon knockdown of BMI1.
- BMI1_DN_MEL18_DN.V1_UP. Genes up-regulated in DAOY cells (medulloblastoma) upon knockdown of BMI1 and PCGF2 genes by RNAi.
- MEL18_DN.V1_UP. Genes up-regulated in DAOY cells (medulloblastoma) upon knockdown of PCGF2 gene by RNAi.
- RAF_UP.V1_DN. Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 MCF-7 cells (breast cancer) stably over-expressing constitutively active RAF1 gene.

4.3.2.3 Using C5 collection: Gene Ontology (GO) gene sets

The execution of GSEA using the C5 collection retrieved the results summarized in the tables 4.7 and 4.8. Specially interesting is the up-regulated detection of:

- Positive regulation of catalytic activity.
- Regulation of multicellular organismal development.
- Regulation of mitotic cell cycle.
- Negative regulation of cell cycle process.
- Negative regulation of mitotic cell cycle.

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
GO, C5	919/1824	36	60	95

Table 4.7: Enrichment in Phenotype for positive correlation using C5.

Gene Set	SIZE	ES	NES	NOM p-val	FDR q-val	LEADING EDGE
GO_POSITIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS	10	0.69	1.82	0.006	0.323	tags=40%, list=9%, signal=40%
GO_MOLECULAR_FUNCTION_REGULATOR	10	0.70	1.81	0.004	0.181	tags=40%, list=6%, signal=38%
GO_POSITIVE_REGULATION_OF_PHOSPHORUS_METABOLIC_PROCESS	8	0.73	1.80	0.000	0.139	tags=50%, list=9%, signal=51%
GO_POSITIVE_REGULATION_OF_PROTEIN_MODIFICATION_PROCESS	8	0.73	1.80	0.000	0.104	tags=50%, list=9%, signal=51%
GO_ENZYME_REGULATOR_ACTIVITY	7	0.80	1.76	0.000	0.134	tags=43%, list=4%, signal=41%
GO_POSITIVE_REGULATION_OF_CATALYTIC_ACTIVITY	11	0.67	1.74	0.010	0.145	tags=36%, list=9%, signal=36%
GO_REGULATION_OF_MULTICELLULAR_ORGANISMAL_DEVELOPMENT	6	0.81	1.71	0.012	0.161	tags=50%, list=10%, signal=52%
GO_REGULATION_OF_HYDROLASE_ACTIVITY	9	0.68	1.69	0.019	0.190	tags=44%, list=9%, signal=44%
GO_POSITIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	6	0.78	1.68	0.014	0.191	tags=50%, list=10%, signal=52%
GO_POSITIVE_REGULATION_OF_MOLECULAR_FUNCTION	12	0.62	1.68	0.021	0.174	tags=33%, list=9%, signal=32%
GO_POSITIVE_REGULATION_OF_TRANSFERASE_ACTIVITY	5	0.78	1.66	0.004	0.180	tags=40%, list=6%, signal=40%
GO_KINASE_ACTIVITY	7	0.74	1.66	0.024	0.167	tags=43%, list=9%, signal=44%
GO_PROTEIN_KINASE_ACTIVITY	5	0.83	1.66	0.012	0.167	tags=60%, list=9%, signal=63%
GO_REGULATION_OF_MITOTIC_CELL_CYCLE	4	0.87	1.64	0.012	0.176	tags=25%, list=0%, signal=24%
GO_PROTEIN_PHOSPHORYLATION	7	0.71	1.64	0.015	0.170	tags=43%, list=9%, signal=44%
GO_REGULATION_OF_GTPASE_ACTIVITY	7	0.71	1.64	0.023	0.164	tags=43%, list=6%, signal=42%
GO_CELL_DIVISION	4	0.86	1.63	0.008	0.172	tags=25%, list=0%, signal=24%
GO_PHOSPHORYLATION	9	0.63	1.62	0.017	0.169	tags=33%, list=9%, signal=33%
GO_NEGATIVE_REGULATION_OF_CELL_CYCLE_PROCESS	3	0.91	1.60	0.008	0.204	tags=33%, list=0%, signal=32%
GO_NEGATIVE_REGULATION_OF_MITOTIC_CELL_CYCLE	3	0.91	1.60	0.008	0.194	tags=33%, list=0%, signal=32%

Table 4.8: First 20 up-regulated gene sets for the C5 collection.

4.3.2.4 Using C2 collection: Curated gene sets

As in the previous sections, the following tables summarize the outcome from running GSEA, this using the C2 collection.

Is interesting to point out the consistency of these results with the achieved with the previous collection as both detected as up-regulated the cell cycle mitotic gene set.

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Curated, C2	908/1598	65	141	175

Table 4.9: Enrichment in Phenotype for positive correlation using C2.

Gene Set	SIZE	ES	NES	NOM p-val	FDR q-val	LEADING EDGE
BERENJENO_TRANSFORMED_BY_RHOA_UP	6	0.85	1.87	0.000	0.037	tags=33%, list=4%, signal=33%
KRIGE_RESPONSE_TO_TOSEDOSTAT_6HR_DN	8	0.77	1.78	0.002	0.089	tags=50%, list=18%, signal=56%
KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_DN	8	0.77	1.78	0.002	0.059	tags=50%, list=18%, signal=56%
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_UP	5	0.89	1.78	0.000	0.045	tags=40%, list=5%, signal=40%
ONKEN_UVEAL_MELANOMA_UP	4	0.93	1.71	0.000	0.093	tags=75%, list=8%, signal=78%
WAMUNYOKOLI_OVARIAN_CANCER_LMP_UP	3	0.98	1.69	0.000	0.113	tags=33%, list=0%, signal=32%
BLALOCK_ALZHEIMERS_DISEASE_INCIPIENT_UP	6	0.78	1.68	0.006	0.108	tags=83%, list=24%, signal=102%
NUYTTEN_NIPP1_TARGETS_DN	5	0.82	1.68	0.004	0.103	tags=60%, list=13%, signal=66%
BLALOCK_ALZHEIMERS_DISEASE_UP	15	0.64	1.67	0.008	0.093	tags=60%, list=24%, signal=66%
MARTORIATI_MDM4_TARGETS_NEUROEPITHELIUM_UP	3	0.96	1.62	0.004	0.168	tags=67%, list=5%, signal=68%
MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	5	0.88	1.61	0.020	0.184	tags=80%, list=8%, signal=83%
KRIEG_HYPOXIA_NOT_VIA_KDM3A	4	0.83	1.61	0.004	0.171	tags=50%, list=6%, signal=51%
SWEET_LUNG_CANCER_KRAS_UP	4	0.85	1.60	0.018	0.170	tags=25%, list=0%, signal=24%
BENPORATH_SOX2_TARGETS	3	0.90	1.59	0.012	0.167	tags=33%, list=0%, signal=32%
PENG_GLUCOSE_DEPRIVATION_DN	4	0.85	1.59	0.012	0.157	tags=50%, list=10%, signal=53%
REACTOME_CELL_CYCLE	3	0.91	1.58	0.006	0.168	tags=33%, list=0%, signal=32%
REACTOME_CELL_CYCLE_MITOTIC	3	0.91	1.58	0.006	0.158	tags=33%, list=0%, signal=32%
CHESLER_BRAIN_QTL_CIS	2	1.00	1.57	0.000	0.174	tags=50%, list=0%, signal=49%
YAGI_AML_WITH_T_8_21_TRANSLOCATION	4	0.85	1.56	0.018	0.172	tags=25%, list=0%, signal=24%
PUJANA_BREAST_CANCER_LIT_INT_NETWORK	3	0.87	1.56	0.018	0.173	tags=33%, list=0%, signal=32%

Table 4.10: First 20 up-regulated gene sets for C2.

4.3.2.5 Using C7 collection: Immunologic signatures

As seen in the table 4.11, there is no enriched gene set that passes the FDR cut-off.

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Immunologic, C7	1707/3175	0	41	106

Table 4.11: Enrichment in Phenotype for positive correlation using C7.

Chapter 5

Conclusions

It is worth to start these conclusions reminding that the main objective of this study was to analyze the similarities between the gene expressions regulated by Cyclin D1 in MCL and DDR using Machine Learning. That objective has been accomplished through an integrative pipeline where classical statistical methods used in biology has been combined with Data Mining and Machine Learning techniques.

The first important observation from the data extracted from GSEA is the up-regulation of hypoxia gene sets in more than one collection, i.e. in H and C2.

Hypoxia is a condition where low levels of oxygen are supplied to a cell tissue. It is used in cancer treatment to predict the response of a tumor to a specific treatment and it is associated to the resistance of a therapy. Studies like the one conducted by Possik et al. [29] show that hypoxia sensitizes melanomas to targeted inhibition of the DDR, contributing in this way to the tumor expansion.

Another interesting observation is the up-regulation of Apoptosis. Apoptosis is a series of molecular steps that ends up in leading the cell to its death. This process is used by the body to eliminate abnormal or unnecessary cells. This process may be blocked by cancer cells.

Studies like the carried out by Greijer and van der Wall [30] show the importance of hypoxia and apoptosis resistance as a fundamental mechanism of tumor progression. A better understanding on this two conditions might lead to better treatments for MCL.

Continuing with the up-regulated gene sets found in this study, GSEA reports an up-regulation of the Notch signaling pathway. As seen in the study conducted by Li et al. [31], Notch signaling plays a critical role in the development of different forms of cancer, and due to its importance in tumorigenesis and metastasis, blocking Notch signaling pathway may be considered as a potential therapy for cancer treatment, and by extension MCL. Furthermore, recent studies like the one performed by Yuan et al. [32] show how Notch inhibitors may improve chemotherapy response, being a great promise for cancer control.

The next up-regulated gene set that worth the attention of this study is the regulation of cell cycle and specially the regulation of mitotic cell cycle. This process consists of a series of steps where chromosomes and other cell materials are duplicated for its posterior usage on the split of the cell into two daughter cells. The found influence of CCND1 in this process for MCL and DDR could be suggested as a target for further study, with the aim to determine if a possible therapy can be obtained. This idea is inline with the concluded in the study done by Bakhoum et al.: "Cancer cells coopt the mitotic DNA damage response to further propagate chromosomal instability. This offers untapped therapeutic opportunities to target genomic instability in cancer."[33]

In conclusion, the results obtained in this study suggested that targeting of Notch pathway and studying potential common mechanisms of hypoxia and apoptosis resistance would be of great interest for future studies on potential treatments of MCL. This in silico conclusion needs to be further validated by experimental studies on those processes that would shed light on the common mechanisms of DNA damage response and MCL development.

Chapter 6

Future Developments

One of the purposes of this study was to include more Feature Selection algorithms in the pipelines, but due to the lack of time this objective is left for future improvements.

The idea of integrating more than one Feature Selection algorithm was to execute several of them in parallel and combine their results. This combination can be done matching the common genes that are selected by each algorithm and perform GSEA over that new set of genes.

Another point worth to comment is the optimization of the model created by the Random Forest algorithm. In this study no optimization has been performed, as the main objective was to extract the most important features and a considerable number of features was going to be selected, but if a more accurate or reduced selection of genes is desired, this point could be considered.

For this study, only two data sets were evaluated, the addition of more data sets could be also considered.

Finally, further experimental studies would be required to validate this in silico analysis.

Bibliography

[1] Emily Clough and Tanya Barrett. The gene expression omnibus database. *Methods in molecular biology (Clifton, N.J.)*, 1418:93–110, March 2016.

[2] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. Ncbi geo: archive for functional genomics data sets–update. *Nucleic acids research*, 41:D991–D995, January 2013.

[3] Integrated genomic profiling in mantle cell lymphoma [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse21452], April 2011.

[4] Atm regulates a dna damage response post-transcriptional rna operon in lymphocytes [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse25848], January 2011.

[5] Gsea user guide [http://software.broadinstitute.org/gsea/doc/gseauserguideframe.html], December 2012.

[6] Justin Lamb, Sridhar Ramaswamy, Heide L Ford, Bernardo Contreras, Robert V Martinez, Frances S Kittrell, Cynthia A Zahnow, Nick Patterson, Todd R Golub, and Mark E Ewen. A mechanism of cyclin d1 action encoded in the patterns of gene expression in human cancer. *Cell*, 114:323–334, August 2003.

[7] Reena. John, N. Malathi, C. Ravindran, and Soumya. Anandan. Mini review: Multifaceted role played by cyclin D1 in tumor behavior. *Indian Journal of Dental Research*, 28(2):187–192, June 2017.

[8] Robert Albero, Anna Enjuanes, Santiago Demajo, Giancarlo Castellano, Magda Pinyol, Noelia García, Cristina Capdevila, Guillem Clot, Helena Suárez-Cisneros, Mariko Shimada, Kennosuke Karube, Mónica López-Guerra, Dolors Colomer, Sílvia Beà, José Ignacio Martin-Subero, Elías Campo, and Pedro Jares. Cyclin d1 overexpression induces

global transcriptional downregulation in lymphoid neoplasms. *The Journal of Clinical Investigation*, 128(9):4132–4147, August 2018.

[9] Suchismita Mohanty, Atish Mohanty, Natalie Sandoval, Thai Tran, Victoria Bedell, Jun Wu, Anna Scuto, Joyce Murata-Collins, Dennis D Weisenburger, and Vu N Ngo. Cyclin d1 depletion induces dna damage in mantle cell lymphoma lines. *Leukemia & lymphoma*, 58:676–688, March 2017.

[10] M Klier, N Anastasov, A Hermann, T Meindl, D Angermeier, M Raffeld, F Fend, and L Quintanilla-Martinez. Specific lentiviral shrna-mediated knockdown of cyclin d1 in mantle cell lymphoma has minimal effects on cell survival and reveals a regulatory circuit with cyclin d2. *Leukemia*, 22:2097–2105, November 2008.

[11] Katrin Tiemann, Jessica V Alluin, Anja Honegger, Pritsana Chomchan, Shikha Gaur, Yen Yun, Stephen J Forman, John J Rossi, and Robert W Chen. Small interfering rnas targeting cyclin d1 and cyclin d2 enhance the cytotoxicity of chemotherapeutic agents in mantle cell lymphoma cell lines. *Leukemia & lymphoma*, 52:2148–2154, November 2011.

[12] Gabriele Di Sante, Agnese Di Rocco, Claudia Pupo, Mathew C Casimiro, and Richard G Pestell. Hormone-induced dna damage response and repair mediated by cyclin d1 in breast and prostate cancer. *Oncotarget*, 8:81803–81812, October 2017.

[13] Mathew C Casimiro, Gabriele Di Sante, Xiaoming Ju, Zhiping Li, Ke Chen, Marco Crosariol, Ismail Yaman, Michael Gormley, Hui Meng, Michael P Lisanti, and Richard G Pestell. Cyclin d1 promotes androgen-dependent dna damage repair in prostate cancer cells. *Cancer research*, 76:329–338, January 2016.

[14] Annalisa Barla, Giuseppe Jurman, Roberto Visintainer, Margherita Squillario, Michele Filosi, Samantha Riccadonna, and Cesare Furlanello. A machine learning pipeline for identification of discriminant pathways. *Springer Handbook of Bio-/Neuroinformatics*, January 2014.

[15] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, September 2005.

[16] Joanna Zyla, Michal Marczyk, January Weiner, and Joanna Polanska. Ranking metrics in gene set enrichment analysis: do they matter? *BMC Bioinformatics*, 18(1):1, May 2017.

[17] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3:140, January 2007.

[18] Rabindra Kumar Singh and M. Sivabalakrishnan. Feature selection of gene expression data for cancer classification: A review. *Procedia Computer Science*, 50:52 – 57, April 2015.

[19] Azadeh Bashiri, Marjan Ghazisaeedi, Reza Safdari, Leila Shahmoradi, and Hamide Ehtesham. Improving the prediction of survival in cancer patients by using machine learning techniques: Experience of gene expression data: A narrative review. *Iranian journal of public health*, 46:165 – 172, February 2017.

[20] Mei Sze Tan, Siow-Wee Chang, Phaik Leng Cheah, and Hwa Jen Yap. Integrative machine learning analysis of multiple gene expression profiles in cervical cancer. *PeerJ*, 6:e5285, July 2018.

[21] Chihyun Park, JungRim Kim, Jeongwoo Kim, and Sanghyun Park. Machine learning-based identification of genetic interactions from heterogeneous gene expression profiles. *PloS one*, 13:e0201056, July 2018.

[22] Arati A Inamdar, Andre Goy, Nehad M Ayoub, Christen Attia, Lucia Oton, Varun Taruvai, Mark Costales, Yu-Ting Lin, Andrew Pecora, and K Stephen Suh. Mantle cell lymphoma in the era of precision medicine-diagnosis, biomarkers and therapeutic agents. *Oncotarget*, 7:48692–48731, July 2016.

[23] Raphael E Steiner, Jorge Romaguera, and Michael Wang. Current trials for frontline therapy of mantle cell lymphoma. *Journal of hematology & oncology*, 11:13, January 2018.

[24] Michael Schieber, Leo I Gordon, and Reem Karmali. Current overview and treatment of mantle cell lymphoma. *F1000Research*, 7, July 2018.

[25] Martin Dreyling, Simone Ferrero, and European Mantle Cell Lymphoma Network. The role of targeted treatment in mantle cell lymphoma: is transplant dead or alive? *Haematologica*, 101:104–114, February 2016.

[26] D. Raymond. Using artificial intelligence to combat information overload in research. *IEEE Pulse*, 10(1):18–21, January 2019.

[27] Nathan Lawlor, Peiyong Guan, Alec Fabbri, Krish Karuturi, and Joshy George. *multiClust: multiClust: An R-package for Identifying Biologically Relevant Clusters in Cancer Transcriptome Profiles*, 2019. R package version 1.14.0.

[28] Alexey Sergushichev. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*, 2016.

[29] Patricia A Possik, Judith Müller, Carmen Gerlach, Juliana C N Kenski, Xinyao Huang, Aida Shahrabi, Oscar Krijgsman, Ji-Ying Song, Marjon A Smit, Bram Gerritsen, Cor Lieftink, Kristel Kemper, Magali Michaut, Roderick L Beijersbergen, Lodewyk Wessels, Ton N Schumacher, and Daniel S Peeper. Parallel in vivo and in vitro melanoma rnai dropout screens reveal synthetic lethality between hypoxia and dna damage response inhibition. *Cell reports*, 9:1375–1386, November 2014.

[30] A E Greijer and E van der Wall. The role of hypoxia inducible factor 1 (hif-1) in hypoxia induced apoptosis. *Journal of clinical pathology*, 57:1009–1014, October 2004.

[31] Li Li, Ping Tang, Shun Li, Xiang Qin, Hong Yang, Chunhui Wu, and Yiyao Liu. Notch signaling pathway networks in cancer metastasis: a new target for cancer therapy. *Medical oncology (Northwood, London, England)*, 34:180, September 2017.

[32] Xun Yuan, Hua Wu, Hanxiao Xu, Huihua Xiong, Qian Chu, Shiying Yu, Gen Sheng Wu, and Kongming Wu. Notch signaling: an emerging therapeutic target for cancer treatment. *Cancer letters*, 369:20–27, December 2015.

[33] Samuel F Bakhoum, Lilian Kabeche, Duane A Compton, Simon N Powell, and Holger Bastians. Mitotic dna damage response: At the crossroads of structural and numerical cancer chromosome instabilities. *Trends in cancer*, 3:225–234, March 2017.

Appendix A

Repositories

The code developed during the execution of this study is hosted in the following repository:

- https://github.com/amilan/Thesis-DS-dev

In addition to that, the LaTeX project for this document is available at:

- https://github.com/amilan/Thesis-DS