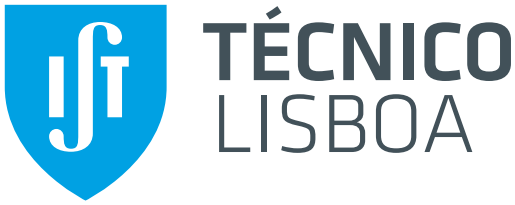


Network-based Regularisation for Survival Analysis

Pedro Jorge Estrela Martinho

Thesis to obtain the Master of Science Degree in

Electrical and Computers Engineering

Supervisor(s): Prof. Susana de Almeida Mendes Vinga Martins
Prof. Alexandra Sofia Martins de Carvalho

Examination Committee

Chairperson: Prof. António Manuel Raminhos Cordeiro Grilo
Supervisor: Prof. Susana de Almeida Mendes Vinga Martins
Member of the Committee: Prof. Nuno Luís Barbosa Morais

November 2018

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Dedicated to my family and girlfriend

Acknowledgments

Thank you to...

... my mom and dad for all the financial support during all my studies and extracurricular activities that made me the man I am today; for every minute discussing my inside struggles, every hour waiting for my return from working late in so many projects or study sessions, every day of constant support and months of patience and dedication.

... my brother for all his patient on my moments of despair and strength when I most need it; his constant presence when writing and developing this work had a substantial impact on my attitude facing this challenge. The motivational speeches made the difference.

... my girlfriend that patiently supported my decisions through this intense and meaningful journey. My lovely lady was present on all the up and down moments despite all the challenges she faced on this period. Moreover, she inspired me to consider this theme for this project given her expertise on the field, more specifically, radiotherapy. A special thank you to her, my brother and my mother for the text revision.

... my research advisers, Prof. Susana Vinga and Prof. Alexandra Carvalho for all the thoughts, materials and support on this project.

... all the Systems Biomedicine members that participated in the weekly meetings giving essential insights as the investigation was evolving. Particular gratitude to André Veríssimo, Marta Lopes and Prof. Eunice Carrasquinha that are experts on the area and always guide me when needed.

... all my friends that walked the same path as me during these long five years. Friends that helped me when I needed them the most and are part of how I am today.

... the staff in Instituto Superior Técnico, from professors to laboratory assistants, for all the contributions during my university journey.

All these people had an active impact on this chapter of my life. A chapter that certainly was not easy, but gave me many happy moments and clear definition of the path I want to take from now on. A journey concluded with this project that hopefully will help other investigations on this field of study.

Resumo

Um dos maiores desafios do século XXI é a prevenção, diagnóstico e tratamento de doenças oncológicas. Para estudar os principais fatores de risco é comum recorrer-se a dados de sobrevivência dos pacientes. Estes conjuntos de dados estão frequentemente associados à expressão genética do individuo, sofrendo a maldição da dimensionalidade. Métodos como o LASSO e Elastic Net têm-se mostrado eficientes para lidar com problemas com as mesmas características. No entanto, resultam regularmente em modelos complexos que podem ser biologicamente pouco relevantes.

Como solução, neste trabalho, é apresentada uma metodologia que melhor restringe o espaço de solução, favorecendo os genes mais relevantes tendo em conta datasets públicos. É considerada uma rede de relações entre proteínas para explorar um novo método de regularização, com base em medidas de centralidade, nomeadamente o grau e a intermediação. Com a restrição apresentada, são obtidas soluções que, no geral, consideram genes que são biologicamente mais interessantes, tendo uma forte presença em diversas investigações oncológicas.

Os resultados obtidos indicam que a metodologia proposta resulta de facto em modelos mais simples e com melhores resultados. Além disso, permite obter genes que não estão ainda associados ao tipo de cancro em estudo, mas manifestam-se como potenciais candidatos a ter em conta. A aplicação desta metodologia em diversos datasets com as mesmas características em conjunto com uma maior validação científica, poderá levar à determinação de novos genes significativos no estudo da expressão de diversos tipos de cancro. Além disso, resulta na construção de modelos simples e mais robustos.

Palavras-chave: Regressão Cox, Regularização, Redes, Expressão de genes, Proteínas.

Abstract

One of the principal challenges of the 21st century is the prevention, diagnosis and treatment of oncological diseases. To study the dominant risk factors, it is common to rely on patient survival data. These data sets are often associated with the genetic expression of the individual, suffering the curse of dimensionality. Methods such as LASSO and Elastic Net have proven to be efficient in dealing with problems with the same characteristics. However, these sometimes result in relatively complex models that might not be biologically significant.

As a solution, this thesis presents a methodology that best restricts the solution space, favouring the most relevant genes taking into account public datasets, from the The Cancer Genome Atlas (TCGA). It is considered a network of relations between proteins to explore a new method of regularisation, based on measures of centrality, namely degree and betweenness. With the restriction presented, solutions are obtained which, in general, consider genes that are biologically more interesting, having a strong presence in several oncological investigations.

The results indicate that the proposed methodology results in simpler models with better results. Besides, it allows obtaining genes that are not yet associated with the type of cancer under study but manifest themselves as potential biomarker candidates to take into account. The application of this methodology in several datasets with the same characteristics together with a greater scientific validation could lead to the determination of new significant genes in the study of the expression of several types of cancer. Furthermore, it leads to the construction of simple and more robust models.

Keywords: Cox Regression, Regularisation, Networks, Gene Expression, Proteins.

Contents

Acknowledgments . . . vii
Resumo . . . ix
Abstract . . . xi
List of Tables . . . xv
List of Figures . . . xvii

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Topic Overview . . . 2
1.3 Objectives . . . 3
1.4 Thesis Outline . . . 4

2 Background **5**
2.1 Survival Analysis and Cox Regression . . . 6
2.2 Networks Properties . . . 10
2.2.1 Degree Centrality . . . 12
2.2.2 Betweenness Centrality . . . 13
2.2.3 Closeness Centrality . . . 14
2.3 Regularization Methods . . . 14
2.3.1 LASSO, Ridge and Elastic Net Regressions . . . 15
2.3.2 Net-Cox Regression . . . 17
2.3.3 DegreeCox Regression . . . 18
2.4 STRING Dataset, TCGA and BioMart . . . 18

3 Proposed Methodology **23**
3.1 Computational Model . . . 23
3.2 Centrality Metrics Complexity . . . 25
3.3 Protein-Gene Mapping . . . 27
3.4 Penalty Factor . . . 27
3.5 Regression Coefficients Computation . . . 28

4 Network Properties of STRING **29**
4.1 Centrality Measures . . . 29
4.1.1 Weighted Network . . . 31
4.1.2 Unweighted Network . . . 32
4.2 Mapping and Penalty Factor . . . 39

5 Results **45**
5.1 Breast Cancer Dataset . . . 45
5.2 Validation Metrics . . . 46
5.2.1 Analysis of the Models Performance . . . 46
5.2.2 Analysis of the Selected Genes Significance . . . 47
5.3 Selected Regression Coefficients Analysis . . . 47
5.4 Results Discussion . . . 54

6 Conclusions **57**
6.1 Achievements . . . 57
6.2 Future Work . . . 58

Bibliography **59**

List of Tables

2.1 Survival analysis examples. 6
2.2 Lung cancer survival dataset sample. 8
2.3 Protein-protein interaction features. 21

4.1 Top 15 proteins regarding weighted and unweighted degree. 34

5.1 Model types according to the penalisation associated. 46
5.2 Models number of selected genes, p-value and c-index with different values of train/test split, α an μ. 49
5.3 Genes selected by all the top models and their documented function on *Homo sapiens*. . . 55

List of Figures

1.1 Need of Bioinformatics given the Growth of Biological Data. 2

2.1 Left-censored, right-censored and failure time illustration. 7
2.2 Kaplan-Meier curves for lung cancer survival sample. 9
2.3 Königsberd's seven bridges problem illustration. 11
2.4 Google and Apple patents network. 12
2.5 Degree, betweenness and closeness centrality measures. 14
2.6 Geometric interpretation of LASSO, Ridge and Elastic-net regressions. 16
2.7 Net-Cox and DegreeCox network regularizers. 18
2.8 Protein synthesis diagram. 19
2.9 The 20 most frequently mutated human cancer genes' proteins. 21

3.1 Proposed methodology over the STRING network to reach the final regression coefficients. 25

4.1 STRING protein-protein network focus on *Homo sapiens*. 30
4.2 Weighted degree distribution in $\log_{10}$ scale with $\theta < 150$. 31
4.3 CDF plot regarding weighted network degree with $\theta < 150$. 32
4.4 Unweighted degree distribution with $\log_{10}$ scale with different θ values. 33
4.5 CDF plots regarding unweighted network degree, having different θ value. 33
4.6 Unweighted degree vs weighted degree. 35
4.7 Betweenness distribution in $\log_{10}$ scale. 35
4.8 Closeness distribution in $\log_{10}$ scale. 36
4.9 Venn diagram on the top 250 proteins regarding the degree, betweenness and closeness metrics. 36
4.10 Re-scaled betweenness vs re-scaled unweighted degree. 37
4.11 DBet and $\text{DBet}_{\log}$ distance distribution. 38
4.12 Logarithm of re-scaled betweenness vs logarithm of re-scaled unweighted degree. 38
4.13 Venn Diagram on the top 250 proteins regarding the degree, betweenness, DBet and $\text{DBet}_{\log}$ metrics. 39
4.14 Centrality measures distribution regarding genes. 40
4.15 Heatmap over the 30 top genes selected considering the degree, betweenness and $\text{DBet}_{\log}$ metrics. 41

4.16 Penalty factor considering degree centrality and different μ values. 42
4.17 Penalty factor considering betweenness centrality and different μ values. 43
4.18 Penalty factor considering $\mathrm{DBet}_{\log}$ distance and different μ values. 43
4.19 Resume of the applied methodology to get to the final metrics. 44

5.1 Distribution of events over time considering survival and censored times. 45
5.2 Boxplots with whiskers with maximum 1.5 interquartile range focusing on the number of genes selected, c-index, p-value, and percentage of genes with no hallmarks. 50
5.3 Kapain-Meier curves considering the best models according to the different penalty factor vectors. 52
5.4 Venn diagram considering the non zero coefficients of the best selected models. 52
5.5 Heat-map on the top 30 regression coefficients for the best selected models. 53
5.6 Hallmark heat-map considering the intersected regression coefficients. 54

Chapter 1

Introduction

1.1 Motivation

Over the years the scientific knowledge has been reached through a well defined scientific process. To bring new theories and discoveries to the scientific community, one needs to generate many hypotheses that will go over tests and then are rejected, accepted or readjusted. This process is still an effective method. However, without the invention of the computers, a bottleneck would be reached due to the complexity of the new hypotheses yet to be presented. In every field of knowledge, there is a strong relationship between the amount of data and the complexity of the phenomena under study. A scientist might spend a lifetime creating and testing some few hypotheses, while a computer, with some lines of code, could test some thousands of hypotheses in less than a second [1]. Machine learning is the origin of this change allowing a scientist to create and test complex models (hypotheses) out of the given data, with the usage of a computer.

Not only did computers allow human beings to test hypotheses much faster, but also to store and work more data. Thanks to all the technological advances, much more data is possible to gather and collect, currently known as big data. The access to a large amount of information is leading to higher interest from many powerful entities to use it properly. A wide range of industries and companies are focused on finding new ways to predict future events based on collected data. The process of considering, testing and implementing this kind of approaches is mandatory for companies to have success among their competitors. Moreover, there are industries, such as energy and the health-care that, with the use of machine learning models, can increase their profits by reducing waste and provide better services to patients by improving their treatment processes.

The number of biological databases is increasing exponentially as presented in Figure 1.1. There is a vast quantity of useful information that needs to be handled and structured, enhancing the role of bioinformatics. The usage of this information leads to significant insights to improve individuals' treatment, perform better diagnostics or detect unusual body reactions. However, dealing with such datasets presents significant computational challenges. An example of these types of datasets are the ones collected by The Cancer Genome Atlas (TCGA). It is a project that "aims to catalogue and discover

major cancer-causing genomic alterations to create a comprehensive 'atlas' of cancer genomic" [2].

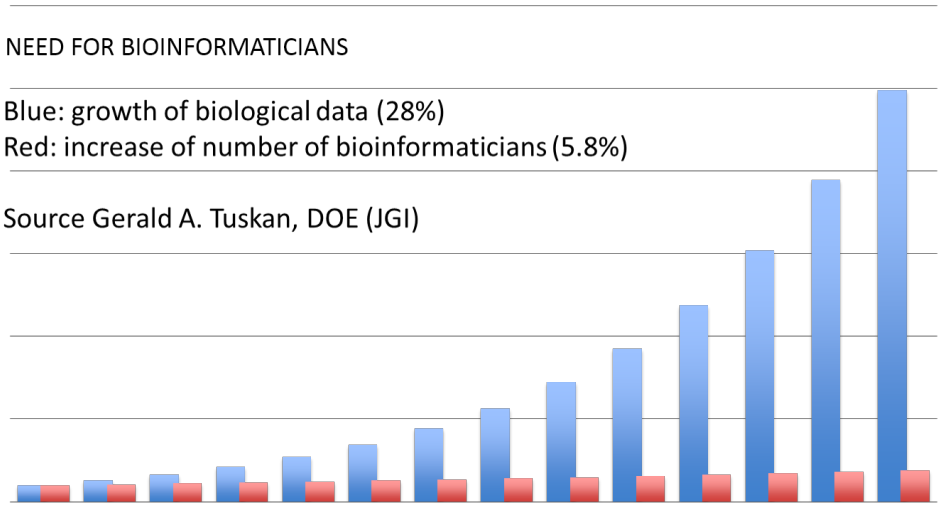


Figure 1.1: Need of Bioinformatics given the Growth of Biological Data. Histogram to understand the comparison between rates of biological data growth and the number of bioinformatics over time (source [3]).

With the development of machine learning techniques to deal with the most well known computational challenges, the models' construction by itself can bring knowledge to the scientific community, leading to major improvements for the individuals' health and diagnostic. Moreover, the increasing interest in data science, all over the world, is inspiring the development of reliable models that, shortly, might combine into one single algorithm that will learn nearly anything out of well-structured data: The Master Algorithm [1].

There are many clinical trials based on a time-to-event endpoint, frequently considering survival analysis. "Survival analysis comprises of the methods used to estimate the rates associated with time-to-an-event data, compare the rates between groups, and assess how other factors impact these rates" [4]. A particular case of survival analysis is the cancer patients in advanced stages, being the event of interest, frequently, death. Some interesting work has been developed using machine learning techniques on survival data in order to better estimate the risk of a given individual. These types of studies are having a significant impact on clinical trials, allowing better diagnosis and understanding of the main factors associated with the increase of the individual's risk.

1.2 Topic Overview

As stated, the machine learning approach is here to stay and is the bioinformatics' challenge to take the best knowledge they can out of the provided data. In fact, these words are not the most accurate. To be precise, the challenge is first to work and prepare the data and then find the best combination between model and parameters, to reach a solution. At that point, the computer will do all the work and learn the best model out of the given data. It is up to us to explore new hypotheses (different data pre-processes, models and parameters), take conclusions and, if possible, readjust based on the results and intuition.

There are different scopes within machine learning, being one of them the classification of new observations through regression models. Many distinct types of datasets can be used in order to achieve classification models. In this project, it will be considered survival data, more specifically, concerning cancer patients where the event of interest is death. Typically, when dealing with cancer datasets, gene expression data needs to be considered, being inserted into the big data category.

Even though many genes have already been studied and documented, modelling their behaviour concerning gene expression to predict the individual risk and cancer development is still difficult to accomplish. With machine learning algorithms, reasonably good models have been obtained. Nevertheless, there are some problems due to the dataset dimension: the number of features significantly outcomes the number of individuals in the study.

Multiple features to consider often leads to overfitted models and for that cases feature selection techniques have been developed [5, 6]. Even though the results are promising, the over-fit problem might persist. To undertake this problem, Zhang et al. and Veríssimo et al. proposed to further constraint the solution space, based on rich networks that model relations between genes [7, 8].

Taking profit from previous works developed by many other specialists, this thesis project purposes a new procedure to measure the gene importance based on a protein-protein interaction network. Using this information to promote a confined solution space is an encouragement to access a more generalised model with genes selected that have great biological relevance and, ultimately, can be associated with cancer investigations.

1.3 Objectives

Many regularisation methods consider only the dataset input and can construct strong models. Nevertheless, it has been proved that using a-priory knowledge can lead to better results and that is what is expected from the presented project. It is proposed the usage of a protein-protein interaction network from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [9] to get a more generalised and simpler model. It is necessary to define the best metrics to extract meaningful knowledge out of the network. For that reason, the exploration and comparison between the main centrality metrics over this network to define the protein's importance is one of this project ambitions.

Moreover, after defining the protein relevance, the relation between the genes responsible for their creation and presence in cancer studies will be analysed. The usage of this information as a penalisation over the solution space will hopefully result in a simpler model with less but more relevant variables selected with results that are, at least, as good as the ones obtained with the known techniques. This study also hopes to validate the usage of STRING datasets for this type of approaches, motivating others to use it on future works.

1.4 Thesis Outline

This thesis is divided into the following chapters: Introduction, Background, Proposed Methodology, Network Properties of STRING, Results and Conclusions.

The Introduction makes a brief overview of the thesis' theme, focusing on the motivations for the development of this work as well as the main objectives of the thesis project.

In the Background, the state-of-the-art is presented as well as the main concepts that need to be understood to follow the presented method and material. Moreover, the sources of the protein-protein network and the protein-gene mapping are also presented, along with the explanation of the relation between genes and proteins.

The Proposed Methodology covers the steps taken to reach the used models as well as the pre-process applied over the considered data. Furthermore, a closer focus is given to the betweenness metric because it has been proved to be an interesting metric regarding biological networks and is for the first time used as a penalty factor on this type of problem.

In the next chapter, Network Properties of STRING, the considered protein-protein network is analysed, focusing on different characteristics of the network and centrality metrics distribution. The penalty factor distribution is also presented in this chapter as it depends on the considered centrality measures.

The chapter Results presents the performance of the models training and testing procedures over a survival dataset focused on breast cancer from TCGA. All the steps and decisions to get to reach the final models are explained in detail. The respective results' analysis is also presented in this chapter.

Next, the Conclusions chapter focus on the achievements and the discussion of possible future work.

Chapter 2

Background

There are many computational challenges when dealing with clinical datasets due to a large number of variables considered and the complex behaviour of the living organisms. A particular case regarding clinical analysis is the survival data (life-tables) analysis, meaning the study of the time between entering a study (or other baseline condition) and experiencing a subsequent event of interest [10]. This type of datasets is possible to analyse and shape through the Cox Proportional Hazard Model (Cox PH Model) [11].

More specifically, the analysed dataset in this project concerns cancer patients having their gene expression information, which is attached to the dimensionality problem: the number of variables is much higher than the number of observations, sometimes, with two orders of magnitude difference. In this type of scenarios, getting generalised models is extremely difficult due to the large number of variables to be considered in the final model and few observations to sustain the model's hypothesis. Moreover, finding the balance between the number of features selected and models performance can be challenging. The higher the number of variables selected, the higher the probability of having an overfitted model.

There have been many different techniques applied to work on this problem, some with promising results. Regularisation methods like Least Absolute Shrinkage and Selection Operator (LASSO) and their derivatives can reduce the number of features to consider, by penalising a least squares regression by the sum of the absolute values (L_1-norm) of the coefficients [5, 12, 13]. The LASSO method is used for both variable selection and regularisation, including, for Cox regression [14]. Although the LASSO and their derivatives are powerful regularisation methods, the usage of the sum of the squared error of the coefficients (L_2-norm) to penalise a least squares regression also leads to positive results. Based on both approaches, Zou and Hastie have presented a more generalised method, in 2005, that considered both L_1 and L_2 norms. With that method one can get sparse solutions while having a grouping effect:"the most correlated predictor tend to be in or out of the model together" [6].

Recently, the idea of using networks to analyse biological behaviour has been proposed, and exciting results and insights have been achieved [15–18]. As stated by Zhang et al., "protein-protein interaction network or co-expression can provide useful prior knowledge to remove statistical randomness and confounding factors from high-dimensional data for several classification and regression models. The major

advantage of these network-based models is the better generalisation across independent studies since the network information is consistent with the conserved patterns in the gene expression data." [7]. Based on these words, the method Net-Cox has been presented and achieved bright results on ovarian cancer survival data, considering gene expression features [7]. Another interesting approach was introduced by Veríssimo et al., which proposed to use the degree centrality from gene co-expression networks and gene functional maps to constraint the model [8].

With those concepts in mind, it will be shown how to use the protein-protein interaction network and select a centrality measure to penalise the solution space further. The protein-protein interaction data used was extracted from the STRING database, that "includes direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases" [9].

All the major concepts are presented in detail in the following chapters along with the state-of-the-art.

2.1 Survival Analysis and Cox Regression

An interesting problem is to predict, based on the individual's gene expression on a specific tissue, whether he is inserted in the high-risk group or low-risk group. In order to do this, survival data is normally considered. These type of datasets are meant to study the period between the time an individual joins the study and the time the event of interest is observed. Some examples over which survival analysis can be performed on are given in Table 2.1.

Example	*Features*	*Event of interest*	*Survival time*
Disease-free cohort	weight, height, etc	heart disease	years
Leukemia patients	gene expression	remission	weeks
Melanoma patients	gene expression	death	months/years

Table 2.1: Survival analysis examples.

As we can see from the given examples, survival analysis is frequently used over medical data with the following primary goals [19]:

- Estimate and interpret survivor and hazard functions;
- Compare survivor and hazard functions;
- Assess the relationship of explanatory variables to survival time.

Typically this type of datasets are composed by the calculated features and the survival time or time-to-event. An important point to consider in the survival analysis is the need to deal with cases that the individual survival time is unknown. This type of observations is called censored data. To deal with these, the datasets also include a variable that specifies if it is an event occurrence or censored data.

The most common type of censored data is the right-censoring, meaning that the event time is earlier than the one of the event occurrence. It may occur because "the person observes the event after the

study ends, is lost to follow-up or withdraws from the study because of death or some other reason." [19]. Another type of censored data, not so frequent, is the left-censoring that happens when an individual has experienced the event of interest, but it was only registered after it happened, not being possible to know the exact event time. In Figure 2.1, all the failure times are equal to ten days. However, scenario A would be classified as left-censoring, scenario B as right-censoring (possible withdraw) and scenario C as no censoring. To deal with this datasets, removing the censored observations may seem a reasonable solution. Nonetheless, this is not accurate because censored observations greatly contribute to the total number at risk up to the time that they ceased to be followed.

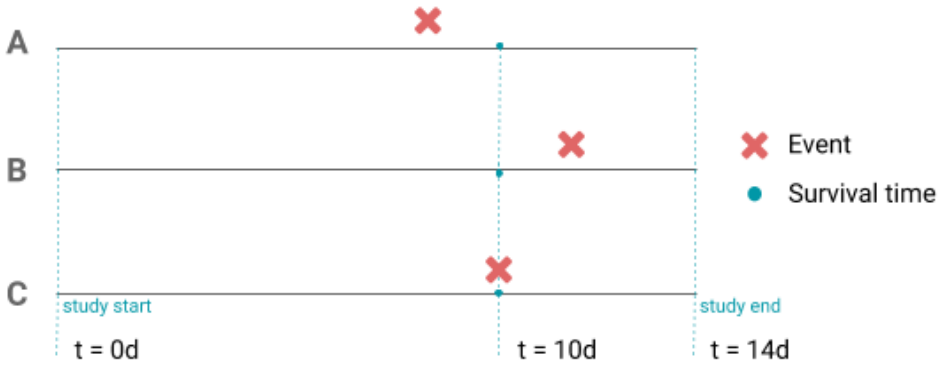


Figure 2.1: Left-censored, right-censored and failure time illustration (adapted from [19]).

In order to properly analyse these incomplete observations, Kaplan and Meier collaborated and came up with Kaplan-Meier curves [20], a powerful tool to deal with differing survival times. The individuals need to be ordered from the lower to the higher survival time value to then obtain the cumulative probability $(y - axis)$ of surviving over time. The lengths of the horizontal lines along the $x - axis$ correspond to the time difference between two consecutive individuals from the same group experiencing an event. For this reason, the Kaplan-Meier curves have a non-continuous nature, meaning that calculating a point of survival can be difficult. The censored observations are going to have a significant impact on these curves because they count as surviving individuals until their censored time. This way, there's a higher cumulative survival value until their censor time leading to a more significant drop in the cumulative value between two successive survival times (event occurrence) [21].

As an example let's consider the built-in lung cancer dataset [22] that ships with the package "survival" from R. The original dataset comprises 10 different variables, however, for this example, only the variables time, status (1 if censored and 2 if the event was verified), age and sex were considered. A total of 228 individuals were considered in the study, yet, only 20 individuals were randomly selected in order to understand the discribed concepts. The selected individuals are displayed in Table 2.2, having a relatively similar distribution concerning sex (8 out of the 20 individuals are females). As for the age, it can be defined two different groups, one with the individuals with more than 60 years old and other with less or equal. Note that the individuals are ordered based on the survival time to facilitate the analysis and obtain the Kaplan-Meier curves. In Table 2.2, within each considered variable, the two groups are specified with different colours:

- Age - green cell if less than 60 years old and yellow if greater or equal to 60;
- Sex - blue if male (1) and red if female (2).

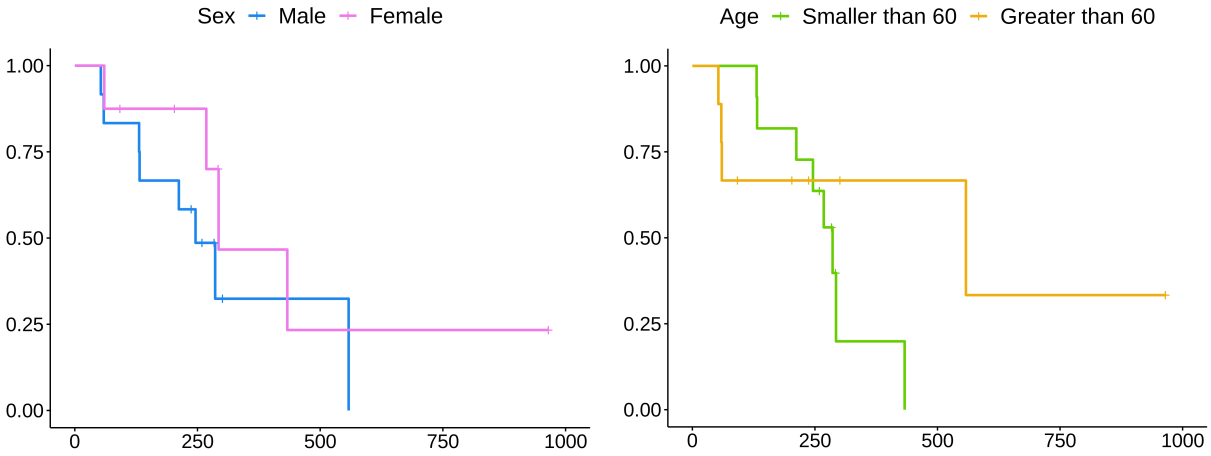
index	*time(days)*	*status*	*age*	*sex*
1	53	2	68	1
2	59	2	73	1
3	60	2	65	2
4	92	1	64	2
5	131	2	50	1
6	132	2	40	1
7	203	1	71	2
8	212	2	49	1
9	237	1	69	1
10	246	2	58	1
11	259	1	58	1
12	268	2	44	2
13	284	1	39	1
14	286	2	53	1
15	292	1	51	2
16	293	2	59	2
17	301	1	61	1
18	433	2	59	2
19	558	2	70	1
20	965	1	66	2

Table 2.2: Lung cancer survival dataset sample.

Considering these two situations, the respective Kaplan-Meier curves have been obtained as shown in Figure 2.2. Even though the number of observations is small, some interesting insights can be taken from that figure. For instance, considering sex, Figure 2.2 (a), the curve associated to the female individuals is above the male curve which means female individuals seem to have smaller death risk when considering lung cancer. Regarding the age, Figure 2.2 (b), it looks like most of the individuals from the older group experience the event sooner since nearly half of them have a survival time of fewer than 100 days. This intuition seems reasonable because the older the person, the weaker their immunologic system. However, this hypothesis requires further study because there are three individuals inserted in the older group that live considerably longer than the average (likely because the evolution of the disease is slower on older persons).

The presented analysis was simply based on the curves appearance, however, further validation is needed to consider the separation between the curves statistically significant. An well know measure is the computation of the p-value using a statistical hypothesis test called the log rank test. It tests the null hypothesis that there is no difference between the population survival curves, meaning, the probability of an event occurring at any time point is the same for each population [23]. The test statistic is calculated considering,

$$\chi^2\left(log\, rank\right) = \frac{\left(O_1 - E_1\right)^2}{E_1} + \frac{\left(O_2 - E_2\right)^2}{E_2}, \tag{2.1}$$



(a) Sex (b) Age

Figure 2.2: Kaplan-Meier curves for lung cancer survival sample.

where O_1 and O_2 are the total numbers of observed events in groups 1 and 2, respectively, and E_1 and E_2 are the total numbers of expected events. The E_1 are computed according to

$$E_1 = \sum_{i=1}^{k} \frac{d_i}{r_i} \cdot r_{1i}, \tag{2.2}$$

with k as the number of individual in study, d_i as the number of occurred events, r_i as the total number of individuals that did not experience the event and r_{1i} as the total number of individuals that did not experience the event on the group 1. The E_2 is computed accordingly to the same formula but considering elements from the group 2 instead of the group 1.

This gives the p-value, being considered statically significant separation between population survival curves if it is smaller than 0.05. On the case of the example presented in Figure 2.2, the p-value for the age curves is 0.34 and for the sex curves is 0.3, not being statistically significant. Of course, this is just an example, and the conclusions are very limited due to the small the number of individuals.

Now that the survival data and respective ways to analyse it have been presented, the way to model this type of inputs is going to be considered. This is achieved through Cox regression, more specifically with Cox PH Model [11] developed by Cox in 1992. There are some other models such as Exponential and Weibull [24]. However, the Cox PH Model has proved to be more robust [4, 19].

Considering the usual survival analysis framework with $((\boldsymbol{x}_1, y_1, \delta_1), ..., (\boldsymbol{x}_n, y_n, \delta_n))$, where n is equal to the number of individuals in the study, $\boldsymbol{x}_i$ is the gene expression profile and y_i is the observed time, being the time of failure if δ_i is 1 or right-censoring if δ is 0. As in regression, $\boldsymbol{x}_i'$ is a vector of potential predictors $(x_{i1}, x_{i2}, ..., x_{ip})$, in this case, considering p genes. The Cox model assumes a semi-parametric from for the hazard

$$h_i\left(t\right) = h_0\left(t\right) e^{\boldsymbol{x}_i'\boldsymbol{\beta}}, \tag{2.3}$$

where $h_i(t)$ is the hazard for patient i at time t, $h_0(t)$ is an unspecified baseline hazard, and $\boldsymbol{\beta}$ represents the regression coefficients, being a fixed, length p vector.

The regression coefficients are obtained by maximising the Cox's log-partial likelihood

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{x}_i' \boldsymbol{\beta} - \log \left(\sum_{j:y_j \geq y_i}^{n} e^{\boldsymbol{x}_j' \boldsymbol{\beta}} \right) \right). \tag{2.4}$$

By considering the partial log-likelihood, all the information between failure times is ignored. Another important point to consider is the fact that this formula assumes that failure time t is unique, $t_1 < t_2 ... < t_n$, however, this is normally not the case. A solution to this problem was presented by Efron as the approximation of the partial likelihood for ties [25].

In the cases where $n > p$, the baseline hazard is ignored. To estimate the baseline hazard, $h_0(t)$, the Breslow estimator is commonly used [26], defined as

$$\hat{h}_0(t) = \frac{1}{\sum_{i=1}^{n} e^{\boldsymbol{x}_i' \boldsymbol{\beta}}}. \tag{2.5}$$

The partial likelihood and the Breslow estimator are induced by the total log-likelihood given by

$$l(\boldsymbol{\beta}, h_0) = \sum_{i=1}^{n} -e^{\boldsymbol{x}_i' \boldsymbol{\beta}} H_0(t_i) + \delta_i \left(\log \left(h_0(t_i) \right) + \boldsymbol{x}_i' \boldsymbol{\beta} \right), \tag{2.6}$$

with

$$H_0(t_i) = \sum_{t_k \leq t_i} h_0(t_k). \tag{2.7}$$

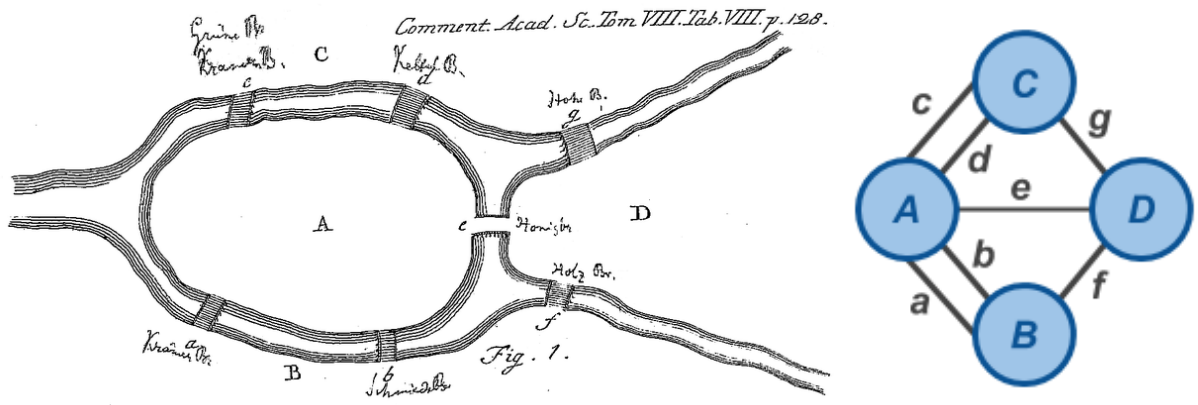
The inference of the optimal regression coefficients is then computed by maximizing the total log-likelihood. Moreover, with the definition of the $\boldsymbol{\beta}$ vector and $h_0(t)$, the patients' hazard relative risk can be computed according to the Eq. (2.3).

2.2 Networks Properties

The presented method strongly focus on the analyse of a big and complex network. For that reason, it is important to explore the main network properties to understand the relevance of the elements involved and the network topology.

Graphs have been studied since the early 18^{th} century with the discussion of the Seven Bridges of Königsberg problem, by Leonhard Euler. This problem schema is presented in Figure 2.3, being B and C both sides of the Pregel River and A and D two large islands (Kneiphof and Lomse). All of them were connected by 7 bridges a, b, c, d, e, f and g and the challenge was to visit all the city passing all the bridges once and only once. Leonhard Euler concluded that it was not possible to achieve a solution to the problem and developed nomenclature to deal with similar problems. His work greatly contributed to this huge research area used in many different fields of knowledge called graph theory [27].

The schema depicted in Figure 2.3 (a) can be represented through a graph, Figure 2.3 (b). The presented representation is the common framework having $G := (V, E)$, with V denoting the set of nodes, in this case, parts of Königsberd city, and E the weighted interaction between them, the bridges. This



(a) Original illustration (source [28])

(b) Graph representation

Figure 2.3: Königsberd's seven bridges problem illustration.

structure is used in many fields of knowledge as, for instance, computer science to represent integrated systems, sociology to present and study how individuals relate to each other and biology to represent how proteins or genes are related. Over the years, there's been a tendency to distinguish between the usage of the graph structure to detail the mathematical object representing the topology of systems and describe physical and biological systems. The later is the one of interest for the problem faced, being commonly called network. When dealing with networks, due to the amount of information to store in a single network, it often leads to huge networks impossible to analyse without clear metrics. One of the most important metrics for this type of networks is the "connectivity distribution, estimated by counting the frequency $P(\lambda)$ of nodes of degree k" [29]. As the connectivity distribution stands for analysing the full network, there are also metrics to analyse the importance of each node in the network, in particular, the centrality measures. The ones considered are closeness centrality, betweenness centrality and degree centrality.

A very interesting work was presented by Travers and Milgram in 1977 [30], verifying that the average number of links to connect any two people in the United States was 6. However, only later, in 1998 this concept was defined by Watts and Strogatz naming this type of graph as small-world network [31]. This types of networks are frequent in systems developed by the human such as the world's banking system [32] or the links between pages in the Wikipedia [33]. All these systems are characterised by the fact that it is possible to connect any two nodes with a relatively low number of connections given the network size.

Another attractive property that usually is associated with this type of networks, in real-world graphs, is the scale-free topology, which means the connectivity distribution follows the power law function [34]. This means the probability $P(k)$ that a vertex in the network interacts with k other vertices decays as a power-law, as presented in

$$P\left(k\right) \sim k^{-\gamma}, \tag{2.8}$$

with γ as the degree exponent that represents "the slop of the logarithmic graph of the frequency of nodes $P(k)$ of degree k versus the node degree itself" [34].

To understand this concept of scale-free topology in small-world networks, the Google and Apple patents creation network, presented in Figure 2.4. In these networks each node corresponds to a person, and the higher the number of patents associated with that person, the bigger the node is. There are connections between nodes and those correspond to cooperation between persons in the network for patent creation. On the right, the Apple network is presented, and it can be seen that it focus on some few persons that have a substantial impact concerning discoveries for the company. The shape of the Apple network is the typical form of a scale-free topology. On the left, it is presented the Google network that has a different topology, having many nodes that have a small/medium contribution in term of patents creation. It is interesting to verify that just by looking at this image, it is clear that the company strategy is very different.

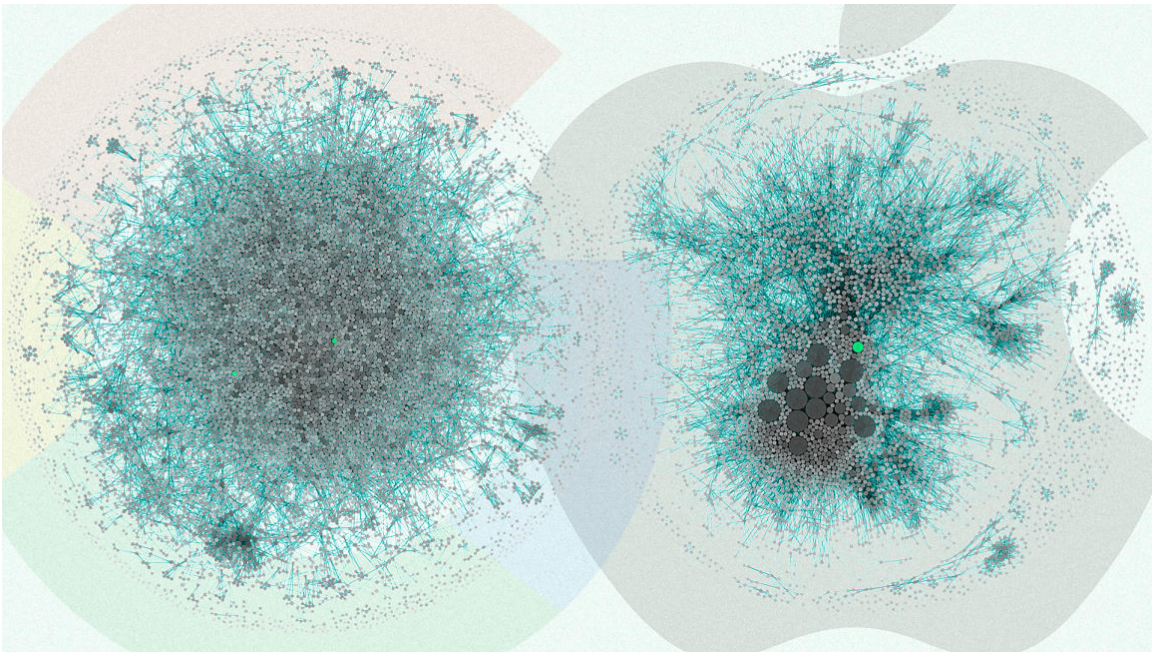


Figure 2.4: Google and Apple patents network (source [35]).

This type of scenarios indicates that large networks self-organise into a scale-free state, resulting in only a few nodes in the network, called hubs, that have a high number of connections (high degree value) and many with few connections (low degree value). The hubs usually are very powerful because of their influence on the network. Taking Figure 2.4 as an example, in the Apple network, elements like Steve Jobs (the green node on the right), had a substantial impact, whether it is positive or negative.

Surprisingly, this goes beyond sociology relations and structures created by the human being. Other complex networks such as genetic, protein and metabolic networks have been analysed and showed the same properties, meaning it is essential to consider them while dealing with biological networks.

2.2.1 Degree Centrality

Focusing on the connections between two nodes in a graph, there can be two different types of graphs: weighted and unweighted. The former corresponds to the cases that each edge in the graph can have a different value based on some variable, for instance, cost or distance. The latter corresponds to a Boolean

merely stating if the two nodes are connected or not in the network. A graph can also be classified as a direct graph if directions are associated with the edges, having $s_{ij} \neq s_{ji}$. On the scope of this project, however, the network that will be considered is undirected, meaning $s_{ij} = s_{ji}$. For that reason, the concepts that will be presented are focused on undirected networks.

When dealing with unweighted networks, the degree of a node d_i is the number of nodes adjacent to it. The degree formula is therefore given by

$$d_i = \sum_{j=1}^{P} a_{ij}, \tag{2.9}$$

with P equal to the total number of nodes, $a_{ij} = 1$ if node i and j are connected and $a_{ij} = 0$ otherwise. Note that the node degree was defined in the early days of graph theory. However, the usage of the degree as centrality measure was only presented in 1974 [36].

In order to work with weighted networks, extensions of the Eq. (2.9) have been proposed. The weighted degree formula is defined as

$$D_i = \sum_{j=1}^{P} s_{ij}, \tag{2.10}$$

where s_{ij} corresponds to the normalised weight of the edge [37].

For both presented metrics, the nodes with high degree value are called the hubs and may be in the path between many other nodes with lower degree value. In scale-free typologies there usually are only a few nodes of this type, being the ones responsible for keeping the longest path of all the shortest paths in a network (diameter) low. An illustration of the degree metric is presented in Figure 2.5 (a), with the nodes diameter directly proportional to their degree value.

2.2.2 Betweenness Centrality

The betweenness is a centrality measure, defined by Freeman in 1977, that is based on the shortest paths between nodes [38]. Considering the node y_i, the betweenness is the frequency of the presence of the node y_i in the shortest paths between every two vertices (y_j,y_k) in the network, with $i \neq j \neq k$. The betweenness centrality B_i is, therefore, given by

$$B_i = \sum_{\substack{j=1 \\ j \neq i}}^{P} \sum_{k=j+1}^{P} \frac{g_{jk}\left(y_i\right)}{g_{jk}}, \tag{2.11}$$

with g_{jk} equal to the number of shortest paths between node y_j and y_k, and $g_{jk}(y_i)$ as the number of shortest paths between y_j and y_i with node y_i present.

This metric is significant because it gives the idea of the "flow" through the vertices in the network. Although the degree might seem a better indicator of node importance in the network (the more connections, the greater the influence in the network), the betweenness will probably still give high values to the high degrees nodes (hubs). Moreover, it will also give high values to nodes with a low degree that make the bridge between two big groups in the network as presented in Figure 2.5(b). Investigations have

been developed to study the relation between centrality metrics. This investigations have shown a strong relationship between the degree and betweenness centrality [39]. Exploring this relation on biological networks can lead to significant insights given the characteristics of both metrics.

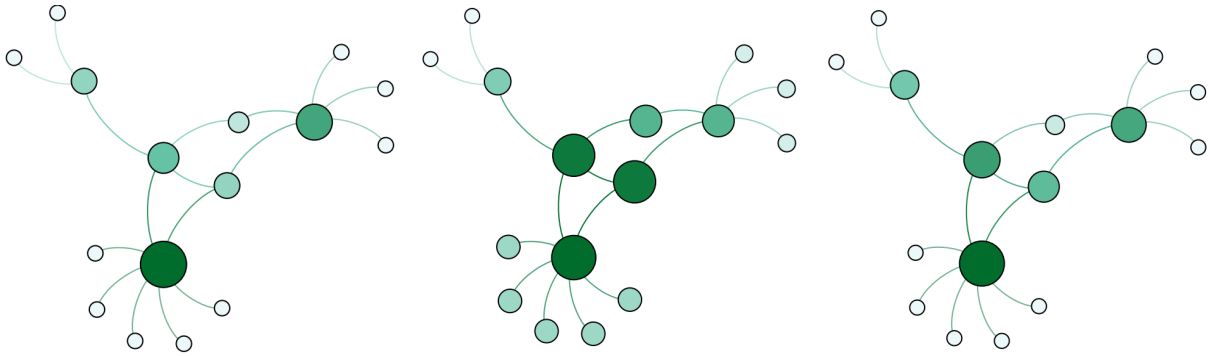
2.2.3 Closeness Centrality

For a specific node y_i, the closeness centrality value corresponds to the inverse of the sum of shortest paths to every node y_j in the network with $i \neq j$ [40, 41]. It is given by

$$C_i^{-1} = \sum_{j \neq i}^{P} g_{ij}, \tag{2.12}$$

having g_{ij} as the distance of the shortest path between node y_i and y_j.

Considering the presented metric, it is expected that nodes in the middle of the network have smaller closeness centrality values. However, this might not be directly related to the node importance in the network. Another important fact to take into account is that this metric can only be used in the case of a connected graph, meaning that from any node y_i it is possible to reach any other y_j with $i \neq j$. If there is no path between any two nodes, that means that all the nodes have at least one node they cannot reach. That scenario results in a situation where $s_{ji} = s_{ji} = \infty$, meaning the centrality is going to be zero to all nodes.



(a) Degree (b) Betweenness (c) Closeness

Figure 2.5: Degree, betweenness and closeness centrality measures (source [8]).

2.3 Regularization Methods

For situations where the number of variables is smaller than the number of individuals ($p < n$), like the one analysed in the example, the Cox model has a reasonably good performance. However, when this is not the case ($p \gg n$) it might lead to a degenerate behaviour (the larger the difference between p and n the worse): most of the regression coefficients tend to $+\infty$, and multiple possible solutions appear. Having this type of situation may lead to severe complications when dealing with new observations to classify, being a well known computational problem.

Some regularisation methods have been presented over the years to constrain the solution space

further. Many of them appeared as a way to get a clear solution considering common regression types, for instance, linear [5], having later adjusts to deal with Cox Model [14]. However, given the importance of extracting significant insights from life-tables, regularisation techniques were developed for the Cox Model such as the Net-Cox and DegreeCox [7, 8].

The total log-likelihood, presented in Eq. (2.6) at page 10, is penalised according to

$$l(\beta, h_0) = \sum_{i=1}^{p} \left(-e^{x_i'\beta} H_0\left(t_i\right) + \delta\left(\log\left(h_0\left(t_i\right)\right) + x_i'\beta\right)\right) - \lambda P\left(\beta\right), \tag{2.13}$$

with λ as the variable that controls how much the solution space is constrained and $P(\beta)$ as the penalisation function according to the regression coefficients, β. Depending on the regularisation technique used, $P(\beta)$ will be estimated in different ways. Their penalisation formula ($P(\beta)$) will be presented in the following pages.

2.3.1 LASSO, Ridge and Elastic Net Regressions

To solve this unwanted, for scenarios with $p \gg n$, Tibshirani has proposed the usage of the L_1 norm penalty in the Cox Model to constrain the solution space [14].

LASSO is widely used because it imposes sparsity in the solutions (many coefficients equal to zero), getting well-defined solutions while making feature selection. Just like LASSO, many other penalties methods have been proposed, such as Fused LASSO and Ridge regression [12, 42]. The Ridge regression considers the L_2 penalty instead of the L_1, which leads to unclear solutions, where differ from zero. Nevertheless, it is still a strong penalisation technique as it handles correlated coefficients better. If two coefficients are strongly correlated the Ridge regression will give equal weight to both while LASSO would probably choose only one of them as non-zero.

Regarding the L_1 norm penalty in LASSO, it constrains the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant, being defined as

$$P(\beta) = \sum_{i=1}^{p} |\beta_i|. \tag{2.14}$$

On the other hand, the Ridge regression constrains the sum of the squared coefficients, having

$$P(\beta) = \sum_{i=1}^{p} \beta_i^2. \tag{2.15}$$

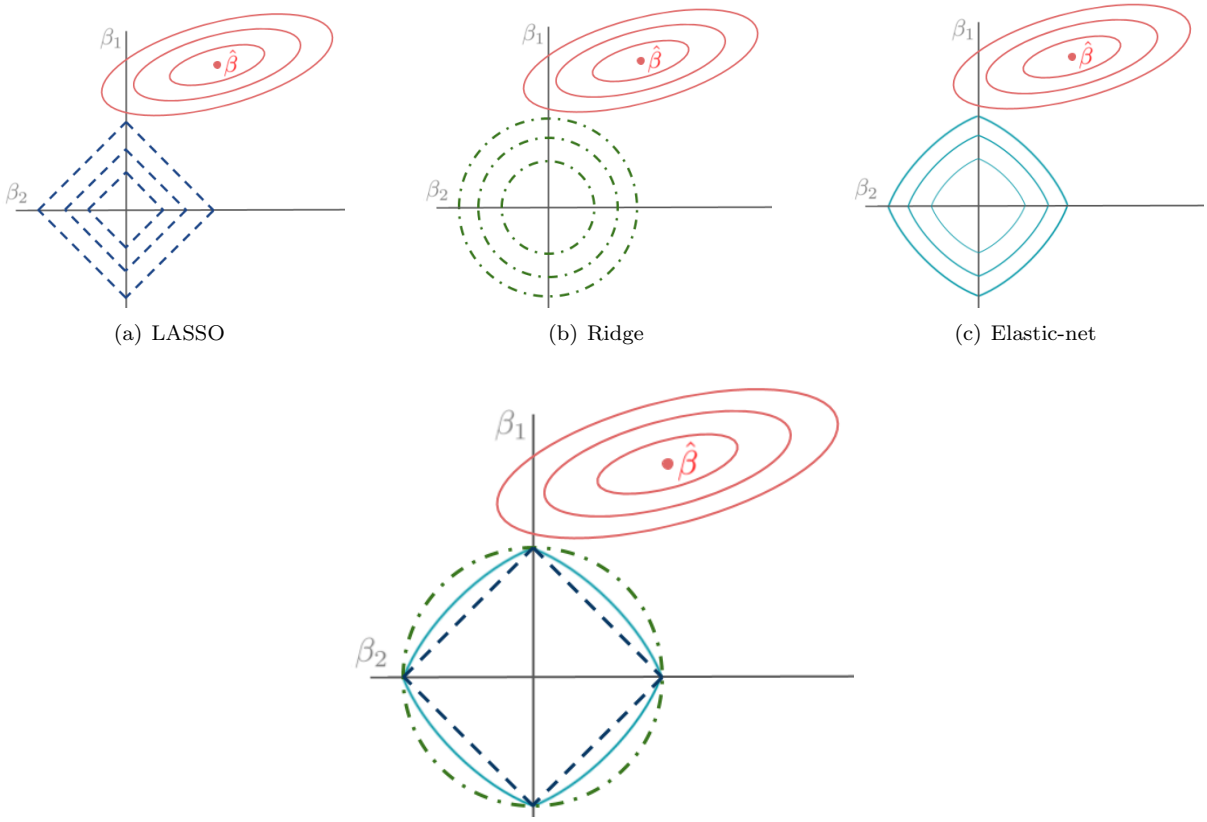
Both penalties had their strengths and weaknesses and based on that, the elastic net method was created, combining the strengths of the two approaches. In this case, LASSO and Ridge regression formulas are joint in a single one having α as a controller between L_1 and L_2 penalties, given a fixed λ, given by

$$\lambda P_\alpha\left(\beta\right) = \lambda \left(\alpha \sum_{i=1}^{p} |\beta_i| + \frac{1}{2}\left(1-\alpha\right) \sum_{i=1}^{p} \beta_i^2 \right). \tag{2.16}$$

With $\alpha = 0$, the constraint applied is like the Ridge regression, and with $\alpha = 1$, the constraint is

equal to LASSO. The ability to balance sparsity and correlation between variables makes the elastic net a much more flexible method for different types of problems/datasets. For instance, by choosing α values very close to 1, for instance 0.95, the solution is going to be very sparse. However, it still keeps regression coefficients different than zero for variables that are strongly related and significant, given the considered problem.

In Fig 2.6 a geometric perspective is presented to understand the differences and similarities between the three regularizers better. The red curves represent the objective function to minimise $f(\beta)$, in this case, the mean squared error (MSE). The red dot in the middle of the curves corresponds to the non-penalised solution. However, this is not the wanted solution for the case $p \gg n$ there is no restriction on the β values leading to multiple possible solutions. Therefore another objective function is added depending on the regularizer applied, $g(\beta)$. The objective is to find the minimum of the sum of this two objectives, which corresponds to the intersection between the curves.



(a) LASSO (b) Ridge (c) Elastic-net

(d) All

Figure 2.6: Geometric interpretation of LASSO, Ridge and Elastic-net regressions (adapted from [6]).

For the sake of the example, only the features β_1 and β_2 are being considered, having the LASSO with $g(\beta) = \lambda(|\beta_1| + |\beta_2|)$ and Ridge regression with $g(\beta) = \lambda(\beta_1^2 + \beta_2^2)$. Now looking at the curves, it is clear to see that the LASSO nearly always implies selecting only one of the coefficients into consideration, in this case, the β_1. On the other hand, when only considering the L_2 penalty, both β_1 and β_2 regression coefficients will have a value assigned even though it is clear that the parameter β_1 is much more significant. Considering the Elastic-net curves, the following scenarios can occur:

- if α is close to 1 the curves are going to be sharper just like in LASSO, only probably selecting β_1;
- if α is close to 0 the curves will approximate the Ridge curve, most likely having β_1 and β_2 regression coefficients with non-zero values.

Having the penalisation function defined as Eq. (2.16), when considering the elastic-net regularization, the log-likelihood is therefore given by

$$l_{L_1L_2}(\beta, h_0) = \sum_{i=1}^{n} \left(-e^{X_i'\beta} H_0(t_i) + \delta \left(\log \left(h_0(t_i) \right) + X_i'\beta \right) \right) - \frac{1}{2}\lambda \left(\alpha|\beta|_1 + (1-\alpha)|\beta|_2^2 \right). \tag{2.17}$$

2.3.2 Net-Cox Regression

In 2013, another interesting idea was proposed based on the fact that many relation networks were being constructed based on previous studies and discoveries. Zhang et al. believed that network-based computational models were "attracting increasing attention in studying cancer genomics because molecular networks provide valuable information on the functional organizations of molecules cells" [7]. Based on that and given that gene expression was being considered, gene relation networks can be used:

- gene co-expression network – based on the given data;
- gene functional linkage – features a large variety of biologically informative perspectives.

Based on those networks, a network constraint to the Cox model was developed, considering both L_2 norm and graph-based constraint. Given a normalised graph weight matrix $\boldsymbol{S}$, it is assumed that related genes should be assigned similar coefficients by respecting the cost term

$$\boldsymbol{\Psi}(\beta) = \frac{1}{2} \sum_{i,j=1}^{p} S_{ij} (\beta_i - \beta_j)^2 = \beta' (\boldsymbol{I} - \boldsymbol{S}) \beta = \beta' \boldsymbol{L} \beta. \tag{2.18}$$

This consideration encourages smoothness among the regression coefficients in the network, having, for any pair of genes connected by an edge, a cost proportional to both the difference in the network and the edge weight.

In regard to Figure 2.7, consider the regression coefficients β_a and β_b value difference was very high as well as the value S_{ab}. This scenario would result in a very high penalisation for both coefficients, meaning that the objective function encourages similar coefficient values to genes connected by edges with greater weight. Aiming for regularising the uncertainty of the network an additional L_2 norm constraint is added to $\boldsymbol{\Psi}(\beta)$. With α as the parameter adjusting between the L_2 norm and the "Lagrangian-norm" constraints, the penalisation function can be rewritten as

$$\lambda P_\alpha(\beta) = \lambda \left((1-\alpha) \beta' \boldsymbol{L} \beta + \alpha |\beta|^2 \right) = \frac{1}{2} \lambda \beta' \left((1-\alpha) \boldsymbol{L} + \alpha \boldsymbol{I} \right) \beta. \tag{2.19}$$

Considering the penalization function given by Eq. (2.19), the total log-likelihood is presented as

$$l_{NetCox}\left(\beta, h_0\right) = \sum_{i=1}^{n}\left(-e^{X_i'\beta}H_0\left(t_i\right) + \delta\left(\log\left(h_0\left(t_i\right)\right) + X_i'\beta\right)\right) - \frac{1}{2}\lambda\beta'\left(\left(1-\alpha\right)\boldsymbol{L} + \alpha\boldsymbol{I}\right)\beta. \tag{2.20}$$

2.3.3 DegreeCox Regression

In the sequence of the work presented by Zhang et al., Veríssimo et al. proposed the DegreeCox[8]. This method considers the same networks as the Net-Cox, yet, the penalisation on the regression coefficients is based on a centrality measure. More precisely, it considers the degree centrality measure for each regression coefficients in the obtained networks. This constraint is given by

$$\Upsilon(\beta) = \sum_{i=1}^{p}\beta_i^2 d_{ii} = \beta'\boldsymbol{D}\beta, \tag{2.21}$$

where $\boldsymbol{D}$ is a diagonal matrix with $D_{ii}^{-1} = \sum_{j=1}^{p} S_{ij}$, i.e., the inverse of the vertex weighted degree.

That means the regression coefficients will be further penalised if they are associated with low degree values. This penalisation method was built on the assumption that nodes with high degree level will have a strong influence in the network, being, therefore, less penalised. For instance, in Figure 2.7, given the number of connections of the regression coefficient β_e the penalisation would be much less than the one applied on β_a, being presumed for the gene expression associated to β_e to be considered in the final model. By adding the DegreeCox constraint to the Cox model, the full likelihood is given by

$$l_{DegreeCox}\left(\beta, h_0\right) = \sum_{i=1}^{n}\left(-e^{X_i'\beta}H_0\left(t_i\right) + \delta\left(\log\left(h_0\left(t_i\right)\right) + X_i'\beta\right)\right) - \frac{1}{2}\lambda\beta'\boldsymbol{D}\beta. \tag{2.22}$$

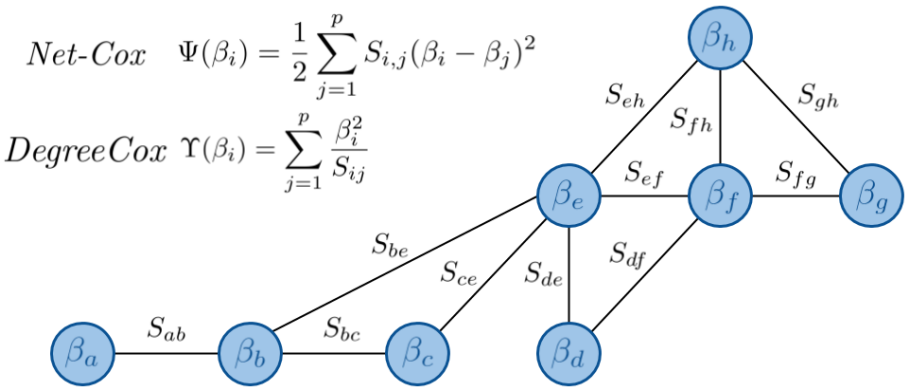


Figure 2.7: Net-Cox and DegreeCox network regularizers (adapted from [8]).

2.4 STRING Dataset, TCGA and BioMart

Based on the evolution of the regression methods over time to deal with datasets with the dimensionality concern, it is clear that the use of networks information is very relevant in this field. Dealing with gene expression dataset involves dealing numbers of features with four orders of magnitude and, if lucky, with little more than a thousand individuals. Therefore, having information regarding biological behaviours

and relationships between entities at a molecule level can be a significant help in the process. There have been many developments in the field to enrich these types of databases over the years; data that is frequently stored in a graph format. Many datasets are available online for the bioinformatics to use and take insights. These datasets use the graph format to translate relations between entities, for instance, proteins or genes.

Genes are distinct sequences of nucleotides that form part of a chromosome and corresponds to the basic unit of heredity. All living things depend on genes as they specify all proteins and functional RNA chains through the interaction between molecular regulators and other substances in the cell. DNA, RNA, proteins and complexes of these are examples of regulators. The interaction between them is responsible for conducting the genes' expression levels.

Proteins are large, complex molecules that play many critical roles in the body. They also determine how the organism looks, how well the body metabolises food or fights infection and sometimes even how it behaves.

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. Over the years, measurements techniques have been developed and improved by scientists since it is a significant matter for life sciences. One of these techniques considers protein degradation, giving emphasis to the relation between gene expression and protein activity [43–45].

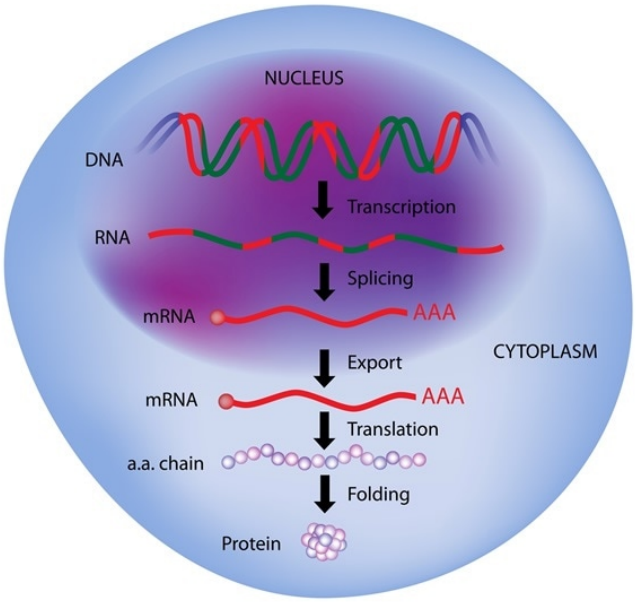


Figure 2.8: Protein synthesis diagram (source [46]).

The journey from gene to protein is controlled within each cell and consists of two significant steps – transcription and translation. The proteins synthesis process is presented in Figure 2.8. It starts with genetic transcription, resulting in a mRNA. After that, the maturation of the mRNA is performed, and finally, protein synthesis is achieved through translation of the mature mRNA.

Each protein-coding gene is responsible for creating proteins needed for the good function of the organism, uncovering a strong relationship between them. Having stablished the bridge between genes,

proteins and gene expression, a network of protein-protein interaction will be used to constrain the solution space further.

The used database was extracted from the STRING, a well-documented and updated collection of data, featuring known and predicted protein-protein interactions for many different organisms [9]. The interactions include direct (physical) and indirect (functional) associations, stemmed from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases. Interactions in STRING are derived from five primary sources:

- Genomic context predictions
- High-throughput lab experiments
- (Conserved) co-expression
- Automated textmining
- Previous knowledge in databases

The dataset considered for training and test the proposed method is extracted from the The Cancer Genome Atlas (TCGA). The TCGA is a collaboration between the National Cancer Institute and National Human Genome Research Institute, that has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer so far. Their datasets comprise 2.5 petabytes of data describing tumor tissue and matched normal tissues from more than 11,000 patients [2]. Their datasets are available to everyone and have been widely used by the research community [8, 47].

The studied dataset concerns survival data of cancer patients, having death as the event of interest and gene expression values as features. These features values are obtained through a process that involves measurement of the gene expression from several cells from a sample of a specific tissue. The values of each gene expression correspond to the average value given all the analysed cells. Those values are achieved through one of the following techniques: Fragments per kilobase (FPKM), Reads per kilobase (RPKM) or Transcripts per million (TPM) [48]. The process involved in each technique will not be specified.

Figure 2.9 corresponds to an example given by the STRING development team, presenting the proteins associated with the 20 most frequently mutated human cancer genes. Each node corresponds to a specific protein, and each line connecting two nodes indicates a channel through those nodes are connected.

Each protein-protein connection is weighted based on seven different *channels* presented in Table 2.3. The table holds a brief description of how channels' values are obtained/calculated and a different colour is assigned according to the channel type: purple for known interactions, yellow for predicted interactions and green for other. Besides these channels, STRING provides the combined score for each protein-protein connection [50]. This score is often higher than the individual sub-scores, expressing increased confidence when several types of evidence support an association.

The STRING database is widely used, covering an extensive range of organisms and proteins. Currently comprises more than two thousand organisms, including the *Homo sapiens*, and nearly ten million proteins. This dataset has been used to support the hypothesis that highly connected proteins have a

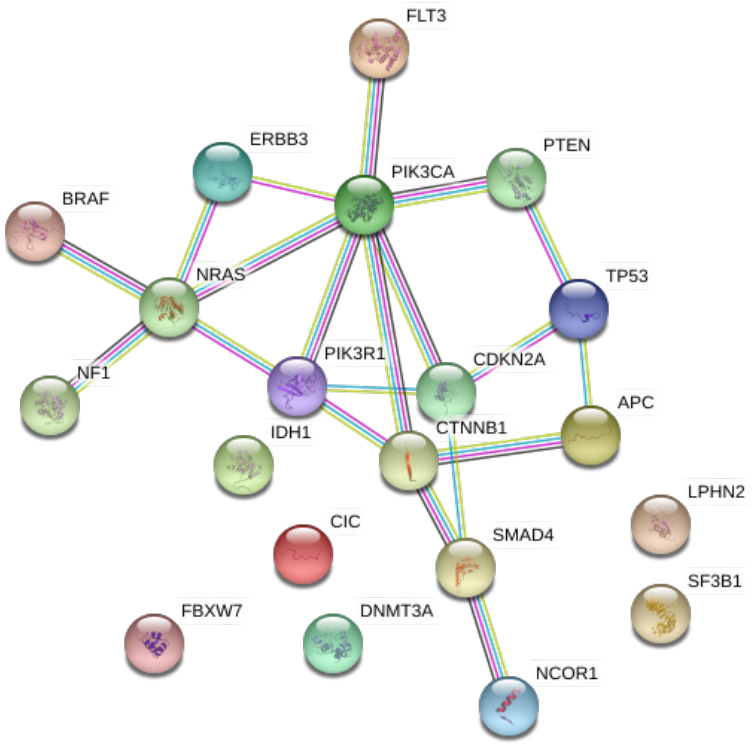


Figure 2.9: The 20 most frequently mutated human cancer genes' proteins (source [49]).

stable steady-state distribution of gene expression [51]. It has also been used for searching candidate genes involved in the immune response to gluten [52].

Channel	Details
Curated Databases	Asserted by a human expert curator - pathway databases.
Experimentally Determined	Comes from experiments in the lab - primary interaction databases organised in the IMEx consortium, plus BioGRID
Gene Co-occurrence	Considers the phylogenetic distribution of orthologs of all proteins in a given organism. If two proteins show high similarity in this distribution, then an association score is assigned.
Gene Neighbourhood	Genes are given an association score where they are consistently observed in each other's genome neighbourhood.
Gene Fusions	Given an association score when there is at least one organism where their respective orthologs have fused into a single, protein-coding gene.
Textmining	Based on mentions of protein names in all PubMed abstracts
Co-expression	Based on gene expression data originating from a variety of expression experiments. Pairs of proteins have a higher score if consistently similar in their expression patterns under a variety of conditions.

Table 2.3: Protein-protein interaction features (source [9]).

The usage of this type of information has proven to be reliable, leading to relevant insights. With time the available information will grow in quantity and quality resulting in a better and richer database. However, the information presented so far focus on proteins and the problem in study concerns gene expression. Therefore, the link between genes and proteins needs to be established.

Given the strong relation between proteins synthesis and genes, many studies focusing on this relation

were developed and published for academic and medical purposes. Base on those type of relations between biological entities, the *BiomaRt* package have been developed, allowing the "access to large amounts of data in a uniform way" [53, 54]. The presented package collects and relates the information stored in rich public datasets, for instance, the Ensembl [55], allowing the mapping between structures at the cell level. As explained, the relations between those type of structures are very complex, and discoveries occur daily. For this reason, the BiomaRt functions are crucial for bioinformatics to establish the relationship between proteins and genes. Furthermore, just like the STRING case, as the years go by, the richer the databases get, leading the *BiomaRt* package to comprise more and better connections.

Chapter 3

Proposed Methodology

In this chapter, the explanation of the path taken to achieve the method in study is exposed. In order to do this, first, the pre-process and analysis of the STRING network had to be performed. Based on these network properties, a penalty factor has been considered giving preference to some specific genes comprised in the dataset used for train and validation.

3.1 Computational Model

As previously presented, many works have been developed to deal with the curse of dimensionality problem in gene datasets. Some of those works, considering networks information to further and "better" constrain the solution space. Firmly based on the work exposed by Veríssimo et al. [8], the proposed method also considers centrality measures as a penalty factor. However, the network in analysis focuses on a protein network instead of gene networks (gene co-expression and gene functional linkage).

The idea behind the method under study is to use the L_1 and L_2 norms penalisation as the Elastic-Net method, yet, with an extra penalty factor v_i, described on

$$\lambda \sum_{i=1}^{p} v_i P_\alpha \left(\beta_i\right) = \lambda \sum_{i=1}^{p} v_i \left(\left(1-\alpha\right) \frac{1}{2} \beta_i^2 + \alpha |\beta|_i \right). \tag{3.1}$$

The usage of the *glmnet* package [56] allows the implementation of this model, having a penalty factor defined by the information provided by the STRING dataset.

With that in mind, the objective is to find the best properties in the network that reflect each node importance and use it to control the level of penalisation of the regression coefficients. As demonstrated in Section 2.4, the STRING dataset comprises a robust network regarding protein-protein interactions as it considers different connection channels. Moreover, for each connection, it is also given an overall score named "combined score". With all the edges information, a biological network can be defined as the adjacency matrix A, given by

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}, \tag{3.2}$$

with a_{ij} equal to the "combined score" between protein i and protein j when considering a weighted network.

It can also be studied the unweighted scenario, where a_{ij} is given by

$$a_{ij} = \begin{cases} 1, & \text{if } i \neq j \text{ and } combined\_score_{ij} > \theta \ (i, j \in \{1, 2, ..., n\}) \\ 0\,, & \text{otherwise} \end{cases}, \tag{3.3}$$

where θ is equal to the threshold applied to the "combined score". This defines whether two proteins are considered connected ($a_{ij} = 1$) or not ($a_{ij} = 0$). Having the matrix A defined, the biological network can be seen as a graph $G := (V, E)$, with V denoting the set of proteins and E the weighted interaction between them.

With the presented graph, centrality measures will be used to deduce the importance of each protein in the network because they reflect important characteristics of the nodes, given the network under analysis. The metrics in study are the ones already presented in the Section 2.3: degree, betweenness and closeness.

The described process is presented in Figure 3.1. In addition to the centrality metrics calculation, the relation between them is going to be explored. A greater focus is given to the betweenness centrality metrics because it has a strong relation with degree centrality, already validated by Veríssimo et al. [8], and covers nodes that control the network flow. For that reason, an explanation of the used algorithm to obtain the betweenness centrality vector is presented in Section 3.2. Note that for each of the presented networks, the three centrality metrics are considered, having a total of six possible protein vectors to study.

The presented network analysis is accomplished to find the best measures to have an accurate definition of the most relevant nodes in the overall system. For that reason, as the results are obtained, some of the considered metrics will be rejected, as illustrated in Figure 3.1.

The relevant metrics are going to follow the presented process to reach the final regression coefficients values. The first step to apply over the obtained protein vectors is the protein-gene mapping. With the accomplishment of this process, exposed in Section 3.3, the vector comprises gene information just as the features vector of the considered train/test data. The penalty factor is then computed considering each of the metrics applying the penalty formula presented with detail in the Section 3.4. Finally, the regression coefficients are obtained according to the process presented in Section 3.5.

The centrality metrics selection process, the penalty factor computation and gene-protein mapping results are presented in the Chapter 4. Then, the regressions coefficients are going to be calculated considering the different penalty factor vectors and parameters. The followed process is presented in the

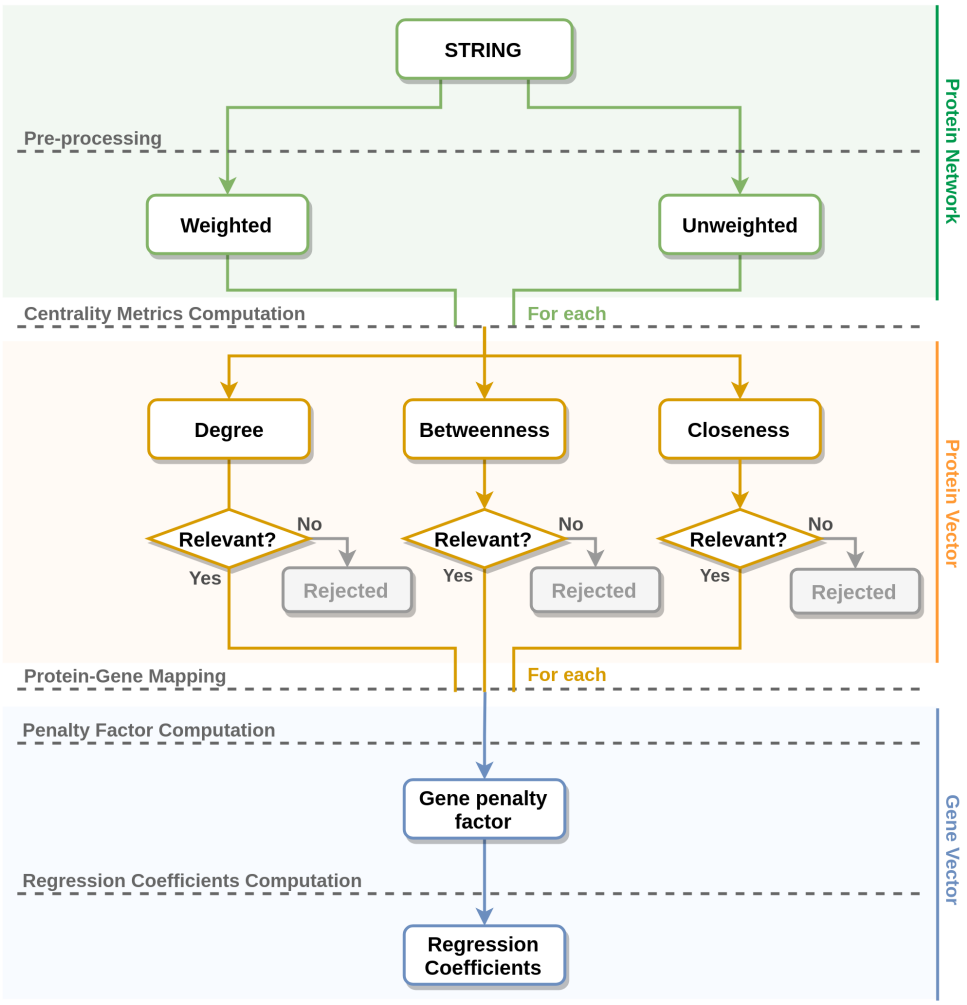


Figure 3.1: Proposed methodology over the STRING network to reach the final regression coefficients.

Section 3.5 and the respective execution and results analysis is exposed in the Chapter 5.

3.2 Centrality Metrics Complexity

Working with biological networks means dealing with a massive amount of information as well as with complex structures. In a protein or gene networks, the number of nodes easily exceeds the thousands and edges can reach the millions. For that reason, it is essential to understand the complexity of the methods not only regarding memory but also time.

The degree metric is a significant metric to consider in networks' analysis and is not difficult to compute. For either weighted and unweighted cases, it consists in going to each node and sum all the edges weights associated with it. Therefore, it is required $\mathcal{O}(n+m)$ time and $\mathcal{O}(n)$ space to obtain this metric values, with n as the number of vertices in the network and m equal to the number of edges between vertices. All the considered centrality metrics were calculated with the usage of the *igraph* package [57].

The calculation of the betweenness centrality is not as simple as the degree one. However, given the fact that it can bring significant insights when associated with other metrics, it is worth tackle the problem. The computation of this metric is associated with a big complexity cost given the involvement

of shortest path calculation. The first algorithm used to calculate this metric required $\Theta(n^3)$ time and $\Theta(n^2)$ space. The high level of resources needed for this metric computation used to be a big limitation when the size of the network was too big, not only because of the space but also the time needed. However, Brandes presented the Algorithm 1 reducing the required resources to $\mathcal{O}(n+m)$ in terms of space and $\mathcal{O}(nm)$ in terms of time [58].

Algorithm 1: Betweenness centrality in unweighted graphs (source [58]).

```
1  C_B[v] ←, v ∈ V;
2  for T in TList do
3      S ← emptystack;
4      P[w] ← empty list, w ∈ V;
5      σ[t] ← 0, tinV; σ[s] ← 1;
6      d[t] ← −1, t ∈ V; d[s] ← 0;
7      Q ← empty queue;
8      enqueue s → Q;
9      while Q not empty do
10         dequeue v ← Q;
11         push v → S;
12         foreach neighbor w o v do
13             // w found for first time?;
14             if d[w] < 0 then
15                 enqueue w → Q;
16                 d[w] ← d[v] + 1;
17             // shortest path t w via v?
18             if d[w] = d[v] + 1 then
19                 σ[w] ← σ[w] + σ[v];
20                 append v → P[w];
21     δ[v] ← 0, v ∈ V;
22     // S returns vertices in order of non − increasing distance from s
23     while S not empty do
24         pop w ← S;
25         for v ∈ P[w] do  δ[v] ← δ[v] + (σ[v]/σ[w]) · (1 + δ[w]) ;
26         if w ≠ s then  C_B[w] ← C_B[w] + δ[w] ;
```

The last studied metric is the closeness centrality that also considers the calculation of shortest paths since it reflects the proximity of the considered node to all other nodes in the network. This fact, by itself, turns the computation of this metric into a complex problem. Nevertheless, the complexity of the problem is the same as the betweenness complexity since both consider the calculation of the shortest paths between every pair of vertices in the network. The best algorithm to calculate all the shortest paths between all the nodes in an unweighted with positive integer weights network was presented by Thorup having also a time complexity of $\mathcal{O}(nm)$.

Note that these algorithms might still not be enough for bigger problems, being necessary methods that estimate this metrics values instead of calculating them precisely. However, given the dimension of the problem and the available computation resources, these algorithms are enough to get results in a reasonable interval of time.

3.3 Protein-Gene Mapping

After the pre-process and study over the STRING data, the bridge between proteins and genes needs to be crossed since the considered survival datasets focus on genes' expression values and not on proteins. Therefore, the respective mapping will be performed, allowing the penalisation of the regression coefficients involved in the Cox regression, considering Eq. (3.1).

The *BioMart* package provides a powerful link between biological databases and microarray data analysis, by bridging proteins and genes information [53, 54]. Unfortunately, some proteins and genes are not fully documented yet, leading to some mismatch between these two entities. As a result, some of the proteins will not have a gene associated, resulting in some of the information obtained about the proteins being ignored.

Moreover, the relationship between proteins and genes is not a one-to-one connection. It is possible to have more than one protein produced by the same gene. Even though there are only a few genes that show this situation, to solve this problem, the genes with more than one protein associated would have the sum of the considered centrality measure values associated to those proteins. The presented solution is adequate because a gene that has more than a protein associated should indeed be considered more relevant.

That mapping between genes and proteins was performed only after the centrality metrics were obtained for each of the proteins. The lack of information regarding the mapping and the need to join information from more than one protein in a single gene would result in a different network if the mapping was applied before these metrics computation. That's why the methodology presented in Figure 3.1 considers always a proteins network and then a proteins network for the validation of considered metrics. Following the inverse order would likely result in an unconnected network that actively harms the centrality metrics computation.

3.4 Penalty Factor

After this pre-processing, it is now feasible to directly work over survival data focused on gene expression. The vector with the "centrality measures" for each gene, will be referred as the gene importance, w, and the objective is to transform this vector to favour the genes with greater importance value. Looking at Eq. (3.1), it can be concluded that, considering a specific gene, the higher the importance, the lower the penalty factor should be. The desired effect can be achieved through the computation of the inverse obtained gene importance. However, this could result in very different ranges of values depending on the used centrality metric.

The higher penalty factor associated with a specific gene, the less likely the respective regression coefficient is going to be considered. For that reason, the higher the centrality metric, the lower the respective penalty factor should be. To have this effect, the penalty factor is given by

$$v_i = \frac{1}{w'_i}, \tag{3.4}$$

with w'_i as the re-scaled centrality metric for the gene i.

The standard procedure would consist in the min-max normalisation of the considered centrality vector, having the respective value between 0 and 1. However, this is not possible given the consideration of the inverse of w'_i. For this reason, the re-scaling process applied over the centrality metric value is given by

$$w' = \frac{w - \min(w)}{\max(w) - \min(w)} + \mu, \tag{3.5}$$

with μ as the parameter that controls the v_i max value ($\frac{1}{\mu}$). This re-scale process results a $\boldsymbol{w}$ between with values between μ e $1 + \mu$.

Note that the parameter μ has a significant impact on the distribution of the penalty vector, being an important parameter to consider on the regression models.

The penalty factor formula has also a impact in Figure 3.1 pipeline definition. The process involve first the protein-gene mapping and only then the penalty factor calculation because of the cases where the more than one protein have more than one gene associated. If these cases were not present, the order of these two operations would be irrelevant.

3.5 Regression Coefficients Computation

As stated, the regression coefficients for the Cox PH Model are obtained according to the Eq.(3.1), being considered different penalty vectors. Within the same centrality metric, it has been shown that the variable μ has a great impact on the penalty vector distribution. This phenom had to be considered in the analysis, being one of the challenges the selection of the best μ value for each of the metrics.

Like the μ value, the α has also a determinant rule in the outcome solution since it has a significant control over the number of regressions considered in the solution. Another important variable that also has a great impact on the results is the train/test, being interesting to verify if a good model is obtained even with few data of if the model is able to avoid overfit when more training data is give.

All these considerations had to be considered in this step, in order to get the best models within each of the centrality metric models. To validate and support the discussion of the proposed method, the Elastic Net will be considered and tested considering the same conditions of the proposed model.

Chapter 4

Network Properties of STRING

The objective of this chapter is to extract relevant insights out of documented protein datasets and use it as the further constraint on the solution space. The STRING dataset was selected to study the proteins relation, being a central "pillar" of this thesis project. Therefore, the STRING network will be analysed in detail, being presented the respective centrality metrics distributions along with other relevant properties.

In that analysis, six different metrics will be studied, having been selected three of them to consider for the final models. However, the resulting properties are assigned to proteins, being necessary the mapping between proteins and genes have. The effects of this mapping process are presented as well as the penalty factor computation. More than one penalty factor distribution will be considered so it can be determined the best one to apply on real data. This analysis is then performed on the Chapter 5.

Note that the discussion over this chapter will focus on the shape of the distributions considering the different centrality metrics. This is the case because a re-scaling factor is applied before calculating the penalty factor due to the significant differences between these metrics range of values.

4.1 Centrality Measures

The file used from the STRING database was *9606.protein.links.detailed.v10.5.txt.gz*, considering only *Homo sapiens* protein-protein interactions [9]. This file contains the values of each channel regarding a connection between any two nodes y_i and y_j with a combined score greater or equal to 150. As previously justified, regarding each connection, all the channels values were ignored except the combined score. The resulting network has 5676527 relations between different proteins (edges) that have an average combined score value equal to 277.6 and never exceeds 999. The total number of proteins considered was 19576, having an average weighted degree of $\langle D \rangle = \sum_{i=1}^{P} \frac{D_i}{P} = 161037$ and unweighted degree of $\langle d \rangle = \sum_{i=1}^{P} \frac{d_i}{P} = 580.007$. It is important to point out that the maximum degree values for the weighted and unweighted network are 2423043 and 7873, which is very high when compared with the mean value. In Figure 4.1 it is presented the considered network with the proteins as the orange circles that have different diameter depending on their unweighted degree value. The edges are presented as grey lines,

yet, given the network dimension, it can only be concluded that there are many connections in the network. That along with the relatively high value of mean unweighted degree values point out the relatively high density of the network.

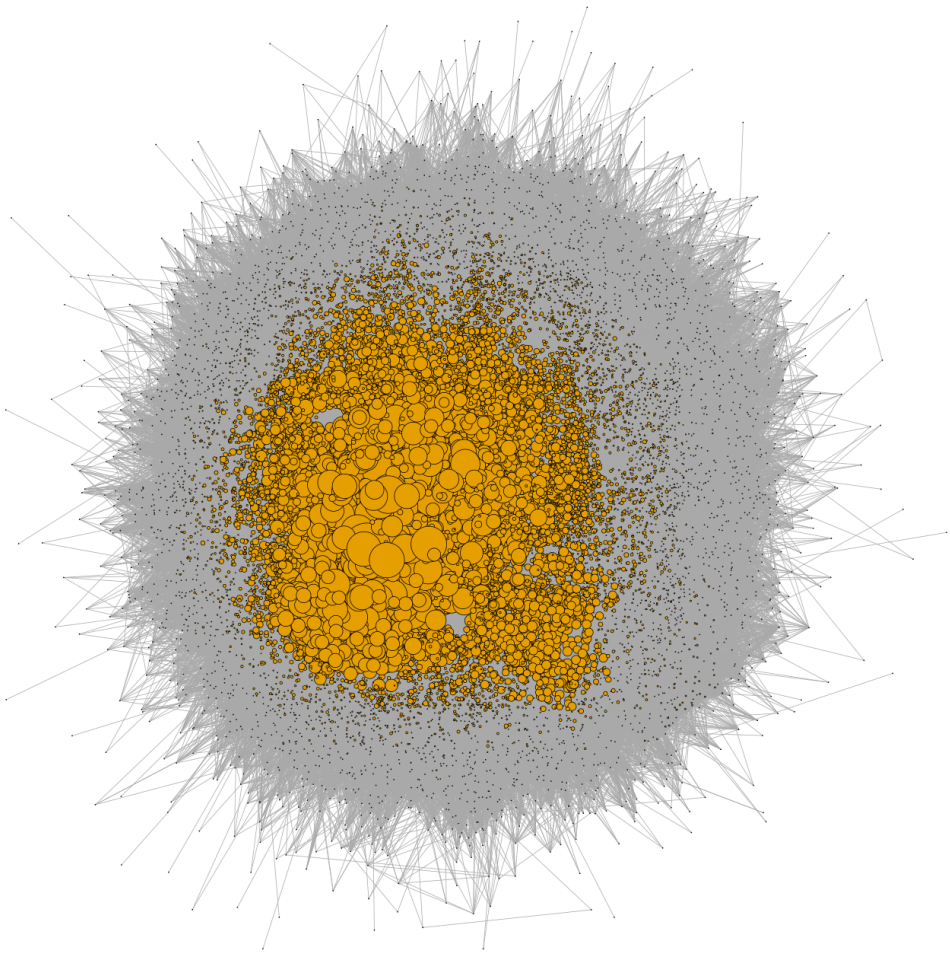


Figure 4.1: STRING protein-protein network focus on *Homo sapiens*.

Another interesting property of this network is that it is not connected, having a strong impact on the calculation of closeness and betweenness. Nevertheless, the graph is composed of only two sub-graphs, and one of those has only two nodes. Those nodes were removed because they would have very low values regarding any metric considered (great penalisation) and would greatly affect the closeness and betweenness values. After that pre-processing step, the average shortest path separating any two nodes in the network shows the value $\langle l \rangle = 2.203$ (shortest path length from a node to itself is always zero), which is very small compared with the network size. The graph density has also been obtain $D' = \frac{2|E|}{|V|(V-1)} = 0.0296$, which means that nearly 3% of the possible edges actually exist in the network. The relatively high-density value along with small $\langle l \rangle$ and Figure 4.1 analysis, gives a strong notion that the network in hand respects a small-world topology. Moreover, given the high values of maximum degree, it seems there might be some few nodes that strongly influence the network behaviour.

4.1.1 Weighted Network

Weighted networks are used when it is important to consider connections of any kind between two entities with a weight attached to that connection. In this particular case, it is considered how well two proteins relate to each other: the higher the value, the stronger the relation is. Therefore, more than just calculate the number of interaction that a protein has with it's neighbours, it can be interesting to consider the"amount" of impact it has on it's surroundings. This can be accomplished by measuring the weighted degree centrality, Eq. (2.10).

Given the STRING information, it was concluded that the only interesting metric to consider would be the degree distribution when using the weighted network. Metrics that comprise shortest path computation in a weighted network are worth considering when dealing with, for instance, distances or time. In the case of relations between entities, this metric is not so relevant. Given that, only on the case of the unweighted network, the three metrics will be considered because it better reflects how entities impact the network.

Degree Centrality

Considering a_{ij} equal to the combined score between node i and j, it is possible to obtain the weighted degree distribution as presented in Figure 4.2. The distribution count axis is in $\log_{10}$ scale to properly analyse the degree distribution and identify the number of hubs network. There is a detail that needs to be mentioned to understand the obtained graph: the count for each bin in the histogram is equal to the real value plus one. The following procedure was applied because there could not be any bin with zero counts due to the $\log_{10}$ scale on the distribution. Therefore, a fictitious node has been inserted for each bin considered just for visualisation purposes, being immediately removed for the following procedures.

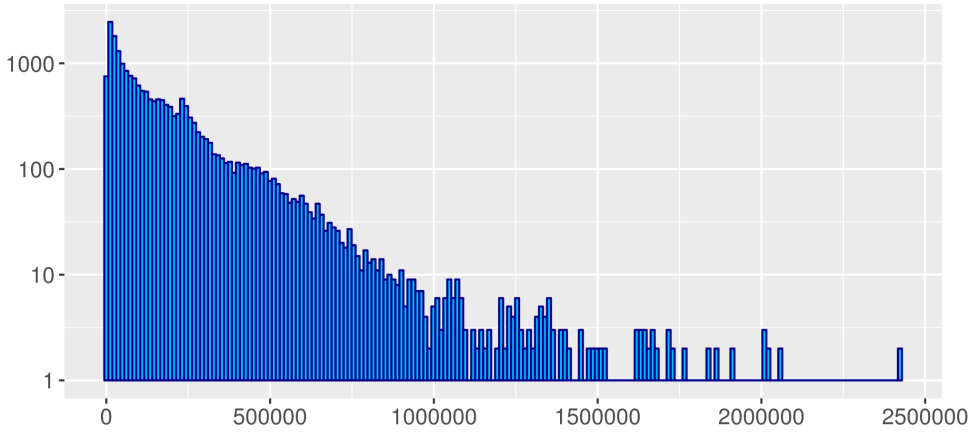


Figure 4.2: Weighted degree distribution in $\log_{10}$ scale with $\theta < 150$.

By analysing the weighted degree distribution, it is clear to see that the number of nodes/proteins with high degree values is, as expected, very low. This type of properties is common in scale-free topology which means, the network degree distribution approximately respects the power law $p(k) \sim k^{-\gamma}$, having $p(k)$ as the probability of a node have that k degree value and γ as a constant. It is, therefore, an important characteristic to consider on biological networks. Zhang and Horvath even stated that "Most biologists

would be very suspicious of a gene co-expression network that does not satisfy scale-free topology at least approximately" [60].

To test this assumption, the *plfit* function from the *poweRlaw* package [61] has been applied in order to see how well does the degree distribution follows the power-law function. This function estimates the x_{min} and γ, being x_{min} the value from which the distribution approximates the power-law and γ the degree exponent. The respective result is presented in Figure 4.3, having been selected $x_{min} = 99953$ with $\gamma = 2.112975$. In this case, the green line does not seem to properly reflect how well the Cumulative Distribution Function (CDF) varies depending on the degree value. A possible reason for the obtained result could be the presence of different degree value ranges that approximately follow the power law distributions and the algorithm method applied is trying to cover both of them.

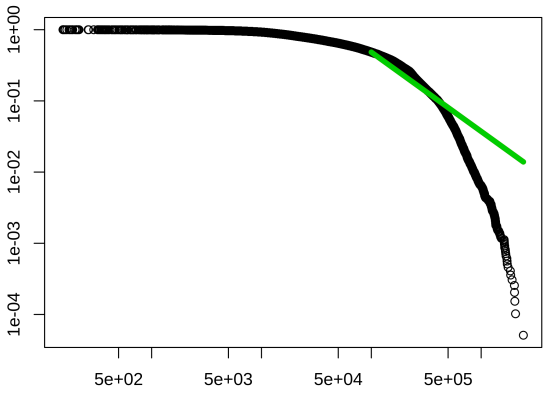


Figure 4.3: CDF plot regarding weighted network degree with $\theta < 150$. The green line correspods to the best powerlaw fit.

With that in mind and given the average shortest path between any two nodes, there are significant pieces of evidence that the exposed STRING-based network is considered a small-world network that approximately follows the scale-free topology for nodes with high degree value. However, further analysis of that hypothesis will be presented.

4.1.2 Unweighted Network

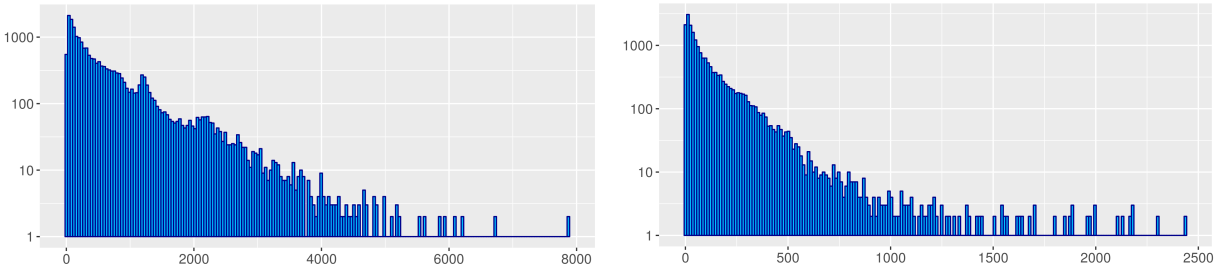
The other obtained network is the unweighted network, that only considers whether two different proteins are connected or not given the STRING combined score value. The a_{ij} is defined by Eq. (3.3), however, it will only be considered $\theta < 150$. The outcome networks on cases with bigger θ values would not be connected networks, strongly harming the betweenness and closeness centrality values. This results on rewriting the a_{ij} formula for unweighted networks as in

$$a_{ij} = \begin{cases} 1, & \text{if } i \neq j \text{ and } combined\_score_{ij} \neq 0 \ (i,j \in \{1,2,...,n\}) \\ 0\,, & \text{otherwise} \end{cases} . \tag{4.1}$$

Considering the unweighted network, the degree, betweenness and closeness measures will be studied as a possible penalty factor.

Degree Centrality

The degree centrality calculation will be applied over this network, considering the Eq. (2.9). The degree distribution is presented in Figure 4.4 (a), being similar to the distribution of the weighted case, Figure 4.2, even though the weighted degree distribution decreases more smoothly.

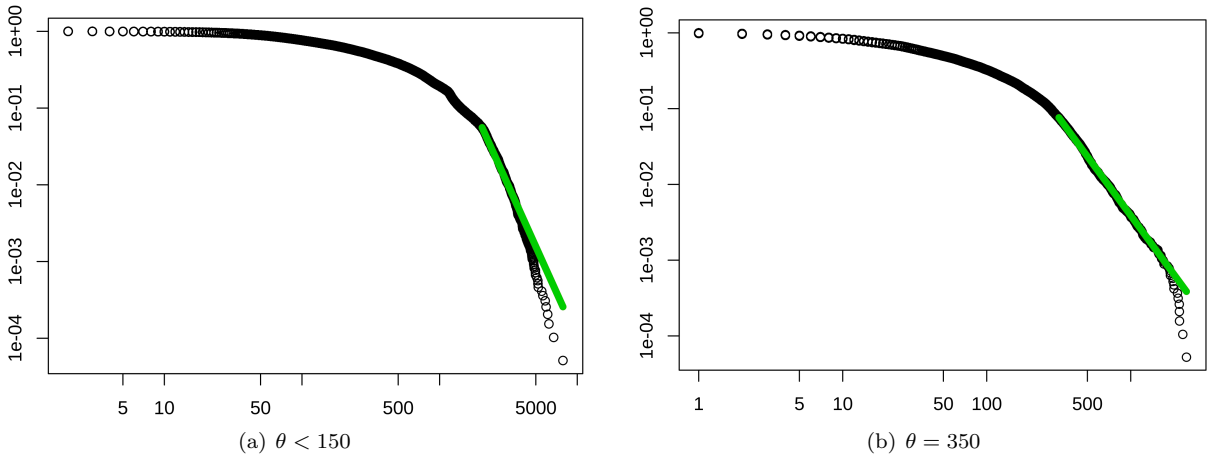


(a) $\theta < 150$ (b) $\theta = 350$

Figure 4.4: Unweighted degree distribution with $\log_{10}$ scale with different θ values.

Even though it is not considered as a penalty factor candidate, the degree distribution considering $\theta = 350$ has been obtained. Taking a closer look to Figure 4.4 (b), the maximum value of the degree vector is decreasing significantly with the increase of theta increases and the number of occurrences of lower degree values increase. Both graphs have been obtained and presented together to emphasise the scale-free topology presence, being even stronger when ignoring weaker connections.

That fact, along with both *plfit* resulting graphs, presented on Figure 4.5, corroborate with the network scale-free topology for high degree nodes. The scenario (a) resulted in $x_min = 2035$ and $\gamma = 4.48$ and the scenario (b) in $x_min = 317$ and $\gamma = 3.59$. These values mean that the network considering $\theta < 150$, Figure 4.5 (a), involves approximately 1087 nodes that respect the obtain power law function. On the case of $\theta < 150$, Figure 4.5 (b), this number increases to 1454 nodes. Given the presented results, it is clear that the θ strongly affect the network properties, being probably interesting to explore this relationship further. However, this goes beyond the scope of this project.



(a) $\theta < 150$ (b) $\theta = 350$

Figure 4.5: CDF plots regarding unweighted network degree, having different θ value. The greenline corresponds to the best powerlaw fit.

To evaluate if it is worth considering both weighted and unweighted degree, Table 4.1 has been

obtained. This table shows the top 15 proteins concerning weighted and unweighted degree ordered in descending order given the respective metric. In this table, a background colour was given to the protein entry if there is an intersection between both column proteins, meaning 12 out of the 15 presented proteins intersect. The tone of the table entries reflects the value of the weighted degree: the darker the blue, the higher the weighted degree. For that reason, on the left, it presented a perfect gradient among the proteins that intersected proteins. On the right, given the proteins are orders by the unweighted degree, the gradient is not so evident but is approximately kept. The different blue tone backgrounds are, therefore, used to emphasise the similarity between the metrics on the selected hubs.

Rank	*Weighted Degree*	*Unweighted Degree*
1^{st}	ENSP00000351686	ENSP00000351686
2^{nd}	ENSP00000229239	ENSP00000229239
3^{rd}	ENSP00000269305	ENSP00000295897
4^{th}	ENSP00000388107	ENSP00000350941
5^{th}	ENSP00000295897	ENSP00000309845
6^{th}	ENSP00000270202	ENSP00000270202
7^{th}	ENSP00000350941	ENSP00000314196
8^{th}	ENSP00000350052	ENSP00000269305
9^{th}	ENSP00000250971	ENSP00000250971
10^{th}	ENSP00000344818	ENSP00000272298
11^{th}	ENSP00000309845	ENSP00000263967
12^{th}	ENSP00000263967	ENSP00000263025
13^{th}	ENSP00000314196	ENSP00000298910
14^{th}	ENSP00000298910	ENSP00000349467
15^{th}	ENSP00000215832	ENSP00000215832

Table 4.1: Top 15 proteins regarding weighted and unweighted degree.

Furthermore, the plot unweighted degree vs weighted degree has been obtained, Figure 4.6, being clear that the relation between both variables is very strong. The two variables have a correlation value of 0.954, meaning that both would result on a very similar contribution from both centrality metrics. For that reason and given the relevance of other metrics considered on the unweighted network, the weighted network will not be considered.

Betweenness

The betweenness centrality is a very interesting and robust metric that reflex the amount of "flow" that passes through each node in the network. It has a strong relationship with the degree [39] and covers some specific cases that the degree alone cannot detect. In cases that nodes make the connection between big groups in the network, the degree value might be small. Nevertheless, the betweenness is going to have a high value because those nodes are the bridge between big groups, having a huge "flow" passing through them.

The betweenness distribution is presented in Figure 4.7, being clear that, just like the degree distribution, few nodes have high values of betweenness. These values genuinely stand out, being probably interesting to relate this metric values with the degree distribution and take some conclusions. If the

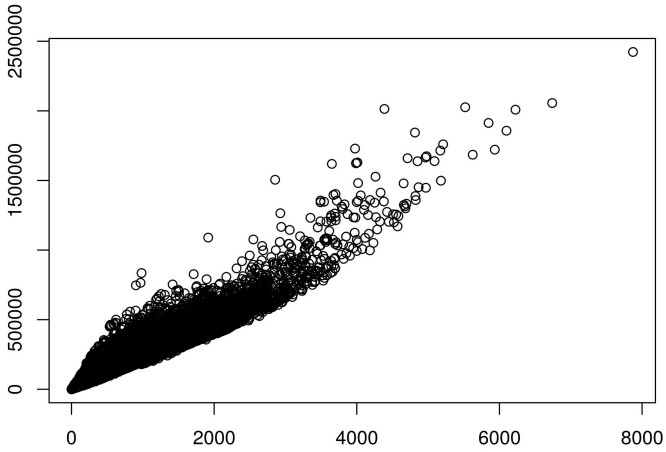


Figure 4.6: Unweighted degree vs weighted degree. The *x-axis* corresponds to the unweighted degree and *y-axis* to the weighted degree.

nodes considered hubs also have high values of betweenness, their influence in the network must be powerful. Therefore, the study between degree and betweenness metrics will be further analysed.

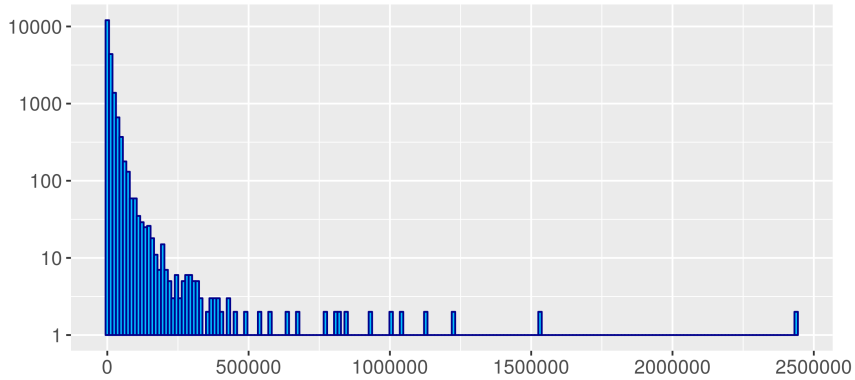


Figure 4.7: Betweenness distribution in $\log_{10}$ scale.

Just like the degree was considered a very relevant centrality metric for the regression coefficients in the Veríssimo et al. work [8], the betweenness will be considered here. This approach was never tested in this conditions, is one of the most exciting results to analyse in the Chapter 5 that covers the models' performance.

The leaves of the network result on a betweenness value equal to zero, because they are not present in any shortest path. All those nodes will not be considered because they strongly influence the metrics that consider both degree and betweenness measures. Ignoring these proteins do not consist into a problem because these nodes would be actively penalised on the final models given their relevance being nearly null considering all the centrality metrics. As the result of the application of this rule 71 nodes were removed, being considered 19503 proteins for now on.

Closeness

The closeness evaluates how adjacent a specific node is to all the nodes in the network. In Figure 4.8, the closeness distribution is presented in $\log_{10}$ scale and the respective distribution is very different from

the betweenness and degree.

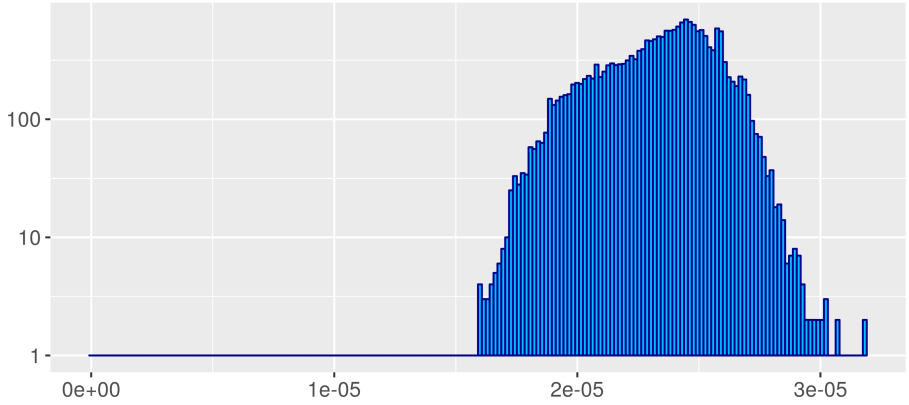


Figure 4.8: Closeness distribution in $\log_{10}$ scale.

Some nodes stand out regarding closeness values. However, the closeness distribution is very smooth when compared to the obtained considering degree and betweenness. Apriori, this result makes the closeness centrality less interesting to consider, given the previously obtained distributions.

At this point, to better understand the best metrics to use, a Venn Diagram was made on the 250 top proteins regarding each considered metric: degree, betweenness and closeness. This diagram is presented on Figure 4.9, showing that 168 proteins are present in all vectors, which seems to be a good indicator since they all measure the nodes centrality. However, note that nearly all closeness vector intersects the degree vector, looking like there is a strong relationship between these two metrics. The correlation between both metrics is 0.785 which indeed corresponds to a significant relationship between them, being relevant to exclude one of them. Knowing that the degree has already been studied over gene networks [8], the selected metric between this two was the degree. Furthermore, as presented, the degree distribution approximately follows the power law function, favouring only some few powerful nodes (hubs).

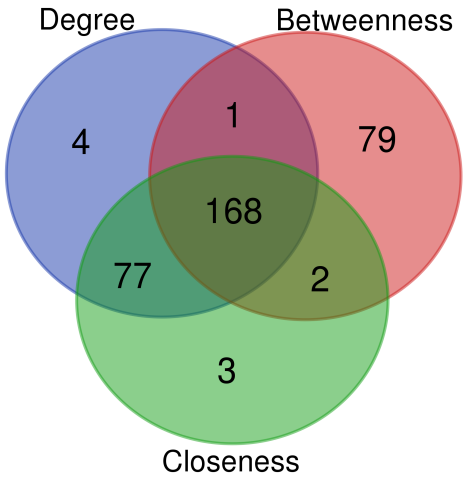


Figure 4.9: Venn diagram on the top 250 proteins regarding the degree, betweenness and closeness metrics.

Notice that the betweenness vector select many nodes that do not intersect any of the other vectors. That fact along with this metric similarity with the degree concerning distribution makes it an attractive

metric to consider a penalty factor.

Even though the relation is not as high as the one closeness and degree, the betweenness has a correlation value of 0.614 with the degree. For that reason and given 81 out of the top 250 proteins do not intersect, make it worth considering their relation to combine both strengths.

Betweenness vs Degree

In order to understand how degree and betweenness measures are related, the graph degree vs betweenness has been obtained after a re-scaling method applied to both metrics, Figure 4.10. From the analysis of this image, it can be verified that some nodes in the network stand out from the others regarding network relevance. The nodes that have high values of degree are usually associated with high values of betweenness.

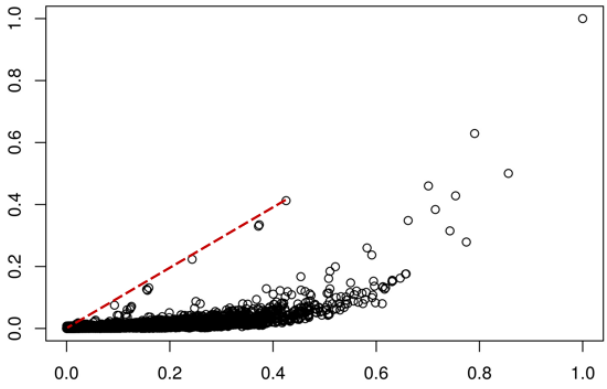


Figure 4.10: Re-scaled betweenness vs re-scaled unweighted degree. The *x-axis* correspond to the re-scaled unweighted Degree and *y-axis* to the re-scaled betweenness.

Nonetheless, in some cases, the node only stand out for one of these metrics, not being caught when considering the other metric even though it can be relevant. With that in mind, it is purposed the DBet distance metric can be established taking advantage of these relationships between degree and betweenness centrality and their impact on the network. This formula would consider both degree and betweenness through

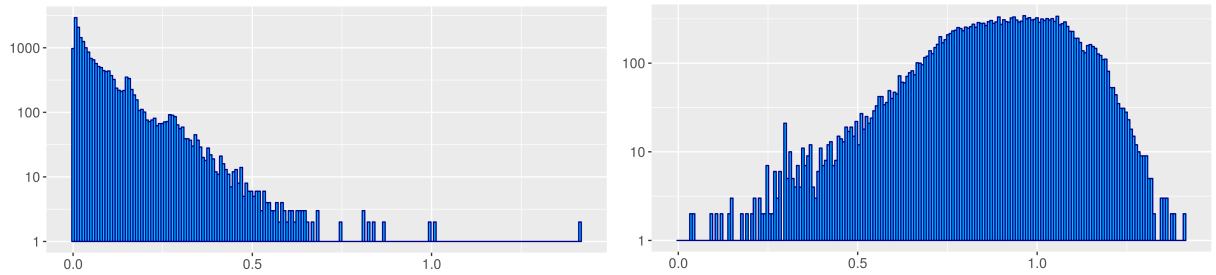
$$DBet = \sqrt{d'^2 + B'^2}, \tag{4.2}$$

with d' and B', respectively, corresponding to the re-scaled degree and betweenness. The re-scaling formula applied is given by

$$d' = \frac{d - min(d)}{max(d) - min(d)}, \tag{4.3}$$

where d corresponds to the degree centrality. The same re-scaled process is applied over betweenness centrality, B. From a geometric view, this corresponds to the distance to the origin focusing a specific node in Figure 4.11, illustrated with a red dotted line. The DBet metric is analysed because it benefits both metrics and will give the higher values to the nodes that emerge regarding both metrics.

The DBet distribution is presented in Figure 4.11 (a), having a shape that reflects both betweenness and degree distributions properties. By analysing the three distributions, it is possible to notice that the decrease in Figure 4.11 (a) is very similar to one presented in Figure 4.4 (a) and, at the same time, the gap between the nodes with high and low value is wider just like is can be seen in Figure 4.7. Nevertheless, by taking a closer look to the degree vs betweenness distribution on Figure 4.10, it is clear that the values of the degree have a much more significant impact on the DBet definition. The large magnitude of values on the betweenness metric make its' contribution close to zero for much of the points, risking not having the desired effect.



(a) DBet (b) DBet_{log}

Figure 4.11: DBet and DBet_{log} distance distribution.

Following the same line of thought, another metric is proposed by studying an exponential relation between both metrics. The respective graph of logarithmic degree distribution vs logarithmic degree after the re-scaled applied (Eq. (4.3)) is presented in Figure 4.12. The relation is stronger, having a correlation coefficient of 0.885. Again, the proposed method considers the distance of the respective node point in the graph to the origin, being defined by

$$DBet_{\text{log}} = \sqrt{\log(d')^2 + \log(B')^2}. \tag{4.4}$$

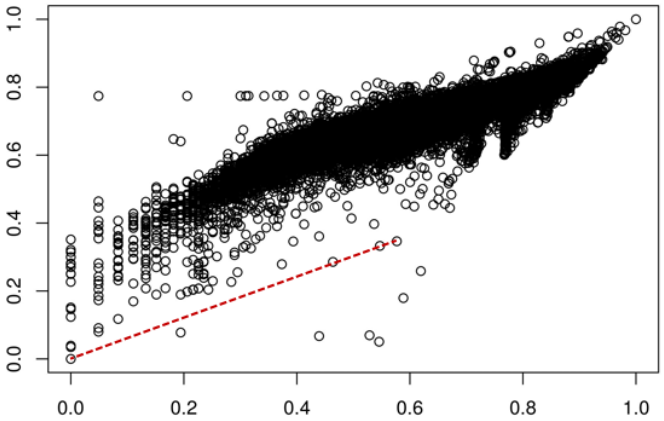


Figure 4.12: Logarithm of re-scaled betweenness vs logarithm of re-scaled unweighted degree. The *x-axis* corresponds to the logarithm of re-scaled unweighted degree and *y-axis* to the logarithm of re-scaled betweenness.

The distribution obtained with this new metric DBet_{log} is presented in Figure 4.11 (b). As it can be observed, the distribution is not so sharp as the betweenness and degree distributions, and that is because

logarithmic values are considered. Figure 4.13 presents a Venn Diagram focusing the top 250 proteins regarding the metrics degree, betweenness, DBet and $DBet_{log}$. From the analysis of this diagram, some fascinating facts are observed, being clear that the $DBet_{log}$ distance might be much more interesting to consider than the DBet. The latter has practically no intersections with the proteins selected considering the betweenness only. On the other hand, the $DBet_{log}$ looks much more interesting to analyse further given that is evenly distributed in terms of intersection with the degree and betweenness top 250 proteins. Apart from the intersection with both of them, 45 intersect only the degree top proteins, and 34 intersect only betweenness top proteins.

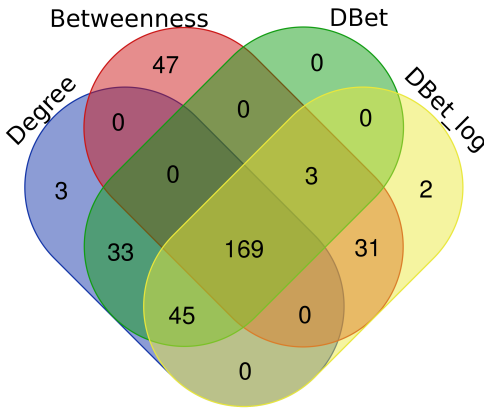


Figure 4.13: Venn Diagram on the top 250 proteins regarding the degree, betweenness, DBet and $DBet_{log}$ metrics.

Given the results presented, it can be concluded that the strongest candidates to use as a penalty factor are betweenness and degree. To support this two metrics, another two have been created based on them to consider both metrics into one. Cases like nodes that make the bridge between two big groups would have no expression if considering only degree, for instance. DBet and $DBet_{log}$ distances, therefore, are considered relevant candidates, that can be also be analysed concerning penalty factor distribution. However, considering Figure 2.4, it has been shown that the 247 intersect out of the 250 considered intersect the proteins selected by the degree metric. It can be concluded that the degree values strongly influence the DBet values.

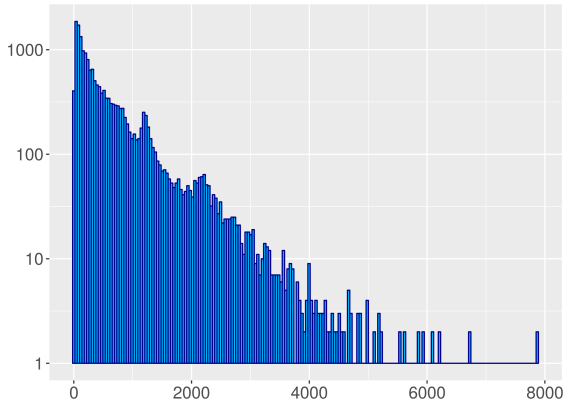
Given the DBet distribution properties, the metrics that will be considered as penalty factor are degree, betweenness and $DBet_{log}$ distance. Those metrics were directly applied on the original STRING network because the mapping protein-gene would result in an incomplete network with less information. With this first analyses, a connected biological network was considered and, a selection over the centrality metrics has been to determine the proteins relevance.

4.2 Mapping and Penalty Factor

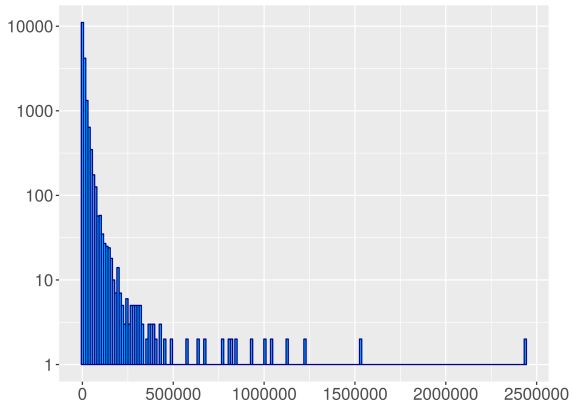
The bridge between proteins and genes needs to be crossed. In order to do that, there is the *BioMart* package, a powerful link between biological databases and microarray data analysis [53, 54]. Unfortunately, some proteins and genes need better documentation so that all the relations can be established.

As a result, some of the proteins will not have a gene associated, and some proteins can have the same gene associated. Focusing on the first point, from the 19503 proteins considered on the STRING network, 18241 have a gene associated. From those, 70 have problems related to the association of the same gene to more than on protein: 32 genes with two proteins associated and two genes with three proteins. After the pre-process over the data, already explained in Section 3, the mapping is performed, having 18205 genes' centrality metrics.

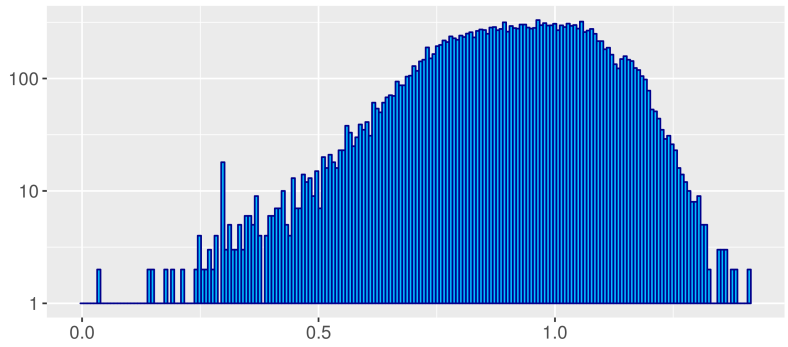
From Figure 4.14, it can be seen that most of the data from the network are preserved. The number of resulting genes is very significant and, the usage of the protein interaction STRING network means that those are protein-mapping genes. The respective centrality metrics distributions are presented on Figure 4.14 and, as expected, they are very similar to the one considered on proteins because most of the relations were one-to-one.



(a) Degree



(b) Betweenness



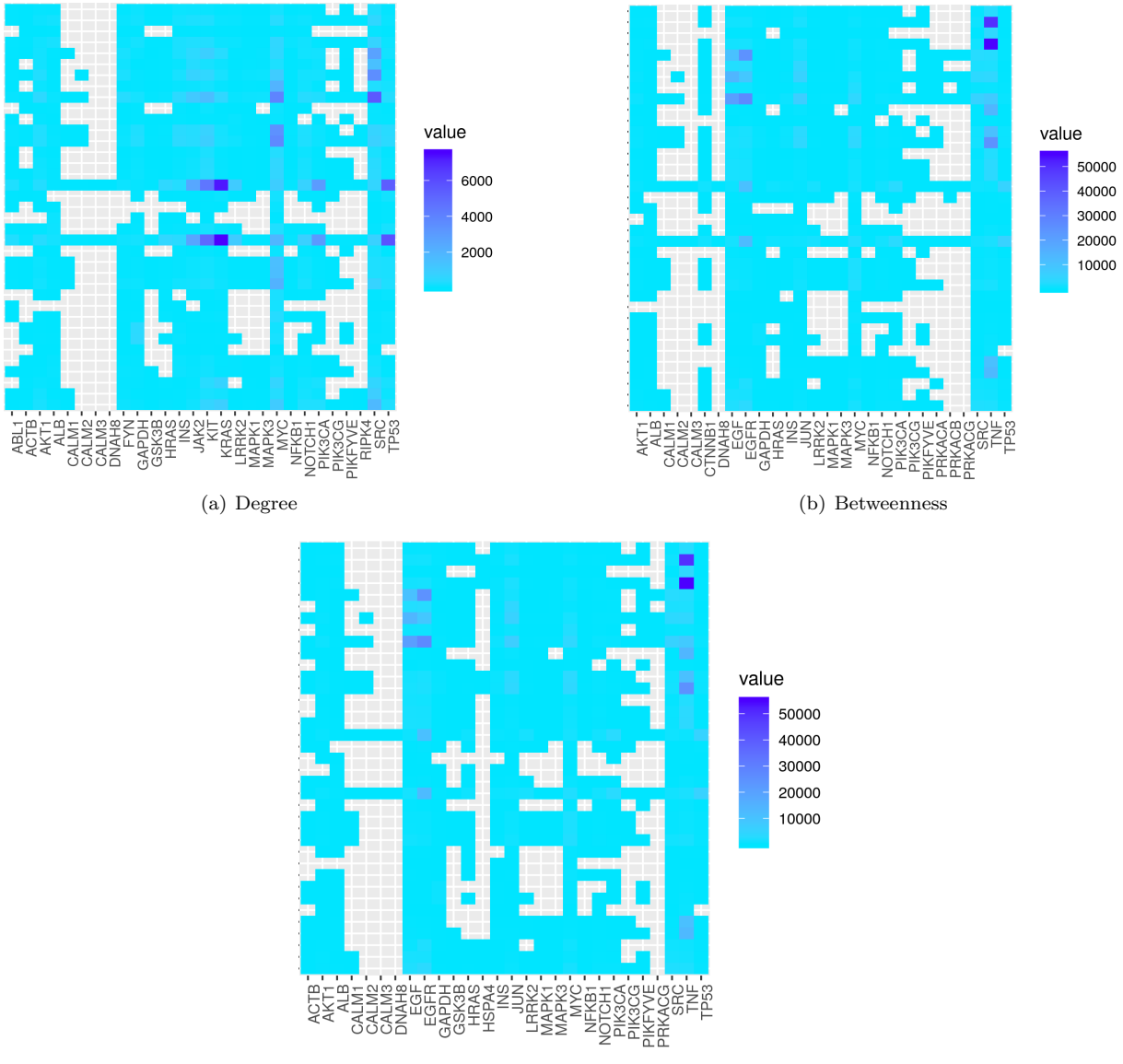
(c) DBet_{log} distance

Figure 4.14: Centrality measures distribution regarding genes.

In order to validate the importance of the genes concerning the presence in cancer-based investigations, the Cancer Hallmarks Analytics Tool (CHAT) has been used [62]. This approach considers text mining techniques over cancer-related references from PubMed. The presented tool gives a total number of hits/presences of the considered gene on those texts, organising in according to 37 different channels. The genes with a high number of hits are considered to be strongly related to cancer. To obtain these metrics values, the *glmSparseNet* package [63] will be used. The usage of the *hallmaks* function returns a heat-map with hallmarks counts per gene: the darker the blue colour, the higher the number hallmarks associated to that gene.

Note that this validation metric is very useful, but not absolutely reliable. The fact that it is based on the presence of genes names in article abstracts or bodies might be fallacious because the reference of a gene on those works does not directly mean it is associated with the cancer under study in the article. However, multiple hits of a gene in many different texts regarding different channels, give a significant level of confidence that the gene is relevant to consider when studying cancer problems.

Even though this is a small sample of genes, the heat-maps for the top 30 genes considering the three different metrics is presented in Figure 4.15. To understand how many genes obtained from the STRING dataset have hallmarks associated, all the 18205 genes so far considered were run through the CHAT, been concluded that approximately $36,63\%$ had no hallmarks registered.



(a) Degree

(b) Betweenness

(c) $\text{DBet}_{\log}$ distance

Figure 4.15: Heatmap over the 30 top genes selected considering the degree, betweenness and $\text{DBet}_{\log}$ metrics.

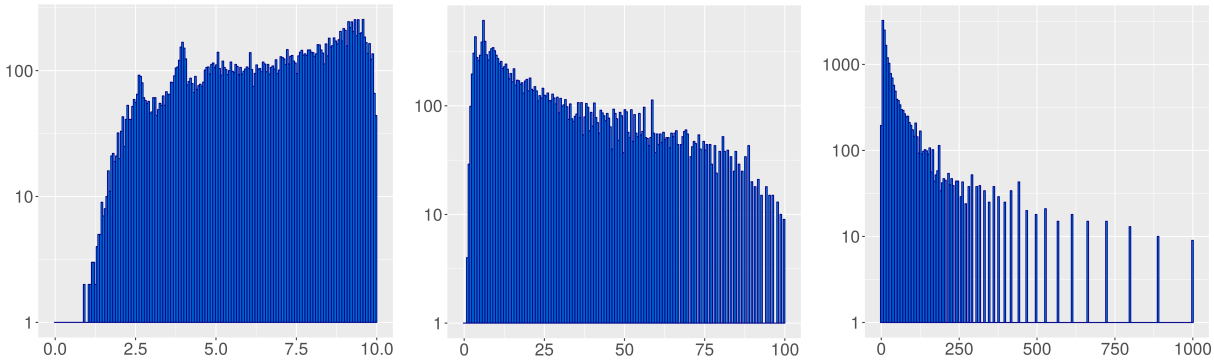
By analysing the obtained heat-maps, it can be concluded that all these centrality metrics allow us to get attractive genes regarding hallmark hits since 28 out of top 30 have hallmarks associated. Not

only that but must of the genes appear in multiple channels, and some of them have very high values. The top number of hits considering the betweenness and DBet_{log} is approximately 50000 and considering the degree is 7000. These results favour the use of the STRING information to work over oncology investigation. Moreover, the usage of betweenness and degree centrality seem to allow an engaging extraction of the involved entities relevance.

With all the pre-processing applied so far, it is now possible to directly work over gene expression datasets. Moreover, with the centrality measures selection and respective values, the importance of the genes is defined, and it can be shown the penalty factor distribution depending on μ. As stated, the μ variable is responsible for controlling the level of influence of penalty factor, defining the max penalty factor value: the lower the μ, the higher the max penalty value. In Figures 4.16, 4.17 and 4.18, the penalty distribution is presented for the degree, betweenness and DBet_{log} metrics, respectively, considering μ equal to 0.1, 0.01 and 0.001.

It is clear that the μ value significantly influence this penalty factor values, being essential to consider the different values of μ on the models' training and testing. As the μ value influences this distribution so does the type of metric used.

By looking at Figure 4.16, it can be verified that low values of μ reflect the desired distribution, having only some few nodes that are little penalised. As the μ value increase, this distribution gets smoother, and the penalty is higher, enhancing more sparse solution. Note that for low values of μ, there are some isolated bins with on the high penalty values bins. This phenomenon is verified because the degree is a discrete measure.

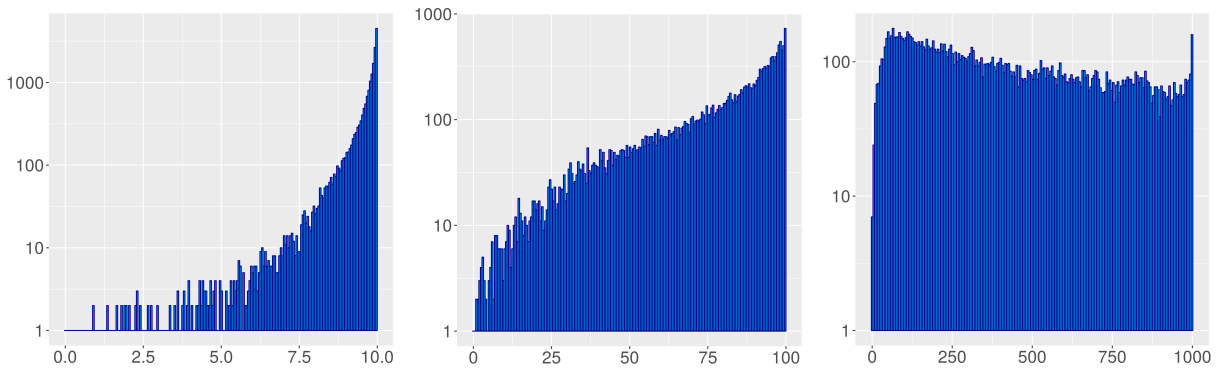


(a) $\mu = 0.1$ (b) $\mu = 0.01$ (c) $\mu = 0.001$

Figure 4.16: Penalty factor considering degree centrality and different μ values.

Focusing on the betweenness scenario, the penalty factor distribution has an exciting behaviour as μ increases. In the case of μ equal to 0.1, presented in Figure 4.17 (a), the number of nodes that have lower values of penalisation is relatively small. Focusing in Figure 4.17 (c), it is clear that the distribution of the penalty factor is much constant. These distribution properties make this penalty vector very interesting to consider if the betweenness gives genes that are indeed relevant to the faced problem.

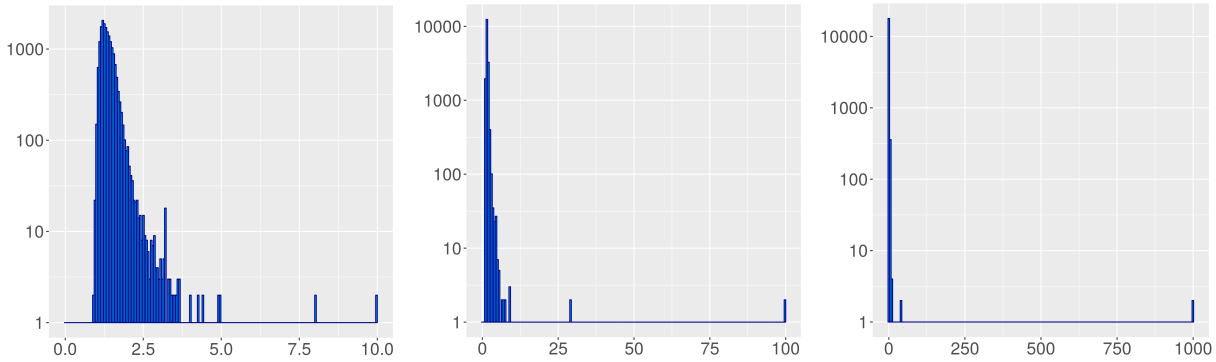
The DBet_{log} penalty distributions, presented in Figures 4.18 are not as significant as the ones obtained with the degree and betweenness. This is probably because there are only some few genes that have a significantly lower value given the range of values of this metric. These few nodes lead have high



(a) $\mu = 0.1$ (b) $\mu = 0.01$ (c) $\mu = 0.001$

Figure 4.17: Penalty factor considering betweenness centrality and different μ values.

penalisation values associated to them, and all the other nodes are inserted in a small range of penalisation values. This distribution is less evident when considering lower μ values, being possible to get model worth considering while focusing on this metric.



(a) $\mu = 0.1$ (b) $\mu = 0.01$ (c) $\mu = 0.001$

Figure 4.18: Penalty factor considering $\text{DBet}_{\log}$ distance and different μ values.

Now that the centrality metrics have been studied and rejected/accepted, it is possible to represent with the diagram presented in Figure 4.19. In the following chapter, the study over the obtained regression coefficients will be performed along with the results discussion.

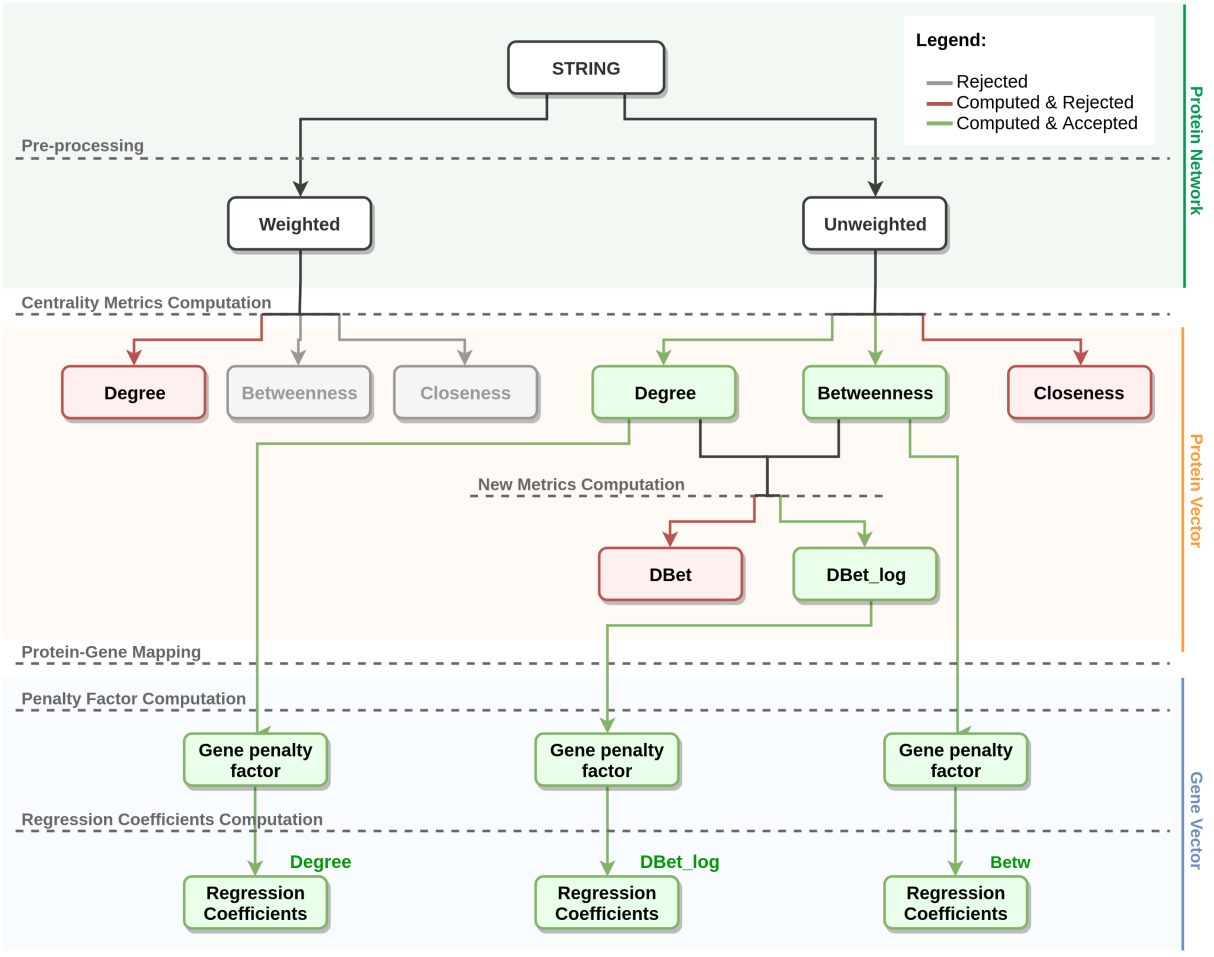


Figure 4.19: Resume of the applied methodology to get to the final metrics.

Chapter 5

Results

Now that the pipeline to get the penalty factor is presented, the regression method under study will be applied on real data from the TCGA to obtain the respective Cox models. The dataset overview, validation metrics and results are exposed in the following sections.

5.1 Breast Cancer Dataset

The Breast Invasive Carcinoma (BRCA) dataset from TCGA will be used to test the presented method focusing the different centrality metrics. The used data is survival type, displaying the gene expression levels and clinical data from patient follow-up. It also comprises a variable that indicates whether the event of interest was verified or censored, being this event the death of the subject.

The BRCA data in analysis involves information from 1036 individuals and 55882 genes, from which only 19868 were considered because they are protein-coding genes according to the Consensus Coding Sequence and Ensembl databases [55, 64]. The gene expression levels of these patients are presented in Fragments Per Kilobase Million (FPKM) expression units.

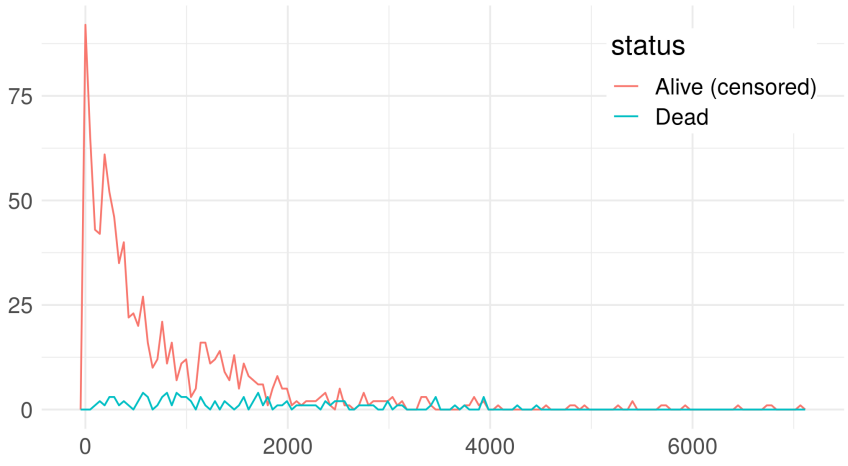


Figure 5.1: Distribution of events over time considering survival and censored times.

In Figure 5.1 it can be observed the number of censored cases (red) and events (blue) per interval

of time. The number of censored observations outcomes the number of events, yet, as presented, this information is still beneficial for the construction of Cox PH models. The dataset is composed of 932 censored observations and 104 events.

Before using the dataset, a selection of the genes was necessary. There were 1783 genes from the original dataset and 118 genes from the STRING that were not considered because they did not intersect. Since only the intersected genes were considered, this resulted in a total of 18085 genes that are going to be used on the construction of the final models. The loss of information was not substantial given the high number of intersected genes, resulting in more than 90% of the protein-coding genes that were kept. However, the dimensionality curse is still evident given that the number of variables is above an order of magnitude from the number of observations.

5.2 Validation Metrics

To validate the presented method, the Elastic Net model will be handled as the baseline model. The usage of the Ridge regularisation would not be interesting to consider because the objective is to obtain a clear solution with only some non-zero regression coefficients. Models close to the LASSO (high α values) were considered. However, the number of selected genes is too small for an interesting analysis of the models' performance. For that reason, most of the selected α values, considering the Eq. (3.1) at page 23, are going to be low to have more non-zero coefficients in the final models and better understand the penalty factor influence on the results. Besides the different penalty factors and α values analysed, the models are going to be computed based on different train/test splits.

To facilitate the analysis of the models in this chapter, a name has been associated to each model type according to the used penalisation factor. This information is summarised in Table 5.1.

Model	**Penalisation Associated**
ENet	Null
Degree	Based on the degree centrality
Betw	Based on the betweenness centrality
DBet_{log}	Based on the exponential relation between degree and betweenness centrality

Table 5.1: Model types according to the penalisation associated.

The respective models' analysis will focus on two different points of view. First the analysis of the models' performance and then over the significance of the selected genes. The former were analysed based on:

- Kaplan-Meier curves along the log-rank test – p-value;
- Concordance c-index.

5.2.1 Analysis of the Models Performance

The Kaplan-Meier [20] curves were already presented, and an example has been exposed in the Section 2.1. Their usage is for the analysis of the survival curves for groups with different characteristics

being used the log rank test to validate the separation between the obtained curves, which corresponds to the p-value computation. In this scenario, however, the groups are not determined yet. The insertion of the new observation into a low or high group risk will be performed with the obtained Cox regression. This insertion is achieved by first computing the hazard relative risk value of each of the new observations. After that, the median is computed, and the individuals with a lower value than the median are inserted in the low-risk group, and the others are inserted in the high-risk group. Then, the low and high risk curves are obtained and the better the separation between the curves, the better the models.

Simultaneously, it will be considered the concordance c-index [65], a measure that contemplates all permissible pairs of individuals under study and compares if their survival time is aligned with the hazard relative risk. Pairs where both individuals are censored or when only one is censored and has a shorter time than the uncensored are not considered valid. The concordance is increased by 1 with every pair that is inserted in one of these cases:

- Individual with higher risk has shorter survival time;
- Hazard risks and survival time are the same;
- One individual is censored and has a lower risk.

If that is not the case, the concordance is only increased by 0.5. The c-index value is finally obtained by dividing the count by the number of permissible pairs, meaning that the closer c-index is to 1, the better the model is.

5.2.2 Analysis of the Selected Genes Significance

To further understand the performance of each of the models, the analysis over the selected genes has also been implemented. The first step was to investigate the intersection of the genes selected by the different models, by the usage of Venn Diagrams. These diagrams were used to understand which of the genes had a big impact since the more relevant ones are likely to be selected by most of the considered models. After that, the selected genes will be crossed with CHAT, which have been explained in the final part of Chapter 4.1. With CHAT information it was possible to understand the percentage of genes that were selected by the models that are present in literature related to oncology investigations.

5.3 Selected Regression Coefficients Analysis

Given the problem in hands, many variables need to be explored to find the best model to predict whether a person belongs to a high-risk group or not. The variables that have to be defined for the model are the train/split ratio, α and μ, and tuning these parameters is essential to achieve the best regression coefficients values.

To validate the model and understand its performance as more train data is given, three different split ratios of data were used: 70%, 75% and 80%. Another important parameter to be considered is the penalty factor v_i, the one which influence was intensely analysed, given that it is the major contribution

of this work. As presented, this vector depends on the focused centrality measure and the μ value, being tested with six different values of the latest: 0.5, 0.1, 0.05, 0.01, 0.005 and 0.001. The last variable to define has a great impact on the considered regularisation formula present in Eq. (3.1): the α value. As stated, it is responsible for controlling the balance between the usage of L_1 and L_2 norms, varying from a non-sparse (α approximately 0) to sparse solution (α approximately 1). Four different α values were tested: 0.05, 0.1, 0.15 and 0.2. All these α values combinations with different μ and train/test ratio were considered to understand the models' performance variation with these parameters and find the best models.

To avoid an over-fited model and understand which is the best λ value, presented on the Eq. (3.1) at page 23, it will be taken into a 10-fold cross-validation. For each of the folds, 1000 λ values are considered and tested with cross-validation and the best λ is selected based on the minimum log-likelihood deviation considering all the different folds. The usage of cross-validation allows a λ selection substantiated on different splits on the train data, avoiding over-fitting the models since more hypotheses are considered.

All the models resulting from the different combinations of parameters followed the presented process, being trained and tested with BRCA dataset already introduced in Section 5.1. To validate the models' performance, were computed the number of genes selected, the c-index and the p-value for each of the parameters combination. All the obtained results are presented in Table 5.2. Nevertheless, for most of the results discussion, it was considered greater μ values since lower values resulted too sparse solutions to take relevant conclusions.

Focusing on the train/test split, it was concluded that the 0.8 ratio value results on models that stand out negatively. All of the resulting models are not statistically significant given the high values of the p-value. Nevertheless, within the cases with train/test split equal to 0.8, it can be observed that the Degree and Betw models have a more significant separation between the curves for μ equal to 0.05 and 0.01. Even though the values are not substantial, they prove that the usage of these penalty factors leads to a more general model that can better avoid overfit scenarios. It is also interesting to verify that the models that show the best results are the ones that considered train/test split equal to 0.7. This fact might be verified because the considered separation naturally favour the majority of the models or because the models start overfitting when higher values of this parameter are used.

Considering the μ values influence on the models' results, it is possible to verify that they significantly impact on the metrics having an essential rule on the number of selected variables. However, for high values of this parameter, the results are not so different than the ones obtained with the Elastic Net. The penalty vector on these scenarios varies between 0.66(6) and 2 for μ equal to 0.5 and 0.9090(90) and 10 for μ equal to 0.1, not having nodes with a great penalisation. As the μ decreases, the range of values of the penalisation vector significantly increases and some exciting models start to stand out, even though the number of considered regression coefficients is smaller. In general, it was observed that this decrease is not harming the performance of the models, leading in fact to some scenarios with better results.

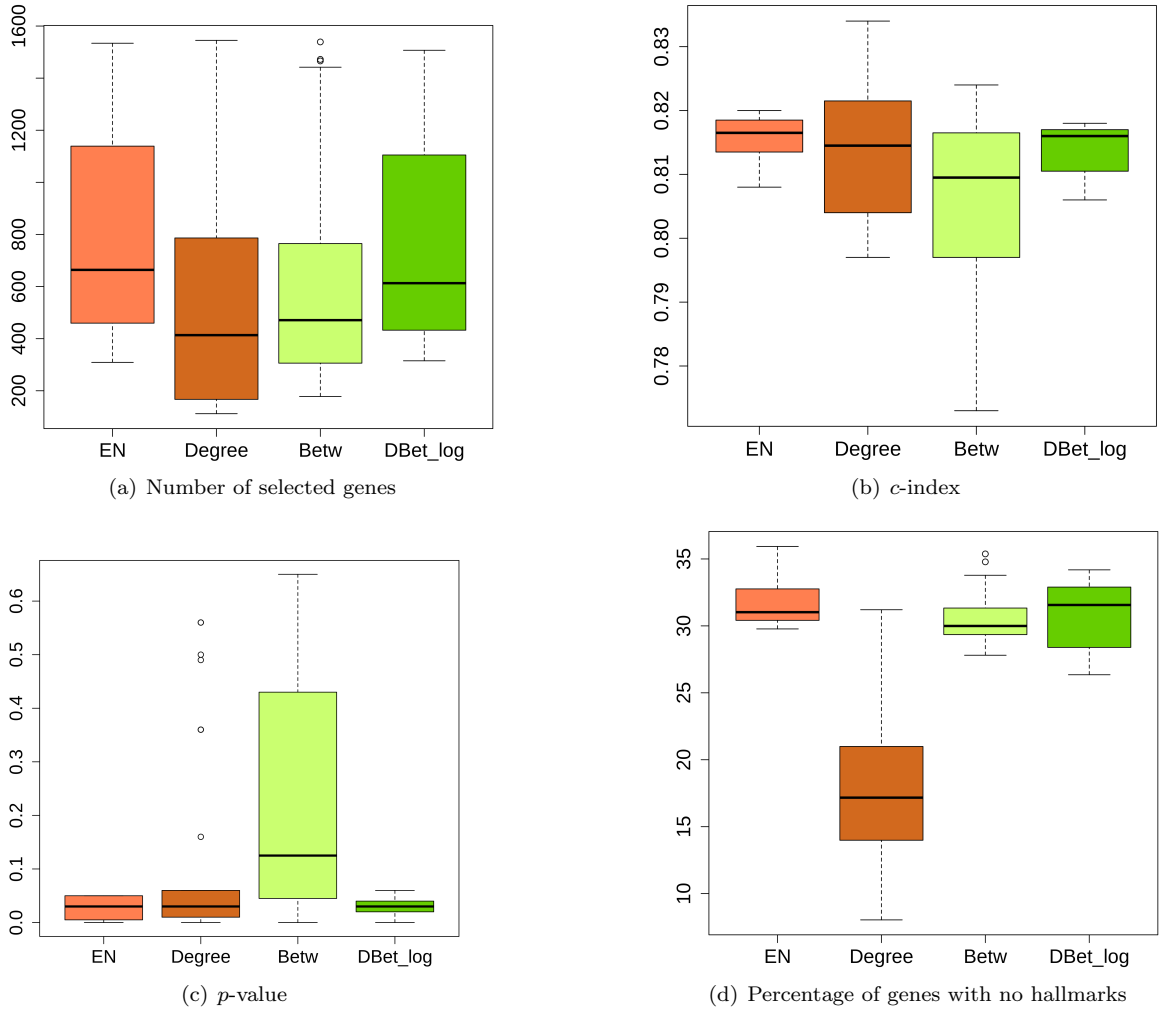
The ENet models have a good performance for all the scenarios except on the case of the train/test split equal to 0.8. However, the number of variables selected is considerably higher than the models in a study, which is "unfair" and may mean that the chosen genes are not so significant.

μ	train	α	# of genes selected				p-value (%)				c-index				No hallmarks genes (%)			
			ENet	Deg.	Betw	DBet	ENet	Deg.	Betw	DBet	ENet	Deg.	Betw	DBet	ENet	Deg.	Betw	DBet
0.5	0.7	0.05	1534	1545	1539	1507	0.37	0.47	0.37	0.29	0.819	0.819	0.820	0.818	31.16	28.16	31.38	31.79
		0.1	**859**	805	860	770	**0.36**	2.81	0.41	1.54	**0.820**	0.820	0.820	0.818	**32.13**	28.32	31.98	32.73
		0.15	551	575	548	561	0.85	1.11	0.86	2.45	0.817	0.820	0.817	0.817	33.39	29.91	33.39	33.16
		0.2	423	423	424	430	3.09	2.93	3.13	4.78	0.815	0.820	0.815	0.815	35.93	31.21	35.38	34.19
	0.75	0.05	1419	1373	1426	1492	3.16	3.65	3.13	4.22	0.818	0.814	0.817	0.817	30.59	27.75	30.65	31.50
		0.1	777	578	776	718	4.53	3.20	4.47	4.19	0.816	0.810	0.816	0.814	30.89	23.70	31.19	29.39
		0.15	496	294	491	475	4.53	3.09	4.51	4.30	0.812	0.805	0.813	0.810	30.24	20.07	29.94	28.21
		0.2	309	220	301	315	4.51	2.25	4.74	3.24	0.808	0.805	0.807	0.806	29.77	21.36	29.57	26.35
	0.8	0.05	1231	1140	1244	1219	71.59	71.59	72.77	71.59	0.812	0.811	0.811	0.809	30.63	25.97	31.03	30.44
		0.1	613	583	624	666	71.59	71.59	72.77	72.77	0.817	0.816	0.816	0.813	29.04	24.01	29.17	30.48
		0.15	439	393	446	421	72.77	73.68	72.77	71.59	0.817	0.820	0.815	0.810	29.39	21.12	28.48	28.74
		0.2	322	302	318	287	72.01	73.68	70.82	92.98	0.816	0.816	0.812	0.808	29.81	23.51	28.30	26.48
0.1	0.7	0.05	1534	1334	1466	1468	0.37	1.68	0.32	1.13	0.819	0.815	0.821	0.816	31.16	22.94	31.51	31.95
		0.1	859	725	808	763	0.36	0.28	4.33	2.36	0.820	0.819	0.823	0.817	32.13	21.24	32.30	32.63
		0.15	551	463	595	557	0.85	0.35	9.18	2.43	0.817	0.820	0.820	0.818	33.39	19.22	33.78	33.39
		0.2	423	192	437	410	3.09	1.37	11.70	4.78	0.815	0.827	0.821	0.817	35.93	16.15	34.78	33.66
	0.75	0.05	1419	983	1472	1480	3.16	3.56	3.18	5.20	0.818	0.801	0.816	0.817	30.59	22.28	31.18	31.08
		0.1	777	486	753	675	4.53	4.25	4.49	3.42	0.816	0.802	0.811	0.812	30.89	17.28	30.15	28.89
		0.15	496	192	482	438	4.53	16.14	4.49	3.29	0.812	0.807	0.808	0.809	30.24	13.54	28.83	27.17
		0.2	309	115	343	326	4.51	49.93	4.65	3.58	0.808	0.806	0.806	0.808	29.77	11.30	28.28	26.99
	0.8	0.05	1231	1074	1166	1223	71.59	73.68	72.77	70.82	0.812	0.807	0.808	0.807	30.63	20.67	31.39	30.09
		0.1	613	367	623	633	71.59	26.60	72.77	71.59	0.817	0.800	0.810	0.810	29.04	14.99	29.21	30.65
		0.15	439	244	470	443	72.77	21.05	44.57	70.82	0.817	0.803	0.810	0.812	29.39	12.71	28.94	28.67
		0.2	322	111	335	309	72.01	31.19	44.57	92.16	0.816	0.810	0.808	0.811	29.81	12.61	26.57	29.45
0.05	0.7	0.05	1534	1206	**1413**	1455	0.37	1.97	**0.47**	1.13	0.819	0.817	**0.821**	0.816	31.16	21.48	**31.28**	31.89
		0.1	859	559	763	766	0.36	0.76	6.98	2.36	0.820	0.821	0.823	0.817	32.13	19.14	32.90	32.64
		0.15	551	275	553	541	0.85	1.21	12.75	2.70	0.817	0.824	0.821	0.817	33.39	17.46	32.73	32.90
		0.2	423	178	417	401	3.09	0.63	13.04	2.74	0.815	0.832	0.824	0.817	35.93	14.61	32.85	33.92
	0.75	0.05	1419	934	1442	1474	3.16	3.91	4.82	5.52	0.818	0.797	0.813	0.816	30.59	21.09	30.86	31.00
		0.1	777	413	720	666	4.53	3.56	5.82	3.44	0.816	0.800	0.808	0.811	30.89	15.98	30.14	28.38
		0.15	496	152	445	440	4.53	35.77	5.60	3.27	0.812	0.804	0.803	0.809	30.24	11.84	28.09	28.41
		0.2	309	115	355	326	4.51	48.61	10.56	3.58	0.808	0.804	0.805	0.808	29.77	10.44	28.73	27.30
	0.8	0.05	1231	930	1164	1206	71.59	41.04	72.77	70.82	0.812	0.804	0.805	0.808	30.63	19.89	30.50	30.18
		0.1	613	336	643	630	71.59	21.05	44.57	71.59	0.817	0.800	0.807	0.810	29.04	13.69	30.02	30.48
		0.15	439	158	442	451	72.77	10.42	64.60	70.82	0.817	0.810	0.811	0.811	29.39	12.03	28.96	29.05
		0.2	322	108	330	295	72.01	21.35	65.34	92.16	0.816	0.811	0.813	0.810	29.81	12.96	27.88	28.48
0.01	0.7	0.05	1534	1067	910	1442	0.37	2.17	12.40	1.13	0.819	0.817	0.815	0.816	31.16	20.62	31.54	31.62
		0.1	859	462	411	**758**	0.36	0.76	13.61	**2.36**	0.820	0.822	0.812	**0.818**	32.13	18.18	30.41	**32.85**
		0.15	551	233	294	552	0.85	0.37	21.27	2.70	0.817	0.830	0.810	0.817	33.39	15.02	29.93	33.51
		0.2	423	**144**	218	396	3.09	**0.63**	15.42	2.74	0.815	**0.834**	0.809	0.816	35.93	**13.89**	29.36	34.09
	0.75	0.05	1419	798	897	1488	3.16	4.29	13.41	5.52	0.818	0.797	0.791	0.816	30.59	20.18	29.77	30.91
		0.1	777	414	501	665	4.53	4.73	46.68	3.44	0.816	0.800	0.789	0.812	30.89	15.46	29.34	28.57
		0.15	496	155	367	435	4.53	35.77	64.27	3.29	0.812	0.803	0.787	0.809	30.24	10.97	29.97	28.05
		0.2	309	112	295	322	4.51	55.68	60.83	3.58	0.808	0.804	0.781	0.808	29.77	8.04	30.17	27.64
	0.8	0.05	1231	794	886	1212	71.59	38.70	29.69	72.91	0.812	0.801	0.800	0.808	30.63	18.51	29.46	30.36
		0.1	613	270	435	636	71.59	9.40	29.69	71.59	0.817	0.807	0.805	0.810	29.04	11.85	28.74	30.03
		0.15	439	167	317	450	72.77	9.40	52.36	70.82	0.817	0.808	0.807	0.811	29.39	11.38	28.08	29.33
		0.2	322	109	250	298	72.01	20.34	52.36	92.16	0.816	0.808	0.804	0.810	29.81	12.84	27.20	28.86
0.005	0.7	0.05	1534	1034	683	1440	0.37	2.17	13.65	1.13	0.819	0.817	0.813	0.816	31.16	20.89	29.87	31.67
		0.1	859	440	338	760	0.36	2.28	11.44	2.36	0.820	0.822	0.812	0.818	32.13	17.05	29.88	32.90
		0.15	551	234	233	552	0.85	0.37	28.46	2.70	0.817	0.830	0.809	0.817	33.39	14.53	28.76	33.51
		0.2	423	143	187	390	3.09	0.63	36.61	2.74	0.815	0.833	0.807	0.816	35.93	13.99	27.81	33.85
	0.75	0.05	1419	792	767	1485	3.16	5.60	46.68	5.52	0.818	0.797	0.787	0.816	30.59	20.33	29.99	30.98
		0.1	777	313	460	673	4.53	5.87	64.53	3.44	0.816	0.799	0.784	0.812	30.89	14.38	30.00	28.83
		0.15	496	155	311	437	4.53	35.77	61.43	3.62	0.812	0.804	0.777	0.809	30.24	10.97	30.23	28.15
		0.2	309	112	251	326	4.51	55.68	64.53	3.58	0.808	0.804	0.776	0.808	29.77	8.04	28.69	27.91
	0.8	0.05	1231	795	738	1208	71.59	38.70	27.78	72.91	0.812	0.801	0.803	0.808	30.63	18.62	29.68	30.30
		0.1	613	272	390	639	71.59	9.40	27.78	71.59	0.817	0.807	0.806	0.810	29.04	11.77	29.23	30.20
		0.15	439	167	286	448	72.77	9.40	50.62	70.82	0.817	0.808	0.807	0.811	29.39	11.38	26.92	29.46
		0.2	322	110	200	297	72.01	20.34	56.51	92.16	0.816	0.808	0.803	0.810	29.81	12.73	28.50	28.96
0.001	0.7	0.05	1534	1013	574	1441	0.37	2.17	12.97	1.13	0.819	0.817	0.812	0.816	31.16	20.83	29.97	31.65
		0.1	859	434	279	751	0.36	0.88	48.10	2.36	0.820	0.822	0.809	0.817	32.13	17.05	28.67	32.89
		0.15	551	233	208	551	0.85	0.37	45.46	2.70	0.817	0.829	0.806	0.817	33.39	14.59	28.37	33.39
		0.2	423	143	178	385	3.09	0.63	27.54	2.70	0.815	0.833	0.806	0.816	35.93	13.99	29.21	34.03
	0.75	0.05	1419	781	669	1482	3.16	5.60	41.12	5.52	0.818	0.797	0.781	0.816	30.59	19.85	31.09	30.97
		0.1	777	307	378	672	4.53	5.87	50.28	3.44	0.816	0.800	0.777	0.812	30.89	14.33	29.10	28.87
		0.15	496	156	272	436	4.53	35.77	50.16	3.62	0.812	0.804	0.775	0.809	30.24	10.26	29.78	27.98
		0.2	309	112	225	326	4.51	55.68	64.39	3.58	0.808	0.805	0.773	0.808	29.77	8.04	29.78	27.91
	0.8	0.05	1231	804	624	1204	71.59	38.70	54.74	72.91	0.812	0.802	0.806	0.808	30.63	18.53	30.61	30.40
		0.1	613	270	357	641	71.59	9.40	52.42	71.59	0.817	0.807	0.807	0.810	29.04	11.85	29.13	30.27
		0.15	439	166	238	449	72.77	12.89	33.10	70.82	0.817	0.809	0.803	0.811	29.39	11.45	26.89	29.40
		0.2	322	108	175	301	72.01	20.34	36.25	92.16	0.816	0.810	0.802	0.810	29.81	12.04	27.43	28.90

Table 5.2: Models number of selected genes, p-value and c-index with different values of train/test split, α an μ. The rows shadowed in light blue correspond to the best models considering each of the model types. The ENet model has $train/testsplit = 0.7$ and $\alpha = 0.1$, the Degree model has $train/test\,split = 0.7$, $\mu = 0.01$ and $\alpha = 0.2$, the Betw model has $train/test\,split = 0.7$, $\mu = 0.05$ and $\alpha = 0.05$ and the $\text{DBet}_{\log}$ model has $train/test\,split = 0.7$, μ equal 0.01 and $\alpha = 0.1$. The resulting values are highlighted with bold font. The rows with the green background correspond to the considered models for the boxplots presented in Figure 5.2.

Another important metric that was measured for all the selected models is the percentage of the non zero regression coefficients with no hallmark hits, also presented in Table 5.2. This value was obtained with the usage of the CHAT, a tool already presented and used in Section 4.2. Regarding all the features used in the train of the models, it has been observed that 36.46% of the examined genes have at least a hallmark associated.

To have a better understanding of the influence of the centrality type on the models' performance, the boxplots with whiskers with maximum 1.5 interquartile range have been obtained for all of the studied metrics, Figure 5.2. However, the models that consider a train/test split equal to 0.8 were not taken into account because they strongly harmed all the models' performance under analysis. The used models for this analysis have their parameters shadowed in green in Table 5.2.



(a) Number of selected genes

(b) c-index

(c) p-value

(d) Percentage of genes with no hallmarks

Figure 5.2: Boxplots with whiskers with maximum 1.5 interquartile range focusing on the number of genes selected, c-index, p-value, and percentage of genes with no hallmarks.

Regarding the number of genes selected by the models presented on Figure5.2 (a), it was verified, as expected, that fewer variables are selected for the cases that include the penalty factor. The Degree and Betw models stand out as they select fewer coefficient regressions. On the case of the DBet_{log}, the number of genes chosen boxplot is similar to the one obtained with the ENet models, likely because the penalisation for most of the genes is small according to Figure 4.18 at page 43.

With respect to the c-index and p-value measures (Figure 5.2 (b) and (c)), the Betw models have shown a weak performance. From Table 5.2 analysis, the betweenness performance does not look so bad, surpassing the best ENet and $\text{DBet}_{\log}$ models when bigger μ values were considered. However, with the μ value decrease, the models' performance drops, leading to weaker models.

The remaining models have similar results regarding c-index and p-value. However, it is possible to observe that some outliers in the degree models concerning p-value, showing significant higher values. This scenario is the same as the one verified for the betweenness models, yet, the impact on the model performance is not so evident. The lower performance is verified because the penalisation applied by both α and penalty factor considered is too high, leading to over-constrainted models with the worst performance.

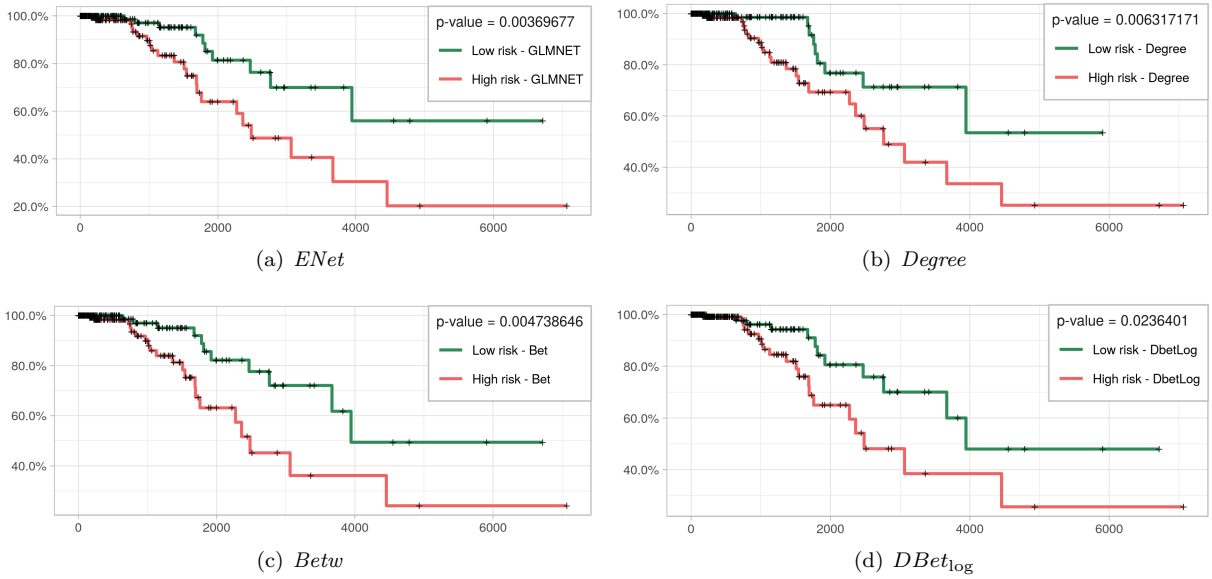
Considering all the metrics, the Degree models have showed the best results. These models exhibit better performance (highest c-index) and also have the characteristic of comprising fewer variables, resulting in simpler models. And finally, the percentage of regression coefficients that do not have hallmark hits is considerably smaller than the obtained with the other models. Although must of the models have a smaller percentage than the random selection of genes from the ones considered, the only models that stand out are the Degree models, having a median of approximately 17%. The Betw models follow the Degree ones since its best models also surpass the ENet and $\text{DBet}_{\log}$. Moreover, the number of considered variables and percentage of nodes with no hallmarks, in most of the scenarios, is smaller.

The performance of the models considering the $\text{DBet}_{\log}$ penalisation is not significant. The obtained results were very similar to the ENet ones, having the advantage of selecting fewer nodes and having most of the p-values varying within a smaller range. This fact does not mean that the obtained metrics were not attractive, but perhaps the penalty factor vector computation could be performed in a way that better benefits the most relevant genes.

In Table 5.2, the best models considering each of the centrality metrics are shadowed in light blue, with the respective values in bold. The selection of each of the models had their c-index and p-value as the primary selection factor since they measure the model performance. Nevertheless, when models have similar results regarding these two metrics, the selected models will be the ones with less non-zero regression coefficients (simpler models). Based on those considerations, the best models within each considered penalty factor are:

- ENet: train/test split ratio equal to 0.7 and $\alpha = 0.1$
- Degree: train/test split ratio equal to 0.7, μ equal 0.01 and α equal 0.2;
- Betw: train/test split ratio equal to 0.7, μ equal 0.05 and α equal 0.05;
- $\text{DBet}_{\log}$: train/test split ratio equal to 0.7, μ equal 0.01 and α equal 0.1.

The respective Kaplan-Meier curves for the high and low-risk groups considering the test data are presented in Figure 5.3 being clear the separation between both curves in all the cases. This distinct separation between the curves were expected given the obtained p-values for these models.



(a) *ENet* (b) *Degree*

(c) *Betw* (d) $DBet_{\log}$

Figure 5.3: Kapain-Meier curves considering the best models according to the different penalty factor vectors.

Nevertheless, the $\text{DBet}_{\log}$ show worst results because their curves are closer to each other. It is also interesting to note that the separation is not perfect for any of the models since they show some problems with the separation both curves at smaller times, having some crossings between the low-risk and high-risk curves.

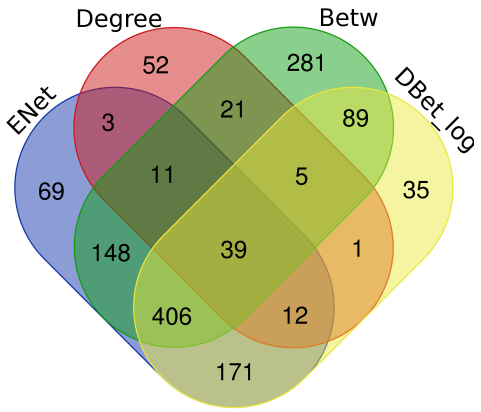
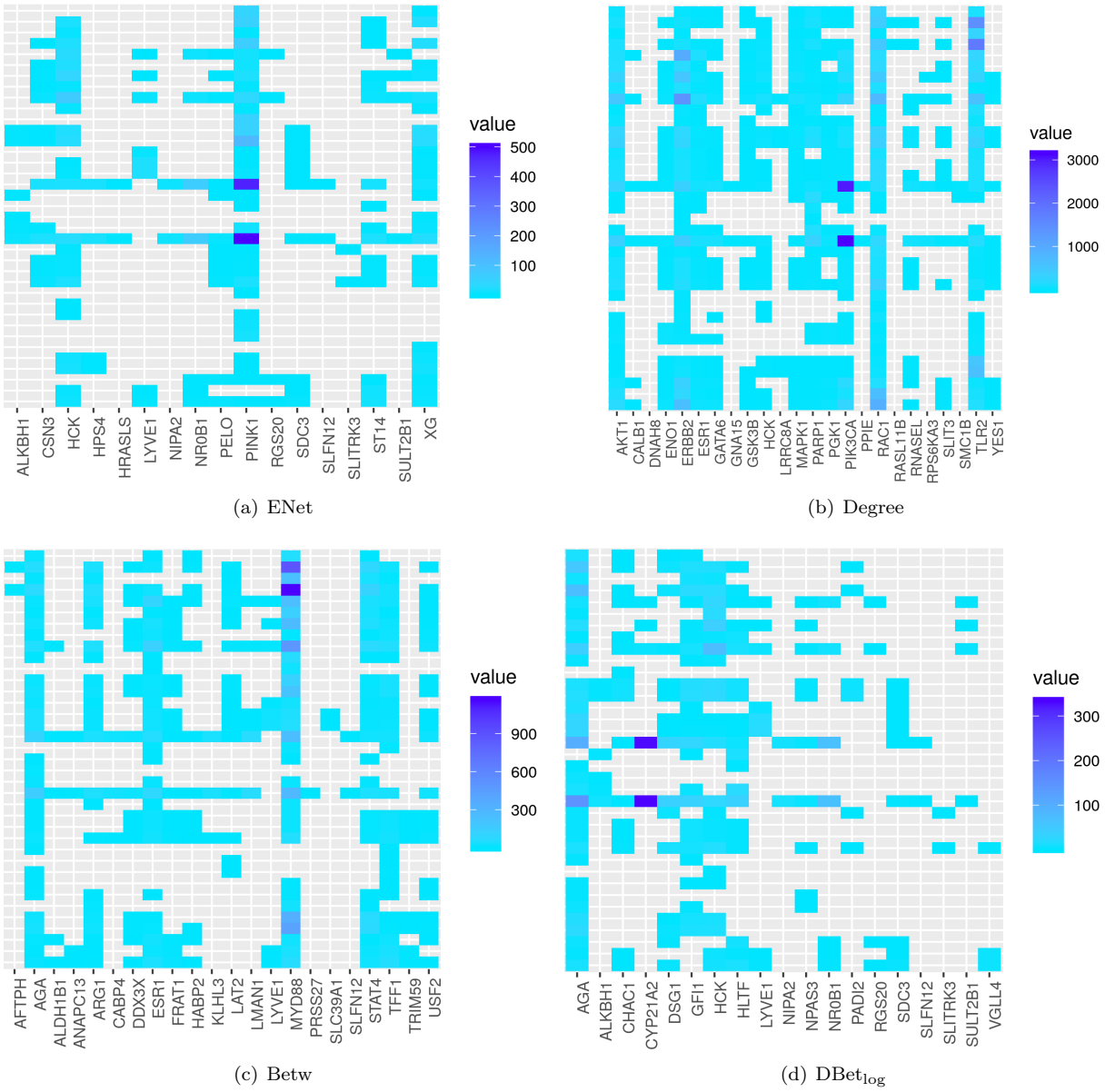


Figure 5.4: Venn diagram considering the non zero coefficients of the best selected models.

Considering the selected regression coefficients by each of the models, a Venn diagram was computed to analyse their intersections. As it can be verified by the analysis of Figure 5.4, all the models have many non-zero regression coefficients that are also considered by the ENet. The best models considering Degree, Betw and $\text{DBet}_{\log}$ penalties, have, respectively, 45.14%, 60.40% and 82.85% of intersection with the non-zero regression coefficients selected by the same method with no penalty factor associated.

To further analyse the selected regression coefficients by each model, the top 30 considering their absolute value have been obtained since they are the ones that have bigger impact on the determination of the individuals survival value. After this selection, the heatmaps returned from the CHAT presented in Figure 5.5 were obtained to understand their presence on cancer investigations.



(a) ENet (b) Degree

(c) Betw (d) $\text{DBet}_{\log}$

Figure 5.5: Heat-map on the top 30 regression coefficients for the best selected models.

The Degree models resulted is the most interesting scenario regarding hallmark hits given the fact that 24 out of 30 had at least a hit and most of the hallmarked genes are present in more than one channel. Also, two genes have more than 2000 hits in at least one of the considered channels. The ENet top 30 genes are less attractive concerning hallmark hits: 13 out of 30 with no hallmark hits. The Betw and $\text{DBet}_{\log}$ show more exciting hallmark heatmaps, yet, they are not as relevant as the Degree ones.

The proposed method encourages the usage of well-known genes according to many other studies since the used tools focus on text-mining techniques. For that reason, this methodology might also be interesting to find new potential candidates for breast cancer analysis. Concentrating on the obtained regression coefficients selected by top models, it has been shown that a significant number of genes have no hallmarks associated. Those may be genes that have a firm rule on determining the risk of breast cancer patients.

With that in mind, from the Venn Diagram on Figure 5.4 analysis, it can be seen that all the top

models select a total of 39 genes that intersected. Those genes have been taken under analysis, and the heatmap returned by the CHAT tool was computed, presented in Figure 5.6. From these 39 genes, 30 have hallmark hits which is a significant portion. Nevertheless, there are still nine genes that have no hallmarks and were considered relevant according to all the models. Those genes were penalised based on different metrics and were still selected which strongly emphasise their rule on breast cancer survival risk determination.

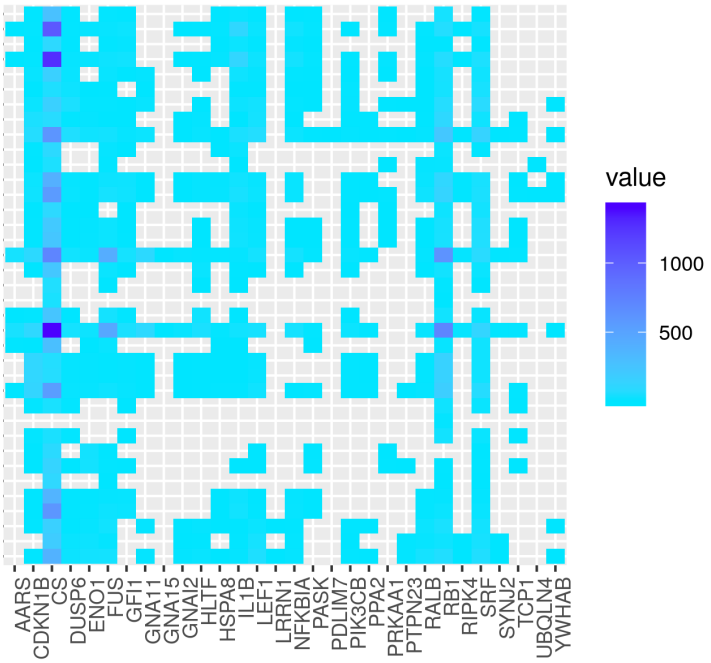


Figure 5.6: Hallmark heat-map considering the intersected regression coefficients.

Based on the studies manually annotated and reviewed by UniProtKB curators and OMIM databases [66, 67], Table 5.3 has been obtained, comprising the nine genes with no hallmarks and the respective documented function on *Homo sapiens*. These genes might be essential to carefully analyse when studying breast cancer patients not only because of their presence as regression coefficients but also, in some of the cases, because of their functions. For instance, the gene *ANKRD52* is associated to the recognition of recognition of phosphoprotein substrates and "Dysregulation of phosphorylation signalling is implicated in a wide variety of diseases" according to Sawyer et al. [68]. Another example is the *ZBTB11* that may be associated with the transcriptional regulation, a process that can be strongly related to cancer expression [69].

5.4 Results Discussion

The presented results enhance the fact that the usage of a priori biological knowledge from public databases leads to more interesting models. The best models within the Degree and Betw models were able to achieve better results while considering fewer variables than the ENet models. Not only was the number of select variables smaller but it was also biologically more significant since they have a more substantial presence on cancer studies according to the CHAT. The only models that did not reach the

Gene	Function
ANKRD52	Associated with the recognition of phosphoprotein substrates.
ANKRD53	Required for normal progression through mitosis. Involved in chromosome alignment and cytokinesis via regulation of microtubules polymerisation.
ATP5F1B	Mitochondrial ATP synthase catalyses ATP formation, using the energy of proton flux through the inner membrane during oxidative phosphorylation.
CYB5R4	Plays a critical role in protecting pancreatic beta-cells against oxidant stress, possibly by protecting the cell from excess buildup of reactive oxygen species (ROS).
LRRC4C	May promote neurite outgrowth of developing thalamic neurons.
TTC13	No well-documented function.
UBA7	Activates ubiquitin by first adenylating with ATP its C-terminal glycine residue and after that linking this residue to the side chain of a cysteine residue in E1, yielding a ubiquitin-E1 thioester and free AMP.
ZBTB11	May be involved in transcriptional regulation.
UBTF	Recognises the ribosomal RNA gene promoter and activates transcription mediated by RNA polymerase I through cooperative interactions with the transcription factor SL1/TIF-IB complex. It binds specifically to the upstream control element.

Table 5.3: Genes selected by all the top models with no hallmarks hits and their documented function on *Homo sapiens* (source [66, 67]).

expected results were the $\text{DBet}_{\log}$ models. The exponencial relation between the degree and betweenness centrality were considered relevant, yet, the results were very similar to the ENet ones. The reason for the poor performance of these models is likely to be defenition of penalty factor formula that was not the best to favour this specific metric.

As it has been shown, most of the exhibited work strongly focused on well known public datasets/tools: STRING, CHAT and BiomaRt. All of these consider similar metrics and data which might also justify the fact that the STRING proteins had so many genes mapped and even the high number of genes with hallmarks hits. Rather than belittle the presented results, this fact makes the introduced methodology an valid approach to find new candidate genes associated with the cancer type under study.

By applying the presented pipeline, the model is "bias", tending to select more important genes according to public datasets. For that reason, the selected genes that are still not present on those datasets as relevant on cancer analysis might be the ones of interest. This method can, therefore, be used to discover new potential candidates as shown with the results presented in Table 5.3.

Chapter 6

Conclusions

Over this chapter, the obtained conclusions are presented focusing the objectives of the thesis project. In the first section, the main achievements are discussed being followed by the possible future works on the introduced methodology.

6.1 Achievements

The regularisation methods for Cox PH Models are still a subject that concerns many scientists all over the world. The ability to estimate the survival of the individuals depending on the patients' gene expression data is frequently associated with the curse of dimensionality. The ability to determine the most relevant coefficients for this regression has been proved to be very difficult, and the best solution is yet to discover.

In the presented work, the hypothesis of using a priori information regarding the biological structures involved was proposed and studied. The STRING network was investigated with the objective of selecting the best metrics that reflect the network elements relevance. Centrality metrics have been considered, being concluded that the degree and betweenness are the best for the desired effect. It has also been shown that there is a exponential relation between the degree and betweenness. For that reason, another metric that considers this relation, $\mathrm{DBet}_{\log}$, has been explored to measure the genes relevance. All these metrics were used for the penalisation of the solution space considering the Cox PH models.

It was also concluded that the STRING dataset focus on very well documented proteins since most of them have their associated protein-coding gene. That fact along with the evidence that the top proteins regarding the centrality metrics are associated with genes with many hallmarks validates the usage of the STRING network on oncological investigations.

Given the obtained regression models, the primary objective of the project was achieved: get simpler models with less but more relevant genes selected while keeping the model performance. The information extracted from the STRING dataset allowed a relevant restriction of the solution space, leading, in a significant number of cases, to a sparser solution with the same performance as the Elastic Net. Moreover, the genes selected by the purposed method tend to have a more significant presence in cancer studies.

Furthermore, it has been concluded that the usage of the presented pipeline might also be relevant to find new genes that have an important role on the determination of breast cancer survival. The presented models tend to favour the genes that have already been proved to be relevant in many different types of cancer. Even so, some of the frequently selected genes are still not associated with any cancer study, being likely interesting to consider them on further analysis by a specialist in the field.

With this thesis project, it has been proved that the usage of network-based regularisation over oncological patients survival data to get Cox regression models, result on simpler models with greater biological meaning according to public datasets. Moreover, the present methodology can also be used as a tool to find interesting genes that are not yet associate with cancer investigations.

6.2 Future Work

This area of knowledge is a large road, and many steps are being taken every day. Regarding the present method, further explorations can be taken into account to explore different parameters and possibly achieve more interesting and useful results.

The penalty factors considering the STRING network can be obtained according to other centrality metrics or merely considering a different pre-process over the dataset. The STRING has shown a great potential to find relevant genes, and the combination of this information with other public datasets might consolidate the metric and possibly have knowledge of more genes.

Given the poor results obtained with the $\mathrm{DBet}_{\log}$, it would be interesting to find a better metric or more robust penalty formula to favour genes according to their betweenness and degree value simultaneously.

Another interesting aspect that could be considered in future works is the usage of the patients' clinical data. Those features typically have a strong relationship with the way the body functions and might lead to more robust models with higher performance.

An important step to further validate this method is its use on other relevant datasets covering different types of cancer. The results presented here are promising, but they should be explored on many other datasets to prove that the achieved regressions are indeed better and more straightforward.

The worked developed so far as also prove to be a relevant method to find potential gene candidates with a strong relation with cancer under study. The exhabited hypothesis, however, needs further exploration and validation over different datasets and requires the revision of a curator.

Bibliography

[1] P. Domingos. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books, 2015.

[2] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology*, 19(1A):A68, 2015.

[3] E. Brodsky. Who will make sense of all that data? `https://www.linkedin.com/pulse/who-make-sense-all-data-elia-brodsky/`, Oct 2015. Accessed: 06 Octobre 2018.

[4] S. L. Pugh. Essence of survival analysis. *Neuro-Oncology Practice*, 4(2):77–81, 2017.

[5] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[6] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[7] W. Zhang, T. Ota, V. Shridhar, J. Chien, B. Wu, and R. Kuang. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Computational Biology*, 9(3):e1002975, 2013.

[8] A. Veríssimo, A. L. Oliveira, M.-F. Sagot, and S. Vinga. Degreecox–a network-based regularization method for survival analysis. *BMC Bioinformatics*, 17(16):449, 2016.

[9] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, page gkw937, 2016.

[10] S. J. Walters. What is a cox model? *Statistics*, 1999.

[11] D. R. Cox. Regression models and life-tables. *Royal Statistical Society. Series B*, pages 527–541, 1992.

[12] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[13] W. Cheng, X. Zhang, Z. Guo, Y. Shi, and W. Wang. Graph-regularized dual lasso for robust eqtl mapping. *Bioinformatics*, 30(12):i139–i148, 2014.

[14] R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16 (4):385–395, 1997.

[15] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(1):140, 2007.

[16] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[17] T. Hwang, H. Sicotte, Z. Tian, B. Wu, J.-P. Kocher, D. A. Wigle, V. Kumar, and R. Kuang. Robust and efficient identification of biomarkers by classifying features on graphs. *Bioinformatics*, 24(18): 2023–2029, 2008.

[18] Z. Tian, T. Hwang, and R. Kuang. A hypergraph-based learning algorithm for classifying gene expression and arraycgh data with prior knowledge. *Bioinformatics*, 25(21):2831–2838, 2009.

[19] D. G. Kleinbaum and M. Klein. *Survival Analysis A Self-Learning Text*. Springer, 2001.

[20] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[21] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. Voelker, B. Nussenbaum, and E. W. Wang. A practical guide to understanding kaplan-meier curves. *Otolaryngology—Head and Neck Surgery*, 143(3):331–336, 2010.

[22] C. L. Loprinzi, J. A. Laurie, H. S. Wieand, J. E. Krook, P. J. Novotny, J. W. Kugler, J. Bartel, M. Law, M. Bateman, and N. E. Klatt. Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group. *Journal of Clinical Oncology*, 12 (3):601–607, 1994.

[23] V. Bewick, L. Cheek, and J. Ball. Statistics review 12: survival analysis. *Critical Care*, 8(5):389, 2004.

[24] G. J. McLachlan and D. McGiffin. On the role of finite mixture models in survival analysis. *Statistical Methods in Medical Research*, 3(3):211–226, 1994.

[25] B. Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977.

[26] N. E. Breslow. Discussion of professor cox's paper. *Journal of the Royal Statistical Society B*, 34: 216–217, 1972.

[27] N. Biggs, E. K. Lloyd, and R. J. Wilson. *Graph Theory, 1736-1936*. Oxford University Press, 1986.

[28] L. Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 8:128–140, 1736.

[29] M. G. Grigorov. Global properties of biological networks. *Drug Discovery Today*, 10(5):365–372, 2005.

[30] J. Travers and S. Milgram. An experimental study of the small world problem. In *Social Networks*, pages 179–197. Elsevier, 1977.

[31] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684): 440, 1998.

[32] F. Allen and A. Babus. Networks in finance. *The Network Challenge: Strategy, Profit, and Risk in an Interlinked World*, 367, 2009.

[33] A. Capocci, V. D. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical Review E*, 74(3):036116, 2006.

[34] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[35] M. Wilson. The real difference between google and apple. `https://www.fastcompany.com/3068474/the-real-difference-between-google-and-apple`, Feb 2017. Accessed: 04 Octobre 2018.

[36] J. Nieminen. On the centrality in a graph. *Scandinavian Journal of Psychology*, 15(1):332–336, 1974.

[37] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752, 2004.

[38] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[39] C.-Y. Lee. Correlations among centrality measures in complex networks. *arXiv preprint physics/0605220*, 2006.

[40] A. Bavelas. A mathematical model for group structures. *Applied Anthropology*, 7(3):16–30, 1948.

[41] A. Bavelas. Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730, 1950.

[42] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[43] A. Hershko. The ubiquitin system for protein degradation and some of its roles in the control of the cell division cycle. *Cell Death and Differentiation*, 12(9):1191, 2005.

[44] A. Ciechanover. The ubiquitin–proteasome pathway: on protein death and cell life. *The EMBO journal*, 17(24):7151–7160, 1998.

[45] A. L. Haas, J. Warms, A. Hershko, and I. A. Rose. Ubiquitin-activating enzyme. mechanism and role in protein-ubiquitin conjugation. *Journal of Biological Chemistry*, 257(5):2543–2548, 1982.

[46] S. Cheriyedath. Gene expression: An overview. https://www.news-medical.net/life-sciences/Gene-Expression-An-Overview.aspx, Aug 2018. Accessed: 05 Octobre 2018.

[47] D. A. Gutman, L. A. Cooper, S. N. Hwang, C. A. Holder, J. Gao, T. D. Aurora, W. D. Dunn Jr, L. Scarpace, T. Mikkelsen, R. Jain, et al. Mr imaging predictors of molecular profile and survival: multi-institutional study of the tcga glioblastoma data set. *Radiology*, 267(2):560–569, 2013.

[48] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17(1):13, 2016.

[49] 20 items (human) - string interaction network. https://string-db.org/, 2017. Accessed: 04 Octobre 2018.

[50] C. Von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. String: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic acids Research*, 33(suppl_1):D433–D437, 2005.

[51] N. Kravchenko-Balasha, A. Levitzki, A. Goldstein, V. Rotter, A. Gross, F. Remacle, and R. Levine. On a fundamental structure of gene networks in living cells. *Proceedings of the National Academy of Sciences*, page 201200790, 2012.

[52] V. Abadie, L. M. Sollid, L. B. Barreiro, and B. Jabri. Integration of genetic and immunological insights into a model of celiac disease pathogenesis. *Annual Review of Immunology*, 29:493–525, 2011.

[53] S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, 4(8):1184, 2009.

[54] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.

[55] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2017.

[56] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[57] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

[58] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25 (2):163–177, 2001.

[59] M. Thorup. Undirected single-source shortest paths with positive integer weights in linear time. *Journal of the ACM (JACM)*, 46(3):362–394, 1999.

[60] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[61] C. S. Gillespie. Fitting heavy tailed distributions: the powerlaw package. *arXiv preprint arXiv:1407.3492*, 2014.

[62] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[63] A. Veríssimo, E. Carrasquinha, M. B. Lopes, A. L. Oliveira, M.-F. Sagot, and S. Vinga. Sparse network-based regularization for the analysis of patientomics high-dimensional survival data. *bioRxiv*, page 403402, 2018.

[64] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, et al. The consensus coding sequence (ccds) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, 2009.

[65] F. E. Harrell Jr, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 3(2):143–152, 1984.

[66] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl_1):D514–D517, 2005.

[67] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 32(suppl_1):D115–D119, 2004.

[68] N. Sawyer, B. M. Gassaway, A. D. Haimovich, F. J. Isaacs, J. Rinehart, and L. Regan. Designed phosphoprotein recognition in escherichia coli. *ACS Chemical Biology*, 9(11):2502–2507, 2014.

[69] J. M. Thomson, M. Newman, J. S. Parker, E. M. Morin-Kensicki, T. Wright, and S. M. Hammond. Extensive post-transcriptional regulation of microrna and its implications for cancer. *Genes & Development*, 20(16):2202–2207, 2006.