Network-base Regularisation for Survival Analysis

Pedro Martinho
pedro.e.martinho@ist.utl.pt

Alexandra M. Carvalho
alexandra.carvalho@tecnico.ulisboa.pt

Susana Vinga
susanavinga@tecnico.ulisboa.pt

***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.**

***Index Terms*—Cox Regression, Regularisation, Networks, Gene Expression, Proteins.**

I. Introduction

Over the years the scientific knowledge has been reached through a well defined scientific process. To bring new discoveries to the scientific community, one needs to generate 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 hypothesis yet to be presented. Scientists might spend a lifetime creating and testing some few hypotheses, while a computer could test some thousands of hypothesis in less than a second. Machine learning is the origin of this change.

A wide range of industries and companies are focused on finding new ways to predict future events based on collected data (big data). Industries, such as energy and the healthcare, use machine learning models to 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, therefore, machine learning is here to stay and is the bioinformatics' challenge to take the best knowledge they can out of the provided data. There is a vast quantity of useful information that needs to be handled and structured, enhancing the role of bioinformatics. However, dealing with such datasets presents significant computational challenges. There are many clinical trials based on a time-to-event endpoint, frequently considering survival analysis. Some interesting work has been developed using machine learning techniques on survival data in order to better estimate the risk of a given patient. 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.

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. This often leads to overfitted models and for that cases feature selection techniques have been developed [1, 2]. 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 [3, 4].

Taking profit from previous works developed by 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.

The used network is from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [5], being necessary to explore the best metrics to extract meaningful knowledge out of it. The exploration and comparison between the main centrality metrics over this network to define the protein's importance is, therefore, 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.

II. Theoretical Background

A particular case regarding clinical analysis is the survival data analysis, meaning the study of the time between entering a study (or other baseline condition) and experiencing a subsequent event of interest. This type of datasets is possible to analyse and shape through the Cox Proportional Hazard Model (Cox PH Model) [6].

More specifically, the problem in analysis concerns cancer patients having their gene expression information, which is attached to the dimensionality curse. 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. There have been many different techniques applied to work on this problem, some with promising results [1, 2].

Recently, the idea of using networks to analyse biological behaviour has been proposed, and exciting results have been achieved when using this information on regression models [3, 4]. Based on those approachs, it will be shown how to use the protein-protein interaction network and select a centrality measure to penalise the solution space further.

A. Survival Analysis and Cox Regression

Survival datasets are meant to study the period between the time an individual joins the study and the time the event of interest is observed. The analysis over this type of dataset is frequently used over medical data to access the relationship of explanatory variables to survival time and estimate/compare survivor functions [7].

Typically, survival data is composed by the calculated features and the survival time or time-to-event. Given that it is frequent to have the individual survival time is unknown, survival analysis has to deal with these cases. This type of observations is called censored data and the datasets also include a variable that specifies if it is an event occurrence or censored data. To properly analyse these incomplete observations, the Kaplan-Meier curves are often considered [8], a powerful tool to deal with differing survival times. To validate the separation between different survival curves it is also considered log rank test.

To model this type of curves, the Cox PH Model [6] is frequently considered since it has proven to be more robust [9]. 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

$$h_i(t) = h_0(t)\, e^{\boldsymbol{x}_i'\boldsymbol{\beta}}, \tag{1}$$

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 β vector is obtained by maximising the Cox log-partial likelihood

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{x}_i'\boldsymbol{\beta} - \log \left(\sum_{j:y_j \geq y_i}^{n} e^{\boldsymbol{x}_j'\boldsymbol{\beta}} \right) \right). \tag{2}$$

Note that this formula assumes that failure time t is unique, $t_1 < t_2 ... < t_n$. To estimate the baseline hazard, $h_0(t)$, the Breslow estimator is commonly used [10], defined as

$$\hat{h}_0(t) = \frac{1}{\sum_{i=1}^{n} e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}. \tag{3}$$

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

$$l(\boldsymbol{\beta}, h_0) = \sum_{i=1}^{n} -e^{\boldsymbol{x}_i'\boldsymbol{\beta}} H_0(t_i) + \delta_i \left(\log(h_0(t_i)) + \boldsymbol{x}_i'\boldsymbol{\beta} \right), \tag{4}$$

with

$$H_0(t_i) = \sum_{t_k \leq t_i} h_0(t_k). \tag{5}$$

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. 1.

B. Networks Properties

The presented method strongly focus on the analyse of a big and complex network. When dealing with networks, it is common to consider the a graph framework having $G := (V, E)$, with V denoting the set of nodes, and E the weighted interaction between them. It is a structure used in many fields of knowledge, describing relationships between entities. When dealing with networks, due to the amount of information to store, their size is often huge, being necessary to use centrality metrics, to extract relevant information. The presented analysis and metrics focus only on undirected graphs.

1) ***Degree Centrality****:* Focusing on the connections between two nodes in a graph, there can be two different types of graphs: weighted and unweighted. When dealing with unweighted networks, the degree of a node d_i is the number of nodes adjancent to it. The degree formula is given by

$$d_i = \sum_{j=1}^{P} a_{ij}, \tag{6}$$

with P equal to the total number of nodes, $a_{ij} = 1$ if node i and j are connected and $a_{ij} = 0$ otherwise. To work with weighted networks, extensions of the Eq. (6) have been proposed. The weighted degree formula is defined as

$$D_i = \sum_{j=1}^{P} s_{ij}, \tag{7}$$

where s_{ij} corresponds to the weight of the edge.

For both presented metrics, the nodes with high degree value are called the hubs and are normally in the path between many other nodes with lower degree value.

*2) **Betweenness Centrality**:* 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 $i \neq j \neq k$ [11]. It is, therefore, given by

$$B_i = \sum_{\substack{j=1 \\ j \neq i}}^{P} \sum_{k=j+1}^{P} \frac{g_{jk}(y_i)}{g_{jk}}, \tag{8}$$

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, catching some important nodes in the networks that the degree metric cannot detect.

*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 $i \neq j$ [12]. It is given by

$$C_i^{-1} = \sum_{j \neq i}^{P} g_{ij}, \tag{9}$$

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

It is important to take into account that this metric can only be used in the case of a connected graph. If that is not the case, there will be scenarios with $s_{ji} = s_{ji} = \infty$, meaning the centrality is going to be zero to all nodes.

C. Cox Regularization Methods

The dimensionality curse ($p \gg n$) is a problem for Cox regression models since it might lead to a degenerate behaviour. Some regularisation methods have been presented over the years to constrain the solution space further [1]. In recent year, it has been proposed models that comprising network-based regulasization techeniques, such as the Net-Cox and DegreeCox [3, 4].

The total log-likelihood, Eq. (4) 2, is penalised based on

$$l(\boldsymbol{\beta}, h_0) = \sum_{i=1}^{n} \left(-e^{\boldsymbol{x}_i' \boldsymbol{\beta}} H_0(t_i) + \delta(\log(h_0(t_i)) + \boldsymbol{x}_i' \boldsymbol{\beta}) \right) - \lambda P(\boldsymbol{\beta}), \tag{10}$$

with λ as the variable that controls how much the solution space is constrained and $P(\beta)$ as the penalisation function, deferring according to the used method.

*1) **LASSO, Ridge and Elastic Net Regressions**:* The LASSO is is widely used regression method fo cases with $p \gg n$ since it imposes sparsity in the solutions(well-defined solutions), by the usage of the L_1 norm penalty [1]. The Ridge regression, on the other hand, considers the L_2 penalty, which leads to unclear solutions. Nevertheless, it is still model as it handles correlated coefficients better. Based on these models strengths, the Elastic Net method was created. The 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

$$\lambda P_\alpha(\beta) = \lambda \left(\alpha \sum_{i=1}^{p} |\beta_i| + \frac{1}{2}(1-\alpha) \sum_{i=1}^{p} \beta_i^2 \right). \tag{11}$$

*2) **Net-Cox Regression**:* Zhang et al. proposed a method based on gene relation 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

$$\boldsymbol{\Psi}(\beta) = \frac{1}{2} \sum_{i,j=1}^{p} S_{ij} (\beta_i - \beta_j)^2 = \beta' (\boldsymbol{I} - \boldsymbol{S}) \beta = \beta' \boldsymbol{L} \beta. \tag{12}$$

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. 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 is rewritten as

$$\lambda P_\alpha(\boldsymbol{\beta}) = (1-\alpha) \boldsymbol{\beta}' \boldsymbol{L} \boldsymbol{\beta} + \alpha |\boldsymbol{\beta}|^2 = \frac{1}{2} \lambda \boldsymbol{\beta}' ((1-\alpha) \boldsymbol{L} + \alpha \boldsymbol{I}) \boldsymbol{\beta}. \tag{13}$$

*3) **DegreeCox Regression**:* In the sequence of the work presented by Zhang et al., Veríssimo et al. proposed the DegreeCox [4]. 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

$$\Upsilon(\beta) = \sum_{i=1}^{p} \beta_i^2 d_{ii} = \beta' \boldsymbol{D} \beta, \tag{14}$$

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.

D. STRING Dataset, BioMart and TCGA

Each protein-coding gene is responsible for creating proteins needed for the good function of the organism, uncovering a strong relationship between them. For that reason, a protein relation network is considered. The network under study is from the STRING, a well-documented and updated collection of data, featuring known and predicted protein interactions for more than two thousand organisms, and nearly ten million proteins. "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" [5].

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

published. 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" [13]. The presented package collects and relates the information stored in rich public datasets, for instance, the Ensembl, allowing the mapping between structures at the cell level. These relations are complex and constantly updated, making the *BiomaRt* functions crucial for bioinformatics to establish the relationship between proteins and genes.

The dataset considered for training and test the proposed method is extracted from the TCGA, 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 and comprise information from more than 11,000 patients. Their datasets are public and have been widely used by the research community [4, 14].

III. Proposed Method

Firmly based on the work exposed by Veríssimo et al. [4], 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. 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

$$\lambda \sum_{i=1}^{p} v_i P_\alpha \left(\beta_i\right) = \lambda \sum_{i=1}^{p} v_i \left((1-\alpha)\, \frac{1}{2} \beta_i^2 + \alpha |\beta|_i \right). \tag{15}$$

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. Within the STRING information, for each connection, it is given an overall score named "combined score". With all the edges information, a biological network can be defined as the adjacency matrix A, 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

$$a_{ij} = \begin{cases} 1, & \text{if } i \neq j \text{ and } combined\_score_{ij} > \theta \\ 0, & \text{otherwise} \end{cases}, \tag{16}$$

where θ is equal to the threshold applied to the "combined score". 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, the pipeline presented in Figure 1 will be applied in order to study the most promising centrality measures to get the best regression models.

A. Centrality Metrics Computation

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.

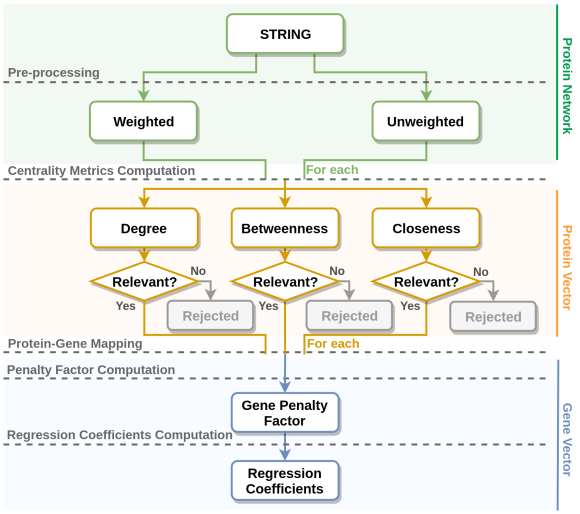


Fig. 1. Proposed methodology over the STRING network to reach the final regression coefficients.

The computation betweenness was associated with a big complexity cost given the involvement of shortest path calculation. Nevertheless, Brandes presented an algorithm that reduces the required resources to $\mathcal{O}(n+m)$ in terms of space and $\mathcal{O}(nm)$ in terms of time [15].

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. It has been proven that the best algorithm to calculate all the shortest paths between all the nodes in an unweighted with positive integer weights network also a time complexity of $\mathcal{O}(nm)$ [16].

B. 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. The *BioMart* package provides a powerful link between biological databases and microarray data analysis, by bridging proteins and genes information [13, 17]. Unfortunately, some proteins and genes are not fully documented yet, leading to some mismatch between these two entities. Moreover, the relationship between proteins and genes is a many-to-many connection being possible to have more than one gene responsible for the production of a specific protein and more than one protein produced by the same gene. To solve the later cases, the genes with more than one protein associated would have the sum of the considered centrality measure values associated to those proteins.

C. Penalty Factor Computation

The higher penalty 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

$$v_i = \frac{1}{w'_i}, \tag{17}$$

with w'_i as the re-scaled centrality metric for the gene i. The re-scaling process applied over the centrality metric value is given by

$$w' = \frac{w - min(w)}{max(w) - min(w)} + \mu, \tag{18}$$

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.

D. Regression Coefficients Computation

As stated, the regression coefficients for the Cox PH Model are obtained according to the Eq.(15), 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.

IV. 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 network will be analysed in detail, being presented the respective centrality metrics distributions along with other relevant properties. The effects of the protein-gene mapping process are also presented as well as the penalty factor computation.

A. Centrality Measures

The used STRING netwrok comprises information of *Homo sapiens* protein interactions [5], being composed by 5676527 edges, with an average combined score value equal to 277.6, minimum value of 150 and never exceeds 999. The total number of proteins considered was 19576, and the average shortest path separating any two nodes in the network shows the value $\langle l \rangle = 2.203$, 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. To have a deeper understanding of the proteins influence in the network

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. (7). Given that betweenness and closeness metrics consider shortest paths in their computation and the network reflect entity relations, it has been concluded that only the degree centrality is worth consider in this scenario.

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 Fig.2. The distribution count axis is in $\log_{10}$ scale to properly analyse the degree distribution and identify the number of hubs network.

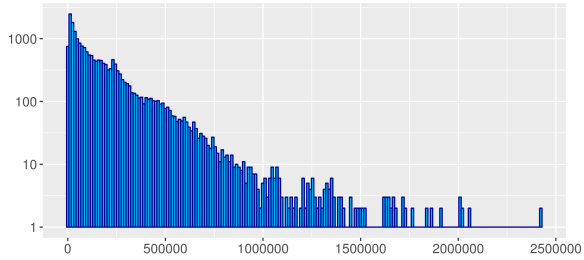


Fig. 2. Weighted degree distribution in $\log_{10}$ scale with $\theta < 150$.

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.

To test this assumption, the *plfit* function from the *poweRlaw* package [18], estimating that the distribution indeed follows a powerlaw distribution for nodes with degree higher than 99953, having $\gamma = 2.112975$. 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.

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. (16) at page 4, with $\theta < 150$. The outcome networks on cases with bigger θ values would not be connected networks, strongly harming the betweenness and closeness centrality values.

The degree distribution is presented in Fig.3, being similar to the distribution of the weighted case, Fig.2, even though the weighted degree distribution decreases more smoothly. Again, the distribution approximates a power law distribution for high values of degree. For nodes with degree higher than 2035, it approximates a power law distribution with $\gamma = 4.48$.

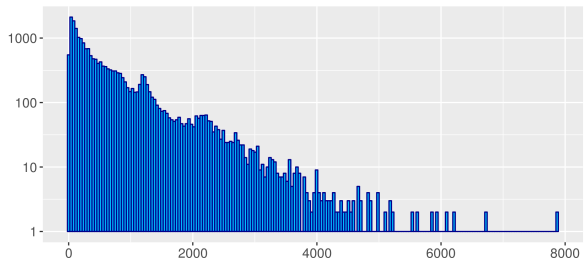


Fig. 3. Weighted degree distribution in $\log_{10}$ scale with $\theta < 150$.

The correlation value between unweighted and weighted degree is 0.954, which is a very high. For that reason and given the relevance of other metrics considered on the unweighted network, was the weigheted degree metric was rejected.

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 [19] and covers some specific cases that the degree alone cannot detect.

In Fig.4, is presented the betweenness distribution, being clear that, just like the degree distribution, few nodes have high values of betweenness. These values genuinely stand out, being interesting to relate this metric values with the degree distribution and take some conclusions.

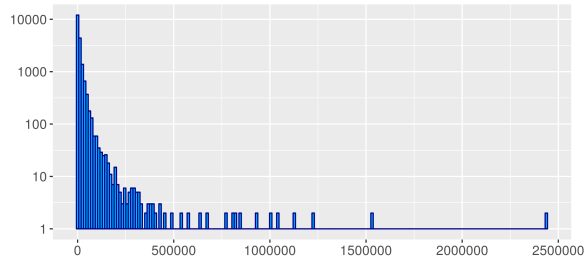


Fig. 4. Betweenness distribution in $\log_{10}$ scale with $\theta < 150$.

The closeness evaluates how adjacent a specific node is to all the nodes in the network. In Fig.5, the closeness distribution is presented in $\log_{10}$ scale and the respective distribution is very different from the betweenness and degree.

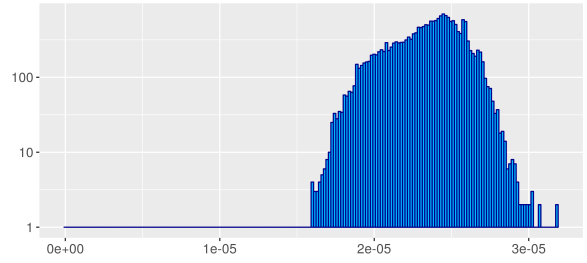


Fig. 5. Closeness distribution in $\log_{10}$ scale with $\theta < 150$.

Some nodes stand out regarding closeness values. However, the closeness distribution is very smooth when compared to the obtained considering degree and betweenness. A priori, 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 metric, Fig.6. Note that nearly all closeness vector intersects the degree vector, indicating a strong relationship between these two metrics. The correlation between both them is 0.785 which is high, being relevant to exclude one of them. Knowing that the degree has already been studied over gene networks [4], the selected metric between this two was the degree.

Fig. 6. Venn diagram on the top 250 proteins regarding the degree, betweenness and closeness metrics.

Notice that the betweenness vector selected 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 and, yet, 81 out of the top 250 proteins do not intersect, make it worth considering their relation to combine both strengths. The linear relation between degree and betweenness was not low, nevertheless, the exponential relation is much more interesting: 0.885. In the Fig.7 it is presented the logarithm of re-scaled betweenness vs logarithm of re-scaled unweighted degree. With this relation in mind, it is purposed the $\text{DBet}_{\log}$ distance metric.. This formula would consider both degree and betweenness through

$$DBet_{\log} = \sqrt{d'^2 + B'^2}, \tag{19}$$

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

$$d' = \frac{d - min(d)}{max(d) - min(d)}, \tag{20}$$

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 Fig. 7, illustrated with a red dotted line.

The distribution obtained with this new metric $\text{DBet}_{\log}$ is presented in Fig.8. As it can be observed, the distribution is

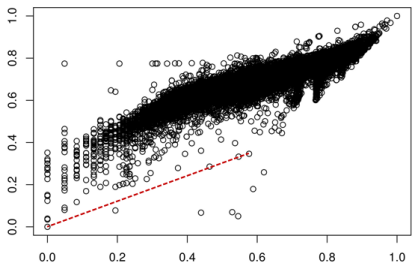
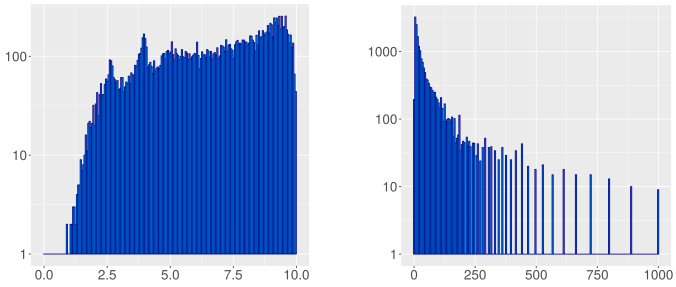


Fig. 7. Logarithm of re-scaled betweenness vs logarithm of re-scaled unweighted degree. The *x-axis* corresponds to the logarithm of re-scaled unweighted degree and *y-axis* to the logarithm of re-scaled betweenness.



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

Fig. 9. Penalty factor considering degree centrality and different μ values.

not so sharp as the betweenness and degree distributions, and that is because logarithmic values are considered.

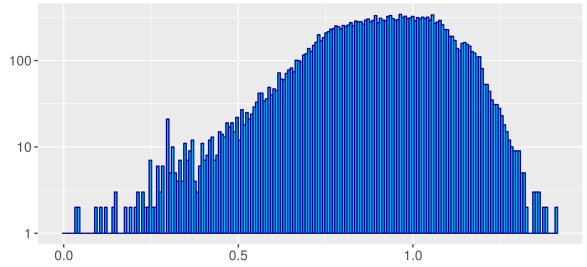


Fig. 8. $\text{DBet}_{\log}$ distribution in $\log_{10}$ scale with $\theta < 150$.

Given the results presented, it can be concluded that the strongest candidates to use as a penalty factor are betweenness and degree. The proposed $\text{DBet}_{\log}$ has also shown an interesting distribution. Not only that but is a metrics that can consider both degree and betweenness at once: focusing on the top 250 proteins, it presented 47 proteins with intersection with only the betweenness and 31 with only the degree. The metrics that will be considered as penalty factor are degree, betweenness and $\text{DBet}_{\log}$ distance.

B. Mapping and Penalty Factor

The bridge between proteins and genes needs to be crossed by using the *BioMart* package. The 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, passing through the process presented in Section III-B. Most of the information was kept on this process and only some few proteins information needed to be joint in a single gene, being kept the different metrics main characteristics.

The penalty factor was calculated considering the different metrics and different μ values. From the analysis of penalty factor distribution on the degree scenario with different μ values, Fig. 9, the μ parameter influence is clear.

V. 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.

A. Breast Cancer Dataset

The Breast Invasive Carcinoma (BRCA) survival dataset from TCGA will be used to test the presented method focusing the different centrality metrics. The dataset comprises the patients gene expression levels and clinical data. It 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 [20, 21].

Before using the dataset, a selection of the genes, according to the ones consider in the STRING, was necessary. 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, meaning that more than 90% of the protein-coding genes that were kept.

B. Validation Metrics

To validate the presented method, the Elastic Net (ENet) model will be handled as the baseline model. On the models under analysis it will be selected low α values, considering the Eq. (15), to have more non-zero coefficients in the final models and better understand the penalty factor influence on the results.

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 the log-rank test (p-value), concordance c-index [22] and number of selected genes.

To further understand the performance of each of the models, the analysis over the selected genes has also been implemented. First the intersection of the selected genes by the different models is analysed and them the most relevant genes will pass over the Cancer Hallmarks Analytics Tool (CHAT) has been used [23]. This tool considers text mining techniques over cancer-related references from PubMed, being possible to understand the percentage of genes that are present in literature related to oncology investigations.

C. Selected Regression Coefficients Analysis

To obtain the best models to predict whether a person belongs to a high-risk group or not, many combinations of train/split ratio, α and μ values were considered. The parameters values combinations are presented in the Table I, having the number of genes selected, the c-index and the p-value for each models according to the different combinations.

These values were obtained after the best λ value was obtained with a 10-fold cross-validation process to avoid overfited models. 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.

It was concluded that the 0.8 train/test split ratio value results on models that stand out negatively., since all of the resulting models are not statistically significant. This fact might be verified because the models start overfitting when higher values of this parameter are used. The μ show a strong impact on the models' results. However, for high values of this parameter, the results are not so different than the ones obtained with the ENet since the nodes have a considerable small penalisation. As the μ decreases, the range of values of the penalisation vector increases and some exciting models start to stand out, even though the number of considered regression coefficients is smaller. The ENet models have a good performance for the majority of the scenarios, yet, 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.

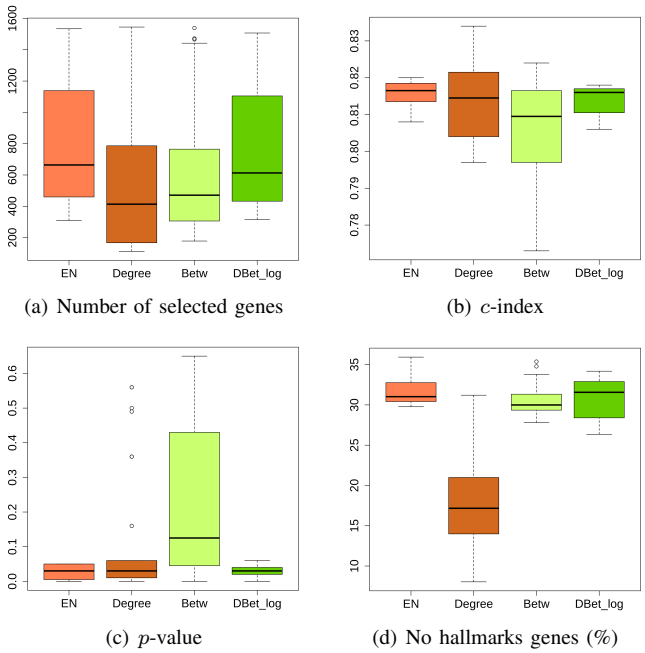
Another important used metric for all the selected models is the percentage of the non zero regression coefficients with no hallmark hits, also presented in Table I. Note that 36.46% of all the examined genes have at least a hallmark associated. These values were obtained with the usage of the CHAT.

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 (models shadowed in green), Figure 10. 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.

Focusing on the number of genes selected in Fig.10 (a), the Degree and Betw models stand out as they select fewer coefficient regressions. The $\text{DBet}_{\log}$ does not stand out and, likely because the penalisation for most of the genes is small since the performance is always very similar to the ENet models.

With respect to the c-index and p-value measures, Figure 10 (b) and (c), the majority of the models have similar results regarding c-index and p-value. However, it is possible to observe some outliers in the Degree and Betw models. This lower performance is verified because the penalisation applied by both α and penalty factor considered is too high for some of the models.

Considering all the metrics, the Degree models have showed the best results. These models exhibit better performance



(a) Number of selected genes (b) c-index

(c) p-value (d) No hallmarks genes (%)

Fig. 10. 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.

(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, a median of approximately 17%. The Betw models follow the Degree ones since its best models also surpass the ENet and $\text{DBet}_{\log}$ and, the number of considered variables and percentage no hallmarkss genes, in most of the scenarios, is smaller.

In Table I, the best models considering each of the centrality metrics are shadowed in light blue, with the respective values in bold. Considering the selected regression coefficients by each of the models, a Venn diagram was computed, Fig. 11. The best models considering Degree, Betw and $\text{DBet}_{\log}$ penalties, have, respectively, 45.14%, 60.40% and 82.85% of intersection with the ENet selected genes, which reflex influence of the penalty factor.

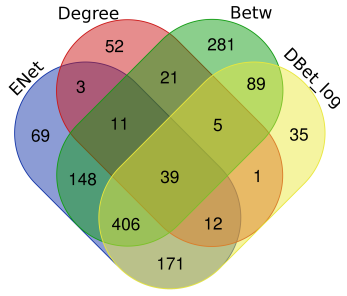


Fig. 11. Venn diagram considering the non zero coefficients of the best selected models.

The proposed method encourages the usage of well-known

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 I

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 10.

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. With that in mind, from the Venn Diagram on Figure 11 analysis, it can be seen that all the top models select a total of 39 genes that intersected. Focusing

on those genes, and using the CHAT, it was concluded that 30 from 39 genes 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.

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" [24]. Another example is the *ZBTB11* that may be associated with the transcriptional regulation, a process that can be strongly related to cancer expression [25].

VI. Conclusions

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.

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. 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.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

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

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

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

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

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

[13] 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.

[14] 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.

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

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

[17] 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.

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

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

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.