UTRECHT UNIVERSITY

INTERNSHIP REPORT

Comparison of Pathway Activity Measures for Predicting Clinical Outcome of Cancer Patients

Author:
Joanna VON BERG

Supervisor:
Dr. Holger FRÖHLICH

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Algorithmic Bio-informatics group
Bonn-Aachen International Institute for Information Technology

July 21, 2017

Utrecht University

Abstract

Faculty of Science
Bonn-Aachen International Institute for Information Technology

Master of Science

Comparison of Pathway Activity Measures for Predicting Clinical Outcome of Cancer Patients

by Joanna VON BERG

In this work, a Cox proportional hazards model in combination with a sparse group lasso was employed to predict survival of breast cancer patients and glioblastoma multiforme patients. To this end, mRNA, CNV and methylation data was summarized in patients x pathways matrices. The summarization was done using four different methods; average, ssGSEA, first principal component and top 50% average so as to compare and choose the method that shows the best prediction performance. The first principal component performed the best for the breast cancer data set, with a median C-index of 0.63. The average method performed the best for the glioblastoma multiforme data set, with a median C-index of 0.57. For both diseases, the top four most important pathways for predicting survival were investigated further in a small literature study. For all these pathways an association with the disease in question was found.

Acknowledgements

I would like to thank *Holger Fröhlich* for giving me the chance to do this project in his lab, the weekly discussions during the group meeting and for reminding me again and again of the goal of my project, every time I was about to try out some interesting new method that I found out about.

I would like to thank *Ashar Ahmad* for the daily supervision and for always being there if I had a question.

I would like to thank *Jeroen de Ridder* for the online support, and the talks we had in Utrecht. It was a good practice for me to talk about my project with someone who was not involved in the research himself.

I would also like to thank my closest friends for keeping me motivated, especially through the writing process, by showing me they really believe in me.

Contents

Chapter 1

Introduction

Predicting the future is something humankind has been attempting to do for as long as we know. While this probably brings to mind a woolly image of a figure clad in robes hunched over a crystal ball, nowadays we are getting better and better at actually predicting the future based on history. We might apply this idea in our personal lives, where we learn to put away the loaf of bread after lunch because last time we didn't, it attracted mice. We can also apply it in our professional lives, for instance when a doctor learns after years of experience that that particular cough is usually a sign of something more serious than a common cold. Instead of relying on an individual's capability to learn during their life and somehow apply their learned knowledge in a correct way, without being biased by their personal experiences, we can also choose to let computers do the learning. For now at least, computers are not sentient and are thus not clouded by emotions. They don't get tired, and they do exactly what you tell them to do. This is what the research field of machine learning is based on: discovering new knowledge in data, using algorithms to either describe the data or predict outcomes for new data entries for which the outcome is unknown. Machine learning, or data mining, has seen an increase in commercial interest in the last decade. This can mainly be attributed to the realization of many companies that the large amounts of data that they have collected on their customers contains information that can be used to make more profit if that information can be made explicit. [10] Within the field of machine learning there are two main ways of learning the structure of the data: supervised and unsupervised. If labels are available for each data entry, and we want to be able to predict the label for new unknown data entries, we will apply a supervised learning method that takes a look at the labels and at how the data values and class label relate to each other. If we mainly want to learn the structure of the data and divide the data into different groups based on the values of several variables for each data entry, we choose an unsupervised learning method. [27] Data, in the form of two-dimensional tables that give the values of several variables for multiple entries (patients in this case), is not the only thing that can be supplied to machine learning algorithms to guide prediction of the class label. In order to improve prediction, we can also supply prior knowledge. This would give us results that are in line with what we already know about a certain subject, but does result in new findings. For instance, prior knowledge on how the cell regulates itself on different layers of regulation can help in guiding the algorithm how the data can be understood. These different layers of regulation in the cell are represented on the one hand by data like gene expression, gene methylation, miRNA expression and other omics data. On the other hand, interactions between biological molecules can be depicted by regulatory networks, signal transduction pathways or protein-protein interaction networks for example. If we can find a way to transform the data from two-dimensional tables that contain values for expression of columns of genes for each patient-row to tables that contain values for activity of columns of

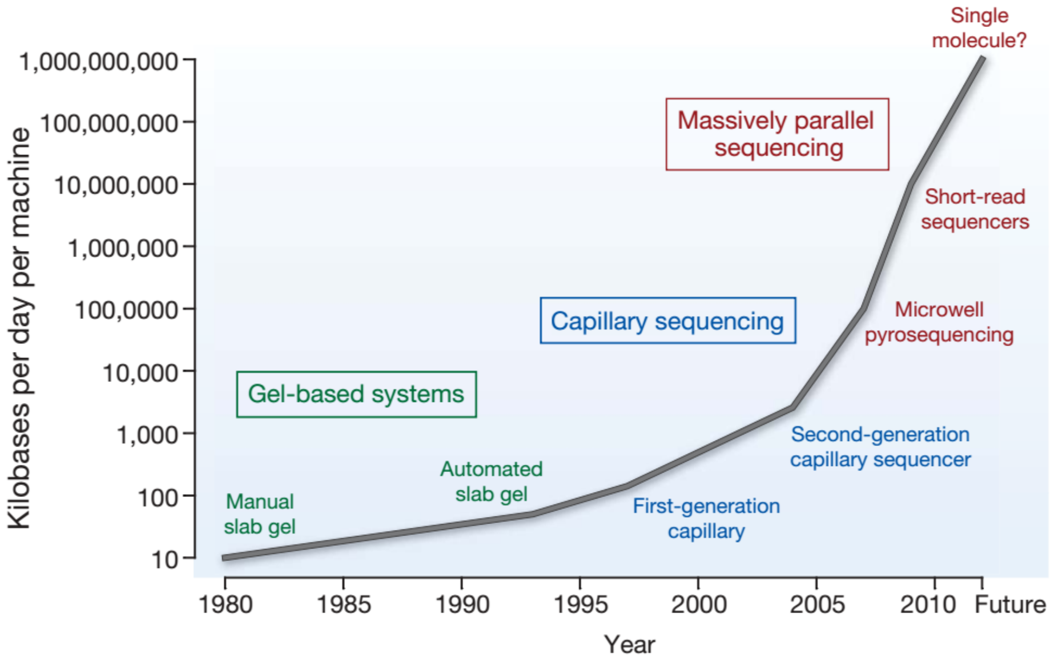

FIGURE 1.1: When DNA sequencing was still done manually in the eighties, only a few dozen kilobases could be read by one machine each day. Capillary sequencing meant a ten-fold increase in efficiency. The introduction of next generation sequencing, or massively parallel sequencing, had taken us into the tens to hundreds of millions of kilobases by the end of the zeroes of the 21st century. This figure was first published by Stratton et al [38].

pathways, we have found a way to use prior knowledge on cellular regulation to better display the data before we use it to train a machine learning model. [14] Since the invention of next generation sequencing (NGS) more genetic data has become and continues to become available. See figure 1.1 for an indication of the volume of data that could be generated by one machine per day, and how it changed over the last few decades. Researchers have used this data and also the older micro-array data to for instance predict survival of patients [2], predict protein functional category [17], and predict important genes in cancer [3]. Up until now, most machine learning in the medical field has been done using just gene expression and without considering pathways or some other way to process the data so as to incorporate prior knowledge. Although this approach was shown to possess some predictive power, it is very plausible that data from different levels of cellular regulation give a more comprehensive view on what is truly going on inside the cell and thus give a better prediction performance.

The growing amount of data on not only gene expression but also for instance the proteome and methylome, can aid in completing the picture that we used to get from gene expression alone. For disease does not only manifest in gene expression, but on all regulatory levels. A disorder might start with a somatic (i.e. obtained during life) mutation in a coding or non-coding part of the genome, but will inevitably show up in the other -omes. For example, in the case of a coding mutation, the protein sequence and structure will have changed, and in the case of a non-coding mutation in a promoter region changes in protein expression will occur for the gene in question. Changes in DNA methylation might occur if there is a mutation in a gene that is associated with that process, or if the gene that is being methylated is made more accessible to modification by a mutation. By combining this new data with prior knowledge on how the cell is regulated and how different molecules interact with each other, we hope to be able to get the most information and hopefully, the most predictive power. Next to giving a more accurate view of cellular regulation by integration of several omics data sources, condensing the data on pathway level also decreases the total number of variables by putting genes together that belong together (namely, in the same pathway).

Genetic variations are typically very sparse; in one genetic locus there will usually only be a variation in some people. This high level of sparseness calls for the extraction of higher level features so the data becomes less sparse and more meaningful. By condensing the data on pathway level, we hope to get a better view of the disease so we can give a better prognosis. Specifically, we want to know which method is the best in summarizing omics data in such a way that the prediction performance benefits the most. Simultaneously, the complexity of the model will decrease because the number of pathways is lower than the number of genes in the original data. Less variables in a machine learning setting is a good thing, because there is a smaller chance of overfitting: a problem that can occur when the model is trained on not enough data entries for the number of variables and results in poor prediction performance on a testing set. [27] We will be comparing different techniques to summarize omics data on pathway level, and train models based on this summarized data. Each technique will be evaluated in a cross-validated manner, using the C-index as a metric. [16] See section 3.4 for an explanation of the C-index used in this work. The aim of this research is to find, from a set of given techniques, the best one to summarize omics data on pathway level and integrate the different omics data sets.

Chapter 2

Data processing

For this research data generated by the The Cancer Genome Atlas (TCGA) Research Network was used [1]. TCGA is a collaboration between the United States' National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) that collected different omics data from patients of 33 different cancer types. The Broad institute has developed a Firehose API that provides the TCGA data, processed in such a way that it is ready for analysis. The data was accessed from there. A data firehose API is an application programming interface that streams all available data in real time (as far as possible) to its end users. It is therefore more up-to-date then a regular database that may or may not be accessible through an API. [2]

2.1 Cancers

Cancers are a collection of diseases that are characterized by abnormal cells dividing in an uncontrolled manner. These abnormal cancerous cells can form a tumour, and can also use the blood and lymph systems to travel through the body and form a new tumour somewhere else (this process is called metastasis). The first described cases of cancerous tumours (as far as we know now) were found in the Edwin Smith papyrus, that is estimated to be written around 1600 years before the common era. Nowadays, we recognize over 100 different types of cancer, each with their own symptoms, risk factors and treatments. [38] The first pointers that suggested that the genetic sequence might be a very important factor in the development of cancers, were studies performed by von Hansemann and Boveri around 1900. [38] The final proof came when Shih et al transformed genes from mouse tumours into healthy mouse cells and saw that the healthy cells had become cancerous. Since this discovery in the 1980's researchers have been looking for which genes are responsible for which cancer and specifically which base mutation causes it.

2.1.1 Breast cancer

Being the second biggest cause of death by cancer for women [19], breast cancer (breast invasive carcinoma) is one of the most intensively researched cancers. Breast cancer patients usually don't die because of the primary tumour, but because of the metastases that so frequently occur for this type of cancer. Researchers have been trying to find the genes responsible for the occurrence of the disease, and also trying to find genetic variables that indicate clinical outcome. [2]

[1] http://cancergenome.nih.gov/

[2] https://www.pubnub.com/blog/2014-11-14-what-is-a-data-firehose-api/, accessed on 2^{nd} of July 2017

2.1.2 Glioblastoma multiforme

Even in this day and age, where we are making so much progress in the medical field, glioblastoma tumours remain one of the deadliest cancers and one of the most difficult ones to treat. [1] With around three in 100,000 people getting the diagnosis glioblastoma multiforme (GBM) each year, we need to strive to make progress in the understanding and treatment of this disease. GBM is a type of glioma; a glial cell that has become cancerous. Glial cells are found in both the central and peripheral nervous systems and support and/or protect the neurons, depending on the exact type of glial cell. In 51% of all cases of glioma, the type of cancer is GBM. The most common treatment of GBM is surgery; it is proven that taking away more than 98% of the tumour leads to double survival (up to a maximum of 11-12 months). [1] Overall, median survival from the moment of diagnosis is a mere 14 months.

2.2 Omics

The Broad Institute TCGA Firehose was used to access the omics data and the clinical data. The run from the 28^{th} of January 2016 was chosen, which was the most recent data at the start of this project. The application *firehose_get* was downloaded from their website and was used to download the data.

All patient ID's were put in the same format: divided by dots, in all capitals. After the separate omics data had been processed (see the subsections below for more info on how the different data was processed), they were combined to get a matrix of samples x omics_genes. Upon doing this and checking the overlap in patients that show up in every omics dataset, we saw that the overlap was too low to train a model. So it was decided to only use mRNA, CNV and methylation data. See the discussion for a motivation for this choice. For BRCA, 214 patients were used in the final cross-validations. For GBM, 269 patients were used in the final cross-validation.

2.2.1 Gene expression micro-array data

The gene expression micro-array data (mRNA) was missing some values, these were imputed using the *impute* R package.[3] Imputation was done using the k nearest neighbours algorithm. The other omics data was not missing any values. The gene expression data was also the only datatype that was corrected for possible batch effects arising from the different hospitals the samples came from. Batch effects are differences in the distribution of different data sets that come from different sources. These differences in distribution can become a problem when the non-biological confounding feature 'batch' of the sample turns out to be an important factor in classification. [20] Batch effects were only corrected for gene expression, because the algorithm was developed for micro-array data so it can not simply be used on the other omics data. See chapter 5 for a further discussion of this choice.

2.2.2 Micro-RNA expression

miRNA's target mRNA's. Godard et al [11] show that first looking up the target genes of the miRNA's under consideration and then doing pathway analysis on the resulting gene targets is biased because a certain gene can and will show up more

[3] Hastie T, Tibshirani R, Narasimhan B and Chu G (2017). impute: Imputation for microarray data. R package version 1.50.1

than once in a certain pathway. The bias comes from the fact that genes that are known to be associated with disease are researched more and will show up more often as the target of micro-RNA's. If analysis is carried out in this way, the expression of miRNA's that target known disease genes will have an unfairly/unrealistically large influence on the resulting new features (namely, a value per pathway). Therefore, they suggest to convert the pathway gene lists to pathway miRNA lists, and do pathway analysis directly on the miRNA data. This prevents miRNA's being 'counted double' since every miRNA will only show up once in a pathway list.

2.2.3 Copy number variation

Copy number variation data is usually presented in a table that gives the genomic region, and the copy number that was found for that region. It is a computationally heavy task to convert this format to a samples x genes table. Fortunately, such a table that is ready for further analysis was already available on the Broad's institute Firehose.

2.2.4 Methylation

The methylation data, which was originally in the form of beta-values, was converted to M values using the *beta2m* function in the *lumi* R package. Beta-values are the absolute DNA methylation values, which are called that way because they tend to resemble a beta-distribution (a type of continuous probability distribution). [4] Using Beta-values can lead to statistical problems, so they can be log-transformed to M values which are all between -1 and 1. However, even though M values are better for doing statistical data analysis, some biological interpretability is lost in this conversion.

2.3 Clinical data

The clinical data from TCGA was processed to only contain a column with patient ID's and two columns called 'days_to_death' (if the patient died during the period of follow-up) and 'days_to_last_followup' (if the patient was lost to follow-up). This was converted to a data-frame with the patient ID's as row names, one column called 'days' which is the combination of 'days_to_death' and 'days_to_last_followup', and one column indicating if the patient died during follow-up or if they were lost to follow-up.

2.3.1 Censoring

Survival data can, and will often be, censored. This means that the patient in question did not die during the period of follow up and their day of death is thus unknown. Censoring can result in problems during analysis because simple regression is not possible with these "larger or equal to" values.

2.3.2 TCGA patient ID's

The Cancer Genome Atlas uses two different ways to identify samples; an older and a newer set of identifiers. The older version is called a bar-code, and is build up of

[4] http://www.nature.com/nrg/journal/v13/n10/glossary/nrg3273.html, accessed on 3rd of July 2017

series of characters that identify e.g. the hospital the sample came from. The TCGA bar-codes are built up as follows: in consequence there are series of characters that denote the project (TCGA in this case), the tissue source site, the participant, the sample, the vial, the portion, the analyte, the plate number and the centre where the samples are analysed.

2.4 Pathway data

To be able to apply the summarization methods that give an activity measure per pathway, we first need to know which genes belong to which pathway. This information will be gotten from the Kyoto Encyclopedia for Genes and Genomes (KEGG) [34]. For miRNA, this list of pathways and their corresponding genes needs to be converted to a list of pathways and their corresponding associated miRNA's. The *PATH2EG()* function in the *org.Hs.eg.db* R library was used to get a list of 229 KEGG pathways and their associated genes (denoted by their Entrez gene identifiers). A lookup table was made to convert Entrez gene ID's to the HGNC ID's that were used to denote the genes in the omics data matrices. The *biomart* library was used for this.

Chapter 3

Methods

All data processing and programming were done using the programming language R, on an Rstudio [39] server running under Linux.

3.1 The different methods that will be compared

In order to get new variables that act as a measure for pathway activity, several methods will be applied to the different omics data sets. This will transform the original sample x genes matrices in sample x pathways matrices. All methods need to be supplied with the original sample x genes matrix and the pathway list of genes/miRNA's. For all methods it applies that absolute values were used for CNV and methylation data. This was done, because otherwise up- and down-regulation could be cancelled out and a value of zero means 'no change' or 'no effect' in these data types. This means that simply averaging the data would result in seemingly no or very small effects while in fact, some genes might be up-regulated and others might be down-regulated. Therefore the absolute values were used. After for instance averaging all the genes in a certain pathway, this results in a "ratio of changed genes" per pathway. Nota bene, by using this method of handling the data, a change by $-x$ counts just as much as a change by $+x$.

3.1.1 Average

As was tested out by Hwang [18], simply taking the average of the gene expression per pathway gives a better classification performance than expected by chance. This will also be applied to the other omics sources in this research. Hwang applied several methods to only gene expression, whereas in this research several omics data will be used. Also, Hwang performed binary classification on a few different datasets, whereas in this research prediction of survival times will be performed. A function was written in R for the application of this method. It iterates over the list of pathways, and for every pathway it looks up the pathway member genes/columns in the data matrix. It then calculates the average over all these genes, and returns one column with the pathway activity for each patient. Because the result of this method for each patient is independent of values or class labels of other patients, the calculation can be done once before cross-validation.

3.1.2 Top 50 percent average

Another method that was tested out by Hwang uses per pathway only the top 50% of the pathway that are most strongly correlated with the class label (a binary class in Hwang's case, integer survival data in this case). Hwang subjected all genes to a Student's t test and ordered them on decreasing p-value. In this research, since we are dealing with survival data, a Cox proportional hazard model was trained per gene column and the p-values per gene were saved in a vector. When iterating over the pathways, in every step the corresponding p-values were looked up and only the top 50% percent of the genes in that pathway were used to calculate the average. Based on the paper by Hwang [18], a function was written in R for the application of this method. Similarly to the average method, it iterates over the list of pathways, and for every pathway it looks up the pathway member genes/columns in the data matrix. Contrastingly, it doesn't take all member genes for calculation of the average, but only that half that correlates the strongest with the class label. This correlation is calculated by performing Cox regression on each gene (using only the current training data). The result of this method does depend on the class labels of other patients, therefore it needed to be incorporated into the cross-validation loops.

3.1.3 Principal component analysis

Also investigated by Hwang was to perform principal component analysis (PCA) and to use the first principal component of the gene expression values in a particular pathway. Principal component analysis transforms the original variables/dimensions of the data into new dimensions (the principal components) that indicate the directions in which the variance in the data is the largest. See figure 3.1 for a two-dimensional example of principal component analysis where the original x and y axes are transformed into the first (red) principal component and the second (blue) principal component. The first principal component gives the single dimension in which the variance is the largest. Each successive principal component is orthogonal to the previous one, and gives the next dimension in which the variance is the greatest. Therefore, PCA can be used as a dimensionality reduction technique by keeping only a subset of the principal components (i.e. the first few). By using only the first, the most variance is kept with only one variable per pathway. A function was written in R for the application of this method. Again, we iterate over the list of pathways, and we take all member genes. But instead of calculating the average over (a part of) the genes, we perform principal component analysis (by using the *prcomp()* function) and calculate the projection of the original data on the first principal component (by using the *predict()* function). Because PCA is performed on multiple samples, this methods is also incorporated in the cross-validation and PCA is only performed with the training set. All samples are projected on the resulting first principal component.

3.1.4 ssGSEA

In regular gene set enrichment analysis (GSEA), the correlation between genes and the class label is calculated and the genes are sorted in a list based on the strength of this correlation. For each gene set (member genes of a certain pathway in this case) this method looks to see whether the genes are found more at the top or at the bottom of the correlation list. Based on this, an enrichment score is calculated. This score is the pathway activity assessment score for this method. In single set gene set enrichment analysis (ssGSEA), a variant of regular GSEA that was developed by Barbie et al. [4], the pathway scores for a particular sample can be calculated using only the expression values for that sample. Instead of ranking the genes based on their correlation with the class label, the genes are ranked based on their absolute expression (so the ranking can be different for each sample). [4] The *GSVA* library was downloaded, and the *gsva()* function was used with the parameter 'method'

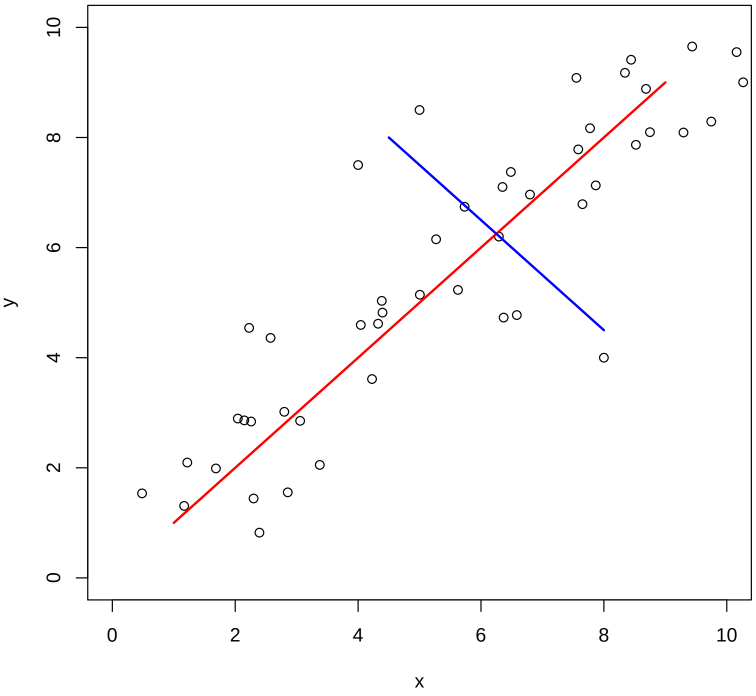

FIGURE 3.1: In this simple 2-dimensional example of PCA, the red line is the first principal component and the blue line is the second principal component. Every principal component is orthogonal to the previous.

equal to "ssgsea" to employ the desired method developed by Barbie et al. [4] This calculation was done before cross-validation, as the result for each sample is only dependent on the values of that sample.

3.2 Cox proportional hazards model

A proportional hazards model models the effect of a unit increase in one of the variables on the hazard score. The hazard score, in this case where we are dealing with survival data, is the risk that a certain patient will die before a certain time. A unit increase in a variable has a multiplicative effect on the hazard score, as stated by the proportional hazards condition. If this condition holds, then the hazard scores can be computed without considering the survival function, which is very convenient. This realization by Sir David Cox meant that survival models can also be trained when the survival function is not available. [7]

3.2.1 Censoring

When a patient dies during the follow-up period (that is, between the time of diagnosis and the end of the research period) it is known exactly when they died and thus how many days they survived after their data of diagnosis. When the follow-up ends and the patient is still alive, we only know that they died or will die after a certain time (the ending date of the follow-up). The Cox proportional hazards model takes this into account, so that also the censored patients can be used in the analysis.

3.2.2 Assumptions

It is assumed that the proportional hazards condition is applicable for the data the Cox model is trained on. However, this is usually not the case, because this would mean that every variable is multiplicatively related to survival outcome and that the relative hazard difference between two patients is constant. Also, especially for biological data, it is very difficult to assess whether this condition holds or not.

3.3 Sparse group lasso

In the 90's of the last century, Tibshirani [40] came up with an alternative to linear regression and called it the 'lasso' to reflect the fact that it works similar to the loop of rope in that it selects only the variables that are important for the class label and gives the un-important variables a coefficient equal to zero. Simon et al [36] (Tibshirani is actually one of the authors of this paper) developed a version of the lasso that creates groups of variables based on prior knowledge about variables that belong together (in this case, all data from one omics source). Their method allows for both intra- and inter-group sparsity; a group as a whole can have a different coefficient from another group, but also the genes within a group can have different coefficients.

3.4 Uno's C-index

After regression of the testing set, we want to be able to compare the result of regression, i.e. the hazard score, with the actual survival data. This is not so straightforward however, since the hazard score can be seen as a probability that a patient will die sooner rather than later, and the survival outcome is the number of days that the patient actually survived. These different data types can be compared by using the concordance index, or C-index, as developed by Heagerty et al [16], see formula 3.1. This index gives the probability that a patient j with a lower survival time than another patient i gets a higher hazard score by the trained model.

$$C := \Pr(n_j > n_i | T_j < T_i) \tag{3.1}$$

C is the C-index, n is the hazard score that was assigned by the model, and T is the number of days that the patient (i or j) survived. In the case of right censoring, one can use a truncated version of the C-index. See formula 3.2.

$$C := \Pr(n_j > n_i | T_j < T_i, T_j \leq \tau) \tag{3.2}$$

The same variables as in the previous formula are used, plus τ which stands for the censoring time. Uno et al [41] developed a version of the C-index that does not rely on the assumptions of a Cox proportional hazards model and also does not simply throw away censored data. See formula 3.3 for Uno's C-index.

$$C_{Uno} := \frac{\sum \delta I(T_j < T_i) I(n_j > n_i)}{\sum \delta I(T_j < T_i)} \tag{3.3}$$

δ is shown below, I is the indicator function that takes on the value 1 if the condition inside the brackets is met and 0 otherwise.

$$\delta = \frac{\Delta_j}{\tilde{G}(\tilde{T}_j)^2} \tag{3.4}$$

δ is the inverse probability that observation j is censored. Δ_j is indicates whether the observation is censored, $\tilde{T}$ is the censored survival time, and $\tilde{G}(\cdot)$ estimates the survival function for the censored survival time (Kaplan-Meier estimator [23]).

3.4.1 Interpretation of the C-index

In the absence of censoring, the C-index is equivalent to a metric that is more widely used in data mining, the area under the receiver-operating curve (AUROC, sometimes shortened to AUC). In the case of continuous class labels, they can be made binary by choosing a certain threshold τ that divides the data into 'lower than τ' and 'equal or larger than τ'. If classification is performed using this threshold, the false positive ration (FPR) and the true positive ratio (TPR) can be plotted as one point in a curve. Multiple thresholds make a continuous curve, that is called the receiver-operating curve (ROC). The integral of, or the area under, the ROC can be interpreted as the probability that a sample with class label 1 will be assigned higher than a sample with class label 0. [15] For both the C-index and the AUC it goes that a value of 0.5 is equal to the prediction performance that can be expected if class labels are assigned by chance. Lower than 0.5 is thus worse than assigning class labels randomly. What actually constitutes a good or a bad AUC or C-index depends heavily on the data and class labels under consideration.

3.5 Training models in a cross-validated manner

10-fold cross validation was performed 10 times for each diseases-method combination. See figure 3.2 for an explanation on 4-fold cross-validation. During each iteration a C-index was calculated. The C-indices were averaged per cross-validation, and the resulting 10 averages were plotted in a box-plot. The cross-validation was parallelized using the *doMC* and *foreach* libraries. For each iteration, the random seed was set to a known number (the product of a constant number, picked at random once and not changed afterwards, and the iterator) using Pierre L'Ecuyer's implementation. [26] This ensures that the same folds were chosen for each method. In the inner for-loop, we iterate over the ten different folds, each time choosing one of them as testing set and the other nine as training set. It is important to choose the same folds when comparing the different methods, because the choice of testing and training set can have an influence on prediction performance. By using known seeds, this bias is corrected for. For the next step, the original samples x genes matrix needs to be converted to a samples x pathways matrix if the current method is dependent on class labels. This has to be done again for every new training set. Then, a model is trained using the *cvSGL()* function in the *SGL* library. In this 10-fold cross-validated version of the sparse group lasso ten Cox proportional hazards models are trained, using the survival data of the training set, by setting the parameter "type" equal to cox. "alpha" was set to 0.5 and the groups were set to the three types of omics data used for this research (mRNA, CNV and methylation). After training these ten models, the *predictSGL()* function was used to test the best model using the testing set and its class labels. The best model in this case is the model with the highest negative log likelihood (the "lldiff" feature of the resulting cvSGL object contains all negative log likelihoods, so the index of the maximum value of this vector was chosen). Finally, Uno's implementation of the C-index was used (*UnoC()* function of the *SurvAUC* package) to calculate the result.

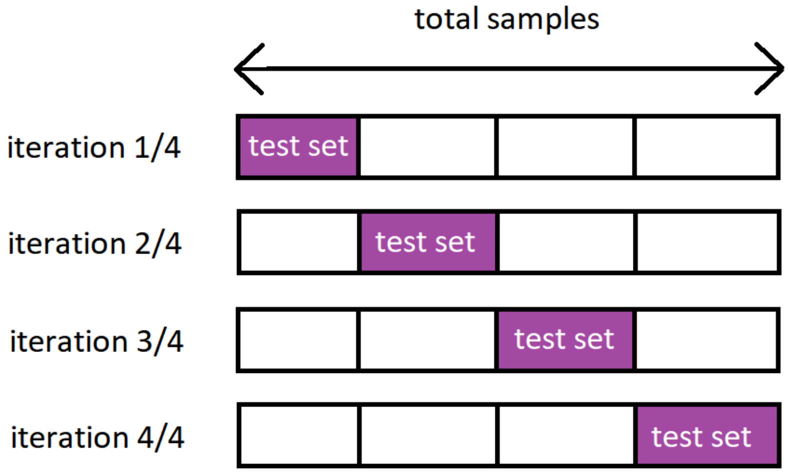

FIGURE 3.2: 4-fold cross-validation explained. In each iteration, a different set of samples is used as the test set and is thus left out of the training of the model. In each iteration, the test set is used to assess the prediction performance of the model.

Chapter 4

Results

4.1 Breast cancer

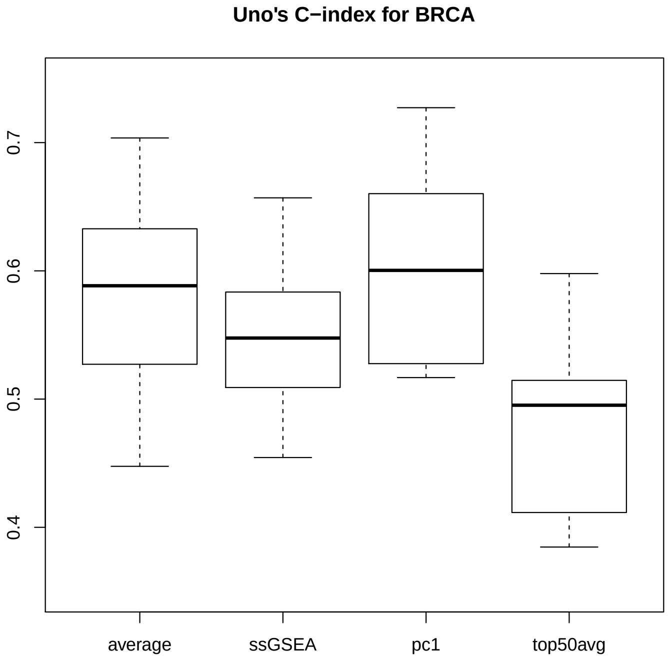

FIGURE 4.1: Results of 10 times 10-fold cross-validation in the form of Uno's C-index on 214 breast cancer patients. The same patients were used for all methods.

method	C-index
average	0.587
ssGSEA	0.570
pc1	0.631
top50	0.487

TABLE 4.1: Median values for the C-index for BRCA

The breast cancer dataset consists of 214 patients with mRNA, methylation and CNV data. 184 (86%) of these patients were censored. The median survival time of all uncensored patients is 1556 days. Cross-validated Uno's C-indices were obtained for the average, ssGSEA, pc1 and top 50 average methods. Box-plots show the average

C-indices of 10-fold cross-validation (averaged over the folds, per repetition) in figure 4.1. All C-indices fall between 0.35 and 0.70. While the median values for the three methods differ, there is a large overlap in the total distribution of C-indices which indicates that there is no significant difference in prediction performance between these methods for the breast cancer data used for this research.

4.1.1 Most important pathways

Based on median values, the first principal component method seems to perform the best with a median C-index of 0.631. Therefore, an overall Cox model was trained on all BRCA data to investigate which pathways were assigned the highest coefficients. In figure 4.2 the coefficients of this overall model can be seen for the pc1 values per pathway for each omics data source. Because a sparse group lasso was used to train the model, there is a global difference in coefficient scale between the different omics data sources. mRNA data is clearly assigned a lower weight than the other omics data sources.

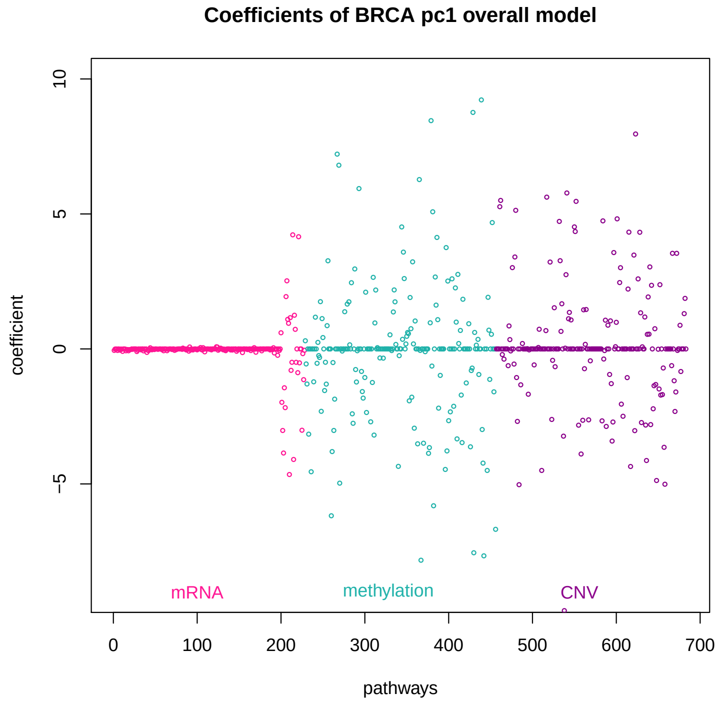

FIGURE 4.2: The coefficients assigned to the different pathways by the final BRCA model. For each omics data source, there are 229 pathways.

Description of top four most important pathways

To get an idea of which pathways are assigned the highest coefficients overall, the pathway coefficients were summed over all omics data sources (see figure 4.3). The absolute values of the summed pathway activities were ordered from highest to lowest, and the top four pathways are indicated with bigger, coloured dots in the plot. In the legend their KEGG indicators can be found. Pathway **00300** is the biosynthesis pathway for the amino acid lysine. Changes in lysine modification in histones and other proteins has been associated with breast cancer. [29] **00532** is glycosaminoglycan biosyntesis. Glycosaminoglycans are polysaccharides that can be attached to proteins to form a glycoprotein, but can also exist freely in the extracellular matrix. Glycosaminogalycans have been associated with metastasis in breast cancer; by changing the micro-environment around the tumour they can favour metastasis. [33] **00062** is the fatty acid elongation pathway. Fatty acid chains, when combined into a fat, make up the cell membrane and intracellular membranes. González-Bengtsson et al [12] found that the expression of one particular fatty acid elongase (an enzyme that performs elongation), Elovl2, has a higher expression in breast cancer cells. **00512** is the biosynthesis of Mucin type O-glycans. O-glycosylation is the modification of serine or threonine residues by attaching a glycan (sugar molecule) to the oxygen atom in these residues. Mucin O-glycans are a specific type of O-glycan that can be branched. Their glycoproteins are mainly found on cell surfaces and in bodily fluids. [1] MUC1 is one particular glycoprotein that has been found to be over-expressed in breast cancer. [37]

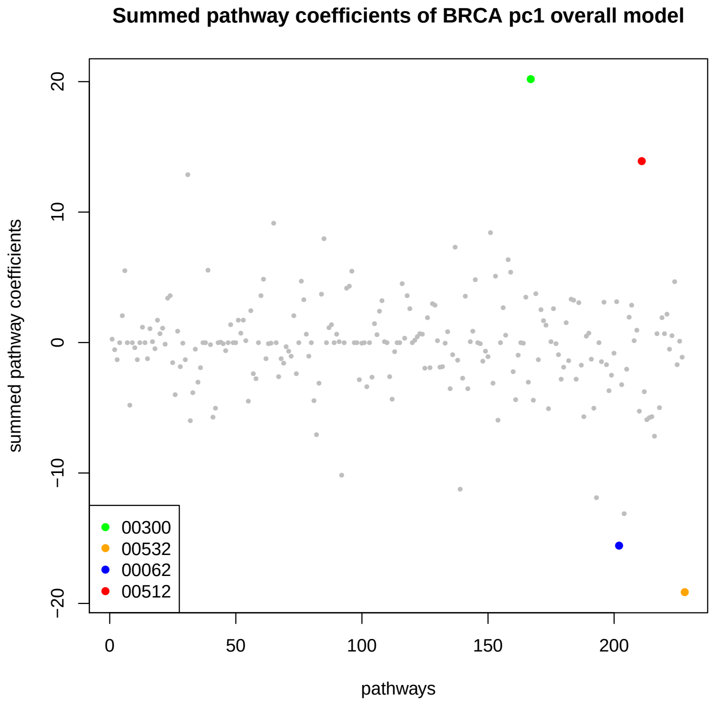

FIGURE 4.3: The coefficients for the 229 pathways, summed over the omics data as seen in figure 4.2.

[1] http://www.kegg.jp/kegg-bin/show_pathway?map=map00512&show_description=show, accessed on 19th of July 2017

4.2 Glioblastoma multiforme

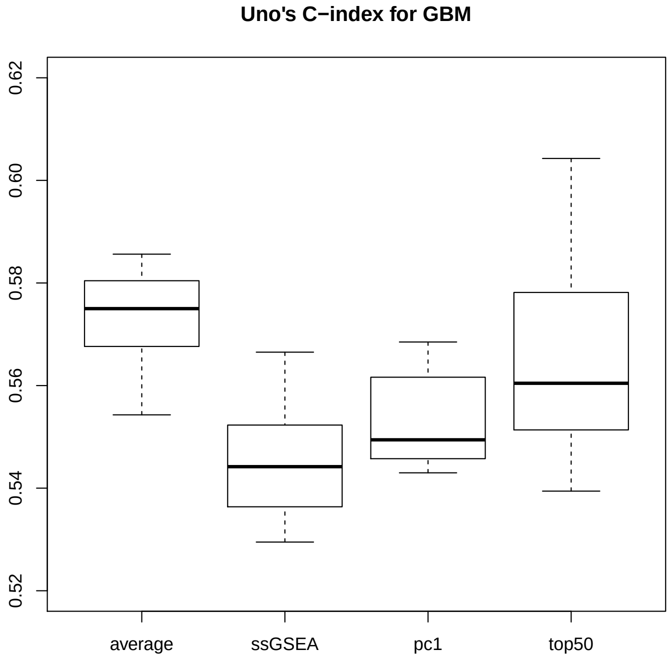

FIGURE 4.4: Results of 10 times 10-fold cross-validation in the form of Uno's C-index on 269 glioblastoma patients. The same patients were used for all methods.

method	C-index
average	0.574
ssGSEA	0.548
pc1	0.563
top50	0.565

TABLE 4.2: Median values for the C-index for GBM

The glioblastoma dataset consists of 269 patients with mRNA, methylation and CNV data. 67 (25%) of these patients were censored. The median survival time of all uncensored patients is 393 days. Cross-validated Uno's C-indices were obtained for the average, ssGSEA, pc1 and top 50 average methods. Box-plots show the average C-indices of 10-fold cross-validation (averaged over the folds, per repetition) in figure 4.1. All C-indices fall between 0.53 and 0.61. Again, as with the BRCA results, there is a large overlap between the C-index distributions of the different methods.

4.2.1 Most important pathways

Based on median values, the average method seems to perform the best with a median C-index of 0.574. Therefore, an overall Cox model was trained on all GBM data to investigate which pathways were assigned the highest coefficients. In figure 4.5 the coefficients of this overall model can be seen for the average values per pathway for each omics data source. Because a sparse group lasso was used to train the model, there is a global difference in coefficient scale between the different omics data sources. However, for GBM it seems like the different omics data did not get a very different group weight, as was the case for the BRCA data.

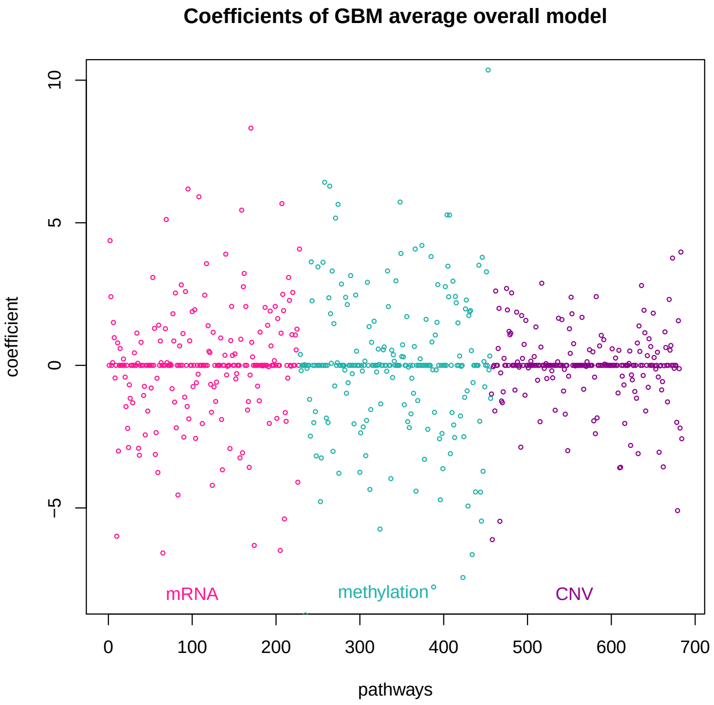

FIGURE 4.5: The coefficients assigned to the different pathways by the final GBM model. For each omics data source, there are 229 pathways.

Description of top four most important pathways

Just like was done for BRCA, the pathway coefficients were summed over all omics data sources (see figure 4.6) and the absolutely highest scoring pathways are indicated with bigger coloured dots. **03010** is the set of genes that make up the ribosome. Yong et al [43] found that over-expression of two ribosomal proteins in particular, RPS11 and RPS20, predicts poor survival in newly diagnosed GBM patients. **05216** is a pathway that is important in the origin of thyroid cancer. Please note that there was no specific glioblastoma pathway in the KEGG pathway list that was used in this research. There is however a glioma pathway, but this one ended up at the 97th place in the ordered pathways. A hypothesis has been raised that proposes a potential role of thyroid hormones on glioblastoma formation. [32] **04122** is the sulfur relay system, that consists of the ubiquitin pathway and sulfur transfer. Both are involved with protein modifications that are highly associated with cellular regulation. Proteins that are tagged with poly-ubiquitin chains are degraded in the proteasome. One of the strategies that a cancerous cell can use, is to make sure tumour suppressors (proteins that protect against cancerous activity) are being degraded by this mechanism of ubiquitination. Lignitto et al [28] have shown that one particular ubiquitin ligase (a protein that attaches ubiquitin groups to other proteins), praja2, targets an element of the Hippo signaling pathway. This down-regulates this tumour-suppressing pathway and promotes glioblastoma growth *in vivo*. **00072** is the pathway responsible for synthesis and degradation of ketone bodies. Ketone bodies are three small water-soluble ketones (compounds with a double-bonded oxygen). They are produced by the liver during periods of starvation (also for instance on low carb diets, and on prolonged intense exercise) as an alternative energy source. Interestingly, most cancerous cells and in particular glioblastoma cells can not use ketone bodies as a source of energy. [13] That's why one of the treatments that is sometimes suggested to glioblastoma patients is to go on a restricted calorie diet to promote ketone body formation. This restricts the supply of glucose, so that the only energy source available is ketone bodies. This ensures that the healthy cells still get energy but the tumour is starved. [30] So the high activity of this pathway is probably an effect of a ketogenic diet.

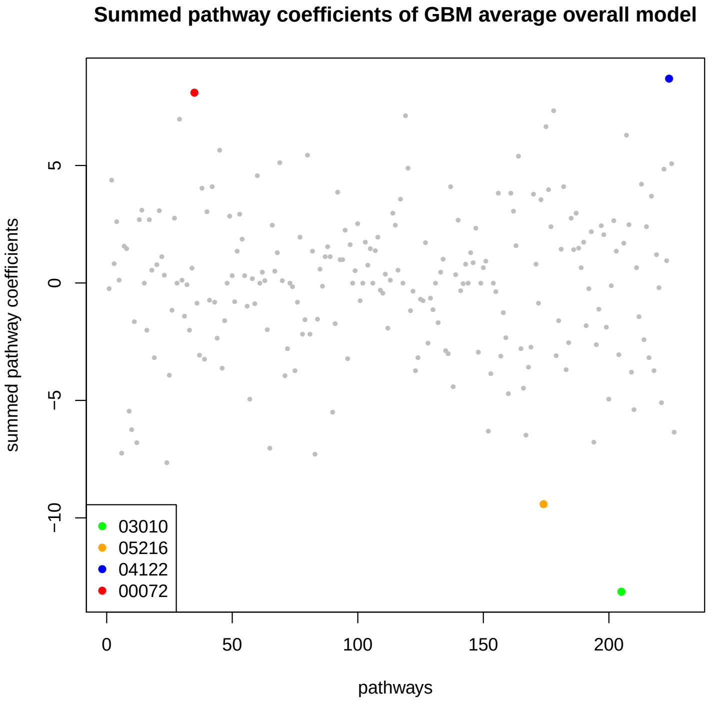

FIGURE 4.6: The coefficients for the 229 pathways, summed over the omics data as seen in figure 4.5.

Chapter 5

Discussion

5.1 Correcting batch effects

The possibility to not correct batch effect was considered, as opposed to only correcting gene expression. The choice was made to still correct the batch effect, even if it was only for mRNA. However, because different omics samples from the same patient still come from the same hospital, if there is a batch effect that arises from the hospital where the sample came from, it will probably still show up in the other omics data sources.

5.2 Leaving miRNA data out

We had mRNA, CNV, methylation and miRNA data at our disposal. However, the choice had to be made to leave out miRNA. Even though adding miRNA data would have made sense on a biological level, this would mean that only data for 61 BRCA-patients could have been used for cross-validation (this would have been using miRNA-seq data, since there was no micro-array data for miRNA for BRCA). That is too little for reliable prediction and this would have meant we would have been overfitting the data. For GBM, none of the patients were in all omics data sets.

5.3 The pathway list used

For this research 229 KEGG pathways were used. It could be interesting to repeat this work with pathway lists from other sources, like Reactome [9] or ConsensusPathDB [22].

5.4 Using topological info gives a more realistic view

In this work, gene sets were used, which are simply lists of all the genes that are associated with a certain pathway. However, pathways are much more intricate than just some proteins that interact with each other. Not all proteins in one pathway interact with all the others, and the ones that do have an interaction interact in a certain way (inhibition and stimulation are examples of interaction types), in a certain direction (for example, one protein can inhibit another one, but not the other way around) and with a certain interaction strength. Work has been done on integrating undirected and directed pathways with omics data. One promising example would be the work of Koumakis et al [24] where they extract so-called 'sub-graphs' from pathways, determine their status (active or inactive) and use these statuses as new features.

5.5 Pathways are not static

In this work, pathways are assumed to be static. However, re-routed pathways are one of the hallmarks of cancer [35], so especially since we are working with cancer data here this is something to be considered. Gene fusion is one of the causes of pathways changing, where interacting domains from different proteins are combined in a new protein. This alters the way in which this protein interacts with other molecules. Latysheva et al [25] have found that, compared to normal genes, fusion genes tend to have a greater number of binding partners and are more important 'hubs' in the protein interaction network. Gene fusion and the resulting change in pathways is thus a significant thing to consider in further work. There is no easily accessible database of different pathways for different disease states available, but work on inferring pathways from data for different types of cancer has been done, among others, by Kaiser et al. [21]

5.6 In complex traits, associated genomic regions tend to be spread out

In this work, we have been focusing on using pathways to condense the data in such a way that we get more info with less variables. This is based on the idea that there are core genes and core pathways that are highly associated with a trait and thus summarizing the data in such a way that similar genes are put together would be beneficial. However, recent work by Boyle et al [5] indicates that while this might be true for Mendelian diseases (diseases in which one locus is responsible for the occurrence of the disease), for complex traits this is usually not the case. In complex disease, the biggest part of the heritability is spread throughout the genome and not near genes with disease-specific functions. Another striking conclusion that has been made since genome-wide association studies (GWAS) became more commonplace, is that most variants found for complex traits affect gene regulation and not the protein-coding part of a gene. In this work, we have also looked at gene regulation in the form of copy number variation and gene methylation.

5.7 The reliability of clinical data

One patient can choose when they want to go to the doctor, if they suspect a health issue. Let's say they have cancer but don't know it yet. The same person might go as soon as possible, while another person with very similar omics data might wait for a few weeks. This can skew the predicted survival outcome and make the model less reliable. But this might not be such a large issue, because the C-index implementation by Uno only considers the order in which the patients died and not their exact survival time.

5.8 The epigenome goes beyond DNA methylation

Epigenetic traits (heritable traits that can not be explained by the DNA sequence) like DNA methylation and histone modification are important for cellular regulation. Changes in the epigenome are associated with cancer and are thus important to consider when working with cancer data. [8] In this work, we have considered

DNA methylation, but there is so much more when it comes to epigenetics: researchers now recognize four different DNA modifications and 16 different histone modifications. Cohen et al [6] have developed a method to assess activity of several epigenetic pathways based on gene expression, copy number variation and DNA methylation changes.

5.9 Gradient boosting as an alternative to Cox proportional hazards model

Mayr et al [31] propose a new way of predicting survival outcome where they directly boost the concordance index. This results in better prediction performance than with a lasso penalized Cox regression model. In gradient boosting first a very simple model is trained on a subset on the samples (and sometimes also a subset of the variables). After that, the deviation of the predictions with the actual class label is considered. This deviation is minimized in the next step, where a new simple model is trained. This goes on until a set number of iterations or until the desired outcome is considered good enough. The advantage of boosting the C-index is that selection of the model is done by directly choosing the model with the best C-index, instead of looking at the log likelihood of the model and afterwards calculating the C-index as is done in this work. Another advantage compared to Cox regression is that the proportional hazard assumption doesn't have to be true. The reason that the choice was still made to use Cox regression in this work is because it is easier to implement and a Cox model can still be useful albeit not statistically correct. Furthermore, the main goal of this research was to be able to compare the different methods for data summarization on pathway level and not to find the model with the best overall prediction performance.

5.10 Combining data sets leads to better prediction performance

In this work, only data form TCGA was used. However, van Vliet et al [42] show that combining breast cancer data sets has a beneficial effect in 73% of the cases. This can be partially attributed to larger data sets, but also to the fact that different data sets represent different parts of the patient population. Further work would benefit from a combination of data from different sources. However, TCGA is one of the few data repositories where several omics data is available for patients.

Chapter 6

Conclusion

This research has shown that summarizing omics data on pathway level instead of leaving it on gene level can be beneficial in a machine learning setting. Prediction performances of a C-index around 0.6 were reached for BRCA and GBM patients. For both diseases considered here the top four most important pathways were found in literature to be associated with the disease in question. This shows that not only could survival be predicted for BRCA and GBM patients, but the final models were also interpretable in a biological sense. However, there are some things that could be considered for further work: the data could have been processed differently, and directed pathways could have been used instead of mere gene lists. The progress made here can be viewed as a proof of concept when it comes to integration of diverse omics data and pathway information. Its feasibility has been shown and the methods can be used to improve further attempts at predicting several features based on omics data.

Bibliography

[1] Cory Adamson et al. "Expert Opinion on Investigational Drugs Glioblastoma multiforme : a review of where we have been and where we are going Glioblastoma multiforme : a review of where we have been and where we are going". In: 3784.January 2016 (2009). DOI: 10.1517/13543780903052764.

[2] Amin Allahyar and Jeroen De Ridder. "FERAL: Network-based classifier with application to breast cancer outcome prediction". In: *Bioinformatics*. 2015. ISBN: 13674811 (Electronic). DOI: 10.1093/bioinformatics/btv255.

[3] Sepideh Babaei et al. "Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion". In: *BMC Bioinformatics* 14 (2013). URL: http://www.biomedcentral.com/1471-2105/14/29.

[4] David A Barbie et al. "Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1." In: *Nature* 462.7269 (2009), pp. 108–12. ISSN: 1476-4687. DOI: 10.1038/nature08460. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2783335&tool=pmcentrez&rendertype=abstract.

[5] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. "An Expanded View of Complex Traits: From Polygenic to Omnigenic". In: *Cell* 169.7 (2017), pp. 1177–1186. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.05.038. URL: http://dx.doi.org/10.1016/j.cell.2017.05.038.

[6] Adam L Cohen et al. "Genomic pathway analysis reveals that EZH2 and HDAC4 represent mutually exclusive epigenetic pathways across human cancers". In: *BMC Medical Genomics* 6.1 (2013), p. 35. ISSN: 1755-8794. DOI: 10.1186/1755-8794-6-35. URL: http://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-6-35.

[7] David R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220.

[8] Mark A. Dawson and Tony Kouzarides. "Cancer epigenetics: From mechanism to therapy". In: *Cell* 150.1 (2012), pp. 12–27. ISSN: 00928674. DOI: 10.1016/j.cell.2012.06.013. URL: http://dx.doi.org/10.1016/j.cell.2012.06.013.

[9] Antonio Fabregat et al. "The reactome pathway knowledgebase". In: *Nucleic Acids Research* 44.D1 (2016), pp. D481–D487. ISSN: 13624962. DOI: 10.1093/nar/gkv1351.

[10] a Feelders, H Daniels, and M Holsheimer. "Methodological and practical aspects of data mining". In: *Information & Management* 37.5 (2000), pp. 271–281. ISSN: 03787206. DOI: 10.1016/S0378-7206(99)00051-8.

[11] Patrice Godard and Jonathan Van Eyll. "Pathway analysis from lists of microRNAs: Common pitfalls and alternative strategy". In: *Nucleic Acids Research* 43.7 (2015), pp. 3490–3497. ISSN: 13624962. DOI: 10.1093/nar/gkv249.

[12] Amanda González-Bengtsson et al. "Estrogen enhances the expression of the polyunsaturated fatty acid elongase Elovl2 via ERα in breast cancer cells". In: *PLoS ONE* 11.10 (2016), pp. 1–18. ISSN: 19326203. DOI: 10.1371/journal.pone.0164241.

[13] Maja M. Grabacka et al. "Fenofibrate induces ketone body production in melanoma and glioblastoma cells". In: *Frontiers in Endocrinology* 7.FEB (2016), pp. 1–13. ISSN: 16642392. DOI: 10.3389/fendo.2016.00005.

[14] Frederik Gwinner et al. "Network-based analysis of omics data: The LEAN method". In: *Bioinformatics* (2016), btw676. DOI: 10.1093/bioinformatics/btw676. URL: http://bioinformatics.oxfordjournals.org/lookup/doi/10.1093/bioinformatics/btw676.

[15] Blaise Hanczar et al. "Small-sample precision of ROC-related estimates". In: 26.6 (2010), pp. 822–830. DOI: 10.1093/bioinformatics/btq037.

[16] P J Heagerty and Y Zheng. "Survival model predictive accuracy and ROC curves 1". In: *Biometrics* 61.0006-341X (Print) (2005), pp. 92–105. ISSN: 0006-341X. DOI: 10.1111/j.0006-341X.2005.030814.x.

[17] Marc Hulsman, Christos Dimitrakopoulos, and Jeroen De Ridder. "Scale-space measures for graph topology link protein network architecture to function". In: *Bioinformatics* (2014). ISSN: 14602059. DOI: 10.1093/bioinformatics/btu283.

[18] Seungwoo Hwang. "Comparison and evaluation of pathway-level aggregation methods of gene expression data". In: *BMC Genomics* 13.Suppl 7 (2012), S26. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-S7-S26. URL: http://www.biomedcentral.com/1471-2164/13/S26/S26.

[19] Ahmedin Jemal et al. "Cancer Statistics , 2009 BOTH SEXES FEMALE BOTH SEXES ESTIMATED DEATHS". In: *CA Cancer J Clin* 59.4 (2009), pp. 1–25. ISSN: 1542-4863. DOI: 10.1002/caac.20073.Available.

[20] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127. ISSN: 14654644. DOI: 10.1093/biostatistics/kxj037.

[21] Jacob L. Kaiser, Cassidy L. Bland, and David J. Klinke. "Identifying causal networks linking cancer processes and anti-tumor immunity using Bayesian network inference and metagene constructs". In: *Biotechnology Progress* (2016). ISSN: 15206033. DOI: 10.1002/btpr.2230.

[22] Atanas Kamburov et al. "The ConsensusPathDB interaction database: 2013 Update". In: *Nucleic Acids Research* 41.D1 (2013), pp. 793–800. ISSN: 03051048. DOI: 10.1093/nar/gks1055.

[23] E. L. Kaplan and P. Meier. "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452.

[24] Lefteris Koumakis et al. "MinePath: Mining for phenotype differential subpaths in molecular pathways". In: *PLoS Computational Biology* Accepted.Oct. To appear (2016), pp. 1–40. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005187.

[25] NS Latysheva et al. "Molecular Principles of Gene Fusion Mediated Rewiring of Protein Interaction Networks in Cancer." In: *Mol Cell.* 63.4 (2016), pp. 579–92. ISSN: 10972765. DOI: 10.1016/j.molcel.2016.07.008. URL: http://dx.doi.org/10.1016/j.molcel.2016.07.008.

[26] Pierre L'Ecuyer. "Tables of linear congruential generators of different sizes and good lattice structure". In: *Mathematics of Computation* 68.225 (1999), pp. 249–261. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-99-00996-5. URL: http://www.ams.org/journal-getitem?pii=S0025-5718-99-00996-5.

[27] Maxwell W. Libbrecht and William Stafford Noble. "Machine learning applications in genetics and genomics". In: *Nature Reviews Genetics* 16.6 (2015), pp. 321–332. ISSN: 1471-0056. DOI: 10.1038/nrg3920. URL: http://www.nature.com/doifinder/10.1038/nrg3920.

[28] Luca Lignitto et al. "Proteolysis of MOB1 by the ubiquitin ligase praja2 attenuates Hippo signalling and supports glioblastoma growth." In: *Nature communications* 4.May (2013), p. 1822. ISSN: 2041-1723. DOI: 10.1038/ncomms2791. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3674242&tool=pmcentrez&rendertype=abstract.

[29] Lanxin Liu et al. "Genetic alterations of histone lysine methyltransferases and their significance in breast cancer." In: *Oncotarget* 6.4 (2015), pp. 2466–82. ISSN: 1949-2553. DOI: 10.18632/oncotarget.2967. URL: http://www.ncbi.nlm.nih.gov/pubmed/25537518%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4385864.

[30] Joseph Maroon et al. "Restricted calorie ketogenic diet for the treatment of glioblastoma multiforme." In: *Journal of Child Neurology* 28.8 (2013), pp. 1002–8. ISSN: 1708-8283. DOI: 10.1177/0883073813488670. URL: http://www.ncbi.nlm.nih.gov/pubmed/23670248.

[31] Andreas Mayr, Benjamin Hofner, and Matthias Schmid. "Boosting the discriminatory power of sparse survival models via optimization of the concordance index and stability selection." In: *BMC bioinformatics* 17 (2016), p. 288. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1149-8. URL: http://www.ncbi.nlm.nih.gov/pubmed/27444890%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4957316.

[32] Pawel Nauman. "Thyroid hormones in the central nervous system (CNS) and their effect on neoplasm formation, particularly on the development and course of glioblastoma multiforme — research hypothesis". In: *Endokrynologia Polska* 64.4 (2013), pp. 319–27. ISSN: 0423-104X. DOI: 10.5603/EP. URL: http://www.ncbi.nlm.nih.gov/pubmed/24002961.

[33] Dragana Nikitovic et al. "The motile breast cancer phenotype roles of proteoglycans/glycosaminoglycans". In: *BioMed Research International* 2014 (2014). ISSN: 23146141. DOI: 10.1155/2014/124321.

[34] Hiroyuki Ogata et al. "KEGG: Kyoto encyclopedia of genes and genomes". In: *Nucleic Acids Research* 27.1 (1999), pp. 29–34. ISSN: 03051048. DOI: 10.1093/nar/27.1.29.

[35] Liem Minh Phan, Sai-Ching Jim Yeung, and Mong-Hong Lee. "Cancer metabolic reprogramming: importance, main features, and potentials for precise targeted anti-cancer therapies." In: *Cancer biology & medicine* 11.1 (2014), pp. 1–19. ISSN: 2095-3941. DOI: 10.7497/j.issn.2095-3941.2014.01.001. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3969803&tool=pmcentrez&rendertype=abstract.

[36] Noah Simon et al. "A Sparse-Group Lasso". In: *Journal of Computational and Graphical Statistics* 22.2 (2013), pp. 231–245. ISSN: 1061-8600. DOI: 10.1080/10618600.2012.681250. URL: http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.681250.

[37] Sarah J. Storr et al. "The O-linked glycosylation of secretory/shed MUC1 from an advanced breast cancer patient's serum". In: *Glycobiology* 18.6 (2008), pp. 456–462. ISSN: 09596658. DOI: 10.1093/glycob/cwn022.

[38] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. "The cancer genome". In: *Nature* 458.7239 (2009), pp. 719–724. ISSN: 0028-0836. DOI: 10.1038/nature07943. URL: http://www.nature.com/doifinder/10.1038/nature07943.

[39] Rstudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA, 2016. URL: http://www.rstudio.com/.

[40] Robert Tibshirani. *Regression selection and shrinkage via the lasso*. 1996. DOI: 10.2307/2346178. URL: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574.

[41] Hajime Uno et al. "On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data". In: *Biophysical Chemistry* 257.5 (2005), pp. 2432–2437. ISSN: 15378276. DOI: 10.1016/j.immuni.2010.12.017.Two-stage.

[42] Martin H van Vliet et al. "Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability." In: *BMC genomics* 9 (2008), p. 375. ISSN: 1471-2164. DOI: 10.1186/1471-2164-9-375.

[43] William H. Yong et al. "Ribosomal proteins RPS11 and RPS20, two stress-response markers of glioblastoma stem cells, are novel predictors of poor prognosis in glioblastoma patients". In: *PLoS ONE* 10.10 (2015), pp. 1–19. ISSN: 19326203. DOI: 10.1371/journal.pone.0141334.