Bioinformatics Program

Technical University of Munich

Ludwig-Maximilians-Universität München

Bachelor's Thesis in Bioinformatics

Computational Drug Repurposing Approach for Transcription Factors using the Connectivity Map Dataset

Thomas Eska

Bioinformatics Program

Technical University of Munich

Ludwig-Maximilians-Universität München

Bachelor's Thesis in Bioinformatics

Computational Drug Repurposing Approach for Transcription Factors using the Connectivity Map Dataset

Computergestützter Drug-Repurposing Ansatz für Transkriptionsfaktoren unter Verwendung des Connectivity Map Datensatzes

Author: Thomas Eska
Supervisor: Prof. Jan Baumbach
Advisors: Gihanna Galindez,
Dr. Tim Kacprowski
Submitted: October 15th 2020

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Abstract

In order to reduce the costs and risks of designing new drugs, researchers are moving towards drug repurposing as a means of finding new therapeutics for diseases. By using already approved drugs on new diseases, several time-consuming steps of the drug development process can be skipped. While there have been several successful knowledge-based attempts at drug repurposing, computational methods can process a large amount of data and propose a multitude of new treatments in a short amount of time.

In this thesis, a computational approach to drug repurposing through examination of disease-specific gene regulatory networks is introduced. By reverse-engineering gene regulatory networks using gene-expression data from the Connectivity Map dataset, activity of transcription factors can be understood. Comparison between networks allows the finding of shared disease mechanism, which in turn identifies drugs used in treatment of one disease as repurposing candidates for another disease.

The method used was able to recover many studied associations between diseases and shows promise in the identification of disease mechanisms. However, there are concerns about the overlap of input data for network inference between different diseases, since networks are partially based on the same data.

Um die notwendigen Kosten und Risiken für die Entwicklung neuer Medikamente zu senken, wenden sich immer mehr Forscher dem Drug Repurposing zu. Indem bereits auf dem Markt zugelassene Medikamente als Kandidaten für andere Krankheiten erforscht werden, können einige zeit- und kostenintensive Schritte der Forschungsarbeit übersprungen werden. Während Methoden, welche auf klinischem Wissen basieren, bereits erfolgreich zum Drug Repurposing verwendet wurden, bieten computergestützte Methoden die Möglichkeit, viele Medikamente in sehr kurzer Zeit zu testen.

In dieser Arbeit wird ein computergestützter Ansatz zum Drug Repurposing vorgestellt. Dafür werden krankheitsspezifische Genregulationsnetzwerke aus Genexpressionsprofilen aus dem Connectivity Map Datensatz inferriert und analysiert. Dies gibt Einblick in die Aktivität von Transkriptionsfaktoren und anderen regulativen Elementen. Durch den Vergleich verschiedener krankheitsspezifischer Netzwerke können Krankheitsmechanismen, welche von verschiedenen Krankheiten geteilt werden, identifiziert werden. Dadurch können Medikamente, welche bereits zur Behandlung einer dieser Krankheiten genutzt werden, als Kandidaten für Drug Repurposing markiert werden.

Die vorgestellte Methode ist in der Lage viele bereits bekannte Verwandschaften zwischen Krankheiten wiederzufinden. Dies zeigt ihre Fähigkeit Krankheitsmechanismen zu identifizieren. Allerdings stellt der Überlapp zwischen den eingesetzten Genexpressionsprofilen ein Problem der Methode dar, da die Genregulationsnetzwerke teilweise auf den selben Daten basieren.

Acknowledgment

I would like to thank my advisor Gihanna Galindez for her invaluable support and guidance. She continuously provided me with answers to my questions and suggestions to solve encountered problems.

Furthermore, I would like to thank Dr. Tim Kacprowski for his reliable feedback and advice throughout my work.

Additionally, I would like to thank Prof. Jan Baumbach for giving me the opportunity to write this thesis and providing a great working environment for students and employees alike.

I am grateful to my family for their continued support and interest in my work.

Contents

List of Tables

List of Figures

List of Abbreviations

CMAP	Connectivity Map
FDA	Food and Drug Administration
FDR	false discovery rate
GRN	gene regulatory network
GSEA	Gene Set Enrichment Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
MI	mutual information
MRMR	maximum relevancy/minimum redundancy
PPI	protein-protein interaction network
TF	Transcription Factor

1 Introduction

1.1 Motivation

Before a new drug is approved and can enter the market, a great amount of time and resources must be invested. A potential drug must first be discovered and tested in a laboratory. After that follow clinical trials, which generally span multiple years. Adams *et al* estimate the complete cost of drug development for a new drug, from discovery to its first market appearance, to be roughly 1.2 billion U.S. dollars [1]. The amount of time required to discover a new drug and ensure its safety and effectiveness can also pose a problem, especially when no other known treatments for a disease are available. A current example is the ongoing SARS-CoV-2 pandemic, which has had more than 32M global cases so far [9]. There is currently no vaccine or effective medication available, which has prompted a global effort to find a working treatment as soon as possible. In order to reduce the time needed to find a cure, many researchers have focused on drug repurposing [10, 41, 45].

Drug repurposing (also called drug repositioning) attempts to lower the cost and time spent on drug development by finding new targets for drugs which are either already approved for market use or in an advanced phase of a clinical trial. The probably most famous case of successful drug repurposing is the usage of sildenafil, also known as viagra, to treat erectile disfunction in men. Originally developed as a treatment for angina, participants in clinical studies quickly noticed the drugs side effect - which then became its main effect [2].
Sometimes drugs, which did not get approved for treatment of a disease, because of their side effects on patients, can be used to treat different, more serious illnesses. An example for this is thalidomide, which used to be prescribed as a treatment for morning sickness in pregnant women from 1957 to 1961. Birth defects in more than 12,000 children were later connected with the intake of the drug, which led to it being taken off the market, and even to the creation of the German "Arzneimittelschutzgesetz", which regulates the approval of pharmaceuticals in Germany [2]. Thalidomide has since been found to be a an effective treatment for leprosy and multiple myeloma and is currently approved for treatment of these two diseases by the US-American Food and Drug Administration (FDA) despite the side effects[17].

In both the examples listed above, the discovery of the drugs other uses was accidental. Nowawdays, there are many ways to systematically search for new uses for existing drugs, both computational and experimental. The main principle, which systematic drug repurposing relies on, is that, if two diseases share a biological mechanism,

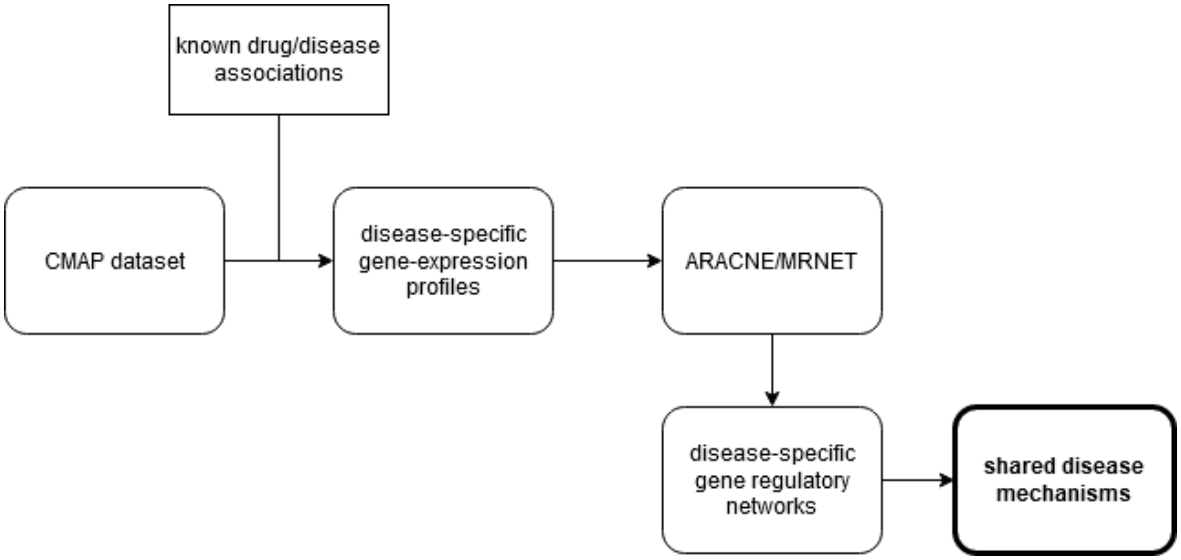

Figure 1.1: Steps used to find shared disease mechanisms. First relevant data is selected from CMAP through known drug/disease associations, the obtained expression profiles are used as input for the ARACNE and MRNET algorithms, together with transcription factor data, to infer GRNs. Comparing GRNs shows shared disease mechanisms.

then a drug that is used to treat one disease can also be used to treat the other. To make use of this, knowledge about disease mechanisms is imperative. A gene regulatory network (GRN) describes the interactions between different genes and gene regulators. Comparing GRNs can give information about the mechanism of different diseases and the regulators involved in them. GRNs can be reverse-engineered from gene-expression data using models based on information theory, correlation, machine learning and more [28].

1.2 Overview

In this thesis, an approach to identify shared mechanisms between diseases for drug repurposing is introduced. After two preliminary approaches using gene expression data to find means of targeting transcription factors, disease-specific GRNs are constructed to discover processes, which are shared by diseases. For this known associations between drugs and diseases were used to extract gene-expression profiles specific to a disease from the Connectivity Map (CMAP) dataset (described in the section Data Sources). These were then used as input to construct GRNs using both the ARACNE and MRNET GRN inference methods. Finally the obtained GRNs were compared to find shared mechanisms between diseases, which would allow for drug repositioning.

1.3 Outline

Chapter 2 gives an overview of Transcription factors and GRNs. Additionally, the concept of mutual information as a method for inferring GRNs and two mutual infor-

mation based inference algorithms are introduced. In Chapter 3, used datasets as well as details of the methods used for drug repurposing are described. Chapter 4 presents results of the drug repositioning workflow and discusses the findings. Finally, Chapter 5 draws conclusions from this thesis and gives prospects for possible future work and extensions.

2 Background

2.1 Transcription Factors

A Transcription Factor (TF) is a protein that regulates the transcription of DNA to RNA. They bind to specific regulatory parts of the DNA and may act as either activators or repressors [26]. TFs are heavily implicated in various diseases, especially in oncology. In fact, TFs account for almost 20% of all oncogenes [25]. Most TFs used to be considered non druggable due to their complexity. With advanced understanding of their structure and function, targeting them has in recent years become possible in a variety of ways [25]. However, these are based on detailed knowledge about the specific protein and its mechanisms. If no such data is available, computational methods can extrapolate some information from gene-expression profiles and GRNs to allow targeting.

2.2 Gene Regulatory Networks

A GRN describes an organism's system of gene regulation. The nodes of a GRN are either regulatory genes, such as TFs, or their interaction partners, while edges describe interactions between genes. By using a network of interactions, a GRN controls the expression of all kinds of genes. Since the expression of regulatory genes is, in general, determined by the expression of other regulatory genes, a complex network of genes is formed. This results in the controlled expression of various genes, which ultimately determines a cells structure, health and more [19]. Knowledge about the regulation of genes may show new means to influence the environment of the cell, for example, for medical treatment. GRNs can be reverse-engineered mathematically by inferring interactions between different genes from gene-expression data. A multitude of algorithms based on different mathematical principles have been published. GRNs can be inferred using regression-based methods [18], probabilistic graphical models, such as bayesian networks [14], or gradient boosting [42]. There are also several methods based on mutual information (MI).

2.2.1 Mutual Information

In information theory, MI between two random variables describes the information which can be gained from one variable about the other. For example, if the variable X were to describe the result of a six-sided die roll and the variable Y were to describe whether the result is even or odd, we would be able to gain information about Y by

looking at the result of X and vice-versa. If, for example, $x = 1$, then we know that $y = ODD$. If $y = EVEN$, then we know that $x\epsilon\{2, 4, 6\}$. If the MI between two variables is zero, they are statistically independent. The MI $I(X;Y)$ for two random variables X and Y is defined as follows:

$$I(X;Y) = \sum_{x\epsilon\chi} \sum_{y\epsilon\Upsilon} P(x,y) log \frac{P(x,y)}{P(x)P(y)} \tag{2.1}$$

Where χ and Υ are the values X and Y can take, $P(X,Y)$ is the joint distribution of X and Y. $P(X)$ and $P(Y)$ are the marginal distributions of X and Y [27].

In the context of gene-expression, MI between a pair of genes is a measure for the influence these two genes have on each other's expression. Through this, gene interaction in a regulatory network can be inferred. In this thesis, two MI based methods, ARACNE and MRNET, were used to infer GRNs. First, both methods calculate the MI for each pairwise combination of genes in the given expression profiles. Afterwards, they apply the gained information in different ways. Both methods are available as part of the R/Bioconductor package *minet* [32, 35, 39].

2.2.2 ARACNE

The first method, ARACNE, uses the MI between a pair of genes as a weight for the edge connecting them and removes an edge if its weight is below a threshold I_0. Additionally, the edge with the lowest weight of each triplet of edges is removed if the weight difference between it and the second lightest edge is above a threshold W_0, since a high weight difference likely indicates an indirect interaction. The remaining edges form a GRN. For example: Between three genes G_1, G_2 and G_3, there can be a maximum of three edges: (G_1, G_2), (G_1, G_3) and (G_2, G_3). $I(G_x, G_y)$ describes the MI between the genes G_x and G_y. ARACNE checks if $I(G_1; G_2) < I_0$. If true, the weight of this edge is set to zero, meaning it is removed from the GRN. This process repeat for each possible pairing. Assuming all three edges have a weight greater than I_0, the weakest edge will be determined in the next step. If $I(G_1; G_2) < I(G_1; G_3) < I(G_2; G_3)$, then the edge (G_1, G_2) is set to zero if $I(G_1; G_3) - I(G_1; G_2) > W_0$. After this algorithm is applied to each triplet of pairs, ARACNE returns a weighted regulatory network. The weight of the edges can be used as a measure for the scale of influence between the involved genes. Since GRNs are sparse [48], ARACNE aims to reduce false positives by removing a multitude potential edges, which may lead to many false negatives. Because of this, ARACNE's express goal is "not to recover all transcriptional interactions in a genetic network but rather to recover some transcriptional interactions with high confidence" [30].

2.2.3 MRNET

The second method, MRNET, is based on maximum relevancy/minimum redundancy (MRMR) feature selection. For each gene in the input expression profiles, this method ranks all other genes as potential interactors. Genes are ranked by first selecting the gene which shares the highest MI with the target gene. Next, a gene which shares high

MI with the target gene, but low MI with the previously selected gene (in subsequent iterations with all previously selected genes) is chosen, thus maximizing relevancy and minimizing redundancy. In order to select this next gene, the algorithm maximizes the score $s = u - r$, where u is the MI between a gene and the target gene and r is the average MI between a gene and the already selected genes. Ideally, a gene will have a high value for u and a low value for r. since all genes are used as target genes for feature selection, two scores s_i and s_j are returned for each pair of genes $(X_i; X_j)$. The final score of a pair $(X_i; X_j)$ is either s_i or s_j, whichever one is higher is chosen. Similarly to ARACNE, pairs (or edges) with a score below a certain threshold are removed, while the rest form a GRN. Just like with ARACNE, the score of the edges can be interpreted as the amount of influence one gene has on another [33].

3 Material and Methods

3.1 Data Sources

3.1.1 Connectivity Map

The CMAP dataset was first introduced by Lamb *et al* in 2006 [24] and has since been used in a variety of successful drug repurposing approaches [6, 36]. The dataset was generated by analyzing gene-expression in cells treated with various types of small molecules. These perturbagens were applied to cells from different cell lines, namely (ss-)MCF7 (breast cancer), PC3 (prostate cancer), HL60 (leukemia) and SKEML5 (melanoma). This resulted in the production of 564 gene-expression profiles with Affymetrix GeneChip microarrays [8]. The dataset was later upscaled; currently, there are about 7000 expression profiles of cells treated with 1190 drugs available. Most experiments were done in the MCF7 cellline and a considerable number were done in the PC3 and HL60 celllines. Additionally, the majority of experiments were done using the Affymetrix GeneChip "HT_HG-U133A" (Figure 3.1). Since the different celllines and microarray platforms can influence an experiment, all programs were run separately for the MCF7, PC3 and HL60 celllines using the GeneChip "HT_HG-U133A". Other celllines and platforms do not contain enough data for reliable results.

The data was adjusted for background intensities using *gcrma* before its first use [47]. Background noise is introduced into microarray experiments through optical noise and non-specific hybridization with arrays. In order to obtain accurate gene-expression profiles, this non-biological noise needs to be filtered out. After this, the influence of different batches on the dataset must also be corrected. Batch effects appear in microarray experiments through changes in the experiment's environment, which can be as minuscule as the starting time of the experiment or the atmospheric ozone level [12]. The program ComBat uses an empirical bayes method to remove the effects of batches [20]. After this step, expression profiles from different batches can be used in the same analysis.

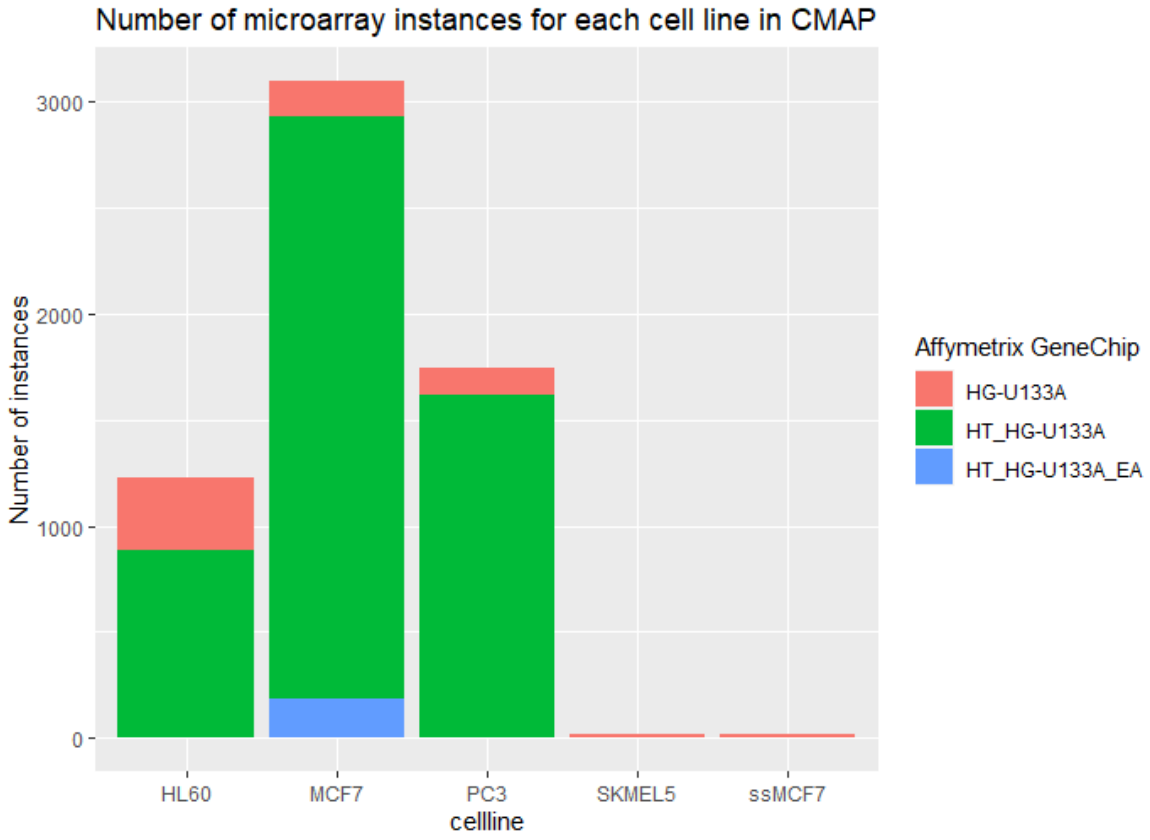

Figure 3.1: The figure shows the number of microarray instances in each of the celllines and on each GeneChip in CMAP.

3.1.2 KnockTF

The KnockTF database contains human gene-expression profiles of TF knockdown and knockout. It provides 570 datasets for 308 TFs sourced from the Gene Expression Omnibus (GEO) [3] and the Encyclopedia of DNA Elements (ENCODE) [7]. In addition to the gene-expression data, KnockTF provides various kinds of analyses, such as differential expression analysis, pathway enrichment using the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene ontology enrichment [13].

3.2 Approaches

3.2.1 First Approaches

To see whether a connection between drugs and TFs could be found using only available gene-expression data without integrating biological knowledge, two preliminary approaches at extracting connected drug/TF pairs were implemented.

3.2.1.1 Gene Set Enrichment Analysis

The first approach attempts to match the gene-expression values from KnockTF and CMAP through Gene Set Enrichment Analysis (GSEA). For a sorted list of gene-expression values and a set of genes, GSEA attempts to determine whether the genes in the set are evenly distributed in the list or tend to be found towards its top or bottom [43]. This shows whether a gene sets expression is mainly up- or down-regulated compared to the overall data.

Gayvert *et al* made use of this method to identify drugs that regulate TF activity. They used gene-expression values from the CMAP dataset to sort the input list of genes and direct target genes of TFs as gene sets for the GSEA program. This resulted in an enrichment score for each drug and TF pair. They further prioritized results using network analysis. Gayvert *et al* were able to rediscover known connections of drugs and TFs and discover new candidates for TF activity regulation using this method [16].

Inspired by their approach, it was attempted to find connections between CMAP data and TF data from KnockTF in this approach. To this end, first gene sets were created from the CMAP data by sorting the differential expression values of all genes for each drug in the database and selecting the top 100 most down regulated genes upon TF knockout. For the gene list input, differential expression values were calculated for each TF using the knockTF data. Differential expression was calculated as the log_2 fold change of each gene:

$$FC = log_2(\frac{mean(S)}{mean(C)}) \tag{3.1}$$

where S is the set of expression values for either drug treatment or TF knockout and C is the set of expression values from the control samples.
Applying GSEA to this input yields an enrichment score, a normalized enrichment score, a p-value and the false discovery rate (FDR) for every instance. The enrichment score shows the degree to which a gene set is overrepresented at either the top or bottom of the sorted gene list for TFs. The score corresponds to a weighted Kolmogorov-Smirnov-like running sum statistic. The normalized enrichment score averages this value to account for the size of the gene set. The p-value estimates the statistical significance of a result, whilst the FDR corrects the p-value for multiple testing by showing a results probability of being a false positive [43].

3.2.1.2 Gene Space Distance

Each gene-expression profile in CMAP contains values for around 13,000 genes. Similarly, entries in KnockTF contain values for 10,000 to 15,000 genes. We can use these genes to build a euclidean space in which each gene is a direction. An example for this is shown in figure 3.2: The space is defined by the two genes *Gene_1* and *Gene_2*. By using the fold changes, calculated as described in equation 3.1, of the two genes as coordinates, drugs and TFs are placed at a specific location in the gene space. While this example only shows a two-dimensional space, a space with as many dimensions as there are genes available can be built. To generate the space used in the analysis, only

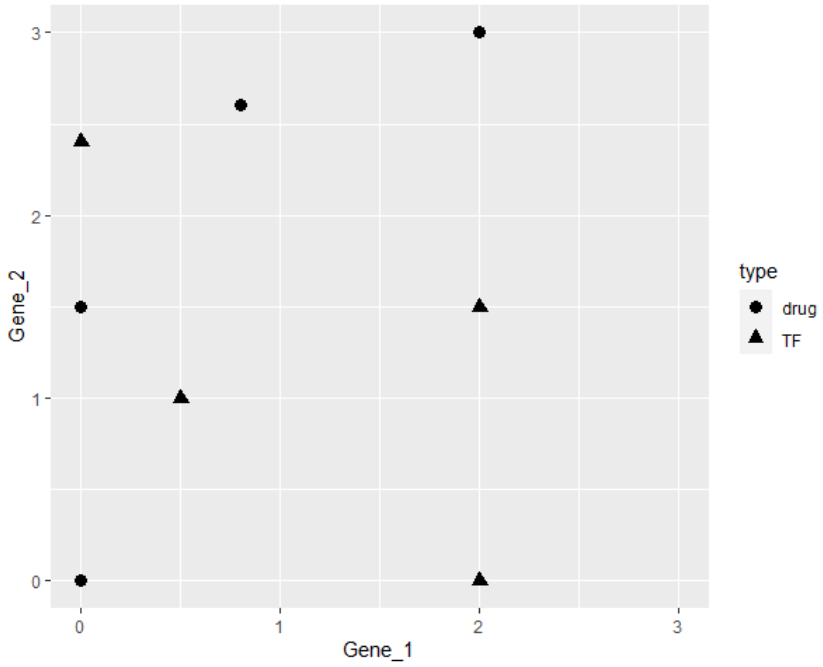

Figure 3.2: The figure shows the space given through the hypothetical genes Gene_1 and Gene_2. Multiple drugs and TFs are mapped to exact locations in the space by their differential expression of the two genes.

genes present in all expression profiles from CMAP and knockTF were used, which resulted in a space with 1485 dimensions. All drugs and TFs in the data can be mapped to an exact location in this space by calculating their differential expression for each gene and using it as the coordinate for the "direction" of this gene. We can calculate the euclidean distance in the space for each drug and TF pair by using the following formula for distance in n-dimensional space:
If $p = (p_1, p_2, ..., p_n)$ and $q = (q_1, q_2, ..., q_n)$ are two points in the n-dimensional space, then the distance $d(p, q)$ between them is defined as:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{3.2}$$

We would assume that if a drug and TF have a short distance in this space, the drug would have a similar effect as the knockout of that TF since the expression values of genes, relative to other genes, are similar. A direct comparison of this is not possible since fold changes of genes in CMAP are inherently lower than fold changes sourced from KnockTF (Figure 3.3). To remedy this, scikit-learns StandardScaler was applied to both datasets [38]. The program scales the values so that they are distributed roughly the same, whilst keeping the position of genes to each other intact.

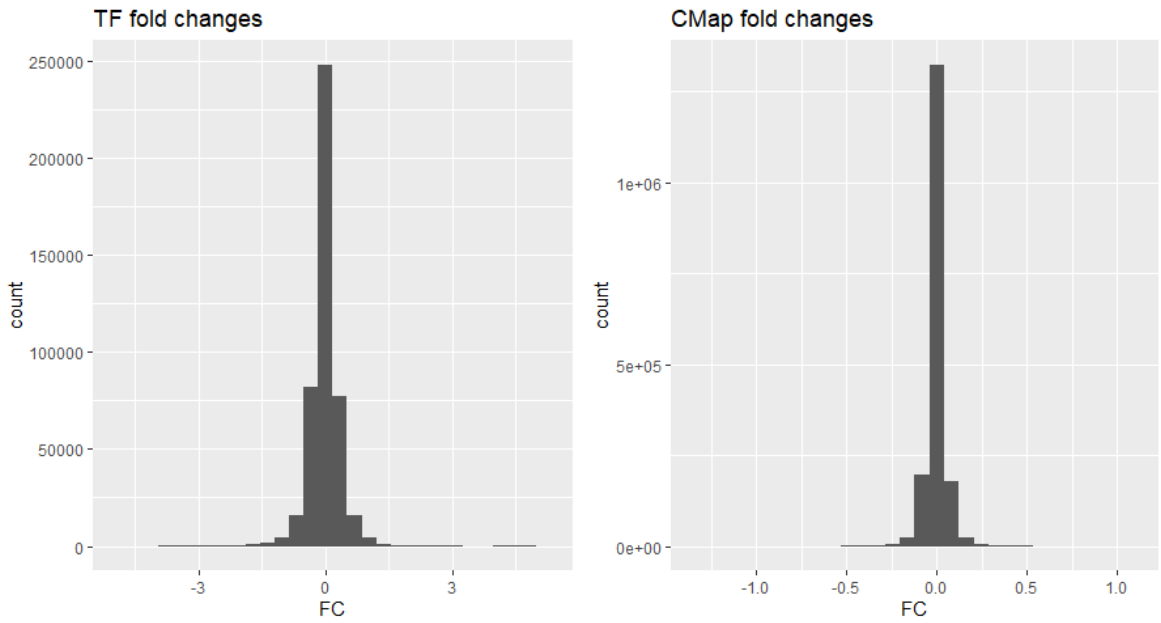

Figure 3.3: Distribution of fold changes in KnockTF and CMAP

3.2.2 Gene Regulatory Network Inference

3.2.2.1 Disease-Specific Networks

The approaches to drug repurposing listed above use the available gene-expression data of drugs and TFs, but they do not take actual biological knowledge into account. In oreder to look deeper into the mechanisms of various drugs, the CMAP data was used to construct disease-specific GRNs for the 33 diseases listed in Table 3.1. While it would be possible to infer a GRN for every single drug, the constructed network would not be reliable as there is not enough data is available for single drugs. Drug-disease associations make more data available for the construction of the GRNs and, instead of drug mechanisms, allow disease mechanisms to be researched directly. In order to extract gene-expression profiles relevant to a disease, drug-disease therapeutic associations were collected from the Comparative Toxicogenomics Database [31] and Repodb (updated 2020 version) [4]. All disease identifiers were converted to the ICD10 Code format using EMBL-EBIs' mapping tool OxO [21]. The resulting 4-character codes were collapsed to the more general 3-character codes. Additionally, various types of drug identifiers were first converted to DrugBank IDs [46] and then to CMAP names. For each disease, all gene-expression profiles from experiments using drugs, which are used to treat this disease, were gathered separately for each of the celllines MCF7, PC3 and HL60. Due to a lack of experiments, the celllines SKEML5 and ssMCF7 were not used. All experiments selected used the Affymetrix GeneChip "HT_HG-U133A" as the microarray platform. These expression profiles were then used as inputs for the GRN inference algorithms ARACNE and MRNET.

Table 3.1: Diseases used in GRN inference

ICD-10 Code	**Disease Name**
C18	Malignant Neoplasm of Colon
C22	Malignant Neoplasm of Liver and Intrahepatic Bile Ducts
C25	Malignant Neoplasm of Pancreas
C91	Lymphoid Leukemia
E11	Type 2 Diabetes Mellitus
E78	Disorders of Lipoprotein Metabolism and Other Lipidemias
F22	Delusional Disorders
F31	Bipolar Disorder
G20	Parkinsons Disease
G21	Secondary Parkinsonism
G30	Alzheimers Disease
G40	Epilepsy and Recurrent Seizures
G41	Status Epilepticus
G43	Migraine
G93	Other Disorders of Brain
I10	Essential (Primary) Hypertension
I20	Angina Pectoris
I21	Acute Myocardial Infarction
I25	Chronic Ischemic Heart Disease
I42	Cardiomyopathy
I48	Atrial Fibrillation and Flutter
I50	Heart Failure
I67	Other Cerebrovascular Diseases
I95	Hypotension
J45	Asthma
K25	Gastric Ulcer
K70	Alcoholic Liver Disease
K74	Fibrosis and Cirrhosis of Liver
L40	Psoriasis
M05	Rheumatoid Arthritis with Rheumatoid Factor
M06	Other Rheumatoid Arthritis
N17	Acute Kidney Failure
R51	Headache

3.2.2.2 Network Analysis

To evaluate correspondence between celltypes, a pairwise comparison of edge weights within a disease was made using GRNs generated with each of the three celllines. This was done for both ARACNE and MRNET. Afterwards, the results of the two algorithms in each cellline were also compared. Additionally, the results were compared to GENIE3, a regression based GRN inference method [18] which uses random forests. Following this, the number of shared edges between GRNs for different diseases were analyzed to find shared mechanisms and compared to known data from MSigDB and OmnipathDB. MSigDB provides different types of gene sets derived from many sources. Amongst them are gene sets associated with various diseases, which were compared to the most connected genes (hubs) in disease-specific GRNs [29]. OmnipathDB contains information on proteins and interactions, which are aggregated from 34 sources [44]. Among those, interactions of TFs with their targets were analyzed. The interactions are sourced from DoRothEA, a database which provides various information of TF interactions, including a confidence score for each interaction [15].

4 Results and Discussion

4.1 Results

4.1.1 First Approaches

4.1.1.1 Verification

Verification poses a challenge in this project, since there is no gold standard for drug repositioning using TFs. For this reason, a method which uses network proximity as a verification was implemented for the first two approaches. The method is based on a protein-protein interaction network (PPI) and uses distance between nodes in the network as a measure for how likely a TF and drug are associated. The PPI used in this analysis was built from a combination of protein-protein interactions and gene regulatory interactions. The protein interactions were obtained from RepotrialDB, which in turn obtained them from the Integrated Interactions Database (version 2018-11) [23]. The gene regulatory interactions were sourced from OmnipathDB [44] The formula for network proximity is sourced from Cheng *et al* [5]. For a set of drug interaction partners T and a TF s, proximity is defined as follows:

$$d(s,T) = \frac{1}{|T|}\sum_{t \epsilon T} d(s,t) \tag{4.1}$$

where $d(s,t)$ is the length of the shortest path between s and t.
Each drug can have multiple associated interaction partners in the PPI, which form the set T. Thus proximity is calculated as the average distance between a protein in T and the TF. A low network proximity indicates an association between two genes.

4.1.1.2 Gene Set Enrichment Analysis

Among the four scores returned by GSEA, the normalized enrichment score and the FDR are used to find significant results since they take geneset size and the multiple testing problem into account. When applying GSEA to the PC3 cellline, 796 pairs of drugs and TFs are identified as significant, with a threshold of 0.1 for the FDR and a threshold of -4 for the normalized enrichment score. When comparing the network proximity score of these pairs to 10,000 randomly drawn results, it was found that the distribution of scores was very similar (Figure 4.1). Varying the thresholds for significance did not improve the results. While Gayvert *et al* were able to find drug/TF pairs with closer proximity through GSEA, similar results were not achieved in this implementation [16]. The exact cause for this difference is unclear, but could be one of

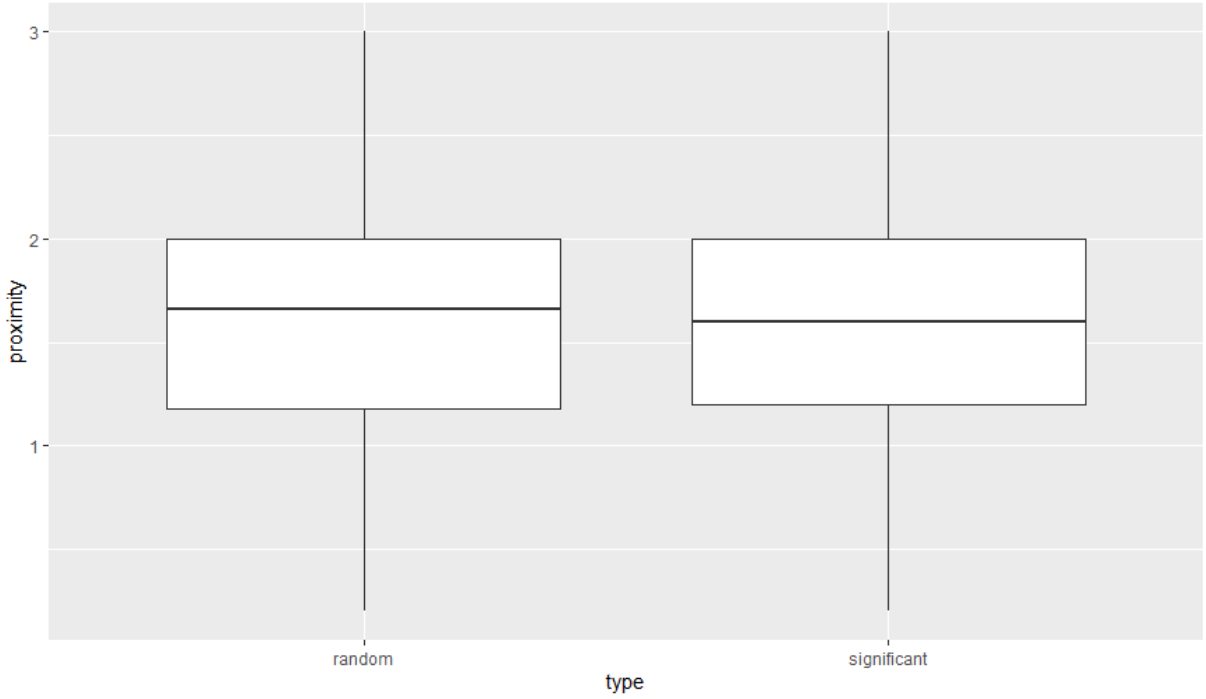

Figure 4.1: Network proximity of significant drug/TF pairs from GSEA compared to random pairs

the following reasons: The input for the GSEA program was changed from using drug-specific gene lists and TF-specific gene sets to using TF-specific gene lists and drug-specific gene sets. This may have caused a more drug-centered analysis. Additionally, the PPI used by Gayvert *et al* may contain more or different interactions than the one used here.

4.1.1.3 Gene Space Distance

Calculating the euclidean distance between every TF in the KnockTF data and every drug in the CMAP dataset results in 366,520 distance values. Analyzing the results in the PC3 cellline shows that, while the largest observed distance is 168.95, the majority of values is below 25, with a mean of 14.58 (Figure 4.2). Moreover, the 4760 values which are greater than 150 describe the distance of all drugs to just four TFs. The same is true when considering all distance values greater than 50: 8330 pairs of drugs and TFs contain only seven different TFs. This shows that some TFs are very far away from all CMAP drugs in the gene space. Additionally, most TFs have a very small range of distance values. The Average TF has a standard deviation of only 0.22 for distances to drugs, meaning the difference in distance to different drugs is very small. This raises the question whether the gene space distance between a TF and a drug is a meaningful measure of their relationship. Nevertheless, the above mentioned network proximity method was applied to each drug/TF pair. Taking the aforementioned results into account, it does not make sense to compare network proximity and absolute distance. Instead, the proximity of each TF's five nearest neighbors by gene space distance was compared to the proximity of 10,000 random drug/TF pairs (Figure 4.3). The proximity of the nearest neighbors was significantly lower than that of

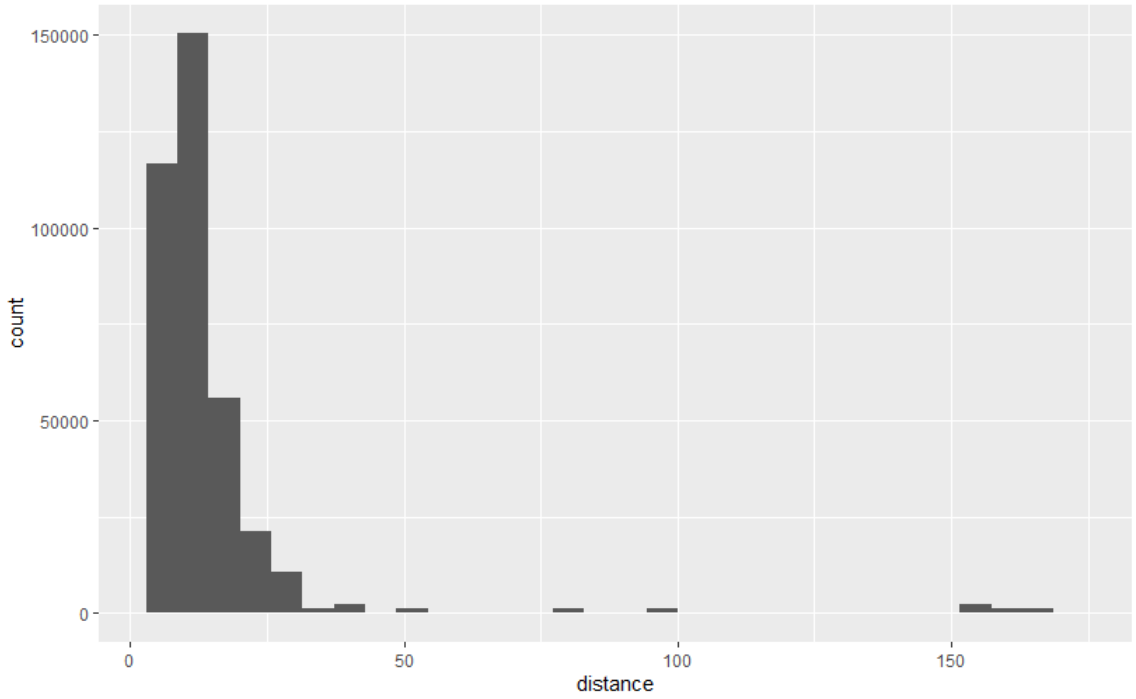

Figure 4.2: Distribution of gene space distances

randomly selected pairs, showing that information about the relationship of a drug and TF could be gained from their distance in the gene space. However, because of the minimal differences in distances between different drugs to the same TF, one needs to be careful when using this method. Additionally, since only 1485 genes were used in the construction of the gene space, a lot of information has been left out of the analysis.

4.1.2 Gene Regulatory Network Inference

4.1.2.1 Comparison in Celllines

After applying the first steps described in section 3.2.2, specific GRNs were constructed for 33 diseases. Since a GRN was constructed for each of the three celllines (MCF7, PC3, HL60) and two algorithms (ARACNE, MRNET), 198 GRNs were constructed in total. Since the Affymetrix GeneChip "HT_HG-133A" measures expression values for 12,990 genes and both algorithms returned a weight for every pair of genes the number of edges N in a GRN is, in theory, $N = \frac{1}{2}(12,990^2 - 12,990) = 84,363,555$. Since in some cases there were no gene-expression values for certain genes, the actual number may differ in individual GRNs. The beginning of the analysis assessed whether each algorithm was stable, when its output was compared across different celllines. This comparison is shown here exemplary for the GRNs constructed for the disease C18 (malignant neoplasm of colon). Comparisons for other diseases can be found in the supplementary results in the appendix. Figure 4.4 shows the good correlation between different celllines when using ARACNE to infer GRNs. Although there are quite a few edges that have a weight of zero in one cellline and not the other, this is to be expected

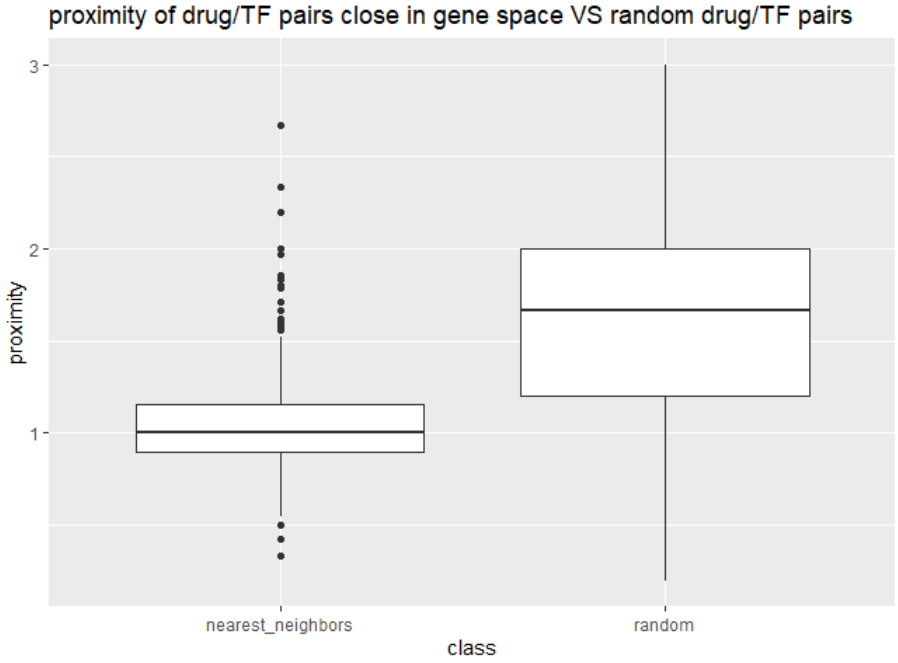

Figure 4.3: Proximity of Gene space nearest neighbors VS random pairs

due to the large number of edges which are weighted zero. In the example of C18, out of more than 83M edges returned, only 119,598 edges, about 0.1 percent, have a weight above zero. Figure 4.5 showcases this by comparing only the top 20,000 strongest edges of each GRN. The very low weights are due to the conservative nature of the algorithm, which tries to avoid false positives [30]. For this reason, it makes sense to only consider gene pairs with a weight above zero in the analysis. By only comparing edges with a non-zero weight in both GRNs, we can see that ARACNE results are fairly consistent across celllines. On the other hand, GRNs inferred with MRNET did not show much correspondence between different celllines (Figure 4.6). This indicates that MRNET is not able to capture disease mechanisms as well as ARACNE does, which makes the algorithm ill-suited for this method of drug repurposing. When inspecting the correspondence of ARACNE and MRNET results within the same cellline, we

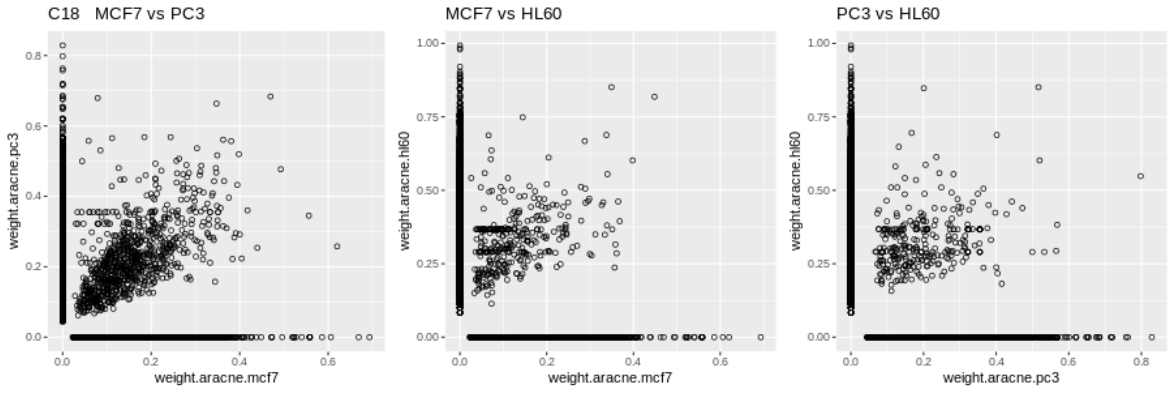

Figure 4.4: Comparison of GRNs for the disease C18 constructed by ARACNE in each of the three celllines MCF7, PC3 and HL60

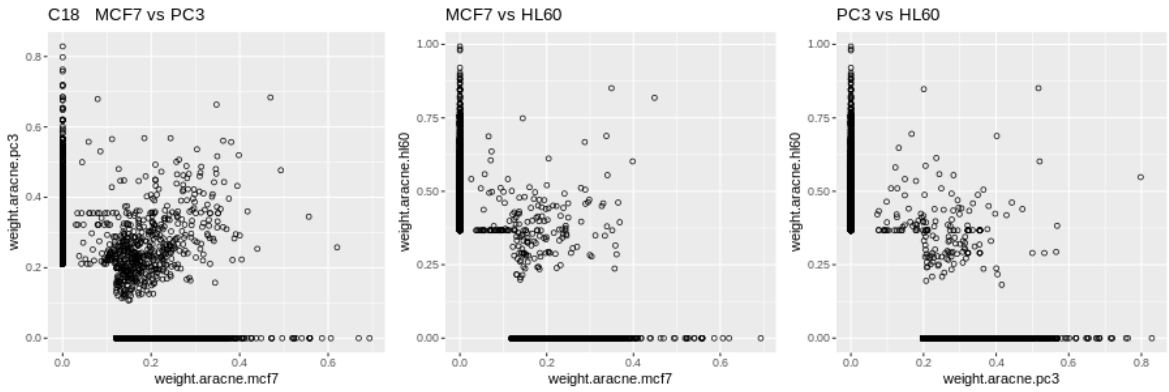

Figure 4.5: Comparison of the top 20,000 edges of GRNs constructed with ARACNE in each of the three celllines MCF7, PC3 and HL60

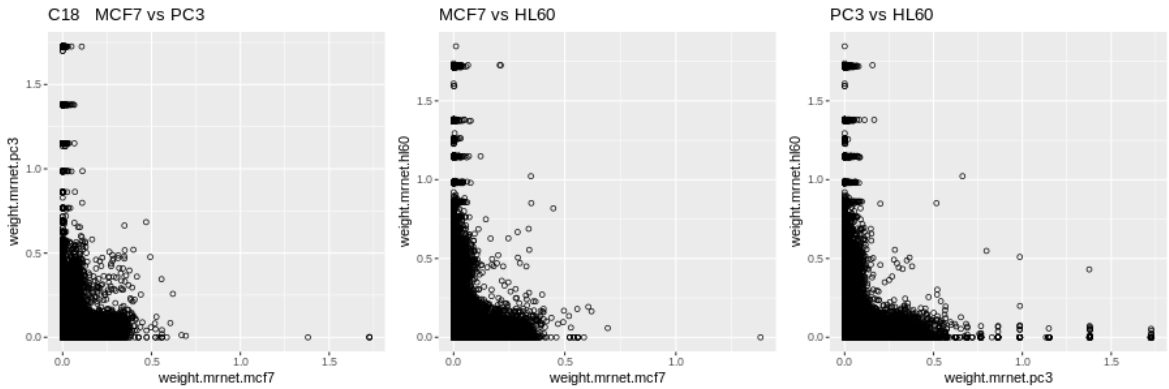

Figure 4.6: Comparison of GRNs for the disease C18 constructed by MRNET in each of the three celllines MCF7, PC3 and HL60

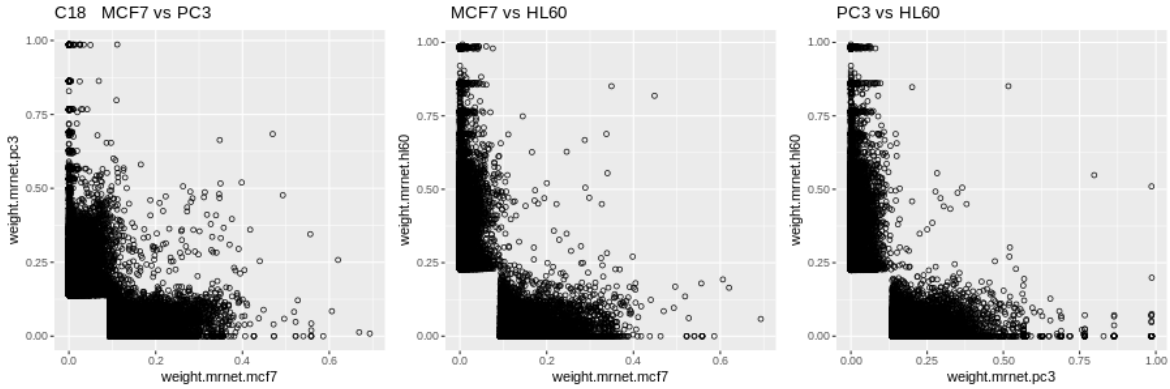

Figure 4.7: Comparison of the top 20,000 edges of GRNs constructed with MRNET in each of the three celllines MCF7, PC3 and HL60

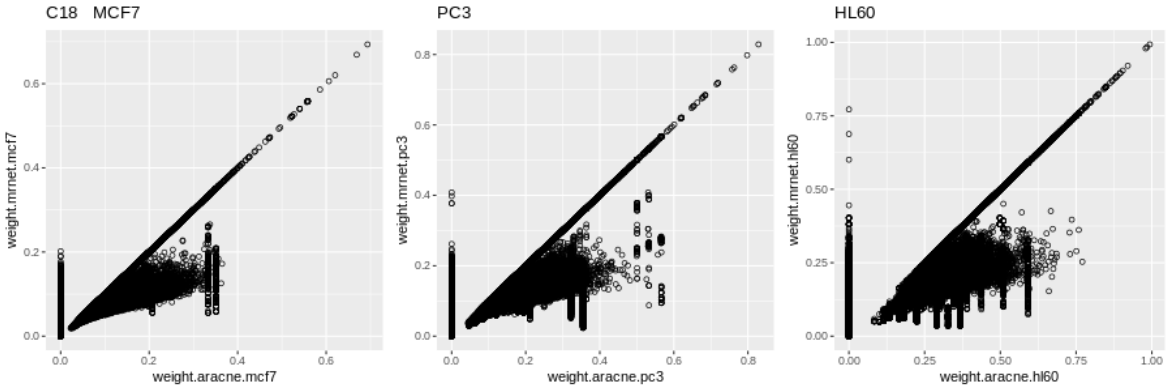

Figure 4.8: Comparison of GRNs for the disease C18 constructed by ARACNE to the results of GRNs constructed by MRNET in each of the three celllines MCF7, PC3 and HL60

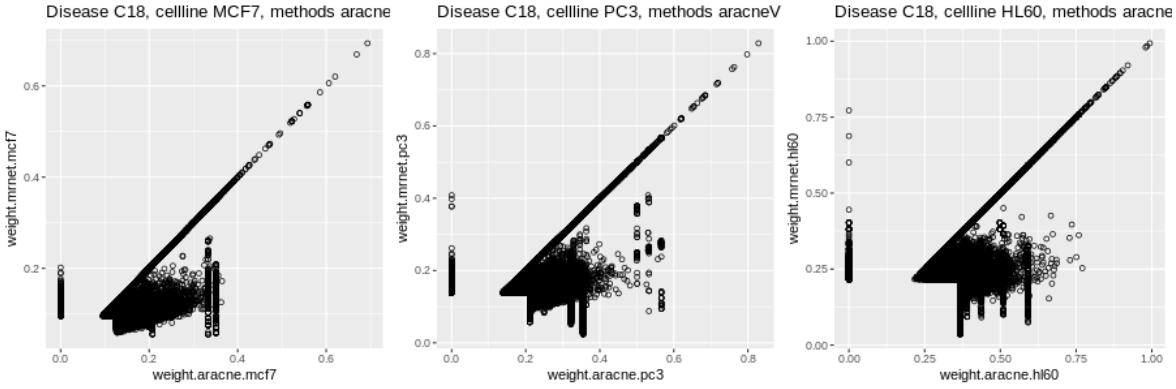

Figure 4.9: Comparison of the top 20,000 edges of GRNs constructed with ARACNE vs MRNET

observed a strong correspondence between their edge weight (Figure 4.8). Since both algorithms are based on MI, the similar results were expected. However, when taking the above observations into account, there should be more deviation from the diagonal since results from ARACNE are relatively stable between celllines and results from MRNET are not. Nevertheless, from here on out use of the MRNET algorithm was abandoned and only GRNs built using ARACNE were considered.

4.1.2.2 Comparison to GENIE3

Due to the results above, only ARACNE was used for subsequent analysis and the expression profiles from the MCF7, PC3 and HL60 celllines were combined. This will add more information to each GRN and should, in theory, result in more accurate findings. Combining the celllines is possible because, as has been shown in the last section, ARACNE's results are stable between celltypes. Running the algorithm on the new, combined dataset results in 33 GRNs to analyze. These 33 GRNs were compared to disease-specific GRNs generated with GENIE3, a regression based GRN inference algorithm [18]. The GRNs from GENIE3 were provided by Gihanna Galindez from the Chair of Experimental Bioinformatics (TU Munich) and also use the three main celllines of the CMAP data as input. Figure 4.10 shows a comparison between GRNs created with ARACNE and GENIE3 for the disease C18. Note that values

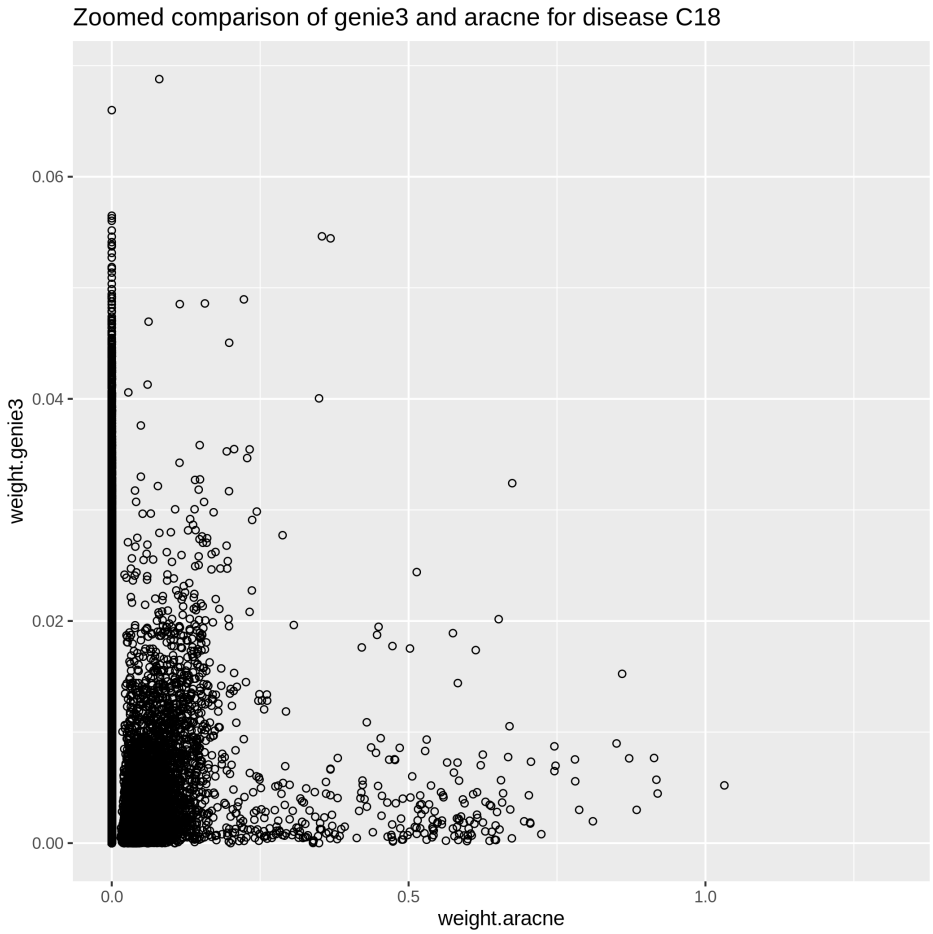

Figure 4.10: Comparison of edge weights between ARACNE and GENIE3 for C18 from all celllines.

between these two algorithms should only be compared relative to the values from each algorithm since GENIE3 assigns weights in a far lower range than ARACNE. Once again there are a multitude of edges with a weight of zero in ARACNE, which have fairly high weights in GENIE3. On the other hand, GENIE3 does not assign many absolute zero weights. Although some edges with low weights in GENIE3 have a high value in ARACNE, in general GENIE3 acts less conservative. Presumably, the lower correspondence between the two algorithms is rooted in their different methods of inferring GRNs as well as ARACNE's tendency to give false negative values in regard to interactions rather than false positives.

4.1.2.3 Finding Shared Disease Mechanisms

In order to detect shared mechanisms between diseases, the 33 disease-specific GRNs generated with ARACNE were compared to each other. Only edges with a weight of at least 0.1 were considered for this analysis. In order to get a sparse GRN with strong interactions, this threshold was chosen after considering the overall distribution of edge weights in the GRNs (Figure 4.11 shows an example). In an effort to find interactions that are shared between diseases, the number of shared edges between each pair of diseases was compared. Using this method, we found a number of disease

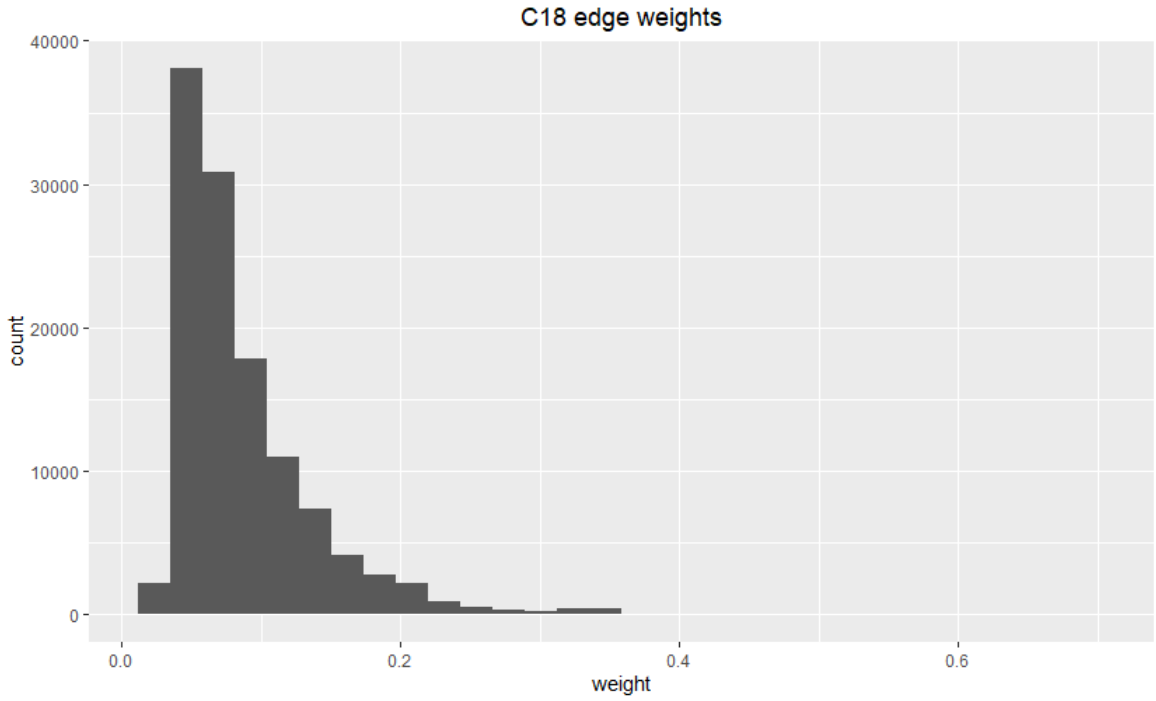

Figure 4.11: Distribution of edge weights in the GRN for C18 inferred with ARACNE using all celllines. Edges weighted zero are excluded.

clusters with a high number of shared edges (Figure 4.12). Some of these are known to be related, for example, the diseases C18, C22, C25 and C91 are all different kinds of malignant neoplasms and are therefore closely associated. In a pairwise comparison of edges, these diseases share between 1838 (C18, C25) and 3948 (C22, C25) edges, whilst the mean number of shared edges between diseases is 997.4. Other examples of diseases with a high number of shared edges that are known to be associated include F22 and F31, K70 and K74 as well as L40, M05 and M06. The first pair is classified as part of the mental and behavioral disorders as defined in ICD-10, while K70 and K74 are both diseases which affect the liver [37]. M05 and M06 are two types of arthritis. L40 (Psoriasis) has long been known to be comorbid with arthritis, in the form of psoriatic arthritis [34]. In a study from 2016, Raheel *et al* found that rheumatoid arthritis patients had an increased risk of developing malignant neoplams compared to other subjects [40]. This comorbidity may be the reason for the high number of interactions shared between M06 (Rheumatoid Arthritis) and the malignant neoplams C22 (Malignant neoplasm of liver and intrahepatic bile ducts) and C25 (Malignant neoplasm of pancreas). Most other high-ranking disease pairs are also known to be associated. These examples show that it is possible to discern diseases with shared mechanisms using disease-specific GRNs. When taking a look at other diseases with a high number of shared edges, we find that L40 (Psoriasis), J45 (Asthma) and C91 (Lymphoid Leukemia) share a high number of edges. While it has been found that psoriasis patients have a higher risk of developing asthma [11], we were not able to find any known associations of lymphoid leukemia with either asthma or psoriasis. This would make medications for these diseases potential candidates for drug repurposing. When limiting this analysis to only shared edges containing TFs, the results mirror

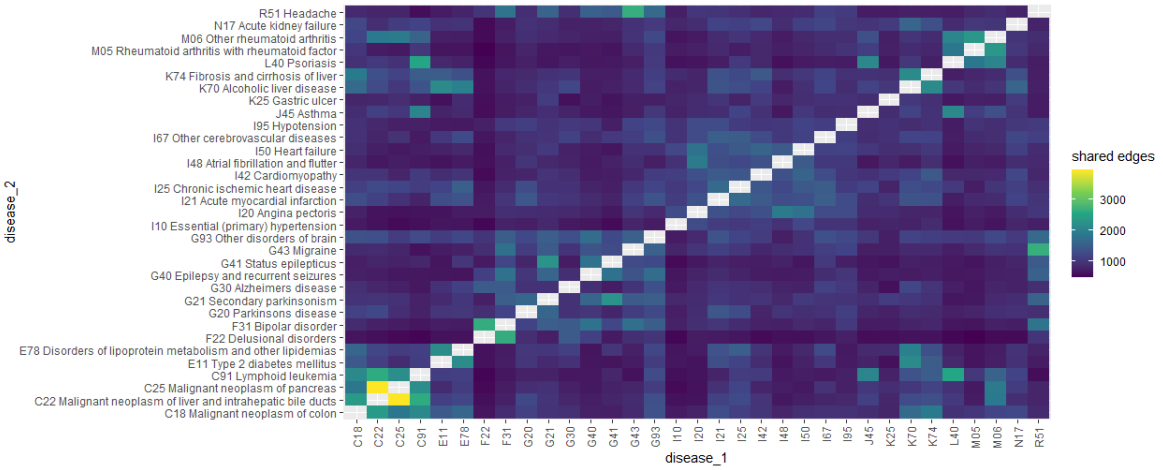

Figure 4.12: The figure shows the number of edges each pair of diseases share in their GRNs inferred with ARACNE

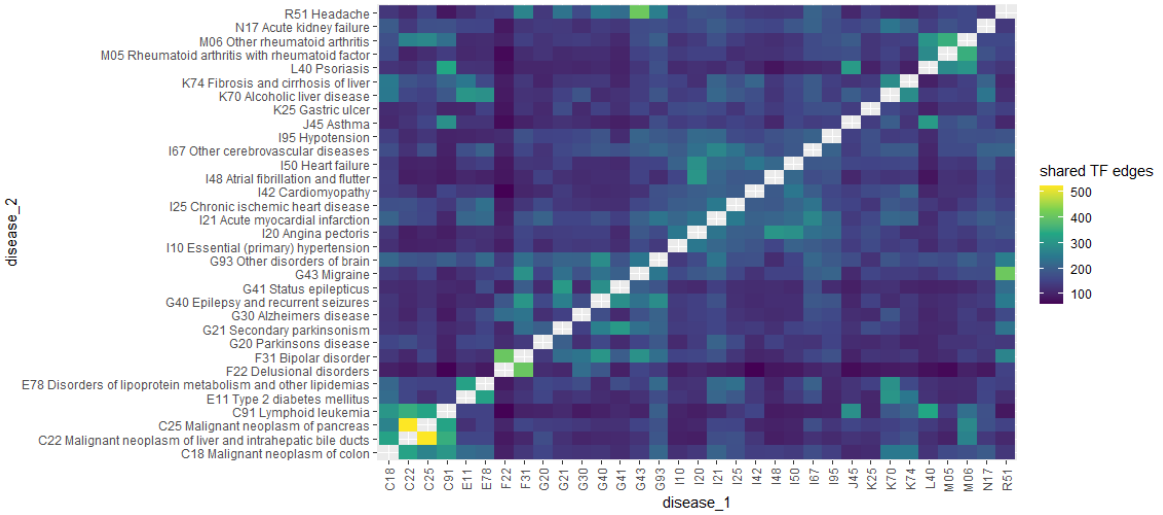

Figure 4.13: Number of shared edges containing TFs between disease-specific GRNs

the findings above (Figure 4.13). This makes sense, since TFs are the regulators of the GRN. It also showcases the strong involvement of TFs in the disease mechanisms and makes it possible to identify specific TFs for drug repurposing.

4.1.2.4 Downstream analysis

For some select diseases, hubgenes were compared to gene sets involved in the disease which were sourced from MSigDB. For this, hubgenes were defined as the top 30 most connected genes of each disease. When comparing these to known gene sets from MSigDB, we find little to no overlap between them. When comparing the hubgenes of C91 (lymphoid leukemia) to the three MSigDB datasets *PEPPER CHRONIC LYMPHOCYTIC LEUKEMIA UP*, *PEPPER CHRONIC LYMPHOCYTIC LEUKEMIA DN* and *WP MICRORNA NETWORK ASSOCIATED WITH CHRONIC LYMPHOCYTIC LEUKEMIA*, no overlap was found. The same goes for comparing the hub-

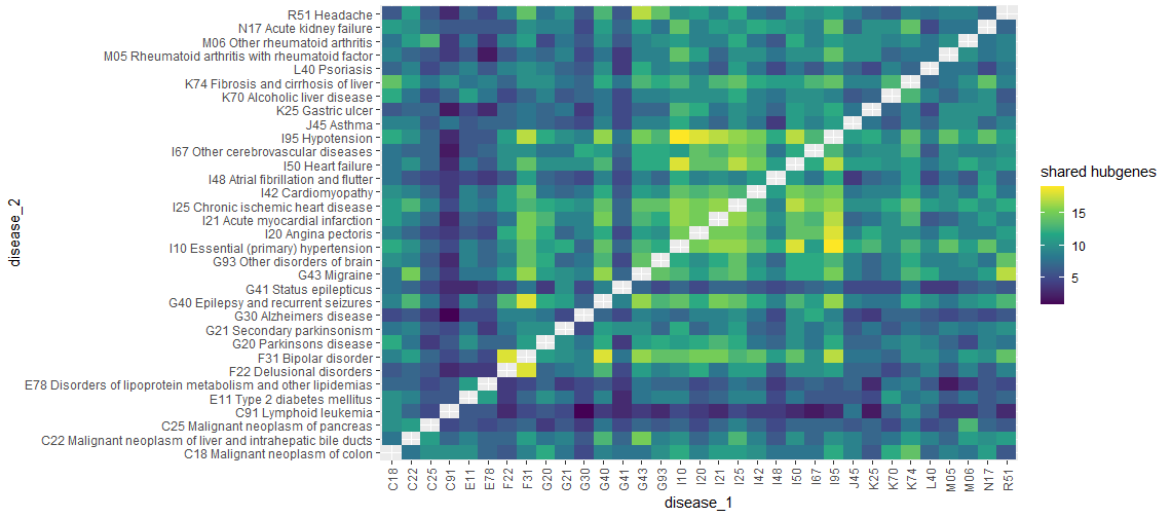

Figure 4.14: Hubgene overlap of GRNs compared between diseases

genes of J45 (asthma) to the *KEGG ASTHMA* geneset, as well as E11 (Type II diabetes mellitus) to the sets *KEGG TYPE II DIABETES MELLITUS* and *WP TYPE II DIABETES MELLITUS*. The reason for this may be that the hubgenes are less disease-specific. When comparing hubgenes shared by diseases, we find them to be less specific for most related diseases. However, diseases of the circulatory system appear to share a high number of hubgenes compared to other diseases (Figure 4.14). Comparing edges of the GRNs to known TF interactions in OmnipathDB shows little overlap between the data. Of the 279,581 interactions sourced from omnipathDB, a disease-specific GRN shares 34 interactions on average. Among those, the vast majority are of level C and D, meaning the DoRothEA database has assigned them a low level of confidence. This raises further concerns about the results of the methods, since few known TF interactions are recovered.

4.2 Discussion

It has been shown that ARACNE has good correspondence between the three celllines MCF7, PC3 and HL60 when inferring disease-specific GRNs. On the other hand, MRNET does not. Since both algorithms use the same method of calculating MI, the difference must lie in the feature selection process. Edges from ARACNE either keep the MI of their gene pair as their weight or are "removed" by setting their weight to zero. Edges from MRNET are assigned a weight, based not only on the mutual information of the involved gene pair, but also on other genes interacting with the gene pair. It appears that this new weight is not consistent throughout different celllines. While it is not exactly clear what causes this, it shows that MRNET should not be used on data stemming from multiple celllines. Since using more data will increase the accuracy of MI, it was decided to abandon the MRNET approach and combine data from the MCF7, PC3 and HL60 celllines to generate new GRNs using ARACNE. ARACNE is able to capture disease-specific interactions in its networks. This is confirmed by the high number of edges shared by diseases, which are already known to be associated,

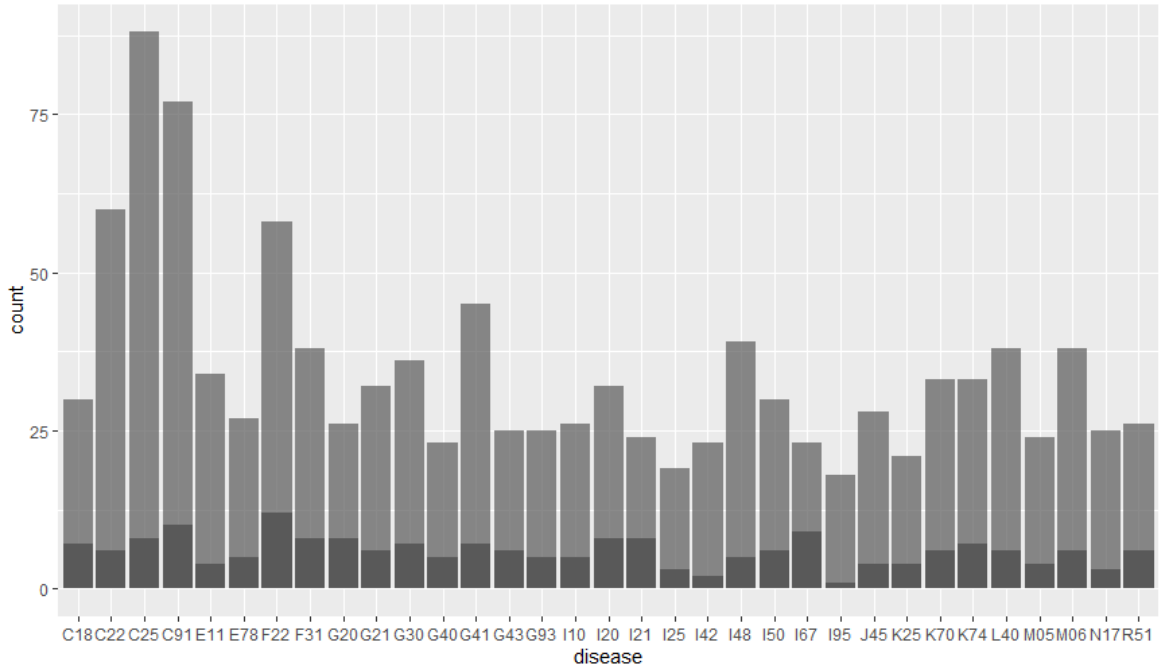

Figure 4.15: Number of edges which are both in the Disease-specific GRNs as well as OmnipathDB. Dark grey shows interactions with a confidence level of A or B, while light grey shows interactions with a confidence level of C or D.

compared to the number of edges shared between diseases, which are not known to be associated. However, it is important to ask whether it is possible to recover those known associations due to the fact that ARACNE detects their disease mechanisms or due to the fact that they are already reflected in the input data. For this, the number of shared drugs (and in turn expression profiles) was compared to the number of shared edges in the resulting GRN for all pairs of diseases (Figure 4.16). It was found that there is a strong correlation between the two variables, especially for pairings which do not share many drugs. When setting a threshold of at most five shared drugs and at least 1500 shared GRN edges, we find only seven out of 338 pairings to be relevant. Out of these pairs shown in Table 4.1, five contain the disease G21 (secondary parkinsonism) and three contain the disease G40 (epilepsy and recurrent seizures). G21, G40, G41 and G93 are all diseases of the nervous system, which indicates that ARACNE is able to recover shared disease mechanisms, even when only a low number of drugs are shared between two diseases. However, the scale is far smaller than previously assumed. It should also be noted that no pairs, which share only one or zero drugs, came close to the threshold of 1500 shared edges. While there may be cases which fulfill this requirement if more diseases were compared this raises concerns over whether it is truly possible to find previously unknown relationships between diseases using this method.

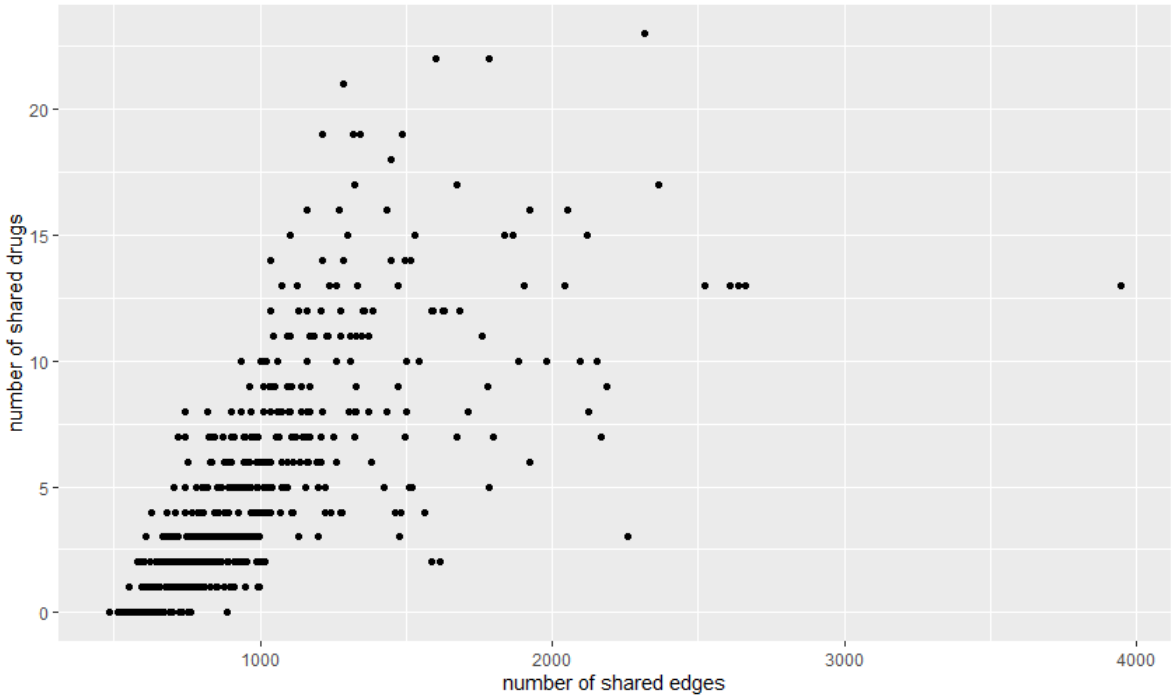

Figure 4.16: Number of shared drugs in ARACNE input compared to number of shared edges in resulting GRNs for each pair of diseases

Table 4.1: Disease pairs with a low number of shared drugs and high number of shared edges

Disease 1	Disease 2	number of shared edges	number of shared drugs
F31	G21	1520	5
G21	G40	1616	2
G21	G41	2258	3
G21	G93	1512	5
G21	R51	1585	2
G40	G41	1785	5
G40	R51	1564	4

5 Conclusion and Outlook

In this work I have made use of the CMAP dataset to find mechanisms shared by different diseases by analyzing disease-specific GRNs. The proposed method aggregates relevant gene-expression profiles from the dataset through drug-disease associations from various databases. Then, GRNs are inferred using the ARACNE algorithm and compared with each other to find shared disease mechanisms. While this method recovers known associations between diseases, they already share some associated drugs and therefore input for ARACNE. For this reason, the method should not be used in its current state to find drug repurposing targets. Tests on a larger dataset are required before a conclusion can be reached as to whether the method is able to identify disease-specific mechanisms. One possible solution to do this would be to apply the method to more diseases and compare the results. Additionally, it would be possible to use only part of the drugs associated with a disease as input for ARACNE to avoid overlaps in the input of the compared GRNs. If the results are similar, it would indicate ARACNEs ability to detect shared mechanisms of diseases. However, since the quality of a GRN is linked to the amount of input data, this should not be done using only the original CMAP dataset. To ensure high-quality GRNs, the LINCS-CMAP dataset could be used as a supplement. The LINCS-CMAP dataset is a large-scale gene-expression dataset. While similar to the original CMAP in its design, it contains about 1.3M expression profiles, which would be more than enough data to support the above testing method [22]. However, since a large part of the data is generated computationally, caution should be exercised.

Bibliography

[1] Christopher Paul Adams and Van Vu Brantner. "Spending on new drug development 1". In: *Health economics* 19.2 (2010), pp. 130–141.

[2] Ted T Ashburn and Karl B Thor. "Drug repositioning: identifying and developing new uses for existing drugs". In: *Nature reviews Drug discovery* 3.8 (2004), pp. 673–683.

[3] Tanya Barrett et al. "NCBI GEO: archive for functional genomics data setsâ€"update". In: *Nucleic acids research* 41.D1 (2012), pp. D991–D995.

[4] Adam S Brown and Chirag J Patel. "A standard database for drug repositioning". In: *Scientific data* 4.1 (2017), pp. 1–7.

[5] Feixiong Cheng et al. "Network-based approach to prediction and population-based validation of in silico drug repurposing". In: *Nature communications* 9.1 (2018), pp. 1–12.

[6] Jie Cheng et al. "Systematic evaluation of connectivity map for disease indications". In: *Genome medicine* 6.12 (2014), p. 95.

[7] ENCODE Project Consortium et al. "The ENCODE (ENCyclopedia of DNA elements) project". In: *Science* 306.5696 (2004), pp. 636–640.

[8] Dennise D Dalma-Weiszhausz et al. "[1] The Affymetrix GeneChip® Platform: An Overview". In: *Methods in enzymology* 410 (2006), pp. 3–28.

[9] Ensheng Dong, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time". In: *The Lancet infectious diseases* 20.5 (2020), pp. 533–534.

[10] Ammar D Elmezayen et al. "Drug repurposing for coronavirus (COVID-19): in silico screening of known drugs against coronavirus 3CL hydrolase and protease enzymes". In: *Journal of Biomolecular Structure and Dynamics* (2020), pp. 1–13.

[11] H-Y Fang et al. "Association between psoriasis and asthma: a population-based retrospective cohort analysis". In: *British Journal of Dermatology* 172.4 (2015), pp. 1066–1071.

[12] Thomas L Fare et al. "Effects of atmospheric ozone on microarray data quality". In: *Analytical chemistry* 75.17 (2003), pp. 4672–4675.

[13] Chenchen Feng et al. "KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors". In: *Nucleic acids research* 48.D1 (2020), pp. D93–D100.

[14] Nir Friedman. "Inferring cellular networks using probabilistic graphical models". In: *Science* 303.5659 (2004), pp. 799–805.

[15] Luz Garcia-Alonso et al. "Benchmark and integration of resources for the estimation of human transcription factor activities". In: *Genome research* 29.8 (2019), pp. 1363–1375.

[16] Kaitlyn M Gayvert et al. "A computational drug repositioning approach for targeting oncogenic transcription factors". In: *Cell reports* 15.11 (2016), pp. 2348–2356.

[17] Kamran Ghoreishi. "Thalidomide". In: *Encyclopedia of Toxicology (Third Edition)*. 2014.

[18] Vân Anh Huynh-Thu et al. "Inferring regulatory networks from expression data using tree-based methods". In: *PloS one* 5.9 (2010), pp. 1–10.

[19] Eric H. Davidson Isabelle Peter. *Genomic Control Process: Development and Evolution*. 1st ed. Academic Press, 2015. Chap. Chapter 2: Gene Regulatory Networks. ISBN: 0124047297,9780124047297.

[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.

[21] Simon Jupp et al. "OxO-A Gravy of Ontology Mapping Extracts." In: *ICBO*. 2017.

[22] Alexandra B Keenan et al. "The library of integrated network-based cellular signatures NIH program: system-level cataloging of human cells response to perturbations". In: *Cell systems* 6.1 (2018), pp. 13–24.

[23] Max Kotlyar et al. "IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species". In: *Nucleic acids research* 47.D1 (2019), pp. D581–D589.

[24] Justin Lamb et al. "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease". In: *science* 313.5795 (2006), pp. 1929–1935.

[25] Mélanie Lambert et al. "Targeting transcription factors for cancer treatment". In: *Molecules* 23.6 (2018), p. 1479.

[26] David S Latchman. "Transcription factors: an overview". In: *The international journal of biochemistry & cell biology* 29.12 (1997), pp. 1305–1312.

[27] Erik G Learned-Miller. "Entropy and mutual information". In: *Department of Computer Science, University of Massachusetts, Amherst* (2013).

[28] Wei-Po Lee and Wen-Shyong Tzou. "Computational methods for discovering gene networks from expression data". In: *Briefings in bioinformatics* 10.4 (2009), pp. 408–423.

[29] Arthur Liberzon et al. "Molecular signatures database (MSigDB) 3.0". In: *Bioinformatics* 27.12 (2011), pp. 1739–1740.

[30] Adam A Margolin et al. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context". In: *BMC bioinformatics*. Vol. 7. S1. Springer. 2006, S7.

[31] Carolyn J Mattingly et al. "The Comparative Toxicogenomics Database (CTD)." In: *Environmental health perspectives* 111.6 (2003), pp. 793–795.

[32] Patrick E Meyer, Frederic Lafitte, and Gianluca Bontempi. "minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information". In: *BMC bioinformatics* 9.1 (2008), p. 461.

[33] Patrick E Meyer et al. "Information-theoretic inference of large transcriptional regulatory networks". In: *EURASIP journal on bioinformatics and systems biology* 2007 (2007), pp. 1–9.

[34] JMH Moll and V Wright. "Psoriatic arthritis". In: *Seminars in arthritis and rheumatism*. Vol. 3. 1. Elsevier. 1973, pp. 55–78.

[35] Martin Morgan. *BiocManager: Access the Bioconductor Project Package Repository*. R package version 1.30.10. 2019.

[36] Aliyu Musa et al. "A review of connectivity map and computational approaches in pharmacogenomics". In: *Briefings in bioinformatics* 19.3 (2018), pp. 506–523.

[37] World Health Organization. *International Statistical Classification of Diseases and Health Related Problems, 10th Revision*. 2019.

[38] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020.

[40] Shafay Raheel et al. "Risk of malignant neoplasm in patients with incident rheumatoid arthritis 1980–2007 in relation to a comparator cohort: a population-based study". In: *International Journal of Rheumatology* 2016 (2016).

[41] Sandro G Viveiros Rosa and Wilson C Santos. "Clinical trials on drug repositioning for COVID-19 treatment". In: *Revista Panamericana de Salud Pública* 44 (2020), e40.

[42] Janusz Sławek and Tomasz Arodź. "ENNET: inferring large gene regulatory networks from expression data using gradient boosting". In: *BMC systems biology* 7.1 (2013), p. 106.

[43] Aravind Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[44] Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. "OmniPath: guidelines and gateway for literature-curated signaling pathway resources". In: *Nature methods* 13.12 (2016), pp. 966–967.

[45] Junmei Wang. "Fast identification of possible drug treatment of coronavirus disease-19 (COVID-19) through computational drug repurposing study". In: *Journal of Chemical Information and Modeling* (2020).

[46] David S Wishart et al. "DrugBank 5.0: a major update to the DrugBank database for 2018". In: *Nucleic acids research* 46.D1 (2018), pp. D1074–D1082.

[47] Zhijin Jean Wu and Rafael Irizarry. *Description of gcrma package*. 2010.

[48] MK Stephen Yeung, Jesper Tegnér, and James J Collins. "Reverse engineering gene networks using singular value decomposition and robust regression". In: *Proceedings of the National Academy of Sciences* 99.9 (2002), pp. 6163–6168.

Appendix

Supplementary Results

Comparison of GRN inference between celllines

ARACNE

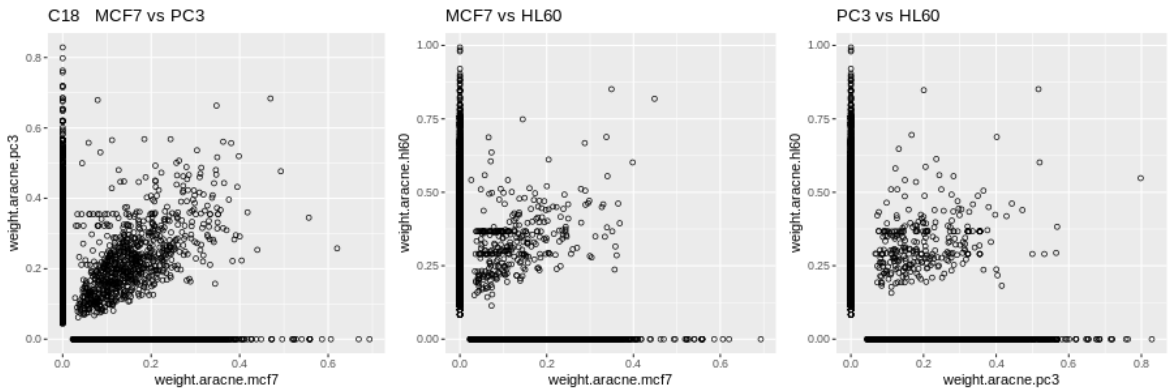

Figure A.1: Comparison between the celllines MCF7, PC3 and HL60 for the disease C18 when using ARACNE

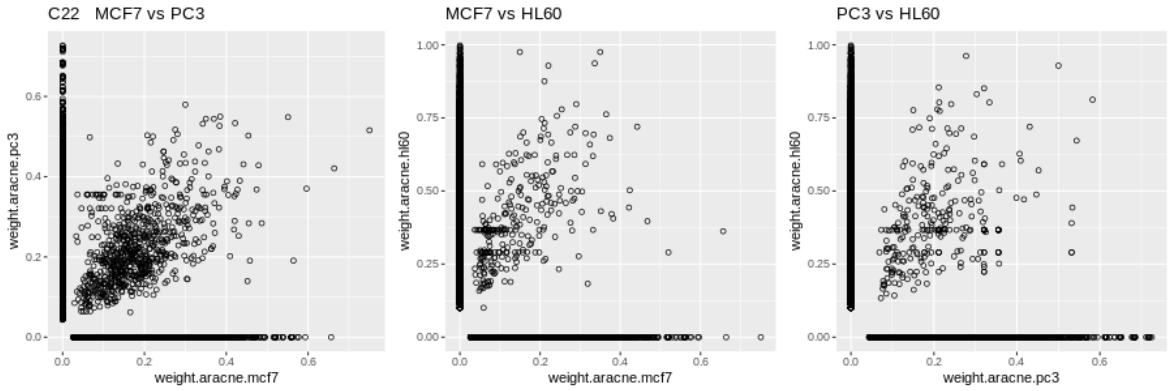

Figure A.2: Comparison between the celllines MCF7, PC3 and HL60 for the disease C22 when using ARACNE

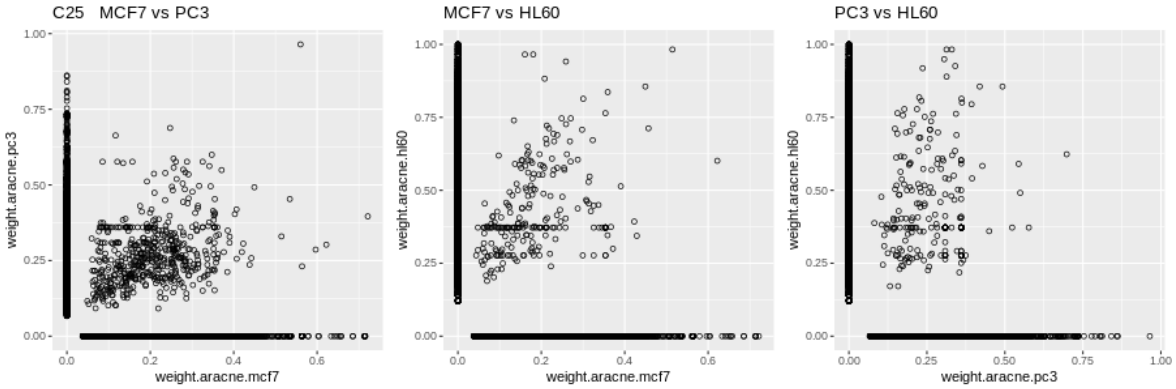

Figure A.3: Comparison between the celllines MCF7, PC3 and HL60 for the disease C25 when using ARACNE

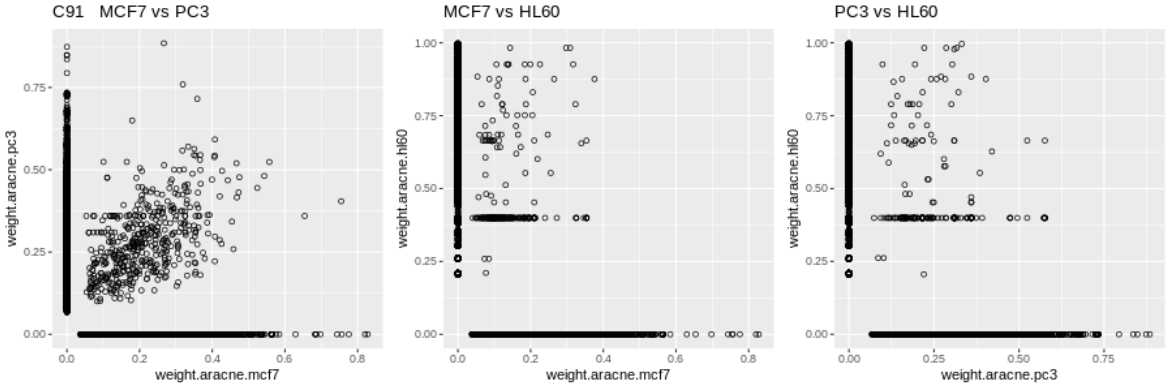

Figure A.4: Comparison between the celllines MCF7, PC3 and HL60 for the disease C91 when using ARACNE

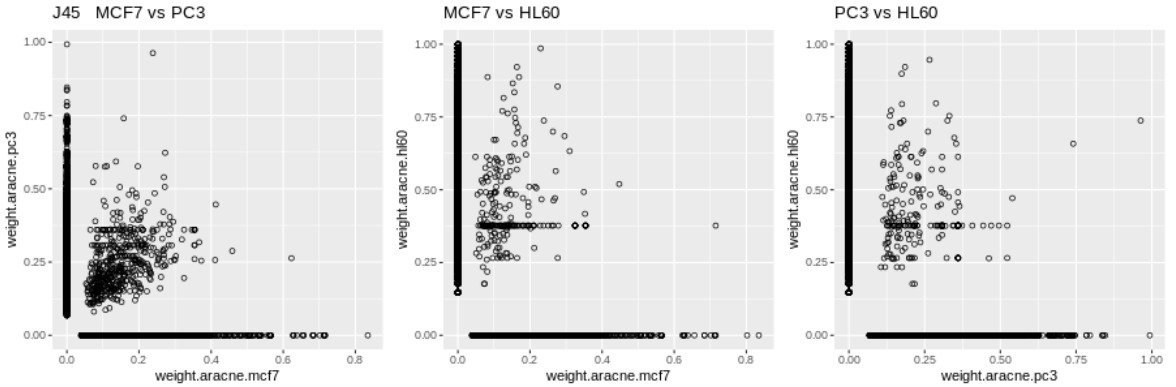

Figure A.5: Comparison between the celllines MCF7, PC3 and HL60 for the disease J45 when using ARACNE

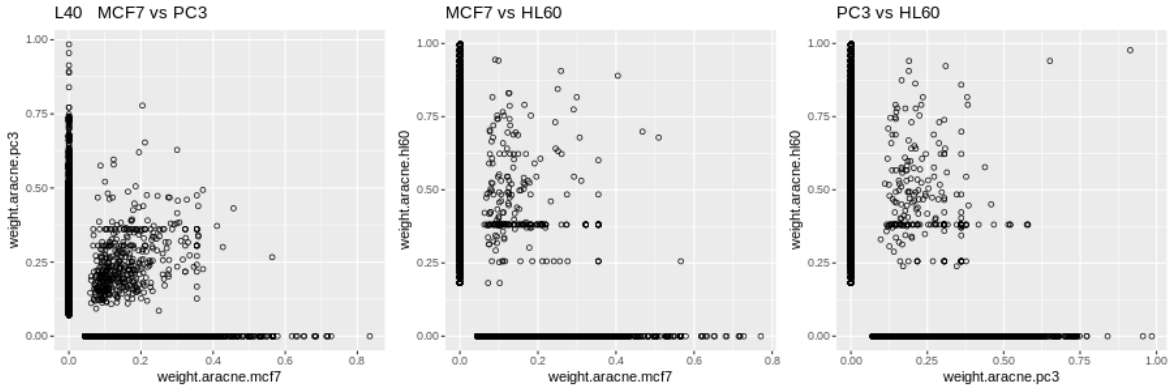

Figure A.6: Comparison between the celllines MCF7, PC3 and HL60 for the disease L40 when using ARACNE

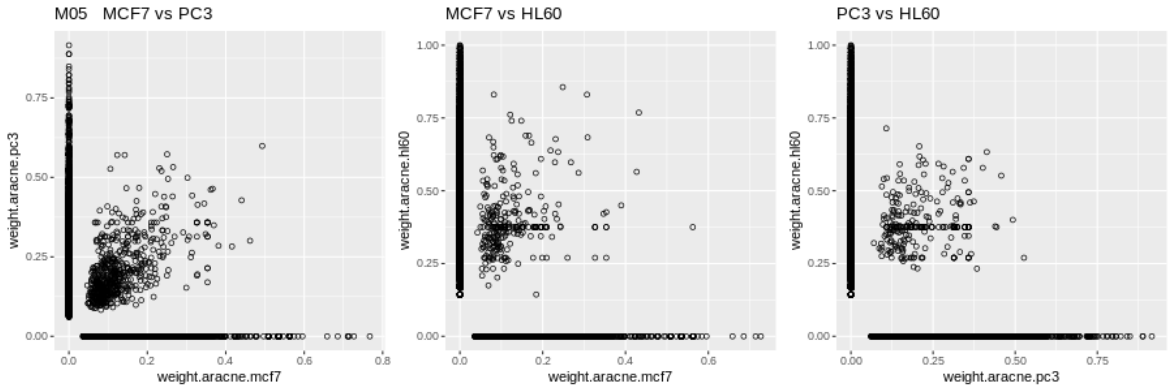

Figure A.7: Comparison between the celllines MCF7, PC3 and HL60 for the disease M05 when using ARACNE

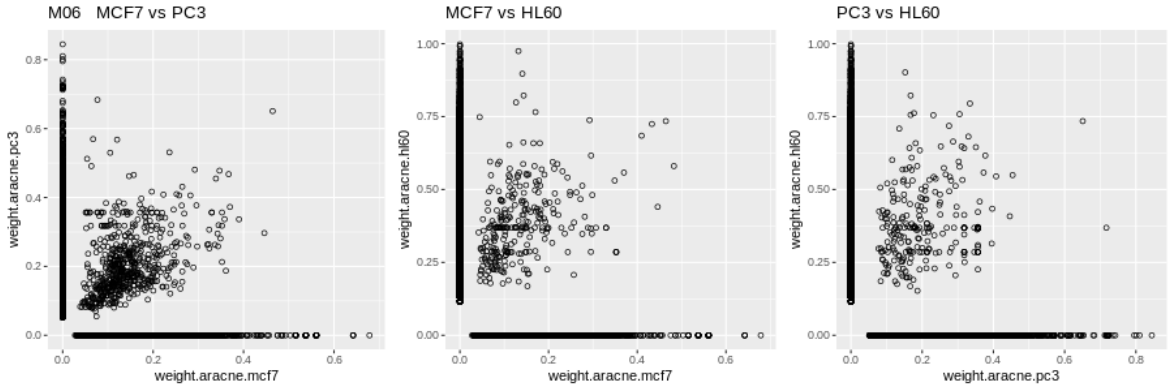

Figure A.8: Comparison between the celllines MCF7, PC3 and HL60 for the disease M06 when using ARACNE

MRNET

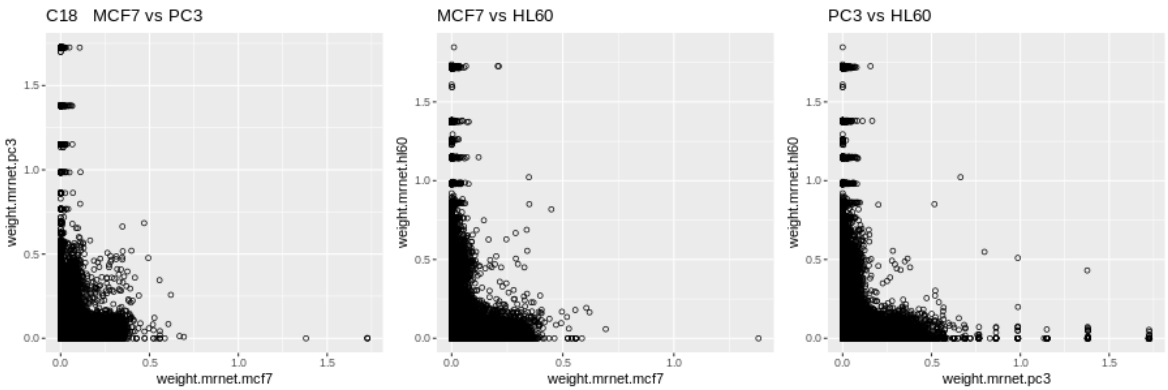

Figure A.9: Comparison between the celllines MCF7, PC3 and HL60 for the disease C18 when using MRNET

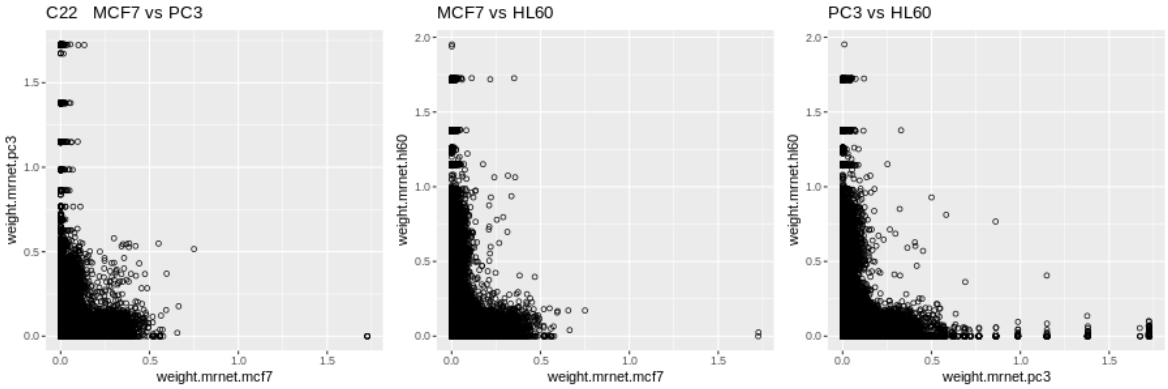

Figure A.10: Comparison between the celllines MCF7, PC3 and HL60 for the disease C22 when using MRNET

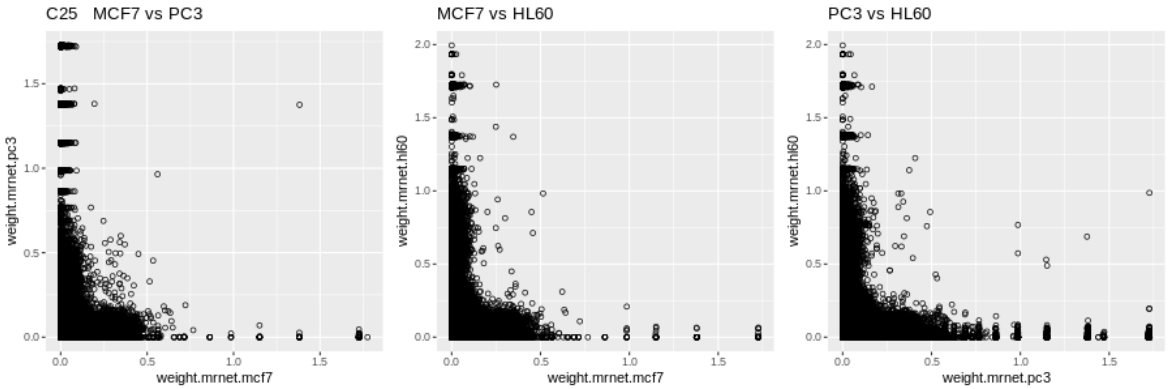

Figure A.11: Comparison between the celllines MCF7, PC3 and HL60 for the disease C25 when using MRNET

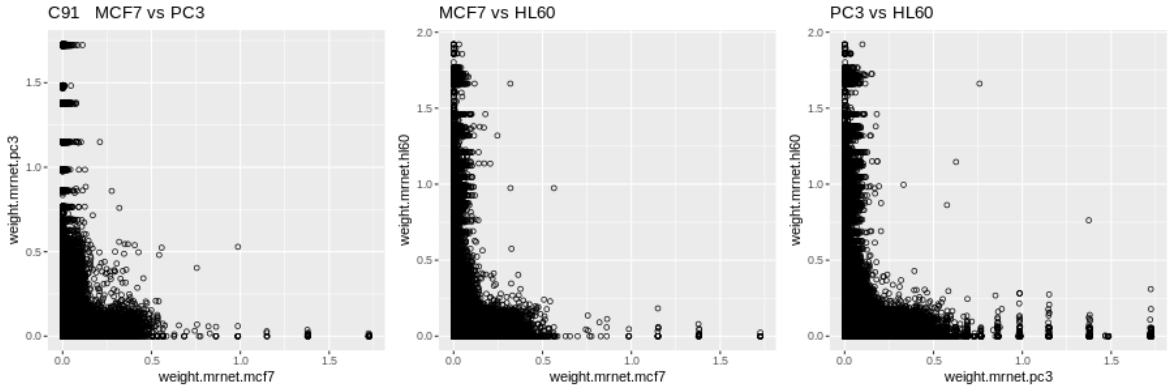

Figure A.12: Comparison between the celllines MCF7, PC3 and HL60 for the disease C91 when using MRNET

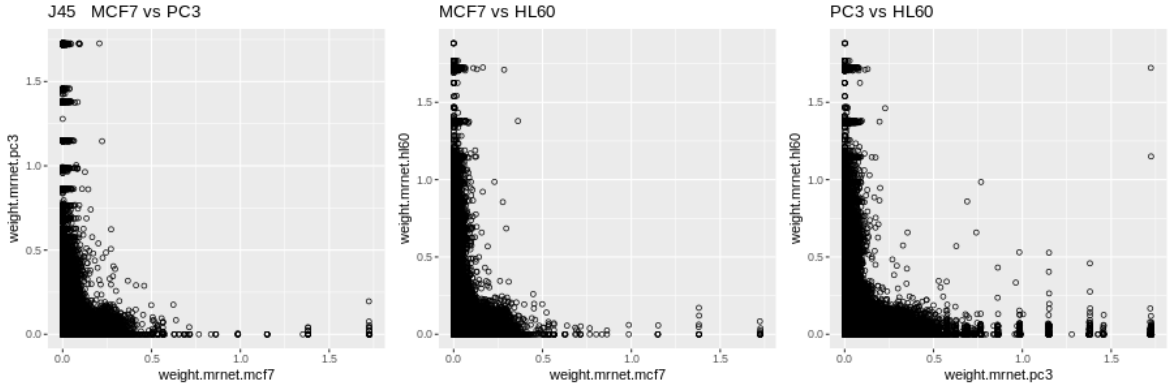

Figure A.13: Comparison between the celllines MCF7, PC3 and HL60 for the disease J45 when using MRNET

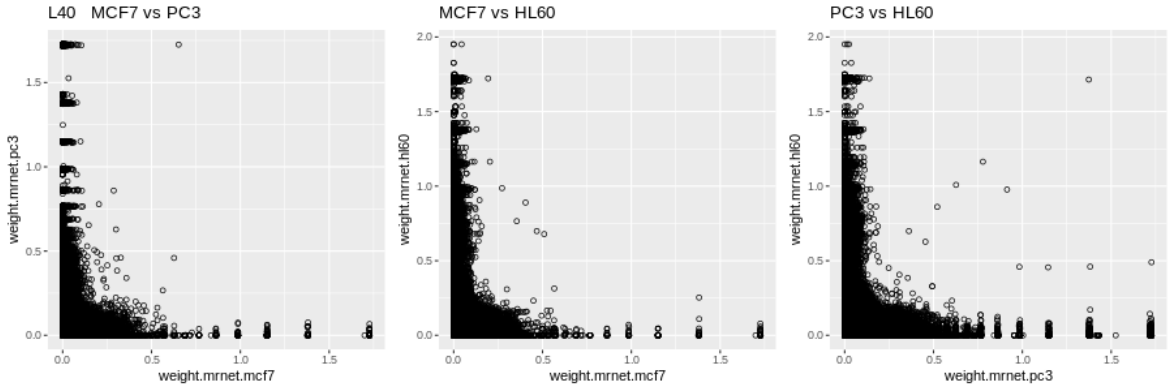

Figure A.14: Comparison between the celllines MCF7, PC3 and HL60 for the disease L40 when using MRNET

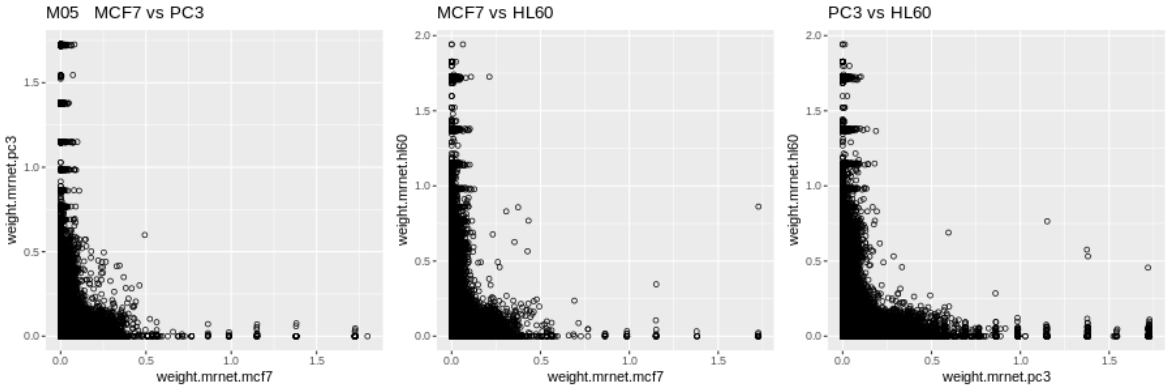

Figure A.15: Comparison between the celllines MCF7, PC3 and HL60 for the disease M05 when using MRNET

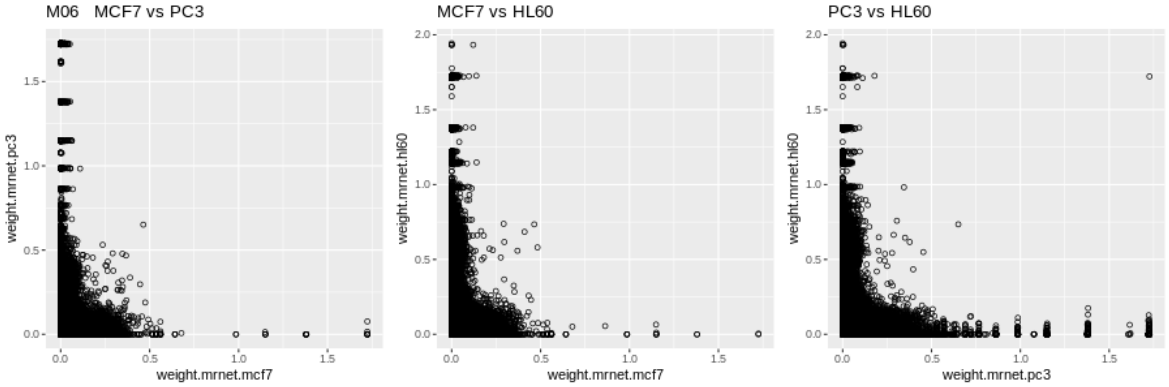

Figure A.16: Comparison between the celllines MCF7, PC3 and HL60 for the disease M06 when using MRNET

Comparison between ARACNE and MRNET

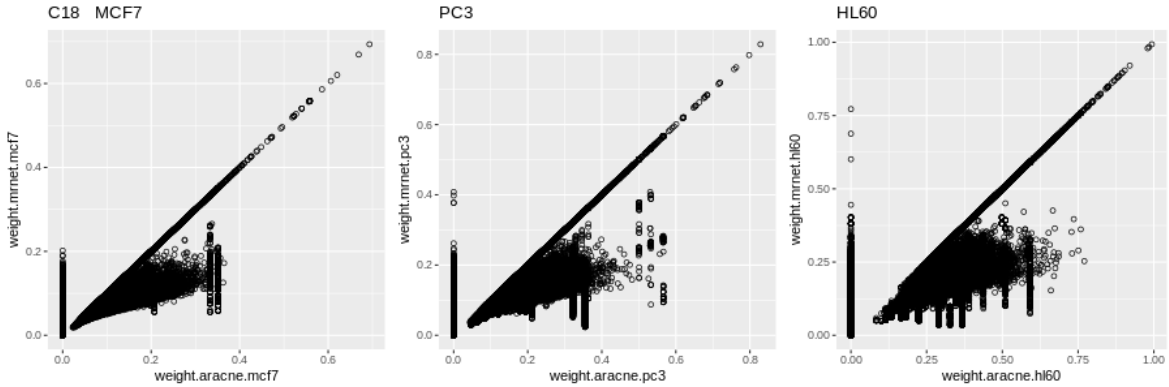

Figure A.17: Comparison between the algorithms ARACNE and MRNET for the disease C18 in each cellline

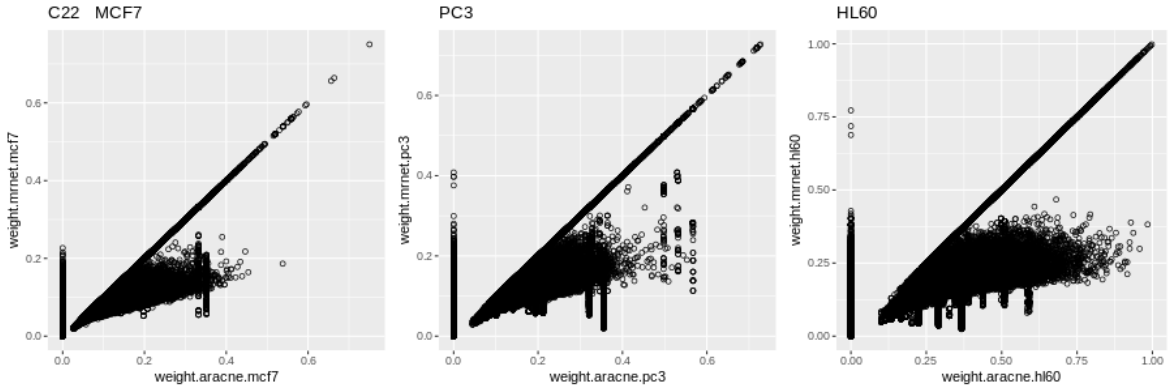

Figure A.18: Comparison between the algorithms ARACNE and MRNET for the disease C22 in each cellline

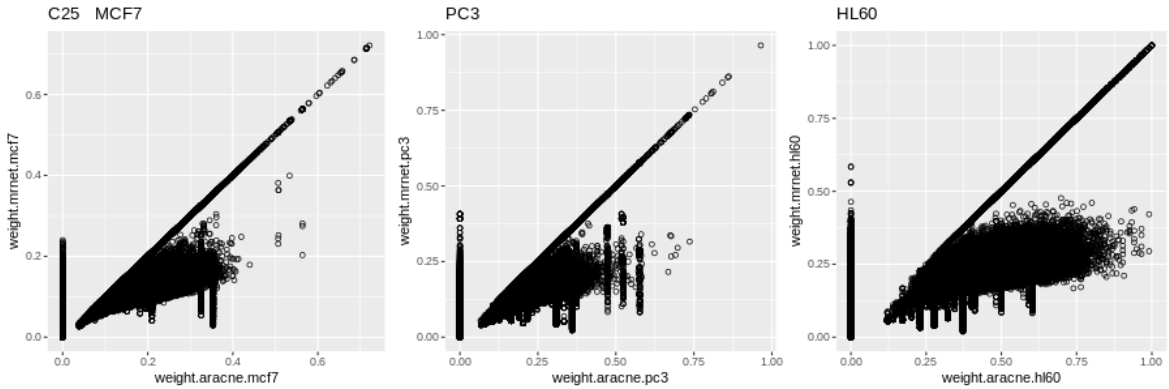

Figure A.19: Comparison between the algorithms ARACNE and MRNET for the disease C25 in each cellline

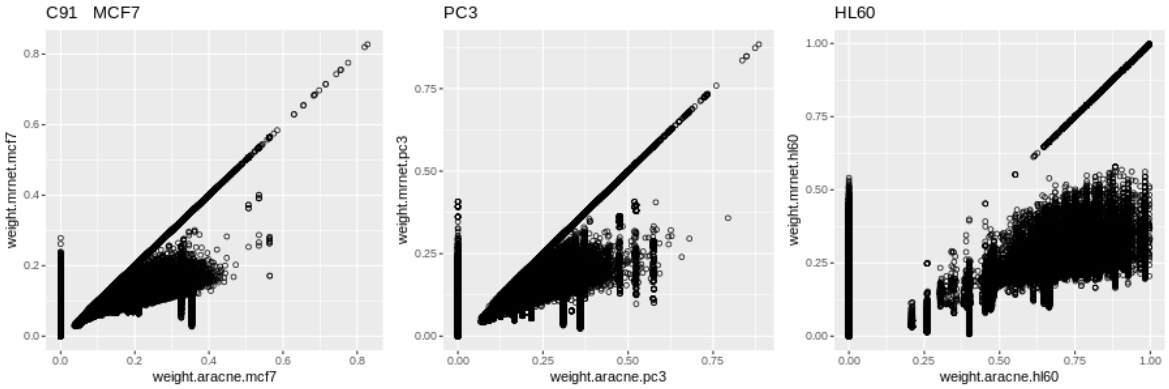

Figure A.20: Comparison between the algorithms ARACNE and MRNET for the disease C91 in each cellline

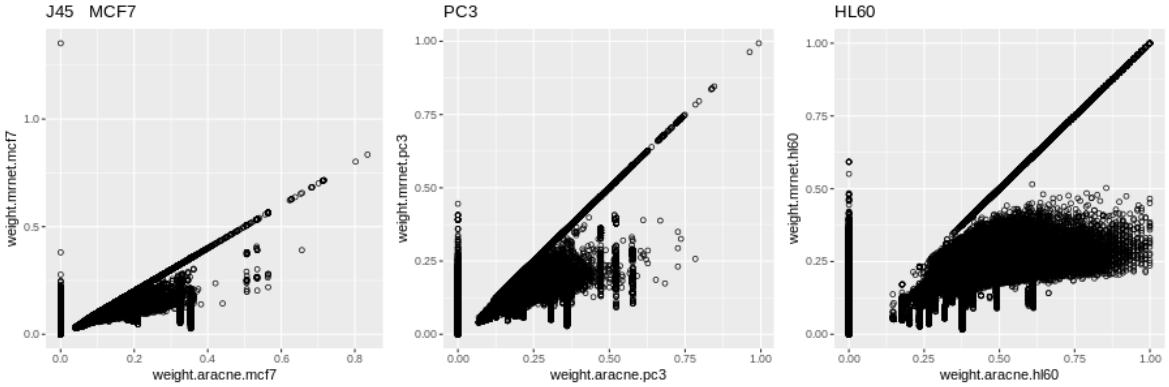

Figure A.21: Comparison between the algorithms ARACNE and MRNET for the disease J45 in each cellline

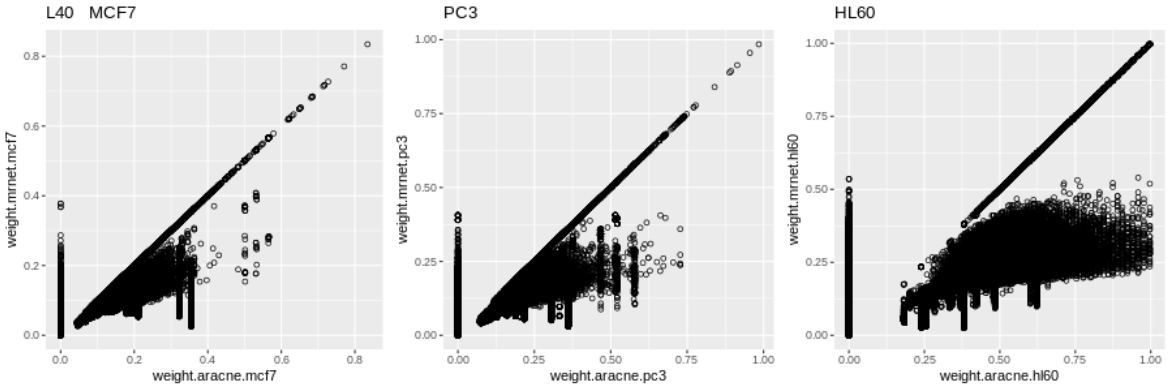

Figure A.22: Comparison between the algorithms ARACNE and MRNET for the disease L40 in each cellline

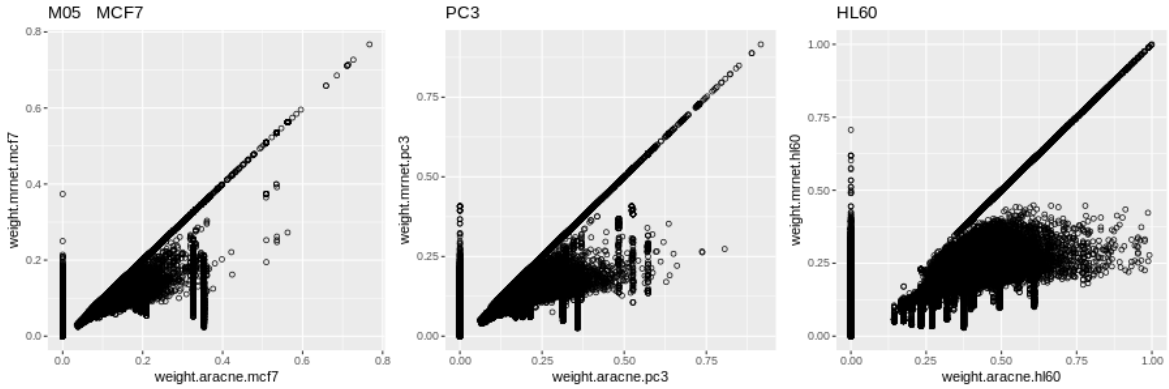

Figure A.23: Comparison between the algorithms ARACNE and MRNET for the disease M05 in each cellline

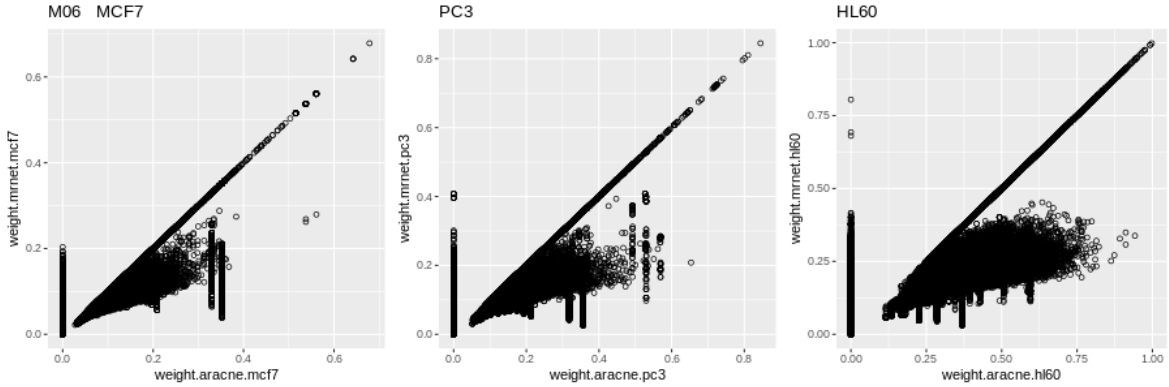

Figure A.24: Comparison between the algorithms ARACNE and MRNET for the disease M06 in each cellline

Comparison between ARACNE and GENIE3

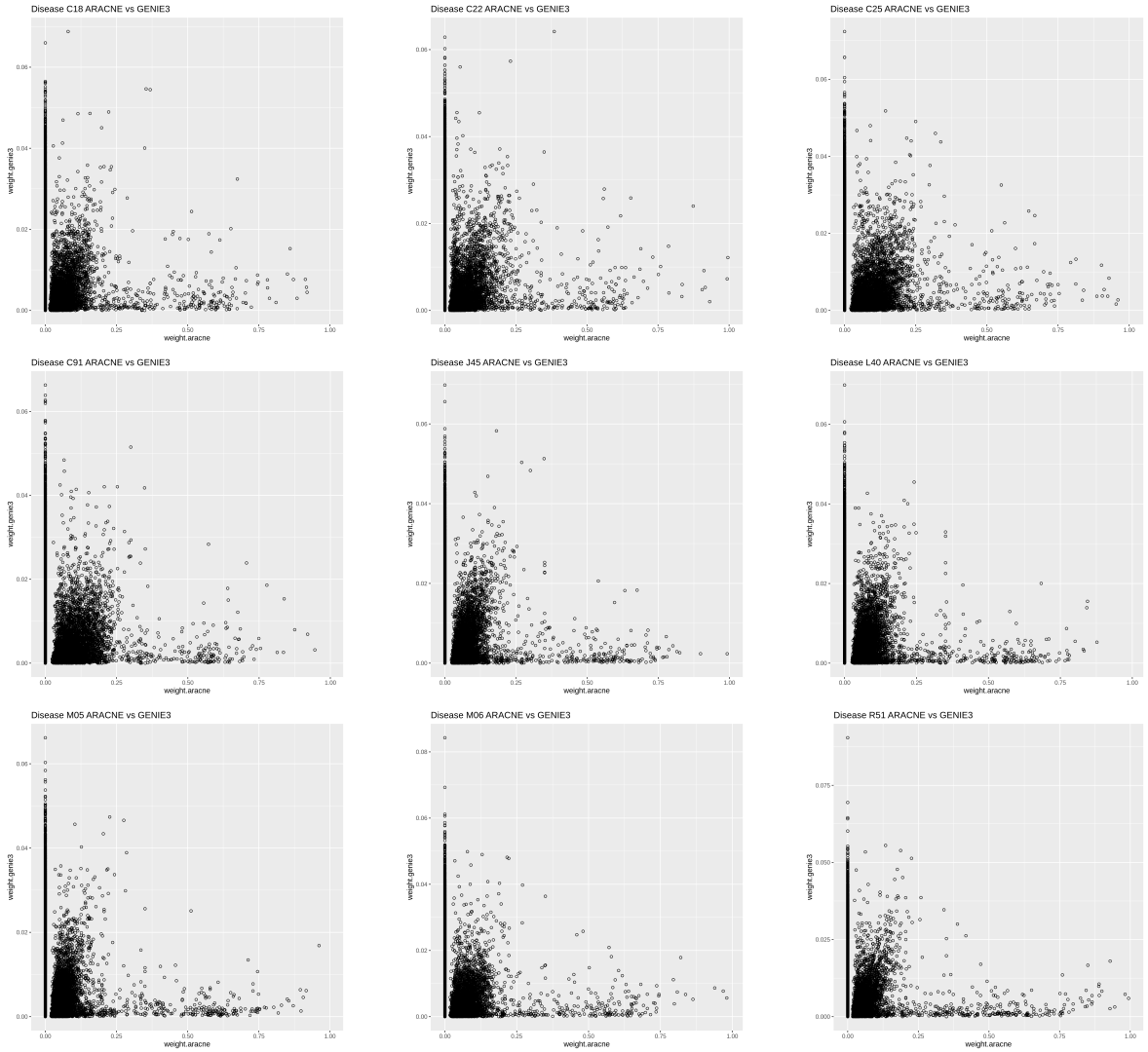

Figure A.25: Comparison of ARACNE to GENIE3 for the diseases C18, C22, C25, C91, J45, L40, M05, M06 and R51 using the combined cellline data

Created Software

The programs created for this thesis can be found at `https://github.com/thomaseska/bachelor-thesis`.