Imperial College London

IMPERIAL COLLEGE LONDON

MPHIL THESIS

Identifying Therapeutic Weak Spots in Cancer Using Network Analysis

---

Author:
Denise Anna THIEL

Supervisors:
Dr. Diego OYARZÚN
Prof. Hector KEUN
Prof. Mauricio BARAHONA

Declaration of Authorship

Declaration of originality
I hereby declare that the work presented in this thesis is my own unless otherwise stated and referenced.

Denise Anna Thiel

Abstract

Mathematical network analysis has been proven to be a useful and powerful tool for biological networks including networks of protein interactions, gene similarity and metabolic interactions. Here I use network analysis to model human cancer and predict which genes or reactions are essential for cancer to allow it to grow or recovery from stress. A general assumption for biological networks is that the centrality of a node is in some way reflective of its biological importance. So I evaluated a wide range of weighted and unweighted node centralities and measures derived from centralities to predict reaction essentiality in metabolic networks. The metabolic networks are Mass Flow Graphs (MFGs), based on the Recon2 reconstruction of human metabolism and constrains on reaction fluxes from the PRIME algorithm. The edge weights in the networks are computed from Flux Balance Analysis (FBA) results in a selection of human cancer cell lines from NCI-60. I could not detect a direct connection between node essentiality and any centrality, but there is a correlation between the overall change of the centrality distribution in the inhibited condition compared to wild type and the inhibited reaction essentiality. With this I have found a promising network measure that can be used to predict possible drug targets. With MFGs a wide range of cellular conditions can be modelled, but only when we know what the cellular objective for FBA is. When cancer cells are put under stress through treatment, they adapt their metabolism to react to the stress. This dynamic process with fluctuating gene expression is difficult to capture in a metabolic network. A better way to analyse the recovery process is to extract which genes are active during which phase. I evaluated seven time points of gene expression data for multiple myeloma cells that were treated with a proteasome inhibitor (PI), which disrupts the protein recycling process. From the pairwise gene expression similarity I constructed network to cluster the genes into groups that are active at the same time. The networks were clustered with a random walk algorithm called Markov Stability and evaluated with gene enrichment analysis. From the resulting clusters, collaborators were able to extract tRNAs that activate a protein called GCN2 that is essential for recovery. Followup experiments showed that a combination of PI and GCN2 is lethal for multiple myeloma as well as a few other cancer cells.

Acknowledgements

This research was initiated by and executed under the supervision of Dr Diego Oyarzún, Prof. Hector Keun and Prof. Mauricio Barahona. The MFG analysis is based on the master's project of Varshit Dusad, with whom we also published the review over using network analysis for metabolic modelling. The gene similarity clustering project comes from a collaboration with the group of Dr. Holger Auner, who produced the data and proposed the research question. Our final approach was developed and executed with Dr. Zijing, a former PhD student from Prof. Mauricio Barahona's group. In a side project with Prof. Hector Keun I got involved in a very interactive and productive collaboration with Dr. Léa Maitre and her colleagues in Barcelona that led to a publication.

I would like to thank my examiners for the early stage assessment and late stage assessment John Pinney, Nick Jones and Philipp Thomas for their positive feedback, and Agnieszka Damasiewicz for her swift and efficient work in all administrative matters.

I am very grateful for the meetings I had with our postgraduate welfare tutor, Gunnar Prüssner, who helped me stay motivated and got me into contact with the groups of Henrik Jensen, Kim Christiansen and Tim Evans, which whom I worked as a GTA. The computer labs with the other GTAs Henry Price, Mads and Gino were a highlight in my week and Henry was a big help in preparation for the class.

The group of Hector Keun was very welcoming, in particular Emily Barnes and Chiharu Wickremesinghe, who gave me a room to stay when I was visiting London. My time in Edinburgh was improved by Arin, Ricardo, Suzanna, Vanessa as well as PhD students from another labs: Diane, Alan, Kevin and Tilly.

Finally and foremostly I am happy that Dr. Mona K Tonn was there to accompany me through all stages of my PhD/MPhil and I am happy that in her I found a good friend.

Contents

1 Introduction **1**
1.1 Essentiality and synthetic lethality 1
1.2 Network modelling 3
1.3 Networks to describe biochemical interactions 4
1.3.1 Metabolic networks 5
1.4 Networks to represent biological data 5
1.5 Objectives and outline 6

2 Mathematical description of networks **8**
2.1 Types of network 8
2.2 Network construction 9
2.3 Network analysis 10
2.3.1 Node properties 10
2.3.2 Community detection 12

3 Analysis of cancer metabolic networks **13**
3.1 Network modelling of metabolism 13
3.1.1 Flux Balance Analysis 14
3.1.2 Mass Flow Graphs 16
3.1.3 Construction of human cancer MFGs 18
3.2 Essentiality prediction of metabolic reactions 18
3.2.1 Essentiality from FBA KO-simulations 19
3.2.2 Correlation between centrality measures and essentiality . . 21
3.2.3 Changes in centrality upon KO 23
3.2.4 New network measure from centrality changes 25
3.3 Validation via E. coli models 27
3.3.1 MFGs for E. coli 27
3.3.2 E. coli predicted essentiality 28
3.4 Conclusions 29

4 Network clustering of gene expression time-series data **30**
4.1 Proteotoxic stress recovery in cancer cells 30
4.2 Processing expression time series data 31

4.2.1 Approaches for time series data 31
4.2.2 Data pre-processing 32
4.2.3 Gene similarity via Gaussian Processes 33
4.3 Clustering a gene similarity network 35
4.3.1 Gene network construction 35
4.3.2 Markov stability clustering 36
4.4 Biological interpretation of gene clusters 40
4.4.1 Enrichment analysis 41
4.4.2 Discovered Vulnerability 44
4.5 Conclusions 44

5 Discussion and Outlook **45**
5.1 Summary 45
5.2 Limitations and strengths 46
5.3 Future Work 46
5.3.1 Similarity networks 46
5.3.2 Reaction essentiality 47
Biological validation 47
Node roles 47
Synthetic lethality 48

Bibliography **49**

List of Figures

1.1 **Diagram of metabolism of a cancer cell** This simplified diagram depicts the core metabolism of a cancer cell using glucose to produce energy. The figure was taken from [4]. 2

1.2 **Synthetic lethality** Often there are alternative pathways that help cells survive when single genes are disrupted. When there is already a mutation present, for example in a cancer cell, the synthetically lethal gene has to make up for it. Otherwise the cell dies even when each gene individually is not essential. Diagram taken from [10]. 3

1.3 **Three examples of widely used network approaches in biology.** A: Gene regulatory networks are directed networks where interactions can be activating (blue) or repressive (red) [15]; B: Protein-protein interaction networks are undirected networks of proteins that can chemically interact with each other [16]; C: Metabolic networks are directed networks of metabolites and reactions that transform them into each other [17]. 4

2.1 **Types of networks** This is a small selection of possible network types. In A the edges have a set direction from source node to target node, the weights in B can be interpreted as distances between nodes, the network C is partitioned into two groups of nodes that do not have any connections inside the groups. 9

3.1 **Diagram of Flux Balance Analysis** In FBA the cone of possible flux solutions is defined through the subject 1) and additional constraints 2) of the optimization problem. The objective function Z is optimized by moving along the edges of the flux cone until the solution cannot be improved anymore. Figure taken from [34]. . . . 15

3.2 **Overview for building and MFG** The Mass Flow Graph is a reaction-based graph that is constructed from the stoichiometric matrix of a system. Optimizing for example biomass production we obtain a flux vector that is used to weight the edges that represent metabolite flow between reactions. 17

3.3 **BT-549 FBA results for all single reaction knock-outs** 20
3.4 **Comparison of node centralities with reaction essentiality** In the table I list the Pearson correlations for all combinations of centralities and essentiality measures. The plots below are selected examples showing the full distribution of reaction's centralities versus essentiality from the cells marked in red in the table. 22
3.5 **Changes in flux and node centrality between SUCD1m-KO and wild type in BT-549.** The ratio of KO-flux over wild type-flux . . . 23
3.6 **Changes in node centralities between treated and wild type BT-549** In order to determine the changes between conditions and the measures representing the difference I plot node centralities from wild type and SDH knock-out against each other. The red line represents a theoretical perfect fit. 24
3.7 **Three essentiality levels** Plotting the pageRank distribution in wild type versus knock-out for three reactions taken from the three severity groups identified in Figure 3.3 shows a growing difference in centrality with increasing change in growth rate. Rho is the correlation coefficient between pageRank distributions. 25
3.8 **PageRank correlation versus essentiality in BT-549 cancer** The new node measure σ of pageRank correlates with the essentiality measure λ with a correlation coefficient of 0.74. The colors correspond to the letahlity groups defined before. Green are non-essential, yellow mildly essential and red are lethal. 26
3.9 **PageRank correlation versus essentiality in four cancer cell lines** The new node measure σ of pageRank correlates with the essentiality measure λ in four other human cancer cell lines. 27
3.10 **FBA results for all single reaction knock-outs in E. coli** For each reaction I perform the inhibited FBA and record the growth rate ratio. The reactions were binned into four groups, compare to human cancer binning. 28
3.11 **PageRank correlation versus essentiality in E. coli** The new node measure σ of pageRank correlates with the essentiality measure λ. . 29

4.1 **Proteasome-Inhibition experimental design** The multiple myeloma cells were treated with a proteasome inhibitor on day 0. By day 2 half of the cells had died, but the remaining cells adapted to the stress and were proliferating again by day 6. RNA-seq experiments were performed on days 0, 1, 2, 4, 6, 8 and 10. The figure was taken from slides from Holger Auner. 31

4.2 **Log-histogram of the expression sum over all days for all genes.** Summing over all days' normalized expression data, we get a bimodal distribution. I employed the logarithmic plot to choose the expression cutoff as the minimum between the two modes, thereby separating randomly detected genes and reliably expressed genes. 33

4.3 **Examples of Gaussian Process regression on single gene expression.** The colored lines are the time courses for different replicates, the black line in the middle is the posterior mean and the grey area around it the 95% confidence interval. 35

4.4 **Gene-similarity graph** Shown are 2542 expressed genes with high variance upon PI-treatment are connected by edges, when their expression profiles are similar. This network is colored by the genes' expression on day 1 relative to day 0. Red: expression went up, blue: expression went down relative to day 0. The figure was created by Dr. Zijing Liu. 36

4.5 **Stability clustering of gene similarity network - choosing the best clustering.** A: The gene expression distance graph was clustered by performing random walks at Markov times from 1 to 100, plotting the resulting number of clusters and the pairwise Variation of information between clusterings. B: For the same clusterings the biological homogeneity index (BHI) was computed. From both these plots the best clustering was chosen from the time, where the number of clusters is stable, the Variation of information low and the BHI high. This figure was created by Dr. Zijing Liu. 39

4.6 **Clustered gene expression network** This distance graph was obtained from gene expression similarity based on GP-regression. The Markov-stability clusters (marked by color and number) were ordered by their position in the graph and enriched gene GO-term functions. This figure was taken from [23] 40

4.7 **Expression profiles of six cluster representatives.** These representative expression profiles were obtained through GP-regression on all genes for each cluster. The six clusters were grouped depending on their initial reaction and long-term behavior. This figure was created by Dr. Zijing Liu and Prof. Mauricio Barahona. 41
4.8 **GO-term enrichment analysis** These are the GO-terms that are enriched in the six gene expression clusters using clusterprofiler. This figure was adapted from Dr. Zijing Liu. 42
4.9 **KEGG-pathway enrichment analysis** These are the KEGG pathways that are enriched in the six gene expression clusters using clusterprofiler. This figure was adapted from Dr. Zijing Liu. 43

Introduction

Cancer is one of the most widespread, deadly diseases worldwide in spite of combined effort across different fields of research. Various treatment options have been developed in cancer ranging from radiation therapy to chemotherapy to surgically removing tumors by cutting it out or using heat to destroy the cells. However, treatment methods are often not sufficiently cancer-specific and accompanied by serious side-effects [1].

More specific treatments can be developed by evaluating the gene activity of cancer versus healthy cells. In the last decade, considerable advances in cancer genomics have led to the identification of an extensive number of cancer-related genes [2],[3]. Unfortunately, the mechanisms in which they exert their carcinogenicity are often not well understood. How does the expression of certain genes change the state of a cell? To answer this question, we have to look at the interactions and function of genes and their involvement in the different pathways of metabolism that are active in a cell. Modelling metabolism additionally lets us incorporate other misregulated mechanisms in the cell that cause cancer as well as external influences. See Figure 1.1 for a simplified view of the central carbon metabolism and other subparts of metabolism. What makes this analysis difficult is that metabolism is robust against small disturbances and cells can adopt alternative pathways to balance out perturbations. Thus it is not enough to look at single agents, but we have to model cancer systems as a whole, taking into account long range interactions and how subgroups have to act together to achieve a common goal. Only with a complete model can we assess which parts are important under which conditions and predict weak points that can be exploited for cancer treatment. In this thesis I explore different approaches to model cancer via mathematical networks and extract important genes or reactions that represent druggable targets.

1.1 Essentiality and synthetic lethality

Two common questions in cancer research are which genes are responsible for carcinogenesis and which can be targeted to selectively remove the cancer. Genes that

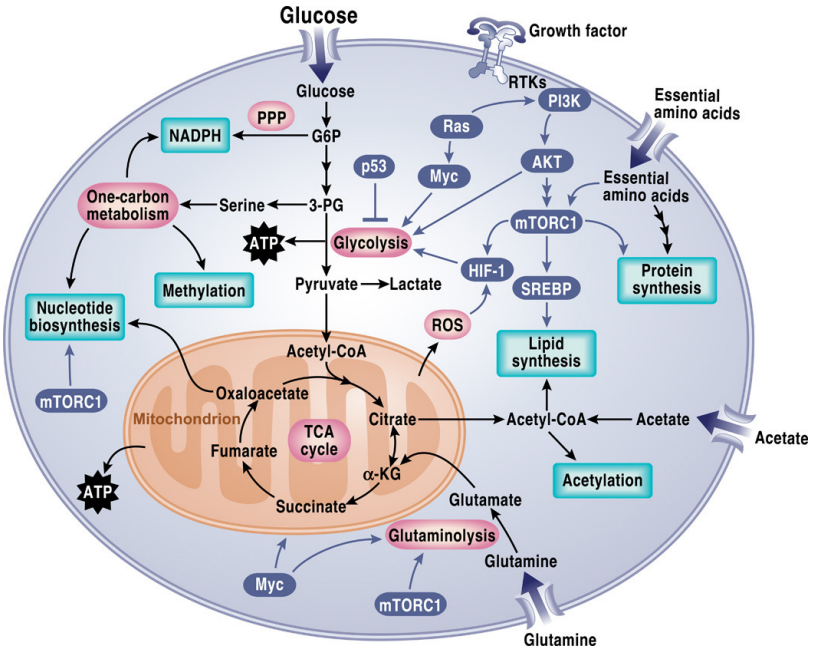

FIGURE 1.1: **Diagram of metabolism of a cancer cell** This simplified diagram depicts the core metabolism of a cancer cell using glucose to produce energy. The figure was taken from [4].

enable cells to grow are called *essential* genes. A typical method to detect essential genes is to inhibit single genes with RNAi or CRISPR [5]. With CRISPR the genetic code can be altered at a specific site to insert, delete or replace nucleotides permanently in a high-throughput fashion. Depending on where in the gene the change is made, the effect can be silent (no effect), the gene might be read less often or the encoded protein could degrade more quickly (knock-down), or the gene might be stopped from being translated into RNA altogether or the resulting protein is misfolded to a degree where it cannot perform its function (knock-out). Metabolites on the other hand have vastly different chemical properties and thus do not allow for high-throughput inhibition methods. The only option is to do mass spectrometry or nuclear magnetic resonance [6] which can resolve only a limited number of molecules. Therefore, metabolites and their interactions have to be studied individually, leading to whole publications to be written about single inhibitors [7].

We can distinguish two types of essentiality. One is universal across cell types and therefore not useful for treatment because healthy cells will be killed along with diseased cells. On the other hand, there are genes that are only essential under certain conditions, depending on the presence of other genes. This leads to the concept of *synthetic lethality* in which the combined inhibition of genes reduces cell fitness, while each gene knock-out separately has no effect. See Figure 1.2 for a

schematic overview. For example, human cells the combined inhibition of pyruvate carboxylase (PC) and succinate dehydrogenase (SDH) impedes the use of the TCA cycle [8], while each inhibition separately does not have a detrimental effect. SDH is often deregulated in cancers [8] but not in healthy cells, so inhibiting PC is a potential treatment. However, beyond a handful of cases little is known about where synthetic lethality can occur and how to use it in treatment. Information about synthetically lethal genes is especially valuable since the combined effect can be specific to malignant cells, keeping the side effects to a minimum [9].

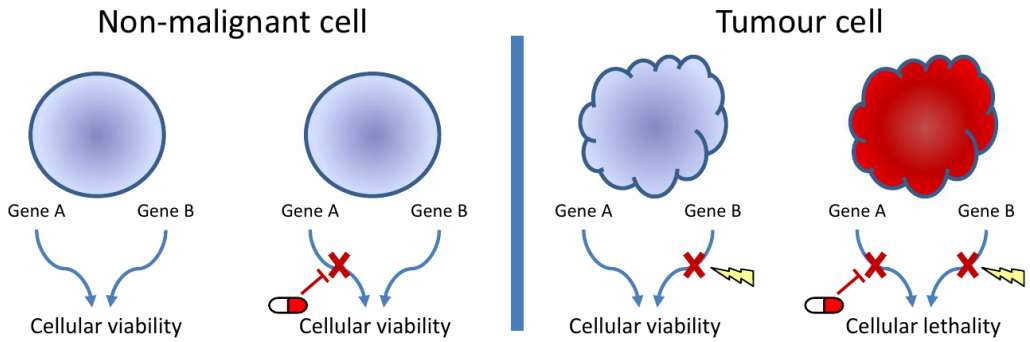

FIGURE 1.2: **Synthetic lethality** Often there are alternative pathways that help cells survive when single genes are disrupted. When there is already a mutation present, for example in a cancer cell, the synthetically lethal gene has to make up for it. Otherwise the cell dies even when each gene individually is not essential. Diagram taken from [10].

1.2 Network modelling

Computational modelling of metabolism promises to identify synthetically lethal candidates in any cell line, along with an interpretation of how the lethality arose. Also it enables us to evaluate combinations of more than two genes from an exponentially growing solution space, while it is not feasible to test every combination of gene knock-out and drug treatment experimentally. A modelling approach that is highly versatile and copes well with big model sizes, are networks. Mathematical network modelling is utilized in many research areas, ranging from social networks, transportation networks, power grids and many more [11].

There are fundamentally two uses of network modelling. On the one hand they are a mathematical representation of a system of agents that interact with each other. These agents can be humans, animals, molecules, etc. For such systems the network is often the simplest but most complete representation of agent interactions.

It lets us analyse how the system functions as a whole and how its agents exert their function. Also it can be a tool for visualizing data. On the other hand networks are a powerful tool for data modelling, where the nodes represent agents, but there is no real-world interaction assignable to the nodes. For example, we can introduce interactions solely based on data similarity. With this we can make predictions about functional similarity of the agents and extract groups that behave similarly but are not immediately visible from the data.

1.3 Networks to describe biochemical interactions

In biology, networks have been successfully used to analyse gene regulatory networks [12], to represent protein interactions [13], or to model metabolic pathways and chromatin interactions [14]. All these networks may vary in their type, size and connectivity but are usually used to extract high-level order information that would be lost with other modelling approaches that only include individual interactions.

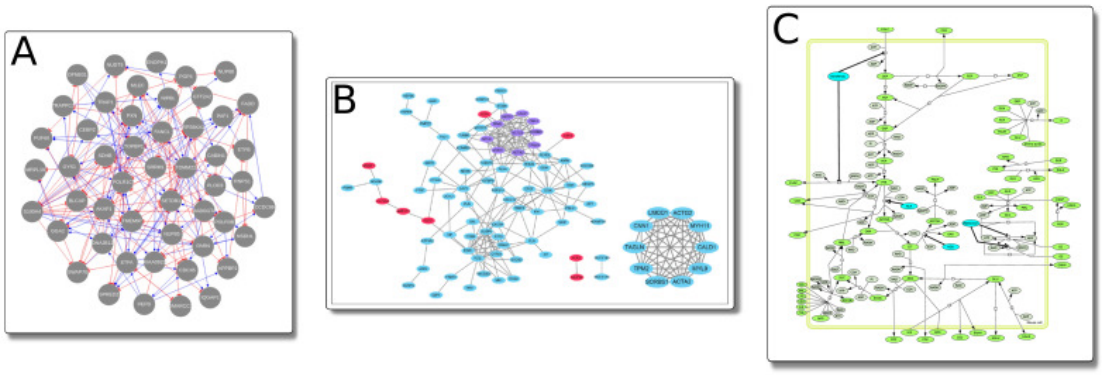

FIGURE 1.3: **Three examples of widely used network approaches in biology.** A: Gene regulatory networks are directed networks where interactions can be activating (blue) or repressive (red) [15]; B: Protein-protein interaction networks are undirected networks of proteins that can chemically interact with each other [16]; C: Metabolic networks are directed networks of metabolites and reactions that transform them into each other [17].

Genes can regulate each others' expression by their RNA or protein binding to another gene's enhancer or promoter region and thereby influencing the activity DNA-polymerases. Often the same protein can bind to many different sites, even of other transcription factor genes, thereby triggering cascades of activation and repression. Through transcription factors, cells can quickly switch their state and

react to external changes. Analysing which genes are regulated together helps us understand their function, and gene regulatory networks are a standard model for analysing the interactions in many different species and diseases [12], [18]. Similarly, protein-protein networks or protein-protein interaction networks (PPIs) help identify which proteins work together or have similar functions [16], [13].

1.3.1 Metabolic networks

Metabolism comprises all biochemical reactions reactions taking place in a living cell and the metabolites transformed in them. This includes all reactions for taking up nutrients, converting them to energy and building blocks, and excreting accumulating waste products. We can discern two possible components for metabolic networks: metabolites and reactions. Therefore they can be modelled as metabolite-based, reaction-based or contain both. Each has a different focus and can be used to answer different questions.

A shared problem that comes with all three approaches is the question of how to treat pool-metabolites. These metabolites are ubiquitous in cells, participate in hundreds of reactions and are therefore highly connected. But for the same reason they don't hold any information about the cell type and end up obscuring the relevance of more specific metabolites. Therefore pool-metabolites are usually arbitrarily excluded from metabolite-based networks altogether, without a universal consensus of which metabolites constitute as pool-metabolites [19].

One of the newest additions to metabolic network modelling are *Metabolic Flow Graphs* (MFG) [20]. MFGs are reaction-based metabolic network that are weighted by the flow of metabolites between reactions. With these weights, pool-metabolites do not pose a problem because they only introduce very weak edges. High-throughput genomic data can be used to predict the flow of metabolites, allowing for highly specific models for different cell lines and conditions.

1.4 Networks to represent biological data

Even if there are no interactions between agents, network modelling can be used to model and analyse data by introducing connections, for example based on the similarity of agents.

Gene similarity can be defined on many levels that are linked to the steps in the *central dogma* [21]. First of all, the similarity of the DNA sequence can be evaluated. The place in the DNA could be an indicator, comparing which other genes are close by or accessible at the same time. Another important characteristic for function is what transcription factors can activate the expression of the gene, leading to the second level in the central dogma, RNA. Here we can compare, which genes have a similar transcript count over time and how similar these transcripts are. The 2D and 3D structure of the transcript can be compared. On the last level of the central dogma, we are interested in the similarity of the encoded protein, should the gene be a coding gene. An abundance of chemical properties can be measured for proteins that together describe it. Of special interest here are functional sites in the protein, that can interact with other molecules. And finally, the localization and function of that protein as listed in gene ontologies [22] can identify if two genes have similar functions.

From similarity networks we can predict which genes or proteins have to work together to perform certain functions by comparing their similarity in being expressed at the same time. This can be used to infer function for genes/proteins that have not been well studied yet or might be difficult to assign a function to. Incorporating data from different species into the same network, we can show if completely novel genes are actually well-conserved and infer their function from the known genes from other species.

1.5 Objectives and outline

In this thesis I use network analysis to extract essential genes and reactions which can be used as drug targets for cancer treatment following two different approaches. In the first, I build Mass Flow graphs to represent metabolism in a number of cancer cell lines, seeking to identify conditionally essential metabolic genes/reactions. In the second, I build a network based on co-expression of genes to predict genes essential during recovery from drug-induced proteotoxic stress.

In Chapter 2, I give a general introduction to networks and formulate the mathematical methods of network analysis including node centralities and module detection, which are the main methods in Chapters 3 and 4.

Chapter 3 describes my main project, evaluating Mass Flow Graphs of cancer metabolism to predict essential reactions in different cancer cell lines and conditions. I compute several centralities and eventually define a new node measure

that is able to predict reaction essentiality without having to run computationally expensive simulations for each reaction. Related to this, I was involved in writing a review about different approaches that can be used when modelling metabolism via networks:

Varshit Dusad, Denise Thiel, Mauricio Barahona, Hector C. Keun, Diego A. Oyarzún. "Challenges and opportunities at the interface of network science and metabolic modelling". In: *Front. Bioeng. Biotechnol.* 25 January 2021

Chapter 4 contains a gene expression similarity network built for cancer cells that have been treated with a proteasome inhibitor. I develop a pipeline to clean the data and filter for genes that are involved in the recovery process. From these genes, we build a network where genes are connected if their expression profiles are similar, and then cluster the gene profiles to predict which genes are involved in the recovery process and thus present possible vulnerabilities when inhibited. The results of the proteasome-inhibition project were published in PNAS [23]:

Paula Saavedra-Garcia, Monica Roman-Trufero, Hibah A Al-Sadah, Kevin Blighe, Elena Lopez-Jimenez, Marilena Christoforou, Lucy Penfold, Daria Capece, Xiaobei Xiong, Yirun Miao, Katarzyna Parzych, Valentina Caputo, Alexandros P Siskos, Vesela Encheva, Zijing Liu, Denise Thiel, Martin F Kaiser, Paolo Piazza, Aristeidis Chaidos, Anastasios Karadimitris, Guido Franzoso, Ambrosius P Snijders, Hector C Keun, Diego Oyarzun, Mauricio Barahona and Holger W Auner. "Global profiling of cancer cell recovery from therapy-induced stress reveals druggable vulnerabilities". In: *Proceedings of the National Academy of Sciences* 118.17 (2021) PNAS

Finally, I summarize my findings in Chapter 5, pointing out the limitations and follow-up analyses that can be done. In particular, I suggest different ways of predicting synthetic lethality, which is one of the main areas of interest for predicting drug targets, but has eluded researchers for decades. MFGs might provide an access point here.

In a side project (not presented in this thesis) I visualized networks of prenatal and postnatal metabolites associated with external exposures, which became part of another publication:

Maitre L, Bustamante M, Hernández-Ferrer C, Thiel D, Lau C-H, et al. "Multi-omics signatures of the human early life exposome". In: *medRxiv* (2021), **doi:** https://doi.org/10.1101/2021.05.04.21256605

Mathematical description of networks

A network or (network) graph $G = (V, E)$ is defined by two sets: the set of nodes, V, and the set of edges, E. Hereby edges represent tuples of nodes in the form $e_m := (v_i, v_j)$. Two common graph notations are the *edge list*, which is a list of all node tuples, and the *adjacency matrix*, A. This $|V|x|V|$ matrix has non-zero entries at A_{ij} if the edge (v_i, v_j) exists, and zero otherwise.

2.1 Types of network

We can distinguish different architectures of networks that require specific notations and tailored analyses. A selection of network types used in this thesis can be seen in Figure 2.1. When A is symmetric, that means that for every edge (v_i, v_j) the reverse edge (v_j, v_i) exists, the graph is called *undirected*, otherwise the graph is *directed* (Fig. 2.1:A). If a network is *weighted* (Fig. 2.1:B), the value of an entry in A can be interpreted as the corresponding edge weight, which are all set to one in an *unweighted* network. For weighed networks, the edge list notation has to change to triplets in the form: $e_m = (v_i, v_j, w_m)$. Networks can be classified as *simple* graphs, where self-loops and multiple edges between the same two nodes are not allowed, or *multigraphs*, where an unlimited number of edges between nodes is allowed, including self-loops. The absence of self-loops can easily be assessed by checking, that the diagonal of the adjacency matrix is zero. Having different types of nodes leads to the idea of multipartite networks, which can be partitioned into several groups of nodes, and edges are only allowed between groups. When a network can be split into exactly two groups of nodes A and B with no interactions between nodes of the same group, this is called a *bipartite* network (Fig. 2.1:C). Bipartite networks have many uses for example in matching problems. But they can also be projected down to a unipartite network with only one type of node, for example from group A. In the projected network nodes are now connected if they both had an edge to the same node from group B in the bipartite version. There will always be a loss of information with the projection, but it allows for more analyses since unipartite graphs are more well-studied.

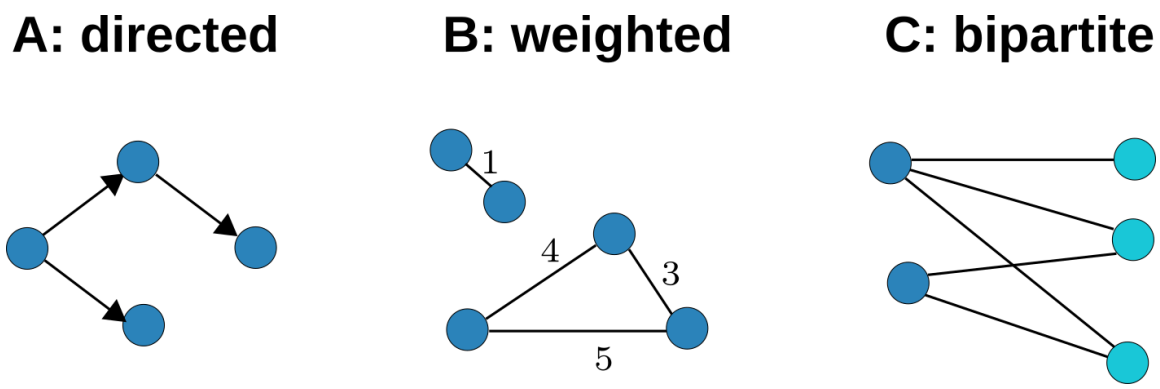

FIGURE 2.1: **Types of networks** This is a small selection of possible network types. In A the edges have a set direction from source node to target node, the weights in B can be interpreted as distances between nodes, the network C is partitioned into two groups of nodes that do not have any connections inside the groups.

Many network measures depend on the concept of *paths* between nodes. A path between two nodes i and j is a succession of edges that starts at node i (the source) and ends in node j (the target) whereby the following edge always starts at the node where the previous edge ended. The length of a path between two nodes i and j is the number of edges connecting those nodes and the length of the shortest path between them can be interpreted as their distance $d(i, j)$. If there exists a path between node i and j, node j is *reachable* from i. If all nodes in a network are reachable from each other, the network is called *connected*. Any subgraph of the network where all nodes are reachable from each other, is called a *connected component* (CC) and the biggest CC is commonly called the giant connected component (GCC).

2.2 Network construction

Constructing a network, we can choose from the aforementioned network types. Though directionality and weights are not always available when building from real interactions.

When the network is built from data, it is often constructed from similarity matrices, for continuous variables in biology often via covariance matrices [24]. Usually this similarity matrix is first converted into a distance matrix by computing an inverse of the similarity matrix. This way similar genes end up close to each other, which makes it more intuitive to look at and easier to analyse for a lot of network algorithms. Still, in both the similarity and the distance matrix there exists an edge value for each pair of nodes, which represents a fully connected network.

Any structural information we might want to extract is obscured by a multitude of weak edges.

Two common ways of pruning the networks, i.e. removing weak edges which in this case are edges between very distant nodes, are *k-nearest neighbors* (kNN) and *ϵ-environment*. For k-nearest neighbors we start with an empty network and only add edges from each node it its k-nearest neighbors. With and ϵ-environment, only edges with a distance value smaller than a chosen cutoff ϵ are included. Tuning these parameters we can optimize the constructed network such that it is sparse enough but complete enough to capture the most important properties hidden in the data.

2.3 Network analysis

Networks contain ample information that can be analyzed ranging from global features like *size* (number of nodes) and overall connectivity to local features for single nodes. I look at node *centralities* to describe the biological agents in my networks. Another area of interest in network analysis is *module detection* or *clustering*. *Modules* (or *clusters*) are subgroups of nodes that are highly connected within the subgroup, while having few connections to the rest of the network. The names 'cluster' and 'module' are often used interchangeably, but for clustering we usually extract a predefined number of clusters, while community detection algorithms can optimize module membership for a variable number of modules.

2.3.1 Node properties

Node *centrality* is a measure for how well-connected a node is, which can be interpreted as a measure for importance of the agent it represents. Some centralities are designed for undirected graphs and others need a directed graph. In the following, I define four commonly used centralities on the adjacency matrix A, where A_{ij} is an edge from node i to node j.

The *degree* centrality of a node is defined as the number of outgoing and incoming edges. In directed networks the degree can be split into *incoming* and *outgoing degree*, whereas in undirected networks it is the sum of all incident edges. These measures are also used in several other centralities. The degree distribution is often used to describe the network structure. Weights can be added by summing

all incident edge weights.

$$\text{indegree}(i) = \sum_{j=1}^{n} A_{ji}, \tag{2.1}$$

$$\text{outdegree}(i) = \sum_{j=1}^{n} A_{ij}, \tag{2.2}$$

$$\text{degree}(i) = \text{indegree}(i) + \text{outdegree}(i). \tag{2.3}$$

A centrality that indirectly uses the degree is the *PageRank*, which iteratively traverses a network to determine how important each node is, usually in directed networks. For the PageRank we can consider an edge-weighted version and node-weighted version.

$$\text{PageRank}(i) = \frac{1-\alpha}{N} + \alpha \sum_{j} A_{ji} \frac{\text{PageRank}(j)}{\sum_i A_{ji}}, \tag{2.4}$$

where α is a scalar damping factor.

Evaluating shortest paths through the network allows for more global measures like betweenness and closeness. The *betweenness* centrality measures how often a node is needed to connect other nodes with each other and can be interpreted as a measure for how well a node connects different subgraphs of the network.

$$\text{betweenness}(k) = \sum_{i \neq j \neq k \in V} \frac{\sigma_{ij}(k)}{\sigma_{ij}}, \tag{2.5}$$

where σ_{ij} is the number of shortest paths from i to j and $\sigma_{ij}(k)$ is the number of shortest paths between vertices i and j that pass through node k.

Another global measure and maybe most intuitive measure for the question of how central a node position in the network is, is the *closeness* of a node. It directly takes into account the distance of a node to all other nodes in the network and is usually normalized by the total number of nodes. The closeness can be normalized by the ratio of all nodes reachable nodes r_i by node i out of all other nodes. For 'incloseness' we turn it around and look at all nodes that can reach node i.

$$\text{incloseness}(i) = \frac{r_i}{N-1} \frac{1}{\sum_j d(j,i)}, \text{outcloseness}(i) = \frac{r_i}{N-1} \frac{1}{\sum_j d(i,j)}. \tag{2.6}$$

If no nodes can reach node i or are reachable from i, then the closeness is zero.

Apart from the standardized known centralities, new centralities can be designed

by combining existing ones, assembling more path information or other network properties. For example, the flow profile of a node can be constructed from the number of all incoming and outgoing paths of a certain length [25]. Another way of constructing a global node feature that does not use paths, is to iteratively collect neighboring nodes' features. This is done in the feature extractor ReFeX [26], which sums and averages features from a node's egonet. The egonet is defined as the subgraph containing the node itself, all its neighbors and all their incoming or outgoing edges.

Apart from using purely graph features, external information from the agents can be added to construct a more generally informed node measure. In protein-interaction networks for example, chemical properties of the proteins could be included, or even information from the genes encoding the proteins.

2.3.2 Community detection

A popular application of networks is to identify subgroups of nodes that are densely connected to each other and therefore strongly influence one another, possibly sharing a common function. *Clusters* or *modules* of nodes that are highly connected to each other but have few connections outside the cluster can be extracted with a variety of *community detection* or *clustering* algorithms. These two terms are often used interchangeably, but for clustering the number of modules is usually preset.

Commonly used clustering methods are *modularity clustering*, *spectral clustering* and algorithms based on *random walks*. For modularity clustering, node assignment is randomly swapped while maximizing the modularity measure introduced by Newman [27]. In spectral clustering the network connectivity and membership of nodes to clusters is deduced via the eigenvalues of the Laplacian matrix of the graph [28]. A random walk on a graph is a stochastic diffusion process [29], where edge weights are used to generate transition probabilities from one node to the next. We can conceptualize the process as a random walker traversing the network from a random starting point using the transition probabilities for each step. Modules are extracted as subgraphs where the random walker is trapped in for a long time. In my work I use a random walk algorithm called *Markov stability* [30].

Analysis of cancer metabolic networks

In this chapter I present an approach to model and evaluate metabolic networks, with the goal of finding a network metric that correlates with essentiality of reactions in cancer cells. The network metric can be employed to quickly simulate any combinations of reaction inhibitions and their effect on the network, thus providing hypotheses for essential reactions that can be tested in the lab.

3.1 Network modelling of metabolism

Metabolism can be modelled through a list of all metabolic reactions and their kinetic properties. Given a vector X of n metabolites we can formulate the m metabolic reactions R

$$R_j : \sum_{i=1}^{n} \alpha_{ij} X_{ij} \rightleftharpoons \sum_{i=1}^{n} \beta_{ij} X_{ij}; j = 1, 2, ..., m. \tag{3.1}$$

Where the coefficients α for consumption and β for production for each reaction are obtained from chemical experiments.

From this we define the $n \times m$ stoichiometric matrix $\mathbf{S}$ with entries $\mathbf{S}_{ij}$:

$$S_{ij} = \beta_{ij} - \alpha_{ij}. \tag{3.2}$$

Thus, $\mathbf{S}$ has negative entries for all metabolites consumed in a reaction and positive entries for all metabolites produced in a reaction.

The turnover of metabolites through reactions, or *flux*, v determines metabolite concentrations x, whose changes can be described through the differential equation:

$$\dot{x} = \mathbf{S}v, \text{ where } v \in \mathbb{R}^m. \tag{3.3}$$

Thus the flux can be used to describe the activity of single pathways up to the overall state of a cell's metabolism. In any given cell the flux is constrained by the availability of enzymes and co-factors facilitating the metabolic reactions.

Lists containing metabolites, reactions, associated genes and flux constraints in a cell are collected into a so-called metabolic reconstruction [31]. Setting up a new metabolic reconstruction requires substantial work, where each reaction has to be experimentally confirmed individually. This is why models for different cell types from the same species are often derived from the same stoichiometric matrix by constraining fluxes with transcriptomic or proteomic data of the facilitating enzymes [9].

3.1.1 Flux Balance Analysis

When studying metabolism ideally as many kinetic factors as possible should be included. For small models, elaborate ODE systems have been built [32] to calculate fluxes, and determine for which flux distributions the system is in steady-state. Unfortunately, genome-scale human metabolism has too many possible reactions to be modelled completely via differential equations with current data [33]. Therefore constraint-based modelling (CBM) emerged as a powerful tool for modelling metabolism. A widespread method for CBM is Flux Balance Analysis (FBA) [34], where the flux through reactions is determined via linear programming. In FBA, metabolism is formulated as an optimization problem where an objective function is maximized along the edges of a polytope of feasible flux distributions. Assuming that the modelled cell metabolism is in steady state [35], we can set $Sv = 0$ in equation 3.3 and use this equation to constrain the solution space. With metabolism heavily relying on feedback loops and alternative pathways, this is not enough to constrain the model enough to get realistic results. Additional constraints on single reaction fluxes v_i (minimal flux a_i and maximal flux b_i) can be inferred through experimental measurement of corresponding enzyme levels. As a proxy, the gene expression of enzyme coding genes is often used to constrain fluxes [36]. For example, in the PRIME algorithm the expression of enzymes is used to first set a global upper and lower bound and then refine single reaction

constraints. See Figure 3.1 for a graphical description of FBA. Typically, FBA optimization is formulated as:

$$\begin{aligned} f &= \max_{v} \; c^T v \\ \text{subject to } \mathbf{S}v &= 0 \\ \text{and } a_i \leq v_i &\leq b_i, \text{ with } i = 1, ..., m. \end{aligned} \tag{3.4}$$

The coefficient vector c selects reactions whose flux contribute to the objective. Typical objectives are maximization of biomass production or ATP synthesis which enable growth and proliferation [35]. Maximization of these objectives is believed to be a valid assumption in cancer cells and microbes, although a cell can have many more objectives at the same time [37] which require different fluxes.

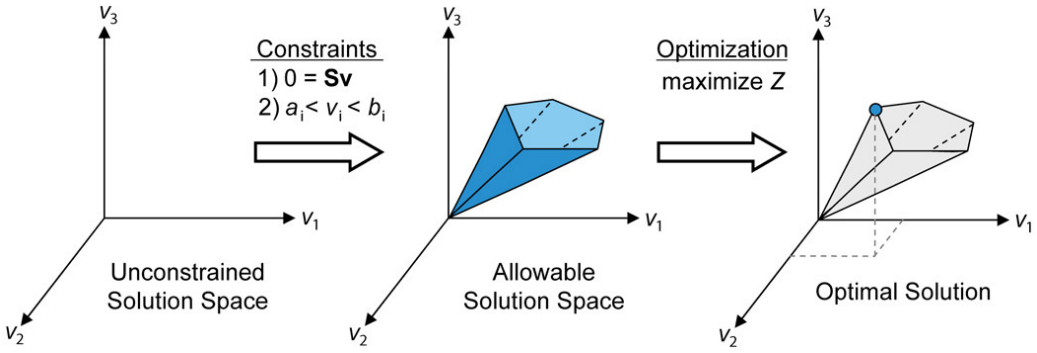

FIGURE 3.1: **Diagram of Flux Balance Analysis** In FBA the cone of possible flux solutions is defined through the subject 1) and additional constraints 2) of the optimization problem. The objective function Z is optimized by moving along the edges of the flux cone until the solution cannot be improved anymore. Figure taken from [34].

The optimal flux vector v^* of the optimization 3.4 is often not unique and we can adjust the FBA by optimizing a secondary objective based on the solution of the first. The first objective does not take into account that creating the enzymes and running the reactions takes energy that the cells minimize while achieving their primary goal. We can simulate this by minimizing the norm of the flux vector v while setting $c^T v$ to the solution of the first FBA optimization. For example:

$$\begin{aligned} \min |v^*| \\ \text{subject to } \mathbf{S}v^* &= 0, \\ c^T v^* &= f \\ \text{and } a_i \leq v_i^* &\leq b_i; i = 1, ..., m. \end{aligned} \tag{3.5}$$

This two-step FBA is called *parsimonious FBA* and not only makes solution more realistic but also computationally more reproducible.

Integrating FBA predictions in a metabolic network enables us to describe the metabolism of the same cell type in different conditions, e.g. nutrient availability, drug treatment or gene knock-out, via mathematical tools from network analysis. The predicted metabolic state can be used to classify medical samples of tumours and predict survival rates [38], while genes essential for the modelled objective can act as biomarkers and targets for new treatment approaches in diverse cancer cells [38].

3.1.2 Mass Flow Graphs

Mass Flow Graphs (MFG) [20] are weighted, directed reaction-based networks. Nodes in MFGs represent active reactions that are connected by a directed edge if one reactions produces a metabolite that us used by another. The predicted FBA fluxes from a certain condition are used to calculate edge weights describing the mass flow of shared metabolites. The edge weight from reaction R_1 to R_2 is defined through the total mass flow of all metabolites X_k shared between these two reactions, assuming that that metabolites are uniformly taken up by all consuming reactions:

$$\text{weight}_{R_1 \to R_2} = \sum_{k=1}^{n} (\text{amount of } X_k \text{ produced by } R_1) \frac{(\text{amount of } X_k \text{ consumed in } R_2)}{(\text{total consumption of } X_k)}. \tag{3.6}$$

See Figure 3.2 for a diagram of how MFGs are constructed. This modelling approach has the advantage that widely shared co-factors such as water and NAD contribute little to the weights and do not have to be arbitrarily excluded.

To obtain the variables in equation 3.6, we split reversible reactions and therefore their fluxes into forward and backward. First we split the flux:

$$v = v^+ - v^- = v^+ - \text{diag}(r)v^-, \tag{3.7}$$

where v^+ are the fluxes through forward reactions and v^- are fluxes through reversible reactions that have a backwards net flux. The m-dimensional reversibility vector r is provided with the reconstruction. In this reversibility vector, the j^{th} entry $r_j = 1$ if reaction R_j is reversible and $r_j = 0$ if not.

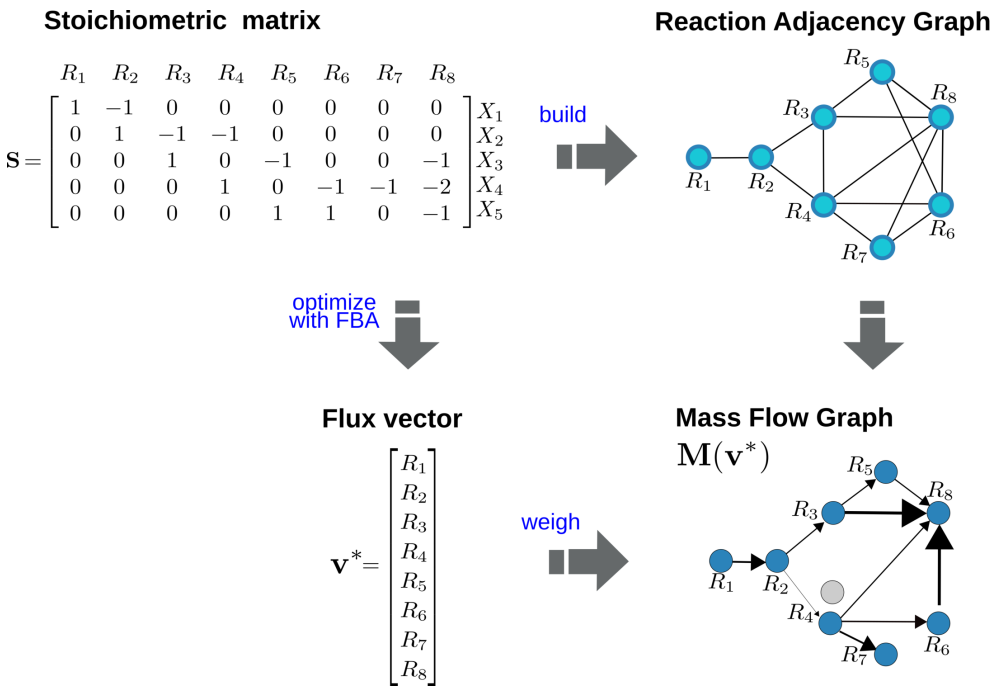

FIGURE 3.2: **Overview for building and MFG** The Mass Flow Graph is a reaction-based graph that is constructed from the stoichiometric matrix of a system. Optimizing for example biomass production we obtain a flux vector that is used to weight the edges that represent metabolite flow between reactions.

The change in metabolite concentrations can be rewritten as

$$\dot{x} = \mathbf{S}v = \underbrace{[\mathbf{S} \quad -\mathbf{S}] \begin{bmatrix} \mathbf{I}_m & 0 \\ 0 & \text{diag}(r) \end{bmatrix}}_{\mathbf{S}^{2m}} \begin{bmatrix} v^+ \\ v^- \end{bmatrix} = \mathbf{S}^{2m} v^{2m}, \tag{3.8}$$

where $v^{2m} = \begin{bmatrix} v^+ & v^- \end{bmatrix}^T$ is the unfolded $2m$-dimensional vector of reaction rates, $(\mathbf{I})_m$ is the $m \times m$ identity matrix, and $\mathbf{S}^{2m}$ is the unfolded stoichiometric matrix containing m forward and m backward reactions.

Similar to the flux decomposition, we can split $\mathbf{S}$ into production and consumption matrices:

$$\text{Production: } \mathbf{S}^{2m+} = \frac{1}{2}(\text{abs}(\mathbf{S}^{2m}) + \mathbf{S}^{2m}), \tag{3.9}$$

$$\text{Consumption: } \mathbf{S}^{2m-} = \frac{1}{2}(\text{abs}(\mathbf{S}^{2m}) - \mathbf{S}^{2m}). \tag{3.10}$$

Finally, for a given flux distribution v^* the adjacency matrix $\mathbf{M}$ of an MFG can calculated, where the entries M_{ij} of the adjacency matrix describe the flow of metabolites from reaction i to reaction j:

$$\mathbf{M}_{ij}(v) = \sum_{k=1}^{n} \mathbf{S}_{ki}^{2m+} v_i^{2m} \left(\frac{\mathbf{S}_{kj}^{2m-} v_j^{2m}}{\sum_{j=1}^{2m} \mathbf{S}_{kj}^{2m-} v_k^{2m}} \right). \tag{3.11}$$

3.1.3 Construction of human cancer MFGs

The models I use for human cancer are based on the genome-scale reconstructions from RECON2 [31], constrained with gene expression data via the PRIME algorithm [36]. They provide models for all cell lines from NCI-60 [39], a collection of 60 well-studied cancer cell lines. In particular I created MFGs for the following five cell lines: BT-549 and MCF7 (breast), HCT-116 (colon), OVCAR-5 (ovarian) and K-562 (leukemia). The initial analyses were focused on BT-549. As additional reaction constraints, I manually set the D-Lactate release to 0.005 and the O2 uptake to -10 to get more realistic solutions.

I performed FBA optimization with the COBRA toolbox in Matlab [40], speeding up the simulations with the optimizer gurobi [41]. As the primary goal the biomass production is maximized, and as a secondary goal the 1-norm of the flux vector is minimized (see equation 3.5) by setting the variable minNorm to 'one' in cobra's optimizeCbModel function. For the cancer models that I used this cuts the number of active reactions in half. This shows that there are many ways to reach the biomass production reactions, but most of them are inefficient.

3.2 Essentiality prediction of metabolic reactions

The main aim of this project is to use MFGs to find weak points in cancer metabolic networks that can be used as potential drug targets. We explore the possibility that nodes that have important positions in the network are essential for cancer growth, like nodes with high centrality in gene networks. The question is, whether there already is a centrality in MFGs that corresponds to measured essentiality or, if this is not the case, whether I can construct another network measure that can predict essentiality.

3.2.1 Essentiality from FBA KO-simulations

There is no high-throughput method to measure reaction essentiality directly. We can only inhibit the involved metabolites or proteins that catalyse the reactions and experiments for essentiality have not been done for all reactions in all conditions. Therefore, there is little biological data to compare our networks with. Gene essentiality can be mapped from essential genes to their enzymes and then to the reactions they catalyse, but this mapping is not one to one (33% of the active reactions have more than on associated gene and 28% have no associated gene), so we have to use a different essentiality measure for comparison.

Here I define the change in growth rate upon reaction inhibition as that reaction's essentiality. To compute it I simulate single reaction knock-outs (KOs) for all reactions that are active in wild type and compute the ratio between wild type and knock-out biomass production. Non-active reactions can be classified as non-essential without running the FBA.

Single KO of reaction j is simulated with an FBA by constraining the flux through that reaction to zero:

$$\text{objective: } f_{\text{KO-j}} = \max c^T v^{\text{KO-j}}, \tag{3.12}$$

$$\text{subject to } \mathbf{S}v^{\text{KO-j}} = 0, \tag{3.13}$$

$$\text{additional constraints: } v_i^{\min} \leq v_i^{\text{KO-j}} \leq v_i^{\max}; i = 1, ..., m \text{ and} \tag{3.14}$$

$$v_j^{\min} = 0 = v_j^{\max}. \tag{3.15}$$

The essentiality λ of reaction j is then computed from the growth rate ratio upon KO of reaction j:

$$0 \leq \lambda_j = 1 - \frac{f_{\text{KO-j}}}{f_{\text{WT}}} \leq 1, \tag{3.16}$$

where f of a given condition is the optimal value of the objective function, in this case the biomass production. For each simulated single knock-out I store the resulting flux distribution for network construction.

A sorted overview of growth rate changes upon knock-out of active reactions in wild type can be seen in Figure 3.3. Apart from about 150 reactions (about half of the active reactions in wild type) there is no change in growth rate between wild type and simulated knock-out. Circa 100 reactions have a lethal effect, where the biomass production reaction is zero since it cannot be reached upon KO.

I binned the reactions into four levels of lethality: no effect, mild, severe and lethal. Reactions whose removal still allowed 90% of the biomass production were included in the 'no effect' group to account for fluctuations. When half of the biomass can still be produced, the knocked-out reaction is classified as mildly lethal, below that as severely lethal:

- no effect: $0.9 < \lambda$
- mild effect: $0.5 < \lambda \leq 0.9$
- severe effect: $0 < \lambda \leq 0.5$
- lethal effect: $\lambda = 0$

Both mild severe groups are very small, with only 10-20 reactions each. The remaining hundred reactions completely disrupt the network upon removal and are classified as lethal.

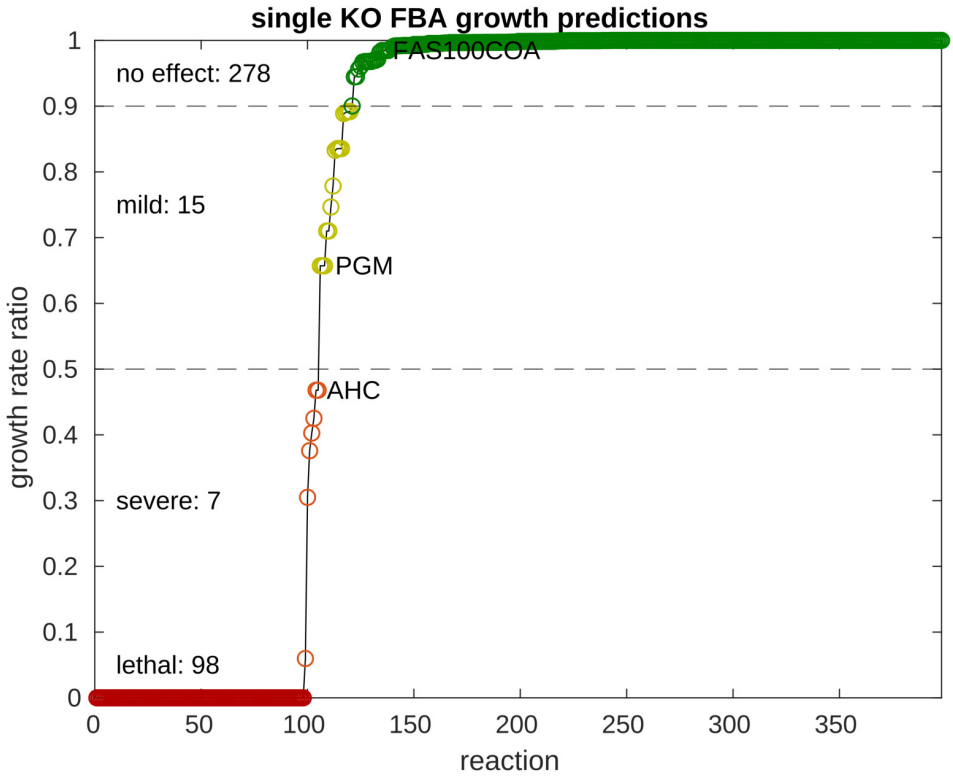

FIGURE 3.3: **BT-549 FBA results for all single reaction knock-outs**
For each reaction I perform the KO-FBA and record the growth rate ratio. Reactions were binned into four groups with different overall effect on the growth rate and therefor survival. Colors range from green for no effect to red for lethal. The three reactions marked in the plot were evaluated in more detail in Figure 3.7.

Additionally, I calculated two alternative essentiality measures for comparison. As the first alternative essentiality measure I compute the pervasiveness of reactions. This is based on the notion that if the same reaction is always needed for the optimal biomass production, regardless of what other reactions are present, it is likely that that reaction is essential for cell growth. The pervasiveness p of a reaction is the percentage of KO conditions where the FBA-predicted flux through that reaction is not zero.

$$p_j = \frac{\sum_{i=1}^{N} \left[v_j^{\text{KO-i}} \neq 0 \right]}{N}, \tag{3.17}$$

where N is the number of KO runs, which is equal to the number of active reactions. The second alternative essentiality measure is the flux through that reaction under wild type conditions v_j^{WT}.

3.2.2 Correlation between centrality measures and essentiality

Node centralities have been shown to represent important biological functions in other kinds of biological networks **[evol_PPI_centr]**. In E. coli metabolic networks it has been shown that centrality directly corresponds to essentiality [42], but this has not worked with the current network models in human cells.

My main focus is to calculate node centralities of MFGs and compare them to essentiality of reactions. Centralities were computed with Matlab's centrality function [43]. For each centrality I evaluated a weighted and an unweighted version.

In Figure 3.4 I evaluated 13 weighted or unweighted centralities of active reaction nodes in the wild type network for their correlation with the node reaction's essentiality. I compare with: 1) λ_j, the growth rate ratio upon knock-out; 2) p_j, the pervasiveness across single reaction knock-outs; and 3) v_j^{WT}, the flux through that reaction in the wild type.

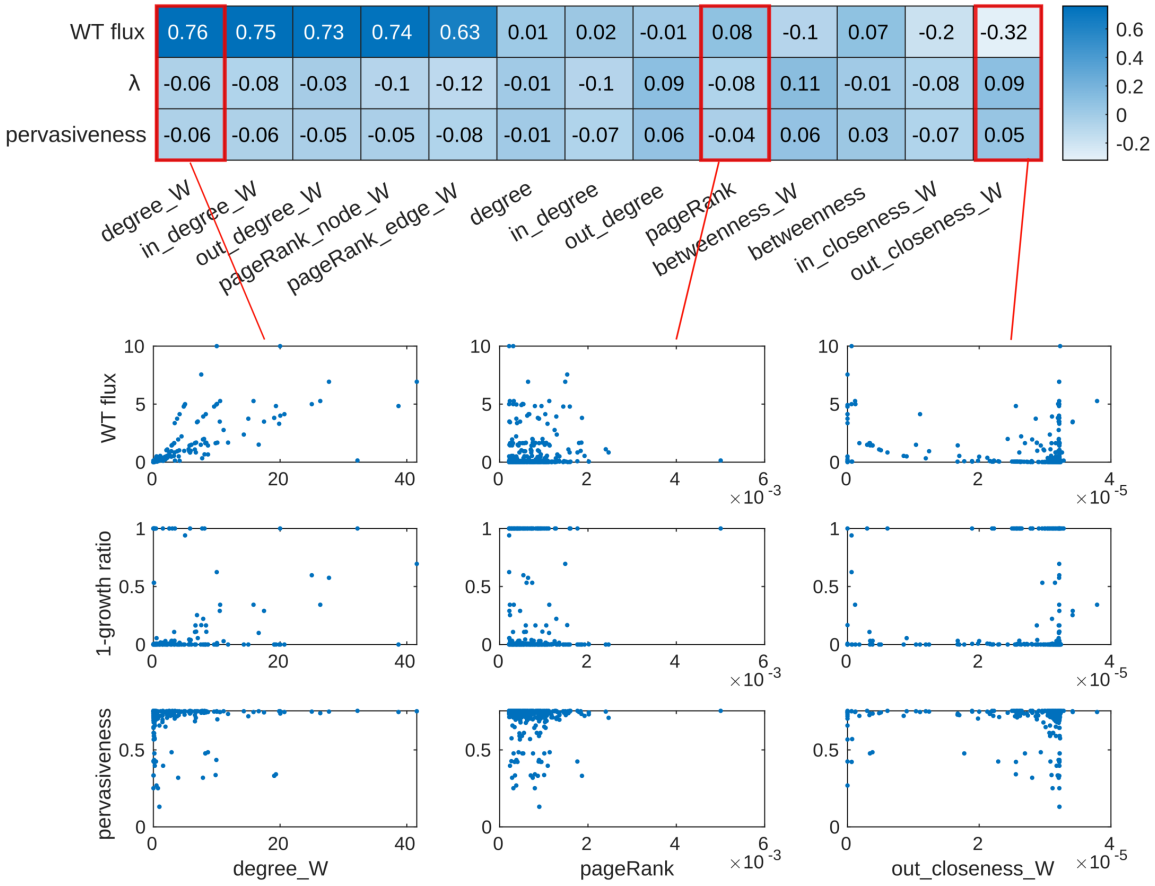

FIGURE 3.4: **Comparison of node centralities with reaction essentiality** In the table I list the Pearson correlations for all combinations of centralities and essentiality measures. The plots below are selected examples showing the full distribution of reaction's centralities versus essentiality from the cells marked in red in the table.

From Figure 3.4 we learn that the only importance measure that centralities correlate with is the reaction flux and the only centralities that are highly correlated to it are those that are computed via edge weights, which in turn depend on the fluxes. Therefor the only high correlations are an artifact. The fact that the same centralities that correlate with the flux do not correlate with the growth rate ratio or pervasiveness indicates that reaction flux is not a useful essentiality measure. And even though λ and pervasiveness have comparable correlation values for centralities, the plots in the lower half of Figure 3.4 shows that they do not capture the same properties. In the following sections I only evaluate λ, since this measure has been successfully applied in the literature.

In conclusion, I did not find any node centrality that is predictive of reaction essentiality. This is not surprising, since it did not work for any previous human

cancer networks. Still, there might be other information contained in MFGs based on centralities that can predict essentiality.

3.2.3 Changes in centrality upon KO

To investigate the usefulness of MFGs and what information is encoded in them, I simulated gene or reaction inhibitions and compared the inhibited networks to the wild type. As an initial analysis I tested if gene inhibition has any effect on the networks that cannot be detected purely by FBA analysis. I compared the node pageRank centrality of wild type networks and Succinate Dehydrogenase knock-out networks, see Figure 3.5. Succinate Dehydrogenase (SDH) is a reaction in the TCA-cycle that is often mutated in cancer and has been shown to be synthetically lethal with pyruvate carboxylase (PC), another commonly mutated reaction in cancer [8]. Upon knock-out the modelled cell line is still able to produce almost wild type levels of biomass, but has to restructure the use of reactions, as can be seen by the large number of reactions that have a flux ratio of zero in Figure 3.5 and a few outliers with very high KO-flux. Interestingly the majority of reactions hardly changes its flux (flux ratio of 1), but sees a wide range of pageRank ratios from 0.6 to 1.5. This tells us that the network structure and in particular the pageRank contains additional information that might be useful for essentiality analysis.

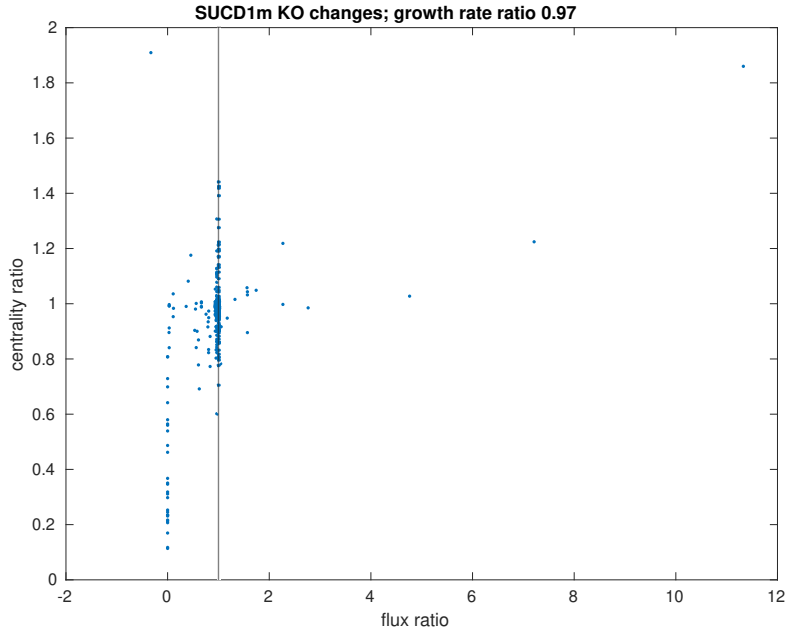

FIGURE 3.5: **Changes in flux and node centrality between SUCD1m-KO and wild type in BT-549.** The ratio of KO-flux over wild type-flux

Next, I compared the node centralities of wild type networks and SDH knock-out networks, see Figure 3.6, to investigate if simulated gene inhibition has any noticeable effect on the networks.

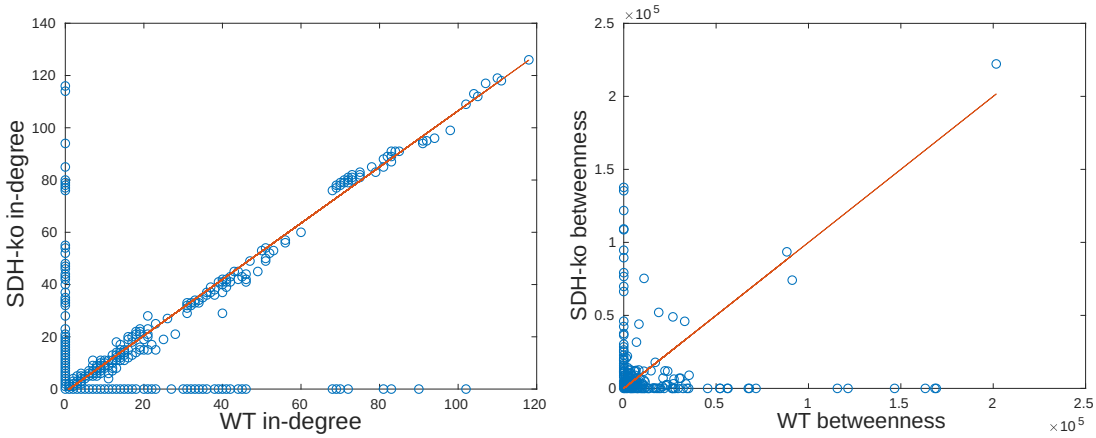

FIGURE 3.6: **Changes in node centralities between treated and wild type BT-549** In order to determine the changes between conditions and the measures representing the difference I plot node centralities from wild type and SDH knock-out against each other. The red line represents a theoretical perfect fit.

Comparison of MFGs is challenging as networks from different conditions do not have the same node sets. With their fluxes reduced to zero, these reactions cannot exchange metabolites and are therefore not connected to the rest of the network. These reactions have to be reintroduced with centrality values of zero for comparisons. Apart from those reactions, the out-degree shows a strong linear relation between untreated and treated (left plot in Figure 3.6). The betweenness on the other hand changes drastically (right plot in Figure 3.6). The majority of nodes with high betweenness are not preserved between conditions. This shows how node importance can change drastically even when only one flux bound was changed. But it also shows that not all centralities can capture that change and that global centralities have properties that could help assess the metabolic state.

Selecting a reaction from each of the no effect, mild effect and severe effect severity groups, I calculated the pageRank centrality in the corresponding knock-out networks. Plotting the pageRank centrality distribution in wild type versus knock-out condition, we see that the difference between the nodes centrality grows with severity of the effect (see Figure 3.7).

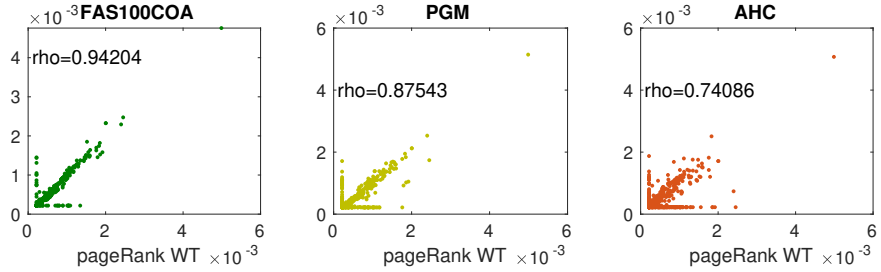

FIGURE 3.7: **Three essentiality levels** Plotting the pageRank distribution in wild type versus knock-out for three reactions taken from the three severity groups identified in Figure 3.3 shows a growing difference in centrality with increasing change in growth rate. Rho is the correlation coefficient between pageRank distributions.

3.2.4 New network measure from centrality changes

Developing the findings from the previous section, we constructed another network measure that is based on changes in node centralities. We observed that the more essential the removed reaction is the more drastically the KO-network changes. To confirm this observation globally, I constructed MFGs for all knock-out models and recorded different centralities along with their correlation to their wild type distribution. With this information we can define the new node measure σ_c from any centrality c of our choice evaluated for all nodes that represent active reactions in the wild type network (V_{active}).

$$\sigma_j^{\text{centr}} = \text{corr}(\text{centr}^{\text{WT}}(V_{\text{active}}), \text{centr}^{\text{KO-j}}(V_{\text{active}})) \tag{3.18}$$

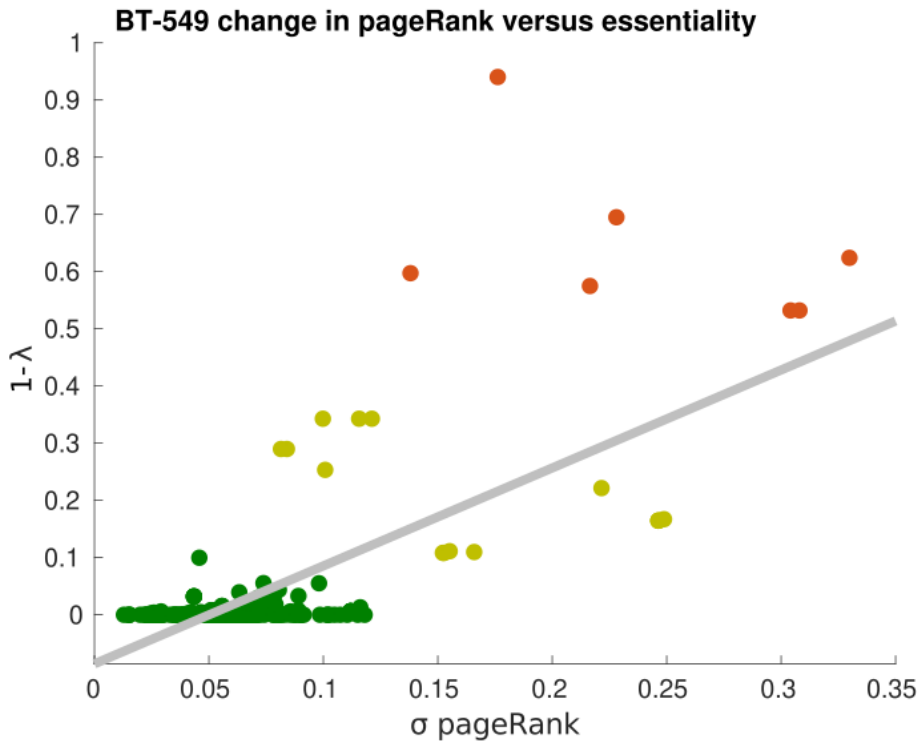

FIGURE 3.8: **PageRank correlation versus essentiality in BT-549 cancer** The new node measure σ of pageRank correlates with the essentiality measure λ with a correlation coefficient of 0.74. The colors correspond to the letahlity groups defined before. Green are non-essential, yellow mildly essential and red are lethal.

Comparing σ^{pageRank} with essentiality λ, we get a strong correlation of $\rho = 0.7419$, see Figure 3.8. With σ we have identified a new network measure that is able to capture reaction essentiality.

When I repeat the procedure in other human cancer cell lines, I get similar results, see Figure 3.9. HCT-116 is a colon cancer, k-562 lung cancer, MCF7 breast cancer and OVCAR-5 an ovarian cancer cell line.

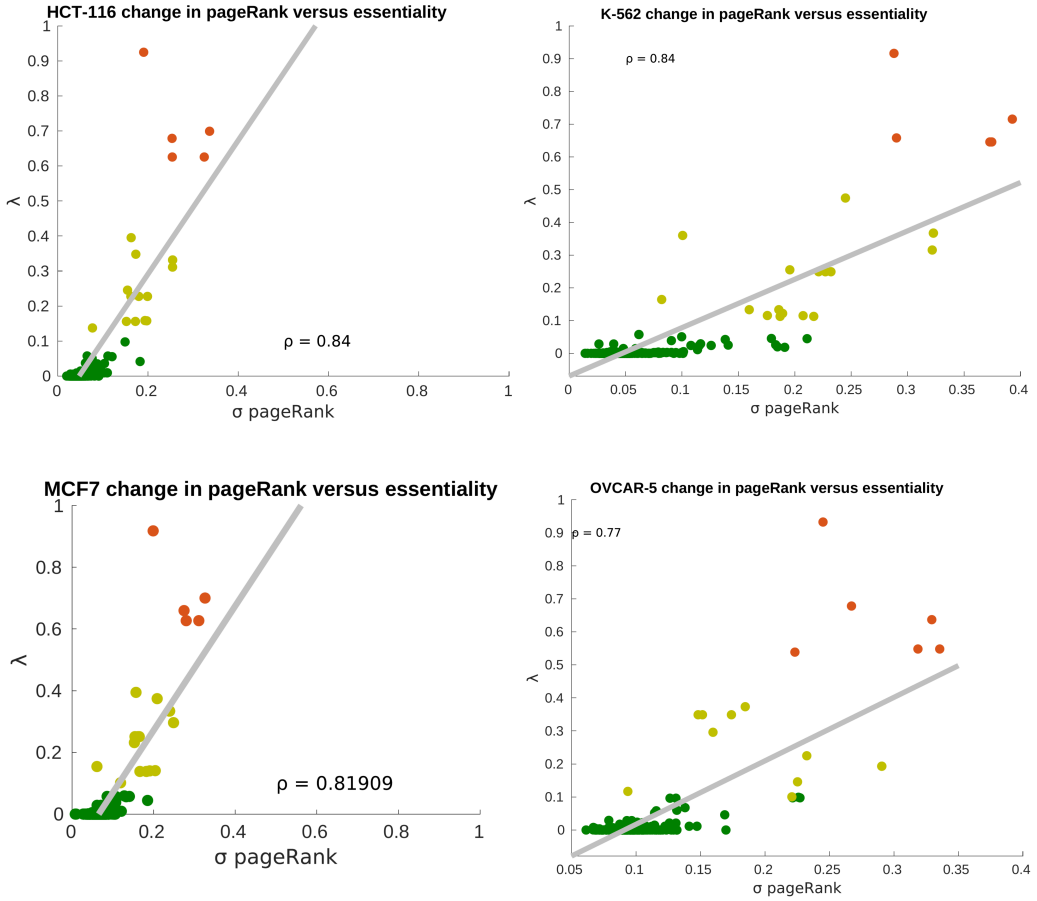

FIGURE 3.9: **PageRank correlation versus essentiality in four cancer cell lines** The new node measure σ of pageRank correlates with the essentiality measure λ in four other human cancer cell lines.

3.3 Validation via E. coli models

As mentioned before, E. coli models are well-described enough to allow for highly accurate prediction of essential reactions [38] via FBA. Here I build MFGs for E. coli and calculate the new sigma measure as described in the previous section. Thus I can evaluate not only the newly proposed essentiality measure, but MFGs in general.

3.3.1 MFGs for E. coli

For MFG construction I utilized the latest genome-scale E. coli model from 2019, iML1515 [44], and an older model from 2011, iJO1366 [45], as a replicate. Again, I

	iJO1366	iJO1366 active	iML1515	iML1515 active
reactions	2251	513	2719	459
genes	1366		1515	
metabolites	1136		1192	

TABLE 3.1: E. coli model sizes

maximize biomass production and minimize the 1-norm of the flux vector in the FBA simulations. This results in the following models 3.1.

3.3.2 E. coli predicted essentiality

Again I perform single knock-out simulations for all active reactions with FBA and record the change in growth rate. The growth change curve for both E. coli models has the same shape as for human cancer, see Figure 3.10, but more reactions are lethal. Also, the lethality of single reactions is different from their human counterparts. iJO1366 has more reactions overall than iML1515, but from inspection by eye, the shared reactions' lethality is about the same.

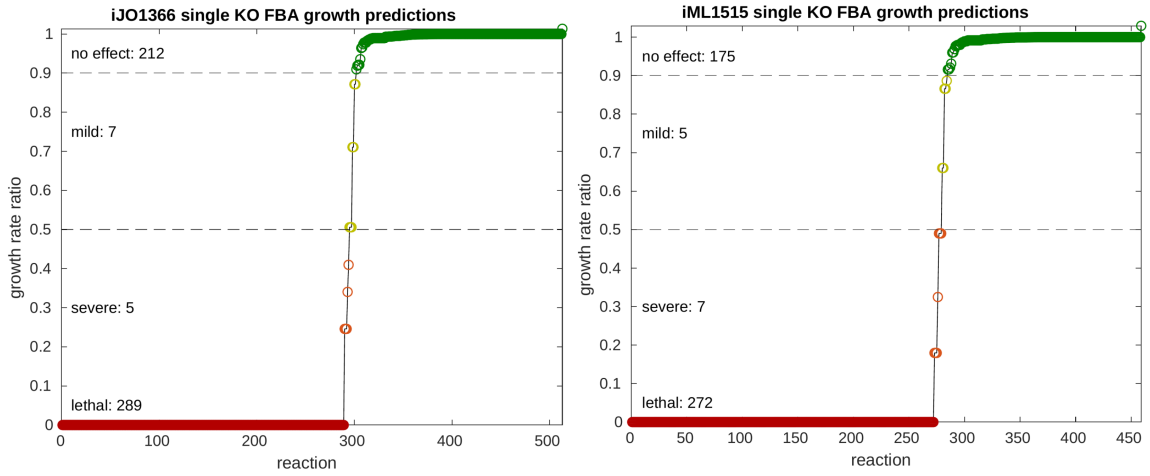

FIGURE 3.10: **FBA results for all single reaction knock-outs in E. coli** For each reaction I perform the inhibited FBA and record the growth rate ratio. The reactions were binned into four groups, compare to human cancer binning.

Like before, I compute the new sigma measure by correlating the wild type and knock-out PageRank of all active reactions. When I correlate σ and the growth rate ratio λ, I get similar correlations as with the human cancer cells, see Figure 3.11.

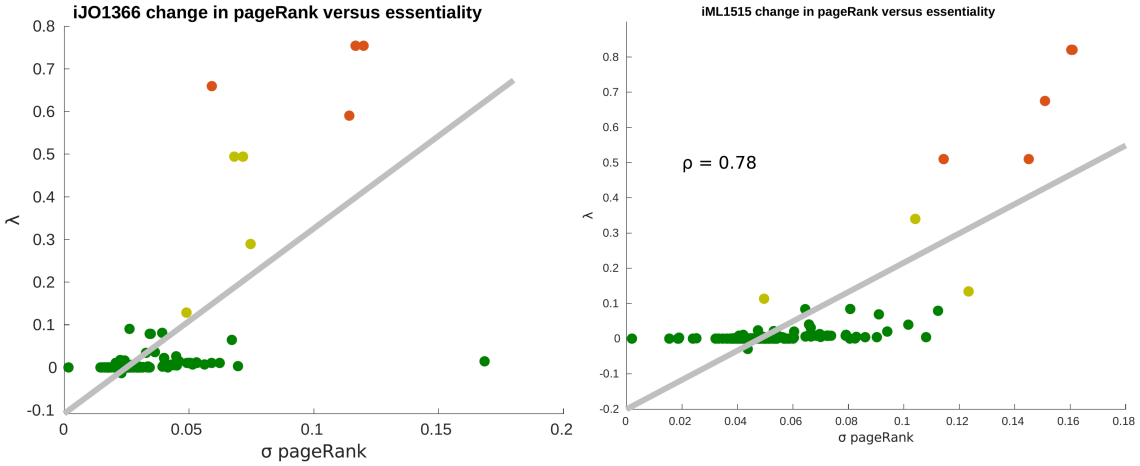

FIGURE 3.11: **PageRank correlation versus essentiality in E. coli** The new node measure σ of pageRank correlates with the essentiality measure λ.

3.4 Conclusions

In this chapter I examined the utility of MFGs for modelling cancer metabolism, especially if they could be used to predict weak spots in cancer metabolic networks that are essential for the integrity of the network and thus good targets for inhibition and treatment. Starting from a metabolic reconstruction I set the constraints for FBA such that I gained realistic fluxes and representative networks for different cancer conditions.

Doing single reaction inhibitions I managed to identify that the change in centrality between wild type and knock-out is correlated with the overall cell survival and thereby reaction essentiality. With this we showed that the relationship between essentiality and centrality does exist, albeit not as directly as hypothesized. Essential reactions do not have high centrality in MFGs, but their removal causes system-wide changes that can be measured via centrality changes.

Network clustering of gene expression time-series data

In this Chapter I evaluate time series data of gene expression. We can utilize networks to model gene expression data in a specific cancer condition. I cluster these using gene similarity networks, taking into account additional biological data, to predict genes that are essential while the cancer cells are recovering from treatment.

4.1 Proteotoxic stress recovery in cancer cells

There are many chemical treatment options in cancer, but often the cancer grows back after an initial phase success. Cancer can be viewed as a colony of disregulated cells that diversifies under evolutionary pressure. While treatment can work on the majority of cells in the colony, if any survive, they will multiply into a new colony that is more resistant to the stress. We need to observe these cells that escape the treatment and the mechanisms that enable them to do so.

One established treatment in multiple myeloma (a type of blood cancer) cells is proteasome inhibition [46]. The proteasome is a chemical complex that recycles proteins, i.e. breaking down unneeded or harmful proteins and extracting amino acids that can be used to build new proteins. Inhibiting the proteasome has proteotoxic effects, especially in cells with high proliferation rates and a large number of misfolded proteins. Therefore it has a more detrimental effect in cancer cells compared to slowly replicating cells, especially in cancers where the cell cycle checkpoints are compromised. Yet only a fraction of cancer cells die with this treatment, leaving it an insufficient method to treat patients [47].

I analysed RNAseq data of proteasome inhibitor-treated bone marrow multiple myeloma cells measured in the Auner lab. Treating with carfilzomib reduced the number of viable cells by 50% by day 2, but after day 6 the number of cells reaches

a pre-treatment amount, see the Figure 4.1. Bulk cell expression data for five replicates at seven time points (day 0, 1, 2, 4, 6, 8, 10) was measured such that short-term effects as well as long-term changes could be captured.

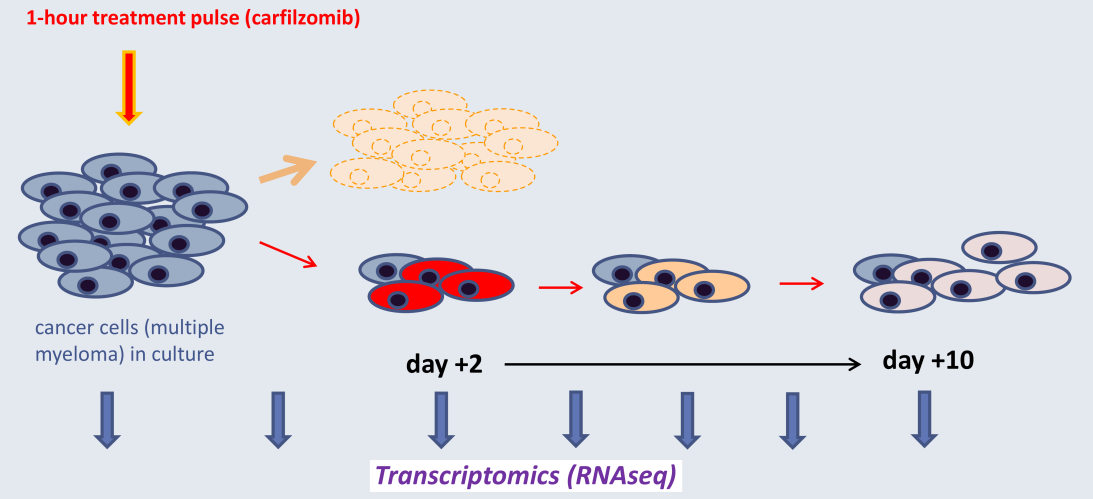

FIGURE 4.1: **Proteasome-Inhibition experimental design** The multiple myeloma cells were treated with a proteasome inhibitor on day 0. By day 2 half of the cells had died, but the remaining cells adapted to the stress and were proliferating again by day 6. RNA-seq experiments were performed on days 0, 1, 2, 4, 6, 8 and 10. The figure was taken from slides from Holger Auner.

Our main goal is to identify genes that drive the recovery process from the initial stress through adaptation to normal proliferation. These genes must have a quick reaction to the treatment and will be selectively active at different times. The quickest and most drastic changes can set in motion by genes causing other genes to change their expression, thereby also synchronizing genes that have to work together to exert their function. We can identify genes that are important during the different phases of recovery by clustering the expression profiles of genes and extracting co-expressed genes.

4.2 Processing expression time series data

4.2.1 Approaches for time series data

In genomics we would like to apply the methods developed in finance and neurology for time series data, but generating data is much more labor-intensive per time point. So usually there are not enough reliable data points for conventional

causality inference methods, even when additional data points are predicted with regression models or similar.

Instead, genes are usually clustered and their influence has to be inferred by hand, adding prior knowledge. Commonly used methods are k-means clustering and principle component analysis (PCA), but again in our case the sparsity of the data is a problem.

4.2.2 Data pre-processing

Before the raw data can be used, it has to be normalized and filtered to exclude random noise and to select only the genes that participate in the recovery process. Starting with normalized gene counts, I chose the expression cutoff as the minimum in the log expression distribution (Figure 4.2). Each gene expression is represented by the sum of normalized counts over all time points. This reduced the number of genes from about 18,000 to a little less than 12,000. For further analyses of the expressed genes I switched to DEseq2-regularized and log-transformed counts as this allows for better comparison of genes expression at different orders of magnitude. To better capture the dynamics after inhibition, I removed the expression data on day 0 (timepoint t_1) and re-normalised the remaining timepoints by subtracting the mean (logspace equivalent of dividing by mean):

$$x_t = \frac{6x_t}{\sum_{t=2}^{7} x_t}, \text{ for } t = 2, .., 7. \tag{4.1}$$

Next I introduced a variance cutoff in order to preselect genes that show a reaction to the inhibition, thereby excluding genes that could dampen the signal, and making the downstream analysis faster. The cutoff was chosen as 0.1, which reduces the number of genes to 2542, a number comparable to what other studies report [48].

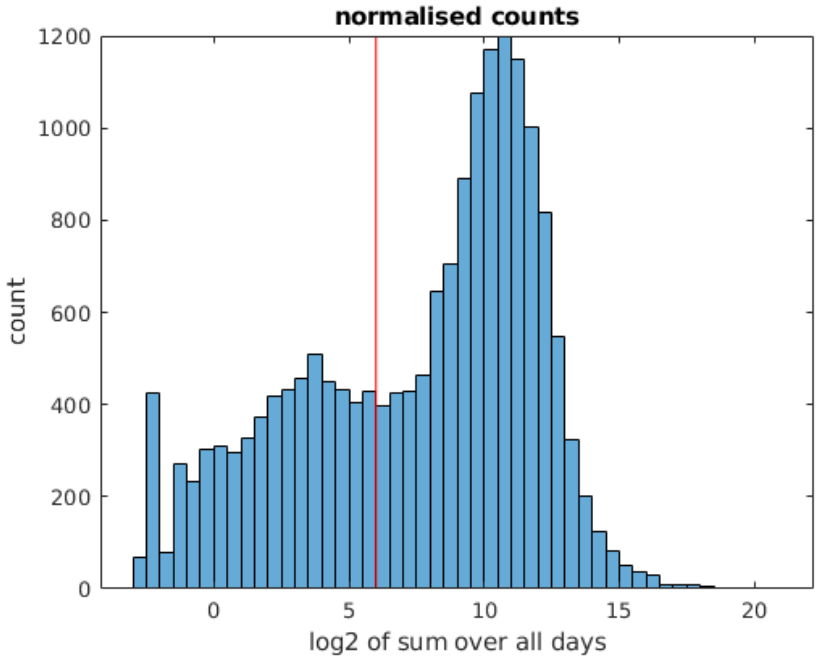

FIGURE 4.2: **Log-histogram of the expression sum over all days for all genes.** Summing over all days' normalized expression data, we get a bimodal distribution. I employed the logarithmic plot to choose the expression cutoff as the minimum between the two modes, thereby separating randomly detected genes and reliably expressed genes.

The resulting matrix with 30 measurements (6 days x 5 replicates) of expression data for 2542 genes was passed on for network construction and clustering.

4.2.3 Gene similarity via Gaussian Processes

I used Gaussian Processes (GPs) [49] to increase the number of data points and define a similarity measure from which to build a gene similarity network. Gaussian Processes have the advantage that they can utilize replicates and prior knowledge, which makes them a good modelling choice for our data. A Gaussian Process is a stochastic process, where every finite linear combination of the random variables has a Gaussian distribution. They are completely defined by their mean and covariance function, which determines how much the process can oscillate between given values.

The regression functions are sampled from:

$$y \sim \mathcal{N}(m(X), K(X,X) + \sigma_n^2 I), \tag{4.2}$$

where m is the mean and K the covariance function.

For computations in matlab I used the GPML Toolbox version 4.2 [50] with the following functions and parameters for exact inference. The mean is $\mu = 0$ and for the covariance function is the isotropic squared exponential covariance function (SEiso):

$$K_{\text{SEiso}}(x, x') = \sigma_f^2 \exp(-\frac{1}{2l^2}(x - x')^T(x - x')). \tag{4.3}$$

Gaussian Processes can be used as a prior probability distribution over functions when performing non-parametric regression. That way we can predict values between measurement and beyond the measured range for each gene. But beyond that GP-regression can be used to define a similarity measure for genes by comparing pairwise regression. The hyperparameters $\theta = (\sigma_f^2, l, \sigma_n^2)$ in the covariance function (σ_f^2 and l) and likelihood function (noise level σ_n^2) were optimized over the entire dataset by maximizing the following likelihood function:

$$\mathcal{L}(y|(f, \sigma_n^2)) = \frac{1}{\sqrt{(2\pi\sigma_n^2)}} \exp(\frac{-(y_i - f_i)^2}{2\sigma_n^2}). \tag{4.4}$$

Adopting the strategy from [51], I define the gene similarity score s as the difference in negative log likelihood between regression of two genes separately and regression of both gene's data points combined 4.5.

$$s_{i,j} = \log p([y_i, y_j]|[X; X], \theta) - \log p(y_i|X, \theta) \log p(y_j|X, \theta) \tag{4.5}$$

The idea behind this is that prediction strength will increase when more data points of the same function are added to one model, instead of trying to model the same function twice with fewer data points. But it will decrease, when the two very different time series are forced into the same regression.

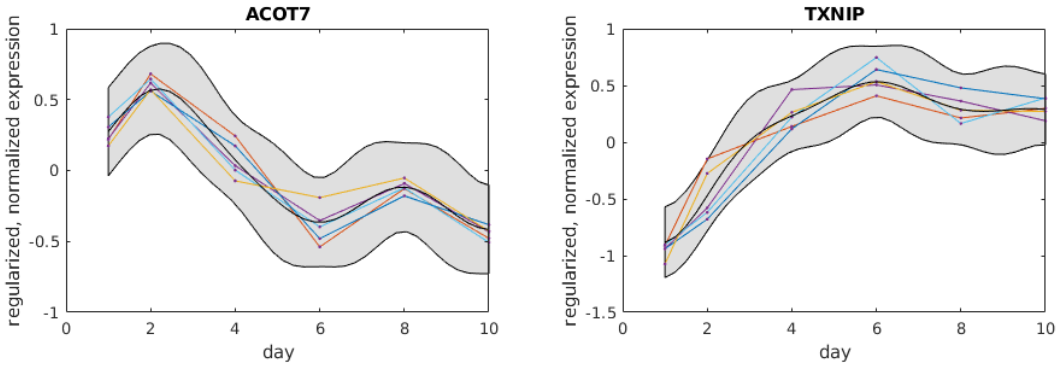

FIGURE 4.3: **Examples of Gaussian Process regression on single gene expression.** The colored lines are the time courses for different replicates, the black line in the middle is the posterior mean and the grey area around it the 95% confidence interval.

4.3 Clustering a gene similarity network

4.3.1 Gene network construction

The similarity matrix contains positive and negative entries, which poses a problem for graph clustering algorithms, which usually expect edge weights to represent distances between nodes. We can easily transform the similarity matrix into a dissimilarity or distance matrix by subtracting each matrix entry from the maximum entry in the matrix:

$$M_{\text{distance}}(i,j) = 10 + max(M_{\text{similarity}}) - M_{\text{similarity}}(i,j). \tag{4.6}$$

Afterwards the diagonal is set to zero again.

$$M_{\text{similarity}}(i,i) = 0, \quad \text{for } i = 1,..,2542. \tag{4.7}$$

From this distance matrix I construct a network by introducing edges for the k-nearest neighbors of each gene with $k = 7$. With this neighborhood the graph is connected enough to link similar genes while being sparse enough to establish subgroups of genes. To make sure that the resulting network has only connected component for clustering, I add the minimum spanning tree of the fully connected graph, resulting in Figure 4.4.

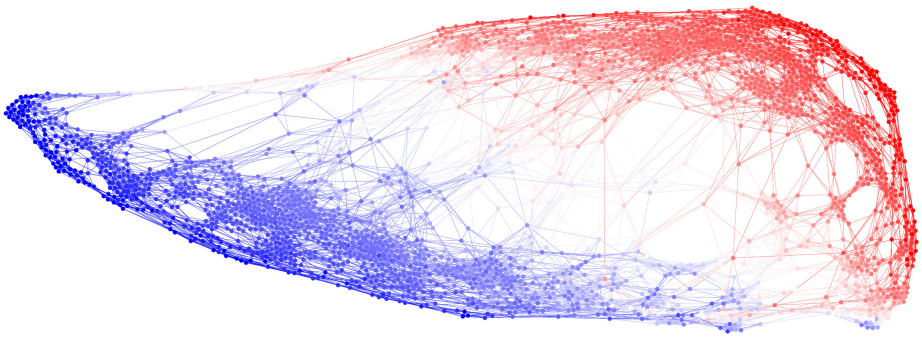

FIGURE 4.4: **Gene-similarity graph** Shown are 2542 expressed genes with high variance upon PI-treatment are connected by edges, when their expression profiles are similar. This network is colored by the genes' expression on day 1 relative to day 0. Red: expression went up, blue: expression went down relative to day 0. The figure was created by Dr. Zijing Liu.

4.3.2 Markov stability clustering

For clustering of the similarity graph into genes with similar expression profiles I chose the Markov Stability (MS) algorithm [30]. It performs several random walks on the network at different Markov times and is thereby able to generate modules of variable sizes, from which I can select the biologically most relevant.

In an unweighted network the transition probability of the random walker from one node to the next can be defined as:

$$p_{ij} = \begin{cases} \frac{1}{\text{degree}(i)}, & \text{if } A_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.8}$$

Since these transition probabilities are independent of the time of visiting, we can define the random walk as a Markov chain with state space V and transition matrix:

$$P = (p_{ij})_{i,j \in V}. \tag{4.9}$$

From P we can compute the stationary distribution π

$$\pi = \pi P \text{ with} \tag{4.10}$$

$$\pi(i) > 0 \forall i \in V \text{ and} \tag{4.11}$$

$$\sum_{i \in V} \pi(i) = 1, \tag{4.12}$$

which lets us determine the long-term behavior of the walker. Clusters are defined as regions in the graph where the random walker is trapped in for a long time. This is evaluated through its position after certain number of steps from random starting point. For longer times the random walker can end up in more remote areas of the graph, thus the clusters become bigger.

Given the adjacency matrix A I extract the number of edges $m = \sum_{i,j} \frac{A_{ij}}{2}$ and its degree vector d ($d_i = \sum_j^n A_{ij}$). From the degree matrix D, where $D_{ii} = d_i$, I construct the one step transition matrix $M = D^{-1}A$ and the stationary distribution matrix $\pi = \frac{D}{2m}$, which is based on the unique stationary distribution $\pi = \frac{d^T}{2m}$.

The optimal clustering for each Markov time is chosen by maximizing the so-called stability function over several runs [30]. The stability function is evaluated at a certain Markov time t for a proposed clustering C, which constitutes a complete graph partitioning into c clusters:

$$r(t, C) = \sum_{s=1}^{c} \sum_{i,j \in C_s} B(t)_{ij}, \tag{4.13}$$

where

$$B(t) = \Pi[(1-t)I + tM] - \pi^T \pi. \tag{4.14}$$

The choice of the best clustering from all evaluated Markov times depends on several factors selected by the user. I take into account three mathematical properties of the clusterings and one biological. Mathematically a clustering is good, if it is preserved across multiple Markov times and different initializations. This property is satisfied by plateaus where the number of predicted clusters (blue line in Figure 4.5 A and B) is constant and the Variation of Information (VI) [52] between clusterings C^t at time points t_1 and t_2 $VI(C^{t_1}, C^{t_2})$ is low. This VI between clusterings is computed across all pairwise Markov times best clusterings (yellow to brown areas in the background of Figure 4.5 A). Additionally the VI among runs at the same Markov time $VI(t)$ (green line) is evaluated as follows:

$$VI(t) = \frac{1}{n(n-1)} \sum_{s=1}^{n} \sum_{s'=1}^{n} VI(C_s(t), C_{s'}(t)), \tag{4.15}$$

where n is the number of clustering runs per Markov time and C_s the clustering obtained from run s.

From a plateau of constant number of clusters and low pairwise VI across many Markov times I select the point where the $VI(t)$ is low. To further narrow down

possible good clustering choices and include biological information in the clustering, I compute the biological homogeneity index (BHI) of gene clusters for all Markov times (red points in Figure 4.5 B). The BHI is a clustering score for how functionally similar the genes in each cluster are [53], obtained from GO-term enrichment analysis on each of the clusterings. Given a clustering C and biological functional groups G (obtained through GO-enrichment), where $G(x)$ is the group that a gene x belongs, the BHI is computed as follows:

$$BHI = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x\neq y\in C_j} I(G(x)=G(y)), \tag{4.16}$$

where k is the number of clusters in the current clustering C, n_j is the number of annotated genes in cluster C_j and the indicator function I is one if x and y are in the same functional group [53]. BHIs were calculated using the R package clusterprofiler [54]. Combining all this information I chose a BHI-selected optimal clustering with six clusters, see Figure 4.6.

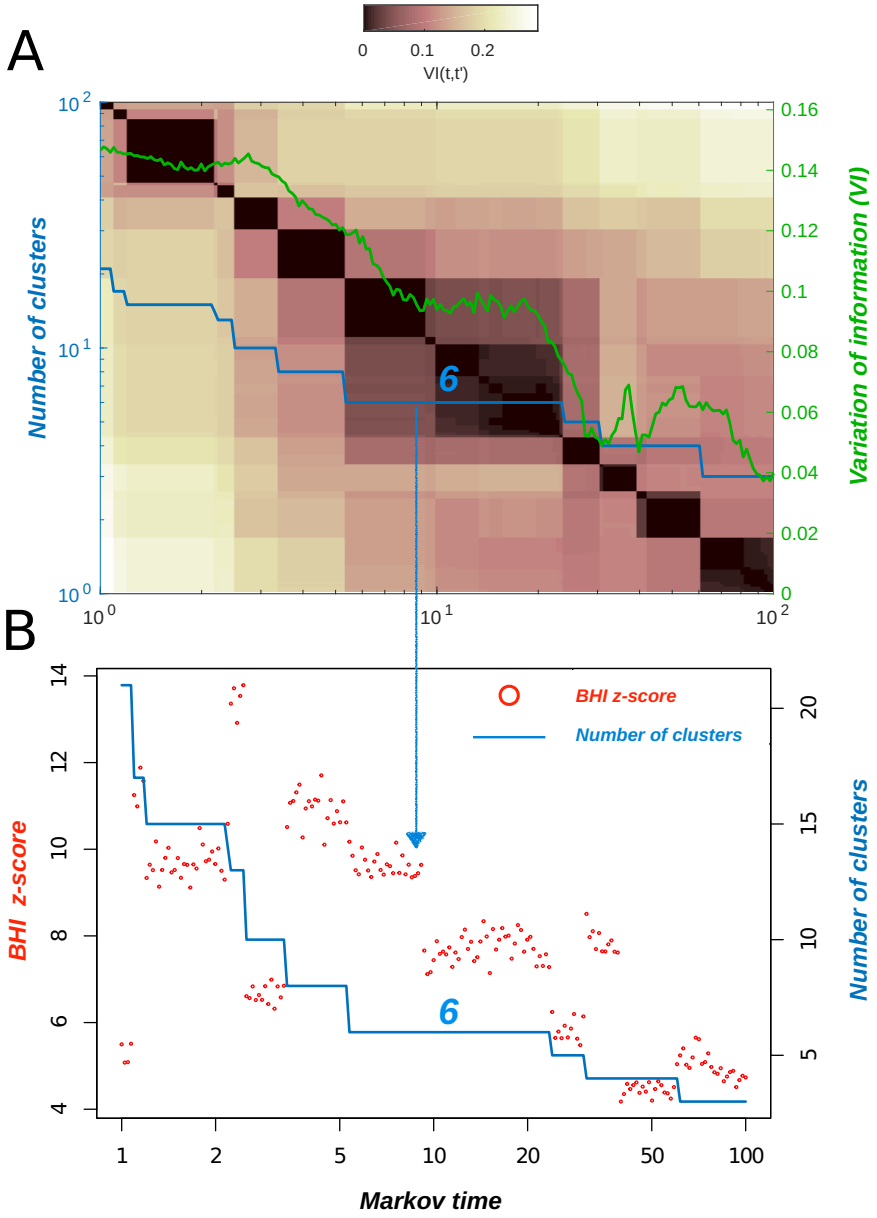

FIGURE 4.5: **Stability clustering of gene similarity network - choosing the best clustering.** A: The gene expression distance graph was clustered by performing random walks at Markov times from 1 to 100, plotting the resulting number of clusters and the pairwise Variation of information between clusterings. B: For the same clusterings the biological homogeneity index (BHI) was computed. From both these plots the best clustering was chosen from the time, where the number of clusters is stable, the Variation of information low and the BHI high. This figure was created by Dr. Zijing Liu.

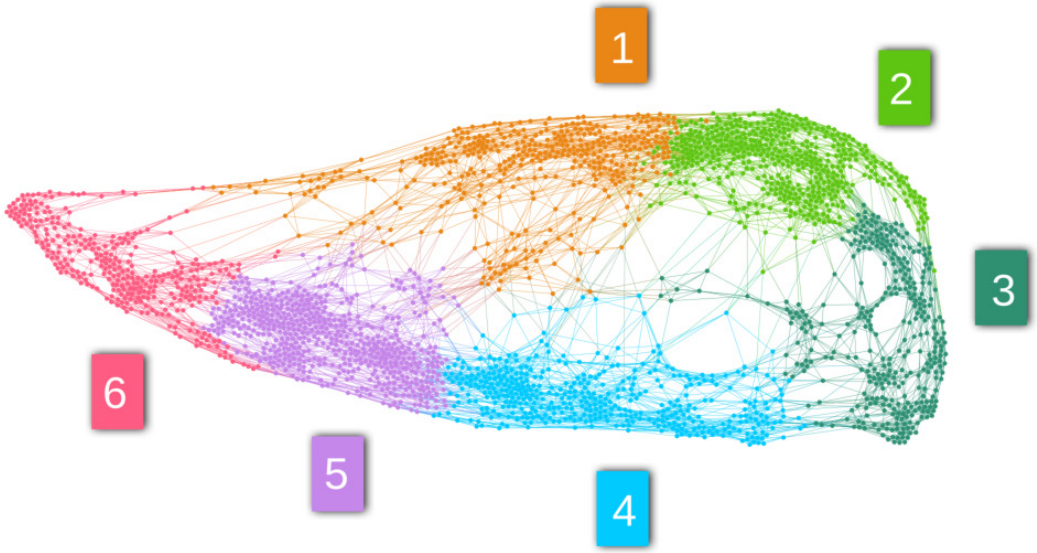

FIGURE 4.6: **Clustered gene expression network** This distance graph was obtained from gene expression similarity based on GP-regression. The Markov-stability clusters (marked by color and number) were ordered by their position in the graph and enriched gene GO-term functions. This figure was taken from [23]

4.4 Biological interpretation of gene clusters

To get the average cluster expression I again run a GP-regression on the few hundred genes per cluster. The representative expression profiles can be seen in Figure 4.7. Using the shape of the cluster profiles, the position of the cluster in the similarity network (see figure 4.6) and information from enrichment analysis, we order the clusters into two groups with three corresponding subgroups each, see Figure 4.7. The expression initially either goes sharply up or down; and in the long run it either returns to its pre-treatment expression, keeps the trend of its initial stress reaction or overshoots beyond its original expression. The time point of the initial peak is slightly shifted between the three trends, from around half a day for clusters 1 and 4, to shortly before day one for clusters 2 and 5 to between day one and two for clusters 3 and 6. This hints at different functions for the genes in clusters with different trends, which we can extract with gene enrichment analysis. It is interesting to note that most changes happen around day one, and that after day six there is no change in expression for these representative profiles. This fits with the observation during the experiments that cells started proliferating by day six.

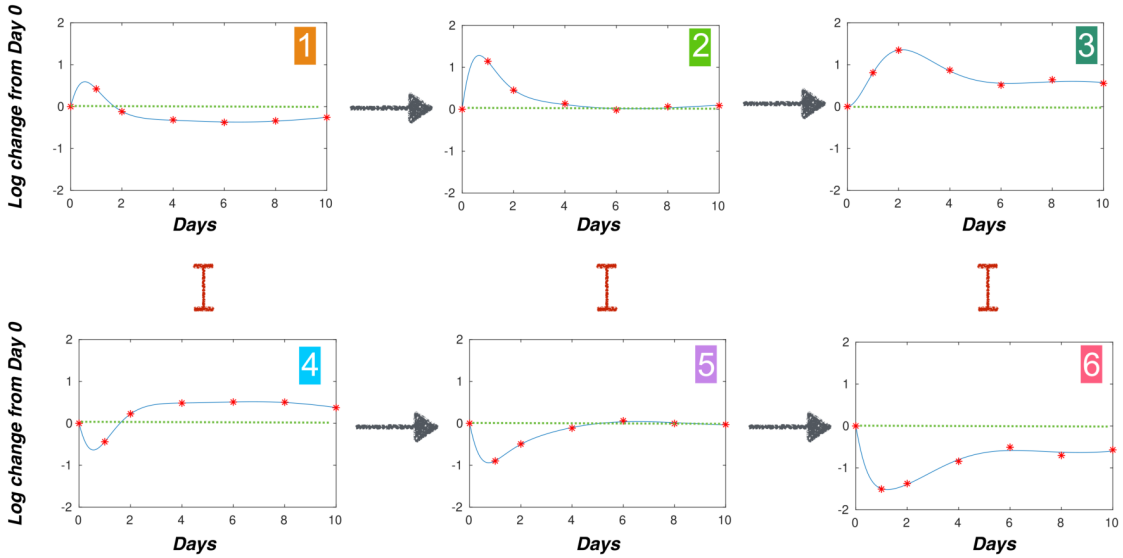

FIGURE 4.7: **Expression profiles of six cluster representatives.** These representative expression profiles were obtained through GP-regression on all genes for each cluster. The six clusters were grouped depending on their initial reaction and long-term behavior. This figure was created by Dr. Zijing Liu and Prof. Mauricio Barahona.

4.4.1 Enrichment analysis

On the six clusters I performed GO-term and KEGG-pathway enrichment analysis, which compares the prevalence of association with terms or pathways for all genes inside a cluster to their prevalence in a background gene set containing all human genes. Again enrichment scores were computed with clusterprofiler with the default 0.05 for GO-term enrichment and a p-value cut-off of 0.5 for KEGG pathway enrichment to allow for hits in every cluster, resulting in Figures 4.8 and 4.9.

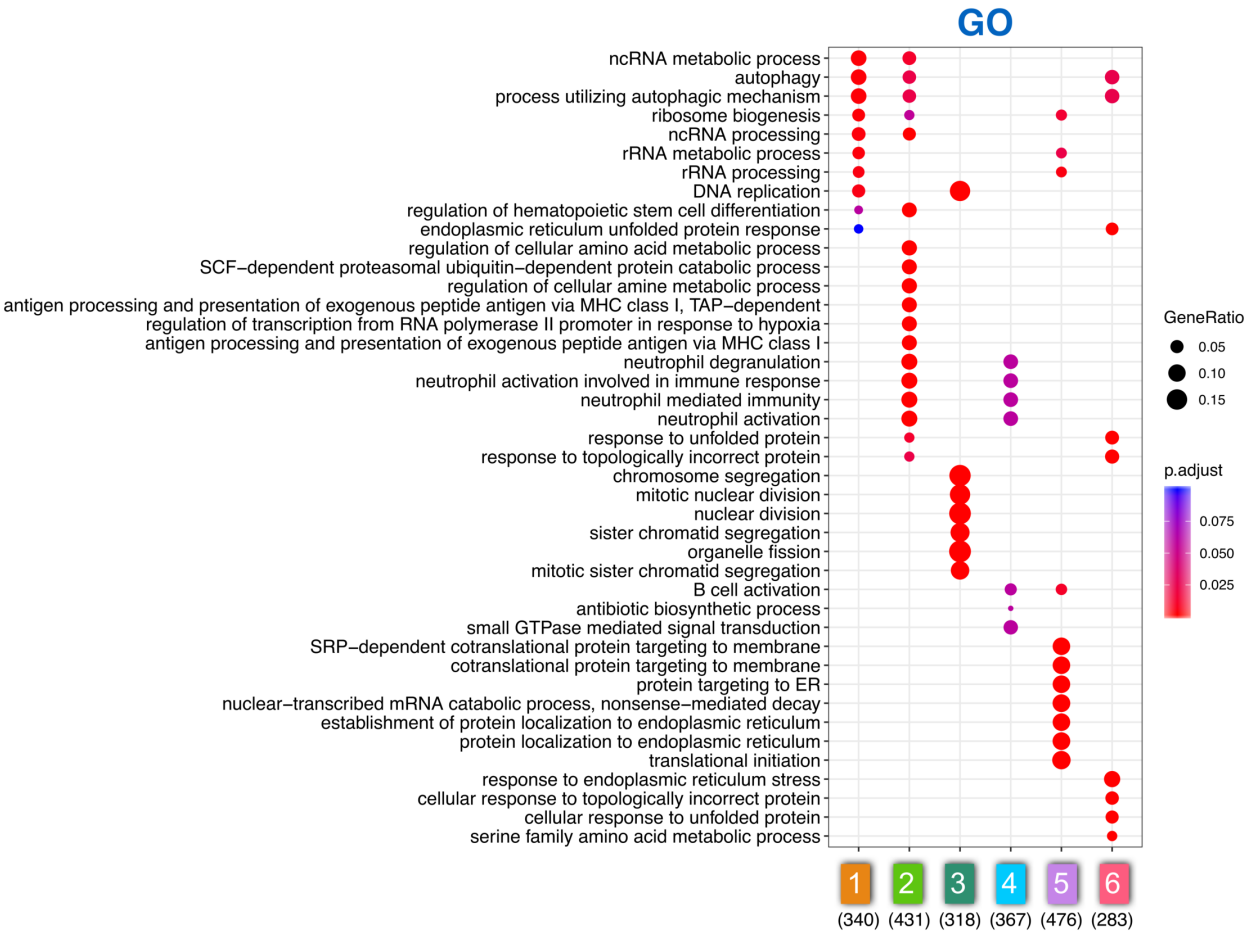

FIGURE 4.8: **GO-term enrichment analysis** These are the GO-terms that are enriched in the six gene expression clusters using clusterprofiler. This figure was adapted from Dr. Zijing Liu.

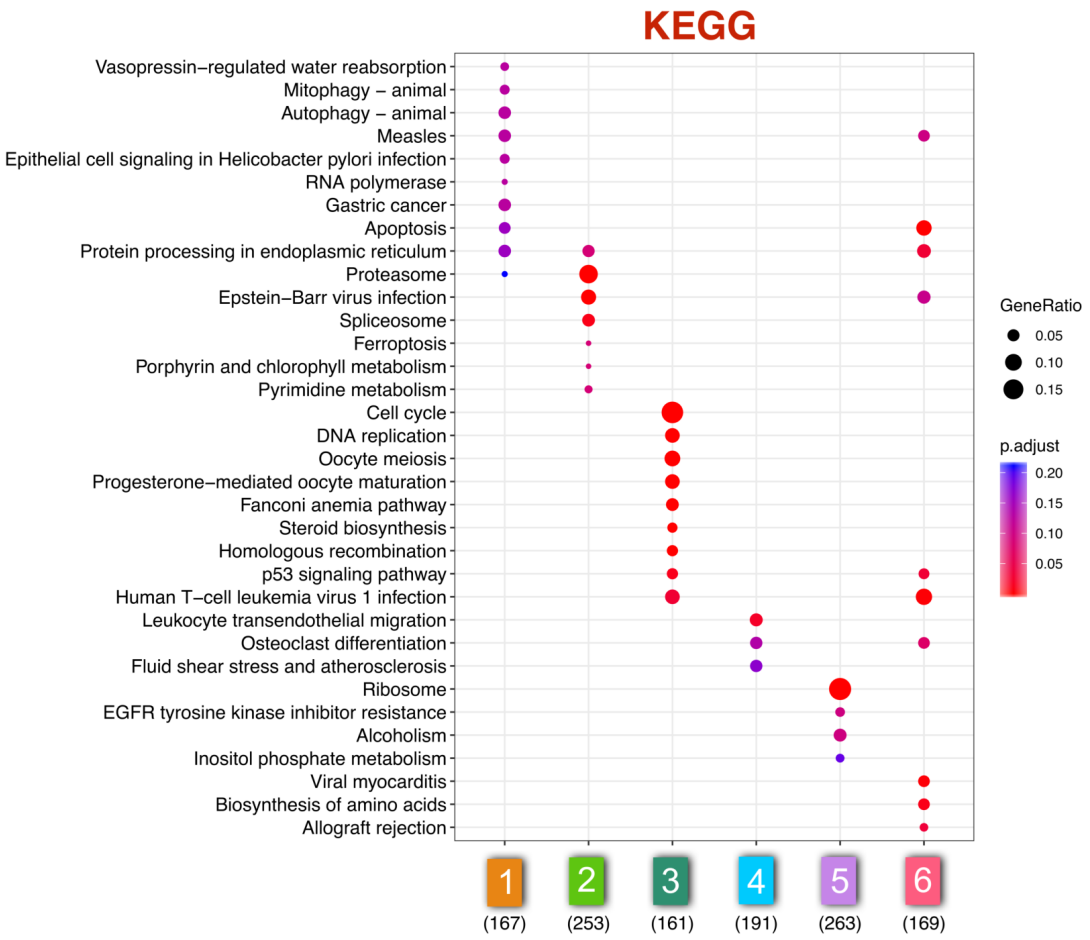

FIGURE 4.9: **KEGG-pathway enrichment analysis** These are the KEGG pathways that are enriched in the six gene expression clusters using clusterprofiler. This figure was adapted from Dr. Zijing Liu.

In accordance with the cells having to deal with proteotoxic stress I found (nc)RNA polymerase activity, autophagy and ribosome biogenesis as enriched GO-terms for clusters 1 and 2. These functions are also found enriched as pathways in cluster 1, but not in cluster 2, which has more pathway activity in ER protein processing, proteasome activity and enrichment in several amino acid metabolic processes. Other enriched terms in cluster 2 are responses to unfolded or misfolded proteins and immune response. Cluster 3 has strongly enriched pathways and terms for cell cycle processes, including nuclear division and chromosome segregation. This indicates that surviving cells are re-initiating proliferation. In cluster 4 (counterpart to cluster 1) there are only few and weak associations, mostly to immune responses. Parallel to cluster 2, there are several enriched ribosome terms in cluster 5, including ribosome biogenesis, protein targeting and localization to the ER as well as translational initiation. But the enriched terms in cluster 6, amino acid processes, response to misfolded or unfolded proteins, ER stress, autophagy and

apoptosis pathway enrichment show the ongoing difficulties when switching back to proliferation.

4.4.2 Discovered Vulnerability

Several of the genes in cluster 6 were found to encode for tRNAs that guide the activity of GCN2, a known kinase that is involved in stress recovery. These results were returned to the experimenters who did several follow-up analyses of the gene expression as well as the other 'omics' datasets. Ultimately, through follow-up experiments inhibiting GCN2, our collaborators were able to identify GCN2 as a promising drug target in various cancers that have to undergo proteotoxic stress recovery or that have similar characteristics to recovering cells [23].

4.5 Conclusions

In this chapter I analysed how networks can be used to cluster gene expression from stress recovery data in order to find essential genes that lead the process. We developed a pre-processing pipeline for gene expression time series data and constructed a gene-similarity network based on Gaussian Process regression. Clustering with Markov Stability revealed six representative groups of genes that are involved with different stages of proteotoxic stress recovery. Enrichment analysis of the genes in those six clusters showed the activity and difficulty of cells dealing with proteotoxic stress and returning to proliferation. This pipeline proved to be useful for the generation of hypotheses of genes that can be targeted during recovery to prevent cancer growth.

Discussion and Outlook

5.1 Summary

In this thesis I presented different methods of network modelling for various biological systems and their application in cancer research. In particular I focused on data modelling of gene expression and metabolic modelling to extract essential genes and reactions that can be used as drug targets.

In microbes like E.coli, FBA on metabolic reconstructions is able to predict essential reactions through flux. But in mammalian cells additional information, usually in the form of genomic data and chemical properties has to be utilized to make predictions. MFG provide a condition-specific way of modelling metabolism as reaction-based networks, that contains mathematical information about cell biology. Main results from Chapter 3:

- None of the node centrality that I tested correlates with reaction essentiality predicted with FBA,
- Centralities and therefore network structure contain additional information that is not captured with the flux,
- Node centralities can change drastically even when only few flux bound are changed,
- A new node measure, called sigma_centrality, that can predict essentiality from inhibited networks and
- These results show that MFGs and therefore reaction-based metabolic networks contain relevant biological information and present a valuable alternative to metabolite-based networks.

With a newly developed data processing pipeline I successfully clustered gene expression time-series data, leading to the identification of GCN2-inhibition as a treatment method across a multitude of cancers. Main results from Chapter 4:

- I developed a pipeline to filter gene expression data from sparse time-series data.

- Using Gaussian processes to build a similarity network and clustering that network with Markov Stability provided a clustering of essential genes during recovery.

5.2 Limitations and strengths

The networks can only be as good as the data provided to build the network. For E. coli the metabolic reconstructions and FBA results are reliable, but human cancer networks like in Chapter 3 are still too big and disregulated to make good predictions. It is not obvious that the FBA objectives of maximizing biomass production and minimizing overall flux is realistic in cancer. Validation of the FBA and network analysis results is still difficult for reactions without known inhibitors or directly mappable genes. Still, as long as this method creates even a few successful predictions, it will have proven its worth. And its specificity and speed make it a valuable prediction method.

While the metabolic networks show the interactions and mechanisms that can make a reaction essential, the gene similarity networks in Chapter 4 do not provide such information. They still require evaluation by hand and biological knowledge to extract causal genes. On top of that, more measurements are needed throughout day 1 and 2 to resolve the regulation patterns in detail. The predicted clusters have been evaluated biologically through enrichment analysis. However, comparison with other clustering methods is necessary to ensure that we obtained the optimal results.

5.3 Future Work

5.3.1 Similarity networks

There are many small decisions that were made when constructing the networks. We should test how similar the constructed networks are when slightly altering the filters and normalizing methods. How robust are the clusters when the network is constructed in a different way? I am most interested in changing the kNN

construction method to using an ϵ-environment. Maybe we could include the clustering into the network construction by gradually increasing ϵ and treating the disconnected subgroups that appear for small ϵ as clusters. This corresponds to hierarchical clustering that can be evaluated at different cutoffs by computing the BHI again. This could even be run as fuzzy clustering, such that we are less reliant on the variance filter and do not force genes that are different from all other genes into clusters where they do not fit. In a complete clustering we are bound to include outliers that interfere with the cluster profile. An interesting method to compare to is WGCNA [14], which was developed for biological analyses. WGCNA uses another similarity measure, network construction and clustering algorithms, which should be compared step by step, including combinations of WGCNA methods and our methods.

5.3.2 Reaction essentiality

The next step in the evaluation of the new σ measure is to investigate reactions where σ and λ do not correspond well. This requires an in-depth literature search for each alternatively classified reaction.

Biological validation

Biomass production prediction from FBA is just a preliminary score to evaluate essentiality, as it is also derived from mathematical simulation. Ultimately we have to compare to biological data to get more realistic values to compare with. CRISPR gene effect data is available from the Achilles project [3] and the Sanger Institute [5] to get essentiality for single genes. These then have to be mapped to the reactions they catalyse, taking into account the possibility that multiple genes can dependently or independently catalyse the same reaction and that the same gene can be involved with several different reactions.

Ultimately, predictions have to be tested in lab with CRISPR or small molecule inhibition combined with mass spectrometry.

Node roles

Ultimately we want to be able to quickly do knock-out simulations on the MFGs without having to run an FBA for each condition. Ideally, we want a network

property that immediately predicts essentiality directly from MFG node measures. One way of extracting such a measure is via machine learning (ML) algorithms. With ML we can train a model that predicts essentiality from centralities or other features and reduce the input data to the most predictive features.

A bachelor's student I supervised was able to predict essential reactions from a wild type MFG with about 90% accuracy [55]. She used two different approaches to predict *node roles* and then use these roles as input vectors for ML. The first approach uses a node feature matrix containing averaged or maximized centralities of the node and its neighboring nodes that are extracted with the algorithm REFEX [26]. The second approach utilizes a vector of incoming and outgoing path lengths, as proposed in the paper [25]. Node essentiality was again defined from FBA knock-outs, so like for the new σ-measure we have to compare with biological data to evaluate the method. Depending on these results I'd abandon the σ-measure and continue predicting essential reactions purely via machine learning.

Synthetic lethality

The ultimate goal is to predict synthetic lethality from MFGs. For this we want to utilize a combination of a the single-node essentiality measures plus a network relationship between two nodes. This add additional 'pair'-data could for example be their distance and a relationship like ancestor/parallel/unrelated which can be extracted from paths through the network. CRISPR data for combinations of genes is rare, but does exist for several cancer and E.coli cell lines. We could again train a ML model to predict essentiality for combinations of genes.

Bibliography

[1] Peter Nygren. "What is cancer chemotherapy?" In: *Acta Oncologica* 40.2-3 (2001), pp. 166–174.

[2] International Cancer Genome Consortium et al. "International network of cancer genome projects." In: *Nature* 464.7291 (2010), pp. 993–8.

[3] A Tsherniak et al. "Defining a cancer Dependency Map". In: *Cell* July 27 (2017).

[4] Ralph J. DeBerardinis and Navdeep S. Chandel. "Fundamentals of cancer metabolism". In: *Science Advances* 2.5 (2016).

[5] Sanger Institute. "Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens". In: *Nature* 568 (2019), pp. 511–516.

[6] Abdul-Hamid M. Emwas. "The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research". In: *Methods Mol Biol.* 1277 (2015), pp. 161–93.

[7] Akito Nakamura et al. "Inhibition of GCN2 sensitizes ASNS-low cancer cells to asparaginase by disrupting the amino acid response". In: 115.33 (2018), E7776–E7785.

[8] S Cardaci et al. "Pyruvate carboxylation enables growth of SDH-deficient cells by supporting aspartate biosynthesis". In: *Nature Cell Biology* 17.10 (2015).

[9] Ori Folger et al. "Predicting selective drug targets in cancer through metabolic networks". In: *Molecular Systems Biology* 7.501 (2011).

[10] Thompson N, Adams DJ, and Ranzani M. "Synthetic lethality: emerging targets and opportunities in melanoma". In: *Pigment Cell Melanoma Res.* 30(2) (2017), pp. 183–193.

[11] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[12] David Alvarez-Ponce et al. "Gene similarity networks provide tools for understanding eukaryote origins and evolution". In: *PNAS* 110.17 (2013), E1594–E1603.

[13] Nahid Safari-Alighiarloo et al. "Protein-protein interaction networks (PPI) and complex diseases". In: *Gastroenterol Hepatol Bed Bench.* 7.1 (2014), pp. 17–31.

[14] Peter Langfelder and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis." In: *BMC Bioinformatics* 9.559 (2008).

[15] Deniz Seçilmiş et al. “Uncovering cancer gene regulation by accurate regulatory network inference from uninformative data”. In: *npj Syst Biol Appl* 6.37 (2020).

[16] Zhao B et al. “dentification of Potential Key Genes and Pathways in Early-Onset Colorectal Cancer Through Bioinformatics Analysis”. In: *Cancer Control* 26.1 (2019).

[17] da Veiga Moreira J et al. “Metabolic therapies inhibit tumor growth in vivo and in silico”. In: *Sci Rep* 9.1 (2019).

[18] M. W. Covertand J N. Xiaoand J T. J. Chen and J. R. Karr. “Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli”. In: *Bioinformatics* 24.2044 (2008).

[19] Gholamreza Bidkhori et al. “Metabolic Network-Based Identification and Prioritization of Anticancer Targets Based on Expression Data in Hepatocellular Carcinoma”. In: *Frontiers in Physiology* 9.916 (2018).

[20] M Beguerisse-Díaz et al. “Flux-dependent graphs for metabolic networks”. In: *npj Systems Biology and Applications* 4.32 (2018).

[21] Francis Crick. “Central Dogma of Molecular Biology”. In: *Nature* 227 (1970), pp. 561–563.

[22] The Gene Ontology Consortium: Michael Ashburner et al. “Gene ontology: tool for the unification of biology”. In: *Nature Genetics* 25.1 (2000), pp. 25–29.

[23] Paula Saavedra-Garcia et al. “Global profiling of cancer cell recovery from therapy-induced stress reveals druggable vulnerabilities”. In: *Proceedings of the National Academy of Sciences* 118.17 (2021).

[24] Shuo Chen et al. “Estimating large covariance matrix with network topology for high-dimensional biomedical data”. In: *Computational Statistics and Data Analysis* 127 (2018), pp. 82–95.

[25] Kathryn Cooper and Mauricio Barahona. “Role-based similarity in directed networks”. In: (2010). arXiv: 1012.2726 [physics.soc-ph].

[26] K. Henderson et al. “It’s who you know: Graph mining using recursive structural features”. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’11. San Diego, California, USA: Association for Computing Machinery, 2011, pp. 663–671. ISBN: 9781450308137.

[27] M. E. J. Newman. “Modularity and community structure in networks”. In: *Proceedings of the National Academy of Sciences* 103.23 (2006), pp. 8577–8582.

[28] Alex Pothen, Horst D. Simon, and Kang-Pu Liou. “Partitioning Sparse Matrices with Eigenvectors of Graphs”. In: *SIAM Journal on Matrix Analysis and Applications* 11.3 (1990), pp. 430–452.

[29] Pons P. and Latapy M. "Computing Communities in Large Networks Using Random Walks". In: 3733 (2005).

[30] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona. "Stability of graph communities across time scales". In: *Proceedings of the National Academy of Sciences* 107.29 (2010), pp. 12755–12760.

[31] N Swainston et al. "Recon 2.2: from reconstruction to model of human metabolism". In: *Metabolomics* 12.109 (2016).

[32] TE Turner, S Schnell, and K Burrage. "Stochastic approaches for modelling in vivo reactions." In: *Comput Biol Chem* 28.3 (2004), pp. 165–78.

[33] Srinivasan S and Mahadevan R. Cluett WR. "Constructing kinetic models of metabolism at genome-scales: A review". In: *Biotechnol J* 10.9 (2015).

[34] J Orth, I Thiele, and B Palsson. "What is flux balance analysis?" In: *Nat Biotechnol* 28.3 (2010), pp. 245–248.

[35] K Yizhak et al. "Modeling cancer metabolism on a genome scale". In: *Molecular Systems Biology* 11.817 (2015).

[36] K Yizhak et al. "Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer". In: *eLife* Nov 21.3 (2014).

[37] Q Zhao et al. "Mapping the landscape of metabolic goals of a cell". In: *Genome biology* 17.109 (2016).

[38] In: ().

[39] G Su et al. "Integrated metabolome and transriptome analysis of the NCI60 dataset". In: *BMC Bioinformatics* 12 (2011).

[40] Laurent Heirendt et al. "Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0". In: *Nature Protocols* 14 (2019), pp. 639–702.

[41] LLC Gurobi Optimization. *Gurobi Optimizer Reference Manual*. 2019. URL: http://www.gurobi.com.

[42] Mahadevan R and Palsson BO. "Properties of metabolic networks: structure versus function". In: *Biophys J* 88.1 (2005), pp. L07–9.

[43] MATLAB. *version 9.5.0 (R2018b)*. Natick, Massachusetts: The MathWorks Inc., 2018.

[44] J. M. Monk et al. "iML1515, a knowledgebase that computes Escherichia coli traits". In: *Nature biotechnology* 35.10 (2017), pp. 904–908.

[45] J. D. Orth et al. "A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011". In: *Molecular systems biology* 7.535 (2011).

[46] Esther A Obeng et al. "Proteasome inhibitors induce a terminal unfolded protein response in multiple myeloma cells". In: *Blood* 107 (2006), pp. 4907–4916.

[47] Philippe Moreau et al. “Proteasome inhibitors in multiple myeloma: 10 years later”. In: *Blood* 120 (2012), pp. 947–959.

[48] In: ().

[49] Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT Press, 2006. ISBN: 0-262-18253-X.

[50] Carl Edward Rasmussen and Hannes Nickisch. “Gaussian Processes for Machine Learning (GPML) Toolbox”. In: *Journal of Machine Learning Research* 11 (2010), pp. 3011–3015.

[51] Zijing Liu and Mauricio Barahona. “Similarity Measure for Sparse Time Course Data Based on Gaussian Processes”. In: *bioRxiv* (2021). DOI: 10.1101/2021.03.03.433709.

[52] M. Meilă. “Comparing clusterings by the variation of information”. In: Lecture Notes in Computer Science 2777 (2003), pp. 173–187.

[53] Susmita Datta and Somnath Datta. “Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes.” In: *BMC bioinformatics* 7: 397 (2006).

[54] Yu G et al. “clusterProfiler: an R package for comparing biological themes among gene clusters.” In: *OMICS: A Journal of Integrative Biology* 16(5) (2012), pp. 284–287.

[55] Lilli Johanna Freischem. “Machine learning for analysing metabolic networks”. MA thesis. School of Informatics at the University of Edinburgh, 2021.