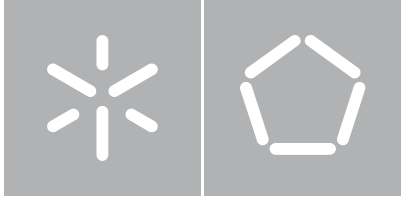
Universidade do Minho
Escola de Engenharia

Jorge Miguel Lourenço Ferreira

Development of integrated models of hepatocyte cells

Outubro de 2017



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Jorge Miguel Lourenço Ferreira

Development of integrated models of hepatocyte cells

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho realizado sob orientação de

Professor Miguel Rocha
Professor Julio Saez-Rodriguez

Outubro de 2017

DECLARAÇÃO

Nome: Jorge Miguel Lourenço Ferreira

Endereço eletrónico: jhmdf@hotmail.com

Telefone: 935838826

Número do Bilhete de Identidade: 13325902

Título da dissertação: Development of integrated models of hepatocyte cells

Orientadores: Miguel Rocha e Julio Saez-Rodriguez

Ano de conclusão: 2017

Mestrado em Bioinformática

Universidade do Minho, __/__/____

Assinatura: _____

Resumo

O metabolismo actua como uma máquina que mantém a funcionalidade da célula como resposta a várias perturbações, mantendo os níveis de metabolitos cruciais e componentes celulares e produzindo energia através da quebra de determinados compostos. Uma melhor compreensão destes mecanismos não se pode restringir ao conhecimento das funções de tecidos ou tipos celulares, também requer um conhecimento sobre as suas interacções. O fígado humano tem um grande número de funções fisiológicas relacionadas com o metabolismo, como a produção de bile, hormonas e vitaminas. Os hepatócitos têm um grande impacto no metabolismo humano, sendo as suas células as metabolicamente mais activas. Um mau funcionamento do metabolismo destas células está associado com algumas doenças, como hepatite, cirrose ou doença hepática gordurosa não alcoólica, onde esta última se encontra associada a obesidade.

Uma via metabólica particular tem sido associada não só com a obesidade, mas também com cancro e diabetes tipo 2, a via metabólica mecânica TOR (mTOR). A sinalização desta via tem efeito na maior parte das funções celulares e regula o crescimento e proliferação. Foi demonstrado que alterações nesta via pode levar a acumulação de gordura em pessoas obesas. Uma melhor compreensão desta via complexa pode ajudar os investigadores para revelar mais informação sobre como esta via funciona e como pode ajudar no tratamento de diversas doenças.

O aumento de dados provenientes de alto débito, devido aos avanços na sequenciação e outras técnicas experimentais, permitiram-nos ter um melhor conhecimento sobre as car-

acterísticas moleculares da célula. Uma ferramenta útil para processar toda esta informação são os Modelos Metabólicos à Escala Genómica (MMEG). Um MMEG é uma lista de reacções balanceadas pela massa, que pode ser relacionada com compartimentos celulares, como o citoplasma. Dados os dados de alto rendimento, MMEG podem ser utilizados para a simulação do metabolismo de um certo tipo celular através de modelação baseados em restrições. Existem vários algoritmos/ferramentas para criar modelos metabólicos específicos para um tecido (baseado em modelos metabólicos humanos, como o Recon2) incluindo o tINIT, MBA ou mCADRE.

Apesar de todos estes métodos ainda apresentarem algumas limitações, os modelos gerados pode simular tecidos humanos e ser um bom ponto de partida para uma melhor compreensão de doenças complexas. Uma limitação importante destes modelos é o facto de apenas representarem a camada metabólica da célula, enquanto para os modelos serem capazes de suportar simulações precisas, outros sub-sistemas (ex: regulação, sinalização) devem ser também tidos em consideração. Estes modelos (modelos integrados) combinam a informação e fluxo de material dos três sistemas previamente descritos, fornecendo assim uma ferramenta robusta com maior poder preditivo.

Abstract

Metabolism acts a machinery by maintaining the functionality of the cell in response to several perturbations, keeping a balance in the levels of crucial metabolites and cell components and producing energy by breaking down certain compounds. A better understanding of these mechanisms cannot be restricted to the knowledge of the function of specific tissues or cell types, it also requires knowledge about their interactions.

The human liver has a high number of physiological functions related to the metabolism, such as the production of the bile, hormones and vitamins. The hepatocytes have a major impact in human metabolism, being the most metabolically active cell types in humans. Malfunction on the metabolism of this type of cells is related to several diseases, like hepatitis, cirrhosis or non-alcoholic fatty liver disease (NAFLD), where the last one is considered a manifestation of obesity.

A particular pathway has been associated not only with obesity, but also with cancer and type 2 diabetes, the mechanistic TOR (mTOR) pathway. Signalling of this pathway has an effect on most of cellular functions and regulates growth and proliferation. It has been shown that alterations in this pathway can lead to fat accumulation in the liver of obese people. A better understanding of this complex pathway may help researchers to unveil more information on how this pathway works and how it can help in the treatment of several diseases.

The increase of high-throughput data, due to the advances in sequencing and other experimental techniques, allowed us to better understand the molecular characteristics

of the cell. A useful tool to process all this information are Genome-scale metabolic models (GSMMs). A GSMM is a list of mass-balanced reactions, which can be related to cellular compartments, like the cytoplasm. Given high-throughput data, GSMMs can be utilized for the simulation of the metabolism of a certain cell type through a constraint-based modelling framework. There are several algorithms/ tools to create tissue-specific metabolic models (based on a generic human model, such as Recon2) including tINIT, MBA or mCADRE.

Although all these methods still face a number of issues, the generated models can simulate human tissues and can be a good starting point for a better understanding of complex diseases. An important limitation of these models is the fact that they only represent the metabolic layer of the cells, while for models to be able to support accurate simulations, a number of other important sub-systems (e.g. regulation, signalling) should also be taken into account. This models (Integrative models) combine the information and material flow of the three previous mentioned sub-systems, delivering a more robust tool with more predictive strength.

Acknowledgments

First, I would like to thank Professor Miguel. He has been a great (if not the greatest) teacher I have had since I started to study. It has helped throughout the whole master degree and gave me the opportunity to go abroad and develop even more my knowledge.

To Julio, who has welcomed me as any person would want to be and helped with a lot with my work.

To all the new friends I made in Aachen, who helped me through all the time I was there and with my work.

To all my friends in Braga, who were always available for me when I needed help and with whom I share good times.

To Sara Correia, who has one hell of a patience with me, who helped me to get motivated and as taught me a lot in this area.

To André Santiago, who has been a dear friend since I joined the masters in Bioinformatics, for all the patience, friendship and pretty much anything a friend can expect.

To Tiago Alves, who has been also a great friend over the past years and to whom I have to thank a lot for making my life greater.

To my family, to whom I do not know how to write how much they have helped me during my whole life. From the good advices, to the times were I made mistakes, they have been there for me since the beginning and I would be be what I am today if it were not because of them.

And finally to Bárbara, who has been my support over the past years. From all the

laughs, good times, love and help, thank you for being at my side for all the moments since our first days.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Context

Metabolism acts a machinery by maintaining the functionality of the cell in response to several perturbations, keeping a balance in the levels of crucial metabolites and cell components and producing energy by breaking down certain compounds. A better understanding of these mechanisms cannot be restricted to the knowledge of the function of specific tissues or cell types, it also requires knowledge about their interactions (Bordbar et al., 2011; Hsu and Sabatini, 2008; Varemo et al., 2013).

The human liver has a large number of physiological functions related to metabolism, such as the production of the bile, hormones and vitamins (Tortora and Derrickson, 2014). The hepatocytes have a major impact in human metabolism, being the most metabolically active cell types in humans. Malfunction on the metabolism of this type of cells is related to several diseases, like hepatitis, cirrhosis or non-alcoholic fatty liver disease (NAFLD), where the last one is considered a manifestation of obesity (Neuschwander-Tetri and Caldwell, 2003).

The increase of high-throughput data, due to the advances in sequencing and other experimental techniques, allowed to better understand the properties of cells and their

behavior. Useful tool to process all this information, mainly related to metabolism, are Genome-scale metabolic models (GSMMs). A GSMM is a list of mass-balanced reactions, which can be related to cellular compartments, like the cytoplasm.

Given high-throughput data, GSMMs can be utilized for the simulation of the metabolism of a certain cell type, through a constraint-based modeling framework. There are several algorithms/ tools to create tissue-specific metabolic models (based on a generic human model, such as Recon2 (Thiele et al., 2013)), including tINIT (Agren et al., 2014a), MBA (Jerby et al., 2010a), mCADRE (Wang et al., 2012b), FASTCORE (Vlassis et al., 2014) or CORDA (Schultz and Qutub, 2016). There have been several biomedical applications for the reconstructed models throughout the years (Kim et al., 2014; Mardinoglu et al., 2014; Park et al., 2012) using these methods, a significant part of which are related to cancer.

Cancer is a disease commonly associated with an unrestrained cell growth and able to invade several parts of the human body (Seely, 1980). The urge for understanding and finding new ways to fight cancer has been a challenge for the past years for the scientific community.

The Genomics of Drug Sensitivity in Cancer (GDSC) project comprises around 1000 different gene expression profiles cell lines of cancer; alongside with this, there is also 265 different drug IC50 values related to them (Yang et al., 2013).

For this work, the associations between the drugs and gene expression were analyzed, as well as with reconstructed models for these cell lines.

With this we want to achieve a better understanding of the importance and possible applications of the reconstructed models for the cancer knowledge and treatment.

1.2 Objectives

Given the context presented above, one of the main aims of this work is the development of reconstructed metabolic models for liver cells, addressing the evaluation of available tissue-specific model reconstruction algorithms. This should allow to obtain a better understanding of its metabolism and changes occurring in both normal and cancer cells.

As a second aim, and after assessing the most suitable algorithm and applying it to a panel of cancer cell lines, we will try to evaluate if reconstructing metabolic models based on gene expression can lend more knowledge. In particular, we verify if it can be complemented by analyzing data on drug sensitivity for a set of rugs administered to the same cell lines. In the final part of the work, we will try to unveil mechanisms of action of the drugs based on the associations obtained.

In more detail, the work will address the following scientific/technological objectives:

- Review the state of the art in the constraint-based modeling of hepatocytes and liver cancer cells, and also on available tissue-specific reconstruction algorithms;

- Based on available experimental data (transcriptomics, proteomics), apply different methods for the metabolic model reconstruction of hepatocyte cell liness (both for normal and cancer ones);

- Create a pipeline for the discovery of associations between genes/reactions with drugs by performing an adequate statistical analysis;

- Achieve a better understanding for the mechanism of action of the drugs and its targets by analyzing the locations of the genes in a simplified network;

1.3 Thesis organization

In the second chapter we describe the state of the art of the work. We will define certain important concepts, such as metabolism, cancer and reconstructed models.

In the third chapter, the materials and methods are presented and discussed. We will describe the data used and the methods used (with a brief explanation of how they work), as well as a separation of the two parts of this work, which will be the reconstruction of liver metabolic models using different algorithms and how the associations between the gene expression/reconstructed models against drugs are calculated.

In the fourth one, we will present and discuss the results in this work. As described before, the work is divided in two parts, so in the first part we present the results for the model reconstruction algorithms, using both liver normal and cancer data. As part of the analysis, we evaluate the similarities and differences in the models, performed a functional analysis and evaluate if they were capable of producing biomass as well as fulfill specific liver tasks. In the second part, we present and discuss the results obtained from the associations generated with the gene expression and the reconstructed models.

In chapter five, we present the conclusions for the work and some future perspectives.

Chapter 2

State of the Art

2.1 Metabolism

Metabolism is a cellular mechanism mainly responsible for the control of the growth and good functioning of the cells. The main goal for unicellular organisms is to be able to grow as much their environment allows, which means that with the right amounts of carbon, energy and nutrients, they are able to generate new cells.

For multicellular organisms, like humans, the availability of required molecules to produce a new cell usually does not represent a problem. Physiological functions, as neural communication or relaxing a muscle, require energy which is obtained from metabolic processes (Muñoz-Pinedo et al., 2012; Vander Heiden et al., 2009). Cells must be able to manage their mechanisms to proliferate.

The human liver is responsible for a major part of the metabolism. It is relevant to the production of bile, certain vitamins and hormones, storage of glycogen or degradation of toxic substances, regulation of certain components of the blood like the plasma or red blood cells and glucose (Tortora and Derrickson, 2014). The liver is composed with two types of cells; parenchymal ones (bile duct and hepatocytes cells) and the non-parenchymal, which comprises the Kupffer, hepatic stellate and sinusoidal endothelial

cells (Kmieć, 2001).

Deregulation of the functions of the liver can lead to several diseases. The main diseases associated with the liver are hepatitis, nonalcoholic fatty liver disease (NAFLD), cirrhosis and hepatocellular carcinoma (HCC) which affects more than half a million people worldwide (Baffy et al., 2012; Finn, 2010; Lanpher et al., 2006). The study of the molecular and cellular mechanisms of the liver can lead to a better understanding of the disorder and improving the knowledge about normal and disease states.

An important mechanism in mammals is the signaling network mTOR (also known as "mechanistic TOR"). This specific pathway controls diverse processes that either control the use of nutrients or the production of energy, which consequently regulates the growth and proliferation of the cells, macromolecule biosynthesis or cytoskeletal organization. Although cancer seems an obvious disorder caused by the deregulation of the previous processes, this pathway is also related with obesity, type 2 diabetes and neurodegeneration.

mTOR is composed by two different complexes of proteins known as mTOR complex 1 (mTORC1, which has six proteins) and mTOR complex 2 (mTORC2, composed by seven proteins). Specifically, while mTORC1 reacts to amino acids, stress, oxygen or growth factors, mTORC2 is responsible to respond to growth factors or regulation of cell survival (Laplante and Sabatini, 2013).

As stated before, the liver is responsible for the control of glucose and lipid homoeostasis, and the mTOR pathway plays an important role in the well functioning of the liver. If the levels of activity of the mTORC1 are low (for example, during fasting), studies have demonstrated that this complex does not impair the activation of the ketogenesis, a biochemical process responsible for the degradation of the fatty acid as opposed when it activity is high, where the hepatic lipogenesis is upregulated, leading to the nonalcoholic fatty liver disease. One serious consequence of it, a disorder where there is a unnatural fat accumulation in the liver, is obesity, which can be implicated in the appear-

ance of others diseases, like cirrhosis or hepatocellular carcinoma (Sengupta et al., 2010; Yecies et al., 2011).

2.2 Cancer

2.2.1 Hallmarks of Cancer

Cancer is a disease that is characterized by an uncontrolled cell growth and, in some cases, is able to invade several parts of the body. There are known over 200 types of cancer (but not in all parts of the human body), being the most frequent ones the breast, lung and colon ones, while it is very rare to have heart or skeletal muscle cancer (Seely, 1980).

The complexity of the research of this disease has been increasing over the years. A number of "hallmarks" were suggested to have a way of describing several features of the disease (Hanahan and Weinberg, 2000, 2011). This can be seen as a "multi-layered" process, where each layer has a specific contribution either in genetic or morphological changes, which contribute to the evolution of cancer cells. Despite the number of different types, there some specific traits which can be "shared" that are acquired during tumor development.

The first six hallmarks of cancer were suggested (Hanahan and Weinberg, 2000) and later revisited (Hanahan and Weinberg, 2011). The first one is the self-sufficiency in growth signals, which is a required process for cells when they are in active proliferative state. Mostly, the signals are transmitted by signaling molecules that can be classified as diffusible growth factors, extracellular matrix components, cell-to-cell adhesion and/or interaction molecules. Although, in a normal condition, convenient diffusible mitogenic factors are required, tumor cells can overcome this need by generating their own growth signals (Hanahan and Weinberg, 2000).

Although this process is crucial, cells must also be able to suppress other signals from their neighborhood that are anti-proliferative (regulated by tumor suppressor genes).

The main two genes associated with this process are the RB (retinoblastoma-associated), which responds both to extra- and intra-cellular signals and decides the fate of the cell, and TP53 proteins, which in their turn, respond to the stress and aberrant intra-cellular function, like DNA damage, lack of nucleotides, nutrients or growth-promoting signals or levels of oxygen, arresting the cell cycle until the optimal conditions are met. In extreme cases, where the damage is irreversible, TP53 proteins can trigger apoptosis. Although they are important genes to take into account, their single absence in some cases has no influence in the outcome of proliferation. This suggests that they are a part of a bigger network and that there are other underlying mechanisms that can fill the gaps and need research (Hanahan and Weinberg, 2011).

The next suggested hallmark for cancer is insensitivity to antigrowth signals. In normal conditions, the cells "communicate" with each other to preserve cellular quiescence and tissue homeostasis, usually through transmembrane cell surface receptors linked to intracellular signaling circuits. To control the cellular growth, cells may be "forced" to stay in the G_0 state, only proliferating again if needed by the organism or when differentiated into specific types of cells. During cancer development, the cells are able to "jump out" of the G_0 state, becoming "insensitive" to anti-proliferative signals and progressing to the G_1 and S phase. But there is more to add, since cancer may also arise from tissues where cells are told to enter a differentiated state (which should be irreversible). However, the change in certain mechanisms may alter this and lead cells to proliferate.

Another hallmark of cancer is escaping the cell's programmed death, known as apoptosis. It is probably the most common mechanism in cancer, and it further helps the uncontrollable expansion of the disease. Normally, the apoptosis' mechanism is present in all cells across the human body and, once activated, it unlocks a very consistent "cascade" of processes to achieve its goal. The disintegration of the cell membrane, the cleavage of cytoplasmatic and nuclear skeleton, the expel of the cytosol, disassemble of chromosomes and disintegration of the nucleus and lastly, the remaining of the cell is assimilated

by other surrounding cells; all of these steps are part of apoptosis.

The apoptotic process can be divided in two components: sensors and effectors. The first is a mechanism that is evaluating if the cell should die or not by overseeing the intra- and extra-cellular environment for anomalies. The latter is "controlled" by the first, so if there is any signal that could instruct the death of the cell, the effectors initiate the apoptotic process. Externally, cell surface receptors that bind to survival or death factors are responsible for triggering apoptosis, while internally, the "mission" is to evaluate abnormalities, like DNA damage, uneven signaling or hypoxia (Evan and Littlewood, 1998).

Putting together all the three past hallmarks, we have the "perfect" combination for an unlimited and uncontrolled proliferation of cancer cells. However, researchers found that even if the previous conditions were met, alone they were not sufficient to lead to tumor growth, since some mammalian cells have a limited number of replications (controlled by the cell), achieving then the process called senescence (Hayflick, 1997). It is thought that this mechanism is independent of cell-to-cell signals described before, and it has to be also disrupted for the tumor to grow into a malignant tumor.

As any other cell in the organism, certain requirements must be achieved. Nutrients and oxygen are vital for cell survival and function. Particularly during human development (organogenesis), the regulated formation of blood vessels is crucial. As for cancer, as a requisite to increase even more the tumor, it must develop angiogenesis (Bouck et al., 1996; Hanahan and Folkman, 1996). As other mechanisms of the cell, it is important to keep a balance between the different signals either to favor or stop angiogenesis. Soluble factors and its receptors on the surface of the endothelial cells are one form of signal to control this process, as well as integrins and adhesion molecules which mediate cell-to-matrix and cell-to-cell associations. This happens during the tumor development when there seems to be a "switch" that makes angiogenesis active and sustained again in tumor cells, specially in the mid-stage lesions, before the appearance of developed tumors. How-

ever, in some forms of cancer like human cervix, breast and skin (Hanahan and Folkman, 1996), it seems that angiogenesis occurs in an earlier stage.

One of the most aggravating characteristics of cancer is the development of metastases, a process where the cells "move" from their original place, invade another tissues and "conquer" new places to continue its proliferation. Due to this, the lethality of the disease skyrockets up to 90% (Sporn, 1996). We can see this as a migration for a new part of the human body, looking for more space and nutrients (an analogy could be made with the nomadism). This hallmark shares similar characteristics with the others, like activation of extra-cellular proteases to adapt to the new environment and physical pairing with the "new" cells.

Although the hallmarks are important processes for a better understanding and characterization of the cancer, there are some important characteristics that enable their rise. Changes in the genome (like mutations) are usually acquired by cancer cells. Although single gene mutations are not an efficient way to create genome instability (due to the mechanism of DNA repair), the structures that oversee the state of the cell are the ones that lead to an increase of the mutations on the genome (Lengauer et al., 1998).

One of the most common alterations in cancer is the alteration of the TP53 tumor suppressor proteins (Levine, 1997). Besides this cell viability check, other genes/proteins associated with other mechanisms like chromosomal segregation during mitosis, have been implied in cancer (Lengauer et al., 1998), giving an advantage in its growth. Another enabler of this could be the leftovers of an apoptotic body (Holmgren et al., 1999), due to the horizontal transfer of genes when the phagocytosis occurs. The genome instability may be considered the main trigger for "appearance" of the hallmarks of cancer.

Although the past six hallmarks have been extensively reported, there are two emerging ones suggested in later work (Hanahan and Weinberg, 2011). One of them is related with how the cancer cells are able to evade the immune system destruction. Our organism has a "constant" surveillance for any kind of abnormalities, including early developed

cancer cells and by some modifications, solid tumors have found a way of tricking it or at least shorten the number of tumor cells eradicated, so it can be seen as an effective obstacle for tumorigenesis and tumor progression. This has been demonstrated in cases were rats immunodeficient for certain components of the immune system developed more times and/or faster when compared to the control . Also, when transplanting tumor in the first stage from immunodeficient rats to immunocompetent ones, the tumor was not able to advance to the second stage, whereas in the reverse case, the tumor began to develop in the immunodeficient one (Kim, 2007; Teng et al., 2008). This behavior of the immune system has been described as "immunoediting", - constant destruction of highly immunogenic cancer cell clones. A side effect of this is "leaving behind" cancer cells which are less immunogenic and later grow into solid tumors (Smyth et al., 2006).

2.2.2 Metabolism and Cancer

Like any other machinery, that to function needs "fuel", cells also need energy. Looking at it from a simple point of view, under aerobic conditions, glucose is first transformed into pyruvate (glycolysis, in the cytosol) and then into carbon dioxide (in the mitochondria). Under anaerobic conditions, glycolysis is preferred and a small amount of pyruvate is transferred to the oxygen-consuming mitochondria.

This was one of the characteristics observed related to the metabolism of cancer cells, where the cells were able to reprogram their glucose metabolism (and therefore the energy production), by facilitating predominantly glycolysis, named "aerobic glycolysis" (Warburg, 1956; Warburg et al., 1927; Weinhouse et al., 1956). Although this process is by far less efficient than the "regular" glycolysis, there are some advantages to cancer cells when opting for the "aerobic" one. When revisiting an old theory (Potter, 1958), Vander Heiden and his colleagues hypothesized that this shift in the pathway for production of ATP can lead to a deviation of some intermediates of the glycolysis for another pathways, mainly related to the generation of nucleotides and amino acids, aiding he production of

macromolecules and organelles needed for the manufacturing of new cells (Vander Heiden et al., 2009). They also suggest that alterations which favor tumor progression lead the cell to a self-governing nutrient uptake and force the metabolism to the proliferative state; on the other hand, when trying to suppress the cancer, the cell restricts pathways that require nutrient for anabolic purposes. This may be possible for the cell if it reverts to its embryonic form or if the cell "evolves" for the facilitating of the metabolism to help cell growth.

It has been also found that there is a second type of population of cells which compensate the waste of the "aerobic" ones, which expels lactate as waste, and have found a way to metabolize it to energy, using part of the citric acid cycle to achieve it (Feron, 2009; Kennedy and Dewhirst, 2010; Semenza, 2008). Although this may seem a "new" feature for the cancer cells, the same process occurs in the muscle cells.

2.2.3 Drugs discovery for cancer treatment

Looking at all the hallmarks, we can begin to think about a more personalized and effective way of treatment for this disease. Looking to the rapid growth of available therapeutics, we can categorize them through the hallmarks. If a drug is targeted for a specific hallmark, it would be expected to lead, at least, to show some progress in the cancer treatment (or at least be efficient in killing it), but that is not always the case.

This may suggest the existence of a core of pathways that are shared among the hallmarks; so, even if a drug targets a specific hallmark, the shared core still can develop resistance to the treatment, overcome the selective pressure (by either mutation, epigenetic reprogramming and others mechanisms), making the tumor grow in such conditions. Even in some cases, there was an unbalance in the "importance" of the hallmarks in order for it to survive. For example, when administrated angiogenesis inhibitors in certain preclinical models, even though the hallmark was successfully suppressed, the models favored the invasion and metastasis, to obtain the requirements for their growth (Azam

et al., 2010; Bergers and Hanahan, 2008; Ebos et al., 2009).

Looking at all the issues that have been described and the amount of information present, informatics opens a new way of processing all the information. Processes like analysis of sequence similarity (both for genes or proteins) or annotation of high-throughput data created a new field of study, Bioinformatics. Allied with computational modeling and statistical analysis, it has achieved the development of ontologies (to merge the medical and "biological" knowledge), applied statistics (for the test of hypothesis) and computational biology (for the generation of models, some cases as described before). Also, text mining tools have also been proved useful to obtain a faster and accurate retrieval of information.

Since the disease is so heterogeneous, the need for an individualized therapy has never been so imperative, and here Informatics could provide a vast help. If we look at the particularities of the disease, such as unique biomarkers to each individual at a molecular level, the genetic evolution of the patient, to access the outcome of each treatment, could be the work, allied with other areas of knowledge, for bioinformatics. Ideally (and possibly in a near future), treatments will not have aggressive side effects, a model for the disease will be generated according to the genomic, proteomic, metabolic (and so on) profile of the patient and the treatments will be more and more efficient (Ochs et al., 2010).

As we stated before, cancer is a complex disease and a mathematical approach could be an important way to lead for a better understanding of it. Although it may not be simple (cancer is supported by nonlinear dynamics, such as growth rate, rate of mutations and others), it could provide insight for more than the empirical, traditional view of the disease. In order to do so, we have to combine both systems. While the main component of research should be biomolecular (both *in vitro* and *vivo*), it should be also coupled with hypothesis approach, modeling the disease with mathematical models for the purpose of construction of frameworks to understand the data, guide new experiments and accelerate

new discoveries (Gatenby, 2010).

With the increase of knowledge granted by this, the number of potential drug target candidates continues to increase every day, but unfortunately it is not translated into new and more effective drugs (Cartwright et al., 2010). The main issue with the introduction of new drugs in the market is the rate failure in clinical trials, mainly in phases II and III (Paul et al., 2010), primarily due to the reasons described before. With the help of system biology and bioinformatics, additional insights can be given to the drug mechanism of action understanding and discovery, in other words, integration of drugs (or a combination of them) with physiological pathways and complex disease systems, like cancer (Fernald et al., 2011). A clear advantage of this approach is the modulation of the dosage (consequently, the significance of several regimens), leading to a decrease in the cost of development of the drug (Wang and Deisboeck, 2014).

2.3 Constraint-based Metabolic Modeling

2.3.1 Principles of constraint-based approaches

Systems biology is a science field based on the construction and *in silico* validation of biological models, using data obtained from experiments. A reconstruction is a set of biochemical reactions that occur in a certain cellular system, like the metabolism, taking into account relations between proteins, transcripts and genes and their respective reactions. This can ultimately be converted to a model with the inclusion of the flux and nutrient flow rates. The main objective of these models is to simulate a cell under different conditions.

To build models, several considerations have always to be made, like the sub-cellular localization of the metabolic reactions and information about gene-protein-reactions (GPR) where it is important to take in consideration the alternative splicing of each gene (Ryu et al., 2015). Basically, a GPR is a description of a reaction. If a gene is involved in an OR condition, the gene will be automatically taken into account to build the relationships;

if it is in an AND condition, the genes that form it can only be taken into account if all the genes have an association found (Figure 2.1).

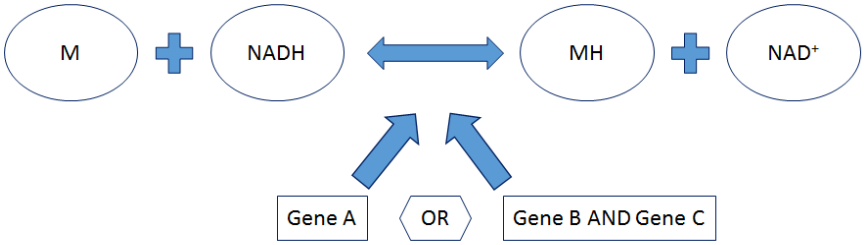


Figure 2.1: Example of a GPR. Considering **M** as a metabolite, this reaction can happen in two different ways. If gene A is present, the reaction occurs in the cell (since it has an OR condition, the gene alone is enough for the reaction to take place). But, if the gene A is missing, only if genes B and C are present, will the reaction take place; if one of them is missing (and it is an AND condition), the reaction will not occur.

Constraint-based metabolic modeling principle resides in the mass balance of metabolites, the assumption of a pseudo-steady state and the use of a stoichiometric matrix to be able to perform simulations using numerical optimization (Orth et al., 2010). It can also be used to simulate different variables (genetic, physicochemical and environmental) that are imported to the model in the form of constraints that are taken in account when the optimization is being performed for the prediction of fluxes (Ryu et al., 2015).

For this analysis, Flux Balance Analysis (FBA) can be used to achieve phenotype simulation, calculating a flux distribution, through an optimization approach, which maximizes an artificial biomass flux (representing cell growth). It takes into account these constraints, alongside with the reversibility and other constraints over maximum and minimum flux values (Kauffman et al., 2003).

2.3.2 Human Metabolic Models

The decrease of the cost of genome sequencing and other high-throughput omics data and the scientific advances in bioinformatics tools have enabled reconstructions, not only for smaller organisms, but also for eukaryotes. The first metabolic human GSMM was

released in 2007, the Recon 1, with a very accurate manual curation for the model to be able to validate each single metabolic reaction (Duarte and Becker, 2007).

Currently, the most extensive human metabolic models are the Recon 2 (Thiele et al., 2013) and HMR 2.0 (Mardinoglu et al., 2014). The Recon 2 is the result of the merge of other metabolic components present in Recon 1, together with the Edinburgh Human Metabolic Network (EHMN) (Hao et al., 2010), HepatoNet1 (Gille et al., 2010a), a module containing information about acyl carnitine and fatty-acid oxidation (Sahoo et al., 2012) and another model with data about the human small intestinal enterocyte (Sahoo and Thiele, 2013). Regarding the HMR 2.0, it has also integrated information from the Recon 1, EHMN, HepatoNet1, iHuman1512 (Agren et al., 2012a) and iAdipocytes1809 (Mardinoglu et al., 2013a), and also information from the four major metabolic databases (KEGG (Kanehisa et al., 2012), HumanCyc (Romero et al., 2005), LIPID MAPS Lipidomics Gateway (Fahy et al., 2009) and REACTOME (Croft, 2013)).

2.3.3 Tissue-specific Metabolic Models

The creation of methods for the integration of omics data for the generation of tissue/cell type-specific metabolic models has been of crucial significance for a better understanding of the biochemical and genetic complexity of the human metabolism (Mardinoglu et al., 2013b). Still, the integration of the omics data to generate tissue-specific models raises significant challenges.

Given the different algorithms to create tissue-specific metabolic models based on a generic human model, we briefly explain the five that will be used in this work. INIT/-tINIT, MBA, mCADRE and FASTCORE were already implemented by Sara Correia. The pseudo-code of all the algorithms described can be viewed in Table 2.1. The addition of the CORDA algorithm is an effort to improve the current tool and knowledge for the liver, and was done as part of this work.

INIT/tINIT

The Integrative Network Interface for Tissues (INIT) algorithm maximizes the matches between reaction states (active or inactive) and data regarding expression or non expression of genes/proteins, returning flux values and a tissue-specific model (i.e. a set of reactions from the original model considered to exist in the tissue). The method solves a Mixed Integer Linear Program (MILP), where binary variables represent the presence of each reaction from the template model in the resulting model. Although the algorithm normally uses proteomic data from HPA, transcriptomics can also be given as an input.

In the definition of the objective function, positive weights are given to reactions with a higher evidence from the input, and negative to the ones who have low or no expression. If there is supportive information (usually metabolomics) that corroborate the presence of a certain metabolite, the necessary reactions may be included in the final model to produce (Agren et al., 2012b). The Task-driven INIT (tINIT) is an extension of the previous algorithm (Agren et al., 2014b). The improvement is based on the possibility to define a metabolic task in agreement with the context of the reconstruction. These may be the consumption or production of a metabolite or activation of the reactions of a particular pathway for the tissue.

MBA

Differently from INIT, the Model-Build Algorithm (MBA) (Jerby et al., 2010b) returns only a final model and no flux values. This algorithm accepts as input a generic metabolic model and two sets of reactions. The first set (C_H) comprises reactions with high support (e.g. literature) for the inclusion on the final model, while the other one (C_M) contains usually information derived from high-throughput data. In an iterative way, all the non-core reactions (based on the previously established sets) are removed in a random order, while the model is tested for consistency. The iteration ends when all the reactions have been submitted for the removal test in the final model. The final model should contain the

whole set of the C_H, a maximum number of reactions from the C_M and the least possible of the remaining non-core ones, normally requested to avoid connectivity issues.

Since the order by which each reaction is tested for removal matters, there is the need to repeat this algorithm several times to obtain a set of models. The final one should be a model based on the ranking of the frequency of the reactions in the set, adding them to the C_H core until a coherent model is found (Jerby et al., 2010b).

mCADRE

The Metabolic Context specificity Assessed by Deterministic Reaction Evaluation (mCADRE) (Wang et al., 2012c) algorithm is quite similar to the MBA, but only requires the reconstruction of a single model. It is initialized by ranking the reactions on the original model using three distinct scores: confidence, expression and connectivity. With the help of a threshold value for the scores, a core of reactions and the order of removal of the non-core ones are established.

The input for the algorithm considers the frequency of expression states in a set of profiles (requiring a change of the data to binary values), instead of levels of expression. Regarding the connectivity, the reactions are ranked by the reactions in the "neighbourhood". For the confidence levels, the reactions are ranked according to the evidences of that reaction in the general metabolic model.

In the process of the reconstruction, if the removal of a non-core reaction does not compromise the production of essential metabolites and the core of reactions, those reactions are removed on the previous order. However, if a particular situation requires it, the elimination of core reactions is possible.

FASTCORE

In a similar approach to MBA (trying not to alter the set of core reactions), the FASTCORE (Vlassis et al., 2014) algorithm uses another strategy by solving two Linear Prob-

lems (LP). The first maximizes the number of reactions in the core, comparing the values of a reaction with a constant, while the other decreases the number of reactions that are absent in the core by minimizing the L_1-norm of the flux vector. Until the core is coherent (the whole set of core reactions is activated with the smallest number of non-core reactions), both problems are being solved alternatively and in a repeated way. For reversible reactions, the algorithm analyses both directions.

CORDA

One of the main features of this algorithm is that it only needs a FBA (which is a Linear Problem) and can provide a faster reconstruction in comparison to other algorithms. As a novel approach, the developers of the algorithm created the *dependency assessment* as a new way to identify the importance of desirable reactions (with higher evidence) in contrast to the one with less information.

They start by modifying the network in four different ways. In the first step, they split the reversible reactions into forward and backward ones. The second step is the addition of a pseudo-metabolite for every single reaction in the model. Reactions who have less evidence will have a higher "cost" associated. On the third step, a reaction is added to the model consuming this pseudo-metabolite.

At last, a positive lower bound is assigned to the reaction being tested, forcing it to carry flux. After these modifications, FBA is performed, while minimizing the flux of the reaction added on the third step. With this step, any reaction with an high cost should not be included, unless it is necessary for the reaction being tested to be able to carry flux itself.

Similar to other algorithms, the reactions are classified in four groups by their confidence (High (HC), Medium (MC), Negative (NC) and Other (OT)). After the definition of the groups, the algorithm starts by including all the reactions present in the HC group into the final model (called as RE in the original paper). Using the *dependency assess-*

ment with the help of FBA, the algorithm tries to find associations between MC and NC reactions with each reaction of the RE (in this moment, it is the same as the HC) and are moved to the model being reconstructed.

The next step tackles the remaining NC reactions that may be associated with the remaining reactions of the MC group and are also transferred to the final model. The reactions of NC group that are left out are blocked (both bounds set to zero). As the third step, MC reactions are tested for the ability to carry flux. If they pass the test, they are moved to the final model. In the final step, the OT reactions that are associated with any reaction from the RE group are also included in the final model (Schultz and Qutub, 2016).

2.3.4 Biomedical applications of constraint-based modeling

Although these models are simply a mathematical representation of a "cell", they have proved that their application can have an high value for biomedical purpose. Characteristics like its ease of implementation or their potential predictive power have made possible the prediction of which genes to manipulate in metabolic engineering (production of shikimic acid and putrescine in *E. coli.*) (Park et al., 2012), predict drug targets (five essential metabolites were considered critical to the *Vibrio vulnificus* CMCP6 and lead to the selection of their chemical analogs) (Kim et al., 2014) and specific cells linked to diseases, for example, the hepatocytes with patients who suffered for nonalcoholic fatty liver disease (with their experiment, they were able to demonstrate that the analysis of chondroitin and heparan sulphates were crucial to diagnose nonalcoholic steatohepatitis and to determine the stage of nonalcoholic fatty liver disease) (Mardinoglu et al., 2014).

Table 2.1: Formulation and description of algorithms of MBA, tINIT, mCADRE and FASTCORE. In the table, R_G represents the list of reactions from the global template model, R_C the set of core reactions on mCADRE, C_H and C_M the core and moderate probability sets used in MBA, r a reaction and the $for(i)$ and the $rev(i)$ represent the *i-th* reaction direction (forward and reverse). In the FASTCORE algorithm, N is the set of all reactions in the model, C is the core set of reactions, and I the set of irreversible reactions. $J \subseteq C$ is a set with the irreversible reactions from C and $P = (N \backslash C) \backslash A$ is a "penalty" set which contains all the non-core reactions that have not been added to A.

MBA

$generateModel(R_G, C_H, C_M)$
- $R_P \leftarrow R_G$
- $R_S \leftarrow R_P \backslash (C_H \cup C_M)$
- $P \leftarrow randomPermutation(R_S)$
- $for(r \in P)$
  - $inactiveR \leftarrow CheckModel(R_P, r)$
  - $e_H \leftarrow inactiveR \cap C_H$
  - $e_M \leftarrow inactiveR \cap C_M$
  - $e_X \leftarrow inactiveR \backslash (C_H \cup C_H)$
  - $if(|e_H| == 0\ AND\ |e_M| < \delta * |e_X|)$
    - $R_P \leftarrow R_P \backslash (e_M \cup e_X)$
  - $endif$
- $endfor$
- $returnR_P$

$endfunction$

tINIT

$min \sum_{i \in R} w_i * y_i$

$s.t.$
- $Sv = b$
- $|v_i| \leq v_{max}$
- $0 < v_i + (v_{max} * y_i) \leq v_{max}$
- $b_j \geq \delta\ j \in Metabolomics$
- $b_j = 0\ j \notin Metabolomics$
- $y_{for(i)} + y_{rev(i)} \leq 1$
- $v_i \geq \delta, i \in RequiredReac$
- $y_i \in 0, 1$
- $w_i, score\ for\ i \in R$

mCADRE

$generateModel(R_G, treshold)$
- $R_P \leftarrow R_G$
- $R_C \leftarrow score(R_P) > treshold$
- $coreActiveG \leftarrow flux(r)! = 0, r \in R_C$
- $R_{NC} \leftarrow R_P \backslash R_C$
- $for(r \in order(R_{NC}))$
  - $inactiveR \leftarrow CheckModel(R_P, r)$
  - $s1 = |inactiveR \cap R_C|$
  - $s2 = |inactiveR \cap R_{NC}|$
- $if(r \notin withExpressionValues\ AND$
  $s1 \backslash s2 <= RACIO\ AND$
  $checkModelFunction(R_p \backslash inactiveR))$
  - $R_P \leftarrow R_P \backslash inactiveR$
- $elseif(|s1| == 0\ AND$
  $checkModelFunction(R_p \backslash inactiveR))$
  - $R_P \leftarrow R_P \backslash inactiveR$
- $endif$
- $returnR_P$

$endfunction$

FASTCORE

$FASTCORE(N, C)$
- $J \leftarrow C \cap I$
- $flipped \leftarrow False, singleton \leftarrow False$
- $A \leftarrow findSparseMode(J, P, singleton)$
- $J \leftarrow C \backslash A$
- $while(J \neq \emptyset)$
  - $P \leftarrow P \backslash A$
  - $A \leftarrow A \cup findSparseMode(J, P, singleton)$
  - $if(J \cap A \neq \emptyset)$
    - $J \leftarrow J \backslash A, flipped \leftarrow False$
    - $else$
      - $if(flipped)$
        - $flipped \leftarrow False, singleton \leftarrow True$
      - $else$
        - $flipped \leftarrow True$
        - $if(singleton)$
          - $\widetilde{J} \leftarrow firstElement(J)$
        - $else$
          - $\widetilde{J} \leftarrow J$
        - $endif$
      - $for(r \in \widetilde{J} \backslash I)$
        - *flip the sign in stoichiometric matrix*
        - *and swap the bounds of reaction r*
      - $endfor$
    - $endif$
  - $endif$
- $endwhile$

$endFunction$

2.4 Omics Data

As stated before, the evolution of the high throughput technologies have improved and generated high amounts of data from various sources, such as genetic, proteomic or metabolic (Duarte and Becker, 2007). We will now describe some of the sources and how the information is generated.

The Gene Expression Barcode (GEB) is a database which contains gene expression information for 131 human tissues (also including disease ones). These data are generated by an algorithm when accessing information from Gene Expression Omnibus (GEO) and ArrayExpress which contain information about microarrays or next-generation sequencing (McCall et al., 2014).

The Human Protein Atlas (HPA) is a database which contains information about proteins from different types of tissue. This data is obtained from immunochemistry on tissue microarrays (Uhlen et al., 2010).

The Genomics of Drug Sensitivity in Cancer (GDSC) project has more than 1000 different cancer cell lines, comprising cancer of epithelial, mesenchymal and haematopoeietic origin for both adults and children. These have been genomically characterized by the Cancer Genome Project at the Wellcome Trust Sanger Institute. They include information on somatic mutations in 75 cancer genes, markers of microsatellite instability, tissue type and transcriptional data, among other information.

Also related to the GDSC database, we are able to find data related to compounds who have anticancer therapeutic properties, as cytotoxic chemotherapeutics or targeted agents. In this collection of drugs, there are already approved compounds, others that are still under clinical development and trials, and even new drugs who are on a initial phase of development. Processes like cell cycle control, DNA damage response and receptor tyrosine kinase signalling or targets like cytoskeleton or the mTOR complex are a few of the examples that are related with cancer biology and that are the target of some drugs.

The sensitivity of a drug in a given cell line is measured using fluorescence-based cell

viability assays followed by a 72 h drug treatment. The values hosted on the database are the half maximal inhibitory concentration (IC50), in other words, how much of a given compound is necessary to inhibit a biological process by half, the slope of the dose-response curve and the area under the curve (AUC) for each experiment (Yang et al., 2013).

Another database containing information about cancer is the The Cancer Genome Atlas (TCGA) which is a collaboration from several organizations in order to map genes that are associated with cancer. However, now it is integrated within the NCI Center for Cancer Genomics (CCG), a new database containing information also from other databases, such as Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative and the Cancer Genome Characterization Initiative (CGCI) (Grossman et al., 2016).

Chapter 3

Materials and Methods

In this chapter, we will describe the methods required to perform this work. First, we selected the data that we were going to use in order to use. For the evaluation of the tissue specific reconstruction, we used liver data (both for normal and cancer cells) and for the second we utilized the cells line from the GDSC. Also, we decided that the template model for the human cell would be Recon1, since it is one of the most well established models available. After the evaluation of the tissue specific algorithm through different parameters, we decided to pick the FASTCORE one to perform a tissue-specific reconstruction for all the cell lines present in the GDSC. With the help of GDSCtools and PyPath package, we tried to uncover new relationships with the IC50s of the drugs that were tested with those cell lines and evaluate the new insights that this approach could give. The Figure 3.1 shows an overview of all the process.

3.1 Models and Data

Three distinct data types were used for this work. The Liver Data is used for the first objective of the work and the GDSC cell lines and Colon Cancer data are utilized for the latter. The Human Genome-Scale Metabolic Model is crucial for the reconstruction of

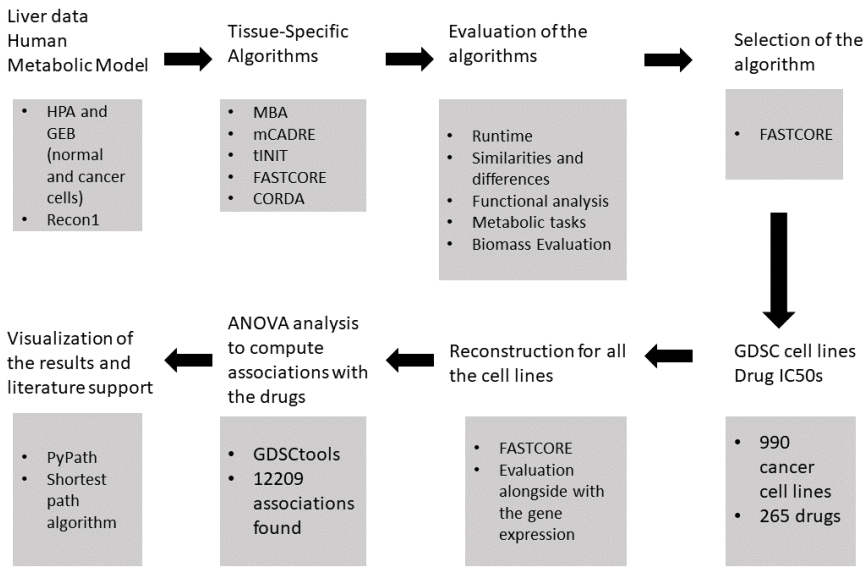


Figure 3.1: Overview of the work. In the top part of the Figure, it is exemplified the evaluation of the tissue-specific algorithms, as in the bottom is illustrated the workflow of the analysis of the drug sensitivity for the reconstructed models

the tissue specific metabolic models throughout the work. The Recon1 is one of the most established and used metabolic models for a human cell. It comprises 3742 reactions, 2766 metabolites, 2004 proteins and 1905 genes.

Two different types of data were used for the first part of the work. The transcriptomics data was extracted from three different samples of HepG2 cell lines from the GSE7307 dataset from Gene Expression Omnibus (GEO), and from the information present in the Gene Expression Barcode (GEB) (McCall et al., 2014) for the hepatocytes from the normal liver tissue data. The data was pre-processed as described by Wang and his colleagues (Wang et al., 2012b).

Regarding the proteomics data, it was retrieved from the Human Protein Atlas (HPA) (Uhlen et al., 2010), containing information about the concentration levels of the proteins. For this work, the version 14 was used for both the HepG2 cell line derived from a hepatocellular carcinoma (Knowles et al., 1980) and hepatocytes from the normal liver tissue data.

For the second one, we used the gene expression from 990 cell lines from GDSC (see section 2.4) that had corresponding IC50 values for 265 different drugs (for the MANOVA analysis, we used the values $1 - AUC$) (Yang et al., 2013).

3.2 Algorithms for the reconstruction of the tissue specific metabolic models

In this study, we reconstructed 20 models comprising all the five algorithms described above, using both conditions and data sources. Due to the fact that we are using the Recon1 metabolic model as the template model for the tissue-specific models, the data used is filtered for the genes present in this model.

All the software tools used and datasets are provided, to allow for full reproducibility of the results, in a software container (using the Docker application). The image and instructions for running it are provided in Docker Hub: `https://hub.docker.com/r/saracorreia/is_cls_hepg2` with the exception of the CORDA one (implemented later on).

For the reconstruction of the MBA models, we generated 50 different models and merged them in a single model (the cut-offs for the creation of the core were "High" and "Medium" for the HPA data and 0.9 and 0.5 for the GEB data).

For both the mCADRE and FASTCORE cores, the cut-offs were either "Medium" or 0.5, for HPA and GEB data, respectively. The cut-offs for the tINIT algorithm were the levels "High", "Medium", "Low" and "Not Detected" were respectively 0.9, 0.5, 0.1 and 0.

Also for the tINIT algorithm, we provided a set of specific metabolic tasks that the cell needs to perform that were given on the original paper for the algorithm (Agren et al., 2014b).

Finally, regarding the CORDA algorithm, the way that "cut-offs" are handled are a bit

different. Here, the algorithm is built around the high confidence reactions, and all the other type (medium, negative or other) can be added to the model, if it is identified as important for the ones present in the reconstruction.

3.3 Analysis of the drug sensitivity for the reconstructed models

3.3.1 Data preprocessing and model generation

For the second part of this work, we decided to evaluate if the reconstruction of tissue models for the GDSC cell lines library could improve the prediction of drug efficiency for the cell lines against the predictions made with gene expression. As shown in the first part of the work, there are several options for model generation with different advantages/disadvantages in terms of time/computation efficiency or more "biological meaning". As later shown in the **Results**, the algorithm that fulfills the most requirements is the FASTCORE one, considering a scenario where many models need to be generated.

Since the data given is "raw", we decided to apply simple statistic procedures to the data. Initially, the genes used were only the ones that were present in the Recon1 (1254 of 1501 present in the model).

The "discretization" of the data, either by gene or cell line (rows or columns of the data matrix), was achieved by computing the "Z-score" or standard score. After calculating the mean and standard deviation of the vector, each element is "scaled" by subtracting the mean and dividing by the standard deviation (Kreyszig, 2006).

Due to the way that the "omics" data are imported, the data was transformed by the following rules: if there were no data for a given gene in a cell line or the Z-score was less than -1, the level would be considered as "Not Detected"; if the Z-score was between -1 and 0, it would be assigned the "Low" level; if higher than 0 and lower than 1, the

"Medium" level would be assigned; finally, if greater than 1, the level would be "High".

Since for this analysis we will also compare to the gene expression alone, the genes that had a level assigned of "Medium" or "High", would get a value of 1; otherwise, the value for that gene is 0.

As mentioned above, the algorithm for the reconstruction of tissue specific models is the FASTCORE. Parameters used for this algorithm were the same as used for the first part, the cut-off being at the "Medium" value. Recon1 was again used as the template model.

After the models for all the cell lines were generated, a matrix was built that had as rows the cell lines and as columns the reactions of Recon1; if a reaction was present in a given model, the value for that cell is 1; otherwise, if the reaction is not in the model the cell would be filled with 0.

3.3.2 ANOVA analysis

We created the four binary matrices, the first one containing gene expression information and discretization by gene, the second discretized by cell line, the third containing information for reactions and the discretization for their reconstruction by gene and the last one discretized by cell line. Letters and numbers will be given to simplify the reading process: G1 - MANOVA analysis with the gene expression normalized by genes; G2 - Same as G1 but the normalization is by cell lines; R1 - MANOVA analysis with the reconstructed models from the gene expression of G1; R2 - same as R1 but the models are reconstructed with the gene expression of G2.

Then, we used the Python package GDSCtools to perform a MANOVA analysis in combination with the IC50 of 265 drugs. The GDSCtools is a free open-source Python library used for testing drug sensitivity on the GDSC panel (Cokelaer et al., 2017). Using this package, we evaluate if the features being tested (in our case, gene expression or reactions from the reconstructed models) can be used as predictors for sensibility of the

drug being analyzed.

In this case, to take into account the multiple factors (tissue origin, microsatellite instability, for example), a more versatile analysis of variance (ANOVA) is implemented. In a simple way, an ANOVA test is performed for each combination of the feature to a given drug.

Since we are dealing with an high number of features, it is needed to perform a error control for multiple testing. For our work, we used the False Discovery Rate (FDR), method to correct the p-values (Lin, 2005). Also, to take into account the variations between the p-values, the effect sizes of the tested statistical interactions are also included, computed by the Cohen models.

When analyzing a set of features, linear models are also applied to the computation. In our work, we used the Ordinary Least Squares (OLS), a method to estimate the unknown parameters in a linear regression model, also used in the pipeline (Cokelaer et al., 2017).

The settings used for this analysis were the default ones, with parameters like regression alpha set to 0.01, with p-value correction method as false discovery rate, and considered significant if less than 25%. Another possible parameter that can be used is effect size where the values tested were 1.1 (there is a drug which is associated to too many features with an effect size near 1.1) and 2.

To compute the associations across all the cell lines and drugs, the ANOVA was done across all features (either gene expression or reactions, but for all the cell lines) and all drugs. The output of the analysis generates a HTML report with several statistics and graphs (e.g. volcano plots) for an easy visualization and interaction with the results. If an association has a FEATURE_delta_MEAN_IC50 below 0, the association found is considered sensible; otherwise, it is resistant.

To be able to compare the information that could be obtained from the reconstruction of the models, we defined five different types of relationships for the comparison between the associations found on the analysis performed on the gene expression and the tissue

models. With this, each type of association defined will be an exact interaction between the feature and the drug. However, since the features of the analysis made with the reconstructed models are the reactions, we decided to track back the genes involved with each reaction Gene-to-Protein Rule (GPR) (Figure 2.1).

With all this in mind, we defined five types of relationships:

- Type 1 - Reaction with a found association, but the genes present in its GPR do not have one.
- Type 2 - Reaction and Gene of its GPR have been associated with the same drug.
- Type 3 - Reaction and Gene of its GPR have been associated with different drug(s).
- Type 4 - Reactions without a GPR (e.g. exchange ones) but with an association to a drug.
- Type 5 - Genes with a found association, but their reaction(s) do not have one.

In addition, features (reactions or genes) who have multiple associations with different drugs can have distinct sensitivity. In these situations, the association is divided into sensitive or resistant. For later queries, the reactants and products of each reaction are included.

The final output of this part of the algorithm is a matrix with 12 columns and the associations as rows, as exemplified in the example_table.csv present in `darwin.diuminho.pt/mscthesis_jmlf`.

3.3.3 Associations between reactions or genes and drugs

With the 2 matrices produced by the last algorithm, we decided to take another approach to try the validation of the associations found. We decided to use the package *PyPath*, a

tool designed to combine the molecular information from multiple online resources. With this, it is possible to generate a network of molecular interactions from the several sources, containing not only proteins, but also RNA, drug compounds and other molecules (Türei et al., 2016).

With this in mind, the objective of using this package is to try to find intermediates that can explain how the drug was associated with the reaction, gene or metabolite.

Since *PyPath* does not work with Recon reactions, we selected the genes associated with them and transcribed them to their respective protein name. As for the drugs, to find how they could impact the pathways of the genes/reactions, we only selected the cases where drug targets were genes (e.g. cases like "DNA replication" as drug target were excluded).

With this, we find the shortest path in the network between a single or set of drugs to a single or set of genes or reactions or metabolites, resulting in an image (.PNG file), where the networks created are represented (Figure 3.2).

Given the input for the search, the networks generated have a specific color code. Nodes that are colored blue are the ones that are related to either the reaction (single or set of genes), metabolite (single or set of reactions and therefore genes) or gene. The green nodes are the "intermediate" genes in between the drug(s) target(s) and the gene(s)/ reaction(s)/ metabolite(s) associated.

Finally, the nodes that are colored as either red, brown or pink are drug targets. Nodes as red mean that the targeted gene is sensitive to the drug found in the association, brown is related to the resistance of the gene to a given drug and pink means that the gene is sensitive and resistant to the given gene(s)/ reaction(s)/ metabolite(s). Arrows shown at red represent inhibition of a gene/protein by another.

An important consideration to take into the following results is that not every gene or drug target are sure to be mapped on the PyPath package, which in the algorithm will lead to an error which will tell that it is not possible to generate the graph with the given

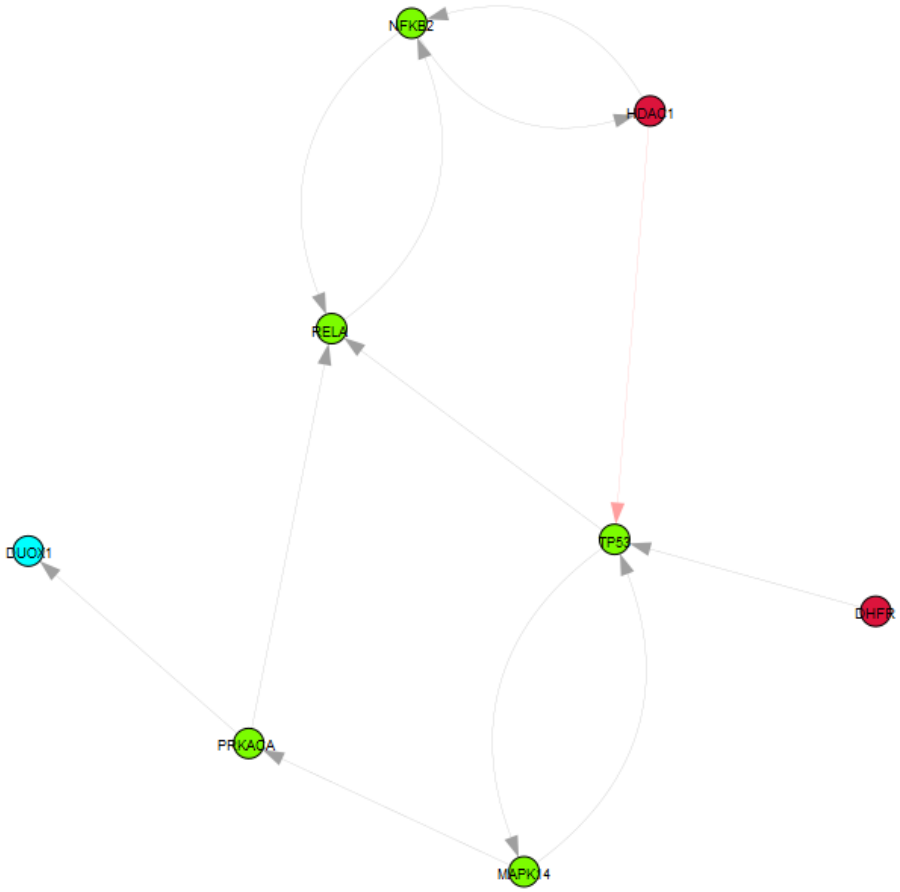


Figure 3.2: Representation of the shortest path between the protein DUOX1 and the drugs that target HDAC1 and DHFR. In the image, we can see that the gene DUOX1 (associated with the synthesis of hydrogen peroxide) has an association found with drugs that target HDAC1 (important in the acetylation of the DNA) and DHFR (catalyzes an essential reaction for de novo glycine and purine synthesis).

input. Also, if the mapping can not be "completed" automatically, it is given the user the possibility of mapping the gene manually (which will be saved for future queries).

With the Table **??** in mind, the algorithm designed for the query can be used in the following ways:

- Type of association - for every query created, we can choose which type of associations we want to select (only if applicable, e.g., if we choose type 4 association, since they are exchange reactions, no genes are involved, so when querying genes, we can not select this type of association).

- Reactions - in this type of query, with a given reaction (from the Recon1), if there is an association found with a drug, the algorithm will pick the related gene(s) and drug target(s) and build the graph; only associations of type 1, 2, 3 and 4 (or all the four) will work.

- Genes - when querying for Genes, the algorithm will select only the genes (from the gene expression) which had an association with the drugs; in this scenario, only type 2, 3 and 5 (or all) associations can be filtered.

- Type of drug - in specific occurrences (like type 3 associations), it is given the user the choice of the drug targets he/she wants to analyze (since in this type of associations, both the associations from the reactions and gene expression have at least one association with a drug); for example, the user may want to evaluate how the drug associated with the reaction has (or not) an association with the gene (from gene expression) and vice-versa.

- Metabolites - for this input, asides from the selection of all the types of associations (since each line contains at least one metabolite), this filter has other options:

  Reactant or Product - with this option, the query for the metabolite will either be performed for one or both.

  Compartment - according to the nomenclature of the model used, this field can be filled with any compartment available (if it is left with an empty string, all the compartments will be taken into account).

Chapter 4

Results

4.1 Analysis of the models for liver cells

Here we are going to present the results and a discussion. First, we will evaluate the algorithms for the tissue-specific reconstruction. This analysis will be made regarding the runtime of them, their number of reactions (normal and cancer cells, and also the shared ones), a clustering to evaluate the distance between them and a functional analysis. Also, we will perform a test to check the number of liver tasks (Gille et al., 2010a) and for the cancer cells see how many precursors the reconstructed models are able to produce, add the missing reactions (to produce the remaining precursors) and assess the production of biomass.

After the evaluation of the algorithms, we reconstructed the GDSC cell lines using FASTCORE. With the gene expression and the reactions from the reconstructed models from the GDSC cell lines, we performed an ANOVA analysis with the IC50s of the drugs administrated for the GDSC cell lines in order to evaluate their sensitivity, using the GDSCtools package. After the associations were calculated, we used the Pypath package to analyze the cellular signalling pathways of the associations found and searched the literature for scientific work that could support some of the findings.

4.1.1 Similarities and Differences in the models

As explained in the previous chapter, we reconstructed 20 models for liver, 10 for normal cells and 10 for cancer cells with the parameterization described. In the previous chapter, one of the most important factors in the algorithms evaluation is the runtime. In this scenario, the one that performed better was the FASTCORE algorithm, normally taking less than 3-5 minutes to be completed. The tINIT one takes about 2-5 more minutes, the mCADRE and the CORDA algorithm takes around one day to be completed. For the MBA, since we have to produce intermediary models, the whole process usually takes more than 1 day.

For a visual understanding of the results, considering the number of reactions in each model, we display Venn Diagrams in Figures 4.1 and 4.2, with the number of reactions that are shared for each set of conditions and different data sources.

Analysing the figures, we can tell that the algorithm that shares the most reactions between the two conditions for both data sources is the tINIT algorithm. This may be due to the nature of the algorithm because, although it can not guarantee that the model is capable of performing all the tasks, if possible, it tries to find a set of essential reactions and ensures that those reactions in the model have flux. For the other algorithms, their percentage of shared reactions is very similar (although the MBA for HPA data has the lowest percentage of shared reactions).

The next step was to execute a hierarchical clustering process of the 20 models. With this method, we will try to identify the relations between conditions, data sources and algorithms.

Taking a closer look to the clustering results (Figure 4.3), we are able to differentiate three separate clusters. The first one includes all the models from the CORDA algorithm and divided by data source. This might happen due to the nature of the algorithm, since it tries to minimize the number of reactions that are needed for the High Confidence (HC) ones to be able to carry flux, thus reactions that are and "support" the HC ones are similar

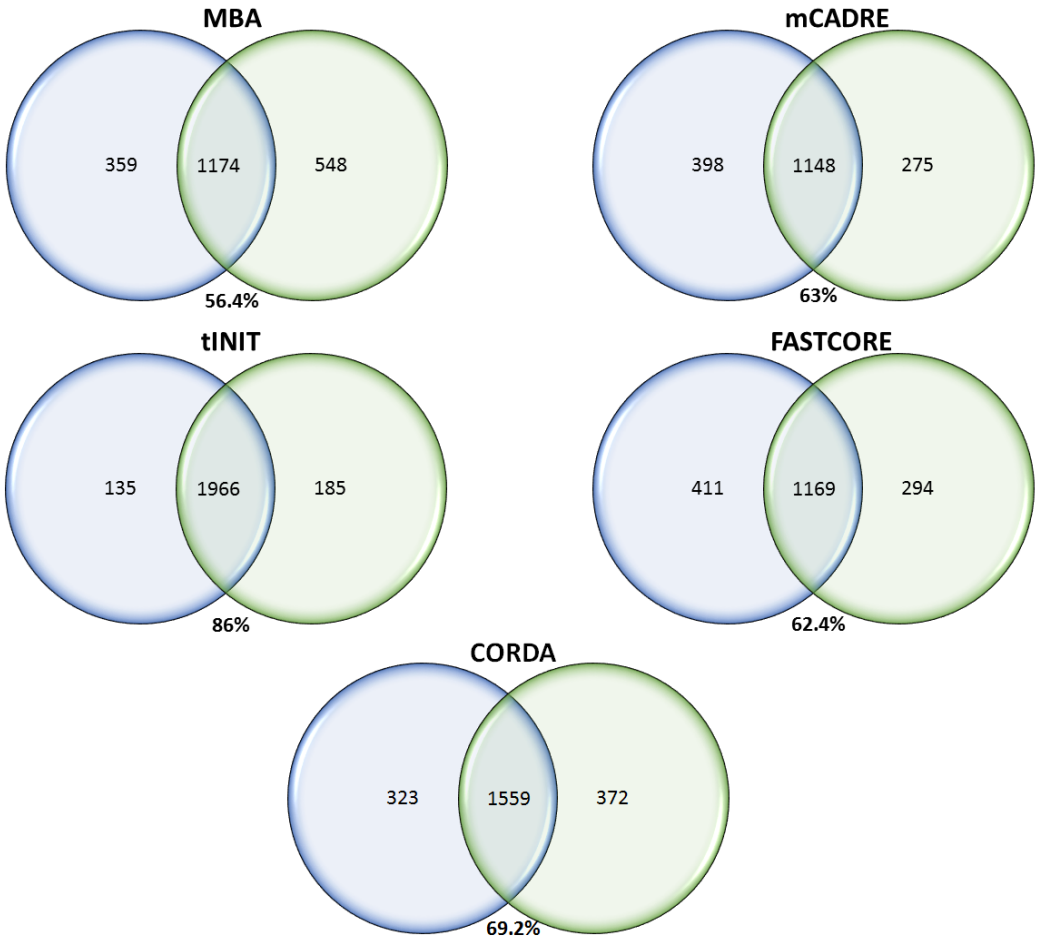


Figure 4.1: Common and exclusive reactions between Normal and Cancer cells from HPA data using MBA, mCADRE, tINIT and FASTCORE algorithms. The values under each Venn diagram represent the percentage of shared reactions for both conditions. The blue one is relative to the Normal model and the green to the Cancer one.

across tissue type.

The second one encompasses the tINIT algorithm. Looking at the sectioning of the four models, we can conclude that there is not a clear separation of any kind. This could be explained since the tasks that are used in the reconstruction of this algorithm can heavily influence the reactions included and may not be suitable, for example, to generate cancer models, requiring a further analysis to find tasks more appropriate.

The other two groups, as one would expect, are clustered by condition, healthy and

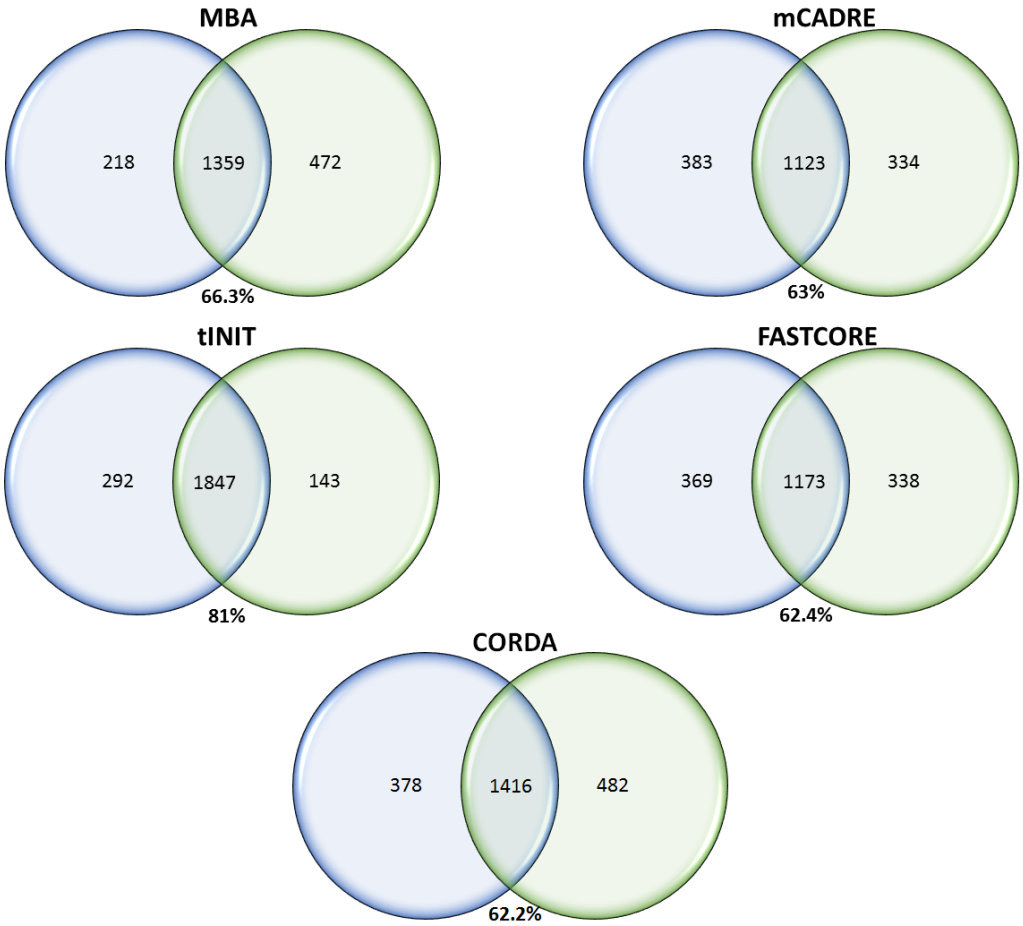


Figure 4.2: Common and exclusive in reactions between Normal and Cancer cells from GEB data using MBA, mCADRE, tINIT and FASTCORE algorithms. The values under each Venn diagram represent the percentage of shared reactions for both conditions.The blue one is relative to the Normal model and the green to the Cancer one.

cancer cells, which shows that there are significant differences between both types of models, regardless of type of data and algorithm. We can further discriminate the groups by data source, since there is a clear separation of the HPA from the GEB ones. Finally, within these sub-clusters, including three models from three algorithms, FASTCORE models are always closer to mCADRE, with MBA further apart. This is expected given the way these different algorithms are designed, as explained in the chapter 2.

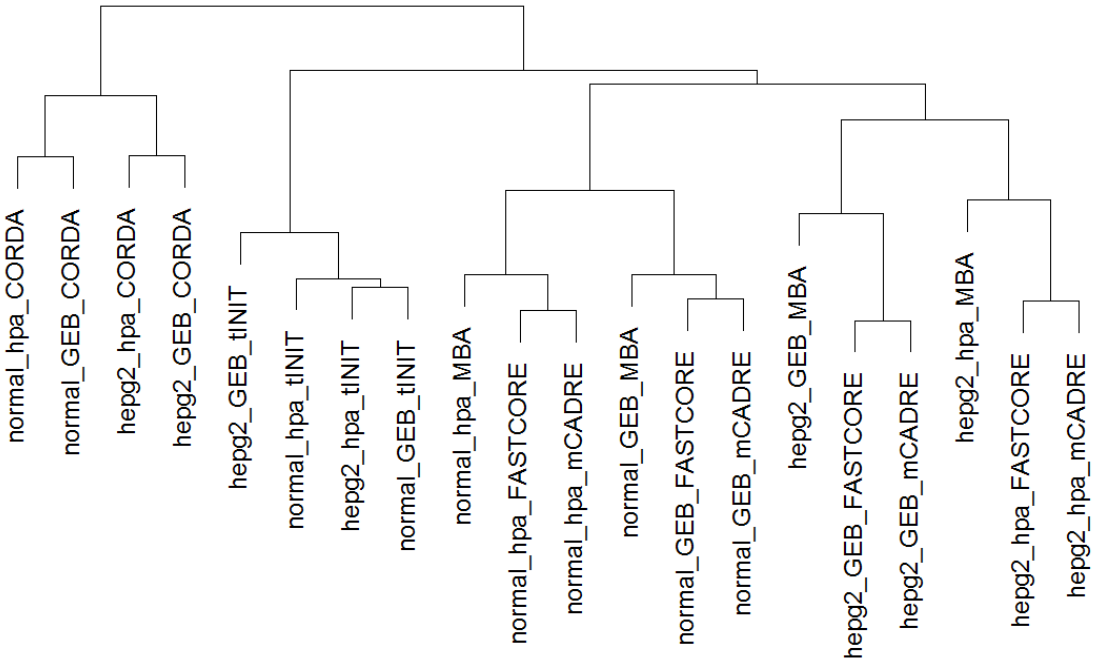


Figure 4.3: Hierarchical Clustering of all the 20 models generated with the method “complete”.

4.1.2 GO analysis

Another test to the models obtained was conducted by performing an enrichment analysis. With the objective of evaluating the processes that are lost and gained by the cancer cells, when compared to the normal ones, a p-value of 0.025 was used while using the *Category* and *GOstats* packages from *Bioconductor*. The analysis was performed by the genes that are different from each model when comparing the same algorithm and data source, but different tissue.

When analyzing the data source from the HPA, the normal MBA models generated “acquired” more biological processes related to the purine metabolism, organophosphate and organonitrogen metabolic processes, while the HepG2 ones favor ion and anion transmembrane transport and also organophosphate and organonitrogen metabolic processes but probably through another pathway.

It has been shown in some works that a shift in the transport of ions has been asso-

ciated with certain hallmarks of cancer (Andersen et al., 2014; Martial, 2016). One of the characteristics in cancer is the alteration of pH which are associated with proton exchanges and water transport, neovascularization with Ca^{2+}, K^+ and Na^+ channels; also, the Na^+, K^+, Ca^+ and Cl^- are associated with the transition to metastasis (see (Andersen et al., 2014)).

Looking at the mCADRE models, the normal model is similar to the ones obtained in the MBA one, while the HepG2 contains even more biological processes and only related to transport, also including hydrogen and inorganic ions one.

Regarding the tINIT ones, the normal model "gained" related to fatty acid metabolism (also unsaturated), mono- and carboxylic and organic acid metabolic processes, though the cancer model has enriched ones related to ATP synthesis, oxidative phosphorylation and nucleotide synthesis. This could be a strange behavior, since usually cancer cells choose glycolysis rather that oxidative phosphorylation in order to produce energy (Hanahan and Weinberg, 2011). It has also been reported that the decrease in the oxidative phosphorylation leads to a decline in the apoptosis process in cancer cells (Chandra et al., 2002; Dey and Moraes, 2000), which is contradictory to the results obtained from the tINIT reconstructed models. However, it has been described that in some cases of acidosis, the cancer cells are able to use oxidative phosphorylation for energy production, thus explaining the ATP synthesis enrichment (Chen et al., 2010; Koivunen et al., 2007; Romero-Garcia et al., 2014).

For the FASTCORE models, processes related to the production of nucleotides, organophosphate and organonitrogen are "acquired" in the normal model, whereas in the HepG2 one has similarities with the tINIT for HepG2 one regarding the ATP synthesis, ions and anions transport and also production of nucleotides, though with different genes/pathways.

In last, for the CORDA ones, both models share the same "enriched" biological processes (but not sharing the same set of genes), such as organonitrogen metabolic processes and carboxylic acid, oxoacid and organic acid metabolic processes. The differences on

the normal one are icosanoid, fatty acid derivative and drug metabolic processes, while the HepG2 one are organophosphate biosynthesis, tricarboxylic acid cycle and sodium transport. Tumor cells have a requirement of an high number of nutrients in order to proliferate and maintain their metabolism (DeBerardinis et al., 2007). Two of the main biological needs of the tumor are the need to produce fatty acids (for lipid biosynthesis) and ribose-5-phosphate (for nucleotide biosynthesis). There are characteristics that both processes share and are important in the context of cancer evolution. Both of them use glucose as carbon source, use intermediates of the Krebs cycle, need NAPDH for its reductive power and anaplerosis (refresh of the intermediates of the Krebs cycle). The increase of fatty acid on the cell (eventually leading to their accumulation) can alter cellular processes such as signal transduction and gene expression, mainly due to cytosolic proteins, enzymes and membrane-targeted proteins, usually leading to an evasion of apoptosis (DeBerardinis et al., 2008). Regarding the nucleotide production, tumor cells use glucose as the main precursor of the ribose-5-phosphate, mainly using the non-oxidative branch of the pentose phosphate pathway (Romero-Garcia et al., 2014).

Looking at the analysis made with GEB data, the MBA models do not differ much from the HPA one. In the normal model, the enriched processes are oxoacid, organic acid and monocarboxylic acid metabolic ones, also with lipid , dicarboxylic acid and amino acid metabolism while the HepG2 ones favor again the trasmembrane transport of ions, anions and cations.

In the mCADRE normal model, processes related with synthesis of ATP and its transport, oxidation-reduction and metabolic processes associated with organonitrogen, carboxylic, oxo- and organic acid are also highlighted. In the HepG2, processes related with transport of protons, hydrogen, iron and proton derived from ATP hydrolysis and phagosome maturation are featured. There has been a report (although it is on breast cancer) which demonstrated that in case of acidification, acidic vesicules (containing phagocytosed extracellular material) could be correlated with the invasive profile of the cancer.

This is also related with the other enriched processes, such as ATP hydrolysis and transport of hydrogen, which could lead to an acidification and facilitate the process (Montcourrier et al., 1994).

Analyzing the tINIT models, the processes are related to the metabolic and biosynthetic ones of carboxylic, oxo-, organic, monocarboxylic, as well as the bile acid biosynthetic process, whereas the HepG2 one is again based on the transport of electrons from NADH, ATP synthesis (also in mitochondria), respiratory chain and oxidative phosphorylation and cellular respiration processes.

For the FASTCORE models, both models share biological processes like small molecule, single-organism and carboxylic acid metabolic ones (again, probably with the use of different genes on the same pathway). For the normal one, the enriched biological processes are carboxylic, organic and oxo- acid and lipid metabolism, single organism biosynthesis and catabolism, anion transport and oxidation-reduction processes. In the HepG2 model, again metabolism of organophosphate and organonitrogen is enriched, as well as cellular lipid and CDP-diacylglycerol metabolic processes.

Lastly, for the CORDA models, both share processes related to small molecule and single organism metabolism. In regard to the normal one, metabolism of several acids like carboxylic, oxo-, organic and alpha-amino one are enriched, and also processes associated with the catabolism of single-organism, small molecule, organic and carboxylic acid. For the cancer model, the biological processes highlighted are biosynthesis (and its metabolic processes) of carbohydrate derivate, organonitrogen and sulfur and biosynthesis of glycosaminoglycan.

4.1.3 Metabolic tasks, precursors and biomass

In another analysis, we decided to evaluate the performance of the models by verifying how many liver-specific metabolic tasks (from (Gille et al., 2010b)) they could complete. From a total of 408 tasks tested, Recon1 can perform 281. Table 4.1 illustrates the per-

centage of tasks that our tissue specific models can perform and the heatmap present in Figure 4.4 shows which subset of the metabolic tasks are performed by each model (some subsets were removed since no models were able to perform any task).

Table 4.1: Percentage of performed tasks by condition and algorithm of the 281 tasks that Recon1 is able to perform.

	MBA	mCADRE	tINIT	FASTCORE	CORDA	**Mean**
Normal_HPA	7.8%	9.6%	87.9%	58.4%	44.1%	**51.4%**
Normal_GEB	55.2%	26.7%	88.6%	71.5%	64.1%	
HepG2_HPA	63%	2.5%	92.5%	40.5%	66.5%	**56.5%**
HepG2_GEB	79.4%	10.3%	76.5%	71.9%	62.3%	

There are several aspects in this analysis. Looking at the Table 4.1, we can see that the algorithm that has an higher percentage of tasks performed is the tINIT and in average the tumor models are able to fulfil around 5% more tasks than the normal ones.

Looking particularly at both models from the HPA data, they mainly differ in two aspects: the normal tissue-specific model is not capable of catabolizing bilirubin and biosynthesizing fatty acids; on the other hand, the cancer model is not able to biosynthesizing creatine. It has been reported that a low level of production of creatine is common in liver cancer patients; Patra and his colleagues hypothesize that low levels of creatine may be associated with a dysfunction of ATP and could be related with cancer (Chen et al., 2009; Patra et al., 2012).

Analyzing at the GEB models, both are not capable of catabolizing bilirubin and transforming fatty acid (which at least the normal tissue should be able to accomplish), the tumor one is not capable of performing detoxification of xenobiotics. Indeed, it has been reported that the way these compounds are metabolized can affect the outcome of the liver cancer (Williams, 1980). However, the tumor model is not able to perform gluconeogenesis and this is different from expected, since one of the treatments applied to this type of cancer is the inhibition of this pathway (Wang et al., 2012a).

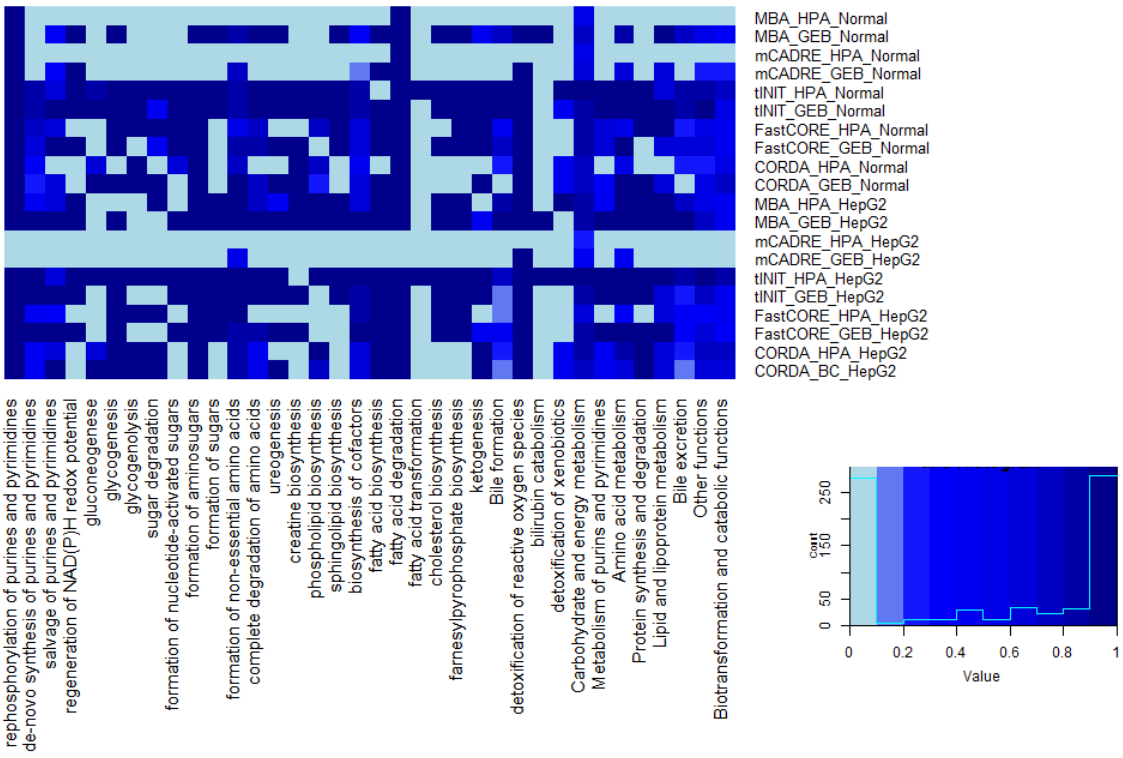


Figure 4.4: Heatmap illustrating the percentage of the subset of metabolic tasks (that can be done by Recon1) performed by each tissue-specific metabolic models.

The algorithm that performed worse was without any doubt the mCADRE one. Even though the best model is the one based on the GEB data for the normal tissue, it is only capable of performing 26.7% of the tasks. The only tasks that all models were able to perform are related to the Carbohydrate and energy metabolism, and not even all of them. With this is mind, this set of models are not of "great help" for this part of the analysis.

Looking at the FASTCORE algorithm, none of them is capable of detoxification of xenobiotics, catabolism of bilirubin, fatty acid transformation or gluconeogenesis. Curiously, both HepG2 models for the FASTCORE algorithm are not able to biosynthesize sphingolipids. In this aspect, they have been controversial in respect to what is their role in cancer, since in some cases they act as tumorigenic and in others as, for example, repress tumor extension (Ségui et al., 2006).

Looking at the MBA models, we can see that the model generated with the HPA data for the normal tissue is the "worst" model of the group. We can also verify that the none of the tissue specific models is capable of performing glycogenolysis or gluconeogenese or even fatty acid transformation. When comparing the models generated from different data, tasks related with the glycogenesis, bilirubin catabolism or ketogenesis were only performed by the GEB models.Finally, when comparing what are the tasks that HepG2 models are capable to perform in comparison to the normal ones, the metabolism of purins and pyrimidines, synthesis and degradation of proteins, both biosynthesis of creatine and phospholipid are only achievable by them.

Finally, analyzing the CORDA models we can see that all of them "fail" in tasks related to regeneration of NAD(P)H, transformation of fatty acids and biosynthesis of cholesterol, sphingolipid and farnesylpyrophosphate. On the other hand, all of them "passed" on the tasks related to rephosphorylation of purines and pyrimidines, glycogenlysis, fatty acid degradation and detoxification of reactive oxygen species (although it may seem contradictory, it seems that cancer may require a balance in reactive oxygen species (Liou and Storz, 2010)). Another interesting result is that both models generated with the HPA data are "not able" to perform ureogenesis, only normal models are able to biosynthesize creatine.

As the final part of the work, we decided to "force" our cancer models to produce biomass. This makes biological sense, since cancer cells evolve to replicate as fast as possible, and, therefore, it is expected that they possess the cellular machinery to obtain all needed precursors.

The Recon1 model does not possess a biomass reaction. We, thus, retrieved this reaction from the Recon2 model (Thiele et al., 2013) and introduced it to the Recon1 model. However, none of our tissue specific models were capable of producing it. The Table 4.2 shows how many biomass precursors each model was able to achieve.

This analysis was achieved by performing an FBA in which the objective function

Table 4.2: Number of precursors that each cancer model had before the inclusion of the new reactions to be able to produce biomass (38 in total).

	MBA	mCADRE	tINIT	FASTCORE	CORDA
HepG2_HPA	29	9	32	23	27
HepG2_GEB	33	14	28	26	24

was the maximization of the excretion of each metabolite, using the RPMI-1640 medium from Folger et al (Folger et al., 2011). tINIT and MBA have the best overall results for the production of the precursors of biomass, with mCADRE being the "least" capable of such task.

Due to this, we decided to add the reactions necessary for each model to fulfill the production of biomass. The following Table 4.3 shows the number of reactions needed to add to each model to be able to produce biomass.

Table 4.3: Number of reactions added to each cancer model to be able to produce biomass.

	MBA	mCADRE	tINIT	FASTCORE
HepG2_HPA	17	28	8	26
HepG2_GEB	9	30	10	16

As expected, the tINIT algorithm models are the ones who need the least number of reactions to be able to produce biomass. MBA reconstructed models need more reactions to be introduced in their models. Again, the mCADRE algorithm showed the highest number of reactions needed. It is also worth noticing that in the general case, the tumor models produce more precursors and need less reactions to be able to produce biomass, which may be biologically plausible.

As the final objective of this work, we decided to perform an FBA to evaluate the differences in the production of biomass for the different cancer models generated (Table 4.4).

Table 4.4: Production of biomass by the cancer models after the integration of the necessary reactions (in mmol.gDW^{-1}.hr^{-1}).

	MBA	mCADRE	tINIT	FASTCORE
HepG2_HPA	0.084	0.003	0.069	0.029
HepG2_GEB	0.084	0.012	0.084	0.084

Since the production of the Recon1 model with the biomass reaction is also 0.084 mmol.gDW^{-1}.hr^{-1}, there are 4 models which can achieve the same amount of biomass production and mCADRE has the lowest overall amount. This shows that the generated cancer models are able to grow at the maximum theoretical level, which is the one defined by the template global GSMM.

Looking at the results, we can make the conclusion that there is no best algorithm, since all of them present advantages and disadvantages. However, since FASTCORE can achieve similar results to the other algorithms (GO analysis, tasks performed) and the time consumption is significantly less, it appears to be a time-efficient and accurate algorithm for the reconstruction of a large collection of gene expression. So, for this reasons, the algorithm will be chosen to perform the second part of the work.

4.2 Reconstructed models for the evaluation of cancer drugs

For the second part of this work, we decided to take an approach to try to find associations between drugs and reaction(s), gene(s) or metabolite(s) (for simplicity, all of them will be defined as features), using ANOVA analysis to infer if a given drug is associated with the features, as well as classifying them either resistant or sensitive to a given drug. As described in the Material and Methods, we have four different matrices (see section 3.3.2) The following Table 4.5 is a summary of the output of the GDSCtools analysis for the

four matrices with no threshold regarding the effect size (the values tested were 1.1 and 2.0).

Table 4.5: Summary of the results obtained through the GDSCtools analysis. The percentage inside parenthesis is related to the total of the feature being analyzed included in that analysis. The number of associations are the ones which their false discovery rate is below 25% and the p-value of the ANOVA is under 0.001.

Data	#Reactions	#Genes	#Drugs	#Associations
G1	-	836 (55.7%)	258 (97.4%)	3169
G2	-	571 (38%)	259 (97.7%)	2334
R1	1121 (30%)	-	239 (90.2%)	3268
R2	1425 (38.1%)	-	138 (52.1%)	3438
Total Reactions	3742	-	-	-
Total Genes	-	1501	-	-
Total Drugs	-	-	265	-
Total Associations found	-	-	-	12209

As stated in the Materials and Methods, the same analysis was performed taking into account different effect size 1.1 and 2. Both Tables 4.6 and 4.7 show the information as shown in Table 4.5.

Looking at the tables, it is clear that the effect size is an important feature regarding the number of associations that are found by the MANOVA analysis. In the total of associations found, there is a decrease of around 93.5% regarding the effect size of 1.1, and around 98% when the value is 2. This occurs due to drugs which are highly associated (Trametinib (targets MEK1 and 2) and Bleomycin (responsible for DNA damage)) and possibly could lead to "less significant" results.

To begin evaluating the impact of the reconstruction of the models on the MANOVA analysis, we will first analyze how the type of the associations can lead to an improvement of the knowledge. We "merged" the analysis from both the ANOVA analysis with the gene expression and the reconstructed models (as explained in **Material and Methods**) to produce two matrices, one for each type of normalization. The following table 4.8

Table 4.6: Summary of the results obtained through the GDSCtools analysis. The percentage inside parenthesis is related to the total of the feature being analyzed included in that analysis. The number of associations are the ones which their false discovery rate is below 25% and the *p*-value of the ANOVA is under 0.001. The effect size is higher than 1.1.

Data	#Reactions	#Genes	#Drugs	#Associations
G1	-	5 (0.3%)	4 (1.5%)	7
G2	-	179 (11.9%)	172 (64.9%)	557
R1	27 (0.7%)	-	20 (7.5%)	37
R2	102 (2.7%)	-	62 (23.4%)	199
Total Reactions	3742	-	-	-
Total Genes	-	1501	-	-
Total Drugs	-	-	265	-
Total Associations found	-	-	-	800

shows the number of each type of associations found, also when taking into account the values 1.1 and 2 for the effect size.

As we can see from the Table 4.8, as we increase the effect size, the percentage of associations of the types from 1 to 4 decreases, in contrast to the type 5 ones (with the exception of the PC). Another interesting fact is that there are only type 4 associations with the normalization by gene and effect size equals to 2. Also, it is important to take notice that the decrease in the total of associations is higher when the normalization is made by gene in comparison to when is made by cell line (from 0 to 1.1 and 2, in the cell line scenario, the decrease is 83.3% and 93.4% where in the normalization by gene is 99% and 99.9%). This is somehow expected, since the increase of the effect size "acts" as a threshold for narrowing down our results.

To try to obtain a better understanding of the biological meaning of these associations, we decided that looking at how possibly the association between the reaction and the drug could happen, by looking at the signaling pathways. To perform this in order to cover all the associations found, we used the Python package **pypath**, which allows us to analyze cellular signaling pathways. Although this analysis could be subjective, we think this is

Table 4.7: Summary of the results obtained through the GDSCtools analysis. The percentage inside parenthesis is related to the total of the feature being analyzed included in that analysis. The number of associations are the ones which their false discovery rate is below 25% and the p-value of the ANOVA is under 0.001. The effect size is higher than 2.

Data	#Reactions	#Genes	#Drugs	#Associations
G1	-	0 (0%)	0 (0%)	0
G2	-	77 (5.1%)	104 (39.2%)	218
R1	3 (0.1%)	-	3 (1.1%)	3
R2	29 (0.8%)	-	18 (6.8%)	46
Total Reactions	3742	-	-	-
Total Genes	-	1501	-	-
Total Drugs	-	-	265	-
Total Associations found	-	-	-	177

Table 4.8: Summary of the number of the associations (and their type) found in the analysis. **PC** stands for the Pancan analysis with the expression normalized by cell line and **PG** for the normalization by genes, **ES** for Effect Size.

Data (ES)	#Type 1	#Type 2	#Type 3	#Type 4	#Type 5	Total
PC (0)	1028 (27.9%)	98 (2.7%)	909 (24.7%)	1101 (30%)	540 (14.7%)	3676
PG (0)	482 (9.7%)	1035 (20.8%)	1126 (22.7%)	1238 (24.9%)	1087 (21.9%)	4968
PC (1.1)	44 (7.2%)	6 (1%)	29 (4.7%)	118 (19.2%)	416 (67.9%)	613
PG (1.1)	5 (10.2%)	6 (12.2%)	1 (2%)	21 (42.9%)	16 (32.7%)	49
PC (2)	15 (6.2%)	2 (0.8%)	4 (1.7%)	30 (12.5%)	190 (78.8%)	241
PG (2)	(%)	(%)	(%)	3 (100%)	(%)	3

the best way to approach the results.

As part of our work, we developed methods to iterate over the results obtained in the previous analysis and build the shortest path between each association. As described in **Materials and Methods**, there are a panoply of parameters (or filters) that can be set to obtain a better understanding of the associations discovered.

The next set of images are going to show some results obtained and how possibly the help of a signaling network can lead us to a better understanding of the mechanisms of action of a drug. In order to obtain a better understanding of how the network could

be deciphered, we decided to take several approaches. Basically, we performed manual queries in Google's Scholar search engine with different keywords, ranging between the gene associated with the reaction to the drug targets and the metabolites involved.

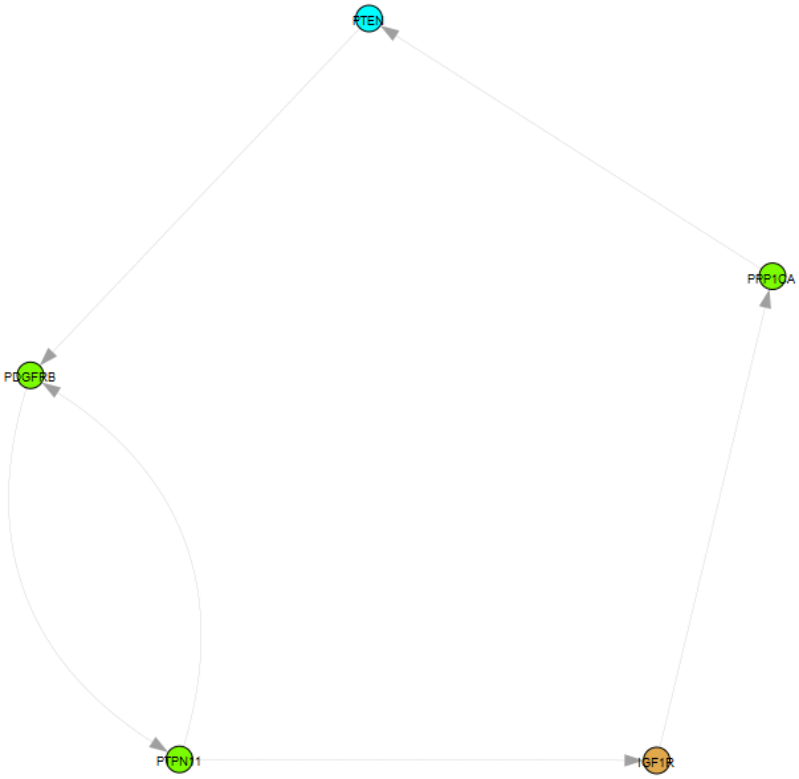


Figure 4.5: Example of a type 2 association. In this case, we can observe that the results obtained from the MANOVA analysis show that IGF1R is associated with the PTEN gene and that some drugs confer both resistance and sensitivity to the IGF1R gene

Looking at Figure 4.5, we can observe an association of type 2 between the PTEN gene and IGF1R (drug target). As Gallardo and his colleagues suggested (Gallardo et al., 2012), both genes can be related in cancer. For us, we consider this case as the drug in question was conferring resistance to gene, and our interpretation of it is that the target was not "affected" by the drug, or at least, it didn't have an impact on the expression of the gene associated.

When looking at Figure 4.6, we can see a more "chaotic" network when compared

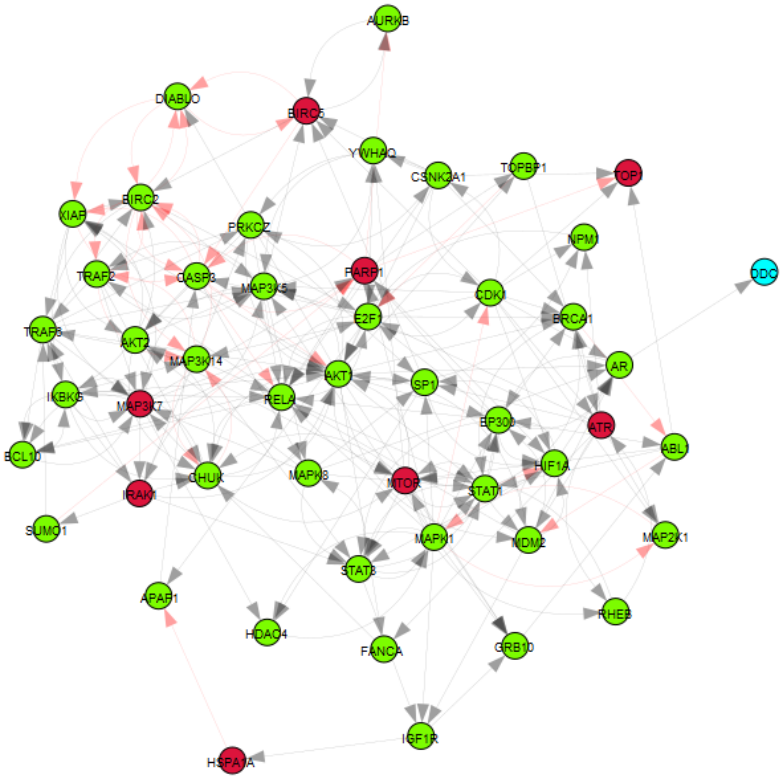


Figure 4.6: Drug targets associated with the reaction that tryptophan to serotonin with the help of the gene DDC. Different that the Figure 4.5, here we only have drug targets that confer sensitivity.

to the previous one. It is harder to analyze (more "intermediate" genes), some of them inhibit others and it is difficult to see how it really affects the gene.

Although this is not optimal and could be eased with text mining with the correct inputs, we decided to tackle the possible combinations (including the metabolites). In a study from Osawa and this colleagues (Osawa et al., 2011), it was found that L-Tryptophan was increasing the levels of lipids in the cell when producing serotonin, which activates mTOR signaling and the latter inhibits autophagy, a mechanism which one of its function is the regulation of the breakdown of stored lipids; and as we can see in the network, there is an association between DDC and mTOR and targeting mTOR leads to sensitivity.

As Soll and his colleagues found (Soll et al., 2010), serotonin promotes tumor growth in Human Hepatocellular Cancer, through the activation of downstream targets of mTOR,

p70S6K and 4E-BP1. In this study, they found that inhibiting a receptor of the Serotonin, HTR2B helped reducing the growth of the cancer. Although we did not find the same receptor, there is a transport reaction mediated by both SLC6A4 or SERT (not present in the network, but present in the tables with all the associations) that had been associated with a drug which targets gamma-secretase and some of the Figure 4.6 and new ones (see Figure 4.7).

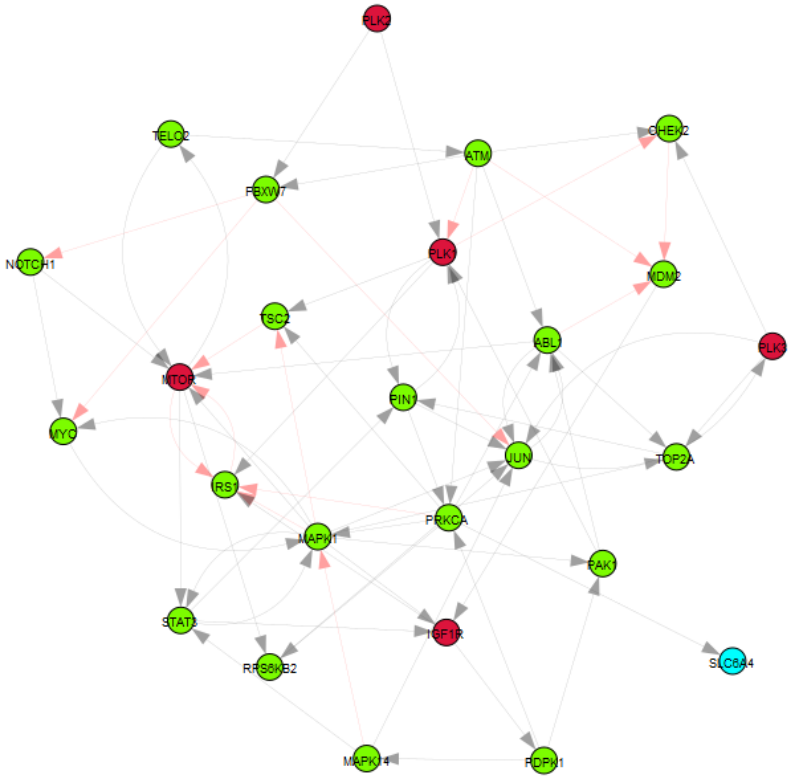


Figure 4.7: Drug targets associated with the reaction transports serotonin with the help of potassium and calcium, with the help of the gene SLC6A4. As in the Figure 4.6, all of the drug targets confer sensitivity.

According to a Gwynne and his colleagues (Gwynne et al., 2017), the serotonergic system usually has an high expression of the SLC6A4 and serotonin is present in human breast cancer, disregarding their molecular and clinical type and also present in lymhphomas and leukemia, protaste cancer, among others (Sarrouilhe et al., 2015).

As we can see from the networks shown, the results obtained from the merge between

the MANOVA analysis (reactions/genes - drugs associations) and the signaling pathway can be possibly used as an useful resource to understand the underlying mechanism of drugs. Although the results should be tested in wet lab, there are some findings that already have literature support, which can also indicate that among all the other associations found, some new interactions can be found.

In the end, this pipeline created shows some interesting results. From this, we understand the value of the addition of another layer of biological knowledge that can increase and maximize the other, as in this case, the reconstruction of the models with the gene expression lead up to new findings.

Chapter 5

Conclusions and Future Perspectives

Metabolism is a crucial part of what makes us human. It is the powerhouse that provides energy and nutrients necessary for the cell to perform its tasks. Looking at the human body, one of the most important organs regarding the metabolism is the liver. Degradation of toxic substances, regulation of the plasma and blood cells are some of its most important features.

Due to the recent growth and evolution of the high throughput technologies (and mainly its cost decrease over the years), it is easier to have more data. With that came the necessity of "adding" informatics with the conventional biology, proving a completely different window time to be able to process all the new information obtained.

Integration of the transcriptomics/proteomics data to provide a better understanding of the metabolism was the initial step for the workflow. Since our current knowledge for liver cancer is reduced, we reconstructed metabolic models using different algorithms for both healthy and cancer cells, from two different data sources. With this, we were able to simulate which are the differences regarding liver cancer. Looking at the algorithms, tINIT was the most successful reconstruction algorithm, sharing more reactions, performing more tasks, needing less reactions to produce biomass. However, this all could be related to the nature of the algorithm and its "proximity" with the template model (Recon1).

Considering this, the FASTCORE one came in second, with similar results and with less computational time (for the reconstruction) and for us, became the better choice for the second part of this work, where a large number of models is needed to be reconstructed.

Cancer is a complex disease, which does not look at age or sex, and has rapidly become one of the most (if not the most) deadliest disease on the world. Due to its heterogeneity, it has been a challenge for the scientific community to fully characterize it, although good progresses have been made. So, with the previous results obtained we reconstructed almost 1000 cell lines of cancer (with gene expression values) with the FASTCORE algorithm and performed a MANOVA analysis (with the GDSCtools Python package) with the administration of 265 different drugs to see if we could gain more knowledge when compared only with the gene expression. With a total of roughly 12.2 thousand associations, about 1.5 thousand were linked directly to the reactions of the models, which leads to believe that this approach could boost the knowledge obtained.

After all the associations were computed, we tried to trace how the drug target and the feature (gene or reaction) interact with each other, with the help of the PyPath Python package. For this, we charted the shortest path between them and see what could be between that could lead to a better understanding of the computed associations. One of the examples found is the possible association that links the serotonin metabolism with several types of cancer.

In this work, we provided a pipeline that could serve as the initial guideline for a new approach for the utility of reconstructed models. Besides the analysis of biological functions, we can also start to use them as a potential additional tool , an "add-on" for the "conventional" genomic analysis. Though, several aspects of the pipeline can be improved, as for example, text mining when trying to link the results obtained with the PyPath or even different algorithms to find the paths between the associations.

While FASTCORE proved to be a interesting choice for the development of the second part of the work, it still has its flaws (as the other ones). The need for new, faster and more

precise reconstruction algorithms, the ability to have more omics input (like metabolomic data) and even if possible, its combinations, should lead to an overall increase of the quality of the models and therefore the knowledge that can be obtained.

Bibliography

R. Agren, S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, and J. Nielsen. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Computational Biology*, 8(5), 2012a. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002518.

R. Agren, S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, and J. Nielsen. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. *PLoS computational biology*, 8(5):e1002518, 2012b.

R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen, and J. Nielsen. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular Systems Biology*, 10(3), 2014a. ISSN 17444292. doi: 10.1002/msb.145122.

R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen, and J. Nielsen. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular systems biology*, 10(3), 2014b.

A. P. Andersen, J. M. A. Moreira, and S. F. Pedersen. Interactions of ion transporters and channels with cancer cell metabolism and the tumour microenvironment. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369(1638):

20130098, 2014. ISSN 1471-2970. doi: 10.1098/rstb.2013.0098. URL http://rstb.royalsocietypublishing.org/content/369/1638/20130098.

F. Azam, S. Mehta, and A. L. Harris. Mechanisms of resistance to antiangiogenesis therapy. *European Journal of Cancer*, 46(8):1323–1332, 2010. ISSN 09598049. doi: 10.1016/j.ejca.2010.02.020.

G. Baffy, E. M. Brunt, and S. H. Caldwell. Hepatocellular carcinoma in non-alcoholic fatty liver disease: An emerging menace, 2012. ISSN 01688278.

G. Bergers and D. Hanahan. Modes of resistance to anti-angiogenic therapy. *Nature Reviews Cancer*, 8(8):592–603, 2008. ISSN 1474-175X. doi: 10.1038/nrc2442. URL http://www.nature.com/doifinder/10.1038/nrc2442.

A. Bordbar, A. M. Feist, R. Usaite-Black, J. Woodcock, B. O. Palsson, and I. Famili. A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC systems biology*, 5(1):180, 2011. ISSN 1752-0509. doi: 10.1186/1752-0509-5-180. URL http://www.biomedcentral.com/1752-0509/5/180.

N. Bouck, V. Stellmach, and S. C. Hsu. How Tumors Become Angiogenic. *Advances in Cancer Research*, 69:135–174, 1996. ISSN 0065230X. doi: 10.1016/S0065-230X(08)60862-3. URL http://www.sciencedirect.com/science/article/pii/S0065230X08608623.

M. E. Cartwright, S. Cohen, J. C. Fleishaker, S. Madani, J. F. McLeod, B. Musser, and S. a. Williams. Proof of concept: a PhRMA position paper with recommendations for best practice. *Clinical pharmacology and therapeutics*, 87(3):278–85, 2010. ISSN 1532-6535. doi: 10.1038/clpt.2009.286. URL http://www.ncbi.nlm.nih.gov/pubmed/20130568.

D. Chandra, J. W. Liu, and D. G. Tang. Early mitochondrial activation and cytochrome c up-regulation during apoptosis. *Journal of Biological Chemistry*, 277(52):50842–50854, 2002. ISSN 00219258. doi: 10.1074/jbc.M207622200.

J. Chen, W. Wang, S. Lv, P. Yin, X. Zhao, X. Lu, F. Zhang, and G. Xu. Metabonomics study of liver cancer based on ultra performance liquid chromatography coupled to mass spectrometry with hilic and rplc separations. *Analytica Chimica Acta*, 650(1): 3–9, 2009.

J. L. Y. Chen, D. Merl, C. W. Peterson, J. Wu, P. Y. Liu, H. Yin, D. M. Muoio, D. E. Ayer, M. West, and J. T. Chi. Lactic acidosis triggers starvation response with paradoxical induction of TXNIP through MondoA. *PLoS Genetics*, 6(9), 2010. ISSN 15537390. doi: 10.1371/journal.pgen.1001093.

T. Cokelaer, E. Chen, F. Iorio, P. Michael, H. Lightfoot, J. Saez-rodriguez, and J. Mathew. GDSCTools for Mining Pharmacogenomic Interactions in Cancer. *bioRxiv*, pages 1–6, 2017. doi: 10.1101/166223. URL `http://www.biorxiv.org/content/early/2017/07/28/166223`.

D. Croft. Building models using reactome pathways as templates. *Methods in Molecular Biology*, 1021:273–283, 2013. ISSN 10643745. doi: 10.1007/978-1-62703-450-0-14.

R. J. DeBerardinis, A. Mancuso, E. Daikhin, I. Nissim, M. Yudkoff, S. Wehrli, and C. B. Thompson. Beyond aerobic glycolysis: Transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proceedings of the National Academy of Sciences*, 104(49):19345–19350, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0709747104. URL `http://www.pnas.org/cgi/doi/10.1073/pnas.0709747104`.

R. J. DeBerardinis, N. Sayed, D. Ditsworth, and C. B. Thompson. Brick by brick: metabolism and tumor cell growth, 2008. ISSN 0959437X.

R. Dey and C. T. Moraes. Lack of oxidative phosphorylation and low mitochondrial membrane potential decrease susceptibility to apoptosis and do not modulate the protective effect of Bcl-x(L) in osteosarcoma cells. *Journal of Biological Chemistry*, 275(10): 7087–7094, 2000. ISSN 00219258. doi: 10.1074/jbc.275.10.7087.

N. Duarte and S. a. Becker. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1777–1782, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0610772104. URL http://www.pnas.org/content/104/6/1777.short.

J. M. Ebos, C. R. Lee, and R. S. Kerbel. Tumor and host-mediated pathways of resistance and disease progression in response to antiangiogenic therapy, 2009. ISSN 10780432.

G. Evan and T. Littlewood. A matter of life and cell death. *Science (New York, N.Y.)*, 281(5381):1317–22, aug 1998. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/9721090.

E. Fahy, S. Subramaniam, R. C. Murphy, M. Nishijima, C. R. Raetz, T. Shimizu, F. Spener, G. van Meer, M. J. Wakelam, and E. A. Dennis. Update of the LIPID MAPS comprehensive classification system for lipids. *J Lipid Res*, 50 Suppl:S9–14, 2009. ISSN 0022-2275. doi: R800095-JLR200[pii] $\backslash$r10.1194/jlr.R800095-JLR200. URL http://www.ncbi.nlm.nih.gov/pubmed/19098281$\delimiter"026E30F$nhttp://www.jlr.org/cgi/reprint/50/Supplement/S9.pdf.

G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman. Bioinformatics challenges for personalized medicine, 2011. ISSN 13674803.

O. Feron. Pyruvate into lactate and back: From the Warburg effect to symbiotic energy fuel exchange in cancer cells, 2009. ISSN 01678140.

R. S. Finn. Development of molecularly targeted therapies in hepatocellular carcinoma: where do we go now? *Clinical cancer research : an official journal of the American Association for Cancer Research*, 16(2):390–7, 2010. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-09-2084. URL http://www.ncbi.nlm.nih.gov/pubmed/20068087.

O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Molecular systems biology*, 7(1), 2011.

A. Gallardo, E. Lerma, D. Escuin, A. Tibau, J. Muñoz, B. Ojeda, A. Barnadas, E. Adrover, L. Sánchez-Tejada, D. Giner, F. Ortiz-Martínez, and G. Peiró. Increased signalling of EGFR and IGF1R, and deregulation of PTEN/PI3K/Akt pathway are related with trastuzumab resistance in HER2 breast carcinomas. *British Journal of Cancer*, 106 (8):1367–1373, 2012. ISSN 0007-0920. doi: 10.1038/bjc.2012.85. URL http://www.nature.com/doifinder/10.1038/bjc.2012.85.

R. A. Gatenby. Mathematical Modeling in Cancer. In *Biomedical Informatics for Cancer Research*, pages 139–147. Springer US, Boston, MA, 2010. doi: 10.1007/978-1-4419-5714-6_7. URL http://link.springer.com/10.1007/978-1-4419-5714-6{_}7.

C. Gille, C. Bölling, A. Hoppe, S. Bulik, S. Hoffmann, K. Hübner, A. Karlstädt, R. Ganeshan, M. König, K. Rother, M. Weidlich, J. Behre, and H.-G. Holzhütter. HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Molecular systems biology*, 6:411, 2010a. ISSN 1744-4292. doi: 10.1038/msb.2010.62. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2964118{&}tool=pmcentrez{&}rendertype=abstract.

C. Gille, C. Bölling, A. Hoppe, S. Bulik, S. Hoffmann, K. Hübner, A. Karlstädt, R. Ganeshan, M. König, K. Rother, et al. Hepatonet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Molecular systems biology*, 6(1), 2010b.

R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, 375(12):1109–1112, 2016. ISSN 0028-4793. doi: 10.1056/NEJMp1607591. URL http://www.nejm.org/doi/10.1056/NEJMp1607591.

W. D. Gwynne, R. M. Hallett, A. Girgis-Gabardo, B. Bojovic, A. Dvorkin-Gheva, C. Aarts, K. Dias, A. Bane, and J. A. Hassell. Serotonergic system antagonists target breast tumor initiating cells and synergize with chemotherapy to shrink human breast tumor xenografts. *Oncotarget*, 8(19):32101–32116, may 2017. ISSN 1949-2553. doi: 10.18632/oncotarget.16646. URL http://www.ncbi.nlm.nih.gov/pubmed/28404880http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5458271.

D. Hanahan and J. Folkman. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis, 1996. ISSN 00928674.

D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000. ISSN 0092-8674. doi: 10.1007/s00262-010-0968-0. URL http://www.ncbi.nlm.nih.gov/pubmed/10647931.

D. Hanahan and R. A. Weinberg. Hallmarks of cancer: The next generation, 2011. ISSN 00928674.

T. Hao, H.-W. Ma, X.-M. Zhao, and I. Goryanin. Compartmentalization of the Edinburgh Human Metabolic Network. *BMC bioinformatics*, 11:

393, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-393. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2918583{&}tool=pmcentrez{&}rendertype=abstract.

L. Hayflick. Mortality and immortality at the cellular level. A review. *Biochemistry. Biokhimiia*, 62(11):1180–90, nov 1997. ISSN 0006-2979. URL http://www.ncbi.nlm.nih.gov/pubmed/9467840.

L. Holmgren, a. Szeles, E. Rajnavölgyi, J. Folkman, G. Klein, I. Ernberg, and K. I. Falk. Horizontal transfer of DNA by the uptake of apoptotic bodies. *Blood*, 93(11):3956–3963, 1999. ISSN 0006-4971. doi: 199993:3956-3963.

P. P. Hsu and D. M. Sabatini. Cancer cell metabolism: Warburg and beyond. *Cell*, 134 (5):703–707, 2008. ISSN 00928674. doi: 10.1016/j.cell.2008.08.021.

L. Jerby, T. Shlomi, and E. Ruppin. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Molecular systems biology*, 6:401, 2010a. ISSN 1744-4292. doi: 10.1038/msb.2010.56. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2964116{&}tool=pmcentrez{&}rendertype=abstract.

L. Jerby, T. Shlomi, and E. Ruppin. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Molecular systems biology*, 6(1), 2010b.

M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1), 2012. ISSN 03051048. doi: 10.1093/nar/gkr988.

K. J. Kauffman, P. Prakash, and J. S. Edwards. Advances in flux balance analysis, 2003. ISSN 09581669.

K. M. Kennedy and M. W. Dewhirst. Tumor metabolism of lactate: the influence and therapeutic potential for MCT and CD147 regulation. *Future Oncology*, 6(1): 127–148, 2010. ISSN 1479-6694. doi: 10.2217/fon.09.145. URL http://www.futuremedicine.com/doi/10.2217/fon.09.145.

H. U. Kim, S. Y. Kim, H. Jeong, T. Y. Kim, J. J. Kim, H. E. Choy, K. Y. Yi, J. H. Rhee, and S. Y. Lee. Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. *Molecular Systems Biology*, 7(1):460–460, 2014. ISSN 1744-4292. doi: 10.1038/msb.2010.115. URL http://msb.embopress.org/content/7/1/460.abstract$\delimiter"026E30F$nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3049409{&}tool=pmcentrez{&}rendertype=abstract.

R. Kim. Cancer Immunoediting: From Immune Surveillance to Immune Escape. In *Cancer Immunotherapy*, pages 9–27. 2007. ISBN 9780123725516. doi: 10.1016/B978-012372551-6/50066-3.

Z. Kmieć. Cooperation of liver cells in health and disease., 2001. ISSN 0301-5556. URL http://www.ncbi.nlm.nih.gov/pubmed/11729749.

B. B. Knowles, C. C. Howe, and D. P. Aden. Human hepatocellular carcinoma cell lines secrete the major plasma proteins and hepatitis B surface antigen. *Science (New York, NY)*, 209(4455):497–499, 1980. ISSN 0036-8075. doi: 10.1126/science.6248960. URL papers3://publication/uuid/163E4872-283A-4E73-B852-DAE720426708.

P. Koivunen, M. Hirsilä, A. M. Remes, I. E. Hassinen, K. I. Kivirikko, and J. Myllyharju. Inhibition of hypoxia-inducible factor (HIF) hydroxylases by citric acid cycle intermediates: Possible links between cell metabolism and stabilization of HIF.

Journal of Biological Chemistry, 282(7):4524–4532, 2007. ISSN 00219258. doi: 10.1074/jbc.M610415200.

E. Kreyszig. Advanced Engineering Mathematics, 2006. ISSN 00255572. URL http://www.jstor.org/stable/3612523?origin=crossref.

B. Lanpher, N. Brunetti-Pierri, and B. Lee. Inborn errors of metabolism: the flux from Mendelian to complex diseases. *Nature reviews. Genetics*, 7(6):449–460, 2006. ISSN 1471-0056. doi: 10.1038/nrg1880. URL http://dx.doi.org/10.1038/nrg1880.

M. Laplante and D. M. Sabatini. mTOR signaling in growth control and disease. *Cell*, 149 (2):274–293, 2013. ISSN 1097-4172. doi: 10.1016/j.cell.2012.03.017.mTOR. URL http://dx.doi.org/10.1016/j.cell.2012.03.017.

C. Lengauer, K. W. Kinzler, and B. Vogelstein. Genetic instabilities in human cancers. *Nature*, 396(6712):643–649, 1998. ISSN 0028-0836. doi: 10.1038/25292.

A. J. Levine. p53, the cellular gatekeeper for growth and division, 1997. ISSN 00928674.

D. Y. Lin. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21(6):781–787, 2005. ISSN 13674803. doi: 10.1093/bioinformatics/bti053.

G.-Y. Liou and P. Storz. Reactive oxygen species in cancer. *Free radical research*, 44 (5):479–96, may 2010. ISSN 1029-2470. doi: 10.3109/10715761003667554. URL http://www.ncbi.nlm.nih.gov/pubmed/20370557http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3880197.

A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, I. Nookaew, P. Jacobson, A. J. Walley, P. Froguel, L. M. Carlsson, M. Uhlen, and J. Nielsen. Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Molec-*

ular systems biology, 9:649, 2013a. ISSN 1744-4292. doi: 10.1038/msb.2013.5. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3619940{&}tool=pmcentrez{&}rendertype=abstract.

A. Mardinoglu, F. Gatto, and J. Nielsen. Genome-scale modeling of human metabolism - a systems biology approach, 2013b. ISSN 18606768.

A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, M. Uhlen, and J. Nielsen. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nature communications*, 5(May 2013):3083, 2014. ISSN 2041-1723. doi: 10.1038/ncomms4083. URL http://www.ncbi.nlm.nih.gov/pubmed/24419221.

S. Martial. Involvement of ion channels and transporters in carcinoma angiogenesis and metastasis. *American Journal of Physiology - Cell Physiology*, page ajpcell.00218.2015, 2016. ISSN 0363-6143. doi: 10.1152/ajpcell.00218.2015. URL http://ajpcell.physiology.org/lookup/doi/10.1152/ajpcell.00218.2015.

M. N. McCall, H. A. Jaffee, S. J. Zelisko, N. Sinha, G. Hooiveld, R. A. Irizarry, and M. J. Zilliox. The Gene Expression Barcode 3.0: Improved data processing and mining tools. *Nucleic Acids Research*, 42(D1), 2014. ISSN 03051048. doi: 10.1093/nar/gkt1204.

P. Montcourrier, P. H. Mangeat, C. Valembois, G. Salazar, a. Sahuquet, C. Duperray, and H. Rochefort. Characterization of very acidic phagosomes in breast cancer cells and their association with invasion. *Journal of cell science*, 107 (Pt 9:2381–91, 1994. ISSN 0021-9533. URL http://www.ncbi.nlm.nih.gov/pubmed/7844158.

C. Muñoz-Pinedo, N. El Mjiyad, and J.-E. Ricci. Cancer metabolism: current perspectives and future directions. *Cell death & disease*, 3:e248, 2012. ISSN 2041-4889. doi: 10.1038/cddis.2011.123. URL http:

//www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3270265{&}tool=pmcentrez{&}rendertype=abstract.

B. A. Neuschwander-Tetri and S. H. Caldwell. Nonalcoholic steatohepatitis: Summary of an AASLD Single Topic Conference. In *Hepatology*, volume 37, pages 1202–1219, 2003. ISBN 0270-9139 (Print)\n0270-9139 (Linking). doi: 10.1053/jhep.2003.50193.

M. Ochs, J. Casagrande, and R. Davuluri. *Biomedical informatics for cancer research*. Springer, 2010.

J. D. Orth, I. Thiele, and B. Ø. Palsson. What is flux balance analysis? *Nat Biotechnol*, 28(3):245–248, 2010. ISSN 1546-1696. doi: 10.1038/nbt.1614.What.

Y. Osawa, H. Kanamori, E. Seki, M. Hoshi, H. Ohtaki, Y. Yasuda, H. Ito, A. Suetsugu, M. Nagaki, H. Moriwaki, K. Saito, and M. Seishima. L-tryptophan-mediated enhancement of susceptibility to nonalcoholic fatty liver disease is dependent on the mammalian target of rapamycin. *The Journal of biological chemistry*, 286 (40):34800–8, oct 2011. ISSN 1083-351X. doi: 10.1074/jbc.M111.235473. URL http://www.ncbi.nlm.nih.gov/pubmed/21841000http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3186417.

J. Park, H. Park, W. Kim, H. Kim, T. Kim, and S. Lee. Flux variability scanning based on enforced objective flux for identifying gene amplification targets. *BMC Systems Biology*, 6(1):106, 2012. ISSN 1752-0509. doi: 10.1186/1752-0509-6-106. URL BMCSystemsBiology.

S. Patra, A. Ghosh, S. S. Roy, S. Bera, M. Das, D. Talukdar, S. Ray, T. Wallimann, and M. Ray. A short review on creatine-creatine kinase system in relation to cancer and some experimental results on creatine as adjuvant in cancer therapy, 2012. ISSN 09394451.

S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 2010. ISSN 1474-1776. doi: 10.1038/nrd3078. URL http://www.nature.com/doifinder/10.1038/nrd3078.

V. R. Potter. The biochemical approach to the cancer problem. *Federation proceedings*, 17(2):691–7, 1958. ISSN 0014-9446. URL http://www.ncbi.nlm.nih.gov/pubmed/13562198.

P. Romero, J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, and P. D. Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome biology*, 6(1):R2, 2005. ISSN 1474-760X. doi: 10.1186/gb-2004-6-1-r2. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=549063{&}tool=pmcentrez{&}rendertype=abstract.

S. Romero-Garcia, J. S. Lopez-Gonzalez, J. L. B´ez-Viveros, D. Aguilar-Cazares, and H. Prado-Garcia. Tumor cell metabolism. *Cancer Biology & Therapy*, 12(11):939–948, 2014. ISSN 1538-4047. doi: 10.4161/cbt.12.11.18140. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3280912{&}tool=pmcentrez{&}rendertype=abstract.

J. Y. Ryu, H. U. Kim, and S. Y. Lee. Reconstruction of genome-scale human metabolic models using omics data. *Integr. Biol.*, 7(8):859–868, 2015. ISSN 1757-9694. doi: 10.1039/c5ib00002e. URL http://xlink.rsc.org/?DOI=C5IB00002E$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/25730289.

S. Sahoo and I. Thiele. Predicting the impact of diet and enzymopathies on human small

intestinal epithelial cells. *Hum Mol Genet*, 22(13):2705–2722, 2013. doi: 10.1093/hmg/ddt119. URL http://www.ncbi.nlm.nih.gov/pubmed/23492669.

S. Sahoo, L. Franzson, J. J. Jonsson, and I. Thiele. A compendium of inborn errors of metabolism mapped onto the human metabolic network. *Molecular BioSystems*, 8(10):2545, 2012. ISSN 1742-206X. doi: 10.1039/c2mb25075f. URL http://xlink.rsc.org/?DOI=c2mb25075f.

D. Sarrouilhe, J. Clarhaut, N. Defamie, and M. Mesnil. Serotonin and cancer: what is the link? *Current molecular medicine*, 15(1):62–77, 2015. ISSN 1875-5666. URL http://www.ncbi.nlm.nih.gov/pubmed/25601469.

A. Schultz and A. A. Qutub. Reconstruction of Tissue-Specific Metabolic Networks Using CORDA. *PLoS Computational Biology*, 12(3), 2016. ISSN 15537358. doi: 10.1371/journal.pcbi.1004808.

S. Seely. Possible reasons for the high resistance of muscle to cancer. *Medical hypotheses*, 6(2):133–7, feb 1980. ISSN 0306-9877. URL http://www.ncbi.nlm.nih.gov/pubmed/7393016.

B. Ségui, N. Andrieu-Abadie, J.-P. Jaffrézou, H. Benoist, and T. Levade. Sphingolipids as modulators of cancer cell death: potential therapeutic targets. *Biochimica et biophysica acta*, 1758(12):2104–20, 2006. ISSN 0006-3002. doi: 10.1016/j.bbamem.2006.05.024. URL http://www.sciencedirect.com/science/article/pii/S0005273606002094.

G. L. Semenza. Tumor metabolism: Cancer cells give and take lactate, 2008. ISSN 00219738.

S. Sengupta, T. R. Peterson, M. Laplante, S. Oh, and D. M. Sabatini. mTORC1 controls fasting-induced ketogenesis and its modulation by ageing. *Nature*,

468(7327):1100–1104, 2010. ISSN 0028-0836. doi: 10.1038/nature09584. URL http://www.ncbi.nlm.nih.gov/pubmed/21179166$\delimiter"026E30F$nhttp://www.nature.com/nature/journal/v468/n7327/pdf/nature09584.pdf.

M. J. Smyth, G. P. Dunn, and R. D. Schreiber. Cancer Immunosurveillance and Immunoediting: The Roles of Immunity in Suppressing Tumor Development and Shaping Tumor Immunogenicity, 2006. ISSN 00652776.

C. Soll, J. H. Jang, M.-O. Riener, W. Moritz, P. J. Wild, R. Graf, and P.-A. Clavien. Serotonin promotes tumor growth in human hepatocellular cancer. *Hepatology*, 51 (4):1244–1254, apr 2010. ISSN 02709139. doi: 10.1002/hep.23441. URL http://doi.wiley.com/10.1002/hep.23441.

M. Sporn. The war on cancer. *The Lancet*, 347(9012):1377–1381, 1996. ISSN 01406736. doi: 10.1016/S0140-6736(96)91015-6.

M. W. L. Teng, J. B. Swann, C. M. Koebel, R. D. Schreiber, and M. J. Smyth. Immune-mediated dormancy: an equilibrium with cancer. *Journal of Leukocyte Biology*, 84(4): 988–993, 2008. ISSN 0741-5400. doi: 10.1189/jlb.1107774. URL http://www.jleukbio.org/cgi/doi/10.1189/jlb.1107774.

I. Thiele, N. Swainston, R. M. T. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bölling, S. Bordel, A. K. Chavali, P. Dobson, W. B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novère, N. Malys, A. Mazein, J. A. Papin, N. D. Price, E. Selkov, M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. G. M. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes, and B. Ø. Palsson. A community-driven global reconstruction of human metabolism. *Nature*

biotechnology, 31(5):419–425, 2013. ISSN 1546-1696. doi: 10.1038/nbt.2488. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed{&}id=23455439{&}retmode=ref{&}cmd=prlinks.

G. J. Tortora and B. Derrickson. *Principles of Anatomy and Physiology*, volume 9406. 2014. ISBN 9780470394953. doi: 10.1016/S0031-9406(05)60992-3.

D. Türei, T. Korcsmáros, and J. Saez-Rodriguez. Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods*, 13(12):966–967, 2016.

M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Björling, and F. Ponten. Towards a knowledge-based Human Protein Atlas. *Nature biotechnology*, 28(12):1248–1250, 2010. ISSN 1087-0156. doi: 10.1038/nbt1210-1248.

M. G. Vander Heiden, L. C. Cantley, and C. B. Thompson. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science (New York, N.Y.)*, 324 (5930):1029–33, 2009. ISSN 1095-9203. doi: 10.1126/science.1160809.

L. Varemo, I. Nookaew, and J. Nielsen. Novel insights into obesity and diabetes through genome-scale metabolic modeling. *Frontiers in Physiology*, 4 APR, 2013. ISSN 1664042X. doi: 10.3389/fphys.2013.00092.

N. Vlassis, M. P. Pacheco, and T. Sauter. Fast reconstruction of compact context-specific metabolic network models. *PLoS Comput Biol*, 10(1), 2014.

B. Wang, S.-H. Hsu, W. Frankel, K. Ghoshal, and S. T. Jacob. Stat3-mediated activation of microrna-23a suppresses gluconeogenesis in hepatocellular carcinoma by down-regulating glucose-6-phosphatase and peroxisome proliferator-activated receptor gamma, coactivator 1 alpha. *Hepatology*, 56(1):186–197, 2012a.

Y. Wang, J. a. Eddy, and N. D. Price. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC systems biology*, 6: 153, 2012b. ISSN 1752-0509. doi: 10.1186/1752-0509-6-153. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3576361{&}tool=pmcentrez{&}rendertype=abstract.

Y. Wang, J. A. Eddy, and N. D. Price. Reconstruction of genome-scale metabolic models for 126 human tissues using mcadre. *BMC systems biology*, 6(1):153, 2012c.

Z. Wang and T. S. Deisboeck. Mathematical modeling in cancer drug discovery, 2014. ISSN 18785832.

O. Warburg. Origin of cancer cells. *Oncologia*, 9(2):75–83, 1956. ISSN 0959-8138. doi: 10.1136/bmj.1.4082.694-a. URL http://www.nature.com/doifinder/10.1038/nature09781.

O. Warburg, F. Wind, and E. Negelein. THE METABOLISM OF TUMORS IN THE BODY. *The Journal of general physiology*, 8(6):519–30, mar 1927. ISSN 0022-1295. URL http://www.ncbi.nlm.nih.gov/pubmed/19872213http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2140820.

S. Weinhouse, O. WARBURG, D. BURK, and A. L. SCHADE. On Respiratory Impairment in Cancer Cells. *Science*, 124(3215):267–272, 1956. ISSN 0036-8075. doi: 10.1126/science.124.3215.267. URL http://www.sciencemag.org/cgi/doi/10.1126/science.124.3215.267.

G. M. Williams. The pathogenesis of rat liver cancer caused by chemical carcinogens. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 605(2):167–189, 1980.

W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber,

M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett. Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1), 2013. ISSN 03051048. doi: 10.1093/nar/gks1111.

J. L. Yecies, H. H. Zhang, S. Menon, S. Liu, D. Yecies, A. I. Lipovsky, C. Gorgun, D. J. Kwiatkowski, G. S. Hotamisligil, C. H. Lee, and B. D. Manning. Akt stimulates hepatic SREBP1c and lipogenesis through parallel mTORC1-dependent and independent pathways. *Cell Metabolism*, 14(1):21–32, 2011. ISSN 15504131. doi: 10.1016/j.cmet.2011.06.002.