SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Biotechnology and Biosciences

PhD program in Biology and Biotechnology, Cycle XXXI

Curriculum in Systems Biology

NEW CONSTRAINT-BASED APPROACHES TO TACKLE THE MULTIPLE SIDES OF CELL METABOLIC PLASTICITY AND HETEROGENEITY

Surname DI FILIPPO Name MARZIA

Registration number 810981

Tutor: Dr. Chiara DAMIANI

Supervisor: Prof. Dario PESCINI

Coordinator: Prof. Paola BRANDUARDI

ACADEMIC YEAR 2017/2018

Acknowledgements

I would like to thank all the people that helped and supported me during these three long years.

First of all, I would like to express my sincere gratitude to my supervisor, Professor Dario Pescini, for his patient guidance, his constant support, and all precious academic opportunities he gave me during all these three years.

My sincere thanks also goes to Professor Lilia Alberghina for giving me many golden opportunities for growth.

A special thanks is also due to Dott. Chiara Damiani and Dott. Riccardo Colombo for their very unceasing and precious support.

I am also grateful to Professor Paola Branduardi, Professor Marco Vanoni and Professor Giancarlo Mauri that during these three Ph.D. years, despite all their academic commitments, always showed their willingness.

Outside the academic field, I would like to dedicate this thesis work to my husband, Davide, who constantly supported and encouraged me, without never stop to believe in me.

I would also like to dedicate this thesis work to my parents because it is also thanks to them if I am here today. I will never stop thanking you for giving me the opportunity to accomplish this great achievement.

And last but not least, a special thanks goes to my two sisters, Susanna and Marika, that are my eternal adventuring companions supporting me in any situations.

Abstract

Plasticity, heterogeneity and modelling approach constitute the three pillars on the top of which this thesis investigates the complexity of cell metabolism. The multiple sides of metabolic plasticity have been explored as cell adaptive response to varying conditions, demand and perturbations under both physiological and pathological conditions. By investigating cell populations as homogeneous and heterogeneous systems, new *in silico* predictive models and novel computational constraint-based methodologies have been defined.

This work started from the investigation of cell populations as homogeneous systems, where the average behaviour is described and cell-to-cell differences are temporarily hidden. Reconstructing high-quality genome-scale metabolic models is crucial to computationally address cell metabolism and organize all the available metabolic knowledge of given cells or organisms. Although multiple tools for performing this task already exist, a pipeline for the semi-automatic reconstruction of genome-scale networks has been proposed to solve some current critical issues and generate higher quality models. The application of this approach for the genome-wide metabolic reconstruction of yeast *Zygosaccharomyces parabailii* showed adherence of *in silico* simulations to experimental data and literature findings. Moreover, metabolic plasticity in response to different metabolic regimes has been explored through constraint-based modelling.

The potentialities of genome-scale reconstructions in mirroring the systemic perspective coexist with difficulty in their management. In this work, greater control is achieved by switching to smaller-scale core networks. In particular, core modelling has been exploited as an effective mean to investigate inter-tumoural heterogeneity, and plasticity of the implemented tumour metabolic programs as adaptation to different environmental scenarios.

The effectiveness of homogeneous systems to lower overall system complexity level without compromising biological validity of *in silico* outcomes goes along with the need to address cell-to-cell variations of cell populations. In this regard, classic constraint-based modelling has been extended to deal with heterogeneous systems. A new strategy, called popFBA, has been developed to reconstruct and simulate cell populations metabolism, by putting emphasis on the relationships established among their components. Using as case study the ecosystemic view of cancer populations, popFBA highlighted that the achievement of optimal biomass is consistent with metabolic plasticity of population components under different scenarios together with a cooperative behaviour. At the same time, countless combinations of flux distributions for the individual population components prompted to develop a novel methodology called

single-cell Flux Balance Analysis (scFBA). This metodology integrates single-cell transcriptomics data as further constraints on the individual components through the computation for each reaction of a Reaction Activity Score, which we implemented in a previous computational framework called MaREA. In this way, scFBA efficiently reduced the amount of allowable individual flux distributions, and captured complex networks of interactions between cells of a specific population.

In view of the findings of this research, a deep characterization of metabolic plasticity within cell populations and of the intricate dialogue between cells and their environment can assist the formulation of more rational and personalized strategies. Their devising could enable to hamper disease progression, or to exploit metabolism of given microorganisms for producing relevant chemical compounds.

Contents

Chapter 1

Aim of the thesis

The aim of this thesis work is to investigate cellular metabolism to gain new knowledge about metabolic programs adopted by given cells or organisms.

Metabolism constitutes the closest level of investigation to cell phenotype [1]. Its investigation represents a very good way to join a given genotype to a specific phenotype. Therefore, global metabolic profiling of a cell provides a complete functional readout of cellular state, by connecting its genome to a particular phenotype, keeping into account environmental influences.

Among the existing omics technologies, metabolomics deals with the identification and quantification of metabolome, which includes the entire set of metabolites that are chemically transformed during cell metabolism. The metabolome of a cell rapidly changes in response to various genetic and environmental stimuli. Consequently, any deviation from cell physiological state may indicate possible pathological conditions. Characterization of cell metabolic profiling through metabolomics has become a powerful and widely exploited approach in clinical diagnostics. In particular, metabolomics can be exploited for multiple applications, including the identification of signatures associated to the onset of diseases, drug treatments discovery and development, and personalized medicine [1, 2].

Although knowing cell metabolome is useful, it is not enough to predict cell phenotype. In this regard, metabolites need to be investigated through a more complete and enriched cell view within the context of biochemical metabolic pathways. Cell metabolism relies on a multitude of reactions, whose nonlinearity and high interconnectivity features imply that a simple one-to-one mapping between each single gene activity and a specific phenotype fails to

capture the complexity characterizing this system. On the contrary, cell phenotypes generally result from the interactions between the products of multiple genes. Consequently, every perturbation on a cell does not provoke an effect on a circumscribed area, but on a global level. Indeed, many enzymes have to alter their function for regaining the overall cell homeostasis [3]. Revealing the complete set of interactions contributing to the structure and function of a living cell has become a key challenge for biology in the twenty-first century [4]. The advent of high throughput technologies, *in primis*, moved the focus away from its individual components, paving the way towards the achievement of a system-level knowledge of cell [4, 5]. Nevertheless, despite the zooming extension to all cellular constituents, a set of entities without any interaction is just a mere aggregation of elements that cannot be actually viewed as a system.

A greater attention to the structure and dynamics of the entire set of cellular interactions has been provided with the coming of systems biology. This multidisciplinary approach aims at developing a system-level understanding of complex biological systems. Its ultimate goal is to unravel how all system components together with the existing relationships affect the characteristics and functions of the system itself, i.e. its phenotype [6]. Systems biology offers an excellent way to describe metabolic profiling of a cell through different *in silico* models and simulation approaches. Thanks to its systemic perspective, systems biology goes beyond classical reductionist approach, where system understanding is linked to the separate analysis of its individual components. Furthermore, systems biology also moves away from standard -omics data analysis, which can be considered a revisited version of the original reductionist paradigm, shifting to a system-level understanding of complex systems [4, 7, 8].

Under a systemic perspective, cell homeostasis is strictly linked to the concept of metabolic plasticity. Metabolic plasticity refers to the physiological high intrinsic ability of cells to respond or adapt to changing environmental conditions, to variation in metabolic or energy demand, as well as to different other types of perturbations [9]. Metabolic pathways undergo a complex intracellular regulatory system acting on the synthesis, degradation, or activity of enzymatic proteins [2]. Overall, this implies that strongly regulated metabolic adjustments allow an adaptative response of the organism in order to maintain homeostasis.

Because of metabolic plasticity, populations of isogenic cells located in the same environment may differentiate in terms of metabolic proteins expression level, by generating a heterogeneous phenotypic composition within the population [10]. Cell-to-cell variations that are always present in any population of cells may contribute to promote the investigation of phenotypic heterogeneity as informative readout of population physiology. Moreover, population hetero-

geneity is also intended as a sort of "escape way" to different perturbations, microenvironmental properties, and changing conditions. Consequently, the ensemble behaviour of a population may not represent that of any single cell. This occurs when individual subpopulations cannot be physically isolated because of their functional interactions as unequal contributors to disease progression or response to drug treatments [11]. Indeed, it is worth noticing that metabolic plasticity can be also negatively exploited when, following specific mutations, alteration of cell metabolic state form the basis of many human diseases, including cancer, obesity, and neurodegenerative disease [12, 13, 14].

A deep characterization of metabolic plasticity within cell populations could have positive impacts in the formulation of more rational strategies for counteractive disease progression. Moreover, rational strategies devising could also favour cell adaptation to different scenarios, especially when the metabolism of given microorganisms is exploited for producing relevant chemical compounds.

The systemic approach constitutes a useful mean to tackle cell metabolism. Computational investigation of cell metabolism founded on constraint-based modelling represents, so far, the most common practice. This approach has been originally developed to be applied in the field of metabolic engineering for the optimization of microbial strains or industrially relevant compounds. More recent is its application to acquire new knowledge about physiological metabolic traits of specific target organisms. Although metabolic reconstructions have been produced for multiple organisms, they do not exist for all the known ones. This aspect contributes to limit the investigation of populations that consist of metabolically unknown organisms. In addition, classic constraint-based modelling is limited to the simulation of a single metabolic model. Consequently, this approach is confined to the metabolic characterization of a single cell that is representative of the population to which it belongs. At the same time, the above described multiple sides of metabolic plasticity that may be present within cell populations increasing the overall complexity degree of the system are hidden.

Plasticity, heterogeneity and modelling approach constitute the three dimensions investigated in this thesis work, as shown in Figure 1.1, to deal with complexity of cell metabolism and to overcome critical issues of current approaches. To address this challenge, new *in silico* predictive models will be defined and novel implemented computational methodologies will be introduced by exploiting the potentialities of constraint-based modelling.

In particular, the thesis is organized as follows. In Chapter 2, I will recall cell metabolism, introduce the features associated to its reprogramming within cancer cells for supporting their growth, and I will discuss about heterogeneity of tumour populations. After that, I will discuss the potential of the systemic

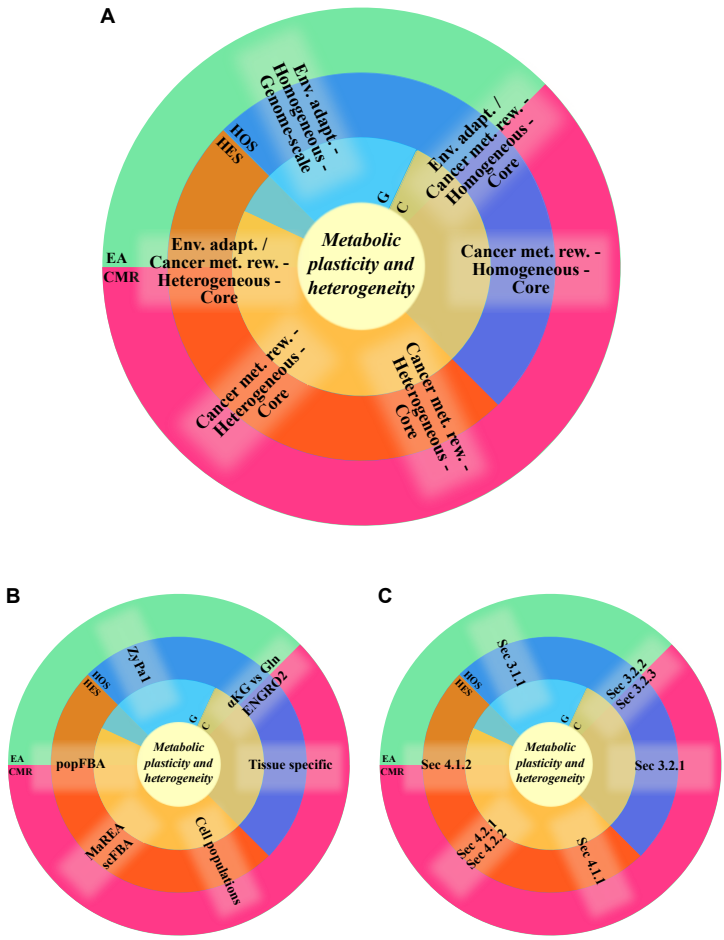


Figure 1.1: Grapical synopsis of this thesis. The three concentric circles represent the three dimensions investigated in this thesis work: plasticity, heterogeneity, modelling approach. From the outermost circle: environmental adaptation (EA) or cancer metabolic rewiring (CMR); homogeneous systems (HOS) or heterogeneous systems (HES); genome-scale models (G) or core models. Each label refers to a specific outcome of this thesis, and it overlaps with specific portions of each circle according to the established cataloguing system. A) The investigation domains. B) The name of each project. C) The section of the thesis where each project is described.

approach proposed by systems biology, as a way to analyze metabolism as a complex biological system. In this regard, I will present metabolic reconstructions that are generally exploited to describe this system, with a particular focus on genome-wide and core networks. Thereafter, I will move to an overview of main modelling approaches for simulating biological systems, by tailoring the discussion to the specific case of metabolic models. Chapter 3 and 4 are similarly organized: after a brief introduction about the topic discussed in each chapter, the subsequent sections will introduce the related projects. In particular, in Chapter 3, the investigation of a given cell or organism metabolism by assuming it as representative of a homogeneous population and exploiting constraint-based modelling will be presented. A computational pipeline implemented for the semi-automatic reconstruction of genome-scale metabolic networks will be introduced to explore metabolic plasticity of a given organism to different metabolic regimes and perturbations. In addition, core constraint-based modelling will be exploited as mean for studying intertumour heterogeneity following metabolic plasticity of different cancer cells adopted for supporting their neoplastic proliferation. Different nutritional conditions and perturbations will be taken into account. Considering phenotypic diversity of cell populations at the intratumour level, in Chapter 4 new computational approaches implemented to shift from an homogeneous to an heterogeneous vision of cell populations will be presented. New methodologies to explore possible metabolic interactions within an heterogeneous tumour population model, and identify phenotypes characterizing the most proliferative subpopulations will be discussed. In addition, a novel approach aiming at integrating single-cell transcriptomics data to enrich the characterization of mechanisms behind metabolic heterogeneity will be introduce. Lastly, in Chapter 5 I will conclude by introducing some future perspectives of this thesis work.

List of publications included in the thesis

Di Filippo, M., Colombo, R., Damiani, C., Pescini, D., Gaglio, D., Vanoni, M. et al. Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models. Computational biology and chemistry, 62, 60-69, 2016. *Personal contribution: AIR, CM, DIM, DMO, PS, WP, RP.*

Di Filippo, M., Damiani, C., Colombo, R., Pescini, D., Mauri, G. Constraint-based modeling and simulation of cell populations. In Italian Workshop on Artificial Life and Evolutionary Computation (pp. 126-137). Springer, Cham., 2016. *Personal contribution: AIR, DIM, PS, RP, WP.*

Damiani, C., Di Filippo, M., Pescini, D., Maspero, D., Colombo, R., Mauri, G. popFBA: tackling intratumour heterogeneity with Flux Balance Analysis. Bioinformatics, 33(14), i311-i318, 2017. *Personal contribution: AIR, DIM, PS, RP, WP.*

Graudenzi, A., Maspero, D., Di Filippo, M., Gnugnoli, M., Isella, C., Mauri, G. et al. Integration of transcriptomic data and metabolic networks in cancer samples reveals highly significant prognostic power. Journal of biomedical informatics, 87, 37-49, 2018. *Personal contribution: AIR, CM, RP.*

Damiani, C., Maspero, D., Di Filippo, M., Colombo, R., Pescini, D., Graudenzi, A., et al. Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. *Accepted to PLoS Comp Biol. Personal contribution: AIR, RP, WP.*

Di Filippo, M., Ortiz-Merino, R. A., Damiani, C., Frascotti, G., Porro, D., Wolfe, K. H., et al. Genome-scale metabolic construction of the stress-tolerant hybrid yeast Zygosaccharomyces parabailii. *Under revision in mSystems. Personal contribution: AIR, CM, DIM, DMO, PS, RP, WP..*

In the personal contribution: AIR corresponds to analyzing and interpreting the simulation results, CM corresponds to curating the model, DIM corresponds to developing and implementing the methodology, DMO corresponds to developing the model(s), PS corresponds to performing the computational simulations, RP corresponds to reviewing of paper in its final form, WP corresponds to co-writing the paper.

Chapter 2

Introduction

2.1 Cell metabolism

Cell metabolism falls within the plethora of biological processes of living cells. The main role of metabolism is to convert nutrients molecules into energy and primary metabolites. Metabolites are chemical species that are directly involved in cellular growth for the synthesis of macromolecules forming the biomass. These macromolecules include proteins, lipids, deoxyribonucleic acids (DNAs), ribonucleic acids (RNAs), and carbohydrates [8, 15]. In many cells, the role of metabolism extends to the production of the so-called secondary metabolites. This class of compounds are small organic molecules that are created from the primary metabolites without being directly involved in the growth, development, and reproduction of cell. Indeed, secondary metabolites are secreted by cell, and are dedicated to perform important tasks, such as communication with other organisms and defense against external stress factors, including xenobiotics, oxidative stress, pharmaceuticals, and pesticides [8, 15].

Biochemical reactions forming cellular metabolism are mostly catalyzed by specific enzymes. Metabolic reactions are grouped to constitute multiple pathways that are, in turn, organized in a highly coordinated network where reactions cooperate to perform the above described biological tasks. As shown in Figure 2.1, generally, metabolic reactions are classified into two categories: catabolism and anabolism. Catabolism corresponds to the degradative phase of metabolism where energy-rich nutrients, including carbohydrates, fats and proteins, are uptaken from the environment, subsequently degraded and finally

converted into other smaller and simpler end molecules. Final products of catabolism are, in turn, reused by cell as precursors of biomass macromolecules. Catabolic pathways are associated to the generation of energy, which is mostly conserved in the high-energy phosphate bonds of adenosine triphosphate (ATP), and of electron carriers like nicotinamide adenine dinucleotidephosphate (NADPH), nicotinamide adenine dinucleotide (NADH), or flavin adenine dinucleotide ($FADH_2$). Partly of the energy generated by catabolic pathways is lost as heat. NADPH coenzyme is then principally exploited within biosynthetic pathways of macromolecules. On the contrary, NADH is mainly re-oxidized through the catabolic processes, and through the oxidative phosphorylation (OXPHOS) pathway by the specific activity of the complex I. The OXPHOS pathway is also called electron transport chain because of its role to transfer electrons through five enzymatic complexes from NADH and $FADH_2$ electron donors to oxygen, by releasing a large amount of energy, which is used to finally produce ATP [16]. Similarly to NADH, $FADH_2$ coenzyme is processed through the OXPHOS pathway by the activity of the Complex II [15].

Anabolism, as opposed to catabolism, represents the biosynthetic phase of metabolism. Anabolic reactions aim at assembling simple and small monomeric precursors into larger and more complex macromolecules, including proteins, lipids, polysaccharides, and nucleic acids. As opposed to catabolism, anabolic reactions generally require energy, which can be provided in the form of the phosphoryl groups that are transferred from ATP molecules, and of the reducing power that is stored in NADH, NADPH, and $FADH_2$ coenzymes [16].

Since about the 1920s to date, many metabolic pathways have been characterized in different living cells, paving the way for the current investigation of cell metabolism. Nevertheless, many portions of cellular metabolism remain currently unknown or still less well defined [15]. The high degree of connectivity that characterizes this system significantly complicates the study of its individual parts. In this regard, the perturbation of a single metabolic pathway may generate an effect on the function of a large part of the complete network [15]. Consequently, the activity of many different metabolic pathways need to be balanced to maintain metabolic homeostasis and ensure the proper function of the cell on the basis of its needs. Moreover, because of high connectivity of metabolism, perturbations of cell functions may lead to an altered metabolism. Therefore, studying metabolic networks is important because it may lead to the identification of novel putative biomarkers and new therapeutic strategies [15].

From the structural point of view, very few metabolic pathways appear as linear. This happens when they include a series of reactions where the product of a chemical transformation acts as substrate only of the next reaction in the

pathway, without participating in any other part of metabolism. Generally, however, metabolic pathways are classified as branched. This means that some intermediates are also involved as substrates or products of other metabolic routes, significantly complicating the structure of the entire system. In addition to linear and branched pathways, some pathways are also classified as cyclic, because the starting component is firstly processed through a series of reactions, and then is finally regenerated. Similarly to the previously discussed branched pathways, cyclic pathways can be also characterized by a series of chemical reactions whose role is to enter at some points in the cycle to replenish or substract metabolic intermediates of the cyclic pathway. These sets of reactions are, respectively, called anapletoric and cataplerotic. A valid example of this last scenario is represented by the Krebs cycle, also known as the tricarboxylic acid (TCA) cycle.

Complex regulatory mechanisms control cells metabolism. These mechanisms ensure appropriate adjustments in the rate of metabolite flow through each metabolic pathway. In this way, metabolites can flow in the right direction and at the correct rate according to cell or organism needs, especially when cells are exposed to varying environmental scenarios. Overall, cells and organisms exist in a dynamic steady state. Disturbance of their steady state trigger these regulatory mechanisms, by allowing cells or organisms to achieve a new steady state, namely the homeostasis [16].

A complex intracellular regulatory system underlies cellular metabolism, to which several factors contribute [16]. The most immediate regulation level of metabolic pathways comes from the availability of the involved substrates in the microenvironment of cells or deriving from other reactions. In addition, enzymes responsible for the catalysis of related reactions must be present so that reactions themselves can occur. In this regard, the expressions of genes encoding for specific enzymes result a crucial factor, and the abundance of these proteins is regulated by several processes, including messenger RNA (mRNA) transcription, splicing, stability, and translation. Once both substrate and enzyme are available, substrate-enzyme binding is regulated by a characteristic affinity that is defined by the dissociation constant K_d. Another measure of the substrate's affinity for a given enzyme is the Michaelis-Menten constant K_m, which corresponds to the concentration of substrate at which the rate of the reaction is half of the maximum one that can be achieved. Low values of K_m indicates high affinity of the molecule for its enzyme, because less substrate is required by the enzyme so that half of the maximum rate of the reaction is reached.

Allosteric regulation represents another factor contributing to the intracellular regulation of metabolic reactions. Some metabolites, in addition to be

chemically transformed among multiple reactions, also play the role of signal molecule of the internal metabolic state of the cell. This regulation is especially needed to adjust the activity of specific enzymes according to the amount of a given molecule in the cell. When a sufficient amount of this molecule is enough for cell requirements, certain allosteric regulators can bind a particular site on the enzyme, called allosteric site. This binding can induce conformational changes of the entire enzymatic structure with a consequent shifting of its state from active to inactive, or vice versa. By exploiting this process, some metabolites can temporarily inhibit the function of one or more enzymes when their activity is not necessary. In general, different post-translational modifications can regulate enzyme function, by locally or globally changing their shapes, or by promoting or inhibiting the binding interactions of substrates and allosteric regulators. Several post-translational modifications are possible, among which the most frequent are the phosphorylation, the acetylation and the methylation. These modifications are controlled through signal transduction processes starting at the level of cell membrane. Signaling processes then lead to the activation or inhibition of specific proteins that are responsible of these modifications by exploiting metabolic products.

Further regulation processes of metabolic pathways are also needed when catabolic and anabolic pathways share the same two end points and most of the enzymatic steps responsible of their interconversion. This scenario is reflected, for example, by fatty acids synthesis and degradation pathways. In this case, it is fundamental that at least one of the steps in these routes is catalyzed by different enzymes in the catabolic and the anabolic direction. In this way, the costly situation where the two opposite pathways happen simultaneously is prevented. It is also avoided the situation where the inhibition of an enzyme acting in the catabolic direction pathway also inhibits the reactions sequence in the opposite direction. These reactions need therefore to be strictly regulated so that when one sequence is active, the other one is inactive. Another contribution to the separate regulation of these types of metabolic pathways also comes from their different localization in the cell. By considering the previous example, fatty acids metabolism is regulated so that their catabolism happens in mitochondria through the beta-oxidation pathway, whereas their synthesis occurs in the cytosol compartment. In this way, separate pools of intermediates, enzymes, and regulators can contribute to the control of metabolic rates of these reactions.

Within multicellular organisms, the regulation of metabolic activities of different tissues occurs at extracellular level. Through the action of several extracellular factors, including growth factors and hormones, metabolic functions can

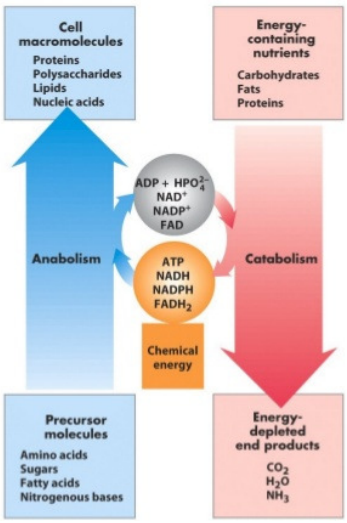


Figure 2.1: Catabolism and anabolism. During catabolism, energy-rich nutrients, including carbohydrates, fats and proteins, are consumed, degraded and finally converted into smaller molecules. Catabolic pathways are also associated to the generation of energy, which is conserved in the form of ATP, and electron carriers like NADPH, NADH, and $FADH_2$. These energy carriers are used within anabolic pathway for assembling small monomeric precursors into larger macromolecules. Image taken from from [16].

be regulated. These extracellular signals can indeed cause changes in the levels of intracellular messengers. Consequently, enzymes activity may be modified by allosteric mechanisms or covalent modifications, such as phosphorylation. Additionally, these extracellular signals can affect enzymes availability by altering their synthesis or degradation rate [16].

2.1.1 Cancer metabolic rewiring

Cancer disease is characterized by a considerable heterogeneity, both at genomic and phenotypic level. This feature is at the basis of remarkable diversity among variuos different tumour types as well as among several subtypes of the same tumour, that extends up to the cells within a single tumour population [17].

Despite this significant level of heterogeneity, in [18] six hallmarks have been

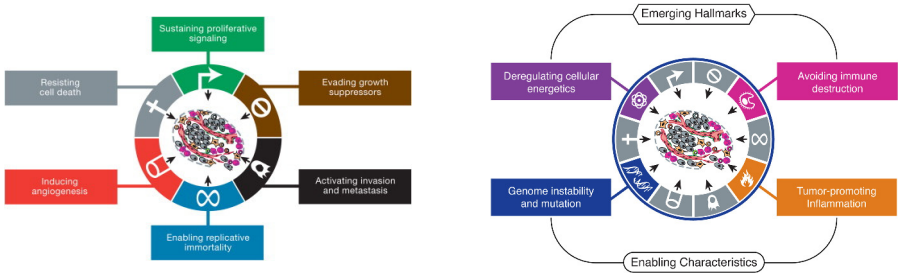


Figure 2.2: The hallmarks of cancer. On the left, core cancer hallmarks, namely self-sufficiency in growth signals, insensitivity to antigrowth signals, evading apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. On the right, two additional cancer hallmarks labelled as emerging hallmarks, namely deregulating cellular energy and avoiding immune destruction. Additionally, two consequential cancer features, labelled as emerging hallmarks, that facilitate the acquisition of both core and emerging hallmarks: genome instability and mutation and tumour-promoting inflammation. Image taken from [18, 19].

proposed as shared features across the plethora of tumours, and as acquired core functional capabilities that allow cancer cells to survive, proliferate, and disseminate. As shown in Figure 2.2, these acquired biological traits are: self-sufficiency in growth signals, insensitivity to antigrowth signals, evading apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis.

One of the emerging hallmarks that may favour the acquisition of these cancer hallmarks may be the development of genomic instability that involves both small structure variations, including increased frequencies of base pair mutation, microsatellite instability, and chromosome instability regarding changes of chromosome number or structure [20]. Due to genomic instability, normal cells progressively evolve towards a neoplastic state that enable them to become tumourigenic and malignant. Recent studies investigate tumours as complex tissues that consists of heterologous cell types interacting among them [21, 22, 23, 24, 25]. Indeed, tumour formation involves the co-evolution of neoplastic cells together with non-transformed cell types that overall form the stroma. In healthy tissue, stroma acts as main barrier against tumorigenesis, but the presence of transformed tumour cells implies a series of changes that induce this environment to support cancer progression. In particular, these changes involve the recruitment of fibroblasts, migration of immune cells, matrix remodelling and the develop-

ment of vascular networks through the angiogenesis process [19, 26].

Recently, in addition to the original defined hallmarks, new ones have been proposed to be important traits assisting tumour growth and development (as shown in Figure 2.2 on the right) [19]. Among them, a reprogramming of cellular energy metabolism replacing the metabolic program that operates in most normal cells, is necessary to support cancer cells growth [17, 19]. As shown in Figure 2.3, under aerobic conditions, normal cells primarily metabolize glucose to carbon dioxide following the oxidation of pyruvate produced by the cytosolic glycolitic pathway, within the mitochondrial TCA cycle. NADH and $FADH_2$ coenzymes that are synthesized by TCA cycle, then fuel OXPHOS pathway to maximize the ATP production. Under this condition, a minimal production of lactate is generated. This preferential behaviour shifts towards glycolysis under anaerobic conditions. In this scenario, normal cells reduce the majority of pyruvate to lactate, by redirecting only a small fraction of pyruvate towards the oxygen-consuming mitochondrial oxidation. Unlike normal cells, energy metabolism of cancer cells behaves anomalously, as was firstly observed in the 1920s by Otto Warburg [27]. In the presence of oxygen, cancer cells mostly reprogram their metabolism towards the glycolysis leading to the so called "aerobic glycolysis". Following this new strategy, glucose is consumed at a higher rate compared to normal cells, and most of glucose-derived carbons are secreted as lactate rather than be completely oxidized within the mitochondrion. This metabolic switch, which is also known as "Warburg effect", seems counterintuitive in energetic terms regarding the net production of ATP per each molecule of glucose. Indeed, the complete oxidation of one molecule of glucose produces 36 ATP molecules as opposed to the value of 2 ATP when glucose is processed through the glycolysis. However, this behaviour, which was initially ascribed to mitochondrial impairments, has been subsequently hypothesized as a way for supporting synthesis of biomass and redox maintenance in cancer proliferating cells [17]. Cancer cells compensate the lower ATP efficiency of aerobic glycolysis by considerably upregulating the glucose transporters for incrementing its uptake into the cytoplasm. The preference towards this path is further supported by the hypoxic conditions that characterize many tumours and that upregulate glucose transporters as well as multiple enzymes of the glycolytic pathway. Glycolytic fueling is also associated with activated oncogenes, such as *RAS* or *MYC*, and mutated tumour suppressors, such as *TP53*. Their alteration within tumour cells is able to confer them the ability to proliferate in an uncontrolled way, to avoid cytostatic controls, and to attenuate the apoptotic process.

The high glycolytic rate provides several advantages for proliferating cancer cells [29]. In support of the explained tumoural metabolic switch, glycolysis

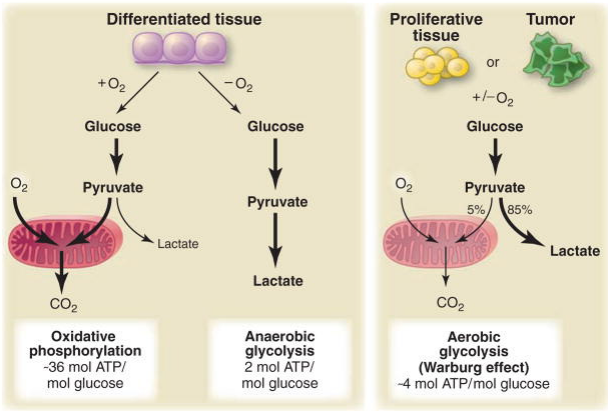


Figure 2.3: Metabolic and energy differences between nonproliferating cells in differentiated tissue under aerobic and anaerobic conditions, and proliferating cancer cells. Image taken from [28].

is not a stand alone pathway. Indeed, it is highly interconnected with several other metabolic routes that are, in turn, involved in the synthesis of macromolecules required for the assembly of new cells. For example, the glycolytic intermediate glucose-6-phosphate, which is synthesized immediately downstream glucose uptake in the cell, can be involved into the pentose phosphate pathway. In particular, it can fuel its oxidative branch for the production of NADPH coenzyme and ribose-5-phosphate, which acts as intermediate of the nucleotides biosynthesis pathway. Similarly, other two glycolytic intermediates, fructose-6-phosphate and glyceraldehyde-3-phosphate, can contribute to the generation of ribose-5-phosphate through the non-oxidative branch of pentose phosphate pathway. Glycolytic intermediates can be also involved in the synthesis of non essential amino acids. In this regard, 3-phosphoglycerate can be shunted into the serine and glycine biosynthesis, whereas pyruvate can be involved in synthesis of alanine, in parallel with its enter into the TCA cycle. Although less directly as in the previous cases, pyruvate is also diverted towards the synthesis in the cytosol compartment of citrate, which acts as precursor for the synthesis of fatty acids and cholesterol.

Tumour cells are also characterized by a considerable high rate of lactate production. From one side, the reduction of pyruvate to lactate can rapidly

replenish the cytosolic levels of NAD^+, enabling the glycolysis to continue. On the other side, the high production rate of lactate is also followed by its secretion in the extracellular environment. Lactate may be actively exchanged as oxidative energy source with neighbouring cells of tumour population that are subjected, due to their location, to a more aerobic microenvironment. In addition, the high level of secreted lactate may be also also exploited as signaling molecule to stimulate angiogenesis within endothelial cells of blood vessels. In this way, cancer growth and motility are incremented, and tumour metastasis are supported. Moreover, lactate molecule may be also related within oxidative cancer cells to a promotion of glutamate uptake and catabolism that, in turn, could be linked to the emergence of cachexia in cancer [30].

TCA cycle is another hub pathway in proliferating cancer cells for the purpose of synthesize biomass precursors [29]. Within normal cells, this pathway is required for substrates oxidation, by deriving the highest possible ATP production rate. Differently from normal cells, intermediates of this cyclic pathway in tumour cells are subjected to efflux and consequent rerouting to energy requiring synthesis of many biomass precursors, including lipids, proteins, and nucleic acids. A key example of this process, which overall is called cataplerosis, is represented by citrate efflux. Once citrate is transported from the mitochondrion into the cytosol compartment, it is converted by the ATP citrate lyase enzyme to oxaloacetate (OAA) and acetyl-CoA. Acetyl-CoA plays a fundamental role within cellular metabolism for its involvement as precursor of lipids synthesis. In this regard, ATP citrate lyase together with another fundamental enzyme involved in the synthesis of fatty acids, that is, fatty acid synthase, are considerably upregulated in cancer cells because of their role in tumour cell proliferation. The propensity of citrate to exit from this cyclic pathway contributes to a "truncated" TCA cycle. This is because the fraction of citrate that goes towards lipids is unbalanced at the expense of the fraction that is oxidized. Other examples of cataplerosis are represented by OAA and α-ketoglutarate (α-KG). These two compounds, processed by transaminase enzyme, are necessary to fuel the synthesis of non essential amino acids that are required during proteins and nucleotides production [31]. The active state of cataplerosis needs of a constant refueling of metabolites that have left the TCA cycle pathway. For that purpose, anaplerosis acts to replenish this cyclic route by mainly refilling the pool of OAA that is required for the mitochondrial formation of citrate. An anaplerotic source of OAA may come from pyruvate carboxylase enzyme, which synthetize OAA from pyruvate with consumption of one molecule of ATP. Nevertheless, it cannot be considered as a core anaplerotic flux for cell proliferation because of its down-regulated activity in some tumours.

A high anaplerotic and cataplerotic flux, together with a high glycolitic rate are strong indicators of cell growth in proliferating cancer cells. At the same time, anaplerosis results to be a more specific signature compared to glycolysis. Indeed, this latter can be also induced by hypoxia or other stressors that are independent from the need of synthesize precursors for cells growth. On the contrary, anaplerosis coupled with its opposite, the cataplerosis, are strictly associated to the usage of TCA cycle as a supply of biomass precursors.

Although glucose metabolism has a key role in fostering cell growth, catabolism of this carbon source is not the only one responsible for a rewired metabolism supporting cancer cells growth. Indeed, glutamine contributes as additional source, being the most abundant amino acid in human serum [17, 32]. Other than its use as nitrogen atoms donor for the synthesis of nucleotides and non essential amino acids, proliferating cells exploit glutamine as carbon atoms provider. In this way, glutamine allows cells to maintain a sufficient anaplerotic flux in order to exploit TCA cycle as hub for the synthesis of biomass precursors [17]. Indeed, following glutamine deamidation to glutamate, this latter is then converted within the mitochondrion by the glutamate dehydrogenase enzyme into the TCA cycle intermediate α-KG. Following this pathway, glutamine indirectly fuels with a domino effect all the other metabolites belonging to TCA cycle.

In many cancer cells, glutamine is processed in the TCA cycle through an alternative way compared to the canonical one. Indeed, glutamine can contribute to the synthesis of acetyl-CoA through the reductive carboxylation pathway. In many cancer cells, especially those subjected to hypoxic conditions, the activity of pyruvate dehydrogenase, which converts mitochondrial pyruvate to acetyl-CoA, is down-regulated. This condition implies lower levels of acetyl-CoA and, consequently, of citrate in the Krebs cycle compared to the α-KG. Following mutations in the isocitrate dehydrogenase enzyme, which physiologically converts isocitrate to α-KG, the corresponding catalyzed reaction may occur in the opposite direction, causing glutamine-derived carbons to fuel the mitochondrial pool of citrate. In this way, this route provides a glucose-independent way for generating the acetyl-CoA that is fundamental for lipogenesis. At the same time, reductive carboxylation allows to conserve glucose for the production of other biomass precursors that cannot be synthesized from other nutrients [33].

2.2 Intratumour heterogeneity

Homeostasis is the balance between cell proliferation and apoptosis such that the overall tissue architecture and function remain constant [34]. Multiple mechanisms regulate these processes by ensuring homeostatic maintenance, including cellular mechanisms, such as cell–cell and cell–extracellular matrix adhesion, and environmental mechanisms, such as metabolic factors, growth factors and stromal cells.

Loss of homeostasis is considered as key features of oncogenic transformation, to which cellular heterogeneity of tumour populations contribute [34]. Tumour initiation process starts from founder cells that evolve by acquiring a series of mutations at each cell division. Subsequent cycles of mutations and selection processes of "driver" mutations providing a selective growth advantage to cells, followed by clonal expansion occur. Depending on the accumulated mutations, tumour populations may contain both clonal mutations that are present in all tumour cells, and subclonal mutations that are heterogeneously present within the tumour [35]. The coexistence in the same neoplastic population of multiple cell subpopulations having different tumorigenic potential constitutes a further level of complexity to the understanding of cancer biology.

Multiple not-indipendent intrinsic and extrinsic factors contribute to cancer heterogeneity within a given tumour and between tumour subtypes [36]. Among intrinsic factors, several genetic mutations, including aneuploidy (i.e., abnormal number of chromosomes), deletions, inversions, translocations, homozygous deletions and gene amplifications, can activate oncogenes or inactivate tumour suppressor genes. Tumorigenesis and tumour progression are also influenced by the origin cell of tumour and by various epigenetic modifications, including DNA methylation, histone and chromatine modification.

The course of neoplastic disease is also influenced by extrinsic factors that constitute cell microenvironment. Ability of tumour cells to recruit an adequate blood supply for the development of the entire population is one of the extrinsic factors. In this regard, blood vessels also provides a specialized niche for cancer stem cells (CSCs), which play an important role in the maintenance of tumour growth. CSCs share with normal stem cells self-renewing capacity and multilineage differentiation properties, but, differently from them, they are highly tumorigenic. Another extrinsic factor is the tumour's ability to recruit diverse stromal cells, including inflammatory cells, fibroblasts, and pluripotent mesenchymal stem cells. By directly interacting with cancer cells, stromal cells play a key role in the promotion of cancer progression.

Two models have been proposed to explain a way in which phenotypic hetero-

geneity could arise within tumour populations [36]. The first one is the clonal evolution theory. According to this model, cancer cell populations progressively evolve due to the accumulation within individual cancer cells of successive heritable genetic and epigenetic mutations. Mutations are then subjected to positive or negative selection depending on whether they confer a competitive advantage or disadvantage to tumour populations. Therefore, in response to selection pressures imposed by tumour microenvironment, the selective outgrowth of clone having more malignant phenotypes occur. New mutations can increase the high interclonal heterogeneity within individual tumours by influencing cell phenotype or function and by contributing to the generation of subclones.

The second model proposed to explain phenotypic heterogeneity of tumour populations is described by the cancer stem cell (CSC) theory. According to this model, cancer populations consists of hierarchically arranged tumorigenic cancer stem cells and their non-tumorigenic progeny. CSCs reside at the apex of cellular hierarchy because of their high tumorigenic and metastatic abilities. Although CSCs are thought to be responsible of driving tumour growth and progression, non-CSCs, being in the great majority within tumour populations, are responsible for expressing many of the phenotypic traits characterizing the tumour as a whole.

CSCs can alternatively generate their progeny through symmetric and asymmetric divisions. Undergoing symmetric division, CSCs generate daughter cells that exhibit the CSC phenotype. On the contrary, through asymmetric division, they generate non-CSCs daughter cells having low tumorigenic and metastatic potential, that can initiate differentiation programs. Divergent phenotypes between CSCs and non-CSCs are regulated by specific stimuli in the microenvironment that can activate given growth factors pathways. Signaling processes can thus affect epigenetic changes in the CSCs and their non-CSC progeny. However, genetic heterogeneity may also contribute to their phenotypic and functional differences. When new mutations occur in tumorigenic cells, clonal evolution and differentiation of tumorigenic into non-tumorigenic progeny can contribute to tumour heterogeneity. Therefore, in some cases, tumours exhibit traits that are generated by both clonal evolution and CSC models [37]. Moreover, recent evidence support the hypothesis of a plastic CSC model where the unidirectional nature of the classic CSC model is substituted by bidirectional CSC-to–non-CSC conversion, where non-CSCs can reacquire a CSC phenotype and contribute to replenish the CSC pool [37].

2.3 Systems biology approach and the role of computational models

The complexity of biological processes occurring in the cell makes it hard to foresee its global behaviour just from the knowledge of its individual parts. As anticipated in Chapter 1, systems biology, in this regard, is intended to be a multidisciplinary approach with the aim of achieving a system-level understanding of complex biological systems. By adopting an integrative approach that joins the reductionist with the holistic view, this new discipline focuses the attention not only on cellular constituents, but, above all, on the relationships existing among them.

In the proposed systemic perspective, the concept of network becomes crucial for describing any complex system at a high abstraction level. Depending on the nature of cellular interactions, several types of networks may be defined, including protein-protein interaction networks, signaling networks, metabolic networks and regulatory networks [4]. However, none of them acts on their own in the cell. On the contrary, they constitute a "network of networks" that as a whole are together responsible for the overall behaviour of cell. In each network, the description of the interactions among its constituting elements is beneficial to get a complete vision of the system. At the same time, however, the high amount of involved items and the non-linearity affecting their interactions, makes impossible to achieve this goal through the pure intuition [7]. Systems biology approach meets these requirements by formally describing the multiple biological processes occurring in the cell following the integration of experimental and computational research. More specifically, this approach relies on mathematical modelling in order to provide both qualitative and quantitative valid predictions about the behaviour of complex biological systems [4, 38].

The concept of model has different meaning among various research areas. In molecular biology, model organisms denote well known and characterized species, having some biological traits that make them well suitable for experimental manipulations. Since they are smaller, simpler, and faster growing than more complex organisms such as humans, model organisms result simplified and tractable systems that could be used to investigate and understand biological processes. By exploiting the conservation of some features over the course of evolution, model organisms are used to investigate specific biological phenomena. Along this line, model organisms provide new insights for the investigated phenomena, which could be used for other species where they would be more difficult to be directly studied [39, 40]. Examples of model organisms includes

Saccharomyces cerevisiae, *Caenorhabditis elegans*, *Mus musculus*, *Escherichia coli* and *Drosophila melanogaster*.

From the computational point of view, the term model changes its meaning. Although it continues to be a simplified representation of reality, an *in silico* model is defined as an abstract representation of a system able to summarize the established knowledge about it in a coherent mathematical formulation. In some cases, more than one mathematical formalism can be exploited to highlight different aspects about the same system. In this regard, mathematical modelling is considered as a selective procedure, because the resulting models represent some aspects of biological reality with the aim of answering to specific tasks according to the purpose of research. Overall, mathematical modelling allows to attain a quantitative understanding of a biological system [38].

Until now, different types of mathematical modelling approaches have been developed, which will be better explained in Section 2.5. Although their diversity, an example of merging of different computational approaches was presented in [41] relative to the development of an integrative whole-cell model for the human pathogen *Mycoplasma genitalium*. This novel approach is based on the combination of multiple computational approaches to model the total functionality of an organism by representing each characterized gene function. The exploitation of this approach for reconstructing whole-cell models for other organisms, including *Mycoplasma pneumoniae*, *Escherichia coli* and H1 human embryonic stem cell, is currently under development.

Generally, the choice of modelling approach is bounded by being appropriate to solve the problem under investigation. Consequently, it strictly depends on the final purpose of the modelling, and on the available biological knowledge to incorporate into the model, deriving both from literature and experimental data [7]. Through iterative cycles of quantitative experimental data generation and mathematical modelling, as illustrated in Figure 2.4, systems biology explores the properties emerging from the interactions established among system components. A cycle of systems biology research begins from the knowledge stored in literature and in various biological databases about a specific biological phenomenon, together with laboratory experiments data. By exploiting this information, a first computational draft model representing the phenomenon of interest is created. Computational outcomes deriving from model simulations are then experimentally validated as a way to identify possible inconsistencies with the established experimental evidence. The validation step of the *in silico* model under reconstruction allows to assess its degree of accuracy. Furthermore, it also helps network refinement through the identification of novel components or interactions whose addition to the model allows to obtain a better match

between computational simulations and experimental results [7, 42].

The usefulness of computational models comes from their ability, first of all, to require a formal representation of the current knowledge relative to the modelled biological system by means of clear and explicit mathematical working hypotheses. Their computational simulations, in turn, enable to formulate new hypotheses and predictions about their working. Usefulness of simulations further goes towards a way for discriminating among alternative explanations about their functioning, and for studying the effect on the global behaviour of various perturbations on components or environmental conditions. Consequently, computational simulations allow to investigate the entire network in different situations. Therefore, computational analyses are also considered as valuable, since they help biologists to predict the behaviour of the system under new conditions, by formulating new hypotheses that can be then experimentally validated [38].

The growing popularity of systems biology approach is witnessed by the increasing number of the generated *in silico* models and computational tools for their simulation and analysis. At the beginning, these models were often developed without complying with certain specifications. However, their lack of promiscuity led to various problems [43]. Among them, the development of models having software-specific format prevented their simultaneous usage from multiple tools. Consequently, a time consuming model recoding was therefore necessary, by limiting model usage. The model usage extensibility to different modelling environments and model representation languages was also compromised by highly personalized reconstruction procedure. The lack of a common format for describing mathematical models led to the inability to exchange models between different simulation and analysis tools. To address this issue, the Systems Biology Markup Language (SBML) has been established as standard exchange language for the representation of mathematical models of biological networks. Moreover, SBML has been developed to be neutral with respect to both multiple programming languages and software encoding [43, 44]. As shown in Figure 2.5, the SBML structure has been constructed based on that of the eXtensible Markup Language (XML), because of its increasing widespread usage as a standard data language for informatics. In particular, SBML inherits the XML structure in terms of the usage of plain text, and of a pair of start/end tags enclosed by the "<" and ">" characters for defining each component that need to be made explicit in the model. Among them, for example, the list of the included biochemical transformations together with their set of reactant and product species, their stoichiometry, cellular localization, kinetic parameters, and reactions rules, whose relevance will be deepen in the next

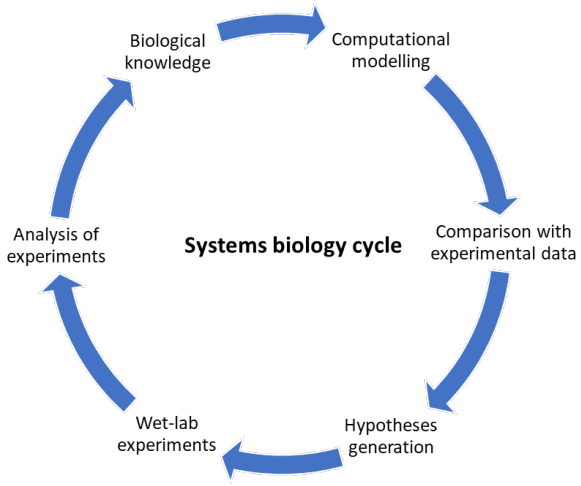


Figure 2.4: Classic iterative cycle of systems biology approach. The definition of new biological knowledge- and experimental data-based mathematical models relative to a given biological phenomenon leads to the formulation of new hypotheses regarding system behaviour. Validation of these predictions through comparison of wet and *in silico* experiments outcomes may help model refinement by identifying new components or interactions to be included in the model itself. Image adapted from [45].

sections, are included.

2.4 Metabolic network reconstructions

Current high-throughput technologies offer the chance to explore all the path from the genome until the phenotype of a cell. In particular: genomics allows to unravel the entire genome of an organism; transcriptomics aims at studying the total set of mRNA transcripts of a cell in a specific condition and moment, globally called transcriptome; the proteomics is the equivalent of the transcriptomics for investigating the set of all the proteins that are translated in a cell in a given condition and moment, named the proteome; and metabolomics deals with the study of the metabolic profile of a cell that translates into the metabolome.

Systems biology aims at exploiting the massive amount of data obtained by omic technologies to address the investigation of genotype to phenotype

```
<?xml version="1.0" encoding="UTF-8"?>
<sbml xmlns="http://www.sbml.org/sbml/level1"
      level="1" version="2">
  <model name="gene_network_model">
    <listOfUnitDefinitions>
        ...
    </listOfUnitDefinitions>
    <listOfCompartments>
        ...
    </listOfCompartments>
    <listOfSpecies>
        ...
    </listOfSpecies>
    <listOfParameters>
        ...
    </listOfParameters>
    <listOfRules>
        ...
    </listOfRules>
    <listOfReactions>
        ...
    </listOfReactions>
  </model>
</sbml>
```

Figure 2.5: Skeleton of a SBML file. All the possible top-level elements showing all components of model definition are depicted. Each level is characterized by a pair of start/end tags enclosed by the "<" and ">" characters. Image taken from [43].

relationship of a given cell or organism [46]. As previously explained, metabolism plays a key role due to its involvement in providing a direct indication of cellular biochemical activity [1]. To gain a systemic perspective on cellular metabolism, the network representation is pivotal. Here, metabolites are depicted as nodes joined by multiple connections when involved in the same reaction [47]. The most appropriate representation of a metabolic network is a bipartite graph. This type of graph contains two disjoint set of nodes, and the only possible connections join two elements belonging to the two distinct sets of nodes. Consequently, links between two elements belonging to the same set are prevented. Networks with different detail levels can be exploited to investigate metabolism, ranging from genome-wide to core networks [8, 38, 48]. As it will better examined in Section 2.5, systems biology offers three main mathematical formalisms for the modelling and simulation of these networks. From one side, mechanism-based modelling is mainly focused on the investigation of the system dynamics. Consequently, this modelling approach allows to achieve a very detailed comprehension of the modelled biological network. Nevertheless, the limited availability of quantitative parameters that are fundamental factors in the mechanicistic view of the system, restrict the applicability of this modelling

approach to small-scale networks. On the other side, interaction-based models aims at investigating both large- and small-scale networks. However, it provides new knowledge just in qualitative terms. Indeed, interaction-based modelling is limited to the analysis of network topological properties through graph theory. The third available modelling formalism is constraint-based modelling that is placed between the two previously mentioned approaches. Indeed, constraint-based modelling allows to reach a more quantitative understanding also of large size systems. The constraint-based modelling, which was initially developed for the optimization of microbial strains and for the maximization of the yields of some compounds for biotechnological applications, represents now the most widely used method in the study of metabolic networks.

Multiple factors can influence the choice of the most appropriate mathematical formalism to be used for the system under consideration [48]. The final objective of the research results the most influential element. A specific objective allows to guide towards the aspect of the system that we want to investigate. The granularity level of the network represents another important factor. This aspect regards the level of abstraction that is chosen for descripting system components and their interactions. According to the chosen level of granularity, the resulting models can vary in terms of their size, which is reflected in the number of components and interactions. In particular, models can be classified as coarse-grained and fine-grained. Coarse-grained models have the highest abstraction level and refer to large scale networks that can be investigated through interaction-based or constraint-based modelling approaches. On the contrary, fine-grained models refer to small size models for which mechanism-based modelling can be exploited.

In addition to all these elements, the experimental data that are available or can be generated for the system under investigation, such as for example metabolites concentrations or metabolite uptake and secretion rate, constitutes another crucial factor for the choice of the most appropriate mathematical formalism. The experimental dataset can be also exploited in the network reconstruction process to discriminate among several hypotheses about the structure of the network under generation.

Computational costs and time that are required to perform the analyses of the models are another factors that can considerably vary passing from the investigation of topology of interaction-based models, to the computation of flux distributions through constraint-based modelling, towards the determination of system dynamics through mechanism-based modelling.

Overall, it is worth noticing that the development of computational models needs an appropriate choice of which components and interactions are to be

included in the model, in a trade-off between including as many molecular details as possible, and avoiding the addition of too many information by creating an unhandled model structure.

The reconstructed network can be represented, in a formal mathematical language, by means of a graph, which consists of a series of items that are connected to each other by links, each one representing the interactions between two components. A graph G is defined by two datasets V and E:

$$G = \langle V, E \rangle$$

The set V, which may be mathematically expressed as $V = \{v_1, v_2, ..., v_n\}$, represents the set of graph components that are also called nodes. The set of edges E, whose mathematical formalization is $E = \{e_1, e_2, ..., e_m\}$, consists of the established interactions among the nodes of the graph. Each element e of the set E, in turn, corresponds to a pair of nodes $e = \{v_i, v_j\}$, where $v_i, v_j \in V$. Moreover, the E set is also defined as the subset of the cartesian product of V by itself, that is, $E \subseteq V \times V$. In general, the cartesian product of two data sets A and B is the set of all the ordered pairs (a, b) where $a \in A$ and $b \in B$.

According to graph theory, a network can be classified as directed or undirected depending on the nature of its interactions. Interactions within directed networks are characterized by a well defined direction, leading each edge to be defined by an ordered pair of nodes. This implies that given two edges $e = \{v_1, v_2\}$ and $e' = \{v_2, v_1\}$, then $e \neq e'$. Directed links in a directed biological network can represent, for example, the direction of the material flow in a metabolic network between specific substrates and products, or the direction of the information flow within regulatory networks from a transcription factor to the regulated gene. On the opposite side, in an undirected graph, interactions are not associated to any direction. This type of networks may be exploited, for example, for the analysis of protein-protein interaction networks, where the directionality of each link loses the previously mentioned meanings and just indicates the only interplay of two elements.

2.4.1 Genome-scale metabolic networks

Thanks to current high-throughput sequencing technologies large amounts of data are rapidly and in an affordable cost generated. Consequently, sequencing of complete genomes for several organisms is growing rapidly, by increasing the possibility to study the genetic composition of virtually any organism [49]. In particular, annotation of the entire genome of a given target organism lays the foundations for investigating its metabolic potential. *In silico* metabolic models

represent the mean to enrich these high throughput data, and achieve a systemic view on all the metabolic capabilities of the investigated organism.

Genome-scale metabolic networks constitute the most comprehensive repository of all the available knowledge concerning the metabolic transformations occurring in a given cell or organism. By integrating the genomic information, these reconstructions allow to follow the flow of information gene→enzyme→reaction: starting from the list of genes of the investigated cell or organism, it is possible to pass to the corresponding encoding enzymes and finally to the catalyzed reactions. Therefore, a direct correlation between the genomic information of a given organism, the corresponding metabolic activity encoded in its genome, and finally a particular phenotype, is depicted. Besides genome annotation, genome-wide networks integrate data deriving from other source of information about the organism of interest. Among them, biological databases, high-throughput data, wet experiments and scientific literature are exploited for the reconstruction of this kind of networks [50].

Each individual reaction included within genome-scale metabolic networks is accompanied by a number of additional information. First of all, it is fundamental to establish the involved substrates and products, together with eventual cofactors to assist the activity of the enzyme catalyzing the reaction. Moreover, chemical reactions must be expressed according to their correct stoichiometry, indicating the exact ratio between all the reactants and all the products so that the reaction can occur. It is also fundamental to define for each reaction its directionality in terms of reversibility or irreversibility, as well as its cellular localization. This last aspect refers to cell compartment where the reaction takes place. Finally, the enzyme responsible for the reaction catalysis need to be declare, coupled with the set of genes encoding for this enzyme. This information is encoded within metabolic networks through the definition for each specific reaction of the corresponding gene-protein-reaction (GPR) rule. A GPR rule is a boolean expression indicating which genes encode for the enzymes catalyzing a specific reaction, and how they are interconnected among them. Consequently, GPR rules express the relation between genes, the corresponding proteins and the reactions where these proteins are involved as catalysts [50]. For example the human enzyme isocitrate dehydrogenase 3 is associated to the following GPR rule: *HGNC:5386 AND HGNC:5385 AND HGNC:5384*. This expression use the boolean operator *AND* to link the indicated genes (in this case expressed through their corresponding HGNC database identifier) by treating them as subunits of the same enzyme. The human glucokinase enzyme is instead associated to the GPR rule *HGNC:23302 OR HGNC:4195 OR HGNC:4922 OR HGNC:4923 OR HGNC:4925*. In this case, the operator *OR* join 5 genes rep-

resenting isoforms of the same enzyme. The definition of GPR rules allows to extend the usefulness of metabolic network as scaffolds for the integration of omics data generated from high-throughput technologies. Indeed, following omics data integration, context-specific models can be created. Morever, omics data integration also gives the possibility to identify metabolic sub-networks, which may sometimes be located in distant parts of the overall metabolic network, whose expression need to be co-altered for maintaining the cellular homeostasis in specific conditions [8].

In the plethora of biochemical transformations that are generally included in genome-scale networks, two further sets of reactions are also included in addition to those coming from the above mentioned sources. These two sets include exchange and transport reactions. Their role is focused on the transport of metabolites that, respectively, need to enter the system from the extracellular environment or vice versa, and to move between different internal cell compartments [50].

The reconstructed metabolic models generally represent an initial draft version that needs to undergo iterative cycles of analysis, validation and refinement, so that a predictive and useful model can be achieved.

Different mathematical modelling approaches can assist in the formulation of novel hypotheses and predictions relative to metabolic models behaviour in physiological conditions and following specific perturbations. Indeed, getting some control over them enables to save biologists costly and time consuming experiments. In this regard, single or multiple gene knockouts, including essential genes, can provide new predictions that can be pivotal in the context of genome engineering to achieve a specific desired phenotype [50]. In addition, secretion or consumption rate of specific metabolites, or the growth ability under changing nutritional conditions can be tested. The subsequent experimental testing of these novel *in silico* insights may result beneficial to assist the iterative process of model refinement. The validation process includes the comparison of *in silico* predictions deriving from model simulations against experimental data to highlight inconsistencies with biological reality. This mismatch may reveal an insufficient degree of information included in the network, which can be explained by incompleteness of data available in literature or biological databases. In addition, missing regulatory processes, which for the purpose of the reconstructed network, are not included within metabolic models, may explain the inconsistencies emerged between computational and experimental evidence. Gaps and thermodynamically infeasible loops [51] can further negatively affect the quality of the reconstructed model. Gap reactions are missing reactions leading to the formation of metabolites, called dead-end metabolites, which are

involved in one or multiple reactions within the model only as substrate or product. The process of gap filling is fundamental within the entire process of network reconstruction because the presence of dead-end metabolites prevent the mass conservation law.

Model validation can also help to infer novel hypotheses about the system under investigation. These hypotheses need to be tested in laboratory through targeted wet experiments, whose results can generate new knowledge that can be integrated within the network, helping its refinement.

Genome-scale metabolic networks are often the product of a community effort that require many years of work during which the original versions of these models typically undergo to incremental improvements and editing. These latter are relative to the curation of the already included biochemical reactions about their stoichiometry and cellular compartmentalization, or the inclusion of new chemical transformations when deemed necessary following the model validation due to inconsistencies with biological reality. Today, many protocols for the reconstruction of these genome-wide models exist, such as that introduced in [52] and other ones that are tailored to the specific set of available initial data [50]. Being as a whole a time consuming procedure, automated strategies of metabolic networks reconstruction result as beneficial for the purpose of accelerating a process that would otherwise be very slow. The first genome-scale metabolic reconstruction was created for the bacterium *Haemophilus influenzae* Rd just after its genome sequencing and annotation was produced [53, 54]. From that time, in the last two decades, other genome-wide models of metabolism have been produced for a plethora of model organisms, spanning from bacteria to higher eukaryotes [55]. The high potential of these reconstructions in unraveling already established and new biological traits for the corresponding organisms justifies their high availability.

2.4.2 Toy and core metabolic networks

Genome-scale metabolic networks, in addition to be repository of the current metabolic knowledge about specific organisms, can serve as scaffolds for the generation of two types of reduced size networks: toy and core metabolic networks [48].

In complete opposition to genome-wide networks, toy networks are characterized by a very simple structure, and by the inclusion of a very limited number of reactions. Overall, toy networks have a scarce adherence to reality. Indeed, the included reactions are generally fake reactions that are not exploited either to describe a specific pathway in molecular terms. These networks aim at high-

lighting the most relevant components of the system under investigation and some major regulatory properties. Moreover, toy networks are often used as proof of concept to test the validity of new computational methodologies or to begin the metabolism investigation of new organisms. A valid example of toy network can be found in [56], where it is exploited to understand the relationship between metabolism rewiring and growth rate.

As opposed to toy networks, core networks are characterized by a higher adherence to reality. Indeed, they include selected reactions that are manually chosen because of their relevance to explore specific metabolic aspects. Core networks are also characterized by the lumping of reactions belonging to a linear pathway without branching towards other metabolic pathways included in the network. Lumped reactions are summarized by a unique fake reactions whose stoichiometry is the net stoichiometry of the merged reactions, without removing biological information from the network. Being so close to biological reality, core networks, differently from toy networks, are suitable for being extended into more complex networks through the inclusion of other reactions or entire pathways.

The intermediate complexity level of core networks make their analyses and the interpretation of simulation results more straightforward if compared to the predictions deriving from genome-scale models simulations. However, although lower complexity of core networks can be beneficial, an over-simplification should be avoided. This would lead to neglect some important elements, with a consequent negative impact on the simulations outcomes.

2.5 Modelling approaches for metabolic networks

As reported in Figure 2.6, three mathematical modelling strategies, ranging from coarse view to very detailed descriptions of biological systems, currently exist: interaction-based, constraint-based and mechanism-based modelling. In the next sections, these approaches will be, in particular, described in the context of metabolic network investigation, as focus of my thesis work.

2.5.1 Interaction-based modelling

Interaction-based modelling is founded on the analysis of network topology, that is, the structure underlying its components and the related interactions. In this modelling approach, quantitative parameters like molecular species amounts, kinetic constants, and interactions stoichiometry are ignored. This leads this

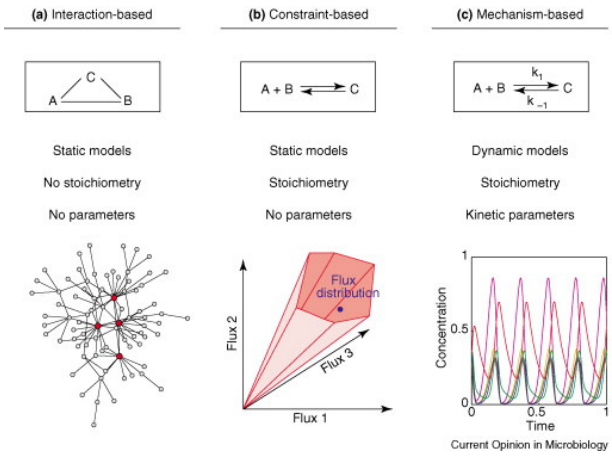


Figure 2.6: Modelling approaches of biological networks. From the left, interaction-based modelling, constraint-based modelling and mechanism-based modelling approach are depicted. For each methodology, the three criteria adopted to distinguish these approaches regard corresponding investigated models. They can be static or dynamic models, they can include or not stoichiometry information, and they can incorporate kinetic parameters. Image taken from [57]

modelling strategy to be particularly suitable for the analysis of large-scale networks. At the same time, interaction-based modelling generates static models because any knowledge about system dynamics over time is considered [48, 57].

Graph theory allows to highlight topological properties of networks.

The most basic topological property is the node degree k, which indicates for each node in the graph the number of other nodes sharing a connection. The concept of node degree is defined at the level of single items. However, it can be extended to the entire network by means of the degree distribution $P(k)$, which indicates the probability that a selected node within the network has exactly k links. The analysis of degree distribution can be beneficial to distinguish among different types of networks.

Bipartite graphs result the most appropriate representation of a metabolic network, where nodes corresponds to both metabolites and biochemical reactions connected by links indicating the involvement of a specific metabolite in a given reaction. Being metabolic networks characterized by a specific directionality of the edges, they can be represented as directed graphs. Therefore, each node is characterized by an "in" and an "out" degree, which denote the number of reactions that, respectively, produce or consume a given metabolite. The assessing of degree distribution $P(k)$ in many metabolic networks revealed the emergence of the so called scale-free topology. This naming is due to the lack of an internal scale as a result of the coexistence in the same network of nodes having widely different degrees. In mathematical terms, in a scale-free network, the degree distribution function approximates a power-law function $P(k) \sim k^{-\gamma}$, where the value of the exponent γ tipically lies between 2 and 3. The meaning of this function is linked to the relevance of hubs within networks following a power-law degree distribution. Hubs are nodes whose degree is very high if compared to the other network nodes. The shape of power-law function highlights the coexistence of few highly connected hubs that are, in turn, linked with numerous small-degree nodes. This scenario is translated within metabolic networks into the simultaneous presence of most metabolites that are characterized by a very low degree because of their involvement in few reactions, coupled with few hubs metabolites that are involved in a high number of chemical transformations. The presence of hubs in the network is relevant because their perturbation can considerably impact the entire structure. In this regard, topological analyses of many metabolic networks also revealed another property called disassortativity. This property indicates the absence of tendency of two hubs to be directly linked. In this way, strong negative effects on the entire network caused by the perturbation of a hub made worse by the connection between two hubs, are avoided.

Another relevant topological property of graph theory is the concept of distance between two components, which is quantified by the so called path length. A path refers to a sequence of connected nodes that enable to move between two nodes, whose length corresponds to the number of linking edges. Given two nodes in the graph, alternative paths can exist. Among them, the shortest paths are relevant because they are characterized by the lowest number of links between the two selected nodes. At biological level, the shortest path length may play an important role, for example, in the context of signaling networks. In this scenario, shortest path length can be a measure of the response speed of the network against an external signal. The mean over all the shortest paths between all pairs of nodes in the graph is called average path length. This measure can be related to the global navigability of the network. In the context of metabolic networks, average path length allowed to infer the ultra-small world property. According to this property, the average path length l is mathematically formulated as $l \sim loglogN$, where N is the number of nodes included in the graph. This expression indicates that short paths link most of pairs of metabolites. Therefore, this feature is related to a fast transmission of information through the network, following which local perturbations could quickly expand through the entire network.

Finally, through graph theory, it is also possible to highlight the presence of modules in metabolic networks. A module corresponds to a semi-autonomous unit designed to perform specific tasks within the overall network. The modularity feature can help to reduce the network complexity following the study of sub-systems in isolation, facilitating the issue of system-level investigation. In the context of metabolic networks, the coexistence of modules, which usually overlap with metabolic pathways, joined the previously discussed scale-free topology, generates the emergence of the so called hierarchical networks.

2.5.2 Constraint-based modelling

Constraint-based modelling represents the most applied computational method to investigate both large and small scale metabolic networks. This approach allows to compute the flux of metabolites through all the included reactions. In order words, constraint-based modelling computes the rate at which every metabolite is consumed or produced by each biochemical transformation.

In nature, any cell operates and evolves under a plethora of several constraints, whose role is to limit the feasible phenotypic states that each cell can achieve [58, 59]. Constraint-based modelling captures the intricate relationship between genotype and phenotype of a cell by simultaneously accounting for the

constraints that are imposed on the phenotype. These constraints derive from the complex relation between the genotype of the cell, its environment and the physico-chemical laws to which cells are subjected [58].

The core assumption of constraint-based modelling is the steady state for the intracellular metabolites. According to this assumption, the consumption rate of each internal metabolite is equal to its production rate. The choice of investigating system behaviour at the equilibrium strictly depends on the fact that cellular metabolism has generally very fast transients. Accordingly, steady state condition is reached in terms of few seconds. Because of fast transients, any perturbation of metabolism will result in a deviation from the steady-state condition that will be very quickly resolved, with a consequent rapid change in the metabolite levels [8].

All the possible functional states achieved by a given metabolic network constitutes the so called "solution space" (graphically shown in Figure 2.7). Solution space denotes a mathematical space including all the candidate solutions of the investigated system, according to the imposed constraints. Overall, it is important to identify and impose the necessary boundaries in order to define a solution space that allows to determine physiologically relevant phenotypes.

The first developed constraint-based modelling approach was the flux balance analysis (FBA). Originally, FBA has been developed to be applied in the field of metabolic engineering with the aim of finding the optimal way to maximize the production of biochemical compounds of industrial interest, such as biofuels [60]. FBA exploits linear programming to identify the optimal flux distribution that maximizes or minimizes a specific objective function under the steady state assumption. This assumption is mathematically expressed as $dX_i/dt = 0$, where X_i is the concentration of a metabolite i belonging to the metabolic model. It follows that time variation of the concentration of each internal metabolite is equal to zero. In line with steady state assumption, mass conservation law is assumed, which states that at steady state the total amount of any compound being produced must be equal to the total quantity being consumed. In short, the total mass of the system remain constant.

The optimization of a given objective function is related to a further assumption of FBA, under which a cell in nature behaves optimally with respect to a specific metabolic function. Different kinds of metabolic objectives may be defined. Among them, maximization of growth rate, minimization of the consumption of specific nutrients, or maximization of the production rate of relevant metabolites, may be assumed.

As shown in Figure 2.8, FBA procedure begins from the reconstructed metabolic network, which consists of stoichiometrically balanced biochemical reac-

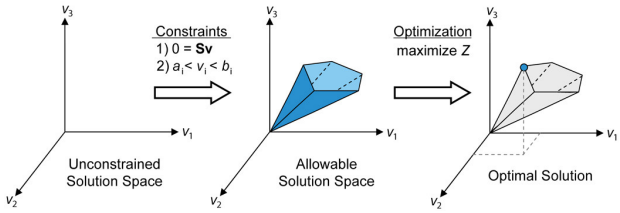


Figure 2.7: The role of constraints within constraint-based modelling. Without imposing any constraint, solution space is unconstrained (left graph). After the imposition of mass balance constraint (1) and capacity constraints (2) on the network, an allowable and restricted solution space is defined (middle graph). Any flux distribution within the blue polytope can be acquired. The optimization of a given objective function allows to identify a single optimal flux distribution, which corresponds to the blue point on the edge of the polytope (right graph). Image taken from from [60].

tions, and, to be compliant with the previously discussed mass conservation law, of the sets of transport and exchange reactions.

After that, all the metabolic reactions included in the model are mathematically represented as a stoichiometric matrix S of size $M \cdot R$, where M is the number of metabolites and R is the number of reactions included in the network. Every element s_{ij} of the stoichiometric matrix S represents the stoichiometric coefficient relative to metabolite i within reaction j. This coefficient assumes positive values when the related metabolite is produced in the reaction, negative when it is consumed, and null when the metabolite does not participate in the reaction. Since most biochemical reactions involves only few metabolites, the matrix S is defined as sparse. This mathematically means that most of its stoichiometric coefficients are equal to zero. The stoichiometric matrix is needed to meet the mass balance constraint (i.e, the steady state) that can be translated in the following equation:

$$\frac{dX_i}{dt} = \Sigma_j s_{ij} v_j = S\vec{v} = 0 \tag{2.1}$$

where s_{ij} corresponds to each element of the stoichiometric matrix S, v_j corresponds to a component of the flux vector $\vec{v}$, which is the output of FBA, and X_i is the concentration of a metabolite i. According to equation 2.1, every vector $\vec{v}$ satisfying the equation $S\vec{v} = 0$ belongs to the null space, or also called Kernel, of the S matrix. The Kernel contains all the possible solutions, that is, all the

feasible flux distributions $\vec{v}$, complying with the steady state assumption.

Typically, metabolic networks include more reactions than metabolites. For this reason, these systems are considered as underdetermined. This situation means that, at this stage, the null space results very large and further constraints need to be imposed (Figure 2.7). Among them, a lower and an upper bound on each individual reaction define a range of flux values that are allowable for that reaction. This kind of constraint is partly related to reaction directionality. Indeed, in case of irreversibility, these bounds vary between 0 and a positive value. This range indicates that flux can pass through the corresponding reaction following the forward direction, with the consequent block of the backward one. Vice versa, a range between a negative value and 0 implies that only the backward direction is active. In case of reversibility, both lower and upper bounds can be set equal to a non-zero value, meaning that both directions can be active.

Spatial and environmental constraints represents other kinds of constraints. Spatial constraints regard the localization of each biochemical transformation and metabolite. Environmental constraints refer to the definition of the extracellular nutritional environment with the consequent modification of flux boundaries of the corresponding exchange reactions according to experimentally measured uptake or secretion rates. In case of complete absence of a specific substrate in the simulated growth medium, the constraints of the corresponding exchange reaction are set to zero, to indicate that no flux can pass through this reaction. Similarly, setting both lower and upper bounds of internal reactions to zero also serves to mimick genes knockouts that are responsible for that reaction. At a higher level, it is in addition possible to modify reactions bounds by exploiting the related GPR rules to integrate transcriptomics or proteomics data.

All the imposed constraints take the form of equalities or inequalities that, as shown in the middle graph in Figure 2.7, overall define a polytope representing all the possible states of the system. The role of constraints is strictly related to define a reduced allowable solution space consisting of all the possible flux distributions. Consequently, when constraints are imposed, all the external points out of the solution space are denied.

The fourth step of FBA is focused on the definition of the objective function. The identification of biologically grounded objective function is not always straightforward. In particular, its knowledge is especially limited when multicellular or multi-organism populations are investigated, since they are generally characterized in nature by multiple objectives. However, even if a plausible objective is identified and formulated in the network, the possibility that the organism is not in an optimal but in a sub-optimal state cannot be excluded.

Besides the role of the constraints in limiting the range of all possible metabolic states, the optimization of a specific objective function allows to retrieve a single optimal flux distribution that lies on the edge of the allowable solution space, as shown in the graph on the right in Figure 2.7. The objective function Z is mathematically expressed as the linear combination of flux vector $\vec{v}$ and of vector $\vec{c}^T$:

$$Z = \vec{c}^T \vec{v} \tag{2.2}$$

Vector $\vec{c}^T$ denotes the transpose of the weights vector $\vec{c}$, which indicates the contribution of each model reaction to the definition of the objective function. Practically, the weights vector includes zero values except in correspondence of reactions that are desired to be optimized.

One of the most frequently defined objective functions is the maximization of biomass synthesis to simulate the maximum amount of biomass, measured in grams, that is made up per unit time according to all the constraints imposed on the system. This objective function is represented in the model by the flux of a biomass production pseudo-reaction that consider all the precursors contributing to create new biomass, including nucleic acids, amino acids, lipids, and carbohydrates. In particular, each building block is associated to a specific stoichiometric coefficient that may be calculated according to the experimental determined macromolecular biomass composition. A way for calculating each coefficient s_m, which will be further discussed in Section 3.1.1, is the following one:

$$s_m = \frac{f_{\mathrm{m}} \cdot f_P \cdot 10^3}{\omega_{\mathrm{m}}}$$

In this equation f_{m} represents the fraction in weight of the monomer m into the macromolecule P, f_{P} represents the fraction in weight of the macromolecule P into the biomass, and ω_{m} is the molecular weight of the monomer m.

In the last step of FBA, linear programming is exploited to solve the system of linear equations deriving from the equation $S\vec{v} = 0$, and the imposed constraints. In general programming problems deal with the efficient use of limited resources to fulfill a specific objective and finding an optimal solution. Linear programming problems represents a subclass of programming problems, which aims at finding a vector of values that optimizes a linear function expressing the objective of the problem that is subjected to linear constraints. Each linear constraint can be expressed as linear equation and/or inequality, and represented by as a hyperplane in the n-dimensional solution space. Globally, the complete set of linear equations depict a convex polyhedron as solution space. Linear

programming algorithms determine within the solution space the solution that optimizes the objective function. The one assumes its minimum at an extreme point of the convex polyhedron. The simplex algorithm is the most used in linear programming. This algorithm is so named because of the simplex that is formed in the n-dimensional space by the set of linear constraints. Through this algorithm, once a first feasible solution satisfying linear constrains has been determined, a series of iterations is to find a new feasible solution where the value of the objective function is smaller or, at least, equal than the value of the objective function at the previous step. This iterative process continues untile a minimum solution of the objective function is reached [61]. So far, various implementation of the simplex algorithm exist, including GLPK [62] and Gurobi [63].

To sum up, the formulation of a FBA problem may be formulated in this way:

$$\text{maximize or minimize } Z = \sum_{i=1}^{R} c_i v_i$$
$$\text{subject to } S\vec{v} = 0,\ \vec{v}_{min} \leq \vec{v} \leq \vec{v}_{max}$$

where c_i indicates the objective coefficient value for the reaction i, v_i represents the flux value of the reaction i; the two vectors $\vec{v}_{min}$ and $\vec{v}_{max}$ represent the lower and upper bounds vectors, expressing for each flux v_i of the vector $\vec{v}$, the range within it can vary. The resulting output is a particular flux distribution $\vec{v}$ that, according to the above mathematical expression, maximizes or minimizes the defined objective function according to all the constraints imposed on the model.

COBRA and CobraPy Toolbox are currently the most used tools, which have been respectively developed in Matlab and Python programming language, for performing constraint-based modelling.

FBA has the advantage of neglecting any information regarding the kinetic parameters of the system, allowing to compute flux distributions, in a short time, even for very large scale metabolic networks. At the same time, disregarding these information, system dynamics, as well as regulatory effects relating to enzyme activation or gene expression, cannot be investigated.

In spite of constraints imposed on metabolic models, the solution identified by simplex algorithm may not be unique. This means that a number of alternative and different flux vectors, i.e. FBA solutions, may give the same optimal objective flux value. These flux vectors are called alternative optimal solutions and represent equally optimal phenotypic states. These alternative optimal solutions can differ among them from a qualitative and/or quantitative

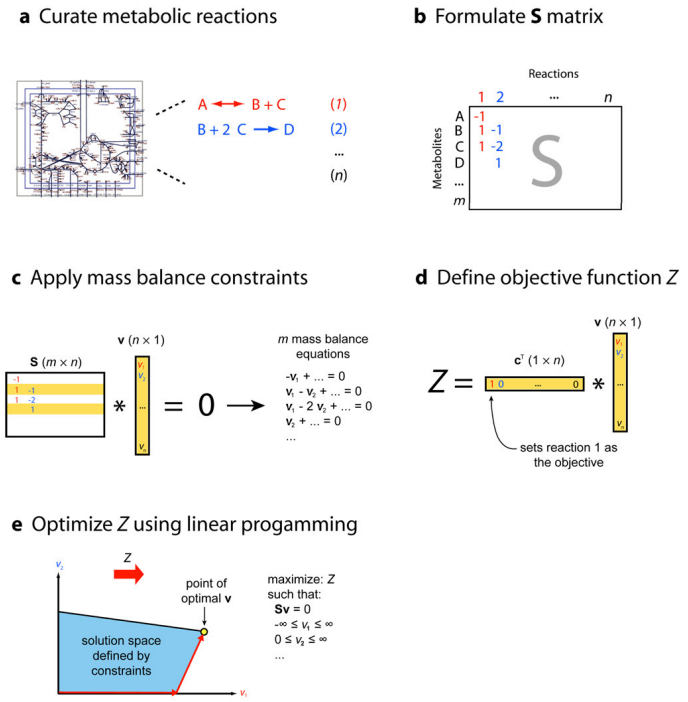


Figure 2.8: Steps of Flux Balance Analysis (FBA). (a) Reconstruction of the metabolic network. (b) Mathematical representation of the metabolic network as a stoichiometric matrix S. Each row of the matrix corresponds to a metabolite and each column to a reaction. (c) Application of constraints on the network. (d) Definition of an objective function Z. (e) Linear programming is used to optimize Z respecting the imposed constraints. Image taken from [60].

point of view. The first case occurs when at least flux value of one reaction varies between different solution, whereas qualitatively different alternative solutions occurs when different metabolic routes are exploited. The emergence of alternative optimal solutions especially occurs within large-scale networks because of redundant pathways that can create alternative ways to move from one point to another one in the network that equally optimize the objective function. In this context, the Flux Variability Analysis (FVA) [64] returns the range of flux variation for each biochemical transformation included in the investigated metabolic model. Allowable interval boundaries are computed by firstly setting the flux of the objective function to its optimal value. After that, two FBA are performed for each reaction of the network to firstly minimize and then maximize it. In this way, this analysis returns for each reaction the minimum and the maximum limit of its flux range. Acting in this way, FVA identifies the possible presence of redundancy in the network under investigation, even if it does not list all the alternative flux distributions for which other computational tools may be exploited. In a first scenario that can emerge from FVA, the range boundaries of a specific reaction coincide. This means that this biochemical transformation is crucial in the network under investigation, and must be used to optimize the chosen objective function given all the constraints imposed on the system. In a second case, the use of this reaction is variable and, consequently, its flux value can vary among alternative optimal solutions, or, in extreme cases, the reaction can either be used or not among alternative solutions. A third situation that can emerge from FVA is when both minimum and maximum allowable flux for a reaction are equal to zero. This result implies that no flux can pass through the reaction because of its null contribution to the optimization of the chosen objective function. When the same scenario recurs without imposing any optimality constraint within the metabolic network, it is possible that blocked reactions are present within the network because of gap reactions. The identification of blocked reactions may result useful for further network refinement, in order to identify and then remove eventual gaps and related dead-end metabolites.

Parsimonious FBA (pFBA) [65] is a variant of classic FBA approach. The underlying assumption of this approach is that, under growth pressure, there is a selection for the fastest growing organisms that minimize the total necessary resources to implement the optimal solution. Consequently, pFBA aims at finding the most stoichiometrically efficient pathways. Among all the feasible solutions, pFBA choose the one that minimize the total flux through all reactions in the metabolic network. To achieve this aim, first of all a new stoichiometric matrix S' is extracted for the metabolic network, where all reversible reactions are split into two irreversible reactions. Classic FBA is then carried out to compute the

optimal value of the reaction choosen as objective function. Finally, keeping this flux value fixed, the sum of all reactions fluxes is minimized. pFBA can be thus summarized as follows:

$$\begin{gathered} \text{minimize } \sum_{j=1}^{m} v_j^{'}, \\ \max v_{OF} = v_{OF,lb}, \\ S^{'}\vec{v^{'}} = 0,\ 0 \leq \vec{v_j^{'}} \leq \vec{v_{max}} \end{gathered}$$

where m is the number of irreversible reactions in the network under investigation, v_{OF} is the flux value of the objective function, and $v_{OF,lb}$ is the lower bound for the objective function reaction.

2.5.3 Mechanism-based modelling

Mechanism-based modelling, differently from previously discussed approaches, allows to obtain detailed quantitative predictions about the dynamics of biological systems. More in detail, it allows to determine the temporal evolution of each molecular species included in the system under investigation [48, 57]. This outcome needs of a series of data whose unavailability compromise the execution of these simulations. In particular, it is fundamental to know the functional interactions established among the species included in the simulated network. Moreover, all kinetic constants for each chemical reaction belonging to the model, and the initial molecular concentration of all the involved species to define the initial state of the system must be known. The requirement of these parameters limits the applicability of mechanism-based modelling to small-scale models because of the difficulty in retrieving these quantitative parameters for an extensive number of reactions [48]. This approach allows to test the behaviour of the system, by simulating the effect of different perturbations. For example, by varying kinetic constants of specific reactions, the effect on the system of temperature or catalyst variation on reaction rate can be simulated. The alteration of the initial state of the system is another type of perturbation to evaluate the effect on the system of the variation of medium composition.

Systems of ordinary differential equations (ODEs) represent the most used mathematical frameworks for simulating the dynamics of biological systems through mechanism-based modelling. These equations describe, in a formal way, the time variation of each molecular species. Cellular species concentrations are described by means of continuous variables, and are the result of the influence of each chemical reaction processing that compound. Given all the initial parameters previously introduced, the dynamics of the system can be

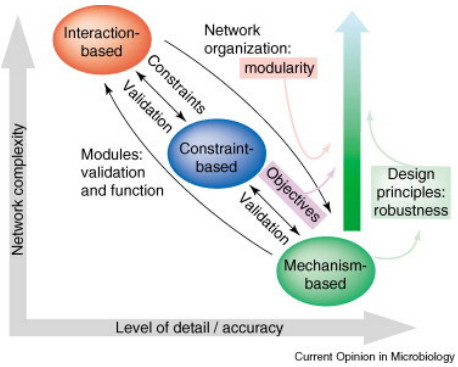


Figure 2.9: Graphical cataloguing of interaction-based, constraint-based and mechanism-based modelling according to their level of detail and accuracy, and network complexity. Black arrows indicate te possible interactions among the three modelling approaches. The green arrow refers to a possible strategy integrating the strengths of present modelling methods with modularity, optimality and robustness. The final outcomes are mechanistic and genome-scale models that allows to gain new knowledge at system-level. Image taken from [57]

simulated by numerically solving the corresponding set of differential equations.

As previously discussed, constraint-based modelling is located halfway between interaction-based and mechanism-based approaches on several points. As shown in Figure 2.9, interaction-based and mechanism-based approaches can be positioned on opposite sides in terms of the achievable detail and accuracy level of the simulated model, and of the network size they can handle. Indeed, interaction-based modelling, without being computationally demanding, is generally exploited for networks having a significant degree of complexity but a low level of detail. On the contrary, mechanism-based modelling is generally exploited for the study of small-scale networks because of the high number of required parameters that are often missing at genome-wide level. Consequently, this computational approach can handle networks whose covered level of detail, and consequent accuracy, are very high, whereas due to their size the complexity level is very low. At the same time, because of the required parameters, mechanism-based modelling results more computationally demanding.

Chapter 3

New constraint-based methods for homogeneous metabolic systems

The works that are presented in this chapter are focused on the investigation of cell populations metabolism, by assuming them to be homogeneous systems. This means that metabolism of a single cell or organism as representative of the entire population to which it belongs is characterized.

Cell-to-cell variability is known to be constantly present in any population of cells [11]. By assuming the system under investigation as homogeneous, we are not disregarding genetics, environmental and spatial dishomogeneities within the population. On the contrary, we are just pointing to hide cell-to-cell differences to lower the overall complexity level of the investigated system. In this regard, I will show in the next sections how computational analyses have benefited from this assumption without invalidating the biological validity of the *in silico* outcomes.

As widely discussed in Chapter 2, the fundamental building block that is essential to address cell metabolism investigation from a computational point of view is the reconstruction of a high quality and curated metabolic network. In this regard, genome-scale metabolic networks represent the basis for investigating the metabolic potential of a given cell or organism since they organize all the available knowledge about their metabolic transformations [50].

In the last two decades, genome-wide reconstructions of metabolism have

been produced for a plethora of model organisms, spanning from bacteria to higher eukaryotes [55]. Nevertheless, genome-scale networks are not yet available for multiple known organisms. For this reason, I will present in Section 3.1.1 a computational pipeline for the automatic reconstruction of genome-scale metabolic networks for specific target organisms. For the implementation of this methodology, the available knowledge stored in biological databases for the target organism coupled with the corresponding sequenced and assembled genome are exploited. In particular, I will show an application of this methodology for the reconstruction of the genome-wide metabolic network of the yeast *Zygosaccharomyces parabailii*, based on its recent whole-genome sequencing and annotation [66], and novel wet-lab data from chemostat cultivation. Constraint-based modelling revealed adherence of computational simulations of our model to both experimental data and literature evidence.

The reconstruction of genome-scale metabolic networks is not a novel task since several tools currently exist. These approaches differ among them on several points, including the automation level and the biological databases used to generate the final network, such as for example BiGG [67], KEGG [68], modelSEED [69], and BioCyc [70]. One of the latest proposed approach, named CarveMe, has been recently presented in [71]. The reconstruction pipeline implemented in this work is very different from our approach. In particular, CarveMe aims at automatically reconstructing a universal metabolic model that can be then used as template to generate organism-specific networks. In this work, the BiGG database is exploited as main source to build the universal model. This database integrates data deriving from a few eukaryotic genome-scale metabolic reconstructions. In this way, BiGG limits the applicability of CarveMe approach to a very reduced portion of living organisms, including *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae*. Differently from CarveMe, our reconstruction approach exploits KEGG as main biological database, resulting in a higher coverage of both prokaryotic and eukaryotic organisms. BiGG database is also limited compared to KEGG in terms of size and scope. Indeed, the mapping of BiGG metabolic reactions into KEGG global pathway map revealed a lack of coverage of pathways associated to secondary metabolism. Since it is known from literature the relevance of secondary metabolites as important player of several eukaryotes functions within plants, yeast and human [72], our strategy of using KEGG database allows to generate more realistic phenotype predictions. Another feature of CarveMe approach regards the biomass synthesis reaction included in the network. In this regard, it provides a simulation-ready model including a universal biomass reaction. Although this strategy seems to be a compromise when an organism-specific biomass compos-

ition is not available, the inclusion of generic data could negatively affect the outcomes of model simulations by providing false predictions. Therefore, our strategy of using organism-specific data, or at least information deriving from its close neighbours, is another key element to achieve high quality genome-scale metabolic networks.

Another recent semi-automatic reconstruction method is the RAVEN (Reconstruction, Analysis and Visualization of Metabolic Networks) toolbox, which is a MATLAB-based toolbox for constraint-based metabolic modelling [73]. In [74], a new version called RAVEN Toolbox 2.0 presenting new features is introduced. In particular, RAVEN 2.0 was improved in order to assist de novo semi-automated draft model reconstructions for a given organism of interest. By taking as input a FASTA format file with whole-proteome sequences of the target organism, two draft models are generated following the integration of knowledge from KEGG and MetaCyc pathway databases. MetaCyc is a curated metabolic pathway database that contains only experimentally determined pathways from all domains of life. Once created, KEGG- and MetaCyc-derived draft models are combined into an integrated genome-scale model.

RAVEN 2.0 generates a genome-scale model consisting of a larger number of reactions compared to the first draft model generated through our developed approach. Although the potentialities of this toolbox in generating a larger metabolic network, the increased coverage does not necessarily entail a more realistic network. In this regard, some criticisms arise from our testing of this reconstruction pipeline with the whole-genome annotation of our target organism *Zygosaccharomyces parabailii*.

According to RAVEN 2.0 pipeline, transport reactions should be included into the model due to the availability of curated transport enzymes in MetaCyc database. Although very few transport reactions are included, all network metabolites are assigned to a unique compartment called "System". This implies that the included transport reactions are fake reactions consisting of an empty set of reactants and products due to the involvement of the same metabolites that causes an empty stoichiometry for the reaction. Therefore, differently from our pipeline, RAVEN 2.0 is not able to generate compartmentalized genome-scale metabolic models, limiting the applicability of the approach to a reduced portion of living organisms. Moreover, the authors of RAVEN 2.0 declares that the integrated genome-scale model derived from KEGG and MetaCyc database is just a draft of the final model that needs additional manual curation to form a high-quality reconstruction. Finally, similarly to our approach, any information regarding the relationships linking the genes involved in each reaction is not included in the network. Consequently, the GPR rule of each reaction is

constructed by joining the involved genes by means of the OR boolean operator, without forcing genes to co-exist.

Since a larger curation phase is needed on the generated model from RAVEN 2.0 toolbox, we are prone to consider our reconstruction pipeline a better and more curated alternative approach.

The systemic perspective proposed by systems biology is well represented by genome-scale metabolic networks. However, the comprehensiveness of these reconstructions co-exists with the difficulty in their managing. In this regard, the high amount and interconnectivity of included reactions limit a straightforward interpretation, from a biological point of view, of the outcomes resulting from their computational simulations, due to the difficulty in rationalizing such amount of data. Moreover, the further inclusion in these networks of wrong or incomplete knowledge about the organism under investigation, may results in errors, such as metabolic gaps, which are often incorrectly filled through the addition of exchange reactions. However, in this way, isolated subnetworks are more prone to be formed, with the consequent increased risk of compromising the simulations outcomes.

Greater control on these reconstructions can be achieved by shifting the focus from genome-scale to core metabolic networks. Therefore, I will show in Sections 3.2.1- 3.2.3 how to reconstruct core metabolic networks by using genome-scale metabolic networks as scaffold for selecting and then extracting the reactions that are regarded as most relevant for dealing and exploring a specific biological task. In this thesis work, in particular, I will present three applications of core modelling as effective means for uncovering system-level properties of cancer metabolic rewiring. Core networks relative to human central carbon metabolism have been extracted from recent human genome-wide metabolic networks, by also exploiting literature knowledge to assist the reconstruction process. Given the controllable size of core metabolic networks, a more accurate curation phase can be carried out regarding the directionality of the included reactions, as well as the presence of gap reactions, by relying on multiple biological databases and recent versions of human genome-wide metabolic reconstructions. Core modelling helped in revealing heterogeneous metabolic rewirings supporting neoplastic proliferation among multiple types of tumours, and under alternative nutritional conditions and different perturbations.

A common procedure is followed for the reconstruction of all the investigated core metabolic models. The two required inputs are a genome-scale metabolic model about the target organism, and the set of metabolic pathways that are relevant to include in the network according to the biological problem under investigation. The reconstruction workflow involves five steps. Firstly, according

to the input list of target pathways, the corresponding reactions are extracted from the initial genome-scale model in order to create a first draft of the final network. The reaction list further includes a biomass composition formulation according to the corresponding stoichiometry provided in the original genome-scale model or derived from experimental data. Depending on the target organism, it could be necessary to add transport reactions to move metabolites that are assigned to multiple compartments. In the third phase, an accurate manual curation of the draft network is performed to identify and fill network gaps, and to check the correctness of reactions in terms of their stoichiometry and directionality. In this regard, pathway databases and up-to-date metabolic reconstructions are exploited. The network curation is then followed by the lumping of reactions belonging to a linear pathway without any branching to a unique reaction that is characterized by the net stoichiometry of the reactions set. In the last step, all the dead-end metabolites are identified and for each one an exchange reaction is included. Depending on the role of the dead-end metabolite in the network, an entry or an exit exchange reaction is included.

A systematic reduction procedure, called redGEM [75], of genome-scale models for constructing core metabolic models has been recently presented to deal with the lacking of consistent criteria about developing of core models, which takes the issue in a similar manner to our. In this work, the authors proposed a bottom-up approach that, starting from a genome-scale model, a set of target metabolic subsystems, medium components and available physiological data, generates a reduced model while maintaining a consistency with the corresponding genome-scale reconstruction in terms of the generated knowledge. After the identification of the core network, redGEM exploits the algorithm called lumpGEM [76] to identify all the alternative minimal sized subnetworks that are able to produce a cellular metabolite from a defined set of metabolites. More in detail, lumpGEM is used to include into the core network all the alternative balanced reactions for the synthesis of the biomass building blocks. This algorithm derives for each subnetwork a unique lumped reaction that includes the overall stoichiometry of the subnetwork.

Another recent pipeline to perform the same task is NetworkReducer[77]. Contrary to redGEM, NetworkReducer is a top-down reduction procedure that aims at reducing in an automatic way genome-scale metabolic reconstructions in order to obtain smaller core models and capture given metabolic modules of interest. NetworkReducer takes as input a large-scale network together with a list of and a list of protected metabolites, reactions, functions and phenotypes that must be retained in the reduced network. From these inputs, the network reduction algorithm initially performs a preprocessing step to check the feasib-

ility of the protected functions in starting network and to remove non-protected blocked reactions. After that, a pruning step is perform to iteratively remove non-protected reactions with smallest flux range until no further reaction can be deleted without violating the user-defined protected elements. Finally, a compression step is applied to collapse reactions from linear pathways if they are not protected.

Both these systematic reduction approaches are indeed interesting while only NetworkReducer was available at the time of the publication of our core model reduction work [78]. Nevertheless, the underlying criterion for NetworkReducer tool is opposite than ours. This requires an initial high quality genome-scale model from which the final core model quality strictly depends. On the contrary, our approach results closer to redGEM and lumpGEM tools. However, these tools have been just recently presented and our reconstruction pipeline has been developed independently from them.

3.1 From genome annotation to genome-wide metabolic models

In this section, I present the computational pipeline that we developed for the automatic reconstruction of genome-scale metabolic networks. Although this procedure is mostly automated, some steps still require a manual supervision. For this reason, we defined the implemented pipeline as semi-automated, even if we are currently working to make it fully independent from the manual intervention. All the computational tools have been implemented in the open source license Python programming language by exploiting, when possible, the existing libraries, and, otherwise, writing *ex novo* the code. Overall, the algorithm underlying the entire reconstruction procedure can be summarized in seven steps:

1. **Annotation of the target organism genome.** The first step aims at annotating the complete sequenced and assembled genome, as one of the input data of this procedure, in order to highlight all the encoded metabolic functions. In this regard, the BLAST tool, which is part of the NCBI database, is exploited to perform annotations of all genes products of the investigated organism based on the sequence homology with proteins having a known function within phylogenetically close species. The Python library called Biopython [79] was very useful for that purpose because of its ability to allow a programmatically access to several databases.

Therefore, we exploited this tool to access NCBI and KEGG databases, automatically perform the annotation, and finally extract for the retrieved list of homolog sequences all the available information that are present in KEGG.

2. **Population of the first draft model.** The second step aims at exploiting the information obtained from KEGG database, especially those regarding genes assignment to a specific EC number and functional hierarchy, in order to filter out all the genes whose function is outside the metabolic field. Having access to these data, it is then possible to extract all the chemical reactions associated to the filtered metabolic genes and constitute a list of candidate metabolic reactions to populate a first draft of the final model. The annotated genes from the first step lacking a reference to KEGG have been processed to check the presence of "false negative genes" , namely genes that have been erroneously catalogued as KEGG identifier missing. The information about the organism name where the homology occurs for a given gene is fundamental for the purpose of automatically converting the NCBI items into the corresponding KEGG ones. However, in some cases, the different nomenclature about organism names between these two databases implies that the automatic conversion of the identifiers fails, and, consequently, false negatives come out. For this reason, this step requires a manual supervision, following which these genes are then handled as previously described. Regarding genes that for sure miss a KEGG database reference, as much information as possible about their role are extracted by exploiting similarities with the other already processed genes, in terms of their annotated function.

3. **Refinement of the draft model.** Once all the set of identified metabolic reactions are joined together, the third step aims at refining the draft model using information from genome-scale metabolic models of phylogenetically close organisms in terms of reaction stoichiometry and compartmentalization. In particular, this latter is not provided by KEGG database. Moreover, reactions from these genome-scale models whose GPR rule is fulfilled according to the list of target organism genes are also included in the model draft. When the identified homolog genes linked to a specific reaction belong to organisms without an associated genome-scale metabolic model in literature, the stoichiometry of the reaction is taken from KEGG database, whereas its localization is assigned to the cytosol compartment, unless otherwise stated in the Uniprot database [80]. Manual supervision is also required to curate reactions that in the KEGG

database are catalogued as "general reaction" because generic substrates and products are included.

4. **Addition of the necessary transport reactions.** The fourth step aims at adding to the draft model transport reactions for metabolites assigned to multiple compartments. If this information is currently lacking for the organism under investigation, reversible and unbounded transport reactions are included to join the same metabolites that are located in different compartments within the network under reconstruction.

5. **Integration of experimental data.** This optional fifth step is performed depending on the availability of chemostat and medium composition data. In the first case, an exchange reaction is added for every metabolite whose production or consumption rate has been experimentally determined from chemostat cultivation experiments, and their boundaries are constrained according to measured values. In the second case, the experimental growth medium is *in silico* mimicked through the definition of the set of corresponding exchange reactions, whose maximum uptake rate is constrained according to the concentration of the corresponding metabolite within the experimental medium.

6. **Gap filling.** The sixth step aims at finding candidate reactions to fill gaps in the network. The corresponding dead-end metabolites, by definition, are involved one or more times within the network, but only as substrates or as products. Consequently, dead-end metabolites are consumed but never produced, or vice versa. The list of possible gap filler reactions that is automatically retrieved from KEGG is processed to verify that the joined dead-end metabolites belong to an equal compartment into the model under construction, but without being involved both as substrates or products. Moreover, it is necessary to check that the reaction is not already included into the draft model.

7. **Inclusion of the exchange reactions to achieve mass-balance in the system.** Following the gap filling step, it is possible that some gaps are not filled and dead-ends metabolites persist within the network. This may happen when some portions of metabolism of the target organism have not been yet characterized. In alternative, gaps cannot be filled when biological databases, which are use as primary source of information for network reconstruction, result inaccurate in terms of information they contain. The last scenario may occur when any chemical reaction exist

in addition to those already included in the network for processing one of the identified dead-end metabolites. This is the case of metabolites that need to be consumed or released from/to the extracellular environment. Nevertheless, to fulfill the mass balance constraint, dead-ends metabolites must not be present within the network. For this reason, the seventh and final step of the reconstruction process aims at performing a new search for these metabolites, and for each one an entry or exit exchange reaction is included into the network. In case of experimental data missing to constrain the lower bounds of entry exchange reactions, the value is inferred through the Particle Swarm Optimization (PSO) in order to minimize the distance between the experimental and the computational biomass yield. The choice of using the PSO to perform this task is motivated by the higher performance of this optimization algorithm with respect to other existing ones, e.g. genetic algorithms, as discussed in [81], and because it was already in-house available.

The entire protocol is thoroughly presented in the currently under submission paper that is reported in Section 3.1.1. As anticipated, we applied the described pipeline to reconstruct the genome-scale metabolic model of the hybrid yeast *Zygosaccharomyces parabailii*, by adapting the reconstruction process to the available data. Among them, the whole-genome sequence and annotation recently published for *Z. parabailii* [66], and chemostat experimental data performed at two different dilution rates, where the metabolism of the investigated organism has been experimentally characterized as opposed since a fully respiratory and respiro-fermentative regime emerged. In particular, the two steady states were defined in terms of principal metabolites that have been found to be consumed or secreted in the extracellular environment, and data about the experimental growth medium. Moreover, the macromolecular biomass composition that has been experimentally determined at the two considered dilution rates allowed us to reconstruct a specific biomass synthesis pseudo-reaction in terms of both its components and stoichiometry.

The decision of opting for this specific hybrid yeast came from its experimentally described high tolerance to different types of stress, including organic acids like acetic acid. Acetate is released together with other toxic compounds when lignocellulose is used as a substrate for the development of sustainable bioprocesses. Once released, acetate act as an inhibitory compound for most microbial cell factories, but *Z. parabailii*. Moreover, this organism has been previously demonstrated to be a flexible platform for the production of recombinant proteins and non-natural metabolites, including lactic acid [82], and ascorbic

acid [83]. These features together with its hybrid nature make this organism a very attractive host for chemicals production. Therefore, we decided to characterize its metabolism at the genome-scale level.

The implemented reconstruction protocol returned the "ZyPa1" model, consisting of 3096 reactions, 2091 metabolites and 2413 genes. Moreover, we associated each reaction to genes that are for it responsible. Since any information regarding the relationships linking the genes involved in each reaction is so far not available, we constructed the GPR rule of each reaction by joining the related genes with the OR boolean operator. In this way, for now, we are not forcing genes to co-exist so that the reaction can occur. Because of the two considered dilution rates coupled with the corresponding different sets of experimental data, we generated two versions of ZyPa1 model, which we refer as ZyPa1_0.1 and ZyPa1_0.3.

The comparison of our model with the two genome-scale reconstructions of *S. cerevisiae* and *K. lactis*, which are two phylogenetically close organisms of *Z. parabailii*, revealed a doubled number of genes assigned to ZyPa1 model compared to the other two networks. This result could be partially attributed to the interspecies hybrid nature of *Z. parabailii*, which presents two copies of most genes deriving from each of the two independent parental genomes, together with possible expansion of copy number of some genes.

Regarding FBA simulations, by setting biomass synthesis reaction as the objective function to be maximized, this first version of the model proved to be able to describe the experimental biomass yield at a quite good level. Moreover, the reaction deletion analysis revealed a different reduction of the biomass synthesis flux value when we simulated the *in silico* deletion of the reaction catalyzed by the L-serine hydro-lyase enzyme. This reaction, which is connected to the metabolism of L-Cysteine, L-Methionine and L-Serine amino acids, may represent a metabolic rewiring under a fermentative regime when source of sugars is depleted and amino acids can be used as energy and carbon sources. Finally, the model correctly replicated the already known high tolerance of *Z. parabailii* to the acetate organic acid. The ability to follow the co-consumption and catabolism of acetate and glucose, without a significant impact on the biomass yield has been described by the model. This *in silico* evidence suggested that catabolism of this organic acid can contribute to its detoxification, due to its toxicity for most microbial cell factories.

3.1.1 Genome-scale metabolic reconstruction of the stress-tolerant hybrid yeast *Zygosaccharomyces parabailii*

Di Filippo M, Raúl A. Ortiz-Merino, Chiara Damiani, Gianni Frascotti, Danilo Porro, Kenneth H. Wolfe, Paola Branduardi, Dario Pescini

Manuscript currently under submission to *mSystems*.

DOI: 10.1101/373621

Abstract

Genome-scale metabolic models play an increasing role in the success of industrial applications exploiting the highly diverse yeasts' metabolic traits. Among these eukaryotes, the hybrid yeast *Zygosaccharomyces parabailii*, member of the *Z. bailii* sensu lato clade, is particularly stress-tolerant to high osmotic pressure and organic acids, thus representing a very attractive host for chemicals production. Moreover, its recent genome assembly revealed how a full sexual cycle was restored, supporting the hypothesis that whole-genome duplication is the mechanism by which interspecies hybrids can regain fertility. We here present the first genome-scale metabolic model of *Z. parabailii*. Novel data on cellular composition and extracellular fluxes were produced to constrain the model simulation. The model proved able to reproduce the co-consumption of acetate and glucose: by using it as a predictive platform, the correlation between genotype to phenotype can result in novel knowledge and exploitation.

Introduction

Current genome sequencing technologies allow a fast and cheap overview into the genetic composition of virtually any organism, but connecting such genotypes to observed phenotypes remains a challenge. The reconstruction of genome-scale metabolic networks provide structured frameworks to represent the biochemical transformations within a target organism as complex genotype-phenotype relationships. Afterwards, different modeling approaches can be used to simulate, understand, predict and eventually control the behavior of such genome-scale metabolic networks. Flux Balance Analysis (FBA) is a widely used constraint-based modeling approach that relies on linear programming and the optimization of a given objective function (e.g., maximization of growth) for the determination of the metabolic model flux distribution [48, 60]. This approach is based on the assumption that organisms operate under a series of constraints limiting their possible functions, and leading to the definition of allowable cell phenotypes from defined metabolic networks [84].

In the last two decades, genome-wide reconstructions of metabolism have been produced for a plethora of model organisms, spanning from bacteria to higher eukaryotes [55]. The original versions of these models typically undergo incremental improvements. In particular, 10 different versions of the *Saccharomyces cerevisiae* metabolic network have been produced to date, by implementing cellular compartments, curated reactions, standard nomenclature, and even transcriptional regulation, as reviewed in [85]. However, less extensive efforts

have been dedicated to other so-called non-conventional or non-*Saccharomyces* yeast species, despite their relevance for biotechnological applications, as well as for basic and biomedical research. The few non-conventional yeast species for which genome-scale metabolic reconstructions have been made to date are: *Pichia pastoris* [86, 87, 88, 89], commonly used as a host for recombinant protein production, *Scheffersomyces stipitis* [90, 91, 92], relevant for biomass bioconversion, the fission yeast *Schizosaccharomyces pombe*, especially relevant for cell cycle control studies [93], the opportunistic pathogens *Candida glabrata* and *C. tropicalis* [94, 95], used for drug target prediction and lipid production, *Yarrowia lipolytica* [96, 97, 98, 99], relevant for its ability to produce and accumulate lipids potentially useful for many applications among which biofuel production, and *Kluyveromyces lactis* [100, 101], remarkable for its characteristic lactose metabolism and its use for heterologous protein production.

These models have however all been reconstructed starting from a reference *S. cerevisiae* model. Nevertheless, it is important to mention that yeast biodiversity is huge, coherently with their phylogenetic distance. In addition to that, many interspecies hybrids have been discovered [66], and, as expected, their phenotypic traits are the result of how the complex genome is translated in metabolic functions and networks. Some of these hybrids are very important at industrial level, as for *Saccharomyces pastorianus* in brewing [102]. Therefore, generally speaking, diverse *in silico* models are necessary to comprehend yeasts biodiversity, and this comment becomes even more relevant in the case of hybrid yeasts, since they often express traits that are emerging properties of their peculiar genomic asset.

Species belonging to the genus *Zygosaccharomyces* are known for their exceptional tolerance to different types of stress. In the case of *Z. bailii sensu lato*, high tolerance to high sugar concentration, low water availability, low pH and weak acids make this yeast notorious as frequent agent of food spoilage [103, 104]. However, the same tolerance traits can also be seen as beneficial for the microbial-based production of otherwise aggressive compounds such as lactic acid [105]. In addition, when lignocellulose is used as a substrate for the development of sustainable bioprocesses, different toxic compounds are released, among which acetic acid [106]. Also in this case, *Z. bailii* has been demonstrated to be particularly robust, being this robustness ascribable to diverse mechanisms and structural adaptations [107].

Among the species belonging to the *Z. bailii sensu lato* clade, we focused on *Zygosaccharomyces parabailii*, which we previously demonstrated as a flexible platform for the production of recombinant proteins and non-natural metabolites, including lactic acid [82], and ascorbic acid [83]. It is characterized by a

hybrid nature [108, 109], as also revealed by our recent genome assembly and annotation, which was used to interprete transcriptional profiles upon lactic acid stress [66]. Moreover, *Z. parabailii* can help in understanding the molecular basis of how genome doubling and restoration of a complete life cycle including a sexual recombination occurred during the evolution of a possible ancestor of the yeast *S. cerevisiae*. Indeed, we recently described that *Z. parabailii* derives as a hybrid from the mating of two divergent haploid parental species. This hybrid, initially sterile, regained fertility as a consequence of a damage to one copy of its mating-type locus [66].

Our model can be therefore pivotal for making progress into the simulation and understanding of the high stress tolerance of this yeast species, including the description of the possible different contribution of the two divergent haploid parental genomes to the nature of the resulting hybrid. Moreover, this approach can pave the way for unveiling how to better exploit *Z. (para)bailii* for industrial purposes.

Results

Reconstruction of *Z. parabailii* metabolic model

We started the reconstruction process from the high-quality whole-genome sequence and annotation of *Z. parabailii* ATCC60483 [66]. We used the information in the KEGG database to filter out non-metabolic genes, and to extract the enzymatic reactions associated to EC numbers identified by our previous functional annotation [68, 110]. In this way, we selected 807 metabolic genes that were used to build the first draft of the *Z. parabailii* genome-scale metabolic network. We then obtained the stoichiometry and localization for the *Z. parabailii* reactions from the *S. cerevisiae* Yeast7 and the *K. lactis* iOD907 models when possible [100, 111]. Otherwise, the stoichiometry was derived from the KEGG database and the compartmentalization information was obtained from Uniprot or manually curated. Transport and exchange reactions were added to the draft metabolic network assuming a "matriosca"-like structure for the cellular compartments. In this way, the extracellular environment represents the most external level containing the cytosol, which in turn includes all the other considered compartments of the model, which are mitochondrion, endoplasmic reticulum, lipid particle, cell envelope, nucleus, peroxisome, Golgi apparatus and vacuole. If a metabolite is available in multiple compartments, we added an exchange reaction only for the metabolite in the outermost compartment. Since no data are available for constraining these reactions, when the exchange

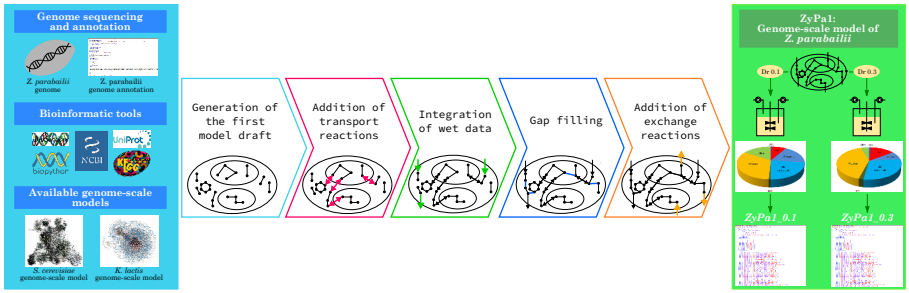


Figure 3.1: Graphical representation of the model reconstruction procedure.

reactions imply the consumption of a given metabolite from the extracellular environment, we set the lower bound to $5*10^{-5}$ mmol gDW^{-1} h^{-1} according to the coefficient inferred through the Particle Swarm Optimization (PSO) to minimize the distance between the experimental and the computational biomass yield. The proposed reconstruction process is summarized in Figure 3.11.

Overall, the reconstruction process resulted in a model consisting of 3096 reactions (1743 of which are internal reactions, 619 are transport reactions and 734 are exchange reactions), 2091 metabolites and 2413 genes. Each reaction was associated to a GPR rule by joining the related genes with the OR boolean operator. Reactions and metabolites included in the model are localized over 10 different cellular compartments. The final version of the model, which we refer as ZyPa1, also includes the biomass synthesis reaction that we reconstructed according to the macromolecular biomass composition that we experimentally determined at the two dilution rates of 0.1 and 0.3 h^{-1}. Because of the different stoichiometry in the two biomass synthesis reactions, we generated two versions of ZyPa1 model, one for each dilution rate, which we refer as ZyPa1_0.1 and ZyPa1_0.3, correspondingly.

We constrained the extracellular environment of the two ZyPa1 models according to experimental data. In particular, the exchange rates of glucose, oxygen, carbon dioxide, glycerol, acetic acid, acetoin and ethanol have been constrained based on the data obtained from chemostat cultivation at the two dilution rates of 0.1 and 0.3 h^{-1}. They have been determined by scanning different dilution rates comprised in this interval. In Figure 3.2B, the experimentally determined metabolic profiles are given. A second set of constraints on the extracellular environment concerned the *in silico* medium composition

in order to replicate the growth on the experimental medium, by adapting it to the model composition. Therefore, we integrated an exchange reaction for ammonium, biotin, (R)-pantothenate, nicotinate, myo-Inositol, thiamine, pyridoxal, 4-aminobenzoate and iron.

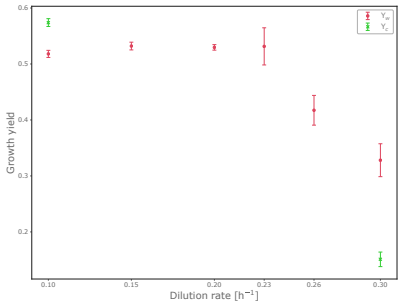
We checked for the presence of blocked reactions into the two ZyPa1 models in order to find all points in the network through which no flux pass, meaning that, following a FVA simulation, minimum and maximum flux of these reactions are equal to zero when no objective function is set. Currently, 75 and 72 reactions, *i.e.* about 2% of the total number of reactions, emerged as blocked respectively in the ZyPa1_0.1 and ZyPa1_0.3 models, with the only difference between them regarding the three exchange reactions of glycerol, acetoin and ethanol that we imposed equal to zero in ZyPa1_0.1 model according to the experimental data. The identified blocked reactions are caused by gaps of knowledge in the metabolic pathways of the target organism. However, since a genome-scale metabolic model, as such, integrates all the knowledge about metabolism of a given organism, these reactions are integral part of the model itself even if they are not, currently, able to carry flux.

The ZyPa1 model was deposited in BioModels [112] and assigned the identifier MODEL1807110001.

Comparison of model with other genome-scale models

We compared the ZyPa1 model with the genome-scale models used as reference during the reconstruction process, *i.e.* Yeast7 and iOD907, in order to highlight similarities and differences among the three networks. First, we calculated the size of the three networks in terms of involved reactions, metabolites and genes. As can be seen from Figure 3.3A, the number of reactions and metabolites that are included in the ZyPa1 model is closer to the corresponding value in *S. cerevisiae* than in *K. lactis*. However, it is worth to mention that the number of genes assigned to ZyPa1 network doubles that of Yeast7 and iOD907 models. This discrepancy is partially due to the interspecies hybrid nature of *Z. parabailii* as this organism has two copies of most genes, one deriving from each of the two independent parental genomes. In addition, some gene expansions occurred in the *Z. parabailii* genome. For example, in haploid *S. cerevisiae* cells, 3 genes encoding for the pyruvate decarboxylase activity, named *PDC1*, *PDC5* and *PDC6* have been annotated and described. Due to the hybrid nature of *Z. parabailii*, it was reasonable to expect to find 6 PDC genes, but the observed number of 12 suggests that in some cases expansion of genes number could justify the value we identified.

A



B

	Experimental data [mmol gDW⁻¹ h⁻¹] [g l⁻¹]		In silico constraints [mmol gDW⁻¹ h⁻¹]	
qO_2	2.81±0.03	5.95±0.11	2.81±0.03	5.95±0.11
qCO_2	2.67±0.08	8.69±0.02	2.67±0.08	8.69±0.02
Glucose	18.03±0.06	11.46±0.72	1.07±0.01	5.07±0.45
Glycerol	0	0.21	0	0.18±0.01
Acetic acid	0.75±0.13	0.18±0.04	0.13±0.02	0.24±0.05
Acetoin	0	0.06±0.02	0	0.05±0.02
Ethanol	0	2.82±0.43	0	4.88±0.81
	Dr 0.1	Dr 0.3	Dr 0.1	Dr 0.3

Figure 3.2: **Comparison between experimental and computational data.** (A) Computation of biomass yield at different dilution rates. Red circle points show the experimental biomass yield (Y_w) computed at the six dilution rates of 0.10, 0.15, 0.20, 0.23, 0.26 and 0.30 h^{-1}. Error bars for each point have been calculated according to experimental errors. Green x markers show the computational biomass yield (Y_c) computed at the two dilution rates of 0.10 and 0.30 h^{-1}. Error bars associated to each point is calculated according to glucose uptake flux variability. (B) Experimental data from chemostat cultivation at the two dilution rates of 0.10 and 0.30 h^{-1} are reported in the red part of the table. Different color are used to distinguish values that are measured in mmol gDW^{-1} h^{-1} (black) from values that are measured in g l^{-1} (blue). Computational constraints imposed on the two ZyPa1 models are reported in the green part of the table. All values are expressed as mmol gDW^{-1} h^{-1}.

A

Model	Reactions	Metabolites	Genes
ZyPa1	3096	2091	2413
Yeast7	3497	2223	909
iOD907	2180	2205	906

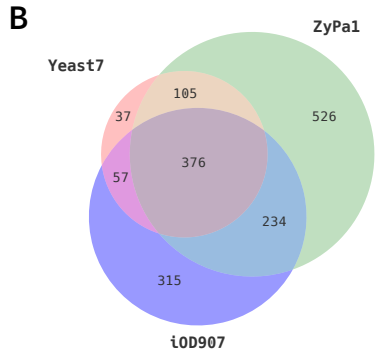


Figure 3.3: **Comparison of ZyPa1 model with genome-scale metabolic models of *S. cerevisiae* and *K. lactis*.** (A) Overview of the reactions, metabolites, and genes in the three models. (B) Venn diagram showing the overlapping reactions among the three models by using the related KEGG identifier. An unknown portion for each of the three network emerged for reactions having a missing reference to the KEGG database.

We extended the comparison among the three networks also at the level of the included reactions, and of pathways where they are involved. In Figure 3.3B, a Venn diagram shows the overlapping reactions among the three models by using, when available, the related KEGG identifier. An unknown portion for each of the three network emerged for reactions having a missing reference to the KEGG database. This portion is equal to 72% in Yeast7, 50% in iOD907 and 52% in our ZyPa1 model. By performing the comparison only on the known part of each network, we observed that ZyPa1 model does not rely as much on Yeast7, but rather on iOD907 model since the overlapping with *K. lactis* model is much higher compared to that with *S. cerevisiae* network.

For the set of reactions with a KEGG identifier, we collected all the metabolic pathways where they are involved. In Figure 3.4, the circle plot shows internally the hierarchy of all KEGG pathways related to metabolism. Each dead-end node corresponds to a specific pathway and the differently colored sections are needed for highlighting pools of pathways that are related one with each other. Around the outer circle, the relative frequency for each single pathway in the three genome-scale models over the total reactions number of the models themselves is shown. It is possible to observe variability among the three networks in terms of pathways distribution, and for some group of path-

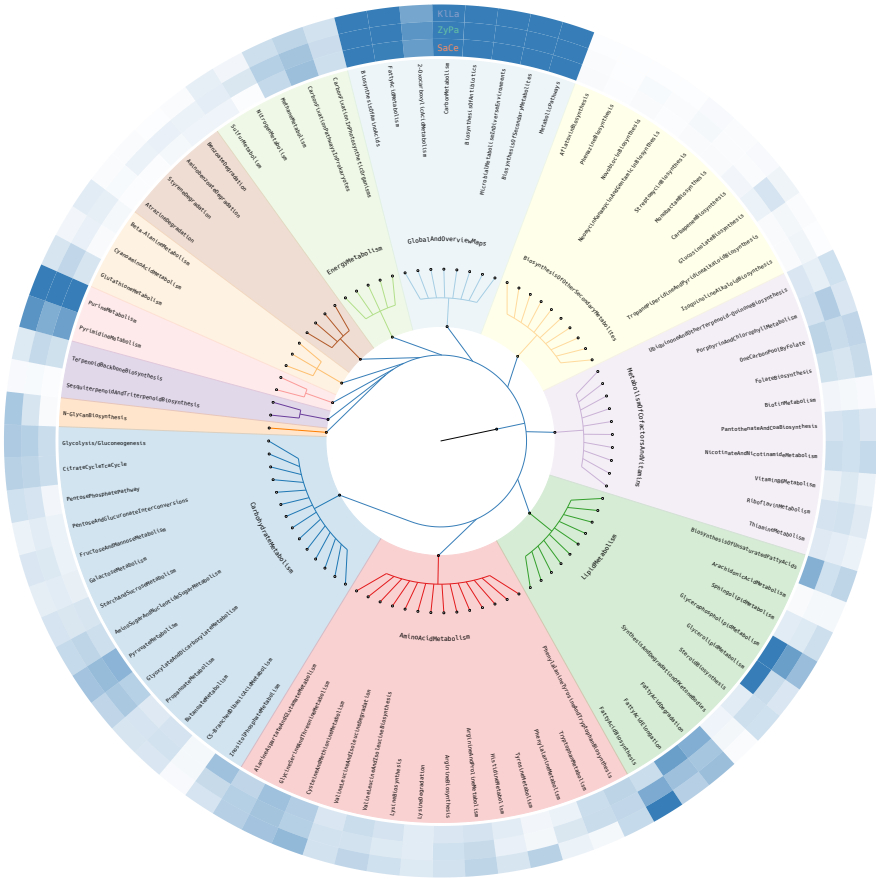


Figure 3.4: **Metabolic pathways distribution among ZyPa1, Yeast7 and iOD907 genome-scale models.** The circle plot shows the hierarchy of all KEGG metabolism related pathways included in the three models. Differently colored sections indicates pathways that are related one with each other. Each dead-end node corresponds to a specific pathway and the three outer concentric circles, starting from the innermost level, show the corresponding relative frequency in the genome-scale models of *S. cerevisiae* (SaCe), *Z. parabailii* (ZyPa) and *K. lactis* (KlLa).

ways the differences are more pronounced. This is the case of the group labeled as "LipidMetabolism". Focusing the attention on the comparison between *S. cerevisiae* and *Z. parabailii*, it is relevant to report that differences in phospholipids composition were proposed to be responsible of the different acetic acid membrane permeability in the two yeasts. More in detail, sphingolipids have been determined to be several times higher in *Z. bailii* than in *S. cerevisiae* and it was shown that upon acetic acid stress the membrane remodeling is more important in the non-*Saccharomyces* yeast and further exacerbates this difference [113]. By looking at the map, it does not seem to evidence a difference in the abundance of genes in the specific category of sphingolipids. Nevertheless, differences can be observed for other lipid categories: this can indicate a diverse ability of the two yeasts to remodel the lipid composition. Indeed, in a different study [114] it was demonstrated that lactic acid stress did not evoke important variations of the sphingolipid content of *Z. parabailii* cells, but a significant decrease in lipid acyl chain length occurring in the stationary phase of growth. This modification could contribute to lower membrane fluidity of organic acids in their undissociated form.

Phenotype prediction

We assessed the *in silico* metabolic capabilities of the two ZyPa1_0.1 and ZyPa1_0.3 models by carrying out FBA simulations, and setting biomass synthesis reaction as the objective function to be maximized. The main readout resulting from the FBA is the calculation of the biomass yield that we obtained as ratio between the biomass synthesis flux value and the glucose exchange reaction flux value converted to grams.

From the FBA simulations, we computed the yield value for both ZyPa1 models and we compared them with the corresponding experimental values. As shown in Figure 3.2A, respectively at the dilution rate of 0.1 and 0.3 h^{-1}, a biomass yield of 0.57 compared to the corresponding experimental value of 0.52, and a biomass yield of 0.15 compared to the corresponding experimental value of 0.33 emerged revealing good agreement of the simulations with experimental data. The computational biomass yield was calculated by taking into account the glucose uptake variability resulted from the FVA simulations indicated by the error bars in the plot.

Reaction deletion analysis

We performed a reaction deletion analysis of both ZyPa1 models to investigate which deletions have an effect on the biomass synthesis reaction flux value. It is worth noticing the impact caused by the *in silico* deletion of the R01290 reaction, which is catalyzed by the L-serine hydro-lyase enzyme and corresponds to the addition of L-Homocysteine to L-Serine, with the release of L-Cystathionine and one water molecule: L-Serine + L-Homocysteine <=> L-Cystanthionine + H_2O. The simulation in both ZyPa1 models of R01290-catalyzing enzyme deletion implied a different reduction of biomass synthesis flux, in particular, of 23% in ZyPa1_0.1 model and a 39% in ZyPa1_0.3 model. This reaction is directly connected to the synthesis and metabolism of L-Cysteine, L-Methionine, along with metabolism of L-Serine, as substrate of R01290 reaction.

The small discrepancy between the stoichiometric coefficients associated to L-Cysteine, L-Methionine and L-Serine into the biomass synthesis reaction of both ZyPa1 models does not justify the observed reduction of biomass production flux value resulted from the knock out of R01290 reaction. A possible explanation for the result of this perturbation can be related to the ability of *Z. parabailii* to produce various aliphatic and aromatic alcohols known as fusel alcohols [115]. During food fermentation in yeast, fusel alcohols are produced from amino acid catabolism by using the so called Ehrlich pathway. A first transamination reaction generates the α-keto acid that is then decarboxylated to the corresponding aldehyde by α-keto acid decarboxylases. Several pyruvate decarboxylases (PDC) enzymes, among which those that generally participate in alcoholic fermentation by converting pyruvate to acetaldehyde can account for the reaction. A higher copy number of PDC genes in *Z. bailii* than in *S. cerevisiae* genome together with a less efficient Crabtree effect suggests that PDC genes in *Z. bailii* may be beneficial for the conversion of amino acids and reducing sugars to aldehydes by the Ehrlich pathway rather than for favoring alcoholic fermentation in this organism. Aldehydes are subsequently converted to higher alcohols (the fusel alcohols) and acids by, respectively, alcohol dehydrogenases and aldehyde dehydrogenases. The fewer copy number of genes encoding alcohol dehydrogenases in *Z. bailii* compared to *S. cerevisiae* could be the reason why it produces more aldehydes and less alcohols compared with S. cerevisiae. This route may represents a rewiring, also partial, of metabolism when sugars source is depleted and amino acids can be then used as energy and carbon sources. Our finding that *in silico* deletion of R01290 reaction causes a higher reduction of biomass synthesis flux value at dilution rate 0.3 than 0.1 h^{-1} when the metabolism is respiro-fermentative is in line with the usage of Erlich

pathway during fermentation, but it needs to be validated through appropriate *in vivo* experiments.

Ability of *Z. parabailii* to consume acetate

Z. parabailii is highly tolerant to organic acids. Moreover, its ability to consume organic acids also in the presence of glucose has been described [116]. This ability to consume acetate in the presence of glucose and oxygen is of particular interest as acetate is released from pretreated lignocellulosic biomass and acts as an inhibitory compound for most microbial cell factories but not for *Z. parabailii*.

We assessed the ability of *Z. parabailii* to simulate this behavior by identifying in the model all the reactions where acetate is involved as substrate or product. Then, we carried out a parsimonious FBA (pFBA) for analyzing if flux of these identified reactions is changed when acetate is consumed from the extracellular environment. In this simulation, we did not force an uptake of acetate in the models, but we just gave them the possibility to consume this carbon source from the external environment. This means that we fixed the upper bound of acetate exchange reaction to zero and we cyclically increased its lower bound from 0 to a value of -10 mmol gDW^{-1} h^{-1}, where the negative value is a convention used for the exchange reactions indicating a consumption of a given metabolite from the extracellular compartment into the model. We choose to perform pFBA because in order to investigate the organic acid tolerance behavior of *Z. parabailii*, this approach returns the flux distribution that minimize the total flux of sources for reaching a given objective. Moreover, the solution that of the classic FBA involved a null acetate uptake flux value that is however possible according to FVA output.

From this analysis, we firstly checked if an uptake of acetate takes place in *Z. parabailii* models, and we then verified if biomass synthesis or the acetate involving reactions change their flux value according to an uptake of this metabolite.

We observed that the yield of both versions of the ZyPa1 model do not vary when this carbon source enters the cell. This result suggested two alternative hypotheses. The first one is that in both Zypa1 models acetate is not being consumed from the extracellular environment, leading to an unchanged flux distribution. According to the second hypothesis, acetate is entering the cell but without contributing to growth. By analyzing flux through acetate exchange reaction and the 21 model reactions where acetate is involved, we are more prone to consider the second hypothesis as true. Indeed, following the FVA, we observed that a consumption of acetate is possible in both models.

We chose to investigate the flux distribution where the acetate uptake flux was the highest value for each tested intake level between 0 and 10 mmol gDW^{-1} h^{-1}. Given the highly tolerance to organic acids in the presence of glucose, the Figure 3.5 briefly shows the catabolism of these two metabolites. Once acetate is consumed, we observed that it contributes to both the cytosolic and mitochondrial Acetyl-CoA (AcCoA) pool, which is involved in multiple pathways, including the fatty acids biosynthesis, the metabolism of some amino acids and the Krebs cycle. At the experimental condition of pH5, acetic acid is about 50% in its indissociated form. Therefore, it enters the cell mainly by simple diffusion: the higher intracellular pH (pH_i) determines its deprotonation, causing negative effects both for acidification and for the negative effects of the counteranion. The *in silico* evidence of the acetate consumption without a significant impact on the biomass yield suggested that the catabolism of this organic acid can contribute to its detoxification, possibly providing ATP for the energy-consuming proton extrusion and 2C units for the remodeling of membrane lipids (see above).

Discussion

In this work, for the first time we reconstructed the metabolic model of the stress tolerant hybrid yeast *Z. parabailii* ATCC60483 based on the recent genome assembly and annotation coupled with wet data obtained with chemostat cultivation at two different steady states, one fully respiratory and the second respire-fermentative. The two steady states were defined in terms of principal metabolites and macromolecular composition of the biomass.

This first version of the model proved to be able to describe the experimental biomass yield at a quite good level, and through the reaction deletion analysis revealed a different impact in terms of biomass synthesis flux reduction caused by the *in silico* deletion of the reaction catalyzed by the L-serine hydro-lyase enzyme. This reaction, which is connected to the metabolism of L-Cysteine, L-Methionine and L-Serine, may represent a metabolic rewiring when source of sugars is depleted and amino acids can be used as energy and carbon sources. It is worth noticing that when GPR rules associated to reactions in the model will be curated, it will be possible to run different simulations, among which the gene deletion analysis, which exploits the knowledge of genes involved in a specific reaction and their relationships. This will confer to the model the capability of prediction about knock out of specific genes in order to perform certain tasks.

This first version of ZyPa1 model also showed the ability to follow the acetate catabolism at different dilution rates. The *in silico* evidence for an acetate

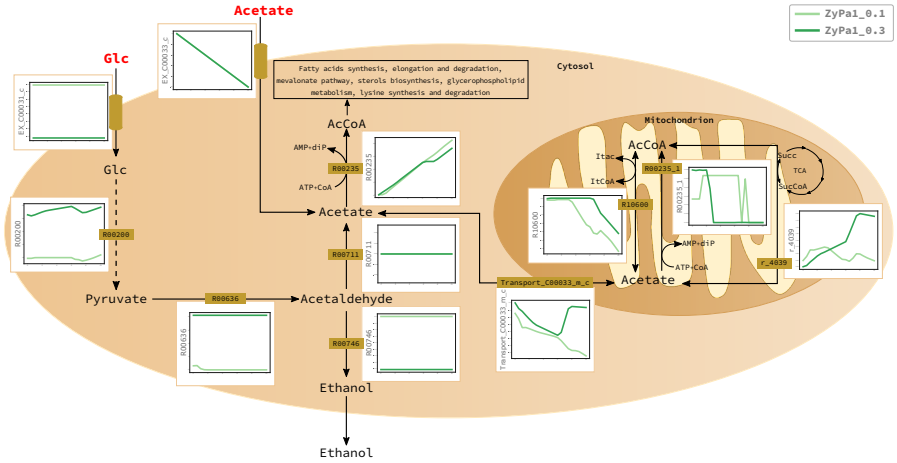


Figure 3.5: **Effect of increasing acetate consumption rate from the extracellular environment between 0 and 10 mmol gDW^{-1} h^{-1} on ZyPa1 models flux distributions.** Only the acetate involving reactions sensitive to acetate intake, and representative reactions from glucose catabolism are reported. In each plot, the x-axis shows the acetate uptake flux value, and y-axis shows the flux value for the reaction indicated along the axis. The light and dark green lines refer, respectively, to ZyPa1_0.1 and ZyPa1_0.3 models. Plots with a negative slope of the lines and, consequently, a negative flux value along the y-axis for the related reaction indicates that the reverse direction of the reaction is followed.

consumption without a significant impact on the biomass yield supports the hypothesis that the main contribution of this catabolism is in the detoxification of this organic acid as an inhibitory compound often released from pretreated lignocellulosic biomass.

It is worth to mention that the majority of the processes related to the production of bulk chemicals are run in batch or fed-batch and it is important also in this case to use wet data for asking the model for predictions. In this regard, the reconstructed genome-scale model can be used as a scaffold for implementing a reduced version of the network that can focus the attention on specific small-scale pathways. In this context, the investigation of the system dynamics can be performed by means of the mechanism-based modeling approach giving a detailed description of specific cellular processes with also the greatest predictive capability about the functioning of biological systems at the molecular level.

In conclusion, we believe that ZyPa1 model can be a useful instrument both for fundamental and applied studies. The genome-scale metabolic reconstruction can be integrated with differential -omic data to create context-specific models, to decipher, as example, how *Z. (para)bailii* robustness correlates with metabolic networks and with the modulation of key structural components (as we underlined for the plasma membrane composition). Moreover, such an integration can help to describe what we recently observed under lactic acid stress, revealing that the genes of each homolog pair tend to diverge in expression to a significantly greater extent than under control conditions [105]. These integrations, together with the differential analysis of gene abundance for diverse categories of functions can inspire novel experiments and help in the description as well in the exploitation of the peculiarity of this yeast.

Materials and Methods

Genome-scale metabolic model reconstruction

The reconstruction procedure consists of seven major steps:

1. **Annotate the genome of the target organism and highlight its metabolic functions.** The *Z. parabailii* genes were automatically obtained with an improved version of the Yeast Genome Annotation Pipeline based on homology and synteny information from 20 different yeast species from the Saccharomycetaceae family [117]. These gene predictions were curated using Illumina RNAseq data using the Trinity / Pasa pipeline followed by manual inspection [110]. Afterwards, Enzyme Commission (EC)

numbers were obtained using Blast2GO as previously reported [66, 118]. These EC numbers were inferred based on sequence homology keeping track of the organisms from which this information is taken.

2. **Populate a first draft model by connecting genes associated with an EC number into corresponding metabolic reactions.** This was done with the Python package Bioservices [119], which was used to connect to the KEGG database [68, 120] to extract all the enzymatic reactions and related functional hierarchy information for every *Z. parabailii* gene with an assigned EC number. The KEGG functional hierarchy was used here to filter out all the EC numbers from enzymes whose function is not related to metabolism.

3. **Refine the draft model using information from phylogenetically close organisms for which genome-scale metabolic models are available.** When possible, we inferred reaction compartmentalization, which is not represented in the KEGG database, and stoichiometry from the iOD907 model for *K. lactis* and the Yeast7 model for *S. cerevisiae*, taken here as references [100, 111]. This process is guided by the EC-homology information stored from step 1. Moreover, reactions from the reference models whose Gene-Protein-Reaction (GPR) rule is fulfilled are also included in the model draft. The GPR rules exploit a boolean expression for indicating which genes are involved in a given reaction and how they are interconnected. The satisfaction criterion implies that if a rule has the form "gene A AND gene B", both genes in the reference model must be homologs of the draft model genes. On the contrary, if the rule has the boolean operator OR, just one of the two genes is necessary to be homolog. All the reactions from the reference models that are not associated to any GPR rule, excluding transport and exchange reactions, are *a priori* included into the draft model.

   Special considerations were taken for reactions in the draft model involving genes whose EC numbers where not inferred either from *S. cerevisiae* or *K. lactis*. In these cases, the stoichiometry of the biochemical reaction was taken from the KEGG database, whereas the localization was assigned to the cytosol compartment unless otherwise stated in the Uniprot database [80]. Manual curation was performed for reactions in the KEGG database that are catalogued as "general reaction" involving generic substrates and products. For example, the KEGG reaction with identifier R01532 corresponds to the equation $Nucleoside\ triphosphate + H_2O <=> Nucleotide$

+ *Diphosphate*. This reaction is labelled as general because its main substrates and products, namely Nucleoside triphosphate and Nucleotide, are generic compound that can correspond to ATP, CTP, UTP, GTP or TTP in case of Nucleoside triphosphate, and to AMP, CMP, UMP, GMP or UMP in case of Nucleotide. In these cases, the KEGG database suggests a group of specific reactions that are all included into the network under construction.

4. **Add transport reactions for metabolites present in multiple compartments.** To our knowledge, transport reactions have not been fully characterized in *Z. parabailii*. Therefore, a reversible and unbounded transport reaction (i.e. lower and upper bound are set equal to -10^6 and 10^6 mmol gDW^{-1} h^{-1}) is always added between the cytosol and any other compartment for every metabolite. For example, if a metabolite m is located in the cytosol (m_{cyt}), in the mitochondrion (m_{mit}) and in the peroxisome (m_{per}), two transport reactions are added to the model, namely $m_{\mathrm{cyt}} \leftrightarrow m_{\mathrm{mit}}$ and $m_{\mathrm{cyt}} \leftrightarrow m_{\mathrm{per}}$. If a metabolite m is located in multiple compartments, such as, the mitochondrion (m_{mit}) and the peroxisome (m_{per}), but neither is in the cytosol compartment, the metabolite m_{cyt} is first added to the model and then two transport reactions are included, namely $m_{\mathrm{cyt}} \leftrightarrow m_{\mathrm{mit}}$ and $m_{\mathrm{cyt}} \leftrightarrow m_{\mathrm{per}}$.

5. **Integrate experimental chemostat cultivation and medium composition data.** This step is optional as it depends on the availability of chemostat data and information about the medium composition. The data from chemostat cultivation is integrated into the model by adding an exchange reaction, which is a transport reaction between the intracellular and the extracellular environment for every metabolite whose production or consumption rate has been experimentally determined. Lower and upper bounds of these reactions are then constrained according to the measured values. Before integrating available experimental concentration values as exchange reaction constraints, these values were transformed into corresponding consumption or production rates q_m using the as following equation:

$$q_m = \frac{X_m \cdot \mathrm{D} \cdot 10^3}{\Omega \cdot \omega_{\mathrm{m}}}$$

where X_m is the concentration of the metabolite m measured in g l^{-1}, D is the dilution rate measured in h^{-1}, Ω is the dry cell weight measured in g l^{-1},

and ω_{m} is the molecular weight of the metabolite m measured in g mol^{-1}. A factor of 10^3 was used to convert mol gDW^{-1} h^{-1} into mmol gDW^{-1} h^{-1}, which is the standard measure unit of the FBA. The model can be further constrained by defining the set of exchange reactions from the extracellular environment in order to replicate the experimental growth medium by adapting its metabolite composition to the draft model. The lower bound of each exchange reaction l_m^E is set as follows:

$$l_m^E = \frac{X_m \cdot \mathrm{D} \cdot 10^3}{\omega_{\mathrm{m}}}$$

where X_m is the concentration of the metabolite m within the experimental medium measured in g l^{-1}, D is the dilution rate measured in h^{-1}, and ω_{m} is the molecular weight of the metabolite m measured in g mol^{-1}. As above, a factor 10^3 is used to pass from mol gDW^{-1} h^{-1} to mmol gDW^{-1} h^{-1}.

6. **Gap filling.** Gaps are missing reactions leading to the formation of metabolites involved in just one reaction within the model as substrate or product, or in more than one reaction but only as substrate or product. These metabolites are generally called dead-end metabolites. The gap filling process consists of three steps:

   - Dead-end metabolites are identified and catalogued as substrate or product, and their compartment localization is stored.
   - The Bioservices package is used to automatically find in the KEGG database the reactions where each dead-end metabolite is involved
   - Reactions involving dead-end metabolites as substrates are intersected with those involving dead-end metabolites as product to determine common identifiers. In case of a positive match, both metabolites are required to be localized into the same compartment while checking that the reaction was not already included into the model.

7. **Include exchange reactions to achieve mass-balance in the system.** This last step involves a new search for dead-end metabolites for which an entry or exit irreversible exchange reaction is added to the draft model. If the unique reaction where a dead-end metabolite is involved is reversible, we added a reversible exchange reaction regardless of the role of the dead-end metabolite a substrate or product. The same criterion

is followed for dead-end metabolites involved in more than one reaction: if all the involved reactions are irreversible, an exchange reaction is added to the model in the direction of uptake for substrate and secretion for product. When at least one of the involved reactions is reversible, an unbounded exchange reaction is included into the model.

Definition of the biomass synthesis reaction

The biomass synthesis reaction takes into account all the experimentally identified components where each element m is associated to a stoichiometric coefficient s_m and computed as follows:

$$s_m = \frac{f_{\mathrm{m}} \cdot f_P \cdot 10^3}{\omega_{\mathrm{m}}}$$

where f_{m} represents the fraction in weight of the monomer m into the macromolecule P; f_{P} represents the fraction in weight of the macromolecule P into the biomass, and ω_{m} is the molecular weight of the monomer m.

Glycerol, trehalose, chitin and 1,3-beta-D-glucan, take part in biomass formation according to the growth reaction composition of the *S. cerevisiae* and *K. lactis* genome-scale models. However, as information about weight percentages was only available for glycerol, we added to the model a mock reaction where carbohydrates contribute to the formation of a generic compound called "carbohydrates" that we included into the biomass reaction associated to a stoichiometric coefficient calculated similarly as above:

$$s_{\mathrm{carb}} = \frac{(f_{\mathrm{carbTot}} - f_{\mathrm{glycerol}}) \cdot f_{\mathrm{carbTot}} \cdot 10^3}{\omega_{\mathrm{dcarb}}}$$

where f_{carbTot} is the fraction in weight of all the carbohydrates into biomass, f_{glycerol} is the fraction in weight of glycerol into carbohydrates, whereas ω_{carb} is the molecular weight of the fictitious "carbohydrates" species.

Since we had no information about single deoxynucleotide and nucleotide composition but we had their total percentages in weight, we added to the model two mock reactions for the formation of total DNA and RNA. In these mock reactions, the sum of the corresponding deoxynucleotides and nucleotides contribute to the formation of two generic compounds that we called "DNA" and "RNA". These were included into the biomass reaction, and associated to a stoichiometric coefficient as follows:

$$s_{\mathrm{DNA}} = \frac{f_{\mathrm{DNA}} \cdot 10^3}{\omega_{\mathrm{DNA}}}, \qquad s_{\mathrm{RNA}} = \frac{f_{\mathrm{RNA}} \cdot 10^3}{\omega_{\mathrm{RNA}}}$$

where $f_{\mathrm{DNA/RNA}}$ represents the fraction in weight of total DNA or RNA into biomass, and $\omega_{\mathrm{DNA/RNA}}$ is the molecular weight of the fictitious DNA or RNA compounds.

As the total ATP cost necessary for the cell growth has not been experimentally determined, it was set to the value used in the biomass synthesis reaction of the *S. cerevisiae* and *K. lactis* genome-scale models.

Computation of the biomass yield

The biomass yield Y corresponds to the amount of biomass produced per grams of consumed carbon source, giving information about substrate conversion efficiency. This parameter is calculated as ratio of model simulations deriving flux of the biomass synthesis reaction over that of the glucose exchange reaction that is necessary to convert from mmol gDW^{-1} h^{-1} into grams. To sum up:

$$Y = \frac{v_b^E}{|v_g^E| \ \cdot \ \omega_{\mathrm{g}} \ \cdot \ 10^{-3}}$$

where v_b^E is the flux value of the biomass synthesis reaction, $|v_g^E|$ corresponds to the glucose exchange reaction flux value which is taken as absolute value to remove the negative sign due to exchange reactions convention, and ω_{g} is the glucose molecular weight.

Classic and parsimonious Flux Balance Analysis

The Flux Balance Analysis (FBA) is a constraint-based approach which exploits the linear programming to identify the optimal flux distribution that maximizes or minimizes a specific metabolic objective [60]. FBA relies on a steady state assumption, according to which time variation of each of the internal metabolites concentration is equal to zero. This means that $dX_i/dt = 0$, where X_i is the concentration of the metabolite i. The flux distribution represents the output of this analysis and shows the rate at which each reaction of the model occurs at steady state.

The application of constraints on the system under evaluation is necessary to reduce the set of candidate flux distributions, defining in this way an allowable solution space where any flux distribution may be equally acquired by the model. The maximization or minimization of a specific objective function Z allows to get a further narrowed feasible solutions space, and to identify a single optimal flux distribution.

The set of metabolic reactions included in the model is mathematically represented as a stoichiometric matrix S of size $M * R$, where M is the number of metabolites and R is the number of reactions included in the model. Given the stoichiometric matrix S, where each element $s_{i,j}$ represents the stoichiometric coefficient of each metabolite within each reaction, the formulation of a FBA problem may be formulated in this way:

$$\text{maximize or minimize } Z = \sum_{i=1}^{R} c_i v_i$$
$$\text{subject to } S\vec{v} = 0,\ \vec{v}_{min} \leq \vec{v} \leq \vec{v}_{max}$$

where c_i indicates the objective coefficient value for the reaction i; v_i represents the flux value of the reaction i; the two vectors $\vec{v}_{min}$ and $\vec{v}_{max}$ represent the lower and upper bounds vectors, expressing for each flux v_i of the vector $\vec{v}$, the range within it can vary. An interval between negative and positive values means that flux through the corresponding reaction can flow in the backward or in the forward direction.

The parsimonious FBA (pFBA) [65] is a variant of the classic FBA approach that applies the same principles, but it set as objective function the minimization of the sum of all fluxes. Therefore, pFBA assumes that there is a selection for the fastest growing strains that minimize the total necessary resources to implement the optimal solution.

Although the FBA only returns a single flux distribution, the constraints imposed on the system under investigation do not always allow to obtain a unique solution, but may confine the solution space to a feasible set of alternative optimal flux distributions in which the same optimal flux value of the objective function may be reached through different but equally possible ways. In this context, the Flux Variability Analysis (FVA) [64] returns the range of flux variability of each reaction, i.e. the allowable minimum and the maximum fluxes by each model reaction.

All the FBA and FVA simulations has been performed by using the COBRApy Toolbox functions [121] and the Gurobi solver.

Reaction deletion analysis

The gene deletion analysis is based on the knowledge for all reactions of the model under investigation of the corresponding gene-protein-reaction (GPR) rule, and implies that the deletion of each gene included in the model is cyclically simulated. This means that flux through all the reactions in which this gene is involved, according to the corresponding GPR rule, is blocked. The reaction

deletion analysis is conceptually similar to the gene deletion analysis. It is performed by cyclically setting to zero both lower and upper bounds of each reaction in the model, preventing flux passing through it, without addressing the corresponding GPR rule. A FBA running is then performed to analyze the effect of each deletion on the objective function flux value.

Particle Swarm Optimization

The Particle Swarm Optimization (PSO) is a population-based stochastic algorithm, originating from studies of birds flocking, which aims at finding the set of parameters that maximize a specific objective [122]. The key concept behind the PSO is that an initial population of several candidate solutions, called particles, are randomly located inside a constrained search space by cooperating one with each other for finding the global best solution. Each candidate solution is then evaluated by the PSO algorithm by considering how much the particle minimizes its distance from the objective function, determining in this way the fitness value of that solution. Each particle, which is characterized by both a position in the search space and a velocity, keeps track of its personal best position that it has achieved so far in the search space. At each iteration of the algorithm, the velocity of the particle updates its current position, and the fitness is then revaluated. The individual best positions of each particle and the global best position achieved among all particles in the swarm are updated by comparing the current fitness values with the previous ones, and replacing them if better fitnesses are obtained. The PSO algorithm continues until stopping condition is reached. The final solution corresponds to the best fitness value achieved among all particles in the swarm, referred to as global best fitness, and the particle that achieved this fitness, referred to as global best candidate solution.

We implemented the PSO algorithm by using Python as wrapper. We carried out this analysis by setting a swarm size of 32 particles, 2000 iterations, and a range between 10^{-6} and 1 for the coefficients to estimate. We computed the fitness function by using the least square method to calculate the minimal distance between the computational and the experimental biomass yield values.

Experimental methods

Glucose-limited aerobic chemostat cultivations. *Z. parabailii* strain ATCC 60483 was used for bioreactor fermentation. Chemostat cultivations were performed in Biostat-B fermenters (B-Braun). A defined medium with vitamins

and trace metals was used [123]. The glucose concentration in the reservoir medium was about $20\,\mathrm{g l^{-1}}$. A constant working volume of 1300 ml was maintained via an effluent line coupled to a peristaltic pump. A dissolved oxygen concentration above 50% of saturation was maintained by an air flow of $1.3\,\mathrm{l^{-1}min^{-1}}$ (1 v/v/m) and a stirrer speed of 1000 rpm. The temperature was maintained at 30 °C and the culture pH at 5.0 by automatic addition of 2 M KOH. The dilution rates were set at $0.1\,\mathrm{h^{-1}}$ (fully respiratory metabolism) or at $0.3\,\mathrm{h^{-1}}$ (respiro-fermentative metabolism). Cultures were assumed to be in steady state when at least seven volume changes had passed since the last change in growth conditions and the culture did not exhibit metabolic oscillations. The dry cell weight (Ω) was determined by filtering 10 ml of culture broth through pre-dried 0.22 μm membranes filters. The filters were washed with demineralized water and dried to constant weight in a microwave oven [124]. For every steady state, during chemostat cultivation, the percentages of O_2 and CO_2, present in the off-gas were measured with a gas analyzer (Bioindustrie Mantovane, with Peltier system for O_2 reading and infrared for CO_2).

Analysis of extracellular metabolites. 2 ml of culture broths were centrifuged for 2 minutes at maximum speed in microcentrifuge. Supernatants were stored at −20 °C until analysis. At the time of the analysis, the samples were diluted with H2O milli-Q according to the need and loaded onto a HPLC apparatus (Jasco), equipped with UV (210 nm) and RI (refractive index) detectors. A BioRad Aminex HPX-87H column was used for metabolites separation. The column was maintained at 35 °C, the flow at $0.5\,\mathrm{ml min^{-1}}$ and the mobile phase was H_2SO_4 0.005 N.

Lyophilization of samples for macromolecular biomass analysis. When the chemostat cultures reached the steady state of growth, an aliquot of culture broth of about 100 ml (a variable amount, needed to recover about 1 g of biomass) was collected by centrifugation at 5000 g for 7 minutes at 4 °C; the supernatants were discarded and the pellets were resuspended in about 45 ml of 20 mM Tris-HCl (pH 7.6) each. Aliquots of 2 ml of sample were transferred to microfuge tubes and, after a second spin cycle of 2 minutes at maximum speed in microcentrifuge, the pellets were quickly frozen by immersion in a equilibrated bath of 50% ethanol and dry ice. The frozen samples were temporarily stored at −80 °C and then loaded on the centrifugal evaporator Eppendorf Concentrator 5301 for about 1 h and 30 min. The lyophilized samples where stored at −20 °C.

Biomass elemental and macromolecular analyses. The analyses were carried out using the lyophilized samples obtained from the chemostat cultures at dilution rates of $0.1\,\mathrm{h^{-1}}$ and $0.3\,\mathrm{h^{-1}}$. The biomass elemental analysis was performed in outsourcing by REDOX S.r.l. (Monza, MB, Italy).

For total protein determination, lyophilized biomasses were lysed before proceeding with the determination of the total proteins. The lysis efficiency has been verified by microscopic observation and by the Micro BCA (Thermo) colorimetric assay. Cells were collected as previously described and resuspended in lysis buffer (Tris-HCL 25 mM, pH 7.5 EDTA 25 mM) containing protease inhibitors (PMSF 0.5 mM and complete protease inhibitors, Roche). The samples were resuspended in 1 ml of lysis buffer and microfuge tubes were prepared, containing aliquots of 600 μl of sample and 600 μl of glass beads were distributed in microfuge tubes placed in a pre-chilled rack, loaded onto the TissueLyser II apparatus (Qiagen) and subjected to three stirring cycles at maximum power (30 Hz), lasting 5 minutes each. For the quantification of total proteins the Micro BCA assay (Thermo) was used. The lysate samples were diluted 1:1000 with H_2O milli- Q and a 1 ml aliquots were added to 1 ml of reagent (prepared by mixing solutions A, B and C provided in a ratio of 25:24:1). The samples and the BSA standards, for the calibration curve set up, were incubated at 60 °C for 1 h. After cooling, samples were transferred into cuvettes for spectrophotometric determination at 562 nm. The determination of the amino acid and lipid composition were performed in outsourcing by the laboratories of Analysis and Peptide Synthesis of the Centres Cientifics I Tecnològics of Barcelona (Spain)and by the Research Center on Metabolism (CEREMET) of the University of Barcelona (Spain), respectively. For total carbohydrates determination, the lyophilized pellets were resuspended in H_2O milli-Q and diluted appropriately to obtain 1 ml aliquots containing about 0.1 mg of cells. Each sample was placed in a 15 ml tube with 5 ml of phenol 5% and 1 ml of H2SO4 96% and incubated at 90 °C for 10 min. Solutions at increasing concentrations of glucose were similarly treated in order to then generate a calibration curve and quantify the carbohydrates present in the samples. Quantification was performed by spectrophotometric analysis at 488 nm. For the determination of glycogen content, samples were resuspended in 10 ml of 0.6 MHCl incubated 1 h at 100 °C, then brought to room temperature, filtered with 0.22 μm filters and subjected to an enzymatic assay (D-glucose–Hk enzymatic assays, Megazyme International Ireland). This assay measures the glucose released from the lysis of glycogen. For the total RNA assay the orcinol method was used [125]. The lyophilized pellets were resuspended in 3 ml of cold 5% trichloroacetic acid (TCA) and then centrifuged at 3500 rpm at 4 °C for 7 min, washed twice in the same solution. The residual pellets were left at −20 °C for about 12 hours, then slowly thawed in melting ice and finally resuspended in 3 ml of perchloric acid (PCA) 0.3 M and placed at 90 °C for 30 minutes. After cooling, the samples were centrifuged at 3500 rpm for 7 min and the supernatants were collected for RNA analysis.

The reagent for the orcinol assay was assembled by dissolving 1 g of crystalline orcinol monohydrate and 0.5 g of FeCl hexahydrate in 100 ml of 37% HCl. The reaction mix was assembled in 15 ml glass tubes containing an aliquot of the sample (from 50 to 200 μl), adjusted to 1.5 ml with H_2O milli-Q and 1.5 ml of orcinol reagent. The tubes were gently stirred and placed at 90 °C for 20 min, covered by glass beads to limit the evaporation. After cooling the samples, absorbance at 660 nm was measured at the spectrophotometer. The calibration curve was prepared with known concentrations of standard *S. cerevisiae* RNA. The total DNA was assayed using the Qubit apparatus (Invitrogen) and the kit associated with it HS-DNA. The lyophilized pellets were resuspended in 1 ml of lysis buffer, lysed by TissueLyser II (Qiagen), with the same methods described above and then diluted 1:1000 and 1:100 in the TNE buffer. The assay was carried out according to the supplier's protocol.

3.2 From genome-wide to core metabolic models

In this section, I present the three previously mentioned applications of core modelling for the investigation of cancer metabolic rewiring.

Recent studies in the field of cancer research contributed to enrich biological knowledge about the role of a global metabolic reprogramming in supporting the proliferative requirements of cancer cells [17, 126]. Moreover, new evidence about the presence of multiple metabolic wirings behind the globally altered tumour metabolism emerged, shifting the attention from the investigation of the metabolic peculiarities specific of the tumoural condition compared to the normal one, to the variability of metabolic programs among multiple cancer types. Indeed, extensive intertumour heterogeneity has been observed among tumours originating from different tissue and cell types in terms of genetic and phenotypic variations, patterns and extent of genomic instability, prognosis, aggressiveness and sensitivity to cytotoxic therapies [127]. Due to intertumour heterogeneity, varying rate of nutrients uptake and consumption coexist with several metabolic programs. Accordingly, tumour cells may have a dual capacity for glycolytic and oxidative metabolism, or they can rely more heavily on mitochondrial respiration for their growth, where less pyruvate is converted into lactate and the majority is oxidized to acetyl-CoA and then metabolized into the TCA cycle. In addition, tumour cells may also have a propensity for the aerobic glycolysis, where pyruvate is preferentially oxidized into lactate at the expense of its mitochondrial processing [17, 126]. Consequently, intertumour heterogeneity pushes the need for a more personalized medicine to develop new and increasingly diversified biomarkers.

In view of the above, in Section 3.2.1, we exploited the potentiality of core modelling to identify and investigate the role of metabolic plasticity of cancer cells among three different types of tumours, namely liver, breast and lung cancer. We manually reconstructed these tissue-specific core networks using as scaffolds the corresponding genome-scale networks that are stored in the Human Metabolic Atlas (HMA) database, with a particular focus on the pathways that are known to play a key role in cancer growth and proliferation. The HMA database contains genome-scale networks for multiple normal and cancer types deriving from a generic and aspecific human metabolic network, called Human Metabolic Reaction (HMR). HMR network has been generated by incorporating information from previously published human genome-scale metabolic reconstructions and various biological databases. The tissue or cell type specific networks of HMA database constitute specific active portions of the generic HMR network according to corresponding proteomics, transcriptomics

and metabolomics data. We also exploited the universal HMR network to reconstruct a core model called "HMRcore" that we used to define a reference model in the comparative FBA with the reconstructed cancer models about metabolic capabilities of tumour cells with respect to the normal ones. The choice of using the same reference model instead of making a comparison between each cancer model against its corresponding healthy one, allowed to make the three cancer models, in turn, comparable among themselves and perform a more meaningful differential analysis.

The analysis of FBA derived flux distributions identified several metabolic alterations characterizing cancer cells compared to the healthy counterpart concerning several aspects, including growth rate, emergence of Warburg effect, lactate secretion, glucose and glutamine uptake, TCA cycle and OXPHOS wirings. These distinctions between reference and cancer condition may represent putative targets for the development of therapies against cancer cells growth and proliferation. In addition, two further analyses revealed other potential anti-cancer drug targets by highlighting the structural differences that within the reconstructed core metabolic networks are responsible for a reversion of the tumoural phenotype towards the reference one, and the fragility points within the cancer models corresponding to reactions that, if perturbed, cause a negative effect on the system.

The constraint-based analysis did not just discriminate between reference and cancer conditions, but it also highlighted heterogeneity in terms of flux values among the three investigated tumours. In particular, although flux distribution analysis pointed out one of the most common cancer traits, that is, a reduced flux through all the components of the OXPHOS pathway, the emerged flux deregulation occurs to a lesser extent in lung cancer model compared to liver and breast tumours. In this regard, as demonstrated in [128], the mitochondrial respiration is a crucial trait of lung cancer cells, and it contributes to promote their progression and development.

The two works presented in Sections 3.2.2 and 3.2.3 are still manuscripts in preparation and regards the utilization of core modelling to explore plastic responses of tumour cells to different nutritional conditions and perturbations. In particular, the work presented in Section 3.2.2 is focused on the investigation of alternative nutrients in promoting enhanced proliferation of *K-ras*-transformed NIH3T3 mouse fibroblasts (NIH-RAS). In this regard, two different scenarios have been explored. In the first one, which is considered the standard condition, cells are grown under the presence of glutamine. In the second case, cells are grown under glutamine deprivation coupled with the simultaneous supply in the medium of an alternative source consisting of α-KG and the four non-essential

amino acids alanine, aspartate, asparagine and proline. To this end, we exploited the core metabolic model "HMRcore" that we presented in Section 3.2.1, by adapting it to the purpose of this work. Moreover, a wide experimental investigation of these two scenarios assisted the setting of model constraints.

FBA model simulations under the two investigated environments highlighted a very good agreement of the *in silico* growth ratio compared to the experimental counterpart. Indeed, in line with the experimental emerged evidences, we observed the inability of α-KG combined to NEAA to fully rescue the glutamine missing in the medium in terms of cell inability.

In addition, we also observed that this growth discrepancy is considerably narrowing by removing the specific constraint on glutamate secretion rate that we previously imposed according to experimental data. This outcome led to hypothesize that an accumulation of unexploited glutamate could occur under the glutamine missing condition, together with its consequent secretion in the extracellular environment and reduction of the biomass synthesis rate. This assumption is in line with the experimental data, where glutamate has been observed to be outside released only in the glutamine deprivated medium. Looking at the flux values distributions in more detail, we could ascribe this behaviour to a similar and low activity of the glutamine synthetase (GS) enzyme between the two investigated scenarios. GS enzyme is involved in the cytosolic conversion of glutamate back to glutamine at the expense of one ATP molecule. The emerged behaviour of GS-catalyzed reaction in the two conditions indicates that this hypothesized activity level of GS could be probably sufficient in the standard condition, but not in the glutamine missing one, to process cytoplasmatic glutamate and avoid its secretion in the extracellular environment.

Following this outcome, we tested the model sensitivity to the progressive flux reduction of GS-catalyzed reaction under both considered environmental scenarios. This perturbation did not cause any effect in the standard condition since FVA pointed out a null minimum limit for the flux range of this reaction. On the contrary, in the condition where glutamine is not supplemented in the medium, the gradual reduction of the flux through the GS-catalyzed reaction caused a progressive decrease of the flux through the biomass synthesis reaction together with an increase of the maximum allowable glutamate secretion rate. This result supported our hypothesis relative to a putative link between GS enzyme activity and glutamate secretion rate in the extracellular environment. In addition, flux analysis after this perturbation also revealed that, in line with experimental data, under the condition of missing glutamine, the canonical forward direction of TCA cycle is increasingly preferred over the reductive carboxylation pathway that characterize the standard condition.

The work presented in Section 3.2.3 has continued the investigation of cancer metabolic rewiring through core modelling. More in detail, this work is focused on the reconstruction a core model of human central carbon metabolism, called ENGRO2, that we compared with a previous network version presented in [31], called ENGRO1. In this way, we explored the role of important previously not included elements, including cell compartments, mitochondrial shuttles and essential amino acids metabolism. In this regard, we extended the core metabolic model "HMRcore" that we presented in Section 3.2.1. Constraint-based simulations of ENGRO2 confirmed main ENGRO1 predictions especially regarding the positive contribution of glucose and glutamine for the maximization of biomass synthesis. In addition, a considerable part of the work has been dedicated to evaluating the contribution of different essential amino acids in the context of optimal tumour biomass. Preliminary computational results revealed differential abilities of the various essential amino acids to support cell growth and to compensate for total glutamine depletion in the medium.

3.2.1 Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models

Marzia Di Filippo, Riccardo Colombo, Chiara Damiani, Dario Pescini, Daniela Gaglio, Marco Vanoni, Lilia Alberghina, Giancarlo Mauri

Computational biology and chemistry 2016; 62:60-69

DOI: 10.1016/j.compbiolchem.2016.03.002

Abstract

The metabolic rearrangements occurring in cancer cells can be effectively investigated with a Systems Biology approach supported by metabolic network modeling. We here present tissue-specific constraint-based core models for three different types of tumors (liver, breast and lung) that serve this purpose. The core models were extracted and manually curated from the corresponding genome-scale metabolic models in the Human Metabolic Atlas database with a focus on the pathways that are known to play a key role in cancer growth and proliferation. Along similar lines, we also reconstructed a core model from the original general human metabolic network to be used as a reference model.

A comparative Flux Balance Analysis between the reference and the cancer models highlighted both a clear distinction between the two conditions and a heterogeneity within the three different cancer types in terms of metabolic flux distribution. These results emphasize the need for modeling approaches able to keep up with this tumoral heterogeneity in order to identify more suitable drug targets and develop effective treatments. According to this perspective, we identified key points able to reverse the tumoral phenotype towards the reference one or vice-versa.

Keywords: Cancer metabolic rewiring; Network reconstruction; Core metabolic model; Flux Balance Analysis.

Introduction

Over the past decades, a large quantity of information has been gathered about the main metabolic differences between cancerous cells and their healthy counterparts [129]. Fundamental differences in the metabolism of normal cells and rapidly proliferating cancer cells were already observed in 1924, when Otto Warburg detected a reprogramming of glucose metabolism in tumor cells also known as "Warburg effect" [28, 130]. According to this phenomenon, cancer cells metabolize large quantities of glucose into lactate by a fermentative metabolism regardless of oxygen availability, as opposed to normal cells, which mainly rely on mitochondrial oxidative phosphorylation to generate the energy required for cellular processes. More recently, it has been suggested that a global *cancer metabolic rewiring* is performed in order to more effectively support the neoplastic proliferation replacing the metabolic program that operates in normal cells, which involves also a stimulated utilization of glutamine by reductive carboxylation [33] and an alteration of redox processes [131].

Given the complexity and nonlinearity of metabolism, a focus narrowed on

the single molecules or mechanisms responsible for carcinogenesis fails to capture the complexity of cancer metabolic rewiring. Attention on the complex network of interactions in which they are involved, typical of Systems Biology [7, 132, 133, 134], is therefore required for the purpose. In this regard, mathematical models, which are formal and simplified representation of real systems, help to handle this complexity, maintaining at the same time a high level of accuracy and detail [57].

In particular, in the context of metabolic networks, genome-scale modeling represents an increasingly used approach [48], as it has been shown to be successful in modeling the entire human metabolism [135, 136, 137, 138], as well as cancer cells metabolism for the prediction of selective drug targets [139, 140, 141, 142]. In spite of the potentialities linked to their comprehensiveness, these reconstructions are difficult to control and often include errors, such as improper filling of network gaps (i.e. missing metabolic reactions that prevent the production or consumption of one or more metabolites). Moreover, the interpretation of the outcomes of their simulation is not always straightforward due to the difficulty in rationalizing such amount of data. On the other hand, *core models*, which limit the scope to specific aspects of metabolism, require more assumptions and a higher level of abstraction; nevertheless, they can be more accurately curated, their underlying assumptions are explicit, they are simpler, and thus their analysis might be more effective in uncovering system-level properties of cancer metabolic rewiring.

In view of this, starting from already existing genome-scale metabolic models, we aim at reconstructing manually curated core models that zoom-in on cancer metabolic rewiring. In order to study the role of the Warburg effect and more in general the role of all cancer metabolic alterations in supporting the neoplastic proliferation, we did not model a generic cancer cell, but rather we focused on the three most harmful neoplasias, in terms of incidence, mortality and prevalence, according to the last estimations in GLOBOCAN [143, 144]: liver, breast and lung tumors.

In fact, even though some metabolic features are associated with most cancer cells, several variations may emerge in specific tumor types [142, 145]. Cancer specific models may therefore be useful in identifying specific drug targets, paving the way towards a more personalized medicine.

Materials and methods

Flux Balance Analysis

Flux Balance Analysis (FBA) [60, 146, 147, 148] is a constraint-based method that allows to quickly calculate the flux of metabolites (i.e., the rate at which every metabolite is consumed or produced by each reaction) through all the reactions of a metabolic network, without requiring any information about kinetic parameters or metabolite concentrations.
FBA assumes that the time variation of internal metabolite concentrations is equal to zero (pseudo-steady state assumption) and proceeds with an iterative application of constraints (reaction thermodynamics and capacity constraints) on the evaluated system [60] in order to exclude flux distributions from the space of allowable ones [149].
Assuming than that the cell behavior is optimal with respect to a certain objective, FBA exploits linear programming to identify an optimal flux distribution within the defined solution space according to an objective function (as e.g. production of biomass or ATP yield) [60].

Human Metabolic Atlas

To investigate the metabolic processes of specific human cancer types, we exploited the Human Metabolic Atlas (HMA), a database which contains genome-scale metabolic networks for 69 cell types and 16 cancer types in addition to a generic and aspecific human metabolic model: the Human Metabolic Reaction (HMR) database, which has been constructed by merging elements of previously published generic genome-scale human metabolic models (Recon 1 [135], EHMN [136]) and by incorporating information from different databases such as HumanCyc [150] and KEGG [68, 151]

The specific models represent portions of the generic HMR that are expressed (or "active") in each tissue or cell type, according to biological evidence. In particular, cell type specific proteomic data contained in the Human Proteome Atlas (HPA) database [152] and tissue specific gene expression data and metabolomic data from the Human Metabolome Database (HMDB) [153, 154, 155] were integrated by applying the INIT (Integrative Network Inference for Tissues) algorithm [140]. Based on these different kinds of "omics" data, INIT assigns weights to the reactions in the HMR according to their different levels of evidence in the specific tissue or cell type. An optimization process is then performed with the aim of maximizing as much as possible the reactions fluxes with a high weight (since the corresponding enzymes have a high expression

level) while minimizing the others. Reactions that carry flux in the obtained optimal flux distribution are assigned to the tissue or cell specific model.

Metabolic network reconstruction process

Starting with the genome-scale networks, we manually reconstructed constraint-based core metabolic models for the three different types of cancers and for a generic aspecific reference cell, by extracting those metabolic pathways having a relevant role in supporting cancer cells growth and proliferation [131, 156].

The reconstruction process of these models was not limited to the selection of the reactions of interest, but also involved the integration of all the elements required to perform FBA, such as transport and exchange reactions.

The former are essential to allow metabolites transport between compartments, while the latter define the enviromental constraints and may be subdivided into sinks (defining the cell nutrients) and demands (defining compounds secreted by the cell). Exchange reactions were also used to represent non modeled metabolites pertaining to pathways not included here.

As a final step, we performed an accurate manual curation of the developed core models in order to check the presence of incorrect reactions, as well as network gaps. This curation phase strongly relied on the biological database KEGG and on the most up-to-date human metabolic reconstruction Recon 2 [138].

Topological analysis of the reconstructed metabolic models

A topological analysis has been performed on the reconstructed core models with the aim of examining their structural properties, by using the software Cytoscape [157, 158], and more precisely its plugin NetworkAnalyzer. In order to characterize the topology of the metabolic networks, we calculated the node degree distribution and the clustering coefficient for every metabolic network. We refer the reader to the review in [4] for an exaustive description of the above mentioned metrics.

Differential analysis of flux distributions

Following the reconstruction of the core metabolic models, we estimated the optimal flux distribution for each of them with FBA. In order to compare the capabilities of cancer cells with respect to normal cells to cope with the same metabolic requirements, given an identical medium composition, we did not

impose specific flux values for nutrient uptakes, but we imposed the same arbitrary value (1000), while we left the flux in the allowed direction unbounded for all internal reactions. For the same reason, although specific biomass formation pseudo-reactions representing the conversion of biomass precursors into biomass are provided in the original genome-wide models, we used the same formulation as objective function for both types of cells.

The simulations have been performed using the COBRA (COnstraint Based Reconstruction and Analysis) Toolbox [159, 160] and the GLPK linear programming solver. Each obtained flux distribution has been evaluated to check the absence of possible thermodinamically infeasible loops by using the algorithm developed in [161].

Once that flux distributions free from thermodynamical infeasible loops were obtained, we compared each cancer model with respect to the reference model, computing the fold change for each reaction as the ratio of the corresponding flux value in the cancer model versus its flux value in the reference one. It is worth stressing that the choice of using the same reference model, rather than comparing each tumor against its corresponding healthy model, is essential in order to make the three cancer models comparable among themselves and making the differential analysis meaningful.

Identification of critical reactions

We carried out two different analyses on the reconstructed core models. The aim of the first one is to highlight the reactions responsible for a reversion of the tumoral phenotype toward the reference model, whereas the second one has as purpose the identification of all the possible fragility points in the cancer models, which are able to cause, if inhibited, a negative effect on the system.

Identification of reactions responsible for the phenotype reversion This analysis aims to identify the structural differences in the core metabolic networks that are mainly responsible for their dissimilarities in terms of flux distributions.

In attempting to reach this goal, it is necessary to identify the reactions that are differentially present in the four core models.

We determined the set of reactions that are present only in the reference model and we added them in a stepwise fashion to the cancer models in order to investigate the effect of each perturbation. We also identified the reactions shared by all tumor models, but not present in the reference one. Such reactions were included in the reference model to evaluate whether a flux distribution

typical of cancer cell could be obtained. Along similar lines, we removed these reactions from the tumor models to assess if their removal caused a reversion of the tumoral phenotype towards the reference one.

Identification of cancer networks fragility points In order to identify fragility points in the reconstructed cancer models, we examined the "extent" of the Warburg effect in each of them through the quantification of two indexes, as proposed in [162]:

- the glycolytic to oxidative ATP flux ratio (AFR)
- the ratio of the glycolytic versus oxidative capacity (EOR), computed as the fraction of extracellular acidification rate (i.e. lactate secretion flux value), over the oxygen consumption rate (i.e. oxygen consumption flux value).

Given that high AFR and EOR ratios denote tumors with a "high" Warburg level (as explained in [162]), we queried those reactions that, if inhibited, reduce the Warburg effect, by decreasing the AFR and EOR ratios. The simulation of each metabolic reaction inhibition (and then, of the corresponding enzyme) has been performed by constraining the flux through each reaction to zero.

We also searched for those reactions that, after inhibition of the corresponding enzyme, lead to a reduction of the biomass synthesis rate. It has indeed been suggested that the above mentioned indexes are positively associated to cancer cell migration, but may have no effect on cell proliferation [162], in which we are interested.

Results and discussion

Core metabolic tissue-specific cancer models

The three reconstructed core metabolic tissue-specific cancer models consist of three cellular compartments (cytosol, mitochondria and external environment), and include those metabolic pathways having a relevant role in cancer cells growth [131, 156], namely glycolysis, pentose phosphate pathway (PPP), tricarboxylic acid cycle (TCA cycle), oxidative phosphorylation, glutamine metabolism, amino acid synthesis, urea cycle, folate metabolism and palmitate synthesis. Sink reactions for glucose, glutamine and oxygen have been added in the reconstructed core models for defining the medium composition. The modifications with respect to the original genome-wide models emerging from the curation process are listed in Table 3.1.

Original reactions	Revised reactions	Compartment
–	3-phospho-D-glycerate ⇒ 2-phospho-D-glycerate	Cytosol
Acetyl-CoA + H_2O + OAA ⇔ Citrate + CoA	Acetyl-CoA + H_2O + OAA ⇒ Citrate + CoA	Mitochondria
Isocitrate + NAD^+ ⇒ AKG + CO_2 + H^+ + NADH	Isocitrate + NAD^+ ⇔ AKG + CO_2 + H^+ + NADH	Mitochondria
Isocitrate + $NADP^+$ ⇒ AKG + CO_2 + H^+ + NADPH	Isocitrate + $NADP^+$ ⇔ AKG + CO_2 + H^+ + NADPH	Mitochondria
AKG + Leucine ⇒ 4-methyl-2-oxopentanoate + Glutamate	4-methyl-2-oxopentanoate + Glutamate ⇔ AKG + Leucine	Cytosol
AKG + Isoleucine ⇒ 2-oxo-3-methylvalerate + Glutamate	2-oxo-3-methylvalerate + Glutamate ⇔ AKG + Isoleucine	Cytosol

Table 3.1: Revised reactions after the curation phase of the core models, performed consulting KEGG database and the human metabolic reconstruction Recon 2. The first reaction is the result of a gap correction found within the glycolysis pathway, which in the starting genome-scale models, has been erroneously filled with two exchange reactions for the metabolites 3-phospho-D-glycerate and 2-phospho-D-glycerate. A revision of the directionality has been done for the other five reactions. In particular, the third and fourth reactions have been corrected within the cancer models because it is known that in tumors, unlike normal cells, the enzyme that is responsible for these reactions (isocitrate dehydrogenase) works mainly in the reverse direction.

The detailed lists of the reactions included within each of the final core models are available in Supplementary materials (S1_HMR_CORE.xls, S2_LIVER_CANCER_COR S3_BREAST_CANCER_CORE.xls, S4_LUNG_CANCER_CORE.xls). Table 3.2, instead, shows the number of metabolites and reactions of the core models, as compared to their genome-wide counterparts. Models are also accessible in the BioModels Database [163] with the identifiers MODEL1502100000, MODEL1502100001, MODEL1502100002 and MODEL1502100003.

From the topological analysis, as expected [4], it emerged that all the four models exhibit a hierarchical topology, that is a structural organization of the network integrating the scale-free property and the presence of modules. In the context of metabolic networks, modules usually overlap with metabolic pathways, while the scale-free topology is characterized by the presence of few species taking part in a high number of reactions, called hub (e.g. cofactors as ATP or NADH), and a huge number of metabolites taking part in a small number of reactions. Moreover, these models are marked by the presence of the ultra small-world property [4], a feature that is correlated to a fast transmission of the information through the network. As a consequence of this property, local perturbations in metabolites levels can quickly affect the entire network. At last, it has also emerged from the analyzed networks that the most part of the interactions are established between hub and species having few interactions. This property, called disassortativity, reflects the fact that elimination of a hub implies a strong negative effect on the entire network.

	Genome-scale		Core	
Model	# reactions	# metabolites	# reactions	# metabolites
HMR database	8180	6011	274	252
Liver Cancer GW	4386	4020	257	242
Breast Cancer GW	4299	3955	243	233
Lung Cancer GW	3809	3653	235	230

Table 3.2: Number of reactions and metabolites for each of the genome-scale and core metabolic models

Tissue-specific cancer redistributions of metabolic flux

An overall understanding of the biological mechanisms behind cancer cells growth and proliferation requires a complete and accurate analysis of the metabolic flux redistribution that they undergo [164]. In this work, we exploited FBA to quantify the metabolic flux through reactions of reference and tumor cells and, above all, to emphasize both the up- and down-regulations emerging between cancer and reference models, as long as the considered metabolic pathways are concerned. As objective function, we maximized the biomass production pseudo-reaction that was equally associated to all cancer genome-scale models, which has been adapted to encompass only the subset of metabolites needed for biomass production that are involved in the pathways here considered, each of them characterized by a proper stoichiometric coefficient, as showed in Table 3.3.

Biomass production reaction
0.4202 Alanine[c] + 0.00328 AMP[c] + 2.4523 Aspartate[c] + 0.00328 Arginine[c] + 0.03272 CMP[c] + 0.00805 dAMP[c] + 0.00537 dCMP[c] + 0.00537 dGMP[c] + 0.00805 dTMP[c] + 0.67628 Glutamine[c] + 0.5942 Glutamate[c] + 0.49244 Glycine[c] + 0.03709 GMP[c] + 0.06369 Proline[c] + 0.13132 Serine[c] + 0.01963 UMP[c] + 0.00427 Cysteine[c] + 0.01313 Isoleucine[c] + 0.0394 Leucine[c] + 0.00657 Methionine[c] + 0.04268 Asparagine[c] + 0.01313 Phenylalanine[c] + 0.0197 Tyrosine[c] + 0.03755 Cholesterol[c] + 0.04 Palmitate[c] → Biomass[s]

Table 3.3: Biomass production reaction used as objective function. This reaction includes the metabolites needed for biomass synthesis that are involved in the pathways here considered, each of them characterized by a proper stoichiometric coefficient. "c" and "s" correspond, respectively, to cytosol and external environment compartments.

A visual representation of the resulting "active" network is provided in Figure 4.6 for the reference model, whereas the main outcomes of the differential analysis are summarized in Figure 4.7, in which red and green chromatic scales

highlight, respectively, the detected up-regulations and down-regulations, in terms of flux value, for the cancer condition with respect to the reference one.

A clear distinction between the two studied conditions in terms of growth can be observed. Indeed, the biomass production rate in cancer cells resulted higher than in reference ones, with a positive fold change of about 1.2-1.4 fold. This is an expected outcome considering that cancer cells are known to show abnormal proliferation and no contact inhibition [19]. It is meaningful to underline that, to support this higher proliferation rate, cancer cells must undergo a complete metabolic rewiring which affects all the pathways having a key role in cancer growth (i.e. the pathways considered in our analysis).

Going more into detail of the emerged dissimilarities, we detected in all three tumor models the presence of the Warburg effect (Figure 4.9). When comparing tumoral against reference cells, the flux analysis highlighted a strong dependence of the tumor cells on glycolytic metabolism and a shift from generation of ATP through OXPHOS pathway to glycolysis, even under a normoxic condition. Since it is known that glycolysis is less efficient than OXPHOS, relatively to the number of ATP molecules generated per unit of glucose consumed [165], cancer cells are expected to be constrained to increase their glucose uptake. Our cancer models are in line with this observation, as shown by a 4-7 fold upregulation of glucose uptake.

A consequence of this phenomenon is that most of the pyruvate produced, instead of entering in the TCA cycle, is redirected to lactate which is then excreted to the extracellular environment. Lactate secretion in cancer cells corresponds to the most known way to regenerate NAD^+ from NADH, balancing in this way the elevated glycolysis rate and allowing it to persist. Moreover, lactate secretion may be able to confer another advantage to tumor cells, enhancing their invasiveness by disrupting normal tissue architecture, and promoting an acidic tumor microenvironment to evade tumor-attacking immune cells [130].

At last, the relevance of an up-regulated glycolysis seems to be also linked to the pivotal relevance of this pathway from a biosynthetic point of view. In fact, the high glucose carbon flux through this pathway allows to divert some glycolytic intermediates toward the synthesis of the amino acids serine and glycine, and the pentose phosphate pathway [29, 166]. Both resulted up-regulated, respectively, by about 1.4—1.8 and 1.2–1.4 fold, since they allow the biosynthesis of some precursors required to assemble new cells [19].

The TCA cycle is another crucial pathway in the tumoral condition. Indeed, it emerged that, in addition to the generation of reducing equivalents for the oxidative phosphorylation, the TCA cycle also acts as a "hub" for biomass precursors production [167]. Several of its intermediates are used to synthes-

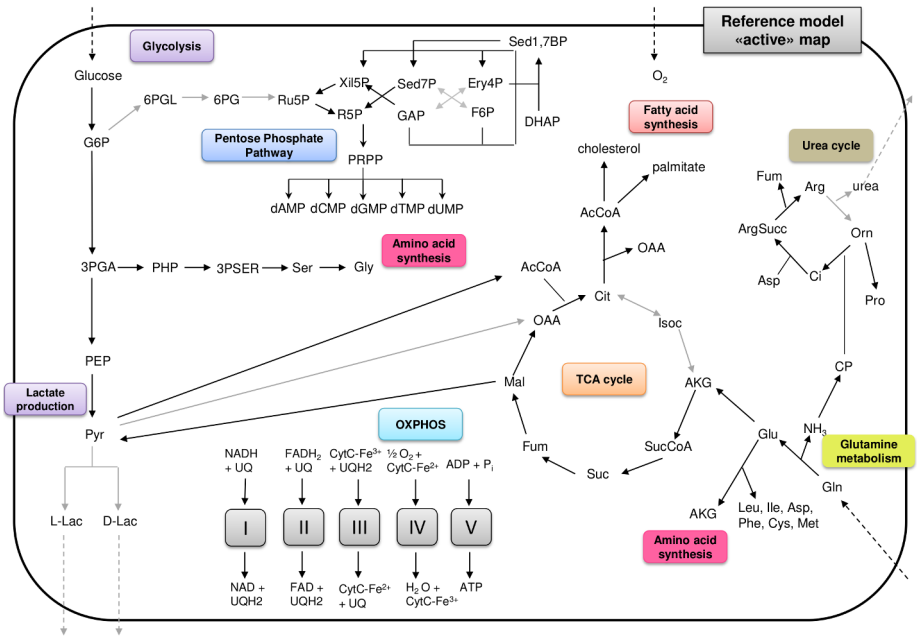


Figure 3.6: Schematic representation of the "active" network relative to the reference core model. This map shows, in a simplified view, the information concerning flux distribution in the reference core model within metabolic pathways under investigation in our work, namely glycolysis, pentose phosphate pathway, tricarboxylic acid cycle (TCA cycle), oxidative phosphorylation, glutamine metabolism, urea cycle, amino acid and fatty acid synthesis. The grey and black arrows correspond, respectively, to reactions having a null and a positive flux. The direction of each black arrow is set depending on the obtained flux value in the corresponding reaction. For reasons of space, cellular compartments are not included. A list of the abbreviations used in the map is provided in Appendix A.

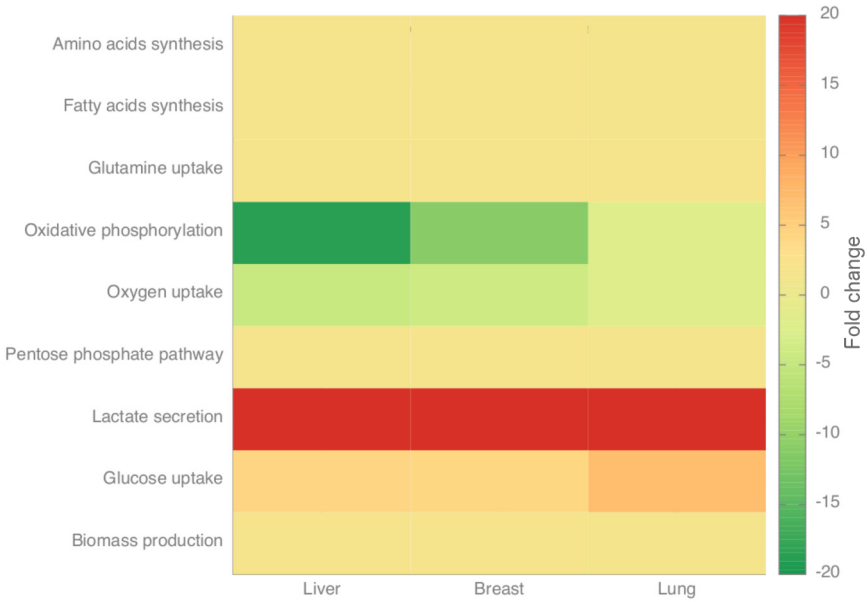


Figure 3.7: Main results obtained from the Flux Balance Analysis. Red and green chromatic scales correspond, respectively, to the detected up-regulations and down-regulations in the cancer cells models with respect to the reference model as result of fold change computation. From the figure it emerges that tumor cells, compared to the reference one, reprogram the metabolic pathways to satisfy their increased needs for the synthesis of macromolecular precursors essential for biomass production during tumor growth. Exploiting the FBA approach a heterogeneous behavior among the three investigated tumors also emerged. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

ize molecules such as lipids and proteins, and, for this reason, a continuous efflux of metabolic intermediates from this pathway, such as oxaloacetate, α-ketoglutarate and citrate, is needed. In particular, the reductive carboxylation way, exploited to produce this last metabolite, has assumed a great relevance inside our analysis. Once the citrate is transported from the mitochondria to the cytosol, its metabolism provides acetyl-CoA for the synthesis of fatty acids (simplified in our models with the palmitate synthesis) and cholesterol, both required for the generation of the lipidic membranes. This justifies the up-regulation observed for the lipogenic enzymes ATP citrate lyase and fatty acid synthase of about 1.2-–1.4 fold. The differential analysis of the fluxes revealed that, in contrast to what happens in the reference model (in which, as shown in Figure 4.6, the generation of citrate starts from the glycolysis-derived pyruvate, following the canonical direction of the TCA cycle), in cancer cells the conversions of mitochondrial pyruvate to acetyl-CoA and of oxaloacetate to citrate do not take place (null flux) (Figure 4.8). This implies, as a consequence, that inside the TCA cycle the level of citrate compared to that of the α-ketoglutarate becomes very low and, probably, this is the reason why cancer cells consider more advantageous having a reverse TCA cycle as the major pathway of citrate formation. Indeed, in each of the three tumor core models, a glutamine-dependent reductive carboxylation reaction occurs, catalyzed by the mitochondrial enzyme isocitrate dehydrogenase (IDH) which, differently from the reference cells, works in the reverse direction converting the α-ketoglutarate to citrate [167].

Glutamine has a key role in cancer growth and proliferation [168]. This has been confirmed in cancer fluxes analysis by the higher glutamine uptake (about 1.4 fold). This metabolite, together with glucose, represents a carbon and energy source to support many cellular processes, like reductive carboxylation and several biosynthetic pathways. Indeed, glutamine serves to maintain the pools of non essential amino acids, to provide a source of carbons for fatty acids synthesis, to produce aspartate (which is a required precursor for the synthesis of both nucleotides and the amino acids asparagine and arginine) and, lastly, providing a source of oxaloacetate, to replenish the TCA cycle intermediates (i.e. performing anaplerosis, a crucial feature for cell growth and proliferation since it confers the ability to use the TCA cycle as a provider of precursors for the biomass synthesis). Moreover, a high anaplerotic flux is a more specific indicator of cell growth compared to a high glycolytic flux, since the latter is also stimulated by hypoxia and other stresses independently of the need to synthesize the biomass precursors [29].

A last typical cancer trait highlighted by the flux distribution analysis is

a down-regulated oxidative phosphorylation which, together with an increased glycolysis, represents another aspect of the Warburg effect (Figure 4.10) [28]. In particular, a reduced activity of all the components of the OXPHOS, with a complete inhibition of the complex I activity and a sharp drop in the fluxes associated to other complexes occurred. Worth of note, downregulation of oxidative phosphorylation in lung cancer models was not as substantial as in the other two types of tumors: liver and breast. Indeed, as recently demonstrated by Hooda et al. in [128], the mitochondrial respiration is crucial in lung cells cancer and it promotes their progression and development.

The fluxes distributions obtained for each core model are available in Supplementary materials, Table S5.

Critical reactions

Stuctural differences responsible for cancer metabolic rewiring Our core models have been analyzed in order to identify critical reactions able to cause a strong flux variation with the consequent reversion of the model phenotype, constituting potential anti-cancer targets.

Let us first focus on the reactions that appear only in the reference model and are absent in the cancer models (listed in Supplementary materials, Table S6).

When analyzing the effect of their addition, one by one, to cancer models, it emerged that four of them are able to cause a perturbation in all three tumor types. These reactions are: 1) inorganic phosphate transport between cytosol and mitochondria; 2) water transport between cytosol and mitochondria; 3) the mitochondrial irreversible conversion from isocitrate to α-ketoglutarate catalyzed by NAD-specific form of isocitrate dehydrogenase (IDH) enzyme; 4) the mitochondrial irreversible conversion from isocitrate to α-ketoglutarate catalyzed by NADP-specific form of IDH enzyme.

1) The inorganic phosphate transport is based on a H^+/P_i cotransport catalyzed by the SLC25A3 enzyme. Its addition to cancer models produced a reduction of glucose uptake (between about 1.3 and 1.6 fold) and lactate secretion (between about 1.5 and 3 fold), and an enhancement of the ATP production by OXPHOS pathway (about 1.2 fold). Instead, the biomass synthesis rate showed no variation of its value and, therefore, the addition of this reaction causes a reduction of the Warburg effect but has no effect on cancer cells proliferation. This same reaction, even when removed from the reference model, produces an increase of the Warburg effect but has

no consequence on cell proliferation rate. We remark that the decoupling between Warburg effect and proliferation has already been suggested (see [162] for further reference).

2) The second critical reaction is the water transport, catalyzed by the AQP8 enzyme. Its inclusion in cancer models caused a lowering of glucose uptake (between about 1.1 and 1.3 fold) and lactate secretion (between about 2 and 7 fold), and an increase of the ATP production by OXPHOS pathway (between about 1.2 and 2.5 fold). As in the previous case, there has been no decrease in biomass production rate. Its removal from the reference model reflected what we observed with the inorganic phosphate transport reaction.

What has emerged so far highlights the importance of transport reactions, towards which the focus is not generally posed in constraint-based analyses. In principle, the lack of these reactions in the cancer models would suggest a null expression level of the corresponding protein but a search within the HPA database has shown that their expression level is registered as medium-low. Therefore, some reactions, such as these two transports, are not removed from the models for the expression level of the corresponding protein, but due to the fact that after the optimization process they have a null flux.

3) and 4) The addition of the irreversible conversion from isocitrate to α-ketoglutarate catalyzed by NAD- and NADP-specific forms of IDH enzyme, has produced in cancer models a decrease of glucose uptake flux with a reduction or no production of lactate, and an increase, more evident in breast and lung cancer than in liver tumor, of ATP production by OXPHOS pathway. Also in this case the flux associated with biomass synthesis has not changed its value after the perturbation.

After that, we identified all the reactions included in the three tumor models or in some cases at least in two of them, but that are absent in the reference model (see Supplementary materials, Table S6). When investigating the effect of their addition to the reference model, we noticed that the demand reaction for α-ketoglutarate causes a reversion from an active OXPHOS to a more glycolytic phenotype even if the lactate production does not occur. In addition, this exchange reaction also produces an increase in cell proliferation rate of about 1.4 fold. The reference model showed a more "cancerous" behavior even after the replacement of the two irreversible reactions catalyzed by isocitrate dehydrogenase enzyme, with their reversible form. Indeed, after this modification, we

observed a higher glucose and glutamine uptake, respectively of about 4 and 1.5 fold, production of lactate, a sharp drop of the OXPHOS pathway rate and an increase of the cell proliferation rate of about 1.5 fold. Lastly, although to a lesser extent, also the conversion reaction from oxaloacetate to phosphoenolpyruvate catalyzed by PCK1 enzyme caused a slight flux variation in the reference model. Indeed, even if the glycolytic flux remains very low and there is no production of lactate, a reduction of the ATP production by OXPHOS pathway (about 1.3 fold) and a nearly null variation of the biomass synthesis rate occur.

In the light of the considerations made on all the identified reactions that singularly caused a great flux variation, we tried to join together their activities. In the reference model, we combined the removal of inorganic phosphate and water transport reactions with the addition of the α-ketoglutarate exchange reaction and of PCK1-catalyzed conversion reaction from oxaloacetate to phosphoenolpyruvate and the reversible form of IDH-catalyzed reactions. This combination produced a full reversion of the reference model towards a cancer phenotype (see Supplementary materials, Table S7). Indeed, we measured, with respect to the original reference model, an increase of glucose and glutamine uptake, respectively, of about 4 and 1.4 fold, a high production of lactate, a reduced activity of all the components of the OXPHOS pathway with a complete inhibition of the complex I activity associated with a sharp drop of the flux associated with the reaction catalyzed by ATP synthase enzyme (about 13 fold), and a biomass synthesis rate 1.4 fold higher.

This same set of reactions is involved in the reversion of the cancer models towards a phenotype very similar to that of the reference model (see Supplementary materials, Table S7). In particular, in breast and lung tumor models, we combined the addition of inorganic phosphate and water transport reactions with the removal of the α-ketoglutarate exchange reaction, the addition of the irreversible form of IDH-catalyzed reactions and the removal (in breast cancer) or addition (in lung cancer) of the PCK1-catalyzed reaction. After these changes, we observed a decrease of both glucose and glutamine uptake, respectively, of about 4–5 fold and 1.3–1.4 fold, no production of lactate, a completely functioning OXPHOS pathway with an up-regulation of ATP synthase activity of more than 10 fold in breast cancer and of about 1.4 fold in lung cancer (in which the ATP synthase flux is already high in the initial model), but also of the other complexes, and a decrease in biomass production rate of about 1.1–1.4 fold. The liver cancer represents an exception due to the fact that this same set of changes produced different results as compared to lung and breast cancers: although the perturbation similarly results in an interruption of lactate production, the glucose uptake rate enhances by 1.5 fold, the OXPHOS pathway fluxes

remain very low, and neither the glutamine uptake nor the biomass synthesis rate is modified. Therefore, we decided to focus our attention on other reactions that are present in the reference model but not in the liver cancer model. We observed that the addition of the L-glutamate 5-semialdehyde transport reaction between cytosol and mitochondria, together with the previous changes, causes a complete reversion of the liver cancer model. Indeed, a decrease of both glucose and glutamine uptake, respectively, of about 4 fold and 1.4 fold, no production of lactate, a completely functioning OXPHOS pathway with a deep up-regulation of ATP synthase activity and, at last, a decrease in biomass production rate of about 1.4 fold occur.

Remarkably, the heterogeneity of the responses of different cancer types to a given perturbation supports the importance of developing cancer type-specific models.

Fragility points of cancer networks Since a high number of marked differences have emerged from the analysis of fluxes distribution in reference and cancer core models, it is interesting to identify which reactions are responsible for the flux rewiring between the two conditions under consideration. Given that we are studying cancer cells growth, we decided to focus on all the metabolic reactions that, after an inhibition, are able to cause a reduction of the flux associated with the biomass synthesis, because they could potentially represent drug targets for cancer treatment. Therefore, we performed single "*in silico*" inhibitions for each reaction of the three cancer core models and we then assessed the effect of these perturbations on flux distribution within the model. Among the identified reactions, we considered those that imply a strong reduction of the biomass synthesis rate (by at least 40–50%), which are likely potential targets to counteract the tumor growth.

In every cancer core model, the highest reduction of biomass has been caused by reactions belonging to the glycolytic pathway (see Supplementary materials, Table S8). Here, we observed that, after the inhibition of the glycolytic enzymes, cancer cells attempt to reorganize fluxes in order to try to maximize their biomass. However, these reactions are crucial for cancer growth and proliferation and, therefore, the outcome is a deep lowering of the biomass synthesis rate and the fluxes corresponding to the reactions involved in the synthesis of other biomass precursors.

Interruption of glycolysis has produced, as further effect, the annulment of the Warburg effect. Indeed, glycolytic reactions correspond also to the top elements in the list of reactions that, if inhibited, cause a decrease of the two

indices, AFR and EOR, connected to the quantification of the Warburg effect. In this context, we observed no production and secretion of lactate and, at the same time, a marked increase of the mitochondrial respiration. Cancer cells must bypass the missed production of ATP through glycolysis (which constitutes the main source for the production of ATP in tumor cells), enhancing the OXPHOS rate. However, clearly, this is not the optimal way for cancer cells to maximize the biomass synthesis and, thus, their growth.

The decrease in tumor biomass production rate caused by the inhibition of glycolytic enzymes supports several features already illustrated in literature. A number of studies focused their attention on the correlation between glycolysis and cancer growth. Indeed, it has been observed that an impaired glycolysis impacts tumor growth in a decisively negative and harmful way, constituting therefore a potential target for the anti-cancer therapy [165]. Many of the glycolytic enzymes and intermediates play a role in several non-glycolytic processes, acting for example in the up-regulation of cell cycle, in the maintenance of cellular redox balance, in the chemoresistance and to antagonize the proapoptotic machinery, facilitating in this way cancer cells growth and survival. Moreover, in 2007 Bonnet et al. [169] demonstrated that the reversal of the glycolytic phenotype toward a more active OXPHOS, as we showed in our results, can induce cancer cells death.

In conclusion, in cancer cells an up-regulated glycolysis gives cancer cells an advantage mainly for three reasons: 1) even though this pathway produces a lower quantity of ATP molecules compared to OXPHOS, the rate of ATP production may be 100 times faster with glycolysis [165]; 2) beyond ATP production, an up-regulated glycolysis allows to maintain a high lactate production and secretion, increasing the invasiveness and metastatic potential of the cancer cells; 3) the accumulation of glycolytic enzymes may promote some of the pathways required to increase the tumor mass during growth and proliferation,which are the PPP pathway and the amino acids serine and glycine synthesis.

The inhibition of glycolytic enzymes with specific inhibitors can sharply reduce cancer growth and proliferation. To date, an example of inhibitor (recently entered in the early phase of the clinical trials [165]) is the pyruvate analog 3-bromopyruvate (3-BrPA), which shows a high ability to prevent the tumor glycolysis, as well as a high specificity and selectivity, both in vitro and in vivo, for the glycolytic enzyme glyceraldehyde 3-phosphate dehydrogenase.

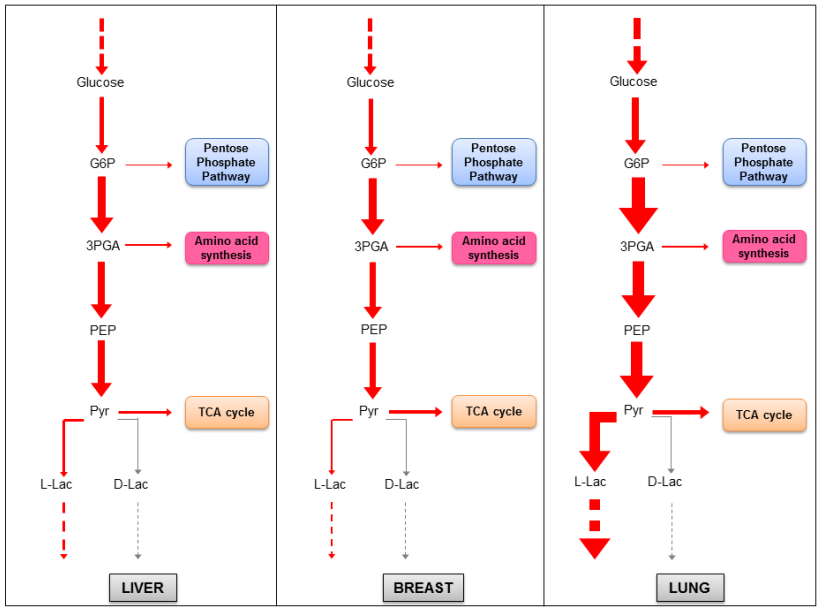


Figure 3.8: Graphical representation of the metabolic flux redistribution in glycolysis within the three investigated cancer types. An heterogeneous up-regulation of the entire metabolic pathway (indicated by red arrows) emerged in cancer cells with respect to the reference cell. The grey arrows correspond to reactions having a null flux. The thickness of each arrow is proportional to the flux value of the corresponding reaction scaled by 100 for the sake of representation. A list of the abbreviations used in the map is provided in Appendix A. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

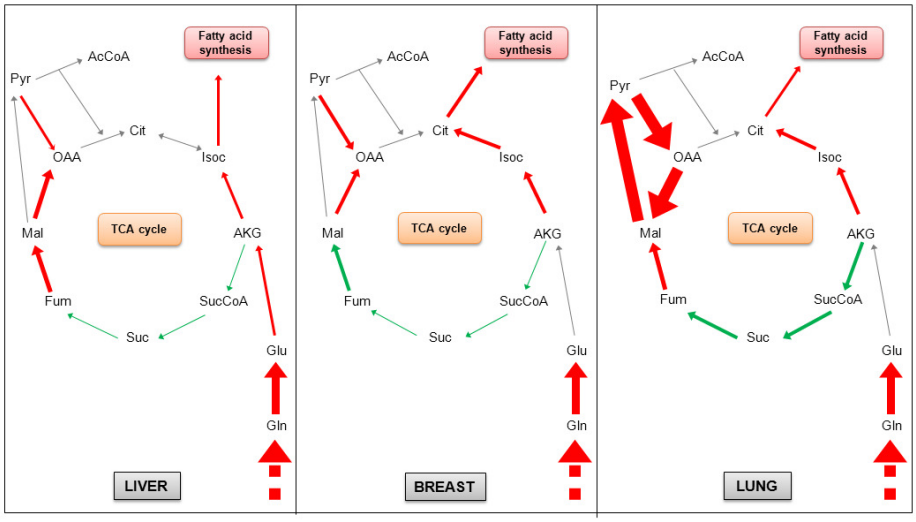


Figure 3.9: Graphical representation of the metabolic flux redistribution in TCA cycle within the three investigated cancer types. An heterogeneous regulation of the entire metabolic pathway emerged in cancer cells with respect to the reference cell. The red and green arrows correspond, respectively, to the emerged up- and down-regulations. The grey arrows correspond to reactions having a null flux. The thickness of each arrow is proportional to the flux value of the corresponding reaction scaled by 100 for the sake of representation. A list of the abbreviations used in the map is provided in Appendix A.

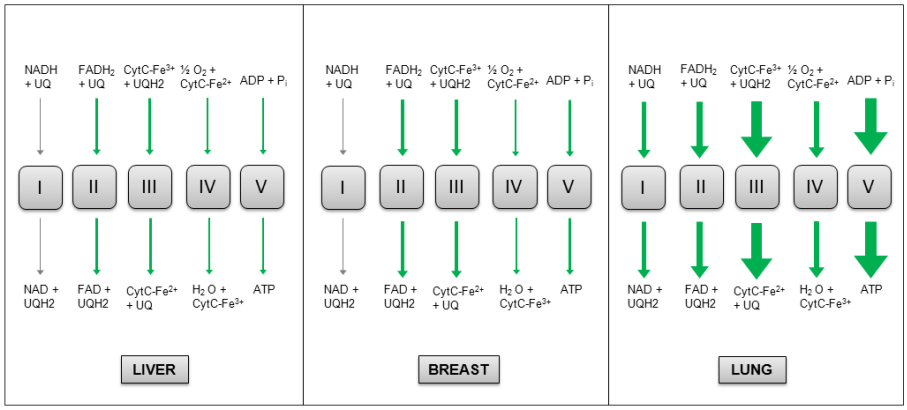


Figure 3.10: Graphical representation of the metabolic flux redistribution in oxidative phosphorylation within the three investigated cancer types. An heterogeneous down-regulation of the entire metabolic pathway (indicated by green arrows) emerged in cancer cells with respect to the reference cell. The grey arrows correspond to reactions having a null flux. The thickness of each arrow is proportional to the flux value of the corresponding reaction scaled by 100 for the sake of representation.A list of the abbreviations used in the map is provided in Appendix A.

Conclusions

In this work, we pointed out the potentiality of the mathematical modeling of complex biological systems in relation to the understanding of cancer metabolic rewiring. In order to identify and study the role of all metabolic alterations supporting the neoplastic proliferation, we performed a comparative analysis between a reference and a cancer condition starting from the automatically generated genome-scale networks of the Human Metabolic Atlas database concerning a generic reference cell and three different types of tumor. In the context of cancer cells metabolism, genome-scale modeling represents an increasingly used approach. However, these networks are often not adequately curated and include errors, such as metabolic gaps, that imply the need to perform a further manual curation phase (as occurred in our work). Starting from the chosen genome-scale models, we manually reconstructed core constraint-based metabolic models that zoom-in on cancer metabolic rewiring, considering only specific aspects of metabolism. Unlike genome-scale models, core models are much less difficult to control and their curation may be more accurate.

A constraint-based approach, the Flux Balance Analysis, allowed to identify several metabolic alterations in cancer cells models, which may represent potential targets for the development of therapies able to counteract cancer cells growth and proliferation. Indeed, following the flux distributions analysis in the four core models, we observed a clear distinction between reference and cancer condition under several aspects: growth rate, emergence of the Warburg effect, lactate secretion, glucose and glutamine uptake, TCA cycle, and the oxidative phosphorylation flux values. Two further analyses on the reconstructed core models allowed to detect other potential drug targets, both analyzing the structural differences in core metabolic networks that are responsible for a reversion of the tumoral phenotype towards the reference one, and identifying all the fragility points within the cancer models corresponding to reactions that, if perturbed, cause a negative effect on the system.

The constraint-based models also highlighted a heterogeneity in terms of flux values not only between reference and cancer conditions, but also among the three different cancers, concerning the pathways shown in Figs. 4.9-4.10, which are glycolysis, TCA cycle and OXPHOS. This result strengthens the need to focus the attention on several types of tumors rather than on a generic cancer cell, because, although some metabolic features are associated to the three cancers, relevant variations emerged among them. These ones are of great interest in the medical field for the identification of cancer type - specific drug targets in order to develop more effective treatments. The emerged heterogeneity high-

lights how, given a specific phenotype of interest (cancer, in our case), three different sub-phenotypes may be identified.

On the other hand, the analysis has revealed that some caution is required when dealing with tissue-specific models that have been automatically reconstructed from *−omics* data. It may indeed happen that, when looking for the optimal "active network", which include the least number of reactions catalyzed by proteins having a low expression level and the greatest number of reactions catalyzed by proteins having a high expression level, some reactions will be removed from the network, even if they are not necessarily inactive in the corresponding tissue (as it is the case for the reaction 3-phospho-D-glycerate ⇒ 2-phospho-D-glycerate and for inorganic phosphate and water transport reactions mentioned, respectively, in Section 3.2.1 and 3.2.1).

In this regard, an approach that is complementary to the reconstruction of distinct tissue-specific networks - and that may represent a possible future development of this work - is the ensemble-evolutionary FBA (eeFBA) approach introduced in [170], which allows to identify ensemble of possible flux distributions of a generic metabolic network that are compatible with different cancer phenotypes.

At last, since we have made the three cancer models comparable among themselves using a common reference model, and in order to identify specific peculiarity linked to a certain type of tissue that is not highlighted using the reference model, a further extension of our work will be the comparison between each of our three cancer core models and their corresponding healthy tissue-specific models recently published within the Human Metabolic Atlas database [171].

3.2.2 Dissecting glutamine roles in promoting proliferation in transformed mouse fibroblasts

Manuscript in preparation

Introduction

The enhanced growth phenotype that characterize K-Ras-transformed cells relies on a deep metabolic rewiring that includes a strict dependence from glutamine consumption and increased oxidative stress [131]. Given the increasing attention to the dependency of cancer cells on the consumption of glutamine, we investigated in this work the role of this carbon and nitrogen source in promoting the enhanced proliferation of K-ras-transformed NIH3T3 mouse fibroblasts (NIH-RAS). In particular, we explored two scenarios: a "standard condition" characterized by the presence of glutamine, and the "-GLN + α-KG + NEAA condition", where glutamine deprivation is complemented with other nutrients having analogous function, which are the α-KG for representing the carbon source, and the four non essential amino acids (NEAA) proline, alanine, aspartate and asparagine, that represent both the carbon and the nitrogen source.

The *in silico* analysis was preceded by an extensive experimental investigation that we exploited for imposing new boundaries on the model, or, in some cases, *a posteriori* for carrying out the validation step of our core model. The experimental co-supplementation of NIH-RAS cells with α-KG and NEAA under glutamine deprivation, occurred in equimolar amounts to the glutamine supplied in the standard condition. Nevertheless, while in this latter medium all the provided glutamine is uptaken, the NIH-RAS cells grown in the condition -GLN + α-KG + NEAA revealed the ability to just consume α-KG, aspartate and asparagine, but no evidence about the uptake of the two other amino acids, alanine and proline, emerged. We exploited this first experimental observation to differentiate in the core model the constraints relative to the exchange reactions of the four NEAA, by allowing the uptake just of aspartate and asparagine. The comparison of the experimentally determined cell viability in the two investigated scenarios highlighted that the combination of α-KG and NEAA only partly rescues the glutamine missing.

Redox unbalance has been observed in cells grown in the -GLN + α-KG + NEAA medium compared to those grown in the standard condition. The transcriptomic analysis revealed that, under glutamine deprivation, the fatty acids and cholesterol synthesis pathways, which are NADPH demanding processes, are downregulated. The low level of activity of these metabolic routes has been hypothesized to be dependent on the observed lower content of NADPH under the -GLN + α-KG + NEAA condition. In turn, this emerged redox unbalance could be closely linked to the activity of the mitochondrial glutamate dehydrogenase enzyme, which converts the glutamate into α-KG in parallel with the reduction of $NADP^+$ to NADPH. Under the -GLN + α-KG + NEAA

condition, the relevance of this reaction considerably decreases because of the α-KG supply in the extracellular environment, with the consequent no strictly longer requirement of the glutamate dehydrogenase, and the unbalance of the $NADP^+$/NADPH cellular ratio.

In line with literature findings, an altered glutamine supply in the medium of NIH-RAS cells has been subsequently observed to negatively impact the deoxyribonucleotides synthesis, with a consequent reduced cell proliferation [ref]. However, when deoxyribonucleotides are supplemented as extracellular source in the medium under the -GLN + α-KG + NEAA condition, the growth ability is just partially rescued.

Finally, the fluxome analysis of NIH-RAS cells grown in standard and in -GLN + α-KG + NEAA conditions has been performed to investigate the destiny of glucose, another fundamental nutrient of cancer cells, when α-KG and NEAA are supplied under glutamine deprivation. This experimental analysis revealed that, under standard medium, glutamine is highly consumed and used as anaplerotic source for the TCA cycle and for the synthesis of the NEAA, except for the alanine, whose production derives from the glucose. Moreover, in this condition, glutamine does not contribute at all to the synthesis of lactate, whose generation fully derived from the glucose. In the -GLN + α-KG + NEAA condition, the glutamine deprivation produced a rewiring of the entire cellular metabolism towards an increased dependence on glucose that, secreting less lactate, is redirected towards the production of amino acids, including serine, glycine and glutamate, with a considerable secretion of this latter in the extracellular environment. The uptaken asparagine amino acid from the extracellular environment only acts as biomass precursor, without contributing to the synthesis of any other molecule. On the contrary, aspartate, once enter the cells, is processed through the transaminase enzyme to produce glutamate. Under standard condition, glutamine undergoes reductive carboxylation as a way to synthesize lipids that are fundamental precursors of biomass. However, the fluxome analysis revealed that, under glutamine deprivation, α-KG that is consumed by the NIH-RAS cells is processed through the TCA cycle by following the canonical clockwise direction.

Reconstruction of the core metabolic network

In this work, we exploited the core metabolic model "HMRcore" that we presented in Section 3.2.1, by adapting it to the aim of this specific research. In particular, we checked for the presence in the model of the biochemical reactions relative to the metabolism of the four considered NEAA by adding, when re-

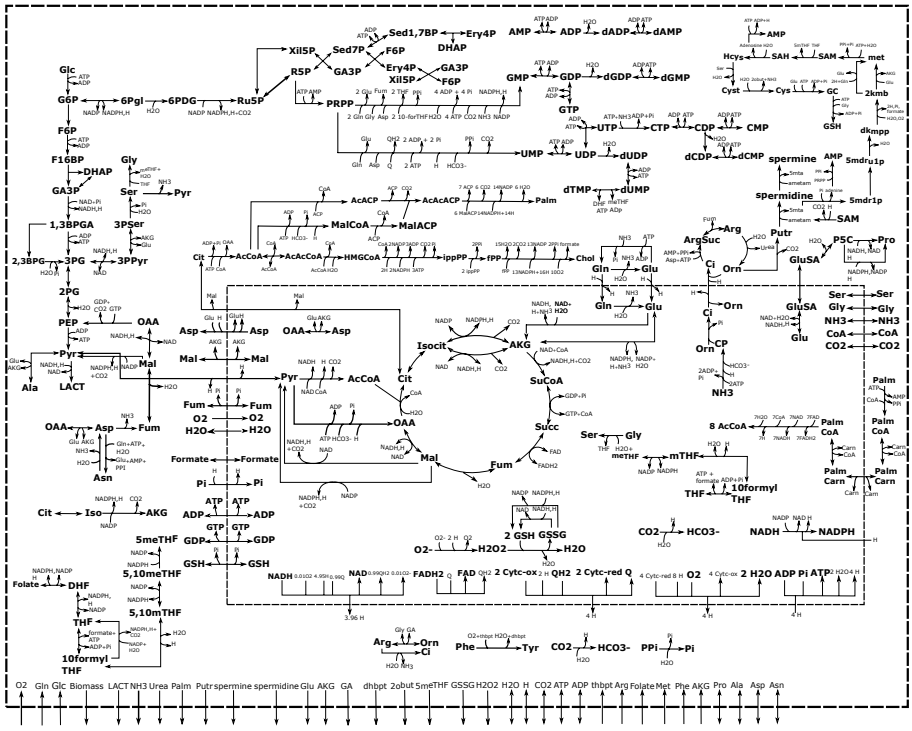


Figure 3.11: Graphical representation of the readapted HMRcore network. A list of the abbreviations used in the map is provided in Appendix A.

quired, the necessary reactions. Moreover, we included the exchange reactions for each NEAA and for the α-KG. In this model editing, according to biological insights derived from wet experiments, we further included the exchange reaction for the glutamate bounded in the direction of its secretion in the extracellular environment. Finally, we used the experimentally measured uptake rate of glucose, glutamine, oxygen, α-KG and the four NEAA to constrain the boundaries of the related exchange reactions.

The map of the final network is reported in Figure 3.11.

Constraint-based simulations

We performed FBA and FVA simulations on our core metabolic model by setting the boundaries reported in Table 3.4 in compliance with the experimental data, and choosing as objective function the maximization of the biomass synthesis reaction. By imposing these constraints, we obtained that in the standard condition the flux value of the biomass synthesis reaction was equal to 1.98 mmol gDW^{-1} h^{-1}, which is five times higher compared to the corresponding one in the -GLN + α-KG + NEAA condition, where its flux value was equal to 0.39 mmol gDW^{-1} h^{-1}. This first outcome was in very good agreement with the experimental NIH-RAS cells growth ratio of four between the standard and the -GLN + α-KG + NEAA condition.

Reaction	**STD**	**–GLN+α-KG+NEAA**
Uptake Glc	-25	-25
Uptake Gln	-4	0
Uptake α-KG	0	-1.6
Uptake Asp	0	-2.8;-2.4
Uptake Ala	0	0.2;0.8
Uptake Asn	0	-3.2;-2.4
Uptake Pro	0	1.2;2
LDH flux	45	33
Secretion of Glu	1;1000	8;1000

Table 3.4: Range of flux boundaries used for the exchange reactions reported in the column "Reaction" under standard condition (STD) and glutamine deprivation condition supplemented with α-ketoglutarate and non essential amino acids (–GLN+α-KG+NEAA). All the values are expressed in mmol gDW^{-1} h^{-1}. A given flux boundaries range indicated as a single value means that lower and upper bound coincide.

By removing any specific constraint on the glutamate secretion rate, we obtained very similar flux values for the biomass synthesis reaction in the two investigated scenarios. In particular, a flux value of 2.20 mmol gDW h^{-1} and 2.15 gDW h^{-1} have been returned, respectively, in the standard condition and in the -GLN + α-KG + NEAA condition. Following this result, we hypothesized that in the -GLN + α-KG + NEAA condition there is an accumulation of glutamate that is not utilized and therefore is secreted in the extracellular environment. A possible reason behind this result can be attributed to a link with the glutamine synthetase (GS) enzyme, whose role is to convert glutamate

back to glutamine at the expense of one molecule of ATP. However, by analysing transcriptomics data, the GS enzyme did not result as differentially expressed between the two investigated conditions. Folliwing this result, we formulated two hypotheses. In the first one, a too high threshold value is chosen for discriminating significant fold changes of expression level in the two nutrients scenarios. In the second hypothesis, this outcome could be caused by an equal and, at the same time, low expression level of GS enzyme that, being active in both conditions, does not fall within the list of differentially expressed genes. However, its low activity level could be responsible of the glutamate secretion under the -GLN + α-KG + NEAA condition. This last hypothesis resulted the most likely scenario. Indeed, both transcriptomics data analysis and FBA simulations has enabled to exclude the first hypothesis. Indeed, in the standard condition, even if an experimental glutamate secretion has not been observed, the flux of the GS-catalyzed reaction is not completely null. On the contrary, FVA revealed a high flux variability range between a minimum of 0 mmol gDW^{-1} h^{-1} and a maximum of 12.69 mmol gDW^{-1} h^{-1}. In the -GLN + α-KG + NEAA condition, the flux of this reaction fell between a positive minimum value of 0.18 mmol gDW^{-1} h^{-1} and a maximum of 154.93 mmol gDW^{-1} h^{-1}. This outcome means that, under this environmental condition, GS-catalyzed reaction must be used to synthesize the glutamine since it is not supplied in the medium. In the standard condition, the emerged flux range instead indicates that, among the alternative optimal solutions, GS-catalyzed reaction could also be blocked.

The role of GS-catalyzed reaction under the -GLN + α-KG + NEAA condition

Given the hypothesized relevance of the GS-catalyzed reaction, we looked into this matter by evaluating, under the standard and the -GLN + α-KG + NEAA condition, the effect caused by the progressive reduction of its flux value on the biomass synthesis reaction. In the standard condition the biomass synthesis rate did not suffer any effect. Due to the null minimum limit of GS-catalyzed reaction flux range, the optimal biomass synthesis flux value may coexist with a null flux of the GS-catalyzed reaction. On the contrary, as shown in Figure 3.12A, in the -GLN + α-KG + NEAA condition, we observed that the gradual reduction of the flux through the GS-catalyzed reaction caused a progressive decrease of the flux through the biomass synthesis reaction until zero when flux passing through the GS-catalyzed reaction is null. As shown in Figure 3.12B, we also found that, by reducing the flux through the GS-catalyzed reaction, the maximum allowable flux value of the glutamate secretion reaction spontaneously

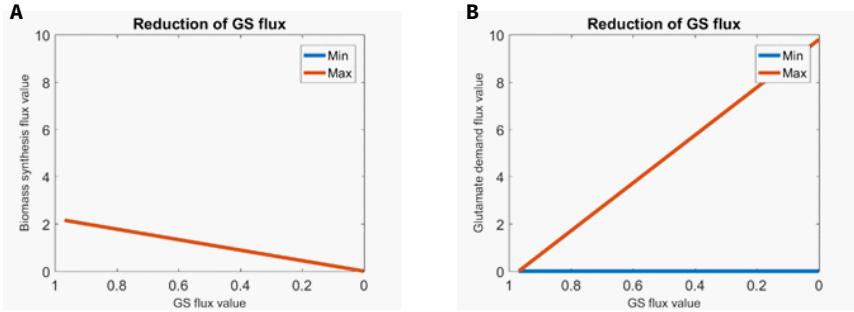


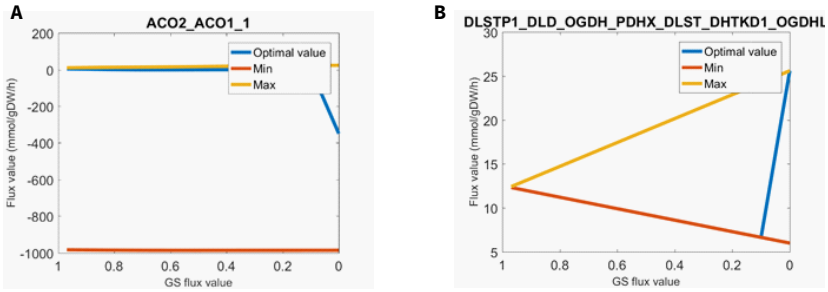
Figure 3.12: (A) Biomass synthesis and (B) glutamate secretion reaction flux value following the progressive reduction of the flux through the GS-catalyzed reaction under the -GLN + α-KG + NEAA condition. Blue and red lines correspond, respectively, to the minimum and maximum flux value of the corresponding reaction. Flux values are expressed as mmol gDW^{-1} h^{-1}.

increases. This outcome strengthens the previously discussed hypothesis of a putative link between the activity of the GS enzyme and the glutamate secretion in the extracellular environment.

As shown in Figure 3.13, under the -GLN + α-KG + NEAA condition, we also observed that when the flux through the GS-catalyzed reaction is gradually reduced, the minumum flux of the aconitase-catalyzed reaction, which reversibly converts the citrate into isocitrate, progressively increas. Analogous situation occurred for the maximum flux of the reaction catalyzed by the α-KG dehydrogenase, which converts the α-KG into succinyl-CoA. The flux variation of these two reactions belonging to the TCA cycle means that the canonical forward direction of that cyclic pathway is increasingly preferred over the reductive carboxylation pathway characterizing the standard condition.

Integration of transcripts fold change of differentially expressed genes

We tried to integrate the experimental fold changes of the emerged differentially expressed genes into our core model. For that purpose, we performed a FVA analysis under the standard condition without imposing any optimality condition and removing the boundary on glutamate production. In this way, consistently with the constraints specific of this nutritional condition, we obtained the allowable minimum and maximum fluxes for each reaction in the



C

Reazione	Flusso GS = 0.969	Flusso GS = 0
Aconitasi	Min: 10.698 Max: -982.909	Min: 24.2 Max: -985.782
AKGDH	Min: 12.321 Max: 12.425	Min: 6 Max: 25.6

Figure 3.13: (A) Aconitase- and (B) α-KG dehydrogenase-catalyzed reactions following the progressive reduction of the flux through the GS-catalyzed reaction under the –GLN + α-KG + NEAA condition without supplementing deoxynucleotides in the medium. Blue line corresponds to the optimal value of the corresponding reaction returned by the Flux balance analysis. Red and yellow lines correspond, respectively, to the minimum and maximum flux value of the corresponding reaction returned by the flux variability analysis. Flux values are expressed as mmol gDW^{-1} h^{-1}. (C) Detail of flux range of aconitase- and α-KG dehydrogenase-catalyzed reactions when GS-catalyzed reaction flux is equal to 0 and to 0.969 mmol gDW^{-1} h^{-1}.

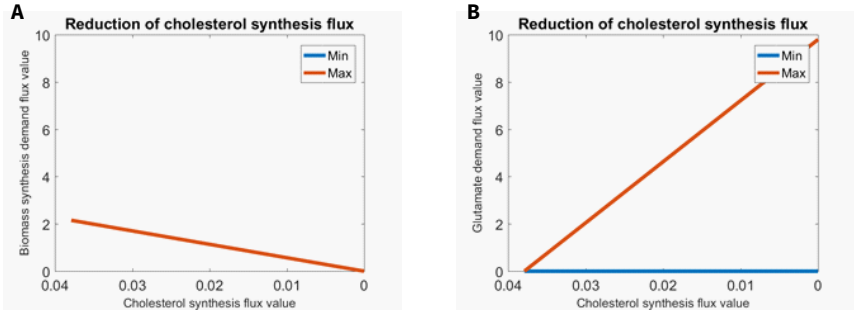


Figure 3.14: (A) Biomass synthesis and (B) glutamate demand reactions flux value following the progressive reduction of the flux through the cholesterol synthesis reaction under the -GLN + α-KG + NEAA condition. Blue and red lines correspond, respectively, to the minimum and maximum flux value of the corresponding reaction returned by the flux variability analysis. Flux values are expressed as mmol gDW^{-1} h^{-1}.

model, regardless of any objective function. We exploited these flux ranges to constrain lower and upper bound of each reaction under the -GLN + α-KG + NEAA condition. We then integrated the experimental fold changes to modify the flux range of each reaction according to the emerged difference with respect to standard condition. Following a FBA simulation, we obtained a flux value for the biomass synthesis reaction of 0.11 mmol gDW^{-1} h^{-1}.

Given the experimentally observed down-regulation of the cholesterol synthesis, we progressively reduced the flux through the reaction responsible for performing this task in the core model. As shown in Figure 3.14, this perturbation causes the progressive reduction of the biomass synthesis rate (plot on the left) with a concomitant progressive increase of the glutamate secretion rate (plot on the right).

In silico extracellular supplementation of deoxyribonucleotides

Under the -GLN + α-KG + NEAA condition, deoxyribonucleotides synthesis is negatively impacted. However, the subsequent supplementation of an extracellular pool of deoxyribonucleotides can ripristinate the NIH-RAS cells growth.

We exploited this experimental observation to validate our core model. In this regard, we performed FBA simulations under the -GLN + α-KG + NEAA

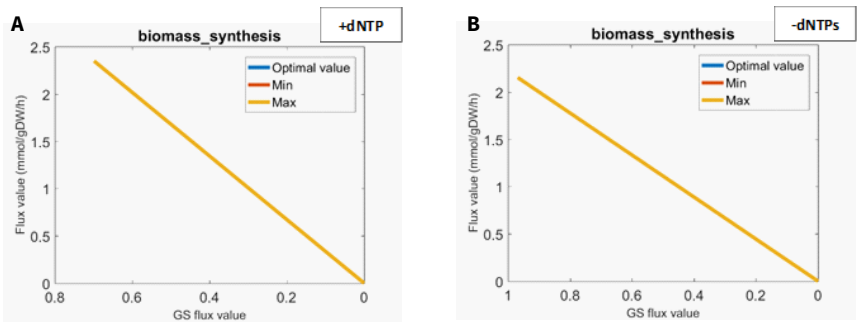


Figure 3.15: Biomass synthesis reaction flux value following the progressive reduction of the flux through the GS-catalyzed reaction under the -GLN + α-KG + NEAA condition (A) by supplementing and (B) not supplementing deoxynucleotides in the medium. Blue line corresponds to the optimal value of biomass synthesis reaction returned by the Flux balance analysis. Red and yellow lines correspond, respectively, to its minimum and maximum flux value returned by the flux variability analysis. Flux values are expressed as mmol gDW^{-1} h^{-1}.

condition by gradually reducing the flux through the GS-catalyzed reaction. After that, we compared the two conditions where the deoxynucleotides are in one case missing and in the other case present in the medium.

As shown in Figure 3.15, we observed that when deoxynucleotides are provided in the extracellular environment the biomass flux is of 2.35 gDW h^{-1}. On the contrary, the lack of deoxynucleotides in the medium led to a considerable reduced flux value of 1.5 gDW h^{-1}. This biomass synthesis rate reduction of 36% supports the experimentally observed positive role of extracellular deoxynucleotides on the growth rate.

3.2.3 Reconstruction of human core model of central carbon metabolism: ENGRO2

Manuscript in preparation

Introduction

In [31], the constraint-based approach has been exploited to gain new knowledge about the logic of metabolic reprogramming in promoting tumour cells proliferation. In this regard, a core model of human central metabolism, called "ENGRO1", has been developed, by including the most known pathways that are required for biomass production starting from the two most relevant cellular nutrients, which are glucose and glutamine. The final aim was to evaluate the possible metabolic rewirings for the purpose of biomass formation, under different nutritional conditions. It was observed that when available glucose and glutamine carbon atoms cannot be fully oxidize because of insufficient availability of oxygen, biomass production takes advantage from the utilization of glutamine through the reductive carboxylation pathway. Moreover, both glucose and glutamine contribute to the synthesis of lactate.

Starting from the outcomes presented in [31] and the developed core model ENGRO1, we reconstructed a more extended and curated constraint-based core model of the human central metabolism that we called "ENGRO2".

Reconstruction of ENGRO2 model

The reconstruction of ENGRO2 model was based on a step-wise manual procedure that started from the elements already present in ENGRO1 model followed by a progressive inclusion of specific pathways or reactions according to their relevance in literature for cancer cells.

The most invasive change we implemented in the model was the compartmentalization of reactions and metabolites that ENGRO1 does not consider since just the internal and the extracellular environments were included. As previously anticipated in Chapter 2, many studies support the evidence of an altered expression of some mitochondrial carriers in multiple cancer cells that, most probably, arise as an adaptation to their current metabolic state and the consequent new requirements [172]. Therefore, in addition to the extracellular compartment, we divided the internal model side into cytosol and mitochondrion, and we catalogued all the so far included reactions depending on their localization according to literature knowledge. This first important change implies the necessity to add specific transport reactions in order to link biochemical transformations occurring within cytosol and mitochondrial matrix.

Another important change introduced in the model regarded the inclusion of the required cofactors within the stoichiometric equation of the considered reactions. In particular, we included coenzyme A (CoA), inorganic phosphate

(P_i), proton (H^+), water (H_2O) and carbon dioxide (CO_2). Cofactors are additional chemical components that are required by some enzymes for assisting their activity. Recent findings went beyond their classic role, by correlating altered levels of some cofactors to the emergence of various side effects. For example, an altered CoA homoeostasis was found to be largely influential not only on the metabolic reactions involving this cofactor, but also on the emergence of neurodegeneration. In particular, an altered CoA homoeostasis influences the process of protein acetylation, which is a post-translational modification that affect the protein function. Moreover, the role of CoA as been also extended to anti-apoptotic signalling molecule [173]. Other experimental evidences highlighted a positive involvement of the cofactors transport in promoting the cancer cells. In this regard, the overespression of the aquaporins, which are the enzymes that move water molecules between different compartments, has been associated to tumour cells migration, proliferation, angiogenesis, and to the tumour grade [174]. The emergence of many chronic diseases, including type 2 diabetes mellitus, cancer, cardiovascular diseases, and obesity, has been associated to high uptake levels of P_i, and high concentrations in the serum. Phosphate is mainly intaken due to diets increasingly richer in phosphorus, including restaurant meals, fast foods, and cheap foods [175]. For these reasons, we deemed it necessary to consider the contribution of these elements into the network under reconstruction.

We then proceeded the reconstruction procedure of ENGRO2 by making explicit all the reactions included in the oxidative phosphorylation (OXPHOS) pathway. A lumped version of the entire route was included in ENGRO1 model in the form of two reactions explaining the NADH and $FADH_2$ oxidation through transfer of electrons from these two reducing agents to oxygen. The stoichiometry of the complex I-like reaction also included the generation of reactive oxygen species (ROS) that are known to be generated by a deficient Complex I activity following specific mutations affecting subunits of this enzymatic complex [176, 177]. In ENGRO2 network, we included five separate reactions representing the reaction catalyzed by each OXPHOS complex. Because of the ROS production from the Complex I, it was necessary to add the ROS detoxification pathway by means of glutathione.

The one-carbon metabolism has been proved to be required by cells for nucleotide synthesis, methylation and reductive metabolism, supporting the high proliferative rate of cancer cells [178]. It includes a set of reactions that are centered around folate and methionine cycles generating one-carbon units, also known as methyl groups, that are used for the purine and pyrimidine nucleotides synthesis, which are essential for DNA and RNA production. Consequently, the

high proliferation rate characterizing tumour cells causes their dependence on nucleotides synthesis pathway and led them to be addicted to this pathway. Furthermore, another scope of one-carbon metabolism in cancer cells is linked to its usage for DNA methylation and the consequent gene expression regulation. For example, it impacts on tumour-suppressor gene promoters expression, RNA and protein methylation to, respectively, regulate gene translation and protein function. In view of the relevance emerged for one-carbon metabolism in cancer cells, we enriched the ENGRO2 model by including both the folate and the methionine cycles.

In the first version of the model ENGRO1, the oxidative branch of the pentose phosphate pathway was just partially included until the synthesis of the phosphoribosyl pyrophosphate (PRPP). PRPP is produced from the ribose-5-phosphate through the ribose-phosphate pyrophosphokinase enzyme, and is fundamental for cell biomass synthesis as entry point of nucleotides biosynthesis. Following the above described implications of the one-carbon metabolism in purines and pyrimidines production, we extended the pentose phosphate pathway by including the complete nucleotides biosynthesis route. Moreover, we also included all the non-oxidative branch of this pathway because of its relevance in reconverting the intermediates of this pathway into glycolytic metabolites.

Fatty acid oxidation pathway, also known as beta-oxidation pathway, mainly occurs within mitochondria. It involves 7 cyclical series of reactions, whose role is to reduce fatty acids length through the removal of two carbon atoms per cycle. At each cycle, NADH, $FADH_2$ and acetyl-CoA are generated. However, in the last cycle just two acetyl-CoA are produced because of the catabolism of a four-carbon molecule. Subsequently, NADH and $FADH_2$ deriving from fatty acids catabolism enter as reducing agents in the OXPHOS for producing ATP. Recent evidence highlighted a crucial role of beta oxidation in tumour cells for providing them fueling growth sources, and benefiting tumour survival especially under metabolic stress, such as following glucose or oxygen deprivation [179]. Moreover, fatty acid oxidation contributes to the total cell NADPH pool, because of acetyl-CoA that, once produced, enters the TCA cycle and is converted with the OAA to citrate. This latter is transported into the cytoplasm where it is involved in NADPH-producing reactions, including the conversion of malate to pyruvate through the malic enzyme, and the oxidation of isocitrate into α-KG catalyzed by the isocitrate dehydrogenase. As already explained in the previous Chapter, the high relevance of NADPH is linked to its role in providing redox power for tumour cells to counteract the oxidative stress. For these reasons, we considered necessary to include the mitochondrial beta-oxidation pathway in ENGRO2 for degradating fatty acids that in the model are modeled in the

form of palmitate.

Another recent findings we exploited for the reconstruction of ENGRO2 regarded a disregulation of the polyamines metabolism and its requirement under the neoplastic condition [180]. Polyamines belong to a family of molecules deriving from the ornithine, including putrescine, spermine and spermidine. These molecules are implicated in various important cellular functions, including nucleic acid and chromatin structure maintenance, ion channel regulation, protein synthesis, substrates for transglutaminase reactions, and free radical scavengers to protect nucleic acids from damage. Moreover, more beneficial for cancer cells, polyamines has recently been proved to be involved in cell migration and metastasis. In view of these evidence, we added reactions belonging to the polyamines metabolism in ENGRO2 model.

Finally, the extension of ENGRO1 to ENGRO2 model also included the addition of all the reactions belonging to metabolism of non-essential and essential amino acids. These latters have the peculiarity that they cannot be synthesized by the organism. On the contrary, they need to be supplemented in the diet and then consumed from the extracellular environment as further medium nutrients.

Following all these implemented changes, the final version of ENGRO2 core model consists of 373 reactions and 327 metabolites, whose graphical representation, for reasons linked to its size, is splitted between Figure 3.16 and 3.17, depicting, respectively, all the inserted reactions, and the metabolism of the essential amino acids. This step-wise reconstruction process further included all the necessary transport and exchange reactions, and an accurate manual curation phase to check the presence of stoichiometrically incorrect reactions or network gaps, by relying on the biological database KEGG and on the most up-to-date human metabolic reconstruction Recon 2.

In a next phase, we will use this core model as scaffold for integrating transcriptomic data relative to multiple cancer cell lines to create cancer cells line-specific core models.

Computational simulations of ENGRO2 model

We performed constraint-based simulations of ENGRO2 model for investigating the metabolic strategies adopted by the model when under cancerous condition it is subjected to different perturbations. In this way, it is possible to explore the role of specific pathways, and identify the major network players supporting cancer growth under various conditions. Considering the maximization of growth as most plausible objective of cancer cells, we carried out FBA and FVA simulations by choosing as objective function the maximization of the biomass

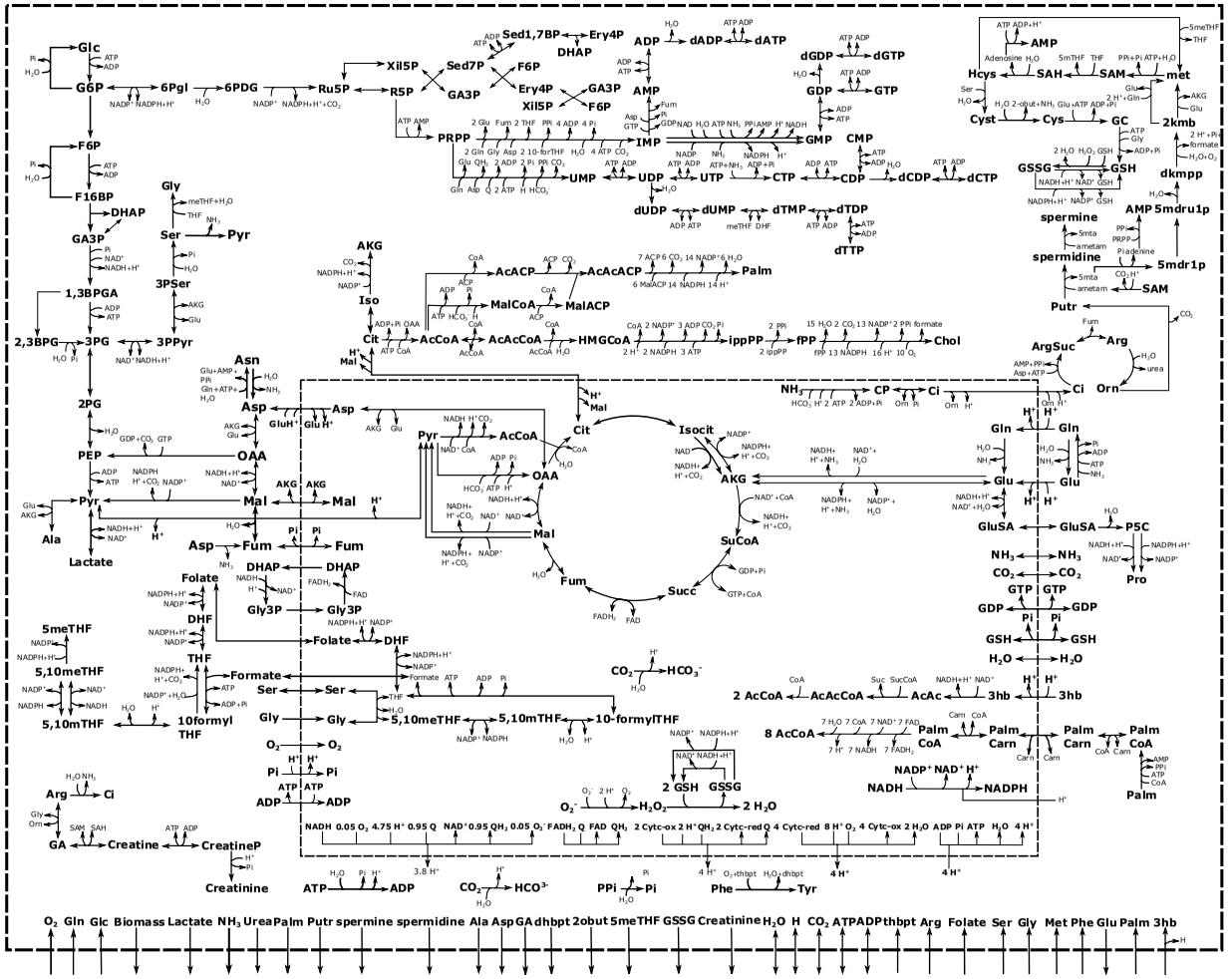


Figure 3.16: Graphical representation of the ENGRO 2 core model. Only the central carbon metabolism is depicted. A list of the abbreviations used in the map is provided in the Appendix A.

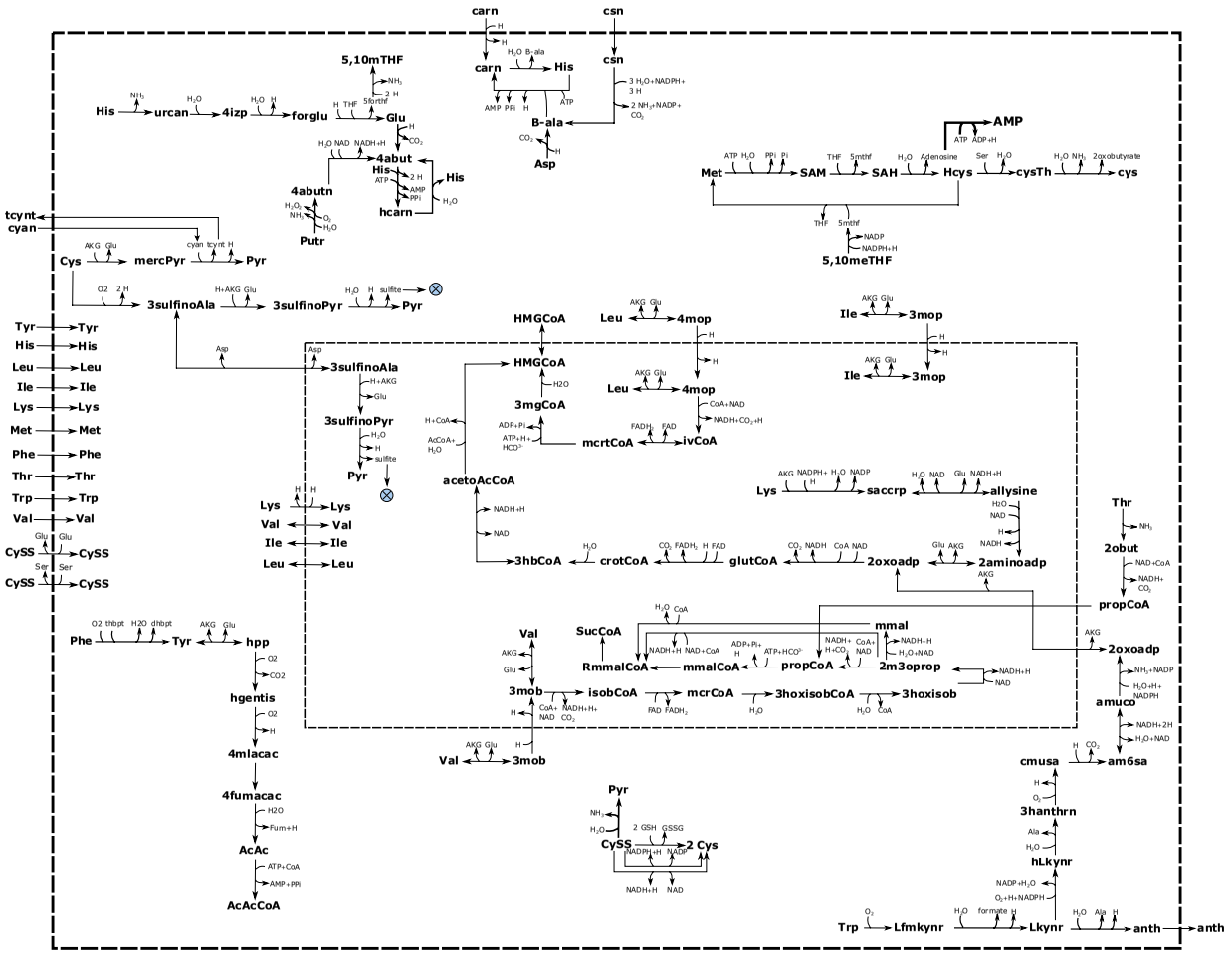


Figure 3.17: Graphical representation of the ENGRO 2 core model. Only essential amino acids metabolism is depicted. A list of the abbreviations used in the map is provided in the Appendix A.

synthesis reaction. In this regard, we exploited the macromolecular synthesis requirement for proteins, DNA, RNA, lipids, and carbohydrates that is expressed in [138], to construct the biomass synthesis pseudo-reaction. We included in this reaction all the required biomass precursors m, by setting the corresponding stoichiometric coefficient s_m as follows:

$$s_m = \frac{f_{\mathrm{m}} \cdot f_P \cdot 10^3}{\omega_{\mathrm{m}}}$$

where f_{m} represents the fraction in weight of the monomer m into the macromolecule P; f_{P} represents the fraction in weight of the macromolecule P into the biomass, and ω_{m} is the molecular weight of the monomer m.

We imposed constraints on the extracellular environment in order to replicate the growth on the medium that is in paralled used for the wet experiments. The *in silico* medium resulted composed by the main nutrients glucose, glutamine, together with arginine, cystine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, tyrosine, valine and folic acid. Being interested on the ratio among the various nutrients more than to their absolute values, we used the concentration of each component in the experimental medium as temporary lower bound of the corresponding exchange reaction in the *in silico* environment. With these constraints, we then performed a parsimonious FBA (pFBA) to compute the minimum consumption rate of each exchange reaction that is, at the same time, consistent with an optimal tumour biomass synthesis rate. We exploited the outcome returned from pFBA for these reactions to set their final lower bound.

Another crucial nutrient that must be present in the *in silico* environment is the oxygen, whose boundary is one of the essential factor contributing to the determination of the tumoural condition. In [181], intermediate oxygenation values for multiple tumour and normal tissue are provided, by giving an average value of 4.6 as fold change reduction of oxygen levels under hypoxia versus normoxia condition. We exploited this value to reduce the lower bound of the oxygen uptake rate previously returned by the pFBA [181].

Once all the constraints are set, we carried out a flux variability analysis (FVA) to calculate the flux value range for each model reaction. In Table 3.5, flux ranges of the previously constrained exchange reactions are reported.

Main ENGRO1 predictions are confirmed in ENGRO2

To investigate the metabolic phenotype resulting from ENGRO2 *in silico* simulations, we carried out an in depth analysis of the resulting flux distributions,

Table 3.5: Outcome of flux variability analysis (FVA) simulation of ENGRO2 model for the exchange reactions reported in the "Exchange reaction" column. The negative sign is a convention used for the exchange reactions to indicate the consumption of the corresponding metabolite. Flux values are expressed as mmol gDW^{-1} h^{-1}. A single flux value instead of a flux range means that minimum and maximum boundaries of flux range coincide.

Exchange reaction	**FVA flux range** $mmolgDW^{-1}h^{-1}$
Biomass	0.586
Glucose	-2.491;-2.491
Glutamine	-3.428;-4
Oxygen	-2.546;-2.546
Arginine	-0.128;-0.397
Cystine	-0.2;-0.2
Histidine	-0.051;-0.051
Isoleucine	-0.136;-0.136
Leucine	-0.258;-0.258
Lysine	-0.252;-0.252
Methionine	-0.064;-0.064
Phenylalanine	-0.098;-0.098
Threonine	-0.163;-0.163
Tryptophan	-0.004;-0.004
Tyrosine	-0.055;-0.055
Valine	-0.187;-0.187
Folic acid	0

whose graphical representation is reported in Figures 3.18-3.21. Moreover, we compared the outcomes of these analyses with those obtained with the previous ENGRO1 model. We observed that once glucose is internalized in the cytosol, it is processed through the glycolysis that branches at the level of the 3-phosphoglycerate because of the flux splitting towards the production of biomass precursors serine and glycine, and the synthesis of lactate. In line with that presented in [31], glucose did not resulted the only contributor to the cytosolic pool of lactate, since its synthesis also derived from another important nutrient, namely glutamine. In particular, following the uptake of this carbon and nitrogen source, other than contributing to the cytosolic glutamate pool and the synthesis of nucleotides, glutamine mainly acts within the mitochondrion compartment. The optimal biomass synthesis rate is consistent with exploitation of reductive carboxylation for supporting the growth and accompanying the redirection of glutamine to lactate. More in detail, glutamine is firstly converted into glutamate that is then used as anaplerotic source because of its transformation into the TCA cycle intermediate α-ketoglutarate (α-KG) through the glutamate dehydrogenase enzyme. Proliferative wirings are characterized by the ability to process glutamine through a reverse flux through the reaction catalyzed by the aconitase enzyme, which is responsible for the conversion of citrate to isocitrate. A positive link between reductive carboxylation and flux redirection towards the synthesis of lactate is also confirmed by testing the sensitivity of ENGRO2 to a high and low glutamine uptake rate. According to [31], we observed that at high glutamine uptake the produced lactate over the consumed glucose ratio exceeds the standard value of 2, meaning that part of glutamine is converted to lactate. In this regard, as shown in Figure 3.22, under low glutamine uptake rate this lactate over glucose ratio decreases.

As previously mentioned, reductive carboxylation is also exploited for supporting the growth. In this regard, this particular wiring allows to contribute to the synthesis of citrate, which is then transported within the cytosol through the citrate/malate antiport by enabling palmitate and cholesterol production. As observed in ENGRO1 model, cancer metabolic rewiring is also characterized by a branched TCA cycle because the mitochondrial α-KG originating from the consumed glutamine, in addition to be processeed through reductive carboxylation, takes a clockwise path until the oxaloacetate (OAA). OAA then undergoes a transamination reaction that allows to convert this molecule into aspartate, which is subsequently transported outside the mitochondrion through the aspartate/glutamate antiport carrier. Within the cytosol, aspartate acts as biomass precursor and substrate for the synthesis of asparagine amino acid. Moreover, it can undergo another transamination reaction in the direction of

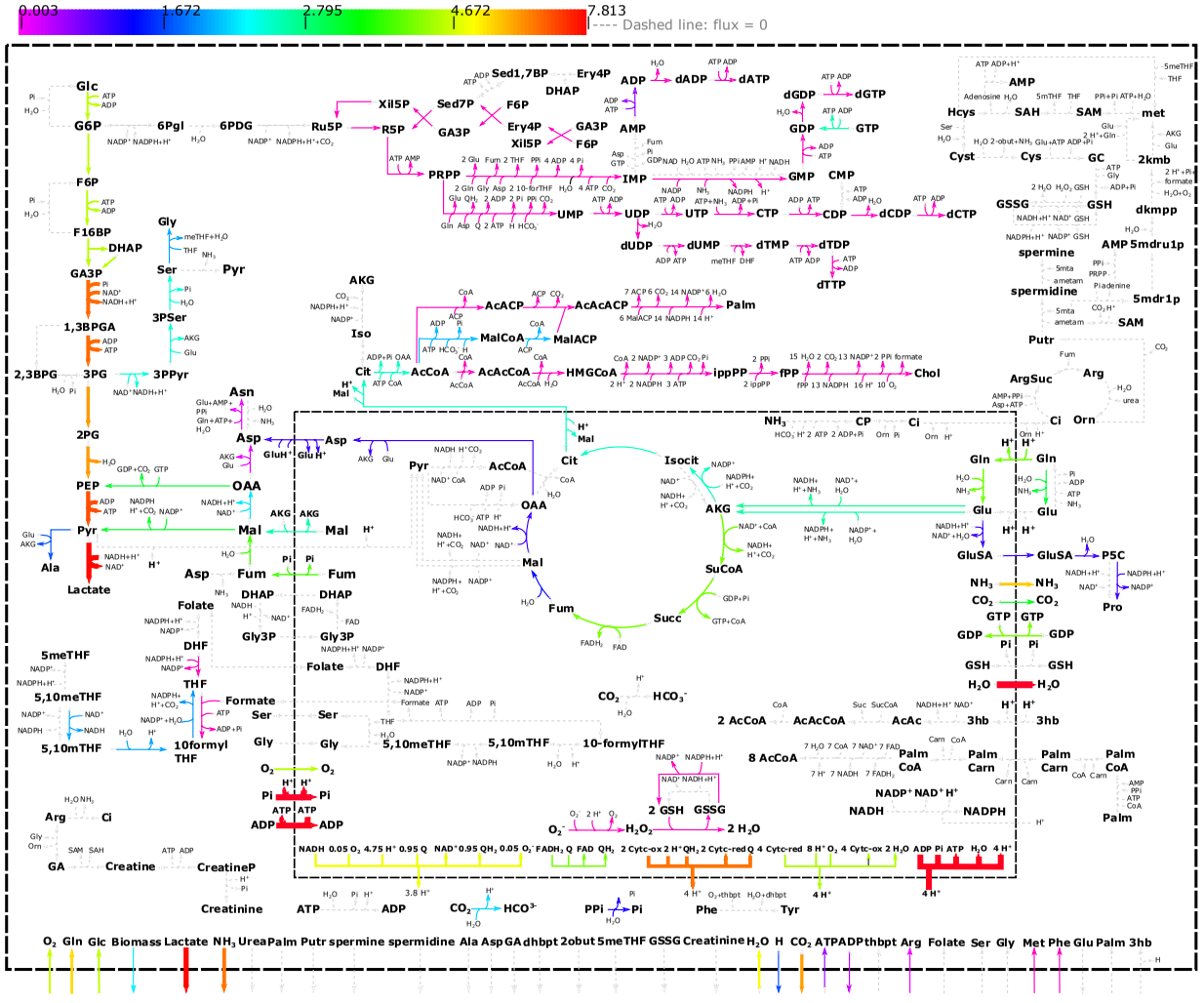


Figure 3.18: Outcome of the parsimonius Flux Balance analysis (pFBA) of ENGRO2 model. Results are shown for the first part of the model. Color and width of each arrow is proportional to the corresponding flux value from pFBA according to the chromatic scale on the top of the figure. Dashed and gray arrows refers to reactions whose flux value is null. A list of the abbreviations used in the map is provided in the Appendix A.

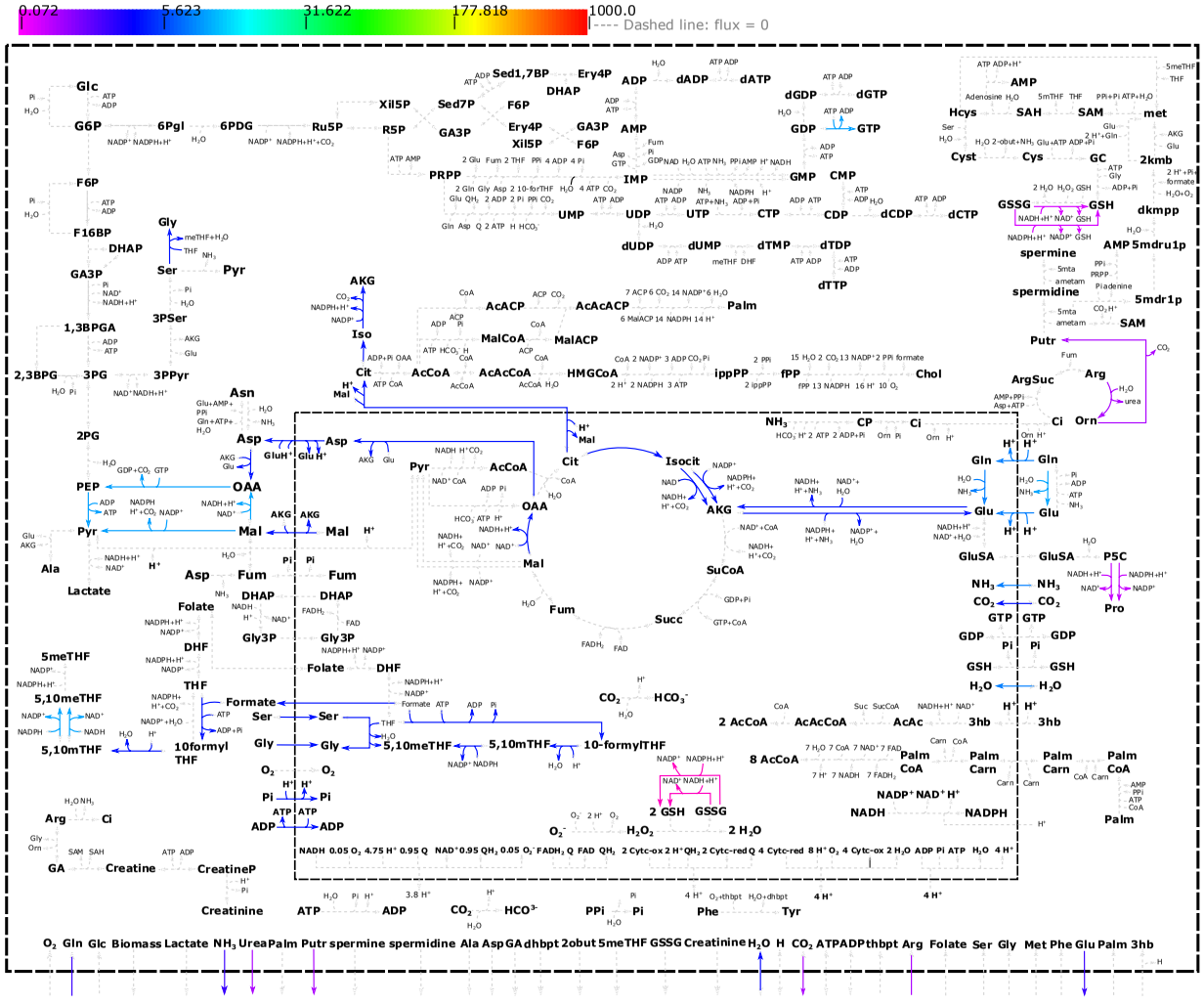


Figure 3.19: Outcome of the flux variability analysis (FVA) of ENGRO2 model. Results are shown for the first part of the model. Color and width of each arrow is proportional to the corresponding flux range from FVA according to the chromatic scale on the top of the figure. Dashed and gray arrows refers to reactions whose flux range is null. A list of the abbreviations used in the map is provided in the Appendix A.

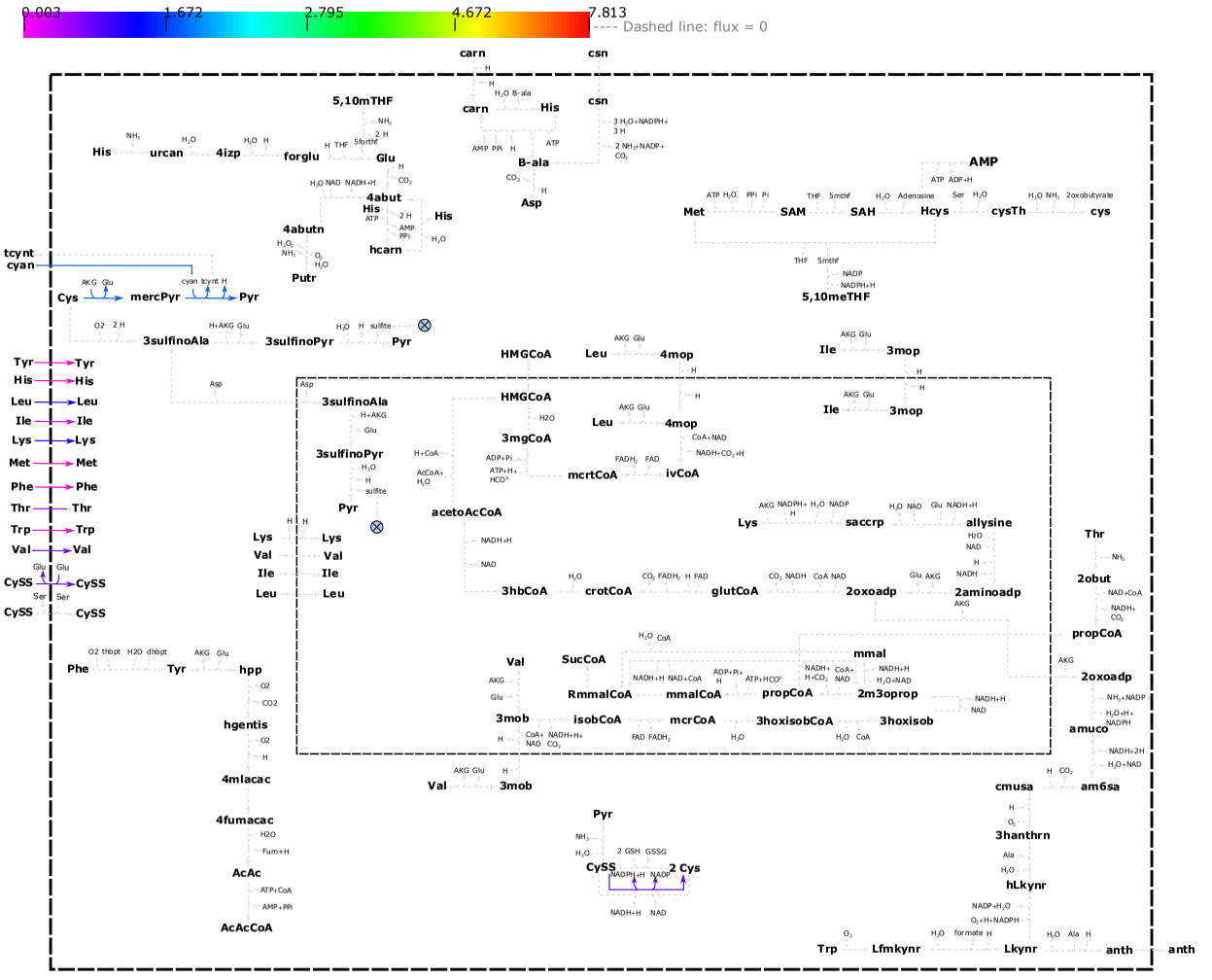


Figure 3.20: Outcome of the parsimonius Flux Balance analysis (pFBA) of ENGRO2 model. Results are shown for the second part of the model. Color and width of each arrow is proportional to the corresponding flux value from pFBA according to the chromatic scale on the top of the figure. Dashed and gray arrows refers to reactions whose flux value is null. A list of the abbreviations used in the map is provided in the Appendix A.

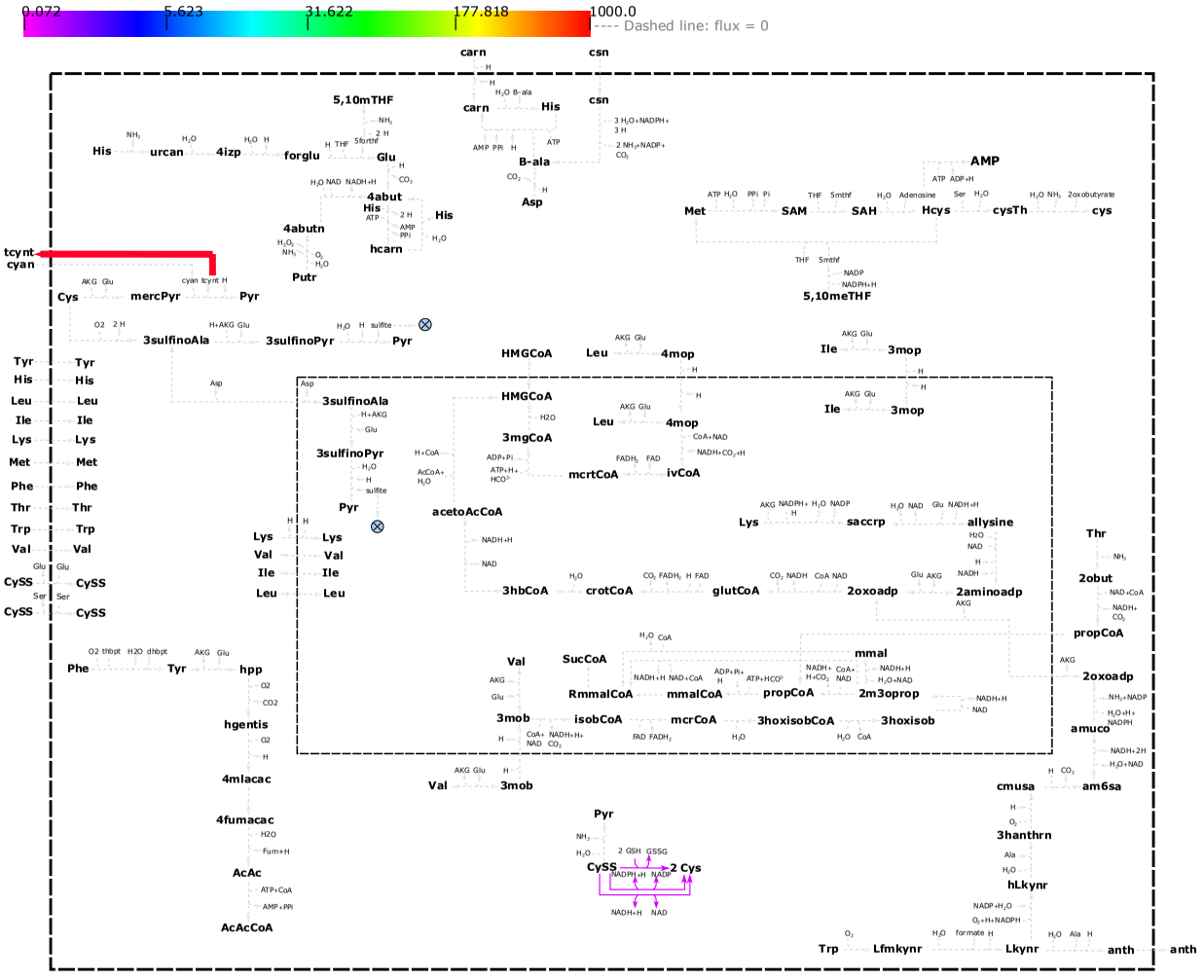


Figure 3.21: Outcome of the flux variability analysis (FVA) of ENGRO2 model. Results are shown for the second part of the model. Color and width of each arrow is proportional to the corresponding flux range from FVA according to the chromatic scale on the top of the figure. Dashed and gray arrows refers to reactions whose flux range is null. A list of the abbreviations used in the map is provided in the Appendix A.

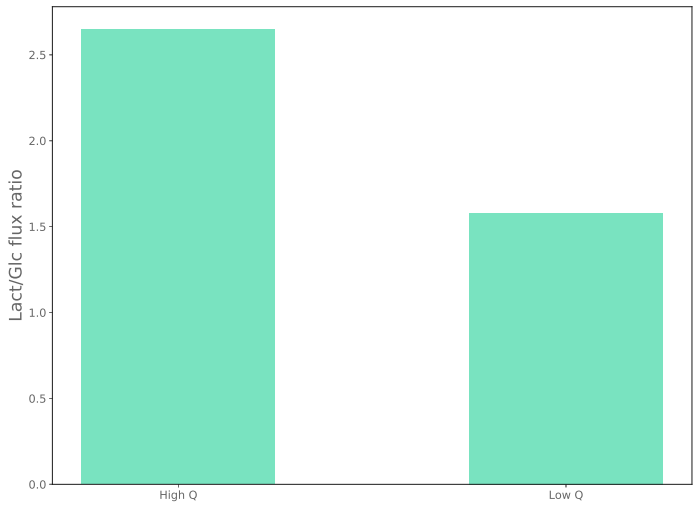


Figure 3.22: Ratio of lactate production flux value over glucose consumption flux value at high (High Q) and low (Low Q) glutamine uptake flux value.

OAA production that may be substrate for the synthesis of phosphoenolpyruvate (PEP). Following this route, glutamine indirectly contribute to the generation of cytosolic lactate. The mitochondrial NADH and $FADH_2$ that are produced during the TCA cycle are re-oxidized through the OXPHOS with the concomitant production of ROS. ROS are modelled in the network as superoxide anion and are removed by the coordinated action of glutathione peroxidase and glutathione reductase. Finally, according to ENGRO1 simulations outcomes, the ammonia that is not used is preferentially excreted in the extracellular environment rather than be removed by using the ATP demanding urea cycle pathway.

We then tested the sensitivity of the model towards the deprivation of the two most crucial nutrients, namely glucose and glutamine. As shown in Figures 3.23 and 3.24, the model resulted totally dependent on the glucose source since under its full deprivation the resulting biomass synthesis flux value was null. However, the total deprivation of glutamine just cause a 26% reduction of biomass synthesis rate. This behaviour resulted as opposed than that showed by ENGRO1 model. A possible explanation is the presence in the *in silico* medium of all the essential amino acids, which, as it will be better shown in Section 3.2.3, are able to differentially compensate for glutamine deprivation.

Reaction deletion analysis

We then performed a reaction deletion analysis to investigate which *in silico* deletions cause an effect on the biomass synthesis flux value. As shown in Figure 3.25, by excluding reactions causing a null biomass synthesis flux value because of their direct involvement in the growth, we observed that the most lethal deletion regarded the reaction catalyzed by the alanine transaminase enzyme. Its *in silico* causes 96% reduction of biomass synthesis flux value due to the role of alanine as biomass precursor. Although its direct involvement into biomass, the *in silico* knock out of this reaction did not cause a 100% biomass rate reduction. Thanks to the catabolism of the essential amino acid tryptophan an alternative and less efficient way to fuel the cytosolic pool of alanine is provided.

Although the hypoxia condition within the model, the mitochondrial oxidative phosphorylation resulted an important growth contributor. In this regard, *in silico* deletions of ATP synthase water transport between cytosol and mitochondrion by means of the aquaporine AQP8 halve the biomass synthesis flux value. Moreover, simulation of Complex I and II of OXPHOS pathway resulted in about 20% reduction of biomass synthesis rate.

Another important group of reactions whose *in silico* deletion negatively

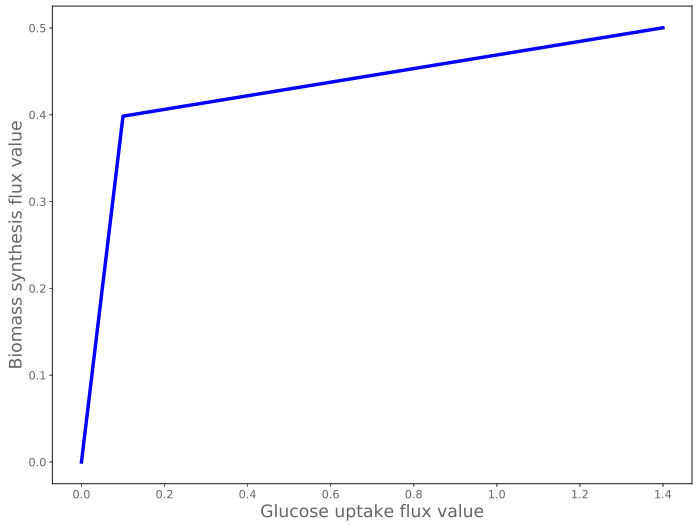


Figure 3.23: Biomass synthesis reaction flux variation following the progressive increasing of glucose uptake flux value. Flux values are expressed as mmol gDW^{-1} h^{-1}.

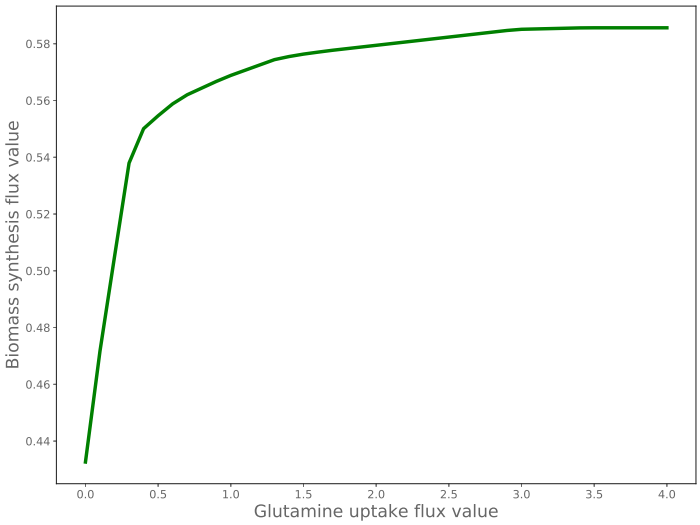


Figure 3.24: Biomass synthesis reaction flux variation following the progressive increasing of glutamine uptake flux value. Flux values are expressed as mmol gDW^{-1} h^{-1}.

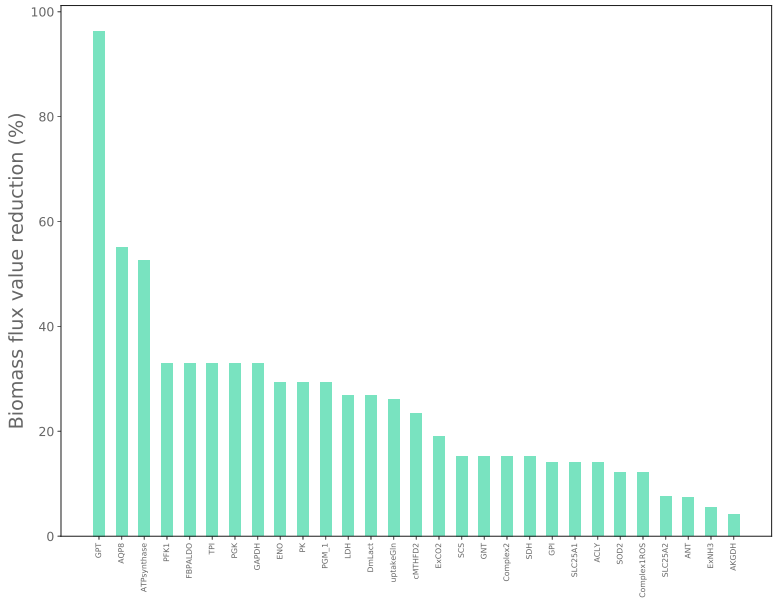


Figure 3.25: Single reaction deletion analysis of ENGRO2 model. Height of each column is set according to the reduction percentage of the biomass synthesis reaction following the *in silico* knock out of the corresponding reaction. Only reactions whose *in silico* knock out produced a positive variation and different from 100% of biomass synthesis reaction are shown.

affect biomass synthesis rate, falls within the glycolytic pathway. When these reactions are singularly deleted, the non oxidative branch of pentose phosphate pathway can be used. However, this alternative rewiring resulted a less efficient way to achieve the optimal biomass flux value.

Lactate production represents another impacting biomass reactions, since its deletion cause a reduction of the 25% of the biomass synthesis flux value. Similarly, the *in silico* deletion of TCA cycle reactions negatively affect the biomass synthesiss rate.

Contribution of essential amino acids to cancer cells growth under glutamine deprivation

In the previous simulations, all the essential amino acids are included in the medium. To understand their role in the network, we then evaluated the response of the model to their addition under total glutamine deprivation.

As reported in Table 3.6, the addition of isoleucine, leucine and valine amino acids is able to cause under glutamine deprivation about 50% compensation of the original biomass synthesis flux value. Graphical mapping of flux distributions relative to the simultaneous addition of the three amino acids under glutamine deprivation is reported in Appendix B, Figures B.1-B.4. By investigate their individual and in pairs role, we observed that the only supplementation of valine in the extracellular medium is able to compensate for glutamine deprivation, since the inclusion of isoleucine and leucine in the medium does not remedy its missing. An in depth analysis of the intracellular flux values pointed out that, following the consumption of valine, the reactions belonging to its metabolism allow the production of the mitochondrial TCA cycle intermediate succinyl-CoA together with the cofactors NADH and $FADH_2$. This outcome causes the flux rewiring towards a more oxidative metabolism that is justified by the increase of OXPHOS flux values. The reactions belonging to this pathway are needed to reoxidize the mitochondrial NADH and $FADH_2$ with the consequent electrons transfer to oxygen in parallel to protons pumping across the inner mitochondrial membrane. The generated proton gradient is utilized by the ATP synthase to produce new molecules of ATP that can contribute to the biomass synthesis. This flux rewiring towards a more oxidative metabolism also affects the cytosolic compartment. Indeed, pyruvate is not totally channelled into lactate production, since it is transported within the mitochondrion where it is converted partly to acetyl-CoA through the pyruvate dehydrogenase, and partly to OAA through the pyruvate carboxylase. These two end products are both substrates of the citrate synthase-catalyzed reaction, which is responsible for the mitochondrial citrate production. Citrate synthesis is fundamental due to the role of this metabolite in the cytosol as precursor of two biomass precursors, namely palmitate and cholesterol. We observed that the mitochondrial OAA is not totally redirected towards the citrate synthesis. On the contrary, it is also rewired in the cytosol through the glutamate/aspartate shuttle where it contributes to fuel cytosolic pyruvate and lactate pools. The succinyl-CoA coming from valine consumption and catabolism implies that the classic reaction of TCA cycle involved in the production of this metabolite catalyzed by the α-KG dehydrogenase, is no longer necessary. The succinyl-CoA follows the

Table 3.6: Individual and combined addition in ENGRO2 medium of isoleucine, leucine and valine amino acids under total glutamine deprivation. "Biomass flux" column indicates the biomass synthesis flux value when parsimonius flux balance analysis (pFBA) is performed after the corresponding perturbation reported in "EAA" column compared to the original biomass synthesis reaction flux value. Flux values in this column are expressed as mmol gDW^{-1} h^{-1}. "Biomass compensation" column refers to the same information reported in the middle column but expressed as percentage reduction.

EAA	**Biomass flux** $mmolgDW^{-1}h^{-1}$	**Biomass compensation** %
Ile	0	0
Leu	0	0
Val	0	0
Leu + Val	0.206	49.7
Ile + Leu	0	0
Ile + Val	0.206	49.7
Ile + Leu + Val	0.206	49.7

canonical direction of the TCA cycle until the fumarate synthesis. Fumarate pool is split between its role as mitochondrial pyruvate contributor through the malic enzyme, and its transport into the cytosol. In this case, fumarate can contribute to lactate production or, in alternative, it can be transported again into the mitochondrion through the malate/α-KG shuttle.

Another set of investigated essential amino acids consists of histidine and lysine. As reported in Table 3.7, their supplementation in the extracellular environment under glutamine deprivation causes 93.5% recovery of the original biomass synthesis flux value. However, this outcome just derives from the effect of the histidine. Indeed, we observed that the individual supplementation of lysine under glutamine deprivation causes a null biomass synthesis flux value, since it only contributes as biomass precursor without remedy the glutamine missing. On the contrary, histidine is able to compensate for that important carbon and nitrogen source. In particular, histidine catabolism fuels cytosolic glutamate pool, which is then processed by the glutamine synthetase (GS)-catalyzed reaction for the production of glutamine. Moreover, cytosolic glutamine can be transported within the mitochondrion through the aspartate/glutamate shuttle. From the analysis of flux distribution, we observed that the GS-catalyzed reaction has a positive flux without variability. This type of flux range indicates

Table 3.7: Individual and combined addition in ENGRO2 medium of histidine and lysine amino acids under total glutamine deprivation. "Biomass flux" column indicates the biomass synthesis flux value when parsimonius flux balance analysis (pFBA) is performed after the corresponding perturbation reported in "EAA" column compared to the original biomass synthesis reaction flux value. Flux values in this column are expressed as mmol gDW^{-1} h^{-1}. "Biomass compensation" column refers to the same information reported in the middle column but expressed as percentage reduction.

EAA	**Biomass flux** $mmolgDW^{-1}h^{-1}$	**Biomass compensation** %
His	0.39	93.5
Lys	0	0
His + Lys	0.39	93.5

the mandatory usage of this reaction in order to replenish the cytosolic pool of glutamine, which is fundamental for the synthesis of multiple biomass precursors. However, we also observed that glutamate and glutamine transport reactions between cytosol and mitochondrion are both inactive in terms of flux values. The aspartate/glutamate shuttle results a more beneficial route compared to the transport of glutamate coupled with protons. At mitochondrial level, a branched TCA cycle with a slight increase of the OXPHOS reactions fluxes occur. In particular, the branched TCA cycle implies that fatty acids and cholesterol just derives from histidine metabolism and not from glucose. Once consumed, glucose is processed through the glycolysis until the production of lactate. Graphical mapping of flux distributions relative to the simultaneous addition of the two amino acids under glutamine deprivation is reported in Appendix B, Figures B.5-B.8.

After the evaluation of histidine and lysine role within the model, we simulated the addition of methionine and cystine, as before, under te total depletion of glutamine. As reported in Table 3.8, this perturbation caused a complete recovery of the original biomass synthesis flux value after the glutamine deprivation. In particular, cystine is responsible for that response, since null biomass production is obtained when only methionine is provided. An in depth flux distribution analysis highlighted that methionine is consumed from the extracellular environment but it just contribute in the network as biomass precursor. On the contrary, the cystine, which represent the non reactive form of the cysteine, when is supplemented in the medium, is internalized in the cytosol and converted to cysteine. The reactions involved in its catabolism produce cytosolic

Table 3.8: Individual and combined addition in ENGRO2 medium of methionine and cystine amino acids under total glutamine deprivation. "Biomass flux" column indicates the biomass synthesis flux value when parsimonius flux balance analysis (pFBA) is performed after the corresponding perturbation reported in "EAA" column compared to the original biomass synthesis reaction flux value. Flux values in this column are expressed as mmol gDW^{-1} h^{-1}. "Biomass compensation" column refers to the same information reported in the middle column but expressed as percentage reduction.

EAA	**Biomass flux** $mmolgDW^{-1}h^{-1}$	**Biomass compensation** %
Met	0	0
CySS	0.756	100
Met + CySS	0.756	100

pyruvate and glutamate and consume cytosolic α-KG. The pyruvate production implied a rewiring towards a more oxidative metabolism characterized by a more active mitochondrial respiration. In this regard, we observed increased flux values of pyruvate transport within the mitochondrion and the subsequent pyruvate dehydrogenase and pyruvate carboxylase-catalyzed reactions. The end products of this two reactions, namely acetyl-CoA and OAA, are both substrate of citrate synthase that is involved in the production of citrate. The TCA cycle is not completely active because the produced citrate is in part directed towards the cytosol for the production of palmitate and cholesterol, and in part follow the canonical direction of the TCA cycle until the α-KG that is required as substrate for the cysteine metabolism. Consequently, the part of this cyclic pathway from the consumption of α-KG is inactive to avoid an inefficient consumption of that metabolite. The malate/α-KG and aspartate/glutamate shuttles deal with the exit of α-KG within the cytosol in order to continue the catabolism of cysteine. Graphical mapping of flux distributions relative to the simultaneous addition of the two amino acids under glutamine deprivation is reported in Appendix B, Figures B.9-B.12.

As reported in Table 3.9, by supplementing phenylalanine, threonine, tryptophan and tyrosine in the medium under glutamine deprivation, we obtained a fully recovery of the biomass synthesis flux value. However, different contribution is provided by each one of these individual essential amino acids. In particular, following glutamine deprivation, phenylalanine and threonine alone are not able to produce a positive biomass synthesis flux value. On the contrary, tyrosine and tryptophan are, respectively, able to ripristinate 52% and 65% of

the original biomass synthesis flux value. However, the effect changed when we evaluated the role of these amino acids in pairs. In particular, although phenylalanine is not able at all to restore the biomass synthesis flux, when it is provided in pairs with tyrosine, it considerably changes its role. Indeed, in this case, phenylalanine amplifies the effect produced by the addition of tyrosine, causing a fully recovery of the biomass synthesis rate. However, if phenylalanine is provided with tryptophan, it does not cause any advantage to the effect provided by the tryptophan alone. In addition, the concomitant supplementation of tyrosine and tryptophan is able to cause a fully recovery of biomass synthesis rate.

When the four amino acids are provided together (graphical mapping of flux distributions relative to the simultaneous addition of the four amino acids under glutamine deprivation is reported in Appendix B, Figures B.13-B.16), everyone is consumed. In particular, phenylalanine, tyrosine and threonine just work as biomass precursors. On the contrary, tryptophan catabolism causes a net production of cytosolic alanine, which through the alanine transaminase is converted back to glutamate and pyruvate. Overall, a more oxidative metabolism occurs where the cytosolic pool of pyruvate instead of being directed towards the lactate is channelled into the mitochondrion towards a more cyclic TCA cycle and a higher respiration rate. The same situation recurs when only tryptophan is supplemented in the medium because its catabolism contribute to fuel alanine pool, and a more oxidative metabolism is followed. By analysing the flux distribution resulted from the supplementation of tyrosine, we observed that its catabolism involved a net consumption of cytosolic α-KG, oxygen and ATP and the production of cytosolic glutamate, fumarate and acetoacetyl-CoA. Because of the production of this latter metabolite, we consider tyrosine an alternative nutrient that is responsible for the production of palmitate and cholesterol. Indeed, acetoacetyl-CoA is the precursor of these metabolites. Consequently, we observed that the cytosolic citrate that derives from the TCA cycle is directed to the production of α-KG in order to continue the catabolism of tyrosine instead of being directed towards the production of palmitate and cholesterol. Moreover, differently from the other situations, beta-oxidation pathway is active. Consequently, we observed higher rate of the reactions belonging to the OXPHOS pathway. These reactions are necessary to reoxidize the NADH and $FADH_2$ that are generated from the catabolism of the palmitoyl-CoA. Moreover, due to the mitochondrial acetyl-CoA produced from the beta-oxidation pathway, it follows that PDH-catalyzed reaction has a null flux. At the mitochondrial level, moreover, we observed that TCA cycle follows its canonical direction, even if alternative solutions are characterized by a branched TCA cycle. Furthermore,

the mitochodrial shuttles malate/citrate, aspartate/glutamate, malate/α-KG and fumarate/ H^2 are active to sustain the synthesis of α-KG and allowing the continue catabolism of tyrosine and the disposal of deriving intermediates. As reported in Table 3.9, the situation changes when phenylalanine is provided together with the tyrosine. In this case, we observed a 100% recovery of biomass synthesis flux value under the total glutamine deprivation. Both phenylalanine and tyrosine are consumed but just tyrosine is catabolize following the same traits previously described in the perturbation where only tyrosine is supplemented in the medium. However, the presence of phenylalanine allows a higher consumption rate of tyrosine and a consequent increase of all the related fluxes extent, by inducing in this way the original biomass recovery. A similar effect of synergy to that producd by phenylalanine and tyrosine is caused by the co-supplementation of tyrosine and tryptophan, whose consumption produced a full recovery of the original biomass synthesis rate. Indeed, we observed that while tyrosine just acts as a biomass precursor, tryptophan is consumed with a twice uptake rate compared to the situation where this amino acid is alone provided in the medium. Tryptophan is then catabolized by following the same metabolic traits described before.

Consumption of an external source of palmitate

Recently, experimental evidences relative to the strict dependence of cancer cells from the de novo synthesis of palmitate-derived lipids rather than on an external source of fatty acids came out [182]. Following this new knowledge, we evaluated the sensitivity of ENGRO 2 model to an external source of palmitate, which as previously said, represents the precursor for all the downstream synthetize fatty acids.

In the previous analysis, we observed that glutamine represents the main source used in the standard condition to produce palmitate and cholesterol through the reductive carboxylation pathway. However, the supplementation of palmitate in the extracellular environment makes glutamine a less essential source for the overall requirements of the network. However, under total glutamine deprivation, palmitate is not able to recovery the original biomass production rate. Indeed, we observed a positive consumption rate of palmitate that once in the cytosol just contribute as biomass precursor without being catabolize in the mitochondrion through the beta-oxidation pathway. At the same time, no growth advantage is conferred by free availability of fatty acids, since the uptake of exogenous palmitate slightly decreases the biomass synthesis rate. This result was in line with the experimental evidence reported

Table 3.9: Individual and combined addition in ENGRO2 medium of phenylalanine, threonine, tyrosine and tryptophan amino acids under total glutamine deprivation. "Biomass flux" column indicates the biomass synthesis flux value when parsimonius flux balance analysis (pFBA) is performed after the corresponding perturbation reported in "EAA" column compared to the original biomass synthesis reaction flux value. Flux values in this column are expressed as mmol gDW^{-1} h^{-1}. "Biomass compensation" column refers to the same information reported in the middle column but expressed as percentage reduction.

EAA	**Biomass flux** $mmolgDW^{-1}h^{-1}$	**Biomass compensation** %
Phe	0	0
Thr	0	0
Tyr	0.217	52.3
Trp	0.269	64.8
Phe + Thr	0	0
Phe + Tyr	0.431	100
Phe + Trp	0.269	64.8
Thr + Tyr	0.217	52.3
Thr + Trp	0.269	64.8
Tyr + Trp	0.508	100
Phe + Thr + Tyr + Trp	0.508	100

in [182]. Moreover, the inactivity of the beta-oxidation following an external supplementation of palmitate is in line with the evidence highlighted in [183]. In this work, the authors pointed out that an exogenous source of fatty acids can substitute for de novo synthesis in promoting cell proliferation. Moreover, exogenous fatty acids can also attenuate the cancer-specific toxic effect of lipogenesis inhibitors. Furthermore, a limited access to environmental lipids may render the cancer cells more sensitive to the inhibitors of lipogenesis. A putative reason behind the inactivity of the beta-oxidation pathway can be due to the hypoxia condition that characterize most cancer cells, which prevents the NADH and $FADH_2$ reoxidation that are produce by the beta-oxidation through the oxidative phosphorylation pathway. It is worth noticing that an increase oxygen uptake rate concurrently with an external supply of palmitate involves a global metabolic rewiring towards a more oxidative metabolism characterized by the usage of the beta-oxidation pathway that producing acetyl-CoA as main product supplies the mitochondrial pool of this metabolite that is needed to fuel in a clockwise manner the TCA cycle and in turn the cytosolic production of cholesterol. Following these outcomes, it is evident that, depending on the boundary conditions, different metabolic trais emerge for supporting the cancer cells growth. Consequently, metabolic heterogeneity of cancer cells need to be considered in order to identify different and effective drug targets and develop more effective treatments.

Validation of ENGRO2 model

We concluded ENGRO2 analysis by performing a validation step with recent literature works.

The first analysed work, which is presented in [184], is focused on the development of an inverse agonist for the nuclear receptor liver-X-receptor (LXR) that regulates the expression of some key genes in the glycolysis and lipogenesis. Given the propensity of tumour cells to rely on aerobic glycolysis and de novo lipogenesis for sustaining their proliferation, the inverse agonist could represent an effective cancer treatment approach.

We performed a series of simulations on ENGRO2 model by acting on the genes that are regulated by the LXR receptor, and, in particular, on the boundaries of the reactions that these genes catalyzed. The *in silico* knock out of a gene is possible by setting to zero both the lower and upper bound of the related reaction, preventing flux passing through it. In the glycolytic pathway, we performed the *in silico* knock out of the reactions catalyzed by the GCK1 enzyme which acts on the phosphorylation of glucose to glucose-6-phosphate.

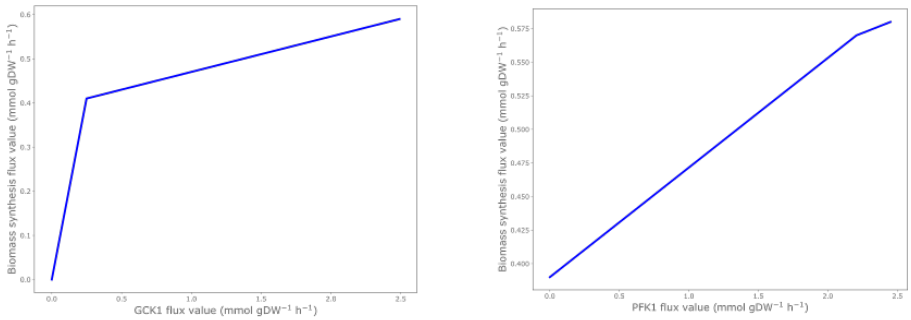


Figure 3.26: Biomass synthesis flux value following flux variation of GCK1- and PFK1-catalyzed reactions between 0% and 100% of the corresponding original flux value under standard condition. Flux values are expressed as mmol gDW^{-1} h^{-1}.

Moreover, we simulated the deletion of PFK1 that catalyze the conversion of fructose 6-phosphate to fructose 1,6-bisphosphate. Finally, we deleted LDH that is involved in the conversion of pyruvate to lactate with the parallel oxidation of NADH to NAD^+. In this first set of individual simulations, we obtained a null biomass synthesis flux value only blocking the flux through the GCK1-catalyzed reaction. Indeed, the other two perturbations produce a gradual reduction of the biomass synthesis rate until, respectively, a 33.5% and 26.6% reduction of the biomass production rate when these reactions are completely knock-out. This outcome is justified by the presence in the network of alternative routes that the model can follow to optimize the defined objective function. In line with the evidence emerged in [184], we observed that the simultaneous knock out of all the three reactions induce a null flux value of the biomass synthesis reaction that translates in the total growth arrest.

Regarding the de novo lipognesis pathway, we considered two genes involved in this route. The first one is the acetyl-CoA carboxylase (ACC), which catalyzes the first step in the fatty acids synthesis through the carboxylation of acetyl-CoA to malonyl-CoA. The second gene is the fatty acid synthase (FASN1) which is involved in the synthesis of palmitate from acetyl-CoA and malonyl-CoA. Being these two reactions the only way through which fatty acids, as precursors of biomass, can be synthesized in the model, it follows that their *in silico* both single and simultaneous knock out causes a gradual decreasing of the biomass synthesis rate until reaching the complete growth arrest.

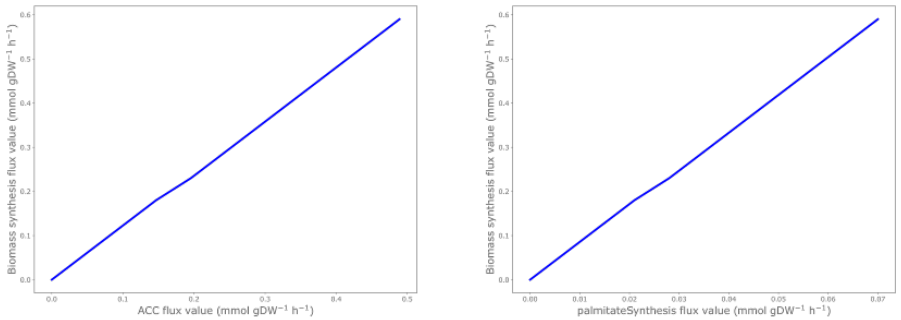


Figure 3.27: Biomass synthesis flux value following flux variation of ACC-catalyzed reaction and palmitate synthesis reaction between 0% and 100% of the corresponding original flux value under standard condition. Flux values are expressed as mmol gDW^{-1} h^{-1}.

Based on the experimental evidence emerged in [185], we investigated in ENGRO 2 model the role of the creatine as potential anticancer compound. The creatine is generally synthesized in mammalian cells in a two-step process. In the first one, the L-arginine:glycine amidinotransferase (AGAT) enzyme catalyzes the transamidation of guanidine group from arginine to glycine yielding guanidinoacetic acid and ornithine. In the second step, then, guanidinoacetic acid is methylated by N-guanidinoacetate methyltransferase (GAMT) to produce creatine, through the methyl group donated by S-adenosylmethionine. Creatine is finally transported into the blood circulation and reaches different creatine-requiring target tissues, such as muscle, brain and heart. In [185], the attention has been focused on the creatine kinase (CK) enzyme, which is responsible for the reversible phosphorylation of creatine by ATP. In this regard, it has been showed that the creatine content and the activity of CK progressively decreased with the progression of malignancy, by reaching almost undetectable levels at the final stage of tumour development.

In view of these evidence, we simulated the sensitivity of ENGRO 2 to an increasing flux value passing through the CK-catalyzed reaction. As shown in Figure 3.28, by forcing flux going through this reaction, we immediately noticed a gradual reduction of the biomass synthesis flux value. Along the lines suggested in [185], phosphorylation of creatine causes a progressively increasing consumption of ATP that is necessary for the activity of CK, by sequestering significant amount of ATP that is necessary for tumour cells growth. From an in depth

analysis of flux distribution, we observed a considerable increase of flux values of AGAT and GAMT-catalyzed reactions. Consequently, since an increase requirement of arginine and glycine is needed in order to synthetize creatine, flux of reactions belonging to the urea cycle enhanced for producing arginine. Moreover, glycolysis mantains its original flux until the branch towards the synthesis of serine and glycine, whose flux significantly increase to fuel the cytosolic pool of glycine that is also needed for the creatine yield. The guanidinoacetate that is require by the GAMT-catalyzed reaction causes the activation of the reactions belonging to metabolism catabolism. These reactions are required to synthesize S-adenosylmethionine and to consume the S-adenosylhomocysteine in order to guarantee the production of creatine. However, this task involves the consumption of 2 molecules of ATP, by contributing in this way to decrease the biomass synthesis flux value. Therefore, the high amount of ATP requested by the creatine synthesis may be the cause of the anticancer effect hypothesized for the creatine.

In [186], the role of Argininosuccinate synthase (ASS1) enzyme in cancer cells is investigated. This enzyme catalyze in the urea cycle the conversion of nitrogen from ammonia and aspartate to urea. In particular, ASS1 downregulation has been proposed as mechanism for supporting cancer cells proliferation, by providing a way for connecting the urea cycle enzymes and pyrimidine synthesis. In this regard, loss of ASS1 activity has been reported in multiple cancers. The authors demonstrated that a decreased flux through the reaction catalyzed by ASS1 enhances the flux through CAD (carbamoyl-phosphate synthase 2, aspartate transcarbamylase, dihydroorotase complex) enzyme. In this way, pyrimidines synthesis is facilitated because a higher pool of aspartate is now available. Indeed, aspartate is substrate of both ASS1 and CAD-catalyzed reactions.

In line with these experimental evidence, we simulate a gradual increase of flux passing through the ASS1-catayzed reaction. Following this perturbation, we observed that biomass synthesis flux value decreases. This outcome is due to the sequester of aspartate in the reaction catalyzed by ASS1, and the parallel removal of this metabolite from the substrate pool of CAD-catalyzed reaction. Being this latter reaction directly involved in the synthesis of pyrimidines, it follows that less nucleotides are synthesized and less biomass is produced. To distinguish the contribution of ASS1 expression to the proliferation from other metabolic changes of cancer cells, we substituted the real reaction catalyzed by ASS1 enzyme with a fake one that is characterized by the consumption of aspartate. In this case, as shown in Figure 3.29, we observed less reduction of the

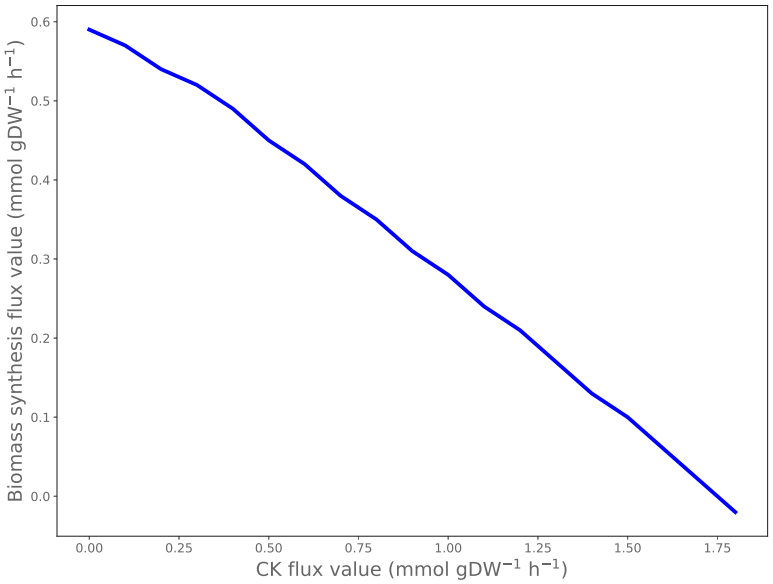


Figure 3.28: Sensitivity of biomass synthesis flux value to flux variation through the CK-catalyzed reaction. Flux values are expressed as mmol gDW^{-1} h^{-1}.

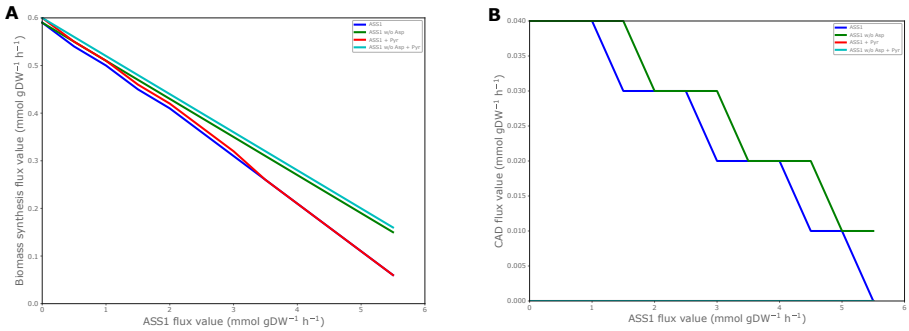


Figure 3.29: Sensitivity of biomass synthesis flux value to flux variation through the ASS1-catalyzed reaction by using the original reaction (ASS1, blue line), a fake version where contribution of aspartate as substrate of the reaction is removed (ASS1 w/o Asp, green line), the original reaction coupled with the supplementation of pyrimidines in the medium (ASS1 + Pyr, red line), and a fake version where contribution of aspartate as substrate of the reaction is removed coupled with the supplementation of pyrimidines in the medium (ASS1 w/o Asp + Pyr, light blue line). Flux values are expressed as mmol gDW^{-1} h^{-1}.

biomass synthesis rate and of the CAD-catalyzed reaction flux value compared to the previous scenario.

Moreover, in [186], the authors hypothesized that ASS1 overexpression negatively affects cell proliferation through the aspartate deviation from the production of pyrimidines. Therefore, they showed that supplementation of an extracellular source of pyrimidines in the medium significantly restores the proliferation of ASS1 overexpressing cells to a similar level as the original one. However, contrary to what the authors said, we noticed that supplementing the medium with pyrimidines the proliferation of ASS1 overespressing cells is just very slightly restored.

Chapter 4

New constraint-based methods for heterogeneous metabolic systems

Cancer is catalogued as a heterogeneous disease because multiple subtypes characterized by their own distinct histopathological and biological features exist [36, 187]. For this reason, the development of more personalized and targeted therapies to specifically treat patients and obtain better outcomes requires an in depth understanding of specific driving forces behind different cancer subtypes [188].

The intertumour heterogeneity discussed in Chapter 3 is just a small part of the whole heterogeneity affecting cancer disease. In this regard, multiple tumoural subtypes deriving from the same origin primary site can show relevant biological differences at either histologic or molecular level. This kind of heterogeneity, which is called interpatient tumour heterogeneity (Figure 4.1A), implies that patients affected by the same cancer type do not behave similarly at clinical level. The influence of multiple host factors, including tumour microenvironment, germline variants, and somatic mutations that are specifically present within the tumour of each individual patient, can influence the overall treatment response [189].

As shown in Figure 4.1, cancer heterogeneity is not confined to a diversity among patients. Indeed, different subclones having different genetic, epigenetic, and phenotypic features can characterize different regions of the same primary

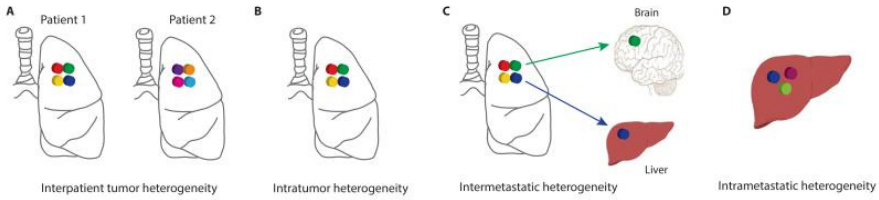


Figure 4.1: Multiple types of tumour heterogeneity. (A) Interpatient tumour heterogeneity: multiple tumoural subtypes can derive from the same origin primary site. (B) Intratumour heterogeneity: multiple heterogeneous subclones coexist within a given primary tumour. (C) Intermetastatic heterogeneity: Different subclones from the same primary tumour can generate multiple and heterogeneous metastatic lesions. (D) Intrametastatic heterogeneity: multiple and heterogeneous subclones can coexist within a single metastatic lesion. Image taken from [189].

tumour. This kind of heterogeneity is called intratumour heterogeneity (Figure 4.1B) cells [35].

Phenotypic heterogeneity at intratumour level is influenced by several selection pressures. Among them, tumour-stromal cells interactions, variation of extracellular hypoxia or acidity level, infiltrating stromal and immune cells, and remodelling of extracellular matrix can influence tumour evolution and its response to various treatments. Under the influence of these factors, subclones of primary tumours having an intrinsic resistance can support the tumour mass growth whereas sensitive clones to selection pressures may be negatively affected. Moreover, selection pressures may also generate new subclones having acquired somatic alterations that can promote cell survival and proliferation, in addition to metastatic tumour formation [189].

Intratumour heterogeneity constitutes the basis for secondary levels of heterogeneity. Different subclones of the same primary tumour can cause the emergence of multiple and heterogeneous metastatic lesions. This kind of heterogeneity, known as intermetastatic heterogeneity (Figure 4.1C), further extends to an intrametastatic heterogeneity (Figure 4.1D), because of metastases ability to acquire new mutations and evolve independently with each cell division [35, 189].

These multiple types of heterogeneity makes it necessary to characterize cancer cell populations and figure out how subpopulations may promote or impede cancer progression. In this way, more effective and personalized therapies can be formulated.

Tumour heterogeneity at genotypic and phenotypic level affects interactions between multiple tumoural subclones, and interactions with stromal cells and local microenvironment. The intricate dialogue between tumour cells and their environment selects for clones that are best phenotypically adapted to survive. Moreover, these interactions can also drive the behaviour of healthy cells as assistants of cancer progression [34]. Intratumour heterogeneity pushes the need for exploring more in detail the set of interactions that occurs within cancer ecosystem. The concept of ecosystem is based on a holistic view of cell populations, where cancer progression results from the interactions between cells, and between cells and their microenvironment. Therefore, a deep understanding of these interactions would allow to hamper tumour progression.

The works that I will present in this chapter are focused on the investigation of cell population metabolism, by shifting from an homogeneous to an heterogeneous vision of the system where cell-to-cell variations are explicitly addressed. Exploiting classic constraint-based modelling and the ecosystemic perspective of cell populations, I will discuss about novel strategies that we developed to explore metabolic plasticity of multiple clones within the same cell population. Particular emphasis will be put on the strategies adopted under varying scenarios, and the relationships established among various components.

4.1 From average behaviour to population model

Constraint-based modelling, although it currently represents the most applied computational technique to study cell metabolism, is a limited approach because it just allow to consider cell populations as homogenous systems. However, as previously discussed in Section 2.2 and in the introduction of this chapter, cell populations, including cancer cell populations, cannot be homogeneously considered in order to investigate metabolic plasticity of their components and their consequent heterogeneous phenotypes. In this regard, I present in Section 4.1.1 a new strategy to address the issue of intratumour heterogeneity in order to study metabolism of cell populations and investigate the relationships among their components. We assumed that the heterogeneity emerging from a given population can be due to a mismatch between the objectives of the individual members and that of the entire population. This novel methodology, which is thought to be complementary to the classic investigation of average cell or standard metabolomics, aims at identifying the best strategies adopted within a population by various subpopulations for promoting the overall growth. In addition, this approach aims at identifying how distinct subpopulations, each

one with its characteristic metabolic features, interact with each other within the same cancer population. As a proof of principle, we tested our methodology with a toy-model of central carbon metabolism that we reconstructed based on the current knowledge about metabolic pathways that are most involved in cancer metabolic rewiring. We studied the potentialities of a population model consisting of multiple interacting components having the same topology and stoichiometry, and sharing the same nutrients supply.

Thanks to this strategy, we highlighted that a single entity model, which is the basic element of classic constraint-based simulations, can be used to represent, at a first level, the entire cancer population by providing a snapshot of the average behaviour of the population itself. In this way, the identification of possible metabolic interactions between subpopulations is temporarily hidden. At a second level, the population model corresponds to a "network of metabolic networks", each one representing the individual subpopulations.

We formalized the proposed strategy in a new computational methodology called "popFBA" that is presented in Section 4.1.2. Overall, it represents an extension of classic flux balance analysis (FBA) approach to investigate metabolic phenotypes of given heterogeneous cell populations, with a particular focus on the relationships established among their components. By continuing to exploit cancer populations as case study, we assumed that a tumour mass consists of several types of cells, including stromal and cancer cells. Due to spatial proximity, the nutrients that are exchanged by the various tumour subpopulations with plasma are different from those exchanged with other subpopulations within the population itself. For these reasons, we included in the model separate exchange reactions with plasma and with tumour microenvironment. Given the previously presented ability of classic FBA on a single metabolic model to predict the net flux distribution of a cell population, we deduced that a single metabolic model represents a black box of the investigated population, and is unable to give information about the complex network of interactions occurring within it. Therefore, in this new methodology we exploited the single metabolic model as a building block to automatically reconstruct the population model, which consists of identical copies of the single metabolic network, all having identical stoichiometry and capacity constraints and sharing the plasma supply of nutrients. By means of linear programming optimization, we explored all the possible interactions among the multiple clones that are consistent with the achievement of the optimal tumour mass biomass. In this work, in particular, we exploited "coreHMR" model introduced in [78] (see Section 3.2.1) as building block to construct a population of 10 replicas of this core model, by optimizing the maximization of the biomass producted by the overall popula-

tion. Moreover, we assumed equal bounds for the reactions of all the clones, and we simulated a plasma supply of the three most important sources, namely glucose, glutamine and oxygen, in addition to an internal exchange of lactate, glutamine, glutamate and ammonia.

According to the previously emerged results, although the global purpose of tumour cells popolation is an enhanced proliferation rate, this whole objective does not constitute the common aim of all the individual components of this system. In this regard, popFBA approach highlighted within the population model the presence of different metabolic wirings adopted by individual subpopulations. Focusing more in detail on the interactions of the population model that are consistent with the achievement of the optimal tumour biomass, we observed that a cooperative behaviour among its subpopulations emerged involving an exchange of the four considered metabolites. In particular, we assessed the correlation between the four metabolites within the tumour microenvironment and the consequent biomass synthesis flux value of a given subpopulation with the aim of characterizing the metabolism of the most proliferative subpopulations. In all the investigated scenarios, a recurrent behaviour of the population persists where the most proliferative subpopulations consume and oxidise the lactate that is secreted in the tumour microenvironment by less proliferative subpopulations, by using it as energy source. This computational outcome has been well described in human tumours by the "reverse Warburg effect", which is based on the existence of a stromal-cancer lactate shuttle characterized by tumour stromal cells that undergo aerobic glycolysis producing lactate that is then used as energy source by the adjacent high proliferative cancer cells [24, 25, 190]. Therefore, in this case, tumour contain two subpopulations differing in their energy-generating pathways, but symbiotically working in order to fuel tumour growth. The first one is the "Warburg effect" subpopulation which consists of hypoxic and glucose-dependent cells that secrete lactate in the tumour microenvironment. In the second subpopulation based on the "reverse Warburg effect", cells import and utilize the lactate produced by their normoxic neighbours as main energy source through TCA cycle and OXPHOS pathways. Following this outcome, popFBA approach showed its ability to highlight "symbiotic" but also "parasitic" metabolic relationships between cancer and stromal cells in complex cancer populations.

Finally, in this work, we also observed how different coexisting subpopulations may follow different metabolic paths when alternative nutrients are exchanged with plasma, or spatial diffusion phenomena are considered by simulating a dishomogeneous distribution of oxygen provision across the clones according to a gradient.

4.1.1 Constraint-based modeling and simulation of cell populations

Di Filippo M†, Damiani C†, Damiani C, Colombo R, Pescini D, Mauri G

Communications in Computer and Information Science 2016; 708:126-137

DOI: 10.1007/978 – 3 – 319 – 57711 – 1_11

Abstract

The intratumor heterogeneity has been recognized to characterize cancer cells impairing the efficacy of cancer treatments. We here propose an extension of constraint-based modeling approach in order to simulate metabolism of cell populations with the aim to provide a more complete characterization of these systems, especially focusing on the relationships among their components. We tested our methodology by using a toy-model and taking into account the main metabolic pathways involved in cancer metabolic rewiring. This toy-model is used as "individual" to construct a "population model" characterized by multiple interacting individuals, all having the same topology and stoichiometry, and sharing the same nutrients supply. We observed that, in our population, cancer cells cooperate with each other to reach a common objective, but without necessarily having the same metabolic traits. We also noticed that the heterogeneity emerging from the population model is due to the mismatch between the objective of the individual members and the objective of the entire population.

Introduction

A reprogramming of cellular energy metabolism has recently been included within the hallmarks [19] of cancer. An overall rewiring of metabolism is indeed fundamental to most effectively support the uncontrolled and enhanced growth characterizing all tumor cells.

An attention on the single molecules that are responsible for cancer onset fails to handle the non-linearity and complexity of cancer metabolic rewiring [132]. For this reason, metabolomics aims at concurrently identifying and quantifying the full set of metabolites that are present within a given cell or tissue type at a given time, thus providing a snapshot of the cell phenotype [48, 191].

As a matter of fact, information and knowledge can be extracted from these large collections of data only by rationalizing and integrating them into computational predictive models. In this regard, constraint-based modeling has been by far the most applied technique to the study of metabolism. It indeed represents the best compromise between the purely qualitative information provided by graph-theory based topological models and the mechanistic details provided by kinetic modeling, which is currently impracticable for networks on a genome-scale. In particular, Flux Balance Analysis (FBA) – which exploits Linear Programming to identify the distribution of the metabolic flux that optimizes a metabolic objective – has extensively been applied to cancer research, as maximization of growth rate may accurately describe the objective driving cancer

evolution [139, 140, 141]. Classic FBA is limited to the simulation of a single (or average) cell that is representative of the metabolism of the entire population this cell belongs to. This is a major drawback if we consider that a cell population is not necessarily homogeneous and various metabolic phenotypes may be generated. In fact, the heterogeneity characterizing cancer disease is not limited to the one existing among individual tumor types, but multiple sources of intratumor heterogeneity leading phenotypic differences among cells belonging to the same population exist. Unfortunately, many anti-cancer treatments are not able to deal with intratumor heterogeneity, drastically reducing their efficacy [188, 192]. As a consequence, single-cell metabolomics techniques are currently under development as a promising strategy to unravel metabolic heterogeneity among cells belonging to the same tumor, which metabolomics hides as a result of average measurements of population behavior, by investigating singularly the role of distinct cell types within a given population. However, these kind of experiments are still at an early stage and numerous technical limitations remain to be solved [193].

To address the issue, we propose here an extension of constraint-based modeling to study metabolism of cell populations in order to provide a more complete characterization of these systems and to investigate relationships among their components. We assume that the heterogeneity emerging from a given cell population is due to the fact that the objectives of the individual members do not correspond to the objective of the entire population. As a proof of principle, we test our methodology with a toy-model of cancer metabolism that has been reconstructed based on the current knowledge on the metabolic pathways most involved in cancer metabolic rewiring.

Flux balance analysis and flux variability analysis

Flux Balance Analysis allows to calculate the optimal flux distribution, which is the rate at which each of the R reactions of a network occurs at steady state.

By relying on a steady state assumption, according to which concentration of each of the M metabolites belonging to the network remains constant over time, FBA does not require any knowledge on enzymatic kinetic or metabolite concentrations. The application of further constraints on the system is used to reduce the number of putative flux distributions defining an allowable solution space in which any flux distribution may be equally acquired by the network. The optimization (maximization or minimization) of a specific objective function (e.g. maximization of adenosine triphosphate (ATP) or biomass production, minimization of nutrients utilization) allows to narrow the set of feasible solutions and

to identify a single optimal flux distribution.

Given a $M \times R$ stoichiometric matrix S, whose element $s_{i,j}$ takes value $-\alpha_{ji}$ if the species S_i is a reactant of reaction j, $+\alpha_{ji}$ if species S_i is a product of reaction R_j and 0 otherwise - where $-\alpha_{ji}$ is the stoichiometric coefficient of reactant/product i in reaction j - the problem is postulated as a general Linear Programming formulation:

$$\begin{aligned} &\text{maximize or minimize } Z = \sum_{i=1}^{R} w_i v_i \\ &\text{subject to } S\vec{v} = \vec{0},\ \vec{v}_{min} \leq \vec{v} \leq \vec{v}_{max}. \end{aligned} \tag{4.1}$$

where w_i is the objective coefficient for flux v_i in vector $\vec{v}$; whereas the vectors $\vec{v}_{min}$ and $\vec{v}_{max}$ specify, respectively, the lower and upper boundaries of the admitted interval of each flux v_i. A negative value v_i conventionally indicates flux trough the backward reaction. To achieve mass balance in an open system, exchange of a given nutrient A with the environment is defined by unbalanced reactions in the form of $A <=> \emptyset$. For a more comprehensive description of FBA, the reader is referred to [60].

Frequently, although FBA only returns a single flux distribution, the constraints imposed on the system under investigation do not allow to obtain a unique solution, but may confine the solution space to a feasible set of alternative optimal flux distributions in which the same optimal flux value of the objective function may be reached through different but equally possible ways. In this context, Flux Variability Analysis (FVA) [64] returns the range of flux variability of each reaction, i.e. the allowable minimum and the maximum fluxes by each reaction, but it does not identify all the alternative optimal solutions. This task can be performed by exploiting recursive mixed integer linear programming (MILP) optimization, as proposed in [194].

A proposal for using the constraint-based approach to model cell populations

Metabolic networks reconstructed starting from genome annotation are today available for different organism, spanning from micro-organisms [195] to human metabolism. These networks may encompass virtually all reactions that can be catalyzed by the enzymes encoded by a given genome, or only a portion of them [196]. In order to fill the existing gap between the understanding of single cells function (represented by a single metabolic network) within a given tissue and

their role when they are interacting with each other within a population, we propose to replicate N copies of the reference metabolic network with univocal names for metabolites and reactions, so to obtain a $(M \cdot N) \times (R \cdot N)$ stoichiometric matrix. For the exchange of intracellular nutrients with the environment (the extracellular matrix) of each of the N networks, the unbalanced reactions $A_i <=> \emptyset$ are replaced by reactions in the form $A_i <=> A_{medium}$ where A_i refers to metabolite A in network i, whereas A_{medium} refers to metabolite A in the extracellular matrix, to mimic the fact that cells in the population share the same resources. To achieve mass balance of the population model as an open system, a set of E exchange unbalanced reactions for metabolites within the extracellular matrix must be included. Note that the set of metabolites that the cells exchange with the extracellular matrix and the set of metabolites that the cell population share with the external environment do not necessarily coincide. A schematic representation of the population model is provided in Fig. 4.6. Once the $(M \cdot N) \times (R \cdot N + E)$ stoichiometric matrix and the vector of objective coefficients are obtained for the population model, standard FBA can be applied to obtain the distribution of flux across the N cells that maximize the population objective. For this purpose, a Python algorithm was implemented to automatically replicate a number of times any constraint-based model and generate the above defined population model. The resulting model is then exported to the Systems Biology Markup Language (SBML) format in order to be made suitable for simulation by any software that allows to perform linear programming optimization (e.g., COBRApy package [121], COBRA Toolbox [160]).

Results

As a proof of concept of our methodology, we constructed a generic and non-compartmentalized toy-model, that we refer to as "single entity core model", based on the current knowledge on the metabolic pathways most involved in cancer metabolic rewiring. This model consists of 45 reactions and 40 metabolites and includes the following metabolic pathways: glycolysis, production and consumption of lactate, tricarboxylic acid cycle (TCA cycle), oxidative phosphorylation (OXPHOS), pentose phosphate pathway (PPP), palmitate synthesis and beta-oxidation of fatty acids. Uptake reactions for the nutrients glucose and oxygen have been added as constraint to the model for defining the cell medium composition, and the maximization of the ATP yield has been chosen as objective function of the system, as we are focused on the reprogramming of energy metabolism of cancer cells.

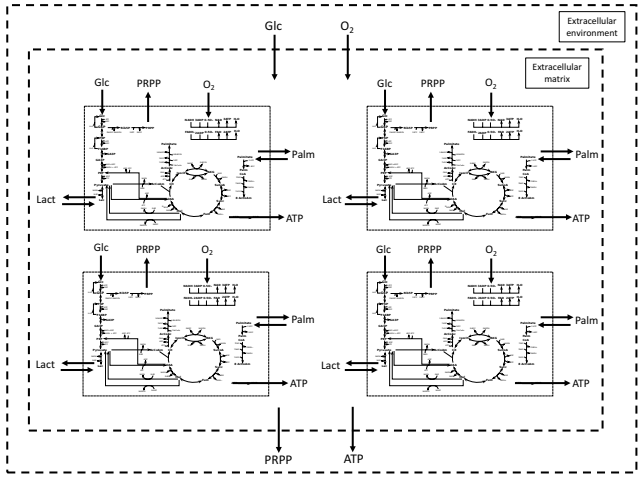


Figure 4.2: Schematic representation of the population model. A single entity model is used as building block and replicated N times to obtain a network of metabolic networks. All the members belonging to the population model have the same topology and stoichiometry, share the same nutrients (in our case, glucose and oxygen) supply, and have the same reactions to exchange some metabolites with other components of the population (within a compartment referred to as "Extracellular matrix"), or with the external environment (referred to as "Extracellular environment"). Abbreviations: Glc, glucose; Lact, lactate; PRPP, phosphoribosyl pyrophosphate; Palm, palmitate; ATP, adenosine triphoshate, O_2, oxygen.

We used this toy-model as building block for constructing the "population model" characterized by the interaction among individual components, all having the same stoichiometry and sharing the same glucose and oxygen supply. As for the single entity model, we chose the maximization of the overall ATP production as objective function of the whole system. We therefore investigated the potentialities of the constraint-based approach in the simulation of both the single entity and the population models in order to understand if this approach is able to highlight some differences between the two models in terms of their resulting flux distributions. The two models under investigation have the same objective function, equal exchange (sink and demand) reactions and the same boundaries on nutrients uptake.

We applied FBA to obtain the ATP production yield – computed as the ratio between the objective function flux value and the number of entities included in the model – as a function of the simulated number of entities, including the classic case of one single entity. We observed that the objective function flux value of the population model increases proportionally with the increase in the number of its components (Fig. 4.7). The computed yield is, therefore, constant and is not affected by the number of entities. This outcome confirms that FBA on individual metabolic networks well approximates the average cell of an optimal population. In fact, the net flux distribution of the different cells perfectly mirrors the flux distribution obtained as a solution of the single FBA model (Fig. 4.9, panel A). However, the population model allows to investigate the tumor population at a different level, elucidating the ways in which the average behavior can be achieved, how the individual cells may differ in their metabolism, and how different subpopulations of cells may interact with each other to attain the common goal.

Metabolic heterogeneity within population models

We shifted the focus toward a more in-depth study of how the flux distribution identified in the single entity model distributes among multiple cells within the population model. We wanted to understand whether FBA approach could highlight the heterogeneity factor that we know to be a long-established knowledge of cancer populations, or, in alternative, if the different components belonging to the system just share out the common good. In other words, we tested if a cooperative behavior could arise within cancer population or if all tumor cells behave the same way for achieving the common goal, which is an enhanced and uncontrolled growth and proliferation. In this regard, we used the toy-model to generate a population model consisting of 10 components, which are assumed

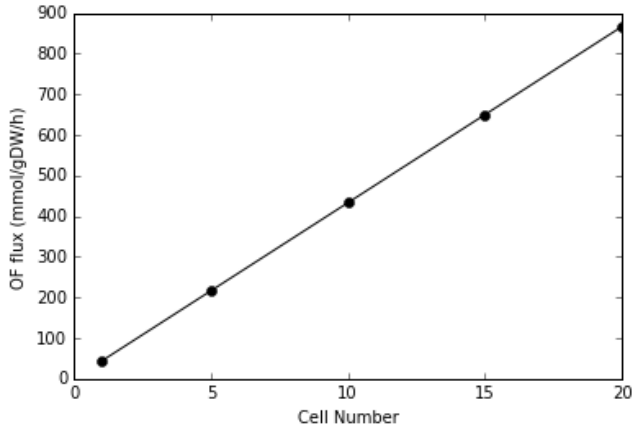


Figure 4.3: Variation of objective function flux value of the population model according to the increasing in the number of its components. The plot shows how the objective function flux value of the population model ("OF flux") increases proportionally with the increase in the number of its members ("Cell Number").

to be single cells that are representative of the metabolism of distinct subpopulations this cells belong to, all having the same topology and stoichiometry, and sharing the same glucose and oxygen supply. We performed FBA simulations on this system maximizing its overall ATP production and then we exploited FVA analysis in order to explore the variability range of the system across the alternative ways for obtaining the same objective function flux value.

Given the same maximum amount of glucose and oxygen to the system, the reached steady state is characterized by a particular ratio between glucose and oxygen uptake flux value of 1:6, which is known to be the correct ratio so that one glucose molecule is completely oxidize by oxygen. We observed that glucose uptake flux value is adjusted based on the quantity of oxygen that is available in the medium, and that all the entities constituting the population under investigation seek to maximize the common good for satisfying the common aim. This aspect, showed by the analysis of the flux distribution of the population model, emerged together with the observation that maximization of the ATP production by the population model is obtained following the interaction between two distinct subpopulations which show a very different ATP production rate and differ in their energy generating pathways (Fig. 4.9). The first subpopulation,

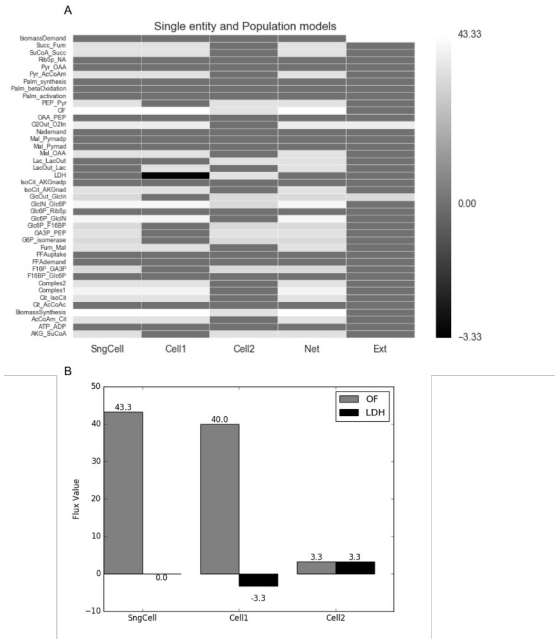


Figure 4.4: Results obtained from the execution of Flux Balance Analysis on the single entity model and on the population model, by giving in both cases the same maximum amount of nutrients (glucose and oxygen). Panel A, Heatmap showing the flux values for the reactions of both the single entity and the population models. The column SngCell contains the flux values of the internal reactions belonging to the single entity model, the column Cell1 contains the flux values of the internal reactions of the first identified subpopulation of the population model, the column Cell2 contains the flux values of the internal reactions of the second identified subpopulation of the population model, the column Net contains the net flux values of the internal reactions of the two different subpopulations, whereas the column Ext contains the flux values for the exchange reactions of both the models. The color of each cell is proportional to the flux value of the corresponding reaction according to the gray chromatic scale on the right-hand side of each heatmap. Panel B, Bar plot showing the flux values for the objective function (referred to as "OF") and lactate production (referred to as "LDH") reactions in both the single entity and population models. Columns "Cell1" and "Cell2" refer to the two subpopulations of the population model.

which corresponds to the hypoxic cancer cells, is composed by glucose-dependent cells that convert glucose into lactate that is then secreted in the medium; the second subpopulation, which corresponds to the better-oxygenated cancer cells, imports the lactate produced by the first subpopulation by using it as energy source instead of glucose, and is characterized by an active TCA cycle and respiratory chain. The flux distribution analysis showed that these two subpopulations do not contribute in an independent manner to the achievement of the common goal, but they cooperate with each other deriving mutual benefit from this interaction.

With changing environmental conditions as in Fig. 4.8, by perturbing the glucose to oxygen ratio and forcing the system towards more tumoral environmental conditions (i.e. constraining glucose uptake reaction flux to a higher level than that we found previously), the system reached different steady states having in common the fact that an increasing glucose uptake corresponds to a lowering of the objective function value. This happens because both there is not enough oxygen to completely oxidize glucose, and the individual can produce lactate whereas the entire population is not able. In addition to this result, we constantly noticed that, among the interacting subpopulations within the system, those that are responsible for the secretion of lactate in the medium, also produce ATP at a lower rate compared to the subpopulations in which lactate is consumed (Fig. 4.8, panel E). The analysis of flux variability, through FVA, showed that there is not just one single possible way by which different components belonging to the population model can interact with each other. On the contrary, for the purpose of maximizing the chosen objective function, three different scenarios (Fig. 4.8, panels B, C and D), which represent alternative optimal solutions, emerged. This outcome strengthens the importance of the heterogeneity factor within cancer populations as a strategy developed for evolutionary reasons in order to resist to anti-tumor treatments.

Conclusions

To investigate heterogeneity within cellular populations and as a complementary approach to either single cell or standard metabolomics, we investigated the potentialities of a population model that is characterized by multiple interacting components, all having the same topology and stoichiometry, and sharing the same nutrients supply. These two elements were necessary for developing a methodology that allows to identify within a population model which are the best strategies able to promote cancer population growth and how many distinct subpopulations, characterized by different types of metabolism, interact

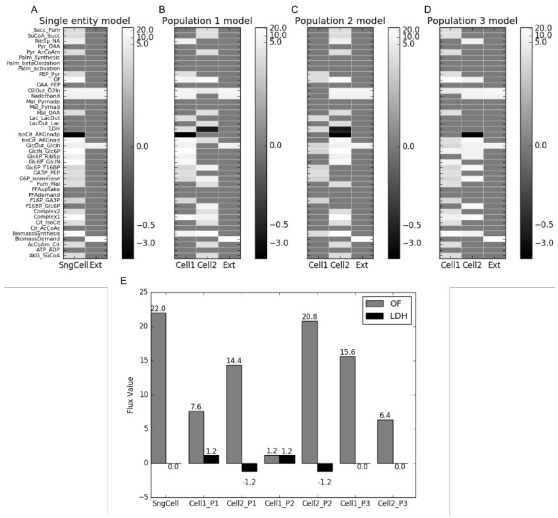


Figure 4.5: Results obtained from the execution of Flux Balance Analysis on the single entity model and on the population model after a perturbation of the glucose/oxygen ratio. In all the heatmaps the color of each cell is proportional to the flux value of the corresponding reaction according to the gray chromatic scale on the right-hand side of each heatmap. Panel A, Heatmap showing the flux values for the reactions of the single entity model. The column SngCell contains the flux values of the internal reactions, whereas the column Ext contains the flux values of the exchange reactions. Panels B-C-D, Heatmaps showing the three alternative and equally optimal flux distributions identified in the population model. The column Cell1 contains the flux values of the internal reactions of the first identified subpopulation, the column Cell2 contains the flux values of the internal reactions of the second identified subpopulation, whereas the column Ext contains the flux values for the exchange reactions of the model. Panel E, Bar plot showing the flux values for the objective function (referred to as "OF") and lactate production (referred to as "LDH") reactions in both the single entity and population models. The graph shows that following the maximization of ATP production, a heterogeneity at objective function flux value level emerged in the population model between the two subpopulations of each of the three identified alternative optimal populations (columns "Cell1_P1" and "Cell2_P1", columns "Cell1_P2" and "Cell2_P2", columns "Cell1_P3" and "Cell2_P3"). The bar plot also shows, in all cases, that between the two interacting subpopulations of each population, the one that is responsible for the secretion of lactate in the medium produces ATP at a lower rate compared to the subpopulation in which lactate is consumed

with each other within the same tumor population. The advantage of performing FBA simulations on a population model compared to that on single entity model is the possibility of identifying distinct subpopulations having different phenotypes, but coexisting within the same system, and the possibility of better understanding the heterogeneity degree within a cancer population.

Through our approach, we came to the conclusion that the entire cancer population can be represented, at a first level, through a single entity model which provides a snapshot of the average behavior of the cell population, and at a second level, through a "network of metabolic networks", each of them representing the individual subpopulations. Indeed, just knowing the average behavior results in a limited outlook because the heterogeneity that might emerge within cancer population is not considered. Exploiting FBA method on a network consisting of multiple interacting components allowed us to observe that cancer cells cooperate with each other to reach a specific objective, and that they do not need to have the same metabolism type in order to reach the optimal value of objective function. Through our methodology we explored another level of complexity owned by cancer disease: the objective of the system does not correspond to the objectives of the individual entities since different subpopulations have different role within tumor tissue.

Since rewiring of energy-generating pathways and enhanced growth are closely related, the results here obtained following the ATP production maximization, may be generalized to the case of maximization of biomass production in cancer population. Accordingly, we can say that also the main metabolic trait that unifies all cancer cells (i.e., an uncontrolled and enhanced proliferation), is not the common objective for all individual cells belonging to the system. As stated by the cancer stem cell theory, the tumor growth is not driven by all cells belonging to the cancer population, but it is mainly sustained by only a specific portion of the tumor that consists in the so-called cancer stem cells [197, 198].

Further analyses on more complex metabolic models will be performed to further validate our methodology and to investigate whether the observation that, among interacting subpopulations of a population model, those that are responsible for lactate consumption produce biomass at an higher level, holds even for more biologically grounded and comprehensive metabolic models. In principle, our modeling approach may be applied to genome-scale models. FBA is indeed computationally efficient even for very large networks, while the FVA computation can be sped up by parallel implementations. For example, the computation time to perform FVA on a model consisting of 2593 reactions and 2414 metabolites, by means of the COBRA Toolbox parallelized FVA function [160], is 27.34 seconds (Laptop Windows 10 - 64 bit Intel(R) Core i7-4710HQ CPU 2.50GHz -

RAM: 16.0GB) and the computation time grows linearly with the model's size. Nevertheless, the actual problem of working with genome-scale models is the non-straightforward interpretation of the simulation outcomes, with particular regard to the typical large variability of optimal solutions, which may hinder the interpretation of the cooperation phenomena. Core metabolic models of specific aspects of metabolism may thus be more effective in uncovering system-level properties of cancer populations [78]. In conclusion, we would like to emphasize that the approach discussed here, is not tailored to just analyzing cancer cells populations, but it may be suitable for exploring, in general, how the interactions among more than one component (such as different types of healthy cells, bacteria, yeasts) may influence the overall behavior of a population for which a mismatch between the objective of the individual members and that of the entire population is assumed.

4.1.2 popFBA: tackling intratumour heterogeneity with Flux Balance Analysis

Damiani C†, **Di Filippo M**†, Damiani C, Pescini D, Maspero D., Colombo R., Mauri G

Bioinformatics 2017; 33(14):i311-i318

DOI: 10.1093/bioinformatics/btx251

Abstract

Motivation: Intratumour heterogeneity poses many challenges to the treatment of cancer. Unfortunately, the transcriptional and metabolic information retrieved by currently available computational and experimental techniques portrays the average behaviour of intermixed and heterogeneous cell subpopulations within a given tumour. Emerging single-cell genomic analyses are nonetheless unable to characterise the interactions among cancer subpopulations. In this work, we propose *popFBA*, an extension to classic Flux Balance Analysis (FBA), to explore how metabolic heterogeneity and cooperation phenomena affect the overall growth of cancer cell populations.

Results: We show how clones of a metabolic network of human central carbon metabolism, sharing the same stoichiometry and capacity constraints, may follow several different metabolic paths and cooperate to maximise the growth of the total population. We also introduce a method to explore the space of possible interactions, given some constraints on plasma supply of nutrients. We illustrate how alternative nutrients in plasma supply and/or a dishomogeneous distribution of oxygen provision may affect the landscape of heterogenous phenotypes. We finally provide a technique to identify the most proliferative cells within the heterogenous population.

Availability: the popFBA MATLAB function and the SBML model are available at https://github.com/BIMIB-DISCo/popFBA

Introduction

Populations of tumour cells display considerable phenotypic diversity both at the intertumour and intratumour level. Along with genetic and epigenetic factors, differential trophic supply and variations in the tumour microenvironment contribute in particular to intratumour metabolic heterogeneity and to the emergence of a complex cancer population architecture [127]. Intratumour heterogeneity increases the repertoire of possible cellular responses to a drug and foster the adaptive nature of cellular behaviours [199], compromising the efficacy of cancer therapies.

Although single cell-based technologies represent a promising approach for a more in-depth understanding of single cell behaviour within solid tumours, cancer populations are composed of both tumour and stromal cells that interact with each other by establishing a network of interactions that cannot be deciphered from the analysis of each of these individual components alone [200]. Computational methodologies that allow to identify the possible cooperations

that can be established to enhance the overall growth of the tumour mass and to investigate the mechanisms underlying them are therefore desired as they may facilitate the development of more effective cancer treatments. In particular, modelling efforts to integrate different sources of data and to determine the metabolic phenotype of cooperating cells would provide information that cannot be obtained directly from the genotype, transcriptome, proteome nor the metabolome alone [201].

Constraint-based modeling and especially Flux Balance Analysis (FBA) represents so far the most applied technique for studying metabolism and has effectively been exploited as a scaffold for 'omic' data integration [48]. In particular, many methods have been introduced for the integration of transcriptomic data into constraint-based models of metabolism [202]. FBA is performed on a single metabolic network and provides the (optimal) net flux distribution of possibly different metabolic populations, hindering the identification of possible metabolic interactions between subpopulations.

In [203], we preliminary showed that clones of a constraint-based toy metabolic network can cooperate to maximise the ATP production of a total population. In this work we propose a method to explore the space of possible interactions between heterogenous cell populations within a putative tumour - given some constraints on plasma supply of nutrients - as well as a technique to identify the most proliferative sub-phenotypes.

Approach

Because the classic Flux Balance Analysis approach on a single metabolic model predicts the optimal net flux distribution of a population [203], the single metabolic model just represents a sort of black box of the investigated population, and on its own is unable to inform about the complex interactions occurring inside it. Therefore, we propose a new methodology aimed at investigating metabolic phenotypes of different subpopulations belonging to the same tumour mass, especially focusing on the relationships among them.

Assuming that a tumour mass may be composed of different types of cells (including stromal cells) and that, for reasons of spatial proximity, the communication with the plasma (in terms of nutrients exchange) of the different components may differ with respect to the communication with other cells whiting the populations, we separately modelled exchanges with plasma and exchanges with the tumour microenvironment.

In particular, the single metabolic model is used as a building block for constructing, in an automatic way, the population model characterised by multiple

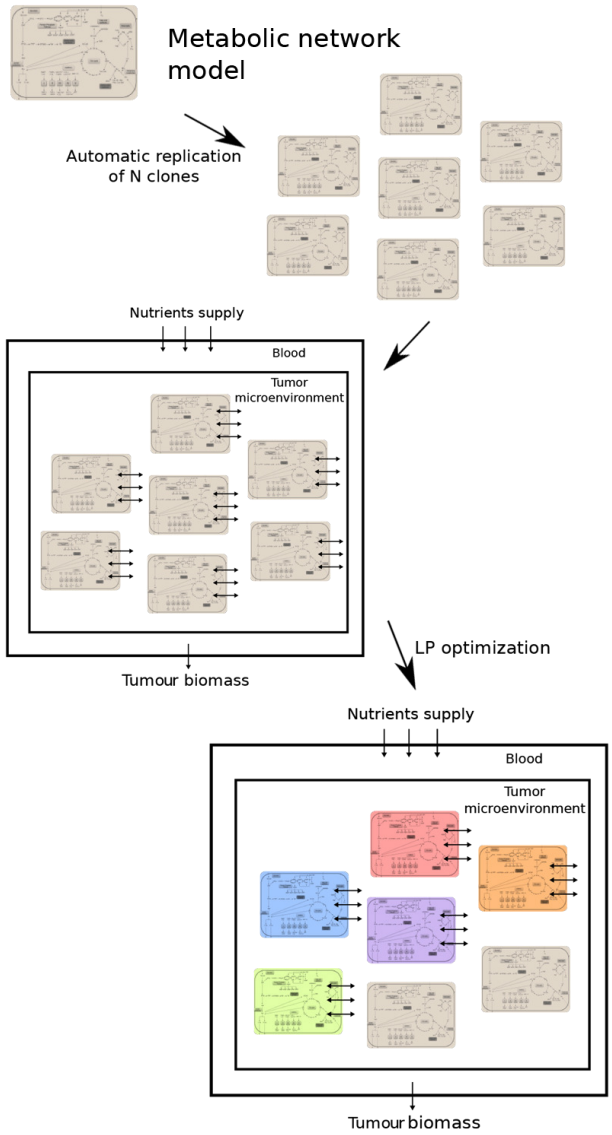


Figure 4.6: Graphical representation of popFBA methodology.

clones, all having identical stoichiometry and capacity constraints and sharing the plasma supply of nutrients. Exploiting linear programming optimisation, we can investigate both the cooperation among different clones that is consistent with the achievement of the optimal growth rate of the entire tumour mass, and identify the strategies adopted by most proliferative clones.

The proposed approach is schematically described in Figure 4.6 and formally defined in the next section.

Methods

Metabolic Network Model

A metabolic network is formalised by specifying a set $\mathcal{X} = \{X_1, \ldots, X_M\}$ of metabolites, and the set $\mathcal{R} = \{R_1, \ldots, R_N\}$ of chemical transformation taking place among them. Reactions are defined as:

$$R_j : \sum_{i=1}^{M_j^s} \alpha_{ji} X_i \longleftrightarrow \sum_{i=1}^{M_j^p} \beta_{ji} X_i, \tag{4.2}$$

where M_j^s and M_j^p are respectively the number of reactants and products of reaction j - relatively to the case in which the reaction proceeds in the forward direction (from left to right) - and $\alpha_{ji}, \beta_{ji} \in \mathbb{N}$ are stoichiometric coefficients associated, respectively, with the i-th susbtrate and the i-th product of the j-th reaction.

For the analyses reported in this work we used the model of central carbon metabolism, which we refer to as 'COREHMR' , extracted from the HMR model [204] and introduced in [78], composed of 243 metabolites and 271 reactions. To adapt the model to popFBA analyses, we set the mitochondrial isocitrate dehydrogenase-catalysed reactions as reversible; we introduced a consumption of ATP within the biomass reaction; we introduced a cell maintenance reaction ($ATP \longrightarrow ADP$); and finally we structurally removed the thermodynamically infeasible loops detected with the algorithm developed in [**?**].

FBA

The assumption underlying Flux Balance Analysis is that metabolic networks will reach a steady state: the concentration of each metabolite is assumed constant: $d[X_i]/dt = 0 \; \forall X_i \in \mathcal{X}$. The stoichiometric constraints lead to a bounded solution space of all feasible flux distributions, which can be further restricted

by specifying maximum and minimum fluxes through any particular reaction. These boundaries on admissible fluxes allow to define constraints on the reversibility of reactions, to impose experimentally measured flux ranges, and to set the extent of nutrients supply. The exchange of matter with the environment is represented as a set $\mathcal{E} = \{E_1, \ldots, E_{N_{ext}}\}$ of N_{ext} unbalanced reactions (exchange reactions), enabling a predefined set of metabolites (including the pseudo-metabolite representing biomass) $\mathcal{Y} = \{Y_1, \ldots, Y_{N_{ext}}\} \subset \mathcal{X}$ to be inserted in or removed from the system, through reactions as E_j:

$$E_j : Y_j \longleftrightarrow \emptyset \tag{4.3}$$

The first step to perform FBA is the derivation of the stoichiometric matrix S, of size $M \times (N + N_{ext})$, whose element s_{ji} takes value $-\alpha_{ji}$ if the species X_i is a reactant of reaction R_j, $+\beta_{ji}$ if the species X_i is a product of reaction R_j and 0 otherwise.

FBA is then applied to determine the rate v_i at which each reaction in $\mathcal{R} \cup \mathcal{E}$ occurs, that is, the flux distribution $\vec{v} = (v_1, v_2, \ldots, v_{N+Next}) = (r_1, \ldots, r_N, e_1, \ldots, e_{N_{ext}})$ that maximises or minimises the objective function $Z = \sum_{i=1}^{N+Next} w_i v_i$, where w_i is the weight that quantifies the contribution of reaction i, while r_i and e_i are, respectively, the flux value associated to an "internal" or an exchange reaction.

In order to simulate tumour growth, in this work we maximise the flux $e_{biomass}$ through the biomass exchange reaction, thus $Z = e_{biomass}$. The optimisation problem is postulated as a general Linear Programming (LP) formulation:

$$\begin{aligned} &\text{maximize or minimize } Z \\ &\text{subject to } S\vec{v} = \vec{0}, \\ &\vec{v}_L \leq \vec{v} \leq \vec{v}_B \end{aligned} \tag{4.4}$$

$\vec{v}_L$ and $\vec{v}_B$ are vectors specifying the lower and upper bound respectively for each fluxv_i of $\vec{v}$. A negative lower bound indicates that flux is allowed in the backward reaction.

To solve the above problem we exploited the GLPK solver within the COBRA Toolbox [160]. For a more comprehensive description of FBA, the reader is referred to [60].

popFBA

In order to investigate the role of cooperation within a population sharing a common environment, in this work we devised popFBA, an extension to FBA

able to cope with the presence of several subpopulations exchanging a defined set of metabolites. Given e metabolic network A defined as $A = (\mathcal{X}, \mathcal{R}, \mathcal{E})$, popFBA maximizes the total biomass of $Npops$ clones A^c of A, which can cooperate by exchanging nutrients in the TUMOUR MICROENVIRONMENT. For each clone A^c, let $\mathcal{X}^c = \{X_i^c\}$ be the set of its metabolites, $\mathcal{R}^c = \{R_j^c\}$ the set of its internal reactions, with $j = 1, \ldots, N$ and $c = 1, \ldots, Npops$. To correct for the fact that in a population model a metabolite is not removed from the systems but becomes a metabolite in the TUMOUR MICROENVIRONMENT, each reaction $E_j \in \mathcal{E}$ of A is transformed in a *cooperation reaction* C_j^c with the form

$$C_j^c : Y_j^c \longleftrightarrow Y_j' \tag{4.5}$$

It is also necessary to define the new set of TUMOUR MICROENVIRONMENT metabolites $\mathcal{Y}' = \{Y_i'\}$ with $i = 1, \ldots, N_{ext}$, together with a new set of N_{blood} exchange reactions $\mathcal{B} = \{B_1, \ldots, B_{N_{blood}}\}$ to allow a subset of metabolites $\mathcal{K} = \{K_1, \ldots, K_{N_{blood}}\} \subset \mathcal{Y}'$ to be exchanged with the BLOOD SUPPLY:

$$B_j : K_j \longleftrightarrow \emptyset \tag{4.6}$$

The population model P is then defined by the union set of the metabolites $\mathcal{X}^P = \bigcup_c \mathcal{X}^c \cup \mathcal{Y}'$, of the internal reactions $\mathcal{R}^P = \bigcup_c \mathcal{R}^c$, of the COOPERATION REACTIONS $\mathcal{C}^P = \bigcup_c \mathcal{C}^c$ and of the population exchange reactions $\mathcal{B}$.

A stochiometric matrix S^P is then built for all reactions in $\mathcal{R}^P$, $\mathcal{C}^P$ and $\mathcal{B}$ and for all metabolites in $\mathcal{X}^P$ and $\mathcal{Y}'$.

The final size of matrix S^P is $(Npops \cdot M + N_{blood}) \times (Npops \cdot (N + N_{ext}) + N_{blood})$.

To obtain the matrix S^P, we implemented a MATLAB function that automatically replicates a number of times any (COBRA Toolbox compliant [160]) SBML model to obtain the above defined population model, in a suitable form to then undergo constraint-based analyses.

Linear programming is then applied as per Equation 4.12 to determine the flux distribution $\vec{v} = (v_1, \ldots, v_{Npops \cdot (N+N_{ext})+N_{blood}}) = (r_1^1, \ldots, r_N^1, \ldots, r_1^{Npops}, \ldots, r_N^{Npops}, c_1^1, \ldots, c_1^{Npops}, \ldots, c_{N_{ext}}^{Npops}, b_1, \ldots, b_{N_{blood}})$ that maximises the biomass exchange flux $b_{biomass}$, with v_i representing any flux i of the population model, and for each clone c, r_i^c representing the i-th internal flux, c_i^c representing the i-th a cooperation flux and b_i a plasma exchange flux.

Sampling in the region of optimal solutions

Linear programming only returns a single optimal solution. However, many alternative optimal flux distributions may exist. Although methods have been

proposed for enumerating alternative optimal solutions [194], an exhaustive enumeration is not practicable for popFBA, due to the interchangeability of the flux distributions of the *Npops* clones.

To cope with this problem, we set the boundaries of the biomass exchange flux $b_{biomass}$ to the optimal value obtained with popFBA and we sampled the admissible solutions. The dominant algorithm of choice to uniformly sampling inside the region of allowed solutions is the so-called "Hit-and-Run" (HR) [205], according to which an initial valid point is moved repeatedly inside the space according to probabilistic rules.

In this work, we also exploited a recently proposed alternative approach [170, 206]: the simplex method with a random set of objective functions to be maximised. The maximisation of each of these objective functions gives a corner in the space of solutions. In [206] random objective functions were generated by selecting random pairs of reactions. To maximise variability of sampled solutions, we instead let any number of reactions to take part in the objective function Z as in [170]. The fraction τ of considered reactions is randomly drawn with uniform probability in (0, 1]. To any selected reaction is then assigned a random weight w_i uniformly tossed from the interval $(0, 1]$ where w_i takes value 0 with probability τ and a random value with uniform probability in $[0, 1]$ with probability $1 - \tau$.

For both methods, we controlled for repetitions in the sampled points.

Assessing subpopulations heterogeneity

To assess the heterogeneity of the metabolism of the *Npops* clones within a given optimal solution, we compared their flux distributions both quantitatively and qualitatively. To avoid taking into account clones that carry no flux, we disregarded the CLONEs that are not cooperating for tumour growth, by filtering out CLONEs that do not have a least one non-zero flux through any of the COOPERATION REACTIONs.

To count the number of quantitatively different SUBPOPULATIONs we considered clones a and b to be different if they differ by at least the value of one flux (rounded at the fourth digit): if $\exists\ i : v_i^a \neq v_i^b$. To count the number of qualitatively different SUBPOPULATIONs, the counting process is performed, rather than on $\vec{v'}$, on a discretized version $\vec{d'}$ of vector $\vec{v'}$, whose component d'_i is obtained as follows:

$$d'_i = \begin{cases} 1 & v_i > 0 \\ -1 & v_i < 0 \\ 0 & v_i = 0 \end{cases} \tag{4.7}$$

In this way two clones are considered as different *iff* they follow at least one different metabolic path, that is, a different route or a different flux direction.

Results

popFBA reveals the existence of cooperation and metabolic heterogeneity within cancer population models

We applied popFBA to 10 clones ($Npops = 10$) of the COREHMR model. In a first experiment we simulated a PLASMA supply of glucose, glutamine and oxygen, the three nutrients provided with the same order of molar magnitude, in accordance with the magnitude found when scanning data in literature. Although the molar concentration of glutamine may be lower as compared to that of glucose and oxygen, in our system glutamine represents the unique nitrogen source, so it is reasonable to increase its uptake flux to account also for other nitrogen sources that are generally available in the plasma. As a first approximation, we assumed equal boundaries for the reactions of the 10 clones, thus disregarding spatial diffusion phenomena. The 10 clones can cooperate, via COOPERATION REACTIONs described in the Methods section, by exchanging lactate, glutamine, glutamate and ammonia. Although these metabolites can be secreted in the microenvironment compartment, they cannot be disposed (as waste products) in the human PLASMA. This experimental setting, which we refer to as the reference condition, is better detailed in Figure 4.7A. In this condition, we sampled $2 \cdot 10^4$ different optimal solutions (10^4 points with the HR sampling methods and 10^4 with the CB sampling method), that are compatible with the same optimal tumour biomass.

We assessed the quantitative heterogeneity of the clones in each of the sampled solutions. Remarkably, in all sampled solutions, we observed that all 10 clones behave differently. This diversity results in a different biomass synthesis flux value (the growth rate) for distinct clones. This heterogeneity may partially be the result of slightly different phenotypes, following the very same metabolic paths but at different rates. We wanted therefore to assess the number of clones that follow different metabolic paths, that is, that differ in the set of reactions that is active and/or in the direction of the flux through such set of reactions. The distribution of the number of qualitatively different subpopulations is reported in Figure 4.7B and it shows that, although SUBPOPULATIONs of quantitatively different clones may overlap from a qualitative point of view, for both sampling methods, the clones typically all follow different metabolic routes.

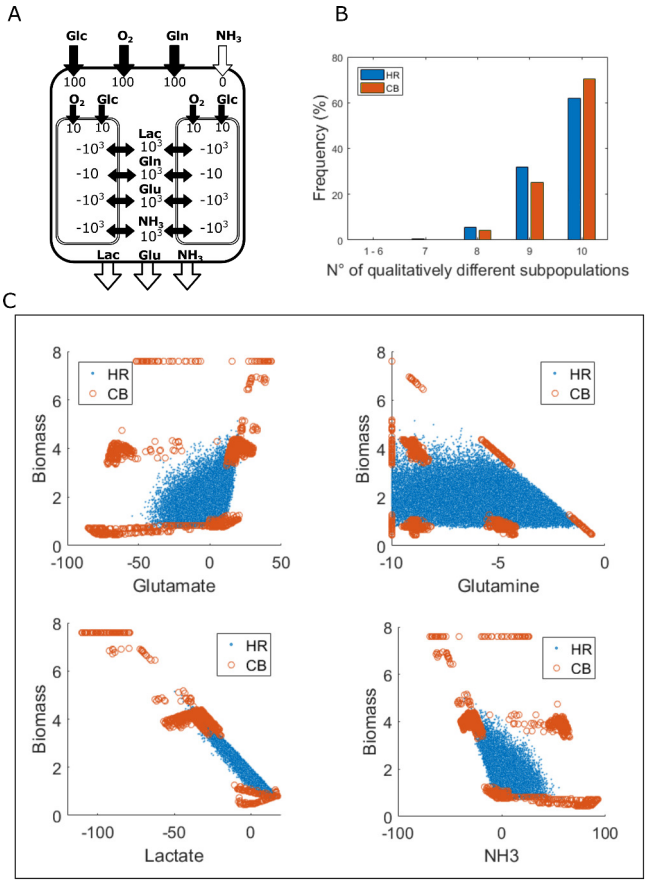


Figure 4.7: Reference condition. A) Experimental setting of the boundaries imposed on the release/consumption of metabolites in/by the plasma, and on the cooperative reactions. Black filled arrows indicate allowed reactions. B) Histogram relative to the number of qualitatively different subpopulations obtained with HR (blue bars) and CB (orange bars) sampling methods. C) Scatter plots obtained with HR (blue points) and CB (orange circles) sampling methods relative, in clockwise order, to the correlation between glutamate exchange and biomass synthesis, between glutamine exchange and biomass synthesis, between NH3 exchange and biomass synthesis and between lactate exchange and biomass synthesis. Abbreviations: Glc, Glucose; O2, Oxygen; Gln, glutamine; NH_3, ammonia; Lac, lactate; Glu, glutamate.

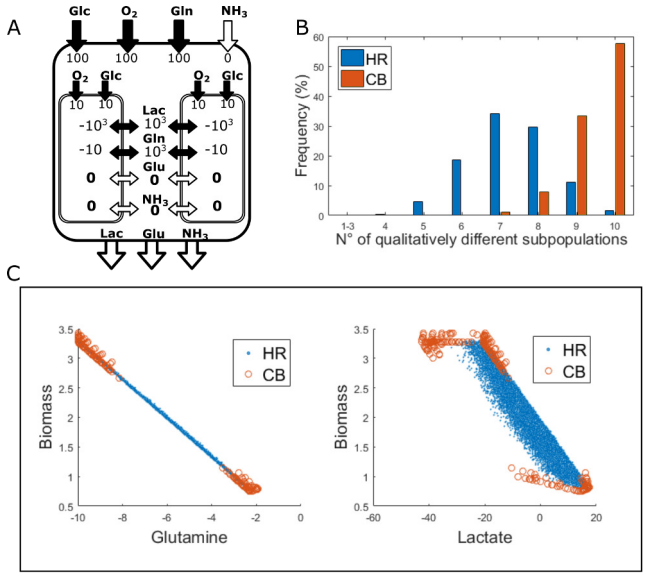


Figure 4.8: Cooperation reactions variation condition. A) Experimental setting of the boundaries imposed on the release/consumption of metabolites in/by the plasma, and on the cooperative reactions. Black filled arrows indicate allowed reactions. B) Histogram relative to the number of qualitatively different subpopulations obtained with HR (blue bars) and CB (orange bars) sampling methods. C) Scatter plots obtained with HR (blue points) and CB (orange circles) sampling methods relative to the correlation between glutamine exchange and biomass synthesis (on the left), and between lactate exchange and biomass synthesis (on the right).

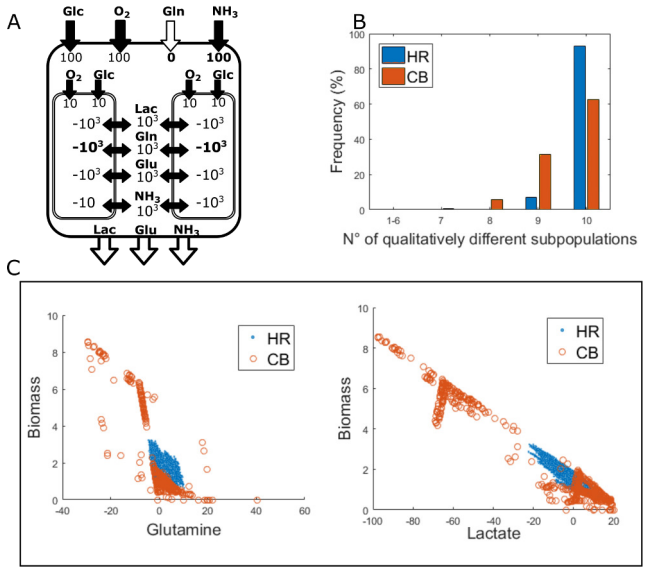


Figure 4.9: Changing nitrogen source condition. A) Experimental setting of the boundaries imposed on the release/consumption of metabolites in/by the plasma, and on the cooperative reactions. Black filled arrows indicate allowed reactions. B) Histogram relative to the number of qualitatively different subpopulations obtained with HR (blue bars) and CB (orange bars) sampling methods. C) Scatter plots obtained with HR (blue points) and CB (orange circles) sampling methods relative to the correlation between glutamine exchange and biomass synthesis (on the left), and between lactate exchange and biomass synthesis (on the right).

Once we established that popFBA is able to highlight the possible heterogeneity of subpopulations of cells belonging to the same tumour population, we shifted the attention toward a more in-depth investigation of which types of interactions are compatible with the achievement of the maximum tumour biomass. The observed heterogeneity results indeed form cooperative behaviours among different subpopulations. In fact, as it can be observed in the scatter plots in Figure 4.7C, a secretion (negative values) or consumption (positive values) of metabolites in the tumour microenvironment, namely of lactate, glutamate and ammonia emerged among the different subpopulations. Among the four allowed cooperations, glutamine is the only metabolite that is always just consumed and it is not exchanged with other subpopulations.

Identification of most proliferative subpopulations

To identify the most proliferative phenotypes among the heterogenous subpopulations identified above, we computed the Pearson Correlation Coefficient (ρ) between the flux of each of the four cooperation reactions (exchange of glutamate, glutamine, lactate and ammonia via the tumour microenvironment) and the biomass production rate of the corresponding clones, consistently with the achievement of the optimal tumour biomass.

We found that the correlation with the biomass synthesis flux, obtained with HR (and CB), is -0.32 (-0.4) for glutamine; +0.35 (+0.44) for glutamate; -0.55 (-0.65) for ammonia; and -0.96 (-0.99) for lactate. Notice that a negative ρ implies a positive correlation between nutrient consumption and biomass, while a positive ρ implies a positive correlation between nutrient secretion and biomass.

The ρ values, along with the corresponding scatter plots (Fig. 4.7B), clearly indicate that the most proliferative subpopulations are those consuming the lactate available in the tumour microenvironment. To evaluate whether lactate consuming subpopulations are oxidising this carbon source, we also analysed the correlation between the lactate exchange and oxygen consumption fluxes, obtaining a ρ of +0.98 (+0.99), confirming that the most proliferative subpopulations are characterised by an active mitochondrial metabolism which is plausibly exploited to completely oxidise the consumed lactate.

The set of nutrients exchanged with plasma affects possible cooperations

The observation that the glutamine supplied with plasma is consumed by all 10 popFBA clones, with no clone producing and exchanging it with the tumour microenvironment, was expected. Glutamine represents indeed a unique source of nitrogen supplied by plasma in the above simulations, and nitrogen is mandatory for the synthesis of amino acids and thus of biomass. The correlation between glutamine and biomass is however modest, because the clones are able to extract the nitrogen from glutamine and to exchange it in the form of NH_3 or glutamate via the tumour microenvironment. Unexpectedly, the correlation with biomass is indeed negative for NH_3, but it is positive for glutamate.

We observed that, when the cooperation flux for both NH_3 and glutamate is prevented (experimental setting in Figure 4.8A and results in Figures 4.8B-C), the correlation between glutamine and biomass becomes indeed close to -1 (-0.99 for both the HR and the CB method). In this situation the negative correlation

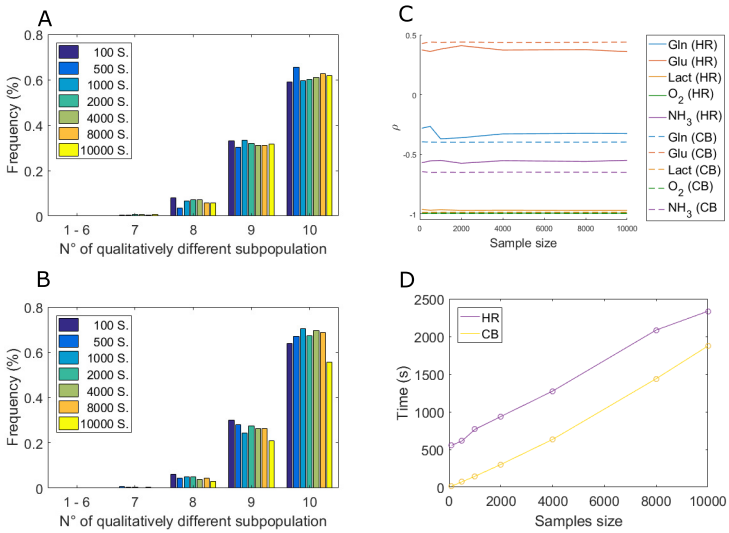


Figure 4.10: A) Distribution of the number of qualitatively different subpopulations obtained for 7 independent samples of different size S ($S \in \{100, 500, 1000, 2000, 4000, 8000, 10000\}$) obtained with the HR method. Parameter $nStepsPerPoint$ of the $sampleCbModel$ MATLAB function was set to default value (200); parameter $nWarmupPoints$ was set to 6000; parameters $nFiles$ and $nPointsPerFile$ were set in a way to maintain the ratio between their default values. B) Distribution of the number of qualitatively different subpopulations obtained for 7 independent samples of different size $\{100, 500, 1000, 2000, 4000, 8000, 10000\}$ obtained with the CB method. C) Values of ρ between biomass and each of the exchange fluxes with the intratumour microenvironment (Gln, Glu, Lact, O_2, NH_3), obtained with HR (solid lines) and CB (dashed lines) as a function of sample size. D) Computation time as a function of samples size for HR and CB.

between lactate and biomass is preserved (-0.97 for HR and -0.96 for CB).

We also wanted to investigate whether the source of nitrogen in the plasma supply may affect the possible cooperative behaviours. We tested the situation in which the nitrogen source provided by the plasma to the tumour mass is not represented by glutamine, but by the ammonia (Fig. 4.9A), which may account for the nitrogen deriving from other amino acids in real cells. In this situation, as shown in Figure 4.9C, the exchange of glutamine among different subpopulations becomes possible. On the contrary, the exchange of NH_3 becomes not possible, as it is now exclusively consumed by the 10 clones (data not shown). Notably, we still observed a high correlation between lactate consumption and biomass formation: ρ =-0.89 (-0.98), as per Figure 4.9C.

Once the effect of an alternative nitrogen source on the internal transport reactions was investigated, we also analysed the possibility to have an outflow of either glutamate or ammonia or lactate from the tumour microenvironment compartment toward the plasma, maintaining the reference experimental setting. We observed that a slight increase in the tumour biomass value is caused by both glutamate (2%) and lactate secretion (2%) in the plasma.

Interestingly, we observed that the exit of glutamate from the tumour microenvironment towards the plasma does not prevent an exchange of glutamate among different subpopulations. On the contrary, as compared to the reference condition in Figure 4.7, this situation results in an enhancement of the correlations between glutamate exchange and biomass synthesis rate – from +0.35 (0.44) to +0.47 (+0.61); between glutamate exchange and lactate exchange – from -0.29 (-0.39) to -0.47 (-0.61); and between glutamate and oxygen – from -0.26 (-0.36) to -0.44 (-0.59). These results indicate that the subpopulations that are responsible for a glutamate secretion in the tumour microenvironment are characterised by a consumption of both lactate and oxygen.

Comparison of HR and CB sampling methods

In the scatter diagrams in Figures 4.7C, 4.8C and 4.9C it is evident how the two sampling methods (HR and CB) differ in the way they explore the space of possible cooperations among metabolic clones. HR uniformly samples within the space of optimal solutions, returning points that are close to one another inside the space of allowed solution, whereas, as already pointed out in [206], the CB method returns solutions corresponding to the corners in the region of allowed flux distributions. This difference in the two methods does not particularly affect the correlation coefficients between the release/consumption of metabolites within the tumour microenvironment and the growth rate of a given sub-

population. It can be indeed observed in Figure 4.10C that the ρs obtained with the two methods are very similar. However the HR method may in some cases underestimate the number of qualitatively different SUBPOPULATIONs. Although the distribution of the number of qualitatively different SUBPOPULATIONs of the two methods is very similar in the experiments presented in Figure 4.7B and 4.9B, it substantially diverges in the experiment in Figure 4.8B. A possible explanation of this phenomenon is that the experiment reported in Figure 4.8B refers to a case in which the clones have less possibilities to cooperate (glutamate and ammonia cooperation is prevented) and thus tend to be more similar to one another (indeed also the CB method finds more cases with a lower number of different subpopulations as compared to experiments in Figures 4.7B and 4.9B), and that the HR method amplifies this effect.

As expected, the computation time of the two methods grows linearly with the sample size (Fig. 4.10D) and it is slightly higher for HR because of the fixed initial time to create the warm up points. In both cases, the computed ρs stabilise for a sample size greater than $\sim 4 \cdot 10^3$. Also the shape of the distribution of the number of different subpopulations (Figure 4.10A and B for HR and CB respectively) is not particularly affected by sample size, although some oscillations are possible especially for the CB case, while the frequency of any number of different subpopulations becomes stable for samples greater than $\sim 8 \cdot 10^3$. All in all these considerations confirm that the size of the sample chosen for our analysis ($10^4 + 10^4$) was reasonable.

Simulation of spatial diffusion phenomena with popFBA

To prove the viability of our popFBA approach to take into account spatial diffusion phenomena, we performed an experiment in which, differently from the experimental settings presented above, the 10 clones are not identical in terms of access capability to the PLASMA supply.

To mimic a simplified oxygen gradient, we set the boundaries of oxygen uptake of the 10 clones according to a linear decreasing function (as illustrated in Figure 4.11A). Assuming that subpopulations are radially stratified, clones with a given availability of oxygen may represent subpopulations of cells at an equal distance from the blood vessel.

The scatterplots in Figure 4.11B clearly show that the subpopulations that consume less oxygen exhibit higher rates of lactate production. Interestingly, the lactate produced by those subpopulations is consumed by cells with high oxygen consumption rates, as indicated by high negative values of the lactate cooperation flux in correspondence with high negative values of oxygen uptake.

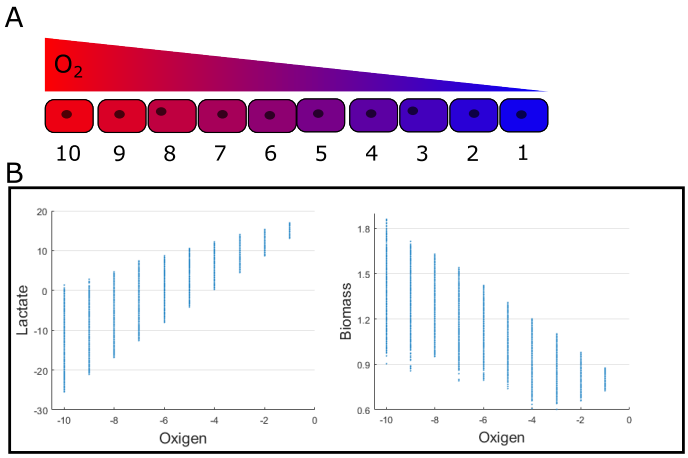


Figure 4.11: Simulation of oxygen spatial diffusion. A) Schematic representation of the performed experiment. The red-to-blue chromatic scale is used to mimic an oxygen gradient through the tumour population, respectively, from an aerobiosis to an hypoxic environment. The values on the bottom of the figure refer to maximum uptake flux value of oxygen imposed on each clone. B) Scatter plots obtained with HR sampling methods relative to the correlation between lactate exchange and oxygen uptake (on the left), and between biomass synthesis and oxygen uptake (on the right).

As expected, oxidative cells (high oxygen consumption) show higher growth rates, as compared to fermentative ones (high lactate production), as shown in Figure 4.11B.

Conclusions

We introduced popFBA, an extension of FBA to take into account intratumour heterogeneity and interactions among different cell populations within the same tumour. We applied popFBA to a model of 10 clones of the metabolic network of human central carbon metabolism, simulating a PLASMA supply of glucose, glutamine and oxygen, assuming equal boundaries for the reactions of the 10 clones and an internal exchange of lactate, glutamine, glutamate and ammonia. We sampled different optimal solutions, by using both the HR sampling method and the CB sampling method, that are compatible with the same optimal tumour biomass.

We observed that popFBA reveals the existence of metabolic heterogeneity and cooperation within the population model. Indeed, by assessing the quantitative heterogeneity of the clones in each of the sampled solutions, we observed that all 10 clones behave differently and are characterised by a different growth rate. Moreover, by assessing the distribution of the number of qualitatively different subpopulations, we found that the subpopulations are following different metabolic routes. This observed heterogeneity is the result of a cooperative behaviour among subpopulations: a consumption or secretion of lactate, glutamine, glutamate and ammonia (the four exchanged metabolites) emerged indeed among the different subpopulations.

Following the observation that popFBA approach is able to point out, if it is present, a cooperative behaviour among multiple subpopulation of the same population, we investigated which types of interactions are compatible with the achievement of the optimal tumour biomass.

In order to characterise the metabolism of most proliferative subpopulations we assessed the correlation between the exchanges of metabolites within the tumour microenvironment and the growth rate of a given subpopulation. In this regard, we showed that, although the two sampling methods (HR and CB) differ in the way they explore the space of possible solutions, this difference has no particular effect on the computed correlation coefficients. Therefore the two methods are equally effective when the goal is to determine the subpopulation with a propensity for growth.

Remarkably, in all of the investigated scenarios, we observed that the lactate exchange within tumour microenvironment is negatively correlated with respect to the biomass synthesis flux. This means that the most proliferative subpopulations consume the lactate that is secreted in the tumour microenvironment compartment by less proliferative subpopulations, by using it as energy source. Because a high positive correlation between the lactate exchange and oxygen consumption fluxes emerged, we deduced that the most proliferative subpopulations are also characterised by an active mitochondrial metabolism that is exploited to completely oxidise the consumed lactate.

Notably, this robust result is in line with the experimental evidence [190, 207] of the existence of a stromal-epithelial lactate shuttle in human tumours, a phenomenon named as "reverse Warburg effect" due to the fact that tumour stromal cells undergo aerobic glycolysis producing lactate that is used as energy source by the adjacent high proliferative cancer cells. Therefore, this "symbiotic" but also "parasitic" relationship between cancer cells and stroma cells fuels the overall biomass of the entire tumour mass.

The agreement of our results with the above mentioned experimental data

supports the reliability of our approach and it also confirms the need for computational and experimental approaches able to take into account the specificity of the subpopulations within a tumour rather than observing the average behaviour. In fact, several works based on the average behaviour of cancer populations pointed out that cancer cells are characterised by a high secretion of lactate in the medium [17, 182], but our approach suggest that these indications might not be precise.

We can conclude that popFBA approach is able to highlight the metabolic plasticity of the tumour mass with respect to the adaptation of its components (i.e. different subpopulations belonging to the tumour) to changing external but also internal scenarios.

Taking inspiration from the plethora of existing methods to integrate transcriptomic data into classic FBA [202], we plan in the next future to define a method to incorporate single cells transcriptomic data into popFBA models, in order to pave the way to the integration of the increasing availability of single cells RNAseqs into computational models.

4.2 From population model to single-cell behaviour

A further step has been done in the investigation of metabolic plasticity of cell populations. Although the potentialities revealed by popFBA approach in Section 4.1.2 in capturing both phenotypic heterogeneity and cooperation among various components of population model, countless combinations of flux distributions of the individual subpopulations emerged. Therefore, finding a way to reduce the possible combinations of individual flux distributions by adding further constraints on the individual components of the population model represented our challenging task.

In this regard, in Section 4.2.1, I present a novel data-integration framework named Metabolic Reaction Enrichment Analysis (MaREA). MaREA aims at integrating transcriptomics RNA-Seq data into into metabolic networks by exploiting the gene-protein-reaction (GPR) rules, without requiring metabolic measurements, in order to compare metabolism of samples in distinct subgroups or different experimental conditions. For that purpose, MaREA computes for each reaction of a given metabolic network a Reaction Activity Score (RAS). RAS score describes the extent of each reaction activity in a given condition, as a function of the expression of the genes encoding for the subunits and/or the isoforms of the associated enzyme(s).

RAS scores are not used as constraints or objective functions in FBA simulations. On the contrary, these scores are static representation of transcriptional deregulations of metabolic reactions of different patient cohorts or conditions (e.g., physiological vs. pathological condition). Thanks to RAS score, MaREA can be used to rank reactions according to their activity variation across different phenotypes and/or experimental conditions. Moreover, it also provides a graphical visualization of how deregulated paths are interconnected, by enriching the map of human metabolic routes with RAS variation. Finally, MaREA, providing a new unsupervised clustering tool, allows to stratify samples according to their metabolic activity.

The reaction enrichment performed via MaREA allows to identify the metabolic patterns underlying the phenotypic and functional properties observed in different sample subgroups, as in the case of patients with distinct cancer subtypes, by visualizing the up-/down-regulated reactions directly on the metabolic networks. MaREA showed its effectiveness in identifying, ranking and visualizing the enriched metabolic reactions between normal and cancer samples. In particular, well-known features of cancer deregulation have been reproduced

and new experimental hypotheses have been generated. Moreover, by applying MaREA to two distinct datasets of The Cancer Genome Atlas database relative to lung and breast cancer, it was possible to stratify, in an unsupervised way, cancer patients in clusters with similar metabolic activity. Metabolic clusters of samples identified by MaREA displayed significantly different survival expectancy, as retrieved from clinical data, for both breast and lung cancer, by proving a crucial role of metabolic reprogramming in cancer aggressiveness. Compared to lung cancer where up to 9 sample subgroups with significantly different prognosis have been identified, metabolic differences within breast cancer in terms of patient prognosis were less relevant. This outcome suggested a less relevant role of metabolic heterogeneity in breast cancer. Nevertheless, MaREA identified two metabolic clusters in breast cancer characterized by significantly different prognosis, showing a relevant overlap with signatures of two of the molecular subtypes of breast cancer that have been defined based on the relative expression of specific genes [208].

When cluster analysis is performed by using a well curated core model focused on central carbon metabolism rather than a genome-wide network, we observed that prognostic power of MaREA improved. Model curation to which I contributed has been performed in terms of the GPR rules associated to metabolic reactions.

Finally, considering RAS scores instead of the only transcripts allows to effectively summarize metabolic deregulations when two conditions are compared. In this regard, MaREA enables to identify, rank and visualize critical metabolic reactions that can be targeted by using metabolic drugs, instead of targeting the expression of individual genes.

Combining the potentialities revealed by this methodology with those previously discussed of popFBA approach, we then developed a new computational framework called single-cell Flux Balance Analysis (scFBA) in order to respond to the previous challenging task of reducing individual flux distribution within population models.

In this regard, scFBA is based on the addition of constraints on the single-cell flux distributions to obtain single-cell specific metabolic models. To this end, scFBA aims at translating single-cell transcriptomes into single-cell fluxomes by exploiting the computation of RAS scores for all the reactions of each single components of the population model in order to characterize the metabolism of heterogeneous cell populations.

In this work, we applied the scFBA methodology to scRNA-seq datasets relative to lung adenocarcinoma and breast cancer.

We firstly observed that the integration of scRNA-seq data efficiently reduces

the space of optimal solutions as compared to the situation where no information on single cell trancriptome is available with the consequent ability of each cell to alone contribute to the 100% of the total population biomass. Indeed, after the scRNA-seq data integration, flux value of biomass synthesis for each cell correspond to a certain fraction of the total population biomass. Moreover, we observed that cluster analysis performed on the metabolic genes expression values, and on the fluxes predicted by scFBA reveal in the second case a better cells clustering in few well-separated groups compared to the lots of singletons generated in the first situation. This outcomes means that fluxes can be better clustered than their transcript counterparts.

Another important results exhibited from the scFBA methodology was its ability to identify the possible interactions among cells within the population, as already pointed out by popFBA approach. However, differently from its predecessor, a more complex network of interactions is established, where, in particular, several cells consumes the lactate and palmitate that are secreted by other groups of cells, but with a greater dispersion of growth rates values than those showed with the popFBA. In line with experimental evidences, when we compared the scenarios where the two major exchanged metabolites, namely palmitate and lactate, are, in turn, externally provided as fuels to the population, or must be secreted in the external environment, a good agreement emerged. Indeed, the uptake of an external source of palmitate does not confer any advantage to biomass production rate, in accordance with experimental evidence under which cancer cells rely on de novo synthesis of palmitate-derived fatty acids rather than on a free availability of lipids. However, we observed that, in the baseline setting when palmitate is not supplied in the external environment, this carbon source is exchanged and then internalized by some of them. In these cells, palmitate directly contributes to the biomass synthesis without being processed through the beta-oxidation pathway. In line with literature findings, an external source of fatty acids can serve to bypass the *de novo* fatty acids synthesis, avoiding the toxic effect provoked by lipogenesis inhibitors. In addition to this finding, we also observed that that a set of genes that are directly involved in the synthesis of palmitate stops being lethal when this source is exogenously supplied. This computational evidence is justified by experimental evidence under which a limited access to environmental lipids may render cancer cells more sensitive to the inhibitors of lipogenesis. Contrary to palmitate, the lactate is not necessarily essential for the purpose of growth. However, the interruption of glucose utilization through glycolysis became lethal when lactate uptake is not allowed, suggesting a possible role of the lactate as alternative carbon source.

4.2.1 Integration of transcriptomic data and metabolic networks in cancer samples reveals highly significant prognostic power

Graudenzi A, Maspero D, **Di Filippo M**, Gnugnoli M, Isella C, Mauri G, Medico E, Antoniotti M,Damiani C

Journal of biomedical informatics 2018; 87:37-49

DOI: 10.1016/j.jbi.2018.09.010

Abstract

Effective stratification of cancer patients on the basis of their molecular make-up is a key open challenge. Given the altered and heterogenous nature of cancer metabolism, we here propose to use the overall expression of central carbon metabolism as biomarker to characterize groups of patients with important characteristics, such as response to *ad-hoc* therapeutic strategies and survival expectancy.

To this end, we here introduce the data integration framework named *Metabolic Reaction Enrichment Analysis* (MaREA), which strives to characterize the metabolic deregulations that distinguish cancer phenotypes, by projecting RNA-seq data onto metabolic networks, without requiring metabolic measurements. MaREA computes a score for each network reaction, based on the expression of the set of genes encoding for the associated enzyme(s). The scores are first used as features for cluster analysis and then to rank and visualize in an organized fashion the metabolic deregulations that distinguish cancer sub-types.

We applied our method to recent lung and breast cancer RNA-seq datasets from The Cancer Genome Atlas and we were able to identify subgroups of patients with significant differences in survival expectancy. We show how the prognostic power of MaREA improves when an extracted and further curated core model focusing on central carbon metabolism is used rather than the genome-wide reference network.

The visualization of the metabolic differences between the groups with best and worst prognosis allowed to identify and analyze key metabolic properties related to cancer aggressiveness. Some of these properties are shared across different cancer (sub)types, e.g., the up-regulation of nucleic acid and amino acid synthesis, whereas some other appear to be tumor-specific, such as the up- or down-regulation of the phosphoenolpyruvate carboxykinase reaction, which display different patterns in distinct tumor (sub)types.

These results might be soon employed to deliver highly automated diagnostic and prognostic strategies for cancer patients.

Keywords: Metabolic networks, RNA-seq data, Genome-wide models, Sample stratification, Cancer metabolism.

Introduction

Alterations of energy metabolism play a relevant role in several pathologies, such as metabolic syndrome, ageing, cancer, diabetes and neurodegeneration [209].

Current research on human metabolism typically relies on genome-wide reconstructions of human metabolic networks [48], such as *Human Metabolic Reaction* (HMR) [210] and Recon [138, 211]. These models include most metabolic reactions that may occur in a generic cell, as well as Gene-Protein-Reaction (GPR) associations, which are logical formulas that describe how gene products concur to catalyze a given reaction.

Several strategies have been proposed to integrate *-omics* data into metabolic networks by exploiting GPRs, in order to derive context-specific models, e.g., the active metabolic network in a given cell or tissue [202, 212, 213].

Such approaches are usually conceived in the framework of constraint-based modeling and, in particular, of *Flux Balance Analysis* (FBA) [60]. FBA relies on *linear programming* techniques to compute the flux through each reaction under a *steady state assumption*, and requires metabolic measurements to constrain nutrient exchange. Unfortunately, the simultaneous presence of metabolic measurements and distinct -omics data on the same patient is rarely available in public databases, such as the *The Cancer Genome Atlas* (TCGA) [214]. Besides, the same metabolic constraints hardly hold for all patients within a single dataset, and more so in highly heterogenous diseases, such as cancer.

Moreover, FBA poses many modeling challenges, e.g., the definition of an appropriate objective function and unfeasibility problems (see Section 4.2.1 for a more detailed discussion).

To address many of these issues, we here introduce a novel data-integration framework named *Metabolic Reaction Enrichment Analysis* (MaREA) (Figure 4.12), which focuses on *transcriptional deregulation* of metabolic reactions, rather than on metabolic flux estimation. That is, MaREA processes transcriptional data, such as RNA-seq, without requiring metabolic measurements.

For each reaction of a given metabolic network, MaREA computes a *Reaction Activity Score* (RAS), which describes the extent of its activity in a given condition, as a function of the expression of the genes encoding for the *subunits* and/or the *isoforms* of the associated enzyme(s). The RAS provides a more refined information than the mere list of genes associated with a reaction, without requiring the setting of any arbitrary threshold, or to binarize data (i.e., gene present or absent), as required by other approaches, such as [215]. Analogous scores are have been employed to integrate continuous gene expression data in constraint-based simulations [202, 212, 213]. However, in MaREA the RAS is not used to define constraints or objective functions in FBA simulations. Instead, it is used as a static representation of the metabolic deregulation of a given sample, which can be directly used to compare different sample sets, e.g., different patient cohorts, or physiological vs. pathological condition.

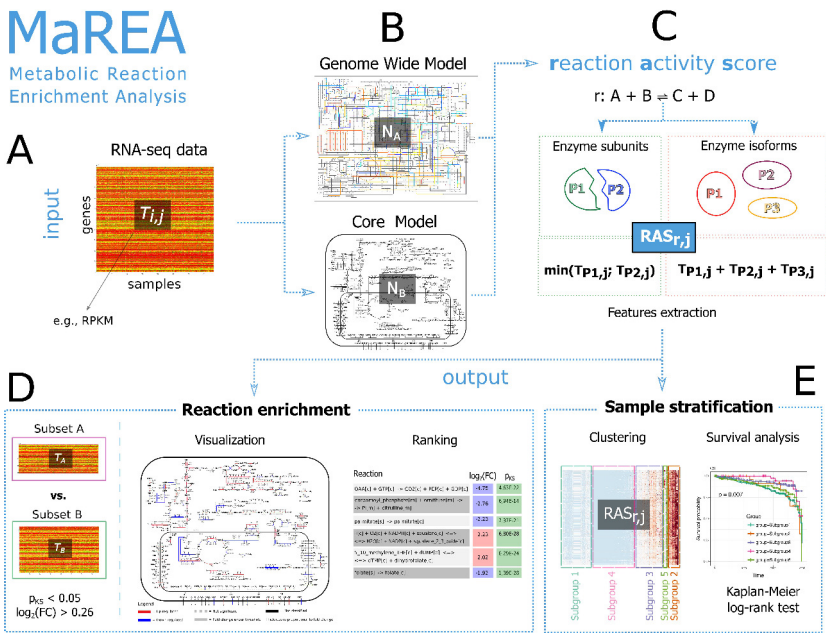


Figure 4.12: **MaREA pipeline.** **(A)** MaREA takes as input a n sample $\times$ m genes matrix T which includes the normalized read count of each gene from a given cross-sectional RNA-seq dataset. **(B)** MaREA can use different input metabolic reaction networks, e.g., the genome-wide Model, or any subset of it. **(C)** A *Reaction Activity Score* (RAS) is defined for any reaction r in the input network and any sample, by distinguishing the case of reactions involving enzymes composed by different subunits (in this case the RAS is computed as the minimum of the transcript level of the genes encoding the subunits), and that of reaction catalyzed by different enzyme isoforms (in this case the RAS is computed as the sum of the transcript level of the genes encoding the isoforms). **(D)** Given two distinct subsets T_A and T_B of the original dataset, the RASs of a given reaction in the two cases are compared; if the p-value of the Kolmogorov-Smirnov test is significant (< 0.05 - default) and the $\log_2$ fold-change is larger than 0.263 (default), that reaction will be marked in the final graph as up- or down-regulated. Accordingly, a reaction ranking can be provided. **(E)** MaREA can stratify patients by employing the RAS as feature on standard clustering methods, e.g., k-means. Survival analyses, such as log-rank test on Kaplan-Meier curves, can finally provide a prognostic validation of the clusters (notice that the image is intended for explanatory purpose only and does not reproduce any real case study).

Moreover, the features extracted by MaREA can be used to stratify samples in an unsupervised manner (*Metabolic Feature Extraction*). Such stratification might provide relevant prognostic indications, as shown in the case studies in Section 4.2.1.
In summary, MaREA can be used to:

i) rank reactions according to their activity variation across different phenotypes and/or experimental conditions.

ii) Enrich the map of human metabolic routes with the RAS variation, providing a clear and user-friendly visualization of how deregulated paths are interconnected.

iii) Efficiently stratify samples according to their metabolic activity, hence providing a new unsupervised clustering tool, with testable clinical relevance, which can be assessed, e.g., via standard survival analyses.

In order to test our approach, we applied MaREA to the investigation of cancer metabolic heterogeneity. The heterogeneity of cancer genotypes and phenotypes hinders the identification of targets for effective treatments and is a major cause of tumor relapse [216]. Therefore, it is common practice to statistically compare the gene expression of patient cohorts, based on clinical observations and/or molecular features, in order to understand how the hallmarks of cancer can be (alternatively) achieved in terms of gene expression regulation. A particularly relevant hallmark for cancer treatment is the the *metabolic reprogramming* of cancer cells [17, 182].

In particular, we applied MaREA to two distinct publicly available datasets in the TCGA database [214]: (*i*) the TCGA-BRCA dataset on breast cancer [217], and (*ii*) the TCGA-LUAD dataset on lung adenocarcinoma [218]. In the analyses, we employed both the genome wide-model Recon 2.2 [211] and a manually curated subset of it, corresponding to the model of central carbon metabolism (HMRCore), previously used in [78, 219].

Because the TCGA-BRCA dataset includes both cancer and normal biopsies for a subset of the samples, we first used MaREA to show its effectiveness in identifying, rank and visualize the enriched metabolic reactions between normal and cancer samples. This allowed to reproduce well-known features of cancer deregulation and produce new experimental hypotheses. Finally, we used MaREA to stratify cancer patients in distinct metabolic clusters with respect to both the TCGA-BRCA and the TCGA-LUAD datasets. Standard survival analysis highlighted statistically significant prognostic predictions for the identified clusters.

MaREA is freely available as a user-friendly MATLAB tool, which allows to process input transcriptomic data, e.g., RNA-seq, in various file formats, and input metabolic networks in COBRA-compliant file format [160], e.g., SBML (see "Availability" Section at the end of the paper).

Materials and Methods

Input

MaREA takes as input any RNA-seq dataset in the form of a $n \times m$ matrix T, where n is the number of genes and m is the number of samples of the considered cohort (see Figure 4.12-A). Each element $T_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ corresponds to the *normalized read count* of gene i in sample j such as, for instance, the *RPKM* (Reads per Kilobase per Million mapped reads).

MaREA then filters T according to a specific input reaction network N, e.g., the genome-wide metabolic network Recon 2.2 [211] or any possible subset of it. In particular, we define the set of reactions as $\mathcal{R} = \{r \in N\}$. Therefore, T is filtered by retaining only the rows corresponding to genes that are associated with enzymes involved in the reactions included in $\mathcal{R}$ (see Figure 4.12-B).

GPR logical formulas include *AND* and *OR* logical operators. AND rules are employed when distinct genes encode different *subunits* of the same enzyme, i.e., *all* the subunits are *necessary* for the reaction to occur. OR rules describe the scenario in which distinct genes encode *isoforms* of the same enzyme, i.e., either isoform is *sufficient* to catalyze the reaction.

For example, the succinate-Coenzyme A ligase enzyme is formed by the subunits alpha (gene SUCLG1) and beta gene (SUCLG2) and catalyzes the reaction $Pi + succinyl\text{-}CoA + GDP \leftrightarrow CoA + succinate + GTP$. The gene-enzyme rule for this reaction is therefore: SUCLG1 *AND* SUCLG2. Conversely, ACACA and ACACB are respectively fully functional enzyme for the reaction *acetyl-coenzyme A ligase carboxylase*, thus the rule is ACACA *OR* ACACB. Such logical operators can of course be combined to depict multi-protein catalytic complexes or more complex situations involving both subunits and isoforms. For instance, ribonucleotide reductase is formed by two subunits: the catalytic (M1) and the regulatory one. The latter exists in two isoforms (M2 and M2B). The rule for this enzyme will therefore be RRM1 *AND* (RRM2 *OR* RRM2B).

Reaction Activity Score (RAS)

To avoid the definition of arbitrary thresholds on the transcript level, we do not resolve the logical expressions in a Boolean fashion, but we employ a *Reaction*

Activity Score (RAS), for each sample $j = 1, \ldots, m$, and each reaction $r \in \mathcal{R}$ (see Figure 4.12-C). In particular, in order to compute the RAS we distinguish:

- Reactions with AND operator (i.e., enzyme subunits).

$$RAS_{r,j} = \min(T_{i,j} : i \in \mathcal{A}_r), \tag{4.8}$$

  where A_r is the set of genes that encode the subunits of the enzyme catalyzing reaction r.

- Reactions with OR operator (i.e., enzyme isoforms).

$$RAS_{r,j} = \sum_{i \in O_r} T_{i,j}, \tag{4.9}$$

  where $\mathcal{O}_r$ is the set of genes that encode isoforms of the enzyme that catalyzes reaction r.

In case of composite reactions, we respect the standard precedence of the two operators. Let $|\mathcal{R}|$ be the cardinality of the set of reactions, the final output is therefore a $|\mathcal{R}| \times m$ matrix M, where each element $M_{r,j}$ is the RAS computed for reaction r in sample j.

Similarly to what has been proposed in [220] to improve FBA predictions, the intuition underneath the introduction of the RAS is that enzyme isoforms contribute *additively* to the overall activity of a given reaction, whereas enzyme subunits *limit* its activity, by requiring all the components to be present for the reaction to occur.

Clearly, we are here adopting a deeply simplified approach to reaction network modeling, by neglecting, for instance, the great heterogeneity of reaction kinetic constants and protein binding affinities, of translation rates, and any possible post-transcriptional regulation effect that might occur within a cell. An optimal choice would be to weigh all the reactions according to such quantities, yet direct measurements or robust estimates are very rarely available, especially for genome-wide models. Therefore, at first approximation, we here assume that all enzyme isoforms and subunits contribute uniformly to the reaction activity of a given reaction, as we expect that this choice does not affect the up-/down-regulation interplay observed at the network level.

Reaction Enrichment: Visualization and Ranking

One important feature of MaREA is the ability to identify and visualize in an explicit way the metabolic routes that are up- or down-regulated in different sample sets and/or experimental conditions (see Figure 4.12-D).

Given two distinct RNA-seq datasets, or two partitions of the same dataset, T_A and T_B, and an input metabolic reaction network N, we first compute the RAS matrices M_A and M_B. For each reaction $r \in N$ we then perform a non-parametric two-sample *Kolmogorov-Smirnov* (KS) test with a default p-value threshold equal to 0.05, to verify whether the distributions of RASs over the samples in the two sets are significantly different.

In that case, we compute the log_2 fold-change of the average RAS_r in the two groups. Because KS-test considers as significantly different distributions with the same mean, but different standard deviation, by default we consider as relevant only log_2 fold-change larger than $\log_2(1.2) = 0.263$ (i.e., corresponding to a 20% variation of the average RAS). In line with the philosophy of GSEA [221], we use a relaxed threshold for the fold-change because even a difference of 20% in genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway. The significance threshold on the p-value of the Kolmogorov-Smirnov test, on the other hand, ensures that expression distributions of the two groups are indeed different. Notice that MaREA allows the user to choose different values for the RAS p-value and the fold-change threshold.

Next, MaREA uses the significant RAS fold-changes to: *i*) determine a ranking of the most relevant up- and down-regulated reactions in the two sets, *ii*) map such quantities over the input metabolic network N, by respectively coloring in red/blue the up-/down-regulated reactions, and by setting the edge thickness as proportional to the RAS fold-change. The reactions that will either display non-significant p-value (either due to identical distributions or to statistically insufficient sample size) or a RAS fold-change below the threshold will not be included in the ranking and will be marked in gray color on the metabolic network.

RAS-based Sample Stratification

Another advantage of our approach is that it is possible to employ the RAS as an effective feature to identify sample subgroups (or *clusters*) that share similar metabolic properties (see Figure 4.12-E). In particular, MaREA includes a k-means clustering [222], which uses the RASs of all reactions $r \in R$ (normalized on patients) to identify sample clusters with distinct metabolic behaviours. Clusters can be compared by means of the reaction enrichment procedure described above, by ranking the significantly different reactions in the distinct clusters and visualizing the RAS fold-changes on the input network.

When data is available, clusters can also be tested via standard *survival*

analysis, such as the *log-rank* test on *Kaplan-Meier* curves, hence providing an orthogonal validation of the clustering results with clinical relevance. We show how clustering on RASs indeed can produce significant prognostic predictions.

Metabolic Network

To compute the RAS of cancer samples in a given TCGA dataset at the genome-wide scale, we used the GPRs included in the most up-to-date genome-wide network of human metabolism: Recon 2.2 [211]. In particular, in order to visualize MaREA results at the genome-wide level, we modified the graphical attributes of the model map in *xml* format obtained from the *Virtual Metabolic Human* (VMH) – `https://vmh.uni.lu` – which is readable by the tool *Cell Designer* [223].

To focus, instead, on central carbon metabolism, we used the metabolic core model (HMRcore) introduced in [78]. For the sake of completeness, we included in the model mitochondrial palmitate degradation and gluconeogenesis. As the original version of the model does not include information on GPRs, such rules have been extracted and manually curated from Recon 2.2 [211] and included in the HMRcore model. In particular, we verified the correctness of GPR rules taking into account the information retrieved from the *Human Protein Atlas* [171] for the protein tissue location, from *UniProtKB* [224] for the enzyme complex composition, and from *KEGG* [68] to check gene/enzyme association. We checked for possible inconsistencies within Recon 2.2 and 29 GPRs were corrected. All the corrections made on Recon 2.2 GPR associations are reported in Supplementary Table S1. Notably, we corrected for the rules for Complex I-V of the respiratory chain, which were too strict in their original formulation. In particular, most patients in the BRCA dataset would have a null RAS for Complex IV, as gene COX7B2 is not expressed in most patients. It is important to notice that if we perform a FBA simulation with a null upper bound for Complex IV reaction, we do not obtain any optimal solution with the HMRcore model. The original rule considers indeed many genes that are actually isoforms and not subunits, thus requiring an OR and not an AND operator. In particular, in the original formulation COX7B does not allow to substitute for COX7B2.

The final version of the HMRcore model includes 264 reactions with a GPR rules and 405 metabolic genes that are associate to them. Genes are identified with the *HGNC ID* provided by the *HUGO Gene Nomenclature Committee* [225]. The SBML of the model is provided in Supplementary file S1. The tabular description of the model is provided in Supplementary file S2.

It should be noted that not every reaction in the metabolic models is asso-

ciated with a gene-enzyme rule: for instance some reactions have been included to fill the gaps in steady state computations, but we lack knowledge on the associate genes. In detail, 4742 (263) reactions over 7785 (314) are associated with a gene-enzyme rule, in the genome-wide (core) model. For such reactions it was possible to compute the RAS.

Datasets

We applied the MaREA pipeline to two TCGA cohorts.

- The breast cancer dataset (TCGA-BRCA) published in [217], which also includes healthy/control samples. We downloaded the dataset via the cBioPortal [226]. This dataset includes the expression profile (RNA Seq V2 RSEM) of biopsies taken from 817 patients. For 105 of them, the expression profile of the normal tissue is also included.

- The lung adenocarcinoma dataset (TCGA-LUAD, provisional) published in [218], downloaded via the cBioPortal [226]. The dataset includes the expression profile (RNA Seq V2 RSEM) of 586 biopsies from 584 patients. In our analyses we used the data of patients with only one sample.

Because the above datasets identify genes with Entrez IDs, we automatically converted them into *HGNC* IDs. We found a correspondence for 1654 (391) genes over the 1673 (404) included in the genome-wide (core) metabolic model. We opted to neglect missing genes in the computation of the RASs, as done in [202]. However, we were still able to compute a reliable RAS for most reactions associated with a GPR, as missing genes were involved in reactions with an AND operator only in three cases: namely, in the GPRs that involve mitochondrial genes (Complex I, III, IV and V), which are are not detected by RNAseq. The MaREA tool provides the user with an alternative option to handle these cases: rules with AND operator, involving missing genes, can be disregarded tout-court.

Other approaches

As specified in the Introduction (Section 4.2.1), metabolic networks are typically used in FBA simulations. In addition to the aforementioned problem regarding the scarce availability of both metabolic and transcriptomic data on the same patient, another challenge when dealing with FBA is the definition of an appropriate objective function.

Although maximization of metabolic growth may approximate well the objective of cancer cells [31], this assumption should not be extended to normal cells, nor generalized to other pathologies. Moreover, integration of transcriptomic data into FBA constraints often prevents the identification of a non-null solution [202, 212, 213] and requires the *ad hoc* release of some constraints. The impact of this problem is particularly evident when dealing with RNA-seq data, as they are more prone to include null values than transcriptomic data obtained otherwise, e.g., with microarrays, and thus to translate into null constraints. A major example is given by the gene that encodes for the transporter of water from mitochondria to cytosol (AQP8), which may often take value 0. Yet, this reaction is necessary to reach a steady state in which the respiratory chain is used.

Null constraints leading to infeasible solutions may also derive from errors in GPR association rules. Due to their large scope, curation errors are frequent in genome-wide metabolic networks. Genome-wide models are also known to be prone to the presence of thermodynamically infeasible cycles [51], hence they may predict unrealistic flux distributions.

Because of all these reasons, a modeling expert is thus required to guide the biologist among the multitude of current methodologies for transcriptomics integration in constraint-based models, as they are known to provide heterogenous predictions [202, 213].

MaREA was conceived to overcome such limitations, providing a simpler framework for trascriptomic data integration in metabolic networks, with relevant translational impact.

The spirit of MaREA is somehow in line with that of Gene Set Enrichment Analysis method (GSEA) [221], which seeks to characterize the sets of up- or down-regulated genes in different phenotypes [221]. The underlying rationale is that even a mild, but concerted, variation in the expression of a set of genes involved in a certain cellular (metabolic) function might be as relevant as a much larger variation in the individual activity of a single gene.

However, MaREA markedly differs from GSEA. The typical GSEA analysis outcome provides generic indications on the deregulated functions of a cell, or on specific functional behaviors when focusing on particular gene sets, derived, for instance, from the Reactome pathway database [227].

The GSEA enriched sets are list of genes involved in comprehensive, thus potentially quite large, metabolic pathways (as, e.g., "DNA replication"), failing to provide details on which specific metabolic routes are favored in a given condition. In particular, metabolic functions can be alternatively achieved by metabolizing different nutrients and/or by following different catabolic and ana-

bolic routes, in a complex and largely undeciphered interplay. For this reason, simply knowing weather a certain function is up- or down-regulated might not be sufficient to shed light on how such function might be achieved in distinct cancer phenotypes.

More recent approaches aim at building sets of genes to be enriched according to the information included in genome-wide metabolic networks. Specifically, *Metabolic Reporter Analyses* try to provide knowledge about variations in metabolite concentrations, starting from sets of genes that are classified according to the associated metabolites [228]. However, such methods do not provide information about which reactions are up- or down-regulated, and thus hinder the identification of putative targets for cancer treatment.

MaREA can provide a finer resolution to the analysis of metabolism than GSEA and reporter metabolites analyses, by enriching individual metabolic reactions in distinct experimental conditions. Moreover, MaREA improves over current methods based on genome-wide models, by providing a list of curated Gene-Protein-Reaction associations for human central carbon metabolism, and an easy-to-interpret map of corresponding central carbon pathways for visualization of results.

Similarly to the recently introduced PARADIGM approach (PAthway Recognition Algorithm using Data Integration on Genomic Models [229]), MaREA extracts a key metabolic feature (the RAS) for each sample. However PARADIGM relies on integrating data from multiple sources and requires curated pathway interactions among genes, hence both the input data and final objective are significantly different.

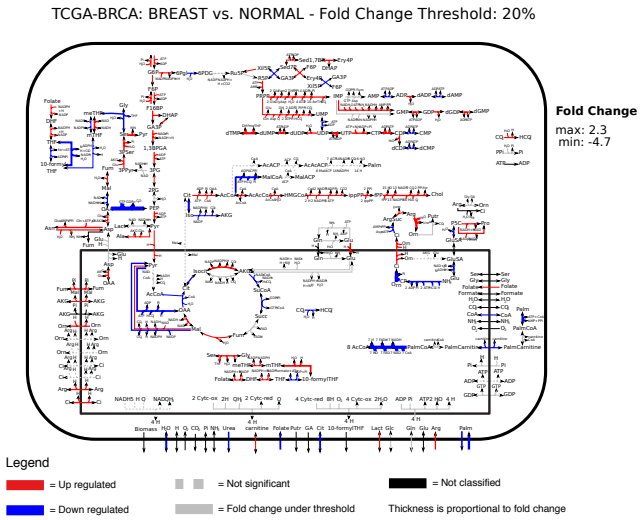
As specified above, MaREA can be used to stratify samples in an unsupervised manner. Distinct approaches make use, for instance, of the information on enriched signaling pathways [230] or of that on mutational profiles [231, 232, 233, 234] to classify cancer samples and subtypes.

Results

Breast Cancer vs. Normal (TCGA-BRCA)

In order to evaluate the overall usefulness of MaREA results, we first applied it to a well-known and characterized case-study, the comparison between cancer and normal metabolism. We performed two steps for this analysis: *Reaction Enrichment* and *Reaction Ranking*.

A



B

Reaction	Pathway	Gene Rule	Up-/ Down-Regulated	log2(FC)
3mb2coa[m] + gly[m] <=> CE2026[m] + coa[m] + h[m]	Glycine, serine, alanine and threonine metabolism	HGNC:13734	Down	-5,31
2 k[c] + 2 na1[e] + srtn[e] <=> 2 k[e] + 2 na1[c] + srtn[c]	Transport, extracellular (serotonine)	HGNC:11050	Up	4,95
glyc3p[c] + nad[c] <=> dhap[c] + h[c] + nadh[c]	Glycerophospholipid metabolism	HGNC:4455	Down	-4,85
gtp[c] + oaa[c] -> co2[c] + gdp[c] + pep[c]	Glycolysis/gluconeogenesis (PEP carboxykinase)	HGNC:8724	Down	-4,76
ala_D[e] + gln_L[c] <=> ala_D[c] + gln_L[e]	Transport, extracellular (alanine, serine, glycine, threonine)	HGNC:11026 and HGNC:11058	Down	-4,75
for[c] + nad[c] -> co2[c] + nadh[c]	Fatty acid oxidation	HGNC:3978	Down	-4,70
nad[c] + retinol_9_cis[c] <=> h[c] + nadh[c] + retinal_cis_9[c]	Vitamin A metabolism	HGNC:9940	Down	-4,40
lac_L[e] + na1[e] <=> lac_L[c] + na1[c]	Transport, extracellular (lactate, acetate,	HGNC:19119	Up	4,39
coumarin[c] + h[c] + nadph[c] + o2[c] -> h2o[c] + hcoumarin[c] + nadp[c]	Cytochrome metabolism	HGNC:2608 or HGNC:2610	Up	4,37
egme[r] + h2o[r] -> ecgon[r] + h[r] + meoh[r]	Alkaloid synthesis	HGNC:17	Down	-4,24

Figure 4.13: **Reaction enrichment and ranking: breast cancer vs. normal samples. (A)** HMRcore map enriched by MaREA: Reactions up-regulated in breast cancer sample set are marked in red, reactions up-regulated in normal sample set are marked in blue. A list of the abbreviations used in the map is provided in the Appendix A. Thickness of the edges is proportional to the fold-change. Non-Classified reactions, i.e., reactions without information about the corresponding gene-enzyme rule, are marked in black. Dashed gray arrows refer to non-significant deregulations according to the Kolmogorov-Smirnov test. Solid gray arrows refer to reactions with a log2 fold-change below 0.263. **(B)** A reaction ranking is provided, by listing the 10 reactions with largest log2 fold-change of the RAS (absolute value) in the two conditions. The reaction formula, the corresponding pathway, the gene rule, the up-/down-regulation flag and log2 fold-change are shown in the table.

Reaction Enrichment

The reactions that have been identified by MaREA as significantly up- or down-regulated in cancer - with at least a 20% increase/decrease - and the magnitude of the deregulation are mapped on the central carbon metabolic network (HMRcore) in Figure 4.13, as well as on the genome-wide metabolic network in Supplementary Figure S1. It can be observed (Figure 4.13) that the pathway of glycolysis is over-expressed in cancer. Extensive utilization of glucose is indeed a well established trait of breast and of cancer cells in general [235]. Cancer cells need glucose to feed the metabolic requirement of enhanced proliferation, with particular emphasis on (1) *de novo* synthesis of nucleotides for genome replication; (2) synthesis of amino acids for protein synthesis; (3) synthesis of fatty acids to support the expansion of cellular membranes; (4) ATP generation for energetic requirements.

Accordingly, MaREA returned the following metabolic "modules" as largely up-regulated in cancer: (1) synthesis of nucleotides from Phosphoribosyl-pyrophosphate (PRPP in Figure 4.13); (2) metabolism of the non-essential aminoacids serine (Ser), glycine (Gly), alanine (Ala), asparagine (Asn), aspartate (Asp), arginine (Arg) and proline (Pro); (3) synthesis of cholestrol (Chol) from citrate (Cit).

As far as ATP production is concerned, the interpretation of the situation portrayed by MaREA is, as expected, less straightforward. Cancer cells are believed to rely more on fermentation of glucose to lactate rather than on oxidation of glucose in the mitochondria (inner box of the map in Figure 4.13), despite the presence of oxygen: a phenomenon well-known as the Warburg Effect [17, 182]. However, in contrast to Warburg's original hypothesis that damaged mitochondria are at the root of this phenomenon, the ability of mitochondria to carry out oxidative phosphorylation is not defective in most tumors [17]. In line with these studies, on the one hand lactate secretion seems to be up-regulated in cancer (reaction crossing the external box in Figure 4.13). On the other hand, the respiratory chain (represented by the 4 reactions at the bottom of the mitochondrial box, which are catalyzed by protein Complexes I, II, III and IV, plus the reaction that oxidize succinate to fumarate, while reducing ubiquinone to ubiquinol, catalyzed by Complex II) is not significantly down-regulated in this breast cancer dataset, whereas, complex V is slightly up-regulated.

Remarkably, the results shown in Figure 4.13 suggest that the working-mode of the TCA cycle may be abnormal in breast cancer, as highlighted in [31]. In particular, up regulation of NADPH-dependent isocitrate dehydrogenase, which catalyzes the reductive carboxylation of α-ketoglutarate (AKG) to isocitrate,

may be linked with the mutations often reported for these enzyme in breast and other cancer types [236]. It has been suggested that this enzyme may support reductive glutamine metabolism in cancer and a branched TCA cycle flux mode [31].

The agreement of the results in Figure 4.13 with the obvious traits of cancer metabolism supports the reliability of our approach, which might shed light on less established traits. Deregulations of breast cancer metabolism identified by the approach, which may be worth of note are, among others: (1) deregulation of beta-oxidation of palmitate; (2) upregulation of folate metabolism; (3) deregulation of Phosphoenolpyruvate carboxykinase, which converts oxaloacetate (OAA) into phosphoenolpyruvate (PEP).

Reaction Ranking

After filtering out the reactions whose activity does not differ between cancer and normal samples, MaREA allows to rank the remaining reactions according to the extent of their up- or down-regulation. Supplementary Table S3 reports the following bits of information for each reaction included in the genome-wide model: the log fold change, the reaction formula, the pathway in which the reaction is involved and a description of its role. As genome-wide models include several reactions that are associated with the very same GPRs, typically involving transporters/enzymes with low substrate specificity, in Figure 4.13 we report the top 10 deregulated reactions with different GPRs.

These reactions include 3 up-regulated reactions and 7 down-regulated ones. Most of them are associated with a single gene, including, consistently with the results obtained for the HMRcore model: fatty acids oxidation and PEP carboxykinase, which are significantly down-regulated; lactate (or substrates pertaining to the same family) transport, which is up-regulated in cancer. Single-gene top deregulated reactions not included in the HMRcore model relate to deregulated vitamine A, glycine and alkaloid metabolism and to up-regulated transport of serotonine. Notably, in accordance with this results, it has been reported [237] that serotonine promotes tumor growth and survival in breast cancer, and that vitamin A [238] plays a role in cancer treatment and prevention.

Two top deregulated reactions are associated with a pair of genes linked by an OR and AND respectively: (1) down-regulated antiporter of the aminoacids alanine, serine, glycine and threonine with glutamine; (2) up-regulated Cytochrome P450 2A6, which is involved in the metabolism of many xenobiotics.

The down-regulation observed for the former reaction is in line with recent

studies that have linked the resistance of specific cancer cell lines to amino acid analogs anticancer drugs to a decreased expression of the corresponding transporter [239, 240].

The up-regulation identified for the latter reaction (P450 2A enzyme) is worth of note, as P450 enzymes may be involved in carcinogens activation in breast cancer. Environmental carcinogens have been identified in the etiology of breast cancer. For example, CYP2A6 protein detected in the breast can activate nitrosamines and food mutagens to their ultimate carcinogens and thus could play a role in the initiation of breast cancer [241]. Moreover, this enzyme can metabolize clinically important drugs, such as the tamoxifen [242], which represents the most widely used hormonal therapy for breast cancer, and the coumarin [243, 244], whose metabolism was proven to produce some metabolites having estrogenic and cytotoxic activities.

Comparison with GSEA Results

The GSEA and MaREA approaches are not directly comparable, as they present several differences in goals, input data, parameters, variables and outputs. For instance, MaREA computes an individual activity score for each sample, whereas GSEA only considers expression fold-changes between pairs of experimental conditions. In order to provide an overview of how the information produced by the two complementary approaches may differ, without any claim about which approach should be preferred, we disregarded the addition multiple test correction (FDR) used by GSEA, and we considered the gene-sets that pass the nominal p-value test, with the same threshold used in MaREA standard settings (i.e., $p = 0.05$). We did not set any threshold on the minimum size of gene-sets.

To run GSEA we used two kind of gene sets: (1) curated gene sets based on Reactome, as directly provided by the GSEA tool; (2) gene-sets reconstructed by us, which correspond to the genes associated with each reaction in the genome-wide model and which are provided in *gmt* format (GSEA compliant) in Supplementary File S2.

The first kind of gene-sets represent a typical application of GSEA to gene-sets involved in broad metabolic functions. The second kind is directly comparable to the sets used to compute the RAS by our approach. It should be mentioned that the second type includes many single gene-sets (i.e., size 1) because many reactions are catalyzed by enzymes associated with a single gene. For this reason, we set the minimum set size to 1 in the GSEA options.

The application of GSEA to Reactome gene-sets returned 144 gene sets significantly enriched in cancer, and 60 gene-sets significantly enriched in normal, at

nominal p-value < 0.05. When ranking the obtained gene-sets according to the returned Enrichment Score (ES), we observed that, as expected, the first 10 gene sets enriched in cancer refer to generic metabolic functions (in particular: cell cycle and mitosis, asparagine glycosylation, DNA replication and chromosome maintenance, HIV infection and kinesins). The results highlight how MaREA should be used as a complement to GSEA analysis, in order to provide a more fine-grained analysis of metabolic deregulations.

Conversely, the application of GSEA to the more fine-grained datasets, based on Recon 2.2 reactions, returned a number of reactions significantly deregulated much lower than that returned by MaREA. MaREA returned 3339 reactions as significantly up- or down-regulated by at least 20% (p-value < 0.05), whereas GSEA returned 105 gene sets significantly enriched in cancer, and 110 gene-sets significantly enriched in normal, at nominal p-value < 0.05. This discrepancy is mainly due to the presence of gene-sets including a single-gene, which are reasonably penalized by the GSEA approach. For instance, the single-gene sets associated with serotonin and vitamin A metabolism, which were ranked in the top 10 deregulated reactions by MaREA and might play a role in cancer according to literature, do not pass the nominal p-value test in GSEA. It is worth noticing that also the top-ranked reactions in MaREA that involve genes in OR (Cytochrome metabolism) or in AND (extracellular transport of alanine, serine, glycine and threonine) do not pass the significance test in GSEA.

Taken together, these results indicate that MaREA provides a more complete and refined portray of metabolic deregulations. Moreover, as opposed to GSEA, MaREA computes an independent score for each sample (the RAS), which can be used to cluster samples in an unsupervised fashion. We illustrate such application of MaREA in the next section.

Sample Stratification via MaREA

MaREA can also be used to stratify samples into distinct metabolic subgroups or clusters, also when the presence and number of such clusters is unknown. To this end, in order to provide an experimental validation of the stratification results, we also employed other known measures, such as, e.g., the *survival probability* of patients.

To provide an example application and compare different clustering features, we performed unsupervised clustering on two distinct cancer datasets: (1) the TCGA-BRCA dataset on breast cancer and (2) the TCGA-LUAD dataset on lung adenocarcinoma (see Section 4.2.1, Datasets, for details).

In particular, we performed a k-means unsupervised clustering over two distinct metabolic networks as input – (1) HMRCore and (2) Recon 2.2 – with different inputs $k \in \{1, 2, \ldots, 9\}$. We repeated the clustering using three different distance measures, respectively based on: *(a)* normalized RAS[1], *(b)* RNA-seq data of all metabolic genes, *(c)* the predicted fluxes.

To predict fluxes via FBA computation, we used the E-Flux method [245], which uses the RAS to set constraints on the flux boundaries, and we optimized for growth. The RAS was first normalized across patients as proposed in [246]. We allowed each extracellular nutrient considered in the baseline version of Recon 2.2 to be uptaken/secreted (upper and lower bound set to -1 and 1 respectively, as in [245]).

For each case, we performed $n = 100$ bootstrap iterations, with random centroid assignments, selecting as optimal the clustering run displaying the maximum inter-cluster distance. We then tested the resulting sample clusters against the survival probability (as retrieved from clinical data in the original datasets [217, 218]), via a log-rank test on the Kaplan-Meier curves with respect to *overall survival* (OS), *disease-free interval* (DFI), and *disease-specific survival* (DSS).

Metabolic Subgroups of Breast Cancer (TCGA-BRCA)

By applying MaREA to the TCGA-BRCA dataset, we found statistically significant differences ($p < 0.05$) in the Kaplan-Meier curves in the following cases:

($i - a$) HMRcore using RAS and $k = 2$, for OS, DFI, and DSS.

($ii - a$) Recon 2.2 using RAS and $k = 2$ for DFI.

($i - b$) HMRcore using RNA-seq of metabolic genes

and $k = 2$ for DFI, and DSS; and $k = 9$ for OS, DFI, DSS.

All the other cases displayed not significant differences in survival expectancy ($p > 0.05$), including those based on clusters identified on fluxes.

In Figure 4.14 we show the most significative result, obtained with ($i - a$) HMRcore with RAS with $k = 2$ (Subgroup 1: 630/817 samples, Subgroup 2: 187/817 samples). One can see that the curves of the two clusters never overlap, leading to a significant log-rank test ($p = 0.046$).

[1] To avoid possible biases due the differences in RAS range and distribution across reactions, we normalizes the RAS value of each sample by dividing by the maximum RAS for that reaction.

This result indicates that the up-/down-regulation patterns, as encoded by the RAS values, might indeed be used to split samples in metabolic groups with significantly different prognosis. It is also worth noticing that, by looking at the composition of the two subgroups with respect to the well-established PAM 50 classification [247] (available for 481 on 817 samples), the subgroup with the worst prognosis (Subgroup 2) is largely constituted (i.e., $\sim 70\%$) by samples belonging to the Basal-like group, which are present in a very small percentage in Subgroup 1 (105 Basal-like samples in 107, i.e., the $\sim 98\%$, belong to Subgroup 2), and show almost no samples from Luminal A subgroup. Subgroup 1, instead, is dominated by Luminal A (200 Luminal A samples in 201, i.e., the $\sim 99.5\%$, belong to Subgroup 1) followed by Luminal B and Her 2 subtypes in different proportions.

This result first suggests that there exists a detectable metabolic *signature* of Basal-like cancer samples, and this is, to the best of our knowledge, a novel result, worth of further investigations. Besides, it is worth noticing that the differences observed at the metabolic level indeed translate into distinct survival probabilities, which standard classification may fail to capture.

More in detail, by looking at the reactions significantly up-/down-regulated with respect to the case Subgroup 1 (best prognosis) vs. Subgroup 2 (worst prognosis) portrayed in Figure 4.14 for the core model – and in Supplementary Figure S2 and Table S4 for the genome-wide model – one can see that many reactions that are enriched in cancer against normal are also enriched in worst against best prognosis, including glycolysis, nucleotide synthesis and serine metabolism. Remarkably some metabolic pathways that are not significantly deregulated in cancer are significantly up-regulated in the worse prognosis subgroup, with particular regard to palmitate biosynthesis.

Metabolic Subgroups of Lung adenocarcinoma (TCGA-LUAD)

In the case of lung adenocarcinoma, the results of sample stratification via MaREA are even more striking. In Figure 4.15 one can see the Kaplan-Meier overall survival curves and the corresponding p-value of the log-rank test, with respect to the stratification obtained in the 6 aforementioned scenarios:

$(i-a)$ HMRcore with RAS.

$(i-b)$ HMRcore with RNA-seq of metabolic genes.

$(i-c)$ HMRcore with fluxes.

$(ii-a)$ Recon 2.2 with RAS

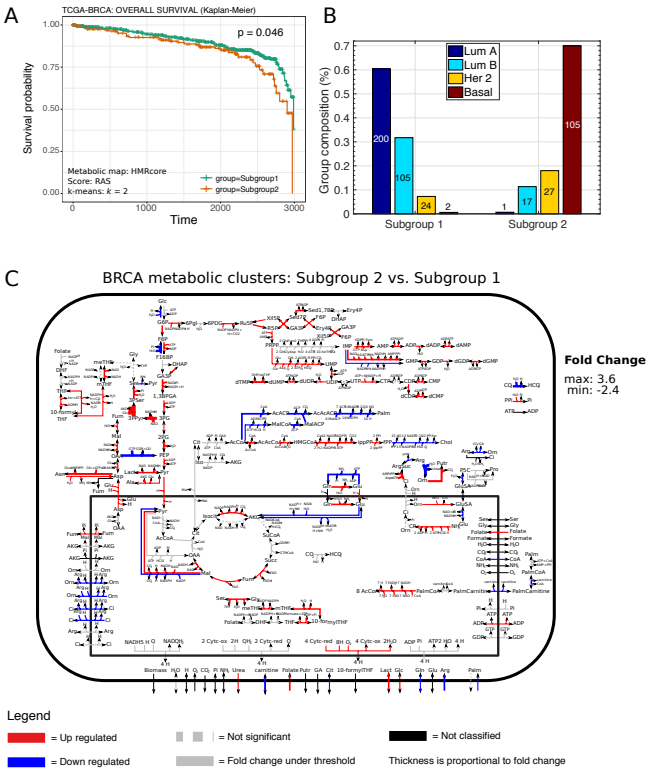


Figure 4.14: **Breast cancer metabolic clusters. (A)** The Kaplan-Meier curves (time unit = days) and the p-value of the log-rank test with respect to the two metabolic clusters of the TCGA-BRCA samples identified by MaREA. The result was obtained using a k-means unsupervised clustering ($k = 2$) with HMRcore and RAS. Subgroups 1 and 2 are shown). **(B)** Composition of the two TCGA-BRCA metabolic clusters identified by MaREA with respect to the PAM50 breast cancer classification. The number of samples in each group is displayed on the bars. **(C)** Enriched map of HMRcore with respect to the TCGA-BRCA metabolic clusters 1 and 2. A list of the abbreviations used in the map is provided in the Appendix A). Red arrows refer to reactions up-regulated in Subgroup 2 (worst), whereas blue arrows refer to reactions up-regulated in Subgroup 1 (best). Black arrows refer to "Non Classified" reactions, i.e., reactions without information about the corresponding gene-enzyme rule. Dashed gray arrows refer to non significant deregulations according to the Kolmogorov-Smirnov test. Solid gray arrows refer to reactions with a log2 fold change below 0.263.

$(ii-b)$ Recon 2.2 with RNA-seq of metabolic genes.

$(ii-c)$ Recon 2.2 with fluxes

with respect to $k \in \{1, 2, \ldots, 9\}$; scenarios with $p <= 0.01$ are marked in green, with $0.01 < p <= 0.05$ in yellow, with $p > 0.05$ are not shown.

In all cases, we were able to retrieve clusters with significantly different prognosis for at least some values of k, the best results being obtained in cases $(i-a)$ HMRcore with RAS, $(i-b)$ HMRcore with RNA-seq of metabolic genes, and $(ii-b)$ Recon 2.2 with RNA-seq of metabolic genes, which show highly significant p-values ($p << 0.01$) for most values of k (similar results are obtained for DFI and DSS curves - not shown here).

In this regard, we expect the stratifications with larger k to be effective in deciphering the inherent heterogeneity of lung adenocarcinoma, whereas fewer and larger clusters might partially hide this effect. Notice that also in this case the clusters identified on fluxes display the worse predictive power on survival expectancy, at least for higher values of k.

Even though the stratification performances of RAS and RNA-seq data of metabolic genes are both remarkable and somehow comparable with respect to survival outcome, we here remark that only with the former approach it is possible to enrich and rank the reactions that distinguish such groups and which, accordingly, might be targeted by metabolic drugs.

For instance, in Figure 4.16 we show the metabolic map enriched via MaREA by comparing Subgroup 4 (best prognosis) and Subgroup 3 (worst), in the case of clustering obtained with HMRcore and RAS with $k = 4$ (such stratification provides a good trade-off between a sufficiently large k and sufficiently large clusters).

Finally, it is worth noting that some key up-/down-regulation patterns observed when comparing worst versus best prognosis BRCA clusters are conserved in the LUAD case, such as up-regulation of glycolysis, nucleotide synthesis and amino acid metabolism. This result would suggest that key regularities of cancer metabolic deregulation, directly linked with metabolic growth (i.e., DNA and protein synthesis), might be shared across distinct cancer types, and deserves further investigations.

Conclusions

We have here introduced MaREA, a computational pipeline that integrates transcriptomic data into metabolic networks, to compare the metabolism of samples in distinct subgroups or different experimental conditions.

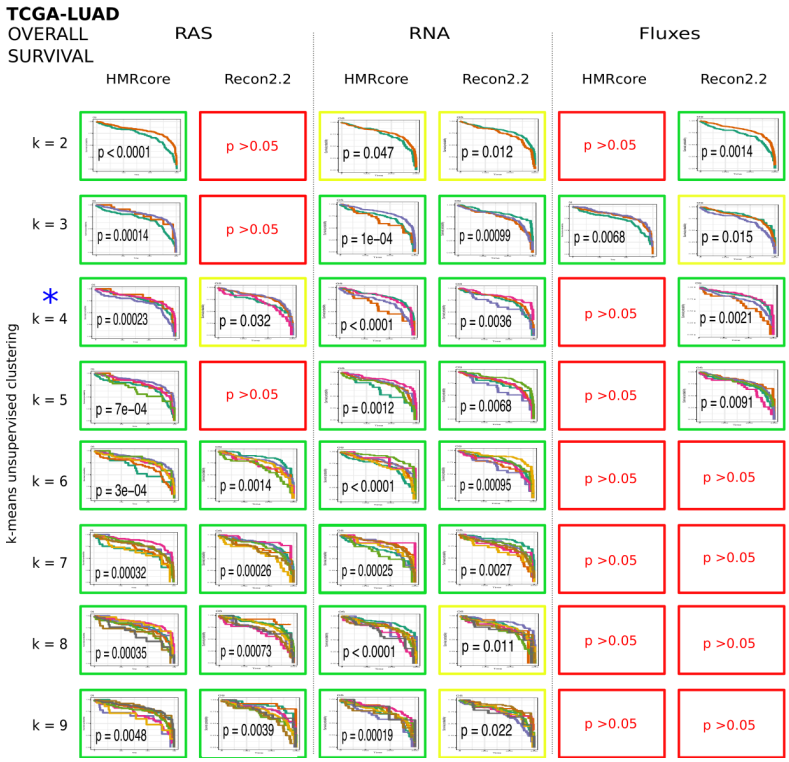


Figure 4.15: **Survival analysis of lung adenocarcinoma metabolic clusters.** The Kaplan-Meier overall survival (OS) curves displaying a p-value of the log-rank test: $p \leq 0.01$ (green) or $0.01 < p \leq 0.05$ (yellow) are shown, with respect to the k-means unsupervised clustering performed on the TCGA-LUAD dataset on breast cancer [218]. 6 cases are considered (from left to right): $(i-a)$ HMRcore with RAS, $(ii-a)$ Recon 2.2 with RAS, $(i-b)$ HMRcore with RNA-seq of metabolic genes, $(ii-b)$ Recon 2.2 with RNA-seq of metabolic genes, $(i-c)$ HMRcore with fluxes, $(ii-c)$ Recon 2.2 with fluxes, for $k \in \{1, 2, \ldots, 9\}$ (time unit = days). The blue star marks the case used to enrich the map in Figure 4.16, i.e., $k = 4$, HMRcore with RAS.

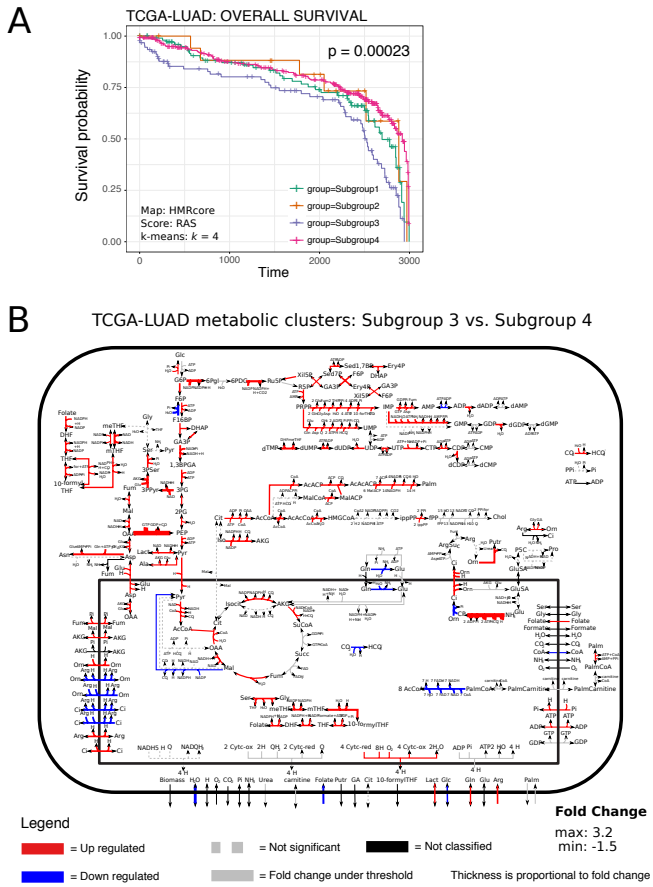


Figure 4.16: **Lung adenocarcinoma metabolic clusters. (A)** The Kaplan-Meier curves (time unit = days) and the p-value of the log-rank test with respect to the two metabolic clusters of TCGA-LUAD samples identified by MaREA with k-means unsupervised clustering ($k = 4$) using HMRcore and RAS are shown. **(B)** Enriched HMRcore map with respect to metabolic clusters 4 (best prognosis) and 3 (worst prognosis), identified by k-means unsupervised clustering ($k = 4$) with HMRcore map and RAS. The list of the abbreviations used in the map is provided in Appendix A). Red arrows refer to reactions up-regulated in Subgroup 3 (worst), whereas blue arrows refer to reactions upregulated in Subgroup 4 (best) Black arrows refer to "Not Classified" reactions, i.e., reactions without information about the corresponding gene-enzyme rule. Dashed gray arrows refer to non significant deregulations according Kolmogorov-Smirnov test. Solid gray arrows refer to reactions with a log2 fold change below 0.263.

In particular, the reaction enrichment performed via MaREA allows to identify the metabolic patterns underlying the phenotypic and functional properties observed in different sample subgroups, as in the case of (but not limited to) patients with distinct cancer subtypes. This is an important advantage of MaREA, especially when the estimation of metabolic fluxes is not possible or is scarcely reliable. The interpretation of the results is then favored thanks to the effective visualization of up-/down-regulated reactions directly on the metabolic networks.

Furthermore, MaREA can effectively stratify samples in clusters with similar metabolic activity, in unsupervised fashion. Its prognostic power can be evaluated via standard survival analysis.

The case studies on TCGA cancer datasets proved that MaREA can reproduce known properties and traits of metabolic networks in different scenarios. For instance, MaREA allowed to identify the key metabolic paths that distinguish normal from tumor samples, but it also provided cues to formulate and test new experimental hypotheses. Moreover, the metabolic clusters of samples identified by MaREA displayed significantly different survival expectancy, as retrieved from clinical data, for both breast and lung cancer, and this proved that metabolic reprogramming plays a crucial role in cancer aggressiveness.

In the Lung Adenocarcinoma case study (TCGA-LUAD dataset), up to 9 sample subgroups with significantly different prognosis have been identified, by exclusively taking into account transcriptional regulation of metabolic reactions. In the breast cancer case study (TCGA-BRCA dataset), the metabolic differences in terms of patient prognosis were less relevant, suggesting that metabolic heterogeneity might play a milder role in breast cancer as compared to lung cancer. However, MaREA was able to identify two BRCA metabolic clusters characterized by significantly different prognosis and that show a striking overlap with Luminal A and Basal-like standard signatures.

When comparing distinct clustering features, RASs and transcripts proved to be more effective than fluxes in identifying sample subgroups with different prognosis, probably due to the lack of proper constraints on extracellular fluxes in FBA computation.

Besides, clustering based on the core model displayed a better prognostic power than that based on the genome-wide model, at least when employing the RAS, despite a dramatic reduction of the dataset dimensionality. This result highlights the importance of the curation process of the GPR associations that was performed for the core model, and which could be further refined.

We also recall that a major benefit of using the RAS, rather than only the transcripts, lays in its effectiveness in summarizing metabolic deregulation

when comparing two cohorts or experimental settings. In fact, MaREA allows to identify, rank and visualize critical metabolic reactions, which might more easily be targeted, e.g., with metabolic drugs, rather than targeting the expression of individual genes.

However, as MaREA does not provide information on metabolic fluxes, we believe that MaREA results should be complemented with metabolic measurements and flux simulations, when possible, to provide an all-encompassing picture of metabolism and its deregulation.

Availability of data and material

MaREA is provided as Matlab tool and is freely available at this link: https://github.com/BIMIB-DISCo/MaREA, along with the source code and the input models. The datasets to reproduce the case studies presented in this paper can be downloaded at this link. The TCGA-BRCA and TCGA-LUAD datasets used in this work are available for download via the cBioPortal for Cancer Genomics http://www.cbioportal.org/.

Supplementary Material

- Supplementary Figure S1 — Recon 2.2 metabolic map, enriched via MaREA: case breast cancer samples vs. normal samples
- Supplementary Figure S2 —Recon 2.2 metabolic map, enriched via MaREA: case best vs. worst survival outcome subgroups, with respect to breast cancer dataset, as clustered via k-means with k=3
- Supplementary Table S1 — Table (Excel) including the annotated list of corrections made to Recon 2.2 GPR associations
- Supplementary Table S2 — Tabular description (Excel) of the curated HMRcore model
- Supplementary Table S3 — Excel file including the RAS Fold-Change and p-values with respect to the comparison between breast cancer samples vs. normal samples
- Supplementary Table S4 — Excel file including the RAS Fold-Change and p-values with respect to the comparison between the worst vs. best survival outcome breast cancer subgrups, as clustered via k-means with k=2

- Supplementary File S1 — HMRcore model in SBML format
- Supplementary File S2 — gmt file with gene-reaction association formatted to be usable in GSEA analysis

4.2.2 Integration of single-cell RNA-seq data into population models to characterize cancer metabolism

Chiara Damiani, Davide Maspero, **Marzia Di Filippo**, Riccardo Colombo, Dario Pescini, Alex Graudenzi, Hans Victor Westerhoff, Lilia Alberghina, Marco Vanoni, Giancarlo Mauri

Under revision from *PLoS Comp Biol.*

DOI: 10.1101/373621

Abstract

Metabolic reprogramming is a general feature of cancer cells. Regrettably, the comprehensive quantification of metabolites in biological specimens does not promptly translate into knowledge on the utilization of metabolic pathways. By estimating fluxes across metabolic pathways, computational models hold the promise to bridge this gap between data and biological functionality. These models currently portray the average behavior of cell populations however, masking the inherent heterogeneity that is part and parcel of tumorigenesis as much as drug resistance.

To remove this limitation, we propose single-cell Flux Balance Analysis (scFBA) as a computational framework to translate single-cell transcriptomes into single-cell fluxomes.

We show that the integration of single-cell RNA-seq profiles of cells derived from lung adenocarcinoma and breast cancer patients into a multi-scale stoichiometric model of a cancer cell population: significantly 1) reduces the space of feasible single-cell fluxomes; 2) allows to identify clusters of cells with different growth rates within the population; 3) points out the possible metabolic interactions among cells via exchange of metabolites.

The scFBA suite of MATLAB functions is available at https://github.com/BIMIB-DISCo/scFBA, as well as the case study datasets.

Author Summary

Cytotoxicity of chemotherapeutic agents and resistance to targeted treatments are the main reasons why cancer is still one of the top causes of death. As tumor cells are intrinsically resistant to therapies that target signaling pathways, targeting the metabolic hallmarks of cancer holds promise for more incisive treatments. Regrettably, the heterogeneity of cancer metabolism hinders the identification of effective treatments. To fully uncover the metabolic heterogeneity within tumors, characterization of metabolic programs (metabolic flux distributions) at the single-cell level is required. To fill the gap between current technologies for genomics and future technologies for fluxomics, both at the single-cell and the genome-wide scale, we propose to integrate cancer data from: 1) single-cell transcriptomics and 2) bulk metabolomics, into a multi-scale stoichiometric model, to deliver for the first time metabolic fluxomes at the single-cell level. To this end, we introduce a new paradigm for flux balance analysis and data integration in cancer metabolism to: 1) characterize metabolic heterogeneity, not only at the inter-, but also at the intra-tumor level 2)

identify the metabolic interactions between cancer populations, whose role in resistance to metabolic treatments has been recently recognized 3) predict the collective response to drug targeting of metabolism.

Introduction

Cancer is a heterogeneous, multi-factorial and essentially genetic disease, in which various types of mutations alter the functioning and interactions of genes, causing cancer cells to proliferate in an uncontrolled manner. Despite the plethora of cancer-related mutations, a reduced number of recognizable phenotypic hallmarks [18, 19] have been identified.

Metabolic rewiring, in particular, is a general feature of cancer cells, which reprogram their metabolism to feed their unrestrained proliferation, as it requires high amounts of energy and building-blocks [248]. The design principles underlying the causative role of metabolism in promoting growth as a function of the nutritional constraints are starting to be investigated [31, 249] and the idea of targeting the distinctive features of cancer metabolism has received considerable attention [250].

Unfortunately, a single metabolic program cannot be used to globally define an altered tumour metabolism [17], as cancer cells, even within the same tumour, may cope with the above metabolic requirements by engaging different metabolic pathways [251]. Such variability produces different dependencies on exogenous nutrients, and reflects into heterogeneous responses to metabolic inhibitors [252].

Furthermore, in solid tumours, cancer cells are embedded within the tumour microenvironment (TME), a complex network of fibroblasts, myofibroblasts, myoepithelial cells, vascular endothelial cells, cells of the immune system and extracellular matrix. TME also includes chemical gradients of oxygen and nutrients: the complex interaction of all these elements plays a major role in tumour metabolic heterogeneity [26]. The metabolic interplay that occurs among cancer cells and TME - supported by experimental evidence on how malignant cells may extract high-energy metabolites (e.g., lactate and fatty acids) from adjacent cells [253, 254] - contributes to treatment resistance [255]. Therefore, effective therapeutic strategy should incorporate knowledge of intra-tumour metabolic heterogeneity and cooperation phenomena within cancer cell populations.

Knowledge about the utilization of metabolic pathways requires quantification of metabolic fluxes (i.e., the rate at which a substance is transformed into another through a given reaction or pathway). As quantification of the full complement of cell metabolites (metabolomics) alone does not provide information on internal fluxes [256], they can be measured only indirectly, mainly via

several isotope-labeled metabolomics experiments coupled with metabolic flux analysis. A more functional option is Flux Balance Analysis (FBA), which uses linear algebra algorithms to solve a mass balance problem, given constraints on the flux of some relevant reactions, with particular regard to extracellular fluxes (rate of intake and secretion of metabolites) [60]. An advantage of FBA is that, as opposed to intracellular fluxes, extracellular fluxes can also be approximated from measurements of the concentration of metabolites in the spent cell culture medium at different time points. FBA has proved able to correctly predict the growth yield of microorganisms when extracellular fluxes are constrained [84, 257, 258]. Several approaches have been proposed to set further constraints on internal fluxes, by exploiting other -omics data, such as transcriptomics or proteomics data. Protein abundances should be used cautiously as proxy for flux rates, as the availability of substrates ultimately determine the reaction rate. Transcript levels should be used even more prudently than protein abundances, because many factors beyond transcript concentration contribute to determine the expression level of a protein, such as translation and degradation rates, spatial locations and post-transcriptional regulation. However, the differences between the mRNA level observed in different conditions (at the steady state) explain most of the variation in concentration of the associated proteins [259, 260]. Therefore, coupling the information of transcripts with that on extracellular fluxes, within the FBA steady-state modeling framework, provides a reasonable solution to the problem of predicting intracellular fluxes.

Nevertheless, extracellular fluxes can be hardly measured at the single-cell level. They cannot be approximated from the spent cell culture medium, because it portrays the extracellular fluxes of the bulk. They might be measured with isotope-labeling and metabolic flux analysis techniques, but this would require single-cell metabolomics.

Unfortunately, single-cell metabolomics is still at its embryonic stage, mainly because of limitations in working with minute amounts of material [193, 261, 262]. This is a major problem, because flux distributions are currently estimated from metabolic measurements retrieved from bulk samples, which often contain intermixed and heterogeneous cell subpopulations, thus overlooking possible cooperation and compensation phenomena. In the simplest example, let us imagine that two populations exist: one that secrete lactate and another that consumes all the lactate produced by the former. The predicted flux distribution will represent the sum of the two populations, hence the flux through lactate production/consumption will be regarded as inexistent, portraying a behavior far from the real one. This limitation might partially be overcome by analyzing a subpopulation of cells supposedly having a more homogeneous

metabolism, for example, by using fluorescence-activated cell sorting techniques to isolate it according to specific physical properties. However, when dealing with populations of cells derived from tumours, it is difficult to assess the relative composition of intermingled cancer cells and of stromal elements within the tumor architecture, and fluorigenic markers may not robustly correlate with metabolic phenotype. Hence, stratification, in terms of metabolic function and of consequent susceptibility to metabolic drugs, of heterogeneous populations taken, for example, from biopsies, xenografts or organoids, requires characterization of cellular metabolism at the single-cell level.
At this purpose, we here introduce the first computational framework scFBA (single-cell Flux Balance Analysis) to predict single-cell fluxomes and possible metabolic interactions among them, starting from (bulk) extracellular fluxes and single-cell transcriptomes. As compared to single-cell metabolomics, the field of single-cell transcriptomics has indeed progressed to a deeper resolution [249, 263].

Current methods to integrate (bulk) transcriptomic data into constraint-based models mainly fall into two categories, as reviewed in [202, 212, 213, 264]: approaches that aim to determine the active metabolic network in order to extract context-specific metabolic networks from generic ones (e.g., [265, 140]), and methods that aim to improve the prediction of the context-specific metabolic flux distribution (e.g., [220, 245, 246, 266, 267, 268]). However, these methods cannot be directly extended to single-cell modeling, because of the lack of information of extracellular fluxes at the single-cell level discussed above. Even if it was possible to measure single-cell extracellular fluxes, there is no protocol to measure transcriptome and fluxes of the very same single-cell (the cell is destroyed after either analyzing the trascriptome or metabolome).

For this reason, we propose to use a multi-scale model, in order to solve a unique mass balance problem to identify the possible combination of single-cell steady states that concurrently satisfies constraints on single-cell transcriptomes and extracellular bulk fluxes. In order to determine the flux distribution of each network, each cell is not considered is isolation: it is allowed to interact with other cells in the population, via release/uptake of metabolites into/from the TME.

Datasets obtained from cancer biopsies or patient-derived xenografts are the ideal candidate to capture the heterogenous composition of tumors. As currently no information on scRNA-seq and extracellular fluxes on the very same sample is publicly available, as a proof of principle, we here applied scFBA to lung adenocarcinoma (LUAD) patient derive xenograft scRNA-seq, collected by [269], while testing different sets of constraints for the extracellular fluxes. To

prove the robustness and applicability of our method, we also applied scFBA on further independent breast cancer datasets collected by [270].

Methods

Approach

Although some attempts to study the cooperation between different metabolic populations have been put forward [271, 272, 273], mostly focused on microbial communities, these methods require indeed *a priori* knowledge about the specific metabolic requirements and objectives of the intermixed populations. Unfortunately, even though metabolic growth may approximate the metabolic function of some cell populations, we cannot assume that each cell within an *in vivo* cancer population proliferates at the same rate, nor that it proliferates at all. A major example is given by the different proliferation rates of stem and differentiated cells [274]. For this reason, differently from other approaches [273], we do not impose that the population dynamics is at steady-state (and hence that cells all grow at the same rate), although we do continue to assume that the metabolism of each cell is. Conversely, scFBA aims at portraying a snapshot of the single-cell (steady-state) metabolic phenotypes within an (evolving) cell population at a given moment, and at identifying metabolic subpopulations, without *a priori* knowledge, by relying on unsupervised integration of scRNA-seq data.

We have previously shown how Flux Balance Analysis of a population of metabolic networks (popFBA)[219] can in line of principle capture the interactions between heterogenous individual metabolic flux distributions that are consistent with an expected average metabolic behavior at the population level [219]. However, the average flux distribution of a heterogenous population can result from a large number of combinations of individual ones, hence the solution to the problem of identifying the actual population composition is undetermined. To reduce this number as much as possible, we here propose to exploit the information on single-cell transcriptomes, derived from single-cell RNA sequencing (scRNA-seq), to add constraints on the single-cell fluxes.

An identical copy of the stoichiometry of the metabolic network of the pathways involved in cancer metabolism is first considered for each single-cell in the bulk. To set constraints on the fluxes of the individual networks, represented by the single-cell compartments of the multi-scale model, we took inspiration from bulk data integration approaches that aim to improve metabolic flux predictions, without creating context-specific models from generic

ones [220, 245, 246, 266, 267, 268]. At the implementation level, we use continuous data, rather than discrete levels, to overcome the problem of selecting arbitrary cutoff thresholds. At this purpose, some methods (e.g. [140, 265]) use expression data to identify a flux distribution that maximizes the flux through highly expressed reactions, while minimizing the flux through poorly expressed reactions. To limit the problem of returning a flux distribution (or a content-specific model) that does not allow to achieve sustained metabolic growth, we use instead the "pipe capacity" philosophy embraced by other methods, such as the E-Flux method [245, 246], of setting the flux boundaries as a function of the expression state. These methods tend to use relative rather than absolute expression values. For instance, the original formulation of E-flux [245] sets relative boundaries in relation to the most expressed reactions. In order to avoid comparing enzymes with different gene-protein translation rates, which may also largely differ in their kinetic parameters (e.g. binding affinity) and in the number of associated isoforms/subunits, we prefer to normalize boundaries in relation to the condition/cell/tissue in which a given reaction is mostly expressed, as done in a more recent version of the E-flux method [246] and in other continuous methods [266, 267, 275]. Specifically, we distribute the total (bulk) possible flux of each reaction proportionally to the activity score of that reaction in each cell.

To compute a score for reactions that involve many genes, similarly to other approaches [220, 245, 246, 276], we assume that enzyme isoforms contribute additively to the overall activity of a given reaction, whereas enzyme subunits limit its activity, by requiring all the components to be present for the reaction to occur. Alternative but similar approaches, such as [265], consider the maximum value (instead of the sum) of isoform values.

A scheme of the scFBA approach is depicted in Figure 4.17.

popFBA

Here we briefly recall the popFBA approach. For a more comprehensive description, the reader is referred to [219].

Starting from a template metabolic network map A, corresponding to a generic single-cell and defined as $A = (\mathcal{X}^A, \mathcal{R}^A, \mathcal{E}^A)$ - where $\mathcal{X}^A = \{X_1, \ldots, X_M\}$ is the set of metabolites in network A , $\mathcal{R}^A = \{R_1, \ldots, R_N\}$ the set of biochemical reactions taking place among them and $\mathcal{E}^A = \{E_1, \ldots, E_{N_{\text{ext}}}\}$ is a set of N_{ext} unbalanced reactions (exchange reactions), enabling a predefined set of metabolites (including the pseudo-metabolite representing biomass) $\mathcal{Y} = \{Y_1, \ldots, Y_{N_{\text{ext}}}\} \subset \mathcal{X}^A$ to be inserted in or removed from the system - the popFBA procedure first

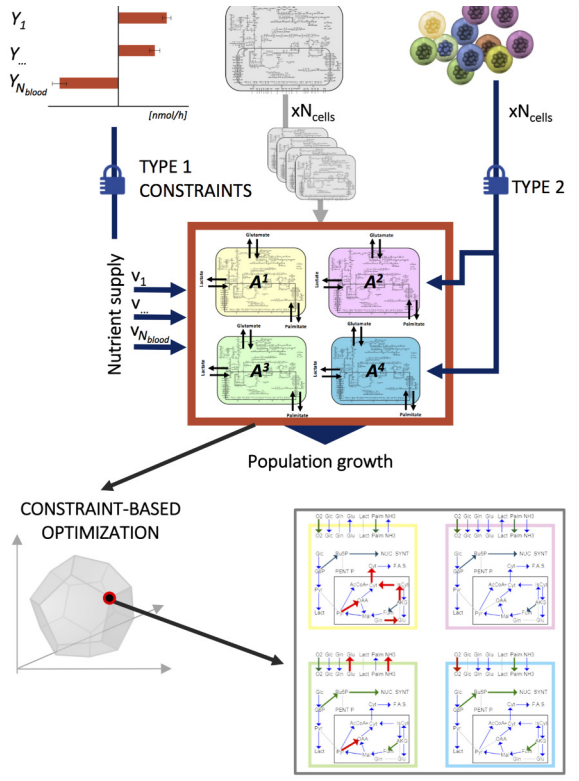


Figure 4.17: Graphical representation of scFBA. Extracellular fluxes and sc-transcriptomes are translated respectively into type 1 and 2 heterogenous constraints (see Materials and Methods) imposed on a initially homogenous population of N_{cells} replicates of metabolic network A. The output is a heterogenous set of flux patterns that may predict sc-fluxes.

builds a population model composed of N_{cells} replicates A^c of network A, each one corresponding to a single-cell c, $c = 1, \ldots, N_{\text{cells}}$, and which can cooperate by exchanging nutrients in the tumour microenvironment.

For each single-cell c, $A^c = (\mathcal{X}^c, \mathcal{R}^c, \mathcal{C}^c)$ is its metabolic network, where:

- $\mathcal{X}^c \equiv \mathcal{X}^A$ is the set of its metabolites;
- $\mathcal{R}^c \equiv \mathcal{R}^A$ is the set of its internal reactions;
- $\mathcal{C}^c = \{C_j^c\}$, with $j = 1, \ldots, N_{\text{ext}}$, is a set of cooperation reactions, defined as reactions that allow to exchange metabolites among single-cells via a shared environment that represents the TME compartment. Cooperation reactions are built by transforming each exchange reaction $E_j \in \mathcal{E}^A$ into a cooperation reaction C_j^c with the form:

$$C_j^c : Y_j^c \longleftrightarrow Y_j' \tag{4.10}$$

Accordingly, a new set of metabolites pertaining to the TME compartment $\mathcal{Y}' = \{Y_i'\}$ with $i = 1, \ldots, N_{\text{ext}}$ must also be defined.

Because original exchange reactions have been replaced by cooperation reactions, a new set of N_{blood} exchange reactions $\mathcal{B} = \{B_1, \ldots, B_{N_{\text{blood}}}\}$ is defined, which allows a subset of metabolites $\mathcal{K} = \{K_1, \ldots, K_{N_{\text{blood}}}\} \subset \mathcal{Y}'$ to be exchanged with the external environment, e.g., the blood supply:

$$B_j : K_j \longleftrightarrow \emptyset \tag{4.11}$$

The population model P is then defined by: (i) the union set of the metabolites $\mathcal{X}^P = \bigcup_c \mathcal{X}^c \cup \mathcal{Y}'$; (ii) the internal reactions $\mathcal{R}^P = \bigcup_c \mathcal{R}^c$; (iii) the cooperation reaction s $\mathcal{C}^P = \bigcup_c \mathcal{C}^c$; (iv) the population exchange reactions $\mathcal{B}$.

A stochiometric matrix S^P is then built for all reactions in $\mathcal{R}^P$, $\mathcal{C}^P$ and $\mathcal{B}$ and for all metabolites in $\mathcal{X}^P$ and $\mathcal{Y}'$. The final size of matrix S^P is $(N_{\text{cells}} \cdot M + N_{\text{blood}}) \times (N_{\text{cells}} \cdot (N + N_{\text{ext}}) + N_{\text{blood}})$. Once the population model is obtained, the total biomass of the N_{cells} single-cells is maximised by means of linear programming, as in standard FBA [60].

The solution of popFBA represents the flux distribution $\vec{v} = (v_1, \ldots, v_{N_{\text{cells}} \cdot (N+N_{\text{ext}})+N_{\text{blood}}}) = (r_1^1, \ldots, r_N^1, \ldots, r_1^{N_{\text{cells}}}, \ldots, r_N^{N_{\text{cells}}}, c_1^1, \ldots, c_{N_{\text{ext}}}^1, \ldots, c_1^{N_{\text{cells}}}, \ldots, c_{N_{\text{ext}}}^{N_{\text{cells}}}, b_1, \ldots, b_{N_{\text{blood}}})$ that maximises the biomass exchange flux $b_{biomass}$, with v_i representing any flux i of the population model, and for each single-cell c, r_i^c representing the i-th internal flux, c_i^c representing the i-th cooperation flux and b_i an exchange flux

with blood. The optimization problem is postulated as follows:

$$\begin{aligned} &\text{maximise } b_{biomass} \\ &\text{subject to } S\vec{v} = \vec{0}, \vec{v}_L \leq \vec{v} \leq \vec{v}_U \end{aligned} \tag{4.12}$$

$\vec{v}_L$ and $\vec{v}_U$ are vectors specifying the lower and upper bound respectively for each flux v_i of $\vec{v}$. A negative lower bound indicates that flux is allowed in the backward reaction. To solve the above problem we exploited the Gurobi solver within the COBRA Toolbox [160].

Input and data pre-processing

scFBA takes as input a template metabolic network map A, as in popFBA, plus a scRNA-seq dataset in the form of a $N_{\text{genes}} \times N_{\text{cells}}$ matrix T, where N_{genes} is the number of genes and N_{cells} is the number of single-cells under study. Each element $T_{g,c}, g = 1, \ldots, N_{\text{genes}}, c = 1, \ldots, N_{\text{cells}}$ corresponds to the normalized read count of gene g in cell c such as, for instance, the TPM (Transcripts Per Kilobase Million).

The risk of the presence of false negatives in RNA-seq, and in particular scRNA-seq, is an established problem. Although a totally safe solution does not exist, scFBA allows to employ the information on bulk expression profile, when available, to manage the risk, by envisioning the following scenarios.

- If a gene has a zero read count in the bulk, as well as in each single-cell, we cannot totally exclude the possibility of a false-negative in the bulk, but we are confident in excluding a false-negative due to low concentrations of scRNA-seq, thus we can assume that such gene is off in all cells. We directly delete this set of genes G_{off} from the template metabolic network A, by solving the Gene-Protein-Reaction (GPR) association rules with a true-false logic (Cobra Toolbox [160] function: *geneDeletionAnalysis*), which results in removing reactions for which their expression is essential (AND operator). We refer to the obtained subnetwork of A as A^*.

- If a gene has non-zero read count in the bulk, but a zero read count in each single-cell, there is a sharp inconsistency between bulk and scRNA-seq that indicates that we cannot trust scRNA-seq for this gene. In this situation, we prefer to lose information on single-cell heterogeneity and rely on the bulk value: we replace the read count for that gene in each cell with the bulk read count.

- If a gene has non-zero read count in the bulk, and zero read count in some of the single-cells, we cannot be sure that the gene is actually not expressed in those cell, but we can exclude that there is a problem with the detection of that specific gene and we can hypothesize that is at least poorly expressed as compared to other cells. As a compromise between a more conservative strategy and the need to preserve information on cell heterogeneity, we retain the single-cell read count for these genes, but we do not prevent completely flux through the associated reactions, when setting boundaries of the reaction as a function of their expression. As we will illustrate in Section 4.2.2, we set the flux bound to a small value ϵ.

Reaction Activity Scores

We define a Reaction Activity Score (RAS), for each single-cell $c = 1, \ldots, N_{\text{cells}}$, and each reaction $j \in \mathcal{R}$, based on Gene-Protein-Reaction association rules (GPRs). GPRs are logical formulas that describe how gene products concur to catalyze a given reaction. Such formulas include AND and OR logical operators. AND rules are employed when distinct genes encode different subunits of the same enzyme, i.e., all the subunits are necessary for the reaction to occur. OR rules describe the scenario in which distinct genes encode isoforms of the same enzyme, i.e., either isoform is sufficient to catalyze the reaction.

In order to compute the RAS we distinguish:

- Reactions with AND operator (i.e., enzyme subunits).

$$RAS_j^c = \min_g (T_{g,c} : g \in S_j) \tag{4.13}$$

  where S_j is the set of genes that encode the subunits of the enzyme catalyzing reaction j.

- Reactions with OR operator (i.e., enzyme isoforms).

$$RAS_j^c = \sum_{g \in I_j} T_{g,c} \tag{4.14}$$

  where I_j is the set of genes that encode isoforms of the enzyme that catalyzes reaction j.

In case of composite reactions, we respect the standard precedence of the two operators.

scFBA

The first step of the scFBA approach is the creation of a multi-scale population model, composed of N_{cells}, according to the popFBA described in Section 4.2.2, but starting from the template metabolic network A^*. As described in section 4.2.2, A^* is a subnetwork of the generic model A which integrates the transcriptional information that holds for all cells in the bulk.

Once the population model is obtained, the scFBA approach imposes two kinds of constraints:

type 1 constraints on the extracellular fluxes of the overall population model P, i.e., the upper and lower bound of the N_{blood} exchange reactions in set $\mathcal{B}$, ideally according to metabolic measurements;

type 2 constraints on internal fluxes of each single-cell c, i.e. for the reactions in $\mathcal{R}^c, with c = 1, \ldots, N_{\text{cells}}$, and for its set of cooperation reactions C^c, according to their RAS, whenever the computation of a RAS is possible, i.e., when a GPR exists for the reaction, along with the transcript values of at least one of the involved genes.

In order to project the information of the activity score of a given reaction j, in a given cell c, R_j^c, onto its flux "pipe capacity":

- we first estimate the possible flux that reaction R_j^c might carry, when only constraints on extracellular fluxes (type 1) are set, whereas the internal fluxes (type 2) are still unbounded and the system is not required to make biomass, i.e., we compute the maximal flux in both the forward (F_f) and backward direction (F_b) of each reaction. To do so, we perform a Flux Variability Analysis [64], with no optimality required (Cobra Toolbox [160] function: *fluxVariability*). We define $F_j^c = \max(|F_f|, |F_b|)$.

- we then compute the relative reaction activity score of R_j^c in each $c = 1, \ldots, N_{\text{cells}}$, with respect to the total activity of reaction j, as follows:

$$\overline{RAS}_j^c = \frac{RAS_j^c}{\sum_{c=1}^{N_{\text{cells}}} RAS_j^c}, \tag{4.15}$$

- finally, we assign an upper bound (U_j^c) to reaction R_j^c, as portion of F_j^c which is proportional to the activity score ($\overline{RAS}_j^c$) of reaction R_j^c. Namely,

we remap the values RAS_j^c, $j = 1, \ldots, N; c = 1, \ldots, N_{\text{cells}}$ in the interval $[\epsilon, F_j^c]$, as follows:

$$U_j^c = \epsilon + (F_j^c - \epsilon) \cdot \overline{RAS}_j^c \tag{4.16}$$

We set the upper bound of reactions having $\overline{RAS}_j^c = 0$ to a small value ϵ rather than to 0 to mitigate the impact of false-negatives. Note that $\sum_{c=1}^{N_{\text{cells}}} U_j^c \approx F_j^c$. We remind that the set of genes G_{off} is instead fully deleted from the model. As baseline value, in this study we set $\epsilon = 10^{-3}$, but we assess how its variation may affect the results by scanning the values: $\{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$.

- if reaction R_j^c is considered irreversible, we assign a zero lower bound ($L_j^c = 0$) to reaction R_j^c, otherwise we assign a lower bound $L_j^c = -U_j^c$. The reason why, when dealing with reversible reactions, we avoid setting different values for backward and forward reaction, by assigning to F_j^c the maximum value between F_f and F_b, is that the RAS reflects the gene expression of its competent enzyme, which may equally work in either direction.

  Once the P model is constrained (with both type 1 and type 2 constraints), Linear Programming, as well as other standard constraint-based methods were applied.

Datasets

In this work, we mainly use the following 3 LUAD datasets obtained from the NCBI Gene Expression Omnibus (GEO) data repository under accession number GSE69405.

LCPT45 Composed of 34 cells acquired from a xenograft, obtained by sub-renal implantation in mice of a surgical resection of a 37-mm irregular primary lung lesion in the right middle lobea of a 60-year-old untreated male patient.

H358 Composed of 50 cells from NCI-H358 bronchioalveolar carcinoma cell line.

LCMBT15 Composed of 49 cells acquired acquired from a xenograft, obtained by sub-renal implantation in mice of a surgical resection of a metachronous brain metastasis acquired from a 57-year-old female after standard chemotherapy and erlotinib treatments.

We repeated all the analyses on the following independent breast cancer datasets (GEO access number: GSE75688), including scRNA-seq data of single-cell suspensions of cancer tissues obtained on the day of the surgery of untreated breast cancer patients [270]:

BC04 Composed of 59 human epidermal growth factor receptor 2 positive (HER2+) cells.

BC03LN Composed of 55 lymph node metastases of human estrogen receptor positive (ER+) and (HER2+) cells.

Each of the 5 datasets includes the gene expression level of more than 20.000 genes in the form of Transcript Per Kilobase Milion (TPM). We filtered out a few cells with less than 5000 genes detected. For each dataset, we retained only the metabolic genes included in HMRcore model (418 genes). The dataset transcripts are identified by Ensembl ID, which we automatically converted into HUGO Gene Nomenclature Committee (HGNC) ID. Datasets also contain the expression profile of the bulk samples, which we used to pre-process data as described in Section 4.2.2.

Metabolic network model

From the computational perspective, the scFBA approach is suitable for simulation of genome-wide metabolic networks, such as [277, 278]. However, in view of previous analyses [31], in order to have more control on the analyses and make the interpretation of results more straightforward, at this stage, we preferred to focus on a more handful and carefully reconstructed core metabolic network. We used, as template network A, the metabolic core model HMRcore introduced in [78] and used in [219, 276]. As exchange of fatty acids between cells in tumours has been recently reported [254, 279], we included the possibility to exchange palmitate via the TME and, accordingly, mitochondrial palmitate degradation and gluconeogenesis. Given the importance of reactive oxygen species (ROS) metabolism observed in [31], we also inserted ROS production and removal pathways. As the original version of the model does not include information on GPRs, such rules have been extracted from Recon 2.2 [277] and included in the HMRcore model. We decided to disregard the GPRs associated to the complexes I to IV of the electron transport chain in scFBA computations, because it unrealistically requires up to 81 genes (AND rule). However the flux through complexes I to IV should be modulated by the constraints on complex V (ATP synthase).

The final version of the HMRcore model includes 315 reactions (of which 263 are associated with a GPR) and 418 metabolic genes. The SBML of the model is provided in https://github.com/BIMIB-DISCo/scFBA.

Experimental setting The choice of the nutrients exchanged with biofluids should ideally be dictated by metabolic measurements on exo-metabolome. As we do not have this information, in the baseline experimental setting, we considered as main exogenous nutrients (which the overall population can uptake from 0 up to $Npop \cdot 100 nmol/h$) [219] those that are the main nutrients of cancer cells according to literature, as motivated in [31]: glucose, glutamine, oxygen: glucose, glutamine, oxygen and arginine. Along similar lines, we considered as nutrients that can be secreted by cancer cells in the tumor microenvironment those that are mainly reported in literature, and which may play a role in metabolic cooperation: glutamate [280], NH_3 [164, 281, 282], lactate [25, 190, 207], and palmitate [254, 279]. In order to be able to discern the advantage of cooperation from that of the mere secretion of metabolites, we considered both a cooperation reaction and a secretion reaction for these nutrients.

Results

Integration of RNA-seq data efficiently reduces the space of optimal solutions

We first applied scFBA to the 5 datasets described in Section 4.2.2, assuming maximization of total (population) biomass synthesis rate as objective function. All five population models displayed a non negligible maximal growth rate, something that cannot be taken for granted when integrating transcriptomics into FBA models [202, 212, 213]. In Figure S1 and accompanying Text S1, we also show that, if ϵ takes value 0, the scFBA problem is still feasible, but we obtain very small values for the fluxes.

In order to highlight that the scFBA approach efficiently reduces the space of optimal solutions, we compared the variability of the biomass production flux of each of the N_{cells} single-cells simulated within the population model, for each of the 5 datasets under study. We report in in Figure 4.18 the results for the two datasets relative to the primary tumors, and in Figure S2 for the other datasets. When no information on cells' transcriptome is employed (as in standard popFBA settings [219]), the type 2 constraints of the metabolic network are identical for all cells. This implies that each cell is capable of contributing alone to 100% (10^0) of the objective function value (i.e., the biomass of the total

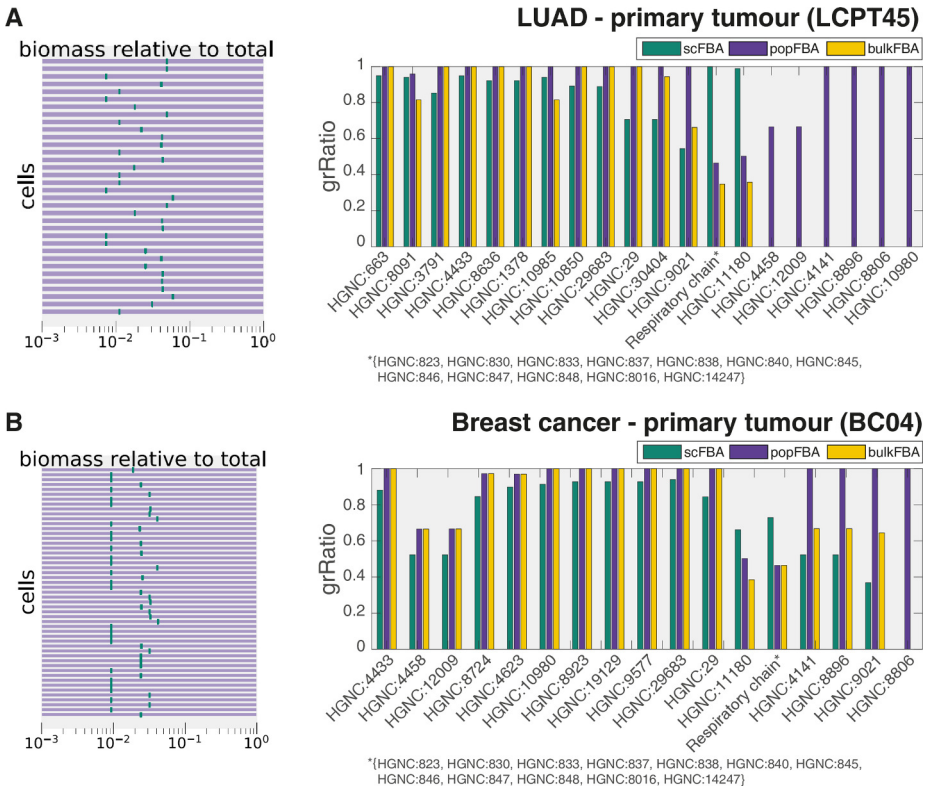


Figure 4.18: scFBA vs. popFBA. A) Dataset LCPT45. Variability of the fraction of the biomass synthesis flux (logarithmic scale) for each cell over the population growth rate (left panel) before (purple) and after data integration (green). Effect of gene deletion (bars in right panel) on the population growth rate before (popFBA), after data integration (scFBA), and for the template metabolic network A^* (bulkFBA). When $grRatio = 0$ (essential gene), the corresponding bar is not displayed. B) Same information as in A for BC04 dataset.

population). As depicted in Figure 4.18 (left plots) and S2 (left plots), the biomass flux value of each cell, within the set of optimal solutions, spans indeed from 0 to 100% of the total biomass (purple rectangles). On the contrary, after scRNA-seq data integration, as performed via scFBA, the biomass flux of each cell can only take a specific (optimal) value, corresponding to a certain fraction of the total biomass (green rectangles, which results in a single line, because the maximum and minimum optimal flux values coincide.)

To show how this volume reduction in the space of alternative optima may actually affect predictions, we performed a single gene deletion analysis with and without scRNA-seq data integration (scFBA and popFBA, respectively). When a single gene is deleted, the reactions for which the expression of such gene is essential (i.e., reactions exclusively associated to the gene, or reactions associated to that gene and other genes with an AND operator) are removed from the network (i.e., from the set $\mathcal{R}^c, \forall c$). After removal, the population model is newly optimized for total biomass production, and the growth ratio ($grRatio$) of the new biomass over the previous one is computed.

Figure 4.18 and S2 (right bar plots) report the $grRatio$ observed for those genes deletions that displayed a different effect before and after data integration. Notice that when the $grRatio$ equals 0, the corresponding bar is not displayed at all. To verify that the differences between scFBA and popFBA are not a mere consequence of the removal of reactions (in scFBA) that are inactive in all cells of the bulk from the template metabolic network A, we include in the plots the prediction of the isolated template metabolic network A^*. We refer to this third simulation setting as bulkFBA. However, bulkFBA includes information on on-off reactions only. It is not possible to modulate the flux capacity of reactions as a function of gene expression, because it not possible to compute relative expression values.

Remarkably, some genes that are redundant ($grRatio = 1$) in popFBA settings (i.e., with no scRNA-seq data integration) may even become essential in scFBA settings (i.e., with scRNA-seq data integration) ($grRatio = 0$). This is the case, in lung adenocarcinoma, of the following genes: HGNC:10980, which encodes enzymes responsible for glutathione/phosphate, fumarate/phosphate or α-ketoglutarate/malate antiports; HGNC:8806, which encodes for a subunit of pyruvate dehydrogenase; HGNC:8896, which encodes for an isoform of phosphoglycerate kinase and HGNC:4141, which encodes for an isoform of glyceraldehyde 3-phosphate dehydrogenase. In breast cancer, only gene HGNC:8806 falls into this category. Conversely, some genes that display a significant effect ($grRatio \approx 0.5$) in popFBA become instead redundant in scFBA. This is the case, in lung adenocarcinoma, of the genes that encode for ATP synthase

(HGNC:823, 830, 833, 837, 838, 840, 845-848, 14247, 8016), indicating that the integration of scRNA-seq data forces a (suboptimal) flux distribution for cancer cells which, consistently with the well-known Warburg effect, does not rely largely on ATP synthase for ATP production, thus resulting in a milder effect when the reaction is depleted. Worth of note, although these genes are not completely redundant ($grRatio < 1$) in breast cancer, the deletion of the respiratory chain has indeed a mild effect in both tumors, as well as in the other cancer datasets reported in Figure S2.

It is apparent, from Figure 4.18 and S2 (right bar plots), that bulkFBA provides intermediate results between scFBA and popFBA. Some genes that are redundant in PopFBA, are lethal in both scFBA and bulkFBA. This is the case for example in LCPT45 of gene HGNC:10980 (Mitochondrial dicarboxylate carrier). This result is expected, given that its isoform has been deleted according to bulk data in both simulations. Conversely, some gene deletions that have a significant effect according to bulk data have no effect when also single-cell data are considered, in particular the genes encoding for components of the respiratory chain. Also, the effect of the deletion of pyruvate kinase (HGNC:9021) is smaller in scFBA than in bulkFBA of scFBA. On the other hand, some deletions may show some effect only when scRNA-seq are considered. This is particularly true for genes that are involved in cooperation mechanisms among cells, as for instance gene HGNC:29, in both datasets, whose product promotes the secretion of palmitate, which can be taken up by other cells.

scFBA extracts useful features from transcript signals.

As previously mentioned, single-cell fluxes are expected to be less noisy than transcript signals, which are typically analyzed by means of multi-variate statistical analysis [249] and, therefore, the former might be used to better identify cell clusters that might represent distinct metabolic subpopulations. To confirm this hypothesis, we performed a cluster analysis on the expression values (scRNA-seq) of the metabolic genes and compared the results with those of a cluster analysis performed on the fluxes predicted by scFBA. To this end, we performed both hierarchical and k-means cluster analysis. In order to avoid reactions with typical high flux-value, or genes with high expression, to induce a bias on clustering results, we first remapped the flux (transcript) values of each reaction (gene) j in the interval $[0, 1]$: value 0 is assigned to the cell showing the lowest value for a given flux (transcript), 1 to the the one showing the highest value.

Figure 4.19 and Figure S3 report the results of the hierarchical clustering

analysis (distance metric: euclidean), for transcripts (left column) and fluxes (middle column), respectively for the two primary tumors and for the other 3 datasets under study. From the dendrograms and heat-maps, one can see that cells cluster in a few well-separated groups, when the extracted features (the fluxes) are considered, whereas they cluster more in "singletons", when the original features (the transcripts) are used. For example, when observing the fluxes computed for the red LUAD dataset LCPT45 (panel A in Figure 4.19), it is apparent that two major (and a minor) groups of cells can be identified, corresponding respectively to the blue and red-coloured leaves in the dendrogram. We evaluated which are the most different fluxes among the two major groups, by using the Z score test of statistical significance (File S1). Remarkably, the two groups significantly differ in their growth rates (Z-scores: 3.2). 82 reactions significantly differ between the two groups with a 99% confidence level. This set mostly include pathways directly or directly linked with biomass synthesis, such as biosynthesis of fatty, amino and nucleic acids.

To quantitatively compare the clustering of transcripts and fluxes, we first performed a k-means clustering with different number of clusters k, by considering $n = 100$ bootstrap iterations (with random centroid assignments) and by selecting the clustering resulting in the maximum inter-cluster distance. We then assessed the clustering goodness, by means of traditional "elbow" and "silhouette" evaluation methods. We refer to text S2 for details about these approaches.

The elbow method in the right column of Figure 4.19, indicates that a elbow is observed at $k = 3$ for the fluxes relative to the primary tumour datasets, hence the optimal number of clusters is 3, which corresponds indeed to the k identified by the hierarchical clustering analysis for these two datasets.

In Figure S3 (panel D), we evaluated the silhouette for the dataset LCPT45 transcripts (left) and fluxes (right) for $k = 3$, i.e., the value identified from the elbow analysis, which also corresponds to the highest average silhouette value, when varying k in $\{2, \cdots, 6\}$ (data not shown).

Consistently with the more ready recognition of major clusters in the flux diagrams noted above, the drop in the sum of squared errors (SSE) is much stronger and the average silhouette value is considerably higher in the flux case than in the transcripts case (where the average coefficient is close to 0) indicating that the calculation of fluxes leads to a better clustering as compared to the evaluation of transcripts. All in all, the results of the cluster analyses indicate that fluxes can be better clustered than their transcript counterparts. Remarkably, it can also be noticed that the LUAD primary tumour xenograft (see the sharper "elbow" and the clustergram in Fig. 4.19A) fluxes better partition into

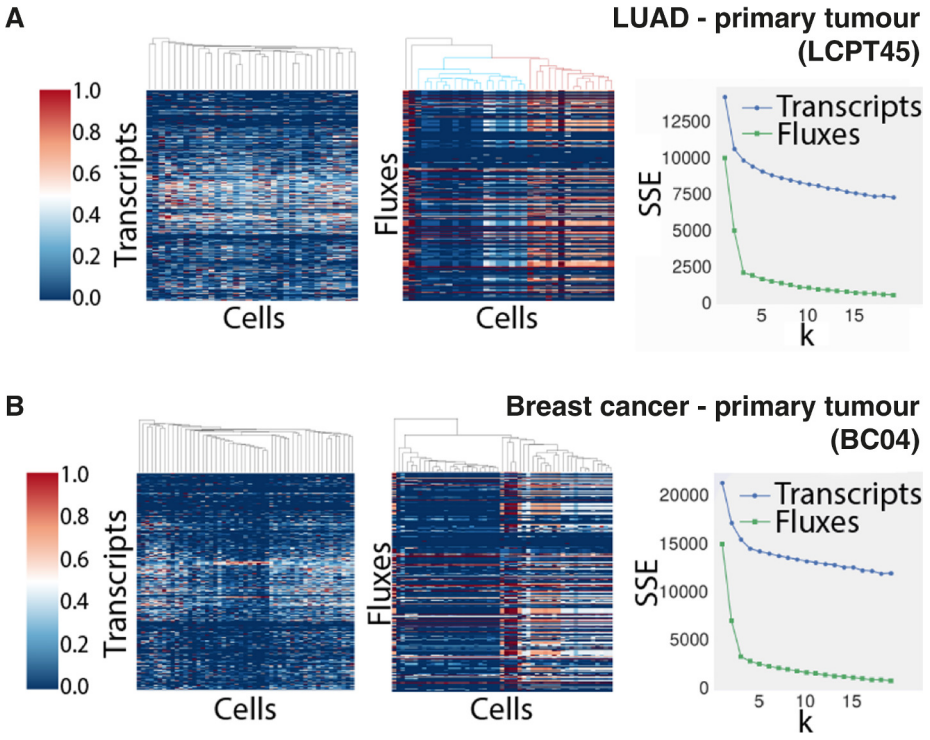


Figure 4.19: Clustering of transcripts vs. fluxes. A) LCPT45 dataset. Clustergram (distance metric: euclidean) of the transcripts of the metabolic genes included in metabolic network (left) and of the metabolic fluxes predicted by scFBA (middle). Right panel: elbow analysis comparing cluster errors for $k \in \{1, \cdots, 20\}$ (k-means clustering) in both transcripts (blue) and fluxes (green). B) Same information as in A for the BC04 dataset.

clusters than the fluxes of the cell line (Fig. S3, H358) and of the secondary tumour xenograft (Fig. S3, LCMBT15). This result is in line with the data reported in [269], indicating that a binary separation of cells is more evident in LCPT45. We detected a similar difference between the clustering results of primary and secondary tumour of the independent breast cancer datasets (BC04 in Fig. 4.19 and BC03LN in Fig. S3). Indeed, the former population is more heterogeneous in the binary sense than the latter [270].

scFBA captures interactions between cells

The main rationale behind solving a unique mass balance problem for many cells together, given constraints on the extracellular fluxes of the bulk, rather than many separate mass balance problems, is that the nutrient consumption and secretion rates (extracellular fluxes) can be readily determined or approximated from measurements of the concentration of metabolites in the cell culture media at different time points for the bulk only. Another major side benefit of this approach is that it allows to identify the possible interactions among cells within a population, as pointed out in [219].

We verified that, after data integration, some cells secret metabolites that are up-taken by other cells. The heat map in Figure 4.20 shows the (normalized) flux values of cooperation reactions for the LCPT45 dataset: a positive value means that the cell is secreting the metabolite in the tumour microenvironment, whereas a negative flux that the cell is uptaking it from the tumour microenvironment. It can be observed that a complex network of interactions is established among cells. In particular, a consistent group of cells consumes the lactate and palmitate that are secreted by other groups. The scatter plots in Figure 4.20 show the dispersion of the fluxes of uptake/secretion from/into the TME for lactate and palmitate and how they couple with different growth rates, portraying a relationship far more complex than that depicted with popFBA (no scRNA-seq integration [219] and no exchange of palmitate allowed).

Metabolic interactions between cancer-associated fibroblasts and cancer cells, mediated by palmitate [254, 279] and lactate [25, 190, 207] have been recently reported. At the same time, it has been shown that metabolic heterogeneity can arise in genetically homogeneous cells as simple as the budding yeast *Saccharomyces cerevisiae* [283]. scFBA has the potential to highlight possible metabolic heterogeneity also within a genetically homogeneous population of cancer cells. The validation of the predicted interactions requires however non-trivial *ad hoc* experiments. As current techniques do not allow for easy determination of metabolites at the single-cell level, the heterogenous population of alive

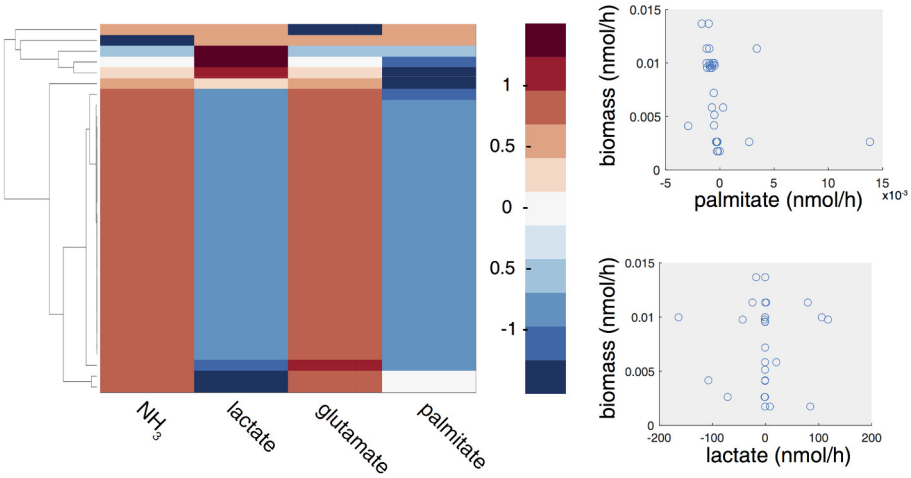


Figure 4.20: Metabolic cooperation in LCPT45 population. Left: Clustergram of the fluxes of cooperation reactions for NH_3, lactate, glutamate and palmitate. Negative fluxes (blue shades) indicate an uptake, whereas positive fluxes (red shades) indicate a secretion of the corresponding metabolite. Right: Scatterplot of the biomass flux values of each cell in the population vs. palmitate (top) or vs. lactate cooperation flux (bottom).

cells should be first sorted to separate it into the sub-populations identified by scFBA. However, to sort cells based on fluorescent labeling (Fluorescence activated cell sorting), further analyses are necessary to possibly identify markers differentially expressed by the sub-populations. Less direct approaches might be taken, for example, by measuring the growth rate of wild-type cell populations and mutant populations in which the cooperation has been prevented (e.g. by blocking secretion or uptake of involved metabolites) and comparing with the model predictions. At this purpose, it should be assessed whether the metabolic interactions identified by scFBA are actually advantageous for tumor growth or are just related to the entropy of the system, i.e., to the fact that a configuration in which interactions among heterogeneous phenotypes take place is more likely than a configuration of identical and independent phenotypes. We address the issue in the following paragraphs.

Effect of cooperation on growth The optimal values for the cooperation fluxes reported in Figure 4.20 display a larger variability than the optimal growth rates (Fig. 4.18), even though much lower as compared to popFBA settings (by at least 60%). Therefore, we verified that the interaction among cells, given their transcriptomes, improves the capability of the overall population to achieve metabolic growth, while also correcting for the possible presence of thermodynamically infeasible loops [161]. At this aim, we compared the population growth rate of the case in which cooperation reactions are allowed, with the case in which they are not, given the same constraints (type 1 and type 2). As the mere secretion of metabolites (such as lactate) in the external environment (e.g., the blood) can improve growth rate, under given boundary conditions (e.g., limiting oxygen [31]), in order to allow for a meaningful comparison, in our experimental setting metabolites that can be secreted in the TME can also be secreted directly in the blood supply. By doing so, when the cooperation reaction is removed, the cell can still rid off of excess metabolites, which cannot however be taken up by other cells.

Remarkably, we observed that the ratio of the total biomass obtained in the absence of cooperation reactions over that in their presence may be lower than 1, implying that removal of cooperation limits the capability to achieve growth. In particular, we observed a ratio of: 0.90 for the LCPT45 dataset; 0.99 (H358); 0.99 (LCMBT15); 0.76 (BC04); 0.95 (BC03LN).

Intriguingly, but not surprisingly, the impact of cooperation prevention is higher on those datasets corresponding to more heterogeneous populations (LCPT45 and BC04). Intuitively, cells specialized in different metabolic programs are more likely to benefit from cooperation, as compared to similar cells.

Effect of cooperation on ATP production For the sake of simplicity, in this study we assumed an optimal growth rate for the overall population, yet other assumptions may be readily investigated with the scFBA approach. Among others, it is common practice in constraint-based modeling to optimize for ATP production [203, 284]. As a proof of principle, we repeated the analysis on the effect of cooperations when the objective function is the total ATP produced by the population. We obtained the following ratios for the 5 datasets: 0.77 (LCPT45); 0.97 (H358); 0.93 (LCMBT15); 0.99 (BC04); 0.87 (BC03RLN). The observed discrepancies in the extent of the effects of cooperation inhibition on growth and energy productions are worth of interest and would deserve further investigation. Notably, both the energy production and growth rates of the H358 (cell line) population, which is expected to be homogeneous, are not

affected by cooperation prevention.

Boundary conditions affect scFBA predictions

Both in popFBA and in scFBA the cells were able to collaborate metabolically. As the integration of scRNA-seq data greatly reduced the space of feasible FBA solutions, those data encode information on how nutrient utilization should be distributed amongst the individual cells. Some cells that can no longer carry out a certain part of a pathway let their neighbors do this. However, it should still matter which nutrients are available to all cells. For a deeper characterization of given cancer populations, exo-metabolomic measurements to constrain the population boundary conditions would thus be needed. An exhaustive sensitivity analysis of scFBA results to boundary conditions is out of the scope of this work. However, it is interesting to compare the conditions in which the two major metabolites involved in cooperation (i.e., lactate and palmitate) are externally supplied to the population or must be produced endogenously.

Notably, we observed that uptake of exogenous palmitate does not affect the biomass production rate, indicating that no growth advantage is conferred by free availability of lipids. This result is in line with experimental evidence that cancer cells rely on *de novo* synthesis of palmitate-derived lipids [248]. However, in the baseline setting (no external palmitate supplied), we observed a group of cells that uptake the palmitate secreted by others (Fig. 4.20). We verified that, once internalized in those cells, palmitate is not processed by the beta-oxidation pathway, but directly contributes to the biomass synthesis, supporting the evidence reported in [183], that an exogenous source of fatty acids can substitute for *de novo* synthesis in promoting cell proliferation and attenuate the cancer-specific toxic effect of lipogenesis inhibitors. It has also recently been pointed out that a limited access to environmental lipids may render the cancer cells more sensitive to the inhibitors of lipogenesis [183]. In line with these findings, it can be observed, with regard to the LCPT45 population (Figure 4.21A), that a set of genes stops being essential when exogenous palmitate is supplied. As expected, this set mainly includes genes directly involved in the synthesis of palmitate, namely: citrate synthase, fatty acid synthase and pyruvate dehydrogenase. The expansion of the plot in Figure 4.21A (left) shows that the latter (pyruvate dehydrogenase, ID: HGNC:8808) is essential for each cell within the population. It should indeed be noted that when the synthesis of palmitate is prevented in all cells, and exogenous palmitate is not supplied, also cells that used to rely on the palmitate synthesized by other cells must be affected.

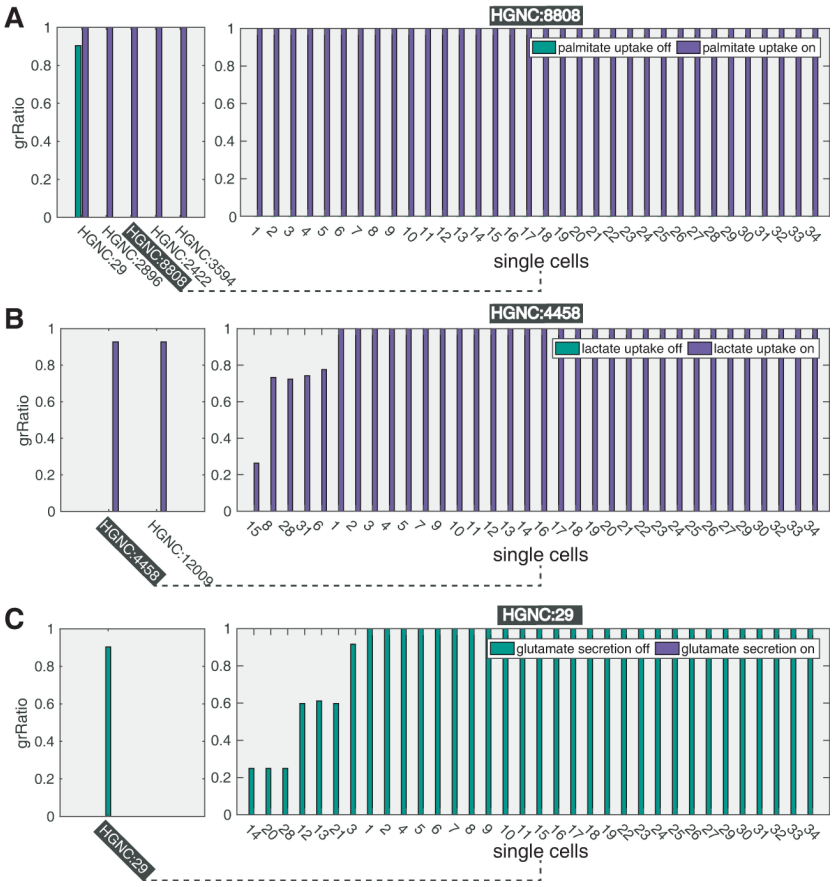


Figure 4.21: Impact of boundary conditions on gene-deletion predictions for LCPT45 dataset. A) Left: effect of gene deletions on the population growth rate, when exogenous palmitate uptake is allowed (purple bars) and when is not (green bars). Only genes with differential effect are reported. A missing bar indicate an essential gene ($grRatio = 0$). Right: effect of the deletion of gene HGNC:8808 on the growth rates of each single-cell. B) Left: effect of gene deletions on the population growth rate when exogenous lactate uptake is allowed (purple) and when is not (green). Right: effect of the deletion of gene HGNC:4458 on each single-cell. C) Left: effect of gene deletions on the population growth rate when endogenous glutamate release is allowed (purple) and when is not (green). Right: effect of the deletion of gene HGNC:29 on each single-cell.

As opposed to palmitate, the metabolite lactate is not strictly required for growth. However, it can be observed in Figure 4.21B that the deletions of genes encoding for glucose-6-phosphate isomerase (HGNC: 4458) and for triosephosphate isomerase 1 (HGNC: 12009) - two important steps for the utilization of glucose through glycolysis - are not essential when lactate uptake is allowed, suggesting that lactate may be able to replace glucose as carbon source. Interestingly, when lactate uptake is prevented, the plot expansion in Figure 4.21B (left) shows that the gene HGNC: 4458 is essential in many but not all cells.

Also the set of metabolites allowed to be released, e.g., in the blood may affect the effect of gene deletions. For instance, if glutamate secretion is prevented, the deletion of the gene that encodes palmitate secretion becomes essential, as shown in Figure 4.21C. Remarkably, it has been reported that secretion of lipids facilitates tumour progression [285], whereas inhibitors of glutamate release have been proposed as new targets for breast cancer-induced bone-pain [280]. scFBA may enable to shed light on how the disposal of carbons through these two metabolites relates with the utilization pattern of exogenous nutrients.

Discussion

We have here introduced scFBA to solve the problem of reconstructing the potential single-cell fluxome, starting from single-cell transcriptomes, by taking into account environmental constraints, as well as cell-cell interactions. Importantly, scFBA is able to point out the metabolic interactions that are established within a cell population.

scFBA integrates sc-transcriptomics data with (bulk) extracellular fluxes of the same cancer cell population, by means of a computational approach inspired to complex systems science [286]. A limitation of our approach is that it uses mRNA levels as a proxy of the maximal velocity of reactions ($Vmax$), thus neglecting the many factors that contribute to determine the expression level of a protein [260], as well as the role played by binding affinities of proteins in determining the $Vmax$. However scFBA does not predict the single-cell fluxome as a linear function of the assumed Vmax: the constraints on mass-balance and on the availability of substrates, as determined by the rates of consumption/secretion of nutrients by the entire population and by the requirement for the tumour mass to grow, are simultaneously taken into account. scFBA might also be implemented by using sc-proteomics rather than transcriptomics, should the former become available.

Although we do not explicitly model spatial organization, the constraints on single-cell transcriptome should implicitly preserve the information on the us-

age/secretion of most nutrients of each cell in its original position. The method seems also to neglect communication between cells through growth factors. In reality, it mostly does not if the growth factors act by changing transcription. But if they act by phosphorylating enzymes, then this is not taken into account. In its current form scFBA can however already be implemented to investigate why metabolic drugs are often ineffective and to provide indications for more effective treatment.

As a proof of principle, we have successfully applied the methodology to LUAD and breast cancer datasets. We have shown that the integration of scRNA-seq greatly reduces the space of feasible solutions that sustain metabolic growth of the overall population, which is prerequisite of tumour growth. This reduction allowed us to restrict the set of candidate drug targets, by eliminating targets that may seem obvious for the bulk, but do not work for heterogeneous populations, and, on the other hand, by revealing targets whose relevance can be appreciated only if cooperation among heterogeneous cells is accounted for. We have also illustrated that scFBA is valuable to extract features (i.e., the sc-flux values) from scRNA-seq data, in order to identify metabolic clusters of cells, which may be used to investigate other fingerprints of the cancer metabolic deregulation.

Although popFBA assumes that the cells achieve optimal growth, this assumption is mitigated in scFBA, by taking into account the transcriptional constraints. Moreover, we have shown how alternative objective functions, such as ATP maximization, may be investigated. Sub-optimality may also be taken into account by using sampling methods [170].

In this study, we have used scRNA-seq obtained with protocols based on C1 Single-Cell Auto Prep System, which have the advantage of allowing to remove dead cells before sequencing but may suffer from including low numbers of cells per sample (34 - 55 cells), as compared to modern emerging technologies which allow to obtain the single-cell transcriptome of thousands of cells at the cost of a bulk experiment [287], with an improved number of genes/transcripts per cell. In the future, our approach may be readily generalized to this kind of data. As illustrated in Figure S4, the time of a scFBA computation increases linearly with the number of simulated cells and with the size of the template metabolic network. Alternatively, bootstrap-like methods can be adopted to reduce the number of cells considered in a single computation, while parallelizing the simulation of many smaller systems.

A major challenge, when dealing with scRNA-seq, is the presence of false-negatives. As a first approximation, we have used a bulk RNA-seq expression filter, where single-cell expression values of transcripts never detected in scRNA-

seq but detected in bulk RNA-seq are replaced by the bulk values. For more reliable genes, we preserve information on cell heterogeneity, but we mitigate the risk of false-negative, setting the bound of the associated reactions to a small value ϵ rather than completely removing it. The choice of the value of ϵ is partially arbitrary. However, we verified that main results of our work are robust with respect to this choice (S1 Text and S1 Figure). The methodology might be refined, by combining it with more sophisticated data pre-preprocessing techniques, which may also take into account the specific quality parameters of the dataset.

Finally we have shown how constraints on the nutrient consumption and secretion rates (extracellular fluxes) of the specific sequenced population may affect scFBA predictions. As opposed to intracellular fluxes, (bulk) extracellular fluxes might be readily estimated, e.g., by approximation from metabolite concentrations in spent medium (exo-metabolome), when culturing patient-derived cells.

Hence, measurements of both extracellular fluxes and single-cell transcriptional information of the same heterogeneous cancer populations are needed to make scFBA predictions fully reliable. These datasets can be realistically obtained by culturing population of cancer cells and then processing the cells with scRNA-seq technologies and analyzing their spent medium with biochemical analyzers. Experiments under a controlled setting, e.g., a co-culture of metabolically characterized cancer cell lines, may first be performed to validate and tune the capability of scFBA to identify and measure the prevalence of different metabolic subpopulations of cells. The application of scFBA to analyze datasets more representative of tumor heterogeneity, obtained, for instance, by culturing cells from human biopsies, xenografts or organoids, will then pave the way to cancer personalized medicine.

Supplementary Information

Figure S1. Sensitivity of scFBA results to ϵ for LCPT45 dataset. A) Left: histogram of biomass produced by each single cell when $\epsilon = 0$. Right: Total biomass produced by the population of cells as a function of ϵ. The inset reports the same curve zoomed in on low ϵ values. B) Clustergram (distance metric: euclidean) of the effect of single gene deletions performed on scFBA for different values of ϵ, popFBA and bulkFBA. Growth ratio (grRatio) = 1 indicates totally redundant genes, while grRatio = 0 indicates lethal genes. C) Elbow analysis comparing cluster errors for $k = 1, \ldots, 20$ (k-means clustering). Each curve refers to a different values of ϵ. D) Impact of cooperation among

single cells for different values of ϵ. Curves refer to the ratio of total biomass (blue curve) and ATP (orange curve) produced by population models when cooperation reactions are blocked as compare to when they are allowed.

Figure S2. scFBA vs. popFBA. A) Dataset H358. Variability of the fraction of the biomass synthesis flux (logarithmic scale) for each cell over the population growth rate (left panel) before (purple) and after data integration (green). Effect of gene deletion (bars in right panel) on the population growth rate before (popFBA), after data integration (scFBA), and for the template metabolic network A^* (bulkFBA). When $grRatio = 0$ (essential gene), the corresponding bar is not displayed. B-C) Same information as in A for LCMBT15 and BC03LN datasets.

Figure S3. Clustering of transcripts vs. fluxes. A) H358 dataset. Clustergram (distance metric: euclidean) of the transcripts of the metabolic genes included in metabolic network (left) and of the metabolic fluxes predicted by scFBA (middle). Right panel: elbow analysis comparing cluster errors for $k \in \{1, \cdots, 20\}$ (k-means clustering) in both transcripts (blue) and fluxes (green). B-C) Same information as in A for the datasets LCMBT15 and BC03LN. D) Silhouette analysis for LCPT45 transcripts (left) and fluxes (right), when $k = 3$. Red dashed lines indicate the average silhouette for the entire dataset.

Figure S4. scFBA computation time. The linear relationship between the time for an FBA (and thus a scFBA) optimization and the size of the network is well established. We estimated the computation time required to perform a complete model reconstruction, from a template metabolic network to a population model with RASs integrated, for different number of cells (1, 10, 100, 1000 and 10000). We tested both our HMRcore metabolic netowork (panel A) and the genome-wide model Recon2.2 [277] (panel B). The former included 315 reactions and 256 metabolites, the latter is composed of 7785 reactions and 5324 metabolites. We were not able to reach the maximum population model size (10000 cells) with Recon2.2 due to insufficient RAM for 1000 cells. We also verified the feasibility of an FBA optimization for HMRcore and 10000 cells considered (2940021 reactions and 2350021 metabolites in total). The optimization required about 321 seconds. All tests were performed using a PC Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz 64-bit capable, with 32 GB of RAM DDR3 1600 MT/s.

Text S1. Description of sensitivity of scFBA results to ϵ.

Table S1. Comparison of the fluxes of the two main clusters in Figure 3A-middle ϵ.

Table S1. Comparison of the fluxes of the two main clusters in Figure 3A-middle ϵ.

Chapter 5

Conclusions and future perspectives

Cells are complex systems characterized by multiple genes, proteins and metabolites. For many decades, molecular biology has individually investigated cell components without taking into account the set of interactions existing among them. In this scenario, the acquired knowledge is necessary, but not sufficient, to understand how a given cell genotype is converted into a specific phenotype. Indeed, since, rarely, a biological function just comes out from the behaviour of single genes or molecules, the complexity of the entire system is underestimated. On the contrary, most of the time, a biological function emerges from the complex set of numerous interactions established among cell components.

With the advent of high-throughput technologies, the focus has shifted from the individual cellular components to the system-level view of cells underlying the systems biology approach. Systems biology aims at a holistic understanding of a biological system, whose global properties cannot be inferred from the sum of those belonging to its individual components. Moreover, the relationships existing among all components of the system become extremely fundamental to understand how they affect all the characteristics and functions of the system itself.

In this thesis work, the attention has been focused on the level closest to cell phenotype, that is, cell metabolism. Metabolic profiling of a cell provides a comprehensive and functional view of the biochemistry connecting its genome to a particular phenotype following the interplay between cell and its envir-

onment. The aim of this thesis was to take advantage of the potential of the systemic approach proposed by systems biology, as a way to study metabolism as complex biological system. In particular, I defined new *in silico* predictive models and implemented novel computational methodologies, exploiting the classic constraint-based modelling as main computational tool, in order to investigate the multiple sides of heterogeneity affecting cell metabolism and contributing to its high degree of complexity.

In Chapter 3, I started the journey towards the investigation of metabolic heterogeneity by treating cell metabolism as a homogeneous system. This means that the focus is on the characterization of metabolism of an individual cell, as representative of the population to which it belongs. In this way, the average metabolic profiling of the population is described, by considering and, at the same time, hiding cell-to-cell differences that are known to be constantly present in any population of cells [11]. This assumption has turned out to be useful to lower the overall complexity level of metabolism as system, without compromising biological validity of our *in silico* outcomes.

The reconstruction of a high quality and curated mathematical model is a crucial factor to address cell metabolism from a computational point of view. In this context, genome-scale metabolic networks represent the basis for investigating cell metabolic potential, because of their key feature of embracing all available knowledge about biochemical transformations taking place in a given cell or organism. Although genome-wide reconstructions of metabolism have been produced for multiple organisms, they are not yet available for all the known ones. Therefore, in Section 3.1.1, I introduced a computational pipeline for the automatic reconstruction of genome-scale metabolic networks for specific organisms of interest. For the implementation of this methodology, the available knowledge stored in biological databases for the target organism, coupled with the corresponding sequenced and assembled genome, have been exploited. Although the problem is not new and multiple tools for performing this task already exist, our approach may be more beneficial on several points. First of all, the trade-off between automation and manual intervention is minimised in favour of an almost full automatic reconstruction pipeline. Moreover, the choice of biological databases from which informations are extracted, results in greater coverage in terms of living organisms, both prokaryotic and eukaryotic, and metabolic pathways that can be investigated. Finally, our reconstruction approach aims at including organism-specific data and knowledge, or at least deriving from phylogenetically closest organisms, by assuming conservation of biological functions between neighbours. On the contrary, as happened in some

cases, general (one-size-fits-all) data about biomass composition and universal metabolic models as templates for reconstructing organism-specific networks, have been used. Although this strategy seems to be a compromise in cases of missing organism-specific data, it increases the risk of including wrong metabolic reactions and compromising the outcomes of network simulations. By resolving these issues, we believe that simulations of models reconstructed by exploiting our computational pipeline can generate more realistic phenotype predictions.

In this thesis, I showed an application of this methodology for the reconstruction of the genome-wide metabolic network of the hybrid yeast *Zygosaccharomyces parabailii*, called ZyPa1. Through constraint-based modelling, adherence of *in silico* simulations to experimental data and literature findings emerged. Indeed, following Flux Balance Analysis simulations, ZyPa1 model proved to be able to describe the experimental biomass yield at a quite good level. Moreover, in line with experimental evidence presented in [116], the model revealed the ability to follow the co-consumption and catabolism of acetate and glucose, without impacting the biomass yield. This *in silico* evidence suggested that catabolism of acetate can contribute to its detoxification, since it is released from pretreated lignocellulosic biomass and can act as an inhibitory compound for most microbial cell factories, except for *Z. parabailii*. Finally, constraint-based simulations allowed to explore the plasticity of *Z. parabailii* metabolism in response to different metabolic regimes. In this regard, analysis of critical reactions differentially impacting on the growth rate suggested a specific metabolic rewiring that is currently under experimental validation to check its consistency with biological reality. In view of the obtained outcomes, we are confident that ZyPa1 model can be pivotal for making progresses into the understanding of the high stress tolerance of this yeast species, including the description of the possible different contribution of the two divergent haploid parental genomes to its resulting hybrid nature. Moreover, our model can also pave the way for understanding how to better exploit *Z. parabailii* for industrial purposes. In this regard, Zygosaccharomyces parabailii has already proved to be a flexible platform for the production of recombinant proteins and non-natural metabolites, including lactic acid [82], and ascorbic acid [83].

The systemic perspective offered by systems biology approach is well represented by genome-scale metabolic networks. Nevertheless, the comprehensiveness of these reconstructions co-exists with the difficulty in their managing. This aspect regards biological interpretation of the outcomes resulting from their simulations that often results not so straightforward, and the presence of errors due to wrong or incomplete knowledge about the organism under investigation. Consequently, despite the relevance of genome-scale metabolic networks,

a greater control can be achieved by switching to small-scale networks, called core metabolic networks. In Sections 3.2.1-3.2.3, I presented three applications of core modelling as effective means for uncovering system-level properties of cancer metabolic rewiring. Indeed, a reprogramming of energy metabolism within cancer cells has been recently catalogued as one of their hallmarks in order to sustain growth and proliferation. In particular, I showed how to reconstruct core networks concerning human central carbon metabolism by exploiting recent genome-wide metabolic networks as scaffold, and literature knowledge to assist the generation and curation phase of these networks.

New evidence about multiple metabolic wirings adopted by cancer cells in addition to global altered tumour metabolism has lead us to investigate the variability existing among tumours originating from different tissues, also known as intertumour heterogeneity [127]. Pushed by necessity of a more personalized medicine and the development of new and increasingly diversified biomarkers, we reconstructed core models for three types of tumours, namely liver, breast and lung cancer, and for a reference cell. Core modelling and constraint-based simulations revealed heterogeneous metabolic rewirings supporting neoplastic proliferation compared to the reference condition. In addition, heterogeneity of metabolic rewirings among the three investigated tumours in terms of their flux distributions emerged. This variability concerned, in particular, a deregulation of the oxidative phosphorylation pathway, which occurs to a lesser extent in the lung cancer model compared to breast and liver tumours. In accordance with experimental findings demonstrated by [128], the mitochondrial respiration is crucial in lung cancer cells because of its role in promoting their progression and development. Further analyses for the detection of other cancer type-specific potential targets have been performed. In particular, we investigated structural differences that are responsible for a reversion of the tumoural phenotype towards the reference one. This analysis revealed the importance of transport reactions in addition to internal metabolic transformation for phenotype reversion. More strikingly, the joint action of the same set of reactions, is involved in the metabolic traits changing of reference model toward a cancerous state, and vice versa. However, the liver tumour represents an exception due to the fact that this set of changes produced different results as compared to lung and breast cancer. In this regard, the heterogeneity of the responses of different cancer types to a given perturbation supports the importance of developing cancer type-specific models and more personalized anti-neoplastic therapies. Finally, other potential targets has been highlighted from the investigation of cancer networks fragility points. In every tumour model, the highest biomass synthesis rate decrease has been achieved by performing *in silico* inhibition of glycolytic

reactions. This computational outcome is supported by several literature works focused on the correlation between glycolysis and cancer growth [165, 288].

Core modelling also helped to investigate variability of metabolic responses to different nutritional conditions and perturbations. In this regard, in Section 3.2.2, we explored the behaviour of *K-ras*-transformed NIH3T3 mouse fibroblasts (NIH-RAS) when fueled with two different chemically similar sources: glutamine and $\alpha - KG$ in addition to four non essential amino acids (NEAA), namely proline, alanine, aspartate and asparagine. For the purpose of this work, we adapted the core model reconstructed in Section 3.2.1, by constraining exchange reactions boundaries according to wet-lab data. The outcomes of constraint-based simulations confirmed their consistency with experimental data, with particular regard to the non complementarity of the two investigated nutrients in terms of cells growth rate. However, we observed that this growth discrepancy is considerably narrowing by removing the specific constraint on the glutamate secretion rate that we previously imposed according to experimental data. A possible accumulation of unexploited and then secreted glutamate could occur under glutamine missing condition, resulting in a reduction of the biomass synthesis rate. Both transcriptomics data and constraint-based simulations led to ascribe this behaviour to a similar and low activity of the glutamine synthetase (GS) enzyme between the two investigated scenarios. GS enzyme is involved in converting cytosolic glutamate back to glutamine. This assumed activity level of GS could be probably sufficient in the standard condition, but not in the glutamine missing one, for processing the cytoplasmatic glutamate and avoiding its secretion in the extracellular environment. At the same time, this low level of activity of GS under glutamine deprivation could be enough to replenish the cytosolic pool of glutamine that is needed in multiple metabolic pathways related to growth, including nucleotides and amino acids synthesis. In support of this computational hypothesis, anti-variation of GS-catalyzed reaction and glutamate secretion fluxes emerged.

In Section 3.2.3, I reported the third application case of core modelling as mean for uncovering system-level properties of cancer metabolic rewiring. This work is focused on the reconstruction of a core model for the human central carbon metabolism, called ENGRO2. We compared ENGRO2 with a previous version presented in [31], called ENGRO1, in order to investigate the role of important elements previously not included, such as cell compartments, mitochondrial shuttles and essential amino acids metabolism. In addition, this work aims at exploring the heterogeneity of metabolic programs implemented by cancer cells to address specific perturbations on the network, such as gene knock out, variation in the extracellular environment composition, and the addition of

a given pathway to explore its role in the network. Although most of the computational outcomes are still under experimental validation, ENGRO2 model confirmed main ENGRO1 predictions presented in [31]. In particular, glucose and glutamine contribution for the maximization of biomass synthesis rate has been corroborated. However, a less negative effect on the growth rate resulted from ENGRO2 simulations under glutamine deprivation. This result is due to the presence in ENGRO2 network of essential amino acids, which revealed their differential ability to compensate for glutamine missing in the medium. Finally, the experimental evidence presented in [184, 185, 186] have been exploited to further validate our ENGRO2 model.

In Chapter 4, I presented new constraint-based methodologies to shift from the investigation of average population to an heterogeneous vision of cell population, where cell-to-cell variations are explicitly addressed. To deepen this issue, we considered that tumours evolve by acquiring a series of mutations over time, followed by selection processes for mutation providing a growth advantage for tumour progression. Overall, cancer populations are made of a heterogeneous mixture of cells, where tumour cells interact with each other, with stromal cells, and with their local microenvironment. Therefore, the cancer heterogeneity discussed in Chapter 3 extends to the intratumour level, where genetic and epigenetic factors together with trophic supply and variations in the tumour microenvironment contribute to generate this type of heterogeneity with a consequent emergence of a cancer-ecosystem view. According to this perspective, the outcome of the strategy adopted by a cancer cell to survive within the overall population does not depend only on its approach used for growing, but also on the strategies used by the other constituents of the population. Consequently, a better understanding of the interactions established within tumour populations could be used to hamper and potentially reverse tumour progression.

Based on the limits of classic constraint-based modelling concerning the simulation of an average cell of a population, in Sections 4.1.1 and 4.1.2, I presented a new strategy, called popFBA, focused on the reconstruction and simulation of cell population metabolism, by putting emphasis on the relationships established among their components. Thanks to this ecosystem view, this approach highlighted that a cooperative behaviour within the investigated population model together with heterogeneity in terms of adopted metabolic strategies, are consistent with the achievement of the optimal tumour biomass. Through popFBA strategy, we observed plasticity of population clones metabolism under different nutrients exchanged with plasma, or when a dishomogeneous distribution of oxygen is provided. In particular, among the explored scenarios, phenotypes

of the most proliferative clones result characterized by a recurrent consumption and consequent oxidation of extracellular lactate that is secreted in the tumour microenvironment by the less proliferative clones. This *in silico* outcome has been experimentally demonstrated by the "reverse Warburg effect" highlighted within human tumour populations as a symbiotic but also parasitic metabolic relationship between cancer and stromal cells [24, 25, 190].

Despite the ability of popFBA methodology to simulate multiple scenarios and capture interactions between cells, the countless combinations of flux distributions of individual population components make the solution to the problem undetermined. Consequently, the challenging task was to reduce the amount of putative combinations by adding further constraints on the individual components. For that purpose, in Section 4.2.1, I presented the novel computational framework for data integration, called Metabolic Reaction Enrichment Analysis (MaREA). MaREA aims at integrating RNA-Seq data into metabolic networks by calculating for each reaction the Reaction Activity Score (RAS) exploiting the corresponding gene-protein-reaction (GPR) rule. The usage of RAS scores instead of the only transcripts allowed to characterize transcriptional deregulations of metabolic reactions under different conditions. Furthermore, this strategy offered the possibility of ranking reactions according to their activity variation, and visualizing the most critical ones directly on metabolic networks. MaREA revealed its ability in discriminating normal and cancer samples, by reproducing well-known features of cancer deregulation and generating new hypotheses. By applying this methodology to two distinct datasets of The Cancer Genome Atlas database relative to lung and breast tumours, cancer patients can be, in an unsupervised way, stratified in distinct clusters characterized by similar metabolic activity. In this way, subgroups of patients differing in terms of survival expectancy can be identified. In the specific case of breast cancer dataset, MaREA identified just two clusters with significantly different prognosis. Nevertheless, these clusters showed a considerable overlap with signatures of two of the molecular subtypes deriving from breast cancer classification [208]. Finally, we performed a comparison of MaREA prognostic power by using a well curated core model focused on central carbon metabolism and a complete genome-wide metabolic network. In the first case, an improvement of its prognostic power has been observed, because of the model curation in terms of GPR rules associated to metabolic reactions.

By combining the potentialities revealed by MaREA and popFBA methodologies, we developed a new approach called single-cell Flux Balance Analysis (scFBA) and introduced in Section 4.2.2. scFBA has been implemented to address the previously discussed task of reducing the amount of individual flux

distributions within population models. scFBA aims at integrating single-cell RNA-seq data (scRNA-seq) within population models by exploiting the computation of RAS scores. In this way, scRNA-seq data act as further constraints on flux boundaries of distinct cells within the population, in order to identify the possible combinations of single-cell steady states. Thanks to this approach, the space of optimal solutions resulted efficiently reduced as compared to the situation depicted by popFBA. Indeed, solutions where each cell in the population can alone contribute to total biomass of the entire tumour mass have been prevented. Moreover, the conversion of single-cell transcriptomes into single-cell fluxomes also allowed to identify metabolic clusters of cells characterized by different growth rates, which can be exploited to better investigate intratumour heterogeneity. Finally, similarly to popFBA, scFBA methodology revealed the ability to identify possible interactions among cells within the population. As opposed to its predecessor, scFBA does not allow to characterize the interactions network of all the possible cell populations, but to capture interactions between cells of a unique and specific population.

Future developments of this thesis work will involve the utilization of the computational tools here implemented for additional applications.

Firstly, I will focus on making the reconstruction pipeline of genome-scale metabolic networks discussed in Chapter 3 as fully automated, by acting on the few steps that still require a manual intervention. In this way, the method can be used by further reducing the computational time for obtaining the final network. Moreover, I will also focus on a variant of this pipeline. This new version will exploit as input clinical data and, more in detail, a list of specific genes and/or metabolites whose expression or level resulted altered from the medical record of a specific patient. The final outcome will be the entire network connecting the above mentioned inputs to understand the set of pathways where these genes or metabolites come from and are involved. In this way, this tool could assist the clinician in understanding how these elements can contribute to the emergence of specific dismetabolisms, such as cancer, but also diabetes, arteriosclerosis, alzheimer and parkinson disease.

As discussed in Section 2.5.2, the solution space computed through FBA, and including all the possible functional states achieved by a given metabolic network, is expected to decrease following the application of constraints. One among the most frequently imposed constraints corresponds to the reaction directionality, which is generally derived from multiple sources, including already existing models and metabolic pathway databases. However, inconsistencies between these sources often is in danger of impairing simulation outcomes [289].

Basically, all biochemical reactions are reversible and their direction is strictly dependent on the corresponding Gibbs energy. The Gibbs energy of a reaction is based on the standard Gibbs energies of formation and on the concentration values of its reactants. Consequently, changes in reactants concentration can reverse the direction of a given reaction with the by, in turn, influencing simulation the output of computational simulations. Although the biological validity of *in silico* outcomes from all the presented works in this thesis, the relevance of this information within metabolic network leads us to consider, as future perspective of this work, the integration of thermodynamic constraints on the feasible set of biologically active pathways within our model. Over the years, several approaches have been exploited to introduce thermodynamic constraints on the network and reduce the feasible bounds of reaction fluxes. Among them, thermodynamics-based flux analysis (TFA) is the most recent one [290]. TFA couples reaction directionalities in the network under investigation to the integration of quantitative metabolomics data. In this way, only flux distributions free from thermodynamically infeasible reactions are generated, while providing reactions flux values together with metabolites activities ranges and reactions free energy change. In addition, the ability to estimate displacement of reactions from thermodynamic equilibrium is also supplied. Given the potententialities of TFA, as shown in [290], I will exploit this approach to further reduce to reduce the feasible flux solution space of FBA problems and increase the predictive ability of our *in silico* models.

As extensively discussed throughout this thesis, an overall rewiring of metabolism in cancer cells is essential to support their ability to growth and proliferate at higher rate compared to their normal counterparts. However, although the positive outcomes shown in Chapters 3 and 4 resulting from the modelling of cancer metabolic networks, metabolism does not constitute the unique contributor to the determination of cell phenotype. The high cell growth rate characterizing cancer cells, implies the necessity of sufficient amounts of nutrients and free energy to support the synthesis of all the building blocks needed to form biomass. Several studies highlighted that this unusual consumption rate of nutrients is regulated by signal transduction pathways. When cells are instructed to proliferate by signaling processes initiated by extracellular growth factors, cells increase their nutrients uptake rate, and reprogram their metabolic pathways for the purpose of growing. However, a unidirectional relationship between cellular signaling and metabolism could lead to rapid cell collpse. Indeed, following particular conditions, the activity of a specific enzyme or the availability of key nutrients in the extracellular environment could be limited. When this situation occurs, a cell needs to sense its overall status and, consequently, ex-

erts a feedback control on specific signal transduction networks to slow down its growth rate if the metabolic state cannot support production of biomass. This feedback control from cell metabolism may be exercised by means of a series of post-translational modifications, including acetylation, glycosylation and methylation, of the output protein of signal transduction cascades. This bidirectional dialogue allows cells to coordinate their growth according to their metabolic activity [291]. In view of the above, another future development of this thesis work will regard the development of a computational strategy for integrating metabolic networks with signal transduction networks. In this way, it will be possible to study the crosstalk between these two big research fields, by continuing to consider the issue of interpatient heterogeneity for the purpose of developing a more personalized medicine. Moreover, in this context, the whole body physiology-based pharmacokinetic (PB-PK) modelling could be used as additional basis to expand for investigating the crosstalk signaling-metabolism. PB-PK modelling, in recent years, has been increasingly applied in the process of drug development in order to generate predictions about the impact caused by the genetic variability on the drug pharmacokinetic [292]. In particular, the four processes of absorption, distribution, metabolization and excretion of endogenous and exogenous compounds are described from a quantitative point of view within mammalian organisms. The overall system is organized as a series of compartments corresponding to the different organs of body, each one characterized by a specific volume (or weight) and blood flow rate. All these compartments result as communicating through the circulating blood system [293]. Contrary to pharmacokinetics, pharmacodynamics aims at investigating the action mechanism of a specific drug and its effect on the organism. By combining these two levels, namely pharmacokinetics and pharmacodynamics, the final objective could be obtaining an increasingly personalized approach based on the genetic of the patient. Hopefully, this strategy could allow to obtain more precise indications about the individual or the cocktail of targets to counteract. Moreover, it could be possible to obtain more accurate information about the effectiveness of the drug in the patient under consideration, and regarding the drug dose range in order to obtain positive effects against the progression of the disease.

Appendices

Appendix A

List of abbreviations

1,3BPGA: 1,3-Bisphospho-D-glycerate
2,3BPGA: 2,3-Bisphospho-D-glycerate
10-formyl-THF: 10-Formyltetrahydrofolate
2aminoadp: 2-aminoadipate
2PG: 2-phospho-D-glycerate
2kmb: 2-Oxo-4-methylthiobutanoic acid
2m3oprop: 2-Methyl-3-oxopropanoic acid
2obut: 2-Ketobutyric acid
2oxoadp: 2-oxoadipate
3hanthrn: 3-hydroxyanthranilate
3hb: 3-Hydroxybutyric acid
3hbCoA: 3-Hydroxybutyryl-CoA
3hoxisobCoA: 3-Hydroxyisobutyryl Coenzyme A
3hoxisob: 3-Hydroxyisobutyric acid
3mgCoA: 3-Methylglutaconyl-CoA
3mob: α-Ketoisovaleric acid
3mop: 3-Methyl-2-Oxovaleric Acid
3PG / 3PGA: 3-phospho-D-glycerate
3PPyr / PHP: 3-phosphonooxypyruvate
3PSer / 3PSER: 3-phosphoserine
3sulfinoAla: 3-Sulfinoalanine
3sulfinoPyr: 3-sulfinylpyruvic acid
4abut: gamma-Aminobutyric acid
4abutn: 4-Aminobutyraldehyde

4izp: 4-Imidazolone-5-propionic acid
4mlacac: Maleylacetoacetic acid
4fumacac: 4-Fumarylacetoacetic acid
4mop: 4-Methyl-2-oxopentanoic acid
6PDG / 6PG: 6-phospho-D-gluconate
6Pgl/6PGL: glucono-1,5-lactone-6-phosphate
5FORTHF: 5-Formiminotetrahydrofolic acid
5mdr1p: 5-Methylthioribose 1-phosphate
5mdru1p: 5-Methylthioribulose 1-phosphate
5mta: 5'-Methylthioadenosine
5meTHF: 5-methyltetrahydrofolate
5,10meTHF / meTHF: 5,10-Methenyltetrahydrofolate
5,10mTHF / mTHF: 5,10-Methylenetetrahydrofolate
AcAc: Acetoacetic acid
AcAcACP: acetoacetyl-ACP
AcAcCoA: acetoacetyl-CoA
AcACP: acetyl-ACP
AcCoA: acetyl-CoA
acetoAcCoA: Acetoacetyl-CoA
ACP: acyl carrier protein
ADP: adenosine 5'-diphosphate
AKG: α-ketoglutarate
Ala: alanine
am6sa: 2-aminomuconate semialdehyde ametam: S-Adenosylmethioninamine
AMP: adenosine 5'-monophosphate
amuco: 2-Aminomuconic acid
anth: anthranilate
Arg: arginine
ArgSuc / ArgSucc: argininosuccinate
Asn: asparagine
Asp: aspartate
ATP: adenosine 5'-triphosphate
BRCA: Breast Cancer
B-ala: beta-alanine
Carn: carnitine
carn: carnosine
CDP: cytidine 5'-diphosphate
Chol: cholesterol
Ci: citrulline

Cit: citrate
CMP: cytidine 5'-monophosphate
cmusa: 2-amino-3-carboxymuconate semialdehyde
CO_2: carbon dioxide
CoA: coenzyme A
CP: carbamoyl-phosphate
CreatineP: Phosphocreatine
crotCoA: crotonyl-CoA
CTP: cytidine 5'-triphosphate
cyan: Hydrogen cyanide
Cys: cysteine
CySS: cystine
Cyst / cysTh: cystanthionine
Cytc-ox / CytC-Fe3+: ferricytochrome c
Cytc-red / CytC-Fe2+: ferrocytochrome c
D-lac: D-lactate
dADP: 2'-deoxyadenosine 5'-diphosphate
dAMP: 2'-deoxyadenosine 5'-monophosphate
dCDP: 2'-deoxycytosine 5'-diphosphate
dCMP: 2'-deoxycytosine 5'-monophosphate
dGDP: 2'-deoxyguanosine 5'-diphosphate
dGMP: 2'-deoxyguanosine 5'-monophosphate
DHAP: dihydroxyacetone phosphate
dhbpt: 4a-Carbinolamine tetrahydrobiopterin
DHF: 7,8-Dihydrofolate
dkmpp: 5-(methylthio)-2,3-Dioxopentyl phosphate
dTMP: 2'-Deoxythymidine-5'-monophosphate
dUDP: 2'-Deoxyuridine-5'-diphosphate
dUMP: 2'-Deoxyuridine-5'-monophosphate
Ery4P: erythrose-4-phosphate
F16BP: fructose 1,6-bisphosphate
F6P: fructose 6-phosphate
FAD: flavin adenine dinucleotide
$FADH_2$: flavin adenine dinucleotide reduced
FBA: Flux Balance Analysis
FDR: False Discovery Rate
for: formate
forglu: Formiminoglutamic acid
fPP: farnesyl diphosphate

Fum: fumarate
G6P: glucose 6-phosphate
GA: guanidinoacetate
GA3P/GAP: glyceraldehyde 3-phosphate
GC: gamma-Glutamylcysteine
GDP: guanosine 5'-diphosphate
Glc: glucose
Gln: glutamine
Glu: glutamate
GluSA: glutamate 5-semialdehyde
glutCoA: glutaryl-CoA
Gly: glycine
Gly3P: Glycerol 3-phosphate
GMP: guanosine 5'-monophosphate
GSEA: Gene Set Enrichment Analysis
GSH: Glutathione
GSSG: Glutathione disulfide
GPR: Gene-Protein-Reaction
GTP: guanosine 5'-triphosphate
H: proton
H_2O: water
H_2O_2: Hydrogen peroxide
HCO_3^-: hydrogencarbonate
hcarn: Homocarnosine
HCys: Homocysteine
hgentis: Homogentisic acid
HGNC: HUGO Gene Nomenclature Committee
His: histidine
hLkynr: 3-hydroxy-L-kynurenine
HMGCoA: hydroxymethylglutaryl-CoA
hpp: 4-Hydroxyphenylpyruvic acid
Ile: Isoleucine
IMP: 2'-inosine-5'-monophosphate
ippPP: isopentenyl diphosphate
Iso: isocitrate
isobCoA: Isobutyryl-CoA
Isocit: isocitrate
ivCoA: Isovaleryl-CoA
KS: Kolmogorov-Smirnov

L-lac: L-lactate
Lact / LACT: lactate
Leu: Leucine
Lfmkynr: N-formyl-L-kynurenine
Lkynr: L-kynurenine
Lys: Lysine
Mal: Malate
MalACP: malonyl-ACP
MalCoA: malonyl-CoA
MaREA: Metabolic Reaction Enrichment Analysis
mcrCoA: Methacrylyl-CoA
mcrtCoA: 3-Methylcrotonyl-CoA
mercPyr: mercaptopyruvate
Met: Methionine
mmal: Methylmalonic acid
mmalCoA: Methylmalonyl-CoA
NAD: nicotinamide adenine dinucleotide
NADH: nicotinamide adenine dinucleotide reduced
NADP: nicotinamide adenine dinucleotide phosphate
NADPH: nicotinamide adenine dinucleotide phosphate reduced
NH_3: ammonia
O_2: oxygen
$O_2{}^2$: superoxide anion
OAA: oxaloacetate
Orn: ornithine
P5C: 1-pyrroline-5-carboxylate
Palm: palmitate
PalmCarnitine / PalmCarn: palmitoylcarnitine
PalmCoA: palmitoyl-CoA
PARADIGM: PAthway Recognition Algorithm using Data & Integration on Genomic Models
PEP: phosphoenolpyruvate
Phe: Phenylalanine
P_i: phosphate
PP_i: diphosphate
Pro: proline
propCoA: Propionyl-CoA
PRPP: phosphoribosyl pyrophosphate
Putr: putrescine

Pyr: pyruvate
Q/UQ: ubiquinone
QH_2/UQH_2: ubiquinol
R5P: ribose 5-phosphate
RAS: Reaction Activity Score
RmmalCoA: (R)-methylmalonyl-CoA
RPKM: Reads per Kilobase per Million mapped reads
Ru5P: ribulose 5-phosphate
saccrp: Saccharopine
SAM: S-Adenosylmethionine
SAH: S-Adenosylhomocysteine
SBML: Systems Biology Markup Language
Sed1,7BP: sedoheptulose 1,7-bisphosphate
Sed7P: sedoheptulose 7-phosphate
Ser: serine
Succ / Suc: succinate
SuCoA / SucCoA: succinyl-CoA
TCGA: The Cancer Genome Atlas
tcynt: Thiocyanate
thbpt: Tetrahydrobiopterin
THF: 5,6,7,8-Tetrahydrofolate
Thr: Threonine
Trp: Tryptophan
Tyr: tyrosine
UDP: uridine 5'-diphosphate
UMP: uridine 5'-monophosphate
urcan: Urocanic acid
UTP: uridine 5'-triphosphate
Val: Valine
VMH: Virtual Metabolic Human
Xil5P: xylulose-5-phosphate

Appendices

Appendix B

ENGRO2 model metabolic maps

In all the reported maps, color and width of each arrow is proportional to the corresponding flux value from parsimonious FBA or flux variability analysis according to the chromatic scale on the top of the figure. Dashed and gray arrows refers to reactions whose flux value is null. A list of the abbreviations used in each map is provided in the Appendix A.

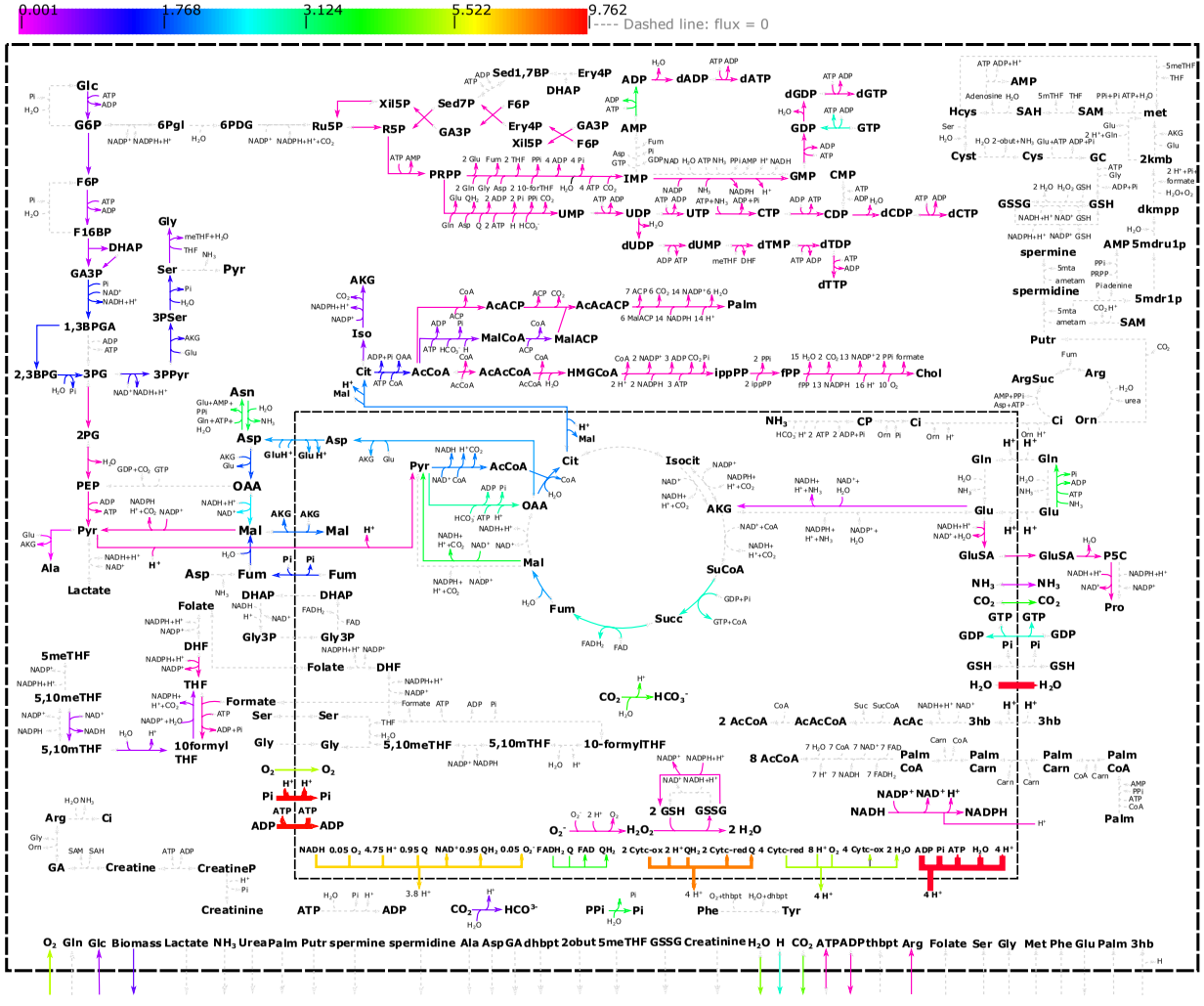


Figure B.1: Parsimonious FBA of ENGRO2 model when isoleucine, leucine and valine are provided in the medium under glutamine deprivation (first network part).

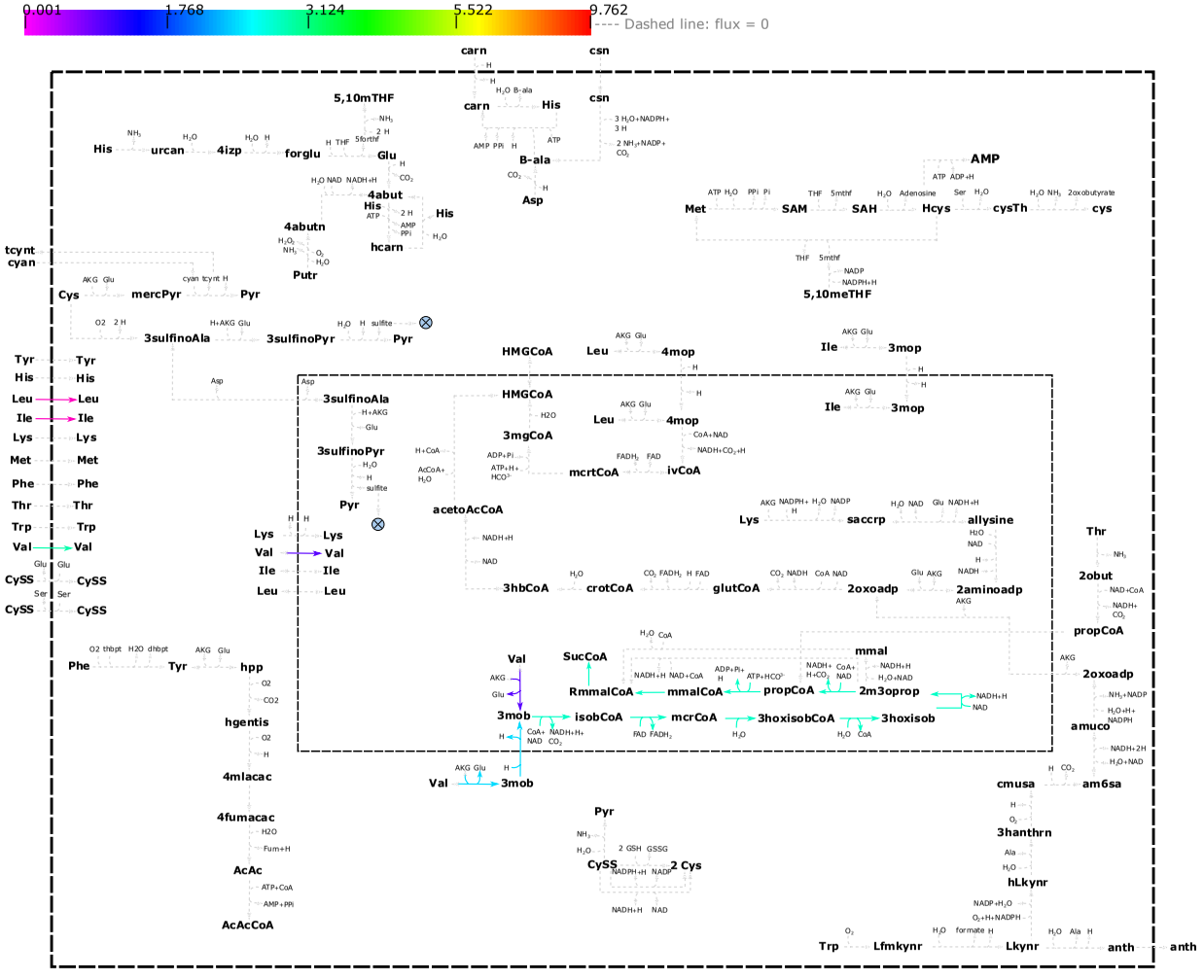


Figure B.2: Parsimonious FBA of ENGRO2 model when isoleucine, leucine and valine are provided in the medium under glutamine deprivation (second network part).

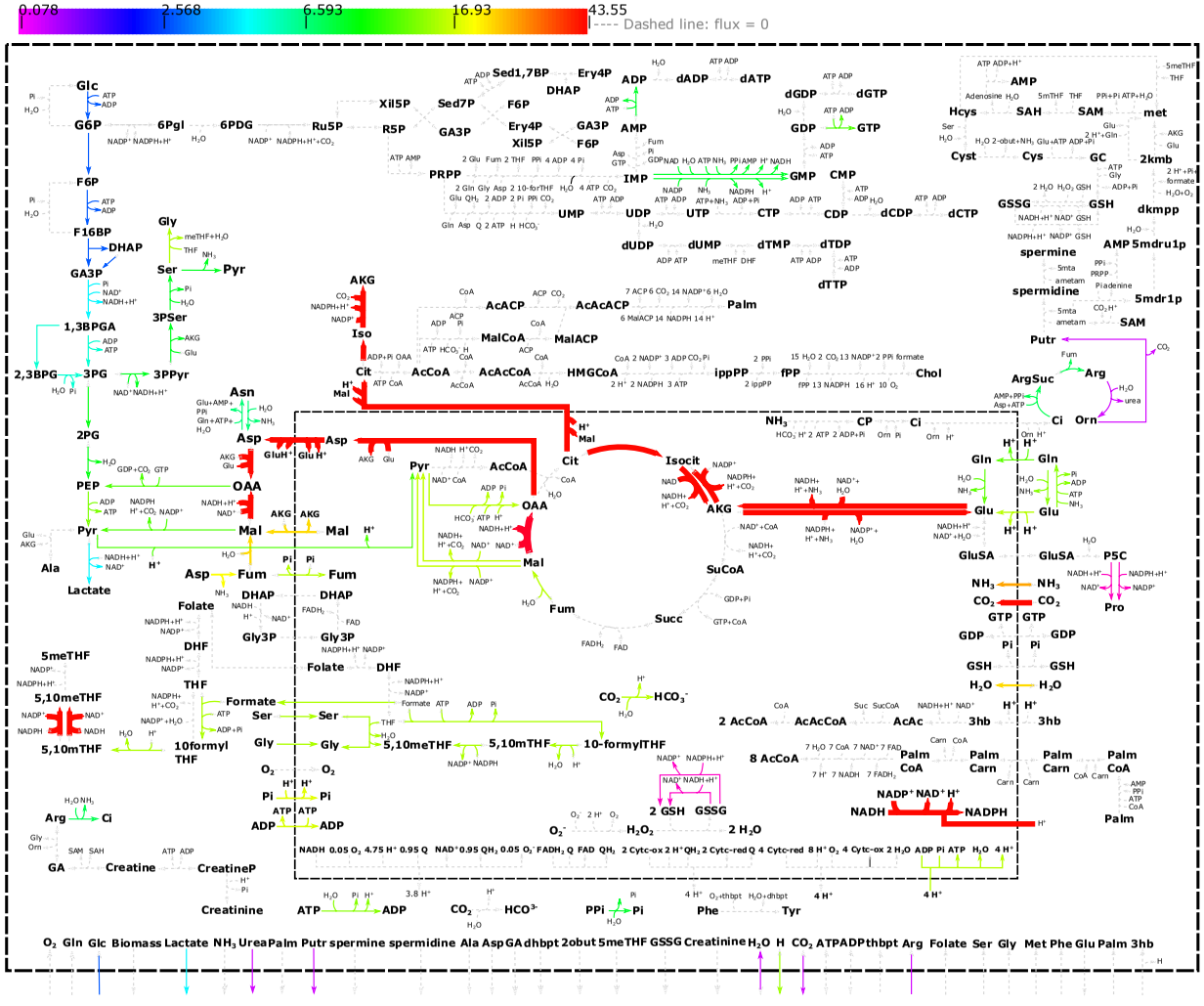


Figure B.3: Flux variability analysis of ENGRO2 model when isoleucine, leucine and valine are provided in the medium under glutamine deprivation (first network part).

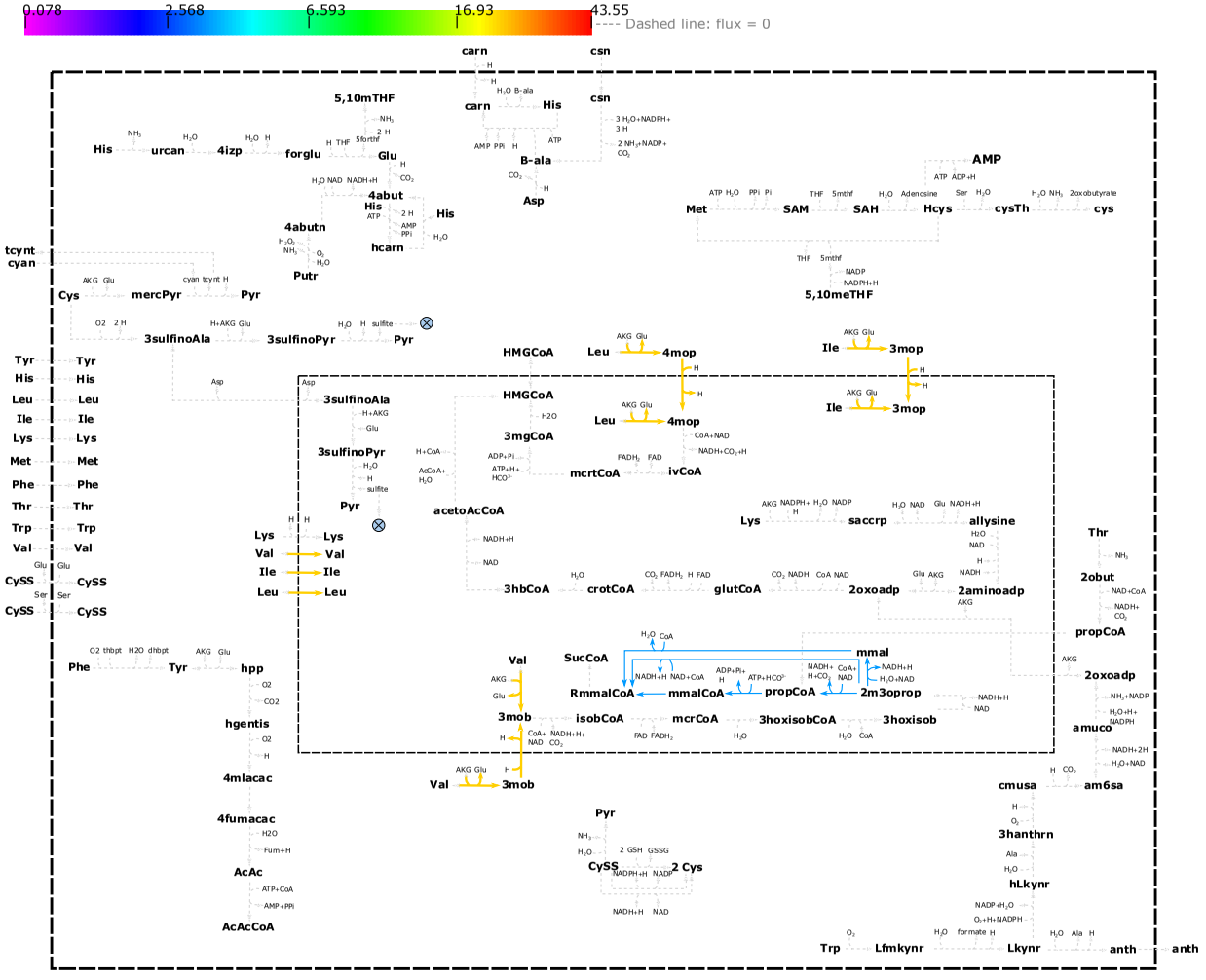


Figure B.4: Flux variability analysis of ENGRO2 model when isoleucine, leucine and valine are provided in the medium under glutamine deprivation (second network part).

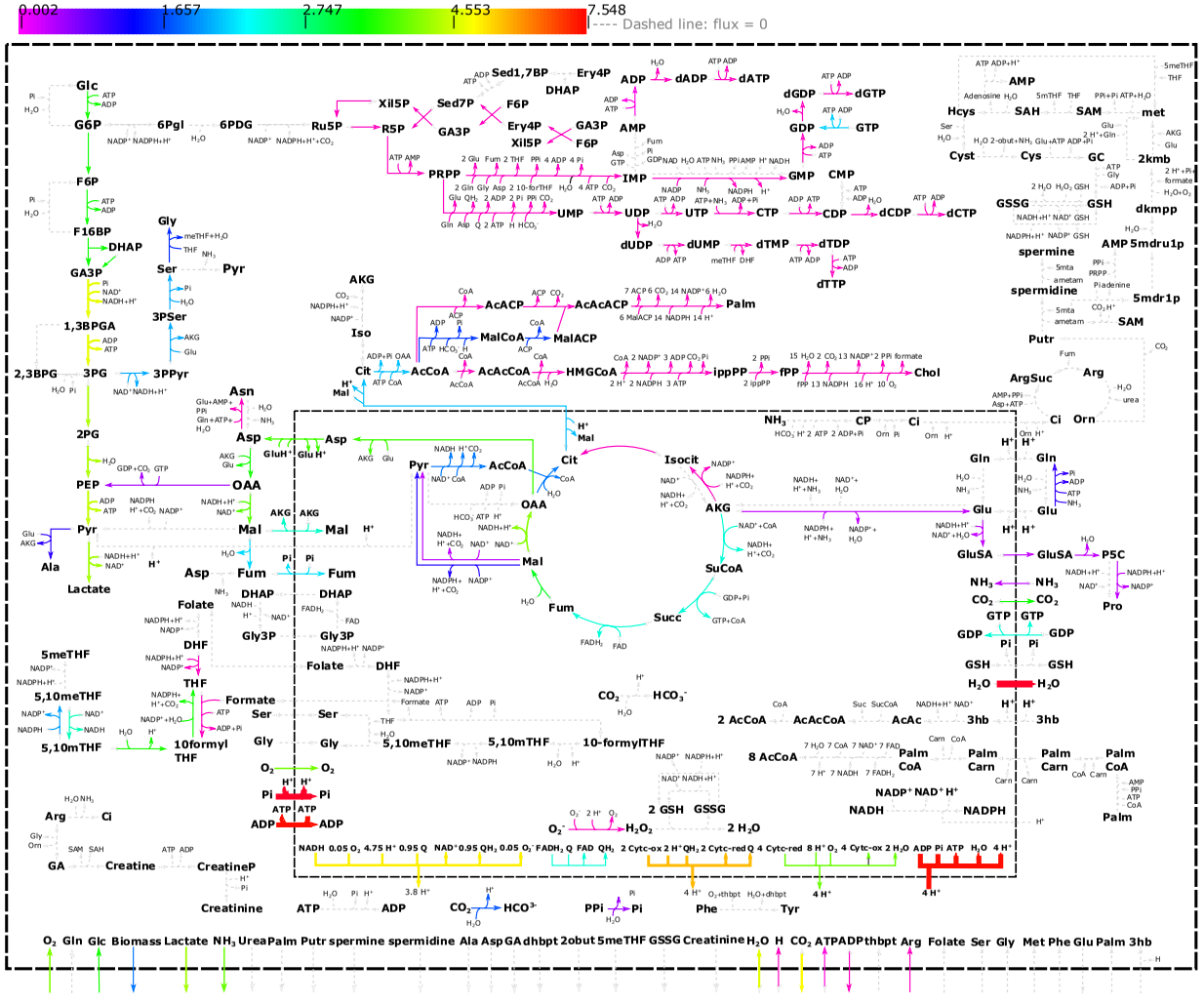


Figure B.5: Parsimonious FBA of ENGRO2 model when histidine and lysine are provided in the medium under glutamine deprivation (first network part).

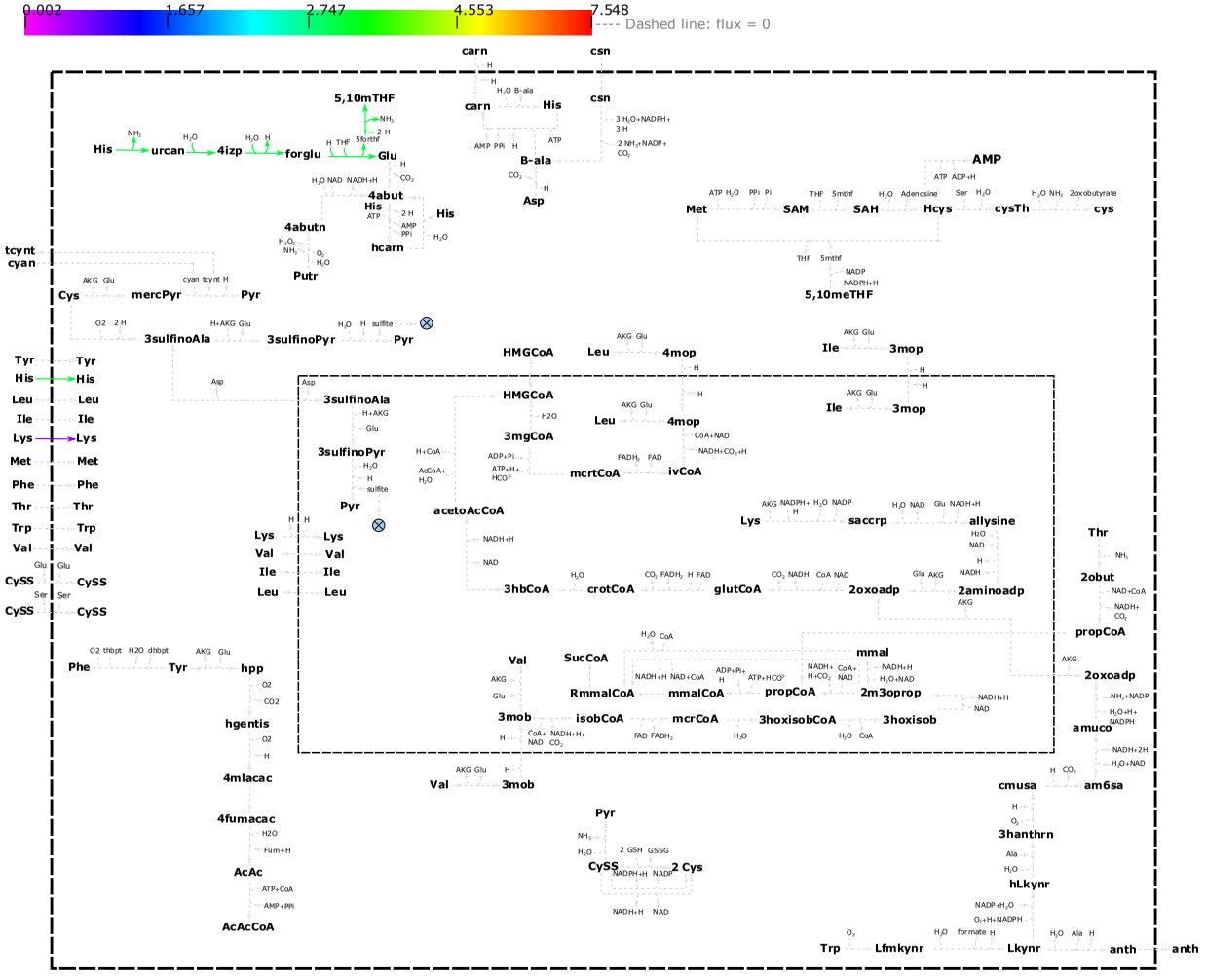


Figure B.6: Parsimonious FBA of ENGRO2 model when histidine and lysine are provided in the medium under glutamine deprivation (second network part).

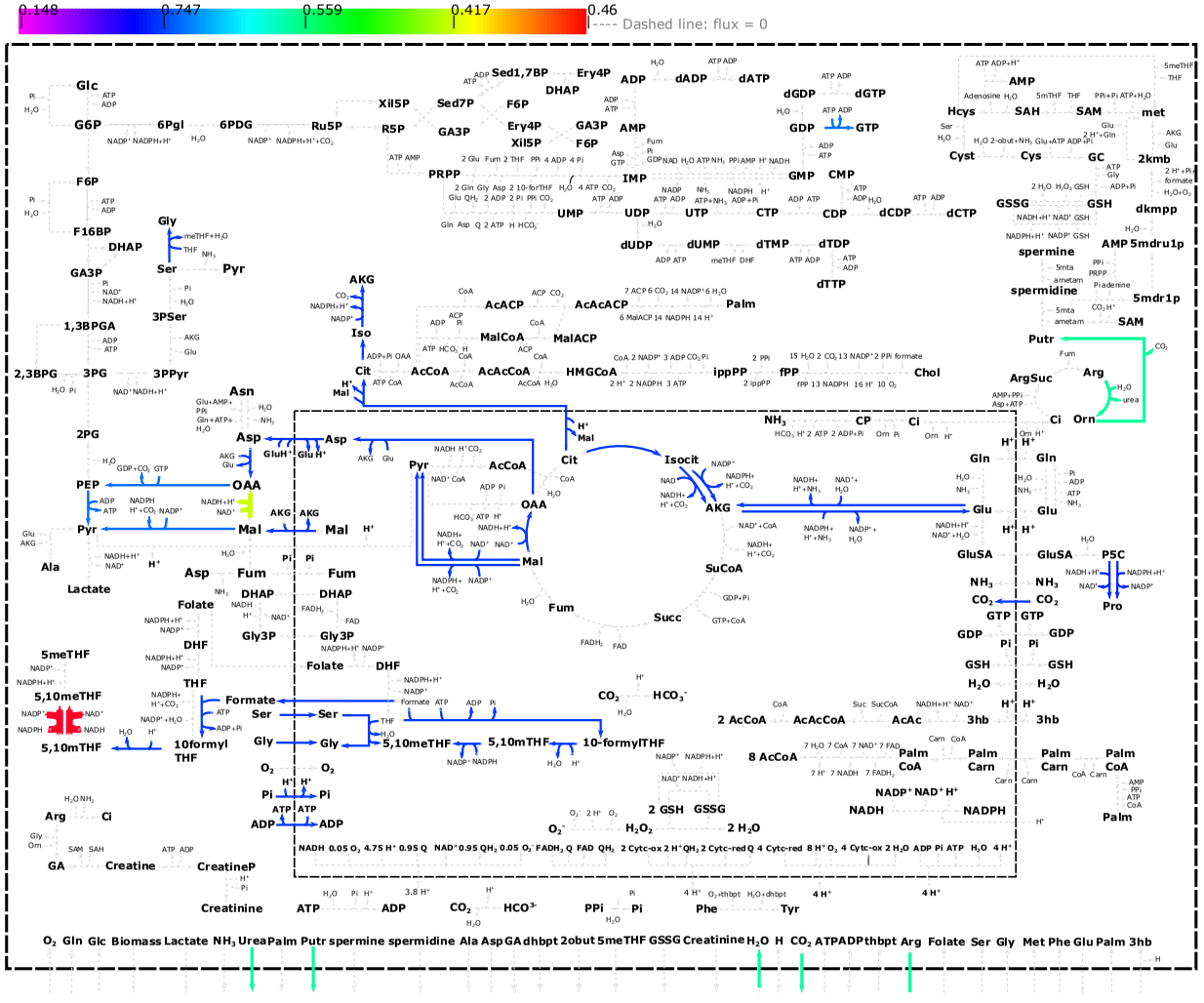


Figure B.7: Flux variability analysis of ENGRO2 model when histidine and lysine are provided in the medium under glutamine deprivation (first network part).

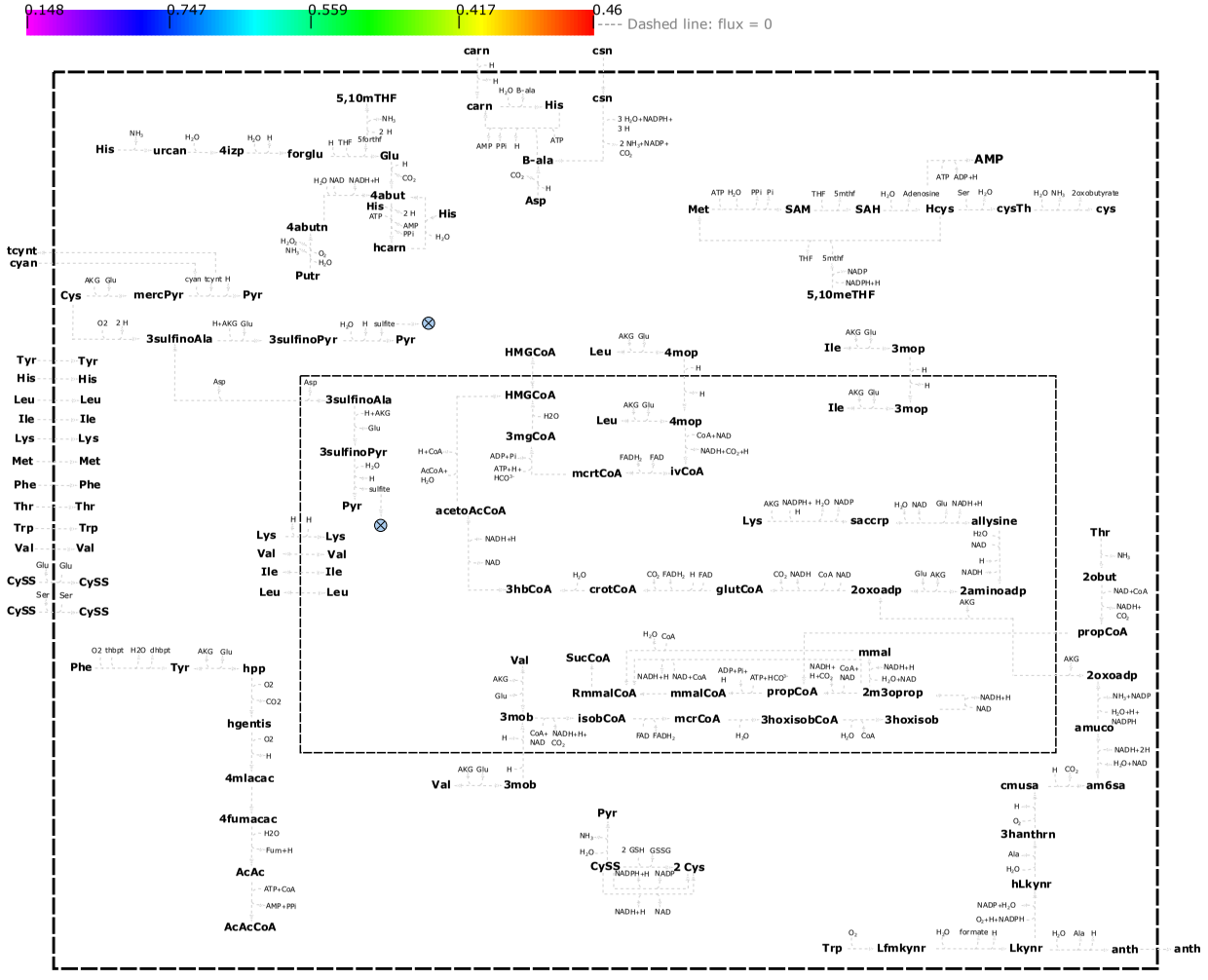


Figure B.8: Flux variability analysis of ENGRO2 model when histidine and lysine are provided in the medium under glutamine deprivation (second network part).

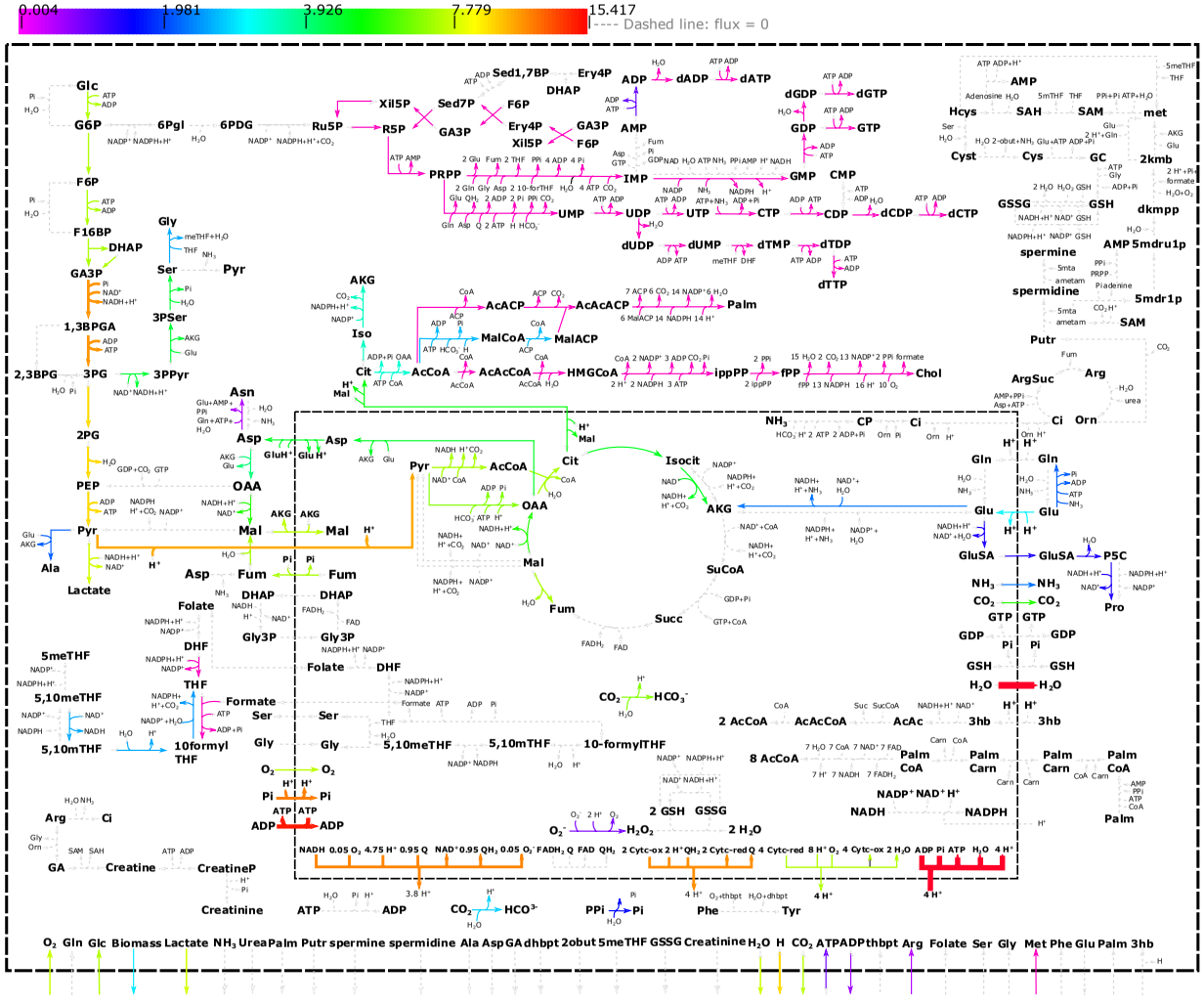


Figure B.9: Parsimonious FBA of ENGRO2 model when methionine and cystine are provided in the medium under glutamine deprivation (first network part).

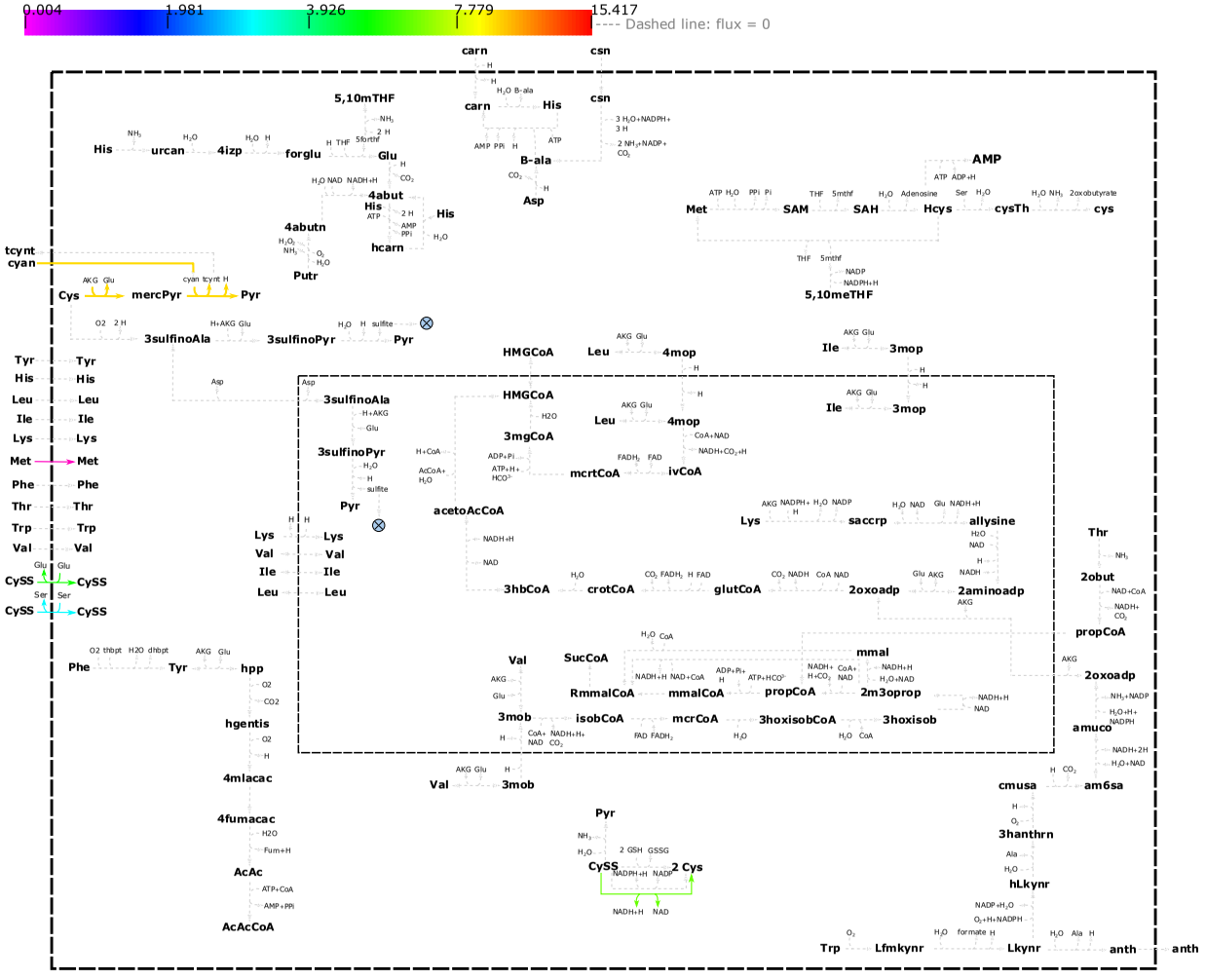


Figure B.10: Parsimonious FBA of ENGRO2 model when methionine and cystine are provided in the medium under glutamine deprivation (second network part).

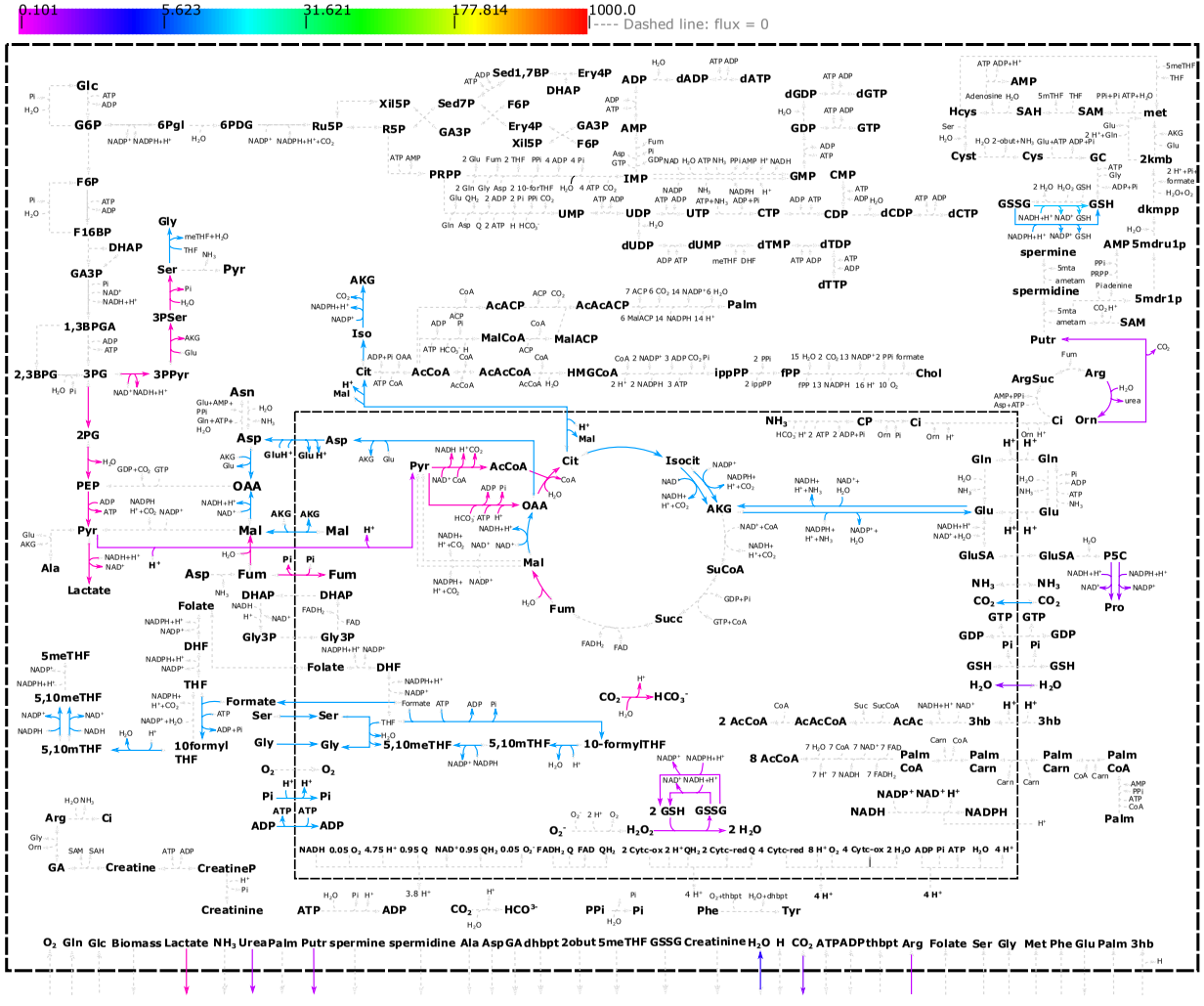


Figure B.11: Flux variability analysis of ENGRO2 model when methionine and cystine are provided in the medium under glutamine deprivation (first network part).

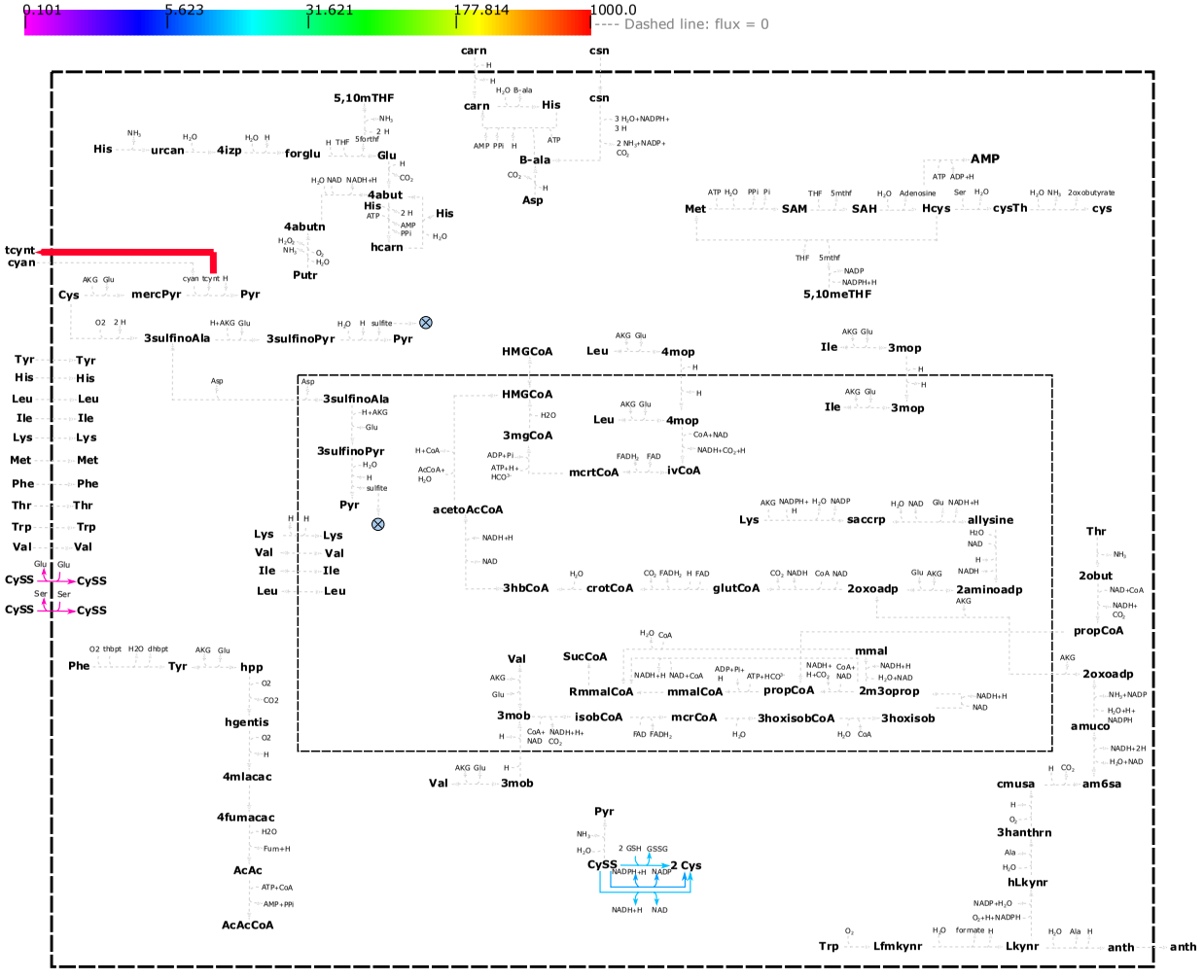


Figure B.12: Flux variability analysis of ENGRO2 model when methionine and cystine are provided in the medium under glutamine deprivation (second network part).

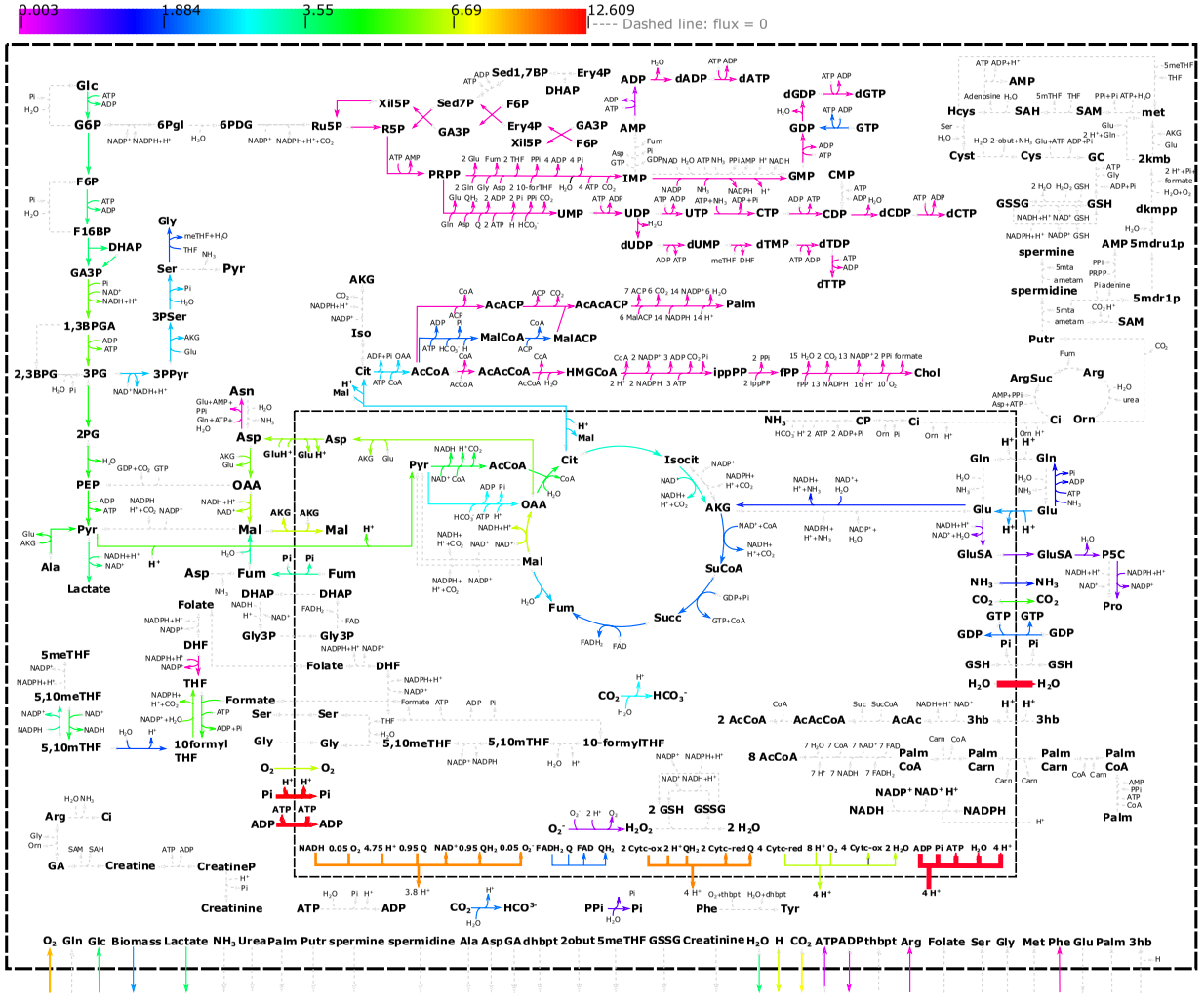


Figure B.13: Parsimonious FBA of ENGRO2 model when phenylalanine, threonine, tyrosine and tryptophan are provided in the medium under glutamine deprivation (first network part).

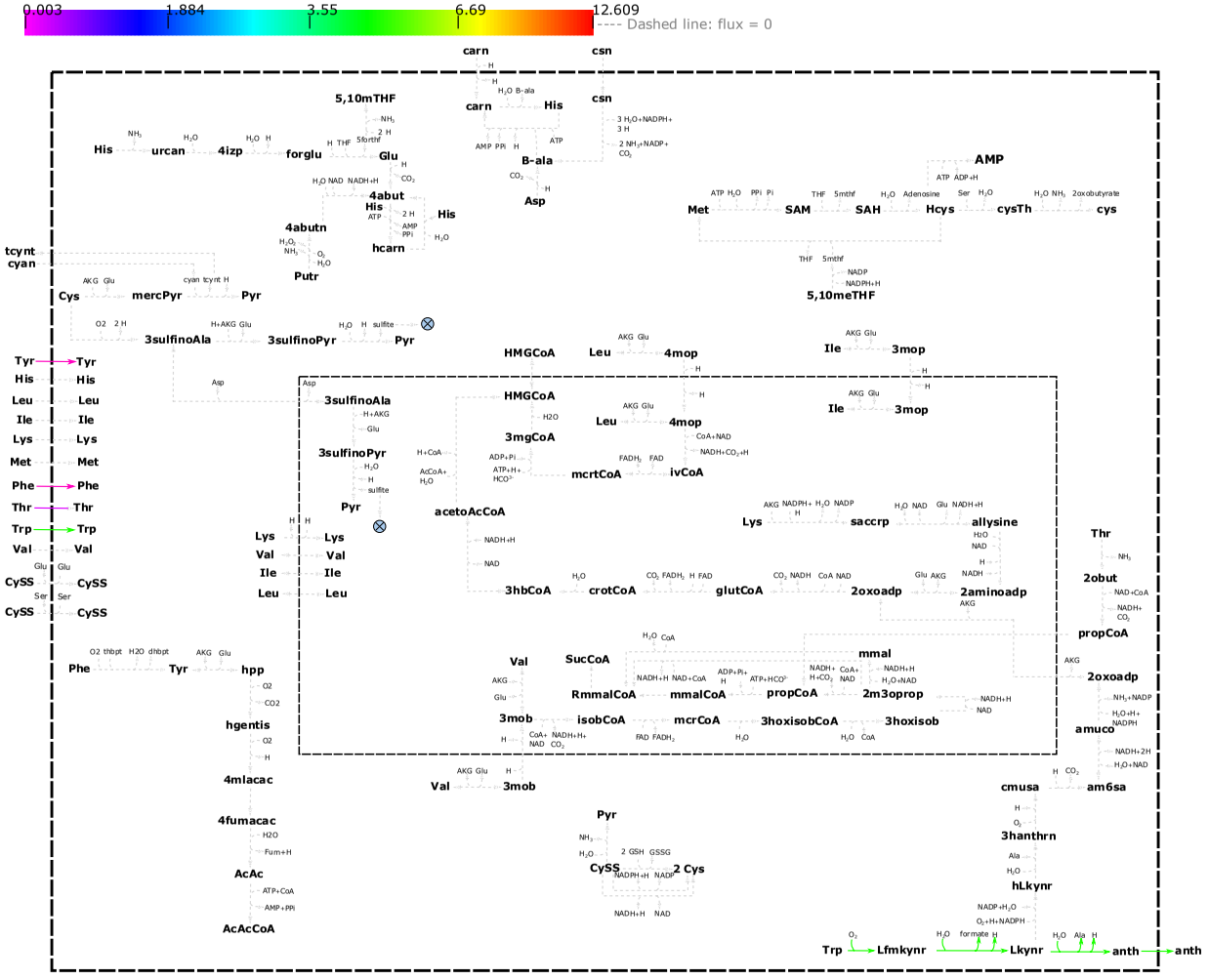


Figure B.14: Parsimonious FBA of ENGRO2 model when phenylalanine, threonine, tyrosine and tryptophan are provided in the medium under glutamine deprivation (second network part).

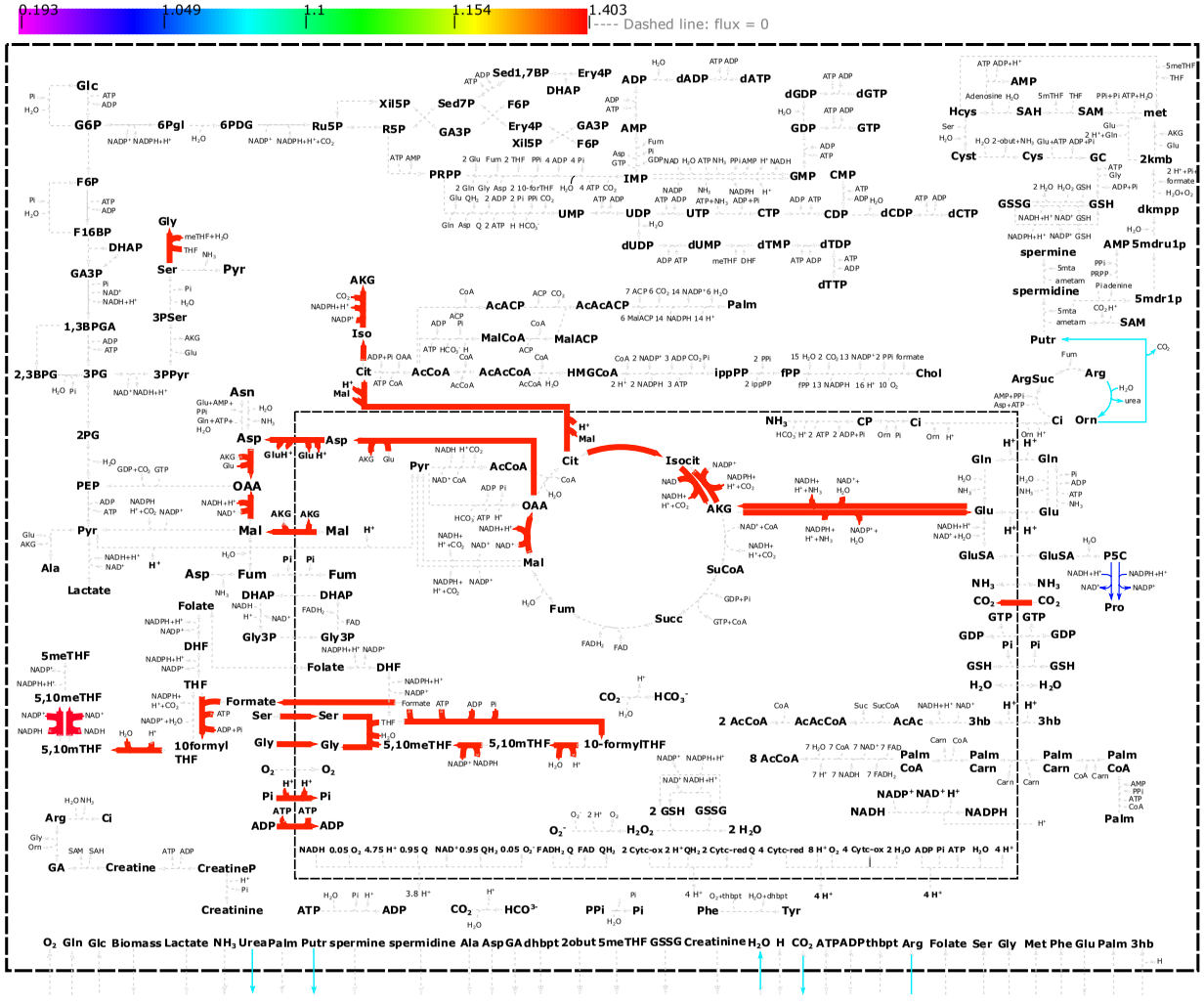


Figure B.15: Flux variability analysis of ENGRO2 model when phenylalanine, threonine, tyrosine and tryptophan are provided in the medium under glutamine deprivation (first network part).

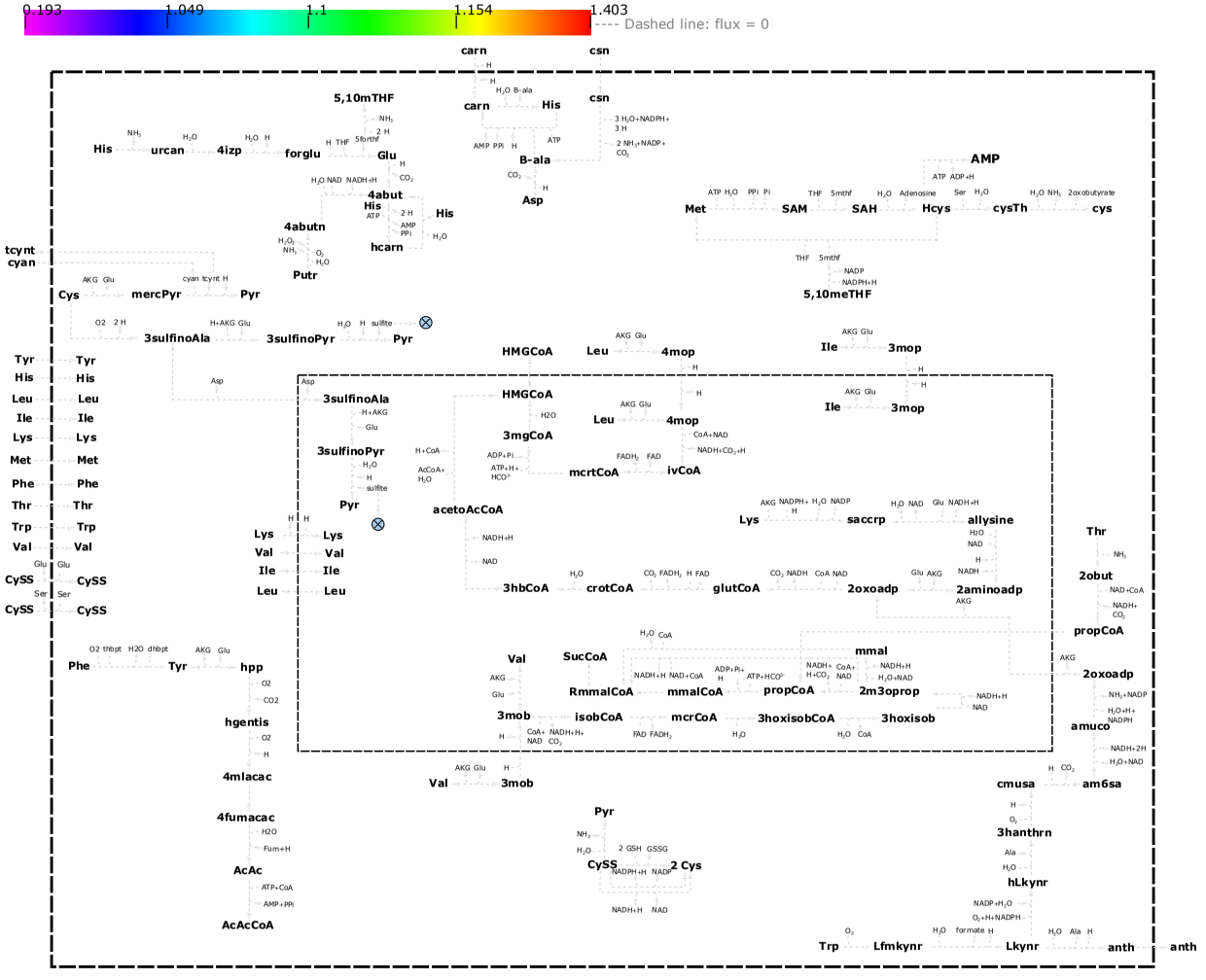


Figure B.16: Flux variability analysis of ENGRO2 model when phenylalanine, threonine, tyrosine and tryptophan are provided in the medium under glutamine deprivation (second network part).

Bibliography

[1] Gary J Patti, Oscar Yanes, and Gary Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature reviews Molecular cell biology*, 13(4):263, 2012.

[2] MS Monteiro, M Carvalho, ML Bastos, and P Guedes de Pinho. Metabolomics analysis for biomarker discovery: advances and challenges. *Current medicinal chemistry*, 20(2):257–271, 2013.

[3] Ben Lehner. Modelling genotype–phenotype relationships and human disease with genetic interaction networks. *Journal of Experimental Biology*, 210(9):1559–1566, 2007.

[4] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101, 2004.

[5] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206, 2002.

[6] Néstor V Torres and Guido Santos. The (mathematical) modeling process in biosciences. *Frontiers in genetics*, 6:354, 2015.

[7] Hiroaki Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[8] Jens Nielsen. Systems biology of metabolism: a driver for developing personalized and precision medicine. *Cell metabolism*, 25(3):572–579, 2017.

[9] Bret H Goodpaster and Lauren M Sparks. Metabolic flexibility in health and disease. *Cell metabolism*, 25(5):1027–1036, 2017.

[10] Emrah Şimşek and Minsu Kim. The emergence of metabolic heterogeneity and diverse growth responses in isogenic bacterial cells. *The ISME journal*, 12(5):1199, 2018.

[11] Steven J Altschuler and Lani F Wu. Cellular heterogeneity: do differences make a difference? *Cell*, 141(4):559–563, 2010.

[12] Ralph J DeBerardinis and Navdeep S Chandel. Fundamentals of cancer metabolism. *Science advances*, 2(5):e1600200, 2016.

[13] Parul Singla, Animesh Bardoloi, and Anuj A Parkash. Metabolic effects of obesity: a review. *World journal of diabetes*, 1(3):76, 2010.

[14] Claudio Procaccini, Marianna Santopaolo, Deriggio Faicchia, Alessandra Colamatteo, Luigi Formisano, Paola De Candia, Mario Galgani, Veronica De Rosa, and Giuseppe Matarese. Role of metabolism in neurodegenerative disorders. *Metabolism*, 65(9):1376–1390, 2016.

[15] Jens Nielsen. Systems biology of metabolism. *Annual review of biochemistry*, 86:245–275, 2017.

[16] David L Nelson, Albert L Lehninger, and Michael M Cox. *Lehninger principles of biochemistry*. Macmillan, 2008.

[17] Jason R Cantor and David M Sabatini. Cancer cell metabolism: one hallmark, many faces. *Cancer discovery*, 2(10):881–898, 2012.

[18] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[19] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[20] Yixin Yao and Wei Dai. Genomic instability and cancer. *Journal of carcinogenesis & mutagenesis*, 5, 2014.

[21] Catherine C Park, Mina J Bissell, and Mary Helen Barcellos-Hoff. The influence of the microenvironment on the malignant phenotype. *Molecular medicine today*, 6(8):324–329, 2000.

[22] Daruka Mahadevan and Daniel D Von Hoff. Tumor-stroma interactions in pancreatic ductal adenocarcinoma. *Molecular cancer therapeutics*, 6(4):1186–1197, 2007.

[23] L Kopfstein and G Christofori. Metastasis: cell-autonomous mechanisms versus contributions by the tumor microenvironment. *Cellular and Molecular Life Sciences CMLS*, 63(4):449–468, 2006.

[24] Tania Fiaschi, Alberto Marini, Elisa Giannoni, Maria L Taddei, Paolo Gandellini, Alina De Donatis, Michele Lanciotti, Sergio Serni, Paolo Cirri, and Paola Chiarugi. Reciprocal metabolic reprogramming through lactate shuttle coordinately influences tumor-stroma interplay. *Cancer research*, pages canres–1949, 2012.

[25] Patrizia Sanità, Mattia Capulli, Anna Teti, Giuseppe Paradiso Galatioto, Carlo Vicentini, Paola Chiarugi, Mauro Bologna, and Adriano Angelucci. Tumor-stroma metabolic relationship based on lactate shuttle can sustain prostate cancer progression. *BMC cancer*, 14(1):154, 2014.

[26] Melissa R Junttila and Frederic J de Sauvage. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*, 501(7467):346–354, 2013.

[27] O Warburg, K Posener, and E Negelein. Über den stoffwechsel der tumoren biochemische. *Zeitschrift*, 152:319–344, 1924.

[28] Matthew G Vander Heiden, Lewis C Cantley, and Craig B Thompson. Understanding the warburg effect: the metabolic requirements of cell proliferation. *science*, 324(5930):1029–1033, 2009.

[29] Ralph J DeBerardinis, Julian J Lum, Georgia Hatzivassiliou, and Craig B Thompson. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell metabolism*, 7(1):11–20, 2008.

[30] Iñigo San-Millán and George A Brooks. Reexamining cancer metabolism: lactate production for carcinogenesis could be the purpose and explanation of the warburg effect. *Carcinogenesis*, 38(2):119–133, 2017.

[31] Chiara Damiani, Riccardo Colombo, Daniela Gaglio, Fabrizia Mastroianni, Dario Pescini, Hans Victor Westerhoff, Giancarlo Mauri, Marco Vanoni, and Lilia Alberghina. A metabolic core model elucidates how enhanced utilization of glucose and glutamine, with enhanced glutamine-dependent lactate production, promotes cancer cell growth: The warburq effect. *PLoS computational biology*, 13(9):e1005758, 2017.

[32] Daniela Gaglio, Christian M Metallo, Paulo A Gameiro, Karsten Hiller, Lara Sala Danna, Chiara Balestrieri, Lilia Alberghina, Gregory Stephanopoulos, and Ferdinando Chiaradonna. Oncogenic k-ras decouples glucose and glutamine metabolism to support cancer cell growth. *Molecular systems biology*, 7(1):523, 2011.

[33] Christian M Metallo, Paulo A Gameiro, Eric L Bell, Katherine R Mattaini, Juanjuan Yang, Karsten Hiller, Christopher M Jewell, Zachary R Johnson, Darrell J Irvine, Leonard Guarente, et al. Reductive glutamine metabolism by idh1 mediates lipogenesis under hypoxia. *Nature*, 481(7381):380, 2012.

[34] David Basanta and Alexander RA Anderson. Exploiting ecological principles to better understand cancer progression and treatment. *Interface focus*, 3(4):20130020, 2013.

[35] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[36] Nemanja D Marjanovic, Robert A Weinberg, and Christine L Chaffer. Cell plasticity and heterogeneity in cancer. *Clinical chemistry*, 59(1):168–179, 2013.

[37] Corbin E Meacham and Sean J Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328, 2013.

[38] Aravind Venkatesan. Application of semantic web technology to establish knowledge management and discovery in the life sciences. 2014.

[39] Pandey Govind. Model organisms used in molecular biology or medical research. 2011.

[40] James J Russell, Julie A Theriot, Pranidhi Sood, Wallace F Marshall, Laura F Landweber, Lillian Fritz-Laylin, Jessica K Polka, Snezhana Oliferenko, Therese Gerbich, Amy Gladfelter, et al. Non-model model organisms. *BMC biology*, 15(1):55, 2017.

[41] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival Jr, Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, 2012.

[42] Anisha Goel, Meike Tessa Wortel, Douwe Molenaar, and Bas Teusink. Metabolic shifts: a fitness perspective for microbial cell factories. *Biotechnology letters*, 34(12):2147–2160, 2012.

[43] Michael Hucka, Andrew Finney, Herbert M Sauro, Hamid Bolouri, John C Doyle, Hiroaki Kitano, Adam P Arkin, Benjamin J Bornstein, Dennis Bray, Athel Cornish-Bowden, et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[44] Michael Hucka, Frank T Bergmann, Stefan Hoops, Sarah M Keating, Sven Sahle, James C Schaff, Lucian P Smith, and Darren J Wilkinson. The systems biology markup language (sbml): language specification for level 3 version 1 core. *Journal of integrative bioinformatics*, 12(2):382–549, 2015.

[45] David Gomez-Cabrero and Jesper Tegnér. Iterative systems biology for medicine–time for advancing from network signatures to mechanistic equations. *Current Opinion in Systems Biology*, 3:111–118, 2017.

[46] Vay Liang W Go, Christine TH Nguyen, Diane M Harris, and Wai-Nang Paul Lee. Nutrient-gene interaction: metabolic genotype-phenotype relationship. *The Journal of nutrition*, 135(12):3016S–3020S, 2005.

[47] Eivind Almaas. Biological impacts and context of network theory. *Journal of Experimental Biology*, 210(9):1548–1558, 2007.

[48] Paolo Cazzaniga, Chiara Damiani, Daniela Besozzi, Riccardo Colombo, Marco S Nobile, Daniela Gaglio, Dario Pescini, Sara Molinari, Giancarlo Mauri, Lilia Alberghina, et al. Computational strategies for a system-level understanding of metabolism. *Metabolites*, 4(4):1034–1087, 2014.

[49] Kenneth J Kauffman, Purusharth Prakash, and Jeremy S Edwards. Advances in flux balance analysis. *Current opinion in biotechnology*, 14(5):491–496, 2003.

[50] Edward J O'Brien, Jonathan M Monk, and Bernhard O Palsson. Using genome-scale models to predict biological capabilities. *Cell*, 161(5):971–987, 2015.

[51] Daniele De Martino, Fabrizio Capuani, Matteo Mori, Andrea De Martino, and Enzo Marinari. Counting and correcting thermodynamically infeasible

flux cycles in genome-scale metabolic networks. *Metabolites*, 3(4):946–966, 2013.

[52] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93, 2010.

[53] RD Fleischmann, MD Adams, O White, RA Clayton, EF Kirkness, AR Kerlavage, CJ Bult, JF Tomb, BA Dougherty, JM Merrick, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, 1995.

[54] JS Edwards and BØ Palsson. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J Biol Chem*, 274(25):17410–6, 1999.

[55] L Safak Yilmaz and Albertha JM Walhout. Metabolic network modeling with model organisms. *Current Opinion in Chemical Biology*, 36:32–39, 2017.

[56] Douwe Molenaar, Rogier Van Berlo, Dick De Ridder, and Bas Teusink. Shifts in growth strategies reflect tradeoffs in cellular economics. *Molecular systems biology*, 5(1):323, 2009.

[57] Jörg Stelling. Mathematical models in microbial systems biology. *Current opinion in microbiology*, 7(5):513–518, 2004.

[58] Nathan E Lewis, Harish Nagarajan, and Bernhard O Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291, 2012.

[59] Aarash Bordbar, Jonathan M Monk, Zachary A King, and Bernhard O Palsson. Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*, 15(2):107, 2014.

[60] JD Orth, I Thiele, and BØ Palsson. What is flux balance analysis? *Nat Biotechnol*, 28(3):245, 2010.

[61] Saul I Gass. Linear programming. *Encyclopedia of Statistical Sciences*, 6, 2004.

[62] Glpk (gnu linear programming kit), 2018.

[63] LLC Gurobi Optimization. Gurobi optimizer reference manual, 2018.

[64] R Mahadevan and CH Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng*, 5(4):264–276, 2003.

[65] NE Lewis, KK Hixson, TM Conrad, JA Lerman, P Charusanti, AD Polpitiya, JN Adkins, G Schramm, S O Purvine, and D et al. Lopez-Ferrer. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol*, 6(1):390, 2010.

[66] RA Ortiz-Merino, N Kuanyshev, S Braun-Galleani, KP Byrne, D Porro, P Branduardi, and KH Wolfe. Evolutionary restoration of fertility in an interspecies hybrid yeast, by whole-genome duplication after a failed mating-type switch. *PLoS Biol*, 15(5):e2002128, 2017.

[67] Jan Schellenberger, Junyoung O Park, Tom M Conrad, and Bernhard Ø Palsson. Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics*, 11(1):213, 2010.

[68] M Kanehisa and S Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[69] Christopher S Henry, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977, 2010.

[70] Peter D Karp, Richard Billington, Ron Caspi, Carol A Fulcher, Mario Latendresse, Anamika Kothari, Ingrid M Keseler, Markus Krummenacker, Peter E Midford, Quang Ong, et al. The biocyc collection of microbial genomes and metabolic pathways. *Brief. Bioinformatics*, 2017.

[71] Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*, 46(15):7542–7553, 2018.

[72] Michael S Siddiqui, Kate Thodey, Isis Trenchard, and Christina D Smolke. Advancing secondary metabolite biosynthesis in yeast with synthetic biology tools. *FEMS yeast research*, 12(2):144–170, 2012.

[73] Rasmus Agren, Liming Liu, Saeed Shoaie, Wanwipa Vongsangnak, Intawat Nookaew, and Jens Nielsen. The raven toolbox and its use for

generating a genome-scale metabolic model for penicillium chrysogenum. *PLoS computational biology*, 9(3):e1002980, 2013.

[74] Hao Wang, Simonas Marcišauskas, Benjamín J Sánchez, Iván Domenzain, Daniel Hermansson, Rasmus Agren, Jens Nielsen, and Eduard J Kerkhoven. Raven 2.0: A versatile toolbox for metabolic network reconstruction and a case study on streptomyces coelicolor. *PLoS computational biology*, 14(10):e1006541, 2018.

[75] Meric Ataman, Daniel F Hernandez Gardiol, Georgios Fengos, and Vassily Hatzimanikatis. redgem: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models. *PLoS computational biology*, 13(7):e1005444, 2017.

[76] Meric Ataman and Vassily Hatzimanikatis. lumpgem: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites. *PLoS computational biology*, 13(7):e1005513, 2017.

[77] Philipp Erdrich, Ralf Steuer, and Steffen Klamt. An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. *BMC systems biology*, 9(1):48, 2015.

[78] Marzia Di Filippo, Riccardo Colombo, Chiara Damiani, Dario Pescini, Daniela Gaglio, Marco Vanoni, Lilia Alberghina, and Giancarlo Mauri. Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models. *Computational biology and chemistry*, 62:60–69, 2016.

[79] PJA Cock, T Antao, JT Chang, BA Chapman, CJ Cox, A Dalke, I Friedberg, T Hamelryck, F Kauff, and B et al. Wilczynski. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[80] UniProt Consortium et al. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 46(5):2699, 2018.

[81] Daniela Besozzi, Paolo Cazzaniga, Giancarlo Mauri, Dario Pescini, and Leonardo Vanneschi. A comparison of genetic algorithms and particle swarm optimization for parameter estimation in stochastic biochemical systems. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 116–127. Springer, 2009.

[82] Paola Branduardi, Minoska Valli, Luca Brambilla, Michael Sauer, Lilia Alberghina, and Danilo Porro. The yeast zygosaccharomyces bailii: a new host for heterologous protein production, secretion and for metabolic engineering applications. *FEMS yeast research*, 4(4-5):493–504, 2004.

[83] Michael Sauer, Paola Branduardi, Minoska Valli, and Danilo Porro. Production of l-ascorbic acid by metabolically engineered saccharomyces cerevisiae and zygosaccharomyces bailii. *Applied and environmental microbiology*, 70(10):6086–6091, 2004.

[84] ND Price, JL Reed, and BØ Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol*, 2(11):886, 2004.

[85] H Lopes and I Rocha. Genome-scale modeling of yeast: chronology, applications and critical perspectives. *FEMS Yeast Res*, 17(5), 2017.

[86] Bevan KS Chung, Suresh Selvarasu, Andrea Camattari, Jimyoung Ryu, Hyeokweon Lee, Jungoh Ahn, Hongweon Lee, and Dong-Yup Lee. Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast pichia pastoris for strain improvement. *Microbial cell factories*, 9(1):50, 2010.

[87] Seung Bum Sohn, Alexandra B Graf, Tae Yong Kim, Brigitte Gasser, Michael Maurer, Pau Ferrer, Diethard Mattanovich, and Sang Yup Lee. Genome-scale metabolic model of methylotrophic yeast pichia pastoris and its use for in silico analysis of heterologous protein production. *Biotechnology journal*, 5(7):705–715, 2010.

[88] Luis Caspeta, Saeed Shoaie, Rasmus Agren, Intawat Nookaew, and Jens Nielsen. Genome-scale metabolic reconstructions of pichia stipitis and pichia pastoris and in silico evaluation of their potentials. *BMC systems biology*, 6(1):24, 2012.

[89] Zahra Azimzadeh Irani, Eduard J Kerkhoven, Seyed Abbas Shojaosadati, and Jens Nielsen. Genome-scale metabolic model of pichia pastoris with native and humanized glycosylation of recombinant proteins. *Biotechnology and bioengineering*, 113(5):961–969, 2016.

[90] Balaji Balagurunathan, Sudhakar Jonnalagadda, Lily Tan, and Rajagopalan Srinivasan. Reconstruction and analysis of a genome-scale metabolic model for scheffersomyces stipitis. *Microbial cell factories*, 11(1):27, 2012.

[91] Ting Liu, Wei Zou, Liming Liu, and Jian Chen. A constraint-based model of scheffersomyces stipitis for improved ethanol production. *Biotechnology for biofuels*, 5(1):72, 2012.

[92] Màrius Tomàs-Gamisans, Pau Ferrer, and Joan Albiol. Integration and validation of the genome-scale metabolic models of pichia pastoris: a comprehensive update of protein glycosylation pathways, lipid and energy metabolism. *PloS one*, 11(1):e0148031, 2016.

[93] Seung Bum Sohn, Tae Yong Kim, Jay H Lee, and Sang Yup Lee. Genome-scale metabolic model of the fission yeast schizosaccharomyces pombe and the reconciliation of in silico/in vivo mutant growth. *BMC systems biology*, 6(1):49, 2012.

[94] Nan Xu, Liming Liu, Wei Zou, Jie Liu, Qiang Hua, and Jian Chen. Reconstruction and analysis of the genome-scale metabolic network of candida glabrata. *Molecular BioSystems*, 9(2):205–216, 2013.

[95] Pranjul Mishra, Gyu-Yeon Park, Meiyappan Lakshmanan, Hee-Seok Lee, Hongweon Lee, Matthew Wook Chang, Chi Bun Ching, Jungoh Ahn, and Dong-Yup Lee. Genome-scale metabolic modeling and in silico analysis of lipid accumulating yeast candida tropicalis for dicarboxylic acid production. *Biotechnology and bioengineering*, 113(9):1993–2004, 2016.

[96] Nicolas Loira, Thierry Dulermo, Jean-Marc Nicaud, and David James Sherman. A genome-scale metabolic model of the lipid-accumulating yeast yarrowia lipolytica. *BMC systems biology*, 6(1):35, 2012.

[97] P Pan and Q Hua. Reconstruction and in silico analysis of metabolic network for an oleaginous yeast, Yarrowia lipolytica. *PLoS One*, 7(12):e51535, 2012.

[98] Martin Kavšček, Govindprasad Bhutada, Tobias Madl, and Klaus Natter. Optimization of lipid production with a genome-scale model of yarrowia lipolytica. *BMC systems biology*, 9(1):72, 2015.

[99] EJ Kerkhoven, KR Pomraning, SE Baker, and J Nielsen. Regulation of amino-acid metabolism controls flux to lipid accumulation in *Yarrowia lipolytica*. *NPJ Syst Biol Appl*, 2:16005, 2016.

[100] O Dias, R Pereira, AK Gombert, EC Ferreira, and I Rocha. iod907, the first genome-scale metabolic model for the milk yeast *Kluyveromyces lactis*. *Biotechnology Journal*, 9(6):776–790, 2014.

[101] Oscar Dias, Andreas K. Gombert, Eugénio C. Ferreira, and Isabel Rocha. Genome-wide metabolic (re-) annotation of *Kluyveromyces lactis*. *BMC Genomics*, 13(1):517, Oct 2012.

[102] Brigida Gallone, Stijn Mertens, Jonathan L Gordon, Steven Maere, Kevin J Verstrepen, and Jan Steensels. Origins, evolution, domestication and diversity of saccharomyces beer yeasts. *Current opinion in biotechnology*, 49:148–155, 2018.

[103] M Stratford, H Steels, G Nebe-von Caron, M Novodvorska, K Hayer, and DB Archer. Extreme resistance to weak-acid preservatives in the spoilage yeast Zygosaccharomyces bailii. *Int J Food Microbiol*, 166(1):126–34, 2013.

[104] N Kuanyshev, GM Adamo, D Porro, and P Branduardi. The spoilage yeast Zygosaccharomyces bailii: Foe or friend? *Yeast*, 34(9):359–370, 2017.

[105] L Dato, P Branduardi, S Passolunghi, D Cattaneo, L Riboldi, G Frascotti, M Valli, and D Porro. Advances in molecular tools for the use of Zygosaccharomyces bailii as host for biotechnological productions and construction of the first auxotrophic mutant. *FEMS Yeast Res*, 10(7):894–908, 2010.

[106] Leif J Jönsson, Björn Alriksson, and Nils-Olof Nilvebrant. Bioconversion of lignocellulose: inhibitors and detoxification. *Biotechnology for biofuels*, 6(1):16, 2013.

[107] Margarida Palma, Joana F Guerreiro, and Isabel Sá-Correia. Adaptive response and tolerance to acetic acid in saccharomyces cerevisiae and zygosaccharomyces bailii: A physiological genomics perspective. *Frontiers in microbiology*, 9:274, 2018.

[108] M Hulin and A Wheals. Rapid identification of Zygosaccharomyces with genus-specific primers. *Int J Food Microbiol*, 173:9–13, 2014.

[109] SO Suh, P Gujjari, C Beres, B Beck, and J Zhou. Proposal of Zygosaccharomyces parabailii sp. nov. and Zygosaccharomyces pseudobailii sp. nov., novel species closely related to Zygosaccharomyces bailii. *Int J Syst Evol Microbiol*, 63(Pt 5):1922–9, 2013.

[110] RA Ortiz-Merino, N Kuanyshev, KP Byrne, JA Varela, JP Morrissey, D Porro, and P Wolfe, KH Branduardi. Transcriptional Response to Lactic

Acid Stress in the Hybrid Yeast Zygosaccharomyces parabailii. *Appl Environ Microbiol*, 84(5), 2018.

[111] HW Aung, SA Henry, and LP Walker. Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. *Ind Biotechnol*, 9(4):215–228, 2013.

[112] Vijayalakshmi Chelliah, Nick Juty, Ishan Ajmera, Raza Ali, Marine Dumousseau, Mihai Glont, Michael Hucka, Gaël Jalowicki, Sarah Keating, Vincent Knight-Schrijver, et al. Biomodels: ten-year anniversary. *Nucleic acids research*, 43(D1):D542–D548, 2014.

[113] L Lindberg, AXS Santos, H Riezman, L Olsson, and M Bettiga. Lipidomic profiling of saccharomyces cerevisiae and zygosaccharomyces bailii reveals critical changes in lipid composition in response to acetic acid stress. *PloS one*, 8(9):e73936, 2013.

[114] N Kuanyshev, D Ami, L Signori, D Porro, JP Morrissey, and P Branduardi. Assessing physio-macromolecular effects of lactic acid on zygosaccharomyces bailii cells during microaerobic fermentation. *FEMS Yeast Res*, 16(5), 2016.

[115] Y Xu, Y Zhi, Q Wu, and R Du. Zygosaccharomyces bailii is a potential producer of various flavor compounds in chinese maotai-flavor liquor fermentation. *Front Microbiol*, 8:2609, 2017.

[116] NV Narendranath, KC Thomas, and WM Ingledew. Effects of acetic acid and lactic acid on the growth of saccharomyces cerevisiae in a minimal medium. *Journal of Industrial Microbiology and Biotechnology*, 26(3):171–177, 2001.

[117] E Proux-Wéra, D Armisén, KP Byrne, and KH Wolfe. A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. *BMC Bioinformatics*, 13:237, 2012.

[118] S Götz, JM García-Gómez, J Terol, TD Williams, SH Nagaraj, MJ Nueda, M Robles, M Talón, J Dopazo, and A Conesa. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*, 36(10):3420–35, 2008.

[119] T Cokelaer, D Pultz, LM Harder, J Serra-Musach, and J Saez-Rodriguez. Bioservices: a common python package to access biological web services programmatically. *Bioinformatics*, 29(24):3241–3242, 2013.

[120] M Kanehisa, Y Sato, M Kawashima, M Furumichi, and M Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*, 44(D1):D457–62, 2016.

[121] A Ebrahim, JA Lerman, BØ Palsson, and DR Hyduke. Cobrapy: constraints-based reconstruction and analysis for python. *BMC Syst Biol*, 7(1):74, 2013.

[122] RC Eberhart and J Kennedy. Particle swarm optimization, proceeding of ieee international conference on neural network. *Perth, Australia*, pages 1942–1948, 1995.

[123] C Verduyn, E Postma, WA Scheffers, and JP Van Dijken. Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation. *Yeast*, 8(7):501–517, 1992.

[124] JP Van Dijken, J Bauer, L Brambilla, P Duboc, JM Francois, C Gancedo, MLF Giuseppin, JJ Heijnen, M Hoare, HC Lange, et al. An interlaboratory comparison of physiological and genetic properties of four saccharomyces cerevisiae strains. *Enzyme Microb Technol*, 26(9-10):706–714, 2000.

[125] Laura Popolo, Marco Vanoni, and Lilia Alberghina. Control of the yeast cell cycle by protein synthesis. *Experimental cell research*, 142(1):69–78, 1982.

[126] Géraldine Gentric, Virginie Mieulet, and Fatima Mechta-Grigoriou. Heterogeneity in cancer metabolism: new concepts in an old field. *Antioxidants & redox signaling*, 26(9):462–485, 2017.

[127] Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338, 2013.

[128] Jagmohan Hooda, Daniela Cadinu, Md Maksudul Alam, Ajit Shah, Thai M Cao, Laura A Sullivan, Rolf Brekken, and Li Zhang. Enhanced heme function and mitochondrial respiration promote the progression of lung cancer cells. *PLoS One*, 8(5):e63402, 2013.

[129] Jorrit J Hornberg, Frank J Bruggeman, Hans V Westerhoff, and Jan Lankelma. Cancer: a systems biology disease. *Biosystems*, 83(2-3):81–90, 2006.

[130] Jason W Locasale and Lewis C Cantley. Altered metabolism in cancer. *BMC Biol*, 8:88, 2010.

[131] L Alberghina and D Gaglio. Redox control of glutamine utilization in cancer. *Cell Death Dis*, 5(12):e1561, 2014.

[132] Lilia Alberghina and Hans V Westerhoff. *Systems biology: definitions and perspectives (topics in current genetics)*. Springer, 2005.

[133] Hiroaki Kitano. Perspectives on systems biology. *New generation Computing*, 18(3):199–216, 2000.

[134] Hiroaki Kitano. Systems biology: Toward system-level understanding of biological systems. In Kitano H., editor, *Foundations of Systems Biology*, pages 1–36. MIT Press, 2001.

[135] Natalie C. Duarte, Scott A. Becker, Neema Jamshidi, Ines Thiele, Monica L. Mo, Thuy D. Vo, Rohith Srivas, and Bernhard Ø Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *PNAS*, 104(6):1777–1782, 2007.

[136] Hongwu Ma, Anatoly Sorokin, Alexander Mazein, Alex Selkov, Evgeni Selkov, Oleg Demin, and Igor Goryanin. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, 3:135, 2007.

[137] Tong Hao, Hong-Wu Ma, Xue-Ming Zhao, and Igor Goryanin. Compartmentalization of the Edinburgh human metabolic network. *BMC Bioinformatics*, 11:393, 2010.

[138] Ines Thiele, Neil Swainston, Ronan MT Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, Monica L Mo, Ottar Rolfsson, Miranda D Stobbe, et al. A community-driven global reconstruction of human metabolism. *Nat Biotechnol*, 31(5):419–25, 2013.

[139] Ori Folger, Livnat Jerby, Christian Frezza, Eyal Gottlieb, Eytan Ruppin, and Tomer Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Molecular Systems Biology*, 7(1):501, 2011.

[140] Rasmus Agren, Sergio Bordel, Adil Mardinoglu, Natapol Pornputtapong, Intawat Nookaew, and Jens Nielsen. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. *PLoS Comput Biol*, 8(5):e1002518, 2012.

[141] Rasmus Agren, Adil Mardinoglu, Anna Asplund, Caroline Kampf, Mathias Uhlen, and Jens Nielsen. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular Systems Biology*, 10(3):721, 2014.

[142] Francesco Gatto, Intawat Nookaew, and Jens Nielsen. Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. *National Acad Sciences*, 111(9):E866–75, 2014.

[143] J Ferlay, I Soerjomataram, M Ervik, R Dikshit, S Eser, C Mathers, MPD Rebelo, D Forman, and F Bray. GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, Fr Int Agency Res Cancer, 2013.

[144] Freddie Bray, Jian-Song Ren, Eric Masuyer, and Jacques Ferlay. Estimates of global cancer prevalence for 27 sites in the adult population in 2008. *International Journal of Cancer*, 132(5):1133–1145, 2013.

[145] Jie Hu, Jason W Locasale, Jason H Bielas, Jacintha O'Sullivan, Kieran Sheahan, Lewis C Cantley, Matthew G Vander Heiden, and Dennis Vitkup. Heterogeneity of tumor-induced gene expression changes in the human metabolic network. *Nature biotechnology*, 31(6):522–529, 2013.

[146] David A Fell and J Rankin Small. Fat synthesis in adipose tissue. an examination of stoichiometric constraints. *Biochem. J*, 238:781–786, 1986.

[147] Amit Varma and Bernhard Ø Palsson. Metabolic capabilities of *Escherichia* coli: I. synthesis of biosynthetic precursors and cofactors. *Journal of theoretical biology*, 165(4):477–502, 1993.

[148] Amit Varma and Bernhard Ø Palsson. Metabolic capabilities of *Escherichia* coli II. optimal growth patterns. *Journal of Theoretical Biology*, 165(4):503–522, 1993.

[149] Karthik Raman and Nagasuma Chandra. Flux balance analysis of biological systems: applications and challenges. *Brief Bioinform*, 10(4):435–49, 2009.

[150] Pedro Romero, Jonathan Wagg, Michelle L. Green, Dale Kaiser, Markus Krummenacker, and Peter D. Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, 6(1):R2, 2004.

[151] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*, 42(D1):D199–D205, 2014.

[152] Mathias Uhlén, Erik Björling, Charlotta Agaton, Cristina Al-Khalili Szigyarto, Bahram Amini, Elisabet Andersen, Ann-Catrin Andersson, Pia Angelidou, Anna Asplund, Caroline Asplund, et al. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & Cellular Proteomics*, 4(12):1920–1932, 2005.

[153] David S Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, et al. HMDB: the human metabolome database. *Nucleic acids research*, 35(suppl 1):D521–D526, 2007.

[154] David S Wishart, Craig Knox, An Chi Guo, Roman Eisner, Nelson Young, Bijaya Gautam, David D Hau, Nick Psychogios, Edison Dong, Souhaila Bouatra, et al. HMDB: a knowledgebase for the human metabolome. *Nucleic acids research*, 37(suppl 1):D603–D610, 2009.

[155] David S Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, et al. HMDB 3.0 — the human metabolome database in 2013. *Nucleic acids research*, page gks1065, 2012.

[156] Osbaldo Resendis-Antonio, Alberto Checa, and Sergio Encarnación. Modeling core metabolism in cancer cells: Surveying the topology underlying the Warburg effect. *PLoS ONE*, 5(8):e12383, 2010.

[157] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, 2003.

[158] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, et al. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*, 2(10):2366–82, 2007.

[159] Scott A Becker, Adam M Feist, Monica L Mo, Gregory Hannum, Bernhard Ø Palsson, and Markus J Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. *Nat Protoc*, 2(3):727–38, 2007.

[160] Jan Schellenberger, Richard Que, Ronan M T Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox v2.0. *Nat Protoc*, 6(9):1290–307, 2011.

[161] Daniele De Martino, Fabrizio Capuani, Matteo Mori, Andrea De Martino, and Enzo Marinari. Counting and correcting thermodynamically infeasible flux cycles in genome-scale metabolic networks. *Metabolites*, 3(4):946–966, 2013.

[162] Keren Yizhak, Sylvia E Le Dévédec, Vasiliki Maria Rogkoti, Franziska Baenke, Vincent C de Boer, Christian Frezza, Almut Schulze, Bob van de Water, and Eytan Ruppin. A computational study of the Warburg effect identifies metabolic targets inhibiting cancer migration. *Mol Syst Biol*, 10:744, 2014.

[163] Chen Li, Marco Donizelli, Nicolas Rodriguez, Harish Dharuri, Lukas Endler, Vijayalakshmi Chelliah, Lu Li, Enuo He, Arnaud Henry, Melanie I Stefan, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC systems biology*, 4(1):92, 2010.

[164] Ralph J DeBerardinis, Anthony Mancuso, Evgueni Daikhin, Ilana Nissim, Marc Yudkoff, Suzanne Wehrli, and Craig B Thompson. Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proc Natl Acad Sci U S A*, 104(49):19345–50, 2007.

[165] Shanmugasundaram Ganapathy-Kanniappan and Jean-Francois H Geschwind. Tumor glycolysis as a target for cancer therapy: progress and prospects. *Mol Cancer*, 12:152, 2013.

[166] Karsten Hiller and Christian M Metallo. Profiling metabolic networks to study cancer metabolism. *Curr Opin Biotechnol*, 24(1):60–8, 2013.

[167] Hu Z. Mullen A.R. Oxidation of alpha-ketoglutarate is required for reductive carboxylation in cancer cells with mitochondrial defects. *Cell reports*, 7(5):1679–1690, 2014.

[168] Ralph J DeBerardinis and Tzuling Cheng. Q's next: the diverse functions of glutamine in metabolism, cell biology and cancer. *Oncogene*, 29(3):313–324, 2010.

[169] Sébastien Bonnet, Stephen L Archer, Joan Allalunis-Turner, Alois Haromy, Christian Beaulieu, Richard Thompson, Christopher T Lee, Gary D Lopaschuk, Lakshmi Puttagunta, Sandra Bonnet, et al. A mitochondria-K^+ channel axis is suppressed in cancer and its normalization promotes apoptosis and inhibits cancer growth. *Cancer cell*, 11(1):37–51, 2007.

[170] Chiara Damiani, Dario Pescini, Riccardo Colombo, Sara Molinari, Lilia Alberghina, Marco Vanoni, and Giancarlo Mauri. An ensemble evolutionary constraint-based approach to understand the emergence of metabolic phenotypes. *Natural Computing*, 13(3):321–331, 2014.

[171] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[172] Oleksandr Lytovchenko and Edmund RS Kunji. Expression and putative role of mitochondrial transport proteins in cancer. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1858(8):641–654, 2017.

[173] Balaji Srinivasan and Ody CM Sibon. Coenzyme a, more than 'just'a metabolic cofactor, 2014.

[174] Marios C Papadopoulos and Samira Saadoun. Key roles of aquaporins in tumor biology. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1848(10):2576–2583, 2015.

[175] Luigi Sapio and Silvio Naviglio. Inorganic phosphate in the development and treatment of cancer: A janus bifrons? *World journal of clinical oncology*, 6(6):198, 2015.

[176] Félix A Urra, Felipe Muñoz, Alenka Lovy, and César Cárdenas. The mitochondrial complex (i) ty of cancer. *Frontiers in oncology*, 7:118, 2017.

[177] Geou-Yarh Liou and Peter Storz. Reactive oxygen species in cancer. *Free radical research*, 44(5):479–496, 2010.

[178] Alice C Newman and Oliver DK Maddocks. One-carbon metabolism in cancer. *British journal of cancer*, 116(12):1499, 2017.

[179] Q Qu, F Zeng, X Liu, QJ Wang, and F Deng. Fatty acid oxidation and carnitine palmitoyltransferase i: emerging therapeutic targets in cancer. *Cell death & disease*, 7(5):e2226, 2017.

[180] Tracy R Murray-Stewart, Patrick M Woster, and Robert A Casero. Targeting polyamine metabolism for cancer therapy and prevention. *Biochemical Journal*, 473(19):2937–2953, 2016.

[181] SR McKeown. Defining normoxia, physoxia and hypoxia in tumours—implications for treatment response. *The British journal of radiology*, 87(1035):20130676, 2014.

[182] Patrick S Ward and Craig B Thompson. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer cell*, 21(3):297–308, 2012.

[183] Veerle W Daniëls, Karine Smans, Ines Royaux, Melanie Chypre, Johannes V Swinnen, and Nousheen Zaidi. Cancer cells differentially activate and thrive on de novo lipid synthesis pathways in a low-lipid environment. *PloS one*, 9(9):e106913, 2014.

[184] Colin A Flaveny, Kristine Griffett, Bahaa El-Dien M El-Gendy, Melissa Kazantzis, Monideepa Sengupta, Antonio L Amelio, Arindam Chatterjee, John Walker, Laura A Solt, Theodore M Kamenecka, et al. Broad antitumor activity of a small molecule that selectively targets the warburg effect and lipogenesis. *Cancer cell*, 28(1):42–56, 2015.

[185] Subrata Patra, Alok Ghosh, Soumya Sinha Roy, Soumen Bera, Manju Das, Dipa Talukdar, Subhankar Ray, Theo Wallimann, and Manju Ray. A short review on creatine–creatine kinase system in relation to cancer and some experimental results on creatine as adjuvant in cancer therapy. *Amino Acids*, 42(6):2319–2330, 2012.

[186] Shiran Rabinovich, Lital Adler, Keren Yizhak, Alona Sarver, Alon Silberman, Shani Agron, Noa Stettner, Qin Sun, Alexander Brandis, Daniel Helbling, et al. Diversion of aspartate in ass1-deficient tumours fosters de novo pyrimidine synthesis. *Nature*, 527(7578):379, 2015.

[187] Kimberly H Allison and George W Sledge. Heterogeneity and cancer. *Oncology*, 28(9):772–778, 2014.

[188] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England journal of medicine*, 366(10):883–892, 2012.

[189] Mariam Jamal-Hanjani, Sergio A Quezada, James Larkin, and Charles Swanton. Translational implications of tumor heterogeneity. *Clinical cancer research*, 21(6):1258–1266, 2015.

[190] Ubaldo E Martinez-Outschoorn, Zhao Lin, Casey Trimmer, Neal Flomenberg, Chenguang Wang, Stephanos Pavlides, Richard G Pestell, Anthony Howell, Federica Sotgia, and Michael P Lisanti. Cancer cells metabolically "fertilize" the tumor microenvironment with hydrogen peroxide, driving the warburg effect: implications for pet imaging of human tumors. *Cell cycle*, 10(15):2504–2520, 2011.

[191] Jong Min Lee, Erwin P Gianchandani, and Jason A Papin. Flux balance analysis in the era of metabolomics. *Briefings in Bioinformatics*, 7(2):140–150, 2006.

[192] Anwoy Kumar Mohanty, Aniruddha Datta, and Jijayanagaram Venkatraj. Determining the relative prevalence of different subpopulations in heterogeneous cancer tissue. In *Genomic Signal Processing and Statistics,(GENSIPS), 2012 IEEE International Workshop on*, pages 95–96. IEEE, 2012.

[193] Renato Zenobi. Single-cell metabolomics: analytical and biological perspectives. *Science*, 342(6163):1243259, 2013.

[194] Jennifer L Reed and Bernhard Ø Palsson. Genome-scale in silico models of e. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Research*, 14(9):1797–1805, 2004.

[195] Adam M. Feist, Markus J. Herrgard, Ines Thiele, Jennie L. Reed, and Bernard Ø Pallson. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129–143, 2009.

[196] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Reconstruction of genome-scale human metabolic models using omics data. *Integrative Biology*, 7(8):859–868, 2015.

[197] Andriy Marusyk and Kornelia Polyak. Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1805(1):105–117, 2010.

[198] Min-Hyuk Yoo and Dolph L Hatfield. The cancer stem cell theory: is it correct? *Molecules and cells*, 26(5):514–516, 2008.

[199] Xiao-xiao Sun and Qiang Yu. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacologica Sinica*, 36(10):1219–1227, 2015.

[200] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334, 2012.

[201] Elaine Holmes, Ian D Wilson, and Jeremy K Nicholson. Metabolic phenotyping in health and disease. *Cell*, 134(5):714–717, 2008.

[202] Daniel Machado and Markus Herrgård. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput Biol*, 10(4):e1003580, 2014.

[203] Marzia Di Filippo, Chiara Damiani, Riccardo Colombo, Dario Pescini, and Giancarlo Mauri. Constraint-based modeling and simulation of cell populations. In *Italian Workshop on Artificial Life and Evolutionary Computation*, pages 126–137. Springer, 2016.

[204] Adil Mardinoglu, Rasmus Agren, Caroline Kampf, Anna Asplund, Intawat Nookaew, Peter Jacobson, Andrew J Walley, Philippe Froguel, Lena M Carlsson, Mathias Uhlen, et al. Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Molecular systems biology*, 9(1):649, 2013.

[205] Jan Schellenberger and Bernhard Ø Palsson. Use of randomized sampling for analysis of metabolic networks. *Journal of biological chemistry*, 284(9):5457–5461, 2009.

[206] Sergio Bordel, Rasmus Agren, and Jens Nielsen. Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput Biol*, 6(7):e1000859, 2010.

[207] Diana Whitaker-Menezes, Ubaldo E Martinez-Outschoorn, Zhao Lin, Adam Ertel, Neal Flomenberg, Agnieszka K Witkiewicz, Ruth Birbe, Anthony Howell, Stephanos Pavlides, Ricardo Gandara, et al. Evidence for a stromal-epithelial "lactate shuttle" in human tumors: Mct4 is a marker of oxidative stress in cancer-associated fibroblasts. *Cell cycle*, 10(11):1772–1783, 2011.

[208] Robin M Hallett, Anna Dvorkin-Gheva, Anita Bane, and John A Hassell. A gene signature for predicting outcome in patients with basal-like breast cancer. *Scientific reports*, 2:227, 2012.

[209] Gökhan S Hotamisligil. Inflammation and metabolic disorders. *Nature*, 444(7121):860, 2006.

[210] Adil Mardinoglu, Rasmus Agren, Caroline Kampf, Anna Asplund, Mathias Uhlen, and Jens Nielsen. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nature communications*, 5, 2014.

[211] Neil Swainston, Kieran Smallbone, Hooman Hefzi, Paul D Dobson, Judy Brewer, Michael Hanscho, Daniel C Zielinski, Kok Siong Ang, Natalie J Gardiner, Jahir M Gutierrez, et al. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, 12(7):1–7, 2016.

[212] Keren Yizhak, Barbara Chaneton, Eyal Gottlieb, and Eytan Ruppin. Modeling cancer metabolism on a genome scale. *Molecular systems biology*, 11(6):817, 2015.

[213] Sjoerd Opdam, Anne Richelle, Benjamin Kellman, Shanzhong Li, Daniel C Zielinski, and Nathan E Lewis. A systematic evaluation of methods for tailoring genome-scale metabolic models. *Cell Systems*, 4(3):318–329, 2017.

[214] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[215] Vytautas Leoncikas, Huihai Wu, Lara T Ward, Andrzej M Kierzek, and Nick J Plant. Generation of 2,000 breast cancer metabolic landscapes reveals a poor prognosis group with active serotonin production. *Scientific reports*, 6, 2016.

[216] Nicholas McGranahan and Charles Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell*, 27(1):15–26, 2015.

[217] Giovanni Ciriello, Michael L Gatza, Andrew H Beck, Matthew D Wilkerson, Suhn K Rhie, Alessandro Pastore, Hailei Zhang, Michael McLellan, Christina Yau, Cyriac Kandoth, et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519, 2015.

[218] The Cancer Genome Atlas Network and others. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543, 2014.

[219] Chiara Damiani, Marzia Di Filippo, Dario Pescini, Davide Maspero, Riccardo Colombo, and Giancarlo Mauri. popfba: tackling intratumour heterogeneity with flux balance analysis. *Bioinformatics*, 33(14):i311–i318, 2017.

[220] Dave Lee, Kieran Smallbone, Warwick B Dunn, Ettore Murabito, Catherine L Winder, Douglas B Kell, Pedro Mendes, and Neil Swainston. Improving metabolic flux predictions using absolute gene expression data. *BMC systems biology*, 6(1):73, 2012.

[221] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[222] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[223] Hiroaki Kitano, Akira Funahashi, Yukiko Matsuoka, and Kanae Oda. Using process diagrams for the graphical representation of biological networks. *Nature biotechnology*, 23(8), 2005.

[224] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, and Amos Bairoch. Uniprotkb/swiss-prot: the manually annotated section of the uniprot knowledgebase. *Plant bioinformatics: methods and protocols*, pages 89–112, 2007.

[225] Kristian A Gray, Bethan Yates, Ruth L Seal, Mathew W Wright, and Elspeth A Bruford. Genenames. org: the hgnc resources in 2015. *Nucleic acids research*, 43(D1):D1079–D1085, 2014.

[226] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, 2012.

[227] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl_1):D428–D432, 2005.

[228] Kiran Raosaheb Patil and Jens Nielsen. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2685–2689, 2005.

[229] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

[230] Alex Graudenzi, Claudia Cava, Gloria Bertoli, Bastian Fromm, Kjersti Flatmark, Giancarlo Mauri, and Isabella Castiglioni. Pathway-based classification of breast cancer subtypes. *Frontiers in bioscience (Landmark edition)*, 22:1697, 2017.

[231] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108–1115, 2013.

[232] Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, Alex Graudenzi, Ilya Korsunsky, Giancarlo Mauri, Marco Antoniotti, and Bud

Mishra. Capri: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026, 2015.

[233] Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Rebeca Sanz-Pamplona, Luca De Sano, Giancarlo Mauri, Victor Moreno, Marco Antoniotti, and Bud Mishra. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences*, page 201520213, 2016.

[234] Giulio Caravagna, Ylenia Giarratano, Daniele Ramazzotti, Ian Tomlinson, Trevor A Graham, Guido Sanguinetti, and Andrea Sottoriva. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature Methods*, 15(9):707, 2018.

[235] Robert A Gatenby and Robert J Gillies. Why do cancers have high aerobic glycolysis? *Nature Reviews Cancer*, 4(11):891–899, 2004.

[236] Zachary J Reitman and Hai Yan. Isocitrate dehydrogenase 1 and 2 mutations in cancer: alterations at a crossroads of cellular metabolism. *Journal of the National Cancer Institute*, 102(13):932–941, 2010.

[237] Vaibhav P Pai, Aaron M Marshall, Laura L Hernandez, Arthur R Buckley, and Nelson D Horseman. Altered serotonin physiology in human breast cancers favors paradoxical growth and cell survival. *Breast cancer research*, 11(6):R81, 2009.

[238] Elena Doldo, Gaetana Costanza, Sara Agostinelli, Chiara Tarquini, Amedeo Ferlosio, Gaetano Arcuri, Daniela Passeri, Maria Giovanna Scioli, and Augusto Orlandi. Vitamin a, cancer treatment and prevention: the new role of cellular retinol binding proteins. *BioMed research international*, 2015, 2015.

[239] Vadivel Ganapathy, Muthusamy Thangaraju, and Puttur D Prasad. Nutrient transporters in cancer: relevance to warburg hypothesis and beyond. *Pharmacology & therapeutics*, 121(1):29–40, 2009.

[240] Ying Huang and Wolfgang Sadée. Membrane transporters and channels in chemoresistance and-sensitivity of tumor cells. *Cancer letters*, 239(2):168–182, 2006.

[241] Heike Hellmold, Tove Rylander, Malin Magnusson, Eva Reihnér, Margaret Warner, and Jan-Ake Gustafsson. Characterization of cytochrome

p450 enzymes in human breast tissue from reduction mammaplasties. *The Journal of Clinical Endocrinology & Metabolism*, 83(3):886–895, 1998.

[242] Ivan Bièche, Igor Girault, Estelle Urbain, Sengül Tozlu, and Rosette Lidereau. Relationship between intratumoral expression of genes coding for xenobiotic-metabolizing enzymes and benefit from adjuvant tamoxifen in estrogen receptor alpha-positive postmenopausal breast carcinoma. *Breast Cancer Research*, 6(3):R252, 2004.

[243] C Rodriguez-Antona and M Ingelman-Sundberg. Cytochrome p450 pharmacogenetics and cancer. *Oncogene*, 25(11):1679–1691, 2006.

[244] Musiliyu A Musa, John S Cooperwood, and M Omar F Khan. A review of coumarin derivatives in pharmacotherapy of breast cancer. *Current medicinal chemistry*, 15(26):2664–2679, 2008.

[245] Caroline Colijn, Aaron Brandes, Jeremy Zucker, Desmond S Lun, Brian Weiner, Maha R Farhat, Tan-Yun Cheng, D Branch Moody, Megan Murray, and James E Galagan. Interpreting expression data with metabolic flux models: predicting mycobacterium tuberculosis mycolic acid production. *PLoS computational biology*, 5(8):e1000489, 2009.

[246] Aaron Brandes, Desmond S Lun, Kuhn Ip, Jeremy Zucker, Caroline Colijn, Brian Weiner, and James E Galagan. Inferring carbon sources from gene expression profiles using metabolic flux models. *PLoS One*, 7(5):e36947, 2012.

[247] Minetta C Liu, Brandelyn N Pitcher, Elaine R Mardis, Sherri R Davies, Paula N Friedman, Jacqueline E Snider, Tammi L Vickery, Jerry P Reed, Katherine DeSchryver, Baljit Singh, et al. Pam50 gene signatures and breast cancer prognosis with adjuvant anthracycline-and taxane-based chemotherapy: correlative analysis of c9741 (alliance). *NPJ Breast Cancer*, 2:15023, 2016.

[248] Patrick S Ward and Craig B Thompson. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer cell*, 21(3):297–308, 2012.

[249] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

[250] Matthew G Vander Heiden. Targeting cancer metabolism: a therapeutic window opens. *Nature reviews Drug discovery*, 10(9):671–684, 2011.

[251] Mark Robertson-Tessi, Robert J Gillies, Robert A Gatenby, and Alexander RA Anderson. Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. *Cancer research*, 75(8):1567–1579, 2015.

[252] A Pieter J van den Heuvel, Junping Jing, Richard F Wooster, and Kurtis E Bachman. Analysis of glutamine dependency in non-small cell lung cancer: Gls1 splice variant gac is essential for cancer cell growth. *Cancer biology & therapy*, 13(12):1185–1194, 2012.

[253] Stephanos Pavlides, Diana Whitaker-Menezes, Remedios Castello-Cros, Neal Flomenberg, Agnieszka K Witkiewicz, Philippe G Frank, Mathew C Casimiro, Chenguang Wang, Paolo Fortina, Sankar Addya, et al. The reverse warburg effect: aerobic glycolysis in cancer associated fibroblasts and the tumor stroma. *Cell cycle*, 8(23):3984–4001, 2009.

[254] Alice Santi, Anna Caselli, Francesco Ranaldi, Paolo Paoli, Camilla Mugnaioni, Elena Michelucci, and Paolo Cirri. Cancer associated fibroblasts transfer lipids and proteins to cancer cells through cargo vesicles supporting tumor growth. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1853(12):3211–3223, 2015.

[255] Olivier Trédan, Carlos M Galmarini, Krupa Patel, and Ian F Tannock. Drug resistance and the solid tumor microenvironment. *Journal of the National Cancer Institute*, 99(19):1441–1454, 2007.

[256] Chiara Damiani, Riccardo Colombo, Marzia Di Filippo, Dario Pescini, and Giancarlo Mauri. Linking alterations in metabolic fluxes with shifts in metabolite levels by means of kinetic modeling. In *Italian Workshop on Artificial Life and Evolutionary Computation*, pages 138–148. Springer, 2016.

[257] Jeremy S Edwards, Rafael U Ibarra, and Bernhard O Palsson. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125, 2001.

[258] Iman Famili, Jochen Förster, Jens Nielsen, and Bernhard O Palsson. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proceedings of the National Academy of Sciences*, 100(23):13134–13139, 2003.

[259] Jingyi Jessica Li and Mark D Biggin. Statistics requantitates the central dogma. *Science*, 347(6226):1066–1067, 2015.

[260] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the dependency of cellular protein levels on mrna abundance. *Cell*, 165(3):535–550, 2016.

[261] Mattia Zampieri, Karthik Sekar, Nicola Zamboni, and Uwe Sauer. Frontiers of high-throughput metabolomics. *Current opinion in chemical biology*, 36:15–23, 2017.

[262] Marissa Fessenden. Metabolomics: Small molecules, single cells. *Nature*, 540(7631):153–155, 2016.

[263] Olivier B Poirion, Xun Zhu, Travers Ching, and Lana Garmire. Single-cell transcriptomics bioinformatics and computational challenges. *Frontiers in genetics*, 7, 2016.

[264] Anna S Blazier and Jason A Papin. Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in physiology*, 3:299, 2012.

[265] Tomer Shlomi, Moran N Cabili, Markus J Herrgård, Bernhard Ø Palsson, and Eytan Ruppin. Network-based prediction of human tissue-specific metabolism. *Nature biotechnology*, 26(9):1003, 2008.

[266] Joel F Moxley, Michael C Jewett, Maciek R Antoniewicz, Silas G Villas-Boas, Hal Alper, Robert T Wheeler, Lily Tong, Alan G Hinnebusch, Trey Ideker, Jens Nielsen, et al. Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator gcn4p. *Proceedings of the National Academy of Sciences*, 106(16):6477–6482, 2009.

[267] Ali Navid and Eivind Almaas. Genome-level transcription data of yersinia pestis analyzed with a new metabolic constraint-based approach. *BMC systems biology*, 6(1):150, 2012.

[268] Paul A Jensen and Jason A Papin. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics*, 27(4):541–547, 2010.

[269] Kyu-Tae Kim, Hye Won Lee, Hae-Ock Lee, Sang Cheol Kim, Yun Jee Seo, Woosung Chung, Hye Hyeon Eum, Do-Hyun Nam, Junhyong Kim,

Kyeung Min Joo, et al. Single-cell mrna sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome biology*, 16(1):127, 2015.

[270] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, et al. Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications*, 8, 2017.

[271] Patricia do Rosario Martins Conde, Thomas Sauter, and Thomas Pfau. Constraint based modeling going multicellular. *Frontiers in molecular biosciences*, 3, 2016.

[272] Nathan E Lewis, Gunnar Schramm, Aarash Bordbar, Jan Schellenberger, Michael P Andersen, Jeffrey K Cheng, Nilam Patel, Alex Yee, Randall A Lewis, Roland Eils, et al. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nature biotechnology*, 28(12):1279–1285, 2010.

[273] Ruchir A Khandelwal, Brett G Olivier, Wilfred FM Röling, Bas Teusink, and Frank J Bruggeman. Community flux balance analysis for microbial consortia at balanced growth. *PloS one*, 8(5):e64567, 2013.

[274] Benjamin Beck and Cédric Blanpain. Unravelling cancer stem cell potential. *Nature Reviews Cancer*, 13(10):727–738, 2013.

[275] Xin Fang, Anders Wallqvist, and Jaques Reifman. Modeling phenotypic metabolic adaptations of mycobacterium tuberculosis h37rv under hypoxia. *PLoS computational biology*, 8(9):e1002688, 2012.

[276] Alex Graudenzi, Davide Maspero, Marzia Di Filippo, Marco Gnugnolo, Claudio Isella, Giancarlo Mauri, Enzo Medico, Marco Antoniotti, and Chiara Damani. Integration of transcriptomic data and metabolic networks in cancer samples reveals highly significant prognostic power. *Journal of Biomedical Informatics*, 87:37–149, 2018.

[277] Neil Swainston, Kieran Smallbone, Hooman Hefzi, Paul D Dobson, Judy Brewer, Michael Hanscho, Daniel C Zielinski, Kok Siong Ang, Natalie J Gardiner, Jahir M Gutierrez, et al. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, 12(7):1–7, 2016.

[278] Elizabeth Brunk, Swagatika Sahoo, Daniel C Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, Avlant Nilsson, German Andres Preciat Gonzalez, Maike Kathrin Aurich, et al. Recon3d enables a three-dimensional view of gene variation in human metabolism. *Nature biotechnology*, 36(3):272, 2018.

[279] S Beloribi-Djefaflia, S Vasseur, and F Guillaumond. Lipid metabolic reprogramming in cancer cells. *Oncogenesis*, 5(1):e189, 2016.

[280] Jennifer Fazzari, Hanxin Lin, Cecilia Murphy, Robert Ungard, and Gurmit Singh. Inhibitors of glutamate release from breast cancer cells; new targets for cancer-induced bone-pain. *Scientific reports*, 5, 2015.

[281] Guillermo Mariño and Guido Kroemer. Ammonia: a diffusible factor released by proliferating cells that induces autophagy. *Sci. Signal.*, 3(124):pe19, 2010.

[282] Christina H Eng, Ker Yu, Judy Lucas, Eileen White, and Robert T Abraham. Ammonia derived from glutaminolysis is a diffusible regulator of autophagy. *Sci. Signal.*, 3(119):ra31–ra31, 2010.

[283] Johan H van Heerden, Meike T Wortel, Frank J Bruggeman, Joseph J Heijnen, Yves JM Bollen, Robert Planqué, Josephus Hulshof, Tom G O?Toole, S Aljoscha Wahl, and Bas Teusink. Lost in transition: start-up of glycolysis yields subpopulations of nongrowing cells. *Science*, 343(6174):1245114, 2014.

[284] Robert Schuetz, Lars Kuepfer, and Uwe Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli. *Molecular systems biology*, 3(1):119, 2007.

[285] Shannon K Oda, Pamela Strauch, Yuko Fujiwara, Amin Al-Shami, Tamas Oravecz, Gabor Tigyi, Roberta Pelanda, and Raul M Torres. Lysophosphatidic acid inhibits cd8 t-cell activation and control of tumor progression. *Cancer immunology research*, 1(4):245–255, 2013.

[286] C Damiani, R Serra, M Villani, SA Kauffman, and A Colacci. Cell–cell interaction and diversity of emergent behaviours. *IET systems biology*, 5(2):137–144, 2011.

[287] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki,

Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[288] Shanmugasundaram Ganapathy-Kanniappan. Taming tumor glycolysis and potential implications for immunotherapy. *Frontiers in oncology*, 7:36, 2017.

[289] Anne Kümmel, Sven Panke, and Matthias Heinemann. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC bioinformatics*, 7(1):512, 2006.

[290] Christopher S Henry, Linda J Broadbelt, and Vassily Hatzimanikatis. Thermodynamics-based metabolic flux analysis. *Biophysical journal*, 92(5):1792–1805, 2007.

[291] Patrick S Ward and Craig B Thompson. Signaling in control of cell growth and metabolism. *Cold Spring Harbor perspectives in biology*, page a006783, 2012.

[292] Elaina M Maldonado, Vytautas Leoncikas, Ciarán P Fisher, J Bernadette Moore, Nick J Plant, and Andrzej M Kierzek. Integration of genome scale metabolic networks and gene regulation of metabolic enzymes with physiologically based pharmacokinetics. *CPT: pharmacometrics & systems pharmacology*, 6(11):732–746, 2017.

[293] Markus Krauss, Stephan Schaller, Steffen Borchers, Rolf Findeisen, Jörg Lippert, and Lars Kuepfer. Integrating cellular metabolism into a multiscale whole-body model. *PLoS computational biology*, 8(10):e1002750, 2012.