Novel metabolic network reconstruction algorithms for -omics data integration and in-silico gene essentiality analysis in cancer

PhD dissertation

Luis Tobalina Segura

Biomedical Engineering department
TECNUN
University of Navarra

December 2015

Novel metabolic network reconstruction algorithms for -omics data integration and in-silico gene essentiality analysis in cancer

Dissertation Submitted for the Degree of Doctor of Philosophy

Under the supervison of
Francisco Javier Planes Pedreño

Biomedical Engineering department
TECNUN
University of Navarra

December 2015

A mi familia

Acknowledgements

First of all, I would like to thank Francisco J. Planes for offering me the opportunity to be part of this exciting adventure. Thank you for your continuous support and help, your guidance and passion have made this thesis possible.

I want to show my gratitude to CEIT, TECNUN and the University of Navarra for all these years of training and education, first as an undergraduate student and then as a PhD student. In addition, I want to acknowledge the Basque Government for providing me with the financial support necessary to develop this research work.

I am very grateful to all my workmates, from which I have learned a lot, as well as the many students that have been around the department helping the group explore new research directions. All of you have contributed to an incredible work atmosphere. Of all my department colleagues I would like to highlight Jon Pey, Alberto Rezola and Ander Aramburu: it has been a real pleasure to share ideas and hopes with you.

I also enjoyed the meetings with Felipe Prosper and his group, which helped us to better understand the experiments and focus the aim of our research. Also, thanks to Manuel Ferrer for giving us the opportunity to collaborate in a very interesting project.

I would also like to thank Julio Saez-Rodriguez for accepting me as a visiting PhD student in his group at the European Bioinformatics Institute (EMBL-EBI) for a short research stay. It was a very enriching experience and I had the opportunity to meet wonderful people during those three months I spent in Cambridge.

Special thanks to my friends, which have provided me with support and unforgettable moments. I am very lucky to have you by my side. Please, forgive me for not naming you all explicitly. Gracias, Gorka, Eneko, Nacho, Marco e Ismael por vuestra amistad. Eskerrik asko Kuadrilla por todos los momentos que hemos compartido juntos. Eskerrik asko Leire, Lasa, Ainara, Cristian, Itziar, Miren, Lierni eta Jon. Izugarrizko zortea daukat zuek bezalako lagunak izatea.

Finally, I would like to thank my family for their unconditional support. Aita, Ama, Marian, Tieta, muchas gracias por todo vuestro apoyo y amor incondicional. Bihotz-bihotzez, eskerrik asko.

Summary

In order to defeat cancer, we need to understand its biology. The study of metabolism is an active area of research in cancer nowadays. Some Systems Biology techniques used for analyzing cancer metabolism contextualize prior biological knowledge with experimental data before further analysis aimed at finding weak points in cancer cells is conducted. Not only cancer but also bacterial communities living in our bodies have an impact on our health, hence methods for their study are also needed. The purpose of this doctoral thesis is to improve automated network reconstruction techniques and apply them to obtain new insights from experimental data and find essential genes for cancer survival. It also aims to find new methods that can be used to integrate experimental data and improve the prediction accuracy of therapeutic targets.

This thesis introduces two novel fast network reconstruction algorithms. One of them is focused on bacterial communities and the use of metaproteomic and taxonomic data. The other is focused on gene expression data coming from cancer samples. The latter algorithm allows us to evaluate a current in-silico approach used for finding essential metabolic genes against experimentally obtained high-throughput gene essentiality data. Finally, a new method that answers the question of what other reactions in a metabolic network make a given one essential is developed, opening the possibility to new methods of integrating experimental data with metabolic networks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Our knowledge of cancer biology has experienced a significant growth during the last decades, but the decline in cancer mortality has been relatively modest. This decrease has been mainly attributed to screenings and preventive measures. It is therefore a clinical need to develop new treatments with a direct impact on survival rates. At the same time, there is an increasing interest in the study of bacterial communities, because they play an important role in human health. Emerging Systems Biology techniques grow on the huge existing biological knowledge trying to reconcile apparently separated observations into a whole and providing new interpretations that can reveal new treatment strategies.

In both cases, the study of cellular metabolism is of paramount importance. Particularly, in cancer, recent findings show that cancer cells adapt their metabolic processes to enhance proliferation (Kaelin and Thompson, 2010; Vander Heiden et al., 2009). To that end, cancer cells consume additional nutrients and divert those nutrients into macromolecular synthesis pathways. Apart from alterations in glucose metabolism, the so called Warburg effect, more have been reported in the synthesis of nucleotides, amino acids and lipids (Vander Heiden, 2011; Vander Heiden et al., 2010). In addition, relevant mutations in metabolic genes and accumulations of key metabolites have been detected in cancer cells (Dang et al., 2009). In light of these evidences, the study of cellular metabolism in cancer research has been actively reawakened. Holistic systems biology approaches, based on genome-scale metabolic networks and high-throughput "omics" data, open new avenues to exploit metabolic disorders of tumour cells, particularly for addressing different unmet clinical needs in cancer.

1.1. Metabolism

Metabolism describes the chemical reactions that convert nutrients and other molecules into the energy and building blocks that sustain an organism's life (Kaelin and Thompson, 2010). Most of these reactions require the intervention of a particular set of proteins, known as enzymes, to occur at a high enough rate. Without the enzymes catalyzing these reactions, they would occur naturally at very low rates.

Enzymes, like any other protein, are composed by a sequence of aminoacids that fold in space to form three dimensional structures. They are built by ribosomes (a protein complex), that assemble the aminoacid chain following the instructions contained in messenger RNA (mRNA) in a process known as translation. mRNA molecules are, in turn, a sequence of nucleotides that convey information from the DNA contained in the nucleus. The process of obtaining mRNA from DNA is known as transcription. Those regions of the DNA that give rise to mRNA that encode proteins are known as genes. This explanation on how the information flows from DNA to RNA to proteins receives the name of the central dogma of molecular biology.

Going back to the description of metabolism, it is usually divided into two main characteristic processes: anabolism and catabolism. Anabolism is the process by which the cell builds large and complex molecules from simple precursors. This process consumes energy, usually in the form of adenosine triphosphate (ATP), the main energy currency of the cell. Catabolism, on the other hand, is the process by which the cell obtains its required energy. The process involves breaking down nutrient molecules into simpler compounds, storing the energy released in the form of ATP and reduced electron carriers (NADH, NADPH and FADH2).

The study of metabolism has been traditionally accomplished using the concept of pathways, since the biochemical reactions that constitute it depend on each other. Pathways help to understand how a given metabolite can be transformed into another one, the different steps that take part in the process and the additional requirements and byproducts. Some examples of well studied pathways are glycolysis, the TCA cycle or the pentose phosphate pathway. However, pathways are also interrelated, as they may share reactions or metabolites (Figure 1.1). The novel behaviours that may arise from pathway interactions may be overlooked if they are studied in isolation. Hence, during the last years, the study of metabolism using the concept of networks has gained popularity. In particular, the development of genome-scale metabolic networks (GSMNs) containing all the reactions and pathways (or at least a great part of them) known to be part of an organism's metabolism, and the emergence of new analysis methods, such

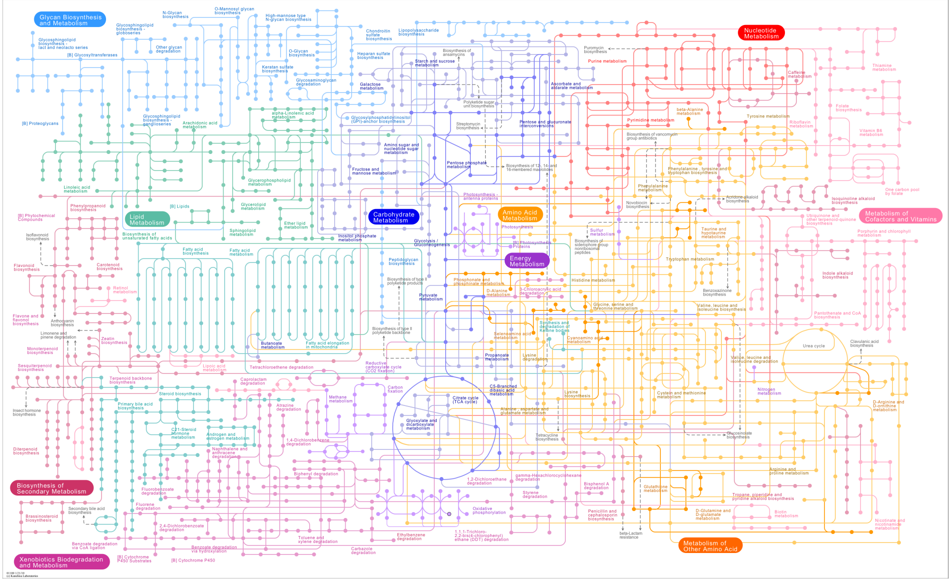

Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Carbohydrate Metabolism
Lipid Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Secondary Metabolism
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid

Figure 1.1: A general overview of metabolic pathways (Kanehisa et al., 2012).

as constraint based modeling, have expanded the possibilities and scope of metabolic research.

1.2. Systems Biology

Systems Biology (SB) aims at a system-level understanding of biological systems (Kitano, 2002a,b), i.e. it tries to understand how biological components interact to give rise to the observed phenotypes. There are four steps in this process (Palsson, 2006). First, the list of biological components part of the system to be studied should be identified. Second, the way these components interact between them should be described. Third, mathematical models are used to describe and analyze the properties of these systems. Fourth, computer models are generated to analyze, interpret and predict the behaviours that can arise from these systems. This process generates hypothesis that can be experimentally tested, and the results obtained from experiments can be used to improve the model.

The first use of the term Systems Biology can be traced back to the 1960s (Noble, 1960), although the origins of the field may go as back as to 1885 (Schneider, 2013). However, it was not until the beginning of the 90s that the field really took off, following the development of high-throughput tools, the generalization on the use of the Internet and increased computational power (Schneider, 2013). Ever since, the need to put the big amount of generated data into context and extract useful insights as well as testable predictions has boosted the interest on Systems Biology research.

1.2.1. Constraint Based Modeling

Several mathematical approaches exist in the field of Systems Biology. For the case of metabolic systems, constraint-based modelling (CBM) techniques have gathered a great deal of attention. These techniques revolve around the stoichiometric matrix (Llaneras and Picó, 2008) and have been fostered by the progress in reconstruction of genome-scale metabolic networks (GSMNs).

An stoichometric matrix $S_{C\times R}$ is a mathematical representation of a metabolic network with C metabolites and R reactions. Each element s_{cr} of the stoichiometric matric S represents the stoichiometric coefficient with which compound c participates in reaction r. More specifically, substrates of a reaction are represented by negative stoichiometric coefficients, while products are represented by positive coefficients.

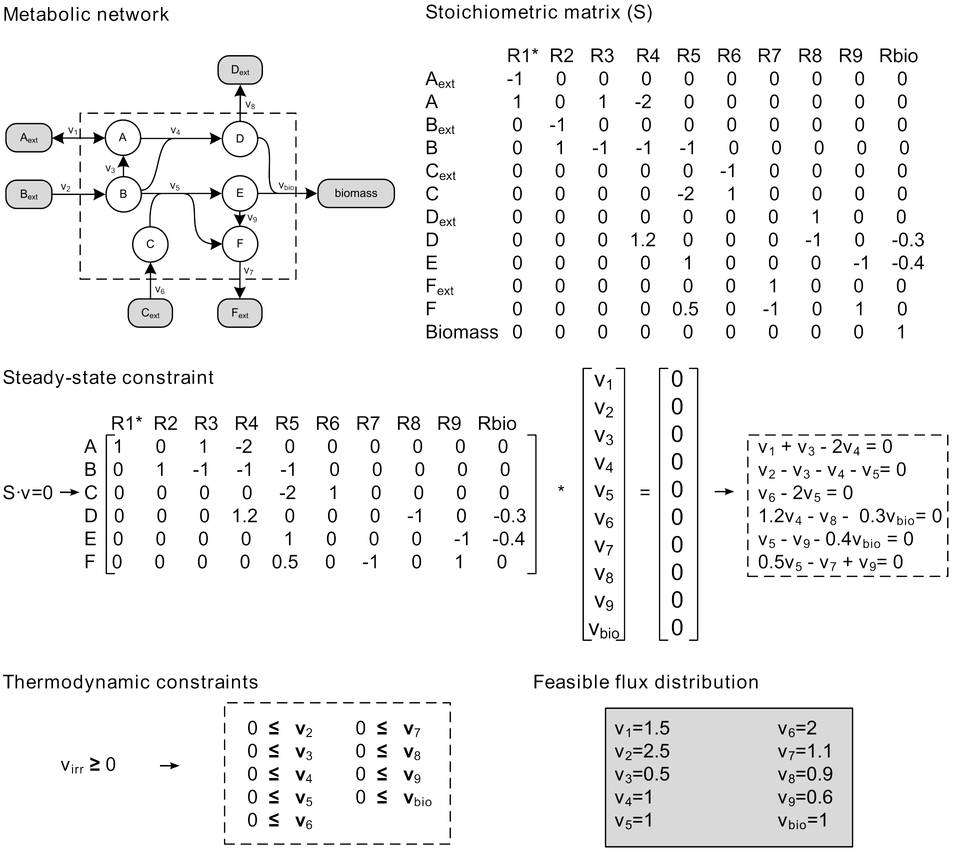
In this framework, the activity of each reaction is represented by a flux variable v_r $(r = 1, ..., R)$. By multiplying the stoichiometric coefficient s_{cr} by the flux of the reaction v_r, we obtain the number of molecules of metabolite c that are consumed (if $s_{cr} < 0$) or produced (if $s_{cr} > 0$) by the reaction.

The stoichiometric matrix and the flux vector are the core elements of the CBM techniques. Upon these, the two basic constraints of most CBM methods are defined, namely the steady-state assumption and the thermodynamic feasibility of fluxes (Figure 1.2).

The steady-state rises from the mass balance assumption within the cell, i.e. that the concentration of metabolites remains constant over time. The consumption rate of a metabolite c must equal its production rate, so that no accumulation or depletion of metabolites is given inside the system boundaries. Those metabolites that happen to accumulate or deplete do not need to fulfill this constraint, but they can be included in this constraint by including reactions with only products or only educts that model inputs and outputs to the system, respectively. Exchange reactions (not to be confused with transport reactions), demand reactions and sink reactions are some examples of this type of reactions.

$$\sum_{i=1}^{R} S_{ci} v_i = 0 \qquad \forall c \in C \tag{1.1}$$

Thermodynamic feasibility constraints impose limits on the values that fluxes can take. In their simplest form, they limit some of the fluxes to being



Stoichiometric matrix (S)

	R1*	R2	R3	R4	R5	R6	R7	R8	R9	Rbio
A_{ext}	-1	0	0	0	0	0	0	0	0	0
A	1	0	1	-2	0	0	0	0	0	0
B_{ext}	0	-1	0	0	0	0	0	0	0	0
B	0	1	-1	-1	-1	0	0	0	0	0
C_{ext}	0	0	0	0	0	-1	0	0	0	0
C	0	0	0	0	-2	1	0	0	0	0
D_{ext}	0	0	0	0	0	0	0	1	0	0
D	0	0	0	1.2	0	0	0	-1	0	-0.3
E	0	0	0	0	1	0	0	0	-1	-0.4
F_{ext}	0	0	0	0	0	0	1	0	0	0
F	0	0	0	0	0.5	0	-1	0	1	0
Biomass	0	0	0	0	0	0	0	0	0	1

Steady-state constraint

$$S\cdot v=0 \rightarrow \begin{array}{c|cccccccccc} & R1^* & R2 & R3 & R4 & R5 & R6 & R7 & R8 & R9 & Rbio \\ \hline A & 1 & 0 & 1 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ B & 0 & 1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ C & 0 & 0 & 0 & 0 & -2 & 1 & 0 & 0 & 0 & 0 \\ D & 0 & 0 & 0 & 1.2 & 0 & 0 & 0 & -1 & 0 & -0.3 \\ E & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & -0.4 \\ F & 0 & 0 & 0 & 0 & 0.5 & 0 & -1 & 0 & 1 & 0 \end{array} * \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \\ v_8 \\ v_9 \\ v_{bio} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow \begin{array}{l} v_1 + v_3 - 2v_4 = 0 \\ v_2 - v_3 - v_4 - v_5 = 0 \\ v_6 - 2v_5 = 0 \\ 1.2v_4 - v_8 - 0.3v_{bio} = 0 \\ v_5 - v_9 - 0.4v_{bio} = 0 \\ 0.5v_5 - v_7 + v_9 = 0 \end{array}$$

Thermodynamic constraints

$$v_{irr} \geq 0 \rightarrow \begin{array}{ll} 0 \leq v_2 & 0 \leq v_7 \\ 0 \leq v_3 & 0 \leq v_8 \\ 0 \leq v_4 & 0 \leq v_9 \\ 0 \leq v_5 & 0 \leq v_{bio} \\ 0 \leq v_6 & \end{array}$$

Feasible flux distribution

v_1=1.5	v_6=2
v_2=2.5	v_7=1.1
v_3=0.5	v_8=0.9
v_4=1	v_9=0.6
v_5=1	v_{bio}=1

Figure 1.2: Toy metabolic network that illustrates the basics of constraint-based modelling. Reproduced with permission from Rezola-Urquía (2013).

non-negative in order to represent the irreversibility of some reactions under physiological constraints. Many GSMN reconstructions already provide information on whether a reaction should be considered reversible (*Rev*) or irreversible (*Irr*).

$$0 \leq v_i \qquad \forall i \in Irr \tag{1.2}$$

Flux Balance Analysis (FBA) (Orth et al., 2010) is one of the most important CBM techniques. Together with the steady-state and thermodynamic feasibility constraints, it introduces an objective function to be optimized, as well as capacity bound constraints on the reaction fluxes. The mathematical representation of thermodynamic feasibility constraints and capacity bound constraints is very similar, and usually they are simply represented as bound constraints on the values of flux variables. It is the objective function, however, the most prominent aspect of FBA. A typical example is the biomass objective function (Feist and Palsson, 2010), which is used to calculate the theoretical maximum rate at which the organism or cell represented by the metabolic network can grow under specific conditions (growth medium can be described through the flux bound constraints). Overall, an FBA problem is a linear programming optimization problem of the following form:

$$\text{maximize } c^T v \tag{1.3}$$

$$\text{subject to:}$$

$$Sv = 0 \tag{1.4}$$

$$v_i^{lb} \leq v_i \leq v_i^{ub} \qquad \forall i \in R \tag{1.5}$$

The solution to an FBA problem is a flux distribution v that optimizes our objective, but this solution may not be unique (Mahadevan and Schilling, 2003). Flux Variability Analysis (FVA) (Mahadevan and Schilling, 2003; Gudmundsson and Thiele, 2010) was designed as a tool to find the maximum and minimum flux values that a reaction can take under some network state (e.g. while the biomass production stays above a 90 % of the maximum). For each reaction, a maximization and a minimization problem is solved, where the objective function is the flux through the reaction of interest. Although linear programs can be efficiently solved, a direct implementation of FVA, which iterates through all the reactions solving one maximization and one minimization each time, can take a good amount of time. Fortunately, the process can be sped up by leveraging the way optimization solvers handle the solution process (Gudmundsson and Thiele, 2010).

A widespread application of FBA is the search of essential genes (Edwards and Palsson, 2000; Folger et al., 2011). Essential genes are defined as those genes whose removal render the cell unable to produce biomass. If this happens, the cell is unable to meet its requirements for growth and it will ultimately die. Using Boolean gene-protein-reaction rules that relate the reactions of the metabolic network to the genes of the organism, we can evaluate which reactions will stop working after a particular gene is deleted. Thus, a gene knock-out is simulated by setting the upper and lower bounds of the corresponding reactions to zero in an FBA calculation, and checking whether the remaining network is still able to produce biomass or not.

1.2.2. Pathway analysis

Metabolic pathway analysis (Schilling et al., 1999) is another approach for the study of genome-scale metabolic networks. The idea of pathways is very attractive from a biological point of view, because they reveal essential functioning parts of metabolism and aid in their understanding. It is very common among biologists to be interested in a specific pathway, such as in a chemical producing pathway, so that they can study it in more detail and engineer it. The emergence of genome-scale networks opened up the possibility to discover novel pathways aided by computational tools followed by experimental validation.

In this framework, the concept of Elementary Flux Mode (EFM) (Schuster and Hilgetag, 1994; Zanghellini et al., 2013) has gained popularity. These can be described as minimal steady-state flux distributions of a metabolic network. A key property of EFMs is that they fulfill the non-decomposability condition, i.e. they cannot be decomposed into smaller flux distributions without contradicting the steady-state condition. This translates into EFMs having a single degree of freedom, i.e. once the flux through one of its reactions is known, the fluxes through the rest of the reactions in the EFM are automatically set. Mathematically, in a network composed of only irreversible reactions (any reversible reaction can be separated into two irreversible reactions), EFMs are defined as the extreme rays of the flux cone P.

$$P = \{v | Sv = 0, v_r \geq 0, \forall r\} \tag{1.6}$$

As can be observed, the flux cone P is also present in FBA constraints. Likewise, linear programming techniques can be used to compute EFMs.

The concept of Elementary Flux Mode is not the only one in metabolic pathway analysis, but it is probably the one that is most extended. Other alternatives include Generating Flux Modes (GFMs) (Pfeiffer et al., 1999), extreme currents (ECs) (Clarke, 1988), extreme pathways (ExPas) (Schilling et al., 2000), or Carbon Flux Paths (CFPs) (Pey et al., 2011).

One of the problems with EFMs is that their number increases exponentially with network size. This poses computational and methodological challenges for their computation. During the last years, several methods have been developed to compute EFMs (de Figueiredo et al., 2009; Kaleta et al., 2009; Kamp and Schuster, 2006; Machado et al., 2012; Pey and Planes, 2014; Pey et al., 2015; Rezola et al., 2011; Terzer and Stelling, 2008; Urbanczik and Wagner, 2005; Quek and Nielsen, 2014), many of which centered around calculating as many as possible. In some cases, however, EFMs of some specific characteristics are the ones we are interested in. In those cases, it would be desirable to have a method that directly calculates them instead of having to compute a large set first and then applying a filtering criterion. A solution to this problem was proposed very recently, using a MILP model that directly calculates EFMs satisfying a desired set of constraints (Pey and Planes, 2014).

One possible way of using EFMs is to integrate them with experimental data (Rezola et al., 2015). This way, we can understand the data in a meaningful biological context without being limited to the classical canonical pathways.

1.2.3. Gene-Protein-Reaction rules

We have mentioned previously that Boolean gene-protein-reaction rules (GPR rules) relate reactions of the metabolic network to the genes of the organism. In this section, we explain these rules that allow us to go from genes to reactions and map experimental gene expression data onto reactions in our metabolic network models.

The theoretical background of these rules comes from the central dogma of molecular biology, which explains how the information contained in the DNA is transcribed into RNA and then translated into proteins. When those proteins are enzymes, which catalyze reactions, we can establish a link between those genes and the reactions being catalyzed. However, the reality is that the relation is not always one-to-one, as different enzymes may catalyze the same reaction and some other reactions may be catalyzed by enzymes complexes (more than one enzyme that bind together to form a new complex with the ability to catalyze the reaction). On top of that, a single enzyme or enzymatic complex may have the ability to catalyze different reactions.

We can encode these occurrences using boolean rules. If the reaction is only catalyzed by one enzyme and that enzyme is only coded by one gene, the relation is straightforward. If the relation is catalyzed by a enzyme complex, the relation is given by an AND rule, where the participants in this AND rule are the genes that contain the sequences for the different subunits of the enzymatic complex. Finally, if the reaction can be catalyzed by different enzymes, we use an OR rule involving the genes that code the different enzymes capable of catalyzing the reaction. Of course, an GPR rule may contain a mix of OR and AND operators, as one reaction may be catalyzed by single enzymes or enzymatic complexes and the enzymatic complexes may, in turn, be composed of alternative subunits (Figure 1.3).

To translate gene classification or expression values into reaction classification or expression values, we can make use of the GPR rules by substituting the AND operator by a min() function and the OR operator by a max() function. The reason is simple, if the rule is of type AND, the element that has the lowest value will be setting an upper bound on the result, whereas if the rule is of type OR, the highest value will be setting a lower bound on the final result.

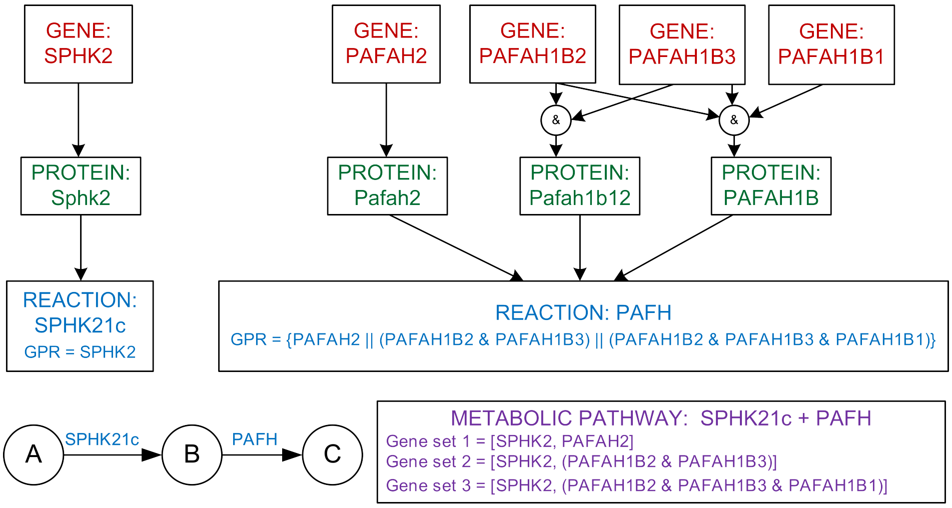

GENE: SPHK2
PROTEIN: Sphk2
REACTION: SPHK21c
GPR = SPHK2
GENE: PAFAH2
GENE: PAFAH1B2
GENE: PAFAH1B3
GENE: PAFAH1B1
&
&
PROTEIN: Pafah2
PROTEIN: Pafah1b12
PROTEIN: PAFAH1B
REACTION: PAFH
GPR = {PAFAH2 || (PAFAH1B2 & PAFAH1B3) || (PAFAH1B2 & PAFAH1B3 & PAFAH1B1)}
A
SPHK21c
B
PAFH
C
METABOLIC PATHWAY: SPHK21c + PAFH
Gene set 1 = [SPHK2, PAFAH2]
Gene set 2 = [SPHK2, (PAFAH1B2 & PAFAH1B3)]
Gene set 3 = [SPHK2, (PAFAH1B2 & PAFAH1B3 & PAFAH1B1)]

Figure 1.3: Gene Protein Reaction (GPR) rules for two different reactions: SPHK21c and PAFH (Schellenberger et al., 2010). Genes (in red) lead to proteins (in green) that catalyze reactions (in blue). A metabolic pathway can be active if one of the gene sets (in purple) related to it is expressed. Reproduced with permission from Rezola-Urquía (2013).

1.3. Omics technologies

Systems Biology approaches have been greatly promoted by the generation of huge amounts of data at different levels of cell biology. The development of new experimental techniques with a high-throughput data output during the last decades have completely changed the biological research landscape. High-throughput molecular data technologies are sometimes referred to as -omics technologies (e.g. genomics, transcriptomics, proteomics or metabolomics).

1.3.1. Transcriptomics

Transcriptomics aims to study all the different kinds of RNA molecules present in a cell, such as mRNA and non-coding RNA. One of its most extended uses is the quantification of gene expression using DNA microarrays (Schena et al., 1995). Microarray technologies are based on the artificial hybridization of cDNA sequences obtained from mRNA of the studied samples in a matrix filled with complementary sequences. These complementary sequences receive the name of probes, and a group of them, targeting all the same sequence, are known as probe-sets. One gene may be interrogated by a group of different probe-sets, but one probe-set may hybridize against

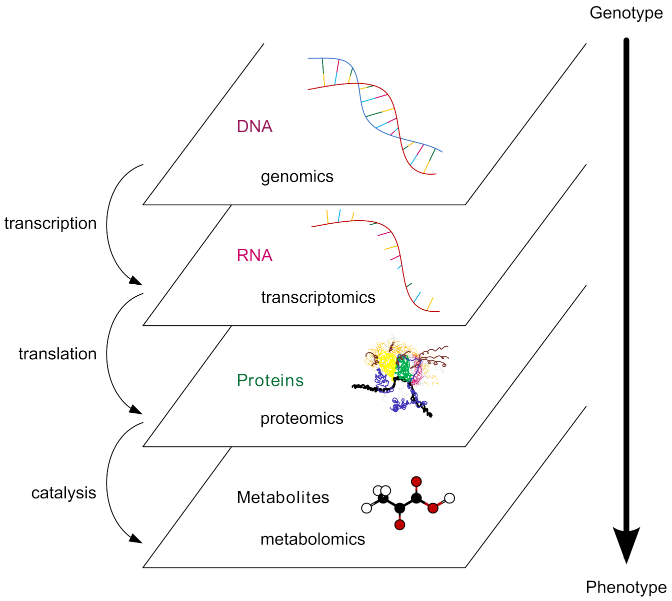


Figure 1.4: Overview of -omics data that can be interrogated at different levels of cell biology. Adapted with permission from Rezola-Urquía (2013).

several genes if its sequence is non-specific. A wide array of methods have been developed to deal with these and other experimental issues inherent to microarrays in order to correctly interpret the data they generate.

The popularity of cDNA microarrays has been enhanced by standardized processing pipelines and public repositories such as Gene Expression Omnibus or ArrayExpress, with thousands of experiments readily downloadable. During the last years, however, RNA sequencing (RNA-Seq) technologies are starting to take over, due to their capacity to interrogate the samples at a finer level of detail. Notwithstanding, the use of microarray experimental data remains attractive thanks to the aforementioned enormous quantity of freely available experimental data.

1.3.2. Proteomics

Proteomics deals with the quantification of all the proteins present in a sample. Unfortunately, the coverage of proteomics technologies is not as wide as that of transcriptomics technologies. High-throughput proteomics experiments normally use mass-spectrometry techniques to quantify them (Aebersold and Mann, 2003).

A mass spectrometer consists of an ion source, a mass analyser and a detector. To use it for detecting proteins, these must be first fragmented and ionized to generate charged molecule fragments. The mass analyser records the mass-to-charge ratio (m/z) of the ionized analytes while the detector keeps track of the number of ions at each m/z value. The result of the experiment is a complex spectrum that must undergo a deconvolution process. Groups of peaks in the spectrum can be related to specific fragments, and it is known how each protein gives rise to a different set of fragments. This way, knowing the abundance of each fragment, proteins in the sample can be identified and quantified.

Proteomics data are not as common as microarray expression data. One of the few centralized databases is the Human Protein Atlas (HPA) (Uhlen et al., 2010), where information about the presence of proteins in several normal and tumoral tissues can be found. The data in HPA however is based on immunohistochemistry assays, an alternative to mass-spectrometry for the measurement of protein expression profiles.

1.3.3. Metabolomics

Metabolomics aims to characterize and quantify the metabolites or small molecules present in a given sample. Like in proteomics, mass-spectrometry related techniques are regularly used (Dettmer et al., 2007). However, metabolomics experiments pose substantially higher challenges because they deal with much smaller molecules, with a great diversity in any given sample and present in concentrations that may vary orders of magnitude between each other. As if this was not enough, residual enzymatic activity after sample collection or oxidation processes may lead to unwanted formation or degradation of metabolites. On top of that, a sample preparation step is usually required, which can cause unwanted losses or the discrimination of some metabolite classes.

An alternative to mass-spectrometry is nuclear magnetic resonance (NMR) spectroscopy (Powers, 2009). NMR is based on the application of magnetic fields to the sample, which play with the magnetic spin configuration of the atomic nucleus. The output of the technique is again an spectrum that must be deconvoluted to infer the identity and amount of the metabolites present in the sample. More than a competitor technique, NMR should be considered as a complementary technique to the mass-spectrometry based methods.

Despite these difficulties, metabolomics is a very necessary technology, as is the closest level to the observed phenotype.

1.4. Goals and outline of this thesis

This thesis aims to improve and develop novel constraint based modeling techniques to reconstruct and contextualize genome-scale metabolic networks using -omics data. Given the importance of metabolism in different areas of biotechnology and human health, this question is receiving much attention in Systems Biology. In addition, we aim to evaluate and improve current methods to conduct FBA-based Gene Essentiality Analysis in cancer, a promising approach for identifying novel therapeutic strategies. The specific objectives are:

- Develop new metabolic network reconstruction and contextualization algorithms. These algorithms will take into account what are the resulting networks going to be used for (what type of data feeds them and what insights would we like to extract from it) and perform much faster than existing algorithms, so that large-scale and computer-intensive studies can be carried out.

- Evaluate the influence of expression data in the obtention of essential genes with constraint based modeling techniques. Evaluate also the accuracy of the results using experimental high-throughput gene silencing data.

- Develop the theoretical basis to design new methods for the direct integration of experimental data with metabolic networks useful for the prediction of essential genes.

The thesis is organized as follows. After Chapter 1, which has provided an overview of some basic concepts the work presented in this thesis relies on, Chapter 2 gives an overview of different network reconstruction methods, providing a general view of how the problem has been approached, and introduces a new fast reconstruction algorithm that uses gene expression data to contextualize metabolic networks amenable to FBA. Chapter 3 introduces a second reconstruction algorithm tailored towards metaproteomic data obtained from bacterial communities. It also provides an interesting study case of samples coming from contaminated soil. Chapter 4 uses the reconstruction algorithm introduced in Chapter 2 to evaluate the results of FBA based Gene Essentiality Analysis and compares them with experimental high-throughput gene essentiality data. Chapter 5 develops a new algorithm for finding reactions that make a specific reaction become essential in an FBA based Gene Essentiality Analysis. This method opens up new possibilities for the integration of experimental data in the context of essentiality analysis. Finally, Chapter 6 summarizes the work and conclusions of the thesis, and outlines possible future lines of research.

Chapter 2

Metabolic Network Reconstruction

In this chapter, we provide a general view of different methods used for network reconstruction. After that, we introduce a new fast reconstruction algorithm that uses gene expression data to contextualize metabolic networks that can be directly analyzed using FBA.

2.1. Introduction

The main input to any CBM method is a metabolic network. The process of building the metabolic model representation of an organism or cell of interest is known as network reconstruction or network contextualization.

Sometimes, the starting point to build the network is an annotated genome and a reference metabolic database. This is common when dealing with bacteria, for example. Thiele and Palsson (2010), described the steps for accurately building a metabolic network, which can be time-consuming, easily taking from 6 months to 2 years. Given the number of existing organisms, methods that help in this process are more than welcome. For instance, the Model SEED provides an integrative and automatic approach that substantially speeds up the time required to obtain a first network draft (Henry et al., 2010).

Other times, we have a good representation of the metabolism of an organism, but we need to contextualize it for a particular situation (Becker and Palsson, 2008). For example, in the case of human metabolism, we have a bunch of models that gather the reactions known to take place in human cells, but it is obvious that each cell type uses only a subset of all those reactions. Thus, in order to capture cell-specific metabolic features, the reference network must be contextualized with available experimental data. Again, the manual process of building a reliable context-specific metabolic network is complex and time consuming, and this has led to research in automatic network reconstruction algorithms.

Given the wealth of transcriptomic data, mRNA expression data is the most frequent type of data used in the different reconstruction methods available in the literature. A non-exhaustive list of this type of methods includes: GIMME (Becker and Palsson, 2008), iMAT (Shlomi et al., 2008), E-Flux (Colijn et al., 2009), MBA (Jerby et al., 2010), PROM (Chandrasekaran and Price, 2010), MADE (Jensen and Papin, 2011), INIT (Agren et al., 2012), or MIRAGE (Vitkin and Shlomi, 2012).

In this chapter, we first give an overview of some of the metabolic network reconstruction and contextualization algorithms available. We show that they follow a similar line of thought in the way they understand that reactions should be included or excluded from the reconstruction based on the available experimental data. However, each one of them possess their own subtleties, given mainly because of the application their authors had in mind when designing them. Next, we introduce a novel fast reconstruction algorithms designed during this thesis. This algorithm was motivated by the need of a fast reconstruction algorithm for the assessment of gene essentiality based FBA that will be discussed in chapter 4. In addition, in Chapter 3 we will introduce a second algorihtm, motivated by the need to integrate metaproteomic data from two bacterial communities in order to discern differences between them.

2.2. Network Reconstruction

Network reconstruction algorithms address the problem starting with a group of reactions that should be present based on previous experimental evidence, typically gene or protein expression levels. These reactions do not usually form a coherent network (Satish Kumar et al., 2007). Indeed, they are not necessarily connected to each other, may form separated clusters or even be isolated from the rest. Thus, reconstruction algorithms fill in the gaps until a coherent network is obtained. Hypothesized reactions come from a database of known biochemical reactions, generally associated with the organism under study. In addition, it is also typical to avoid some reactions in the reconstruction because of experimental evidence of their absence (Shlomi et al., 2008).

Current reconstruction algorithms typically rely on Mixed Integer Linear Programming (MILP). It is also the case that each reconstruction algorithm is usually focused towards the integration of a different type of one or more input experimental information. Because of this, in most cases, the results obtained from each one of them are not easily comparable. Although reconstructed networks are normally used in CBM analysis, not every algorithm returns a network directly amenable to the analysis of interest. For instance, networks to be used in FBA based GEA should be able to produce biomass while they fulfill the steady state condition. However, most reconstruction algorithms are designed to guarantee the latter but not the former.

2.2.1. iMAT

Shlomi et al. (2008) designed an algorithm to integrate tissue specific gene- and protein-expression data with the global human network reconstruction. We refer to it as iMAT because of the web-based tool that implements it (Zur et al., 2010). This method starts by classifying genes into significantly highly and lowly expressed. The ones that do not fall in any of those two groups are considered as moderately expressed. Subsequently, gene-to-reaction boolean rules are used to classify reactions in the network in one of those three groups. Highly expressed reactions are denoted as R_H and lowly expressed reactions as R_L. After that, a mixed-integer linear programming formulation is used to find a steady-state flux distribution that maximizes the number of reactions whose activity is consistent with their expression state (using y^+ and y^- binary variables) and at the same time satisfies stoichiometric and thermodynamic constraints.

$$\max_{v,y^+,y^-} \Big(\sum_{i \in R_H} (y_i^+ + y_i^-) + \sum_{i \in R_L} y_i^+ \Big) \tag{2.1}$$

subject to

$$S \cdot v = 0 \tag{2.2}$$

$$v_{min} \leq v \leq v_{max} \tag{2.3}$$

$$v_i + y_i^+ (v_{min,i} - \epsilon) \geq v_{min,i} \qquad i \in R_H \tag{2.4}$$

$$v_i + y_i^- (v_{max,i} + \epsilon) \leq v_{max,i} \qquad i \in R_H \tag{2.5}$$

$$v_{min,i}(1 - y_i^+) \leq v_i \leq v_{max,i}(1 - y_i^+) \qquad i \in R_L \tag{2.6}$$

$$v \in R^m \tag{2.7}$$

$$y_i^+, y_i^- \in [0, 1] \tag{2.8}$$

As the solution to the MILP problem may not be unique, i.e. there may be several different solutions with the same objective function value, the authors suggested to use a variant of Flux Variability Analysis to account for them (Shlomi et al., 2008). The proposed approach consists in solving the iMAT problem twice for each reaction, one with the reaction forced to be active and another one with the reaction forced to be inactive. The objective values obtained in these two runs of the algorithm are compared and their difference give a confidence score on whether the reaction should be active or not. If both objective values turn out to be the same, the reaction is considered to be in an undetermined activity state.

2.2.2. MBA

The model-building algorithm (MBA) (Jerby et al., 2010) was designed to generate tissue-specific models from the generic human model by integrating different molecular data sources. These data sources can include literature-based knowledge, transcriptomic, proteomic, metabolomic or even phenotypic data. It was conceived as an alternative to iMAT that could overcome the high computational demands imposed by the many tissue-specific data sources on a mixed-integer linear programming formulation. The core idea is to heuristically prune the generic human metabolic network to obtain a sub-network as consistent as possible with the available tissue-specific data.

The method begins by defining a core set of reactions with help from the tissue-specific data. Reactions in this core may have a high probability of actually happening in the target tissue network (C_H) or a moderate probability (C_M). Once these sets are defined, the goal is to find a consistent subnetwork of the initial generic metabolic model that includes all the high-probability reactions (C_H), as many moderate probability reactions (C_M) as possible, and the minimal number of extra reactions required to allow non-zero flux through all the reactions in the final model. A parameter in the objective function weights the parsimoniousness of the model against the inclusion of moderate-probability reactions.

The problem is solved using a greedy heuristic based on iteratively pruning reactions from the generic model. The prunning is done in a random order and consistency is checked after every step. A reaction is only removed if its elimination does not block any high-probability reaction and the model's score is increased. The algorithm is executed several times with different random pruning orders. Finally, all the models are aggregated to obtain the final model. Specifically, the fraction of models in which each reaction appears gives an indication on the confidence with which that reaction should be included in the final model. Starting from C_H, reactions are added iteratively, based on their confidence scores, until a minimal but consistent model is obtained.

2.2.3. GIMME

Becker and Palsson (2008) presented the Gene Inactivity Moderated by Metabolism and Expression (GIMME) method to produce context-specific reconstructions that are most consistent with the experimental data at hand. In addition, one or more Required Metabolic Functionalities (RMF) that the cell is assumed to be able to perform are requested to the contextualized network. Gene expression data is used to guide the reconstruction process.

The method has two steps. First, the maximum flux through all the RMF of interest is computed allowing the use of all the reactions in the network. Second, the set of reactions that best fits the experimental data is obtained through an optimization model, while the RMFs are constrained to operate at above a percentage of the maximum found in the first step.

$$\min \sum c_i \cdot |v_i| \tag{2.9}$$

subject to

$$S \cdot v = 0 \tag{2.10}$$

$$v_{min} \leq v \leq v_{max} \tag{2.11}$$

$$c_i = \begin{cases} x_{\text{cutoff}} - x_i, & \text{if } x_{\text{cutoff}} \geq x_i \\ 0, & \text{otherwise} \end{cases} \qquad \forall i \tag{2.12}$$

In this model, x_i is the normalized expression mapped to each reaction, and x_{cutoff} is a value selected by the user above which a reaction is considered as present. For those reactions with no expression available, the conservative approach of considering them as being above the threshold is taken.

The result of GIMME is a list of reactions predicted to be active (those that surpass x_{cutoff} plus the ones activated by the optimization model) and an inconsistency score (IS) reflected in the objective function value (coming from reactions that are below the cutoff value but are necessary for the constraints of the optimization model). The strategy of trying to leave reactions out of the final model that the optimization model performs is the reason why the method is called "gene inactivation".

2.2.4. INIT

The Integrative Network Inference for Tissues (INIT) algorithm was developed by Agren et al. (2012) in order to obtain cell-type specific models using protein abundance data available at the Human Protein Atlas database (Uhlen et al., 2010). The method can also leverage gene expression data or metabolomic data, e.g. from the Human Metabolome Database (HMDB) (Wishart et al., 2007).

$$\max \left(\sum_{i \in R} w_i y_i + \sum_{j \in M} x_j\right) \tag{2.13}$$

subject to

$$S \cdot v = b \tag{2.14}$$

$$|v_i| \leq 1000 y_i \tag{2.15}$$

$$|v_i| + 1000(1 - y_i) \geq \epsilon \tag{2.16}$$

$$v_i \geq 0, \qquad i \in Irr \tag{2.17}$$

$$b_j \leq 1000 x_j \tag{2.18}$$

$$b_j + 1000(1 - x_j) \geq \epsilon \tag{2.19}$$

$$b_j \geq 0 \tag{2.20}$$

$$x_j = 1, \qquad j \in present \tag{2.21}$$

$$y_i, x_j \in \{0, 1\} \tag{2.22}$$

An interesting feature of this method is that, instead of imposing the steady-state condition for all the internal metabolites, it allows a small net positive accumulation (Eq. 2.20). If the metabolite has been experimentally measured, a positive net production is imposed (Eq. 2.19 and Eq. 2.21), otherwise, the model tries to maximize the amount of metabolites that the network can produce taking into account the evidence available for each reaction (Eq. 2.13). At the same time, the inclusion of each reaction (controlled by binary variables y_i) is weighted (with weights w_i) according to the experimental evidence. When this evidence is against the presence of a reaction, its weight is negative. It is also negative for reactions for which no evidence is available in order to obtain a model that is as parsiomonious as possible.

2.2.5. MADE

Metabolic Adjustment by Differential Expression (MADE) (Jensen and Papin, 2011) seeks to obtain a functional metabolic network that reflects metabolic adjustments between two conditions. It formulates a MILP model to decide on a functional network that reflects the most statistically significant changes in the measured gene expression data. There is one set of independent constraints for each experimental condition ($i = 1, ..., n$).

$$\begin{aligned} & \max \sum_{i=1}^{n-1} f_{i \to i+1}(x) && (2.23) \\ & \text{subject to} \\ & \qquad S \cdot v = b && (2.24) \\ & \qquad lb \leq v \leq ub && (2.25) \\ & \qquad v_{obj} \geq v_{min} && (2.26) \\ & \qquad N \begin{pmatrix} v \\ x \end{pmatrix} = b && (2.27) \end{aligned}$$

The last constraint (Eq. (2.27)) is the representation of the GPR rules (following the SR-FBA formalism introduced in Shlomi et al. (2007)) and their coupling to metabolic fluxes. GPR rules are translated into linear constraints and they are used to allow or block the flux through each reaction. Equation (2.27) is in a very generic form, and details of how it is implemented in detail can be found in Shlomi et al. (2007). Hence, while other methods deal with the GPR rules and the mapping of gene expression values onto reaction values before the optimization step, MADE directly integrates those into the problem formulation. The objective function the problem maximizes is:

$$\begin{aligned} f_{i \to i+1}(x) = & \sum_{x \in I} w(p_{x_{i \to i+1}})(x_{i+1} - x_i) \\ & + \sum_{x \in D} w(p_{x_{i \to i+1}})(x_i - x_{i+1}) \\ & - \sum_{x \in C} w(p_{x_{i \to i+1}}) \Delta_{x_i, x_{i+1}} \end{aligned} \qquad (2.28)$$

where $\Delta_{x_i,x_{i+1}}$ is a binary variable that takes the value 0 when $x_i = x_{i+1}$ and 1 otherwise, and x_i are binary variables related to each gene representing their expression state. The sets I, D and C partition the x variables into increasing, decreasing and constant expression between two conditions, respectively. The weighting function $w(p_{x_{i \to i+1}})$ assigns larger weights to more significant p-values (e.g. $w(p) = -\log p$). The objective function thus measures the discrepancy between the selected changes and the measured expression changes.

2.2.6. FASTCORE

Recently, a fast algorithm for context-specific network reconstruction was presented in Vlassis et al. (2014) under the name of FASTCORE. It takes as input a core set of reactions that should be present in the desired network and searches a subnetwork of the initial network that contains all those reactions defined in the core set and a minimal number of additional reactions so that all the reactions can carry flux in the steady-state. In this aspect, FASTCORE is conceptually similar to MBA.

FASTCORE builds the context-specific networks by iteratively adding a set of *sparse* flux modes of the global network. At every iteration, two linear programs are solved, one for finding a flux distribution with maximal support on the reactions part of the core set (J) and another one to minimize the number of reactions that do not belong to the core set (P) in the obtained flux mode.

As the algorithm aims to include all the reactions that are part of the core set, the reference network must be able to carry flux through all of its reactions. For this, the authors developed a fast network consistency checking algorithm (FastCC).

$$\max_{v,z} \sum_{i \in J} z_i \tag{2.29}$$

subject to:

$$S \cdot v = 0 \tag{2.30}$$

$$v_i^{lb} \leq v_i \leq v_i^{ub} \qquad \forall i \tag{2.31}$$

$$v_i \geq z_i \qquad \forall i \in J \tag{2.32}$$

$$z_i \in [0, \epsilon] \tag{2.33}$$

This LP (Eqs. 2.29 - 2.33) favours flux “splitting” over flux “concentrating” to maximize the number of reactions belonging to the core set participating in the solution. It is enough to solve this problem once in order to know the maximum number of irreversible reactions in J that are able to carry flux while fulfilling the steady-state condition. For reversible reactions, however, it is necessary to take an iterative approach, considering them in one direction first and then in the opposite. This can be achieved by changing the sign of the columns in the stoichiometric matrix that correspond to these reactions. This LP (Eqs. 2.29 - 2.33) is also the first of the two LPs solved by FASTCORE at each iteration.

$$\min_{v,z} \sum_{i \in P} z_i \tag{2.34}$$

subject to:

$$S_N \cdot v = 0 \tag{2.35}$$

$$v_i^{lb} \leq v_i \leq v_i^{ub} \qquad \forall i \tag{2.36}$$

$$v_i \geq \epsilon \qquad \forall i \in K \tag{2.37}$$

$$v_i \in [-z_i, z_i] \qquad \forall i \in P \tag{2.38}$$

$$z_i \geq 0 \qquad \forall i \tag{2.39}$$

This second LP (Eqs. 2.34 - 2.39) minimizes the L_1 norm of the fluxes in a penalty set P (formed by reactions not included in the core), subject to a minimum flux through reactions in the set K. This set K is formed by reactions of the core activated in a single run of the FastCC LP. This is the second LP solved by FASTCORE at each iteration.

The FASTCORE algorithm deals first with the irreversible reactions in the core and then focuses in the reversible ones. Eventually, it goes over some remaining reversible reactions one by one. The speed of this algorithm is remarkable, obtaining reconstructions in a matter of seconds.

2.2.7. The Model SEED

The Model SEED (Henry et al., 2010) is an addition to the SEED framework (Overbeek et al., 2014) that provides genome-scale metabolic network reconstruction capabilities. More than a reconstruction model, it is a reconstruction pipeline that covers the process from the annotation of the genome to the generation of a functioning metabolic model. The steps covered are the following: (i) annotation; (ii) preliminary reconstruction; (iii) auto-completion; (iv) FBA analysis; (v) Biolog consistency analysis; (vi) gene essentiality consistency analysis; and (vii) reaction network optimization (Henry et al., 2010).

The Model SEED is geared towards the reconstruction of bacterial metabolic networks starting from their assembled genome sequences. Depending on the data available, it may be difficult to complete all the steps. Here, we want to focus on the preliminary reconstruction and auto-completion steps, which follow a similar strategy to other reconstruction algorithms already explored in this chapter.

The preliminary reconstruction consists in assembling a preliminary metabolic model composed of spontaneus reactions, transport reactions and enzymatic reactions, according to the annotated genome information. A set of GPR rules linking genes to reactions is also included, as well as a draft biomass reaction.

In the auto-completion step, a minimal set of reactions from the database is identified to be added to the preliminary model, so that the resulting network is capable to produce biomass under the minimal medium growth conditions known for the organism being modeled. The Model SEED uses a database of 12,000 reactions and 15,000 compounds to select the reactions from.

$$\min \sum_{i=0}^{R} (1 + P_{T,i} + P_{K_i} + P_{SS,i} + P_{F,i} - f_{SS,i} - f_{p,i}) z_i \tag{2.40}$$

subject to:

$$S \cdot v = 0 \tag{2.41}$$

$$0 \leq v_i \leq 1000 z_i \qquad \forall i = 1, ..., r \tag{2.42}$$

$$v_{bio} \geq 10^{-3} \text{ g/gCDWh} \tag{2.43}$$

The greatest novelty of this approach is its objective function (Eq. 2.40). The inclusion of any reaction not yet in the model (linked to binary variables z_i) is weighted according to different criteria in order to complete the network with the most likely set of reactions. $P_{T,i}$ is a penalty on the addition of transport reactions. $P_{K,i}$ is a penalty to favour the addition of reactions listed in KEGG. $P_{SS,i}$ favours the addition of reactions mapped to SEED functional roles and subsystems. $P_{f,i}$ penalizes the addition of reactions in a thermodynamically unfavourable direction. $f_{SS,i}$ is a bonus term that favours the addition of reactions involved in subsystems already well represented in the preliminary model. Finally, $f_{p,i}$ is a bonus applied to reactions involved in short linear pathways already well represented in the preliminary model. Thus, the optimization is designed to use known biological components (e.g. pathways, subsystems or functional roles) to guide the reconstruction process.

2.2.8. Other algorithms

So far we have briefly seen some of the network reconstruction algorithms available, but many more exist. Some are modifications of the already mentioned ones, while others integrate more types of data and expand beyond metabolism to also include signaling or regulatory networks or even take a different approach to the integration of the data.

For instance, GIM^{3}E (Schmidt et al., 2013) is an evolution of GIMME (Becker and Palsson, 2008) to also make use of metabolomics data and guarantee that the resulting network uses the detected compounds. The name stands for Gene Inactivation Moderated by Metabolism, Metabolomics, and Expression. MIRAGE (Vitkin and Shlomi, 2012) is similar to MBA (Jerby et al., 2010), but it also looks for the production of biomass components and for the growth-associated dilution of all network metabolites (guaranteeing that all the metabolites in the network can be obtained through transport reactions or synthesis pathways). On top of that, it assigns a continuous score to each reaction based on the input data instead of just classifying them in some predefined sets. These scores aim to reflect the probability of retaining each reaction in the final model and allow for a better use of the data.

Taking another approach to the data integration problem, PROM (Chandrasekaran and Price, 2010) and E-Flux (Colijn et al., 2009) adjust the maximum allowed flux through each reaction using gene expression data. In particular, PROM integrates metabolism with regulatory networks, requiring a large gene expression dataset with genetic and environmental perturbations. Another recently presented algorithm, PRIME (Personalized Reconstruction of Metabolic models) (Yizhak et al., 2014a), on the other hand, explores the idea of modifying reaction bounds using gene expression, but only for those reactions whose expression shows a significant correlation with the growth rate. This way, the algorithm tries to also leverage phenotypic data.

2.3. Fast Network Reconstruction

In this section, we present a multistep strategy based on a linear programming formulation to contextualize a metabolic network with gene expression data, which substantially reduces the computation time found in competing algorithms. This enables more demanding studies requiring a large number of reconstructions. In particular, this algorithm will be used in Chapter 4 to analyze the influence of expression data and random data in the contextualized networks that are used in gene essentiality analysis. The different steps of our reconstruction algorithm are detailed below.

2.3.1. Reaction classification

The first thing to do is to classify the reactions according to the experimental data. The input of the reconstruction algorithm is the reaction classification as highly (H), moderately or medium (M) and lowly (L) expressed. This information can be obtained from gene expression experiments, for example the ones available at the GEO database (Edgar et al., 2002).

In this thesis, we focused on Affymetrix HGU133plus2 arrays, which can be processed using Barcode (McCall et al., 2011). This method is designed to be able to work with just one sample and make it comparable to others, instead of needing several samples at the same time. We preprocessed the data using Barcode's R script, using one sample at a time. We retrieved the z-score values obtained from this algorithm, which is equivalent to processing each sample with fRMA (McCall et al., 2010).

Using gene-probe relationships annotated in hgu133plus2.db R package, the gene value was obtained as the median value of the corresponding probe sets. Each gene value was transformed into present (1) and absent (0) calls using Barcode's criteria. Then, present genes were classified as high (+1) and absent genes as low (-1).

Finally, reactions are classified as highly (H), medium (M) or lowly (L) expressed using gene-protein-reaction rules and the gene expression classification mentioned above (Rossell et al., 2013). Those reactions for which no gene expression is available or that are not related to any gene (e.g. spontaneous reactions) are classified as medium expressed.

2.3.2. Basic Network

Consider a general metabolic network with C compounds and R reactions represented by its stoichiometric matrix S (Palsson, 2006). We denote Irr the set of irreversible reactions. For convenience, each reversible reaction contributes two different irreversible reactions to the total number R. These two irreversible reactions are denoted f and b, forward and backward, respectively, each of which represents the original reversible reaction in one different direction (de Figueiredo et al., 2009). The set of forward and backward steps that arise from reversible reactions are denoted Rev.

The flux through each reaction i ($i = 1, ..., R$) is represented by a continuous variable v_i. After the split of reversible reactions, fluxes can only take non-negative values, bounded by a maximum flux value, v_i^{max} (Eq. 2.44). To later apply FBA-based GEA, we also enforce the steady state condition (Eq. 2.45) and a minimum flux $v_{biomass}^*$ through the biomass reaction (Eq. 2.46). For those compounds taken from or excreted to the medium, exchange reactions were added appropriately.

$$0 \leq v_i \leq v_i^{max} \qquad i = 1, ..., R \tag{2.44}$$

Table 2.1: Weighting schemas used for the reconstruction algorithm ($\alpha = 1000$).

Schema	W^H	W^M	W^L
1	α	1	α^2
2	α	1	α
3	α^2	1	α

$$\sum_{i=1}^{R} S_{ci} v_i = 0 \qquad \forall c \in C \tag{2.45}$$

$$v_{biomass} \geq v^*_{biomass} \tag{2.46}$$

To properly define v_i^{max} for each reaction, we perform a Flux Variability Analysis (FVA) (Gudmundsson and Thiele, 2010) under constraints in Eqs. (2.44)-(2.45). Uptake reaction bounds from the growth-medium under consideration are included in Eq. (2.44).

We also define a continuous variable z_i for each reaction, bounded between 0 and 1 (Eq. 2.47), which may force a minimum flux through its associated reaction, v_i (Eq. 2.48). δ is a constant between 0 and 1 that fixes the lower bound on v_i in relation with the value of z_i with respect to v_i^{max}. The inclusion of v_i^{max} in Eq. (2.48) as calculated by FVA allows us to set an activation threshold independent of the stoichiometric representation. We remark that this set of variables is continuous, as in Vlassis et al. (2014), and not binary, as in a number of previous works (Jerby et al., 2010; Shlomi et al., 2008).

$$0 \leq z_i \leq 1 \qquad i = 1, ..., R \tag{2.47}$$

$$\delta v_i^{max} z_i \leq v_i \qquad i = 1, ..., R \tag{2.48}$$

Our objective is to minimize the number of reactions in L while maximizing those in H. For that, our objective function minimizes the sum of fluxes through reactions belonging to L with a weight W^L, as well as the flux through reactions in M with a weight W^M, while maximizing the number of reactions in H using z variables with a weight W^H (Eq. 2.49). The term δv_i^{max} in Eq. (2.49) allows us to avoid the flux bias introduced by the specific stoichiometric representation of reactions. A proposal for the weights in the model is presented in Table 2.1 and discussed in Section 2.4.

$$\min W^L \sum_{i=1|i\in L}^{R} \frac{v_i}{\delta v_i^{max}} + W^M \sum_{i=1|i\in M}^{R} \frac{v_i}{\delta v_i^{max}} - W^H \sum_{i=1|i\in H}^{R} z_i \tag{2.49}$$

As noted above, it is common to set z_i as a binary variable, but relaxing that constraint, as done here, achieves the same "flux diversification" effect desired (Vlassis et al., 2014). Minimizing the sum of fluxes for L and M is not the same as minimizing the number of reactions in L and M, but it allows us a linear formulation of the problem without negatively influencing the final solution in terms of quality. Overall, with these features, we avoid a mixed binary formulation, harder to solve because of the integrality constraints on some of the variables (Vanderbei, 1996).

Since we have split the reversible reactions into two irreversible steps, but have added no constraint guaranteeing that only one of them is active at a time, solving this problem (Eq. 2.49 subject to Eqs. 2.44-2.48) will give us a solution where all forward and backward steps from reversible reactions in H are active, even if their net flux $(v_f - v_b)$ is zero. Note that this does not occur with reversible reactions in L or M, because minimizing the sum of fluxes already enforces the usage of reversible reactions, if necessary, only in one direction. For this reason, we need an iterative procedure that disentangles whether these reversible reactions in H can certainly be included in the reconstructed network.

On the other hand, the solution resulting from this step directly provides us with the subset of irreversible reactions from H that will be involved in our final reconstruction. For this reason, the flux of irreversible reactions from H that have not been activated in this first step is set to zero for the rest of the iterative process (Eq. 2.50).

$$v_i = 0 \qquad \forall i \mid i \in H, i \in Irr, i \notin D \tag{2.50}$$

Overall, this first step provides a first draft network D that will be expanded in the next steps. Reversible reactions in H with net flux equal to zero cannot be directly included in D and require further analysis to evaluate their presence in the final reconstruction.

2.3.3. Iterative Refinement

The aim of this iterative process is to determine which reversible reaction in H will be part of the final reconstructed network, in particular those with net flux equal to zero in the previous step. During the iterative process, we will gradually increment the number of reactions included in our draft network D. In each iteration, we will set the penalty W^L and W^M of those reactions already included in the solution in previous iterations to zero, as once a reaction is included in the draft, there is no need to penalize it further. Similarly, we will set the W^H bonus of reversible reactions in H already included in the draft from previous iterations to zero. Note that the W^H bonus of irreversible reactions in H is kept to guide the addition of reactions in D. These variations lead to a new objective function, which is represented by Eq. (2.51).

$$\begin{aligned} \min W^L \sum_{i=1|i\in L, i\notin D}^{R} \frac{v_i}{\delta v_i^{max}} + W^M \sum_{i=1|i\in M, i\notin D}^{R} \frac{v_i}{\delta v_i^{max}} \\ -W^H \sum_{i=1|i\in H, i\in Irr}^{R} z_i - W^H \sum_{i=1|i\in H, i\in Rev, i\notin D}^{R} z_i \end{aligned} \tag{2.51}$$

In order to evaluate whether a reversible reaction from H, currently not in D, must be added into the reconstruction, we need to solve the linear program defined by Eq. (2.51) subject to Eqs. (2.44)-(2.48) and (2.50) in two different scenarios: one with flux equal zero in the forward direction, $v_f = 0, f \in H, f \in Rev, f \notin D$, and the other with flux equal zero in the backward direction, $v_b = 0, b \in H, b \in Rev, b \notin D$. If $v_b > 0$ in the first scenario and/or $v_f > 0$ in the second scenario, then this reversible reaction takes part in the final reconstruction, as well as additional reactions from the sets H, M and L required to perform in steady state. We may also need to add other reversible reactions from H currently not in D and, therefore, their analysis will not be further required. In case that $v_b = 0$ in the first scenario and $v_f = 0$ in the second scenario, this reaction is discarded from the final reconstruction. We will refer to this process as Iteration A.

The strategy described above, though general, may require a large number of linear programs, as we need to individually check each reversible reaction. In order to reduce computation time, we introduce an intermediate algorithmic step, based on the concept of reduced cost from linear programming, which allows us to minimize the number of linear programs to be solved. Full details are provided below.

2.3.4. Efficient Implementation

If we knew in which direction was going to work each reversible reaction in a possible solution, we could block the reactions in the opposite direction and solve the previous problem to recover that solution. However, as this is not the case, we will make a guess and then use linear programming theory to further improve it.

We will first solve the linear program defined by Eq. (2.51) subject to Eqs. (2.44)-(2.48) and (2.50) in two different scenarios. In particular, for all the reversible reactions in H not included in D, we will set their fluxes to zero in one direction first ($v_f = 0 \quad \forall f \in H, f \in Rev, i \notin D$) and later in the other ($v_b = 0 \quad \forall b \in H, b \in Rev, b \notin D$), similarly to what is done in the FastCC algorithm in Vlassis et al. (2014). In addition, we will relax the bounds in the other direction, as observed in Eq. (2.52). The solution to this linear program may provide new reactions to the draft network D, but the solution might have been better if we had selected some reversible reactions in the opposite direction.

$$v_f = 0, -v_f^{max} \leq v_b \leq v_b^{max} \qquad \forall (f,b) \in H, \forall (f,b) \in Rev, \forall (f,b) \notin D \quad (2.52)$$

Here, we can make use of linear programming theory to improve upon our solution and eventually reduce the number of reactions that will need to be checked individually. Specifically, we will make use of the concept of reduced cost of a variable. This value indicates how much the objective value would theoretically change if we modify the value of a variable by one unit.

It is important to clarify that, for the determination of reduced costs, it was assumed that the proposed linear problem is solved handling the bounds on variables implicitly. In fact, most available linear programming solvers implicitly handle variable bounds, meaning that those bound constraints are not explicitly added to the constraint matrix. Under these circumstances, the non-basic variables are no longer necessarily zero and their reduced cost can take any real value.

For readers unfamiliar with linear programming, variables are classified as basic or non-basic. Non-basic variables are independent variables and their value is set equal to their upper or their lower bound. By constrast, basic variables are dependent variables and their value is obtained by solving the corresponding system of equations (Vanderbei, 1996).

In the optimal solution, reduced cost of basic variables is zero, while it is usually nonzero for non-basic variables, unless the problem has alternative solutions, where non-basic variables may have a reduced cost of zero. For a minimization problem, the reduced cost of non-basic variables at their lower bound will be positive, and for non-basic variables at their upper bound, it will be negative. These reduced costs can also be interpreted as the shadow prices of the lower and upper bound constraints, respectively, of these variables.

With the solution to our modified problem in our hand, we set the focus on the reduced costs of z_b variables associated to v_b variables, for which we have relaxed the lower bounds. If their reduced cost is positive, a decrease in their value implies a reduction in the objective function value. In this case, if we allow a small negative value for z_b, the corresponding v_b would be able to take a negative value (see Eq. 2.52). This may be sufficient to activate another reaction from H, allowing another z variable to take a positive value and, thus, improving the objective function value. If their reduced cost is zero and they are non-basic variables, we have alternative optimal solutions, implying that a small change in the lower bound of that variable could lead to a different optimal solution and, therefore, we also need to look at these variables.

These backward reactions b (with $b \in H, b \in Rev, b \notin D$) that are non-basic and have a non-negative reduced cost are stored in the set J. Then, backward reactions in J are fixed to zero, and a positive flux for their associated forward reactions is enabled, as shown in Eqs. (2.53) and (2.54).

$$v_f = 0, 0 \leq v_b \leq v_b^{max} \qquad \forall (f,b) \in H, \forall (f,b) \in Rev, \forall (f,b) \notin D, b \notin J \quad (2.53)$$

$$v_b = 0, 0 \leq v_f \leq v_f^{max} \qquad \forall (f,b) \in H, \forall (f,b) \in Rev, \forall (f,b) \notin D, b \in J \quad (2.54)$$

Therefore, we solve the following optimization problem: Eq. (2.51) subject to Eqs. (2.44)-(2.48), (2.50), (2.53) and (2.54). We repeat this process but starting with the reactions in the opposite direction, this is, switching f and b in Eqs. (2.52), (2.53) and (2.54). The whole procedure is repeated until no new reaction from H is added to the network. We refer to this procedure as Iteration B.

Once Iteration B has ended, we may have reactions in J not included in D. However, some of them could possibly be included in the reconstruction. The reason for not having them included during Iteration B is that we should have reversed only a subset of them. Thus, the final step is to apply the procedure described in Iteration A for those reactions that remain included in J but not in D.

2.4. Performance evaluation

The approach presented in Section 2.3 is used here to reconstruct 174 metabolic networks corresponding to cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012). The technical performance of our approach is evaluated and compared with that of iMAT, the most similar approach to the one introduced here (details of the implementation of iMAT used can be found in Appendix A).

To carry out this analysis, we used the human metabolic network Recon 2 (Thiele et al., 2013) as reference network. We also carried out a similar analysis for Recon 1 (Duarte et al., 2007). Both networks provide a biomass reaction, which is directly used in this study. The growth medium was RPMI1640, defined as in (Folger et al., 2011). In addition, reactions were classified as highly (H), medium (M) or lowly (L) expressed using gene-protein-reaction rules and the gene expression classification as described in Subsection 2.3.1.

We implemented the fast network reconstruction algorithm in Matlab, using Cplex optimization software as the underlying software in charge of solving the corresponding linear programs. The computation time needed to solve a single reconstruction problem is in the order of seconds, in par with the performance of FASTCORE (Vlassis et al., 2014). On the instances our method was applied, computation time is generally below 10 seconds on a 64 bit Intel Xeon E5-1620 v2 at 3.70 GHz (4 cores) and 16 GB of RAM (Figure 2.1, Figure 2.2 and Table 2.2). This sets our algorithm as substantially faster than iMAT, where the median time to obtain a solution was around 57 seconds (stopping with a 0.5 % optimality gap).

In our reconstruction algorithm we have several parameters that require being fixed. The most relevant parameters are the weights W^H, W^M and W^L, as there is a conflicting trade-off between reactions in H and L. In particular, the use of all reactions in H may involve a significant number of reactions in L; similarly, a minimum use of reactions in L may imply a limited use of reactions in H. In order to study this trade-off between reaction in H and L, we propose the schemas in Table 2.1, with $\alpha = 10^3$. Schema 1 gives more weight to the minimization of reactions in L over the maximization of reactions in H; Schema 2 provides equal weight, while Schema 3 is the opposite of Schema 1.

When classifying reactions from gene expression data, avoiding the inclusion of reactions in L as much as possible might be more meaningful than trying to force the presence of all reactions in H, as a high gene expression signal does not necessarily translate into a high enzymatic activity. However, the identification of non-expressed genes constitutes a more difficult

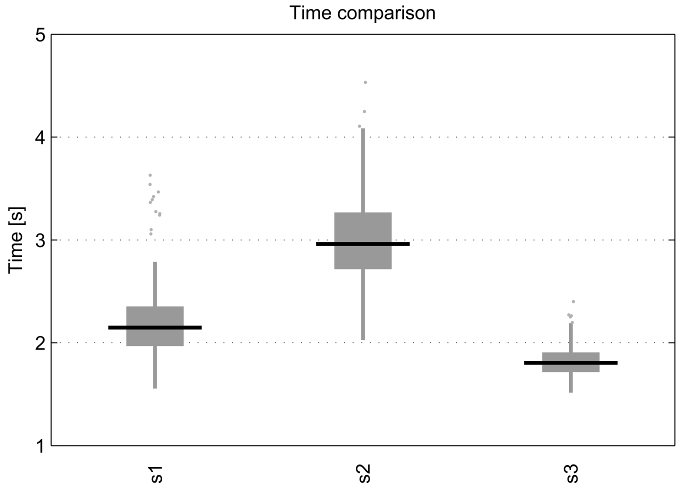


Figure 2.1: Boxplot showing the computation times for reconstructed context-specific networks of selected cancer cell lines using our algorithm under schema 1, 2 and 3, with Recon 2 as the reference metabolic network.

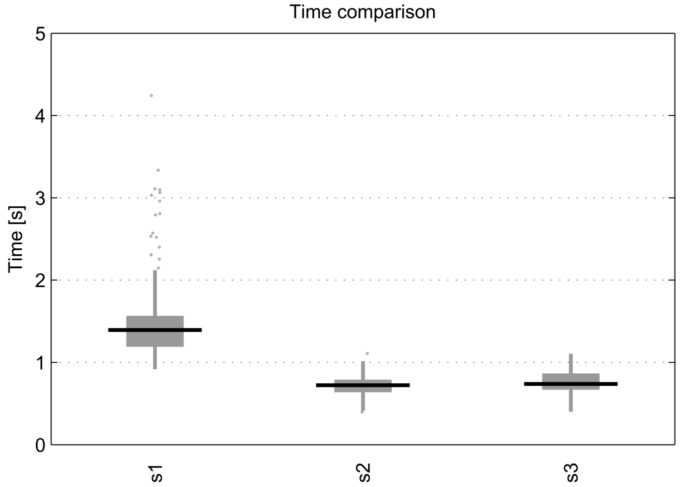


Figure 2.2: Boxplot showing the computation times for reconstructed context-specific networks of selected cancer cell lines using our algorithm under schema 1, 2 and 3, with Recon 1 as the reference metabolic network.

Table 2.2: Average computation time, percentage of included H reactions and percentage of included L reactions under different parameter settings.

Setup	Schema	Time [s]	H %	L %
$\alpha = 10^2$	1	2.14	34.73	1.84
	2	2.99	56.17	8.36
	3	1.8	69.04	20.47
$\delta = 0{,}01$	1	2.57	41.68	1.65
	2	3.51	57.54	8
	3	2.91	70.11	18.53
$v^*_{biomass} = 0{,}10 \cdot v^{max}_{biomass}$	1	2.26	34.63	1.65
	2	3.06	56.22	8.36
	3	1.8	69.06	21.2
General growth medium	1	2.22	41.3	0.42
	2	3.16	61.55	7.98
	3	2.47	75.2	23.7

task (Åkesson et al., 2004). For this reason, an approach closer to Schema 3 has been typically preferred. Notwithstanding, if we had data where non-expressed reactions were identified with high reliability, Schema 1 would be the preferred option.

We compared the performance of our reconstruction approach using the different schemas with iMAT. As can be seen in Figure 2.3, which shows the percentage of reactions classified as H and L that were included using each reconstruction algorithm, the avoidance of L reactions in Schema 1 has an impact on the number of reactions in H included in the model, providing a significantly different solution than Schema 3. We also experimented with other parameters for our algorithm (Figure 2.3, Figure 2.4 and Table 2.2), observing that the conclusions achieved were robust to changes of these parameters.

As expected, Schema 2 is the most similar to iMAT, as both provide equal weight to reactions in H and L. It can be observed that the number of L reactions included is very similar and the number of H reactions included by our algorithm is somewhat lower (Figure 2.5). Overall, both methods obtain similar reconstructions in terms of the number of H and L reactions they include. Thus, we consider our algorithm a valid tool for the task at hand. Note that the maximum possible percentage of H reactions included in the reconstruction does not necessarily reach 100 % as there might be reactions that cannot operate in steady state under the imposed medium conditions.

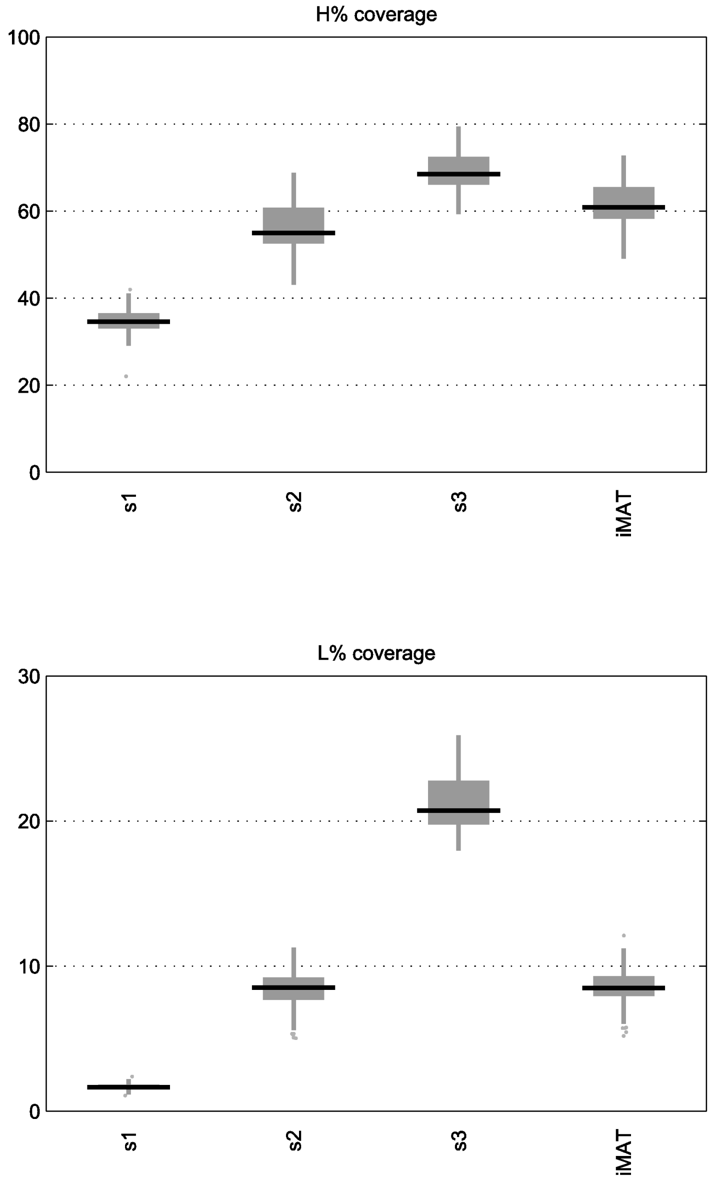


Figure 2.3: Boxplots showing the percentage of H and L reactions included in the reconstructed context-specific networks of selected cancer cell lines using our algorithm under schema 1, 2 and 3 and iMAT, with Recon 2 as the reference metabolic network.

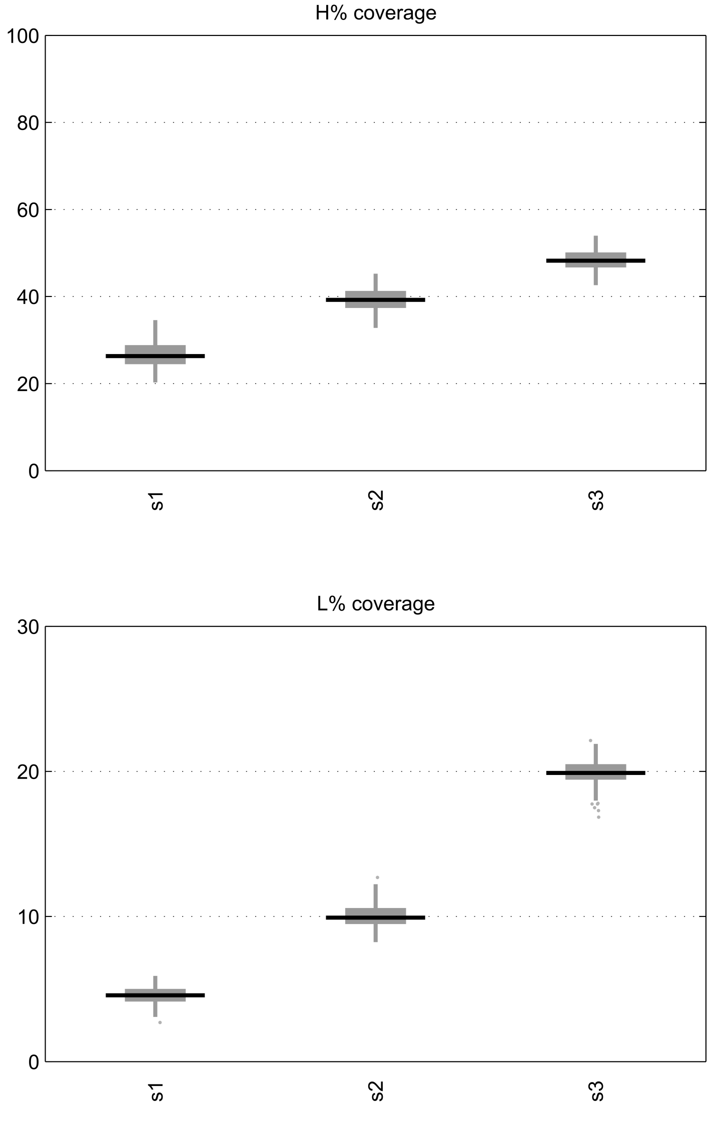


Figure 2.4: Boxplots showing the percentage of H and L reactions included in the reconstructed context-specific networks of selected cancer cell lines using our algorithm under schema 1, 2 and 3, with Recon 1 as the reference metabolic network.

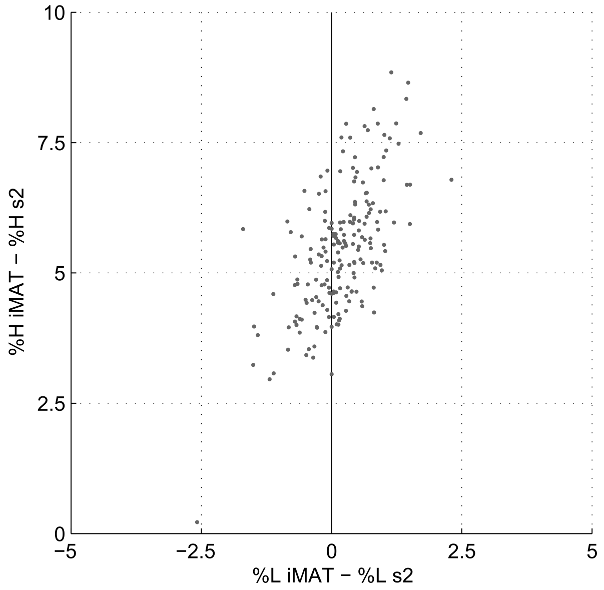


Figure 2.5: Comparison between the percentage of L and H reactions included with iMAT or with our reconstruction algorithm with schema 2 in networks reconstructed starting from Recon 2. The figure shows the difference between the percentage of L reactions included with iMAT and the percentage included with our algorithm versus the difference between the percentage of H reactions included with iMAT and the percentage included with our algorithm. It can be observed that both include a similar percentage of L reactions and iMAT includes a slightly higher percentage of H reactions.

2.5. Conclusion

In this chapter we have given an overview of different metabolic network reconstruction and contextualization algorithms, and we have introduced a novel fast reconstruction method. All these methods have a similar conceptual approximation to the problem, but each one of them includes its own subtleties given by the requirements of their application. Our new method has been designed with speed and FBA compatibility in mind, and it takes advantage of linear programming theory to design a fast iterative approach to network contextualization.

The development of our Fast Network Reconstruction algorithm was motivated by the study done in Chapter 4. To achieve the objective of that study, the need of a fast reconstruction algorithm was a must. The Fast Network Reconstruction algorithm proposed in Section 2.3 builds on the reasoning present in previous approaches (Folger et al., 2011; Jerby et al., 2010; Vlassis et al., 2014). Our approach is conceptually similar to iMAT (Shlomi et al., 2008), but, in line with FastCore (Vlassis et al., 2014), it provides us with network reconstructions in much less time, thanks to the iterative LP strategy used. Figure 2.5 shows that this advantage in computation time does not compromise the quality of obtained contextualized metabolic networks.

In Section 2.4 we have evaluated the behaviour of our algorithm. We observe a clear role of the schema (parameters W^H, W^M and W^L) in the reconstruction, which illustrates the need to decide over the treatment of highly and lowly expressed reactions. Systematic procedures as Barcode (McCall et al., 2011), which was used in this study to process gene expression data (see Chapter 4), are an essential part in this debate. Barcode, for example, is very restrictive for expressed genes and it is more likely to have false negatives. If that is the case, Schema 3 could be the preferred option, as it biases the reconstruction towards including reactions classified as H. However, if we had data that gave us a high confidence on the reactions that should be absent in the final reconstruction (RNA-Seq data could be an example), Schema 1 would be the preferred option, as it makes more difficult to include reactions classified as L in the final network.

Finally, this Fast Network Reconstruction algorithm will allow us to reconstruct in Chapter 4 a high number of networks in order to assess the frequency with which genes appear as essential in randomly generated data.

Chapter 3

Bacterial Community Metabolic Network Reconstruction

In this chapter, we present a novel computational procedure for determining a functional, context-specific metabolic network for bacterial communities using metaproteomic and taxonomic data. The method described herein was motivated by the need to rapidly gain insights on two different bacterial communities obtained from contaminated soil (Tobalina et al., 2015). Besides, recent interest in the study of gut microbiota and its relation with human health also motivates the development of reconstruction methods tailored towards bacterial communities. This study allows us to show the potential of metabolic network reconstruction approaches to increase the insights that can be extracted from the generated data.

3.1. Introduction

Microbes are shaping the world and, by forming communities, are causal of geochemical cycles (Mascarelli, 2009), human health (Kinross et al., 2011) and biotechnological processes (Beloqui et al., 2008). Thus, it is not surprising to find increasing interest in studying how these consortia lead to function (Carter et al., 2012).

The analysis of microbial communities begins by assessing the structure of the population, which is currently often achieved using metagenomic data (Röling et al., 2010). The next step consists of characterizing the metabolic capacity of the microbial community, but this has proven to be a considerable challenge, even when using metatranscriptomic data (Moran et al., 2013). This need has led to the development of metaproteomics, by which at least the abundance of metabolically active molecules can be detected (Seifert et al., 2013). In parallel, methods for analysing these data have arisen and evolved.

From the outset, computational methods have been essential for capitalizing on data to obtain clear and novel insights (Guazzaroni and Ferrer, 2011). Traditional approaches described in the literature typically map data for genes, proteins or metabolites onto well-known pathways or Gene Ontology (GO) terms (Yamada et al., 2011). This enables identifying molecular functions of identified proteins in light of metabolic pathways. However, the high connectivity among biological pathways has shifted the focus to networks (Letunic et al., 2008; Palsson, 2009), which allows us to capture more global properties (McCloskey et al., 2013). Molecular networks integrate different pathways and constitute a more general framework for interpreting "omics" data (Bordbar and Palsson, 2012).

On the single-species level, different computational methods have been developed to analyse "omics" data using genome-scale metabolic networks. In particular, a number of approaches have been specifically designed to incorporate gene and protein expression data, as reviewed in Chapter 2. These approaches start from genome-scale metabolic networks, which are reconstructed from the annotated genome of an organism (Bachmann et al., 2013; Zomorrodi et al., 2012) and a reference metabolic database as input information.

For microbial communities, the reconstruction of metabolic networks is more complicated and faces new challenges. Ideally, each organism can be represented by its own metabolic network and its input/output metabolites define its possible interaction with other members of the consortia. Should this information be available, recently developed constraint-based modelling approaches could be applied. In this situation, methods mentioned above to incorporate "omics" data for single organisms could be extended to deal with bacterial communities.

However, in complex bacterial communities the number of organisms could be extremely high, most typically lacking a high-quality, genome-scale metabolic network, which makes the identification of shared components between organisms even more complicated. For this reason, current approaches have been applied to well-known microbial consortia, including only a limited number of organisms, typically 2 or 3 (dos Santos et al., 2013; Khandelwal et al., 2013; Zomorrodi and Maranas, 2012).

To overcome this issue, the use of a supraorganism or metanetwork has been proposed (Borenstein, 2012), which ignores boundaries for each organism, but models community-level metabolism. In an early work (Greenblum et al., 2012), using a graph theoretical approach, metagenomic data were used to reconstruct the human gut microbiome metanetwork in different conditions, finding key variations in patients with obesity and inflammatory bowel disease.

In this study, we exploit this metanetwork strategy and present a novel computational procedure for obtaining a context-specific metabolic network for a bacterial community using metaproteomic data. In contrast with the work presented in Greenblum et al. (2012), we did not use a graph-theoretical approach, but a constraint-based one, which takes into account stoichiometric relationships. In particular, our approach takes some ingredients from the mathematical optimization model presented in the Model SEED (Henry et al., 2010). However, our approach is fundamentally different: it is designed for bacterial communities, not for a single organism, and focuses on the usage of metaproteomic data, which directly leads to a contextualized network that gives cohesion to identified proteins. We also use the taxonomic assignment of the identified proteins to favour the inclusion of enzymes annotated in the genomes of those organisms in cases where such information is available.

3.2. Methods

Here, we present our computational procedure for determining a functional, context-specific metabolic network for bacterial communities using metaproteomic data. Based on a reference metabolic database (Henry et al., 2010), we seek a functional network that includes the maximum number of measured proteins (highly likely set, H) in a given sample. We may have evidence that some proteins are not expressed (lowly likely set, L) and, therefore, their participation is minimized. Then, we complete the network using enzymes in the database, preferably those annotated in active organisms in the community (medium likely set, M) (Guazzaroni et al., 2013).

We denote the set of enzymes from the reference database not included in H, L, and M as D, namely D involves the subset of non-identified enzymes that are currently annotated for organisms not present in the community. By linking proteins to reactions via Enzyme Commission (EC) numbers (Bairoch, 2000), sets H, L, M and D may also refer to reactions.

When we refer to a functional network, we mean a subset of reactions that are able to produce biomass at steady state under the specified medium conditions. We describe these conditions in detail and introduce the mathematical notation below.

We denote the sets of reactions and compounds in the reference metabolic database as R and C, respectively. The set of reactions is typically classified into reversible and irreversible reactions. For convenience, both reversible and irreversible reactions are divided into two non-negative steps: forward and backward reactions. We define the set $B = \{(f, b) \mid$ reaction f and reaction b are the reverse of each other, $f < b\}$. For each reaction $i \in R$ we define a flux variable, v_i, and a binary variable, z_i, where $z_i = 1$ if $v_i > 0$, 0 otherwise. We denote the stoichiometric coefficient associated with the metabolite $i \in C$ and reaction $j \in R$ as s_{ij}. This information is stored in the stoichiometric matrix, S.

The steady-state assumption implies mass balancing and, therefore, the accumulation/depletion of metabolites inside the system is not possible, as observed in Eq. (3.1). The definition of the boundaries of the system is an important issue. As noted above, in complex bacterial communities, the previous knowledge of shared input/output metabolites is typically not available. For the sake of simplicity, we only include boundaries for the whole community and remove boundaries between individual organisms. Therefore, we obtain a metanetwork in which the identified proteins from various organisms in the community are coexpressed. Using exchange reactions, we then define metabolites that can be taken up from outside (the boundaries of) the system (culture medium conditions) and those that can be excreted outside (the boundaries of) the system, which may prevent the network from utilizing unavailable nutrient sources.

$$Sv = 0 \tag{3.1}$$

As our aim is to obtain a metabolic network that supports growth, we must define a biomass reaction. Given that we are using a metanetwork strategy, the biomass reaction represents a consensus equation for all organisms in the community. Note that determining an appropriate biomass reaction is a challenging task, even for single organisms (Feist and Palsson, 2010). However, using an existing biomass reaction from a different organism is a common practice (Nogales et al., 2008), as many constitutive compounds are shared across a wide range of organisms. Equation (3.2) forces a minimum flux through the biomass reaction ($v_{biomass}$).

$$v_{biomass} \geq 1 \tag{3.2}$$

As fluxes are non-negative, their lower bound is 0, as observed in Eq. (3.3). We also fixed a sufficiently large value for their upper bounds.

$$0 \leq v_i \leq 1000 \qquad \forall i \in R \tag{3.3}$$

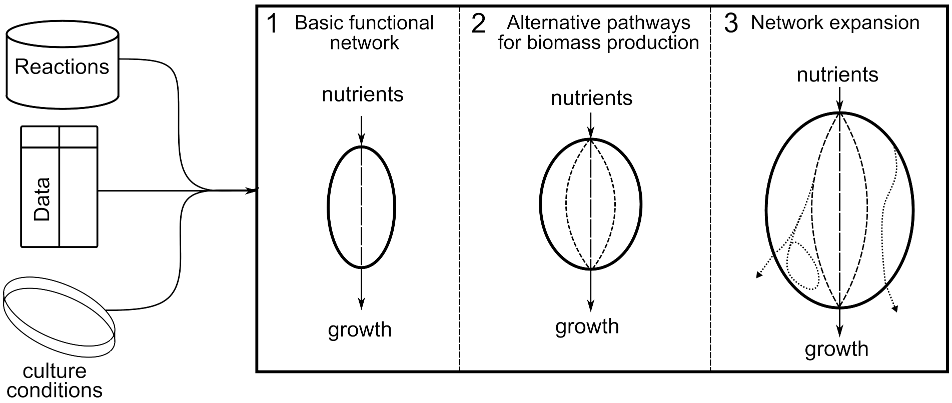


Figure 3.1: Proposed bacterial community metabolic network reconstruction workflow. The context-specific metabolic network reconstruction algorithm starts from a database of reactions, experimental (metaproteomic) data and knowledge about growth medium as input data. It involves 3 steps: 1) construction of a basic network capable of using the available nutrients to produce the compounds necessary for growth; 2) addition of alternative pathways for biomass production; 3) network expansion with pathways not necessarily involved in biomass production.

As noted above, our approach takes some elements of the mathematical optimization model presented in the Model SEED (Henry et al., 2010). However, our purpose is different, as we aim to obtain a context-specific reconstruction for bacterial communities, not for a single organism, based on metaproteomic data. Instead of a general metabolic reconstruction, our aim is to build a network as specific to the observed phenotype as possible, given the measured data. To this end, we include important technical differences. In particular, we use a 3-step iterative procedure based on linear programming and mixed integer linear programming. We describe each of these three steps below. A graphical summary of the complete workflow of our approach can be found in Figure 3.1.

3.2.1. Step 1: basic functional network

Due to regulatory effects, the experimental measurement of proteins is not sufficient to guarantee their activity (Seifert et al., 2013). This is typically observed in the conflicting trade-off between enzymes in the H and L sets (Agren et al., 2012; Becker and Palsson, 2008; Shlomi et al., 2008). In other words, the use of all enzymes in H may involve a considerable number of enzymes in L, whose participation in the reconstruction must be in general avoided (Åkesson et al., 2004). For this reason, we prefer to leave the decision of selecting enzymes in H and L to the optimization model, which incorporates evidence from metaproteomic data in the objective function.

This allows metaproteomic data to influence the resulting network, without directly constraining it.

The objective function is presented in Eq. (3.4). In particular, the flux activity is guided by its penalization, p_i, and bonus, b_i, terms, similar to what is performed in the Model SEED (Henry et al., 2010).

$$min \sum_{i \in R} (p_i - b_i) v_i \tag{3.4}$$

where p_i and b_i are the sums of various concepts. Weights are defined such that, by minimizing Eq. (3.4), which is subject to Eqs. (3.1)-(3.3), we obtain a functional metabolic network in which fluxes in H will prevail, followed by those in M, then those in D and, finally, those in L. As in the Model SEED (Henry et al., 2010), we penalized reversibility changes and favoured the completion of KEGG modules that were substantially covered with metaproteomic data. Finally, in order to avoid the flux bias introduced by the specific stoichiometric representation of each reaction (Brochado et al., 2012), weights were rescaled using maximum flux values obtained from flux variability analysis (Mahadevan and Schilling, 2003). See Appendix B for further details.

It should be noted that the introduction of continuous fluxes in Eq. (3.4), in contrast to the Model SEED (Henry et al., 2010), which includes binary variables (z), does not guarantee the optimal use of metaproteomic data, i.e. the reactions in H. The removal of binary variables converts a highly expensive, mixed integer linear program into a linear program, which can be easily solved; however, optimal solutions obtained from linear programming are extreme points, which, in conjunction with our minimization objective function, generate networks involving a limited number of degrees of freedom. Through Steps 2 and 3 described below, we aim to further exploit experimental evidence from protein expression data.

3.2.2. Step 2: alternative pathways for biomass production

Once Step 1 is solved, we obtain a list of active reactions, N_1. In this second step, we aim to capture alternative pathways for biomass production that are not included in N_1 using the reactions in H, but not in L. To this end, we block each of the reactions in N_1, one-by-one, and resolve the linear program posed in Step 1, i.e. Eq. (3.4), which is subject to Eqs. (3.1)-(3.3). As a result, we obtain a number, card(N_1), of functional networks. The rule here is to merge N_1 with those networks that include additional reactions in H, but not in L. Therefore, we obtain a functional network (N_2) that makes better use of the metaproteomic data for biomass production.

Note that the solution from Step 1 is the result of solving a linear program, whose optimal solution is an extreme point, a solution with zero degrees of freedom. For that reason, all the reactions involved in N_1 are necessary to produce biomass. As relevant alternative routes for biomass production are added into the model using Step 2, the number of degrees of freedom of the resulting solution is increased, while reducing the number of essential reactions/enzymes.

3.2.3. Step 3: network expansion

Once Step 2 is concluded, we may not have included all measured enzymes in N_2. We denote K as a particular set of these enzymes that we aim to include in the reconstruction. Note that K will typically involve all enzymes from H that are not included in N_2, possibly obtaining a maximum use of the metaproteomic data. However, if our purpose is to emphasize the metabolic differences between two conditions under study, K could involve a subset of them, namely, those that are differentially expressed. To achieve this goal, we address one further optimization problem.

We begin from Eq. (3.1) and Eq. (3.3). The constraint on biomass production, Eq. (3.2), is removed, as it is currently satisfied in N_2. We now make use of binary variables, z_i. In particular, Eq. (3.5) relates v and z variables, where M is the maximum flux, and the minimum (non-zero) flux is 1. Equation (3.6) prevents reaction f and its reverse b from being active simultaneously.

$$z_i \leq v_i \leq M z_i \qquad \forall i \in R \tag{3.5}$$

$$z_f + z_b \leq 1 \qquad \forall (f, b) \in B \tag{3.6}$$

Then, for each enzyme, $j \in K$, we introduce a continuous variable, e_j, with a value between 0 and 1, as observed in Eq. (3.7). In Eq. (3.8), if any of the set of reactions, R^j, that are associated with enzyme $j \in K$ cannot be activated, then e_j is necessarily 1; therefore, to maximize the use of the enzymes in K, we must minimize the e_j variables. This is achieved by amending the objective function as in Eq. (3.9). In particular, for the e_j variables, we assign the maximum overall penalty, w_j.

$$0 \leq e_j \leq 1 \qquad \forall j \in K \tag{3.7}$$

$$\sum_{i \in R^j} z_i + e_j \geq 1 \qquad \forall j \in K \tag{3.8}$$

$$min \sum_{i \in R} (p_i - b_i) v_i + \sum_{j \in K} w_j e_j \tag{3.9}$$

Equation (3.9), which is subject to Eqs. (3.1), (3.3), and (3.5)-(3.8), is a mixed linear integer program and empirical evidence shows that it is not an expensive problem (less than 100 seconds in the instances considered in Section 3.3). Active reactions from this optimization problem are added to N_2 and define the final resulting metabolic network, N_3.

3.3. Results

Our approach is applied to draft the metabolic networks of two different, naphthalene-enriched communities (Guazzaroni et al., 2013) derived from an anthropogenically influenced, polyaromatic hydrocarbon (PAH)-contaminated soil with (CN2) or without (CN1) bio-stimulation with calcium ammonia nitrate, NH_4NO_3 and KH_2PO_4 and the commercial surfactant Iveysol$^{®}$. Naphthalene, a model PAH compound, is a common, persistent pollutant in crude oil and industrial chemical manufactures that can be released into the environment (i.e. soils) through anthropogenic activities (Kästner, 2008). Current treatments for naphthalene- and other PAH-contaminated sites involve the use of bio-surfactants and additional electron acceptors as well as nitrogen sources (nitrate and ammonia) to improve the bioavailability and bioremediation of these compounds. It has also been observed that many bacteria are capable of degrading and growing on naphthalene (Guazzaroni et al., 2013; Lu et al., 2011), and their activities might only be limited by environmental conditions. Thus, gaining insight into the mechanisms underlying naphthalene degradation can aid in the design of better remediation strategies.

3.3.1. Reconstruction of CN1 and CN2 functional networks

In our analysis, we only considered proteins with an annotated metabolic function, i.e. with an EC number, namely 570 out of 1234 measured proteins, collectively involving 327 unique EC identifiers. Based on the relative protein concentrations, we classified enzymes found in CN1 and CN2 (Guazzaroni et al., 2013) into the H, L, M and K sets as follows. For one scenario, enzymes listed in that sample were included in the H set, while enzymes that did not appear in that sample, but did appear in the other scenario, were included in the L set. As we were interested in obtaining networks that emphasized the differences between both scenarios, the K set involved up-regulated enzymes in each scenario. In particular, enzymes showing a 1.5-fold change in their relative protein concentrations in one sample compared with the other were considered up-regulated.

Using full-length and partial 16S rRNA gene sequences obtained through a metagenomic approach (Guazzaroni et al., 2013), it was found that 13 and 12 distinct species constituted the CN1 and CN2 communities, respectively, with only two species (*Achromobacter* and *Azospirilum*) conforming to the common set. While *Azospirillum*, *Comamonas*, *Achromobacter* and *Pseudoxanthomonas* species dominated CN1, *Pseudomonas* and *Achromobacter* species dominated CN2. This information was used to aid in the context-specific network reconstruction process. In particular, the set of related genome annotations for CN1 and CN2, which was established on the basis of phylogenetic affiliations (Guazzaroni et al., 2013), was obtained from the KEGG website. The enzymes (ECs) from these genome annotations, which were neither included in H nor L, were included in the M set.

The list of reactions and metabolites was downloaded from the Model SEED database (Henry et al., 2010). The above enzyme lists were translated into reactions lists using their EC numbers annotated in this database. The D set comprised ECs (enzymes) from the Model SEED database not included in H, M and L. When a reaction was associated with more than one EC belonging to different sets, the reaction was assigned to the most favourable set. For example, if a reation could be catalyzed by one enzyme from H and one from L, then the reaction was assigned to H.

A minimal medium based on naphthalene as the only carbon source (as was used in the enrichment cultures; Guazzaroni et al. (2013)) was defined for the reconstruction process. The biomass reaction was taken from a *Pseudomonas* reconstruction that was provided by SEED (rxn12834), as this specie plays a major role in CN2. We also used annotated modules from KEGG.

The computation time for both CN1 and CN2 reconstruction was less than 200 seconds. All computations were performed on a 64-bit Windows XP machine with an Intel Core 2 CPU at 2.4 GHz and 8 GB of RAM. The code was written in MATLAB and CPLEX was used to solve the linear optimization problems.

Randomly perturbing the selected weights with a 10 % uniform noise only changed a few reations, giving rise to very similar networks. For CN1, we used 148 of the 206 enzymes form H and 21 enzymes from L, and we completed the network with 274 and 165 enzymes from M and D, respectively. Similarly, for CN2, we employed 259 of the 311 enzymes from H and 1 enzyme from L, and we completed the network with 267 and 282 enzymes from M and D, respectively. In both cases, the use of enzymes from H was remarkable, corresponding to $> 70\,\%$ of the measured data, which was increased to $\sim 90\,\%$ for up-regulated enzymes (K set). In contrast, the number of enzymes in L required a more careful reading.

As noted above, there are reactions in the Model SEED database that involve more than one EC number and are, therefore, catalyzed by different enzymes. For example, if a reaction is catalyzed by one enzyme from H and one from L, we assume that the flux through this reaction is supported by the enzyme from H, which is consistent with the experimental data. An inconsistency arises when reactions that are exclusively catalyzed by enzymes in L are included in the reconstruction. We found four reactions of this type in CN1, which collectively involved three enzymes of 21. In particular, two of these inconsistent reactions in CN1 were associated with EC 2.5.1.9 (*riboflavin synthase*) and are required to produce FAD (*flavin adenine dinucleotide*), an essential metabolite for biomass production. The third reaction was linked to EC 2.4.1.227 and is required to produce the *peptidoglycan* subunit of *P. putida KT2440*, which is involved in biomass production. As this metabolite is specific for *Pseudomonas*, which is not involved in CN1, the need for this enzymes is unlikely. The fourth reaction is associated with EC 3.5.1.18 and is activated to support up-regulated enzymes. There is only one inconsistent reaction in CN2, which is associated with EC 2.7.7.38 and their activation is due to the same reason as for EC 3.5.1.18. The inclusion of these enzymes in CN1 is not in accord with the evidence from metaproteomic data, which may be attributed to three possible causes: (i) incompleteness of the Model SEED database; (ii) inaccuracy of the biomass reaction or (iii) a lack of resolution in the metaproteomic data. To address this issue, further experimental evidence is required.

With the resulting context-specific networks for CN1 and CN2, we decided to evaluate how single-reaction deletions could hamper their ability to produce biomass. CN2 turned out to be more resilient, as only 22 single-reaction deletions prevented its biomass production capacity, in contrast with 42 single-reaction deletions in CN1. However, the overlap was significant, as 19 of those reactions affected both networks. When deleting enzymes related to a given EC number, 31 instances affected growth in CN2 and 42 in CN1, with 23 of them being the same in both cases. In addition, we substituted naphthalene as the only carbon source with each of the compounds present in the reconstructed networks. CN1 was able to take advantage of 26 compounds to produce biomass, while CN2 exhibited theoretical ability to use 446 compounds. We conducted the same analysis with the nitrogen, phophorus and sulphur sources, finding that CN1 could make use of 166, 114 and 34 compounds, respectively, while CN2 could make use of 270, 212 and 104 compounds, respectively. Although these results should be taken with caution, they suggest that the metabolism of CN2 is more robust and varied than CN1. The fact that the availability of substrates is promoted during the bio-stimulation process used for obtaining a CN2 community (Guazzaroni et al., 2013) might agree with this hypothesis.

3.3.2. CN1 and CN2 pathway analysis

To obtain a global picture of the pathways characterized in the CN1 and CN2 contextualized networks, we resorted to the use of KEGG maps. In particular, to extract the functional differences between CN1 and CN2, we compared the KEGG maps using a score, J_p, derived from the Jaccard distance. In particular, for each KEGG map, we first calculated the Jaccard index, J, between CN1 and CN2, with Eq. (3.11), where A and B represent the set of reactions involved in CN1 and CN2, respectively, in a given KEGG map. Then, we determined the Jaccard distance (Eq. 3.11), which measures the dissimilarity between CN1 and CN2 for a particular pathway. Finally, we multiplied the Jaccard distance by the maximum between the number of reactions that belonged to CN1, but not to CN2, and vice versa (Eq. 3.12). This score gives more importance to pathways where the CN1 and/or CN2 networks show high coverage and share few reactions. An illustration of this process can be found in Appendix B for the "Histidine metabolism" KEGG map. Functional differences between CN1 and CN2 can be analysed via J_p, where the higher the value of J_p, the greater the difference between CN1 and CN2 and, hence, the more relevant the pathway.

$$J = \frac{|A \cap B|}{|A \cup B|} \tag{3.10}$$

$$J_\delta = 1 - J \tag{3.11}$$

$$J_p = J_\delta \cdot \text{máx}(|A \cap \bar{B}|, |\bar{A} \cap B|) \tag{3.12}$$

We ranked the KEGG pathways according to this measure for the CN1 and CN2 metabolic networks. Table 3.1 shows some of the top most different KEGG pathways between CN1 and CN2. We repeated the same analysis in two additional cases: (1) direct use of metaproteomic data form CN1 and CN2, neglecting our network reconstruction approach ("Rank only metaproteomics"); (2) removal of metaproteomic data, only considering CN1 and CN2 taxonomic data and their annotated genomes in our network reconstruction approach ("Rank taxonomics"). As observed in Table 3.1, substantial differences can be found among them, which emphasize the effect of our reconstruction approach, showing a clear contribution of proteomics to genomics data.

Table 3.1: Ranking of KEGG pathways after reconstruction using functional network data for CN1 and CN2. The columns "CN1" and "CN2" indicate the number of reactions involved in CN1 and CN2 reconstructions active in the KEGG pathway under consideration. The "Rank" column indicates the position of KEGG pathways according to descending order of the obtained score. The "Rank only metaproteomics" column indicates the rank obtained for KEGG pathways before the reconstruction process, namely with the score exclusively calculated from metaproteomic data of CN1 and CN2. The "Rank taxonomics" indicates the rank obtained after the reconstruction only with taxonomic data, i.e. with empty H, L and K sets.

KEGGID	Name	CN1	CN2	Score	Rank	Rank only meta proteomics	Rank taxonomics
map00071	Fatty acid metabolism	4	27	19.5926	1	22	1
map00062	Fatty acid elongation	0	15	15	2	42	12
map00330	Arginine and proline metabolism	17	31	14.0541	3	19	10
map00540	Lipopolysaccharide biosynthesis	3	18	12.5	4	54	39
map00760	Nicotinate and nicotinamide metabolism	13	25	12.4667	5	24	31
map00230	Purine metabolism	42	57	12	6	5	23
map00281	Geraniol degradation	0	12	12	7	37	-
map00260	Glycine, serine and threonine metabolism	14	26	9.31034	8	33	21
map00400	Phenylalanine, tyrosine and tryptophan biosynthesis	12	24	8.61538	9	14	79
map00500	Starch and sucrose metabolism	5	10	8.35714	10	62	8
map00523	Polyketide sugar unit biosynthesis	0	8	8	11	20	6
map00620	Pyruvate metabolism	12	23	7.8	12	36	28
map00130	Ubiquinone and other terpenoid-quinone biosynthesis	2	9	7.2	13	2	7
map00364	Fluorobenzoate degradation	7	0	7	14	102	-
map00650	Butanoate metabolism	14	12	6.85714	15	43	33

Table 3.1 shows clear differences between CN1 and CN2. The geraniol degradation pathway (map00281) was predicted to be completely functional in CN2, but inactive in CN1. In CN2, enzymes from H in this pathway were complemented with enzymes form M and D. In contrast, in CN1, enzymes from H were discarded from the reconstruction. On the other hand, the fluorobenzoate degradation pathway (map00364) was filled in to some extent in the CN1 reconstruction, whereas it was inactive in CN2. Note that these differences between CN1 and CN2 cannot be easily obtained from the other two cases considered (see "Rank only metaproteomics" and "Rank taxonomics"). This is particularly relevant since these differences between CN1 and CN2 are experimentally validated below, which shows the predictive power and need of the approach presented here.

3.3.3. Experimental analysis of fluorobenzoate and geraniol metabolism in CN1 and CN2

Given the high rank obtained by the fluorobenzoate and geraniol degradation pathways and its specificity for CN1 and CN2, respectively, we evaluated the correctness of these hypotheses. First, *in silico* stoichiometric analysis showed that CN1 was capable of growing with fluorobenzoate

Table 3.2: Summary of sensitivity analysis to the inclusion of fluorobenzoate and geraniol with different biomass equations. An X indicates that the corresponding compound is part of the reconstruction computed using the corresponding biomass reaction from the Model SEED reaction database.

	CN1		CN2	
Reaction	Fluorobenzoate	Geraniol	Fluorobenzoate	Geraniol
rxn12828	X	-	-	X
rxn13762	X	-	-	X
rxn12821	X	-	-	X
rxn13777	X	-	-	X
rxn11733	X	-	-	-
rxn11918	X	-	-	X
rxn11928	X	-	-	X
rxn12830	-	-	-	X

as the sole carbon source, and the same was observed for CN2 and geraniol, after some minor corrections to the obtained networks. In particular, for CN1 to produce biomass from fluorobenzoate, we needed to allow for the net production of fluoride (F-), which is found in abundance in soils (McQuaker and Gurney, 1977), and in bacterial cultures as cellular degradation by-product (Hidde Boersma et al., 2004). In the case of CN2 and geraniol, we needed to change the direction of reaction rxn07886 (*geranic acid CoA-transferase*) in the SEED Database, which converts geranic acid into *trans*-geranyl-CoA and was originally defined to act in the opposite direction. Based on KEGG (map00281) and existing literature (Clemente-Soto et al., 2014), we found that this reaction is commonly annotated in the direction proposed. Note here that the opposite is not possible, i.e. the growth of CN2 and CN1 on fluorobenzoate and geraniol, respectively, as they are not active in CN2 and CN1, respectively. It is important to clarify that these issues come from inaccuracies in The SEED Model database and not from the algorithm presented here. When doing this modification prior to the reconstruction process, the resulting networks directly grow on fluorobenzoate in CN1 and geraniol in CN2.

Secondly, in order to discard that the relevance of fluorobenzoate in CN1 and geraniol in CN2 is an artefact derived from an inaccurate biomass equation, we conduct a sensitivity analysis with different existing biomass equations, finding that the major conclusions achieved are conserved in most cases (Table 3.2).

Experimental validation assays were conducted to prove the extent of agreement with our computational predictions. For that, we set up CN1 and

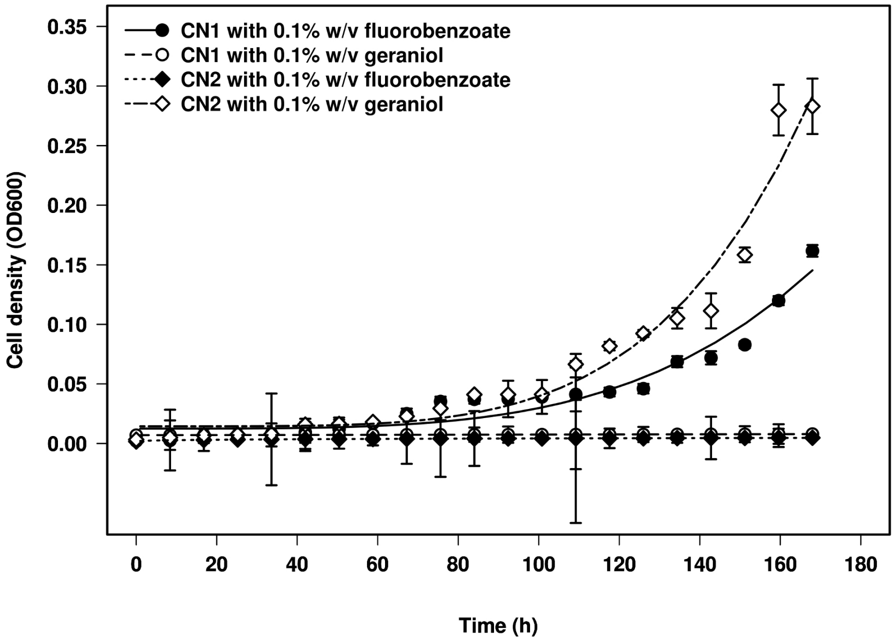


Figure 3.2: Growth curve of CN1 and CN2 enrichment cultures in Bushnell Hass minimal medium in the presence of 0.1 % (*w*/*v*) 3/4-fluorobenzoate and geraniol, respectively at 30°C and 250 rpm. As shown, within the examined time frame, no appreciable growth was observed in clutures of the CN1 and CN2 consortia in the presence of geraniol and fluorobenzoate, respectively.

CN2 enrichment cultures using previously described conditions (Guazzaroni et al., 2013); instead of naphthalene as the carbon source, geraniol and 3/4-fluorobenzoate (0.1 % w/v) were used, and samples were taken at different time points (see Appendix B). Fingerprinting by Gas Chromatography-Mass Spectrometry (GC-MS) was used to confirm the presence of the initial substrates as well as the existence of degradation intermediates in both cultures. A careful inspection of the MS signatures of the initial metabolites known to participate in geraniol (map00281) and 3/4-fluorobenzoate (map00364) degradation (see Appendix B) confirmed the presence of 3/4-fluorocatechol in CN1 and citral and geranic acid in CN2. These findings demonstrated that the fluorobenzoate-degradation pathway occurred or was active in CN1, while the geraniol-degradation pathway is active in CN2. This was also confirmed by measuring the OD_{600} of the enrichment cultures at different time intervals (Fig. 3.2). As shown, CN1 grew only in the presence of fluorobenzoate (0.1 % w/v), whereas CN2 grew only in the presence of geraniol (0.1 % w/v).

3.3.4. Contributions of bacteria to the CN1 and CN2 functional networks

We also attempted to quantify the contributions of particular sets of microbes to the entire reconstructed, context-specific metabolic network, where multiple proteins from multiple organisms are coexpressed. This is an important advance because the complement of proteins used to metabolize recalcitrant pollutants and the specific roles of different bacterial members within a consortium in pollutant (or other potential carbon/energy sources) deconstruction are not well explored.

As the population diversity and structures of the two enrichment cultures were relatively low and well characterized, the taxonomic affiliations of the proteins quantified in the shotgun metaproteomes could be unambiguously established (Guazzaroni et al., 2013). Based on this, for CN1 and CN2, we knew which members of the community were actually expressing the enzymes used to catalyse each reaction in H.

To evaluate the role of each bacterial member in CN1 and CN2 at the functional level, we determined its contribution to each KEGG map. The contribution was determined as the number of times a bacterium appeared in a KEGG map divided by its total number of active reactions. For this analysis, we only took into account the reactions in H and M that were involved in the CN1 and CN2 reconstructed networks. As noted above, the taxonomic affiliation was known for the reactions in H. In contrast, for the reactions in M, different members of the community might be involved in a rection. For simplicity, in these situations, if possible, we assigned an organism that was previously included in the KEGG map via the reactions from H. Full details as to the taxonomic assignment of reactions involved in the the CN1 and CN2 metabolic networks can be found in Supplementary Material IV of Tobalina et al. (2015).

Figure 3.3 shows the contribution of each organism found in both CN1 and CN2 to each KEGG map. Pathways were reconstructed for the most abundant populations, which included composite genomes for populations closely related to sequenced strains of *Achromobacter*, *Azospirillum*, *Comamonas*, *Mesorhizobium*, *Microbacterium*, *Planctomycetes*, *Pseudoxanthomonas*, *Singulisphaera* and *Pseudomonas.*

Identification of genes for naphthalene processing (map00626) and metabolic reconstructions suggested *Achromobacter* followed by *Mesorhizobium* and *Pseudoxanthomonas* in CN1 and mainly *Achromobacter* in CN2 as key groups for naphthalene degradation. In addition, we identified *Achromobacter*, *Azospirillum*, and *Comamonas* in CN1 and *Azospirillum* as well as *Pseudomonas* in CN2 as groups that might primarily metabolize low mole-

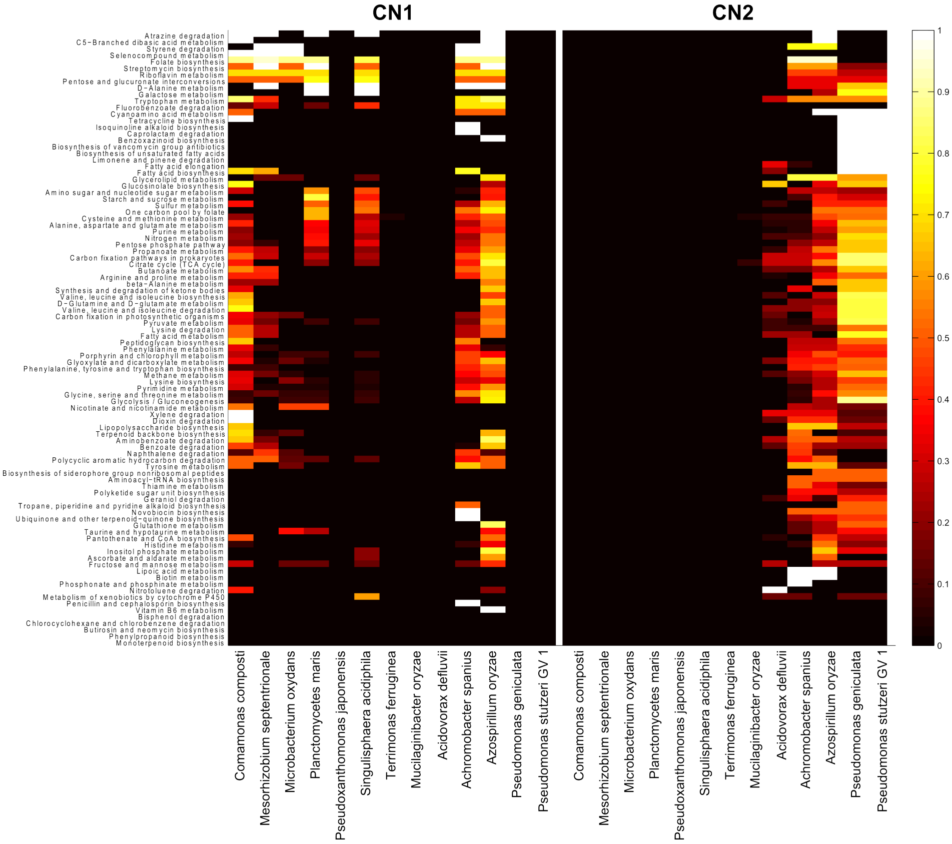


Figure 3.3: Heatmap showing the contributions of the most relevant bacterial members of CN1 and CN2 to the KEGG maps. Relative contributions of each of the 13 distinct species found to constitute the CN1 and CN2 communities (Guazzaroni et al., 2013) are differentiated by a colour code. A high-resolution image can be found in Supplementary Material V of Tobalina et al. (2015).

cular weight molecules produced from naphthalene. It could also be observed that, while metabolic reconstructions indicated a central role played by *Achromobacter* in naphthalene degradation, multiple bacteria participated in the active pathways (see Appendix B for further details; Figs. B.2-B.4). A careful examination of the data presented in Fig. 3.3 clearly leads to a different pathway organization at the organismal level.

3.4. Conclusion

In this study, we make use of the novel computational procedure for obtaining a context-specific functional metabolic network for a bacterial community using metaproteomic data introduced in Section 3.2. Our approach was based on the mathematical optimization model presented in the Model SEED (Henry et al., 2010). However, we adapt this model to incorporate metaproteomic data and obtain a context-specific metanetwork in which the identified proteins from the multiple organisms making up the community are coexpressed. To this end, we also include important technical differences. In particular, we use a 3-step iterative procedure based on linear programming and mixed integer linear programming.

Our approach is an alternative to previously reported methods (dos Santos et al., 2013; Khandelwal et al., 2013; Zomorrodi and Maranas, 2012), where the role of each organism is explicitly represented in a different metabolic compartment and, therefore, their relationships can be directly analysed. These methods require the genome-scale metabolic network of each organism in the community as input data, which, in consortia involving a high number of species as we have here, is typically not available; therefore, we turn to a metanetwork approach, which involves several assumptions. First, we need a consensus biomass equation that represents the metabolic requirements of the community to support growth. With metametabolomics approaches being developed, it is expected that consensus biomass equations will be refined in the near future. Second, a free exchange of metabolites between species is allowed, as boundaries between individual organisms are not defined. However, a metanetwork could serve as a basis to disentangle the role of each organism in the community, as suggested in Section 3.3.4. More sophisticated approaches need to be developed for this task, for example, analysing the role of a single organism in the context of the entire metanetwork.

Our approach was applied to draft the context-specific metabolic networks of two different naphthalene-enriched communities (Guazzaroni et al., 2013). Analysis of the resilience to single-reaction elimination and the ability to grow on different sources suggests that CN2 metabolism is more varied than CN1. Then, we used KEGG maps to obtain a global picture of the reconstructed draft networks. We were able to capture the overall functional differences between CN1 and CN2 at the metabolic level. We showed that CN1 and CN2 utilize different metabolic pathways to synthesize essential metabolites for growth. In particular, we hypothesized an important role for the fluorobenzoate degradation pathway in CN1 and for geraniol metabolism in CN2. Experimental validation was conducted and good agreement with our computational predictions was observed.

On the other hand, we showed that these metabolic differences lead to a different pathway organization at the organismal level. For example, while naphthalene degradation (map00626) seems to be supported by *Achromobacter* in both CN1 and CN2, *Mesorhizobium septentrionale* and *Pseudoxanthomonas japonensis* may be involved in an alternative pathway in CN1. In addition, while metagenomic sequences outlined the broad metabolic capabilities of the abundant populations present in an adapted community, proteomics-guided metabolic reconstructions allowed us to focus on the pathways that were actually expressed and refine the assignment of roles for community members not only in naphthalene degradation but also in the assimilation of the low molecular weight compounds produced from it.

These results show that network-based methods represent a promising strategy for exploiting the value of data and the available bioinformatics tools, allowing us to obtain a better understanding of biological systems. As the available meta-omics data from scientific studies at different levels are increasing, reconstruction procedures will play an important role in disentangling contexts-specific metabolic phenotypes. The approach presented here can be extended to meta-genomic and meta-transcriptomic data and will clearly benefit from the availability of meta-metabolomic data, mainly to address the failure to detect all different enzymes (ECs) that catalyze different reactions. Amending our approach to include these data is straightforward.

Chapter 4

Assessment of FBA based Gene Essentiality Analysis

Gene Essentiality Analysis based on Flux Balance Analysis (FBA-based GEA) is a promising tool for the identification of novel therapeutic targets in cancer. The reconstruction of cancer specific metabolic networks, typically based on gene expression data, constitutes a sensible step in this approach. However, to our knowledge, no extensive assessment on the influence of the reconstruction process on the obtained results has been carried out to date.

In this chapter, we aim to study context-specific networks and their FBA-based GEA results for the identification of cancer specific essential genes. To that end, we used gene expression datasets from the Cancer Cell Line Encyclopedia (CCLE), evaluating the results obtained in 174 cancer cell lines. In order to more clearly observe the effect of cancer-specific expression data, we did the same analysis using randomly generated expression patterns. Our computational analysis showed some essential genes that are fairly common in the reconstructions derived from both gene expression and randomly generated data. However, we also found essential genes that are very rare in the randomly generated networks, while recurrent in the sample derived networks, and, thus, would presumably constitute relevant drug targets for further analysis. In addition, we compare the *in-silico* results to high-throughput gene silencing experiments with conflicting results, which leads us to raise several questions. Notwithstanding, there are findings in the literature that indicate some of the predictions are in the right track.

4.1. Introduction

The results obtained from FBA-based GEA are dependent on the different elements involved in this network reconstruction process, i.e. reference network, defined growth medium, gene expression data and reconstruction algorithm. However, to our knowledge, no extensive assessment evaluating the influence of the metabolic reconstruction process and expression data on the results of gene essentiality analysis has been carried out to date in cancer. To that end, in this chapter, we conducted an extensive study for different types of cancers from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) so as to disentangle the effect of some of these factors in the resulting list of essential genes. In order to more clearly observe the effect of cancer-specific expression data, we did the same analysis using randomly generated expression patterns. In addition, we used high-throughput gene silencing data (Cowley et al., 2014) to extensively test the predictions of the FBA-based GEA approach. Finally, we contrasted literature data about predicted essential genes in two Glioblastoma Multiforme (GBM) cell lines.

4.2. Methods

4.2.1. Gene Essentiality Analysis

Essential genes are defined here as those genes whose removal render the cell unable to produce biomass. Using the Boolean gene-protein-reaction rules incorporated in Recon 2 (Thiele et al., 2013), we can evaluate which reactions will stop working after a particular gene is deleted. Thus, a gene knock-out is simulated by setting the upper and lower bounds of the corresponding reactions to zero in an FBA calculation, and checking whether (or not) the remaining network is still able to produce biomass.

In order to reduce the number of FBA calculations required to check the essentiality of every single gene, we first calculated the maximum biomass possible in the wild-type network and searched for a flux distribution with minimum sum of fluxes through reactions for which gene-to-reaction mapping is defined. If a particular gene knock-out does not affect any reaction in that optimal flux distribution, we can be certain that a new FBA calculation will give us the same solution as in the wild-type network and we can therefore skip such gene knock-out.

4.2.2. Network contextualization

The FBA based GEA is done using context specific metabolic networks. The reference human metabolic network is contextualized using cancer gene expression data by means of the fast network reconstruction algorithm introduced in Section 2.3.

As explained in Section 2.3, the input of the reconstruction algorithm is the reaction classification as highly (H), medium (M) or lowly (L) expressed. This information is obtained from gene expression experiments, in our case collected from GEO database (Edgar et al., 2002).

We used the Gene Expression Barcode (McCall et al., 2011), a robust statistical method developed to predict expressed and non-expressed genes in microarrays, to treat the expression data. This expression processing tool allowed us to build one network from each sample, since it is designed to be able to work with just one sample and make it comparable to others, instead of needing several samples at the same time. In particular, we focused on Affymetrix HGU133plus2 arrays. We preprocessed the data, one sample at a time, using Barcode's R script and retrieved the z-score values obtained from this algorithm. The process is equivalent to processing each sample with fRMA (McCall et al., 2010), an algorithm that is at the core of Barcode.

Because the z-scores retrieved from Barcode were given at the probeset level, using gene-probe relationships annotated in hgu133plus2.db R package, we obtained the gene value as the median value of the corresponding probe sets (one gene may be interrogated in a microarray by different probes, and taking the median value is a robust way of summarizing that information). Each gene value was transformed into present (1) and absent (0) calls using Barcode's criteria. Subsequently, present genes were classified as high (H) and absent genes as low (L). Finally, we used the GPR rules to convert the gene classification into a reaction classification.

If a reaction is associated to a single gene, it is classified as H, M or L depending on the classification of the corresponding gene. If it involves an OR rule, it is classified as H if one of the genes is classified as H. On the contrary, it is classified as L if all the genes are classified as L. If a reaction involves an AND rule, it is classified as H if all the genes are classified as H, while as L if any of the genes is classified as L. Those reactions for which no gene expression is available or that are not related to any gene (e.g. spontaneous reactions) are classified as medium expressed (M). A systematic way to perform this conversion is to assign a numerical value to each category (+1 to H, 0 to M and -1 to L) and to substitute AND and OR rules by min() and max() functions respectively, after which we only need to evaluate the resulting mathematical expression.

4.2.3. Comparison to experimental data

A systematic effort to identify essential genes in different cancer cell types is being carried out in what is known as project Achilles (Cowley et al., 2014). The coverage of this project has grown during the last years (Cowley et al., 2014; Cheung et al., 2011; Luo et al., 2008). They performed high-throughput gene silencing experiments to identify vulnerabilities in different cell lines using a library of lentivirally encoded short hairpin RNAs (shRNAs). The method consisted in infecting cultured cells with a pool of shRNAs, allowing them to proliferate for a period of time and measuring the relative abundance of the shRNAs at the end of the experiment (Luo et al., 2008). Underrepresented shRNAs would target more essential genes than the rest, as surviving cells will only be carrying shRNAs silencing genes not related to proliferation. Comparison of computationally obtained essential genes with the results coming from this type of experiments have been previously used to assess the validity of the approach (Folger et al., 2011).

The processing and interpreting of the data generated by these experiments is quite challenging. In each new paper, new ways of looking into the data are presented. Luo et al. (2008) used the RIGER score to analyze the observed fold changes, while Cheung et al. (2011) and Cowley et al. (2014) used the ATARiS score. The later gives a relative score of essentiality for each gene in one cell line with respect to that gene in the rest of cell lines. This means that a better ATARiS score for a gene in a cell line with respect to the rest of the genes in that same cell line does not necessarily mean that that gene is more essential than the others.

Recently, Hart et al. (2014) proposed to use gold standards of essential and nonessential genes to classify the genes depending on their corresponding observed shRNA fold changes as belonging to essential or nonessential genes. Their approach fits a probability density function for each type of gene using the gold standards that is later used to classify the rest of the genes not included in the gold standards. Specifically, a bayes factor is computed for each gene, indicating how more likely is it to be essential with respect to being nonessential according to the observed fold changes. It is this processing method that we use for the data here.

In order to assess the accuracy of our approach to predict essential genes, we used the high-throughput silencing experiments taken from project Achilles (Cowley et al., 2014). We derived a score for each gene in each cell line following the method introduced in Hart et al. (2014). However, we multiplied the obtained scores by -1 so that the lower the score, the more essential the gene is supposed to be, as it happens with the shRNA fold changes in the high-throughput silencing experiments. We then compared the distribution of the scores of the obtained essential metabolic genes versus the nonessential metabolic genes using a one-sided two-sample Kolmogorov-Smirnov test, as suggested in Folger et al. (2011). This test helps us to see if the obtained essential genes are biased towards lower, more essential scores. However, the bias may be significant but not sufficiently large so, in addition, we measured the proportion of obtained essential genes with a negative score in each scenario.

4.3. Results

4.3.1. Gene essentiality analysis

With the fast reconstruction algorithm introduced in Section 2.3 in our hands, we can address the question of the extent to which the set of essential genes is being affected by context-specific expression data. To further explore this issue, we permuted the metabolic gene expression classification of each sample 10 times and reconstructed the corresponding networks followed by the calculation of their corresponding essential genes, leading to a background of almost 2000 random results.

The cancer gene expression data used for this study comes from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012). The selected CCLE cell lines include samples from 20 different cancer types: bladder, bone sarcoma, breast, colon, endometrial, esophageal, glioblastoma, gastric, leukemia, liver, lung mesothelioma, lung NSCLC, lung SCLC, melanoma, multiple myeloma, ovarian, pancreas, prostate, renal cell carcinoma and soft tissue sarcoma. The choice of this subset of cell lines was made taking into account the available high-throughput gene silencing data from project Achilles (Cowley et al., 2014) (see Table D.1). We also selected some other U251 and U87 samples from GEO to perform an analysis more focused on a specific type of cancer, GBM in this case (Appendix D).

Figure 4.1 shows the result of this experiment for schema 3. As partially expected, there are some genes that are fairly common in any reconstructed network. The most extreme cases are genes that appear as essential whatever the input expression is. These are a direct consequence of the input reference network, the fixed growth medium conditions and the selected biomass reaction. This analysis confirms the extent to which these factors can affect the results.

Note that there also exist some essential genes very frequent in the individual samples but less frequent in the random networks. These would be, a priori, the most interesting ones, as they are more related than the

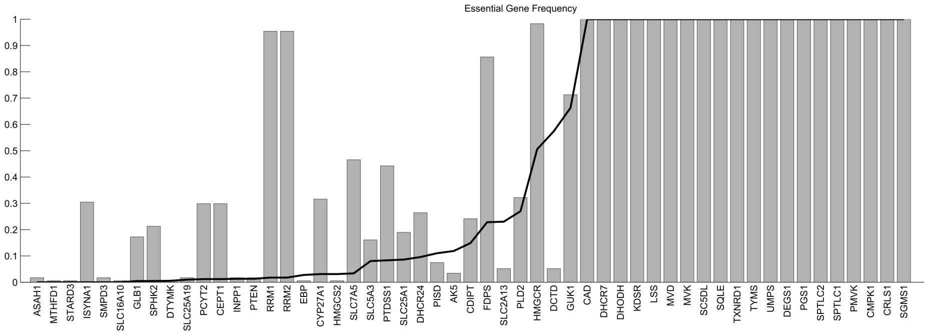


Figure 4.1: Essential gene frequency for reconstructed context-specific networks of selected cancer cell lines using our algorithm with schema 3 and Recon 2 as the base network. The horizontal axis contains the ENTREZ Symbols of the obtained essential genes. The height of the bars indicates the fraction of randomly reconstructed networks in which the corresponding gene appears as essential.

other genes to the particular expression of the samples.

Figure 4.2, Figure 4.3 and Figure 4.4 are analogous to Figure 4.1, but focusing on the GBM samples present in CCLE, U251 samples from GEO and U87 samples from GEO, respectively. As can be observed, very similar conclusions can be extracted.

The most striking fact is that the list of obtained essential genes exclusive of each cancer type is fairly short. Only 6 genes appeared in one cancer type when using our algorithm with schema 3, 22 and 21 if we used schema 1 and 2, respectively. We expected a more diverse set of essential genes for each cancer type. Changing the parameter settings of the reconstruction algorithm, the base network or the culture medium did not significantly affect this observation (see Table 4.1).

Some previous work explored the essentiality concept under very diverse growth medium conditions (Almaas et al., 2005) for some bacterial metabolic networks. They concluded the existence of a core set of reactions needed for biomass production independent of the selected growth medium. Our study leads to very similar insights for the case of network contextualization. The same conclusion was achieved for other parameter settings.

4.3.2. Comparison to high-throughput gene silencing experiments

The results for the Achilles experiments are summarized per gene as explained before, so that the more negative the score is, the more essential the gene is considered. Focusing on samples from the CCLE with matched Achilles experiments, we used a two-sample one-tailed Kolmogorov-Smirnov

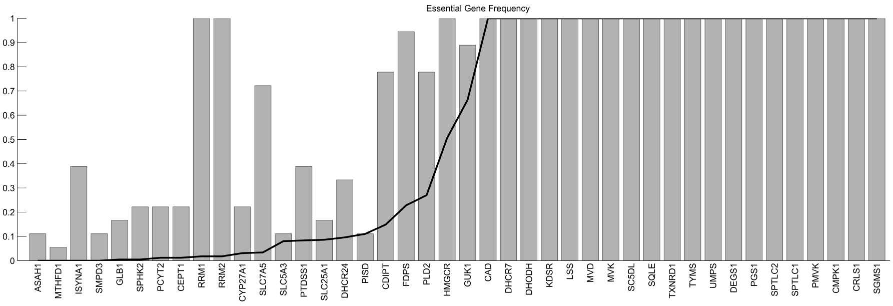


Figure 4.2: Essential gene frequency in networks reconstructed from GBM samples in CCLE with matched Achilles experiments using our algorithm with schema 3 and Recon2 as the reference network. The horizontal axis contains the ENTREZ Symbols of the obtained essential genes. The height of the bars indicates the fraction of samples in which the gene appears as essential. The height of the black line indicates the fraction of randomly reconstructed network in which the corresponding gene appears as essential.

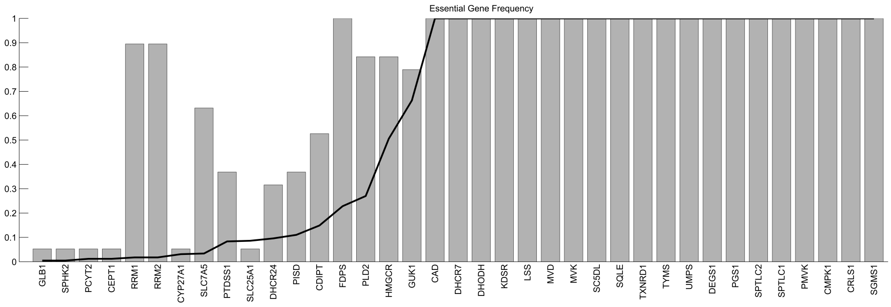


Figure 4.3: Essential gene frequency in networks reconstructed from U251 samples using our algorithm with schema 3 and Recon2 as the reference network. The horizontal axis contains the ENTREZ Symbols of the obtained essential genes. The height of the bars indicates the fraction of samples in which the gene appears as essential. The height of the black line indicates the fraction of randomly reconstructed network in which the corresponding gene appears as essential.

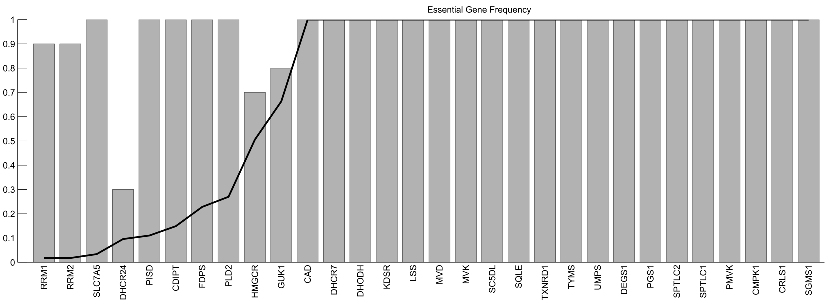


Figure 4.4: Essential gene frequency in networks reconstructed from U87 samples using our algorithm with schema 3 and Recon2 as the reference network. The horizontal axis contains the ENTREZ Symbols of the obtained essential genes. The height of the bars indicates the fraction of samples in which the gene appears as essential. The height of the black line indicates the fraction of randomly reconstructed network in which the corresponding gene appears as essential.

Table 4.1: Number of genes that appear as essential exclusively in samples from one single cancer type. Default settings are $\alpha = 1000$, $\delta = 0{,}10$ and $v^{*}_{biomass} = 0{,}01 \cdot v^{max}_{biomass}$. If parameters are not explicitly stated, default values are assumed.

Setup	Schema	Number of genes exclusive of one cancer type
Default settings	1	22
	2	21
	3	6
$\alpha = 10^2$	1	20
	2	34
	3	11
$\delta = 0{,}01$	1	30
	2	20
	3	14
$v^{*}_{biomass} = 0{,}10 \cdot v^{max}_{biomass}$	1	29
	2	24
	3	4
General growth medium	1	19
	2	19
	3	18
Recon 1	1	21
	2	17
	3	7

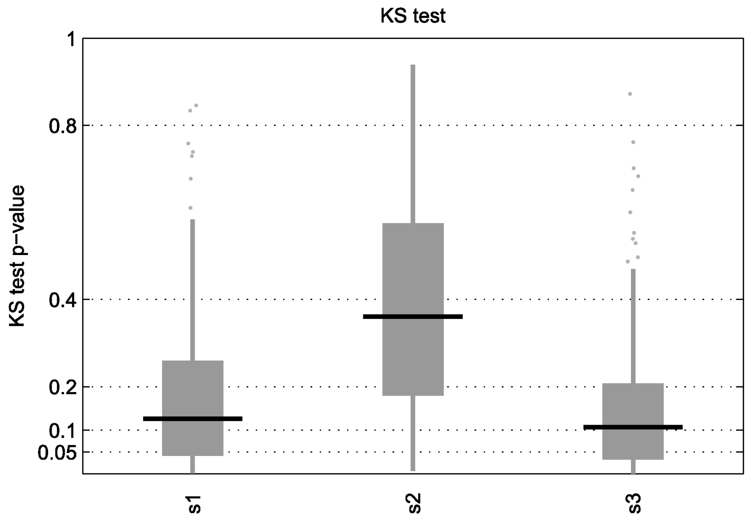


Figure 4.5: KS test p-value of essential genes obtained from the networks reconstructed from CCLE samples with matched Achilles experiment using our algorithm under schema 1, 2 and 3, with Recon 2 as the reference metabolic network.

test between the metabolic essential genes obtained from our *in-silico* approach and the rest of metabolic genes to see if the former had generally lower scores than the later. Figure 4.5 shows the results for this analysis, where our algorithm (under different schemes) does not seem to obtain a high number of results below the 5 % value (which should be later corrected for multiple hypothesis testing using, for example, the false discovery rate (FDR)). Figure 4.6 provides results using Recon 1, where results below the 5 % value are more common. However, we expected the FBA based GEA to give significant results with almost all the samples we tried it on.

In addition, we decided to count the fraction of genes predicted as essential in our *in-silico* approach with a negative score. The reason is that when the score becomes negative, the gene starts to have a higher probability of being essential than non-essential. If the FBA based GEA methodology is capturing essentiality correctly, the list of essential genes obtained should be enriched in genes with negative scores. We observed in Figure 4.7 that the proportion is around 10 to 20 %, regardless of the reconstruction method used. Around 8 to 16 % of all the metabolic genes represented in Recon 2 had a negative score. In other words, the list of essential genes returned by the FBA based GEA approach did not provide a significantly higher number of genes with negative (more essential) scores. Using Recon 1 instead of Recon 2 (Figure 4.8) or trying with different parameter settings for the

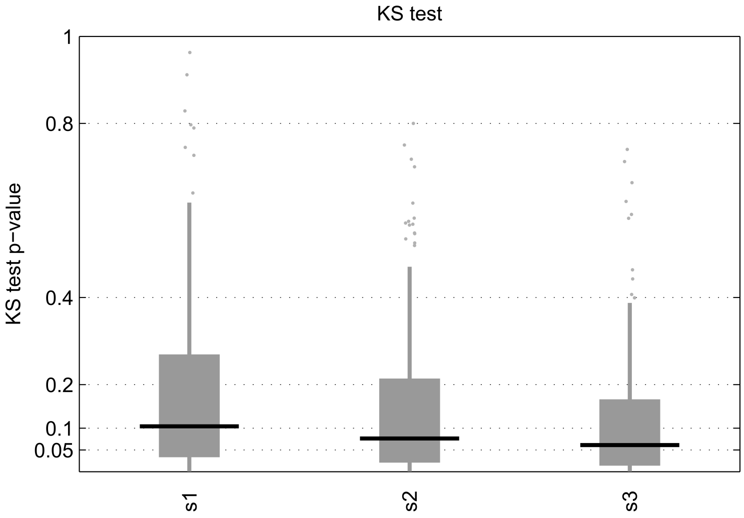


Figure 4.6: KS test p-value of essential genes obtained from the networks reconstructed from CCLE samples with matched Achilles experiment using our algorithm under schema 1, 2 and 3, with Recon 1 as the reference metabolic network.

reconstruction algorithm (Table 4.2) did not alter this conclusion. While Almaas et al. (2005) found that the core reactions in bacterial metabolic networks were enriched in experimentally proven essential genes, our results do not allow us to make the same claim.

Knowing that some genes appear similarly in both random and cancer-specific expression based networks, while others are clearly tied to the cancer-specific expression pattern, we tried to identify a relationship between this and the corresponding experimental essentiality score. However, we were unable to find a significant correlation.

4.3.3. Findings in the literature

Despite the results of our comparison with project Achilles data, we decided to take a look into the essential genes obtained for the glioblastoma multiforme (GBM) type cell lines. Since the data used only considered a single sample for each cancer cell line, we applied the same approach to an additional set of U251 and U87 GBM cell line samples (see Appendix D).

We found interesting the case of the gene PLD2 (ENTREZ ID 5338). This gene appears as essential in 84 % of the networks reconstructed from GBM samples with schema 3, while it only appears in 27 % of the networks reconstructed from random expression data and in 32 % of the networks reconstructed from all the CCLE selected samples. This gene has been shown

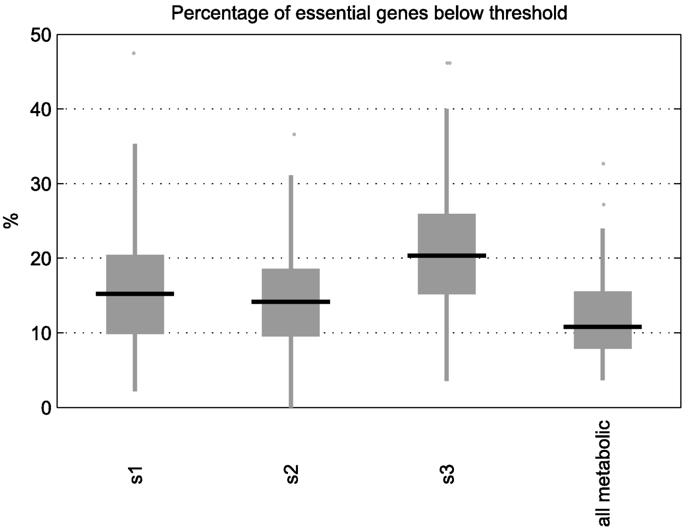


Figure 4.7: Percentage of essential genes obtained from the reconstructed context-specific networks of selected cancer cell lines using our algorithm under schema 1, 2 and 3 that have a negative score, indicating a higher probability of being essential. The base network is Recon 2. The last boxplot indicates this percentage when all the metabolic genes are taken into account.

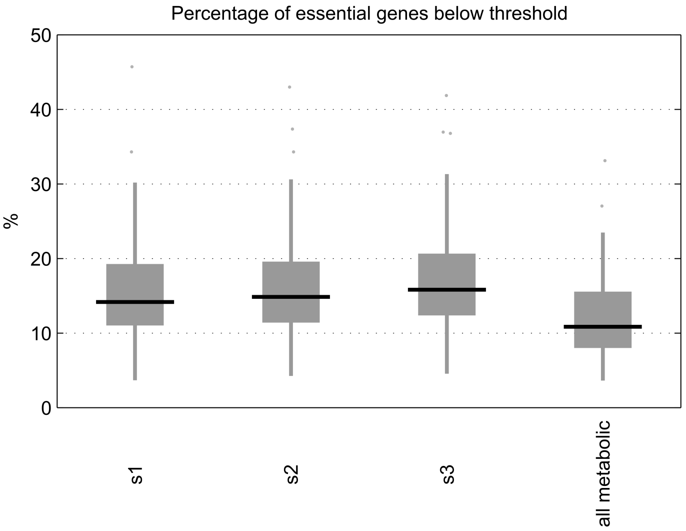


Figure 4.8: Percentage of essential genes obtained from the reconstructed context-specific networks of selected cancer cell lines using our algorithm under schema 1, 2 and 3 that have a negative score, indicating a higher probability of being essential. The base network is Recon 1. The last boxplot indicates this percentage when all the metabolic genes are taken into account.

Table 4.2: Median KS test p-value and percentage of essential genes with negative score under different parameter settings. Default settings are $\alpha = 1000$, $\delta = 0{,}10$ and $v^*_{biomass} = 0{,}01 \cdot v^{max}_{biomass}$. If parameters are not explicitly stated, default values are assumed.

Setup	Schema	KS test p-value	Percentage of essential genes below threshold
	1	0.1508	0.1403
$\alpha = 10^2$	2	0.3852	0.14
	3	0.1373	0.1852
	1	0.105	0.1505
$\delta = 0{,}01$	2	0.3	0.153
	3	0.3276	0.15
	1	0.1275	0.1509
$v^*_{biomass} = 0{,}10 \cdot v^{max}_{biomass}$	2	0.3694	0.1429
	3	0.1068	0.2034
	1	0.4331	0.1429
General medium	2	0.4863	0.1429
	3	0.3638	0.2

to play an important role in cancer (Kang et al., 2014) and to inhibit the proliferation of U251 cells when suppressed (Chen et al., 2014). However, its Achilles score in the U251 cell line is 20.03 and 12.70 in the U87 cell line, which, in principle, would have not attracted our attention (because positive scores predict the gene to be non-essential).

Other interesting genes are RRM1 (ENTREZ ID 6240), RRM2 (ENTREZ ID 6241), FDPS (ENTREZ ID 2224) and HMGCR (ENTREZ ID 3156), although these also appear as essential in many of the networks reconstructed from all the CCLE selected samples (but not in the randomly reconstructed networks). RRM1 and RRM2 form the enzyme ribonucleotide reductase, whose inhibition via GTI-2040 has been shown to have a potent antitumor activity in different cell lines, including U87 (Lee et al., 2003). On the other hand, FDPS inhibition has shown an increased paclitaxel-induced apoptotic cell death in U87 glioblastoma cells (Woo et al., 2010). Finally, inhibition of HMGCR through simvastatin reduced cell growth in U87 cells (Gliemroth et al., 2003) and increased apoptosis in an *in vivo* mouse GBM model (Bababeygy et al., 2009).

The Achilles score for RRM1 in U251 and U87 is -17.8360 and -7.416, respectively; 5.6092 and 8.7197 for RRM2; -4.08 and 6.74 for FDPS; 14.85 and 5.16 for HMGCR. Again, except for RRM1 and FDPS in U251, these genes would probably have not called our attention, but there is literature in their favour.

4.4. Conclusion

In this chapter, we have evaluated the accuracy of the results obtained from Flux Balance Analysis based Gene Essentiality Analysis (FBA based GEA) when networks contextualized with gene expression are used. In order to carry out this research, we have used a new fast metabolic network reconstruction algorithm introduced in this thesis (see Chapter 2.3). We focused on the reconstruction of networks from samples in the Cancer Cell Line Encyclopedia (CCLE) that had an available high-throughput gene silencing experiment in project Achilles.

Current methods that apply gene essentiality analysis, to our knowledge, lack an assessment of their results against random input, which would allow us to distinguish them from those obtained because of algorithmic artifacts or the rigidity or incompleteness of the reference network. Here, we directly addressed this issue by evaluating the impact of cancer-specific gene expression in GEA in comparison with randomly generated expression patterns.

The results show that the resulting list of essential genes has a varying degree of dependence on the reference network, the reconstruction algorithm and context-specific expression data. Most relevant conclusions are discussed in the following paragraphs.

On the one hand, we emphasize the need of a robust gene essentiality analysis, as the general results seem insensitive to the reconstruction process, at least when transcriptomic data is exclusively used. In particular, using randomly generated expression patterns, we found essential genes that are more likely to appear than others, which illustrate a higher dependence on the reference network than on the context-specific expression data.

On the other hand, we found essential genes that are rare in the randomly generated networks and recurrent in cancer samples. These can be considered as highly dependent on the context-specific expression data and are presumably the ones we would find more interesting at first glance. However, the comparison to experimental high-throughput silencing data cannot be considered to agree with the computational results, suggesting the need to consider additional factors.

We believe the selection of the appropriate biomass reactions, or an appropriate metabolic task function, is key in the whole process, as it plays a central part in the definition of essentiality. In fact, the selection of an appropriate biomass function is probably one of the biggest challenges ahead in the FBA based GEA methodology. The use of the biomass function is well founded in bacterial metabolism, but it may not be the best option for modeling some types of cancer. Furthermore, it is likely that this function will differ between different cell lines. More than a biomass reaction we will be probably looking for a cancer-specific metabolic task or objective function, as recently proposed in Agren et al. (2014) and Yizhak et al. (2014b).

Another possibility is that the reference network (from which our reconstructions are derived) misses several reactions that would allow us to capture more differences at the reconstruction level. It is known that the central metabolism and some major pathways are well studied and represented in these networks, but a big part of the real metabolic possibilities of the cells might still be missing. It is also important to note that this approach does not consider in any way the cell rearrangement that may follow a knock-out intervention, which could provide an explanation for some false-positive predictions.

It should also be noted that high-throughput gene essentiality experiments are far from trivial, and their analysis is also complicated. Hence, lists of essential genes based on the analysis of those experimental results would also likely need follow-up experiments to confirm them.

We would like to point out that the p-value of the KS-test does not seem a good strategy to test the relevance of the obtained list of essential genes as it may change depending on the number of obtained essential genes. This can be seen in the results obtained with Recon 1, where the number of obtained essential genes is higher and the p-values of the KS-test are lower. However, checking the proportion of obtained essential genes with a score below a defined threshold, it can be observed that it does not differ much from the proportion taking into account all the metabolic genes.

Despite these discrepancies between the predicted genes and the Achilles data, we were able to find a number of genes in GBM with a higher frequency in the samples than in the random networks which have been shown to effectively reduce its proliferation when targeted (Chen et al., 2014; Lee et al., 2003). We also found some additional genes with interesting references in the literature (Woo et al., 2010; Gliemroth et al., 2003; Bababeygy et al., 2009).

Constraint based modeling approaches have had a great success in the microbial metabolic modeling research. The success has been possible in part thanks to the great number of experiments and information that has been gathered in those organisms. As cancer cell lines are less well characterized than some microorganisms, data from projects like Achilles constitute a very valuable resource to boost the models' performance. Until now, data on the real essentiality of metabolic genes was scarce. Although high-throughput silencing experiment data is complex and may contain inaccuracies, it is our best chance to improve these models.

Overall, we believe that constraint based approaches hold great promise in the study of cancer metabolism. Notwithstanding, for its application towards the identification of essential genes, the technique needs to carefully consider the assumptions it makes and identify the correct formulation of the question at hand. Our framework exposes these issues and can be a key component to solve the final puzzle.

Chapter 5

Minimal Cut Sets through specific reaction knock-outs

In this chapter, we establish the theoretical basis to design new methods for the direct integration of different sources of experimental data (e.g. gene expression or gene silencing data) with metabolic networks aimed at the prediction of essential genes and synthetic lethals. Specifically, we develop a novel mathematical method able to directly evaluate which reactions should be eliminated (if necessary) together with another specific reaction selected by the user in order to completely eliminate the flux through a given target reaction, e.g. the biomass reaction. This method opens the door to new ways of integrating experimental data with metabolic networks in order to obtain information about essentiality without the need to contextualize the network.

5.1. Introduction

The concept of Minimal Cut Sets was introduced in Klamt and Gilles (2004) and refined in Klamt (2006). MCSs were defined as a minimal set of reactions whose removal would render the functioning of a given objective reaction impossible. Their relationship with Elementary Flux Modes and the dualization of the main problem was mentioned in Klamt (2006) and formally exploited in Ballerstein et al. (2012). With this theoretical breakthrough, it became possible to calculate MCSs using algorithms designed to calculate EFMs. For example, the K-shortest EFM enumeration algorithm (de Figueiredo et al., 2009) was used to enumerate MCSs in von Kamp and Klamt (2014).

Efficient calculation of EFMs has been an active area of research in the last years (de Figueiredo et al., 2009; Kaleta et al., 2009; Kamp and Schuster, 2006; Machado et al., 2012; Pey and Planes, 2014; Pey et al., 2015; Rezola et al., 2011; Terzer and Stelling, 2008; Urbanczik and Wagner, 2005; Quek and Nielsen, 2014). Most of the algorithms have focused the attention on calculating as many EFMs as possible. However, one may only be interested in some EFMs that fulfill some constraints. While, in theory, a valid approach would be to calculate all the EFMs and then filter them according to the desired criteria, in practice it is currently impossible in large networks, as the number of EFMs grows exponentially with network size. It is this question that was addressed in a recent paper by Pey and Planes (2014), where a Mixed Integer Linear Programming (MILP) formulation was introduced to directly calculate EFMs satisfying a desired set of constraints. For example, an EFM containing several reactions of interest can be now directly calculated.

For the metabolic modeling community, it will be very valuable to directly calculate MCSs of certain characteristics. For instance, a MCS involving a specific reaction knock-out can be used to determine if it is possible to couple growth with the synthesis of a specific product (Klamt and Mahadevan, 2015). In fact, strain design for the optimization of chemical production, such as biofuels (Erdrich et al., 2014), is one of the main applications of MCSs. It is not necessary to solve a MCS problem to know if the production of a chemical is coupled to growth but, in case it is not, a MCS suggests the modifications necessary to make the coupling a reality. Another possible application of MCSs in the field of human health involves finding a complementary target to an already druggable reaction. Here, MCSs would be viewed as synthetic lethals for treating cancer or other diseases (Folger et al., 2011; Frezza et al., 2011; Kaelin, 2005; Suthers et al., 2009).

In this chapter, we adapt the formulation in Pey and Planes (2014) to achieve this goal. In the process, we stress that not all the EFMs in the dual problem correspond to correct MCSs of the original network (Ballerstein et al., 2012), which implies that not all the algorithms designed for the calculation of EFMs can be directly applied to the calculation of MCSs without careful consideration. The modifications made to the algorithm in Pey and Planes (2014) are mathematically justified and bring interesting insights between the primal and the dual problem of the MCS computation.

5.2. Methods

A metabolic network of m metabolites and n reactions can be represented by an $m \times n$ stoichiometric matrix S, where each column represents a reaction with negative coefficients for the educts and positive coefficients for the products. The activity of the reactions is represented by the flux vector v. Under the steady-state assumption, the sum of fluxes that produce a compound are equal to the sum of fluxes that consume it (Eq. 5.1). Irreversible reactions can only carry positive fluxes (Eq. 5.2), while reversible reactions can carry negative and positive fluxes.

$$S \cdot v = 0 \tag{5.1}$$

$$v_i \geq 0, \qquad \forall i \in Irrev \tag{5.2}$$

We would like to be able to perform a given task, in our case, to carry flux through a given reaction (usually, this reaction is the biomass reaction):

$$t^T \cdot v \geq v^* \tag{5.3}$$

where t is a vector of all zeros except for a 1 in the position of the reaction we want to carry flux through, and v^* is the minimum amount of flux that the reaction should carry.

We are interested in a minimal group of reactions that, if blocked, would render this task impossible. To find those, we define the possible reaction knockout constraints we could impose (Eqs. 5.4 and 5.5).

$$v_i = 0, \qquad \forall i \in Rev \tag{5.4}$$

$$v_i \leq 0, \qquad \forall i \in Irrev \tag{5.5}$$

Note that for the knock-out of irreversible reactions we only limit their upper bound (Eq. 5.5), as their lower bound is already zero because of the irreversibility constraint (Eq. 5.2).

Following the theory in Ballerstein et al. (2012) and von Kamp and Klamt (2014), we formulate the dual problem of the infeasible primal problem defined by the previous constraints (Eqs. 5.1-5.5), to obtain a new feasible and unbounded problem (Eqs. 5.6-5.9).

$$N \cdot \begin{pmatrix} u \\ rp \\ rn \\ w \end{pmatrix} = \begin{bmatrix} S^T & I & -I & -t \end{bmatrix} \cdot \begin{pmatrix} u \\ rp \\ rn \\ w \end{pmatrix} = 0 \tag{5.6}$$

$$- v^* \cdot w \leq -c \tag{5.7}$$

$$rp \geq 0, rn \geq 0, w \geq 0, c > 0 \tag{5.8}$$

$$u \in R^m, rp \in R^n, rn \in R^n, w \in R \tag{5.9}$$

This dual problem can be viewed as a new stoichiometric matrix N with new flux variables u, rp, rn and w, with their respective reversibility constraints. When these variables take values different from zero, they indicate that their associated constraints in the primal problem are active. In particular, u variables are related to steady-state constraints, rp variables are related to constraints limiting the upper bound of a reaction, rn variables refer to constraints limiting the lower bound of a reaction and w variable is linked to Eq. (5.3). An MCS is an EFM in this dual problem that contains w and has minimal support in rp and rn, with the exception of rn variables related to the irreversibility constraints (Eq. 5.2), which do not count for the minimal support (Ballerstein et al., 2012).

Constraint in Eq. (5.7) forces w to have a value different from zero, meaning that its associated constraint (Eq. 5.3) must take part in the solution. This constraint is forcing flux through w, as c is a positive constant that rules out the trivial solution. If we want a MCS involving a specific reaction knock-out, we must also force flux through its related variable.

Recently, Pey and Planes (2014) formulated an optimization model to directly obtain EFMs fulfilling several biological constraints, such as carrying flux through a group of specific reactions. The key insight was to acknowledge in the problem formulation that an EFM has a single degree of freedom. When we have an EFM and the flux through one of its reactions is set, the rest of the fluxes of that EFM are automatically and univocally determined.

Certainly, if we add one reaction activation constraint to the steady-state constraint in a network composed of only irreversible reactions, the extreme points of the feasible region they define coincide with EFMs. In an n-dimensional space, extreme points lie on n linearly independent and binding constraints. They are mathematically represented as basic solutions, where variables are divided into basic and non-basic variables. The value of non-basic variables is set by their binding constraints (for example, a non-basic variables with a non-negative binding constraint will be set to zero). The value of basic variables results from solving the system of linear equations that is obtained once the value of non-basic variables has been fixed.

An EFM can always be described with the steady-state constraints and flux inactivation constraints for those reactions that are not part of the EFM, collectively defining $n - 1$ linearly independent and binding constraints. Once we add one flux activation constraint to this system, all the fluxes are fully determined. Hence, if we want to include a second flux activation constraint and obtain EFMs as solutions, we need to ensure that the new constraint is redundant with respect to the binding constraints that describe the EFM. Redundancy is guaranteed by forcing that this second flux activation constraint can be written as a linear combination of the rest of the linearly independent and binding constraints. The tricky part here is that we do not know *a priori* which reactions are active in the EFM and which ones are not. In Pey and Planes (2014), mixed-integer linear programming was used to select active reactions in the EFM and properly apply the linear combination requirement.

The direct application of the method described in Pey and Planes (2014) to our MCS problem would involve considering N as the initial network, and w and our knock-out related variable rp and rn the reactions we want to activate. Thus, we would make all the variables irreversible, splitting the u variables into two other irreversible variables and follow the implementation described in that work (see Appendix C for the full formulation of the optimization problem obtained following this strategy). The solutions provided by this model constitute Elementary Flux Modes in the network N. However, not all the EFMs in N correspond to MCSs in the original network S (Ballerstein et al., 2012) and, more importantly, not all the EFMs that contain our two target variables correspond to MCSs in the original network S (see Section 5.3 for a toy example illustrating these events).

Following that approach, we obtain solutions that have minimal support in u, rp, rn and w, but MCSs are defined as solutions with minimal support in rp and rn variables related to knock-outs of reversible reactions, that include w. Hence, MCSs correspond to a subset of all the possible EFMs. Interestingly, we realize that MCSs are closer to the concept of Generating Flux Modes (GFMs) than EFMs. GFMs are elements of a convex basis and have minimal support with respect to the set of irreversible reactions (Larhlimi and Bockmayr, 2009; Rezola et al., 2011). In our case, irreversible reactions are rp, rn and w; however, MCSs are minimal with respect to w, rp, and the subset of rn variables related to knock-outs of reversible reactions. This implies that not all the GFMs of the dual problem correspond to

MCSs of the original network (see Section 5.3 for a toy example illustrating this event).

In order to limit the solutions of the dual problem to proper MCSs, we introduce two main modifications. The first one is to force the linear combination constraint for all the columns related to u variables, regardless of their value being different to zero or not. This means that we no longer need to split the u variables into two irreversible variables. As a bonus, these variables no longer need related binary z variables, which are now only associated with rp, rn and w variables. The second modification consists in treating the x variables, which represent the coefficients of the linear combination, as if they were the reaction flux variables of the original problem and constraining their reversibility accordingly. We present below the full mathematical model:

$$\text{minimize} \sum_{i} zp_i + \sum_{i \in Rev} zn_i \tag{5.10}$$

subject to:

$$N \cdot \begin{pmatrix} u \\ rp \\ rn \\ w \end{pmatrix} = \begin{bmatrix} S^T & I & -I & -t \end{bmatrix} \cdot \begin{pmatrix} u \\ rp \\ rn \\ w \end{pmatrix} = 0 \tag{5.11}$$

$$- v^* \cdot w \leq -c \tag{5.12}$$

$$\begin{pmatrix} 0 & d_p & d_n & 0 \end{pmatrix} \cdot \begin{pmatrix} u \\ rp \\ rn \\ w \end{pmatrix} \geq b \tag{5.13}$$

$$\alpha \cdot \begin{pmatrix} zp \\ zn \\ zw \end{pmatrix} \leq \begin{pmatrix} rp \\ rn \\ w \end{pmatrix} \leq M \cdot \begin{pmatrix} zp \\ zn \\ zw \end{pmatrix} \tag{5.14}$$

$$zp_i + zn_i \leq 1, \qquad \forall i \in Rev \tag{5.15}$$

$$\begin{bmatrix} S & 0 \\ I & 0 \\ -I & 0 \\ -t^T & v^* \end{bmatrix} \cdot x = \begin{pmatrix} 0 \\ d_p + \epsilon_p - \delta_p \\ d_n + \epsilon_n - \delta_n \\ 0 + \epsilon_w - \delta_w \end{pmatrix} \tag{5.16}$$

$$M \cdot \begin{pmatrix} 1 - zp \\ 1 - zn \\ 1 - zw \end{pmatrix} \geq \begin{pmatrix} \epsilon_p + \delta_p \\ \epsilon_n + \delta_n \\ \epsilon_w + \delta_w \end{pmatrix} \tag{5.17}$$

$$\begin{pmatrix} rp \\ rn \\ w \end{pmatrix} \geq 0, \begin{pmatrix} \epsilon_p \\ \epsilon_n \\ \epsilon_w \end{pmatrix} \geq 0, \begin{pmatrix} \delta_p \\ \delta_n \\ \delta_w \end{pmatrix} \geq 0 \tag{5.18}$$

$$x_i \geq 0, \qquad \forall i \in Irrev \tag{5.19}$$

$$\begin{pmatrix} zp \\ zn \\ zw \end{pmatrix} \in \{0, 1\} \tag{5.20}$$

$$u \in R^m, rp \in R^n, rn \in R^n, w \in R \tag{5.21}$$

$$x \in R^{n+1} \tag{5.22}$$

$$\epsilon_p \in R^n, \epsilon_n \in R^n, \epsilon_w \in R \tag{5.23}$$

$$\delta_p \in R^n, \delta_n \in R^n, \delta_w \in R \tag{5.24}$$

$$c > 0, b > 0 \tag{5.25}$$

where α and M represent a sufficiently small and large constants, respectively, and dp and dn are vectors of all zeros except for a single 1 in the position related to the knock-out constraint that we want to activate. If the knock-out we want to enforce involves an irreversible reaction we will only set a 1 in dp and leave dn as a vector of all zeros. As in Pey and Planes (2014), variables ϵ_p, ϵ_n, ϵ_w, δ_p , δ_n and δ_w allow the linear combination constraint (Eq. 5.16) to be applied only to active variables with the help of variables zp, zn and zw and their linking constraints (Eqs. 5.14 and 5.17). The optimal solution of this optimization problem corresponds to a MCS that includes our desired reaction knock-out.

The formulation of the problem can be simplified (see Appendix C), but we have written it here in a way that the similarities and differences with the formulation in Pey and Planes (2014) are easier to spot with the aim to ease the understanding of the model.

The reasons why these modifications work have a mathematical underpinning that we proceed to explain in the following paragraphs.

In contrast with Pey and Planes (2014), u variables are not split into two irreversible steps and, therefore, the linear combination (Eq. 5.16) must always apply to them. In particular, since the value 0 is not a bound for u variables and they can take any real value, they should always be considered as basic variables and thus, they must take part in the linear combination constraint. With this modification alone, the algorithm goes from calculating EFMs to obtaining GFMs satisfying several constraints (see Appendix C). However, as noted above, this is not sufficient to obtain MCSs.

The justification for the second modification is somewhat more difficult to grasp. It has its roots in duality theory and it involves the correct manipulation of equality constraints in Eq. (5.11). Dual variables of equality constraints are always unrestricted sign variables, but dual variables of inequality constraints are always either non-negative or non-positive variables. In this context, and because we do not care about their value in the optimal solution, rn variables related to the irreversibility constraints are actually explicit excess variables in Eq. (5.11) and, in order to directly obtain MCSs, they can be removed without altering the solution space, transforming the equalities where they appear into greater than or equal inequalities (see Appendix C). This action has consequences in the linear combination requirement introduced in Pey and Planes (2014), here Eq. (5.16), as now the linear combination coefficient variables x cannot be of unrestricted sign type by default and must be non-negative in these inequalities (see Appendix C). In fact, those x variables are dual variables of the corresponding constraints and they must abide by duality theory. Eventually, we have that the x variables must obey the same irreversibility constraints that the r variables of the original problem do.

These modifications have a surprising consequence in the interpretability of the problem. First, x variables can be assimilated to the flux variables of the original problem and, second, we are forcing $S \cdot x = 0$, which assimilates to the steady-state constraint of the original model. This means that the values of the x variables in a valid solution of our direct MCS problem formulation can be interpreted as a valid flux distribution in the original network.

Finally, we can iteratively enumerate MCSs fulfilling our conditions by introducing a new constraint that eliminates previously obtained solutions (de Figueiredo et al., 2009; von Kamp and Klamt, 2014).

$$\sum_{i} zp_i^k \cdot zp_i + \sum_{i \in Rev} zn_i^k \cdot zn_i \leq (\sum_{i} zp_i^k + \sum_{i \in Rev} zn_i^k) - 1, \qquad k = 1, ..., K \tag{5.26}$$

5.3. Results

In this section, we first introduce a toy example where it can be seen that not every EFM in the dual network corresponds to a MCS in the primal network (Ballerstein et al., 2012) and check that our algorithm is capable of correctly identifying MCSs for each reaction. Then, we apply our algorithm to two networks of different sizes, namely the *E. coli* core metabolic network and the iAF1260 *E. coli* model, comparing the behaviour of our algorithm with that of the enumeration approach presented in von Kamp and Klamt (2014). The model was implemented in Matlab, using CPLEX as the underlying optimization software. The computations were carried out on a 64 bit Intel Xeon E5-1620 v2 at 2.70 GHz (4 cores) and 16 GB of RAM.

5.3.1. Toy example

Figure 5.1 shows a primal toy network and the dual network arising from its corresponding MCS problem. This network has only one reversible reaction (v_2) and we will consider v_8 as the target reaction. The network includes an irreversible reaction, v_9, that is not related in any way to the target reaction v_8, hence, there is no MCSs containing v_9 that blocks the activity of v_8. However, there exists an EFM containing the input reaction for node v_9 in the dual network (e.g. rp_9, rp_6, rn_5, u_B), but that EFM does not contain our target related variable w.

Now we draw the attention to v_6. This reaction participates in a MCS with v_2 and v_4. It is easy to see that the dual network contains an EFM composed of the inputs to v_2, v_4 and v_6 (rp_2, rp_4 and rp_6 respectively), the arcs u_C, u_D and u_E and the target w. However, there is also another EFM that contains the input to node v_6 (rp_6) together with the input to node v_1 (rp_1), the output to node v_5 (rn_5), the arcs u_A, u_C, u_D and u_E and the target w. This EFM would correspond to a MCS composed of v_1 and v_6, which is incorrect because v_1 is already a MCS on its own. It turns out that this second EFM involves less knock-out related dual variables than the former (2 vs. 3, because the output of node v_5, rn_5, does not count), which implies that the direct application of the formulation in Pey and Planes (2014) to the dual network asking for the shortest EFM containing w and

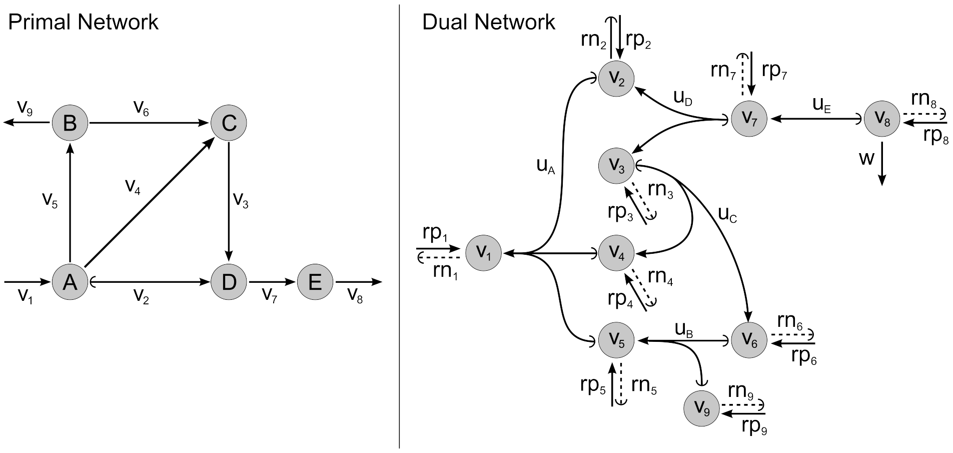


Figure 5.1: Example metabolic network illustrating an original network in the primal and the associated dual network of its MCS problem. The original network only has one reversible reaction, v_2. The target reaction we want to block the flux through is v_8. The external input reactions to the reaction nodes in the dual network correspond to rp variables and the external output reactions to rn variables. MCSs of this network are: $\{v_1\}$, $\{v_7\}$, $\{v_2, v_3\}$, $\{v_2, v_4, v_5\}$ and $\{v_2, v_4, v_6\}$. Two dual EFMs of v_1 knock-out that include w are $\{rp_1, rn_9, u_A, u_B, u_C, u_D, u_E, w\}$ and $\{rp_1, rp_6, rn_5, u_A, u_C, u_D, u_E, w\}$. The former corresponds to a MCS, while the latter does not.

the input to v_6 would obtain this EFM first. This invalidates the direct use of their formulation for obtaining MCSs involving a specific reaction knock-out.

However, the new model introduced in this work correctly captures the MCS, as the incorrect MCS solution is infeasible for it. In particular, the values of the x variables (corresponding to the coefficients of the linear combination in Eq. 5.16) in the case of the invalid EFM are all non-negative except for x_9, which has a negative value that our new formulation forbids. We can assimilate the values of the x variables to a flux distribution in the primal network, resulting in reactions v_6, v_3, v_7 and v_8 carrying flux in the forward direction and reaction v_9 carrying flux in the backward direction. Obviously, this flux distribution is not a valid one, because v_9 is irreversible. Our new model takes this into account and renders this solution infeasible.

Because of the way that MCSs are enumerated using k-shortest in von Kamp and Klamt (2014), from smallest to largest and taking into account only variables related to knock-out constraints, EFMs in the dual network that are not MCSs in the primal network are avoided. In the case of the network in Figure 5.1, the EFM corresponding to the MCS that contains only v_1 is obtained before the invalid EFM that contains v_1 and v_6. Hence, the solution enumeration constraint avoids the calculation of the invalid

EFM, as it contains an already calculated MCS, v_1.

5.3.2. Application to *E. coli* core metabolism

Here, we apply our algorithm to the *E. coli* core network available in the COBRA Toolbox (Schellenberger et al., 2011). This network contains 72 metabolites and 95 reactions. We conducted structural analysis of the reported network, so we did not take into account possible growth medium constraints (e.g. glucose and oxygen supply) or compound production requirements (e.g. ATP maintenance).

As with any other MILP, the time needed to find the optimal solution may widely vary depending on the chosen reaction. This disparity in problem difficulty can already be experienced in this simple network. Taking the biomass reaction as our target for the MCSs, we calculated one MCS for each reaction. We used the reduced formulation, presented in Appendix C, which deals with a lower number of variables than the one introduced in the Section 5.2. The median time to obtain a solution was 0.08, but a small number of reactions required more than 15 seconds to reach the optimal solution. All the MCSs were obtained in less than 60 seconds nonetheless.

We also enumerated 1000 MCSs using a custom implementation of the approach described in von Kamp and Klamt (2014). The median time in this case was 0.28 seconds, but it steadily increased as the number of reactions participating in the MCSs increased (the median time needed to obtain the last 200 MCSs was 0.80 seconds).

If one is interested in collecting many short MCSs, the enumeration approach could be more appropriate than our method. However, if one is interested in the MCSs in which a particular reaction participates, our approach will be more suited. Among the 1000 MCSs calculated with the enumeration approach, there were 18 MCSs involving only 1 reaction, 111 involving 2, 223 involving 3, 396 involving 4, and 252 involving 5. On the other hand, when asking one MCSs for each of the 95 reactions in the *E. coli* core metabolic network, 18 where MCSs on their own, 36 participated in a 2 reaction MCS, 11 in a 3 reaction MCSs, 7 in a 4 reaction MCSs, 10 in a 5 reaction MCS, 2 in a 6 reaction MCS, and 3 in a 8 reaction MCS, while 8 reactions did not participate in any MCS. If our knock-out of interest only participates in high-order MCSs, with the enumeration approach we would have needed a lot of time to obtain it. If it does not participate in any MCS, we would need to enumerate all the possible MCSs to find it out. Our method clearly poses an advantage in these situations.

5.3.3. Application to *E. coli* iAF1260 metabolic network

The *E. coli* iAF1260 (Feist et al., 2007), available for download at the BiGG database (Schellenberger et al., 2010), contains 1668 metabolites and 2382 reactions. As before, structural analysis of the reported network was conducted. We calculate 1000 MCSs with the enumeration approach (von Kamp and Klamt, 2014) and one MCS for each reaction with the reduced formulation of our method (2382 MCSs in total). These MCSs are all related to the biomass reaction. In this case, we only allowed our algorithm a maximum of 5 minutes to find each MCS.

The increased complexity of a genome-scale model evidenced numerical difficulties on the implementation of the MILP models (see Appendix C). Using a custom implementation of the enumeration approach, 42 % of the calculated 1000 MCSs were not actual MCSs. With our method, focusing only in those solutions that were obtained in less than the 5 minute time limit (1508 out of 2382), only 6.76 % were not correct MCSs (but 61.14 % did not have any associated MCS, leaving only 32.1 % MCSs). The *E. coli* core network is already capable of showing incorrectly calculated MCSs in some cases, which can be shown to be artifacts of limited computational precision and not due to a flaw in the theory of MCSs enumeration in von Kamp and Klamt (2014) or in our method. Note however that devising an implementation that can broadly deal with these difficulties is out of the scope of this Chapter and will be addressed in the future.

The enumeration approach yielded MCSs with a maximum of 3 reaction knock-outs, while our approach was capable of discovering MCSs involving up to 5 reaction knock-outs among the solutions returned in less than 5 minutes. More importantly, the size of the MCSs can go as far as to include 121 reactions if one analyzes the solutions returned by the algorithm after hitting the 5 minute time limit that correspond to correct MCSs. These solutions are not guaranteed to be optimal and thus they may not be correct MCSs (although some of the reactions included in that solution will form a proper MCS) or a guarantee that the reaction has no associated MCS if no solution was returned. However, 59.84 % of the solutions returned after hitting the time limit turned out to be correct MCSs (17.73 % of them contained no MCS and 22.43 % were incorrect MCSs). A histogram of the sizes of these MCSs is shown in Figure 5.2.

The median time to obtain each one of the 1000 MCSs using the enumeration approach was 22.5 seconds. In contrast, the median time to obtain a solution with our method was 15.4 seconds. Focusing only on solutions obtained in less than 5 minutes, the median time was 18.5 seconds if the reaction of interest participated in a MCS and 0.05 seconds if it participated

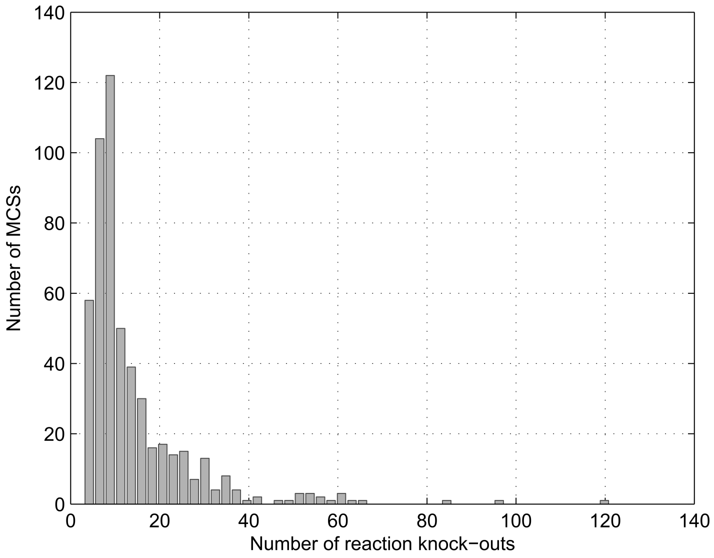


Figure 5.2: Histogram of the number of reaction knock-outs participating in the MCSs returned by the direct MCS calculation algorithm when hitting the time limit on the *E. coli* iAF1260 metabolic network case study.

in none. This case shows that our algorithm is applicable to the genome-scale setting.

5.4. Applications

We envision several applications for the method presented here. In this section, we outline some of them.

In metabolic engineering, it can be used to determine what modifications are necessary to couple the production of a desired compound with the growth of the organism. Note that, in order to know if the production of a compound is coupled to the biomass production, it is not necessary to calculate a MCS that involves the production reaction but, if the answer is negative, calculating such a MCS will suggest modifications to enforce the coupling. If the production reaction goes with some other reactions in a MCS for the production of biomass, blocking those other reactions will force the coupling between the production reaction and the biomass production, since if the production reaction happens to have zero flux, the MCS will be completed and no biomass will be possible.

This method can also aid in metabolic network curation. If information about experimentally obtained essential genes do not match the predictions made by the network, the method will suggest the elimination of reactions that will make the experimentally observed genes also essential in the model.

In the field of human health, if we have a drug that we know limits the activity of a specific reaction, we can search for other reactions that we can also limit in order to strengthen its effect. In other words, we can use MCSs to find synergistic targets of an already druggable reaction.

Finally, and the idea that originally motivated this development, is to use MCSs to determine the essentiality potential of a reaction given experimental evidence of the reactions in the network. Until now, as we have seen in this thesis, in the framework of constraint based modeling, gene essentiality is assessed after network contextualization. Our aim is to find a way to circumvent the need to contextualize the network, so that possible alternative networks to the contextualization problems do not bias our results. The hypothesis is that, if the genes coding the reactions that go in the MCS of a specific reaction have low expression values, that reaction might be more essential than if the expression of the accompanying reactions is higher.

To do this, we can weight the objective function of our method with scores derived from gene expression data or other experimental sources (Eq. 5.27).

$$\text{minimize} \sum_{i} w_i \cdot zp_i + \sum_{i \in Rev} w_i \cdot zn_i \tag{5.27}$$

Obtaining essentiality potential scores for each reaction in the network would allow us to rank them and compare the ranking to experimental gene essentiality data, such as the one provided by project Achilles (Cowley et al., 2014). Moreover, given the importance of the biomass reaction definition for an accurate gene essentiality prediction using constraint based modeling techniques, we can use known essential genes and the essentiality potential approach outlined in this section to test and find an appropriate biomass function.

5.5. Conclusions

In this work, we have introduced an optimization model to calculate MCSs involving a specific reaction knock-out. We have emphasized that not all the EFMs in the dual problem correspond to valid MCSs in the primal problem (Ballerstein et al., 2012). Also, we acknowledge that MCSs are closer to the definition of GFMs of the dual problem than to EFMs, although not all the GFMs in the dual problem correspond to valid MCSs in the primal problem either. As a side effect, this work extends the formulation in Pey and Planes (2014) to the calculation of GFMs satisfying several constraints and deepens in the mathematical understanding of the model.

This formulation is important because it makes possible to find if each reaction can participate in a MCS or not. Without it, the only possible way would be to enumerate all the MCSs (using the method in von Kamp and Klamt (2014), for example) until one incorporating the reaction of interest is found. This may become impractical, as the enumeration process is currently limited by memory and time constraints. Given that some reactions may participate in very high order MCSs, this drawback is evident. Our formulation can also be used to enumerate those MCSs that include a particular reaction knock-out.

Looking at the model and the solution a bit more in depth, we can realize that, whenever the model chooses to activate a particular reaction knock-out, the x variable associated to that reaction that has been knocked-out gets equal to zero. This is in line with interpreting the x variables as fluxes in the primal network: if the reaction has been knocked-out it cannot carry any flux. As an exception, the x variable associated to the reaction whose knock-out we are interested in gets a fixed value different from zero. Thus, all the possible knock-outs of reactions that participate in the flux distribution (x value different from zero) are discarded, except for the one we were interested in. Then, the algorithm chooses the minimal number of knock-outs that make all the other remaining knock-out constraints redundant.

Regarding non-optimal solutions to the MCS problem formulation, if we stop the solution process before reaching optimality, the last feasible solution calculated will not necessarily be a MCS, although it will contain one. The MCS contained in that solution will not be necessarily optimal according to our objective function either. It may exist another MCS involving less reaction knock-outs than the MCS contained in the non-optimal solution.

Future work will include considering how to deal with more general reaction bounds as well as considering multiple reaction knock-out or gene knock-out constraints. The formulation can also be easily extended to consider the calculation of MCSs including several specific reaction knock-outs constraints. Importantly, we will adapt the approach to incorporate experimental data and use it to rank reactions according to an essentiality potential score.

In summary, we believe that this work will boost new research in metabolic engineering around MCSs, including its applications in strain design and human health. Moreover, the insights brought by our formulation could lead to more efficient algorithms for the calculation of MCSs.

Chapter 6

Conclusion

This brief chapter provides a summary of the work presented in this thesis, highlights its main conclusions and discusses future lines of research.

The research carried out in this thesis has been focused on the development of new constraint-based modeling methods that would integrate experimental data with prior knowledge about metabolism. The goal was to gain new insights on the data and improve the prediction of targets for the treatment of cancer. We have developed two new fast metabolic network contextualization algorithms and a new algorithm capable of identifying Minimal Cut Sets that involve a specific reaction knock-out. All these three contributions open up new possibilities of analysis, some of which exemplified in this thesis.

After giving an overview of different metabolic network contextualization approaches, we have introduced a novel fast reconstruction method in Chapter 2. As other reconstruction algorithms, ours was motivated by some specific needs: speed and FBA compatibility of the contextualized network. This new reconstruction algorithm is conceptually similar to iMAT (Shlomi et al., 2008) in the way it reasons about the inclusion or exclusion of each reaction depending on its experimental evidence, and achieves similar speed performance to FastCore (Vlassis et al., 2014), a recently published fast reconstruction algorithm. Additionally, the use of Barcode (McCall et al., 2011) for processing gene expression data allowed us to use our algorithm for deriving one contextualized network from each sample we had access to, while other algorithms generally summarize the data from different samples and build a single network for a group of samples. This fast reconstruction algorithm has made possible the study described in Chapter 4.

In Chapter 3 we have introduced a new metabolic network reconstruction method focused on metaproteomic data and geared towards the reconstruction of bacterial communities. We have applied this method to gain insight into the data obtained from two different naphthalene-enriched bacterial communities. One of the most interesting insights gained by our approach was the discovery of only one of the communities being able to metabolize geraniol while the other was the only one able to metabolize fluorobenzoate. This observation was experimentally validated after noticing it when analyzing the contextualized networks using KEGG pathways, showing that network-based methods present a promising strategy for exploiting the value of experimental data. Given the recent interest on gut microbiota and the study of other bacterial communities in their association with human health, we anticipate this and similar approaches to have a profound impact on future discoveries.

One of the core questions we wanted to address in this thesis has been covered in Chapter 4. There, we have evaluated the accuracy of the results obtained with the FBA based Gene Essentiality Analysis methodology when networks contextualized with cancer gene expression data are used, and compared it to the results obtained from networks contextualized with random expression data. Key to this study has been the network contextualization algorithm introduced in Chapter 2 and the availability of high-throughput gene essentiality experimental data. The results showed that the in-silico obtained essential genes did not have a satisfactory degree of agreement with the high-throughput experimental data. We pointed out to the definition of the biomass reaction as critical in obtaining accurate results. Similarly, we emphasized the need of taking into account cellular rearrangement, which may be neglected when network contextualization algorithms are used in the pipeline. However, as analyzing high-throughput essentiality data presents its own challenges, we also searched in the literature some of the obtained genes, and found some of them to be indeed related with essentiality. These findings, as well as the successful study in Chapter 3, maintains this constraint-based modeling approach as a valuable tool for the study of cancer metabolism and the search of new therapeutic targets that still needs some development in order to be fully functional.

Finally, in Chapter 5 we have developed a new formulation to obtain Minimal Cut Sets that involve a specific reaction knock-out. The new insights obtained in the process, such as the relation between the MCS solution and a related flux distribution that is obtained at the same time, will allow to have a better understanding of the problem and lead to new and better algorithms to solve it. Despite the very theoretical nature of the chapter, the method opens up new possibilities to the integration of experimental data with metabolic networks in order to characterize the importance of each reaction with respect to a given phenotypic objective (e.g. cellular proliferation in the case of cancer) without the need of network contextualization.

6.1. Future lines

The contributions of this thesis pave the way to further developments in constraint-based modeling analyses. On the one hand, it will be interesting to use the Bacterial Community Reconstruction algorithm for analyzing data coming from different samples of gut microbiota, where, given the elevated number of organisms, a meta-network approach is currently the most sensible option. On the other hand, the Fast Reconstruction algorithm in Chapter 2 has been used only with gene expression data, but it can be easily adapted to handle RNA-Seq data, which is becoming more prevalent nowadays. In addition, the fast nature of the reconstruction algorithm would allow us to extend the FBA based Gene Essentiality Analysis done in Chapter 4 to other candidate biomass reactions, including the definition of reactions that describe a given metabolic task different from growth. The comparison of the results to the high-throughput gene essentiality experiments would inform what biomass definitions better represent the cancer metabolic phenotype.

Finally, the method for the direct calculation of MCSs involving a specific reaction knock-out presented in Chapter 5 provides a solid mathematical ground for the development of new approaches, necessary because of the complexity of the problem and its potential applications. Among all the possible applications for the method outlined in Chapter 5, we would like to highlight the one involving the integration of expression data. This idea proposes to rank the reactions in a network according to an essentiality potential score obtained by weighting the objective function of the problem with experimental data. The hypothesis is that, if the genes coding the reactions that go in the MCS of a specific reaction have low expression values, that reaction might be more essential than another one for which the reactions that go with it in the MCS have a higher expression. Furthermore, we could use the information on known essential and non-essential genes to design or test a biomass reaction definition that leads to results that agree with the experimental observations. Adapting current algorithms for this task, given the increasing size of human genome-scale metabolic networks and numerical issues reported in Chapter 5, will be a necessary research direction in the future.

In conclusion, with the growing availability of heterogeneous –omics data, extending the mathematical models presented in this doctoral thesis to make a synergistic use of them will constitute a major goal in the future. This is particularly relevant for understanding molecular basis of complex diseases, such as cancer, where a complete understanding requires multi-omics approaches.

Appendix A

iMAT modification

As discussed in Chapter 2, there are several proposals of reconstruction algorithms in the literature, like MBA (Jerby et al., 2010), MIRAGE (Vitkin and Shlomi, 2012), GIMME (Becker and Palsson, 2008), GIM^{3}E (Schmidt et al., 2013), INIT (Agren et al., 2012), iMAT (Shlomi et al., 2008), MADE (Jensen and Papin, 2011). Most of these algorithms rely on Mixed Integer Linear Programming (MILP) in order to select the active reactions for the contextualized reconstruction according to some predefined optimality criteria. Usually, each reaction is assigned a score as to how likely it is to be present in the reconstruction under consideration. This score can be obtained from genomics, transcriptomics, proteomics or other sources of data, or even from a combination of some or all of them. All the reconstruction methods have their place as some may be better suited for the integration of one type of data than others. Likewise, the results obtained from each one of them are not easily comparable, as each one aims for slightly different things.

In our case, our algorithm is closest (in the way it treats the inclusion of reactions) to iMAT (Shlomi et al., 2008). The original optimization model proposed by iMAT for network reconstruction is described in Chapter 2 by Eqs. (2.1)-(2.8).

These optimization problem aims to strike a balance between the inclusion of H reactions and the exclusion of L reactions. It does not directly consider the inclusion of reactions that are not H or L. Our algorithm (Section 2.3) includes a term to control the inclusion of those reactions (M set) and promote compact solutions.

Aware of the possible existence of alternative solutions, iMAT proposes an iterative solution scheme to assign a confidence score for the inclusion or exclusion of each reaction. This step, however, is not compulsory for our task, and we will only solve the optimization problem once.

iMAT does not require a definition for the growth medium nor the specification of a biomass function, although they can be included into the formulation if desired (Shlomi et al., 2008). When comparing iMAT to our algorithm, we will set the same medium conditions and ask for a minimum biomass production in order to have them in the same conditions. Another modification we will introduce to iMAT is in the definition of ϵ, which will be selected depending on the reaction bounds, as we do in our algorithm.

$$\max_{v,y^+,y^-} \Big(\sum_{i \in R_H} (y_i^+ + y_i^-) + \sum_{i \in R_L} y_i^+\Big) \tag{A.1}$$

subject to:

$$S \cdot v = 0 \tag{A.2}$$

$$v_{min} \leq v \leq v_{max} \tag{A.3}$$

$$v_{biomass} \geq v_{biomass}^* \tag{A.4}$$

$$v_i + y_i^+ (v_{min,i} - \epsilon) \geq v_{min,i} \qquad i \in R_H \tag{A.5}$$

$$v_i + y_i^- (v_{max,i} + \epsilon) \leq v_{max,i} \qquad i \in R_H \tag{A.6}$$

$$v_{min,i}(1 - y_i^+) \leq v_i \leq v_{max,i}(1 - y_i^+) \qquad i \in R_L \tag{A.7}$$

$$v \in R^m \tag{A.8}$$

$$y_i^+, y_i^- \in [0, 1] \tag{A.9}$$

$$\epsilon_{v_{max},i} = \delta \cdot |v_{max}| \tag{A.10}$$

$$\epsilon_{v_{min},i} = \delta \cdot |v_{min}| \tag{A.11}$$

We select $\delta = 0{,}10$, the same value we use in our algorithm. When retrieving the solution, we consider as active any reaction with flux greater than 10^{-8}. If ϵ is lower than 10^{-8}, we set it to that quantity instead.

We use Cplex to solve the optimization problem. In addition, we exit the optimization process when the relative optimality gap is below 0.5 % (gap between the relaxed problem solution and the best integer solution found), as closing the gap completely can be extremely memory and time consuming and adds little to the solution quality.

Appendix B

Supplementary Material for Reconstruction of a naphthalene-degrading bacterial community

This appendix gives additional details on the reconstruction model introduced in Section 3.2 and its application on the reconstruction of a naphthalene-degrading bacterial community analyzed in Chapter 3.

B.1. Objective function weights

Penalty (p_i) and bonus (b_i) terms in Eq. (3.4) of Section 3.2 are the sum of various concepts. All of the reactions in the reference database start from a penalty value of 10. Reactions in set D are penalised with 90 units. In addition, as we treat all reactions as being potentially reversible, the reconstruction algorithm is able to use irreversible reactions in the less favourable annotated direction; however, this choice implies a penalty of 1000. Penalised reversibility changes were also used in the Model SEED approach (Henry et al., 2010). Finally, a penalty of 10000 is given to reactions in the lowly likely set (set L). We do not force their flux to be zero because their absence is always hypothesised with a certain p-value, and with very low probability, they may therefore be active. It should be noted that we also add a small penalty (5) to exchange reactions and transporters that do not act via diffusion.

Concerning bonus terms, reactions in the highly likely set (H set) receive a bonus of 9. We also add a small bonus (5) for reactions defined as

Table B.1: Summary of the penalty and bonus used in Eq. (3.4).

Penalty (p_i)	**Value**	**Description**
$i \in R$	10	Default penalty for all reactions
$i \in D$	90	Penalty for using reactions in D
$i \in Irr_b$	1000	Penalty for using an irreversible reaction in the backward direction
$i \in L$	10000	Penalty for using reactions from the lowly likely set
Bonus (b_i)	**Value**	**Description**
$i \in H$	9	Bonus for using reactions from the highly likely set

spontaneous or described as being guided by diffusion.

Finally, as in the Model SEED approach (Henry et al., 2010), we take into account the coverage of experimental data in annotated modules, e.g., from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2012). These modules group a small number of reactions into common metabolic functions and can be used to provide additional information for the reconstruction. In particular, completion of modules that are well represented with metaproteomic data is favoured during the reconstruction process, which is realised by providing a bonus to the reactions in D equal to 90 multiplied by the average coverage of its related modules. Module coverage is determined by the percentage of reactions from the H set that are active in the module.

A summary of the penalty and bonus terms can be found in Table B.1. In particular, the weights for fluxes in Eq. (3.4) are roughly as follows: $(p_i - b_i) \approx 1$ for reactions in H; $(p_i - b_i) \approx 10$ for reactions in M; $(p_i - b_i) \approx 100$ for reactions in D; $(p_i - b_i) \approx 1000$ for backward irreversible reactions; and $(p_i - b_i) \approx 10000$ for reactions in L.

B.2. Reaction classification example

In order to obtain a context-specific metabolic network for our metaproteomic data, we need a collection of reactions from which to build the draft network. All those reactions are automatically assigned to the set D. Then, according to the data we have measured, that classification may change.

Assume we have the data in Table B.2. Each enzyme has been identified with a different value in each of the samples. In addition, the enzymes have been assigned to some organism taxonomy in each sample. It can be

Table B.2: Toy example of metaproteomic data.

Enzyme	CN1	CN2	Taxonomy CN1	Taxonomy CN2
E1	1.1	0.9	Organism1	Organism1
E2	0.7	0	Organism2	None
E3	2	1.2	Organism3	Organism2

observed that enzymes both E1 and E3 have measures in CN1 and CN2, so the reactions associated to those enzymes are assigned to the set H in both scenarios. On the other hand, E2 has only been identified in CN1, thus it is classified as H in CN1 but as L in CN2. We can also see that E3 has a value more than 1.5 times higher in CN1 than in CN2, so E3 is also assigned to the set K in CN1, but not in CN2. Lastly, we can observe that the enzymes identified in CN1 have been assigned to Organism1, Organism2 and Organism3, while the enzymes identified in CN2 have been assigned to Organism1 and Organism2. With this information, we check the genome annotation of those organisms and see which reactions can be associated to the enzymes encoded in them. Reactions that are still classified as D but can be found in the genomes of Organism1, Organism2 and Organism3 will be assigned to the set M in CN1, while only the ones found in genomes of Organism1 and Organism2 will be assigned to M in CN2.

B.3. Genome information

As noted in Chapter 3, using full-length and partial 16S rRNA gene sequences obtained through a metagenomic approach (Guazzaroni et al., 2013), 13 and 12 distinct species were found to constitute the CN1 and CN2 communities, respectively. This information was used to refine the network reconstruction process. In particular, we downloaded a set of related genome annotations from the KEGG website for CN1 and CN2, established on the basis of phylogenetic affiliations (Guazzaroni et al., 2013). Full details can be observed in Table B.3 and Table B.4.

B.4. Culture medium

Full details for the minimal medium based on naphthalene used in Chapter 3 can be found in Table B.5. The maximum input flux for these compounds was set to 100 units.

Table B.3: KEGG genomes used to help in the reconstruction of CN1.

Taxonomy	**KEGG abbreviation**
Achromobacter	axy
Azospirillum	azl ali abs
Comamonas	ctt
Mesorhizobium	mlo mci mop mam
Microbacterium	mts
Planctomycetes	rba psl plm pbs ipa phm
Pseudoxanthomonas	psu psd
Singulisphaera acidiphila	saci

Table B.4: KEGG genomes used to help in the reconstruction of CN2.

Taxonomy	**KEGG abbreviation**
Achromobacter	axy
Acidovorax	aav ajs dia aaa ack
Azospirillum	azl ali abs
Pseudomonas stutzeri GV 1	psa psz psr psc psj psh
Pseudomonas sp.	ppuu

Table B.5: Allowed medium for the reconstruction process.

SEED ID	Name
cpd00618	Naphthalene
cpd00254[e]	Mg^{2+}
cpd00048[e]	Sulfate
cpd00209[e]	Nitrate
cpd00013[e]	NH_3
cpd10516[e]	Fe^{3+}
cpd00099[e]	Cl^-
cpd00063[e]	Ca^{2+}
cpd00205[e]	K^+
cpd00009[e]	Phosphate
cpd00001[e]	H_2O
cpd00007[e]	O_2

B.5. CN1 and CN2 pathway examination

We used KEGG pathways to analyze CN1 and CN2 reconstructed metabolic networks. KEGG offers the possibility to colour reactions using KEGG reaction identifiers or EC numbers. We decided to use KEGG reaction identifiers, since KEGG is not an organism specific database and, therefore, ECs may involve more reactions than required for the reconstruction (Thiele and Palsson, 2010). It should be noted that as CN1 and CN2 reconstructions were based on the Model SEED database, reactions need to be first translated into their corresponding KEGG reaction identifiers.

As noted in Chapter 3, reactions in H that were included in the reconstruction were coloured red; reactions in M that were included in the reconstruction were coloured green; reactions in D that were included in the reconstruction were coloured blue; and reactions in L that were included in the reconstruction were coloured grey. In addition, metabolites can also be coloured using KEGG compound identifiers. We coloured all the compounds included in the reconstructed network green.

It should be noted that, despite the advantages of using KEGG maps, the colouring process is not straightforward and may include some mistakes. The reason is mainly because an EC number is potentially related to more than one reaction. Therefore, different results can be obtained colouring the pathway using EC numbers or KEGG reaction identifiers.

As an illustration of this process, Figure B.1 shows the KEGG pathway "Histidine metabolism" in CN1 and CN2.

As detailed in Chapter 3, we compared the KEGG maps in CN1 and

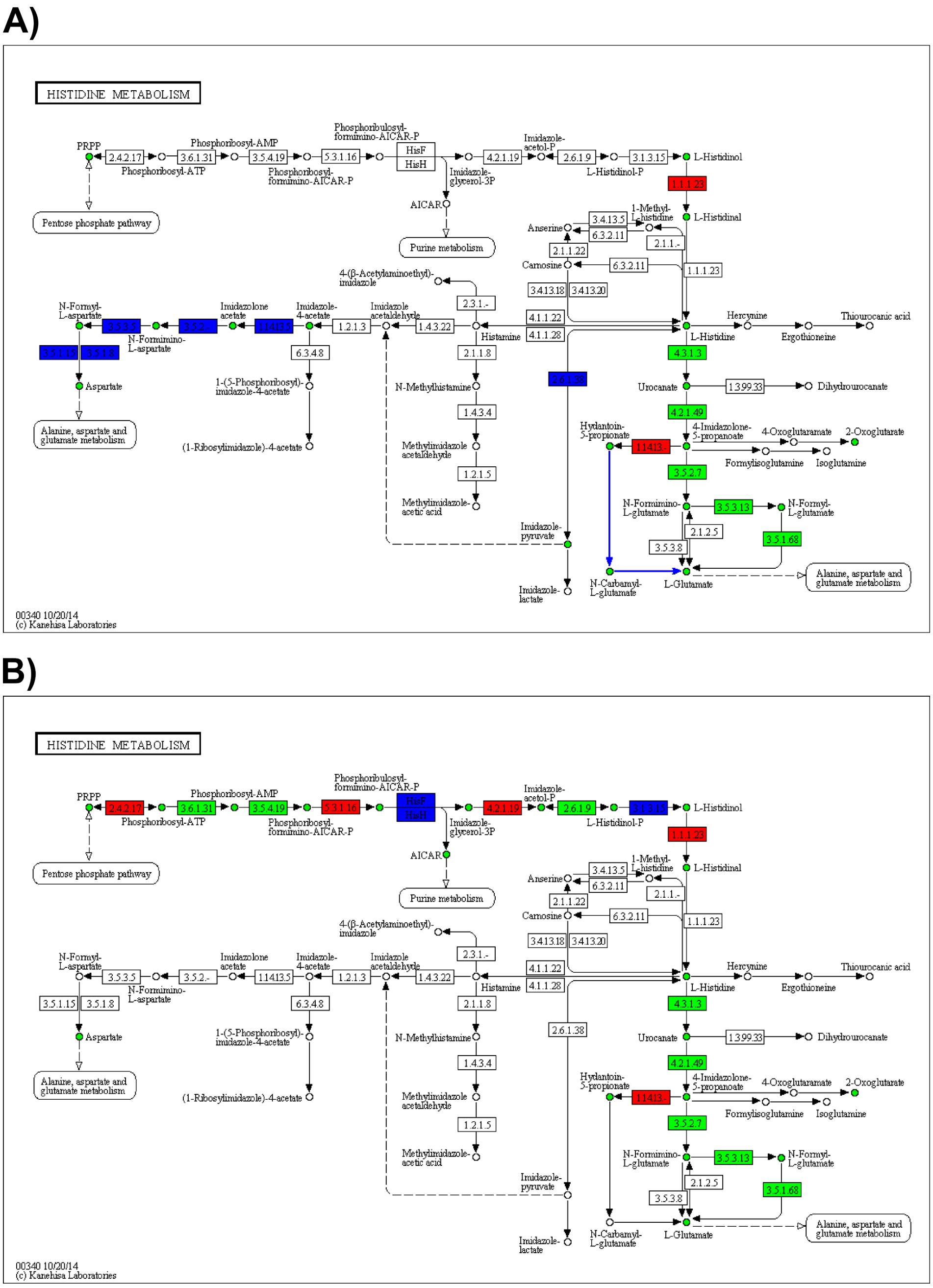


Figure B.1: The KEGG histidine metabolism pathway in A) CN1 and B) CN2.

Table B.6: Ranking of the KEGG pathways prior to and after reconstruction using metaproteomic data for CN1 and CN2. Note: "NaN" implies that no reactions from a map have been included in the CN1 and CN2 metabolic network, and therefore, the score is not applicable

KEGGID	Name	CN1	CN2	Score	Rank	Rank after
map00061	Fatty acid biosynthesis	15	33	9.81818	1	58
map00130	Ubiquinone and other terpenoid-quinone biosynthesis	5	17	8.47059	2	73
map01040	Biosynthesis of unsaturated fatty acids	5	17	8.47059	3	13
map00905	Brassinosteroid biosynthesis	5	15	6.66667	4	NaN
map00230	Purine metabolism	29	46	6.28261	5	6
map00240	Pyrimidine metabolism	17	27	5.7931	6	17
map00040	Pentose and glucuronate interconversions	0	5	5	7	55
map00680	Methane metabolism	13	23	4.34783	8	16
map00402	Benzoxazinoid biosynthesis	1	6	4.16667	9	87
map00480	Glutathione metabolism	8	14	3.73333	10	40
map00670	One carbon pool by folate	5	11	3.27273	11	24
map00030	Pentose phosphate pathway	8	10	3.07692	12	42
map00290	Valine, leucine and isoleucine biosynthesis	9	16	3.0625	13	46
map00400	Phenylalanine, tyrosine and tryptophan biosynthesis	6	12	3	14	9
map00521	Streptomycin biosynthesis	0	3	3	15	25

CN2 using a score, J_p, derived from the Jaccard distance (Eqs. 3.10-3.12). For illustration, in the case of the KEGG pathway "Histidine metabolism" shown in Figure B.1, the total number of reactions that are active in the CN1 and CN2 reconstructions is 22, with only 7 being common to both. Hence, the corresponding Jaccard index is $\sim$ 0,32, and the Jaccard distance is $\sim$ 0,68. The number of reactions in CN1 that are not present in CN2 is 7, while the number present in CN2, but not in CN1, is 8, and, thus, the J_p score assigned to this pathway is $8 \cdot (\sim 0{,}68) =\sim 5{,}45$.

The same analysis was conducted for each KEGG map in CN1 and CN2. For completeness, we also analysed differences prior to the reconstruction, only considering and mapping the metaproteomic data from CN1 and CN2 onto the KEGG maps. Table B.6 and Table B.7 show the top ranking KEGG pathways before and after the reconstruction of both CN1 and CN2. Full details can be found in Supplementary Material III of Tobalina et al. (2015).

In Table B.6 and Table B.7, the CN1 and CN2 columns indicate the number of reactions involved in a particular pathway before and after the reconstruction, respectively. After the reconstruction, some pathways maintained a similar position in the ranking, e.g., map00230, while others occupied completely different positions, e.g., map00071, map00062 and map00281. This illustrates the effect of our reconstruction approach. The reasons for these changes are varied and, therefore, are discussed in the following section in more detail. The first example is the fatty acid elongation pathway (map00062), which was given a rank of 42 prior to reconstruction and was ranked second following reconstruction. Although there are 15 reactions from H in CN1 in this map, the reconstruction algorithm decided to exclu-

Table B.7: Ranking of the KEGG pathways after reconstruction using functional network data for CN1 and CN2.

KEGGID	Name	CN1	CN2	Score	Rank	Rank before
map00071	Fatty acid metabolism	4	27	19.5926	1	22
map00062	Fatty acid elongation	0	15	15	2	42
map00330	Arginine and proline metabolism	17	31	14.0541	3	19
map00540	Lipopolysaccharide biosynthesis	3	18	12.5	4	54
map00760	Nicotinate and nicotinamide metabolism	13	25	12.4667	5	24
map00230	Purine metabolism	42	57	12	6	5
map00281	Geraniol degradation	0	12	12	7	37
map00260	Glycine, serine and threonine metabolism	14	26	9.31034	8	33
map00400	Phenylalanine, tyrosine and tryptophan biosynthesis	12	24	8.61538	9	14
map00500	Starch and sucrose metabolism	5	10	8.35714	10	62
map00523	Polyketide sugar unit biosynthesis	0	8	8	11	20
map00620	Pyruvate metabolism	12	23	7.8	12	36
map00130	Ubiquinone and other terpenoid-quinone biosynthesis	2	9	7.2	13	2
map00364	Fluorobenzoate degradation	7	0	7	14	102
map00650	Butanoate metabolism	14	12	6.85714	15	43

de all of them from this pathway. This does not mean that the measured enzymes are not listed in the corresponding reconstruction; it only means that other reactions associated with those enzymes have been chosen.

We also found that the TCA cycle (map00020) was not complete after reconstruction, despite the fact that all of its enzymes were experimentally identified. In particular, fumarate reductase (EC 1.3.99.1) and 2-oxoglutarate dehydrogenase (EC 1.1.1.42) were absent, although they were certainly part of the reconstructed networks. The problem is that the SEED reaction database does not link these reactions to the corresponding KEGG identifiers because they are written with generic acceptors in KEGG. As noted above, the colouring process applied in KEGG is not straightforward, and it may involve some mistakes.

The reconstruction process can also emphasise differences that were discernible before the reconstruction. A remarkable example of this is provided by the purine metabolism pathway (map00230). Before the reconstruction, the pathway from D-ribose-1P to 1-(5'-phosphoribosyl)-5-amino-4-imidazolecarboxamide (AICAR) in this map was nearly complete in CN2, and it was complete after the reconstruction. In contrast, in CN1, only 3 enzymes in this pathway were measured experimentally but were neglected once the network was reconstructed.

Another interesting fact is that some pathways that are known to be inoperative in CN1 and CN2 did not appear after the reconstruction, as is the case for "Carotenoid biosynthesis" (map00906), "Insect hormone biosynthesis" (map00981) or "Drug metabolism - cytochrome P450" (map00982), among others. These pathways were coloured according to the ECs that were measured, but they were discarded after the reconstruction. It should also be noted that the names of the KEGG maps might be misleading. For example, “Methane metabolism” appeared at the 16th position after the reconstruction (see Table B.6), which seems to contradict the experimental evidence that methane production is not possible in either CN1 or CN2. A more detailed examination verified that the active reactions in this map are not related to methane production.

Additionally, the reconstruction can provide attractive hypotheses for further study. For instance, for the nicotinate and nicotinamide metabolism pathway (map00760), activity associated with the conversion of nicotinate to pyruvate and propanoate was predicted in our CN2 reconstruction to involve only enzymes present in D, and not in H or M.

B.6. Contribution of bacterial members to CN1 and CN2 functional network

As noted in Chapter 3, in order to evaluate the role of each bacterial member in CN1 and CN2 at the functional level, we determined its contribution for each KEGG map. The contribution was determined as the number of times a bacterium appears in a KEGG map divided by its total number of active reactions. Figure 3.3 in Chapter 3 shows the contribution of each organism found in both CN1 and CN2 to each KEGG map.

For this analysis, we only took into account the reactions in H and M involved in the CN1 and CN2 reconstructed network. As noted above, the taxonomic affiliation is known for the reactions in H. In contrast, for the reactions in M, we may have different members of the community involved in a reaction. For simplicity, in these situations, if possible, we assign an organism that was previously included in the KEGG map via the reactions from H. Full details as to the taxonomic assignment of reactions involved in CN1 and CN2 metabolic networks can be found in Supplementary Material IV of (Tobalina et al., 2015).

Here, we present 3 bar graphs that complement the heatmap in the main text. Figure B.2 shows the number of KEGG maps including a particular number of organisms. Figure B.3 shows the number of pathways in which each organism takes part. Finally, Figure B.4 details the number of reactions of CN1 and CN2 metabolic network that can be related to each organism.

We repeated the same analysis but taking into account the reactions in H and not the ones in M to evaluate the role of different bacterial members in CN1 and CN2. Figure B.5 shows that the pathways appear less populated. In particular, no organism is assigned to the naphthalene degradation pathway, which shows the need of reconstruction methods. Figure B.6 and B.7 confirm that the predictions made by the reconstruction algorithm com-

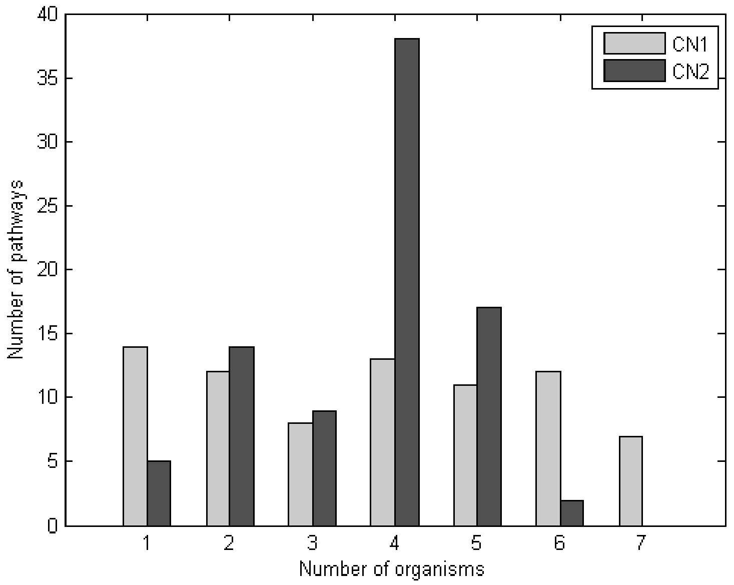


Figure B.2: Number of pathways that involves a particular number of organisms as computed in the heatmap in Figure 3.3 in Chapter 3.

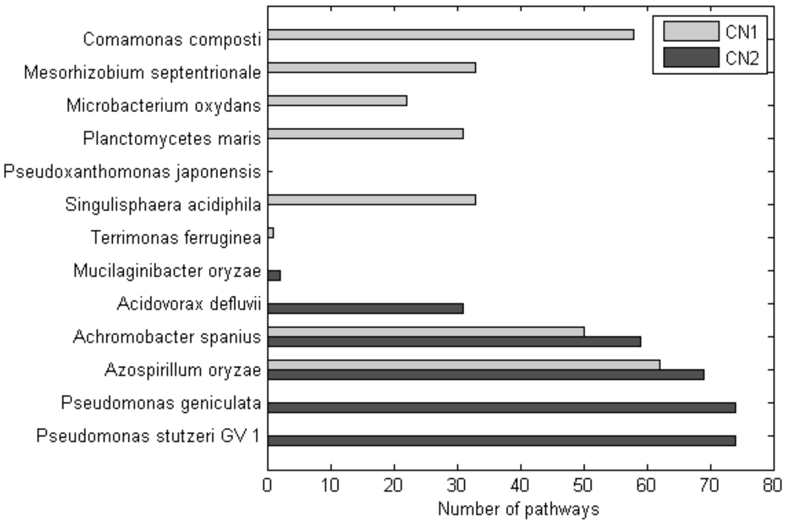


Figure B.3: Number of pathways where an organism takes part as computed in the heatmap in Figure 3.3 in Chapter 3.

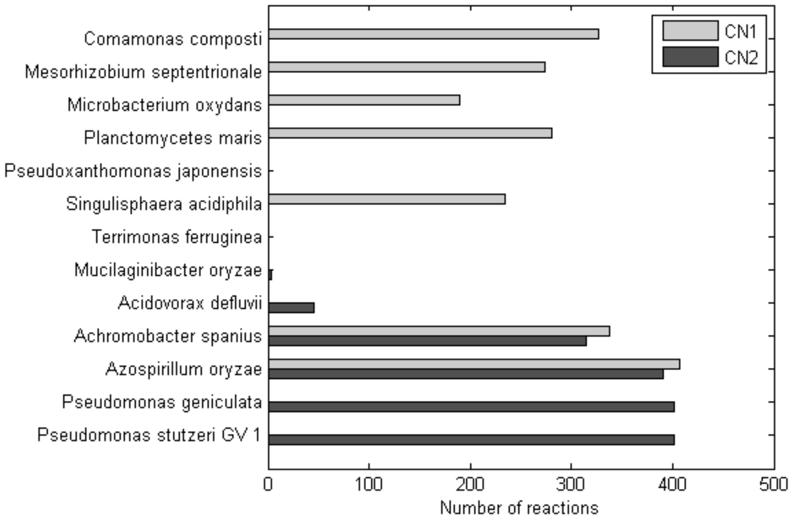


Figure B.4: Number of reactions associated to each organism in the reconstructed networks.

plements experimental data, as the number of interactions among bacterial members in different pathways is substantially decreased when reactions in M are removed.

B.7. Enrichment and mineralization experiments

Chemicals and reagents: The following reagents have been used: O-methoxyamine hydrochloride (Sigma-Aldrich - Taufkirchen, Germany) 15 mg/mL in pyridine (Silylation grade - Taufkirchen, Germany), N,O-Bis(trimethylsilyl)trifluoroacetamide (BSTFA) with 1 % of trimethylchlorosilane (TMCS; Pierce Chemical Co, Rockford, IL, USA), C18:0 methyl ester (Sigma-Aldrich - Taufkirchen, Germany) in heptane (GC-MS grade – Sigma-Aldrich - Taufkirchen, Germany), isopropanol (HLPC-MS grade – Sigma-Aldrich - Taufkirchen, Germany) in addition to standards: 3-fluorobenzoic acid, 4-fluorobenzoic acid, 3-fluorocatechol, 4-fluorocatechol, geraniol, geranial and geranic acid (Sigma Chemical Co.; St Louis, MO, USA or TCI Fine Chemicals, Eschborn, Germany).

The ability of the consortia to grow on 3/4-fluorobenzoate and geraniol (Sigma Chemical Co.; St Louis, MO, USA) as the sole carbon and energy sources was evaluated in 250-ml Erlenmeyer flasks containing 100 ml of Bushnell Hass minimal medium (Sigma Chemical Co.) and supplemented

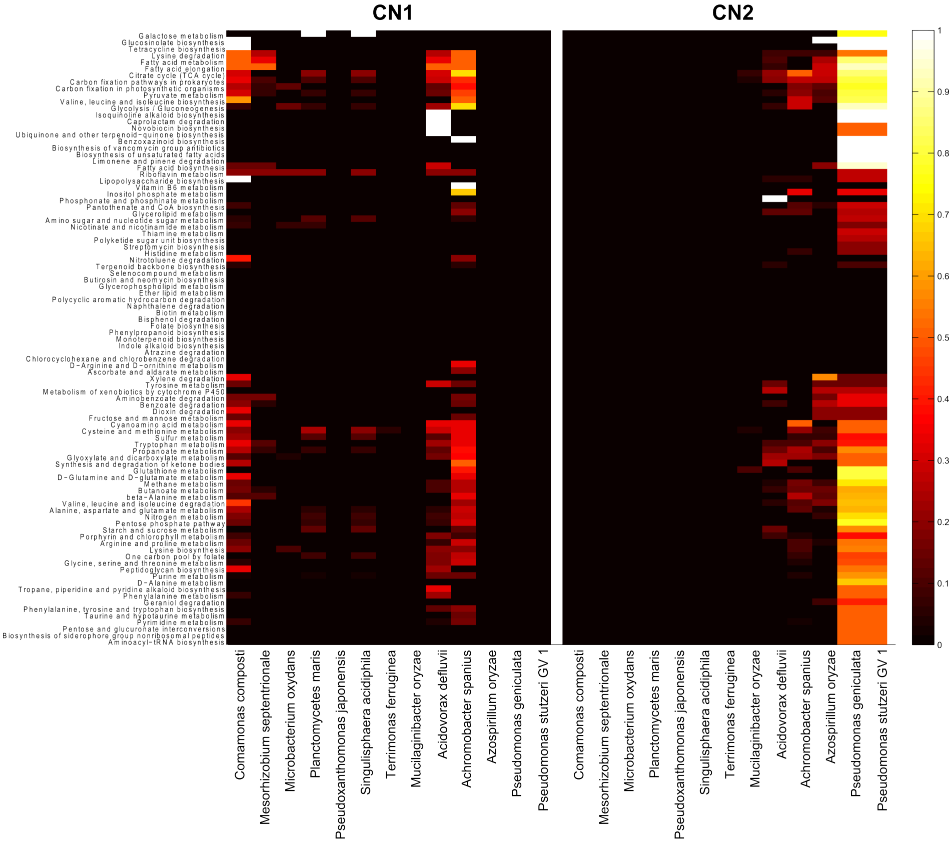


Figure B.5: Heatmap showing the contributions of the most relevant bacterial members of CN1 and CN2 to the KEGG maps, taking only into account reactions in H included in the reconstructions.

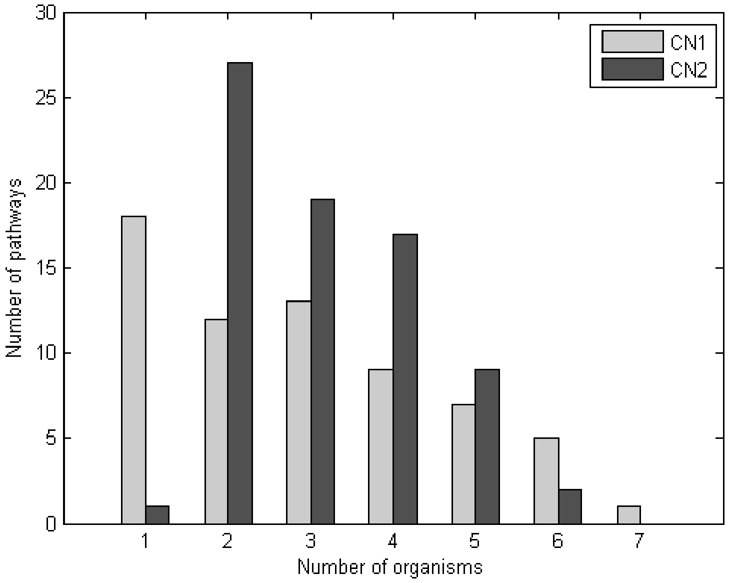


Figure B.6: Number of pathways that involves a particular number of organisms as computed in the heatmap in Figure B.5.

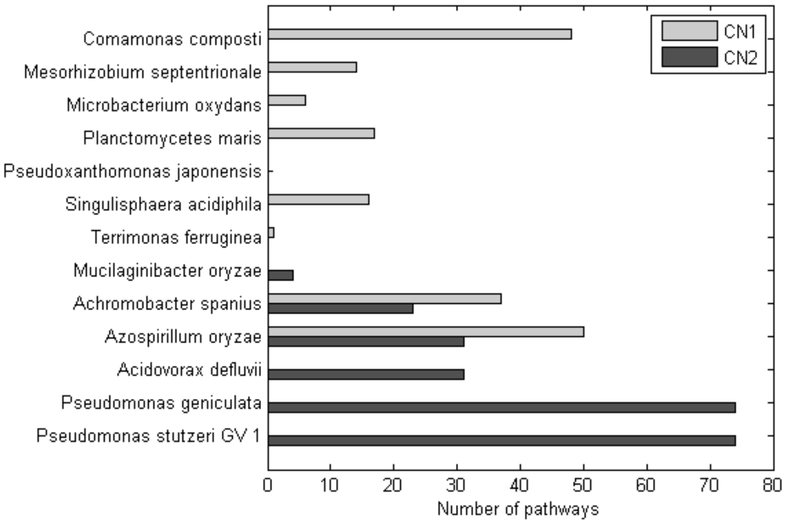


Figure B.7: Number of pathways where an organism takes part as computed in the heatmap in Figure B.5.

with the substrates to 0.1 % (w/v). Inocula (1 % (v/v) of the culture previously grown and maintained in naphthalene; Guazzaroni et al. (2013), used were cells grown in the same medium with 0.1 % (w/v) of naphthalene (Guazzaroni et al., 2013) and washes three times prior to use with the same medium. Enrichment cultures were incubated at 30°C and 250 rpm and growth was measured spectrophotometrically (using a Take3^{TM} microvolume plate and a BioTek Eon spectrophotometer; Izasa, Madrid, Spain) by taking measurements of the culture medium periodically at an optical density of 600 nm. The extent of mineralization was quantified on a solution containing 1 ml of the culture medium (previously separated by centrifugation at 13000 g for 5 min) and 1 ml of a methanol solution prepared as follows. Briefly, microbial cells obtained at time 0, 1 week and 2 weeks were centrifuged from enrichment at 13000 g for 5 min; then, metabolite extraction was performed by adding 1.2 ml of cold (-80°C) methanol using conditions previously described (Pérez-Cobas et al., 2012).

The presence of 3/4-fluorobenzoate and geraniol as well as their putative initial degradation products, e.g. 3/4-fluorocatechol, citral, geranic acid, was determined by using Gas Chromatography-mass analyzer (GC-MS). Samples for GC-MS analysis were prepared from 150 μL volumes of the methanol extract obtained as above. Blanks were prepared to reflect the matrix of samples; they were subsequently treated in the same way as samples. Standards were prepared at a concentration of 100 ppm. All samples were evaporated to dryness using a Speedvac Concentrator (Thermo Fisher Scientific Inc., Waltham, MA) and derivatised using a two stage process: methoxymation and silylation. For methoxymation, 10 μL of O-Methoxyamine hydrochloride (15 mg/mL) in pyridine was added to each sample, following which samples were subjected to three cycles of vortex mixing and ultra-sonication and kept in the dark at room temperature for 16 h. For silylation, 10 μL of BSTFA with 1 % TMCS was added to the solution and samples were again subjected to three cycles of vortex mixing and ultra-sonication before being placed in 70°C for 1 h. Finally, 75 μL of 10 ppm C18:0 Methyl Stearate in heptane (internal standard) was added to each sample and all samples were vortex mixed for 2 min.

The analytical run started with the injection of C18 (10 ppm) followed by four blanks, in order to equilibrate the column. Subsequently, samples were analysed in a randomised order followed by the standards. The run terminated with the injection of the final blank.

The GC-MS system (Agilent Technologies 7890A) consisted of an autosampler (Agilent Technologies 7693) and an inert MSD with Quadrupole (Agilent Technologies 5975). Derivatised samples were injected at volumes of 2 μL through a GC-Column DB5-MS (30 m length, 0.25 mm i.d., 0.25

Table B.8: Metabolomic target analysis of key chemical species participating in the 3/4-fluorobenzoate and geraniol-degradation pathway. The separation and quantification was performed by GC-MS. The average fold-change (F2 vs F0 cultures) corresponding to the area of the peak (calculated on the basis of appropriate standards) per metabolite is given. Note: 3- and 4-fluorochatechol were indistinguishable used assay conditions. Abbreviations as follows: F2: cultures after 2 weeks cultivation in the present of the corresponding chemical; F0: cultures after 1 h cultivation in the present of the corresponding chemical; NC: no change observed (no production of the chemical species).

	Fold change (FC)	
Name	CN1 (F2 vs F0)	CN2 (F2 vs F0)
3/4-Fluorocatechol	283,49	NC
Citral	NC	23
Geranic acid	NC	6

μm film 95 % dimethyl/5 % diphenylpolysiloxane) with a pre-column (10 m J&W integrated with Agilent 122-5532G). The helium carrier gas flow rate was set at 1 mL/min and the injector temperature at 250°C. The split ratio was 1:10 flowing into a Restek 20782 deactivated glass-wool split liner. The temperature gradient was programmed as follows: the initial oven temperature was set at 60°C (held for 1 min), then it was increased to 325°C at the rate of 10°C/min. Finally a cool-down period was applied for 10 min before the following injection. The total analysis time for each sample was 37.5 min. The detector transfer line, the filament source and the quadrupole temperature were set at 290°C, 230°C and 150°C respectively. The electron ionization (EI) source was operated at 70 eV. The mass spectrometer was operated in scan mode over a mass range of m/z 50-600 at a rate of 2 spectra/s. Peak detection and spectra processing were obtained using the Agilent ChemStation Software (G1701EA E.02.00.493, Agilent).

The compound identification was performed by using the NIST 08 Library (National Institute of Standards and Technology, U.S. Department of Commerce), with the ChemStation software (G1701EA E.02.00.493, Agilent). As soon as they were properly characterised in the chromatograms of standards (retention time and spectrum) a target analysis method was created in the ChemStation software (G1701EA E.02.00.493, Agilent) that was used to identify and integrate the corresponding peaks in the chromatograms of samples.

The fold change of the abundance level of the degradation intermediates 3/4-fluorocatechol (for 3/4-fluorocatechol) and citral/geranic acid (for geraniol) in CN1 and CN2 enrichment cultures can be found in Table B.8.

Appendix C

Direct calculation of Minimal Cut Sets involving a specific reaction knock-out

C.1. Direct calculation of Elementary Flux Modes involving two specific reactions

In this section, we apply the formulation of Pey and Planes (2014) to the MCS problem without modifications to the set of constraints. We only modify the objective function to minimize the variables relevant to MCSs.

The reversible u variables are separated into two irreversible variables, up and un, as shown in Eqs. (C.1) and (C.2). These new variables will have related binary variables, zup and zun.

$$u_i = up_i - un_i, \qquad i = 1, ..., m \tag{C.1}$$

$$up_i \geq 0, un_i \geq 0, \qquad i = 1, ..., m \tag{C.2}$$

We want to activate two reactions, w and one specific rp or rn. Equation (C.5) forces w to be active and Eq. (C.6) does the same with the rp or rn of interest. For the solution to be an EFM, it must fulfill the non-decomposability condition (NDC), i.e. it must have a single degree of freedom. As explained in Chapter 5, this can only be achieved if the second flux activation constraint can be written as a linear combination of the equations that describe the EFM. However, because we do not know what reactions will become active beforehand, we need to introduce binary variables that will force the linear combination only when necessary (Eq. C.10).

Overall, the problem to be solved is the following:

$$\text{minimize} \sum_i zp_i + \sum_{i \in Rev} zn_i \tag{C.3}$$

subject to:

$$N \cdot \begin{pmatrix} up \\ un \\ rp \\ rn \\ w \end{pmatrix} = \begin{bmatrix} S^T & -S^T & I & -I & -t \end{bmatrix} \cdot \begin{pmatrix} up \\ un \\ rp \\ rn \\ w \end{pmatrix} = 0 \tag{C.4}$$

$$-v^* \cdot w \leq -c \tag{C.5}$$

$$\begin{pmatrix} 0 & 0 & d_p & d_n & 0 \end{pmatrix} \cdot \begin{pmatrix} up \\ un \\ rp \\ rn \\ w \end{pmatrix} \geq c \tag{C.6}$$

$$\alpha \cdot \begin{pmatrix} zup \\ zun \\ zp \\ zn \\ zw \end{pmatrix} \leq \begin{pmatrix} up \\ un \\ rp \\ rn \\ w \end{pmatrix} \leq M \cdot \begin{pmatrix} zup \\ zun \\ zp \\ zn \\ zw \end{pmatrix} \tag{C.7}$$

$$zup_i + zun_i \leq 1, \qquad i = 1, ..., m \tag{C.8}$$

$$zp_i + zn_i \leq 1, \qquad \forall i \in Rev \tag{C.9}$$

$$\begin{bmatrix} S & 0 \\ -S & 0 \\ I & 0 \\ -I & 0 \\ -t^T & v^* \end{bmatrix} \cdot x = \begin{pmatrix} 0 + \epsilon_{up} - \delta_{up} \\ 0 + \epsilon_{un} - \delta_{un} \\ d_p + \epsilon_p - \delta_p \\ d_n + \epsilon_n - \delta_n \\ 0 + \epsilon_w - \delta_w \end{pmatrix} \tag{C.10}$$

$$M \cdot \begin{pmatrix} 1-zup \\ 1-zun \\ 1-zp \\ 1-zn \\ 1-zw \end{pmatrix} \geq \begin{pmatrix} \epsilon_{up}+\delta_{up} \\ \epsilon_{un}+\delta_{un} \\ \epsilon_{p}+\delta_{p} \\ \epsilon_{n}+\delta_{n} \\ \epsilon_{w}+\delta_{w} \end{pmatrix} \tag{C.11}$$

$$\begin{pmatrix} up \\ un \\ rp \\ rn \\ w \end{pmatrix} \geq 0, \begin{pmatrix} \epsilon_{up} \\ \epsilon_{un} \\ \epsilon_{p} \\ \epsilon_{n} \\ \epsilon_{w} \end{pmatrix} \geq 0, \begin{pmatrix} \delta_{up} \\ \delta_{un} \\ \delta_{p} \\ \delta_{n} \\ \delta_{w} \end{pmatrix} \geq 0 \tag{C.12}$$

$$\begin{pmatrix} zup \\ zun \\ zp \\ zn \\ zw \end{pmatrix} \in \{0,1\} \tag{C.13}$$

$$up \in R^m, un \in R^m, rp \in R^n, rn \in R^n, w \in R \tag{C.14}$$

$$x \in R^{n+1} \tag{C.15}$$

$$\epsilon_{up} \in R^m, \epsilon_{un} \in R^m, \epsilon_p \in R^n, \epsilon_n \in R^n, \epsilon_w \in R \tag{C.16}$$

$$\delta_{up} \in R^m, \delta_{un} \in R^m, \delta_p \in R^n, \delta_n \in R^n, \delta_w \in R \tag{C.17}$$

$$c > 0 \tag{C.18}$$

where d_p and d_n are vectors of all zeros except for a single 1 in the position related to the knock-out constraint that we want to activate. If the knock-out we want to enforce involves an irreversible reaction we will only set a 1 in d_p and leave d_n as a vector of all zeros. Note that, as explained in Chapter 5, the solutions obtained from this formulation are not necessarily MCSs of the original problem.

C.2. Direct calculation of Generating Flux Modes involving two specific reactions

Elements of a convex basis of EFMs are termed Generating Flux Modes (GFMs). As introduced in Larhlimi and Bockmayr (2009), GFMs have minimal support with respect to the set of irreversible reactions. We can adapt the formulation in Pey and Planes (2014) to find GFMs by not splitting reversible reactions into two irreversible reactions and not taking them into account in the objective function. As a consequence of not splitting reversible reactions, the linear combination constraint must always apply to them. Since the value 0 is no longer a bound for those variables and they can take any real value, they must always be considered as being active, which forces them to participate in the linear combination constraint. Reversible reactions do not have a constraint capable of intersecting the solution space and, consequently, cannot reduce the degrees of freedom of the solution, only irreversible reactions do.

C.3. From GFMs to MCSs

GFMs do not correspond one-to-one to MCSs since GFMs have minimal support with respect to the full set of irreversible reactions (w, rp and rn variables in our case) while MCSs have it with respect to a subset of irreversible reactions (w, rp and rn variables related with the knockout of reversible reactions).

In the toy example illustrated in Chapter 5, for instance, we have one GFM involving the following irreversible dual reactions $g_1 = \{rp_1, rn_9, w\}$ and other involving $g_2 = \{rp_1, rp_6, rn_5, w\}$. These GFMs are minimal with respect to the full set of irreversible reactions. However, as rn_5 and rn_9 are not dual variables related with knockouts of reversible variables, it can be shown that g_2 does not correspond to an MCS. If we map the solutions with respect to the dual variables related with knockouts, we would have $g_1^* = \{rp_1, w\}$ and $g_2^* = \{rp_1, rp_6, w\}$, and it is clear now that g_1^* is contained in g_2^*, thus g_2^* cannot be minimal as g_1^* is already an MCS on its own. Therefore, we need to find a way to remove GFMs that do not correspond to MCSs.

Consider the following example network in Figure C.1, which involves 4 irreversible reactions (this network does not correspond to the dual of any MCS problem, but that does not affect the discussion that follows). As all the reactions are irreversible, in this case the set of GFMs is equivalent to the set of EFMs. The network involves 2 GFMs: $e_1 = \{v_1, v_3\}$ and $e_2 = \{v_1, v_2, v_4\}$. Assume now that we want to discuss the subset of GFMs that are minimal with respect to the subset of reactions $K = \{v_1, v_2\}$ and, additionally, involve reaction v_1.

First, as v_3 and v_4 are non-negative variables and we are not interested in their value, we can view them as excess variables that appear explicitly in the constraints. We can remove them from the set of constraints without altering the solution space as long as we modify the constraint type ac-

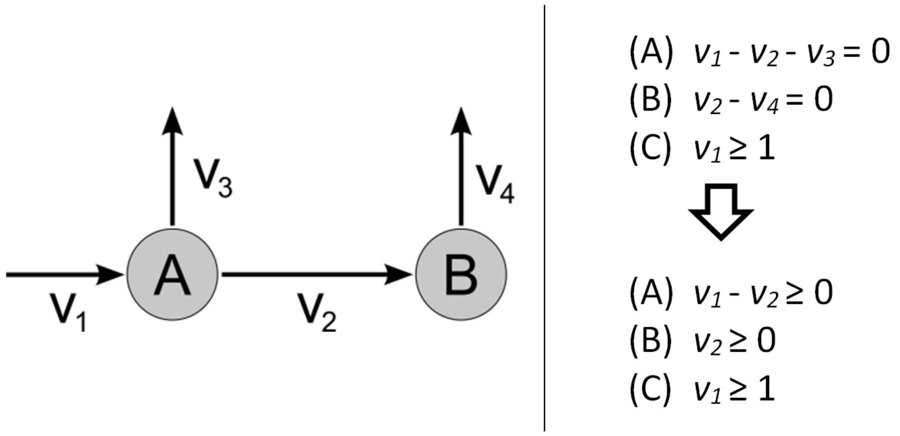


Figure C.1: Example network to illustrate the effect in the formulation of removing variables that can be considered as explicit excess variables (v_3 and v_4).

cordingly. Equality constraints become *greater than or equal to* constraints after the removal of v_3 and v_4, as can be seen in Figure C.1. This is similar to what is done in Eq. (C.19) in the next section, where rn variables associated with irreversible reactions are removed. The resulting feasible region is shown in Figure C.2A by the shaded region. We have 2 extreme points that corresponds to the 2 GFMs mentioned above ($w_1 = \{v_1\}$, $w_2 = \{v1, v2\}$).

Under these constraints, there are only two possible basic solutions for the problem posed: i) either v_1 is active on its own or ii) v_1 and v_2 are both active. Now, we are interested in knowing if the second option would be part of the subset of GFMs that we are looking for (minimal with respect to K and involving the activity of v_1). It is easy to see in this example that, as we have one extreme point (w_1) that exclusively involves v_1, w_2 is not minimal with respect to K and, therefore, the answer to our question is negative. Note the analogy with the example introduced in Chapter 5 and discussed above with GFMs g_1 and g_2.

In order to systematically answer this question, we add the constraint $v_2 \geq 1$, further imposing that it must be redundant with respect to the rest of constraints. A redundant constraint can be obtained as a linear combination of the rest, with non-negative coefficients for *greater than or equal to* inequalities, unrestricted coefficients for equalities and non-positive coefficients for *less than or equal to* inequalities. Note that the coefficients of this linear combination follow the same rules of the dual variables associated to these constraints.

Given that in our example we have *greater than or equal to* constraints, the coefficients used for the linear combination must be non-negative (see x

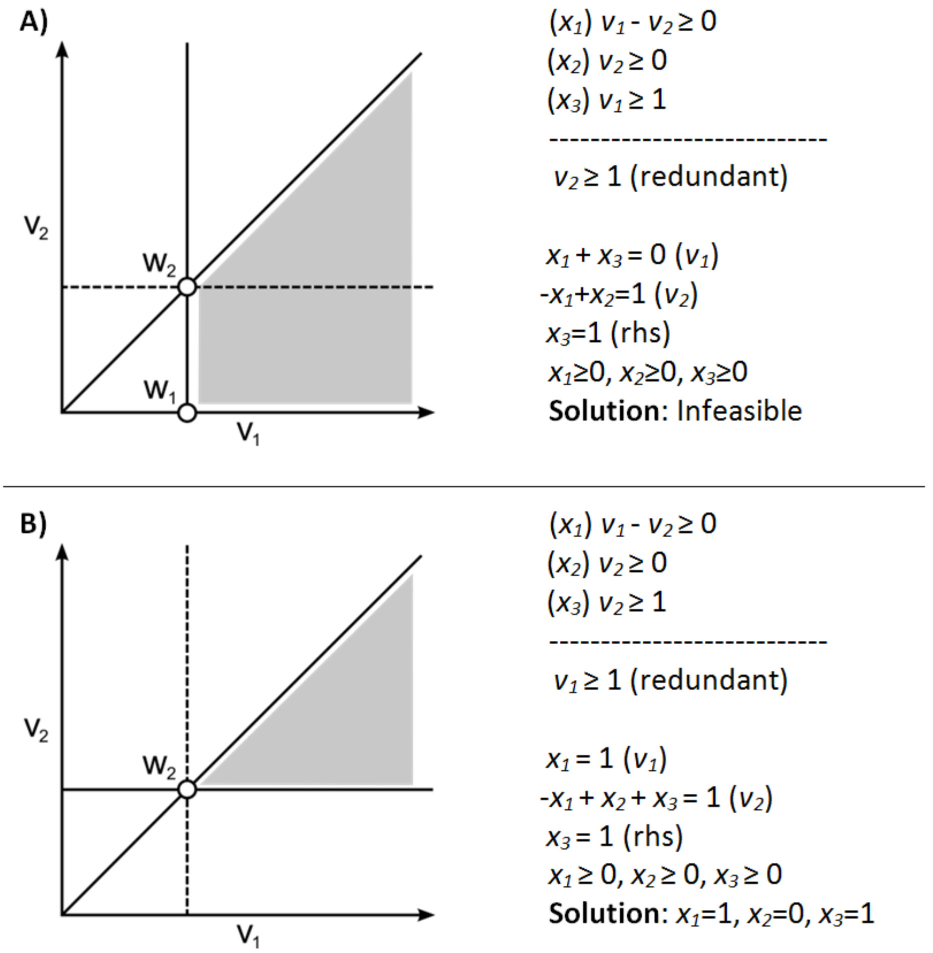


Figure C.2: Solution space (shaded region) generated by the constraints related to the network in Figure C.1 and a flux activation constraint (solid line). An additional flux activation constraint (dashed line) may or may not intersect the existing solution space (case A and B, respectively). We want the additional constraint to be redundant. Only if the additional constraint does not intersect the existing solution space (case B), it is a redundant constraint with respect to the others, i.e. it does not eliminate already existing extreme points nor generate any new ones.

variables in Figure C.2A). In this case, $v_2 \geq 1$ is not redundant with respect to the rest of constraints and the solution is infeasible. Visually, this has the effect of $v_2 \geq 1$ intersecting the feasible region (Figure C.2A, dashed line). This justifies the use of non-negative coefficients (Eq. 5.19) in the linear combination constraints (Eq. 5.16), adapting the methodology presented in Pey and Planes (2014) for computing MCSs.

Assume now that we want to filter the subset of GFMs that is minimal with respect to the subset of reactions $K = \{v_1, v_2\}$ and, additionally, involve the reaction v_2. Again, we treat v_3 and v_4 as explicit excess variables. The resulting feasible region is shown by the shaded region in Figure C.2B. We only have 1 extreme point that corresponds to one of the 2 GFMs mentioned before (w_2).

We would like to consider again whether v_1 and v_2 can be both active and the result be minimal with respect to K. For that, we impose $v_1 \geq 1$ and the linear combination constraint to force its redundancy (see Figure C.2B). In this case, as v_2 cannot operate on its own, we certainly have that $v_1 \geq 1$ is a redundant constraint and, therefore, we obtain a feasible set of x variables. Visually, a redundant constraint does not eliminate already existing extreme points or generate new ones.

The reader should note the analogy of this problem with the formulation presented in this article for MCSs enumeration. In the second example, $v_1 \geq 1$ is implied by the rest of constraints and, therefore, it is redundant. This does not occur in the first case. Certainly, v_2 can also be equal to zero, since v_1 can be active on its own. Therefore, by imposing redundancy, we are removing from the solution space GFMs that are non-minimal with respect to the subset of dual variables related with knockouts.

C.4. Reduced formulation

As stated in Chapter 5 (and discussed above), rn variables related to the irreversibility constraints can be viewed as excess variables explicitly stated in the formulation. Thus, we can remove them from the formulation, transforming the equality constraints in which they appear in Eq. (5.11) into *greater than or equal* inequalities (Eq. C.19).

$$\begin{bmatrix} S_{irr}^T & I_{irr} & 0 & 0 & -t_{irr} \\ S_{rev}^T & 0 & I_{rev} & -I_{rev} & -t_{rev} \end{bmatrix} \cdot \begin{pmatrix} u \\ rp_{irr} \\ rp_{rev} \\ rn_{rev} \\ w \end{pmatrix} \begin{matrix} \geq \\ = \end{matrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{C.19}$$

, where subindices *irr* and *rev* refer to the appropriate columns or rows related to the irreversible and reversible reactions respectively. We now only have *rp* and *rn* variables related to knock-out constraints (Eqs. 5.4 and 5.5). We denote $|Irr|$ and $|Rev|$ the number of irreversible and reversible reactions in the primal system respectively.

$$u \in R^m, rp_{irr} \in R^{|Irr|}, rp_{rev} \in R^{|Rev|}, rn_{rev} \in R^{|Rev|}, w \in R \tag{C.20}$$

$$rp_{irr} \geq 0, rp_{rev} \geq 0, rn_{rev} \geq 0, w \geq 0 \tag{C.21}$$

$$\begin{pmatrix} 0 & dp_{irr} & dp_{rev} & dn_{rev} & 0 \end{pmatrix} \cdot \begin{pmatrix} u \\ rp_{irr} \\ rp_{rev} \\ rn_{rev} \\ w \end{pmatrix} \geq b \tag{C.22}$$

We also adapt the remaining constraints.

$$\begin{bmatrix} S_{irr} & S_{rev} & 0 \\ I_{irr} & 0 & 0 \\ 0 & I_{rev} & 0 \\ 0 & -I_{rev} & 0 \\ -t_{irrev}^T & -trev^T & v^* \\ 0 & 0 & c \end{bmatrix} \cdot \begin{pmatrix} x_{irr} \\ x_{rev} \\ x_w \end{pmatrix} = \begin{pmatrix} 0 \\ dp_{irr} + \epsilon p_{irr} - \delta p_{irr} \\ dp_{rev} + \epsilon p_{rev} - \delta p_{rev} \\ dn_{rev} + \epsilon n_{rev} - \delta n_{rev} \\ 0 \\ b \end{pmatrix} \tag{C.23}$$

$$- v^* \cdot w \leq -c \tag{C.24}$$

$$x_{irr} \in R^{|Irr|}, x_{rev} \in R^{|Rev|}, x_w \in R \tag{C.25}$$

$$\epsilon p_{irr} \in R^{|Irr|}, \epsilon p_{rev} \in R^{|Rev|}, \epsilon n_{rev} \in R^{|Rev|} \tag{C.26}$$

$$\delta p_{irr} \in R^{|Irr|}, \delta p_{rev} \in R^{|Rev|}, \delta n_{rev} \in R^{|Rev|} \tag{C.27}$$

$$x_{irr} \geq 0 \tag{C.28}$$

$$\epsilon p_{irr} \geq 0, \epsilon p_{rev} \geq 0, \epsilon n_{rev} \geq 0 \tag{C.29}$$

$$\delta p_{irr} \geq 0, \delta p_{rev} \geq 0, \delta n_{rev} \geq 0 \tag{C.30}$$

$$\alpha \cdot \begin{pmatrix} zp_{irr} \\ zp_{rev} \\ zn_{rev} \end{pmatrix} \leq \begin{pmatrix} rp_{irr} \\ rp_{rev} \\ rn_{rev} \end{pmatrix} \leq M \cdot \begin{pmatrix} zp_{irr} \\ zp_{rev} \\ zn_{rev} \end{pmatrix} \tag{C.31}$$

$$zp_{rev,i} + zn_{rev,i} \geq 1, \qquad \forall i \in Rev \tag{C.32}$$

$$M \cdot \begin{pmatrix} 1 - zp_{irr} \\ 1 - zp_{rev} \\ 1 - zn_{rev} \end{pmatrix} \geq \begin{pmatrix} \epsilon p_{irr} + \delta p_{irr} \\ \epsilon p_{rev} + \delta p_{rev} \\ \epsilon n_{rev} + \delta n_{rev} \end{pmatrix} \tag{C.33}$$

$$\begin{pmatrix} zp_{irr} \\ zp_{rev} \\ zn_{rev} \end{pmatrix} \in \{0, 1\} \tag{C.34}$$

Note that Eq. (C.23) has an extra row that comes from the right-hand side of Eqs. (C.19), (C.22) and (C.24). Because the only requirement for the values of b and c is that they must be strictly positive, we can substitute that last row in Eq. (C.23) by the following constraint:

$$x_w \geq \phi, \phi > 0 \tag{C.35}$$

This allows the solver to decide if b is greater than c or not.

Finally, we adapt the objective function and the solution enumeration constraint (Eqs. C.36 and C.37, respectively).

$$\text{minimize} \sum_{i \in Irrev} zp_{irr,i} + \sum_{i \in Rev} zp_{rev,i} + \sum_{i \in Rev} zn_{rev_i} \tag{C.36}$$

$$\begin{aligned} &\sum_{i \in Irrev} zp^k_{irr,i} \cdot zp_{irr,i} + \sum_{i \in Rev} zp^k_{rev,i} \cdot zp_{rev,i} + \sum_{i \in Rev} zn^k_{rev,i} \cdot zn_{rev_i} \leq \\ &(\sum_{i \in Irrev} zp^k_{irr,i} + \sum_{i \in Rev} zp^k_{rev,i} + \sum_{i \in Rev} zn^k_{rev,i}) - 1, \qquad k = 1, ..., K \end{aligned} \tag{C.37}$$

C.5. Limiting the maximum size of the MCSs

Although not used in this work, we can add an additional constraint to limit the maximum size (C) allowed for a MCS (Eq. C.38 or C.39).

$$\sum_{i} zp_i + \sum_{i \in Rev} zn_i \leq C \tag{C.38}$$

$$\sum_{i \in Irrev} zp_{irr,i} + \sum_{i \in Rev} zp_{rev,i} + \sum_{i \in Rev} zn_{rev,i} \leq C \tag{C.39}$$

C.6. Computational difficulties

As with any other MILP, numerical difficulties can arise during the solution process. For this reason, we recommend checking the correctness of each MCS. This can be done by calculating the maximum flux through the target reaction with all the reactions included in the MCS set to zero, which should give a maximum flux of zero, and then recalculating this quantity after setting free one of the reactions included in the MCS, which should give a maximum flux greater than zero. If the maximum flux through the target reaction is still zero after setting free one of the reactions included in the MCS, then the solution is not really a MCS. This process is not computationally expensive, as only linear programming FBA simulations are involved.

Whenever we found that the solution to our problem was not a real MCS, we were able to obtain a correct solution by adjusting some of the solver parameters (e.g. the integrality tolerance) or modifying the values of some of the constants in the formulation (e.g. the M value).

Regarding the computation time, as shown in Chapter 5, it can be very variable. It might be desirable to set a time limit for obtaining a solution. By the time the limit is reached, we may or may not have found a solution. If no solution has been found, we cannot assure that the problem is infeasible. If a solution has been found, it will contain an MCS in case it is not already one, but we cannot assure that it is optimal as measured by our objective function. Depending on our application, we may decide to increase the time available to the solution process or to continue the analysis with the answer we got within our time limit.

Appendix D

Table of samples

This appendix contains the list of samples used in the study described in Chapter 4. The "Accession" column contains the GEO database accession number of the selected cell line sample in the Cancer Cell Line Encyclopedia (CCLE). The "Name" column corresponds to the cell line name. Those with "fail" beside the name correspond to experiments that did not pass the quality control in Project Achilles and were not taken into account in this work. Finally, the column "Type" describes the cancer type associated to each cancer cell line.

Table D.1: List of samples used in Chapter 4.

Accession	Name	Type
GSM886837	22-RV1	Prostate
GSM886843	697	Leukemia
GSM886845	786-O	Renal Cell Carcinoma
GSM886850	A172	GBM
GSM886851	A-204	Soft Tissue Sarcoma
GSM886853	A2780 (fail)	Ovarian
GSM886858	A549	Lung NSCLC
GSM886859	A-673	Bone sarcoma
GSM886863	ACHN	Renal Cell Carcinoma
GSM886864	AGS	Gastric
GSM886866	AM-38	GBM
GSM886867	AML-193	Leukemia
GSM886870	AsPC-1	Pancreas
GSM886891	BT-20	Breast
GSM886892	BT-474	Breast
GSM886894	BT-549 (fail)	Breast

Table D.1: (continued)

Accession	Name	Type
GSM886896	BxPC-3	Pancreas
GSM886897	C2BBe1	Colon
GSM886898	C32	Melanoma
GSM886901	CADO-ES-1	Bone sarcoma
GSM886902	Caki-1 (fail)	Renal Cell Carcinoma
GSM886904	CAL-120	Breast
GSM886909	CAL-51	Breast
GSM886914	Calu-1	Lung NSCLC
GSM886918	CaOV-3	Ovarian
GSM886919	CaOV-4	Ovarian
GSM886922	CAS-1	GBM
GSM886925	CFPAC-1	Pancreas
GSM886940	COLO 205	Colon
GSM886947	COLO-704	Ovarian
GSM886948	COLO 741	Melanoma
GSM886949	COLO-783	Melanoma
GSM886956	COR-L23	Lung NSCLC
GSM886962	COV318 (fail)	Ovarian
GSM886963	COV362	Ovarian
GSM886964	COV434	Ovarian
GSM886965	COV504	Ovarian
GSM886966	COV644	Ovarian
GSM886974	DBTRG-05MG	GBM
GSM886978	DK-MG	GBM
GSM886979	DLD-1	Colon
GSM886989	DU4475 (fail)	Breast
GSM886997	EFE-184	Endometrial
GSM886999	EFM-19	Breast
GSM887000	EFO-21	Ovarian
GSM887001	EFO-27	Ovarian
GSM887003	EJM (fail)	Multiple Myeloma
GSM887010	F-36P	Leukemia
GSM887014	FU-OV-1 (fail)	Ovarian
GSM887021	GCIY (fail)	Gastric
GSM887025	GMS-10 (fail)	GBM
GSM887027	GP2d	Colon
GSM887035	HCC1187	Breast
GSM887037	HCC1395	Breast
GSM887046	HCC1954	Breast

Table D.1: (continued)

Accession	Name	Type
GSM887049	HCC2218	Breast
GSM887056	HCC-44	Lung NSCLC
GSM887058	HCC70	Breast
GSM887060	HCC827	Lung NSCLC
GSM887062	HCT116	Colon
GSM887069	HEC-1-A	Endometrial
GSM887080	Hey-A8	Ovarian
GSM887083	HL-60	Leukemia
GSM887086	HLF	Liver
GSM887089	HPAC	Pancreas
GSM887090	HPAF-II	Pancreas
GSM887106	Hs 683	GBM
GSM887117	Hs 766T	Pancreas
GSM887132	Hs 944.T	Melanoma
GSM887138	HT-1197	Bladder
GSM887141	HT-29	Colon
GSM887142	HT55	Colon
GSM887145	HuG1-N	Gastric
GSM887155	HuTu80	Colon
GSM887159	IGR-39	Melanoma
GSM887160	IGROV1	Ovarian
GSM887175	JHOC-5	Ovarian
GSM887176	JHOM-1	Ovarian
GSM887179	JHOS-4 (fail)	Ovarian
GSM887193	K-562	Leukemia
GSM887194	KALS-1	GBM
GSM887197	Kasumi-1	Leukemia
GSM887212	KM12	Colon
GSM887220	KMS-11 (fail)	Multiple Myeloma
GSM887221	KMS-12-BM	Multiple Myeloma
GSM887223	KMS-20 (fail)	Multiple Myeloma
GSM887225	KMS-26 (fail)	Multiple Myeloma
GSM887228	KMS-34 (fail)	Multiple Myeloma
GSM887230	KNS-60	GBM
GSM887232	KNS-81	GBM
GSM887235	KP-2	Pancreas
GSM887237	KP4	Pancreas
GSM887248	KYSE-150	Esophageal
GSM887251	KYSE-30	Esophageal

Table D.1: (continued)

Accession	Name	Type
GSM887253	KYSE-450	Esophageal
GSM887254	KYSE-510	Esophageal
GSM887258	L3.3	Pancreas
GSM887259	L-363	Multiple Myeloma
GSM887262	LAMA-84	Leukemia
GSM887267	LK-2	Lung NSCLC
GSM887270	LN-229	GBM
GSM887274	LoVo	Colon
GSM887276	LP-1 (fail)	Multiple Myeloma
GSM887280	LS411N	Colon
GSM887281	LS513	Colon
GSM887291	MCF7	Breast
GSM887300	MDA-MB-453	Breast
GSM887320	MIA PaCa-2	Pancreas
GSM887326	MKN7	Gastric
GSM887328	MM1-S	Multiple Myeloma
GSM887329	MOLM-13	Leukemia
GSM887337	MONO-MAC-1	Leukemia
GSM887338	MONO-MAC-6	Leukemia
GSM887344	MV-4-11	Leukemia
GSM887347	NALM-6	Leukemia
GSM887349	NB-4	Leukemia
GSM887355	NCI-H1299	Lung NSCLC
GSM887364	NCI-H1437	Lung NSCLC
GSM887373	NCI-H1650	Lung NSCLC
GSM887382	NCI-H1792	Lung NSCLC
GSM887392	NCI-H196	Lung SCLC
GSM887393	NCI-H1975	Lung NSCLC
GSM887398	NCI-H2052	Lung Mesothelioma
GSM887407	NCI-H2122	Lung NSCLC
GSM887411	NCI-H2171	Lung SCLC
GSM887421	NCI-H23	Lung NSCLC
GSM887424	NCI-H2452	Lung Mesothelioma
GSM887428	NCI-H441	Lung NSCLC
GSM887431	NCI-H508	Colon
GSM887435	NCI-H524 (fail)	Lung SCLC
GSM887440	NCI-H660	Prostate
GSM887441	NCI-H661	Lung NSCLC
GSM887443	NCI-H716	Colon

Table D.1: (continued)

Accession	Name	Type
GSM887447	NCI-H82 (fail)	Lung SCLC
GSM887448	NCI-H838	Lung NSCLC
GSM887453	NCI-N87	Gastric
GSM887456	NIH:OVCAR-3	Ovarian
GSM887458	NOMO-1	Leukemia
GSM887465	OAW42	Ovarian
GSM887468	OCI-AML2	Leukemia
GSM887469	OCI-AML3	Leukemia
GSM887470	OCI-AML5	Leukemia
GSM887476	OE33	Esophageal
GSM887478	OPM-2	Multiple Myeloma
GSM887482	OV7	Ovarian
GSM887483	OV-90	Ovarian
GSM887484	OVCAR-4	Ovarian
GSM887485	OVCAR-8	Ovarian
GSM887488	OVMANA (fail)	Ovarian
GSM887496	Panc 03.27	Pancreas
GSM887499	Panc 08.13	Pancreas
GSM887500	Panc 10.05	Pancreas
GSM887521	PSN1	Pancreas
GSM887522	QGP-1	Pancreas
GSM887530	Reh	Leukemia
GSM887540	RKN	Soft Tissue Sarcoma
GSM887541	RKO	Colon
GSM887544	RMG-I	Ovarian
GSM887545	RMUG-S	Ovarian
GSM887548	RPMI 8226 (fail)	Multiple Myeloma
GSM887552	RT-112	Bladder
GSM887563	SEM	Leukemia
GSM887565	SF126	GBM
GSM887566	SF-295	GBM
GSM887574	SJSA-1	Bone sarcoma
GSM887576	SK-CO-1	Colon
GSM887589	SK-MEL-5	Melanoma
GSM887591	SK-MM-2	Multiple Myeloma
GSM887598	SK-OV-3	Ovarian
GSM887606	SNU-1105	GBM
GSM887615	SNU-201 (fail)	GBM
GSM887640	SNU-840	Ovarian

Table D.1: (continued)

Accession	Name	Type
GSM887643	SNU-C1	Colon
GSM887644	SNU-C2A	Colon
GSM887650	SU.86.86	Pancreas
GSM887667	SW1417	Colon
GSM887671	SW1783	GBM
GSM887672	SW 1990 (fail)	Pancreas
GSM887674	SW480	Colon
GSM887675	SW48	Colon
GSM887687	T98G	GBM
GSM887689	TC-71	Bone sarcoma
GSM887691	TE10	Esophageal
GSM887694	TE-15	Esophageal
GSM887702	TE-9	Esophageal
GSM887706	THP-1	Leukemia
GSM887710	TOV-112D	Ovarian
GSM887711	TOV-21G	Ovarian
GSM887714	T.T	Esophageal
GSM887718	TYK-nu	Ovarian
GSM887720	U-251 MG	GBM
GSM887723	U-87 MG	GBM
GSM887731	VCaP	Prostate
GSM887748	YKG1	GBM
GSM887751	ZR-75-30	Breast

Table D.2: U251 glioblastoma cell line samples from GEO Database used in Chapter 4

Accession
GSM713416
GSM713417
GSM844718
GSM844719
GSM803632
GSM803691
GSM803750
GSM697600
GSM697601
GSM697602
GSM697603
GSM697604
GSM697605
GSM502019
GSM502020
GSM502021
GSM502037
GSM502038
GSM502039

Table D.3: U87 glioblastoma cell line samples from GEO Database used in Chapter 4

Accession
GSM887723
GSM795820
GSM795821
GSM795822
GSM795824
GSM795825
GSM795827
GSM862922
GSM862923
GSM862924

Appendix E

Publications

Journal Publications

Luis Tobalina, Jon Pey and Francisco J. Planes (2015) Direct Calculation of Minimal Cut Sets Involving a Specific Reaction Knock-out. *(under review)*

Luis Tobalina, Jon Pey, Alberto Rezola, Francisco J. Planes (2015) Assessment of FBA Based Gene Essentiality Analysis in Cancer with a Fast Context-specific Network Reconstruction Method. *(under review)*

Luis Tobalina, Rafael Bargiela, Jon Pey, Florian-Alexander Herbst, Iván Lores, David Rojo, Coral Barbas, Ana I. Peláez, Jesús Sánchez, Martin von Bergen, Jana Seifert, Manuel Ferrer and Francisco J. Planes (2015) Context-specific metabolic network reconstruction of a naphthalene-degrading bacterial community guided by metaproteomic data. *Bioinformatics* 31 (11): 1771-1779

Jon Pey, Juan A. Villar, **Luis Tobalina**, Alberto Rezola, José Manuel García, John E. Beasley and Francisco J. Planes (2015) TreeEFM: calculating elementary flux modes using linear optimization in a tree-based algorithm. *Bioinformatics* 31 (6): 897-904

Alberto Rezola, Jon Pey, **Luis Tobalina**, Ángel Rubio, John E. Beasley and Francisco J. Planes (2014) Advances in network-based metabolic pathway analysis and gene expression data integration. *Briefings in Bioinformatics* 16 (2): 265-279

Jon Pey*, **Luis Tobalina***, Joaquín Prada J. de Cisneros and Francisco J. Planes (2013) A network-based approach for predicting key enzymes explaining metabolite abundance alterations in a disease phenotype. *BMC Systems Biology* 7:62 *(* equal contribution)*

Conference contributions

Luis Tobalina, Jon Pey and Francisco J. Planes. Direct Calculation of Minimal Cut Sets Involving a Specific Reaction Knock-out. *4th Conference on Constraint-Based Reconstruction and Analysis (COBRA 2015)* September 16-18, 2015 - Heidelberg *(Poster Presentation)*

Luis Tobalina, Jon Pey, Joaquín Prada J. de Cisneros and Francisco J. Planes. Detecting key enzymes responsible for metabolite accumulation using Carbon Flux Paths and Connectivity Curves. *Libro de Actas XXX CASEIB 2012* – 19 a 21 de noviembre – San Sebastián. (ISBN 978-84-616-2147-7) *(Oral Presentation)*

Bibliography

Aebersold, R. and Mann, M. Mass spectrometry-based proteomics. *Nature*, vol. 422(6928), pages 198–207, 2003.

Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I. and Nielsen, J. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Computational Biology*, vol. 8(5), page e1002518, 2012.

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

Almaas, E., Oltvai, Z. N. and Barabási, A.-L. The activity reaction core and plasticity of metabolic networks. *PLoS Computational Biology*, vol. 1(7), page e68, 2005.

Bababeygy, S. R., Polevaya, N. V., Youssef, S., Sun, A., Xiong, A., Prugpichailers, T., Veeravagu, A., Hou, L. C., Steinman, L. and Tse, V. HMG-CoA reductase inhibition causes increased necrosis and apoptosis in an in vivo mouse glioblastoma multiforme model. *Anticancer Research*, vol. 29(12), pages 4901–4908, 2009.

Bachmann, H., Fischlechner, M., Rabbers, I., Barfa, N., Branco dos Santos, F., Molenaar, D. and Teusink, B. Availability of public goods shapes the evolution of competing metabolic strategies. *Proceedings of the National Academy of Sciences*, vol. 110(35), pages 14302–14307, 2013.

Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Research*, vol. 28(1), pages 304–305, 2000.

Ballerstein, K., von Kamp, A., Klamt, S. and Haus, U.-U. Minimal cut sets in a metabolic network are elementary modes in a dual network. *Bioinformatics*, vol. 28(3), pages 381–387, 2012.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehar, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jane-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspesi, P., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R. and Garraway, L. A. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, vol. 483(7391), pages 603–307, 2012.

Becker, S. A. and Palsson, B. O. Context-specific metabolic networks are consistent with experiments. *PLoS Computational Biology*, vol. 4(5), page e1000082, 2008.

Beloqui, A., de María, P. D., Golyshin, P. N. and Ferrer, M. Recent trends in industrial microbiology. *Current Opinion in Microbiology*, vol. 11(3), pages 240 – 248, 2008.

Bordbar, A. and Palsson, B. O. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *Journal of Internal Medicine*, vol. 271(2), pages 131–141, 2012.

Borenstein, E. Computational systems biology and in silico modeling of the human microbiome. *Briefings in Bioinformatics*, vol. 13(6), pages 769–780, 2012.

Brochado, A. R., Andrejev, S., Maranas, C. D. and Patil, K. R. Impact of stoichiometry representation on simulation of genotype-phenotype relationships in metabolic networks. *PLoS Computational Biology*, vol. 8(11), page e1002758, 2012.

Carter, K. K., Valdes, J. J. and Bentley, W. E. Pathway engineering via quorum sensing and sRNA riboregulators—Interconnected networks and controllers. *Metabolic Engineering*, vol. 14(3), pages 281 – 288, 2012.

Chandrasekaran, S. and Price, N. D. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia* *coli* and *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, vol. 107(41), pages 17845–17850, 2010.

Chen, Z., Li, D., Cheng, Q., Ma, Z., Jiang, B., Peng, R., Chen, R., Cao, Y. and Wan, X. MicroRNA-203 inhibits the proliferation and invasion of U251 glioblastoma cells by directly targeting PLD2. *Molecular medicine reports*, vol. 9(2), pages 503–508, 2014.

Cheung, H. W., Cowley, G. S., Weir, B. A., Boehm, J. S., Rusin, S., Scott, J. A., East, A., Ali, L. D., Lizotte, P. H., Wong, T. C., Jiang, G., Hsiao, J., Mermel, C. H., Getz, G., Barretina, J., Gopal, S., Tamayo, P., Gould, J., Tsherniak, A., Stransky, N., Luo, B., Ren, Y., Drapkin, R., Bhatia, S. N., Mesirov, J. P., Garraway, L. A., Meyerson, M., Lander, E. S., Root, D. E. and Hahn, W. C. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences*, vol. 108(30), pages 12372–12377, 2011.

Clarke, B. Stoichiometric network analysis. *Cell Biophysics*, vol. 12(1), pages 237–253, 1988.

Clemente-Soto, A. F., Balderas-Rentería, I., Rivera, G., Segura-Cabrera, A., Garza-González, E. and del Rayo Camacho-Corona, M. Potential mechanism of action of *meso*-dihydroguaiaretic acid on *Mycobacterium tuberculosis* H37Rv. *Molecules*, vol. 19(12), pages 20170–20182, 2014.

Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., Cheng, T.-Y., Moody, D. B., Murray, M. and Galagan, J. E. Interpreting expression data with metabolic flux models: Predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Computational Biology*, vol. 5(8), page e1000489, 2009.

Cowley, G. S., Weir, B. A., Vazquez, F., Tamayo, P., Scott, J. A., Rusin, S., East-Seletsky, A., Ali, L. D., Gerath, W. F., Pantel, S. E., Lizotte, P. H., Jiang, G., Hsiao, J., Tsherniak, A., Dwinell, E., Aoyama, S., Okamoto, M., Harrington, W., Gelfand, E., Green, T. M., Tomko, M. J., Gopal, S., Wong, T. C., Li, H., Howell, S., Stransky, N., Liefeld, T., Jang, D., Bistline, J., Hill Meyers, B., Armstrong, S. A., Anderson, K. C., Stegmaier, K., Reich, M., Pellman, D., Boehm, J. S., Mesirov, J. P., Golub, T. R., Root, D. E. and Hahn, W. C. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific Data*, vol. 1, 2014.

Dang, L., White, D. W., Gross, S., Bennett, B. D., Bittinger, M. A., Driggers, E. M., Fantin, V. R., Jang, H. G., Jin, S., Keenan, M. C., Marks, K. M., Prins, R. M., Ward, P. S., Yen, K. E., Liau, L. M., Rabinowitz, J. D., Cantley, L. C., Thompson, C. B., Vander Heiden, M. G. and Su, S. M. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature*, vol. 462(7274), pages 739–744, 2009.

Dettmer, K., Aronov, P. A. and Hammock, B. D. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, vol. 26(1), pages 51–78, 2007.

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R. and Palsson, B. . Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, vol. 104(6), pages 1777–1782, 2007.

Edgar, R., Domrachev, M. and Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, vol. 30(1), pages 207–210, 2002.

Edwards, J. S. and Palsson, B. O. Metabolic flux balance analysis and the in silico analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics*, vol. 1(1), pages 1–1, 2000.

Erdrich, P., Knoop, H., Steuer, R. and Klamt, S. Cyanobacterial biofuels: new insights and strain design strategies revealed by computational modeling. *Microbial Cell Factories*, vol. 13(1), 2014.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V. and Palsson, B. Ø. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, vol. 3(1), 2007.

Feist, A. M. and Palsson, B. O. The biomass objective function. *Current Opinion in Microbiology*, vol. 13(3), pages 344 – 349, 2010.

de Figueiredo, L. F., Podhorski, A., Rubio, A., Kaleta, C., Beasley, J. E., Schuster, S. and Planes, F. J. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, vol. 25(23), pages 3158–3165, 2009.

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E. and Shlomi, T. Predicting selective drug targets in cancer through metabolic networks. *Molecular Systems Biology*, vol. 7, 2011.

Frezza, C., Zheng, L., Folger, O., Rajagopalan, K. N., MacKenzie, E. D., Jerby, L., Micaroni, M., Chaneton, B., Adam, J., Hedley, A., Kalna, G., Tomlinson, I. P. M., Pollard, P. J., Watson, D. G., Deberardinis, R. J., Shlomi, T., Ruppin, E. and Gottlieb, E. Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature*, vol. 477(7363), pages 225–228, 2011.

Gliemroth, J., Zulewski, H., Arnold, H. and Terzis, A. Migration, proliferation, and invasion of human glioma cells following treatment with simvastatin. *Neurosurgical Review*, vol. 26(2), pages 117–124, 2003.

Greenblum, S., Turnbaugh, P. J. and Borenstein, E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences*, vol. 109(2), pages 594–599, 2012.

Guazzaroni, M.-E. and Ferrer, M. Metagenomic approaches in Systems Biology. In *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches* (edited by F. J. de Bruijn), vol. 1, chapter 54, pages 475–489. Wiley-Blackwell, John Wiley & Sons, Inc, first edition, 2011.

Guazzaroni, M.-E., Herbst, F.-A., Lores, I., Tamames, J., Pelaez, A. I., Lopez-Cortes, N., Alcaide, M., Del Pozo, M. V., Vieites, J. M., von Bergen, M., Gallego, J. L. R., Bargiela, R., Lopez-Lopez, A., Pieper, D. H., Rossello-Mora, R., Sanchez, J., Seifert, J. and Ferrer, M. Metaproteogenomic insights beyond bacterial response to naphthalene exposure and bio-stimulation. *ISME Journal*, vol. 7(1), pages 122–136, 2013.

Gudmundsson, S. and Thiele, I. Computationally efficient flux variability analysis. *BMC Bioinformatics*, vol. 11(1), page 489, 2010.

Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. and Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular Systems Biology*, vol. 10(7), 2014.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B. and Stevens, R. L. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, vol. 28(9), pages 977–982, 2010.

Hidde Boersma, F., Colin McRoberts, W., Cobb, S. L. and Murphy, C. D. A ^{19}F NMR study of fluorobenzoate biodegradation by *Sphingomonas* sp. HB-1. *FEMS Microbiology Letters*, vol. 237(2), pages 355–361, 2004.

Jensen, P. A. and Papin, J. A. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics*, vol. 27(4), pages 541–547, 2011.

Jerby, L., Shlomi, T. and Ruppin, E. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Molecular Systems Biology*, vol. 6, 2010.

Kaelin, W. G. The concept of synthetic lethality in the context of anticancer therapy. *Nature Reviews Cancer*, vol. 5(9), pages 689–698, 2005.

Kaelin, W. G. and Thompson, C. B. Q&A: Cancer: Clues from cell metabolism. *Nature*, vol. 465(7298), pages 562–564, 2010.

Kaleta, C., de Figueiredo, L. F., Behre, J. and Schuster, S. EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. In *Lecture Notes in Informatics-Proceedings*, vol. 157, pages 179–189. 2009.

von Kamp, A. and Klamt, S. Enumeration of smallest intervention strategies in genome-scale metabolic networks. *PLoS Computational Biology*, vol. 10(1), page e1003378, 2014.

Kamp, A. v. and Schuster, S. Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, vol. 22(15), pages 1930–1931, 2006.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, vol. 40(D1), pages D109–D114, 2012.

Kang, D. W., Choi, K.-Y. and Min, D. S. Functional regulation of phospholipase D expression in cancer and inflammation. *Journal of Biological Chemistry*, vol. 289(33), pages 22575–22582, 2014.

Åkesson, M., Förster, J. and Nielsen, J. Integration of gene expression data into genome-scale metabolic models. *Metabolic Engineering*, vol. 6(4), pages 285 – 293, 2004.

Khandelwal, R. A., Olivier, B. G., Röling, W. F. M., Teusink, B. and Bruggeman, F. J. Community flux balance analysis for microbial consortia at balanced growth. *PLoS ONE*, vol. 8(5), page e64567, 2013.

Kinross, J., Darzi, A. and Nicholson, J. Gut microbiome-host interactions in health and disease. *Genome Medicine*, vol. 3(3), page 14, 2011.

Kitano, H. Computational systems biology. *Nature*, vol. 420(6912), pages 206–210, 2002a.

Kitano, H. Systems biology: A brief overview. *Science*, vol. 295(5560), pages 1662–1664, 2002b.

Klamt, S. Generalized concept of minimal cut sets in biochemical networks. *Biosystems*, vol. 83(2–3), pages 233 – 247, 2006.

Klamt, S. and Gilles, E. D. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, vol. 20(2), pages 226–234, 2004.

Klamt, S. and Mahadevan, R. On the feasibility of growth-coupled product synthesis in microbial strains. *Metabolic Engineering*, vol. 30, pages 166 – 178, 2015.

Kästner, M. *Degradation of aromatic and polyaromatic compounds*, pages 211–239. Wiley-VCH Verlag GmbH, 2008. ISBN 9783527620999.

Larhlimi, A. and Bockmayr, A. A new constraint-based description of the steady-state flux cone of metabolic networks. *Discrete Applied Mathematics*, vol. 157(10), pages 2257 – 2266, 2009.

Lee, Y., Vassilakos, A., Feng, N., Lam, V., Xie, H., Wang, M., Jin, H., Xiong, K., Liu, C., Wright, J. and Young, A. GTI-2040, an antisense agent targeting the small subunit component (R2) of human ribonucleotide reductase, shows potent antitumor activity against a variety of tumors. *Cancer Research*, vol. 63(11), pages 2802–2811, 2003.

Letunic, I., Yamada, T., Kanehisa, M. and Bork, P. iPath: interactive exploration of biochemical pathways and networks. *Trends in Biochemical Sciences*, vol. 33(3), pages 101 – 103, 2008.

Llaneras, F. and Picó, J. Stoichiometric modelling of cell metabolism. *Journal of Bioscience and Bioengineering*, vol. 105(1), pages 1 – 11, 2008.

Lu, X.-Y., Zhang, T. and Fang, H.-P. Bacteria-mediated PAH degradation in soil and sediment. *Applied Microbiology and Biotechnology*, vol. 89(5), pages 1357–1371, 2011.

Luo, B., Cheung, H. W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J. S., Beroukhim, R., Weir, B. A., Mermel, C., Barbie, D. A., Awad, T., Zhou, X., Nguyen, T., Piqani, B., Li, C., Golub, T. R., Meyerson, M., Hacohen, N., Hahn, W. C., Lander, E. S., Sabatini, D. M. and Root, D. E. Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences*, vol. 105(51), pages 20380–20385, 2008.

Machado, D., Soons, Z., Patil, K. R., Ferreira, E. C. and Rocha, I. Random sampling of elementary flux modes in large-scale metabolic networks. *Bioinformatics*, vol. 28(18), pages i515–i521, 2012.

Mahadevan, R. and Schilling, C. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, vol. 5(4), pages 264 – 276, 2003.

Mascarelli, A. L. Geomicrobiology: Low life. *Nature*, vol. 459(7248), pages 770–773, 2009.

McCall, M. N., Bolstad, B. M. and Irizarry, R. A. Frozen robust multiarray analysis (fRMA). *Biostatistics*, vol. 11(2), pages 242–253, 2010.

McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J. and Irizarry, R. A. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, vol. 39(suppl 1), pages D1011–D1015, 2011.

McCloskey, D., Palsson, B. O. and Feist, A. M. Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Molecular Systems Biology*, vol. 9, 2013.

McQuaker, N. R. and Gurney, M. Determination of total fluoride in soil and vegetation using an alkali fusion-selective ion electrode technique. *Analytical Chemistry*, vol. 49(1), pages 53–56, 1977.

Moran, M. A., Satinsky, B., Gifford, S. M., Luo, H., Rivers, A., Chan, L.-K., Meng, J., Durham, B. P., Shen, C., Varaljay, V. A., Smith, C. B., Yager, P. L. and Hopkinson, B. M. Sizing up metatranscriptomics. *ISME Journal*, vol. 7(2), pages 237–243, 2013.

Noble, D. Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations. *Nature*, vol. 188(4749), pages 495–497, 1960.

Nogales, J., Palsson, B. and Thiele, I. A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. *BMC Systems Biology*, vol. 2(1), page 79, 2008.

Orth, J. D., Thiele, I. and Palsson, B. O. What is flux balance analysis? *Nature Biotechnology*, vol. 28(3), pages 245–248, 2010.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F. and Stevens, R. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, vol. 42(D1), pages D206–D214, 2014.

Palsson, B. *Systems biology : properties of reconstructed networks*. Cambridge University Press, New York, NY, USA, 2006.

Palsson, B. Metabolic systems biology. *FEBS Letters*, vol. 583(24), pages 3900 – 3904, 2009.

Pey, J. and Planes, F. J. Direct calculation of elementary flux modes satisfying several biological constraints in genome-scale metabolic networks. *Bioinformatics*, vol. 30(15), pages 2197–2203, 2014.

Pey, J., Prada, J., Beasley, J. and Planes, F. Path finding methods accounting for stoichiometry in metabolic networks. *Genome Biology*, vol. 12(5), page R49, 2011.

Pey, J., Villar, J. A., Tobalina, L., Rezola, A., García, J. M., Beasley, J. E. and Planes, F. J. TreeEFM: calculating elementary flux modes using linear optimization in a tree-based algorithm. *Bioinformatics*, vol. 31(6), pages 897–904, 2015.

Pfeiffer, T., Snchez-Valdenebro, I., Nuo, J. C., Montero, F. and Schuster, S. METATOOL: for studying metabolic networks. *Bioinformatics*, vol. 15(3), pages 251–257, 1999.

Powers, R. NMR metabolomics and drug discovery. *Magnetic Resonance in Chemistry*, vol. 47(S1), pages S2–S11, 2009.

Pérez-Cobas, A. E., Gosalbes, M. J., Friedrichs, A., Knecht, H., Artacho, A., Eismann, K., Otto, W., Rojo, D., Bargiela, R., von Bergen, M., Neulinger, S. C., Däumer, C., Heinsen, F.-A., Latorre, A., Barbas, C., Seifert, J., dos Santos, V. M., Ott, S. J., Ferrer, M. and Moya, A. Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut*, 2012.

Quek, L.-E. and Nielsen, L. K. A depth-first search algorithm to compute elementary flux modes by linear programming. *BMC systems biology*, vol. 8(1), page 94, 2014.

Rezola, A., de Figueiredo, L. F., Brock, M., Pey, J., Podhorski, A., Wittmann, C., Schuster, S., Bockmayr, A. and Planes, F. J. Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics*, vol. 27(4), pages 534–540, 2011.

Rezola, A., Pey, J., Tobalina, L., Rubio, n., Beasley, J. E. and Planes, F. J. Advances in network-based metabolic pathway analysis and gene expression data integration. *Briefings in Bioinformatics*, vol. 16(2), pages 265–279, 2015.

Rezola-Urquía, A. *A novel systems biology framework to contextualize metabolic processes using elementary flux modes and gene expression data*. Phd, University of Navarra, 2013.

Röling, W. F., Ferrer, M. and Golyshin, P. N. Systems approaches to microbial communities and their functioning. *Current Opinion in Biotechnology*, vol. 21(4), pages 532 – 538, 2010.

Rossell, S., Huynen, M. A. and Notebaart, R. A. Inferring metabolic states in uncharacterized environments using gene-expression measurements. *PLoS Computational Biology*, vol. 9(3), page e1002988, 2013.

dos Santos, F. B., de Vos, W. M. and Teusink, B. Towards metagenome-scale models for industrial applications — the case of Lactic Acid Bacteria. *Current Opinion in Biotechnology*, vol. 24(2), pages 200 – 206, 2013.

Satish Kumar, V., Dasika, M. and Maranas, C. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, vol. 8(1), page 212, 2007.

Schellenberger, J., Park, J., Conrad, T. and Palsson, B. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, vol. 11(1), page 213, 2010.

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R. and Palsson, B. O. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature Protocols*, vol. 6(9), pages 1290–1307, 2011.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, vol. 270(5235), pages 467–470, 1995.

Schilling, C. H., Letscher, D. and Palsson, B. O. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, vol. 203(3), pages 229 – 248, 2000.

Schilling, C. H., Schuster, S., Palsson, B. O. and Heinrich, R. Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era. *Biotechnology Progress*, vol. 15(3), pages 296–303, 1999.

Schmidt, B. J., Ebrahim, A., Metz, T. O., Adkins, J. N., Palsson, B. . and Hyduke, D. R. GIM^{3}E: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics*, vol. 29(22), pages 2900–2908, 2013.

Schneider, M. Defining systems biology: A brief overview of the term and field. In *In Silico Systems Biology* (edited by M. V. Schneider), vol. 1021 of *Methods in Molecular Biology*, pages 1–11. Humana Press, 2013. ISBN 978-1-62703-449-4.

Schuster, S. and Hilgetag, C. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, vol. 02(02), pages 165–182, 1994.

Seifert, J., Herbst, F.-A., Halkjær Nielsen, P., Planes, F. J., Jehmlich, N., Ferrer, M. and von Bergen, M. Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities. *Proteomics*, vol. 13(18-19), pages 2786–2804, 2013.

Shlomi, T., Cabili, M. N., Herrgard, M. J., Palsson, B. O. and Ruppin, E. Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology*, vol. 26(9), pages 1003–1010, 2008.

Shlomi, T., Eisenberg, Y., Sharan, R. and Ruppin, E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular Systems Biology*, vol. 3(1), 2007.

Suthers, P. F., Zomorrodi, A. and Maranas, C. D. Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Molecular Systems Biology*, vol. 5(1), 2009.

Terzer, M. and Stelling, J. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, vol. 24(19), pages 2229–2235, 2008.

Thiele, I. and Palsson, B. O. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, vol. 5(1), pages 93–121, 2010.

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bolling, C., Bordel, S., Chavali, A. K., Dobson, P., Dunn, W. B., Endler, L., Hala, D., Hucka, M., Hull, D., Jameson, D., Jamshidi, N., Jonsson, J. J., Juty, N., Keating, S., Nookaew, I., Le Novere, N., Malys, N., Mazein, A., Papin, J. A., Price, N. D., Selkov, E., Sigurdsson, M. I., Simeonidis, E., Sonnenschein, N., Smallbone, K., Sorokin, A., van Beek, J. H. G. M., Weichart, D., Goryanin, I., Nielsen, J., Westerhoff, H. V., Kell, D. B., Mendes, P. and Palsson, B. O. A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, vol. 31(5), pages 419–425, 2013.

Tobalina, L., Bargiela, R., Pey, J., Herbst, F.-A., Lores, I., Rojo, D., Barbas, C., Peláez, A. I., Sánchez, J., von Bergen, M., Seifert, J., Ferrer, M. and Planes, F. J. Context-specific metabolic network reconstruction of a naphthalene-degrading bacterial community guided by metaproteomic data. *Bioinformatics*, vol. 31(11), pages 1771–1779, 2015.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Bjorling, L. and Ponten, F. Towards a knowledge-based Human Protein Atlas. *Nature Biotechnology*, vol. 28(12), pages 1248–1250, 2010.

Urbanczik, R. and Wagner, C. An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics*, vol. 21(7), pages 1203–1210, 2005.

Vander Heiden, M. G. Targeting cancer metabolism: a therapeutic window opens. *Nature Reviews Drug Discovery*, vol. 10(9), pages 671–684, 2011.

Vander Heiden, M. G., Cantley, L. C. and Thompson, C. B. Understanding the Warburg effect: The metabolic requirements of cell proliferation. *Science*, vol. 324(5930), pages 1029–1033, 2009.

Vander Heiden, M. G., Locasale, J. W., Swanson, K. D., Sharfi, H., Heffron, G. J., Amador-Noguez, D., Christofk, H. R., Wagner, G., Rabinowitz, J. D., Asara, J. M. and Cantley, L. C. Evidence for an alternative glycolytic pathway in rapidly proliferating cells. *Science*, vol. 329(5998), pages 1492–1499, 2010.

Vanderbei, R. Linear programming: Foundations and extensions. no. 4 in international series in operations research & management. 1996.

Vitkin, E. and Shlomi, T. MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. *Genome Biology*, vol. 13(11), page R111, 2012.

Vlassis, N., Pacheco, M. P. and Sauter, T. Fast reconstruction of compact context-specific metabolic network models. *PLoS Computational Biology*, vol. 10(1), page e1003424, 2014.

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., MacInnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J. and Querengesser, L. HMDB: the Human Metabolome Database. *Nucleic Acids Research*, vol. 35(suppl 1), pages D521–D526, 2007.

Woo, I. S., Eun, S. Y., Kim, H. J., Kang, E. S., Kim, H. J., Lee, J. H., Chang, K. C., Kim, J.-H., Hong, S.-C. and Seo, H. G. Farnesyl diphosphate synthase attenuates paclitaxel-induced apoptotic cell death in human glioblastoma U87MG cells. *Neuroscience Letters*, vol. 474(2), pages 115 – 120, 2010.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M. and Bork, P. iPath2.0: interactive pathway explorer. *Nucleic Acids Research*, vol. 39(suppl 2), pages W412–W415, 2011.

Yizhak, K., Gaude, E., Le Dévédec, S., Waldman, Y. Y., Stein, G. Y., van de Water, B., Frezza, C. and Ruppin, E. Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *eLife*, vol. 3, 2014a.

Yizhak, K., Le Dévédec, S. E., Rogkoti, V. M., Baenke, F., de Boer, V. C., Frezza, C., Schulze, A., van de Water, B. and Ruppin, E. A computational study of the Warburg effect identifies metabolic targets inhibiting cancer migration. *Molecular Systems Biology*, vol. 10(8), 2014b.

Zanghellini, J., Ruckerbauer, D. E., Hanscho, M. and Jungreuthmayer, C. Elementary flux modes in a nutshell: Properties, calculation and applications. *Biotechnology Journal*, vol. 8(9), pages 1009–1016, 2013.

Zomorrodi, A. R. and Maranas, C. D. OptCom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Computational Biology*, vol. 8(2), page e1002363, 2012.

Zomorrodi, A. R., Suthers, P. F., Ranganathan, S. and Maranas, C. D. Mathematical optimization applications in metabolic networks. *Metabolic Engineering*, vol. 14(6), pages 672 – 686, 2012.

Zur, H., Ruppin, E. and Shlomi, T. iMAT: an integrative metabolic analysis tool. *Bioinformatics*, vol. 26(24), pages 3140–3142, 2010.