Aalto University
School of Science
Degree Programme in Computer Science

Maja Ilievska

Identification of metabolic fluxes leading to the production of industrially relevant products

Master's Thesis
Espoo, May 9, 2016

Supervisors: Prof. Juho Rousu

Advisors: Elena Czeizler, Ph.D.
Peter Blomberg

Aalto University
School of Science
Degree Programme in Computer Science

ABSTRACT OF
MASTER'S THESIS

Author:	Maja Ilievska		
Title:	Identification of metabolic fluxes leading to the production of industrially relevant products		
Date:	May 9, 2016	**Pages:**	58
Major:	Bioinformatics	**Code:**	T901-D
Supervisors:	Prof. Juho Rousu		
Advisors:	Elena Czeizler, Ph.D. Peter Blomberg		

In metabolic pathway analysis the focus is on identifying the complete range of paths within a biochemical network. However, most current methods characterizing all potential paths between the selected substrates and product are based either on the enumeration of all elementary flux modes or all extreme pathways. This becomes computationally unfeasible for large reaction matrices. In this work, we propose an alternative approach that identifies a set of potential paths while avoiding an exhaustive enumeration. More specifically, we identify a set of (minimal) flux vectors that produce the desired product and do not accumulate any intermediates while consuming at least one of the specified substrates. Our k-best approach uses linear programming to identify the first k solutions, according to a pre-defined objective function. Furthermore, in order to determine biologically more meaningful flux vectors we define an augmented solution space, where in addition to the flux distribution we incorporate the net consumption/production of external metabolites and the contribution of the null space basis vectors to the given flux distribution.

One of the main aims of this research was to computationally determine the best substrate-path-product combination for industrial scale production. In fact, we were interested in identifying the best carbon source (or the best combination of different carbon sources) that will lead to the highest productivity for a specific product, as well as the best metabolic pathway from the identified sources to the product. A special focus within this work was the identification of an objective function for the enumerated paths, which would return a good set of candidate paths.

The results demonstrate that our k-best method is able to identify a set of candidate pathways for genome-scale metabolic models, where elementary modes and extreme pathway analysis fail to provide a resulting set of pathways. Among the pathways proposed by our enumeration approach there are novel ones with the potential to improve the production processes of the specific product in terms of energetic efficiency.

Keywords: metabolic pathway analysis, flux vectors, right null space, constrained-based stoichiometric modelling, industrial products

Language: English

Acknowledgements

I would like to thank my advisors, Elena Czeizler and Peter Blomberg, for their continuous support and guidance throughout this work. I also want to thank my supervisor Juho Rousu for his support and patience.

Thank you to my friends and family for the friendship and support, and for always being there for me.

Espoo, May 9, 2016

Maja Ilievska

Contents

Chapter 1

Introduction

The technological advancements of the 20th century that led to improvements in transport, energy and food production are less and less able to meet the needs of the increasing population of the 21st century. With environmental issues on the rise and the scarcity of petroleum resources, there is an obvious need for new technologies and resources to replace the existing ones, in a manner that will be sustainable on the long run [11]. From petroleum based industry the focus has recently moved to industrial microbiology, where micro-organisms are utilized as biocatalysts relying on renewable resources as substrates [38]. In order to be viable for large scale production and compete with the established industrial processes, these cell factories have to overcome the challenges of selecting a renewable carbon source, establishing the production process and the purification of the product. If these are to be implemented on large scale, they will result in lowering the carbon emissions and will set the basis of bioeconomy. Bioeconomy refers to economy resulting from scientific research in the field of biotechnology, where the focus is on understanding the processes on genetic and molecular level and utilizing it for directed improvement of industrial processes [2].

Closely related to the development of the microbial cell factories, the field of metabolic engineering has emerged [3], which refers to cyclic process of analysis and engineering of the desired (microbial) strain. Since the emergence of the first rudimental microbial cell factory - the yeast ethanol fermentation for beer production in the 19th century - developments like recombinant DNA technologies, genomics and high-throughput sequencing technologies have led to major improvements in metabolic engineering. Herein, the initial step is the identification of the desired product. Next, potential substrates are analysed and suitable host strain is selected based on the capabilities to perform the desired metabolic conversions, and the affinity to interact with both substrate and product [44]. From here, genetic data can be used for generation of genome-scale metabolic models where we account for the metabolic reactions (fluxes) taking place and their participating metabolites. Finally, high-throughput omics technologies alongside computational modelling and simulation can be used to identify possible engineering targets, such as gene insertions or deletions, over-expression of genes and even introduction of heterologous pathways [24]. Furthermore it is very important to consider the physiological and genetic background of the host organism which are essential for understanding the conditions for cell survival, generation of appropriate intracellular environment and gene regulation [33]. As a result, complex systems level analysis is necessary in order to yield a successful metabolic engineering process. In this work we are interested in developing and applying a computational approach for substrate-path-product combination for industrial scale production. More specifically, we are interested in identifying the best carbon source (or the best combination of different carbon sources) that will lead to the highest yield for a specific product, as well as the best metabolic pathway from the identified sources to the product. Given that we want to expand the set of feasible substrates and products of a certain organism (i.e. S. Cerevisae), we need to extend the native metabolic network of that organism by adding new reactions connecting existent metabolites to certain metabolites of interest. However, since the search space of new reactions that could be added to the network is huge, we need computational methods to predict the best set of novel reactions that will give the highest product yield when added to the initial metabolic network.

Once a metabolic model has been generated based on the genetic and metabolic information, several methods can be employed for identification of desirable production pathways. Usually they incorporate a cellular objective function (e.g. biomass maximization) that generates the path with the most extreme value for the objective. However, in general the cells are subject to a complex combination of physicochemical, topobiological, environmental and regulatory factors that govern the cell behaviour and they have to be taken into account when generating the mathematical model. These limiting factors or constraints are represented in mathematical form in terms of balances and bounds. Balanced are constraints related to conserved quantities such as energy and mass, and bounds are the constraints that limit the numerical range of variables such as flux rates. Once the assumption of pseudo steady state for internal metabolites is adopted, where no accumulation of metabolites occurs, a general so called constraint-based stoichiometric model can be constructed incorporating the bounds and balances [35]. We say that the internal metabolites are in pseudo steady state because their dynamics are very fast and we can assume that they reach the steady state or equilibrium instantaneously, which is not the case for external metabolites outside of the cell [25]. Computational analysis of the stoichiometric models identifies feasible steady states represented through flux distributions. The set of constraints used in the stoichiometric model are not sufficient to generate a unique flux distribution for the given metabolic network. Instead there exists a set of flux distributions that define the so called solution space. This solutions space is to be further examined and potential "good" solutions are drawn from the pool of solutions [26].

When we use the flux balance approach based on an objective function, we might not always identify the solutions of interest and there might be other solutions relevant for the problem of interest [37]. As a result methods such as elementary flux modes (EFMs) [40, 41] and extreme pathways [32] have been developed that can enumerate all potential paths and then rank them by using certain evaluation criteria Nevertheless, most current methods characterizing all potential paths between the selected substrates and product rely on exhaustive enumerations that can generate up to several millions paths for a genome scale models which can become computationally unfeasible at times. Herein we propose an alternative approach that identifies a set of potential paths while avoiding an exhaustive enumeration. More specifically, we identify a set of (minimal) flux vectors that produce the desired product and do not accumulate any intermediates while consuming at least one of the specified substrates. Our k-best approach uses linear programming to identify the first k solutions. The method is based on exclusion constraints that have been implemented for enumeration of shortest elementary flux modes [10]. We have adopted the approach and combined it with our constraint based stoichiometric model. As a result we generate a set of suitable pathways, that connect source metabolites to a target metabolite. As baseline methods we have used elementary flux modes and extreme pathways. We have compared the performance of out method against these established enumeration techniques as well as the number of paths retrieved by each of these approaches.

From industrial perspective we are interested in identifying the minimal number of modification to the original metabolic network for a given organism, including additions of new reactions. Initially, a biological expert has identified the set of reactions to be potentially included to the network and the possible sets of objectives that could be used to evaluate the results in a biologically relevant way. A special focus within this work is the identification of an objective function for the enumerated paths, which would return a good set of candidate paths. An iterative approach has been implemented where we first enumerate potential paths according the preselected objectives and then we analyse the feasibility of the proposed paths. Biochemical analysis terms the following factors as important for the path ranking: smallest number of new enzymes (reactions) and the total overall number of enzymes, cofactor balancing, maximal carbon yield etc. Furthermore, the analysis is extended to overall net reaction and thermodynamic feasibility. The analysis proposed different outcomes depending on the criteria used. Good potential paths were identified with different sets of objectives. Thus it was evident that a further computational analysis and deeper enumeration of pathways is required in order to single out the computational approach that would result in better selection of relevant paths. Some of the existing enumeration methods are based on the right null space analysis where the only input is the stoichiometric matrix and from there the basis vectors are identified that can generate all steady state distributions. It is important to realize which of the basis vectors are biologically most relevant in terms of their contribution to the primary metabolism [6]. We wanted to account for the importance of the basis vectors when identifying flux distributions and therefore we described an augmented solution space. In this augmented solution space, along the flux distribution we optimize the net consumption/production of the external metabolites and the contribution of the null space basis to the given flux distribution.

The work in the thesis is initiated by the need of providing a computational framework that can complement the work of biochemical experts in identifying ways for economically feasible and sustainable, large scale production of industrial products utilizing microbes as the production hosts. The project is undertaken as part of large initiative at VTT Technical Research Centre of Finland, project named Living Factories [1]. The project aims at uncovering the full potential of Synthetic Biology in Finland given that it has the potential to become a major player in shaping our future economies.

This report is structures as follows. Chapter 2 gives the background in metabolic pathway analysis and the standard computational tools for performing the analysis. It also provides the required background and concepts upon which we base our work. In chapter 3 we introduce the data and the detailed analysis we have performed. In here, we formulate our k-best enumeration method and we define the combined solution space. Chapter 4 contains the results obtained by our methods in comparison to the established computational tools. Finally, chapter 6 concludes our work.

[1] http://www.vtt.fi/sites/livingfactories/en

Chapter 2

Background

2.1 Mathematical representation of metabolic networks

The recent advancements in high-throughput technologies have given rise to disciplines such as genomics, transcriptomics, metabolomics and fluxomics, that generate wide range of "omics" data. All these data types can be used for reconstruction of cellular biochemical reaction networks [13]. In the case of metabolic networks, the process of reconstruction identifies all the reactions that belong to the network and their participating metabolites, the result of which provides basis for understanding the underlying cellular mechanisms by observing the reaction interactions.

Chemical reactions that belong to a given metabolic network can be represented by means of chemical equations that capture the stoichiometry of the reactions. The stoichiometry of a reaction provides information on the amount of substance that is produced or consumed by the reaction. To allow for computational analysis of metabolic networks, the complete set of equations can be represented in a matrix form, by the so-called stoichiometric matrix, denoted by S. This in turn, enables for translation of the high-throughput data into mathematical form and thus creates analogy between the mathematical and biochemical properties of the network [6]. In this section we outline the basic properties of metabolic networks in a mathematical framework and we provide insight into their interpretation. More specifically, we describe the basic properties of the stoichiometric matrix upon which many computational methods are based. Subsections 2.1.1, 2.1.2 and 2.1.3 are heavily based on [6]. Next, we give an outline to the major representative methods based on stoichiometric modelling and we discuss their strengths and weaknesses.

2.1.1 Basic properties of stoichiometric matrix

The stoichiometric (S) matrix, is structured in a way that each column corresponds to a single reaction in the network, and each row corresponds to a compound. The elements of the matrix are the stoichiometric coefficients of the reactions, and they represent the number of molecules of each of the compounds (chemical species) that are transformed in a given reaction. If we observe one column in the matrix, we get an overview of the compounds participating in the reaction, and herein if a compound is consumed the coefficient has a negative sign, whereas if it is produced the coefficient is positive. The compounds for which the coefficients are zero, do not participate in the reaction. All the reactions are constraint by the chemical law of elemental balance, implying that the number of chemical elements (carbon, oxygen, etc.) has to be equal between the two sides of the equation. By observing the rows of the matrix, we get an insight into how the reactions are interconnected, as for each metabolite we see the reactions in which it contributes. From mathematical point of view, the S matrix transforms the flux vector into a vector that represents the time derivatives of the concentrations of compounds. The flux vector contains the rates for each of the reactions, and the formulation of this transformation is given according to [6].
Given the vector **v** with n reaction rates

$$\mathbf{v} = (v_1, v_2, v_3, ...v_n) \tag{2.1}$$

and **x**, which represents the concentration of metabolites, and has dimension of m

$$\mathbf{x} = (x_1, x_2, .., x_m) \tag{2.2}$$

we can write

$$Sv = \frac{d\mathbf{x}}{dt} \tag{2.3}$$

where the dimension of S is mxn.
The equation 2.3, represents the system of mass balance equations, which at steady state becomes

$$\mathbf{Sv} = 0. \tag{2.4}$$

The concentration of metabolites inside the cell fluctuate over time, but their dynamics are so fast that we can assume that they reach the steady state instantaneously [14]. During steady state no accumulation or depletion of metabolites can take place, consequently the rates of production and consumption of each metabolite have to be equal. In a biological network when

the entire system of equations is in equilibrium it is said that the concentrations of the metabolites are at steady state.
The biochemical interpretation of the stoichiometric matrix indicates linear transformation of the space of reaction activities into the time derivatives of concentration space, and any biochemical transformation can be depicted by the mass balance equation. The stoichiometric matrix has four fundamental subspaces that provide complete interpretation of biochemical networks in terms of the networks dynamics, steady states and time-invariant characteristics. These four spaces are essential for analysis of biochemical networks and they are the row and column space, and right and left null spaces. Here we will give more detailed formulation on the right null space as this is relevant for providing the background for our computational method.

2.1.2 Right null space

Right null space consists of all vectors that satisfy the equation 2.4. We can construct a matrix R that spans the null space and contains the basis vectors as columns.

$$SR = 0 \tag{2.5}$$

If we denote by r the rank of S, the dimension of the right null space is n-r. The importance of the right null space is in that it spans the entire set of steady state flux distributions. From a biological perspective, each of these flux distributions corresponds to a possible functional or phenotypic state of the network.
In linear algebra, a set of vectors is called a basis of vector space if all the vectors in the set are linearly independent and any other vector that belongs to the vector space can be represented as a linear combination of the basis vectors, whose coefficients are denoted as components or coordinates.

2.1.3 The total stoichiometric matrix

When constructing a biochemical network, we have to identify its boundaries and the interaction with the metabolites that exceed those boundaries. Based on whether the reactions involve metabolites within or outside the system boundaries they are divided into internal and exchange reactions. Consequently the fluxes are named internal fluxes, denoted by v_i, and exchange fluxes b_i. Finally, the concentration vector is broken into two components, the internal concentrations x_i and external concentrations c_i. The metabolites that belong to the internal network are those that have to be balanced and therefore are subjected to the mass balance constraint, whereas the metabolites that belong to the external network represent the products, substrates and cofactors which concentration fluctuates. After assembling the internal metabolic network, the exchange reactions that connect substrates and products to the internal metabolites have to be added. Furthermore, transport reactions and reactions accounting for cofactor balancing are further added. We restructure the stoichiometric matrix in order to account for the changes imposed by the system boundaries. We call this restructured matrix the total matrix, and it is depcited as follows

$$\begin{array}{cc}
 & \begin{array}{ccc|c} & \mathrm{v}_i & & \mathrm{b}_i \end{array} \\
\begin{array}{c} \\ \mathrm{x}_i \\ \mathbf{S}_{tot}: \\ \\ \mathrm{c}_i \\ \\ \end{array} & \left(\begin{array}{ccc|c} & & & \\ & & & \\ & & & \\ -- & -- & -- & - \\ & & & \\ & 0 & & \end{array}\right)
\end{array}$$

where the dashed lines represent the separation areas.
In our work we use the total matrix including the internal and exchange concentrations as well as internal and exchange fluxes. However, given that the general notation of flux vector is v, we will further on refer to the entire set of fluxes, both external and internal as to v.

2.2 Stoichiometric modelling

By definition metabolism is the complex set of chemical and physical processes, dedicated to maintenance of life. It can be fully described and analysed based on the biochemical reactions that constitute the metabolic network. Through systemic approach, i.e. studying metabolic pathways as functional units of metabolic systems, we can learn about cellular behaviour and capabilities of metabolic networks [39]. Stoichiometric modelling is a term used for number of methods directed towards quantitative analysis of metabolic pathways. As the name implies, all these methods are based on the stoichiometric matrix and they all share the assumption of pseudo steady state. The pseudo steady state imposes the constraint of mass-balance (2.4) on the internal metabolites [25]. By assuming the steady state, stoichiometric models neglect the dynamic intracellular behaviour, which accounts for the reaction kinetics. However, research results have shown that in the case of lack of intracellular experimental measurements it can be beneficial to simplify the models by omitting the kinetics, and from there explore the structural properties of the network such as existence of simplest pathways as done in S modelling [4]. Each feasible steady state is represented by a flux vector v. Considering the reaction reversibility along the mass-balance constraint, the space of feasible steady state flux distributions can be mathematically formulated as in the following equation. It is referred to as the flux solution space.

$$P = \{v \in R^n : Sv = 0, Dv \geq 0\} \tag{2.6}$$

where D represents a diagonal matrix with elements D_{ii} equal to one is the flux if irreversible, and zero otherwise.

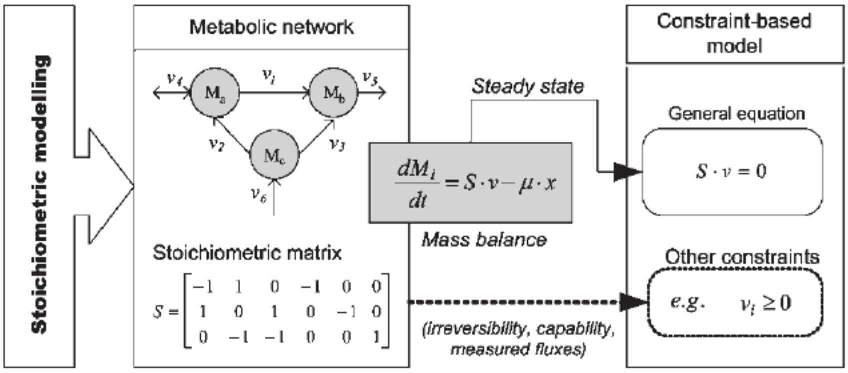
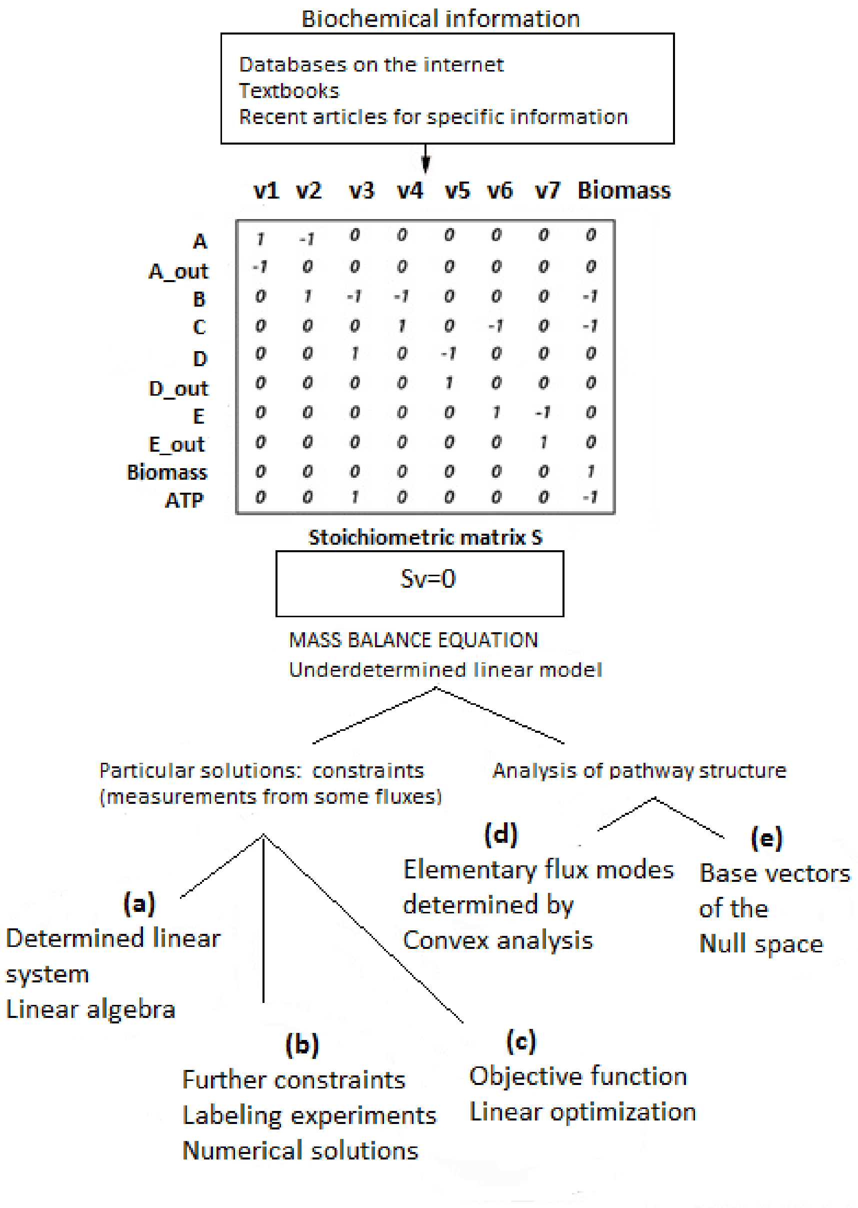


Figure 2.1: Overview of the stoichiometric modelling framework. The initial point is the metabolic network, from which the relevant information on metabolites and reactions is translated into mathematical terms, through the S matrix. Mass balance equation can then be written for each metabolite, which at steady-state equivalents the system of homogeneous linear equations, $Sv = 0$. Further constraints are imposed by the reaction irreversibility, and given the constraint-based model meaningful functional states in the network can be depicted. Figure taken from [25]

Provided the fact that cells are subject to biological constraints that drive their behaviour, the concept of flux space is used as basis for resembling that behaviour in constraint-based modelling approaches. However, noteworthy is the fact that biochemical networks are very complex in nature, consisting of hundreds of connections and governed by interconnected regulatory and control mechanisms, thus all mathematical models to date are based on many simplifications and the true underlying complexity of the cellular processes has not been successfully captured. The constraints can be in general divided in two groups: non adjustable such as those arising from thermodynamic principles like irreversibility of fluxes or enzyme capacities like maximum flux values; and adjustable such as those resembling experimental measurements [25]. Under the given constraints, and the fact that for a typical biochemical network the number of reactions is greater than the number of compounds, we can expect numerous feasible flux distributions to be potentially good solutions. A schematic representation of S modelling is given in Figure 2.1. Depending on the objective of the performed analysis, stoichiometric modelling can be approached from different perspectives. The two major categories are analysis of pathway structure by exploring the whole range of solutions, and analysis of particular solutions under further constraints [14], as illustrated in Figure 2.2. On one hand, we can perform pathway structure analysis based on linear algebra, where we analyse the null space of the S matrix that contains all possible steady state distributions. From here it is possible to focus on the basis vectors, that represent biologically meaningful states and will provide insight into the structural properties of the network at hand. Also, when interested in structural analysis we can perform convex analysis to identify the so called elementary flux modes (EFMs). On the other hand, there are pathway analysis aiming to single out smaller subset of the solution space, based on either experimental measurements of given fluxes or pre-defined objective function. Herein, focusing on the possibilities to explore subset of solutions, flux balance analysis (FBA) is a commonly used framework providing an insight into flux distribution that represent extreme values of the flux space. FBA assumes that cell behaviour is optimal with respect to a given objective, and then the optimal flux distribution is calculated using an optimization procedure [30]. FBA uses linear programming to identify optimal flux distributions through the network by optimizing the objective function. In the following section we will describe in more details the different methods for metabolic path identification and look into their specific characteristics as well as their strengths and weaknesses. Next, we will focus on FBA analysis and identification of objective functions. Furthermore, we will provide the basis of EFMs as a tool for complete enumeration of the flux space, and extreme pathways as subset of EFMs and their properties.

2.3 Path identification

In flux analysis there are two approaches to analyse metabolic networks based on the availability of background information, namely, data driven and hypothesis driven approach. In a hypothesis driven approach, the objective of


Biochemical information
Databases on the internet
Textbooks
Recent articles for specific information

	v1	v2	v3	v4	v5	v6	v7	Biomass
A	1	-1	0	0	0	0	0	0
A_out	-1	0	0	0	0	0	0	0
B	0	1	-1	-1	0	0	0	-1
C	0	0	0	1	0	-1	0	-1
D	0	0	1	0	-1	0	0	0
D_out	0	0	0	0	1	0	0	0
E	0	0	0	0	0	1	-1	0
E_out	0	0	0	0	0	0	1	0
Biomass	0	0	0	0	0	0	0	1
ATP	0	0	1	0	0	0	0	-1

Figure 2.2: Methodologies of stoichiometric modelling. Based on the study objective, different manipulations of the S matrix are possible: (a), (b) and (c) focus on identifying particular solution by imposing further constraints on the system; (d) and (e) are utilized for structural analysis of the pathway. For our specific problem we are interested in (c), (d) and (e). Figure modified from [14]

the study at hand is always known and usually the knowledge on how to achieve the objective is available. Herein it is possible to directly undertake the desired engineering steps and apply FBA with the aim to estimate the flux distribution under the particular hypothesis. On the other hand, when the objective of the study is loosely defined and the exact engineering strategies are not known in advance, we need a way to identify possible engineering targets. If there is enough data for the given problem, a data driven analysis could suggest the strategy to achieve the desired objective. In fact, it can also suggest possible objectives given the input data. Sufficient amount of data in stoichiometric modelling usually requires complete genome-scale metabolic networks of the host species, and optionally additional database of reactions resembling enzymes that could be incorporated into the host species. There are computational methods developed to answer specific questions such as identification of set of deletions that lead to increased production of the desired compound [8], suggesting possible gene insertions and deletions that would allow for production of a non-native compound [34], etc. However, these methods are beyond the analysis performed in this work and they only investigate the optimal production routes in the network in a similar manner as FBA.

When provided a genome-scale metabolic model, one can also further explore the solution space by completely enumerating the set of feasible flux distributions or according to some selection criteria. The existing methods such as EFMs and extreme pathways are utilized in this direction, yet the enumeration performed by these methods is computationally expensive and results in millions of pathways for genome-scale models, an amount rather infeasible for further interpretation. Our approach provides an alternative to these exhaustive enumeration, when one is interested in short-listing subset of (minimal) flux distributions satisfying a set of criteria. The resulting pathways can then be subjected to biochemical evaluation in terms of their suitability and feasibility for industrial implementation.

2.3.1 Flux balance analysis (FBA)

Mathematical models of a metabolic network can be generated based on the S matrix. However, the metabolic system is completely described only after the constraints on the flux values are included in the model. Metabolic fluxes can be assigned lower and upper bounds, limiting the range of values that a single flux can take. These constraints bound the null space of S and shrink the solution space [30].

Set of constraints lie in the cornerstone of FBA, and they can be represented either as balance constraints in a form of equalities, or bounds in the form of inequalities. The invariant components of FBA are the imposed steady state constraint on internal metabolites (as equality) and the flux bounds (as inequalities). Additional constraints can be optionally included. Since the system is represented in terms of set of linear equations, linear optimization can be applied to solve the given system of equalities and inequalities depicted as follows

$$Sv = 0 \quad \text{where} \quad 0 \leq \mathrm{v}_i \leq \mathrm{v}_{i,max} \tag{2.7}$$

The following step is identifying an objective that will quantify the biological phenotype of interest. A linear objective function is described as

$$Z = w^T v = \sum_i w_i v_i \tag{2.8}$$

where w is the vector of weights associated with each of the reaction fluxes. Z represents the objective and it can be maximized or minimized. A detailed illustration of the steps of FBA is presented in 2.3 The most common objective in FBA is optimizing the biomass production. represents the rate at which metabolic compounds are transformed into biomass components such as nucleic acids, proteins and lipids. Another example of FBA objectives are minimization of ATP production when aiming to depict the state of optimal metabolic energy efficiency or maximization of metabolite production when interested in identifying optimal production rates. Due to the optimization, FBA results in a single solution which might not be unique and there might be other optimal or suboptimal solutions with different flux distributions. Flux variability analysis (FVA) is a convenient tool for investigating the range of values that fluxes can take within a specific optimal cell behaviour [15]. However the procedure requires two optimizations per flux, namely minimization and maximization to yield the minimal and maximal value of the flux within the optimal behaviour.

FBA performance, usage and limitations

Unlike the dynamic modelling approach where varying kinetic parameters have to be estimated, FBA relies solely on reaction stoichiometries which are fixed and known in advance. This makes FBA well suited for different analysis of various metabolic networks, including genome scale models. The simplicity of the approach allows for applying FBA in studies interested in metabolic perturbations such as utilizing a combination of substrates or introduction of synthetic reactions. Applications of FBA include exploration

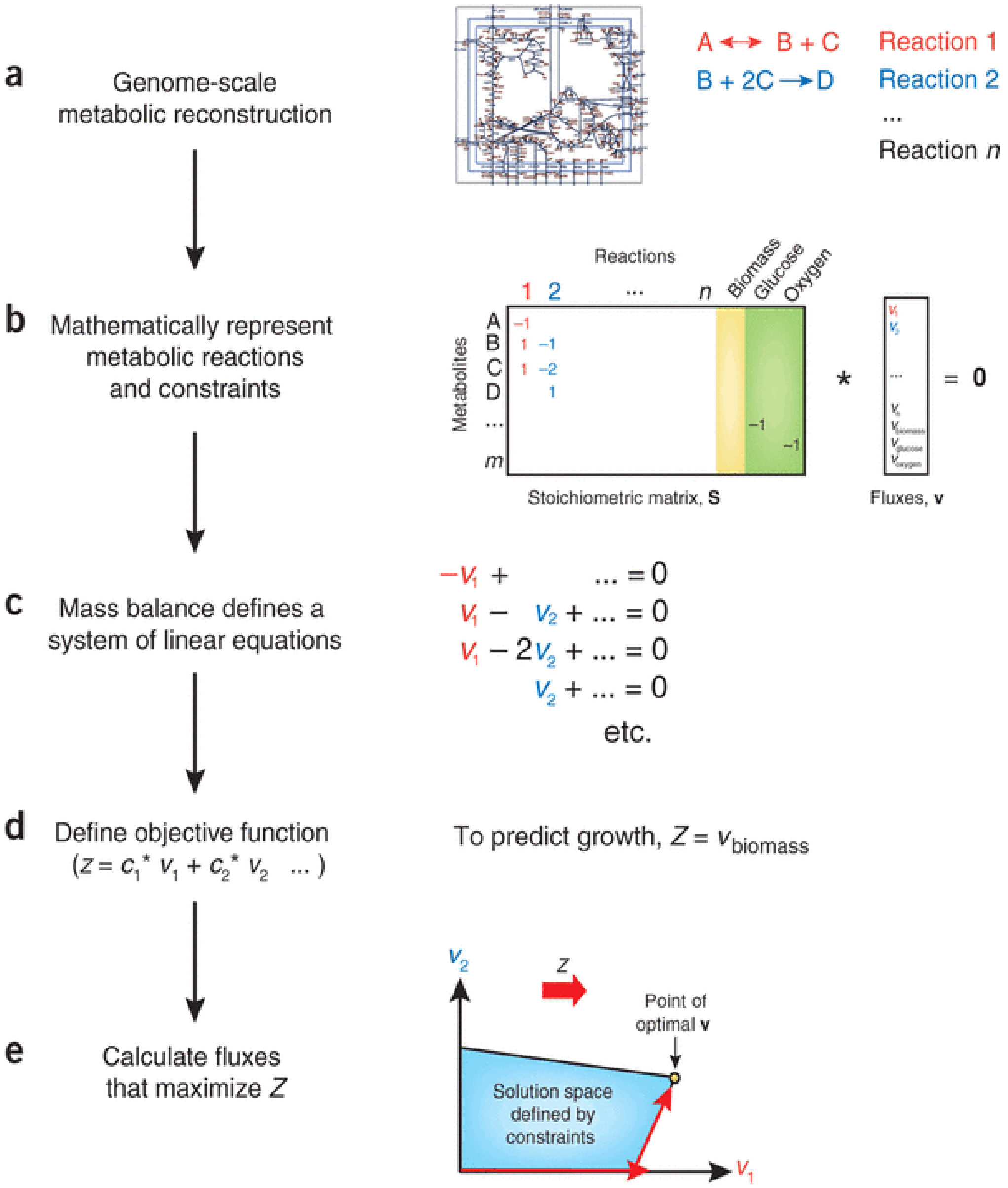

a
Genome-scale metabolic reconstruction
A ↔ B + C
B + 2C → D
Reaction 1
Reaction 2
...
Reaction n
b
Mathematically represent metabolic reactions and constraints
Reactions
1 2 ... n
Biomass
Glucose
Oxygen
Metabolites
A B C D ... m
Stoichiometric matrix, S
Fluxes, v
* = 0
c
Mass balance defines a system of linear equations
−v1 + ... = 0
v1 − v2 + ... = 0
v1 − 2v2 + ... = 0
v2 + ... = 0
etc.
d
Define objective function
(z = c1* v1 + c2* v2 ...)
To predict growth, Z = vbiomass
e
Calculate fluxes that maximize Z
v2
Z
Point of optimal v
Solution space defined by constraints
v1

Figure 2.3: Methodology for flux balance analysis. Figure from [30]

of network capabilities and limitations [21], robustness analysis of metabolic networks [12], etc. When it comes to perturbation analysis, FBA can either predict the outcome of performed perturbation or identify potential perturbation targets when there is an objective but no hypothesis on how to achieve it.

The fact that FBA does not use kinetic parameters, brings certain drawbacks to the approach, such as the inability to estimate metabolic concentrations [30]. FBA is limited to analysis at steady state, making it non feasible for analysis of dynamic changes. Furthermore, it only analyses very limited portion of the solution space based on the pre-defined objective function, therefore it is unable to explore the solution space beyond the objective.

2.3.2 Network-based pathway analysis/ metabolic pathway analysis (MPA)

Major representative methods in quantitative analysis of metabolic networks, developed during the last decade, are elementary flux modes and extreme pathways. The umbrella term of metabolic pathway analysis encompasses these methods with the aim to enumerate the polyhedral cone representing the entire flux space [15].

Elementary flux modes (EFMs) analysis

After representing a metabolic network in terms of its flux space, several methods can be applied to identify relevant pathways in the network. There are 3 properties based on which we can evaluate the ability of a set of pathways to explain the entire solution set: They are (P1) they can generate the flux space according the equation 2.6, (P2) they are the minimal set of vectors that fulfil p1 and (P3) they resemble all non-decomposable pathways in the network [26].

The term of elementary flux mode refers to a minimal set of reactions that can operate at steady state, with directions according to the thermodynamic feasibility [41]. The minimality herein is explained when we have a set of reactions that represent an elementary mode, and there is no subset of that EFM that represents an EFM itself [26]. The thermodynamic feasibility constraint coincides with the flux bounds, and it states that if a reaction i is irreversible, the flux through it v_i must be positive. One way to model this is by splitting the reversible reactions into forward and backward with each proceeding only in positive direction. The flux bound constraints on internal fluxes then become $v_i \geq 0$.

The definition of EFMs satisfies (P1)-that is, each steady state flux distribution can ge generated as non-negative combination of the elementary modes. The minimality of EFMs implies that they also satisfy the third property (P3) as they represent the set of all nondecomposable steady state distributions. However, the EFMs do not fulfil the second property (P2) of network-based pathways.

The major strenght of EFMs analysis is in that it enumerates the entire solution space in unbiased manner, but it has the drawback that the number of EFMs grows exponentially with the size of the network due to which many metabolic networks have been intractable for EFMs analysis [43]. As an example, the network of central metabolism in Escherichia coli, consiting of 110 reactions resulted in over 2 million EFMs. Nonetheless, [19, 42] have shown that the full set of elementary modes can be enumerated by using linear programming. The question posed here is how to identify subset of the most relevant pathways in the set of millions. The application of suitable weighting factors to the EFMs that would yield a meaningful physiological state depicting the outcome of perturbation is a challenging task [15]. Another implication resulting from the steady-state condition in EFMs analysis and the effort to downsize the network complexities, is that exchange fluxes and external metabolites are excluded from the analysis. Hence, the assumptions are that exchange fluxes represent either consumption or production of species by reactions omitted by the model, and the external metabolites are buffered by reactions of the complete system. As a result, all the information on the interactions between consumption and productions of substrates and products are lost. One of the pitfalls of these simplifications is that it can result in EFMs that fulfil the steady-state condition within the analysed sub-network, but does not constitute a part of any steady state path throughout the complete network [19]. In order to circumvent the problem of explosion in number of EFMs, other efforts have been reported working with reduced size of the networks. However, the results have shown that the interpretations might be misleading [19]. Some methods have tackled the problem of computational explosion in EFMs by decomposition of the network in modules [16, 42]. In [16], it is proven that division of biochemical networks into sub-networks enables the complete enumeration of EFMs in genome-scale models. Nonetheless this still does not decrease the number of identified EFMs, thus the challenge in further exploration remains.

Extreme pathways

Enzymes are the basic functional unit in metabolism, and when interested in understanding the driving mechanisms that allow for their activation and in turn navigate the metabolites through the complex net of metabolic reactions, one is seeking to analyse the metabolic pathways [36]. Extreme pathways analysis is convex analysis for computation of set of independent pathways in any metabolic network and they can be used for understanding metabolic functioning and control. In metabolic network analysis the main focus is on understanding the structural properties of the networks. Among the structural properties, which are the invariant characteristics of the networks, the stoichiometry is the most relevant one. It provides information on the network topology and the rates of metabolite conversion in a chemical reaction, which are constant over time and non-dependant on any external factors. A mass balance can be constructed around each metabolite in the network, and under the steady-state constraint the system of metabolic reactions and metabolites can be depicted with the equation 2.4. As mentioned earlier, the null space analysis of the S matrix determines a set of basis vectors at steady-state and they describe the entire solution space. However, the information on thermodynamic reversibility is omitted in the classical null space analysis and therefore, to completely describe the system we need to account for them. After including the inequalities for reaction reversibility, the system expands beyond the limits of linear algebra and that calls for convex analysis. Solution of a set of linear inequalities is a convex set that in the case of metabolic networks represents the edges of the so called flux cone or steady-state cone, and referred to as the extreme pathways. Furthermore, every point within the cone, or every pathway can be written as non-negative linear combination of the extreme pathways. The set of extreme pathways are subset of the EFMs that is systemically independent, which means that no extreme pathway can be represented as a non negative linear combination of any other extreme pathways [32]. Since the extreme pathways are identified in an augmented space, that does not resemble the original space after the reversible reactions are split into forward and reverse ones, some of the properties of the augmented space are not shared with the original space. Regarding the properties P1-P3 of metabolic pathways in determining their ability to generate the entire solution space, enumerated in the beginning of section 2.3.2, the extreme pathways satisfy P1, but not P2 and P3 in the original vector space [26].

Some of the differences in extreme pathways and EFMs are due to the different representation of reversible fluxes. Herein, extreme pathways decouple the reversible fluxes, whereas the EFMs incorporate the information on reversibility with the use of additional rules. Next, systemic independence established in extreme pathways is the reason behind smaller number of extreme pathways compared to EFMs [32]. The conclusion from comparative analysis on extreme pathways and elementary modes, suggest that the set of extreme pathways might be insufficient for complete analysis of metabolic networks as they are only a subset of EFMs and among the remaining EFMs there might be some of high biological importance. However, they can be derived as a non-negative linear combination of the set of extreme pathways [18].

2.4 Linear programming (LP)

Constrained based metabolic network analysis use linear programming to identify feasible flux distributions. In this section we describe linear programming, representation of a linear program, algorithms and tools that are used for solving linear programs.
Linear program in its basic form is defined as a set of constraints and objective function, through equations that are linear in terms of the unknown variables. Linear programming is an optimization method, one of the most widely used for formulation of wide range of real world optimization problems, such as maximizing the business profit based on the company model. Essential to linear programming, are the background knowledge and capability to formulate the problem in a way that captures the fundamental relations of the interacting variables, as well as the skills to interpret the outcome of the optimization in a meaningful way. The popularity of the method initially arises from the fact that numerous allocation problems and economic phenomena can be represented in terms of linear objectives and constraints [27]. Moreover, given the simplicity of linear functions, originally more complex non-linear problems are sometimes represented in simplified linear form. The standard representation of a linear program is given according to [7] as follows

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \leq b \\ \text{and } x \geq 0 & \end{array} \tag{2.9}$$

where x is an n-dimensional column vector with elements corresponding to the variables to be determined, c^T is an n-dimensional row vector, b is an m-dimensional column vector, and finally A represents an mxn matrix with constant elements. In here, m is the number of inequalities and n is the number of unknown variables. The elements of both vectors c and b are fixed real constants. The inequality $Ax \geq b$ depicts the constraints which restrict the solution space, and it represents a convex polyhedron over which the objective is to be optimized. When all of the variables $x_i, i = 1, 2, 3..n$ are bounded and none can reach infinity, the polyhedra becomes a convex polytope. The optimal solution of a linear program is a convex set, and therefore every point inside the polytope can be represented as a convex combination of the solution set [22]. In constraint-based metabolic network analysis, where cells are governed by their topobiological, genetic, regulatory and environmental constraints and where we usually are interested in capturing the fluxes throughout the network under optimal conditions, it makes perfect sense to rely on linear programming. All the constraints imposed are linear in their nature or can be represented in linear form. Once the objective function has been defined, the canonical form of a constrained-based metabolic flux analysis, or more specifically of flux balance analysis becomes

$$\begin{array}{ll} \text{minimize} & c^T v \\ \text{subject to} & Sv = 0 \\ \text{and} & v_l \geq v \geq v_u \end{array} \tag{2.10}$$

where v denotes the vector of fluxes to be determined, c is a known vector representing the contribution of each reaction to the objective function and S is the stoichiometric matrix. The equaltiy $Sv = 0$ depicts the mass-balance at steady state. The lower and upper bounds for each of the reactions are given by the column vectors v_l and v_u respectively.

If a linear program is formulated such that along the canonical form given in equation 2.9, there exists the following constraint x_i integer for $i = 1, 2, ..n$, then we are reffering to an integer linear program (IP). When only some of the variables are restricted to integer values, but not all of them, the problem is called mixed integer linear problem and consequently the method to solve it is called mixed integer linear programming (MILP). In metabolic flux analysis, as we are studying structural properties of the networks it is rather convenient to use integer fluxes. From application of Petri net theory in systems biology [23], we learn that given that the nodes have integer values, the edges can also take integer values. The analogy here is that since the coefficients of the S matrix are integer, we can use integer variables for the fluxes as well. Apart from this, another argument for applying IP over MILP for metabolic flux analysis is that experimental results show that it is computationally costlier to apply MILP for some particular MFA problems [10].

2.4.1 Linear programming optimization tools

When a linear program has been mathematically formulated, multiple software tools can be utilized to solve it. The range of available optimization tools includes both open-source and commercial ones, and we will give a brief overview of the most representative examples.

Different LP solvers have implemented functions that convert between *lp* and *mps* file formats. There are many solvers for linear programming, some of which are free of charge and open-source and other are commercial. Most widely used is however *CPLEX*, part of IBM Optimization Studio - an analytical decision support package for development and evaluation of optimization models based on constraint programming [17]. *Lpsolve* is a free linear programming software that relies on the simplex method and branch-and-bound for integer problems [29]. *Lpsolve* does not have a limitation on the model size, however when presented with a larger problem it can very often take long time for solving it or completely fail to solve. Another open-source framework for solving constraint integer programs is *SCIP* [1].

2.5 Computing the k-shortest elementary modes

In this section, we give an overview of the method for computing the k-shortest elementary modes as developed by Figueiredo et al. [10], and we present the mathematical model of the method. The strength of this method can be seen in its capability to identify special subsets of EFMs by including additional constraints, avoiding complete enumeration of EFMs at the first place. In doing so, the computational complexity is significantly reduced when compared to methods for complete enumeration of EFMs [39, 40]. In addition, it is possible to impose further constraints based on the optimization problem at hand, such as the requirements for producing specific products.

The first stage of the method identifies the shortest elementary mode after solving the optimization problem:

$$\text{minimize} \sum_{i=1}^{n} z_i, \tag{2.11}$$

where z_i is a binary variable corresponding to each reaction in the S matrix, such that $z_i = 1$ if the flux through the i-th, v_i, is > 0, and 0 otherwise. The objective function coincides to the definition of EFMs which states that no subset of reactions of an EFM can perform at steady state, and thus it minimizes the total number of active fluxes. Furthermore, the relation between the binary variable z_i and the flux variable is captured by the following constraints

$$\begin{cases} v_i \leq M z_i, & i = 1, 2, 3..n \\ z_i \leq v_i, & i = 1, 2...n \end{cases} \tag{2.12}$$

The above set of equations guarantee that if z_i is equal to zero, consequently v_i must be zero and vice versa, if z_i is 1 then v_i has to be greater than 1, but smaller or equal to arbitrary large constant M. In addition, the reversible fluxes are decomposed into 2 irreversible fluxes and this is depicted as follows

$$z_\alpha + z_\beta \leq 1 \quad \forall(\alpha, \beta) \in B \tag{2.13}$$

where the set B is defined as $\{(\alpha, \beta)\}$, where α and β are the forward and reverse of the same reaction, and $\alpha < \beta$. This ensures that a reaction can only appear in one direction. Finally, the mass-balance constraint applies to all the internal metabolites as one of the cornerstones of elementary flux modes analysis, and this is included in the optimization problem. The given optimization is solved as ain integer linear program.
Once the shortest elementary mode has been detected, a sequential optimization is applied to identify the subsequent shortest modes. In each following iteration of the procedure,a constraint has to be added that prevents the new solutions from containing already identified flux modes. This is achieved by requiring that the already identified EFMs are not completely included in the solutions generated at the next steps. If we denote the binary form of the shortest EFM by $\tilde{z}_i$; $i \in \{1, 2, ..n\}$, the formulation of the constraints is as given below

$$\sum_{i=1}^{n}(\tilde{z}_i z_i) \leq \sum_{i=1}^{n}(\tilde{z}_i) - 1 \tag{2.14}$$

where z_i represents the binaries of the current EM.
To summarize the procedure, Algorithm 1 presents the pseudo-code for the sequence of optimizations. We start with a an empty list of EMs that we want to populate until the desired number of EMs, defined as $MaxNumEM$ in the code, have been calculated. In each of the iterations, we first solve the linear program as defined by the constraints in equations 2.12 and 2.13 and the objective given in 2.11, and this is equivalent to the identification shortest EFm at the given step. Next we add the EM to the pool of solutions and we impose the exclusion constraint, as formulated in 2.14 to avoid completely overlapping EFMs.

Algorithm 1 k-shortest EMs (Figueiredo et al.)

```
$EMS = \{\}$;
$k = 0$;
while $k < maxNumEM$ do
    $k++$;
    newEM=solveMILP();
    EMS=EMS $\{newem\}$;
    add_exclusion_constr(newEM);
```

2.6 Alpha-spectrum

All allowable steady state flux distributions can be represented as a non-negative linear combination of the extreme pathways. The extreme pathways represent vectors that correspond to the edges of the flux cone. The so called flux cone is the convex polyhedron representing the solution space as defined by the system of linear equalities and inequalities that govern the metabolic system behaviour. Every point inside the solution space can be represented as a linear combination of the extreme pathways. For a given flux distribution, it is possible to reveal the set of extreme pathways that contribute to it and consequently the extent of the contribution. The method that describes this decomposition is called Alpha-Cone Method, developed by Palsson et al. [31]. Herein, a steady state flux distribution is decomposed into extreme pathways and the range of weighting for these extreme pathways is identified. The weightings are termed α and the range of possible values that they can take for a specific flux distribution is termed the α-spectrum. It is described by the equation

$$P\alpha = v, \tag{2.15}$$

where P represents the matrix of extreme pathway vectors, α is the vector of weightings on the pathways and v is the steady state flux distribution. The range of the weightings α_i for each extreme pathway P_i are determined by both minimizing and maximizing their values using LP, while leaving all other extreme pathway weightings free [46]. The contribution of the extreme pathways to a flux distribution is significant for the understanding of the biological importance of the flux distribution, and for the regulation mechanisms of the network that can be tracked through consecutive changes of flux distributions and their corresponding extreme pathways. Extreme pathways can also provide insight into phenotypic capabilities of a metabolic network by distinguishing flux distributions arising from extreme pathways of functional interest. This is possible due to the classification of extreme pathways into three categories, out of which one type only is of functional interest.

Chapter 3

Methods and Data

3.1 Input data and preprocessing

The metabolic models we have analysed in this work were provided by VTT Technical Research Centre of Finland. The reaction networks have been created based on biochemical expert's knowledge for the purpose of investigating production routes to a specific product of interest. The models are of different sizes and we have divided them in three categories, namely, there are small networks consisting of less than 100 metabolites and reactions, medium sized networks with number of reactions and metabolites ranging between 100-500, and finally large networks incorporating genome-scale information, with up to 4000 reactions and metabolites. There are 12 potential substrates that the networks are allowed to utilize and there is always one product per network. Along the S matrix, meta-data are included in the input. The meta-data provides information on lower and upper bounds for the fluxes, KEGG [20] reaction identifiers and classification of metabolites into 4 categories - substrates, products, cofactors and intermediates. Intermediates here refer to the internal metabolites upon which steady state constraint is imposed. Substrates and products represent the external metabolites and the exchange reactions that connect them to the internal metabolites are included. Cofactors, or carriers, in biochemistry are considered "helper molecules" that facilitate biochemical transformation.

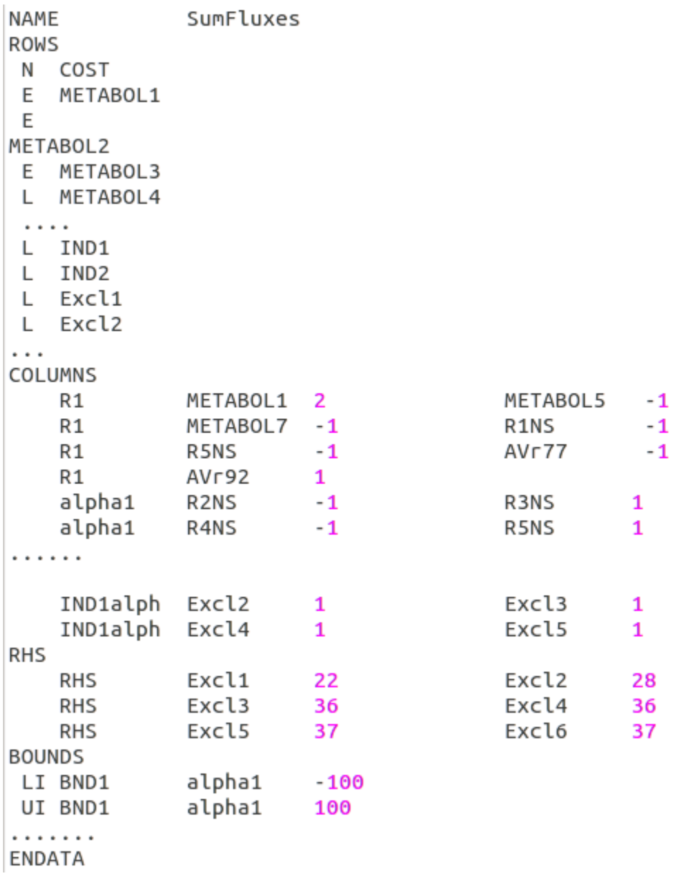
The preprocessing of the data includes conversion of the S matrix into a format required by the optimization software packages that we use in later stage.The most common way of presenting a linear or mixed integer programming problems is the LP format. It is native format for reading and writing linear programming problems for some of the software packages like lpsolve [29]. The file starts with the problem statement that defines whether we want to maximize on minimize the objective function. Next we define the objective function, and then we provide the set of constraints each starting in a new line. Finally, there is a list of inequalities representing the bounds for each of the variables. This is the general structure of LP file, but differences in syntax arise with different tools. On the other hand, MPS files are presented in a slightly different format, yet still frequently used in optimization tasks. MPS format is a column oriented format, where fields start at a specific column in a text file. Next, the file is divided into sections, each starting with a special header. The sections that can appear are ROWS, COLUMNS, RHS, RANGES and BOUNDS. The ROWS section contains the name of the constraints, and a description of whether the constraint is an equality (E), less-than (L), greater-then (G) or whether it is the cost function for the objective (N). Another section - COLUMNS, contains the actual entries of the A matrix from the canonical form given in Chapter 2. For each variable, the column assigns all the non zero constraint coefficients related to the variable. The RHS section specifies the name of the right-hand side vectors and values for each of the constraints. RANGES accounts for constraints restricted between a range of values and BOUNDS specify the limit values for each variable.

For the purpose of data conversion into MPS format, we have used the MPS format exporting tool from Bruno Luong available at MathWorks [1], implemented in MATLAB. The function converts a standard linear programming problem into an MPS format. The input for the function, requires separate matrices for the equalities and inequalities representing the constraints of the linear programming problem, vectors representing right-hand side values for the equalities and inequalities, cost function vector representing the weighting for the objective function, lower and upper bounds for the variables and there are additional optional parameters. The output is an MPS text file, and a sample file generated based on our input data is given in Figure 3.1.

For comparison, an lp representation of the same problem is given in Figure 3.2 This file is generated from the MPS file, with the use of Lpsolve conversion function.

From here, the generated MPS and LP files are used by the optimization tools for identification of flux distributions. The entire work is implemented in MATLAB, and the optimization tools that we have used have callable

[1] http://www.mathworks.com/matlabcentral/fileexchange/19618-mps-format-exporting-tool/content/BuildMPS/BuildMPS.m



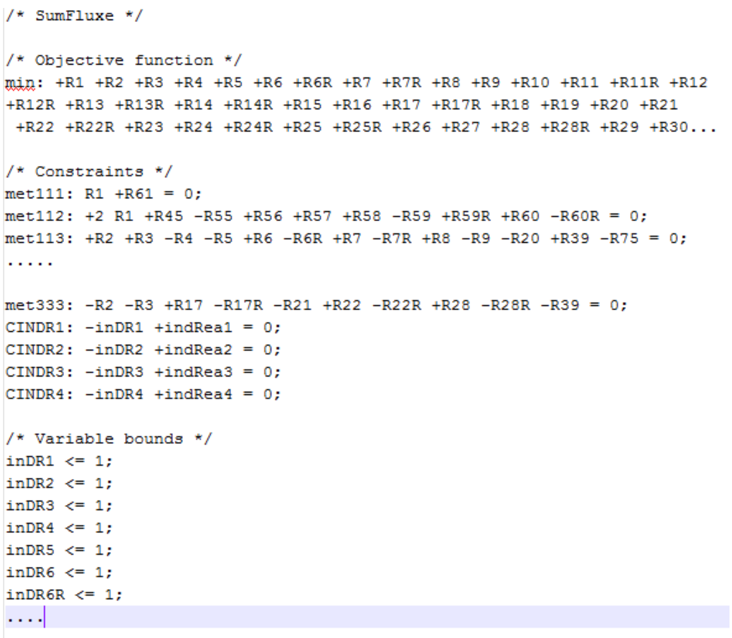
```
NAME          SumFluxes
ROWS
 N  COST
 E  METABOL1
 E
METABOL2
 E  METABOL3
 L  METABOL4
 ....
 L  IND1
 L  IND2
 L  Excl1
 L  Excl2
...
COLUMNS
    R1        METABOL1  2               METABOL5   -1
    R1        METABOL7  -1              R1NS       -1
    R1        R5NS      -1              AVr77      -1
    R1        AVr92     1
    alpha1    R2NS      -1              R3NS      1
    alpha1    R4NS      -1              R5NS      1
......

    IND1alph  Excl2     1               Excl3     1
    IND1alph  Excl4     1               Excl5     1
RHS
    RHS       Excl1     22              Excl2     28
    RHS       Excl3     36              Excl4     36
    RHS       Excl5     37              Excl6     37
BOUNDS
 LI BND1      alpha1    -100
 UI BND1      alpha1    100
.......
ENDATA
```

Figure 3.1: MPS file generated by MPS format exporting tool, based on our S matrix as input. The sections in the file are presented in the following order: NAME, ROWS, COLUMNS, RHS and BOUNDS. In section ROWS the second column gives the (metabolite) name of the row in S. In COLUMNS section the first row belongs to the variable name or reaction name, followed by the metabolites to which it belongs and the S matrix coefficients. RHS second column gives the metabolite names again an their right-hand side constraint. BOUNDS as can be assumed gives the lower and upper bound for each of the reactions.

MATLAB libraries. The CPLEX MATLAB library is included in the original CPLEX Optimization studio distribution, and in order to be employed it required that the CPLEX connector for MATLAB installation folder is added to the MATLAB path. Lpsolve is callable from MATLAB through an external interface or MEX-function and it requires that a driver program is installed [2]. SCIP Optimization suite also has a build-in MATLAB interface.

[2] http://web.mit.edu/lpsolve/doc/MATLAB.htm



```
/* SumFluxe */

/* Objective function */
min: +R1 +R2 +R3 +R4 +R5 +R6 +R6R +R7 +R7R +R8 +R9 +R10 +R11 +R11R +R12
+R12R +R13 +R13R +R14 +R14R +R15 +R16 +R17 +R17R +R18 +R19 +R20 +R21
 +R22 +R22R +R23 +R24 +R24R +R25 +R25R +R26 +R27 +R28 +R28R +R29 +R30...

/* Constraints */
met111: R1 +R61 = 0;
met112: +2 R1 +R45 -R55 +R56 +R57 +R58 -R59 +R59R +R60 -R60R = 0;
met113: +R2 +R3 -R4 -R5 +R6 -R6R +R7 -R7R +R8 -R9 -R20 +R39 -R75 = 0;
.....

met333: -R2 -R3 +R17 -R17R -R21 +R22 -R22R +R28 -R28R -R39 = 0;
CINDR1: -inDR1 +indRea1 = 0;
CINDR2: -inDR2 +indRea2 = 0;
CINDR3: -inDR3 +indRea3 = 0;
CINDR4: -inDR4 +indRea4 = 0;

/* Variable bounds */
inDR1 <= 1;
inDR2 <= 1;
inDR3 <= 1;
inDR4 <= 1;
inDR5 <= 1;
inDR6 <= 1;
inDR6R <= 1;
....
```

Figure 3.2: LP representation generated from the MPS file in Figure 3.1. Sections included are Objective function, Constraints and Variable bounds.

For the optimization tasks related to the small sized metabolic networks we have used the three different platforms, and the results were easily obtained with similar running times. However, for the medium sized and large networks, when the number of variables exceeds 100 we have heavily relied on CPLEX as the other tools have in many instances failed to provide results and the running times were significantly larger. The non-commercial optimizers such as Lpsolve and SCIP Studio do not guarantee that the convergence to an optimal solution for large problems will occur. Detailed comparison of commercial and open-source solvers for linear optimization problems is provided in [28].

3.2 FBA with varying sets of objectives

The main goal of our work is to identify pathways producing a desired industrially relevant product given the potential substrates, product and bag of possible reactions to be included in the selected pathways. Moreover, we want to determine which novel reactions are economically most feasible to introduce, and thus we have the objective of minimizing the fluxes through the reactions. Consequently we seek to identify alternative pathways to the optimal one as identified by the linear programming optimization. The alternative pathways might have either the optimal or suboptimal objective value. Once these pathways are obtained, we want to rank them according to predefined criteria such as number of new enzymes introduced to the pathway as well as the overall number of enzymes (reactions); objective value; ATP to product ratio; carbon yield; thermodynamic efficiency; etc. Additionally we want to perform pathway analysis based on relative stoichiometry of cofactors and net thermodynamic feasibility. As there are number of relevant properties that we are interested in, it is very challenging to identify the best combination of objectives or ranking criteria before applying optimization techniques and performing further analysis. Hence, we are interested in applying different secondary targets for the objective, along the minimization of the sum of absolute fluxes. These secondary objectives can be applied separately or in combination based on prior biological knowledge and experimental results. When we use a multi-objective optimization sometimes the combined objectives are conflicting, i.e. the improvement of one only comes at the cost of the others. For this reason, there has to be a trade-off between the features that we want to optimize and we need to identify the most satisfactory distributions after analysing the proposed pathways.

In addition to the lower and upper bounds for the fluxes and steady state condition on the internal metabolites, we impose also the following constraints:
1) Substrates need to be consumed, meaning that overall net change depicted by the mass-balance equation should be negative. In the case of multiple substrates, the requirement is that at least one of them should be consumed by the internal metabolites.
2) Metabolite acting as product have to be produced, that is, the mass balance for the product has a positive value.

Furthermore, cofactors are free of constraints, but in some of our analysis we are still interested in optimizing some of their net stoichiometry as will be seen in continuation. The net stoichiometry represents the net consumption/production of a given metabolite. From here, we investigate the effect of different sets of additional constraints on the flux distributions across the networks. By far, the most important objective that we are interested in is minimizing the absolute sum of fluxes. In doing so we seek to obtain reasonable set of changes to the native network (for ex. introducing new enzymes, changing cofactors, etc.)

In our initial approach we are interested in applying FBA with a range of objective functions and rank them according to predefined criteria. In addition to identifying the optimal objective value, we also explore suboptimal solution that provide us bigger set of pathways for further analysis. Next, we give an overview of the objective functions that we used to generate lists

of candidate pathways.
The formulation of the integer linear program with single objective of minimizing the sum of absolute fluxes is given in Equation (3.1).

$$\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{n} |v_i| & \\
\text{subject to} \quad & \sum S_j v \leq -1 & \text{for each row j corresponding to substrates} \\
& S_j v \leq 0 \ , & \text{if row j corresponds to substrate} \\
& S_j v > 0 \ , & \text{if row j corresponds to product} \\
& S_j v \geq -20 \ , & \text{if row j corresponds to cofactor} \\
& S_j v = 0 \ , & \text{default} \\
& v_i \text{ integer}, & i = 1, 2, ..., n
\end{aligned} \tag{3.1}$$

Here we require that at least one substrate is consumed and given a substrate - it can only be consumed, as formulated by the first two constraints. Next we require that the product is produced and we select an arbitrary lower bound of -20 for the cofactors since some of the optimizing tools require a right hand side value for each of the constraints. However, in practice we do not have constraints on the cofactors. Later on, we are going to refer to this integer linear program (ILP) as to *solveOptimalILP()*.
Our next objective is to maximize the carbon yield of the product, provided single substrate as input. The carbon yield is calculated as

$$\frac{k_1}{k_2} \frac{S_j v}{S_l v}. \tag{3.2}$$

It is calculated per carbon atom, and k_1 and k_2 are the coefficients corresponding to the number of carbon atoms in the product and substrate respectively. S_j vector is the row in the S matrix that belongs to the product and S_l is the row that belongs to the substrate. Carbon yield optimization initially represents a linear-fractional programming problem [9]. There is a relationship between linear-fractional programming and linear programming, thus the problem can be transformed into a linear form. The transformation is performed according to [9], and it requires the introduction of the variables y and t

$$\begin{cases} y = \dfrac{v}{S_l v} \\ t = \dfrac{1}{S_l v} \end{cases} \tag{3.3}$$

Finally, the formulation of the ILP is the following

$$
\begin{array}{lllr}
\text{maximize} & \frac{k_1}{k_2} S_j y & & \\
\text{subject to} & S_j y > 0, & \text{if row j belongs to the product} & \\
& S_j y \leq -t, & \text{if row j belongs to specific substr.} & \\
& S_j y = 0, & \text{if row j belongs to any other substrate} & (3.4) \\
& S_j y \geq -20t, & \text{if row j belongs to cofactors} & \\
& S_j y = 0, & \text{default} & \\
& S_l y = 1 & & \\
& t \geq 0, & &
\end{array}
$$

where the last 2 constraints are introduced due to the transformation. The other constraints coincide with the constraints in equation (3.1).

Other objective that we have tested are maximization of ATP usage, minimization of ATP production, minimize the net stoichiometry for some of the reducing equivalents, etc. In biochemistry, reducing equivalents are chemical species that transfer the equivalent of one electron in redox reactions, which are opposite of oxidation reactions.

For solving (mixed)integer programs CPLEX uses the function *cplexmilp()* [3], that relies on the branch and cut optimization method, however the implementation of the function is proprietary to the Optimization Studio. For the purpose of exploring the space of optimal and suboptimal solutions, we have used CPLEX built-in function *populate()* [4]. The function generates multiple solutions to a (mixed) integer program. It first finds the optimal solution to the program and simultaneously sets up a branch and cut tree for the next stage. Based on this information, it then generates multiple solutions. The function takes on input, arguments that can control the number of generated solutions. We have required 20 solutions per objective function, however in some cases there are only few optimal solutions that the method suggest. Nevertheless, there is no way to avoid enumeration of solutions involving the same sets of variables, representing scalar multiples of each other; and these are not of interest to our work as we want to identify distinctive flux distributions in terms of the reactions (fluxes) they use.

[3] `http://www.ibm.com/support/knowledgecenter/SSSA5P_12.2.0/ilog.odms.cplex.help/Content/Optimization/Documentation/CPLEX/_pubskel/CPLEX1194.html`

[4] `http://www.ibm.com/support/knowledgecenter/SSSA5P_12.4.0/ilog.odms.cplex.help/refdotnetcplex/html/M_ILOG_CPLEX_Cplex_Populate.htm`

3.3 Enumeration of the *k-best* pathways

Here we present our approach for enumeration of the *k-best* pathways. It is based on the work of Figueiredo et al. [10] as presented in the background chapter, where they derive a procedure for enumeration of the k-shortest EFMs. The formulation of our *k-best* method is an extension of the previously defined *solveOptimalILP()* statement in Equation (3.1), and it enumerates the *k-best* pathways. In contrast to executing a linear program that results in a single optima, here we enumerate pathways that satisfy the optimality criteria and thus represent (sub)optimal flux distributions. We believe this enumeration will result in better candidate set of pathways for industrial implementation, compared to the pathways obtained by the CPLEX "populate" enumeration procedure. This reasoning is motivated by the fact that the CPLEX enumeration resulted in large number of minor variants of only one path instead of multiple paths, as well as multiplies of previous solutions. Thus we aim to identify more biologically relevant pathways by including the exclusion constraint.

The procedure is presented in Algorithm 2. It starts with identification of the first best pathway as defined by the formulation in Equation (3.1), with highest objective value according the given constraints. Next, we incorporate an exclusion constraint to the original optimization problem and we iteratively identify k best solutions. The exclusion constraint that we use, prevents subsequent solution from being supersets of previous solutions and from having exactly the same pattern of active fluxes. In each iteration step, one constraint is added that eliminates the current solution from the set of next solutions, and consequently in the k-th step there are k-1 exclusion constraints. In our method, the value for the parameter k is dependent on the S matrix and it corresponds to its nullity, which is the dimension of the null space. Finally, the structure of our *k-best* method is as follows

Algorithm 2 k-best flux distributions (FD)

```
$FD = \{\}$;
$k = 0$;
while $k < nullity(S)$ do
    $k++$;
    newFD=solveOptimalILP();
    FD=FD $\{newFD\}$;
    add_exclusion_constr(newFD);
```

We are aware that we can not directly compare the solutions attained with this enumeration method to the basis vectors of the null space, as in the first place we do not impose the requirement for linear independence among the obtained solutions. The exclusion constraint that we use only guarantees that, there is pairwise independence between each newly obtained solution and all of the previous solutions. In our analysis, we also experimented by iterating over the procedure until we obtain a solution matrix for which the rank coincides with the rank of the null space, i.e. until we identify linearly independent vectors of the same size as the nullity of S.

3.4 Combined objective in augmented space

Looking at the results from the *k-best* enumeration, we still produce minor variations of single biological pathway for some of the networks. The variability for the flux distributions exits but the varying reactions are only different to each other in a single cofactor which does not bring biological variability. To tackle this issue and increase he biological variability of the proposed pathways, we shift the objective function to the space of external metabolites. We could also examine and optimize the way the null space basis are used for the generation of a single flux vector as done in the alpha-spectrum analysis on extreme pathways [46]. Therefore, alternative approaches to the problem would be to optimize the net stoichiometry for the exchange metabolites or to optimize the way the basis vectors are combined in order to construct a feasible pathway. Working separately with each of these approaches did not yield improved results, and thus we have defined a combined augmented solution space where we incorporate the flux distributions, null space basis and net stoichiometry for exchange species into what would represent an augmented stoichiometric matrix.

For convenience, let us introduce the following representation of the S matrix $S = \begin{bmatrix} R \\ B \end{bmatrix}$, where R contains the rows that belong to internal metabolites in S, and B contains the rows that belong to external metabolites in S matrix. Let us also define the net stoichiometry, $NetSpec$, of the exchange species with the following equation

$$B \times v = netSpec \tag{3.5}$$

The null space basis vectors were calculated based on the internal stoichiometric matrix that only contains the intermediate metabolites, and for this we used the null space MATLAB function. Having defined that, we can write the following set of equations

$$\begin{cases} R \times v = 0 \\ B \times v = netSpec \\ N\times\ \alpha\ \ = \text{v} \end{cases} \tag{3.6}$$

where N is the null space basis and α represents the weighting factors of the basis vectors towards a specific flux distribution. We can rewrite the equation as follows

$$\begin{cases} & R \times v & & = 0 \\ & B \times v & -I \times netSpec & = 0 \\ N\times\ \alpha & -I \times v & & = 0 \end{cases} \tag{3.7}$$

where each of the columns corresponds to the alpha vector, flux vector v and the net stoichiometry netSpec respectively. They are all unknown variables and if combined together they can be written as a single vector named y, and their coefficients can be represented in a matrix named M in the following manner

$$M = \begin{bmatrix} 0 & R & 0 \\ 0 & B & -I \\ N & -I & 0 \end{bmatrix} ;\ y = \begin{bmatrix} \alpha \\ v \\ nSpec \end{bmatrix}$$

(3.8)

Finally, we have our formulation of the integer linear program in the combined space

$$\begin{array}{ll} \textbf{minimize:} & \sum_{i=1}^{ny} |y_i| \\ \textbf{subject to} & My = 0 \end{array} \tag{3.9}$$

3.5 Elementary flux modes and extreme pathways

We have computed elementary flux modes and extreme pathways for the networks for which they were calculable. Our aim in doing so, was to compare the number of EFMs and extreme pathways to the number of pathways generated with our approach, in addition to comparing the running times.

Elementary modes were calculated using the METATOOL platform for computing elementary modes and other structural properties of biochemical reaction networks [45]. At the time of its implementation, the program was the the fastest one for calculating elementary modes. As input the program either takes a standard METATOOL format file as described on the METATOOL web pages [5], or S matrix and a vector specifying the positions of the irreversible reactions.

ExPA, extreme pathway analysis program developed by Palsson et al., [5] was employed for the computation of the extreme pathways. The program is open-source and uses a command line interface. The input is again either the S matrix and information on the number of exchange fluxes or a specific reaction file with one reaction per line of the file. Detailed explanation of the reaction file is given on the website of the method [6].

[5] http://pinguin.biologie.uni-jena.de/bioinformatik/networks/metatool/metatool5.1/metatool5.1.html

[6] http://systemsbiology.ucsd.edu/Downloads/ExtremePathwayAnalysis

Chapter 4

Results and Discussion

In this thesis, our goal was to develop a computational approach for the identification of metabolic pathways connecting substrates to a product. Our methods, k-best pathways and combined space, short-list pathways while avoiding exhaustive enumeration. Furthermore, our approach includes analysis of the identified pathways based on net stoichiometry, cofactor balancing and thermodynamic feasibility. For the sake of comparison, we perform elementary flux modes and extreme pathways analysis.

4.1 Comparison of methods with respect to pathway properties

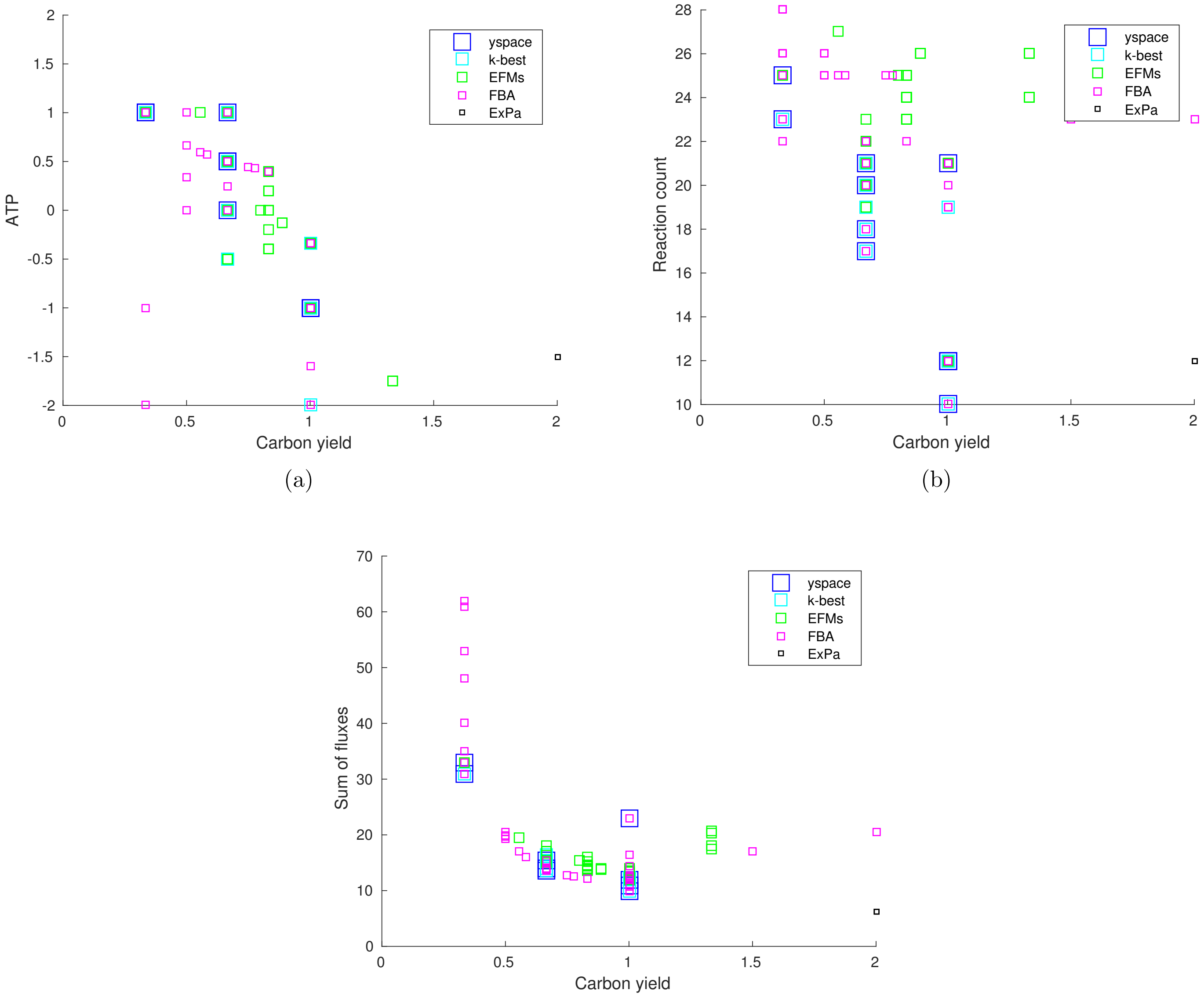
As indicated before, one of our secondary objectives is to optimize carbon yield (CY) as it has a determining effect on the viability of the proposed pathways for industrial production. In Figure 4.1, a scatter plot is given reflecting the carbon yield of the generated pathways in relation to three other parameters: ATP illustrated in Figure 4.1a, reaction count illustrated in Figure 4.1b and sum of absolute fluxes ilustrated in Figure 4.1c.

The results presented give a comparison between the methods in terms of the carbon efficiency of the pathways. The carbon yield is calculated as the number of carbon atoms in the product per single carbon atom in the substrate. The results in Figure 4.1 are based on the small network and the methods that we compare are as indicated in the legend: 'yspace' - enumeration in combined space, 'k-best - k-best enumeration', 'EFMs' - elementary flux modes, 'FBA' - combined results from FBA with different objectives as presented in chapter 3. and 'ExPa' - extreme pathways analysis. Pathways generated with EMFs were filtered so that only the pathways consuming a substrate and producing a product are taken into account. However, all of the ExPa pathways are included and therefore some of them do not satisfy the requirement for substrate consumption or production of product, or neither of them.

We have also discussed earlier the importance of ATP, and we want to maximize the use or minimize the production of ATP so that the overall net stoichiometry for ATP is as close to zero as possible. ATP production or usage is calculated as ATP net stoichiometry per single unit of product. In Figure 4.1a we have termed the results that have carbon yield close to one (or at least higher than 0.6) and ATP close to zero as successful. We can therefore observe that the best pathways were identified by k-best, FBA and EFMs analysis. Next, in Figure 4.1b we are interested in solutions that minimize the number of reactions and maximize carbon yield (CY), thus the best solution has CY of one and 10 reactions, and it is identified by FBA, k-best and yspace. The second best solution is obtained by the same methods and EFMs in addition.
The objective of minimizing the sum of fluxes versus CY in 4.1c reveals that the best solutions are again pinpointed by our methods k-best and yspace as well as by FBA. In here, we are looking for points located close to the right bottom corner of the graph and we have lower sum of fluxes compared to the reaction count due to the fact that some reactions carry fluxes higher than one. Minimal sum of fluxes is 10, with CY of one. The ExPA point with CY of two and sum of fluxes lower than 10 is an outlier, as it refers to a non-product related pathway.

Figure 4.2 illustrates the ATP versus reaction count and sum of fluxes. On the y-axis, positive values represent ATP production and negative points stand for ATP consumption. In both figures we note that the best solutions are converging towards zero ATP and minimal sum of fluxes and number of reactions. In Figure 4.2a a good candidate solutions are the one with zero ATP and reaction count of 18, obtained by our methods. If we are willing to compromise towards ATP consumption of one we get three solutions with reaction count between 10 and 15. At the end it comes down to the biological experts to decide upon the matter. The results are similar in 4.2b.

Figure 4.3 presents the scatter plot of reaction count versus absolute sum of fluxes. The best results in terms of minimal sum of fluxes and reaction count are proposed by our methods as well as ExPa. ExPa solutions with lower sum of fluxes or reaction count than our solutions exemplify pathways with no connection from a substrate to product. If we would like to take



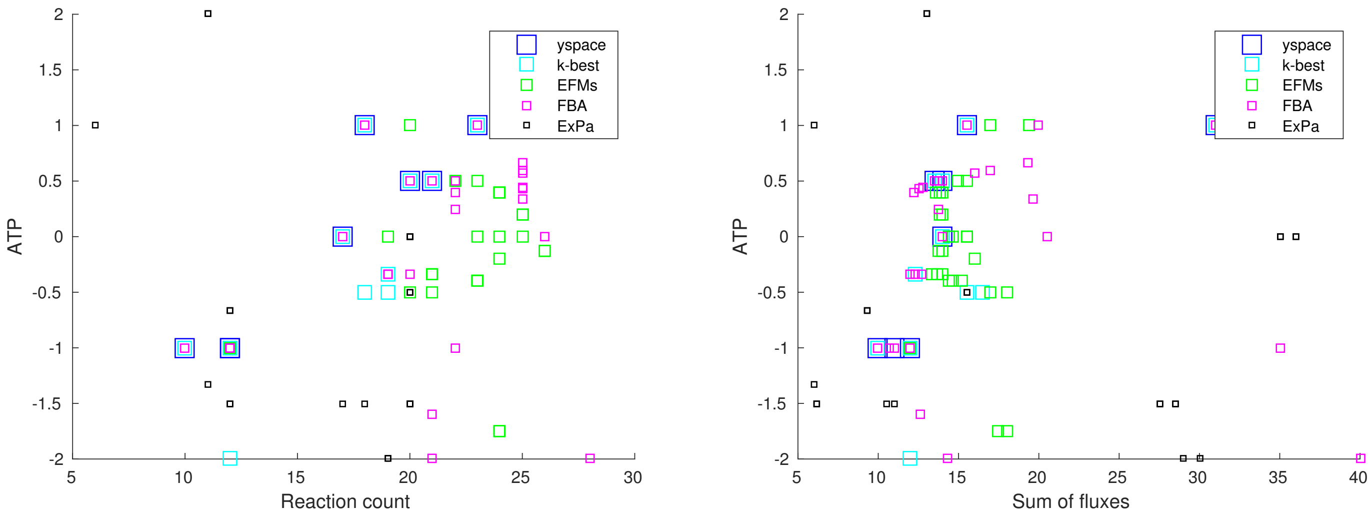
(a)

(b)

(c)

Figure 4.1: Small network, scatter plot representing carbon yield versus (a) ATP (consumption/production), (b) Reaction count and (c) Sum of fluxes.

advantage of these results considering that they offer solutions with minimal reaction count, we could explore ways to connect the pathways to substrate and product. Anyhow, it will probably lead to higher than the optimal reaction count, depending on the number of connections to be introduced to



(a)

(b)

Figure 4.2: Small network, scatter plots of ATP versus (a) reaction count and (b) sum of fluxes.

the pathway.
Lastly, Figure 4.4 illustrates the sum of fluxes versus the reaction count for the medium 4.4a and for the large matrix 4.4b. We only illustrate the performance of our methods here as the ExPa program did not provide the extreme pathways in neither case, probably due to the size of the network, and for the EFMs we did not present the proposed pathways for the medium sized network. We can conclude from Figure 4.4b that for the large network there is a high correlation between the reaction count and sum of fluxes.

4.2 Analysis of pathways identified in combined space

Summary of pathway properties based on net reaction stoichiometry is given in Table 4.1. The analysis was performed with respect to the k-best pathways obtained in combined space with incorporated exclusion constraint. The properties listed have been predefined as relevant pointers for the phenotypic capabilities of the resulting paths. All the values contained in the table are scaled to account for one unit of product. The 'number of active enzymes' column shows the total number of reactions activated at least once in the set

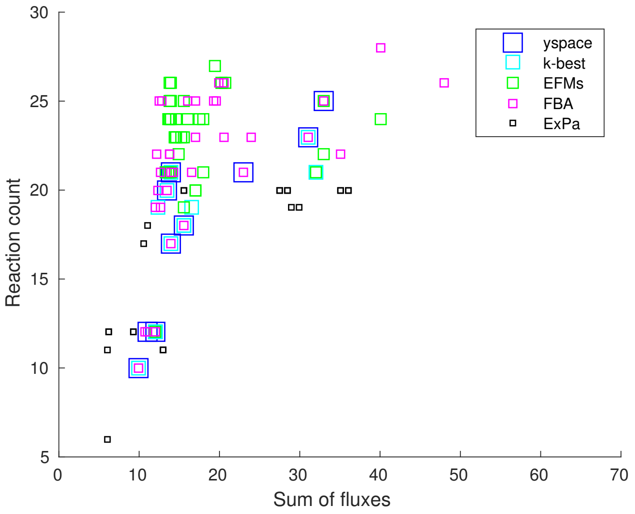
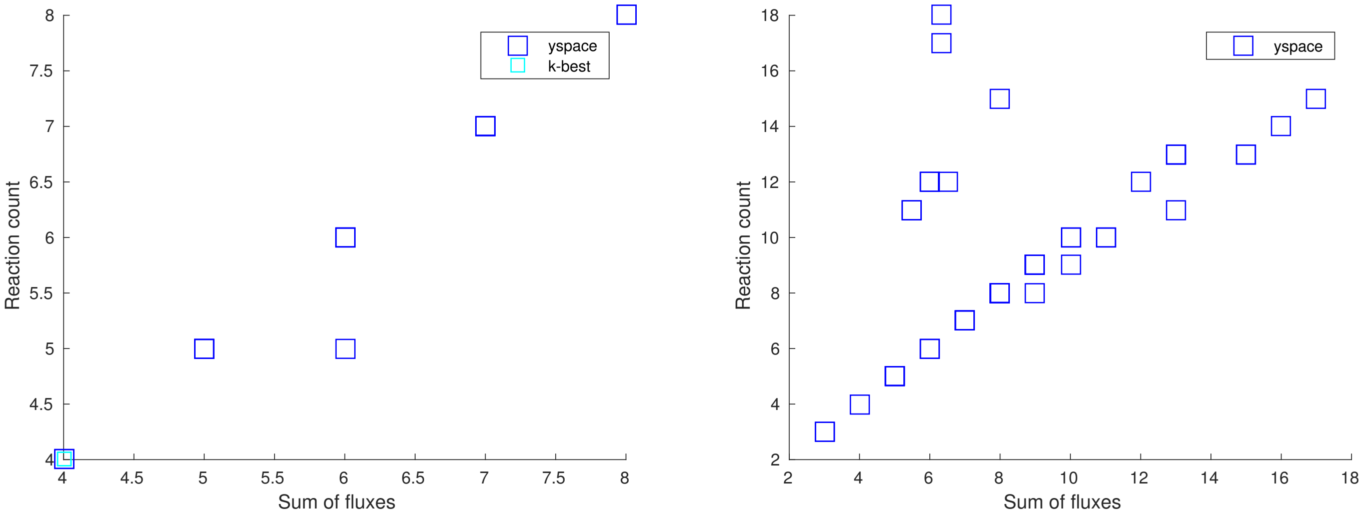


Figure 4.3: Small network, sum of fluxes versus reaction count.



(a)

(b)

Figure 4.4: Medium (a) and Large (b) networks, scatter plots of reaction count versus sum of fluxes.

of pathways for the given network. As can be seen in the table, for the small network all of the reactions have been utilized by the pathways, and as the number of reactions increases the percentage of active reactions drastically decreases. For the large network there are only 161 active reactions, and further analysis could provide insight into functional characteristics of the relevant reactions. Next, the total number of reversible versus irreversible reactions for the networks is given. 'Sum of absolute fluxes' represents the range of the sum of fluxes among the pathways, i.e. the minimal and maximal sum of fluxes; it is evident that again the sum of fluxes is significantly lower for the medium and large networks. However, more detailed exploration of the active enzymes in the large networks would probably yield the conclusion that additional cofactor balancing is required before the pathways are complete. In our work, the pathways are not initially balanced in terms of cofactors, and thus there is the necessity to account for the balancing in the postprocessing phase. The following three properties give the average/-median flux through an active reaction, the total number of reversible/irreversible enzymes active and the average flux through reversible/irreversible reaction. The last three properties, are related to specific metabolites including redox metabolites that have some relevant lower and upper limits that have to be obeyed.

The variation in number of active reactions across the pathways is illustrated in Figure 4.5. The figure compares the number of reversible and irreversible reactions activated in the generated pathways. The number of reversible and irreversible reactions is given in Table 4.1. We can observe from the figure that the variation is lowest for the small matrix for both, reversible and irreversible active reactions. Moreover, the values approach the total

Table 4.1: Analysis of k-best pathways generated in combined space. The properties are list in the first column, and the consecutive columns hold their values for the small, medium and large sizes network.

properties	**48x36**	**201x333**	**1971x3371**
total number of active enzymes(reactions)	36	116	161
number of revers./irrevers. enzymes	22/14	137/183	1220/2151
sum of absolute fluxes	10,36	5,17	3,17
avg./median flux through active reaction	1.3/1	1/1	1.01/1.01
total number of revers./irrevers. enzymes active	23/13	80/36	116/45
avg. flux through revers./irrevers. reaction	1/1	1/1	1/1.03
sum of abs. fluxes for redox metabolites	0,4	0,7	2,12
sum of abs. fluxes where water is a reactant	5,16	0,9	2,12
net sum of fluxes for redox metabolites	0,4	-2,3	-3,4

number of reversible/irreversible reactions which in turn suggests that the same set of reactions are consistently picked by the majority of pathways for this networks, especially for the reversible reactions.

The outliers far below, for the reversible reactions, belong to the two best pathways in terms of the objective. For the remaining two networks, there is greater diversity in number of selected reactions, but also the number of total reactions is much bigger compared to the small network. Surprisingly, the median number of reversible active reactions is higher for the small network, even though we would expect the opposite taken in consideration the size of the networks.

The product carbon yield per given substrate is represented in Figure 4.6, 4.6a for the large and 4.6b for the medium network. The values correspond to a range of 10 different pathways, and for their calculation we have specifically required 10 pathways per substrate. Regarding the substrates, we have only required that a substrate must be consumed by the internal metabolites, however we did not require that each of the substrates has to be consumed. As a result, only a subset of the proposed substrates were taken as input. For the medium network there were five substrates selected out of 12, and for the small network there were two offered substrates and both were consumed in the set of identified pathways. The percentage distribution of pathways consuming a substrate is shown in Figure 4.7. For the large network, we have only calculated 120 pathways, 10 per each substrate and therefore the distribution in that case is uniform and omitted from representation on the graph.

4.3 Performance comparison

We compared the performance of our methods to the performance of the elementary modes and extreme pathways analysis and we show the results in Table 4.2. The second column represents the nullity of the matrix, and this is the value that we use for the parameter k in our k-best method. The EFMs were calculated for the small and medium matrix, but the METATOOL program did not yield results for the large networks. We can see that even for the medium network the number of EFMs is very large. ExPa program only delivered the results regarding the small matrix and the results for larger matrices were not calculable.

Table 4.2: Performance comparison between k-best in combined space, EFMs and ExPa. Number of pathways and the running times for obtaining them.

	48x36	**172x320**	**1971x3371**
null space size	8	210	1672
no. reversible react.	22	137	1220
no. potential substrates	2	12	12
no. EFMs	66	270803	not calculable
EFMs running time (s)	6.5	41722.2 (12h)	/
no. k-best paths	8	210	1672
k-best running time	3	178.7	>24h
no. ExPa	89	/	/
ExPa running time	<1	/	/

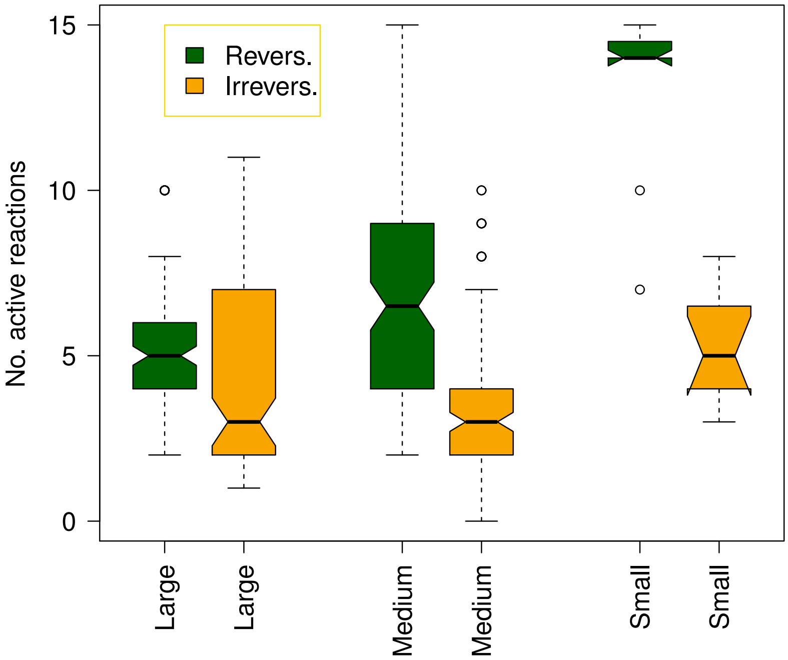
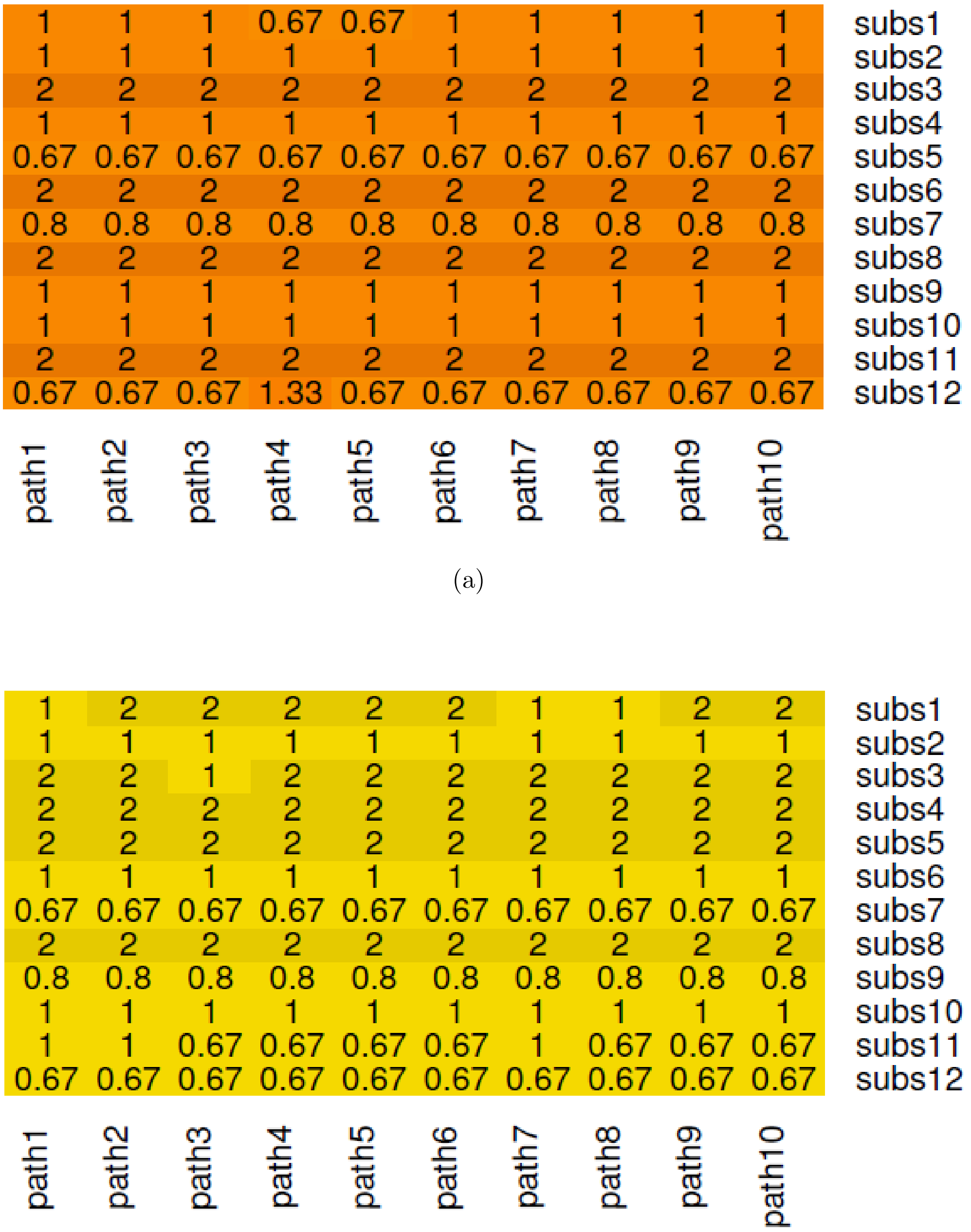


Figure 4.5: Illustration of the distribution of the number of active reactions, both reversible and irreversible, across the generated pathways for the large, medium and small network.



	path1	path2	path3	path4	path5	path6	path7	path8	path9	path10
subs1	1	1	1	0.67	0.67	1	1	1	1	1
subs2	1	1	1	1	1	1	1	1	1	1
subs3	2	2	2	2	2	2	2	2	2	2
subs4	1	1	1	1	1	1	1	1	1	1
subs5	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
subs6	2	2	2	2	2	2	2	2	2	2
subs7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
subs8	2	2	2	2	2	2	2	2	2	2
subs9	1	1	1	1	1	1	1	1	1	1
subs10	1	1	1	1	1	1	1	1	1	1
subs11	2	2	2	2	2	2	2	2	2	2
subs12	0.67	0.67	0.67	1.33	0.67	0.67	0.67	0.67	0.67	0.67

(a)

	path1	path2	path3	path4	path5	path6	path7	path8	path9	path10
subs1	1	2	2	2	2	2	1	1	2	2
subs2	1	1	1	1	1	1	1	1	1	1
subs3	2	2	1	2	2	2	2	2	2	2
subs4	2	2	2	2	2	2	2	2	2	2
subs5	2	2	2	2	2	2	2	2	2	2
subs6	1	1	1	1	1	1	1	1	1	1
subs7	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
subs8	2	2	2	2	2	2	2	2	2	2
subs9	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
subs10	1	1	1	1	1	1	1	1	1	1
subs11	1	1	0.67	0.67	0.67	0.67	1	0.67	0.67	0.67
subs12	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67

(b)

Figure 4.6: Carbon yield for a range of pathways and substrates, (a) large network, (b) medium sized network. The values are scaled to one unit of product. For each of the substrates there are 10 pathways represented with their corresponding carbon yields.

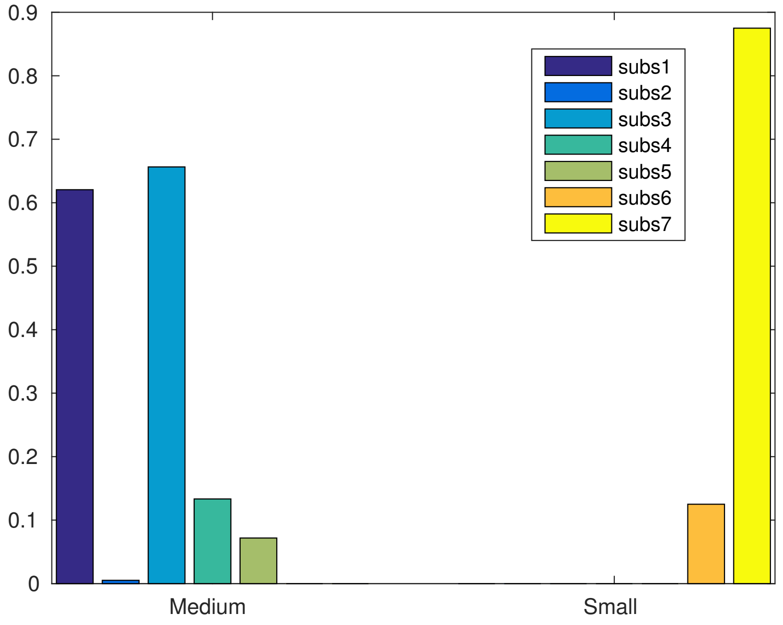


Figure 4.7: Percentage distribution reflecting the fraction of pathways that consumed the given substrate. For the medium network there were 12 substrates available, and some pathways consumed two substrates simultaneously. For the small one there were 2 substrates available and only singe substrates were selected in each pathway.

Chapter 5

Conclusion

The purpose of this work was to provide a computational framework for enumeration of metabolic pathways that are able to convert a set of source metabolites into a target metabolite, through intermediate metabolites. Current enumeration techniques include computation of elementary flux modes and extreme pathways which become computationally very expensive for genome scale metabolic models. We apply an objective function that shrinks the solution space, and we use an exclusion constraint for the enumeration of pre-defined number of pathways, hence reducing the computational complexity of the enumeration process. Moreover, we perform ranking and analysis of the enumerated pathways based on met stoichiometry and thermodynamic feasibility.

As shown in the Results section, when working with large metabolic network consisting of 1971 metabolites and 3371 reactions, both elementary flux modes and extreme pathways analysis failed to enumerate the pathways. Our method on the other hand was able to perform the enumeration of 1672 pathways. In cases, where EFMs and extreme pathways are able to enumerate large scale networks they result in large number of pathways in the range of millions. Hence, our method provides a more efficient exploration of the solution space without the need for complete enumeration of pathways. Moreover, based on the study at hand, different types of constraints can be incorporated into the model. Further improvements of the method could include implementation of exclusion constraint with the requirement of systemic linear independence as the current implementation only guarantees pairwise independence of the generated pathways.

Finally, the results suggest that some of the generated pathways between substrates and product are novel and thus have not been utilized in production processes before. These results have the potential to improve current production processes in terms of energetic efficiency.

Bibliography

[1] Achterberg, T. Scip: Solving constraint integer programs. *Mathematical Programming Computation 1*, 1 (July 2009), 1–41. `http://mpc.zib.de/index.php/MPC/article/view/4`.

[2] Arundel, A., and Sawaya, D. The bioeconomy to 2030.

[3] Bailey, J. E. Toward a science of metabolic engineering. *Science 252*, 5013 (1991), 1668–1675.

[4] Bailey, J. E. Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. *Biotechnology progress 14*, 1 (1998), 8–20.

[5] Bell, S. L., and Palsson, B. O. Expa: a program for calculating extreme pathways in biochemical reaction networks. *Bioinformatics 21*, 8 (2005), 1739–1740.

[6] Bernhard, P. Systems biology: properties of reconstructed networks, 2006.

[7] Bertsimas, D., and Tsitsiklis, J. N. *Introduction to linear optimization*, vol. 6. Athena Scientific Belmont, MA, 1997.

[8] Burgard, A. P., Pharkya, P., and Maranas, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering 84*, 6 (2003), 647–657.

[9] Craven, B. Fractional programming, sigma series in applied mathematics, vol. 4. *Heldermann Verlag* (1988).

[10] 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 25*, 23 (2009), 3158–3165.

[11] Dugar, D., and Stephanopoulos, G. Relative potential of biosynthetic pathways for biofuels and bio-based products. *Nature biotechnology 29*, 12 (2011), 1074–1078.

[12] Edwards, J. S., and Palsson, B. O. Robustness analysis of the escherichiacoli metabolic network. *Biotechnology Progress 16*, 6 (2000), 927–939.

[13] Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., and Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology 7*, 2 (2009), 129–143.

[14] Gombert, A. K., and Nielsen, J. Mathematical modelling of metabolism. *Current opinion in biotechnology 11*, 2 (2000), 180–186.

[15] Horvat, P., Koller, M., and Braunegg, G. Recent advances in elementary flux modes and yield space analysis as useful tools in metabolic network studies. *World Journal of Microbiology and Biotechnology 31*, 9 (2015), 1315–1328.

[16] Hunt, K. A., Folsom, J. P., Taffs, R. L., and Carlson, R. P. Complete enumeration of elementary flux modes through scalable, demand-based subnetwork definition. *Bioinformatics* (2014), btu021.

[17] ILOG, Inc. Ilog cplex: High-performance software for mathematical programming and optimization, 2006. `http://www.ilog.com/products/cplex/`.

[18] Jevremovic, D., Trinh, C. T., Srienc, F., and Boley, D. A simple rank test to distinguish extreme pathways from elementary modes in metabolic networks. *Univ. of Minnesota, Computer Science and Eng. Dept. Tech. Rep* (2008), 08–029.

[19] Kaleta, C., de Figueiredo, L. F., and Schuster, S. Can the whole be less than the sum of its parts? pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Research 19*, 10 (2009), 1872–1883.

[20] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research 44*, D1 (2016), D457–D462.

[21] Kauffman, K. J., Prakash, P., and Edwards, J. S. Advances in flux balance analysis. *Current opinion in biotechnology 14*, 5 (2003), 491–496.

[22] Kelk, S. M., Olivier, B. G., Stougie, L., and Bruggeman, F. J. Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. *Scientific reports 2* (2012).

[23] Koch, I., Junker, B. H., and Heiner, M. Application of petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics 21*, 7 (2005), 1219–1226.

[24] Lee, S. Y., Kim, H. U., Park, J. H., Park, J. M., and Kim, T. Y. Metabolic engineering of microorganisms: general strategies and drug production. *Drug Discovery Today 14*, 1 (2009), 78–88.

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

[26] Llaneras, F., and Picó, J. Which metabolic pathways generate and characterize the flux space? a comparison among elementary modes, extreme pathways and minimal generators. *BioMed Research International 2010* (2010).

[27] Luenberger, D. G., and Ye, Y. *Linear and nonlinear programming*, vol. 2. Springer, 1984.

[28] Meindl, B., and Templ, M. Analysis of commercial and free and open source solvers for linear optimization problems. *Eurostat and Statistics Netherlands within the project ESSnet on common tools and harmonised methodology for SDC in the ESS* (2012).

[29] Michel Berkelaar, Kjell Eikland, P. N. Open source (mixed-integer) linear programming system: lpsolve. `http://lpsolve.sourceforge.net/5.5/`, 2004.

[30] Orth, J. D., Thiele, I., and Palsson, B. Ø. What is flux balance analysis? *Nature biotechnology 28*, 3 (2010), 245–248.

[31] Palsson, B. O., Covert, M. W., and Herrgard, M. Methods and systems to identify operational reaction pathways, June 8 2010. US Patent 7,734,420.

[32] Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., and Palsson, B. O. Comparison of network-based pathway analysis methods. *Trends in biotechnology 22*, 8 (2004), 400–405.

[33] Pfleger, B. F., Pitera, D. J., Smolke, C. D., and Keasling, J. D. Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nature biotechnology 24*, 8 (2006), 1027–1032.

[34] Pharkya, P., Burgard, A. P., and Maranas, C. D. Optstrain: a computational framework for redesign of microbial production systems. *Genome research 14*, 11 (2004), 2367–2376.

[35] Price, N. D., Reed, J. L., and Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology 2*, 11 (2004), 886–897.

[36] Price, N. D., Reed, J. L., Papin, J. A., Wiback, S. J., and Palsson, B. O. Network-based analysis of metabolic regulation in the human red blood cell. *Journal of Theoretical Biology 225*, 2 (2003), 185–194.

[37] Raman, K., and Chandra, N. Flux balance analysis of biological systems: applications and challenges. *Briefings in bioinformatics 10*, 4 (2009), 435–449.

[38] Sauer, M., and Mattanovich, D. Construction of microbial cell factories for industrial bioprocesses. *Journal of Chemical Technology and Biotechnology 87*, 4 (2012), 445–450.

[39] 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 203*, 3 (2000), 229–248.

[40] Schuster, S., Fell, D. A., and Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature biotechnology 18*, 3 (2000), 326–332.

[41] Schuster, S., and Hilgetag, C. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems 2*, 02 (1994), 165–182.

[42] Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I., and Dandekar, T. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to mycoplasma pneumoniae. *Bioinformatics 18*, 2 (2002), 351–361.

[43] Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. Metabolic network structure determines key aspects of functionality and regulation. *Nature 420*, 6912 (2002), 190–193.

[44] Stephanopoulos, G., and Sinskey, A. J. Metabolic engineeringâmethodologies and future prospects. *Trends in biotechnology 11*, 9 (1993), 392–396.

[45] Von Kamp, A., and Schuster, S. Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics 22*, 15 (2006), 1930–1931.

[46] Wiback, S. J., Mahadevan, R., and Palsson, B. Ø. Reconstructing metabolic flux vectors from extreme pathways: defining the α-spectrum. *Journal of theoretical biology 224*, 3 (2003), 313–324.

Appendix A

Appendix: Matlab code

Matlab function for enumeration of k-best pathways in combined space.

```
function [Solutions,fluxVectors]=EnumerateCombined(path_to_product,product)
%function for obtaining first k solutions, k corresponds to the nullity
% of the S matrix

%Input:
% path_to_product - path to the structure file, 'netModifiedvL.mat'
% product - combined product name and size of net


tic;
load(strcat(path_to_product,'netModifiedvL.mat'));
react_mat=product.R;
directions=sign(product.vL+product.vU);
%change to only forward and rev reactions
dirs=directions;
react_mat(:,directions<0)=-1*react_mat(:,directions<0);
directions(directions<0)=1;
%read in constraints
constr=product.CIDclassification;
constraints=cell((size(constr)));

for i=1:size(constr,1)
    switch (constr(i))

        case 1
            constraints(i)=cellstr('substrate');
        case 0
            constraints(i)=cellstr('intermediate');
        case 2
            constraints(i)=cellstr('product');
        case -2
```

```
            constraints(i)=cellstr('cofactor');
    end
end

substrate_size=size(strmatch('substrate',char(constraints)),1);
product_size=size(strmatch('product',char(constraints)),1);
cofactor_size=size(strmatch('cofactor',char(constraints)),1);
intermediate_size=size(strmatch('intermediate',char(constraints)),1);

inter_indices=strmatch('intermediate',char(constraints));
subs_indices=strmatch('substrate',char(constraints));

% R is the matrix of ineternal metabolites, B is matrix of exchange met.
R=react_mat(inter_indices,:);
%null space size should be based on R matrix only
nullity=size(react_mat,2)-rank(R);
B=react_mat;
B(inter_indices,:)=[];
nullSpace=null(R,'r');

% M combined matrix
M=[zeros(size(R,1),size(nullSpace,2)) R zeros(size(R,1),size(B,1)); ...
    zeros(size(B,1),size(nullSpace,2)) B -1*eye(size(B,1)); ...
    nullSpace -1*eye(size(nullSpace,1)) zeros(size(nullSpace,1),size(B,1))];


[rows,columns]=size(M);
%add vars for absolute values (we minimize the sum of abs values),
%indicator vars for the exclusion constraint
M=[M zeros(size(M)) zeros(size(M))];
ineq_s=substrate_size+product_size;
%oneineq constraint for substrate
A=zeros(ineq_s+1+2*columns+2*columns,size(M,2));
%A=zeros(ineq_s+1,size(M,2));
Aeq=M;
k=1;
%keq=1;
b=zeros(1,size(A,1));
beq=zeros(1,size(Aeq,1));
sum_s=zeros(1,size(A,2));
subs_positions=zeros(1,substrate_size);
s_pos=1;

for i=1:size(react_mat,1)
    if(ismember(cellstr('substrate'),constraints(i)))
        A(k,:)=[zeros(1,size(nullSpace,2)) react_mat(i,:) zeros(1,size(B,1))...
        zeros(1,2*columns)];
        b(k)=0;
        sum_s=sum_s+A(k,:);
        subs_positions(s_pos)=k;
        s_pos=s_pos+1;
        k=k+1;
    end
    if(ismember(cellstr('product'),constraints(i)))
        A(k,:)=[zeros(1,size(nullSpace,2)) -1*react_mat(i,:) ...
        zeros(1,size(B,1)) zeros(1,2*columns)];
        b(k)=-1;
        k=k+1;
    end
end

%sum of substrates
A(k,:)=sum_s;
b(k)=-1;
Elenames(k)=cellstr('subCons');
k=k+1;
%abs inequalities
for j=1:columns
    A(k,[j columns+j])=[-1 -1];
    A(k+1,[j columns+j])=[1 -1];
    b([k k+1])=[0 0];
    Elenames([k k+1])=[cellstr(strcat('AVr',num2str(j),num2str(1))) ...
    cellstr(strcat('AVr',num2str(j),num2str(2)))]';
    k=k+2;
end
%indicator constraints
for j=1:columns
    A(k,[columns+j 2*columns+j])=[1 -30];
    A(k+1,[columns+j 2*columns+j])=[-1 1];
    Elenames([k k+1])=[cellstr(strcat('IND',num2str(j),num2str(1))) ...
    cellstr(strcat('IND',num2str(j),num2str(2)))];
    k=k+2;
end
% cost function
L=repmat(-100,1,size(M,2))';
L(nullity+(find(directions>0)))=0;
U=repmat(100,1,size(M,2))';
U(nullity+find(directions<0))=0;
ints=linspace(1,size(M,2),size(M,2));
ints(2*columns+1:end)=[];
bins=linspace(2*columns+1,size(M,2),size(M,2)-2*columns);
cost=zeros(1,size(M,2));
cost(columns+1:2*columns)=ones(1,columns);

%input for cplex
  f=cost';
  Aineq=A;
  bineq=b';
  beq=beq';
  lb=L;
  ub=U;
  lb(size(ints,2)+1:end)=0;
  ub(size(ints,2)+1:end)=1;
  ctype=repmat('I',1,size(ints,2)+size(bins,2));
  [x,~,~,~]=cplexmilp(f,Aineq,bineq,Aeq,beq,[],[],[],lb,ub,ctype);

  Solutions=zeros(nullity,columns);
  Objectives=zeros(1,nullity);
  Solutions(1,:)=x(1:columns);
  Objectives(1)=sum(x(columns+1:2*columns));

 for i=2:nullity

     %exclusion constraint
    A(k,:)=zeros(1,size(A,2));
    A(k,(2*columns+1:end))=x(2*columns+1:end)';
    b(k)=sum(x(2*columns+1:end))-1;

    k=k+1;
    Aineq=A;
    bineq=b';

   [x,~,~,~]=cplexmilp(f,Aineq,bineq,Aeq,beq,[],[],[],lb,ub,ctype);

    Solutions(i,:)=x(1:columns);
    Objectives(i)=sum(x(columns+1:2*columns))
 end

    %extract the reaction fluxes columns only
  fluxVectors=Solutions(:,nullity+1:nullity+size(react_mat,2));

%change fluxes for directions
for i=1:size(directions,1)
    if (dirs(i)==-1)
        fluxVectors(:,i)=-1*fluxVectors(:,i);
    end
end
time=toc;
save('Results.mat','time','fluxVectors','Solutions','Objectives');
```