Study of the evolution of symbiosis at the metabolic level using models from game theory and economics

Martin Wannagat

▶ **To cite this version:**

Martin Wannagat. Study of the evolution of symbiosis at the metabolic level using models from game theory and economics. Bioinformatics [q-bio.QM]. Université Claude Bernard Lyon 1, 2016. English. NNT : . tel-01394107v1

HAL Id: tel-01394107

https://inria.hal.science/tel-01394107v1

Submitted on 8 Nov 2016 (v1), last revised 22 Nov 2016 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre NNT : 2016 LYSE1094

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale ED01
E2M2

Spécialité de doctorat : Bioinformatique

Soutenue publiquement/à huis clos le 04/07/2016, par :
Martin Wannagat

Study of the evolution of symbiosis at the metabolic level using models from game theory and economics

---

Devant le jury composé de :

Charles, Hubert, Prof., INSA Lyon	Président
Bockmayr, Alexander, Prof. Dr., Freie Universität Berlin	Rapporteur
Jourdan, Fabien, CR1, INRA Toulouse	Rapporteur
Neves, Ana Rute, Dr., Chr. Hansen A/S	Rapporteure
Brochier-Armanet, Céline, Prof. UCBL	Examinatrice
Andrade, Ricardo, Dr., INRIA Rhône-Alpes	Examinateur
Sagot, Marie-France, DR, LBBE, INRIA Rhône-Alpes	Directrice de thèse
Stougie, Leen, Prof., Vrije Universiteit, CWI Amsterdam	Co-directeur de thèse
Marchetti-Spaccamela, Alberto, Prof., Sapienza Univ. di Roma	Co-directeur de thèse

UNIVERSITE CLAUDE BERNARD-LYON 1

Président de l'Université	**M. le Professeur F. FLEURY**
Président du Conseil Académique	M. le Professeur H. BEN HADID
Vice-Président du Conseil d'Administration	M. le Professeur D. REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur P. CHEVALIER
Vice-président de la Commission Recherche	M. F. VALLÉE
Directeur Général des Services	M. A. HELLEU

COMPOSANTES SANTE

Faculté de Médecin Lyon-Est - Claude Bernard	Directeur: M. le Professeur J. ETIENNE
Faculté de Médecine et de Maeutique Lyon Sud Charles Mérieux	Directeur: Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur: M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur: Mme la Professeure C. VINCIGUERRA
Institut Techniques de Réadaptation	Directeur: M. le Professeur MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur: Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur: M. F. De MARCHI
Département Biologie	Directeur: M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur: Mme C. FELIX
Département Génie Electrique et des Procédés	Directeur: M. Hassan HAMMOURI
Département Informatique	Directeur: M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur: M. le Professeur G. TOMANOV
Département Mécanique	Directeur: M. le Professeur H. BEN HADID
Département Physique	Directeur: M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur: M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur: M. B. GUIDERDONI
Ecole Polytechnique Universitaire de Lyon 1	Directeur: M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur: M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur: M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur: M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur: M. N. LEBOISNE

Acknowledgements

In 2011 I moved from Lyon to Berlin. Due to a disappointing work experience I contacted Marie-France to explore the possibilities of coming back to Lyon and of working together. She rapidly promised to recruit me as an engineer. Already at this point she made the proposal to do a PhD. Back to Lyon, it took only some weeks to convince myself to do a PhD in her team. I am deeply grateful to Marie-France for this opportunity. I spent great three and a half years in the team and under her supervision. The work with her is characterized by determination and freedom. She trusts in the strength of every Phd student and she encouraged me in moments when the work did not progressed as expected. Many thanks to you, Marie-France, and I am glad to stay in the team for another two years.
I would like to thank Leen and Alberto, my two co-advisors abroad. Given the geographical distance we were able to meet quite often in Rome, Amsterdam and Lyon. Thanks for the discussions and advices. Under different conditions (without my son) I would have liked to spend more time in your beautiful cities. With my partner and my son, we managed already to visit Amsterdam; Leen, thanks again for the baby crib. And we will certainly come to Rome.
The one who introduced me to metabolism is Paulo. His enthusiasm motivated me to work on this subject. With him and Vicente, the master of the counterexamples, I shared several nice moments together and it makes me sometimes sad that these moments are not very frequent anymore due to the large geographical distance. The person I probably worked with most of time is Ricardo, my unofficial co-advisor. He arrived shortly after I started my PhD. He is patient, curious, and has a great capacity of listening and analyzing problems. He will never admit, but maybe I bothered him a bit too often with simple questions or strange ideas and he regretted sometimes the proximity of our desks. Now that he went back to Brazil, I should buy a teddy bear to explain some simple ideas. I can hardly imagine a better co-worker. Also off the job, I enjoyed the time sharing a beer or a glass of wine. You are always invited to come to the mountains. I am sure that we will manage to do a hiking tour next summer.
Many other people in and outside of the team made out of this PhD an unforgettable period of my life: Cecilia, Mariana, Alice, Greg, Alex, Carol, Xavier, Mattia, Christian Gautier, Christian Baudet, Dorota, Marina, Florence, Leandro, Vincent, Hélène, Arnaud, Delphine, Taneli, Blerina, Laura, André, Laurent, Matteo, Rita, Lilia, Bea, Paf, Alexandre, Susan, Janice, Thomas, Camille, Gustavo, Patricia, Lison, Michael, Etienne, Tristan, Renaud, Elie, Jérémy, Guillaume, Thomas and Antoine.
Vielen Dank auch meiner Familie ohne deren Unterstützung ich nicht so weit gekommen wäre. Ganz verstanden haben Sie das Thema meiner Doktorarbeit jedoch nie.
Un grand merci à ma compagne, Julie, qui m'a encouragé à faire un doctorat et qui m'a soutenu tout au long des trois ans. Elle et notre fils Tom sont mes points de repère.

TITRE en français

L'étude de l'évolution de la symbiose au niveau métabolique en utilisant des modèles de la théorie des jeux et de l'économie

RESUME en français

Le terme symbiose recouvre tous types d'interactions entre espèces et peut être défini comme une association étroite d'espèces différentes vivant ensemble. De telles interactions impliquant des micro-organismes présentent un intérêt particulier pour l'agriculture, la santé, et les questions environnementales. Tous les types d'interactions entre espèces tels que le mutualisme, le commensalisme, et la compétition, sont omniprésents dans la nature et impliquent souvent le métabolisme. La libération de métabolites par des organismes dans l'environnement permet à d'autres individus de la même espèce ou de différentes espèces de les récupérer pour leur usage propre. Dans cette thèse, nous étudions comment les interactions entre espèces façonnent l'environnement. Nous examinons les questions de (i) quels sont les besoins minimaux en éléments nutritifs pour établir la croissance, et (ii) quels métabolites peuvent être échangés entre un organisme et son environnement. L'énumération de tous les ensembles minimaux stoechiométriques de précurseurs et de tous les ensembles minimaux de métabolites échangés, en utilisant des modèles complets de réseaux métaboliques, fournit un meilleur aperçu des interactions entre les espèces. Dans un environnement spatialement homogène, les métabolites qui sont libérés dans un tel environnement sont partagés par tous les individus. Le problème qui se pose alors est de savoir comment les tricheurs, les individus qui profitent des métabolites libérés sans contribuer au bien public, peuvent être exclus de la population. Ceci et d'autres configurations ont déjà été modélisées avec des approches de la théorie des jeux et de l'économie. Nous examinons comment les concepts d'ensembles minimaux de précurseurs stoechiométriques et d'ensembles minimaux de composés échangés peuvent être introduits dans ces modèles.

MOTS-CLEFS en français

symbiose; métabolisme; modélisation des réseaux métaboliques; énumeration; ensembles minimaux de précurseurs; ensembles minimaux de fabriques; ensembles minimaux de métabolites échangés; théorie des jeux; économie

Title in english

Study of the evolution of symbiosis at the metabolic level using models from game theory and economics

Abstract in english

Symbiosis, a term that brings all types of species interaction under one banner, is defined as a close association of different species living together. Species interactions that comprise microorganisms are of particular interest for agriculture, health, and environmental issues. All kinds of species interactions such as mutualism, commensalism, and competition, are omnipresent in nature and occur often at the metabolic level. Organisms release metabolites to the environment which are then taken up by other individuals of the same or of different species. In this thesis, we study how species interactions shape the environment. We examine the questions of (i) what are the minimal nutrient requirements to sustain growth, and (ii) which metabolites can be exchanged between an organism and its environment. Enumerating all minimal stoichiometric precursor sets, and all minimal sets of exchanged metabolites, using metabolic network models, provide a better insight into species interactions. In a spatially homogeneous environment, the metabolites that are released to such an environment are shared by all individuals. The problem that then arises is how cheaters, individuals that profit from the released metabolites without contributing to the public good, can be prevented

from the population. This and other configurations were already modeled with approaches from game theory and economics. We examine how the concepts of minimal stoichiometric precursor sets and minimal sets of exchanged compounds can be introduced into such models.

Keywords in english
symbyosis; metabolism; metabolic network modeling; enumeration; minimal stoichiometric precursor sets; minimal stoichiometric factories; minimal sets of exchanged compounds; game theory; economics

Contents

Introduction

This PhD thesis is about the evolution of **symbiosis** at the level of **metabolism** using **models from game theory and economics**. These are the three main components that will be discussed. In 1879, de Bary defined symbiosis as a close association of different species living together Bary (1879). This definition was later reduced to mutualism. However, the "de Bary" definition is accepted in current general biological textbooks (Martin and Schwab, 2012). Species interactions are omnipresent in nature and even across different taxa, *e.g.* among lichens (Schwendener, 1868), between plants and pollinators (Mitchell et al., 2009), heterotrophic coral animals and phototrophic dinoflagellate endosymbionts (Toller et al., 2001), sea anemones and anemonefish (Nedosyko et al., 2014), yucca plants and yucca moths (Pellmyr and Huth, 1994), legumes and nitrogen-fixing bacteria (West et al., 2002; Kiers et al., 2003; Simms et al., 2006), plants and ants (Edwards et al., 2006), plants and mycorrhizal fungi (Bever et al., 2009), fig trees and the fig wasps (Jandér and Herre, 2010), epiphytes that grow on certain woody plants (Schimper, 1888), and between ectoparasites such as lice and ticks living on the skin of domestic animals (Hopla et al., 1994). We certainly do not exaggerate too much by saying that no free-living species is isolated from the others.

Species interactions that comprise microorganisms are of special interest for agriculture, health, and environmental issues. The interaction between anaerobic methane oxidizing archaea and sulfate-reducing bacteria is accounted for the consumption of more than 80% of the ocean methane flux. An important part of the green house gas methane is thus not emitted to the atmosphere (Reeburgh, 2007). Bacterial consortia were shown to be important in remediation of soil and groundwater from pesticides (Dejonghe et al., 2003), heavy metals (Valls and de Lorenzo, 2002), radioactive and inorganic compounds (Glick, 2003). The focus on the human gut microbiota is increasing since the last fifteen years, motivated by its impact on the human physiology, metabolism, nutrition, and immune function. The disruption of this complex web of interactions, that contains up to one thousand microbial species, is associated with obesity, malnutrition, and diseases such as diabetes, inflammatory bowel disease, encompassing ulcerative colitis and Crohn's disease (Guinane and Cotter, 2013). Determining the composition of the microorganisms present in the soil and understanding their interactions with plants seems to be important for soil quality, health, resilience, and sustainable agricultural productivity (Welbaum et al., 2004). Microbial species interactions play a major role in the food industry where mixed-cultures are employed in the fermentation process for the production of cheese, fermented milks, amino acids, and organic acids (Sieuwerts et al., 2008). Consortia of different species are engineered by synthetic biologists for producing various products such as methane-containing biogas, solvents, biohydrogen (Kleerebezem and van Loosdrecht, 2007), enzymes, food additives, antimicrobial substances, and bioethanol (Bader et al., 2010).

Species interactions often act at the metabolic level. Organisms release metabolic compounds to the environment which are then taken up by other individuals (from the same or different species). This leads us directly to the topics discussed in this thesis. The metabolic interactions shape the environment and we are interested in understanding (i) under which

conditions a species can grow and (ii) which compounds can be produced and exported by an organism to the environment. The information about which compounds can be potentially exchanged with the environment build the basis of species interactions.

Chemical compounds are transformed through chemical reactions. The metabolic capabilities of an organism are represented by a metabolic network which is a complex structure due to the fact that compounds can be consumed and produced by several reactions. A metabolic network can be modeled in different manners, *e.g.* by (weighted) directed graphs, bi-partite graphs, hypergraphs, and by a matrix whose entries correspond to the *stoichiometry* that a compound is consumed or produced in a chemical reaction. Negative entries stand for the consumption, while positive entries reflect the production of a compound. To be able to simulate growth in metabolic network modeling, the network is augmented by an artificial reaction (*biomass reaction*) which consumes all chemical compounds in the appropriate amounts that a species is supposed to need to produce one gram of biomass.

Determining the conditions under which a species can grow is not only interesting for studying species interactions. Currently many microorganisms are not cultivable in the laboratory due to lack of knowledge about appropriate growth conditions, *e.g.* nutrients, pH, osmotic conditions, and temperature (Stewart, 2012). There are possibly many alternative nutrient sets that enable growth. We are in particular interested in the minimal ones, that is nutrient sets that must be at least present in the medium (environment) to sustain growth. Having no *a priori* about the quality of one minimal nutrient set, we enumerate all of them, letting the choice to the user to select her or his optimal solution. A minimal set of nutrients, that we call a *minimal precursor set*, is not restricted to the production of biomass. The concept of a minimal precursor set can be generalized such that it allows the production of a given set of target compounds. Minimal precursor sets can be computed, taking stoichiometry into account, for any target compound of interest, *e.g.* a compound whose over-production is desired.

Taking a minimal precursor set as starting point for the production of a set of target compounds (*e.g.* biomass), there are multiple paths that connect the source with the target compounds. Again, the enumeration of all minimal sets of reactions, henceforth called *factories*, that allow the production of a set of targets from a given minimal precursor set is desirable. As we will see, this is a difficult task. In any case, the factories provide the information of which compounds can be produced and thus possibly exported to the environment.

In a given environment, different species interactions can arise, ranging from competition to mutualism. As mentioned above, organisms have several metabolic pathways to convert chemical compounds, *e.g.* ATP can be produced by respiration and fermentation or a mixture of both. The respiration pathway produces about $20 - 30$ times more molecules of ATP per molecule of glucose than by fermentation (Voet and Voet, 2011). However, ATP is produced by respiration at a lower rate compared to fermentation. Respiration thus uses the nutrients from the environment efficiently which would be optimal for growth if all individuals in the population would adopt this strategy. In a spatially homogeneous environment and under the assumption that every individual acts according to its self-interest, then fermentation and thus the depleting of the common resource is the best strategy. This configuration is well-known as the “Tragedy of the commons" in economics and game theory. In the case of mutualistic cross-feeding where two species depend on each other, that is one provides a compound to the other and vice versa, the following dilemma arises. An individual that does not provide the compound for the other species has an advantage compared to other individuals of the same species, because it saves the cost for the production of the compound but still benefits of the compound provided by the other species. The question is how cheaters can be prevented from the population. Another intriguing phenomenon is the evolution of commensalism in *Escherichia coli* when growing on a glucose-limited medium for long periods in continuous

culture. After some generations, two phenotypes evolve, one that partially degrades glucose into acetate which is then exported to the environment, and another strain that grows on acetate (Rozen and Lenski, 2000). What is the selective advantage in degrading glucose partially by several strains compared to the degradation by only one single strain?
This thesis is organized as follows. First, we introduce the biological and mathematical concepts used throughout the manuscript. In chapter 3, we discuss the relationship between minimal stoichiometric precursor sets and an ancestor approach that takes only the topology into account. We provide two methods, even though only one of them is of practical use, for the exhaustive enumeration of minimal stoichiometric precursor sets. In chapter 4, we address the problem of the enumeration of all minimal factories from a given minimal precursor set that enables the production of a set of targets. In this context, we show how minimal cut sets (sets of reactions that, if they were removed or blocked, would prevent the production of the target) can be employed to enumerate a subset of factories. The last chapter is dedicated to game theoretical approaches and models from economics to study species interactions at the metabolic level.

Chapter 1

Biological Concepts

Contents

This chapter contains the main biological concepts that are used throughout the thesis. The first part is devoted to symbiosis followed by a section about metabolism.
First, we provide (i) a definition of the term *symbiosis*, and (ii) some types and examples of symbiotic relationships. We will see that there is a continuum of symbiotic relationships. However, some levels may be distinguished.
In the second part of this chapter, some key concepts of a metabolic network are described, *e.g.* chemical compounds, chemical reactions, and enzymes.

1.1 Symbiosis

At the end of the 19th century, Anton de Bary discovered that lichens are a close association of algae and fungi. In this context, he defined in 1879 the term *symbiosis* as a close association of different species living together (Bary, 1879). His definition includes mutualism, commensalism, and parasitism. Species in a mutualistic interaction provide reciprocal benefits. The interaction may be more or less beneficial for the involved species. In contrast, a parasite benefits at the expense of the host. Parasitoïdism can be further distinguished from parasitism due to the fact that the host is killed or sterilised. In commensalism, one species benefits from the interaction whereas the other species neither benefits nor is harmed by the interaction. It is however debatable if an interaction can be completely neutral to a species (Parmentier and Michel, 2013). Not every species interaction fits well in only one of the latter categories. Symbiosis can thus be better described as a continuum ranging from mutualism to parasitism (Martin and Schwab, 2012; Parmentier and Michel, 2013). One year before de Bary, Albert-Bernhardt Frank had already used the term *Symbiotismus* when he studied the relationship between fungi and the roots of forest trees (Sapp, 2004). Since then, the definition of symbiosis as mutualism, commensalism, and parasitism was contested. Between 1960 and 1990, symbiosis was thought to be equivalent to mutualism. Nowadays, the "de Bary" definition is accepted in most general biological text books (Martin and Schwab, 2012).
Despite the continuum of species interactions, it is possible to categorise symbiosis at different levels. Species living in symbiosis are often distinguished as hosts (the larger organisms) or symbionts (the smaller organisms) where the latter usually benefit more from the interaction (Parmentier and Michel, 2013). Moreover, one can differentiate between *ectosymbiosis* (the

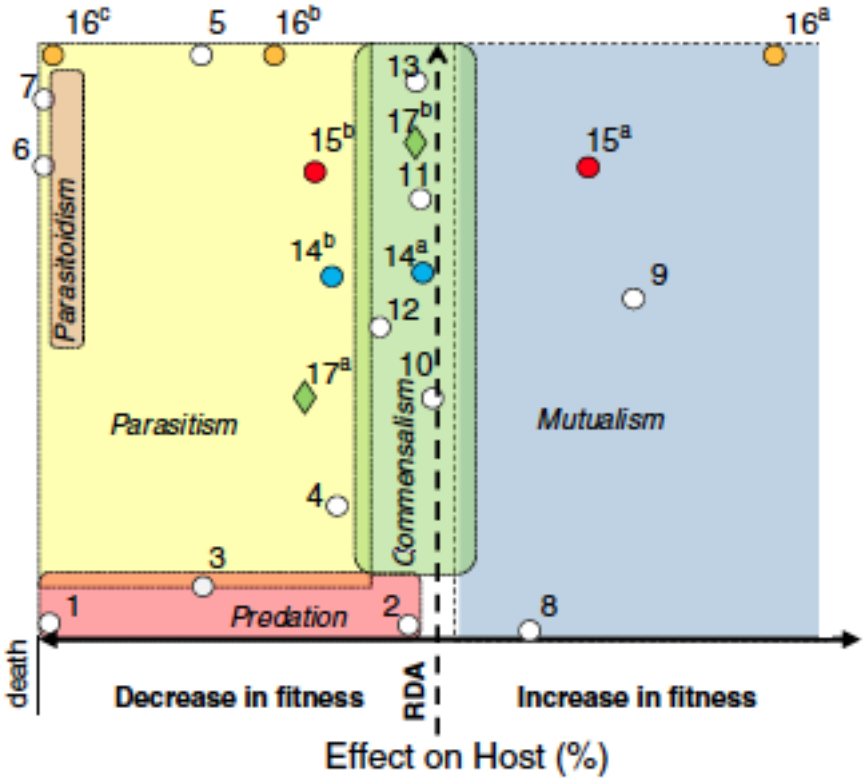


Figure 1.1: Different species interactions classified by the effect on the host (X-axis) and by the **R**elative **D**uration of the **A**ssociation (RDA) on the Y-axis. The different points correspond to species interactions observed in nature and are partially explained in the text. (Figure from Parmentier and Michel (2013))

symbiont lives outside the host) and *endosymbiosis* (the symbiont lives inside the host). Peacock (2011) further divides the term ectosymbiosis into: (i) interactions where one species lives on the surface of others, and (ii) more distant associations which he called *exosymbiosis*. In endosymbiosis, a symbiont can live in the intra- or extracellular space of the host. A symbiotic association can be obligate or facultative for the species involved.

Parmentier and Michel (2013) suggest a scheme (see Figure 1.1) to classify species interaction by (i) the impact on the host, and (ii) the **R**elative **D**uration of the **A**ssociation (RDA). RDA is defined as the ratio of the duration of the association to the life expectancy of the symbiont (Parmentier and Michel, 2013). Different species interactions are classified according to the two factors, *e.g.* the predation of the rabbit by the wolf (point 1 at the bottom left in the Figure 1.1) is characterized by a low RDA and a decrease in fitness for the rabbit (death). Broomrape is a genus of parasitic plants lacking chlorophyll and that are dependent on other plants for their nutrients (point 5). Other species interactions can be classified as parasitoïdism (points 6, and 7), commensalism (points 10, 11, 12, and 13), and mutualism (points 8, and 9). The more interesting aspect of this graph is its capacity to depict the variability of the species interactions during the lifespan of the symbionts (numbers with a superscript). Indeed, it was reported that *Pinnotheres* crabs are able to pass from commensalism to parasitism feeding either on excrement in the pallial cavity or on parts of the gill tissues of the mussel (points 14^a, and 14^b). The *Chlorella* algae is even more extreme in the association with a freshwater *Hydra* by switching between mutualism and parasitism (points 15^a, and 15^b). During the day,

the algae photosynthesize and provide carbohydrates to the *Hydra*. The carbon flow reverses in the night.

The microbiota in the human gut provides the host with several vitamins. It also digests complex polysaccharides, maintains the intestinal epithelial barrier, and makes the host resistant to pathogens in exchange for nutrients. Some species of the microbiota have pathogenic properties that are expressed due to switches in the species composition or changes of environmental conditions. This can transform a mutualistic relationship into disease or death (points 16^a, 16^b, and 16^c) (Parmentier and Michel, 2013). These examples make clear that a relationship between species may evolve and thus can often not be characterized by a single category.

Finally, it is important to stress that symbiosis is ubiquitous in nature. Two examples among many others are lichens, and mycorhrhizas. In the first case, lichens are classically described as a symbiotic association between a photobiont (green algae and/or cyanobacteria) and a mycobiont. However more recently, a third partner was also identified in some cases: the bacteriobiont (associated bacterial communities). Lichens are observed in temperate climate areas as well as in subarctic climate areas. Furthermore, they were observed on many kinds of substrate (rocks, soil, trees). In the second case, mycorrhizas describe a mostly mutualistic relationship between a fungus and vascular plants. One can differentiate between ectomycorrhizas and endomycorrhizas, where in the former, the fungus and the root cells build an intercellular interface, whereas in the latter the fungus penetrates the root's cell wall. To this day, microbial symbionts are found in association with animals, plants, insects, fishes, and birds. Furthermore, it should be highlighted that mitochondria of eukaryotic cells and chloroplasts of plants and protists were free-living bacteria before starting an endosymbiotic relationship with a host cell (Sapp, 2004). Symbiosis can thus be seen as fundamental in nature.

1.2 Metabolic Network

A metabolic network can be seen as a factory used by a cell to survive, grow, and reproduce. The building blocks of such factory are henceforth called *chemical compounds*. An individual production step that transforms some compounds into others is henceforth called a *chemical reaction*. Consider for instance the following reaction:

$$1\,CH_4 + 2\,O_2 \rightarrow 1\,CO_2 + 2\,H_2O.$$

We call *substrates* the compounds methane (CH_4) and oxygen (O_2) that are on the left side, and *products* the compounds carbon dioxide (CO_2) and water (H_2O) that are on the right side. The values before each compound are the stoichiometric coefficients and refer to the quantities of the compounds that are consumed or produced by the reaction. We say that the above reaction transforms one compound of methane and two compounds of oxygen into one compound of carbon dioxide and two compounds of water. The number of carbon ($1\ C$), hydrogen ($4\ H$), and oxygen ($4\ O$) atoms are equal between the substrate and the product side; we say that the reaction is *mass balanced*.

Many reactions in metabolic networks are depicted as above suggesting that they are unidirectional. However theoretically all reactions are reversible such that we should write:

$$\mathrm{S} \underset{k_r}{\overset{k_f}{\rightleftharpoons}} \mathrm{P}.$$

The substrate S is converted into the product P at a rate k_f (forward direction). The product P is transformed into S at a rate k_r (reverse direction). An equilibrium point is reached when k_f and k_r are equal. The constant $k_{eq} = [P]/[S]$ denotes the ratio of product and substrate concentrations at this equilibrium point. If $k_{eq} < 1$, it means that the reaction favors the consumption of the substrate S to produce P. If $k_{eq} > 1$, then the reverse direction is favored (Storey, 2004). There is the following relationship between the constant k_{eq} and the change in free energy of the system (ΔG), and the energy change measured under standard conditions ($\Delta G^{\circ\prime}$, measured at pH 7.25°C and 1M aqueous solution concentration):

$$\Delta G = \Delta G^{\circ\prime} + RT \times ln(k_{eq}),$$

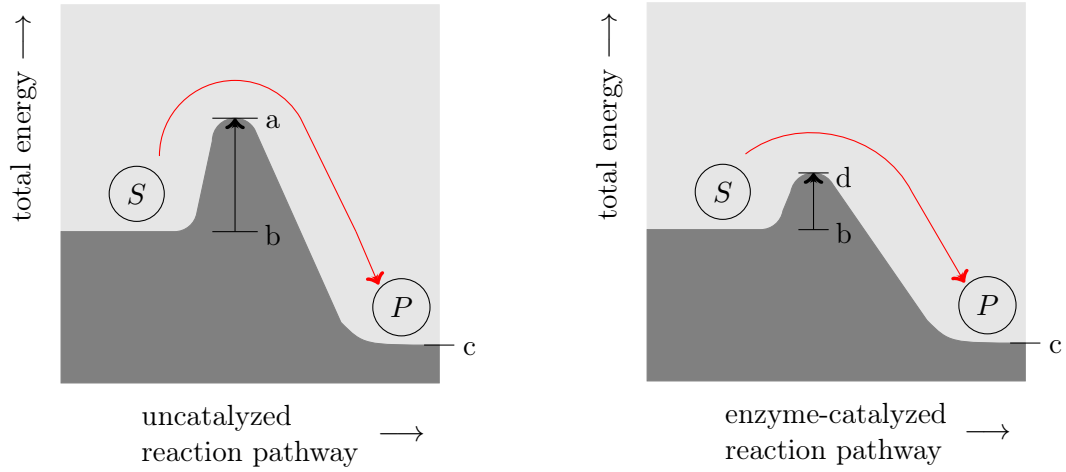
where R is the gas constant, and T is the temperature in degree Kelvin (Storey, 2004). At equilibrium, that is when ΔG equals zero, one can determine $\Delta G^{\circ\prime}$ as follows:

$$\Delta G^{\circ\prime} = -RT \times ln(k_{eq}).$$

The reaction's preference for a direction can also be expressed by the change in free energy of the system (ΔG). W. Gibbs provided a formula that relates ΔG to the change in enthalpy (ΔH) and entropy (ΔS):

$$\Delta G = \Delta H - T\Delta S,$$

where T is the temperature. The enthalpy relates to the internal energy of the system. The entropy can be seen, though in a simplified view, as the degree of randomness or disorder of the molecules of a system (Storey, 2004). When ΔG of a reaction is negative, it means that energy is released and the accumulation of the products is favored (forward direction). A reaction with a positive ΔG instead requires energy and favors the accumulation of the substrates (Storey, 2004; Alberts et al., 2010).
The thermodynamic concepts above pinpoint the favored direction of a chemical reaction. However, it says nothing about the velocity of the reaction. Even a reaction with a negative ΔG needs energy to break the chemical bonds of the substrates before the transformation into products. This energy is called *activation energy*. As depicted in Figure 1.2a, a reaction needs at the beginning some activation energy (energy a minus energy b) to overcome the energy barrier. At this stage enzymes come into the play. Enzymes are able to lower this activation energy (see Figure 1.2b); enzymes catalyse reactions. The activation energy in Figure 1.2b (energy d minus energy b) is smaller than in Figure 1.2a (energy a minus energy b). This enables the substrate S to overcome more easily the energy barrier and hence the reaction happens more often. The activation energy needed to start the reaction is the reason why in practise some reactions are *irreversible*. This is because the activation energy may be very high, such that it happens rarely that the substrates obtain the required energy from their surroundings (Alberts et al., 2010). The issue is depicted in Figure 1.2b. The activation energy of the reaction $P \rightarrow S$ (energy d minus energy c) is greater than the activation energy of the reverse direction (energy d minus energy b). It may therefore be very hard for the compound P to overcome the energy barrier and to be transformed into the compound S.



(a) (b)

Figure 1.2: Activation energy. (a) The activation energy (energy a minus energy b) is required to transform S into P. (b) An enzyme lowers the activation energy (energy d minus energy b). The figure is adapted from Alberts et al. (2010).

Several chemical reactions are conceptually associated together to fulfill a specific task. Such an association is called a *metabolic pathway*. The reactions within a metabolic pathway are generally linked such that the product of a reaction becomes the substrate of another reaction. Cells exhibit alternative pathways for the production of some compounds to face different environmental conditions. Many compounds are furthermore used in several pathways, making a metabolic network a complex and tangled structure. The sum of all reactions is called metabolism (Alberts et al., 2010).

Chapter 2

Mathematical Concepts

Contents

The following chapter is devoted to the mathematical concepts used throughout the thesis. In the first part, the focus is put on the metabolic network modeling. Directed graph models were applied to study the topological aspects of metabolic models (Fell and Wagner, 2000; Wagner and Fell, 2001; Ma and Zeng, 2003); see Lacroix et al. (2008) for a review. A directed graph G is defined as a pair $G = (V, A)$ with a vertex set V and a set of arcs A that consists in ordered pairs of vertices of V. The ordering defines the direction of an arc. Thus, an arc (u, v) with $u, v \in V$ is an arc from u to v. Metabolic networks can be modeled by directed graphs in three different ways. Either the set of vertices V contains the chemical compounds (compound graph), or the chemical reactions (reaction graph) or even both (bipartite graph). The differences are depicted in Figure 2.1. The toy metabolic network in Figure 2.1a consists of three reactions. Note that all substrates of a reaction must be present to use the reaction, *e.g.* a cell can only use reaction r_3 if it possesses the compounds e and c. In the compound graph, the set of vertices represents the compounds. There is an arc between two compounds u, v if a reaction consumes u and produces v. The metabolic network of 2.1a modeled as a compound graph is shown in Figure 2.1b. A reaction graph uses the chemical reactions as vertex set V. There is an arc between two reactions r_1, r_2 if at least one product compound of r_1 is consumed by r_2. See Figure 2.1c for a reaction graph representation of the metabolic network of 2.1a. Both modeling approaches have limitations (Deville et al., 2003; Lacroix et al., 2008). To produce compound c, the metabolic network needs to transform a and b through reaction r_1. In the compound graph (2.1b), it is however possible to produce the compound c from either a or b, respectively. The problem of the reaction graph is that if there is more than one arc towards one vertex (a chemical reaction), one cannot distinguish if the arcs correspond to: (i) alternative ways to produce a substrate, or (ii) ways to produce

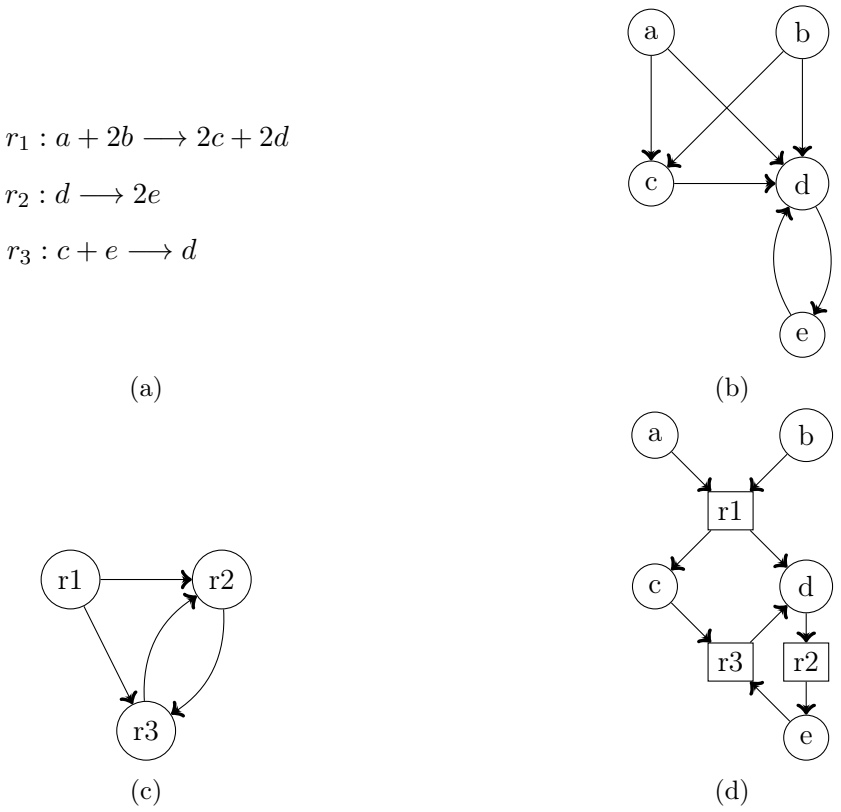


Figure 2.1: Metabolic network (a) modeled as a compound graph (b), reaction graph (c) or a bipartite graph (d).

all substrates of the reaction. In Figure 2.1c, there are two arcs that point towards r_3 which means – knowing the metabolic network – that the reaction r_3 needs a product from r_1 and r_2 in order to have the full set of its substrates. In contrast, the two arcs pointing to the vertex r_2 correspond to two alternatives to produce the substrate of reaction r_2. A bipartite graph avoids the latter problem. Here, the set of vertices is split between compounds and reactions. There is an arc from a compound u to a reaction v if v consumes u. There is also an arc from a reaction v to a compound u if v produces u. Arcs between the same types of vertices are not possible (compound-compound, reaction-reaction). As depicted in Figure 2.1d, one can now distinguish that the substrate of r_2 (compound d) can be produced either by reaction r_1 or r_3. However, the fact that all substrates of a reaction must be present in order for the reaction to happen is still not explicitly modeled. There is still the possibility to produce the compound c from either a or b passing through r_1.
To circumvent the latter issue, directed hypergraphs are used. We will see that a directed hypergraph is a very natural way to model a chemical reaction. The stoichiometry of the reactions can furthermore be incorporated in weighted directed hypergraphs or constraint-based models. Both models are used within the thesis and are presented in this chapter.
The second part of this chapter is dedicated to *game theory* which was first applied in economics. Later, game theory was adopted by biologists. Since then models from biology and economics inspired each other. We describe cooperative, and non-cooperative game theory.
The last part handles two models from economics that are worth to mention as they were already applied to species interaction (Mark W. Schwartz, 1998; Hoeksema and Schwartz, 2003; Wyatt et al., 2014; Tasoff et al., 2015).

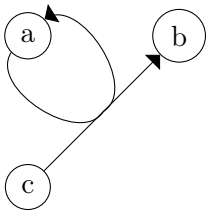


Figure 2.2: An autocatalytic reaction. Compound a is substrate and product of the same reaction: $a + c \rightarrow a + b$

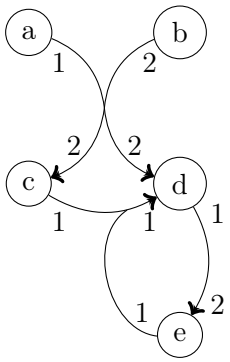
2.1 Metabolic Network modeling

2.1.1 Weighted Directed Hypergraph

To overcome the shortcomings of the above mentioned graph models, we use weighted directed hypergraphs to model metabolic networks. A metabolic network is defined as a pair $\mathcal{N} = (\mathcal{C}, \mathcal{R})$ with a set of vertices $\mathcal{C}$ (representing the chemical compounds), and a set of hyperarcs $\mathcal{R}$ (representing the chemical reactions) that consists of ordered pairs of subsets of $\mathcal{C}$, i.e. $r = (Subs(r), Prod(r)) \in \mathcal{R}$. Topologically, a reaction $r \in \mathcal{R}$ is defined by its substrates $Subs(r) \subseteq \mathcal{C}$ and its products $Prod(r) \subseteq \mathcal{C}$, suggesting the interpretation of a reaction as a directed hyperarc with $Subs(r)$ as the set of tail nodes and $Prod(r)$ as the set of head nodes of a reaction r. In the example metabolic network of Figure 2.1a, $Subs(r_1) = \{a, b\}$ and $Prod(r_1) = \{c, d\}$. Given a subset of reactions $F \subseteq \mathcal{R}$, we denote by $Subs(F)$ and $Prod(F)$ the union of the substrates and products, respectively, of the reactions in F.
In order to include the stoichiometry of a reaction, we can assign a weight to each substrate and product of the reaction. The network $\mathcal{N} = (\mathcal{C}, \mathcal{R})$ with the associated weights can then be seen as a weighted directed hypergraph. An illustration of the example metabolic network of Figure 2.1a is shown in Figure 2.3a. All reactions are precisely described in terms of the sets of substrates and products that take place in a reaction as well as their quantities. Although not used within this thesis, it is worth to mention that an autocatalytic reaction – a reaction that has a compound as substrate and product at the same time – can also be modeled through (weighted) directed hypergraphs (see Figure 2.2).
Metabolic networks were already modeled through (weighted) directed hypergraphs in various problems, *e.g.* subgraph centrality and clustering (Estrada and Rodríguez-Velázquez, 2006; Zhou and Nakhleh, 2011), measure of reciprocity (Pearcy et al., 2014), pathway enumeration (Mithani et al., 2009; Carbonell et al., 2012), and enumeration of minimal topological precursor sets (Cottret et al., 2007; Acuña et al., 2012).

2.1.2 Constraint-based models

Constraint-based models use a matrix representation to characterise a metabolic network $\mathcal{N} = (\mathcal{C}, \mathcal{R})$. This so-called stoichiometric matrix S is of dimension $|\mathcal{C}| \times |\mathcal{R}|$; each compound corresponds to a row, and each reaction corresponds to a column. The cell $S[i, j]$ refers to the consumption (production) of compound i by reaction j and is called stoichiometric value. If $S[i, j]$ is smaller (greater) than zero then the compound i is consumed (produced) by reaction j. A zero entry means that the compound i is not involved in the reaction j. Usually the stoichiometric matrix is sparse. Note that autocatalytic reactions cannot, contrary to (weighted) directed hypergraphs, be represented through the stoichiometric matrix. The stoichiometric matrix associated to the toy metabolic network of Figure 2.1a is shown in Figure 2.3b.



(a)

	r_1	r_2	r_3
a	-1	0	0
b	-1	0	0
c	2	0	-1
d	2	-1	1
e	0	2	-1

(b)

Figure 2.3: Metabolic network modeled as a weighted directed hypergraph (a), or as a stoichiometric matrix (b).

Usually constraint-based models assume *steady state*, that is

$$Sv = 0, \tag{2.1}$$

where the flux vector $v \in \mathbb{R}^{|\mathcal{R}|}$ denotes the flux of every reaction in the network per time unit. An additional constraint can be put on the flux of *irreversible reactions*, namely

$$v_i \geq 0. \tag{2.2}$$

Normally there is not a unique solution to this system of equations because a metabolic network has more reactions than compounds ($|\mathcal{R}| > |\mathcal{C}|$). Each flux vector v represents a capability of the metabolic network to achieve a certain goal. Different analyses of the flux vector can be grouped into three categories: (i) metabolic flux analysis (MFA), (ii) flux balance analysis (FBA), and (iii) metabolic pathway analysis (Trinh et al., 2008). The first two categories aim to find a single flux vector while the approaches in the latter category describe the full flux space.

Metabolic flux analysis takes advantage of the measurements of some external metabolic rates, *e.g.* growth rate, substrate uptake, and product accumulation (Antoniewicz, 2015). Dividing the flux vector v in a measurable flux vector v_m and an unmeasurable flux vector v_u, equation (2.1) can be formulated as:

$$S_u v_u = -S_m v_m, \tag{2.3}$$

where S_u (S_m) refers to the stoichiometric matrix S restricted to the columns of the unmeasurable (measurable) reactions. A large number of measured flux rates (v_m) is usually needed to make S_u invertible which is necessary for the computation of the unmeasurable flux vector v_u (Trinh et al., 2008):

$$v_u = -S_u^{-1} S_m v_m. \tag{2.4}$$

Note that v_u is calculated for a given measured flux rate vector v_m originating from a certain growth condition and the measurement of external metabolic rates. Hence, different growth conditions and/or rate measurements yield a different v_m and thus also a different v_u. Furthermore MFA returns only a single flux vector (Trinh et al., 2008; Antoniewicz, 2015). To

gain insight into the intracellular fluxes, further constraints are added that are obtained from measurements of ^{13}C-labeling tracers (Antoniewicz, 2015).

The goal of flux balance analysis is to find a single flux v that: (i) solves equation (2.1), and (ii) minimizes or maximizes an objective function:

$$Z = c^T v, \tag{2.5}$$

where the vector c of size $|\mathcal{R}|$ corresponds to the coefficient of the reactions in the objective function (Orth et al., 2010). The flux value of a reaction can be further limited by establishing a lower and an upper bound ($lb \leq v_r \leq ub$). An irreversible reaction is modeled requiring a non-negative flux ($v_r \geq 0$). To block a reaction r, *e.g.* an uptake reaction, the constraint $v_r = 0$ is added. The system (2.1), the objective function (2.5), and the inequality constraints on the flux v can be formulated in a linear programming problem:

$$\begin{aligned} \min Z = \quad & c^T v \\ s.t \quad & Sv = 0, \\ & lb \leq v_i \leq ub, \quad \forall i \in \mathcal{R} \end{aligned} \tag{2.6}$$

Solving this linear programming problem provides a single flux v maximizing the objective function. Note that there is usually more than one solution in the solution space of the stated problem. Lee et al. (2000) proposed an algorithm to enumerate all alternate optimal solutions via mixed integer linear programming (MILP). The usually huge solution space can be subdivided into smaller modules making their analysis easier (Kelk et al., 2012). Different objective functions are used in the literature, *e.g.* maximization of biomass production (Feist and Palsson, 2010), ATP production (Pramanik and Keasling, 1997) or minimization of metabolic adjustment (Segrè et al., 2002). Schuetz et al. (2007) compared the ^{13}C-determined *in vivo* fluxes in *Escherichia coli* under different environmental conditions to the solutions obtained from FBA using eleven different objective functions. The authors show that the nonlinear maximization of the ATP yield per unit of flux was the best objective function when *E. coli* grows on a rich glucose medium. Maximizing ATP or biomass yield were the best objective functions under scarce conditions (Schuetz et al., 2007). To summarise: FBA usually provides one out of many optimal flux solutions. The optimization criteria has to be chosen carefully. The above described methods provide a single solution for equation (2.1) requiring either data about fixed reaction rates (metabolic flux analysis) or an objective function (flux balance analysis) to reduce the solution space. The set of all flux vectors that fulfill equation (2.1) combined with the inequality constraint (2.2) defines a polyhedral cone which is called the *flux cone* (Clarke, 1980). There are two methods that provide an inner description of the flux cone based on generating vectors, namely *elementary flux modes* (EFM) (Schuster and Hilgetag, 1994), and *extreme pathways* (ExPa) (Schilling et al., 2000). The concepts named minimal metabolic behaviours (MMB) and the reversible metabolic space (RMS) by Larhlimi and Bockmayr (2009) use an outer description of the flux cone and are based on sets of non-negativity constraints. All these approaches, namely EFM, ExPa, and MMB together with RMS, offer a complete description of the flux cone.

To define an elementary mode, we first provide the definition of the support of a flux vector v as $supp(v) = \{r|v_r \neq 0\}$, that is the set of reactions that have a strict positive flux value in v. A mode is a set of reactions corresponding to the support of a flux vector v that act at steady-state. A mode is called *minimal* if it does not contain another mode. Extreme pathways are a subset of elementary modes. An extreme pathway is systemically independent of other extreme pathways (Schilling et al., 2000).

There are several algorithms and implementations for the enumeration of elementary modes (Schwarz et al., 2005; Kamp and Schuster, 2006; Hoops et al., 2006; Urbanczik, 2006; Klamt

et al., 2007; Terzer and Stelling, 2008; Jevremovic et al., 2011; Quek and Nielsen, 2014; Pey et al., 2014; Hunt et al., 2014). As the enumeration of all elementary modes is a difficult task for large genome-scale metabolic networks (there is a huge number of solutions), different approaches were proposed for the enumeration of subsets of elementary modes, *e.g.* the k-shortest elementary modes (de Figueiredo et al., 2009), elementary modes involving a set of target reactions (David and Bockmayr, 2014), or elementary modes with an optimal biomass yield (Müller and Bockmayr, 2013).

This chapter shows only some constraint-based modeling methods and is far from being exhaustive. Lewis et al. (2012) provide a more complete review including a nice "phylogenetic tree" of the different methods. In this thesis, we use constraint-based and weighted directed hypergraph models for the enumeration of minimal stoichiometric precursor sets and minimal stoichiometric factories.

2.2 Game Theory

Game theory is a mathematical model to analyze interactions between rational individuals (henceforth called *players*). The players have choices (henceforth called *actions*) on how to handle a given situation. A *payoff* is assigned to every player depending on the actions taken by all players. Each player wants to maximize its payoff which however depends on the actions of the other players. Game theory analyzes such situations to find out which actions the players should take to maximize their payoffs, that is to find their best *strategy*. Different games, *e.g.* chess, the card game le Her, dice games, and tic-tac-toe, were analyzed for best strategies since the 17th century (Broom and Rychtar, 2013). However, the book *Theory of Games and Economic Behavior* (von Neumann and Morgenstern, 1944) is seen as the start of formal studies of game theory. The recently died John Forbes Nash, Jr. made his major contribution to non-cooperative game theory with the concept of Nash equilibrium. Reinhard Selten introduced the subgame perfect equilibria (1965), and the trembling hand perfect equilibria (1975). At the same time, John Harsanyi established the distinction between cooperative and non-cooperative game theory (1966). He further introduced the theory of games with incomplete information (Broom and Rychtar, 2013). All these works find a large application in economics which was honored to Nash, Selten, and Harsanyi by the Nobel Prize for Economics in 1994. Some of these concepts will be important for the modelling of species interactions and are thus explained in this chapter.

The works of Darwin, Dusing and Fisher about the reason why natural selection tends to equalise the sex ratio already implicitly used game theoretical concepts. Lewontin was the first to apply explicitly game theory in his book entitled *Evolution and the Theory of Games* (1961). Later, Hamilton and Trivers applied game theory in their works about relatedness and altruism. Maynard Smith and Price (1973) developed the concept of an evolutionary stable strategy which is central to evolutionary biology, and as important as Nash equilibrium (Broom and Rychtar, 2013).

In the following section, we describe non-cooperative and cooperative game theory and some key concepts therein. Game theory was applied to economics before being introduced to biology. The enhancements in the latter field, namely the evolutionary game theoretical approaches, have found a role in economic models (Sandholm, 2010). There seems to be an exchange of ideas between economics and biology. Therefore, at the end of this section, we depict economic models that were already applied to species interactions (Mark W. Schwartz, 1998; Hoeksema and Schwartz, 2003; Wyatt et al., 2014; Tasoff et al., 2015).

		Player 2	
		Left	*Right*
Player 1	*Up*	3, 1	0, 0
	Down	2, 2	2, 2

Figure 2.4: A normal form game in matrix representation.

2.2.1 Non-Cooperative Game Theory

In this section, we provide a formal definition of a non-cooperative game and demonstrate that a game can be represented in two different manners: as a normal-form, and an extensive-form game. There are static and dynamic approaches for the prediction of which action(s) the players will choose. Such a prediction is called a solution of a game. Several solution concepts such as the famous Nash equilibrium and evolutionary stable strategies (both static approaches), as well as replicator and adaptive dynamics are described. Some typical games and their analysis will be briefly presented below.

Game in normal- and extensive-form

In non-cooperative game theory, an n-player game in its *normal* or *strategic form* is defined by $G = (N, \{S_i\}_{i \in N}, \{u_i\}_{i \in N})$, where $N = \{1, \ldots, n\}$ is the set of *players*. Player i has a set S_i of *actions* or *strategies*. If the action set of every player is finite, the game is called *finite*. The *payoff* or *utility function* $u_i : S \to \mathbb{R}$ assigns a value to player i, where S is called the *profile* and is defined as: $S := S_1 \times \cdots \times S_n$. Note that the payoff (utility) of a player depends on the strategies chosen by all players. A normal form game refers to a game that is played once and where the players choose their respective action simultaneously. The players are supposed to know the details of the game (number of players and the action set of all players). Furthermore, the players are assumed to be rational, that is, they try to maximize their respective payoff. A game in normal form can be represented in matrix form as in Figure 2.4. In this 2-player game, the action set of player 1 (also called the *row-player*) consists in $\{Up, Down\}$, and the one of player 2 (also called *column-player*) in $\{Left, Right\}$. The first (second) value in each cell of the matrix corresponds to the utility of the row-player (column-player). The matrix can be read as follows: If the row-player plays *Up* and the column-player plays *Left*, the row-player gets an utility of 3 and the column-player gets an utility of 1. The same reasoning can be applied to the other strategy profiles, namely $\{Up, Right\}$, $\{Down, Left\}$, and $\{Down, Right\}$.
A game in extensive form considers the case when the players take their actions sequentially. An extensive-form game under perfect information means that every player is aware of all previous actions when it is his or her turn to make a decision. Such a game can be defined as follows:

Definition 1. *(Osborne and Rubinstein, 1994) A game in extensive-form is a triple* $G = (N, H, P)$*, with*

- N players,
- *A set* H *of sequences with the following properties:*
  - $\emptyset \in H$,
  - *If* $(a^k)_{k=1,\ldots,K} \in H$ *(where* K *may be infinite) and* $L < K$ *then* $(a^k)_{k=1,\ldots,L} \in H$.

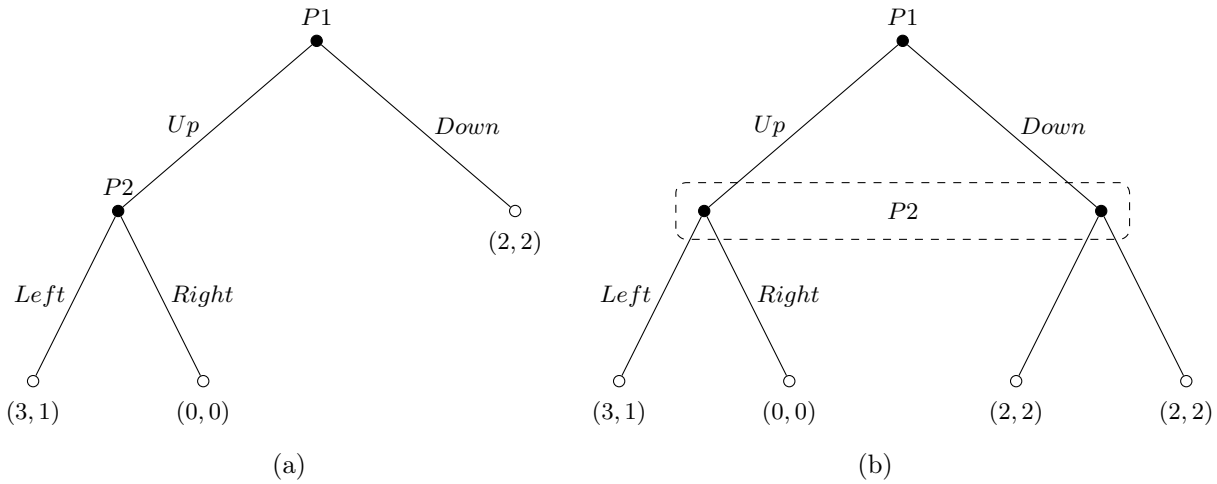


Figure 2.5: Two games in extensive form with perfect (a) and imperfect (b) information.

  - *If an infinite sequence* $(a^k)_{k=1}^{\infty}$ *satisfies* $(a^k)_{k=1,\dots,L} \in H$ *for every positive integer* L *then* $(a^k)_{k=1}^{\infty} \in H$.

  An element of H *is called a* **history**. *The elements* a^k *of a history denote the actions of the players. A history* $(a^k)_{k=1,\dots,K} \in H$ *is* terminal *if it is infinite or if there is no* a^{K+1} *such that* $(a^k)_{k=1,\dots,K+1} \in H$.

- *A function* P *that assigns a member of* N *to each non-terminal history.*

An extensive-form game can be represented by a tree as in Figure 2.5a. At each internal node a player is assigned to take an action; here, player 1 ($P1$) moves first (*Up* or *Down*). If player 1 has chosen *Up*, player 2 moves afterwards (*Left* or *Right*). The payoffs for both players are assigned to terminal nodes, *e.g.* if player 1 plays *Up* and then player 2 plays *Left*, the first (second) player receives a payoff of 3 (1). If player 1 plays *Down* then both players receive each a payoff of 2. In this situation, it is not described by the game if player 2 has not the "right" to act, or if it does not matter if he plays *Left* or *Right* (both receive a payoff of 2 anyway). In the latter case, the tree is a compact representation of the game. Under perfect information, each player has the full information about the actions taken before (the *history*). In this example, player 2 knows if player 1 has chosen the action *Up* or *Down*. In contrast, in a game with imperfect information, the players have only partial information about the history. In Figure 2.5b, player 2 does not know what player 1 did in the first step. This situation is represented by the dashed box.
Every game in extensive form can be represented by exactly one matrix-form game. This can easily be verified for both extensive-form games in Figure 2.5. The payoffs given at each terminal history (leafs of the tree) correspond to the payoffs of a cell in the matrix-form game. The associated actions of the row- and the column-player can then be determined by following the edges in the tree from the terminal history (leafs) to the empty history (root). Taking Figure 2.5a as an example, the payoff vector $(3, 1)$ (bottom left) is assigned when player 2 plays *Left* and player 1 plays *Up*. Instead, if player 2 plays *Right* and player 1 plays *Up*, both get a payoff of zero. When player 1 plays *Down*, both players receive a payoff of two no matter the action of player 2. The matrix of Figure 2.4 therefore represents the extensive-form game of Figure 2.5a. This holds also for the second extensive-form game in Figure 2.5b. Hence, every extensive-form can be represented by exactly one matrix-form game but a matrix-form game can be represented by several extensive-form games.

		Husband	
		F	O
Wife	F	$1, 2$	$0, 0$
	O	$0, 0$	$2, 1$

Figure 2.6: The *Battle of the Sexes* game.

Nash Equilibrium

The Nash equilibrium is a solution concept for strategic-form games. Recall that the payoff function u depends on the simultaneously taken actions of all players. The chosen actions of every player constitute a strategy profile: $S := S_1 \times \cdots \times S_n$. Then, loosely speaking, a strategy profile $x \in S$ is a Nash equilibrium when no player has an incentive to deviate from its strategy, that is no player can get a higher payoff by deviating unilaterally from its strategy. Before providing a formal definition of a Nash equilibrium, we denote by x_i the strategy profile of player i, and by x_{-i} the strategy profile of all players different from i.

Definition 2. *The strategy profile $x^* \in S$ is called a (weak)* **Nash equilibrium** *if $\forall i, x_i \in S_i : u_i(x_i^*, x_{-i}^*) \geq u_i(x_i, x_{-i}^*)$.*

The above definition illustrates that player i has no influence on the actions chosen by the other players: the strategy profile x_{-i}^* is fixed. When player i supposes that the other players play x_{-i}^*, then he cannot be better off than playing x_i^*. If we apply this reasoning to every player $i \in N$, then x^* is an equilibrium point. When we ask for strict inequality in Definition 2, then x^* is called a *strict Nash equilibrium*. In this solution concept, it is not stated how this equilibrium point is reached.
Taking the example strategic-form game of Figure 2.4, we will show whether a strategy profile is a Nash equilibrium. The strategy profile $\{Up, Left\}$ is a strict Nash equilibrium for the following reasons: If player 1 would play *Down* (bottom left cell) instead of *Up*, he would get a smaller payoff ($2 < 3$). If player 2 would deviate from its strategy by playing *Right*, he would get only a payoff of zero (up right cell). These arguments make clear why the strategy profiles $\{Up, Right\}$, and $\{Down, Left\}$ are not Nash equilibria. Considering $\{Down, Left\}$, player 1 is better off playing *Up* when player 2 plays *Left* (payoff $3 > 2$). Regarding the strategy profile $\{Up, Right\}$, both players have an incentive to deviate their respective strategies: player 1 is better off playing *Down* because then his payoff is 2 (cell below right). Player 2 prefers to play *Left* as his payoff would be 1 (cell above left). The strategy profile $\{Down, Right\}$ is a weak Nash equilibrium. No player has an incentive to change its strategy unilaterally. However, it is not a strict Nash equilibrium because for player 2 it does not hold that $u_2(Right, Down) > u_2(Left, Down)$.
Until now, we considered the case where a player has the option to choose only one action from its set of pure strategies. This can however be generalized to *mixed strategies*, where an action can be played with a certain probability. Nash proved that there always exists a mixed strategy Nash equilibrium in a finite strategic-form game (Nash, 1951). The computation of a mixed strategy Nash equilibrium will be demonstrated with the help of the matrix-form game *The Battle of the Sexes*. Here, a couple wants to spend an evening together. There are two options: going to the opera or watching a football match in the stadium. The husband prefers the football match, whereas the wife prefers going to the opera. However, both would be upset if they did not spend the evening together. The matrix in Figure 2.6 represents this situation. Both get a payoff of zero if they do not go to the same event. If they decide to go to the same event, then the wife prefers the opera ($u_W(O, O) > u_W(F, F)$) and the

husband prefers the football match ($u_H(F,F) > u_H(O,O)$). The strategy profiles (F,F) and (O,O) are pure strategy Nash equilibria. In both cases, nobody has the incentive to deviate unilaterally from his or her strategy. However, neither the wife nor the husband can predict the action of the other. Should the husband go to the opera because he thinks that the wife prefers it? What happens if the wife applies the same reasoning?
If there is more than one pure strategy Nash equilibrium, then it is not clear which one is finally chosen by the players. Playing a mixed strategy can resolve this uncertainty. Suppose the husband goes to the football match with probability p, and to the opera with probability $1-p$. Similarly, the wife goes to the football match with probability q, and to the opera with probability $1-q$. Given the wife's mixed strategy, the expected payoff for the husband going to the football match is $u_H(F) = 2 \times q + 0 \times (1-q) = 2q$, and the expected payoff for going to the opera is $u_H(O) = 0 \times q + 1 \times (1-q) = 1-q$. Going to the football match is the best response to the wife's mixed strategy if $u_H(F) \geq u_H(O)$, which is the case if $q \geq \frac{1}{3}$. Analogously, going to the opera is a best response if $q \leq \frac{1}{3}$. So, the husband should go to the football match when $q \geq \frac{1}{3}$ and to the opera if $q \leq \frac{1}{3}$. At $q = \frac{1}{3}$, the husband is indifferent playing any mixed strategy that contains both strategies because his payoff will always be $\frac{2}{3}$. The same calculation for the wife's expected payoffs yield that she is indifferent when the husband goes to the football match with probability $p = \frac{2}{3}$. Thus, at the mixed strategy Nash equilibrium, the husband goes to the football match (opera) with probability $\frac{2}{3}$ ($\frac{1}{3}$) and the wife goes to the football match (opera) with probability $\frac{1}{3}$ ($\frac{2}{3}$). The probability to go to the same event is $\frac{2}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{3} = \frac{4}{9}$, and the expected payoff for the husband and the wife is $2 \times \frac{2}{9} + 1 \times \frac{2}{9} = \frac{2}{3}$.

Evolutionary Stable Strategy

The solution concept of an *evolutionary stable strategy* (ESS), introduced by Smith and Price (1973), is widely used in biology. Here, in its most basic version, random encounters of two individuals from a single infinite population are considered. At each encounter, the selected individuals play a game that can be represented in matrix-form. The payoff entries correspond to the fitness of an individual. Fitness can be defined in several ways, *e.g.* as the reproductive success of an individual over its lifetime. In contrast, inclusive fitness considers the number of gene copies. Even though an individual may not reproduce, it shares identical genes with other individuals that may reproduce and thus increase the number of gene copies in the next generation (Broom and Rychtar, 2013). An individual may participate in several games, that is, the individual can be selected at random many times to participate in a game. Each game is independent of the others. The total payoff for an individual is the average of its payoffs received in all games.
The actions of a player correspond to inherited phenotypes. Mixed strategies (phenotypes) of an individual are interpreted at the population level as follows: A fraction of the population has phenotype A and the remaining part of the population has phenotype B. If two players, one having phenotype A, the other having phenotype B, meet each other in a game, the individual with the higher fitness will win the competition. Thus, natural selection replaces the rationality of the players. Of course, a player does not *choose* a phenotype, but natural selection determines the fittest phenotypes to survive and replaces individuals with a lower fitness. A phenotype A that is a best response to all other phenotypes in the population corresponds to a (weak) Nash equilibrium. However, a population that consists of almost all individuals having phenotype A is not immune to the invasion of a small number of mutants B. This leads us to the definition of an evolutionary stable strategy:

	D	H
D	$\frac{V}{2}, \frac{V}{2}$	$0, V$
H	$V, 0$	$\frac{V-C}{2}, \frac{V-C}{2}$

Figure 2.7: A Hawk-Dove game.

Definition 3. *A strategy A is an* evolutionary stable strategy *if one of the following conditions holds for all strategies $B \neq A$:*

- $u(A, A) \geq u(B, A)$,
- *if* $u(A, A) = u(B, A)$, *then* $u(A, B) > u(B, B)$.

A phenotype A is an evolutionary stable strategy if it is a Nash equilibrium, and in the case when $u(A, A) = u(B, A)$, the individuals with phenotype A have an advantage when playing against B ($u(A, B) > u(B, B)$). Let us consider the second condition, that is we assume that $u(A, A) = u(B, A)$. A small number of mutants B can invade a population of phenotype A if B gains at least as much as A when they encounter another mutant B ($u(B, B) \geq u(A, B)$). The latter condition can be split into two cases. If $u(B, B) > u(A, B)$, the phenotype B is an ESS by definition and thus will spread out in the population. If $u(B, B) = u(A, B)$, then there is no selective advantage for any of the two phenotypes and the fraction of phenotype B increases or decreases by random chance (genetic drift) (Broom and Rychtar, 2013).
The famous Hawk-Dove game of (Smith and Price, 1973) describes the following situation in a population of birds: There are two phenotypes, called *Hawks* and *Doves*, in the population fighting for a resource V. Both phenotypes display aggression when fighting for the resource, but only the Hawks get into the fight. If a Hawk contests another Hawk, there will be a fight that he wins (he gets the resource V) or loses half of the time. In the latter case, the Hawk gets injured (modeled by a cost C). If a Hawk encounters a Dove, the Hawk gets the entire resource as the Dove avoids the fight. The resource is shared equally if two Doves meet. The payoff matrix in Figure 2.7 represents the above described situation.
The outcome of the game depends on the values of the parameters V, and C. If $V > C$, then the phenotype Hawk is a pure ESS because $u(H, H) > u(D, H)$. If $V < C$, Hawks and Doves coexist in the population. There is a mixed ESS consisting in that there are V/C individuals with phenotype Hawk and $1 - V/C$ individuals with phenotype Dove (Osborne and Rubinstein, 1994).

The solution concept of an evolutionary stable strategy is a refinement of the Nash equilibrium. The matrix-form game in Figure 2.8 illustrates an example where a Nash equilibrium is *not* an ESS. Here, the strategy profiles (A, A) and (B, B) are Nash equilibria. However, only the strategy (phenotype) B is also an ESS. The strategy (phenotype) A is not an ESS as the second condition of the Definition 3 does not hold: $u(A, A) = u(B, A)$ but $u(A, B) < u(B, B)$. Thus, a population entirely consisting of individuals with phenotype A can be invaded by B mutants.

Symmetric games classification

We have already seen some well-known games, *e.g.* the Battle of the Sexes and the Hawk-Dove game. Here, we want to provide a classification of *symmetric* strategic-form games and describe some of them more in detail. A game is called symmetric if the payoff for playing a strategy depends only on the other chosen strategies and not on who has chosen the strategies.

	A	B
A	$2, 2$	$1, 2$
B	$2, 1$	$2, 2$

Figure 2.8: Not every Nash equilibrium is also an ESS.

	C	D
C	R	S
D	T	P

Figure 2.9: The generic payoffs in a symmetric strategic-form game.

This means that it is not important if someone acts as row- or column-player; the payoff will be the same. The payoffs of a *symmetric* strategic-form game can be represented by a matrix as shown in Figure 2.9. The actions (**c**ooperate and **d**efect) and the payoff acronyms originate from the Prisoner's Dilemma game (see below), where R corresponds to a reward if both players cooperate; P stands for a punishment if both defect. If one player cooperates and one player defects, then the cooperator receives the so-called sucker's payoff (S) and the defector receives the temptation payoff (T).
It we fix $R > P$, then it is possible to classify 12 different games depending on the values of S and T (see Figure 2.10 from Hummert et al. (2014)). We have already analyzed the Hawk-Dove game (region 2 in Figure 2.10) where $T > S$ (above the diagonal line $T = S$), $T > R$ (above the horizontal line labeled $T = R$), $S < R$ (left from the vertical line $S = R$), and $S > P$ (right from the vertical line $S = P$). Thus, a game with the payoffs $T > R > S > P$ constitutes a Hawk-Dove game.
For a full description of all the 12 games, we refer to the publication of Hummert et al. (2014). Here, we will analyze only two of them.

Prisoner's Dilemma
The Prisoner's Dilemma was first stated in unpublished works by Raiffa (1951), and by Flood and Dresher (1952) before being formalized by Tucker (Osborne and Rubinstein, 1994). Two criminals of a gang are arrested and both are kept in solitary confinement making communication between them impossible. The prosecutor has not enough solid evidence to convict both suspects for the principal criminal act. There is only evidence for a minor crime for which both suspects go one year to jail. So, the prosecutor proposes simultaneously a deal to each suspect: Each suspect has the choice to testify that the other has committed the principal criminal act (defect), or to stay silent (cooperates with the other suspect). If both testify that the other has committed the crime, then both of them go to jail for two years. If only one of them stay silent, then the defector will be released and the cooperator go to jail for three years. If both cooperate, then they will be imprisoned one year for the minor criminal act. The game is depicted in Figure 2.11 where the values correspond to the number of years spent in prison. The best strategy is to defect because it is the best response to the action of the other player: If the other player cooperates, it is best to defect ($0 < 1$). If the other player defects, it is best to defect too ($2 < 3$). Thus, the strategy profile where both players defect constitutes a Nash equilibrium and an ESS. The dilemma consists in that both suspects would be better off when both cooperate. They would go to jail for one instead of two years ($u(C, C) = 1$; $u(D, D) = 2$).

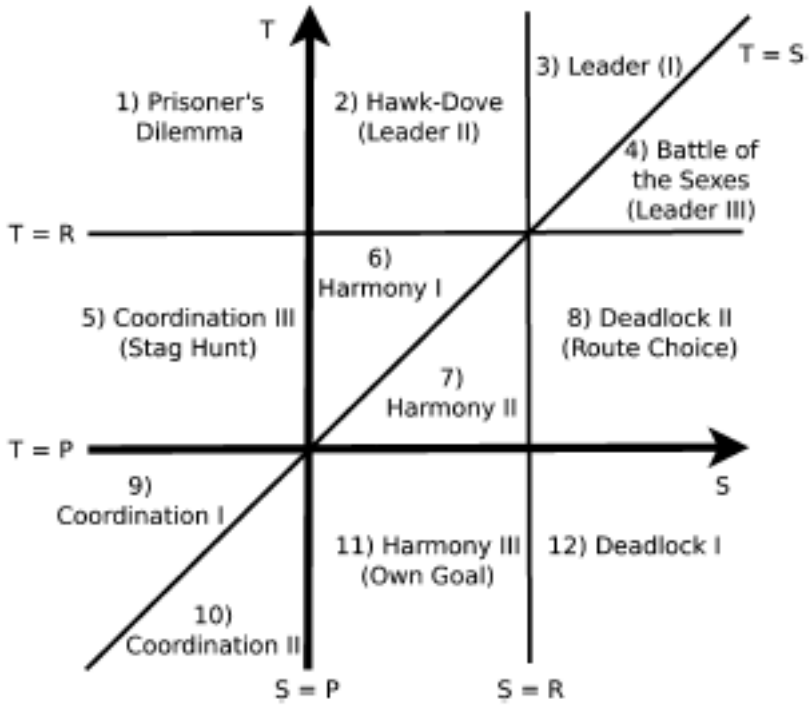


Figure 2.10: Classification of two player games from Hummert et al. (2014).

	C	D
C	1	3
D	0	2

Figure 2.11: The payoffs of the Prisoner's Dilemma.

	Stag	*Hare*
Stag	2	0
Hare	1	1

Figure 2.12: The payoffs of the stag hunt game.

Stag hunt

Two men go to the forest to hunt a stag or a hare, where the latter is less valuable than the first. Both sit on a raised hide waiting for a stag. In the meanwhile, both of them have spotted a (separate) hare. They can hunt a hare or continue waiting for a stag. If a hunter chooses to hunt the hare, he will have a dinner. However, if one of them shoots, the stag would be scared and never show up in front of the hunter's guns; if the other hunter decides not to shoot the hare he will starve in the evening. If they wait and kill a stag later, they share the prey. What should the hunters do? The payoff matrix is depicted in Figure 2.12. The analysis of this game reveals that there are two pure Nash equilibria ($(Stag, Stag)$ and $(Hare, Hare)$), and one mixed Nash equilibrium hunting the stag and the hare with probability of $\frac{1}{2}$.

Dynamic approach

The solutions concept, we saw so far, are static, that is they analyze if a player has an incentive to move away from a particular strategy profile. However, these concepts do not consider how this strategy profile is reached. In this section, we discuss how the strategies that are played in a population evolve over time. Two approaches are described: the replicator and the adaptive dynamics.

Replicator dynamics

Let us assume a population of individuals and a set $S = S_1, \ldots, S_n$ of pure strategies. A $n \times n$ payoff matrix A describes the fitness that an individual gains when it encounters another individual, *e.g.* a_{ij} denotes the payoff of an individual playing S_i when it encounters an individual playing S_j. The relative frequency of each pure strategy $S_i \in S$ at time t is denoted by $x_i(t)$. Thus, $N_i(t) = x_i(t)N(t)$ denotes the number of individuals in the population of $N(t)$ individuals that play strategy S_i at time t. The average fitness of strategy S_i at time t is given by $f_i(t) = \sum_{j=1}^{n} a_{ij}x_j(t)$. This means that the fitness gain of an individual that plays strategy S_i encountering an individual playing strategy S_j depends on the frequency of the latter. An individual with strategy S_i has f_i descendants (also with strategy S_i) in the next generation such that $N_i(t+1) = N_i(t)f_i(t)$ (Broom and Rychtar, 2013). The mean fitness of the population at time t is then given by $\bar{f}(t) = \sum_{i=1}^{n} x_i(t)f_i(t)$ (Hofbauer and Sigmund, 1998).

We can differentiate between discrete and continuous replicator dynamics. In both approaches, we assume asexual reproduction, that is an individual with strategy S_i generates $f_i(t)$ copies of itself in the next generation.

In the discrete replicator dynamics framework, we assume to have discrete non-overlapping

generations. The discrete replicator dynamics is formulated as follows:

$$\begin{aligned} x_i(t+1) &= \frac{N_i(t+1)}{N(t+1)} = \frac{N_i(t) \times f_i(t)}{\sum_{j=1}^{n} N_j(t) \times f_j(t)} \\ &= \frac{x_i(t) \times N(t) \times f_i(t)}{\sum_{j=1}^{n} x_j(t) \times N(t) \times f_i(t)} \\ &= x_i(t) \frac{f_i(t) + \beta}{\bar{f}(t) + \beta} \end{aligned} \quad , \tag{2.7}$$

where β represents a background fitness that is attributed to each individual by default and that is not modeled by the game (Hofbauer and Sigmund, 1998; Broom and Rychtar, 2013). In the continuous replicator dynamics framework, we assume a very large population and overlapping generations. The continuous replicator dynamics is given by (Hofbauer and Sigmund, 1998):

$$\frac{d}{dt} x_i = x_i\big(f_i(t) - \bar{f}(t)\big) \tag{2.8}$$

A stationary (equilibrium) point x can be computed solving (2.8) for $i = 1, \ldots, n$: (i) $x_i(t+1) = x_i(t)$ for the discrete replicator dynamics, or (ii) $\frac{dx_i}{dx} = 0$ for the continuous replicator dynamics (Pelillo, 2009). It holds that every Nash equilibrium of a matrix game is also a stationary point of the replicator dynamics, whereas the opposite does not hold. Furthermore, every evolutionary stable state of a matrix game is an asymptotically stable point of the replicator dynamics (Hofbauer and Sigmund, 1998).

Adaptive dynamics
Replicator dynamics consider the change of frequencies of a fixed set of strategies. On the contrary, adaptive dynamics studies the evolution of a population allowing for rare mutations. It is assumed that the whole population displays the phenotype x of a continuous trait, except a small proportion that plays a slightly different mutant strategy $y = x + h$. If the mutant group can invade x, then the population may evolve towards the fixation of the trait y. It is assumed that selection happens on a faster timescale than mutations. This means that either the resident or the mutant strategy become fixed by natural selection before a new mutation arises (Broom and Rychtar, 2013). The payoff of the mutant strategy y against the resident strategy x is denoted by $A(y, x)$. The relative fitness advantage $A(y, x) - A(x, x)$ is denoted by $W(w, x)$ (Hofbauer and Sigmund, 1998). The adaptive advantage is:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \frac{\partial}{\partial y_i} A(y, x) \tag{2.9}$$

where the derivative is evaluated at $y = x$, for $i = 1, \ldots, n$. This vector, the gradient of $y \to A(y, x)$, points in the direction of the maximal increase of the mutant's fitness advantage (Hofbauer and Sigmund, 1998).

2.2.2 Cooperative Game Theory

The branch of cooperative game theory focuses on *what* a group of players can achieve. It does not matter *how* an individual player acts. The players form coalitions with some binding agreements. Then, a coalition chooses a collective strategy that results in the achievement (utility) of the coalition. There are cooperative games with *transferable utility* and with *non-transferable utility*, where in the latter case, the collective action determines the payoff of each player of the coalition. Games with transferable utility consider only the value that a coalition can achieve. No statement is made about how the value is distributed among the members in the coalition. A cooperative game with transferable utility is formally defined as:

Definition 4. *(Osborne and Rubinstein, 1994) A cooperative game with transferable utility is defined by a pair (N, v), where N denotes a (finite) set of players $N = \{1, \ldots, n\}$. The* characteristic function *v assigns a value to every subset $S \subseteq N$, that is $v : 2^N \to \mathbb{R}$. A subset $S \subseteq N$ is called* coalition.

The players are considered to be rational and they want to maximize their utility. Similar to non-cooperative game theory, a solution concept should not offer an incentive for a player to leave a coalition for another. Here, we describe two solution concepts: the *Core* and the *Shapley value*. Before that, we provide several definitions. An *allocation* $x \in \mathbb{R}^N$ is a division of the value achieved by the coalition S, that is $v(S)$, to its members; player i receives the value x_i.

Definition 5. *An allocation x is* **individually rational** *if every player i receives as least as much as if he would form a coalition that contains only himself, that is $x_i \geq v(\{i\})$.*

Definition 6. *An allocation x is* **efficient** *if the whole value of the coalition is distributed to its members, that is $\sum_{i=1}^{n} x_i = v(N)$.*

Osborne and Rubinstein (1994) define the *marginal contribution* of a player i to a coalition $S \subset N$, with $i \notin S$, as:

$$MC_i(S) = v(S \cup \{i\}) - v(S). \tag{2.10}$$

With these definitions at hand, we say that an *individually rational* and *efficient* allocation x satisfies the *Marginal-Contribution Principle* if the following holds for every player i: $x_i \leq MC_i(S)$ (Brandenburger, 2007).

The Core

We further denote by $x(S)$ the value $\sum_{i \in S} x_i$, that is the sum of the values allocated to the players in a coalition S. This allows us to define the core of a cooperative game as:

Definition 7. *(Brandenburger, 2007) An allocation $x \in \mathbb{R}^N$ is part of the* **core** *of the game (N, v) if x is efficient and for every coalition $S \subseteq N$ it holds that $x(S) \geq v(S)$.*

It can further be stated that an allocation that is in the core of a game is also individual rational and satisfies the Marginal-Contribution Principle (Brandenburger, 2007). Thus, no coalition has the incentive to deviate from an allocation with the core property because it is impossible to make all players of the coalition better off (Osborne and Rubinstein, 1994).
Let us examine two examples. In the first example from Brandenburger (2007), there are three players. Player 1 is a seller who possesses one unit of a good that he values at €4. The other two players are buyers, where player's 2 limit to buy the good is at €9, and player 3 could afford €11. The characteristic function v is described as follows:

$$\begin{aligned} v(\{1,2\}) &= €9 - €4 = €5 \\ v(\{1,3\}) &= €11 - €4 = €7 \\ v(\{2,3\}) &= €0 \\ v(\{1\}) = v(\{2\}) &= v(\{3\}) = €0 \\ v(\{1,2,3\}) &= €7 \end{aligned}$$

If player 1 and 2 build a coalition, the good is sold for €9 and the total gain of this transaction is €5. The same reasoning can be applied straightforwardly to the other coalitions. Attention

must be paid only to the value of the grand coalition. As the seller possesses only one unit of the good, he sells it at the highest price he can get and thus the gain is €7. For an allocation x to be in the core of the game, it must satisfy $x_1 + x_2 + x_3 = 7$ (efficiency), and be such that $x(S) \geq v(S)$ for every $S \subseteq N$. Thus, it must hold that $x_1 + x_2 \geq 5$, $x_1 + x_3 \geq 7$, and $x_i \geq 0$. Therefore, the core consists of the allocations x with $(x_1, x_2, x_3) = (a, 0, 7 - a)$, where $5 \leq a \leq 7$. The second player receives nothing as his marginal contribution to the grand coalition is zero. The core, as in this case, may not consist of a unique allocation. Furthermore, the core can be empty as we will now see with the second example.
In the second example from Osborne and Rubinstein (1994), a group of n people discovered a treasure of gold bars. It needs two persons to carry one piece. This situation can be represented by a cooperative game (N, v), with

$$v(S) = \begin{cases} |S|/2 & \text{if |S| is even} \\ (|S| - 1)/2 & \text{if |S| is odd.} \end{cases}$$

If $|N| \geq 2$ is even, then the core consists in one allocation that distributes one half of a gold bar to everybody. However, the core is empty if $|N| \geq 1$ and odd.

The Shapley value

In this approach, a *unique* payoff vector, the *value*, is assigned to a game. The ith entry of the vector denotes the value or the power of the ith player. Contrary to the previous solution concept, it is guaranteed to assign a unique payoff vector to every game.
The Shapley value can be characterized through axioms as follows:

Definition 8. *(Osborne and Rubinstein, 1994) A value ψ assigns to the characteristic function v an n-tuple $\psi(v) = (\psi_1(v), \ldots, \psi_N(v))$ with $\psi(v) \in \mathbb{R}^N$. Here $\psi_i(v)$ represents the worth (or value) of player i in the game with the characteristic function v. The following **Shapley axioms** must hold for $\psi(v)$:*

1. ***Efficiency**: $\sum_{i \in N} \psi_i(v) = v(N)$.*
2. ***Symmetry**: If $v(S \cup \{i\} = v(S \cup \{j\})$ for every coalition S that does not contain i and j, then $\psi_i(v) = \psi_j(v)$.*
3. ***Dummy Axiom**: If $v(S) = v(S \cup \{i\})$ for every coalition S not containing i, then $\psi_i(v) = 0$.*
4. ***Additivity**: If u and v are characteristic functions, then $\psi(u + v) = \psi(u) + \psi(v)$.*

Alternatively, we can define the Shapley value as:

$$\psi_i(v) = \frac{1}{|N|!} \sum_{r \in R} v(S_i(r) \cup i) - v(S_i(r)), \ \forall i \in N, \tag{2.11}$$

where R is the set of all $|N|!$ permutations of N, and $S_i(r)$ denotes the coalition that contains all players that precedes player i in the permutation $r \in R$. The Shapley value for player i corresponds to the marginal contribution of player i to the coalitions that precede i over all permutations (Osborne and Rubinstein, 1994).
Consider the following game with three players and the characteristic function v, with $v(1, 2, 3) = v(1, 2) = v(1, 3) = 1$, and $v(S) = 0$ otherwise. This game is similar to the first one above. Here, we have one seller (player 1) who sells one unit of a good that he does not value. Furthermore, there are two buyers (players 2 and 3) who are willing to pay €1 each. There are

six permutations over the three players: $(1,2,3)$, $(1,3,2)$, $(2,1,3)$, $(2,3,1)$, $(3,1,2)$, $(3,2,1)$. The marginal contribution of player 1 is 1 in the four permutations where he is at the second or third position. Player 2 has a marginal contribution of 1 in the permutation $(1,2,3)$, and player 3 has a marginal contribution of 1 in the permutation $(1,3,2)$. Thus the Shapley value of this game is $(4/6, 1/6, 1/6)$ which is different from the core that consists in the single allocation $(1,0,0)$.

We refer to Osborne and Rubinstein (1994) for the description of several other solution concepts for cooperative games, *e.g.* the stable set, the Bargaining set, Kernel, and Nucleolus.

2.3 Economic Models

Mutualistic relationships where nutrients or services are exchanged between species resembles a market in economics. It is probably more similar to the original form of trade, called barter, where goods and services were exchanged directly without using an exchange medium such as money. In this section, we describe two selected market models rather than giving a full overview of economic models. In economics, goods and services are valued somehow. There are two different schools of how to assign a value to a good: the labor theory and the subjective theory of value. The first one values a good or a service by the total amount of labor required to produce it. The second assigns a higher value to a good or a service if it is more important to the seller or buyer. The difference is nicely depicted by the example of Böhm-Bawerk (von Böhm-Bawerk, 1891). Herein, a farmer has five sacks of grain which he uses in different manners. With the grains of the first sack, he makes bread to survive. With the second, he makes more bread to be strong for work. The third sack of grains will be used to feed the farm animals. Whisky is made out of the grains of the fourth sack. To have fun, he gives the grains of the last sack to some parrots. What would the farmer do if one sack of grain were lost? He probably would still continue to make bread to survive and to be strong for work. To be able to do the work in the farm, he needs the help of the farm animals which must be fed for this reason. The least important thing to do with the grains is to give them to the parrots. So, he will stop this activity. The different sacks of grain have a different importance, and thus different values, to the farmer. In contrast, it needs the same amount of labor to produce each sack of grain.

2.3.1 Comparative advantage

David Ricardo, adherent of the labor theory of value, developed the concept of *comparative advantage* (Ricardo, 1817) which we explain with the help of an example. Assume two countries, Portugal and England, that produce two goods, cloth and wine, of identical quality. We consider the following amounts of hours of labor for the production of one unit of each good: England needs 100 hours of labor to produce one unit of cloth, and 120 hours of labor to produce one unit of wine. Portugal needs 90 hours of labor to produce one unit of cloth, and 80 hours of labor to produce one unit of wine. Thus, Portugal is more efficient in producing both goods as it needs less labor to produce a unit of them. We say that Portugal has an *absolute* advantage in producing both goods due to a lower amount of labor. However, England has a *comparative* advantage in producing cloth due to a lower opportunity cost of cloth. The opportunity cost of the production of a good A is defined as the amount of good B that must be sacrificed in order to produce another unit of good A. In the present example, England needs 100 hours of labor for the production of one unit of cloth. With the same amount of labor it can produce $\frac{5}{6}$ units of wine. On the other hand, Portugal can produce one unit of cloth or $\frac{9}{8}$ units of wine with 90 hours of labor. Thus, the opportunity cost of

cloth production is smaller in England than in Portugal ($\frac{5}{6} < \frac{9}{8}$). Portugal has a comparative advantage of wine production for the same reasoning.
If we assume autarky, it takes 220 hours of labor for England to produce one unit of both goods. Portugal requires 170 hours of labor to produce the same quantities. If both countries would specialize in the production of the good in which they have a comparative advantage (England produces only cloth, Portugal produces only wine), then the overall production of each good increases. England can produce 2.2 units of cloth ($1 + \frac{6}{5}$), and Portugal produces 2.125 units of wine ($1 + \frac{9}{8}$). Assuming free trade, the countries can exchange the goods such that each country ends up with one unit of each good. England (Portugal) would still have 0.2 (0.125) units of cloth (wine) left that they can consume or trade.
Before the concept of comparative advantage, it was thought that trade is only worthwhile if both countries have an absolute advantage in the production of one good.

2.3.2 General equilibrium theory

The French mathematical economist Marie-Esprit-Léon Walras, adherent of the subjective theory of value, developed the *general equilibrium theory* which aims to determine the prices of many goods considering the supply and demand in several markets (Varian, 2009). To simplify the exposition, which is entirely based on the book of Varian (2009), we will consider two persons (A, and B) and two markets (two goods 1, and 2). Person A consumes both goods in certain amounts, expressed by the *consumption bundle* $X_A = (x_A^1, x_A^2)$, where x_A^1 (x_A^2) denotes A's consumption of good 1 (2). Similarly, the consumption bundle of B is given by $X_B = (x_B^1, x_B^2)$. A pair of consumption bundles (X_A, X_B) is called an *allocation*. If the overall consumption and production of each good is equal then the allocation is feasible:

$$x_A^1 + x_B^1 = \omega_A^1 + \omega_B^1$$
$$x_A^2 + x_B^2 = \omega_A^2 + \omega_B^2,$$

where $\omega_{A,B}^{1,2}$ corresponds to the initial *endowment* of A or B of good 1 or 2. We can conceive the situation as where A and B come to a market place with some amounts of good 1 and 2 (the endowment) which are traded to end up with a *final allocation*. This can be represented with the so-called Edgeworth box (see Figure 2.13) which should be read as follows. The total number of units of good 1 (2) in the economy corresponds to the width (height) of the box. Each point in the box represents an allocation. Person A's origin is in the lower left corner. Person's B origin is in the above right corner. Thus, the distance on the horizontal (vertical) line from the respective origin depicts the quantity of good 1 (resp. 2) that a person holds. The indifference curves of person A (B) are depicted as blue (black) lines. A person is just as satisfied with all consumption bundles that are on the same line, that is the person is indifferent in choosing one or the other.
What happens when persons A and B come with an endowment of goods 1 and 2 (point W in Figure 2.13) to the market? Let us consider the A' and B' indifference curves that pass through W. Person A would prefer to obtain a consumption bundle that is above his indifference curve passing through W. The consumption bundles where B is better off than at the endowment are also above his indifference curve (from the point of view of B's origin). Thus, both persons will agree on a trade that ends up somewhere, *e.g.* at the point M, in the lens-shaped region in Figure 2.13. Both persons are better off at point M than at the endowment. Person A exchanges $|x_A^1 - \omega_A^1|$ units of good 1 against $|x_A^2 - \omega_A^2|$ units of good 2. Similarly, we can read from the Edgeworth box that person B acquires $|x_B^1 - \omega_B^1|$ units of good 1 and gives up $|x_B^2 - \omega_B^2|$ units of good 2.
At such a point M, we can continue the analysis by drawing the indifference curves of A and B that pass through M. If there is no intersection between both indifference curves, as depicted

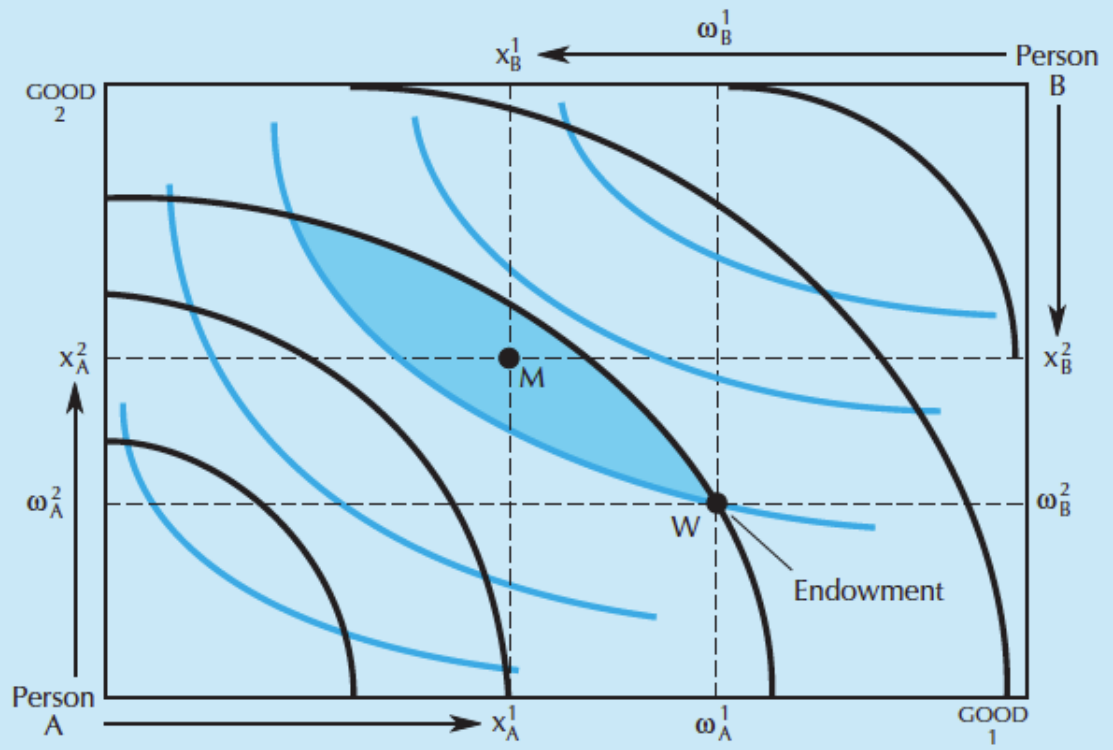


Figure 2.13: (From Varian (2009)) An Edgeworth box where the width (height) measures the total amount of good 1 (2) in the economy. The indifference curves of person A (B) are drawn in blue (black). We refer to the text for further explanations.

by the point M in Figure 2.14, then there exists no allocation that is advantageous for both persons. Here, the region in which person A would be better off than at point M (blue region) is disjoint from the region where person B would be better off (grey region). Thus, there will be no mutual agreement for a trade. Such an allocation is called *Pareto efficient* which means that there is no way to make one person better off without making another person worse off. At a Pareto efficient allocation, the indifference curves of both persons are tangent. Thus, by identifying the points where the indifference curves of both persons are tangent yields the full set of Pareto efficient points, which is called the *contract curve* (see Figure 2.14).

If we are given an endowment W, then all Pareto efficient allocations that are inside the lens-shaped region formed by the indifference curves passing through W, are possible outcomes of mutual beneficial trade. However, there still may be several Pareto efficient allocations that are within this region.

For the remaining, we consider that there are not two persons A and B, but rather two types of consumers A and B. The Edgeworth box can thus be read as the average demands of the consumer types. Furthermore, a person that is of consumer type A is called an *agent* A.

Now, assume that an auctioneer chooses a price p_1 for good 1, and a price p_2 for good 2. The agents A and B come to the market with an amount of both goods (the endowment). Both know the price of the goods and thus they can calculate the worth of their endowments, that is their budget. Both decide how much of each good they would like to buy or sell at the given prices. The total amount of a good that an agent wants to acquire is called the *gross demand*. An agent's *net* or *excess demand* for a good is the difference between the gross demand and the amount of units of this good that he possesses when he comes to the market

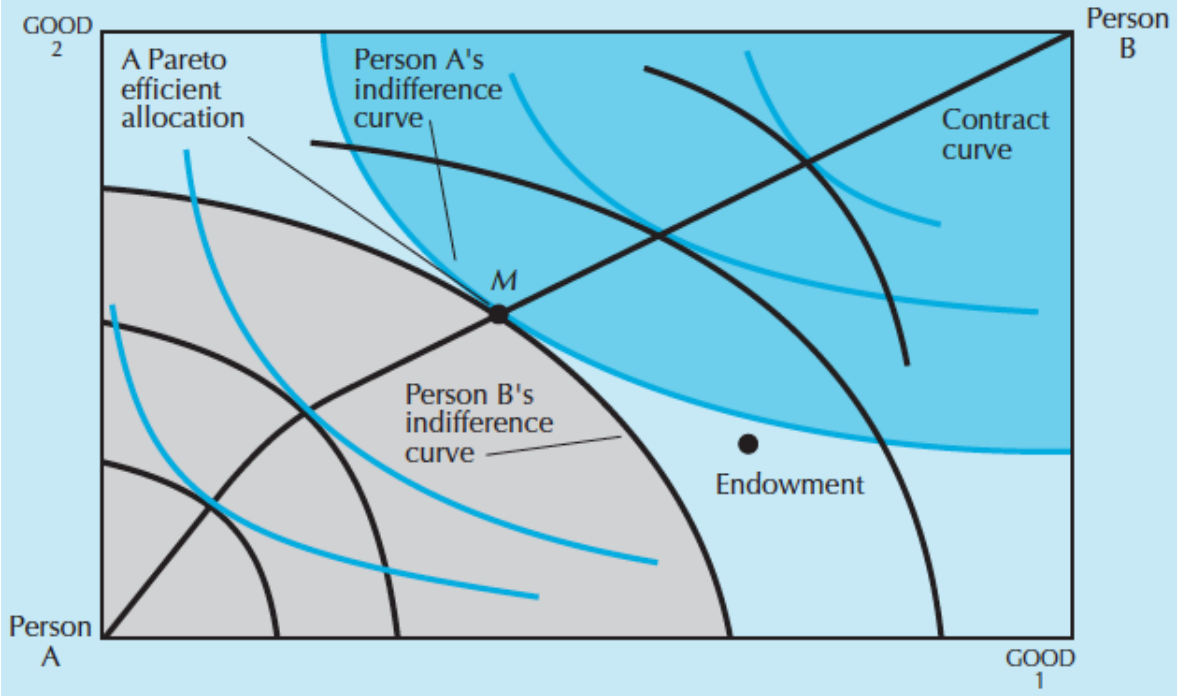


Figure 2.14: (From Varian (2009)) A particular Pareto efficient allocation M on the contract curve which is the set of all Pareto efficient allocations.

(endowment). Thus, the excess demand for good 1 of agent A is:

$$e_A^1 = x_A^1 - \omega_A^1.$$

These terms are depicted in the Edgeworth box in Figure 2.15. The black line in Figure 2.15 corresponds to the budget line for the given prices (p_1, p_2). Two demand bundles are depicted (point (x_A^1, x_A^2) for agent A, and point (x_B^1, x_B^2) for agent B). The market described in Figure 2.15 is in disequilibrium because the total demand of good 2 (1) is greater (smaller) than its supply. We should mention that it is in general assumed that a market clears. A market clears when the supply of a good equals the demand of the good, this means that all produced goods are consumed. If the demand is higher than the supply for a good then the auctioneer would raise the price of this good. On the contrary, if the supply is higher than the demand, then the price would be decreased. These price adjustments are done until the demand equals the supply as depicted in Figure 2.16. Here, the amount of good 1 that A wants to sell equals the amount that B wants to buy. The same holds in the opposite sense for good 2. Thus, each agent chooses its preferred bundle for the given prices and demand equals supply. We say that the market is in *Walrasian equilibrium*. Walras proved that if there are markets for k goods, then it is sufficient to find a set of prices where $k-1$ of the markets are in equilibrium. The market for good k is then automatically in equilibrium (demand equals supply).

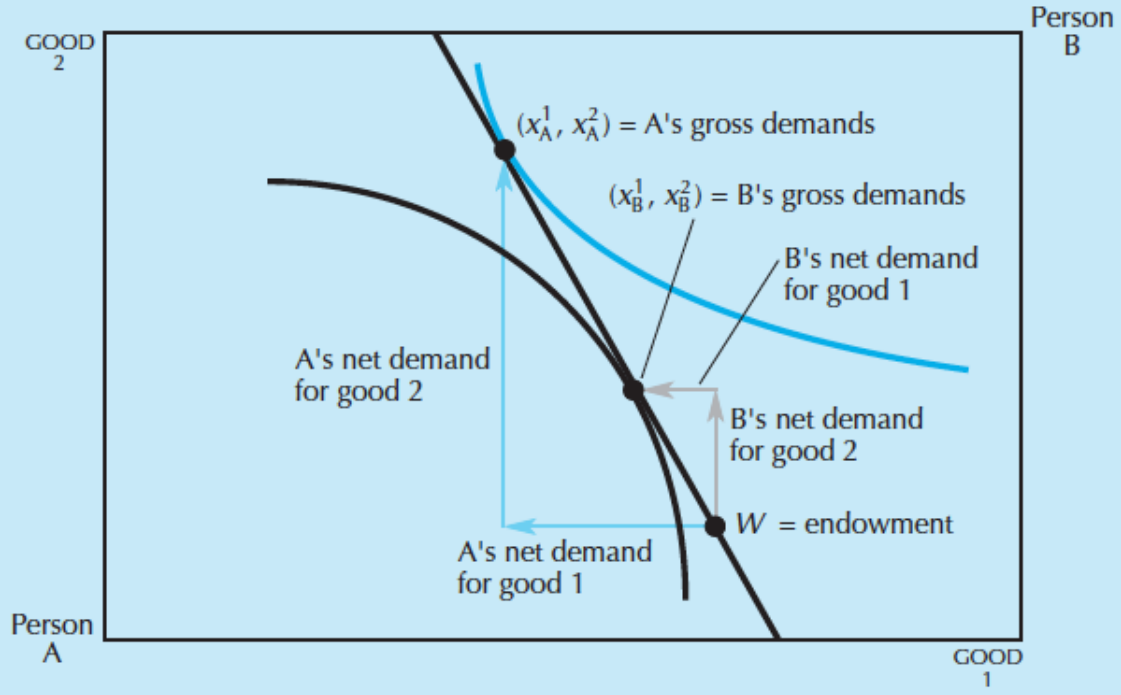


Figure 2.15: (From Varian (2009)) Gross and net demands.

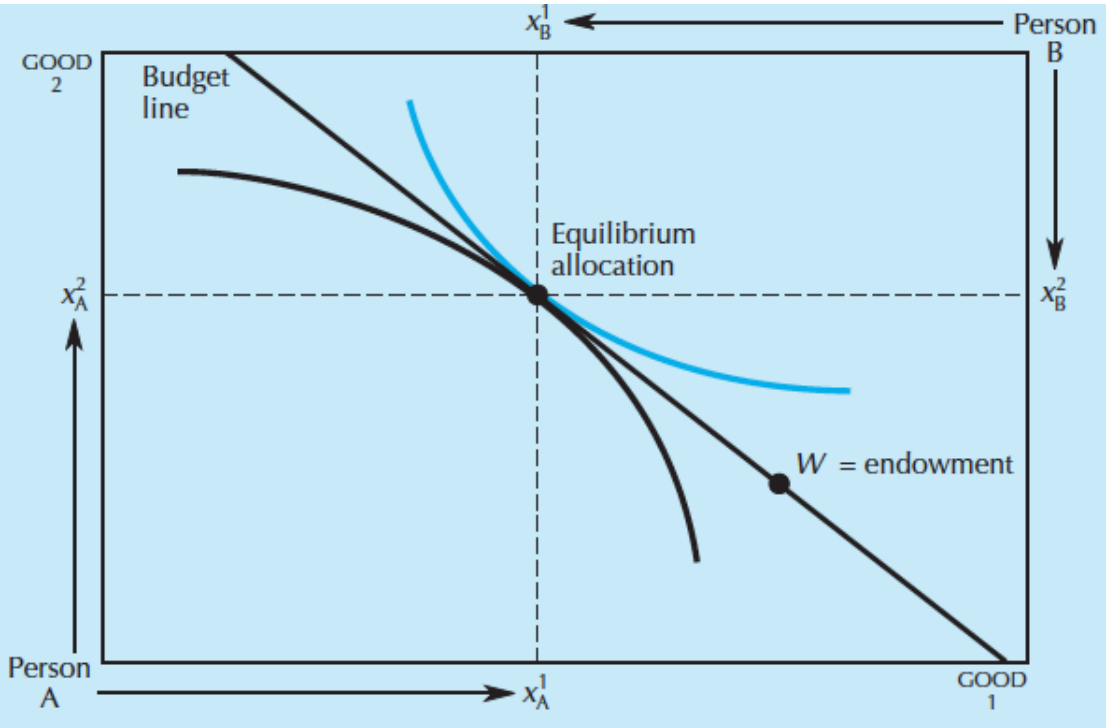


Figure 2.16: (From Varian (2009)) Equilibrium in the Edgeworth box.

Chapter 3

Minimal Precursor Sets

Contents

3.1 Introduction

This chapter is mostly based on the (to Algorithms for Molecular Biology submitted) publication Andrade et al. (2016). Both, the concept of minimal precursor sets and techniques used to enumerate them, are crucial to attack the problem of species interaction which will be covered in chapter 5. Here, we concentrate on the enumeration of minimal precursor sets in a metabolic networks of a single species. The question of which metabolites an organism needs from its environment (henceforth called the *sources*) in order to grow or to produce a given set of metabolites (henceforth called the *targets*) is crucial for both fundamental and applied reasons. This indeed enables to define the growth conditions of organisms in the laboratory, as well as the minimal media necessary for the production of compounds of biotechnological interest (for instance, ethanol). More recently, great interest in establishing which nutrients are exchanged among different organisms in communities such as present in the human gut has also been raised by the interest to develop new strategies for fighting infection that rely on the use of probiotics instead of antibiotics (Lin et al., 2014). However the latter requires that: (1) such exchanges are computed in a very efficient way in genome-scale metabolic networks; (2) all possible minimal sets of sources are identified for a given target set of interest in order

to fully understand the interactions that may take place among the organisms in a community, as well as the alternative niches that may with time develop for some such organisms.

Early attempts at enumerating all *minimal precursor sets* (minimal sets of sources) were based only on topology (henceforth called *topological precursor sets*). Stoichiometry was thus not taken into account, leading to possibly many unfeasible solutions (Romero and Karp, 2001; Handorf et al., 2008; Cottret et al., 2007; Acuña et al., 2012). The algorithm of Romero & Karp was based on a backtrack traversing of the metabolic graph from the target compounds to the seeds while Handorf *et al.* tested the reachability of the target from a heuristically defined collection of sets of sources. Neither enumerated all minimal precursor sets. Cycles, although omnipresent in metabolic networks (*e.g.* Krebs cycle), were not included until the method of Cottret *et al.* (Cottret et al., 2007). However, the latter algorithm could be applied only to small networks due to a high memory requirement; subsequently, Acuña *et al.* (Acuña et al., 2012) allowed the enumeration of all minimal precursor sets of networks of about 1000 reactions. The authors also pointed out that the enumeration of precursor sets and of precursor cut sets could be done simultaneously in quasi-polynomial total time. Precursor cut sets are a set of sources such that, if they are eliminated, then the target set of interest can no longer be produced by any combination of the remaining sources.

The approach of Zarecki *et al.* (Zarecki et al., 2014) takes stoichiometry into account and consists of two steps. First, the size of a set of sources of minimal cardinality that allows the production of a target is determined solving a mixed integer linear programming problem. In a second step, the authors identify a single set of sources of the determined size such that the sum of the molecular weight of the compounds is minimal.

To our knowledge, there are two algorithms that attempt to enumerate all minimal precursor sets with stoichiometry (henceforth called *stoichiometric precursor sets*) (Imieliński et al., 2006; Eker et al., 2013).

Imieliński *et al.* (Imieliński et al., 2006) propose a method that first enumerates all extreme semipositive conservation relations (ESCR), that is the extreme rays of the cone defined by the transposed stoichiometric matrix. The precursor sets are then obtained by the enumeration of hitting sets of the ESCRs. As the authors state, this approach is impractical for genome-scale metabolic networks since it is impossible to enumerate all ESCRs with the current algorithms (Imieliński et al., 2006). Consequently, a method is proposed that enumerates a subset of the ESCRs (those that do not contain water) to obtain (via hitting sets) minimal precursor sets that contain water. These solutions are physiologically minimal (all media contains water), but not necessarily the theoretically minimal.

The method of Eker *et al.* (Eker et al., 2013) is based on logical and linear constraint solving and on computational boolean algebra. The authors formulated two different constraint models, that were called *steady-state* and *machinery-duplicating*. Their steady-state model requires a non-negative net production of all compounds that are on the path from the precursors to the target. Observe that the term steady-state is usually used to denote a slightly different model where all compounds that are on the path from the precursors to the target cannot accumulate.

Their machinery-duplicating model is more restrictive as it requires a strict positive net production of these compounds. Notice that a set of sources that allows the production of the target(s) in the machinery-duplicating model, allows also the production of the same target(s) in the steady-state model. A toy example illustrates the difference between the two models (see Figure 3.1). In this network, we have a source (p), a target (t), internal compounds (a, b, c), and three reactions ($r_1 : p + a \to c, r_2 : c \to b, r_3 : b \to a + t$). Following their steady-state model, the source p is a precursor of t. The compounds a, b, and c have a zero net production when we assign a positive flux value 1 to each reaction. In the machinery-duplicating model, there is no precursor set that enables to produce t: indeed, no flux would

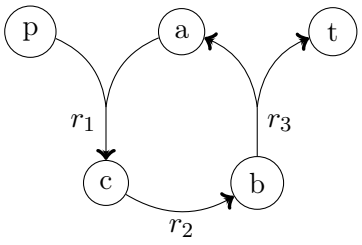


Figure 3.1: Network with one source p and one target t illustrating the difference between the two models used by Eker *et al.* (Eker et al., 2013), and the limitation of the machinery-duplicating model.
The source p is a precursor set for the production of the target if the steady-state model is assumed. In this toy example, the target can not be produced following the machinery-duplicating model.

fulfil the condition of a strict positive net production of a, b, and c. However, this type of cycle resembles the Krebs cycle that plays an essential role in the production of energy in aerobic organisms. To reveal the similarity between the toy example and the Krebs cycle, let the compound a take the role of oxaloacetate (which is regenerated through the Krebs cycle), the source p feed the cycle as acetyl-CoA, the compound b be any compound on the Krebs cycle such as *e.g.* citrate or succinate, and the target t any by-product of the Krebs cycle such as NADH or carbon dioxide. We argue that the machinery-duplicating model is too restrictive as therein cycles of the type shown in Figure 3.1 are not captured.
An approach widely used in flux-balance analysis (FBA) (Watson, 1984) to model an organism's growth condition is to include a so-called biomass reaction, that consumes in the right amounts every compound needed for a cell to duplicate. Such a reaction has a single product, an artificial compound (representing the duplicated cell) that can be modelled as a target compound in our enumeration approach: a cell can grow and duplicate if it can produce this target compound.
The objective of this chapter is twofold. On the theoretical level, we show the relationship between topological and stoichiometric precursor sets and we discuss some complexity results. On the methodological level, we provide two algorithms that enumerate all minimal stoichiometric precursor sets. The first one is based on the above mentioned relationship between topological and stoichiometric precursor sets. Although interesting in terms of theory, it however is not efficient in practice. The second approach, called Sasita, uses a similar approach as in Lee et al. (2000); de Figueiredo et al. (2009); von Kamp and Klamt (2014). Therein the authors enumerate reaction subsets solving recursively mixed integer linear programming (MILP) problems. The reaction subsets correspond to alternate flux distributions to obtain an identical value of the objective function (Lee et al., 2000), the k-shortest elementary flux modes (EFM) (de Figueiredo et al., 2009) or the smallest Minimal Cut Sets (MCS) (von Kamp and Klamt, 2014). All approaches enumerate minimal reaction sets. Here we consider minimality of the set of compounds.
Sasita enables to enumerate all stoichiometrically feasible minimal precursor sets in both models, *steady-state* and *machinery-duplicating*, as in Eker et al. (2013). A natural question is to compare our results with those obtained in Eker et al. (2013). Unfortunately, the algorithm of Eker *et al.* (Eker et al., 2013) was not publicly available for testing. We nevertheless tried to reproduce the authors' results by incorporating their definitions in our method, but we were unable to obtain the same results, even using the same *Escherichia coli* network they made available. More surprisingly, we found a precursor set that is minimal with respect to

the solutions found by Eker *et al.*
In the enumeration of minimal precursor sets for a given target set, Eker *et al.* (Eker et al., 2013) are able to work with genome-scale metabolic networks and are exhaustive, but their method is very time and memory consuming. The authors indeed indicate that it required 3 days of execution on a 24-core (with Hyper threading) 2.67 GHz Intel X5650 Xeon CPU-model processor, using the machinery-duplicating model on an *Escherichia coli* network composed of 2314 unidirectional reactions of which 388 were transport reactions, to enumerate 787 solutions.
We show that we can apply SASITA on big networks like the iJO1366 reconstruction of the *Escherichia coli* K-12 MG1655 with 3646 reactions and 2258 compounds.
In the first Section, we provide basic definitions; in particular, we extend ideas from topological precursor sets as defined in Acuña et al. (2012) in order to incorporate stoichiometry, and we discuss the relationship between topological and stoichiometric solutions. We then describe the SASITA algorithm for enumerating all stoichiometrically feasible minimal precursor sets. In the Result Section, we discuss in detail the comparison with respect to Eker's *et al.* proposal. Here we observe that SASITA is the first publicly available software to enumerate minimal stoichiometric precursors sets both with the steady state and the machinery duplicating model. Experiments show that SASITA can be applied to large genome-scale metabolic networks and we discuss the obtained results.

3.2 Definitions and Properties

A metabolic network is composed of a set of compounds together with the reactions that transform them. The following example represents a metabolic network with four compounds and two reactions (the values before each compound in a reaction are the *stoichiometric coefficients* of the reaction):

$$r_1 : 1.0\, c_1 + 2.0\, c_2 \to 1.0\, c_3,$$
$$r_2 : 3.0\, c_3 \to 1.0\, c_4.$$

In the following, we use directed hypergraphs (Gallo et al., 1993; Ausiello et al., 2001) to model a metabolic network. A metabolic network is characterised by a pair $\mathcal{N} = (\mathcal{C}, \mathcal{R})$, where $\mathcal{C}$ is the set of vertices (representing metabolic compounds) and $\mathcal{R}$ is a set of hyperarcs (representing metabolic reactions). All reactions are considered to be irreversible. Reversible reactions are thus split into a forward and a backward reaction. A stoichiometric matrix S associated to $\mathcal{N}$ is a matrix containing the stoichiometric coefficients of each reaction with the reactions in $\mathcal{R}$ as its columns and the compounds in $\mathcal{C}$ as its rows. We define $\mathcal{X} \subseteq \mathcal{C}$ as the set of *source compounds* and $\mathcal{T} \subseteq \mathcal{C}$ as the set of *target compounds*. For simplicity, we assume that sources are not produced by any reaction; it is easy to verify such condition by adding for each metabolite x in $\mathcal{X}$ a dummy metabolite x' and a dummy reaction producing one x from one x'. Replacing x by their representative x' in the set $\mathcal{X}$ produces an equivalent network (in terms of what a set of sources is able to produce) with the desired property (see Acuña et al. (2012)).
Topologically, a reaction $r \in \mathcal{R}$ is defined by its substrates $Subs(r)$ and its products $Prod(r)$, suggesting the interpretation of a reaction as an hyperarc with $Subs(r)$ as the set of tail nodes and $Prod(r)$ as the set of head nodes of a reaction r. In the above example, $Subs(r_1) = \{c_1, c_2\}$ and $Prod(r_1) = \{c_3\}$. The network $\mathcal{N}$ can then be seen as a directed hypergraph with stoichiometric coefficients associated with each hyperarc. Given a subset of reactions $F \subseteq \mathcal{R}$,

we denote by $Subs(F)$ and $Prod(F)$ the union of the substrates and products, respectively, of the reactions in F.
Given a network $\mathcal{N}$ and the sets of source $\mathcal{X}$ and target $\mathcal{T}$ compounds, we loosely define a precursor set as a set $X \subseteq \mathcal{X}$ that can *produce* all the targets of $\mathcal{T}$ using a subset of the reactions in $\mathcal{R}$. The concept of a *factory* was introduced in Acuña et al. (2012); a *topological factory* is defined as follows:

Definition 9. *A set $F \subseteq \mathcal{R}$ is a **topological factory** from $X \subseteq \mathcal{X}$ to $T \subseteq \mathcal{T}$ if $T \cup Subs(F) \subseteq Prod(F) \cup X$; i.e., if T and every substrate of every reaction in F is either a source or is produced by some reaction in F.*
*A set $X \subseteq \mathcal{X}$ is a **topological precursor set** (TPS) for $\mathcal{T}$ if there exists a topological factory from X to $\mathcal{T}$.*

Extending this definition to include stoichiometry requires that any substrate of a reaction in F, should be either produced at least in a same quantity by one or more reactions also in F, or the substrate should be a source compound. Observe that, if the flux vector $v \in \mathbb{R}^{|\mathcal{R}|}$ denotes the flux of every reaction in the network per time unit, then $Sv \in \mathbb{R}^{|\mathcal{C}|}$ is the vector of net production of all compounds in the network for the flux v. Furthermore, $(Sv)_A$ specifies the net production of the compounds in a set A.

Definition 10. *A **stoichiometric factory** (S-factory) from $X \subseteq \mathcal{X}$ to $T \subseteq \mathcal{T}$ is a set $F \subseteq \mathcal{R}$, such that there exists a flux vector $v \geq 0$ satisfying:*

1. $v_i \begin{cases} > 0 & i \in F \\ = 0 & otherwise, \end{cases}$
2. $(Sv)_{\mathcal{C} \setminus X} \geq 0$,
3. $(Sv)_T > 0$.

*A S-factory from X to T is **minimal** if it does not contain any other S-factory from X to T.*

It is possible to adapt Definition 10 to the steady-state assumption. In elementary mode analysis (Schuster and Hilgetag, 1994; Schuster et al., 2002a; Gagneur and Klamt, 2004; Wagner and Urbanczik, 2005), the set of compounds is split into *internal* and *external* compounds, whereat the steady state constraint is applied only to the former. In the context of precursor sets, for every external compound e, we add an export reaction $r_e : e \to \emptyset$. We replace furthermore the greater-than-or-equal sign by the equal sign in the second constraint of Definition 10. Thus, the steady-state constraint is put on the compounds in $\mathcal{C}$ that are neither in X nor in $\mathcal{T}$ (the latter through the third constraint). Note that the steady state constraint is thus put on the (remaining) external compounds which is fine because of the export reactions.

For the remaining, a stoichiometric factory (S-factory) and precursor set (SPS) refer to the case where accumulation is allowed if not specified otherwise.

Definition 11. *A set $X \subseteq \mathcal{X}$ is a **stoichiometric precursor set** (SPS) **of** T if there exists a S-factory from X to T. A SPS of T is **minimal** if it does not contain any other SPS of T.*

The following **Facts** summarise the main differences between TPSs and SPSs:

1. *Every S-factory is a topological factory. Every SPS is also a TPS.*
2. *Not every topological factory is a S-factory. Not every TPS is a SPS.*

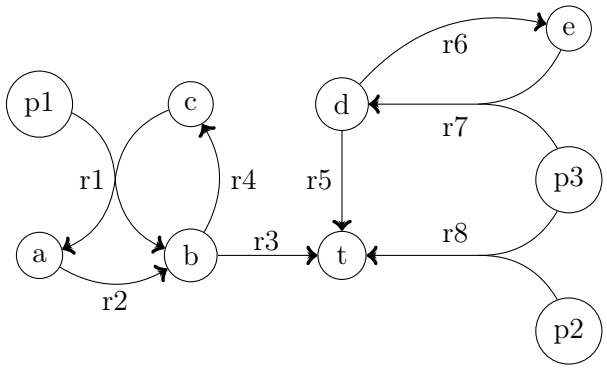


Figure 3.2: Illustration of Facts 2-4. The stoichiometric values are all equal to one. There are two minimal TPSs: $\{p1\}$ (obtained from the topological factory $\{r_1,\ r_3,\ r_4\}$), and $\{p3\}$ (obtained from the topological factory $\{r_7,\ r_6,\ r_5\}$). The source $p2$ does not take part of a minimal topological factory because its consumption involves the consumption of the source $p3$, which forms already a minimal TPS. There are two minimal SPSs: $\{p1\}$ (obtained from the stoichiometric factory $\{r_1,\ r_2,\ r_3,\ r_4\}$), and $\{p2, p3\}$ (obtained from the stoichiometric factory $\{r_8\}$).

Given these facts, it is clear that any S-factory always contains a topological factory. A natural question that arises is whether we can decompose an S-factory into a set of topological factories. We show that this is not true:

3. *There exist minimal S-factories which are not the union of minimal topological factories.*
4. *There exist minimal SPSs which do not consist of a union of minimal TPSs.*

The first fact is a direct consequence of the definitions of SPS and TPS. The remaining facts are illustrated using Figure 3.2 that has two minimal TPSs ($\{p1\}$ and $\{p3\}$), and two minimal SPSs ($\{p1\}$ and $\{p2, p3\}$) of the target set $\{t\}$. Observe that $\{p3\}$ is a (minimal) TPS but is not a SPS (fact 2). The minimal stoichiometric factory from $p1$ to t consists in the set of reactions $r1$, $r2$, $r3$, and $r4$, while the minimal topological factory from $p1$ to t does not contain the reaction $r2$ from a to b (fact 3). The minimal SPSs $\{p2, p3\}$ cannot be obtained as combinations of any minimal TPSs (fact 4). Figure 3.2 gives an intuition about the facts; similar characteristics can be found in real metabolic networks as well.
Figure 3.2 shows an example where it is indeed not possible to obtain the minimal S-factory that contains $r1$, $r2$, $r3$, and $r4$ from minimal topological factories. However, it is true that every minimal S-factory is a union of minimal topological factories of the **many-to-one transformed network** defined as follows:

Definition 12. *Given $\mathcal{N} = (\mathcal{C}, \mathcal{R})$, the* **many-to-one transformation** *of $\mathcal{N}$ is the metabolic network $\Psi(\mathcal{N}) = (\mathcal{C}, \Psi(\mathcal{R}))$ such that for each reaction $r \in \mathcal{R}$ and for each metabolite $a \in Prod(r)$, there is a reaction r_a in $\Psi(\mathcal{R})$ such that $Subs(r_a) = Subs(r)$ and $Prod(r_a) = \{a\}$.*

Given a reaction $r \in \mathcal{R}$ with $Prod(r) = \{a_1, \ldots, a_k\}$, we denote by $\Psi(r) = \{r_1, \ldots, r_k\}$ the set of k reactions in $\Psi(\mathcal{R})$ that correspond to the many-to-one transformation of r, that is, $Subs(r_i) = Subs(r)$ and $Prod(r_i) = \{a_i\}$. Furthermore, we extend this definition to sets of reactions, that is, if $R \subseteq \mathcal{R}$, we denote $\Psi(R) = \cup_{r \in R} \Psi(r)$.
It is clear that all minimal topological factories in $\mathcal{N}$ are also among the minimal topological factories in the many-to-one network $\Psi(\mathcal{N})$ if we would retransform the many-to-one reactions into their original hyper-reactions. However, there are additional minimal topological factories

in $\Psi(\mathcal{N})$ that are not minimal in $\mathcal{N}$ after the retransformation. From this, we claim that every minimal S-factory in $\mathcal{N}$ is a union of minimal topological factories of $\Psi(\mathcal{N})$. The following definition and lemmas will provide the basis for the proof of this statement.

Definition 13. *In a hypergraph, we define a (simple) path $p = (M, R)$ from s to t as a chain of different metabolites $M = (m_0, \ldots, m_n)$ and a chain of different reactions $R = (r_1, \ldots, r_n)$ such that:*

1. $m_0 = s$ *and* $m_n = t$;
2. $m_i \in Subs(r_{i+1}),\ i \in \{0, \ldots, n-1\}$;
3. $m_i \in Prod(r_i),\ i \in \{1, \ldots, n\}$.

Lemma 1. *Given a minimal S-factory H from $X \in \mathcal{X}$ to $T \in \mathcal{T}$ in the network $\mathcal{N}$, for every reaction $r \in H$, there is always at least one path in H from one of the products of r to some metabolite in $\mathcal{T}$.*

Proof. Let us suppose without loss of generality that $|T| = 1$. We are going to prove the lemma by contradiction. Suppose that there is a reaction $r \in H$ such that $Prod(r) = \{p_1, \ldots, p_k\}$, and that for all $p_i \in Prod(r)$ there is no path from p_i to T in H. Since there is no path to T in H that includes r, if $r' \in H$ is a reaction that consumes one product of r, then there cannot be a path to T that includes r' (in fact if such a path exists then there is also a path that includes r). By repeating the same reasoning, consider the set of reactions I_r corresponding to the reactions in H that consume the products of r and of the reactions that consume the products of those reactions and so on. Let $\overline{H} = H \setminus I_r$; we will argue that $\overline{H}$ remains a stoichiometric factory.
Consider S the stoichiometric matrix associated to N and v the positive vector associated with H. Removing r from H means $v_r = 0$. Consider $\bar{v}$, a vector with the same values of v except for the components of v corresponding to the reactions of I_r (the fluxes corresponding to the reactions in I_r) which should be zero. Since any of the reactions of I_r lead to T, we have that:

$$S\bar{v} \geq 0, (S\bar{v})_T \geq 1.$$

Since $\overline{v_k} > 0$ for all $k \in \overline{H}$, $\overline{H} \subset H$ is a stoichiometric factory, which is a contradiction because H is minimal. □

Lemma 2. *Given a set of reactions $H \subseteq \Psi(\mathcal{R})$ and the set of sources $X = Subs(H) \cap \mathcal{X}$, then H is a minimal topological factory from X to $\mathcal{T}$ if and only if the two following statements are true:*

1. *For every metabolite m in $Subs(H) \setminus X$ there is exactly one reaction in H that produces m;*
2. *For every metabolite m in $Prod(H)$ there exists a path from m to $t \in \mathcal{T}$ contained in H.*

Proof. We first prove that if H is a minimal topological factory from X to $\mathcal{T}$, then both statements (1. and 2.) above hold. By definition, H is a topological factory from X to $\mathcal{T}$ if and only if any metabolite in $\mathcal{T}$ and in $Subs(H)$ is a source in X or is produced by some reaction in H. Let m be a metabolite in $Subs(H) \setminus X$. Then by definition there is a reaction r in H that produces m. Suppose however that there is another reaction r' in H also producing m. Then $Prod(H \setminus \{r'\}) = Prod(H)$. Thus, $H \setminus \{r'\}$ would still be a topological factory from X to $\mathcal{T}$ which contradicts the minimality and thus proves Statement 1.

Now let m be a metabolite in $Prod(H)$. We show that there is a path from m to some target in $\mathcal{T}$. By contradiction, suppose that there is no such path. Consider then the following iterative process. Starting from $M = \mathcal{T}$ and $R = \emptyset$, consider all reactions H' of H that produce the metabolites in M. Then add to R all reactions in H' and to M all substrates of H', and repeat the process until no reaction is added. Clearly, for all $m \in M$, either $m \in \mathcal{X}$ or m is produced by some reaction in R, and therefore R is a topological factory from X to $\mathcal{T}$. Since all reactions in $\mathcal{R}$ are also in H and H is a minimal topological factory, $R = H$ and m must have been included in M in some iteration. Clearly from that iteration, we can recover a path from m to some metabolite in $\mathcal{T}$ by going backwards in the described process and therefore Statement 2 above holds.
In order to prove the opposite implication, we first observe that, if both statements are true, then by Definition 1, the set H corresponds to a topological factory from X to $\mathcal{T}$. Therefore, we only need to show that it corresponds to a *minimal* topological factory. Let $H' \subseteq H$ be a minimal topological factory from X to $\mathcal{T}$. By contradiction, suppose that $H' \neq H$, then there is a reaction r in $H \setminus H'$. Let a be the product of r. By hypothesis, there is a path from a to $\mathcal{T}$ in H. However each reaction in the path is the only one in H producing the metabolites composing its products. Clearly the last reaction in the path (which produces a target) must also belong to H'. Thus, at some point in the path, there is a metabolite which is the product of a reaction in H which is not in H', and is the substrate of a reaction in H'. There is thus a substrate of a reaction in H' that is not produced by any reaction in H', which is a contradiction with the fact that H is a topological factory from X to $\mathcal{T}$. Therefore, $H' = H$ and the minimality is proved. □

The following theorem shows that any minimal S-factory is the union of minimal topological factories in the many-to-one network.

Theorem 1. *For any minimal S-factory $H \subseteq \mathcal{R}$ from X to T in $\mathcal{N}$, there exists a set of minimal topological factories $F_1, \ldots, F_k$ from X to T in $\Psi(\mathcal{N})$ such that:*

1. *$F_1, \ldots, F_k \subseteq \Psi(H)$;*
2. *For each reaction r in H there is $i \in \{1, \ldots, n\}$ such that $\Psi(r) \cap F_i \neq \emptyset$.*

Proof. From a given minimal S-factory H, we select a reaction $r \in H$. By Lemma 1, we know that there is a path from at least one product m of r to one target compound t. Clearly, there is a path $p = (M_p, R_p)$ from m to t in $\Psi(\mathcal{N})$. Since $\Psi(\mathcal{N})$ is a many-to-one network, every metabolite in M_p is produced by only one reaction in R_p. We show that p can be extended to a topological factory from X to t. Starting from the set $R_0 = R_p$, we consider the set of metabolites $M_0 = Subs(R_0) \setminus Prod(R_0)$, that is, the set of substrates that are not produced by any reaction in the set. Let c be any metabolite in M_0. In the S-factory H in $\mathcal{N}$, there exists a reaction h that produces c. Let h_c be the many-to-one reaction in $\Psi(h_c)$ that produces c, that is, $Prod(h_c) = \{c\}$. We define $R_1 = R_0 \cup \{h_c\}$ as the new set of reactions and $M_1 = Subs(R_1) \setminus Prod(R_1)$. We repeat this process defining R_{i+1} and M_{i+1} by choosing any metabolite in M_i until M_{i+1} is empty. By construction, the set of reactions $F = R_{i+1}$ satisfies the two properties of Lemma 2, and therefore F is a minimal topological factory from X to t contained in $\Psi(H)$. Repeating this process for every reaction $r \in H$, we obtain a set $F_1, \ldots, F_k$ of topological factories from X to t satisfying the desired properties. □

The theorem suggests that a straightforward idea to enumerate all $SPSs$ is to enumerate minimal topological factories in $\Psi(\mathcal{N})$ and then just build combinations thereof checking their stoichiometric feasibility in $\mathcal{N}$. The combinations can be done in the following way. We check all combinations of k minimal topological factories for feasibility, starting with $k = 1$.

Before incrementing k, we test if there is a minimal SPS (with respect to the already obtained $SPSs$) that can be build from at least $k+1$ minimal topological factories. This however is in general not an efficient approach because (i) many topological factories in $\Psi(\mathcal{N})$ are not part of a SPS, (ii) the powerset of all topological factories in $\Psi(\mathcal{N})$ has to be built to obtain $SPSs$. Issue (i) is illustrated in the network of Figure 3.3a. There are n minimal topological factories in $\Psi(\mathcal{N})$. One contains only $\psi(r_1)$. The other minimal topological factories contain each $\{\psi(r_t), \psi(r_a), \psi(r_b)\}$ and one of the reactions in $\{\psi(r_2), \ldots, \psi(r_n)\}$, respectively. The only SPS consists of p_1 and can be obtained directly from the minimal topological factories of $\Psi(\mathcal{N})$. The enumeration of the minimal topological factories that contain one of the reactions in $\{\psi(r_2), \ldots, \psi(r_n)\}$ may be time consuming, for nothing since none of them yields a SPS. Indeed, the number of minimal topological factories in $\Psi(\mathcal{N})$ can be much higher than the number of SPSs in $\mathcal{N}$. Issue (ii) is depicted in Figure 3.3b. There are n minimal topological factories in $\Psi(\mathcal{N})$. Only the combination of all n minimal topological factories yield a SPS in $\mathcal{N}$. However, all other combinations (that can be huge for large values of n) have to be considered and tested for feasibility.

3.3 Complexity

We now discuss the main complexity results for finding and enumerating $SPSs$. The next theorem shows that deciding whether a set is a SPS can be done efficiently.

Theorem 2. *Given a network $\mathcal{N}$, a subset $X \subseteq \mathcal{X}$ of sources and a target set $\mathcal{T}$, we can decide in polynomial time whether X is a SPS for $\mathcal{T}$.*

Proof. We are going to show that it suffices to solve a linear optimisation problem to decide whether X is a SPS for $\mathcal{T}$.
Consider the network $\overline{\mathcal{N}} = (\bar{\mathcal{C}}, \overline{\mathcal{R}})$ with $\bar{\mathcal{C}} = \mathcal{C} \cup \{t\}$ and $\overline{\mathcal{R}} = \mathcal{R} \cup \{\bar{r}\}$. We add a compound t and a reaction $\bar{r}$ to the network $\mathcal{N}$. The reaction $\bar{r}$ is build as follows: $Subs(\bar{r}) = \mathcal{T}$ and $Prod(\bar{r}) = \{t\}$. Also, the values of all stoichiometric coefficients of $\bar{r}$ are one. Consider now the following optimisation problem:

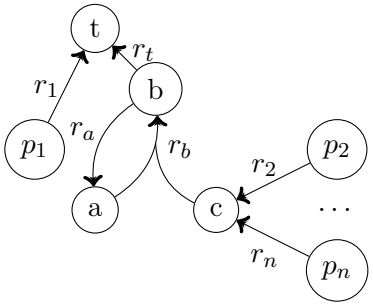
$$\begin{aligned} &\text{Maximize} f(v) = (Sv)_t \\ &s.t. \\ &\qquad (Sv)_{\mathcal{C}\setminus X} \geq 0, \\ &\qquad v_i \geq 0, \qquad\qquad i = 1, \ldots, |\overline{\mathcal{R}}|, \end{aligned} \tag{M1}$$

where S represents the stoichiometric matrix of $\overline{\mathcal{N}}$.
If v^* is a solution to M1 and $f(v^*) > 0$ then the support of v^* is a stoichiometric factory from X to $\mathcal{T}$ and X is a stoichiometric precursor set for $\mathcal{T}$. □

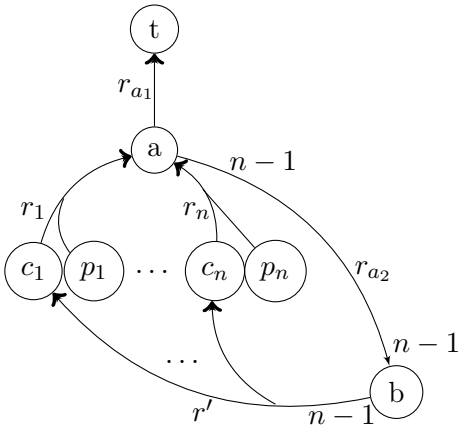
As concerns the problem of enumerating all solutions, we first observe that the proof that enumerating all minimal TPSs cannot be done in polynomial total time (that is, in the size of the input and the number of solutions) unless P=NP given in Acuña et al. (2012) can be immediately applied to show that enumerating all minimal SPSs cannot be done in polynomial total time unless P=NP. The same observation holds for enumerating all minimal cut sets (SCSs), which we define as follows:

Definition 14. *A set $X \subseteq \mathcal{X}$ is a **stoichiometric cut set** (SCS) (**topological cut set** (TCS)), if $\mathcal{X} \setminus X$ is* **not** *a stoichiometric precursor set (topological precursor set).*

We now show that the simultaneous enumeration of minimal SPSs and SCSs can be done in quasi-polynomial time. Notice that in Acuña et al. (2012), a quasi-polynomial time algorithm



(a)



(b)

Figure 3.3: (a) A network with $\mathcal{R} = \{r_t, r_a, r_b, r_1, \ldots, r_n\}$. Reaction r_i with $i = 2, \ldots, n$ consumes p_i and produces compound c. $\mathcal{T} = \{t\}$, $\mathcal{X} = \{p_1, \ldots, p_n\}$. All stoichiometric values are equal to one. There is one minimal SPS ($\{p_1\}$) and n minimal topological factories in $\psi(\mathcal{N})$. One contains only $\psi(r_1)$. The other minimal topological factories contain each $\{\psi(r_t), \psi(r_a), \psi(r_b)\}$ and one of the reactions in $\{\psi(r_2), \ldots, \psi(r_n)\}$, respectively. (b) In this network, the set of compounds is given by $\mathcal{C} = \{a, b, t, c_1, \ldots, c_n, p_1, \ldots, p_n\}$. The compounds $p_1, \ldots, p_n$ are the sources and t is the target. The stoichiometric values are equal to 1 if not stated otherwise. Beside the reactions $r_{a_1} : a \to t$ and $r_{a_2} : a \to b$, there is the reaction r' that consumes $n-1$ b and produces $\{c_1, \ldots, c_n\}$ (1 each). Furthermore, there are n reactions with $Subs(r_i) = \{c_i, p_i\}$ and $Prod(r_i) = \{a\}$, with $i = 1, \ldots, n$. The dots in the Figure illustrate the products $c_2, \ldots, c_{n-1}$ of r' that are not shown for simplicity. The reactions $r_2, \ldots, r_{n-1}$ are not shown for the same reason. There are n minimal topological factories in $\psi(\mathcal{N})$, each containing the reactions $\{\psi(r_{a_1}), \psi(r_{a_2}), \psi(r')\}$ and one of the many-to-one reactions of $\{\psi(r_1), \ldots, \psi(r_n)\}$, respectively. The only minimal SPS contains all sources.

to simultaneously enumerate all TPSs and TCSs was presented by formulating the problem with a monotone boolean formula and then using a result of Gurvich and Khachiyan (1999). Such approach is possible even in the case of SPSs and SCSs.

Theorem 3. *The set of minimal SPSs and the set of minimal SCSs can be enumerated in total quasi-polynomial time.*

Proof. Define the Boolean function $f : 2^{\mathcal{S}} \to \{0,1\}$ as $f(X) = 1$ if X is a SPS and $f(X) = 0$ otherwise. Clearly, this function is *monotone*: if $f(X) = 1$ then $f(Y) = 1$ for any set $Y \supseteq X$. The collection $\mathcal{P}$ of minimal SPSs is the collection of all minimal sets in $\mathcal{S}$ that evaluate to 1 and the collection $\mathcal{C}$ of minimal SCSs is the collection of all minimal sets whose complement in $\mathcal{S}$ evaluates to 0. In the context of monotone Boolean functions, minimal SPSs correspond to the *prime implicants* and minimal SCSs to the *prime implicates* of f. In Gurvich and Khachiyan (1999), a general algorithm is proposed to jointly enumerate prime implicants and prime implicates of any Boolean function. The algorithm and time analysis are rather technical and we only give a brief description of the incremental algorithm applied to our case. Briefly, given two collections of solutions already found, that is, of collections $(\mathcal{P}', \mathcal{C}')$ of SPSs and SCSs, the algorithm finds a set $X \subseteq \mathcal{S}$ such that X does not contain any minimal SPS in $\mathcal{P}'$ and $\mathcal{S} \setminus X$ does not contain any minimal SCS in $\mathcal{C}'$ (or proves that such set does not exist). Since either X is a SPS or $\mathcal{S} \setminus X$ is a SCS, we have found a new solution not in $(\mathcal{P}', \mathcal{C}')$. Such a new solution is found in time $O(n(\tau + n)) + m^{O(\log m)}$ where $n = |\mathcal{S}|$, m is the number of partial solutions already found (*i.e.* $m = |\mathcal{P}'| + |\mathcal{C}'|$) and τ is the time needed to evaluate f. Since τ is polynomial, we conclude the proof. □

3.4 Relation to previous work

The paper in the literature that comes closest to ours is Eker et al. (2013). In fact, one of their definitions coincides completely with our definition of SPS. However, their work concentrates on a more restrictive model, which they call *machinery-duplicating*. The underlying idea of the latter is that each compound involved in a path from the precursor set to the target set should be produced in strictly positive amount, allowing a cell to therefore duplicate itself. We translate their definition by using the concept of factory (cf. Definition 10).

Definition 15. *A* MD-stoichiometric factory *from $X \subseteq \mathcal{X}$ to $T \subseteq \mathcal{T}$ is a set $F \subseteq \mathcal{R}$, if there exists a flux vector $v \geq 0$ with $Y = Subs(F) \setminus X$ satisfying:*

1. $v_i \begin{cases} > 0 & i \in F \\ = 0 & otherwise, \end{cases}$
2. $(Sv)_{\mathcal{C} \setminus X} \geq 0$,
3. $(Sv)_{T \cup Y} > 0$.

A set $X \subseteq \mathcal{X}$ is a MD-stoichiometric precursor set *(MD-SPS) if there exists a MD-stoichiometric factory from X to $\mathcal{T}$.*

Comparing this definition to Definition 10, clearly any MD-SPS is a SPS, but not the other way around. Moreover, not every minimal MD-SPS is a minimal SPS. In their work, the authors claim that for the growth of a colony of cells, one must consider that not only the biomass compounds should be produced in positive amount, but also all the reactants of every reaction with nonzero flux that are not sources. However, as we already mentioned, cycles like the one in Figure 3.1 are considered unfeasible according to the machinery-duplicating model. Yet cycles with this structure are present in real networks and play an important role in metabolism, such as in the urea or the Krebs cycle.

3.5 Enumerating precursor sets via MILP

In the following section, we describe how to enumerate all minimal SPS and $MD-SPS$ using a MILP approach similar to Lee et al. (2000); de Figueiredo et al. (2009); von Kamp and Klamt (2014). The authors of these papers describe methods that enumerate reaction subsets by recursively solving MILP problems. Therein, solutions obtained in a previous step are excluded from the solution space.

3.5.1 Enumeration of minimal SPS

We now present a practical method to enumerate all minimal stoichiometric precursor sets that allow to produce the set $\mathcal{T}$ in a positive amount. We iteratively solve a series of optimisation problems: at each iteration a mixed integer linear programming (MILP) problem is solved to obtain a minimal precursor set X; then we define a new MILP by adding a constraint that removes the obtained solution X and all the sets that contain it from the feasible set. We keep repeating this process until all solutions are found.
We need some additional definitions. For each source compound $x_j \in \mathcal{X}$, we add to $\mathcal{R}$ a reaction, which we call *source-pool reaction*, that produces x_j from nothing (with stoichiometric coefficient 1). We denote this new set by $\overline{\mathcal{R}}$ and the set containing all source-pool reactions by $\overline{\mathcal{R}}_{\mathcal{X}}$. This set of reactions allows to model the availability of the source compounds since the upper bounds on their fluxes are linked to the amount of each source that is available. In the sequel $\overline{S}$ denotes the stoichiometric matrix S obtained by adding the columns given by the set of reactions $\overline{\mathcal{R}}_{\mathcal{X}}$ and v the flux vector, U is an upper bound constant for the values of each flux, ϵ is a vector of size $|\mathcal{T}|$ with an arbitrarily small positive real number in all coordinates, b is the vector of binary variables associated with each compound in $\mathcal{X}$, and we assume that $b_j = 1$ ($b_j = 0$) implies that compound x_j is used (not used) to produce the target.
Given a network $\mathcal{N} = \{\mathcal{C}, \mathcal{R}\}$, a stoichiometric matrix S, a set $\mathcal{X} \subseteq \mathcal{C}$ of sources and a set $\mathcal{T} \subseteq \mathcal{C}$ of targets, we first define the following optimisation problem (3.1) to find *the first minimal solution*:

$$\begin{aligned} \min f = \quad & \sum_{j=1}^{|\overline{\mathcal{R}}_{\mathcal{X}}|} b_j \\ s.t \quad & \left(\overline{S}v\right)_{\mathcal{T}} \geq \epsilon, \\ & \overline{S}v \geq 0 \\ & b_j = 0 \leftrightarrow v_j = 0, \quad \forall j \in \overline{\mathcal{R}}_{\mathcal{X}} \\ & b_j \in \{0,1\}, \quad \forall j \in \overline{\mathcal{R}}_{\mathcal{X}} \\ & 0 \leq v_i \leq U, \quad \forall i \in \overline{\mathcal{R}} \end{aligned} \tag{3.1}$$

Model (3.1) is similar to the first MILP presented in Zarecki et al. (2014). The first set of constraints requires to produce the target at least in a quantity ϵ. Instead of putting a small value for ϵ one could also put *e.g.* the maximum biomass yield. In this sense we enumerate all minimal precursor sets that allows for the maximal production of biomass. The second constraint allows for an accumulation of compounds. To enumerate minimal SPS at steady state, this constraint must be changed into $\overline{S}v = 0$. The third set of constraints (constraints $b_j = 0 \leftrightarrow v_j = 0$) denotes the fact that v_j — the flux associated to the source compound x_j — is positive if and only if $b_j = 1$. These constraints can be formulated as a MILP as follows:

$$\left.\begin{array}{rcl} b_j & \leq & v_j \\ v_j & \leq & Ub_j \end{array}\right\} \text{ for } \forall j \in \overline{\mathcal{R}}_{\mathcal{X}}, \tag{3.2}$$

If $b_j = 1$, we have $v_j \geq 1$, which will force us to have at least one unity of the source compound j.
Since the objective function of (3.1) minimises $\sum_j b_j$, then the optimal solution S^*, is a precursor set of minimum cardinality. We now show how to modify the MILP to obtain all other minimal precursor sets. To this goal let the pair (v^*, b^*) be an optimal solution to Problem (3.1). Let I_{b^*} be the support of b^*; we consider the following constraint:

$$\sum_{j \in I_{b^*}} b_j \leq |I_{b^*}| - 1, \tag{3.3}$$

Constraint (3.3) excludes the solution (v^*, b^*) and all the solutions that contain b^* from the set of solutions of (3.1). Hence, adding to (3.1) constraints in the form (3.3) gives a new instance of the MILP whose solution is a new precursor set that is not included in the previously obtained ones and is minimal (though not necessarily of minimum cardinality). By repeating this procedure, which is a standard technique in mixed integer linear programming, we iteratively enumerate all minimal solutions.
If the obtained problem has no feasible solution, then we claim that we have found all minimal precursor sets.

3.5.2 MILP constraints for $MD - SPS$

In the work of Eker *et al.* (Eker et al., 2013), the machinery-duplicating model is defined through the use of linear constraints and boolean operators. If a set of sources is a MD-SPS, this implies it is a feasible solution according to their model. The authors also present a method to enumerate all MD-SPSs. Suppose we are given a set $\{X_1, \ldots, X_k\}$ of precursor sets that were already found. Their method consists in finding a minimal subset of sources Y that verifies two conditions: (1) Y has at least one source in common with each precursor set in $\{X_1, \ldots, X_k\}$; and (2) the complement of Y must be able to produce the target according to the machinery-duplicating model. If one can find such a subset Y, a minimal precursor set can be obtained by taking the complement $\overline{Y}$ of Y, and finding one minimal subset of $\overline{Y}$. If no Y verifying the above conditions can be found, all minimal precursor sets have been enumerated and the algorithm stops.
Our method could also be adapted to consider the machinery-duplicating model presented by Eker et al. (2013). The machinery-duplicating constraint is defined as:

$$\left(\overline{S}v\right)_j > 0 \vee \bigwedge_{i \in Q_j} v_i = 0, \tag{3.4}$$

where Q_j is the set of indices of reactions that use the compound j as a substrate. This can be reformulated into MILP constraints as:

$$\begin{array}{rcl} \left(\overline{S}v\right)_j & \geq & \bar{\epsilon} - D_j E_j, \\ \sum_{i \in Q_j} v_i & \leq & D_j(1 - E_j), \end{array} \tag{3.5}$$

where $\bar{\epsilon}$ is an arbitrarily small positive real number, D_j is a constant that can take any value greater or equal to $U|Q_j|$ and E_j is an artificial binary variable. Adding (3.5) to Problem (1) allows us to enumerate all minimal stoichiometric precursor sets that respect the machinery-duplicating model.

3.6 Results and Discussion

In this section, we present the experiments we realised and discuss the results we obtained. We start by comparing the method we developed with the one of Eker *et al.* (Eker et al., 2013). We then show the performance of Sasita versus the approach where minimal $SPSs$ are obtained from combinations of minimal topological factories in the many-to-one network. Finally, we apply Sasita to some genome-scale metabolic networks, obtained from Monk *et al.* (Monk et al., 2013). The objective of this last part is both to illustrate how our method can be used and to validate it by reproducing the findings of the authors.
All the experiments were performed using an Intel QuadCore i7-4770 computer with 16GB of RAM memory. The algorithm Sasita is coded in Java (OpenJDK IcedTea) and uses Cplex (IBM ILOG AMPL/CPLEX 12.5.1) for solving the MILP models; the constants are fixed as follows: $\epsilon = 0.5$, $\bar{\epsilon} = 0.5$, $U = 1000.0$. The constraints (3.2) were coded using indicator constraints to avoid numerical instability. The software and all network and input files can be downloaded at http://sasita.gforge.inria.fr.

3.6.1 Comparison between Sasita and Eker *et al.*'s approach

We start by calling attention to the fact that the comparison with the method of Eker *et al.* was difficult due to the fact that it is not publicly available. We also were not able to obtain it upon request. We therefore implemented a version of Sasita that enumerates all minimal $MD-SPSs$ using the contraints given by Equation (3.5). As input we took the metabolic network, the set of sources $\mathcal{X}$ and the set of targets $\mathcal{T}$ provided in the Supplementary Material of Eker *et al.* (Eker et al., 2013). The authors provided also a list of "auxiliary compounds" without which, according to them, their model does not work. No auxiliary compound appears in the minimal precursor sets that are enumerated by Eker *et al.* It is not clear how these compounds are handled in their approach. If we treat such auxiliary compounds as ordinary ones, we are not able to enumerate a single $MD-SPS$ with Sasita. If we add a *source-pool reaction* for each one of the auxiliary compounds, we obtain the minimal $MD-SPS$ $X = \{CCO-PERI-BAC@SULFATE\}$. Eker *et al.* find 787 solutions and all of them contain Sulfate. So the minimal solution X we found is in fact a subset of all their solutions.
We provide at http://sasita.gforge.inria.fr a list of reactions F that form a MD-stoichiometric factory from X to $\mathcal{T}$, the flux values in F, and the stoichiometric matrix restricted to the reactions in F. Furthermore, we show that all substrates of the reactions in F and the target set $\mathcal{T}$ are produced in a positive amount using the reactions in F. Hence, the minimal $MD-SPS$ X fulfils the properties of a precursor set according to the machinery-duplicating model (Eker et al., 2013). Such minimal $MD-SPS$ X is not found by Eker *et al.*, probably because they do some preprocessing on the network that is not described in their paper and that we were not able to obtain upon request.

3.6.2 Comparison between Sasita and combinatorial approach

We ran both approaches, *i.e.* Sasita and a combinatorial approach (called Combi) where minimal $SPSs$ are obtained from combinations of minimal topological factories in the many-to-one network, on several instances. Our objective was to analyse the differences in the running times between both approaches, so we set Cplex into single thread mode for Sasita. Table 3.1 shows clearly that the MILP approach is more efficient than the combinatorial one. The networks of *S. muelleri*, *C. ruddii* and *B. aphidicola* were obtained from MetExplore, filtering out ubiquitous metabolites and pairs of co-factors. We obtained the *E. coli* core model from http://systemsbiology.ucsd.edu/InSilicoOrganisms/Ecoli/EcoliSBML. As

Table 3.1: Our MILP approach (SASITA) versus the combinatorial one

Strain	#compounds/#reactions	#sources/#targets	t_{Sasita}	t_{Combi}
S. muelleri	76/64	9/Pyruvate	<1s	<1s
C. ruddii	128/126	45/Pyruvate	<1s	1s
B. aphidicola	282/245	91/L-histidine	2s	2s
E. coli core†	72/126	14/Biomass core	<1s	42s
E. coli core	78/126	14/Biomass core	<1s	*
E. coli CFT073	1911/2949	26/Biomass core	12s	*
E. coli EDL993	1895/2943	25/Biomass core	13s	*
E. coli K-12	1806/2854	25/Biomass core	12s	*
E. coli Sakai	1895/2942	25/Biomass core	12s	*

† Filtered network (see text). The MILP approach (SASITA) and the combinatorial one (COMBI) are applied to several metabolic networks. For each instance, we provide the size of the network (number of compounds and reactions), the number of source and target compounds, and the time spent by both approaches (t_{Sasita}, t_{Combi}). One asterisk means that the combinatorial approach could not finish within the time limit.

sources we considered all compounds that are not produced by a reaction or those that are produced by reversible reactions only. For the *E. coli* strains, we used the same networks from Monk *et al.* (Monk et al., 2013) and considered as sources the compounds from Table 3.3. We set a time limit to the combinatorial approach of 2 hrs. SASITA is by far more efficient on genome-scale networks where the combinatorial approach did not finish within the time limit. To be able to show an example where both approaches finish and SASITA outperforms COMBI, we removed all compounds from the *E. coli* core network if they are consumed and produced by more than ten reactions. This is the case for M_atp_c, M_nad_c, M_nadh_c, M_h2o_c, M_h_e, M_h_c. The resulting network is denoted by a †. Notice that the number of reactions remains the same because we remove only the above-mentioned compounds from the reactions. The difference in the time spent to solve the problem is remarkable. It takes less than one second with SASITA and 42 seconds with COMBI.

3.6.3 Enumerating minimal precursor sets in genome-scale metabolic networks

In this case, we based our experiments on the work of Monk *et al.* (Monk et al., 2013) who investigated the pan and core metabolic capabilities of 55 *Escherichia coli* and *Shigella* strains based on genome-scale reconstructions of their metabolism. By core is meant the elements shared by all strains and by pan the union of the elements from all strains. As concerns the latter in particular, the authors found the pan to be enriched in alternate carbon metabolic pathways. In order to determine the functional differences among the strains, the authors computed by flux balance analysis (FBA) the growth phenotypes of 385 nutrients (henceforth called the *test metabolites/compounds*), each considered individually as a source of carbon, nitrogen, phosphorus and sulfur, aerobically and anaerobically. To that purpose, an *in silico* minimal medium that contains a sole carbon, nitrogen, phosphorus and sulfur source was defined. The authors then replaced the sole carbon source by each of the 385 test metabolites one at a time. Whether or not these new media constituted a growth condition was tested by

Table 3.2: Differences between solutions found by Monk *et al.* (Monk et al., 2013) and by Sasita.

Network	Matches	Not found
E. coli CFT073	599	0
E. coli O157:H7 EDL933	597	0
E. coli str. K-12 MG1655	607	7
E. coli O157:H7 str. Sakai	597	0

The column "Matches" has the amount of solutions from Monk et al. (2013) for which we found at least one subset. The column "Not found" indicates the amount of solutions from Monk et al. (2013) for which we could not find a correspondence.

FBA. The procedure was repeated for each source in the minimal media, namely for nitrogen, phosphorus and sulfur, as well as for each strain. The resulting metabolic phenotypes indicated strain-specific adaptation to nutritional environments.

Our first goal was to validate our method: we compared the results obtained with Sasita to the ones in Monk *et al.* (Monk et al., 2013). We enumerated and compared the minimal precursor sets allowing for biomass production of the *E. coli* strains, which included commensals as well as both intestinal and extraintestinal pathogens. We used for this the genome-scale metabolic models from Monk *et al.* (Monk et al., 2013). The strains were *E. coli* str. K-12 MG1655 (Commensal), *E. coli* O157:H7 str. Sakai (Enterohemorrhagic *E. coli*, EHEC), *E. coli* O157:H7 EDL933 (EHEC), and *E. coli* CFT073 (Uropathogenic *E. coli*, UPEC). The same 385 compounds tested in Monk et al. (2013) were given as part of the sources for different runs of Sasita. Since Monk *et al.* (Monk et al., 2013) were not interested in minimal solutions, we wanted to check whether our solutions were subsets of their solutions.

The second goal was to explore some solutions that were only found by Sasita in order to illustrate one application of our method. These solutions contain more than one of the test metabolites, after excluding sources from the minimal media (*i.e.* carbon, nitrogen, phosphorus and sulfur). Such solutions were explored as concerns strain-specific growth and their relation to niches and to pathotypes.

For almost all solutions found by Monk *et al.* (Monk et al., 2013), Sasita was capable of correctly finding at least one corresponding minimal subset. There is a small number of solutions (7) found by Monk *et al.* (Monk et al., 2013) and not by Sasita. In Table 3.2, we show the amounts of solutions found and not found for each strain. We confirmed through FBA that there is indeed no feasible flux for those solutions (this is further discussed below). In the remainder of this section, we explain how we realised our experiments and we present our results in more detail.

Two experiments were conducted. In both cases, oxygen was always available and we used as target an artificial compound that is added as an extra product of the core biomass reaction, with stoichiometry of 1.0. Also, the minimal media for *E. coli* CFT073 contained tryptophan, its auxotrophy. We now give a general description of each experiment. The exact list of compounds for each experiment as well as all networks can be found in the Sasita website.

In Experiment 1, we tested growth on minimal media of each strain as defined by Monk *et al.* (Monk et al., 2013) by enumerating the minimal precursor sets using as sources only the compounds from such minimal media. In Table 3.3, we present the list of compounds considered as sources for each strain, for this experiment. For each strain we found two solutions, one aerobic and another anaerobic as expected. This shows that, theoretically, the so-called "minimal media" are in fact not minimal as a whole (i.e. they are minimal in terms

Table 3.3: List of source compounds considered in the first experiment (Minimal medium defined by Monk et al. (2013))

Calcium, Cob(I)alamin, Chloride, Co^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, H^{+}, Potassium, Magnesium, Mn^{2+}, Molybdate, Calcium, Nickel, Selenate, Selenite, Tungstate, Zinc, D-Glucose, Sulfate, Ammonium, Diphosphate, Nicotinate, L-Tryptophan*, O_2

* Present only in simulations with *E. coli* CFT073

of carbon, sulfur, nitrogen and phosphorus sources) and that the considered strains can grow from a proper subset of that set of compounds.
In order to check the ability of each strain to grow on the 385 test compounds as sources of carbon, nitrogen, phosphorus and sulfur, we ran Experiment 2. For each network we considered as input source set a subset of the minimal media compounds plus a subset of the 385 test metabolites.
For subsets of the minimal media compounds, we considered the set of minimal media compounds minus one of the following: glucose, ammonium, phosphate or sulfate respectively. Since we were removing a compound from the minimal media, we included only the test metabolites that could replace the removed one (if we removed glucose, we considered only the set of test compounds that have carbon in their composition and so on). This was done because considering all the 385 test metabolites together leads to a combinatorial explosion of the number of solutions that are unpractical to enumerate with Sasita. In one case, namely when we remove glucose, the set of test metabolites to include was also too big and we needed to split it further in two smaller sets. As a side effect, this split of the input compounds can lead to a loss of some solutions, namely those containing compounds that are in different input sets. We thus may lose some solutions that have more than one minimal media compound replaced by two or more test metabolites. However, our split guarantees that at least all the sets considered in Monk *et al.* (Monk et al., 2013) are possible combinations of our input compounds because the authors replace glucose, ammonium, phosphate or sulfate from the minimal media with only one of the test metabolites. The lists of compounds that are considered as sources in each experiment are presented in tables 3.4 to 3.8.

First goal: Comparison with Monk *et al.*

We found 837 minimal precursor sets for *E. coli* CFT073 and between 11.164 and 13.732 for the other strains (Table 3.9).
This difference is remarkable but not surprising: *E. coli* CFT073 has a tryptophan auxotrophy, and tryptophan itself can be a source of carbon and nitrogen. Since we search for minimal precursor sets and tryptophan is always a source, there is no need for any extra source of carbon and nitrogen, thus reducing the number of solutions for this strain.
All solutions found by Sasita are either a subset of the given minimal media or a subset of the minimal media plus one or more test metabolites. The number of solutions where at least one of the four sources (carbon, nitrogen, phosphorus and sulfur) is replaced by only one compound among the 385 test metabolites is presented in Table 3.10. These solutions contain one test metabolite plus a subset of the minimal media in which the test metabolite replaces one, or more of the following compounds: glucose, ammonium, phosphate and sulfate. They therefore correspond to minimal sets of the solutions found by Monk *et al.* (Monk et al.,

Table 3.4: List of source compounds considered in the second experiment removing glucose from the minimum medium (part 1)

2',3'-Cyclic UMP, Nicotinate, Butyrate (n-C4:0), Co^{2+}, D-Cysteine, dIMP, 2',3'-Cyclic GMP, (S)-Propane-1,2-diol, L-Alanine, L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate-D-alanine, dCMP, dTMP, AMP, 2,3-diaminopropionate, L-Methionine, Cu^{2+}, L-Aspartate, Fe^{3+}, beta D-Galactose, N-Acetyl-D-galactosamine, CMP, cellobiose, Dodecanoate (n-C12:0), Fe^{2+}, D-Glycerate-2-phosphate, Dihydroxyacetone, Acetate, Deoxyuridine, dGMP, Ammonium, N-Acetyl-D-galactosamine 1-phosphate, Cytosine, Mn^{2+}, tungstate, D-Alanyl-D-alanine, O_2, Formaldehyde, Adenine, N,N'-diacetylchitobiose, 2-Deoxy-D-ribose, magnesium, D-Alanine, 5,6,7,8-Tetrahydrofolate, dAMP, L-Leucine, sn-Glycero-3-phosphoethanolamine, potassium, L-Ascorbate, Sulfate, L-Tryptophan*, D-Allose, Decanoate (n-C10:0), 2-Oxoglutarate, Acetoacetate, 3'-GMP, ethanesulfonate, 3'-AMP, Agmatine, Cytidine, D-Fructuronate, 1,4-alpha-D-glucan, Zinc, Acetaldehyde, Deoxyadenosine, 3-Phospho-D-glycerate, Ethanolamine, L-alanine-D-glutamate, Chloride, D-Galactose, D-Glucose 6-phosphate, Fumarate, L-alanine-L-glutamate, L-Cysteine, D-Fructose 6-phosphate, 2',3'-Cyclic AMP, Calcium, L-Fucose, selenite, L-Arginine, L-Arabinose, D-Fructose, N-Acetyl-D-glucosamine(anhydrous)N-Acetylmuramic acid, 2',3'-Cyclic CMP, alpha-D-Galactose 1-phosphate, H_2O, 4-Hydroxy-L-threonine, butanesulfonate, Fe(III)dicitrate, Thiamin, Dopamine, 3-hydroxycinnamic acid, nickel, N-Acetyl-D-glucosamine, H^+, Deoxyinosine, Molybdate, fructoselysine, 3'-UMP, N-Acetylneuraminate, 2-Dehydro-3-deoxy-D-gluconate, Glycerophosphoserine, sn-Glycero-3-phosphocholine, Ethanol, N-Acetyl-D-glucosamine 1-phosphate, 4-Aminobutanoate, N-Acetylmuramate, Allantoin, sn-Glycero-3-phospho-1 -inositol, Phosphate, Glycerophosphoglycerol, Co^{2+}, Deoxycytidine, D-Glucose 1-phosphate, Citrate, Sodium, 3'-cmp, 3-(3-hydroxy-phenyl)propionate, 5-Dehydro-D-gluconate, N-Acetyl-D-mannosamine, L-Asparagine, D-Glucose, Deoxyguanosine, Adenosine, Selenate, Cyanate, Cys-Gly, Formate, dUMP, Cob(I)alamin, L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate, Choline

* Present only in simulations with *E. coli* CFT073

Table 3.5: List of source compounds considered in the second experiment removing glucose from the minimum medium (part 2)

UDPgalactose, Lactose, (R)-Glycerate, L-Glutamine, Tyramine, Sulfate, Hexanoate (n-C6:0), Chloride, Phenethylamine, D-Xylose, Nicotinate, selenite, Ornithine, L-Galactonate, Xanthosine 5'-phosphate, D-Galactarate, Uracil, UDP-N-acetyl-D-galactosamine, octadecenoate (n-C18:1), sulfoacetate, Molybdate, Selenate, Phosphotyrosine, Zinc, D-Mannose, D-Lactate, octanoate (n-C8:0), Raffinose, 4-Hydroxyphenylacetate, Fe^{3+}, Propionate (n-C3:0), Glycerol 3-phosphate, L-Prolinylglycine, Thymidine, D-Mannitol, Calcium, tungstate, Maltotetraose, tetradecanoate (n-C14:0), Isethionic acid, potassium, L-Idonate, L-Glutamate, Co^{2+}, Ribitol, Cob(I)alamin, Melibiose, Maltohexaose, L-Tryptophan*, Guanosine, L-Xylulose, D-Glucosamine 6-phosphate, D-Glucuronate 1-phosphate, H2O, Phosphate, D-Arabinose, D-Arabitol, Propanal, Phenylacetaldehyde, Xanthosine, Pyruvate, UDP-N-acetyl-D-glucosamine, Maltotriose, O_2, Oxaloacetate, H^+, D-Glucuronate, Uridine, UDP-D-glucuronate, Hypoxanthine, O-Phospho-L-serine, L-tartrate, Xanthine, D-Glucosamine, GMP, D-Mannose 6-phosphate, octadecanoate (n-C18:0), nickel, Sucrose, psicoselysine, D-Galactonate, Taurine, Sodium, L-Lyxose, L-Methionine, Glycine, Guanine, magnesium, Inosine, L-Threonine O-3-phosphate, UDPglucose, Thiamin, Maltose, D-Malate, methanesulfonate, Orotate, Trehalose, L-Serine, L-Malate, D-tartrate, L-Leucine, GTP, Phenylpropanoate, Putrescine, Glycerol 2-phosphate, D-Glucarate, Urea, Galactitol, alpha-D-Ribose 5-phosphate, Co^{2+}, D-Galacturonate, Glycerol, D-Sorbitol, NMN, D-Glyceraldehyde, Mn^{2+}, Reduced glutathione, tetradecenoate (n-C14:1), Cu^{2+}, Maltopentaose, IMP, Hexadecanoate (n-C16:0), UMP, myo-Inositol hexakisphosphate, D-Ribose, L-Rhamnose, 2(alpha-D-Mannosyl)-D-glycerate, L-Threonine, D-Glucose, Ammonium, Fe^{2+}, D-Gluconate, L-Lactate, L-Proline, Glycolate, Hexadecenoate (n-C16:1), Succinate, D-Serine

* Present only in simulations with *E. coli* CFT073

Table 3.6: List of source compounds considered in the second experiment removing ammonium from the minimum medium

Cyanate, Inosine, D-Serine, UMP, GTP, dUMP, 2',3'-Cyclic GMP, D-Alanine, Nitrite, 3'-cmp, Adenosine, Thymidine, Cytosine, Sodium, Ethanolamine, D-Alanyl-D-alanine, selenite, Deoxyinosine, Uracil, L-Cysteine, L-Serine, L-Glutamine, D-Glucose, N-Acetyl-D-glucosamine(anhydrous)N-Acetylmuramic acid, Cytidine, 5,6,7,8-Tetrahydrofolate, sn-Glycero-3-phosphoethanolamine, Ammonium, Cu^{2+}, N-Acetyl-D-mannosamine, L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate, 3'-UMP, fructoselysine, nickel, potassium, Calcium, 3'-AMP, Glycine, IMP, 2,3-diaminopropionate, N,N'-diacetyl-chitobiose, CMP, N-Acetylmuramate, Mn^{2+}, dTMP, L-Glutamate, 2',3'-Cyclic UMP, Dopamine, Guanine, Phosphotyrosine, Xanthine, L-Aspartate, dIMP, N-Acetyl-D-glucosamine, UDP, galactose, UDP-N-acetyl-D-galactosamine, 4-Hydroxy-L-threonine, Ornithine, Co^{2+}, H^{+}, tungstate, Taurine, GMP, L-Tryptophan*, Tyramine, dGMP, Urea, Deoxyadenosine, Reduced glutathione, Agmatine, magnesium, Guanosine, UDPglucose, Deoxyuridine, Selenate, Molybdate, L-Asparagine, L-Prolinylglycine, 2',3'-Cyclic AMP, AMP, Nitric oxide, D-Cysteine, Allantoin, L-Threonine, Glycerophosphoserine, Xanthosine 5'-phosphate, Hypoxanthine, Xanthosine, Putrescine, Co^{2+}, L-Arginine, O-Phospho-L-serine, 2',3'-Cyclic, CMP, Orotate, Zinc, Deoxyguanosine, Sulfate, Nicotinate, D-Glucosamine, N-Acetyl-neuraminate, Cob(I)alamin, 4-Aminobutanoate, L-Proline, Nitrate, Uridine, L-Methionine, Adenine, Fe^{2+}, Phenethylamine, L-Leucine, N-Acetyl-D-galactosamine 1-phosphate, L-Alanine, Deoxycytidine, Fe^{3+}, sn-Glycero-3-phosphocholine, L-Threonine O-3-phosphate, Chloride, UDP-D-glucuronate, D-Glucosamine 6-phosphate, UDP-N-acetyl-D-glucosamine, O_2, Phosphate, 3'-GMP, psicoselysine, N-Acetyl-D-glucosamine 1-phosphate, dAMP, Thiamin, L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate-D-alanine, dCMP, L-alanine-L-glutamate, L-alanine-D-glutamate, Cys-Gly, NMN, N-Acetyl-D-galactosamine, H2O, Choline

* Present only in simulations with *E. coli* CFT073

Table 3.7: List of source compounds considered in the second experiment removing phosphate from the minimum medium

D-Glucose 1-phosphate, selenite, D-Glycerate-2-phosphate, 2',3'-Cyclic CMP, Phosphonate, O-Phospho-L-serine, H^+, sn-Glycero-3-phosphocholine, 2',3'-Cyclic UMP, NMN, D-Glucuronate 1-phosphate, UDPgalactose, Glycerol 2-phosphate, Cob(I)alamin, Sodium, N-Acetyl-D-glucosamine 1-phosphate, UDP-N-acetyl-D-glucosamine, 3-Phospho-D-glycerate, AMP, Cu^{2+}, Zinc, D-Glucose 6-phosphate, D-Glucose, alpha-D-Ribose 5-phosphate, magnesium, CMP, Sulfate, Fe^{2+}, Glycerol 3-phosphate, CO_2, 3'-UMP, GMP, Calcium, Xanthosine 5'-phosphate, UDPglucose, nickel, L-Tryptophan*, Phosphate, N-Acetyl-D-galactosamine 1-phosphate, dIMP, myo-Inositol hexakisphosphate, sn-Glycero-3-phospho-1-inositol, dCMP, GTP, L-Threonine O-3-phosphate, dAMP, D-Fructose 6-phosphate, 2',3'-Cyclic GMP, Molybdate, Glycerophosphoglycerol, Co^{2+}, IMP, Chloride, alpha-D-Galactose 1-phosphate, Selenate, dUMP, H2O, UMP, UDP-D-glucuronate, Fe^{3+}, 3'-GMP, Phosphotyrosine, dGMP, 3'-AMP, Glycerophosphoserine, sn-Glycero-3-phosphoethanolamine, dTMP, potassium, 3'-cmp, tungstate, Ammonium, UDP-N-acetyl-D-galactosamine, D-Mannose 6-phosphate, D-Glucosamine 6-phosphate, 2',3'-Cyclic AMP, O_2, Mn^{2+}

* Present only in simulations with *E. coli* CFT073

Table 3.8: List of source compounds considered in the second experiment removing sulfate from the minimum medium

Fe^{3+}, Phosphate, H^+, Molybdate, D-Cysteine, sulfoacetate, L-Cysteine, magnesium, Thiamin, nickel, CO_2, Taurine, Reduced glutathione, Thiosulfate, L-Methionine, potassium, Cys-Gly, butanesulfonate, Isethionic acid, D-Glucose, Zinc, Sodium, ethanesulfonate, Sulfate, Calcium, Ammonium, Cob(I)alamin, Cu^{2+}, tungstate, L-Tryptophan*, Selenate, methanesulfonate, Chloride, O_2, Mn^{2+}, Co^{2+}, H2O, selenite, Fe^{2+}

* Present only in simulations with *E. coli* CFT073

Table 3.9: Number of solutions found for each *E. coli* strain

Strain	Solutions
E. coli CFT073	837
E. coli EDL993	11.164
E. coli K-12	13.732
E. coli Sakai	11.164

Table 3.10: Number of solutions with one test metabolite.

Sources	*E. coli* CFT 073		*E. coli* EDL933		*E. coli* K-12		*E. coli* Sakai	
	O_2	No O_2	O_2	No O_2	O_2	No O_2	O_2	No O_2
C	0	0	104	51	109	61	104	54
C,N	0	0	51	42	52	44	51	42
C,N,P	0	0	37	22	38	22	37	22
C,N,S	0	0	8	0	8	0	8	0
C,P	0	0	14	22	13	21	14	22
C,S	0	0	0	0	2	0	0	0
N	0	0	11	14	14	10	11	14
P	51	51	6	6	7	7	6	6
S	22	0	14	0	10	0	14	0
Minimal Media	1	1	1	1	1	1	1	1
Total	74	52	246	158	254	166	246	161
Total	126		404		420		407	

Number of solutions with one test metabolite and removing one or more sources from the minimal media, namely carbon (C), nitrogen (N), sulfur (S) or phosphorus (P).

2013).

Other differences found in the comparison of our results with those from Monk et al. (2013) are presented and discussed below. Some such differences arise for *E. coli* CFT073 for which, as mentioned, tryptophan can always be a source of carbon and nitrogen. Since we find a minimal solution without any test metabolite and without glucose and ammonium, but with tryptophan, it is a minimal subset of all solutions from Monk et al. (2013) considering the replacement of glucose or ammonium by a test metabolite. Furthermore, a few minimal precursor sets for which Monk *et al.* (Monk et al., 2013) found no growth are present among our solutions because of the different conditions we allowed in our test, namely some sources were available at bigger amounts and the compounds were allowed to accumulate. There were as well 7 solutions for which Monk *et al.* (Monk et al., 2013) found a positive flux and we did not (see Table 3.2). Those solutions are for the K-12 strain. Among these solutions, 6 are related with the compounds 4-Hydroxy-L-threonine and Oxaloacetate from the exchange subsystem, and in fact there are no reactions in the network that use those compounds to produce anything. The remaining solution is the one using Thiosulfate as source of sulfur, and we confirmed no growth by FBA for this condition. There is one last difference, for the solution with the test metabolite Fe(III) dicitrate which allows growth as a carbon source in aerobic and anaerobic conditions for *E. coli* K-12. We explicitly found only the anaerobic solution. Since the aerobic one can be seen as a superset of the anaerobic, it is not a minimal precursor set. This does not happen in the other test metabolites owing to different iron oxidation states in each solution. We thus found different aerobic and anaerobic solutions when we enumerated the minimal precursor sets.

In conclusion, we obtained almost all of the solutions found by Monk *et al.* (Monk et al., 2013), showing that the nutrient sources of alternate catabolic pathways are part of the minimal precursor sets that allow for biomass production in most of the tested *E. coli* strains.

Table 3.11: Solutions with more than one test metabolite.

Test Metabolites	Solutions	*E. coli* strains
2	710	CFT073
2	3.387	EDL993;K-12;Sakai
2	298	EDL993;Sakai
2	656	K-12
3	4.992	EDL993;K-12;Sakai
3	703	EDL993;Sakai
3	2.106	K-12
4	1.027	EDL993;K-12;Sakai
4	299	EDL993;Sakai
4	1.113	K-12
5	19	EDL993;K-12;Sakai
5	21	EDL993;Sakai
5	5	K-12
6	5	EDL993;K-12;Sakai
6	4	EDL993;Sakai

Second goal: Original results using SASITA

The remaining solutions go beyond the analyses performed by Monk *et al.* (Monk et al., 2013) because they contain two or more test metabolites (Table 3.11). Most of these solutions actually have two or three test metabolites (4.961 and 7.801 respectively). In both cases, more than 60% of those solutions are aerobic.

Most solutions in which the two or three test metabolites replace all four sources from the minimal media (glucose, ammonium, phosphate, sulfate) are found in all the three strains, *E. coli* EDL993, *E. coli* K-12 and *E. coli* Sakai. Moreover, there are some minimal precursor sets specific to *E. coli* EDL993 and *E. coli* Sakai which are both EHEC, and others specific to *E. coli* K-12 which is commensal (Table 3.12). The latter are specific to pathotypes and probably indicate adaptations to nutritional environments.

From these results, the pairs of test metabolites in the 6 solutions specific to *E. coli* EDL993 and *E. coli* Sakai are: N-Acetyl-D-galactosamine 1-phosphate with butanesulfonate, ethanesulfonate or taurine, respectively; all aerobic and each one in two solutions with different iron states (line in bold in Table 3.12). These minimal precursor sets are not solutions for the strain *E. coli* K-12. N-Acetyl-D-galactosamine 1-phosphate was shown to give extraintestinal pathogenic strains of *E. coli* a catabolic advantage when compared to commensals, supporting growth in 100% of the cases compared to 67%, respectively (Monk et al., 2013). Furthermore, the enterohemorrhagic strains *E. coli* EDL933 and *E. coli* Sakai were shown to occupy the same niche in the streptomycin-treated mouse intestine (Meador et al., 2014) while *E. coli* EDL933 was shown not to colonise the same niche and does not use the same sugars as carbon source as the commensal *E. coli* K-12 (Fabich et al., 2008; Leatham et al., 2009). The three solutions detailed above and the solutions presented in Table 3.12 therefore represent metabolic capabilities that are specific to the pathogenic strains analysed here when compared to the commensal strain *E. coli* K-12, in agreement with the niches occupied by such strains.

Table 3.12: Solutions with two or three test metabolites.

Test Metabolites	Solutions	*E. coli* Strains
2	674	CFT073
2	556	EDL993; K-12; Sakai
2	**6**	**EDL993; Sakai**
3	2.641	EDL993; K-12; Sakai
3	143	EDL993; Sakai
3	511	K-12

Solutions with two or three test metabolites without any of the four sources from the minimal media (carbon, nitrogen, phosphorus and sulfur). Further details about the row in bold is given in the text.

These results suggest that SASITA can depict pathotype and niche-specific metabolic capabilities which allow broad *in silico* studies of strains or species interactions. For instance, an extension of the analysis presented in this paper to a larger dataset of *E. coli* strains including both pathogenic and commensal biotypes could help predict *in silico* sets of commensal strains that would prevent the colonisation of pathogens due to a consumption by the native microbiota of the nutrients required by the pathogen (see the experimental study of mutant phenotypes in Maltby *et al.* (Maltby et al., 2013)).

Source equivalence classes

The analysis of the collection of minimal precursor sets for a given metabolic network, and a set of sources and targets may become difficult if there is a high number of solutions. Therefore, we grouped sources that are equivalent with respect to the minimal precursor set solutions into *equivalence classes*. Herefore, we refer to the definition of source equivalence of Eker et al. (2013):

Definition 16. *Given a collection of minimal precursor sets A for a set of sources $\mathcal{X} \subseteq \mathcal{C}$ and targets $\mathcal{T} \subseteq \mathcal{C}$, we call the sources $c_1, c_2 \in \mathcal{X}$ to be equivalent with respect to A if and only if*

1. $\forall a \in A$ *with* $c_1 \in a : ((a \setminus \{c_1\}) \cup \{c_2\}) \in A$, *and*
2. $\forall a \in A$ *with* $c_2 \in a : ((a \setminus \{c_2\}) \cup \{c_1\}) \in A$.

Thus, two sources $c_1, c_2 \in \mathcal{X}$ are equivalent, if we obtain a minimal precursor set when replacing c_1 with c_2 in every minimal precursor set in which c_1 occurs, and vice versa. This equivalence relation is reflexive, symmetric, and transitive (Eker et al., 2013). Indeed, every source belongs to exactly one equivalence class. One compound of each equivalence class is chosen as representative compound of the class. When we replace each compound c of a minimal precursor set by the representative compound of the equivalence class of c, we end up with several duplicated minimal precursor sets that can be removed. Thus, we can reduce the number of minimal precursor set solutions without loosing information. Indeed, each representative compound in the remaining (compressed) minimal precursor sets can be replaced by any compound belonging to the same equivalence class to obtain the original minimal precursor sets.

Table 3.13: Comparison of the number of original and compressed solutions found for each *E. coli* strain.

Strain	# Original solutions	# Compressed solutions
E. coli CFT073	837	45
E. coli EDL993	11.164	1648
E. coli K-12	13.732	2167
E. coli Sakai	11.164	1648

Table 3.14: Equivalence classes for *E. coli* CFT073. Each line corresponds to an equivalence class. The compounds that are part of an equivalence class are in the right column. Their atom composition (C,N,P,S) is shown in the left column.

Atoms	Source compounds
C,P	M_minohp_e, M_glcur1p_e, M_man6p_e, M_r5p_e, M_glyc2p_e, M_glyc3p_e
C,P	M_g3pg_e, M_2pg_e, M_f6p_e, M_g1p_e, M_gal1p_e, M_g6p_e, M_3pg_e, M_g3pi_e
C,S	M_isetac_e, M_sulfac_e, M_mso3_e
C,S	M_butso3_e, M_ethso3_e
C,N,P	M_tyrp_e, M_gmp_e, M_pser_DASH_L_e, M_uacgam_e, M_gam6p_e, M_imp_e, M_ump_e, M_udpacgal_e, M_udpgal_e, M_xmp_e, M_udpg_e, M_nmn_e, M_thrp_e, M_gtp_e, M_udpglcur_e
C,N,P	M_cmp_e, M_dgmp_e, M_23cump_e, M_23camp_e, M_dtmp_e, M_3gmp_e, M_dump_e, M_dcmp_e, M_23ccmp_e, M_3ump_e, M_g3pc_e, M_amp_e, M_acgam1p_e, M_3cmp_e, M_3amp_e, M_acgal1p_e, M_dimp_e, M_g3ps_e, M_damp_e, M_23cgmp_e, M_g3pe_e
C,N,S	M_cys_DASH_D_e, M_cys_DASH_L_e, M_cgly_e
C,N,S	M_taur_e, M_gthrd_e

We computed the equivalence classes for each of the collections of minimal precursor sets obtained for the different *E. coli* strains (depicted in Table 3.9). The respective number of compressed minimal precursor sets is shown in Table 3.13. Indeed, the difference in the number of original and compressed solutions is remarkable. There is an up to 18-fold reduction (*E. coli* CFT073).

Equivalence classes that contain a single source are of less interest as the sources therein are equivalent only to themselves. Therefore, only the equivalence classes that contain more than one source are shown in Table 3.14 (for *E. coli* CFT073), Table 3.15 (for *E. coli* EDL993 and *E. coli* Sakai), and Table 3.16 (for *E. coli* K-12). The metabolic networks of *E. coli* EDL993 and *E. coli* Sakai are very similar which is reflected by (i) an identical collection of precursor sets and (ii) identical equivalence classes. Note that a compressed minimal precursor set does not necessarely contain a compound of each equivalence class.

Table 3.15: Equivalence classes for *E. coli* EDL993, and *E. coli* Sakai. Each line corresponds to an equivalence class. The compounds that are part of an equivalence class are in the right column. Their atom composition (C,N,P,S) is shown in the left column.

Atoms	Source compounds
N	M_no3_e, M_no2_e
C	M_fald_e, M_arab_DASH_L_e, M_fruur_e, M_ac_e, M_gal_e, M_for_e, M_2ddglcn_e, M_fru_e, M_dha_e, M_ddca_e, M_etoh_e, M_14glucan_e, M_fum_e, M_fuc_DASH_L_e, M_cit_e, M_acald_e, M_dca_e, M_akg_e, M_ascb_DASH_L_e, M_12ppd_DASH_S_e, M_gal_DASH_bD_e
C	M_mnl_e, M_octa_e, M_tre_e, M_glcr_e, M_man_e, M_glcn_e, M_oaa_e, M_tartr_DASH_L_e, M_tartr_DASH_D_e, M_lac_DASH_L_e, M_glcur_e, M_rib_DASH_D_e, M_lac_DASH_D_e, M_xyl_DASH_D_e, M_xylu_DASH_L_e, M_sbt_DASH_D_e, M_ttdcea_e, M_hxa_e, M_raffin_e, M_glyald_e, M_sucr_e, M_galctn_DASH_L_e, M_lyx_DASH_L_e, M_rmn_e, M_malt_e, M_glyc_e, M_glyc_DASH_R_e, M_succ_e, M_galur_e, M_galct_DASH_D_e, M_malthx_e, M_mal_DASH_D_e, M_glyclt_e, M_galt_e, M_mal_DASH_L_e, M_hdcea_e, M_ocdcea_e, M_lcts_e, M_ppal_e, M_maltpt_e, M_melib_e, M_pyr_e, M_hdca_e, M_ttdca_e, M_ppa_e, M_malttr_e, M_ocdca_e, M_maltttr_e
C	M_3hpppn_e, M_3hcinnm_e
C,N	M_arg_DASH_L_e, M_acgal_e, M_trp_DASH_L_e, M_4hthr_e, M_alaala_e, M_anhgm_e, M_din_e, M_ala_DASH_D_e, M_LalaDglu_e, M_4abut_e, M_acmum_e, M_23dappa_e, M_LalaDgluMdapDala_e, M_cytd_e, M_LalaLglu_e, M_adn_e, M_chtbs_e, M_agm_e, M_ala_DASH_L_e, M_dgsn_e, M_LalaDgluMdap_e, M_acmana_e, M_acnam_e, M_asn_DASH_L_e, M_asp_DASH_L_e, M_dcyt_e, M_acgam_e, M_etha_e, M_dad_DASH_2_e
C,N	M_psclys_e, M_thymd_e
C,N	M_gln_DASH_L_e, M_gly_e, M_ser_DASH_L_e, M_glu_DASH_L_e, M_gam_e, M_orn_e, M_xtsn_e, M_gsn_e, M_ins_e, M_ser_DASH_D_e, M_thr_DASH_L_e, M_pro_DASH_L_e, M_ptrc_e, M_progly_e
C,N	M_xan_e, M_hxan_e
C,S	M_isetac_e, M_sulfac_e, M_mso3_e
C,P	M_glcur1p_e, M_man6p_e, M_r5p_e, M_glyc2p_e, M_glyc3p_e
C,P	M_g3pg_e, M_g6p_e, M_f6p_e, M_g3pi_e, M_g1p_e, M_gal1p_e
C,N,P	M_cmp_e, M_dgmp_e, M_23camp_e, M_3gmp_e, M_23ccmp_e, M_dcmp_e, M_amp_e, M_acgam1p_e, M_3cmp_e, M_3amp_e, M_acgal1p_e, M_dimp_e, M_g3ps_e, M_g3pe_e, M_23cgmp_e, M_damp_e
C,N,P	M_gmp_e, M_pser_DASH_L_e, M_uacgam_e, M_gam6p_e, M_imp_e, M_udpacgal_e, M_xmp_e, M_thrp_e, M_nmn_e
C,N,P	M_udpgal_e, M_udpg_e, M_ump_e, M_udpglcur_e
C,N,P	M_3ump_e, M_23cump_e, M_dump_e
C,N,P	M_g3pc_e, M_dtmp_e
C,N,S	M_cys_DASH_D_e, M_cys_DASH_L_e, M_cgly_e

Table 3.16: Equivalence classes for *E. coli* K-12. Each line corresponds to an equivalence class. The compounds that are part of an equivalence class are in the right column. Their atom composition (C,N,P,S) is shown in the left column.

Atoms	Source compounds
N	M_no3_e, M_no2_e
C	M_3hpppn_e, M_3hcinnm_e
C	M_acac_e, M_fald_e, M_arab_DASH_L_e, M_fruur_e, M_ac_e, M_gal_e, M_for_e, M_2ddglcn_e, M_fru_e, M_dha_e, M_ddca_e, M_etoh_e, M_14glucan_e, M_fum_e, M_fuc_DASH_L_e, M_cit_e, M_dca_e, M_acald_e, M_all_DASH_D_e, M_akg_e, M_ascb_DASH_L_e, M_but_e, M_12ppd_DASH_S_e, M_gal_DASH_bD_e, M_5dglcn_e
C	M_mnl_e, M_octa_e, M_galctn_DASH_D_e, M_tre_e, M_glcr_e, M_man_e, M_glcn_e, M_tartr_DASH_L_e, M_tartr_DASH_D_e, M_lac_DASH_L_e, M_glcur_e, M_rib_DASH_D_e, M_lac_DASH_D_e, M_xyl_DASH_D_e, M_xylu_DASH_L_e, M_sbt_DASH_D_e, M_ttdcea_e, M_hxa_e, M_idon_DASH_L_e, M_glyald_e, M_sucr_e, M_lyx_DASH_L_e, M_galctn_DASH_L_e, M_rmn_e, M_malt_e, M_glyc_e, M_glyc_DASH_R_e, M_succ_e, M_galur_e, M_galct_DASH_D_e, M_malthx_e, M_mal_DASH_D_e, M_glyclt_e, M_galt_e, M_mal_DASH_L_e, M_hdcea_e, M_manglyc_e, M_ocdcea_e, M_lcts_e, M_ppal_e, M_maltpt_e, M_melib_e, M_pyr_e, M_hdca_e, M_ttdca_e, M_ppa_e, M_malttr_e, M_ocdca_e, M_maltttr_e
C,S	M_butso3_e, M_ethso3_e
C,S	M_isetac_e, M_sulfac_e, M_mso3_e
C,N	M_xan_e, M_hxan_e
C,N	M_psclys_e, M_thymd_e
C,N	M_gln_DASH_L_e, M_gly_e, M_ser_DASH_L_e, M_glu_DASH_L_e, M_gam_e, M_orn_e, M_xtsn_e, M_gsn_e, M_ins_e, M_ser_DASH_D_e, M_thr_DASH_L_e, M_pro_DASH_L_e, M_ptrc_e, M_progly_e
C,N	M_arg_DASH_L_e, M_trp_DASH_L_e, M_alaala_e, M_anhgm_e, M_din_e, M_ala_DASH_D_e, M_LalaDglu_e, M_4abut_e, M_acmum_e, M_23dappa_e, M_LalaDgluMdapDala_e, M_cytd_e, M_LalaLglu_e, M_adn_e, M_chtbs_e, M_agm_e, M_ala_DASH_L_e, M_dgsn_e, M_LalaDgluMdap_e, M_acmana_e, M_acnam_e, M_asn_DASH_L_e, M_asp_DASH_L_e, M_dcyt_e, M_acgam_e, M_etha_e, M_dad_DASH_2_e
C,P	M_glcur1p_e, M_man6p_e, M_r5p_e, M_glyc2p_e, M_glyc3p_e
C,P	M_g3pg_e, M_g6p_e, M_f6p_e, M_g3pi_e, M_g1p_e, M_gal1p_e
C,N,P	M_dcmp_e, M_23ccmp_e, M_cmp_e, M_amp_e, M_dgmp_e, M_acgam1p_e, M_3cmp_e, M_3amp_e, M_23camp_e, M_3gmp_e, M_dimp_e, M_damp_e, M_23cgmp_e, M_g3pe_e, M_g3ps_e
C,N,P	M_gmp_e, M_pser_DASH_L_e, M_uacgam_e, M_gam6p_e, M_imp_e, M_xmp_e, M_thrp_e, M_nmn_e
C,N,P	M_udpgal_e, M_udpg_e, M_ump_e, M_udpacgal_e, M_udpglcur_e
C,N,P	M_3ump_e, M_23cump_e, M_dump_e
C,N,P	M_g3pc_e, M_dtmp_e
C,N,S	M_cys_DASH_D_e, M_cys_DASH_L_e, M_cgly_e

3.7 Conclusions and Perspectives

We examined the relationship between topological and stoichiometric precursor sets. We highlighted that stoichiometric precursor sets can be obtained from combinations of minimal topological factories in the many-to-one network. However, this does not lead to an efficient method. We then presented Sasita, an efficient algorithm for the exhaustive enumeration of minimal precursor sets for a given target that takes into account stoichiometry. To the best of our knowledge, there exists only one previous approach for this problem due to Eker *et al.* (Eker et al., 2013) who proposed two different constraint models, steady-state and machinery-duplicating. However, in their computations, the authors use only the latter, that requires a strictly positive net production of the intermediate compounds on the path from the sources to the target. This model may exclude solutions as we showed (Figure 3.1).
In our experiments, we enumerated and compared the minimal precursor sets of nutrient sources of alternate catabolic pathways allowing for biomass production of some *Escherichia coli* strains, comprising commensal and both intestinal and extraintestinal pathogens, using genome-scale metabolic models. We compared our results to those of Monk *et al.* (Monk et al., 2013) in order to have a guideline on part of the solutions we generated, since our approach is different from the one that the authors used, and our results go beyond such comparison as we find solutions that were not obtained by Monk *et al.* (Monk et al., 2013). We found metabolic capabilities that distinguish the strains compared in their ability to catabolise nutrients, and such were specific to pathotypes and niches of *E. coli* strains.
Our method can therefore be used in a wide variety of applications in order to study minimal growth conditions as well as strains and/or species interactions based on their catabolic abilities and their nutritional niches. One valuable application in this context would be to predict patterns of colonisation of commensal and pathogenic *E. coli* strains in the intestine.
Our method can furthermore be used to refine a metabolic network. If growth of an organism is observed for a defined medium in the laboratory, but no minimal precursor set is a subset of such medium, then either the metabolic network lacks reactions, *e.g.* export reactions, or the biomass function is not well formulated.
The execution of all experiments of the comparison with Monk *et al.* (Monk et al., 2013) took altogether around 5 days. The execution times ranged from 20 seconds to 12 hours. Running the experiments in parallel, one could such retrieve the results after less than a day. We cannot claim to be more efficient than Eker *et al.* (Eker et al., 2013) as their software is not available for testing. However, we are guaranteed to enumerate all minimal $SPSs$ and $MD-SPSs$.

We observed that the computation time to obtain the next minimal precursor set in Sasita increases with the number of already computed minimal solutions. In addition, we have shown that determining the source equivalence classes can be used for compression and thus reduces the number of minimal precursor sets. Consequently, having the source equivalence classes at hand before enumerating the minimal precursor sets would probably reduce the computation time. Therefore, for a given set of equivalence classes $C_1, \ldots, C_n$ we can transform the network in the following way. For each equivalence class C_i we introduce a dummy compound $source_{C_i}$. Furthermore, for every source $c \in C_i$, we add a reaction that consumes $source_{C_i}$ and produces c. The compounds in the original set of sources $\mathcal{X}$ are replaced by the dummy sources $source_{C_i}$. Enumerating the minimal precursor sets of the transformed network results in the compressed minimal precursor sets. The beforehand identification of the source equivalence classes may be addressed in the future.
In the current version of Sasita we enumerate minimal precursor sets that allow for the production of the target. It is straightforward to extend our approach to enumerate minimal precursor sets that enable to maximize the production of the target. Even though the

processing time is not too long, but because our approach is deterministic, we consider to establish a database that stores minimal precursor sets for a given metabolic network and a set of sources and targets.

Chapter 4

Minimal Stoichiometric Factories

Contents

4.1 Introduction

A metabolic network presents the capabilities of an organism to transform chemical compounds by means of chemical reactions. An organism produces compounds that are suitable for growth, reproduction and maintenance, and there are usually many alternative ways to produce them. Furthermore, these alternatives often share compounds resulting in a highly connected network that is difficult to analyze. A common practice to model the organism's growth is to add an artificial *biomass reaction* to the metabolic network. This reaction consumes, in the right amounts, all compounds that are needed for growth, and produces one unit of an artificial biomass compound. Thus, the organism needs to produce the biomass compound to be able to grow. How does an organism make use of its metabolic network to produce biomass? Does the metabolic network exhibit a structure that can be explored? Is it possible to break the whole network into smaller subnetworks that then could be analyzed independently? These are the questions we address in this chapter of the thesis.
A traditional metabolic pathway, *e.g.* the glycolysis pathway and the Krebs cycle, is a set of reactions that fulfill a specific task. The reactions that take part in a pathway were experimentally discovered step by step. Drawing all metabolic pathways of an organism in one map provides a nice overview of its metabolism (see Figure 4.1). Even though this is an already quite complex picture, it however represents only a coarse view of the metabolic functions of an organism. This is due to the fact that these pathways were detected independently, neglecting the fact that they share compounds and thus interact.

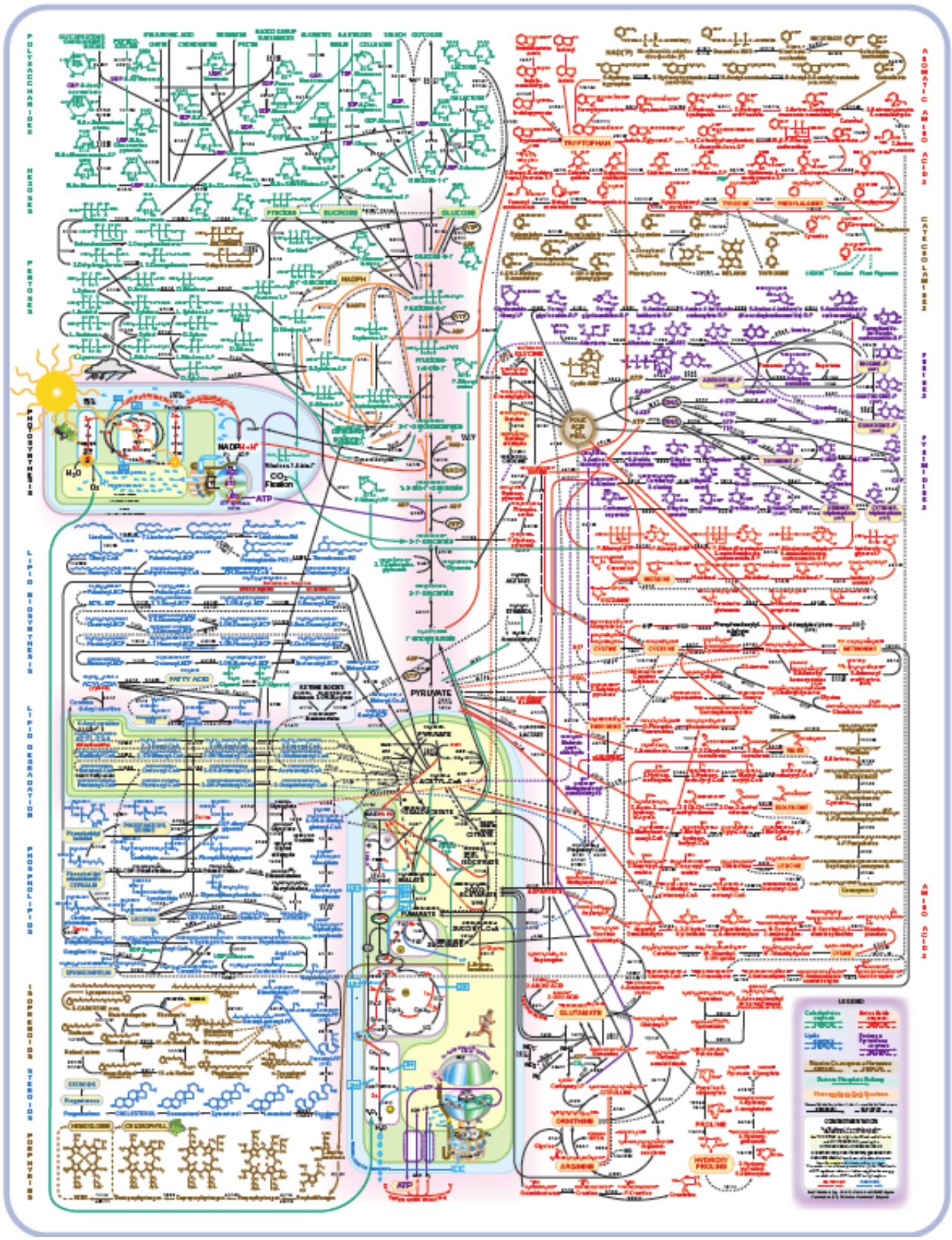


Figure 4.1: Metabolic pathway map (adopted from Sigma-Aldrich (2016))

To gather a more detailed view on the metabolism of an organism, one needs to reconstruct its genome-scale metabolic network which can be done by following *e.g.* the protocol of Thiele and Palsson (2010). The latter consists of 94 iterative, semi-automated steps that are based on the organism's genome sequence and data from the literature. Experiments have to be conducted to get a more precise biomass reaction. One should keep at the back of one's mind that the accuracy of the resulting metabolic network model depends strongly on the available data.

Constraint-based approaches represent a metabolic network by a matrix whose entries correspond to the amounts, the so-called *stoichiometric values*, of a compound that is consumed or produced within a reaction. If a compound is consumed by a reaction, the respective matrix entry is negative. It is positive if it is produced. These approaches usually assume that the metabolic network is in steady-state, that is, the rate of change in the concentration of each compound in the network is zero. Thermodynamics dictate further constraints on the reversibility of the reactions. These stoichiometric and thermodynamic constraints define a convex polyhedral cone which contains all possible flux distributions in the metabolic network at steady state (Clarke, 1980; Gagneur and Klamt, 2004; Larhlimi and Bockmayr, 2009).

Flux balance analysis (FBA) aims to find a single optimal flux in the metabolic network. An often used optimality criterium is the maximization of biomass or ATP production. Schuetz et al. (2007) assessed if the predicted fluxes considering eleven different optimality criteria are consistent with ^{13}C-determined *in vivo* fluxes in *Escherichia coli* under different environmental conditions. It turned out that the nonlinear maximization of the ATP yield per unit of flux was the best objective function when *E. coli* grows on a rich glucose medium. In contrast, maximizing ATP or biomass yield were the best objective functions under scarce conditions (Schuetz et al., 2007). This demonstrates that the single flux obtained from FBA should not be taken as granted that the organism "applies" the same flux *in vivo*. First, the organism may optimize something different. Second, there are several fluxes attaining the same optimal value. To this purpose, Lee et al. (2000) provided an algorithm to enumerate all fluxes with minimal support that reach optimality.

To enumerate a finite basis of all possible fluxes at steady state in a metabolic network, several approaches are proposed in the literature, namely elementary flux modes (EFM) (Schuster and Hilgetag, 1994; Schuster et al., 2002a; Gagneur and Klamt, 2004; Wagner and Urbanczik, 2005), extreme pathways (EP) (Schilling et al., 2000), and minimal metabolic behaviors (MMB) together with the reversible metabolic space (RMS)(Larhlimi and Bockmayr, 2009). They provide either an inner (EFM, EP) or an outer (MMB, RMS) description of the above mentioned flux cone.

A flux mode is a set of reactions that operate at steady-state. Furthermore, the flux values on irreversible reactions are non-negative. A mode is then called *elementary* if it is non-decomposable, that is the removal of a reaction results in an unfeasible flux. To obtain any feasible flux in the cone, non-negative linear combinations of EFMs can be built. This approach is applied *e.g.* to identify new network-based metabolic pathways (Papin et al., 2003), to enumerate all pathways with an optimal yield (Schuster et al., 1999, 2002b), and to assess the network flexibility and robustness (Stelling et al., 2002).

Beside the conditions imposed on elementary modes, extreme pathways fulfil two extra conditions. First, reactions are classified as external or internal. Reversible internal reactions must be split into one reaction for the forward and one reaction for the backward direction; the flux on each internal reaction must be non-negative. Second, an extreme pathway cannot be obtained through non-negative linear combinations of other extreme pathways. Extreme pathways are a subset of elementary modes, and coincide with the latter if all exchange reactions between the environment and the cell are irreversible (Klamt and Stelling, 2003).

Larhlimi and Bockmayr (2009) provide an outer description, that is based on inequality

constraints on irreversible reactions, of the steady-state flux cone. This approach consists of two elements: minimal metabolic behaviors (MMBs) and the reversible metabolic space (RMS). An MMB corresponds to a characteristic set of irreversible reactions that is associated to a minimal proper face of the cone. The RMS is the lineality space of the cone. Any flux vector of the steady-state flux cone can be built through linear combinations of the MMBs and the RMS. The authors show that the size of their representation, calculated as the sum of the number of MMBs and the dimension of the RMS, is smaller than the number of elementary modes or extreme pathways (Larhlimi and Bockmayr, 2009). Since these approaches are computationally expensive, their application is restricted to subnetworks or to the production of certain compounds only (Terzer and Stelling, 2008; Pey et al., 2014). To our knowledge, there exists only one approach that enumerates *all* elementary flux modes of a relatively small genome-scale metabolic network (Hunt et al., 2014). Furthermore, a huge number of solutions (several millions) is generated even for small networks making their analysis difficult.
The approach of Figueiredo *et al.* tries to circumvent these shortcomings and consists in enumerating the k-shortest elementary modes solving iteratively mixed integer linear programs (de Figueiredo et al., 2009). The authors enumerated the ten shortest EFMs in the genome-scale networks of *E.coli* and *C.glutamicum* in a reasonable time.
Structurally important parts of a metabolic network can be detected with *minimal cut sets* (Klamt and Gilles, 2004). A minimal cut set (MCS) is a set of reactions whose inactivation disables a desired function. MCSs are minimal hitting sets of elementary modes. Klamt and Gilles (2004) thus enumerated first the elementary modes that produce biomass and computed the minimal hitting sets. Ballerstein *et al.* demonstrated that MCSs can be computed directly as elementary modes in the dual network (Ballerstein et al., 2012). This enables the enumeration of the k-shortest minimal cut sets of the flux cone (von Kamp and Klamt, 2014) which is a similar approach to the above mentioned k-shortest EFM enumeration.
Another manner to tackle genome-scale metabolic networks is to restrict the analysis to fluxes that reach an optimality criterium as in FBA. It turns out that in this case several reactions have a fixed flux in all elementary modes. The remaining reactions, those with a variable flux, can be grouped into subnetworks (Kelk et al., 2012). Later, Müller and Bockmayr (2013) showed that these subnetworks, which the authors called *modules*, can be computed without first enumerating all elementary modes. These modules can be analyzed independently, *e.g.* by enumeration of EFMs inside a module. An efficient algorithm, based on matroid theory, for the detection of these modules is provided in Müller et al. (2014).

In Chapter 3, we enumerated minimal stoichiometric precursor sets that enable the production of a set of targets. A subset of sources is a stoichiometric precursor set if there is a *stoichiometric factory* from this set to the target set. A natural goal is to enumerate all minimal stoichiometric factories for a given minimal precursor set. This chapter is organized as follows. We first recall some definitions from the previous chapter. The relationship between minimal stoichiometric factories and elementary flux modes will be discussed afterwards. We then provide different algorithms that can be applied to enumerate minimal stoichiometric factories at steady state or allowing for accumulation. These algorithms take advantage of different structural properties of the metabolic network. An application to the genome-scale model of *E.coli* will be discussed in Section 4.4.

4.2 Definitions and Properties

We use the same notation as in Chapter 3, *e.g.* the pair $\mathcal{N} = (\mathcal{C}, \mathcal{R})$ characterizes a metabolic network with its chemical compounds and reactions. A set of sources (precursors) $\mathcal{X} \subseteq \mathcal{C}$ corresponds to the chemical compounds that are present in the medium. We assume that the

sources are not produced by any reaction. In Chapter 3, we showed how to enumerate all minimal stoichiometric precursor sets that enable the production of the set of targets $\mathcal{T} \subseteq \mathcal{C}$. A subset of sources $X \subseteq \mathcal{X}$ is a precursor set if there exists a *stoichiometric factory*, i.e. a set of reactions that consumes all sources in X and produces all targets in $\mathcal{T}$ in a positive amount. Furthermore, all substrates and products that are involved in the reactions of such a factory must have a non-negative net production. We recall some notation of the previous chapter: the flux vector $v \in \mathbb{R}^{|\mathcal{R}|}$ denotes the flux of every reaction in the network per time unit, and $Sv \in \mathbb{R}^{|\mathcal{C}|}$ is the vector of net production of all compounds in the network for the flux v. Furthermore, $(Sv)_A$ specifies the net production of the compounds in a set A. A stoichiometric factory can then be defined as:

Definition 17. *A **stoichiometric factory** (S-factory) from $X \subseteq \mathcal{X}$ to $T \subseteq \mathcal{T}$ is a set $F \subseteq \mathcal{R}$, such that there exists a flux vector $v \geq 0$ satisfying:*

1. $v_i \begin{cases} > 0 & i \in F \\ = 0 & \text{otherwise,} \end{cases}$
2. $(Sv)_{\mathcal{C} \setminus X} \geq 0$,
3. $(Sv)_T > 0$.

A S-factory from X to T is minimal *if it does not contain any other S-factory from X to T.*

In Definition 17, we explicitly allow for an accumulation of the compounds that do not belong to the precursor set X, that is these compounds are not in steady state. The definition can however be easily adapted to the steady state assumption. As in elementary mode analysis, one needs to distinguish *internal* from *external* compounds, denoted by I and E respectively. The steady state constraint is required only on the internal compounds. In the context of stoichiometric factories, at least the compounds of a minimal precursor set X and the target set $\mathcal{T}$ are part of the external compounds, that is $(X \cup \mathcal{T}) \subseteq E$. To account for the fact that there is no (steady state) constraint on external compounds, we need to transform the network in the following way: for every external compound $e \in E$, we add an export reaction $r_e : e \to \emptyset$ to the network. Furthermore, consider the case when the user declares an external compound e as a source and e is produced by a reaction. We then create a dummy compound e' and a reaction that consumes one unit of e' and produces one unit of e. The compound e is replaced by e' in the set of sources $\mathcal{X}$. These network transformations are illustrated in Figure 4.2. In Figure 4.2a, the compound A is declared as external compound (superscript e). In the transformed network on the right-hand side, we add an export reaction r_{ex} that consumes the compound A and produces the empty set. In Figure 4.2b, the compound A is declared as source (superscript $*$) and external compound (superscript e), but A is produced by a reaction. We transform the network in the following way. We add an export reaction r_{ex} that consumes the compound A. We add furthermore a dummy compound A' and a reaction that consumes A' and produces A. The compound A' replaces the compound A in the set of sources $\mathcal{X}$. Thus, the source A' is not produced by any reaction as desired.
After the network transformation, we can revoke the distinction between internal and external compounds and define a stoichiometric factory at steady state similar to Definition 17 as follows:

Definition 18. *A **stoichiometric factory** at **steady state** from $X \subseteq \mathcal{X}$ to $T \subseteq \mathcal{T}$ is a set $F \subseteq \mathcal{R}$, such that there exists a flux vector $v \geq 0$ satisfying:*

1. $v_i \begin{cases} > 0 & i \in F \\ = 0 & \text{otherwise,} \end{cases}$

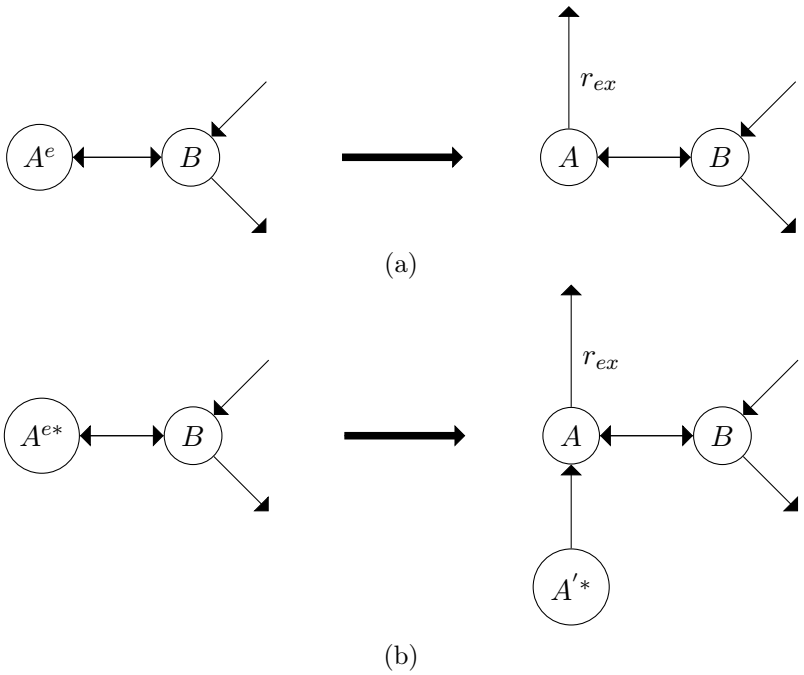


Figure 4.2: Two network transformations are shown. The original network is shown on the left, the transformed network is shown on the right. The classification of internal and external compounds is revoked in the transformed network. (a) The compound A is declared as external compound (superscript e). In the transformed network, we add an export reaction r_{ex} that consumes the compound A. (b) The compound A is declared as source (superscript $*$) and external compound (superscript e), but A is produced by a reaction. We transform the network in the following way. We add an export reaction r_{ex} that consumes the compound A. We add a dummy compound A' and a reaction that consumes A' and produces A. The compound A' replaces the compound A in the set of sources $\mathcal{X}$.

2. $(Sv)_{\mathcal{C}\setminus X} = 0$,
3. $(Sv)_T > 0$.

A S-factory at steady state from X to T is minimal *if it does not contain any other S-factory at steady state from X to T.*

For the remaining, a stoichiometric factory (S-factory) and precursor set (SPS) refer to the case where accumulation is allowed. We will specify explicitly when the steady state assumption for precursor sets and factories holds.

In elementary flux mode analysis (Schuster and Hilgetag, 1994; Schuster et al., 2002a; Gagneur and Klamt, 2004; Wagner and Urbanczik, 2005), the set of compounds is divided into internal and external compounds, where the latter correspond to compounds that are connected through a reaction to the environment. These compounds either are supplied by the environment, or represent waste products of the cell that are exported to the environment. Reversible reactions can be split into a forward and a backward reaction (Terzer and Stelling, 2008). This implies that each reaction must have a non-negative flux value. A minimal flux mode can then be defined as follows:

Definition 19. *Given a metabolic network $\mathcal{N} = (\mathcal{C}, \mathcal{R})$, with a set of internal compounds $I \subseteq \mathcal{C}$, a* ***flux mode*** *is a set $F \subseteq \mathcal{R}$ such that there exists a flux vector $v \geq 0$ satisfying:*

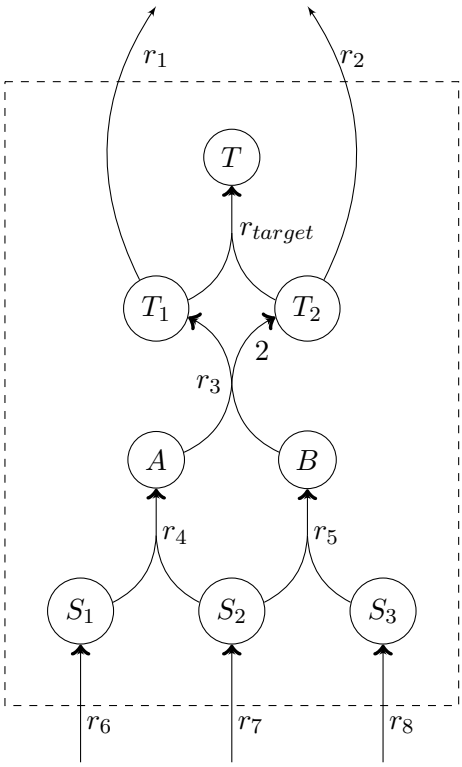


Figure 4.3: All compounds except T are considered as internal compounds. Here, the single steady state stoichiometric factory from $X = \{S_1, S_2, S_3\}$ to $\mathcal{T} = \{T\}$ contains all reactions except r_1, and coincides with the elementary mode that contains r_{target}. There is a second elementary mode that contains all reactions except r_{target}.

1. $v_i \begin{cases} > 0 & i \in F \\ = 0 & otherwise, \end{cases}$
2. $(Sv)_I = 0$.

A flux mode is **elementary** *if it does not contain any other flux mode.*

To be able to compare the concepts of minimal factories and elementary modes, we transform the network and the set of targets in the following way:

Definition 20. *Given a metabolic network $\mathcal{N} = (\mathcal{C}, \mathcal{R})$ and a set of target compounds $\mathcal{T} = \{T_1, \ldots, T_n\}$, we define the* **transformed problem** *as the network $\mathcal{N}' = (\mathcal{C}', \mathcal{R}')$ and the set of targets $\mathcal{T}'$ such that:*

1. $\mathcal{C}' = \mathcal{C} \cup \{T\}$,
2. $\mathcal{R}' = \mathcal{R} \cup \{r_{target}\}$, *with* $Subs(r_{target}) = \mathcal{T}$, $Prod(r_{target}) = \{T\}$,
3. $\mathcal{T}' = \{T\}$

Following Definition 20, we add an artificial reaction r_{target} that consumes all compounds of the target set $\mathcal{T} = \{T_1, \ldots, T_n\}$ and produces a dummy target compound T (stoichiometric values are equal to one). We then replace the compounds $T_1, \ldots, T_n$ by T in the target set $\mathcal{T}$. A transformed network is depicted in Figure 4.3. Here, the single minimal steady state

stoichiometric factory from $X = \{S_1, S_2, S_3\}$ to $\mathcal{T} = \{T\}$ coincides with the elementary mode that contains r_{target}.
Given a transformed network $\mathcal{N}' = (\mathcal{C}', \mathcal{R}')$, a transformed set of targets $\mathcal{T}'$ (according to Definition 20), and a minimal steady state stoichiometric precursor set $X \subseteq \mathcal{X}$ for the target set $\mathcal{T}'$ in the transformed network $\mathcal{N}'$, we can say that, under the condition that the compounds in the target set $\mathcal{T}'$ are considered as external compounds, an elementary mode E that consumes all compounds in X and that produces the target set $\mathcal{T}'$ is also a minimal stoichiometric factory at steady state from X to $\mathcal{T}'$. By definition, elementary modes and minimal stoichiometric factories (at steady state) are minimal sets of reactions, meaning that when a reaction is removed from them, then they are no longer a mode or a stoichiometric factory. Definition 18 and definition 19 require a positive flux on each reaction in a mode or a stoichiometric factory at steady state. If one defines the internal compounds as the set of network compounds minus the set of sources X, that is $I = \mathcal{C} \setminus X$, then the second condition of both definitions 18 and 19 are identical. The last condition on a stoichiometric factory ($(Sv)_{\mathcal{T}'} > 0$) has no explicit correspondent condition in the definition of a mode. However, we require that the elementary mode E produces the set of targets $\mathcal{T}'$ which implies that the reaction r_{target} is part of E as it is the only reaction that produces $\mathcal{T}'$. The constraint that involves putting a positive flux on the reactions results in $(Sv)_{\mathcal{T}'} > 0$ (note that $\mathcal{T}'$ is not consumed by a reaction).

4.3 Enumeration algorithms

In this section, we introduce algorithms that can be applied for the enumeration of (steady state) minimal factories from a *single* minimal precursor set $X \subseteq \mathcal{X}$ to $\mathcal{T}$. We will show that all (steady state) factories share a common structure, they all share some equal reactions or compounds. A structural analysis, based on minimal cut sets, that allows the enumeration of a biologically meaningful subset of minimal (steady state) factories is proposed.
First, we show that we can prune the metabolic network. Then, we describe different concepts of structural analysis that can be applied to factories and steady state factories. Finally, we provide algorithms that can be applied for the enumeration of minimal (steady state) factories.

4.3.1 Pruning

Again, as this is important for understanding the pruning steps, we consider minimal factories from a *single* minimal precursor set $X \in \mathcal{X}$ to the target set $\mathcal{T}$.

There are three pruning steps, two of which can be applied to both kinds of stoichiometric factories. The third pruning step is of importance only for factories allowing for an accumulation.

Removal of topological sources that are not in X

Given a minimal precursor set X of $\mathcal{T}$, all factories from X to $\mathcal{T}$ consume *only* the topological sources $X \subseteq \mathcal{X}$. Hence, topological sources (compounds that are not produced by the network) that are not in X can be removed from the network by removing the reactions consuming these sources. Starting with a network $\mathcal{N}_0 = \mathcal{N}$ and a set of topological sources $X_0 = \mathcal{X}$, we remove from N_0 all reactions that consume topological sources that are in X_0 but not in X to obtain $\mathcal{N}_1$. The removal of reactions may create a new set of topological sources, namely X_1. We repeat this process until $X_i = X$, that is all topological sources of the network are contained in X.

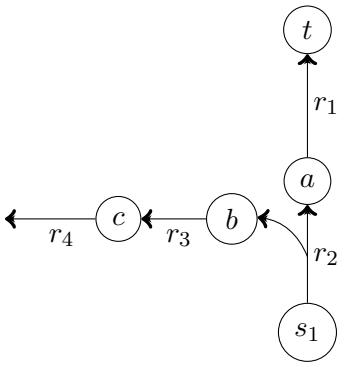


Figure 4.4: A metabolic network to illustrate that Lemma 1 does not hold for steady state factories.

Removal of reactions

Beside the reactions that were removed from the network in the previous pruning step, we can remove reactions that cannot have a positive flux. This can easily be verified by solving the following linear program for every reaction $r \in \mathcal{R}$:

$$\begin{aligned} &(Sv)_{\mathcal{C}\setminus X} \diamond 0 \\ &(Sv)_{\mathcal{T}} \geq \epsilon_1 \\ &v_r \geq \epsilon_2 \\ &v_{r_{rev}} = 0 \\ &0 \leq v_i \leq U, \quad i \in \mathcal{R} \end{aligned} \tag{4.1}$$

The symbol $\diamond$ in the first line can be replaced by a $\geq$ or $=$ sign depending on whether an accumulation of compounds is allowed or not. The target set $\mathcal{T}$ must be produced in a positive amount ϵ_1. We further require a positive flux on the reaction r. If the latter is reversible, we set the flux on the reverse direction ($v_{r_{rev}}$) to zero. This constraint is added to avoid a flux on the forward and the backward direction of a reversible reaction. As we split reversible reactions into a forward and a backward reaction, we require a non-negative flux. If there is no such flux v, then r can be removed from the network.

Path to a target of $\mathcal{T}$

Lemma 1 in chapter 3 states that for every reaction r in a factory H, there is at least one path in H from one of the products of r to a target compound in $\mathcal{T}$. The network in Figure 4.4 illustrates why this does not hold for steady state factories. The factory from s_1 to t comprises the reactions r_1 and r_2 and the compound b accumulates. In the steady state factory from s_1 to t, the compound b must be consumed to achieve a zero net production. Thus the steady state factory comprises all four reactions. However, there is no path from a product of the reactions r_3 and r_4 to the target t.
Lemma 1 provides the justification to remove a reaction r if there is no path from any product of r to a target. The computation of all compounds that have a path to some target can be done using the recursive algorithm 1. In fact, we check which compounds can be reached from a target taking the network reactions in their opposite direction. The parameters of the procedure are: (i) a compound c, and (ii) a set, denoted by C_r, of compounds that were already reached. The latter set is returned if the compound c was already reached in an earlier recursion call. If not, the compound$c$ is added to C_r and the procedure explores the substrates of the reactions that produce c (Reac_p(c)). This recursive procedure is called until the base

case is reached, that is when the compound $c \in C_r$. Calling the procedure for a target $t \in \mathcal{T}$ as $getReachableCompounds(t,\ \emptyset)$ results in the set (C_{r_t}) of compounds that have a path to the target t. Thus, a reaction whose products are in no C_{r_t}, with $t \in \mathcal{T}$, can be removed from the network.

Algorithm 1: $getReachableCompounds(\mathcal{N},\ c\ C_r)$

Input : The network $\mathcal{N} = (\mathcal{C}, \mathcal{R})$, a compound $c \in \mathcal{C}$, and the set of reachable compounds C_r.
Output : The set of reachable compounds C_r

```
if c ∈ C_r then
    return C_r;
C_r ← C_r ∪ {c};
foreach r ∈ Reac_p(c) do
    foreach c' ∈ Subs(r) do
        C_r ← C_r ∪ getReachableCompounds(N, c', C_r);
return C_r;
```

4.3.2 Structural analysis

We present three interrelated concepts of structural analysis of the factories from a given minimal precursor set X to $\mathcal{T}$. All of them can be applied to both kinds of factories.

Essential reactions

Reactions that are in every factory are called *essential*. Removing such a reaction from the network would prevent the production of the set of targets from X. Thus, each such reaction corresponds to a minimal cut set (Klamt and Gilles, 2004) in the pruned network. Verifying if a reaction $r \in \mathcal{R}$ is essential can be done by solving the following LP:

$$\begin{aligned} &(Sv)_{\mathcal{C}\setminus X} \diamond 0 \\ &(Sv)_{\mathcal{T}} \geq \epsilon_1 \\ &v_r = 0 \\ &0 \leq v_i \leq U, \quad i \in \mathcal{R}, \end{aligned} \tag{4.2}$$

where the symbol $\diamond$ in the first line can be replaced by a $\geq$ or $=$ sign depending on whether an accumulation of compounds is allowed or not. The target set $\mathcal{T}$ must be produced in a positive amount ϵ_1. The reaction r is blocked setting the flux v_r to zero. As we split reversible reactions into a forward and a backward reaction, we require a non-negative flux.
If there is no such flux v, then r is an *essential reaction*.

Essential compounds

An *essential compound* is a compound that must be consumed in every factory from X to $\mathcal{T}$. The verification if a compound $c \in \mathcal{C}$ is essential is done by solving the LP:

$$\begin{aligned} &(Sv)_{\mathcal{C}\setminus X} \diamond 0 \\ &(Sv)_{\mathcal{T}} \geq \epsilon_1 \\ &v_r = 0 \qquad\quad r \in \mathrm{Reac_s(c)} \\ &0 \leq v_i \leq U, \quad i \in \mathcal{R}, \end{aligned} \tag{4.3}$$

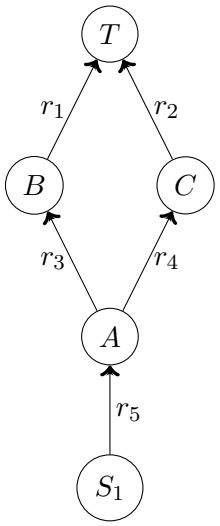


Figure 4.5: A metabolic network to illustrate that not every essential compound is consumed by an essential reaction.

where $\text{Reac}_s(c)$ denotes the reactions that consume the compound c. Thus, we block these reactions. If there is no such flux v, then c is an *essential compound*. The substrates of essential reactions are essential compounds by definition. However, the definition of essential compounds is not restricted to this case. In Figure 4.5, the compounds S_1 and A are essential even though the latter is not consumed by an essential reaction. Here, there are two minimal factories, namely $\{r_5,\ r_3,\ r_1\}$ and $\{r_5,\ r_4,\ r_2\}$. Thus, the reaction r_5 (that produces the compound A) is essential because it is present in all factories. The concept of essential compounds is not new in metabolic network analysis. Essential compounds were already detected in the same manner as in Problem (4.3) but maximizing the growth rate at the same time (Kim et al., 2007; Chung and Lee, 2009). A compound must be producable in a positive amount to be essential according to Imieliński et al. (2005).

Minimal Cut sets

In the previous sections, we provided the concepts of essential reactions and compounds. We demonstrated that an essential compound must not necessarily be consumed by an essential reaction. It also holds that an essential compound need not to be produced by an essential reaction. However, it is true that in every factory, each essential compound c is consumed and produced by at least one reaction. The source compounds in X build an exception as they are not produced by a reaction. They are however essential in every factory from X to $\mathcal{T}$ and must thus be consumed.
We ask whether there are reactions, that may produce or consume an essential compound, but that do not take part of a minimal factory from X to $\mathcal{T}$. Alternatively, we ask whether it is possible to enumerate a subset of (biologically meaningful) minimal factories from X to $\mathcal{T}$ if some of the reactions that consume and produce essential compounds were removed. This leads us to the enumeration of minimal reaction cut sets (Klamt and Gilles, 2004). For a set CS of reactions, starting with singletons, pairs, triples, etc., Klamt and Gilles (2004) propose to check if CS hits all elementary modes. This however requires the computation of all elementary modes beforehand. Acuña et al. (2009) showed that it is sufficient to solve an LP similar to the one in (4.2) setting the flux of all reactions in CS to zero. If there is no flux v then CS is a cut set. This facilitates checking whether a set of reactions is a

cut set or not, however it does not help to enumerate all minimal cut sets. In the paper of de Figueiredo et al. (2009) an algorithm to enumerate the k-shortest elementary modes is provided. Ballerstein et al. (2012) showed how to enumerate all minimal cut sets through the enumeration of elementary modes in the dual network. Combining the latter two approaches is the main idea of von Kamp and Klamt (2014), who provided an algorithm to enumerate the k-smallest cut sets through the enumeration of the k-shortest elementary modes in the dual network. This approach can be used theoretically to enumerate all minimal cut sets by setting k to a high value. In practice, the authors enumerated the 5-shortest minimal cut sets in an *E.coli* genome-scale model. It should be noted that the parameter k has a different interpretation in the approaches of de Figueiredo et al. (2009) and of von Kamp and Klamt (2014). In the former, the authors enumerate k elementary modes. von Kamp and Klamt (2014) enumerate all k-shortest (cardinality) elementary modes in the dual network. So, *e.g.* for $k = 1$, de Figueiredo et al. (2009) enumerate one single elementary mode (the shortest one of length l), and von Kamp and Klamt (2014) enumerate all elementary modes of length l in the dual network. The k-shortest elementary modes in the dual network corresponds to the k-smallest minimal cut sets in the primal network.
In the context of factories from X to $\mathcal{T}$, we take advantage of the essential compounds. We enumerate separately, for each essential compound c, the minimal reaction cut sets among the reactions that produce c and the reactions that consume c, respectively. Therefore, we make use of the method of Ballerstein et al. (2012) combined with the MILP approach used for the enumeration of minimal precursor sets. First, we explain the method of Ballerstein et al. (2012). We then describe our method for the enumeration of elementary modes in the dual network where the support of the elementary modes is restricted to a subset of reactions. First, let us add to the network: (i) a dummy compound $Target$, and (ii) an artificial reaction r_{Target} that consumes all $T \in \mathcal{T}$ and that produces the compound $Target$. Let S denote the $m \times n$ stoichiometric matrix that is associated to a metabolic network $\mathcal{N} = (\mathcal{C}, \mathcal{R})$ with $m = |\mathcal{C}|$ compounds and $n = |\mathcal{R}|$ reactions. Furthermore, t is a $(n \times 1)$ vector where positive entries represent target reactions that must be blocked, *e.g.* the reaction r_{Target} in our case. The primal network can then be represented by the following system of inequalities:

$$\begin{aligned} & Sv = 0 \\ & t^{\mathrm{T}} v \geq 1 \\ & v_i \geq 0, \quad i \in \mathcal{R}_{irrev}, \end{aligned} \tag{4.4}$$

where R_{irrev} represents the set of irreversible reactions. The transposed vector t is denoted by t^{T}. Ballerstein et al. (2012) show that, based on the *Farkas Lemma*, the enumeration of minimal reaction cut sets in the (primal) network is equivalent to the enumeration of elementary modes in the dual network. The dual network can be described as follows:

$$\begin{aligned} & S_{dual} r_{dual} := (S^{\mathrm{T}} \; I \; -\bar{I}_{irrev} \; -t) \begin{pmatrix} u \\ v \\ z \\ w \end{pmatrix} = 0 \\ & u \in \mathbb{R}^m, v \in \mathbb{R}^n, z \in \mathbb{R}^{|\mathcal{R}_{irrev}|}, w \in \mathbb{R} \\ & z \geq 0, w \geq 0, \end{aligned} \tag{4.5}$$

where the stoichiometric matrix is transposed. Thus, the reactions of the primal network become compounds of the dual network and the compounds of the primal network become reactions in the dual network. Furthermore, the $n \times n$ identity matrix is denoted by I. The $n \times |\mathcal{R}_{irrev}|$ identity matrix of the irreversible reactions is denoted by $\bar{I}_{irrev}$. The matrix of the dual network has a dimension of $n \times (m + n + |\mathcal{R}_{irrev}| + 1)$. The variables u and v are

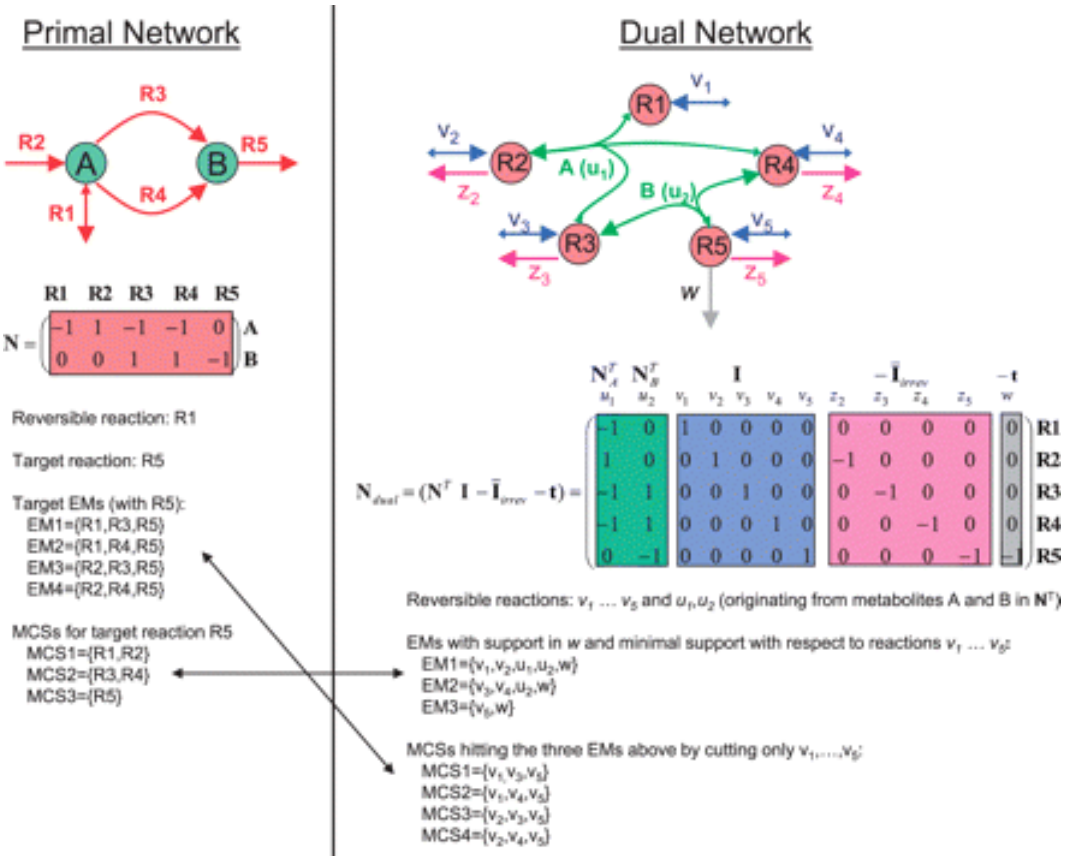


Figure 4.6: The conversion from the primal to the dual network (from (Ballerstein et al., 2012)).

reversible reactions in the dual, where the former correspond to the compounds of the primal network. The import/export reaction variables v are of special interest as the support on these variables in the elementary modes of the dual network determine the cut sets in the primal network. The variables z and w are irreversible export reactions in the dual network and originate from irreversible reactions and the target reaction, respectively. A picture is worth a thousand words, and Figure 4.6 thus illustrates that the conversion from the primal to the dual network is an easy task. Here, the primal network on the left consists in two compounds and five reactions (only $R1$ is reversible). The reaction $R5$ is the target reaction that one would like to prevent. The 2×5 primal stoichiometric matrix is given in the red box on the left. The dual network is depicted on the right and one can see that the reactions of the primal network become compounds in the dual network, and the compounds of the primal network become reactions in the dual network. The dual matrix representation shows this as well: the transposed stoichiometric matrix of the primal network has dimension 5×2 (green box). The columns (the reaction variables u) correspond to the compounds of the primal network and the rows correspond to the reactions of the primal network. The blue part of the matrix represents the reversible exchange variables v. The pink part corresponds to the identity matrix of the irreversible reactions. Note that the first row herein has only zero entries as $R1$ is reversible in the primal network. Lastly, the irreversible export variable w is depicted in grey.

The goal is to enumerate elementary modes in the dual network that have (i) a positive flux on w, and (ii) minimal support on the exchange reaction variables v. In the dual net-

work of Figure 4.6, there are three elementary modes with these properties, *e.g.* $EM1 = \{v_1,\ v_2,\ u_1,\ u_2,\ w\}$, where v_2 imports $R2$, and u_1 consumes $R2$ to produce $R1$, $R3$, and $R4$. The former is exported through v_1. The compounds $R3$ and $R4$ are consumed by u_2 to produce $R5$ which in turn is exported through the reaction w. This elementary mode in the dual network corresponds to the minimal cut set $\{R1,\ R2\}$ in the primal network, because the v variables in the dual network represents the reactions in the primal network.
Enumeration of all elementary modes is a hard problem. von Kamp and Klamt (2014) therefore concentrated on the enumeration of the k-shortest elementary modes in the dual network to obtain the k-smallest cut sets in the primal network. As already mentioned above, reaction i belongs to a minimal cut set in the primal network if $v_i \neq 0$ in an elementary mode in the dual network. As the variables v in the dual network can be negative and positive, the authors transformed the dual network by splitting the reactions that are associated to the variables v into forward and backward reactions ($vp,\ vn \in \mathbb{R}^n$). A binary variable is associated to each reaction direction, that is, a binary variable zp_i for the forward reaction of vp_i, and a binary variable zn_i for the backward reaction of vn_i. Each binary variable captures through a constraint if the respective flux value is positive:

$$zp_i = 0 \leftrightarrow vp_i \leq 0,\ zn_i = 0 \leftrightarrow vn_i \leq 0. \tag{4.6}$$

To avoid that the forward and backward reactions are used (both with a positive flux value) in the same elementary mode in the dual network, the authors add the following constraint:

$$zp_i + zn_i \leq 1. \tag{4.7}$$

It is however not necessary to do this transformation. We can associate two binary variables (zp_i and zn_i) to each variable v_i to capture when the latter has a flux value different from zero. A constraint is built for each binary variable. The variable zp_i, that is associated to a positive flux value of v_i, is equal to one if v_i is greater than 0, and equal to zero otherwise. The constraint can thus be formulated as follows:

$$zp_i = 0 \leftrightarrow v_i \leq 0. \tag{4.8}$$

Similarly, the constraint to capture a negative flux value v_i is given by:

$$zn_i = 0 \leftrightarrow v_i \geq 0. \tag{4.9}$$

Thus, if the flux value of v_i is different from zero, either of the binary variables is equal to one. If v_i is greater (smaller) than zero then zp_i (zn_i) is equal to one.
We are interested in the enumeration of particular elementary modes that have a positive flux value in w (the target reaction) and a minimal support on the variables v. To enumerate these elementary modes in the dual network, we apply an approach similar to the one that was used for the enumeration of minimal precursor sets: we solve recursively mixed integer linear programming (MILP) problems. The solutions that are found in previous iterations are excluded from the feasible set of the current MILP problem. This MILP approach is commonly used to find minimal subsets (Lee et al., 2000; Larhlimi and Bockmayr, 2007; de Figueiredo et al., 2009; von Kamp and Klamt, 2014). As for the enumeration of minimal precursor sets, for every source $x \in X$, we add to $\mathcal{R}$ of the primal network a *source-pool* reaction that produces one unit of x from the empty set. Furthermore we add a target reaction r_{target} that consumes one unit of all targets $t \in \mathcal{T}$ and produces one unit of a dummy compound $target$. The transformed network is denoted by $\overline{\mathcal{N}} = (\overline{\mathcal{C}}, \overline{R})$, with its associated stoichiometric matrix $\overline{S}$. We transform the primal network $\overline{\mathcal{N}} = (\overline{\mathcal{C}}, \overline{R})$ into its dual network $\mathcal{N}_{dual} = (\mathcal{C}_{dual}, \mathcal{R}_{dual})$ as in Equation (4.5) (column t corresponds to the reaction r_{target}). The dual stoichiometric

matrix is denoted by S_{dual}, and the flux vector r_{dual} is composed of the $u,\ v,\ z$, and w vectors. In the sequel, L and U are a lower and an upper bound for the flux values, ϵ is a small positive real number.
Given the dual network $N_{dual} = (\mathcal{C}_{dual}, \mathcal{R}_{dual})$, a stoichiometric matrix of the dual network S_{dual}, and the target reaction w in the dual network, we provide the MILP to find a first minimal solution of minimum cardinality:

$$\begin{aligned}
\min f = \quad & \sum_{j=1}^{|v|} zp_j + zn_j \\
s.t \quad & S_{dual} r_{dual} = 0, \\
& zp_j = 0 \leftrightarrow v_j \leq 0, && \forall j \in v \\
& zn_j = 0 \leftrightarrow v_j \geq 0, && \forall j \in v \\
& zn_j, zp_j \in \{0,1\}, && \forall j \in v \\
& L \leq u_i \leq U, && \forall i \in u \\
& L \leq v_i \leq U, && \forall i \in v \\
& z_i \geq 0, && \forall i \in z \\
& w \geq \epsilon,
\end{aligned} \tag{4.10}$$

where the constraints in the second and third line capture if there is a flux different from zero on the variables v. They can be formulated as indicator constraints in CPLEX. The lower and the upper bounds for the flux values are shown in the last four lines. As we are interested in elementary modes in the dual network that contain the target reaction w in their support, we set $w \geq \epsilon$.
Let the triple (r^*_{dual}, zp^*, zn^*) be an optimal solution of Problem (4.10). Furthermore, let I_{v^*} be the support of v in this optimal solution. The set of reactions of the primal network that corresponds to I_{v^*} constitutes a minimal cut set in the primal network.
To enumerate all elementary modes in the dual network (respectively all minimal cut sets in the primal network), we introduce the following constraint:

$$\sum_{j \in I_{v^*}} zp_j + zn_j \leq |I_{v^*}| - 1. \tag{4.11}$$

The constraint (4.11) excludes the optimal solution (r^*_{dual}, zp^*, zn^*) and all solutions that contain I_{v^*}. We add a constraint in the form of (4.11) to Problem (4.10) to obtain a new instance of the problem. This is done recursively to enumerate all elementary modes in the dual network. If there is no feasible solution, then all elementary modes, and thus all minimal cut sets are found.
Up to this point we considered the steady state condition. The dual network differs slightly if we allow for an accumulation of the compounds.

$$\begin{aligned}
& S_{dual} r_{dual} := (-S^{\mathrm{T}} \ I \ -\bar{I}_{irrev} \ -t) \begin{pmatrix} u \\ v \\ z \\ w \end{pmatrix} = 0 \\
& u \in \mathbb{R}^m, v \in \mathbb{R}^n, z \in \mathbb{R}^{|\mathcal{R}_{irrev}|}, w \in \mathbb{R} \\
& u \geq 0, z \geq 0, w \geq 0,
\end{aligned} \tag{4.12}$$

The enumeration of elementary modes in such a dual network S_{dual} can be done as under the steady state assumption solving recursively the Problem (4.10). The only difference is that the fluxes on u are constrained to be non-negative.

Here, we concentrate on minimal reaction cut sets for the production of the target, taking into account, as variables v, the subnetworks which are given by the reactions that either consume or produce an essential compound. Thus, we determine which reactions that consume *or* produce a single essential compound $i \in \mathcal{C}_{ess}$ must be at least blocked to prevent the production of the target in the complete network. For each essential compound $i \in \mathcal{C}_{ess}$, there are two subnetworks:

1. $SUB_{ci} = \text{Reac}_\text{s}(\text{i})$ (reactions that consume i),
2. $SUB_{pi} = \text{Reac}_\text{p}(\text{i})$ (reactions that produce i).

Computing the minimal reaction cut sets for the production of the target, considering only the reactions in such a subnetwork of size l, requires to introduce into the dual network the variables $v_1, \ldots, v_l$ (instead of $v_1, \ldots, v_n$ if the complete network is considered). Let CS_{ci} and CS_{pi} denote the collection of minimal cut set taking into account the reactions in SUB_{ci} and SUB_{pi}, respectively. Furthermore, let $\mathcal{R}_{ci}$ and $\mathcal{R}_{pi}$ denote the union of the reactions that are in the minimal cut sets CS_{ci} and CS_{pi}, respectively. To simplify the upcoming definition, we declare $\mathcal{R}_{cpEss} \subseteq \mathcal{R}$ to be the union of all reactions that consume or produce an essential compound. Then, Y denotes the set of reactions, consuming or producing an essential compound, that are not present in any minimal cut set over all subnetworks, that is $Y = \mathcal{R}_{cpESS} \setminus \{r \in \mathcal{R}_{cpEss} | r \in (\mathcal{R}_{ci} \cup \mathcal{R}_{pi}), \exists i \in C_{ess}\}$.
We want to find a set $Y' \subseteq Y$ such that, even when all reactions of Y' are removed from the network, there is at least one factory from X to $\mathcal{T}$. All reactions in $\mathcal{R}_{cpEss}$ that are not in Y are part of at least one minimal cut set and are thus in at least one minimal factory. Removing the reactions that are in the set Y reduces the number of factories as can be seen in Figure 4.7. Here, the reactions r_7 and r_8 are not part of any minimal cut set of a subnetwork related to an essential compound. If one removes both reactions, the number of minimal factories decreases from four to two, that is $\{r_1, r_2, r_4, r_6\}$ and $\{r_1, r_3, r_5, r_6\}$.
However, it is true that the removal of all reactions in Y may prevent the production of the target, as depicted in Figure 4.8. Here, there are two factories from S_1 to T: $\{r_1, r_2, r_3, r_5, r_7\}$ and $\{r_1, r_2, r_4, r_5, r_6\}$. The compounds S_1, A, B, and E are essential compounds in all factories . All reactions consume or produce an essential compound, and all reactions except r_6 and r_7 are part of a minimal cut set, *e.g.* $\{r_1\}$ is a minimal cut set of the subnetwork of the reactions that consume the compound B, $\{r_2\}$ is a minimal cut set of the reactions that consume A, $\{r_5\}$ is a minimal cut set of the reactions that consume S_1, and $\{r_3,\ r_4\}$ is a minimal cut set of the reactions that produce E. The reactions r_6 and r_7 do not take part in any cut set of the subnetworks related to the essential compounds. However, both reactions together constitute a minimal cut set of the factories from S_1 to $\mathcal{T}$. If both reactions are removed, then the compound E, and thus T can not be produced anymore.
In the case where the removal of all reactions of Y prevents the production of $\mathcal{T}$ from X, we propose the following approach. We enumerate minimal cut sets among the reactions in Y using the MILP approach from above. The collection of minimal cut sets among Y is denoted by MCS. The collection $\mathcal{R}_{MCS}$ denotes the union of reactions in the minimal cut sets MCS. We enumerate minimal cut sets of increasing cardinality. Before adding the minimal cut set mcs_i of the iteration i to MCS, and the reactions of mcs_i to $\mathcal{R}_{MCS}$, we check if the cardinality of mcs_i has increased compared to the minimal cut set of the previous iteration: mcs_{i-1} (for $i > 0$). If this is the case, we build a set of blocked reactions Y' that consist in the reactions of Y minus the reactions in MCS. To test if there is a factory from X to $\mathcal{T}$, when the flux of the reactions in Y' are set to zero, can be done with the help of the LP formulation (4.3). If there is a flux solution v to this LP, then we stop enumerating minimal cut sets. The set of reactions Y' will be removed from the network. If there is no such flux v, we continue enumerating minimal cut sets among Y.

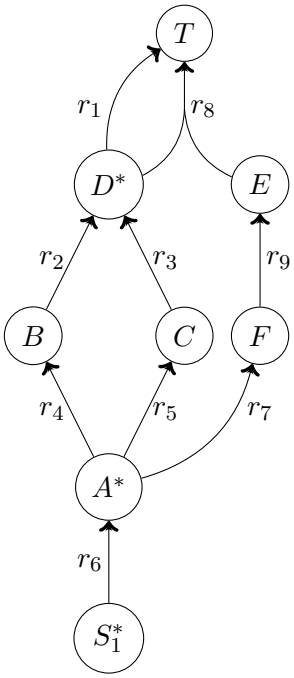


Figure 4.7: A metabolic network with several essential compounds (denoted by a $*$) to illustrate the effect of the removal of reactions that consume or produce essential compounds. The only minimal cut set for the consumption of A consists in $\{r_4,\ r_5\}$. The minimal cut set for the consumption of D consists in$\{r_1\}$. The cut sets for the production of A and D consists in $\{r_6\}$ and $\{r_2,\ r_3\}$, respectively. The reactions r_7 and r_8 do not take part of any minimal cut set of any subnetwork comprising the reactions that consume/produce a single essential compound. The removal of both reactions divides the number of minimal factories by two.

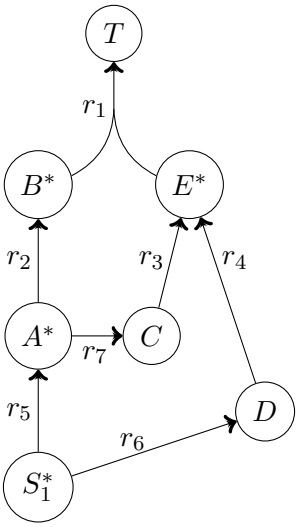


Figure 4.8: A metabolic network with several essential compounds (denoted by a $*$). The reactions r_6 and r_7 do not take part of any minimal cut set of a subnetwork containing the reactions that respectively consume or produce a single essential compound. However, r_6 and r_7 build a minimal cut set for the factories from S_1 to T.

The pseudocode of this approach is given in the Procedure 2, whose parameters are the metabolic network $\mathcal{N}$, the minimal precursor set X, the target set $\mathcal{T}$, and the set of reactions Y. The procedure returns Y if there is a factory from X to $\mathcal{T}$ when the fluxes on the reactions in Y are set to zero (lines $1-3$). On the contrary, minimal cut sets among Y are enumerated (line 8): the function *nextMinCutSet* computes a minimal cut set (minimal with respect to the previously found minimal cut sets in MCS) for the production of the set of targets $\mathcal{T}$ taking into account the reactions in Y. If the cardinality of the current minimal cut set is greater than the previous one, the LP (4.3) tests whether there is a factory from X to $\mathcal{T}$ when the reactions in $Y' = Y \setminus \mathcal{R}_{MCS}$ are blocked (line 11). If this is the case, Y' is returned. On the contrary, the reactions of the current minimal cut sets are added to $\mathcal{R}_{MCS}$ (line 15) and we enumerate the next minimal cut set. We repeat this loop until there is no minimal cut set anymore or all reactions of Y are part of at least one minimal cut set (line 16). In this case, we return the reactions that do not take part of a minimal cut set in MCS (line 17). The reactions that are returned from this algorithm are removed from the network.

Algorithm 2: *getReactionSubset*$(\mathcal{N},\ X,\ \mathcal{T},\ Y)$

Input : The network $\mathcal{N} = (\mathcal{C}, \mathcal{R})$, the minimal precursor set X, the target set $\mathcal{T}$, and the set of reactions Y.
Output : The set of reactions $Y' \subseteq Y$

1 $v \leftarrow LP_{(4.3)}(\mathcal{N}, X, \mathcal{T}, Y)$;
2 **if** $|supp(v)| > 0$ **then**
3 **return** Y;
4 $MCS \leftarrow \{\}$;
5 $\mathcal{R}_{MCS} \leftarrow \{\}$;
6 $i \leftarrow 0$;
7 **repeat**
8 $MCS[i] \leftarrow nextMinCutSet(\mathcal{N}, \mathcal{T}, Y, MCS)$;
9 **if** $i > 0$ && $|MCS[i]| > |MCS[i-1]|$ **then**
10 $Y' \leftarrow Y \setminus \mathcal{R}_{MCS}$;
11 $v \leftarrow LP_{(4.3)}(\mathcal{N}, X, \mathcal{T}, Y')$;
12 **if** $|supp(v)| > 0$ **then**
13 **return** Y';
14 $i++$;
15 $\mathcal{R}_{MCS} \leftarrow \mathcal{R}_{MCS} \cup MCS[i]$;
16 **until** $|MCS[i]| == 0$ || $Y == R_{MCS}$;
17 **return** $Y \setminus \mathcal{R}_{MCS}$;

After removal of the reactions of (a subset of) Y, the pruning steps of Section 4.3.1 can be applied, *e.g.*, because the network may contain topological sources that are not in X (as the compound F in Figure 4.7).

4.3.3 MILP approach

The approach presented in this section is similar to the one described in the previous chapter, and to several methods in the literature (Lee et al., 2000; Larhlimi and Bockmayr, 2007; de Figueiredo et al., 2009; von Kamp and Klamt, 2014). The idea is to find recursively minimal reaction subsets solving a MILP problem at each iteration. Given a metabolic network $\mathcal{N} =$

$(\mathcal{C}, \mathcal{R})$, a stoichiometric matrix S, and a minimal precursor set $X \subseteq \mathcal{X}$ for the set of targets $\mathcal{T}$, we enumerate all minimal factories from X to $\mathcal{T}$. A factory F corresponds to the support of a flux $v \in \mathbb{R}^{|\mathcal{R}|}$. We assume that all reactions are irreversible. Thus, a reversible reaction is split in a forward and a backward reaction. For each network reaction in $\mathcal{R}$, we introduce a binary variable b, which on one hand captures if the flux v_j of reaction $j \in \mathcal{R}$ is positive, and on the other hand allows to exclude already found minimal solutions from the solution space. To find a first minimal solution we solve:

$$\begin{aligned} \min f = \quad & \sum_{j=1}^{|\mathcal{R}|} b_j \\ s.t \quad & (Sv)_{\mathcal{T}} \geq \epsilon, \\ & (Sv)_{\mathcal{C}\setminus X} \diamond 0 \\ & (Sv)_X < 0 \\ & b_j = 0 \leftrightarrow v_j = 0, \quad \forall j \in \mathcal{R} \\ & b_j \in \{0,1\}, \quad \forall j \in \mathcal{R} \\ & 0 \leq v_i \leq U, \quad \forall i \in \mathcal{R}, \end{aligned} \tag{4.13}$$

where the $\diamond$ symbol can be substituted by the $\geq$ or the $=$ sign, dependent on if $S-factory$ or steady state $S-factory$ are enumerated. The set of targets is produced in a positive amount. Note that contrary to the approach for the enumeration of minimal precursor sets, we do not introduce reactions that produce the sources from the empty set. Therefore, the topological sources in X are constrained to be consumed. The constant U denotes an upper bound of the fluxes. The constraint $b_j = 0 \leftrightarrow v_j = 0$ can be modeled through indicator constraints in CPLEX or as follows:

$$\left.\begin{array}{rcl} b_j & \leq & v_j \\ v_j & \leq & U b_j \end{array}\right\} \text{ for } \forall j \in \mathcal{R}. \tag{4.14}$$

These constraints are known as the big M method and let b_j become equal to one when $v_j \geq 1$. Let the pair (v^*, b^*) be the minimal solution of Problem (4.13), and I_{b^*} the support of b^*. Adding the following constraint to a subsequent MILP formulation prevents the enumeration of the same solution and of any superset of it.

$$\sum_{j \in I_{b^*}} b_j \leq |I_{b^*}| - 1, \tag{4.15}$$

The constraint (4.15) is built for each already enumerated minimal solution to obtain a new MILP formulation in the next iteration. If there is no feasible solution to a problem, then we claim that all minimal (steady state) factories have been enumerated.

4.3.4 Combinatorial approach

Theorem 1 in Chapter 3 states that for any minimal $S-factory$ $H \subseteq \mathcal{R}$ from X to T, there exists a set of minimal topological factories $F_1, \ldots, F_k$ from X to T in $\Psi(\mathcal{N})$ such that:

1. $F_1, \ldots, F_k \subseteq \Psi(H)$;
2. For each reaction r in H there is $i \in \{1, \ldots, n\}$ such that $\Psi(r) \cap F_i \neq \emptyset$.

This means that one can first enumerate all topological factories from X to T in the many-to-one network $\Psi(\mathcal{N})$, then map the many-to-one reactions of these topological factories back to their corresponding hyperreactions, and finally combine them to obtain minimal $S-factories$. In this section, we first present an algorithm of Acuña et al. (2012) whose original purpose was to enumerate minimal topological precursor sets. We show that this approach can easily

be adapted to enumerate minimal topological factories from X to $\mathcal{T}$. We explain why this approach is however not suited to enumerate steady state $S-factories$. Finally, we demonstrate how essential compounds can be used to split the problem of enumerating minimal topological factories from X to $\mathcal{T}$ into subproblems.

To enumerate all minimal topological factories in a many-to-one network, we make use of the TRD algorithm presented in Acuña et al. (2012). The TRD algorithm is a graph based approach that uses depth-first search (DFS) and backtracking to compute $\mathbb{P}_M(A)$, the collection of all minimal topological factories for a target set M when A is available, starting with $M := \mathcal{T}$ and $A := \emptyset$. For a given precursor set $X \subseteq \mathcal{X}$, the authors state that A is available if there is a topological factory from $X \cup A$ to M. The algorithm traverses the many-to-one hypergraph from the targets to the sources in DFS manner taking the reactions in their reverse sense. The algorithm computes the minimal topological factories of each element of $M = \{m_1, \ldots, m_k\}$ (target decomposition). The collection of topological factories that produce M is formed in the following way $\mathbb{P}_M(A) = \{F_1 \times \cdots \times F_k | F_i \in \mathbb{P}_{m_i}(A)\}$ that we call cartesian product. In order to obtain the collection of topological factories $\mathbb{P}_{m_i}(A)$, with $i \in \{1, \ldots, k\}$, the algorithm considers all reactions $r_1, \ldots, r_l$ that produce m_i (reaction decomposition). For each reaction, the algorithm calls the target decomposition step for its substrates. The collection of topological factories $\mathbb{P}_{m_i}(A)$ is the union of all $\mathbb{P}_{Subs(r_j)}(A)$ with $j \in \{1, \ldots, l\}$. The target and reaction decomposition steps are called alternately until M contains only one element that is either a source or an element in A.

If the set of available compounds would not be augmented then the algorithm would run into an endless loop due to cycles in metabolic networks. Let us consider the cycle spanned by the reaction $r_1 : X \to Y$ and $r_2 : Y \to X$. Starting with $M := \{X\}$ and $A := \emptyset$, alternate calls of the target and reaction decomposition would result in the following steps $\mathbb{P}_{\{X\}}(\emptyset) = \mathbb{P}_{Subs(r_2)}(\emptyset) = \mathbb{P}_{\{Y\}}(\emptyset) = \mathbb{P}_{Subs(r_1)}(\emptyset) = \mathbb{P}_{\{X\}}(\emptyset)$. As X is neither a source nor in the set of available compounds A, the algorithm continues to calculate $\mathbb{P}_{\{X\}}(\emptyset)$ in an endless loop. Adding m_i to A at the reaction decomposition step avoids cycling when $\mathbb{P}_{Subs(r)}(A \cup \{m_i\})$ is computed. In the toy example above, the target/reaction decomposition steps stop after $\mathbb{P}_{\{X\}}(\emptyset) = \mathbb{P}_{Subs(r_2)}(\{X\}) = \mathbb{P}_{\{Y\}}(\{X\}) = \mathbb{P}_{Subs(r_1)}(\{X,Y\}) = \mathbb{P}_{\{X\}}(\{X,Y\})$ as $X \in \{X,Y\}$. At the target decomposition step, for every m_i all compounds of M that are different from m_i can be added to A ($\mathbb{P}_{m_i}(A \cup (M \setminus m_i))$) as the minimal topological factories of these compounds will be computed in parallel subproblems. This refinement decreases the depth of the search. For further details, we refer to Acuña et al. (2012).

The target and reaction decomposition steps are shown below. At the target decomposition (*tDecomp*) we call the reaction decomposition (*rDecomp*) for each compound that is neither a source nor in the set of available compounds A. Topological factories of the different $m \in M$ are then combined through the cartesian product, if the resulting topological factories respect Lemma 2 of Chapter 3, that is each compound is produced by exactly one many-to-one reaction. Lemma 2 thus imposes the minimality of a topological factory. The algorithm examines every reaction that produces a compound m at the reaction decomposition step. Therefore, *tDecomp* is called for the substrates of a reaction r (line 3). The reaction r is added to all topological factories of $\mathbb{P}_{Subs(r)}(A)$ (line 4).

The procedures *tDecomp* and *rDecomp* differ slightly from Acuña et al. (2012) in that they return minimal topological factories instead of minimal topological precursor sets. Furthermore, one cannot prune the solutions by minimality (Acuña et al., 2012), where a reaction r' is not considered for the production of a compound m in *rDecomp*, if there is a reaction r with $Subs(r) \setminus A \subseteq Subs(r') \setminus A$. This is illustrated in Figure 4.9. Here, in the case of the enumeration of minimal topological precursor sets, r' can be omitted when the reaction decomposition step is called for T ($A = \emptyset$), because $Subs(r) \subseteq Subs(r')$. The minimal topological precursor

Algorithm 3: tDecomp(M, A, X)

Input : The set **M** of target compounds, the set **A** of available compounds, the set of source compounds **X**.

Output : The collection of all minimal topological factories for M.

```
C_f ← {{}};
foreach m ∈ M do
    if m ∉ (X ∪ A) then
        C_m ← rDecomp(m, A ∪ (M \ {m}), X);
        C_f ← C_f × C_m ;
return C_f;
```

Algorithm 4: rDecomp(m, A, X)

Input : The compound **m**, the set **A** of available compounds, the set of source compounds **X**.

Output : The collection of all minimal topological factories for m.

```
C_m ← {{}};
foreach r ∈ Reac_p(m) do
    C_r ← tDecomp(Subs(r), A ∪ {m}, X);
    C_r ← C_r × {r};
    C_m ← C_m ∪ C_r;
return C_m;
```

sets that are found by following the reaction r are contained in those found by following r' (S_1 in this example). However, when minimal topological factories are enumerated, r' needs to be considered as it results in a minimal topological factory ($\{r_1,\ r_2,\ r'\}$) from S_1 to T that would not have been discovered otherwise.

We demonstrate the functioning of the algorithm using the toy network of Figure 4.10. There is one minimal stoichiometric precursor set $\{S_1\}$ for the production of the target compound T. We call the target decomposition for T, and then the reaction decomposition for the same compound as it is neither a source nor an available compound A. The compound T is produced from A which in turn is produced from C or D. The target and reaction decomposition steps are called alternately until the base cases are reached, *e.g.* S_1 and A at the bottom, where the former is a source, and the latter is in the collection of available compounds A (compound A is a substrate of a reaction that was encountered in an earlier iteration). On the backtracking to the target the topological factories are recovered, namely $\{r_2,\ r_{3_C},\ r_4,\ r_1\}$ and $\{r_2,\ r_{3_D}, r_5,\ r_1\}$. These minimal topological factories have to be combined to constitute a (minimal) stoichiometric factory.

Figure 4.10 illustrates that this approach is not suited to enumerate steady state factories. Pay attention to the compound F which must be exported through r_6 to be in steady state. The reaction r_6 is however not recovered by any (minimal) topological factory from X to $\mathcal{T}$. The topological factories correspond to the hyperpaths from the root to the leafs (taking the reaction in the opposite direction) in the recursion tree (of $tDecomp(\mathcal{T}, \emptyset, X)$) which is shown in Figure 4.10b.

Use of essential compounds

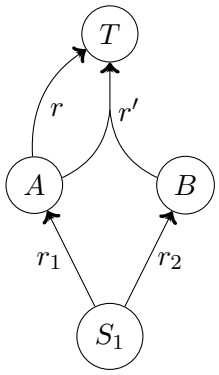


Figure 4.9: Illustration of the fact that the pruning of minimality, as described in Acuña et al. (2012), can not be applied.

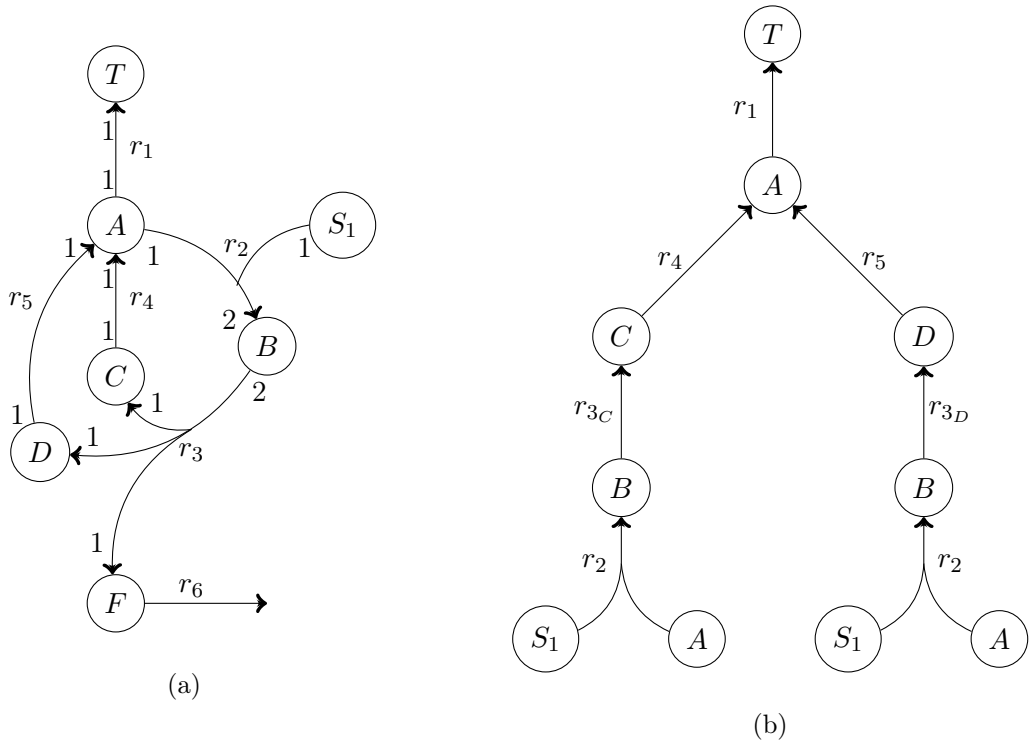


Figure 4.10: The toy network is depicted in 4.10a with the minimal stoichiometric precursor set S_1 for the production of the target T. The recursion tree of the target and reaction decomposition steps in the many-to-one network is shown in 4.10b

Pruning the network already reduces the search space but it is the concept of *essential compounds* that allows us to enumerate, from a practical point of view, all minimal topological factories from X to $\mathcal{T}$ in a network such as the *E.coli* core network (E.coli, 2016). Every $S - factory$ passes through all essential compounds. Let $\mathcal{C}_{ess}$ denote the set of essential compounds. The compounds in the minimal precursor set X are essential compounds that are only consumed in all $S - factories$. The remaining essential compounds are consumed *and* produced. Thus, these compounds can be considered as sources and targets in smaller subproblems. For each compound $c \in \mathcal{C}_{ess} \cup \mathcal{T}$, we enumerate all minimal topological factories from $\mathcal{X}' \subseteq \mathcal{C}_{ess} \setminus \{c\}$ to c in the many-to-one network. In this way, essential compounds enable us to decompose the problem of enumerating all topological factories from X to $\mathcal{T}$ in the many-to-one network into subproblems. Each subproblem can be solved in parallel. The minimal topological factories of the subproblems must be combined to obtain minimal topological factories from X to $\mathcal{T}$. The latter ones have to be combined to obtain minimal $S - factories$ from X to $\mathcal{T}$ according to Theorem 1.
The algorithm is depicted in the procedure 5. In the beginning, the *essential compounds* of the $S - factories$ from X to $\mathcal{T}$ are computed (*getEssentialCmp*, line 1). We enumerate all minimal topological factories for each essential compound in the many-to-one network $\psi(\mathcal{N})$ using the TRD algorithm presented in Acuña et al. (2012) (*tDecomp*, line 4). The obtained minimal topological factories for the essential compounds are then combined to get minimal topological factories from X to $\mathcal{T}$ in $\psi(\mathcal{N})$ (*assembleFactories*, line 5). The latter ones are further combined to obtain minimal $S - factories$ from X to $\mathcal{T}$ in $\mathcal{N}$ (*combineMinTopFac*, line 6). Only minimal $S - factories$ are returned.

Algorithm 5: enumSFac($\mathcal{N}, X, \mathcal{T}$)

Input : The network $\mathcal{N} = (\mathcal{C}, \mathcal{R})$, The minimal SPS X, and the set of target compounds $\mathcal{T}$.

Output : The collection of all minimal $S - factories$ from X to $\mathcal{T}$.

1 $\mathcal{C}_{ess} \leftarrow getEssentialCmp(\mathcal{N}, X, \mathcal{T})$;
2 $F \leftarrow \{\}$;
3 **foreach** $c \in C_{ess} \cup \mathcal{T}$ **do**
4 $\quad F \leftarrow F \cup tDecomp(\{c\}, \emptyset, \mathcal{C}_{ess} \setminus \{c\})$;
5 $F_t \leftarrow assembleFactories(\mathcal{T}, F, \emptyset)$;
6 $F_s \leftarrow combineMinTopFac(F_t, X, \mathcal{T})$;
7 **return** $minimal(F_s)$;

Calling $tDecomp(\{c\}, \emptyset, \mathcal{C}_{ess} \setminus \{c\})$ for every $c \in \mathcal{C}_{ess} \cup \mathcal{T}$ provides all minimal topological factories from $\mathcal{C}_{ess}$ to c, respectively (collection F in the procedure 5). We denote by $F(m)$ the minimal topological factories from $\mathcal{C}_{ess}$ to m. To obtain all minimal topological factories from X to $\mathcal{T}$ in $\psi(\mathcal{N})$ we need to assemble the parts. We denote the compounds that are only consumed in a topological factory f by $Cons(f)$, and the compounds that are produced in f by $Prod(f)$. The procedure *assembleFactories*, that computes the minimal topological factories of M if A is available ($\mathbb{P}_M(A)$), works in a similar way to the target and reaction decomposition. The procedure does not traverse the network by taking the reactions that produce a compound m in the reverse direction, but instead considers the topological factories $f \in F(m)$ that produce a compound m. The procedure is then called recursively for all compounds that are only consumed in f. The collection of available compounds is augmented at each iteration by the compounds that are produced by f and the compounds in M except the focal compound m *inM* (line 5). We call *assembleFactories*$(\mathcal{T}, \emptyset, F)$ to obtain all minimal topological factories from C_{ess} to $\mathcal{T}$. Only the minimal topological factories from X to $\mathcal{T}$ are

kept.

Algorithm 6: assembleFactories(M, A, F)

Input : The set **M** of target compounds, the collection of available compounds A, and the collection **F** of minimal topological factories.

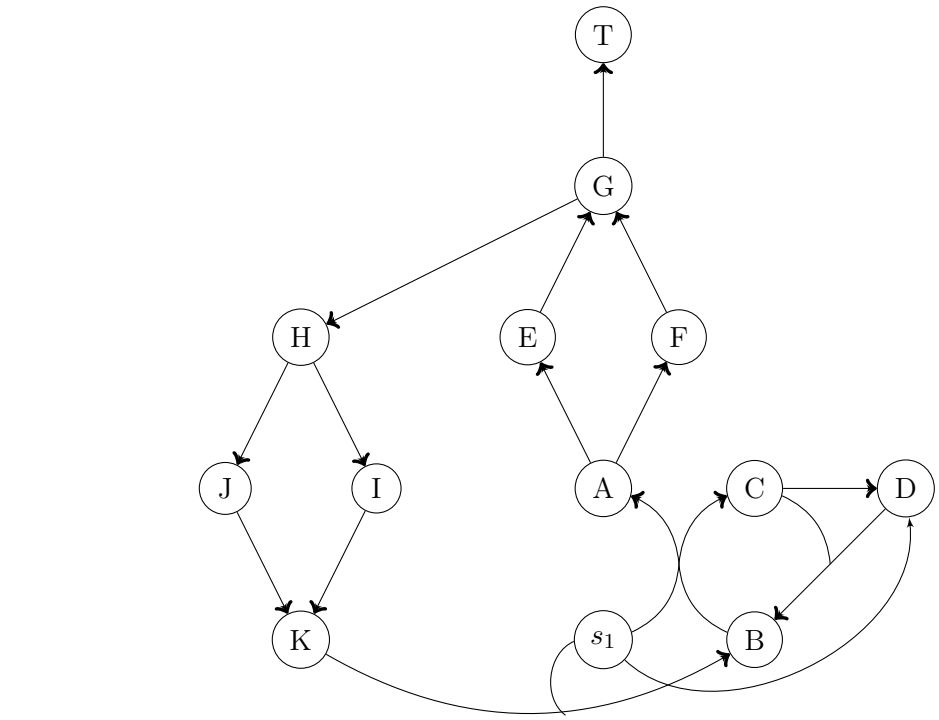
Output : The collection of all minimal topological factories for M.

```
1 C_t ← {};
2 foreach m ∈ (M \ A) do
3     C_m ← {};
4     foreach f ∈ F(m) do
5         C_f ← assembleFactories(Cons(f), A ∪ Prod(f) ∪ M \ {m}, F);
6         C_m ← C_m ∪ {f ∪ C_f};
7     C_t ← C_t ⋈ C_m;
8 return minimal(C_t);
```

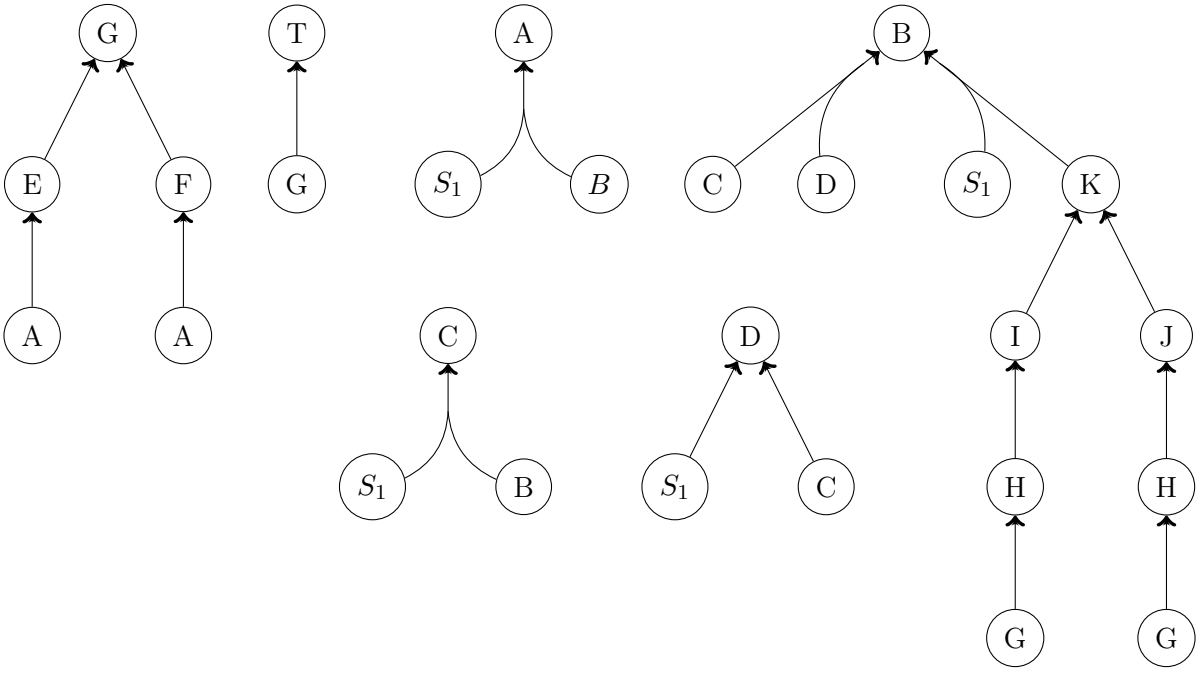
At the end, we combine all minimal topological factories from X to $\mathcal{T}$ in $\psi(\mathcal{N})$ to obtain all minimal $S-factories$ from X to $\mathcal{T}$ in $\mathcal{N}$. We check through linear programming if a topological factory is a $S-factory$. Only minimal $S-factories$ are retained.

We illustrate, with the help of Figure 4.11, the algorithm that takes into account the essential compounds. Here, there is one minimal stoichiometric precursor set S_1 for the production of the target T. The essential compounds of all stoichiometric factories are G, A, B, C, and D. The result of the computation of the minimal topological factories of all essential compounds and the target is shown in Figure 4.11b. These parts are assembled according to procedure 6 to obtain minimal topological factories from $X = \{S_1\}$ to $\mathcal{T} = \{T\}$. In this example, there are eight minimal topological factories (see Figure 4.12 and 4.13). All minimal stoichiometric factories from S_1 to T are shown in Figure 4.14. The minimal topological factories in Figure 4.12d and in Figure 4.13d build already minimal stoichiometric factories. The remaining minimal stoichiometric factories result from the following combinations:

1. the minimal $S-factory$ in Figure 4.14a results from the combination of the minimal topological factories in Figure 4.12a and 4.12c,
2. the minimal $S-factory$ in Figure 4.14b results from the combination of the minimal topological factories in Figure 4.12a and 4.12b,
3. the minimal $S-factory$ in Figure 4.14c results from the combination of the minimal topological factories in Figure 4.13a and 4.13c,
4. the minimal $S-factory$ in Figure 4.14e results from the combination of the minimal topological factories in Figure 4.13a and 4.13b,



(a)



(b)

Figure 4.11: 4.11a In the metabolic network, there is one minimal stoichiometric precursor set S_1 for the target T. The essential compounds in the stoichiometric factories from S_1 to T are G, A, B, C, and D. 4.11b shows the minimal topological factories of the essential compounds and the target T

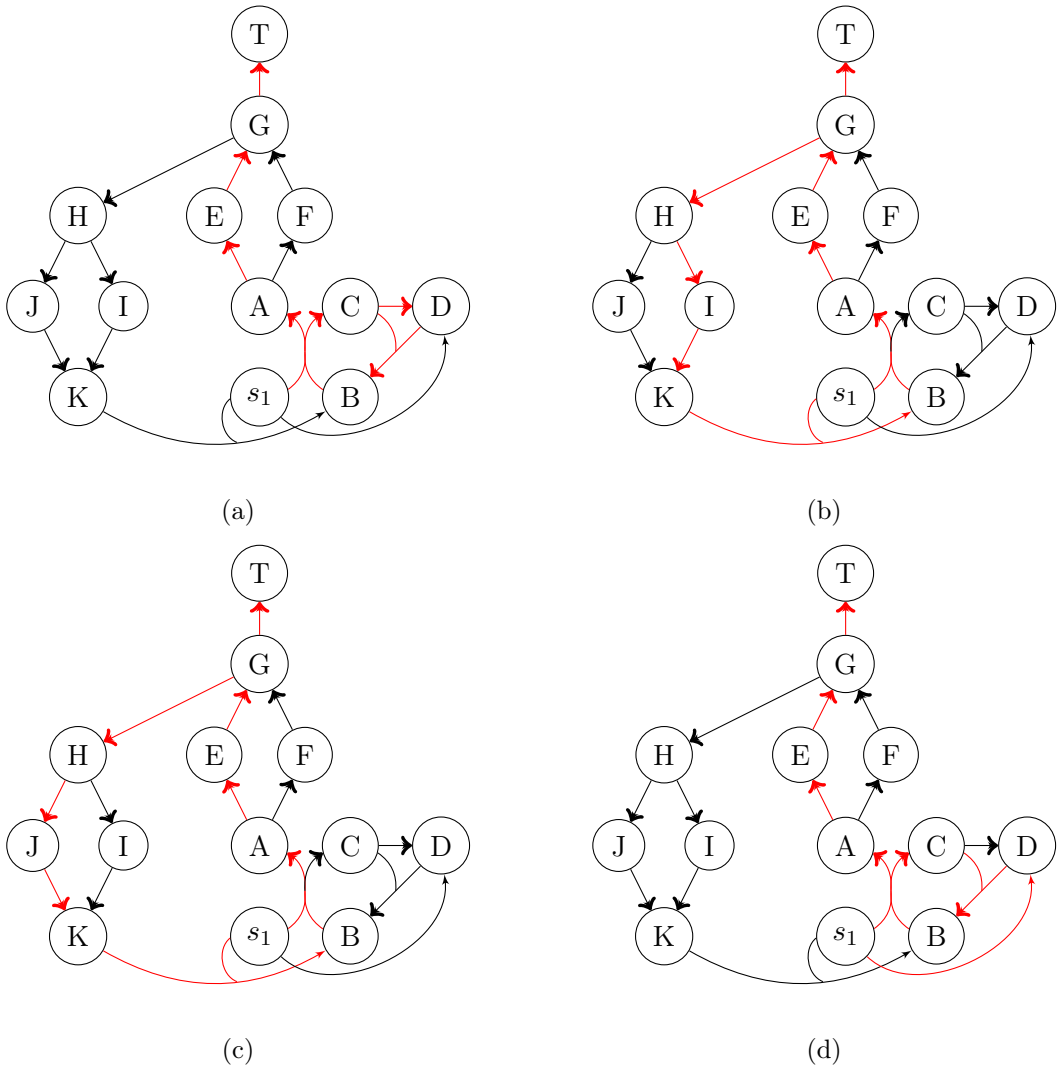


Figure 4.12: Minimal topological factories $1-4$ from S_1 to T

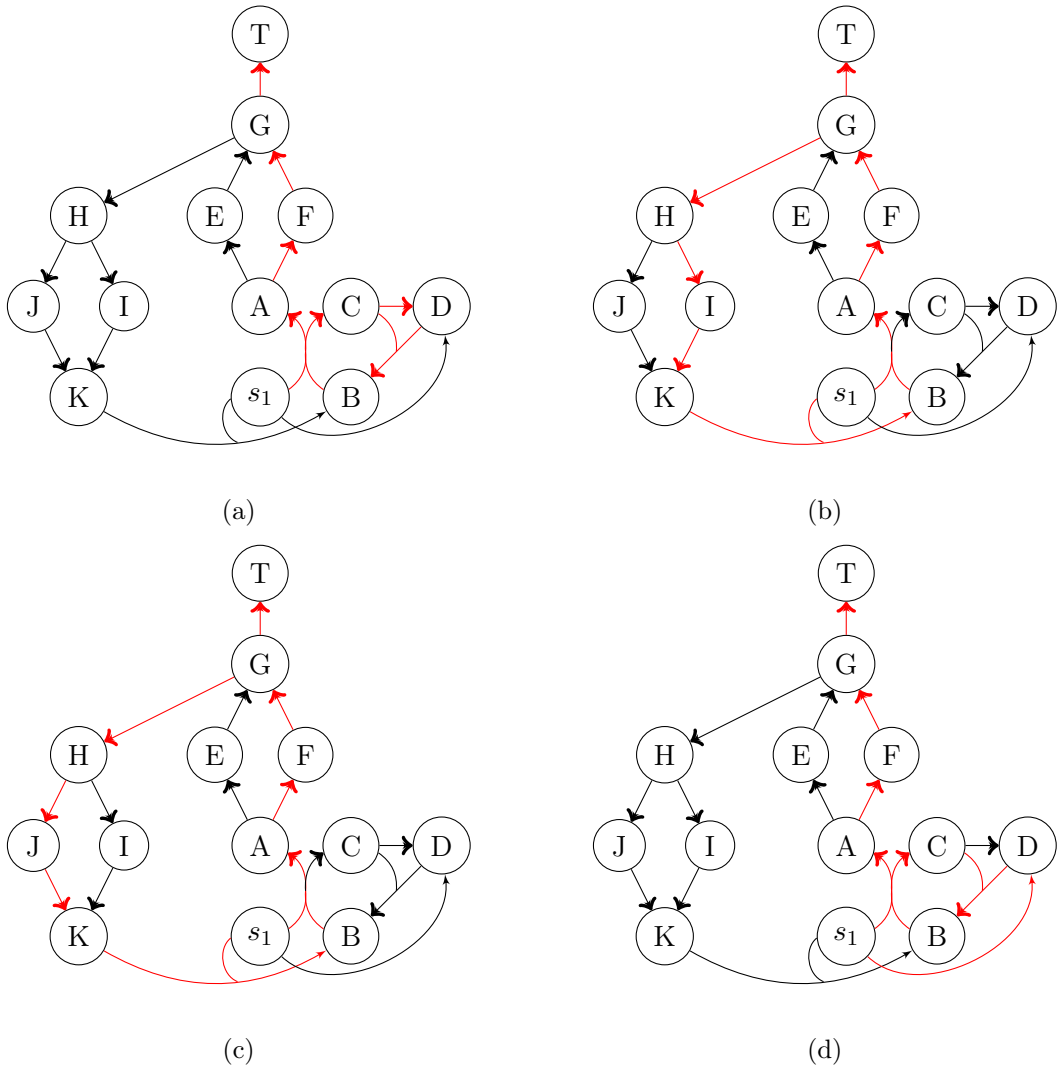


Figure 4.13: Minimal topological factories $5-8$ from S_1 to T

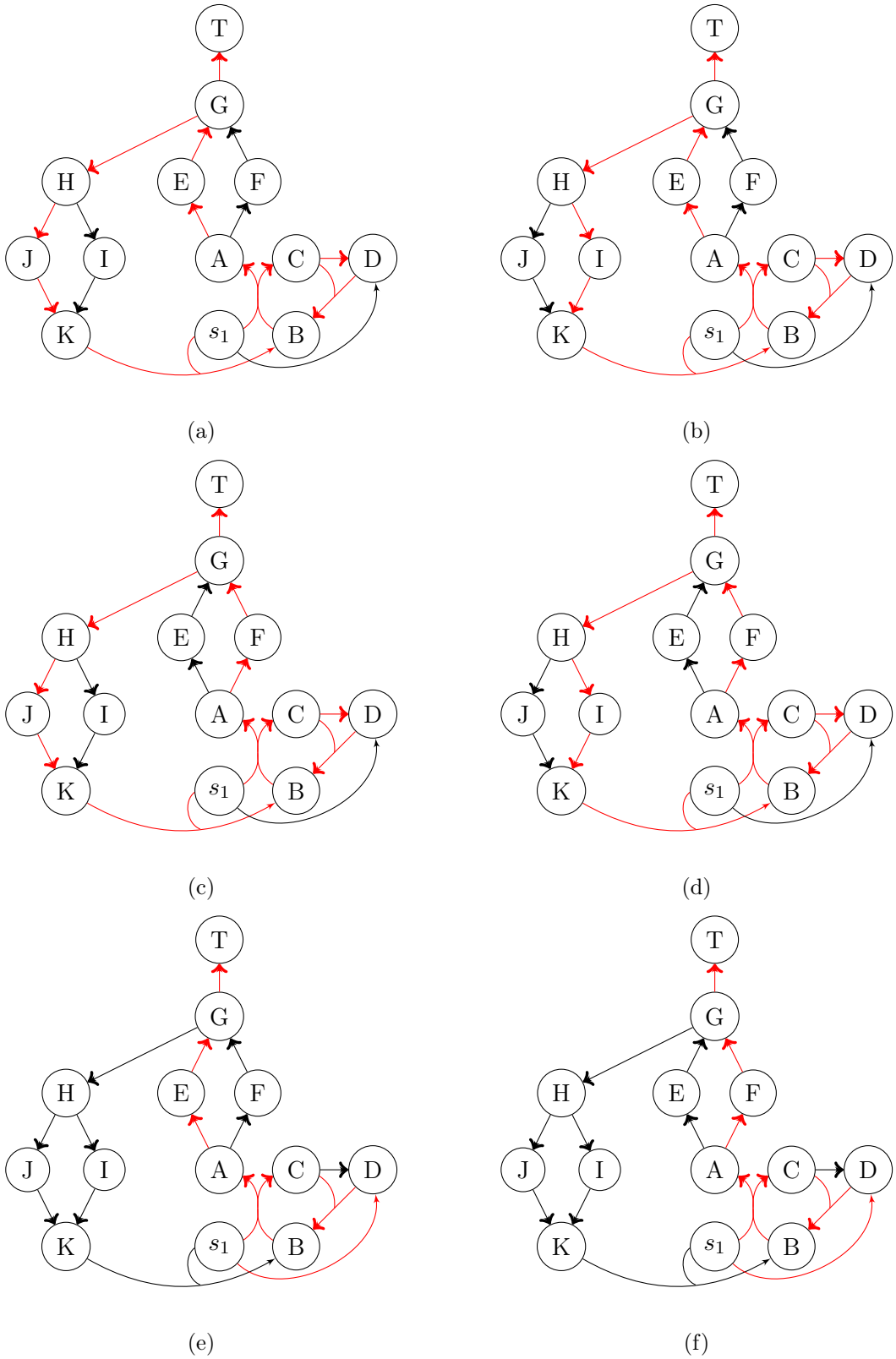


Figure 4.14: All minimal stoichiometric factories from S_1 to T.

Minimal precursor set	steady state	accumulation
pi, glc_D	✓	✓
pi, h2o, fum, etoh	✓	✓
pi, etoh, o2	✓	✓
pi, h2o, fum, h	✓	✓
pi, fum, h, o2	✓	✓
pi, fum, lac_D	✓	✓
pi, h2o, lac_D	✓	✓
pi, h2o, pyr	✓	✓
pi, lac_D, h	✓	✓
pi, succ, h, o2	✓	✓
pi, lac_D, o2	✓	✓
pi, pyr, o2	✓	✓
pi, h2o, ac, akg	✓	✓
pi, ac, h, o2	✓	✓
pi, h2o, fum, akg	✓	✓
pi, akg, o2	✓	✓
pi, pyr, h		✓
pi, ac, akg, h		✓
pi, ac, fum, h		✓
pi, fum, pyr, etoh		✓
pi, ac, fum, akg, etoh		✓
pi, h2o, co2, akg, etoh		✓

Table 4.1: Minimal stoichiometric precursor sets in steady state (second column) and allowing for an acculation (third column). Abbreviations: pi (phosphate), glc_D (D-glucose), h2o (water), fum (fumarate), etoh (ethanol), o2 (oxygen), h (hydrogen), lac_D (D-Lactate), pyr (pyruvate), succ (succinate), ac (acetate), and akg (2_oxoglutarate)

4.4 Results and Discussion

We enumerated all minimal stoichiometric precursor sets that enable the production of biomass (in steady state and allowing for an accumulation, respectively) in the *E.coli* core model (E.coli, 2016) that contains 78 compounds and 77 reactions. We obtained 16 (steady state) and 22 (with an accumulation) minimal precursor sets that are listed in Table 4.1. All minimal precursor sets that fulfil the steady state constraint are also solutions when an accumulation is allowed. For six of the solutions allowing for an accumulation, there is apparently an accumulation of some compound in the network such that steady state can not be attained.

For each minimal precursor set we enumerated all minimal stoichiometric factories using the MILP approach. In a second step, we enumerated the minimal cut sets among the reactions that produce or consume essential compounds (in the factories of a single minimal precursor set) and remove reactions, that are linked to essential compounds but are not in a minimal cut set. We enumerated again all minimal stoichiometric factories in the remaining subnetwork. The number of the obtained minimal stoichiometric factories are shown in Table 4.2. It can be seen that there are less minimal factories allowing for an accumulation compared to the minimal factories at steady state. The reason is that the compounds that accumulate in a factory (with accumulation) can be exported in multiple ways to achieve steady state. We

Minimal precursor set	steady state		accumulation	
	mcs	complete	mcs	complete
pi, glc_D	36	2733	1	71
pi, h2o, fum, etoh	309	1666	20	165
pi, etoh, o2	857	865	14	76
pi, h2o, fum, h	28	1286	2	237
pi, fum, h, o2	447	4273	21	124
pi, fum, lac_D	61	150	1	17
pi, h2o, lac_D	106	106	1	22
pi, h2o, pyr	8	390	1	52
pi, lac_D, h	9	69	1	11
pi, succ, h, o2	28	2342	2	237
pi, lac_D, o2	442	2814	2	54
pi, pyr, o2	97	2298	1	92
pi, h2o, ac, akg	42	144	28	128
pi, ac, h, o2	28	253	14	56
pi, h2o, fum, akg	44	691	3	761
pi, akg, o2	193	1344	5	641
pi, pyr, h			1	14
pi, ac, akg, h			16	117
pi, ac, fum, h			4	171
pi, fum, pyr, etoh			5	15
pi, ac, fum, akg, etoh			32	89
pi, h2o, co2, akg, etoh			7	28

Table 4.2: Minimal stoichiometric factories per minimal precursor set, either in steady state or allowing for an acculation, and either in the pruned network (reactions linked to essential compounds that are not in a minimal cut set (mcs) among these reactions are removed) or in the complete network. Abbreviations: pi (phosphate), glc_D (D-glucose), h2o (water), fum (fumarate), etoh (ethanol), o2 (oxygen), h (hydrogen), lac_D (D-Lactate), pyr (pyruvate), succ (succinate), ac (acetate), and akg (2_oxoglutarate)

observe that for every minimal factory at steady state in the complete network, there is a minimal factory (allowing for an accumulation) that is a subset of the former. This relation does not hold for minimal factories in the network that is pruned by the minimal cut sets approach. This is because the minimal cut sets at steady state and with an accumulation do not coincide.

Unfortunately, the combinatorial approach is not efficient enough, in practice, for the enumeration of all minimal stoichiometric factories allowing for an accumulation in the *E.coli* core model (E.coli, 2016). We are able to enumerate the minimal topological factories from X to the target set $\mathcal{T}$ in the *E.coli* core network in a reasonable time, but we generate a lot of solutions. The bottleneck is thus the combination of the minimal topological factories to obtain minimal stoichiometric factories. Although possible in theory, it is certainly not necessary to make all combinations of the topological factories. This is because some of them do not result in a stoichiometric factory. We however did not managed to find a criteria to exclude combinations beforehand and are thus obliged to test via LP if a combination of minimal topological factories is also a stoichiometric factory.

Minimal precursor set	steady state		accumulation	
	min	max	min	max
pi, glc_D	38.16	310.22	4.38	172.70
pi, h2o, fum, etoh	8.99	79.76	4.03	59.42
pi, etoh, o2	35.52	251.91	4.56	224.27
pi, h2o, fum, h	6.80	67.37	10.23	62.62
pi, fum, h, o2	23.90	316.97	10.00	186.68
pi, fum, lac_D	14.58	28.19	11.77	20.15
pi, h2o, lac_D	10.78	41.03	11.45	29.65
pi, h2o, pyr	5.96	80.47	3.31	53.57
pi, lac_D, h	8.65	28.05	9.21	21.63
pi, succ, h, o2	17.20	283.00	9.43	211.52
pi, lac_D, o2	7.70	272.99	6.30	213.07
pi, pyr, o2	11.77	274.16	5.14	192.33
pi, h2o, ac, akg	3.07	46.18	2.26	29.81
pi, ac, h, o2	9.43	185.03	8.13	52.13
pi, h2o, fum, akg	5.55	116.08	2.67	97.57
pi, akg, o2	12.58	192.26	2.36	132.35
pi, pyr, h			12.40	22.00
pi, ac, akg, h			4.00	25.49
pi, ac, fum, h			7.89	26.97
pi, fum, pyr, etoh			6.28	10.32
pi, ac, fum, akg, etoh			2.37	13.89
pi, h2o, co2, akg, etoh			21.78	27.55

Table 4.3: The minimum and maximum achievable biomass production rate in the complete network (no minimal cut set pruning) among the minimal factories per minimal precursor set and with an upper bound of 10 000 on the input fluxes. Abbreviations: pi (phosphate), glc_D (D-glucose), h2o (water), fum (fumarate), etoh (ethanol), o2 (oxygen), h (hydrogen), lac_D (D-Lactate), pyr (pyruvate), succ (succinate), ac (acetate), and akg (2_oxoglutarate)

We analyzed two properties of the minimal stoichiometric factories. First, we observed that the minimal factories in the network after the minimal cut set pruning are not necessarily among the shortest ones and are thus different from enumerating the $k-$shortest minimal factories (de Figueiredo et al., 2009). The mean length of the minimal factories and the standard deviation in the different groups are as follows: steady state in the subnetwork (minimal cut sets approach) (mean: 51.32, σ : 2.78), steady state in the complete network (mean: 51.22, σ : 2.77), allowed accumulation in the subnetwork (mean: 36.13, σ : 3.13), and allowed accumulation in the complete network (mean: 36.66, σ : 2.81).

Second, for each obtained minimal factory, we maximized the biomass production rate putting an upper bound of 10.000 on the influx of the source compounds. The minimum and maximum values that could be achieved in the complete network are shown in Table 4.3. The minimal stoichiometric factories in steady state attain a higher maximum biomass production rate than minimal stoichiometric factories allowing for an accumulation. The biomass production rate is in general higher for precursor sets containing oxygen (except for the precursor set {pi, glc_D}) which is coherent with the fact that *E. coli* favor aerobic growth conditions.

We further examined the difference of the distribution of maximum biomass production rates

between factories in steady state, and with and without the minimal cut set approach. For this purpose, we round down the maximum biomass production rate of each minimal stoichiometric factory to the next natural number. Counting the occurences of biomass production rates allows a comparison between minimal factories in the complete network and those of a subnetwork (minimal cut set approach). As an example, we show in Figure 4.15a and Figure 4.15b, the distribution of the maximum biomass production rates of the minimal stoichiometric factories from the minimal precursor set {pi, glc_D} in the complete network and in the subnetwork (min cut set approach), respectively. The barplots in Figure 4.15 suggest that the subnetwork contains only minimal stoichiometric factories with a higher biomass production rate. However, this pattern can not be observed throughout all minimal stoichiometric precursor sets. The distribution of the maximum biomass production rate from the minimal precursor set {pi, succ, h, o2} is shown in Figure 4.16. Contrary to what we have seen before, the remaining minimal stoichiometric factories of the subnetwork have a low maximum biomass production rate compared to what is achievable in the complete network (for the same minimal precursor set).
Analyzing the minimal stoichiometric factories, that produce biomass from the minimal stoichiometric precursor set {pi, glc_D} in steady state, obtained from the subnetwork (pruned with the minimal cut set approach) reveals that the reactions of these minimal factories overlap with the metabolic pathways glycolysis, oxidative phosphorylation, citric acid cycle, and pentose phosphate pathway. From the metabolic pathways that are specified in the network file, only the pyruvate metabolism is not represented in the minimal factories. This indicates that the cut set based approach might be useful to identify the metabolic functions of a network.

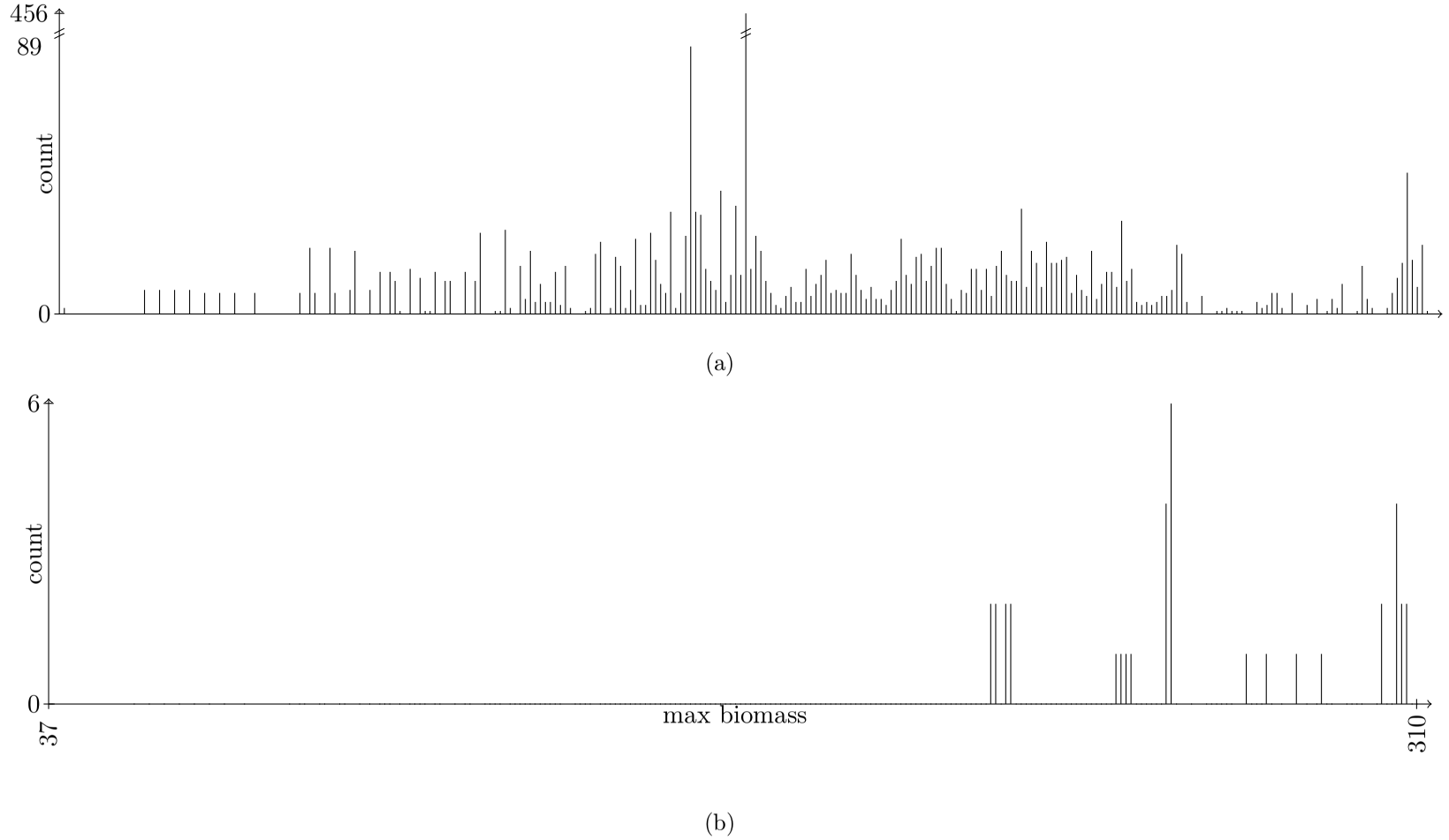


Figure 4.15: The distribution of the maximum biomass production rates of the minimal stoichiometric factories from the minimal precursor set {pi, glc_D}. The x-axis shows the rounded maximum biomass production rate, and the y-axis shows the number of factories with a certain biomass production rate. The scale of the x-axis is equal in both bar plots and is given in 4.15b. The scale of the y-axis is different between 4.15a and 4.15b. (4.15a) complete network; the y-axis and the bar at the production rate of 174 are cut for a nicer visualization. (4.15b) subnetwork after pruning with minimal cut set approach.

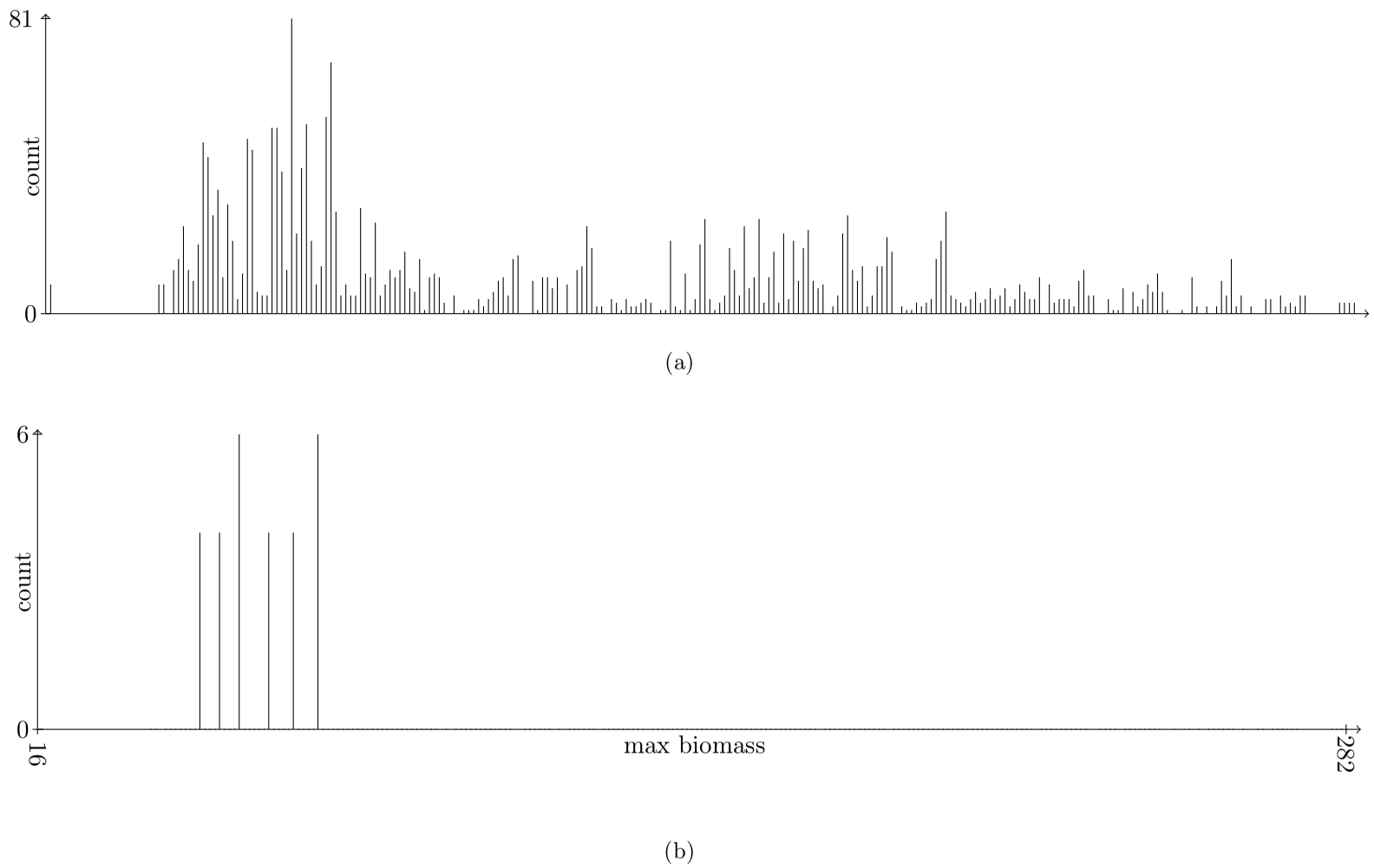


Figure 4.16: The distribution of the maximum biomass production rates of the minimal stoichiometric factories from the minimal precursor set {pi, succ, h, o2}. The x-axis shows the rounded maximum biomass production rate, and the y-axis shows the number of factories with a certain biomass production rate. The scale of the x-axis is equal in both bar plots and is given in 4.15b. The scale of the y-axis is different between 4.15a and 4.15b. (4.15a) complete network. (4.15b) subnetwork after pruning with minimal cut set approach.

4.5 Conclusion and Perspectives

In this chapter, we discussed the problem of minimal stoichiometric factories that consume a given minimal stoichiometric precursor set and produce a set of targets. Two alternative definitions of stoichiometric factories are proposed, one that considers the steady state assumption and the other that allows for an accumulation of compounds. We discussed briefly the relationship between elementary modes and minimal stoichiometric factories at steady state. The relationship between stoichiometric factories and other concepts such as chemical organizations (Dittrich and di Fenizio, 2007) are worth to be studied in the future.
Furthermore, we provided two algorithms for the enumeration of minimal stoichiometric factories. The combinatorial approach is based on the enumeration of all minimal topological factories from the set of sources to the target set. These topological factories are then combined to obtain minimal stoichiometric factories. Unfortunately, this approach is not efficient in practice, even for medium size metabolic networks. This was already discovered when we enumerated minimal stoichiometric precursor sets. However, the bottleneck is shifted more towards the combinations of minimal topological factories. The latter can be enumerated in a reasonable time in a medium size metabolic network (such as the *E. coli* core network) and for a given minimal precursor set and a set of targets. This is mainly due to the following facts: (i) the network can be pruned, and (ii) essential compounds allow the independent enumeration of smaller parts of minimal topological factories which then are assembled to obtain minimal topological factories from the given minimal precursor set to the set of targets. The second approach consists in solving recursively MILP formulations similar to the enumeration of minimal stoichiometric precursor sets. We proposed a method that is based on minimal reaction cut sets (of the production of the target set) among reactions that consume or produce an essential compound (essential in the factories from a given precursor set to the set of targets). This enables to enumerate a subset of minimal stoichiometric factories in many cases. Beside the advantage of enumerating less solutions, a substantial biological motivation must still be found before going a step further to apply the approach to genome-scale metabolic networks. The body of literature about robustness in metabolic networks should be examined closely (Burgard et al., 2003; Klamt and Gilles, 2004; Larhlimi et al., 2011; Clark and Verwoerd, 2012; Gerstl et al., 2015). The idea of taking into account only the reactions linked to essential compounds may also be considered to enumerate some minimal reaction cut sets with a higher cardinality than five in genome-scale networks (von Kamp and Klamt, 2014).

Chapter 5

Evolution of Symbiosis

Contents

In this chapter, we want to characterize species interactions at the metabolic level. We show that the minimal stoichiometric precursor sets allow to distinguish competition, commensalism, and mutualism. Minimal stoichiometric factories enable to get a more detailed view on an interaction among different species for a given minimal precursor set. We further introduce a new method to detect the compounds that are potentially exchanged between the different species. Finally, we review game theoretical and economic models that are used in the context of symbiosis and we demonstrate how the concepts of minimal precursor sets, minimal factories, and minimal set of exchanged compounds may be used therein.

To accomplish these tasks, we need to integrate the metabolic networks of several species into a joint network. We assume an environment that contains n species where each of them is associated with a metabolic network $\mathcal{N}_i = (\mathcal{C}_i, \mathcal{R}_i)$, a set of sources $\mathcal{X}_i$, and a set of target compounds $\mathcal{T}_i$ (*e.g.* biomass), for $i = 1, \ldots, n$. The set of sources contains typically, but not exclusively, topological sources (not produced by the network) and compounds that can be exchanged with the environment by a (reversible) reaction that consumes/produces the empty set. In the latter case, we remove the (reversible) exchange reaction to replace it by another slightly different exchange reaction (see below). We assume further that a chemical compound, *e.g. glucose*, has the same identifier in all networks.

We build the joint network as follows. First, we create a compartment for each metabolic network. These "species compartments" are surrounded by an environmental compartment. We build the union of the sources of all networks, that is $\mathcal{X} = \bigcup_i \mathcal{X}_i$, with $i = 1, \ldots, n$. A source that appears in several sets of sources, occurs once in $\mathcal{X}$. For each source $x \in \mathcal{X}$, we

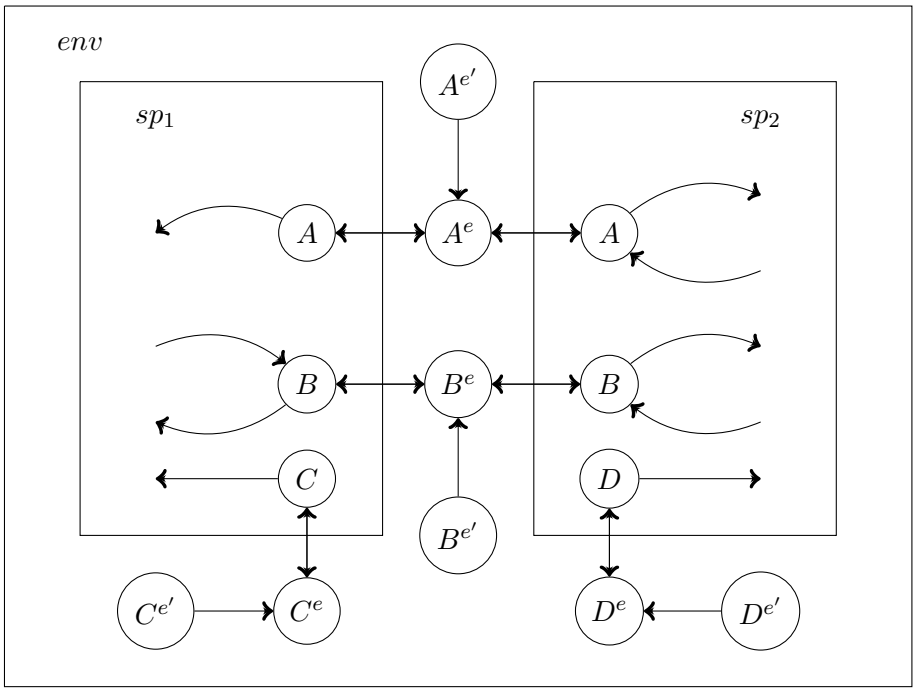


Figure 5.1: Joint network model with two species.

create two compounds $x^e, x^{e'}$ and two reactions: (i) a reversible exchange reaction $x \leftrightarrow x^e$, and (ii) a supply reaction that consumes $x^{e'}$ and produces x^e. The former reaction exchanges the original source compound with the environment. It replaces the original exchange reaction from the empty set (see above). The compound x is replaced by $x^{e'}$ in $\mathcal{X}$. All sources in $\mathcal{X}$ are thus not produced by a reaction in the joint network. This transformation is illustrated for two species in Figure 5.1, where the original set of sources are $\mathcal{X}_1 = \{A, B, C\}$ and $\mathcal{X}_2 = \{A, B, D\}$. Note that compound C (D) can only be exchanged between species 1 (2) and the environment. The set of sources of the joint network contains $A^{e'}, B^{e'}, C^{e'}$, and $D^{e'}$. The original exchange reactions must be examined carefully in this modeling step. Usually the steady state assumption holds in metabolic networks. We observed different techniques that are used to assure the steady state on compounds that are on the border between the cell and the environment. All three techniques depicted in Figure 5.2 (one at the top of Figure5.2a and two at the top of Figure 5.2b) model the exchange of compound A with the environment. In Figure 5.2a, the network contains an additional compound A^b (on the **b**oundary) that is excluded from the steady state constraint. A reversible reaction exchanges both compounds A and A^b with the environment. In Figure 5.2b, the compound A is exchanged with the environment via a reversible reaction (left). On the right of Figure 5.2b, the compound A is excluded from the steady state constraint by declaring it as a compound on the boundary (denoted by A^b). Thus, the compound A^b is imported from the environment if its net production is negative, and it is exported if its net production is positive. To build the joint model, we remove such export reactions (if present) and replace them as described above. The results of this transformation on the three exchange techniques are shown at the bottom of Figures 5.2a and 5.2b. In Figure 5.2a, the compound A^b becomes disconnected from the remaining network and we remove the compounds A^b, A^{b^e}, $A^{b^{e'}}$ from the network. It is still possible to model the exchange of compound A with the environment. The transformation of both techniques at the top of Figure 5.2b results in the same scheme where the exchange of compound A is modeled through the reversible reaction that consumes/produces A^e. The boundary label

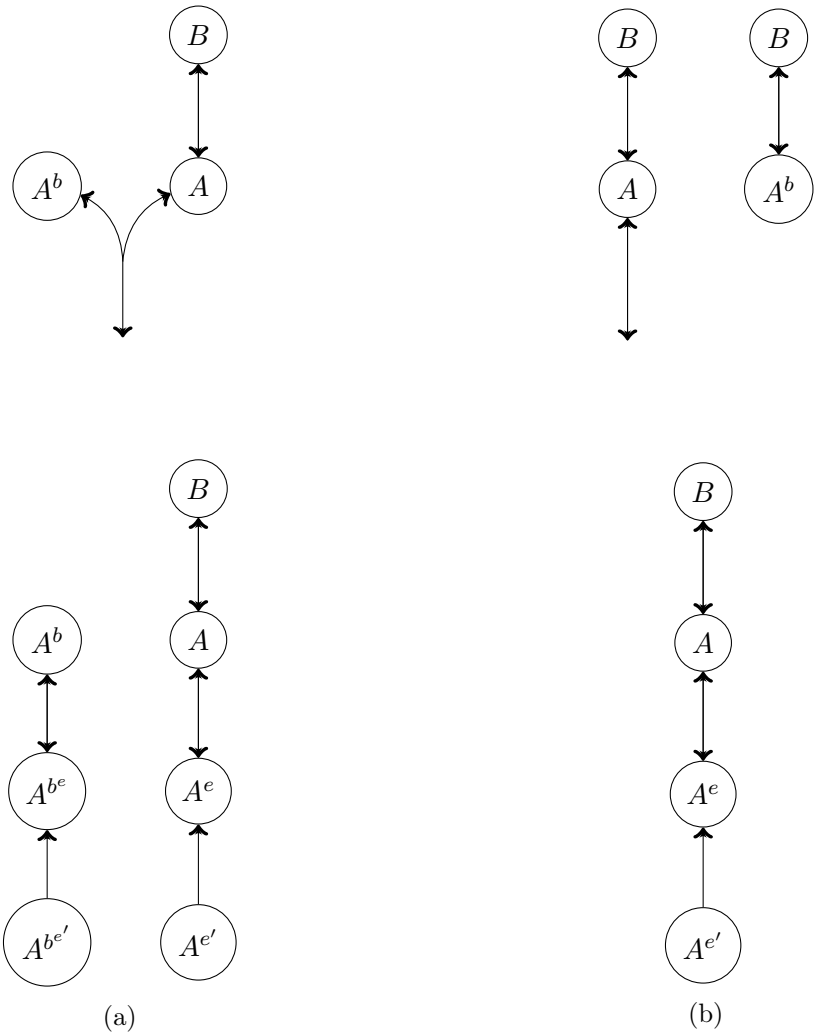


Figure 5.2: Different techniques to assure the steady state of compounds and their implications to the joint model.

(superscript b) is removed from A, which is thus under a steady state constraint in the joint model. In the case where we consider the steady state constraint, we exclude compounds in the environment (superscript e) from this constraint.

5.1 Species interaction characterization

5.1.1 Minimal media

Given the joint network $\mathcal{N} = (\mathcal{C}, \mathcal{R})$, we enumerate all minimal stoichiometric precursor sets that allow the production of all target sets $\mathcal{T}_1, \ldots, \mathcal{T}_n$. As in Chapter 3, for each source compound $x \in \mathcal{X}$, we add a *source-pool reaction* that produces x from the empty set. The collection of source-pool reactions is denoted by $\overline{\mathcal{R}}_{\mathcal{X}}$. The stoichiometric matrix of the transformed network is denoted by $\overline{S}$ .The following problem, that is similar to the problem (3.1)

1, 2	1	2	type of interaction
✓	✓	✓	competition/mutualism
✓	✓	✗	commensalism
✓	✗	✓	commensalism
✓	✗	✗	mutualism

Table 5.1: Each minimal precursor set $X \subseteq \mathcal{X}$ that enables the production of the target sets $\mathcal{T}_1$ and $\mathcal{T}_2$ in the joint network is tested in the network $\mathcal{N}_1$ ($\mathcal{N}_2$) whether it allows the production of $\mathcal{T}_1$ ($\mathcal{T}_2$)

in Chapter 3, illustrates how to obtain the first minimal solution:

$$\begin{aligned} \min f = \quad & \sum_{j=1}^{|\overline{\mathcal{R}}_{\mathcal{X}}|} b_j \\ s.t \quad & \left(\overline{S}v\right)_{\mathcal{T}_i} \geq \epsilon, && \forall i \in \{1,\ldots,n\} \\ & \overline{S}v \geq 0 \\ & b_j = 0 \leftrightarrow v_j = 0, && \forall j \in \overline{\mathcal{R}}_{\mathcal{X}} \\ & b_j \in \{0,1\}, && \forall j \in \overline{\mathcal{R}}_{\mathcal{X}} \\ & 0 \leq v_i \leq U, && \forall i \in \overline{\mathcal{R}}, \end{aligned} \tag{5.1}$$

where the binary variable b_j is associated to the flux of reaction j. If the flux value on reaction j is zero, then b_j takes value zero, and one otherwise. Note that the fluxes must be positive (last line). We enumerate all minimal stoichiometric precursor sets solving (5.1) recursively wherein at each iteration a constraint is added to exclude former minimal solutions from the solution space. We refer to Chapter 3 for further details.
For each obtained minimal precursor set $X \subseteq \mathcal{X}$, and each species $i = 1,\ldots,n$, we check via LP whether X is also a precursor set in $\mathcal{N}_i = (\mathcal{C}_i, \mathcal{R}_i)$ to produce $\mathcal{T}_i$. A prerequisite therefore is that $X \subseteq \mathcal{X}_i$. Let us consider two species and suppose that the target sets $\mathcal{T}_1$, $\mathcal{T}_2$ contain the compounds that are needed to produce the respective biomass. Thus, the production of the target sets is equivalent to growth of the organisms. A minimal precursor set in the joint network corresponds to the configuration where both species together grow in the given medium (consisting in the compounds of the minimal precursor set). To check if a species alone can produce its target set from the same minimal precursor sets signifies to put the organism alone in the medium and to verify its growth. If only two species are considered then all possible outcomes are shown in Table 5.1. The first column denotes that the joint metabolic network of both species can produce both target sets ($\mathcal{T}_1$, $\mathcal{T}_2$). The second and third column show whether species 1 or 2 can produce their respective target set on their own from a given minimal precursor set of the joint model. If both species can produce their respective target sets then the species do not need the presence of the other. We will show in Section 5.2.3 and 5.2.4 that competition and mutualism can arise in this configuration. In the case where only one of the two species can produce its target compounds, then the one that cannot produce its target set is feeded by the other one (commensalism; second and third line in Table 5.1). If both species are not able to produce their target sets on their own, but are able to do so in presence of the other, then they cross-feed each other. We will describe these configurations and the corresponding models from game theory and economics in the next sections.
This approach is similar to the one of Klitgord and Segrè (2010) which investigated minimal media that induce species interactions. The authors however do not enumerate all minimal precursor sets. Furthermore, the interaction, when both species can produce their respective

sets of targets from a minimal precursor set of the joint network (first line in Table 5.1), is classified as neutral. We show that the species interaction in this situation can take the form of competition or mutualism.

5.1.2 Exchanged compounds

To gather more insight into the species interaction on a given medium, one might be interested in all minimal factories that allow the production of the targets (biomass) of both species. This is however a difficult task for genome-scale networks as we have seen in Chapter 4. On a less detailed level, minimal sets of compounds that are exchanged in an interaction on a medium $X \subseteq \mathcal{X}$ may capture the essential of a species interaction. We investigate the latter and prune the network for that purpose. We remove the supply reaction ($A^{e'} \rightarrow A^{e}$) of each source that is not part of X. We furthermore split all reversible reactions of the network into a forward and backward reaction. The collection of reactions $\mathcal{R}_{ex}$ denotes all import and export reactions (reversible exchange reactions between the species compartments and the environment before the split) that are considered to capture the flow of exchanged compounds. Notice that all import and export reactions need to be considered in the case of the steady state assumption; four reactions per exchanged compound are taken into account in this case. When the accumulation of compounds inside a cell is permitted, then it is sufficient to consider only the export reactions (export the compound in the environment) because an exported compound by one species must be imported by another species. On the contrary, in steady-state, a compound may be exported by a species to fulfil the steady state constraint. In this case, the other species does not use this compound and it accumulates in the environment. The compounds in the environmental compartment are excluded from the steady state constraint in the case that we assume steady state.
Given a minimal precursor set $X \subseteq \mathcal{X}$, we find a first minimal set of compounds that are exchanged between the species in the joint network $\mathcal{N} = (\mathcal{C}, \mathcal{R})$ by solving the following MILP:

$$\begin{aligned} \min f = \quad & \sum_{j=1}^{|\mathcal{R}_E|} b_j & & \\ s.t \quad & (Sv)_{\mathcal{T}_i} \geq \epsilon, & & \forall i \in \{1, \ldots, n\} \\ & (Sv)_{\mathcal{C}\setminus X} \diamond 0 & & \\ & b_j = 0 \leftrightarrow v_j = 0, & & \forall j \in \mathcal{R}_{ex} \\ & b_j \in \{0, 1\}, & & \forall j \in \mathcal{R}_{ex} \\ & 0 \leq v_i \leq U, & & \forall i \in \mathcal{R}, \end{aligned} \tag{5.2}$$

where the $\diamond$ symbol can be replaced by the $=$ or the $\geq$ sign dependent on whether steady state is assumed or not. A binary variable b is associated with each import or export reaction; b_i takes the value zero if the flux value v_i of reaction i is equal to zero. Otherwise b_i is equal to one. The source compounds are excluded from the constraint in the second line. In steady state, the compounds in the environment compartment are excluded from any constraint as mentioned above.
Let the pair (v^*, b^*) be an optimal solution of Problem (5.2). Then, the support of b^* is denoted by I_{b*}. Note that under the steady state assumption not all export reactions in I_{b*} contribute to the exchange of compounds with another species. These reactions are part of the solution to assure steady state of the exported compound as illustrated in Figure 5.3. In this figure, the minimal precursor set $X = \{A^{e'}, G^{e'}\}$ sustains growth of both species in the joint metabolic network. Species sp_2 can however not grow alone on X. This configuration represents thus commensalism between species 1 and species 2 where the former exports and the latter imports the compound F. Species 1 needs to export furthermore the compound B from the network to

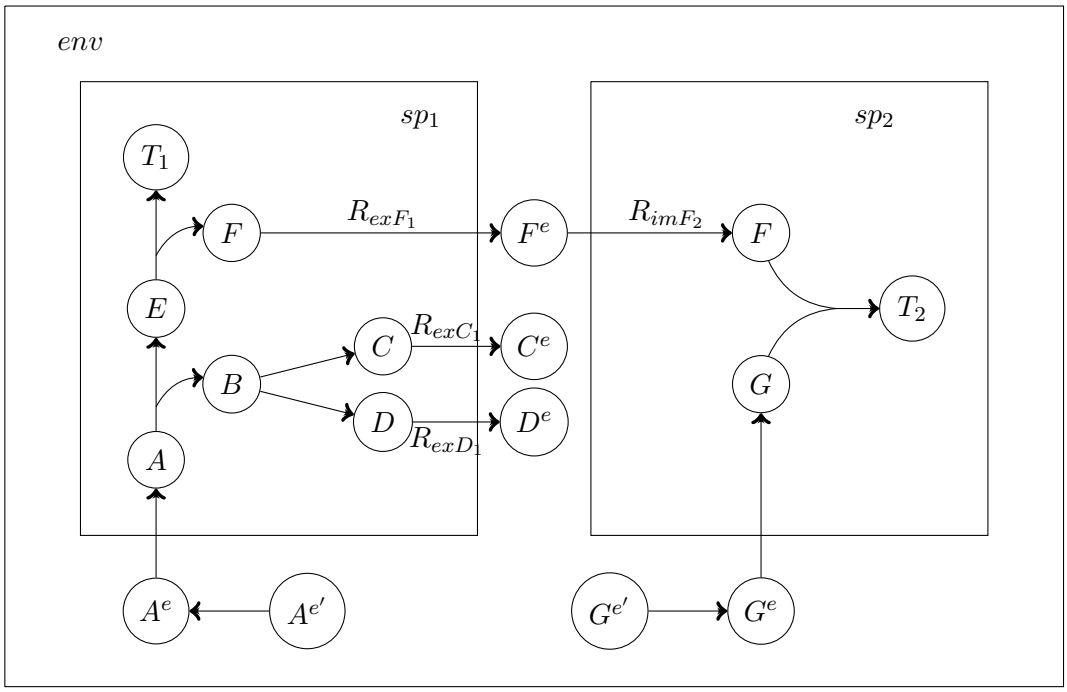


Figure 5.3: Under the steady state assumption, some export reactions are part of a solution of Problem (5.2) to assure steady state of the compounds and not to provide them for exchange with another species. The compounds C or D are exported to the environment to eliminate B from species sp_1. Thus the support of a solution to Problem (5.2) is either $\{b_{R_{exF_1}}, b_{R_{imF_2}}, b_{R_{exC_1}}\}$ or $\{b_{R_{exF_1}}, b_{R_{imF_2}}, b_{R_{exD_1}}\}$

maintain the steady state. This can be achieved through its transformation into the compound C or D and the export of one of the latter. Exporting either C or D sets the associated binary variable to one. However, neither C nor D are consumed by species 2 and thus accumulate in the environment. This may be a useful information, but in our case we need to remove the associated binary variables from the support I_{b*} as we are interested only in the compounds that take part in an exchange between the different species. In the example of Figure 5.3, there are two possible solutions to Problem (5.2) and the associated supports of the binary variables are $\{b_{R_{exF_1}}, b_{R_{imF_2}}, b_{R_{exC_1}}\}$ or $\{b_{R_{exF_1}}, b_{R_{imF_2}}, b_{R_{exD_1}}\}$. Removing one of the export reactions (either C or D) yields in both cases the solution $I_{b*} = \{b_{R_{exF_1}}, b_{R_{imF_2}}\}$.
As for the enumeration of minimal stoichiometric precursor sets in Chapter 3, we add the following constraint to Problem (5.2) to exclude the solution I_{b*} from the solution space:

$$\sum_{j \in I_{b^*}} b_j \leq |I_{b^*}| - 1. \tag{5.3}$$

We solve Problem (5.2) recursively; at each iteration we add a constraint of the form of (5.3) to the next MILP problem to obtain a minimal solution with respect to the solutions found so far. If there is no feasible flux to Problem (5.2) then we claim that all minimal sets of exchanged compounds are found.

Note that we introduce two, respectively four integer variables per exchanged compound to fully describe the imports and exports when accumulation is allowed or the steady state is considered. Actually we do not just enumerate which compounds are exchanged but also in which direction. This may however be impractical if a large number of exchanged compounds

are considered. In this case, we decided to at least detect the exchanged compounds ignoring which species imports or exports these compounds. For each exchangeable compound, we introduce an integer variable that takes the value one if it is produced (that is, it is exported by a species), and zero otherwise. We may miss solutions compared to the method described above in the following case: Suppose that species 1 exchanges the compound A for the compounds B and C from species 2. Alternatively, species 1 provides the compound B for the compound A produced by species 2. The original method detects both solutions because the minimization is actually done on the import and export reactions. The method that enumerates the exchanged compounds ignoring the direction detects only that the compounds A and B are exchanged. We denote the collection of exchangeable compounds by $\mathcal{C}_{Ex}$. The enumeration of all minimal sets of exchanged compounds can be done in a similar way as shown above. We obtain a first minimal set of exchanged compounds of minimum size solving the following MILP problem:

$$\begin{aligned} \min f = & \sum_{j=1}^{|\mathcal{C}_{Ex}|} b_j \\ s.t \quad & (Sv)_{\mathcal{T}_i} \geq \epsilon, && \forall i \in \{1,\ldots,n\} \\ & (Sv)_{\mathcal{C}\setminus X} \diamond 0 \\ & b_j = 0 \leftrightarrow v_j = 0, && \forall j \in \mathcal{C}_{Ex} \\ & b_j \in \{0,1\}, && \forall j \in \mathcal{C}_{Ex} \\ & 0 \leq v_i \leq U, && \forall i \in \mathcal{R}, \end{aligned} \tag{5.4}$$

Let the pair (v^*, b^*) be an optimal solution of Problem (5.4). Then, the support of b^* is denoted by I_{b*}. The constraint (5.3) can be added to subsequent formulations of Problem (5.4) to exclude already found solutions from the solution space.

In the case where both species can grow either together or independently from each other on a given medium (first line of Table 5.1), no compound is necessarily exchanged. However, contrary to Klitgord and Segrè (2010), we do not assume a neutral relationship between the species because (i) it may be advantageous for both species to exchange compounds, *e.g.* to achieve a higher growth yield, or (ii) there might be a competition for the compounds of the medium. If there is only one compound that is used by both species then we state that both species are trapped in the yield versus rate dilemma (Pfeiffer et al., 2001). We will describe this model in more detail in Section 5.2.3. If more than one compound is taken up from the media by both species, then models from economics such as the comparative advantage principle (Ricardo, 1817) or the general equilibrium theory (Varian, 2009) may be applied.

5.2 Modeling species interaction

In the following sections, we examine game theoretical approaches and models from economics that are applied to the species interactions shown in Table 5.1. Notice that mutualism occurs twice (first and last line in Table 5.1). We treat these cases separately because in the configuration where both species can grow independently on the joint minimal medium the mutualism is facultative, whereas mutualism is indispensable in the case where both species can not grow independently on such medium.

5.2.1 Obligate mutualism

In this section, we consider the case where both species are not able to grow alone in a given medium, but both species cohabitating in the same medium sustain growth (lines 4 of Table 5.1)

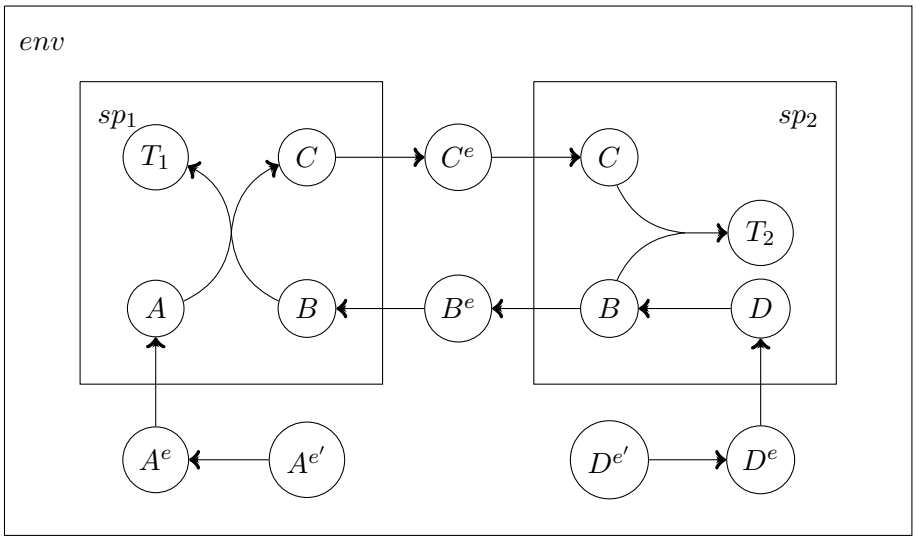


Figure 5.4: An obliged mutualism between two species when growing on the compounds A and D. Only species 2 is in the dilemma to export the compound B to maintain the mutualism with species 1 while B is also used to produce T_2. Species 1 exports the compound C.

	C	D
C	2	0
D	3	1

Figure 5.5: The payoffs of the Prisoner's Dilemma.

In this configuration, there is an exchange of compounds in both directions (if we consider only two species). We can however distinguish the configurations where (i) a species is constrained to export a compound, *e.g.* due to toxicity or to avoid accumulation, or (ii) a species exports a portion of a compound that it could use for its own interest. These configurations are illustrated in Figure 5.4. Here, both species can grow together in the medium $\{A^{e'}, D^{e'}\}$ if they exchange the compounds B and C. However, species 1 needs to export the compound C to avoid its accumulation inside the cell. On the contrary, species 2 is in the dilemma to either use the compound B to produce its target compound or to export a portion of it to the environment whereat the export is associated with an additional cost. A cheater strain of species 2 that does not export the compound B has thus an advantage compared to a cooperative strain in a spatially homogeneous environment. However if no individual of the species 2 cooperates then both species starve. The dilemma of species 2 in Figure 5.4 can be modeled as the prisoner's dilemma in game theory whose payoff matrix is depicted in Figure 5.5.

Many game theoretical approaches study how cooperation can evolve and be maintained in such a dilemma. Before explaining some of them in more detail we show how the configurations (i) and (ii) can be recognized. For this purpose and a given minimal precursor set of the joint network as well as a minimal set of exchanged compounds, we remove from the network all exchange reactions that take part neither in the uptake of the sources of the minimal precursor set nor in the import or export of compounds in the minimal set of exchanged compounds. For each species separately, we check if it can produce its target from the sources of the minimal precursor set and the minimal set of exchanged compounds without exporting compounds to

the environment. If this is the case then the species is in the dilemma to provide compounds to the environment that it needs to produce its target. In Figure 5.4, species 1 can produce the target T_1 from A and B but it exports the compound C to the environment to avoid its accumulation. On the contrary, species 2 uses completety the compounds C and D to produce the target T_2 without being "obliged" to export a compound.

Herein, it would be more beneficial for the players if everybody cooperated (sum of the payoffs is 4). However, defection is the dominant strategy, hence the predicted outcome of the game. Cooperation is nevertheless observed in nature. The question that arises is how cooperation can be established and maintained? It is assumed that a cooperator pays a cost c to receive a benefit b while a defector receives the benefit without paying a cost. Martin A. Nowak (Nowak, 2006) demonstrates that the cost-benefit ratio is a critical value for the establishment of cooperation in a population. He discusses five concepts for the evolution of cooperation.

First, the *Hamilton's rule* (Hamilton, 1964) accounts for the relatedness between the players. The relatedness r is defined as the probability to share a gene. Two players cooperate if their relatedness exceeds the cost-benefit ratio c/b. This idea is also known as *kin selection* or *inclusive fitness* (Nowak, 2006).

Second, *direct reciprocity* requires that the same two individuals of a population meet each other repeatedly. At each encounter, they have the choice to cooperate or defect. Robert Axelrod organized a round-robin computer tournament to identify the best strategy for the repeated prisoner's dilemma. Axelrod (1984) proposed a strategy, called *tit-for-tat*, that cooperates in the first game. In all subsequent rounds, a player that uses the tit-for-tat strategy repeats the action that the opponent player did in the previous round. If the opponent player cooperates in round i, then the tit-for-tat player cooperates in the round $i + 1$. This strategy is however vulnerable against itself if mistakes are possible. If two tit-for-tat players play against each other and one of them defects by mistake in the first round then both players will defect in all subsequent rounds – which is not the maximal achievable payoff. Alternative strategies were proposed such as generous tit-for-tat (cooperates sometimes even if the opponent defected in the last round) or win-stay, lose-shift (a player repeats its action of the last round if he did well and changes otherwise). All possible strategies have in common that they can only lead to cooperation if the probability of another encounter of the same two players is greater than the cost-benefit ratio.

Third, *indirect reciprocity* assumes again that the game is played repeatedly but without the requirement that the two players must meet again. In each game, one player acts as a donor, and the other player acts as a recipient. The donor decides whether he gives money to the recipient or not. A part of the population is able to observe this interaction. It was shown that reputation leads to cooperation if the probability to know the reputation of a player exceeds the cost-benefit ratio (Nowak, 2006). This means that a donor gives money to a recipient more readily if the reputation of the latter is higher. The reputation of a player increases when he gives money in the presence of witnesses. This concept should not play a role in our context because reputation seems not to be important to microorganisms.

Fourth, *spatial structure* or *network reciprocity* may lead to cooperation. A graph can be used to model (i) the individuals of a population, and (ii) the interactions between them, where the vertices correspond to the former and the edges to the latter. A cooperator pays a cost c for each neighbor to receive the benefit b. A defector pays no cost, and its neighbors (adjacent vertices) receive no benefit. Cooperation appears in network clusters if the benefit-cost ratio exceeds the average number of neighbors (Nowak and May, 1992; Nowak, 2006).

Fifth, *group selection* or *multi-level selection* considers that natural selection acts on individuals *and* on groups. The idea is to divide a population into groups. Contrary to defectors, the cooperators help each other within a group. The reproduction of an individual is propor-

tional to its payoff. If the number of individuals in a group exceeds a given threshold, then the group splits into two and another group is replaced to meet the constraint on the total population size. Selection favors defectors within groups as they have a higher payoff and thus a higher reproduction rate compared to cooperators. On the other hand, selection favors cooperators between groups because pure cooperator groups grow faster and thus split more often than mixed or pure defector groups. Group selection allows the evolution of cooperation if $b/c > 1 + (n/m)$, where n is the maximum group size and m is the number of groups (Traulsen and Nowak, 2006; Nowak, 2006).

Archetti *et al.* (Archetti et al., 2011a) discrimate three components in the evolution of cooperation in the example of the interaction between the bacterium *Vibrio fischeri* and the bobtail squid *Euprymna scolopes*, where the former emits light and the latter offers food in return. The amount of the bacterium's luminous emittance in the direction of the sea floor corresponds to the amount of moon light that hits the top of the squid. The bobtail squid is thus more difficult to spot from below as its shadow is camouflaged. Every day at dawn the bacteria is expulsed from the bobtail squid Ruby and McFall-Ngai (1999). This interaction raises the following questions. First, how does the bobtail squid manages to acquire the bacterial species with the desired property (*hidden characteristics* problem). Second, once the bacterium is inside the host, how does the latter assures that the bacteria emits light (*hidden actions* problem)? Third, there is the *collective action* problem that corresponds to the dilemma stated above. Why does an individual bacterium emits light if it would be more beneficial for itself to profit from the nutrients provided by the host without paying the cost for the luminous emittance?

The classical example of the hidden characteristic problem consists in the mating market where females choose partners from a population (Ronald Noë, 1994). Females prefer high-quality males but only the males know their own quality. Thus, the low-quality males are not interested to exhibit their quality. How does the females select the high-quality males? One solution is *signalling* (Grafen, 1990; Maynard Smith and Harper, 2003) where the males display costly phenotypes, *e.g.* the colorful plumage of birds. The cost can be carried out only by high-quality males. If signalling is impossible, then the *screening* concept from microeconomics may be an alternative (Rothschild and Stiglitz, 1976). Herein, the principal makes a proposition consisting in a reward and some cost to the agent who decides to accept or not, *e.g.* the boss of an enterprise makes a job offer that a worker can accept or not. The proposition should be set up in such a way that only high-quality agents accept it. Archetti *et al.* (Archetti et al., 2011b) show that the concept of screening can be applied to the interaction between the bobtail squid and *Vibrio fischeri*. The former makes the proposition that consists in providing food and producing reactive oxygen species (ROS) which are lethal for bacteria. Only bacteria that pay the cost to express the enzyme *luciferase* which consumes O_2 and thus prevents the squid to produce more ROS accept the proposition. A by-product of the enzyme *luciferase* is light which is the desired service of the bobtail squid.

Once the symbiont has entered the host, the question is how the host can prevent the symbiont from cheating. Two concepts exist in the literature: host sanction and partner fidelity feedback. In the concept of host sanction, the host employs punishments to maintain cooperation. On the contrary, partner fidelity feedback describes the configuration where the benefits provided by the symbiont to the host feeds back to the symbiont. In other words, the more the symbiont cooperates, the higher is the fitness of the host which in turn offers a higher benefit to the symbiont (Weyl et al., 2010). Host sanction is widely applied to species interactions between, *e.g.* yucca plants and yucca moths (Pellmyr and Huth, 1994), legumes and nitrogen-fixing bacteria (West et al., 2002; Kiers et al., 2003; Simms et al., 2006), plants and ants (Edwards et al., 2006), plants and mycorrhizal fungi (Bever et al., 2009), and between the fig tree and the fig wasp (Jandér and Herre, 2010). Bull and Rice introduced the

term *partner fidelity* in the context of mutualism and distinguished this concept from partner choice where in the latter the partner is chosen before the observation of a behavior of the symbiont (Bull and Rice, 1991). On the contrary, host sanctions and partner fidelity feedback is effectuated after the observation of the behavior of the symbiont. Weyl *et al.* (Weyl et al., 2010) state that partner choice and host sanction is often wrongly used in the literature and that the interactions above could be better explained by partner fidelity feedback.

Archetti *et al.* (Archetti et al., 2011a) show that cooperation in an N-player dilemma can be maintained without the above discussed mechanism, *e.g.* kin selection, etc. (Nowak, 2006). In an N-player game, cooperators pay a cost to contribute to the public good that is then transformed and redistributed to every player. If the amount of public goods grows linearly with the number of cooperators, then the game is called the N-player version of the prisoner's dilemma (NPD) (Hamburger, 1973). Cooperators pay a cost c, the sum of the contributions is multiplied by a reward factor $r > 1$ and the benefit b is equally distributed to all individuals of the population of size N. If $r/N < 1$ all individuals defect as in the two player prisoners dilemma. On the contrary, if $r/N > 1$, there is no dilemma and everybody cooperates. If one considers non-linear public goods, cooperation is naturally maintained. Archetti *et al.* (Archetti et al., 2011a) argue that non-linear public goods are widely present in nature, *e.g.* the benefit of enzymes is a saturating, sigmoid or step function of the concentration of the enzyme. In the case of a Heaviside step function (its value is zero for arguments that are below a threshold, and the value is one for arguments above a threshold) where the benefit is only achieved if there are at least k cooperators, the game is called a volunteers dilemma if $k = 1$ or a teamwork dilemma if $k > 1$. Again, each individual prefers to defect but if there are not k cooperators then everybody suffers. The equilibrium frequency of cooperators in large groups turns out to be approximately k/N if the cost-benefit ratio is below a certain threshold. An individual cooperates with a certain probability which depends on the cost-benefit ratio, the group size, and k (if $k > 1$). If $k - 1$ other individuals cooperate, then the best strategy is to cooperate, while the best strategy is to defect if there are already k cooperators. The mixed equilibrium is however often not pareto efficient, that is at least one individual can achieve a higher fitness without harming another one. An additional mechanism such as kin selection may improve the fitness. In another publication, Archetti and Scheuring show that cooperation can be maintained in interspecific mutualism if the species trade non-linear public goods (Archetti and Scheuring, 2013).

Another concept that allows the coexistence of cooperators and defectors without mechanisms such as kin selection, punishment or repeated encounters is proposed by (Hauert et al., 2002). Here, a third strategy, a "loner", is introduced with the characteristic that it outcompetes defectors and that it is outcompeted by cooperators. Given that defectors outcompete cooperators, this configuration is similar to the rock-paper-scissor game. In a well-mixed population, there is a cyclical dynamics: if most individuals cooperate, then it is better to defect. When the defectors are the majority in the population, then it is more beneficial to act as a "loner". If most individuals are loners, then the best strategy is to cooperate. Individuals may update their strategy in different ways by adopting (i) the strategy of a random individual with a probability proportional to the payoff difference (if positive), (ii) the strategy of a random individual if the payoff of the latter is higher, or (iii) the best-reply given the current composition of the population. Hauert *et al.* show that there exists either a mixed equilibrium with stable orbits circling around (update mechanism (i) and reward factor $r > 2$), stable oscillations (ii), or damped oscillations that converge to a stable polymorphism (Hauert et al., 2002).

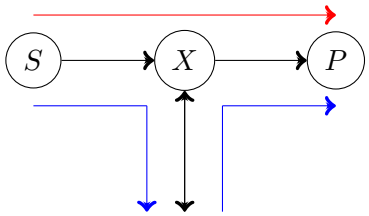


Figure 5.6: A simplified illustration of different strains in the studies of Doebeli (2002) and Pfeiffer and Bonhoeffer (2004). Either a single strain transforms the substrate S into a product P (red arrow) or two specialized strains evolve (blue arrows). One consumes S and exports X while the other consumes X and produces P.

5.2.2 Commensalism

In this section, we consider the case where one species is able to grow alone in a given medium and the other is not. Both species cohabitating in the same medium sustain growth (lines 2 and 3 of Table 5.1)

Doebeli (2002) and Pfeiffer and Bonhoeffer (2004) studied under which conditions commensalism can evolve. They focused on the situation where at the beginning, there is one strain that transforms completely a source (S) into a product (P). The population may evolve into one strain that specializes on the consumption of the substrate S and that exports an intermediate or waste compound X (on the path from S and P), and on the other hand, a second strain that specializes on the consumption of the exported compound X and that produces the product P. This configuration is depicted in Figure 5.6.

Doebeli (2002) proposes a model based on *adaptive dynamics* and the assumption that there is a trade-off between the uptake efficiencies on the primary substrate S and the waste product X. He shows that a monomorphic population (all individuals consume S) evolves gradually towards a population that consists in two resource specialists where one feeds the other as described above. The evolutionary branching is predicted to be more likely in chemostat cultures than in serial batch cultures, because in the latter, the waste product is present at appreciable concentrations only for a relatively short time. The conditions for the specialist on the waste product are thus harsher (Doebeli, 2002).

Pfeiffer and Bonhoeffer (2004) assume that an organism maximizes the rate of ATP production, and minimizes the concentration of enzymes and intermediate compounds on the pathway from the source substrate S to the product P. They simulate the competition between strains that differ in the level of enzyme expression, *e.g.* a single strain that transforms S into P expresses at a certain level the enzyme of the reaction $S \to X$ and $X \to P$, while partial degraders express either $S \to X$ and $X \to$, or $\to X$ and $X \to P$. The population dynamics are expressed by ordinary differential equations for the substrate S, for the intermediate compound X, and for the subpopulation N_i of each strain i. The steady state of the compounds concentrations and the population size is computed for a given condition. The authors determine a strain with highest growth rate under the conditions of the steady state and allow it to invade the population. The compuation of the steady state and the determination of the highest growth rate strain are done recursively until there is no strain that can invade the resident population. This approach is depicted in Figure 5.7 where a strain (blue bullet) is given to the medium (Figure 5.7a). When steady state is reached, a strain with highest growth rate under these conditions is determined (red bullet in Figure5.7b). This strain eventually invades the population and reach another steady state. This is done repeatedly until no strain can invade the population (illustrated by the strain (green bullet) in Figure 5.7c). This characteristic corresponds to an evolutionary stable strategy (see Section 2.2.1) even though

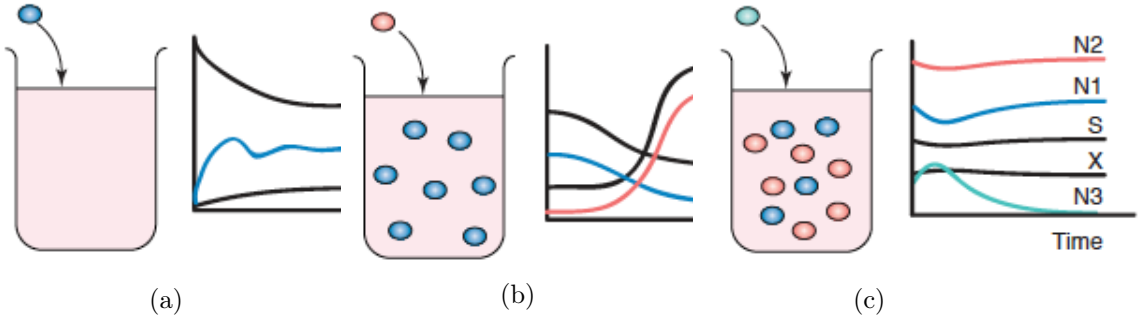


Figure 5.7: (From (Pfeiffer and Schuster, 2005)) Illustration of the approach of Pfeiffer and Bonhoeffer (2004)

no payoff matrix is used in this approach. The authors predict that a split transformation of S into P by several strains is favored at high resource levels, while complete transformation by a single strain is favored at low resource concentrations. A split transformation is more likely if high intermediate concentrations are associated with high costs, *e.g.* toxicity. At intermediate dilution rates in a chemostat, the following coexistences are found: (i) a strain that does the complete transformation together with strains that partially transform S into P, and (ii) strains that partially transform S into P. Contrary to the approach of Doebeli (2002) where the commensal interaction evolves gradually, the interaction evolves in two steps in the method of Pfeiffer and Bonhoeffer (2004): First, a strain that transforms the substrate S and exports the intermediate compound X invades a population of a strain that completely transforms S into P. In a second step, a strain that consumes the accumulated intermediate compound X and produces P invades the population.

5.2.3 Competition

Species that are able to grow individually and in co-culture on a given media (first line of Table5.1) may be trapped in the yield versus rate dilemma presented in Pfeiffer et al. (2001). The issue is described at the example of the degradation of energy rich compounds, *e.g.* glucose, to obtain energy. Organisms have alternative pathways to produce energy in form of ATP. Some energy must be however invested in these pathways, *e.g.* for the production of enzymes. Investing more energy into the pathways increases the rate of ATP production but diminishes the ATP yield. This yield versus rate trade-off can be observed in the sugar degradation which can be accomplished (if oxygen is present) through respiration and fermentation. The former achieves a high yield at low rate whereat the latter less ATP is produced but at a higher rate. An organism that produces more ATP grows faster. Approaches, *e.g.* flux balance analysis, that optimize a given criteria, *e.g.* growth or ATP production, predict respiration in this configuration Pfeiffer and Schuster (2005). On the contrary, the game theoretical approach of Pfeiffer *et al.* (Pfeiffer et al., 2001) assigns fermentation to be the best strategy. This is plausible when one considers that a fermenter takes up rapidly glucose from the media leaving little or nothing for the respirators. Respiration would be the best strategy for the population at a whole (higher growth) and can be seen as a cooperative action. This configuration reflects what is known as the tragedy of the commons or the N-player prisoner's dilemma. It is also notable that fermentation actually reduces fitness (Pfeiffer et al., 2001; Pfeiffer and Schuster, 2005)

5.2.4 Facultative mutualism

We have seen that species that are able to grow on a medium, either individually or in co-culture, may compete for compounds due to the yield versus rate dilemma. Let us consider a slightly different situation where both species need to uptake more than one same compound from the medium, *e.g.* both species need to uptake the compounds A and B. The concept of comparative advantage (Ricardo, 1817) is an economic theory that can be best explained by an example of two countries that produce both two goods A and B. Consider that country 1 has a lower relative opportunity cost (compared to country 2) for the production of good A. In the opposite direction, country 2 has a lower relative opportunity cost for the production of good B. The opportunity cost of good A can be defined as the amount of good B that must be sacrificed to produce another unit of good A. Following the theory of comparative advantage, it is beneficial for both countries to specialize in the production of the good for which they have a comparative advantage and to trade the goods. This holds even if one country has an absolute advantage in the production of both goods. This idea is applied to mutualism in the literature (Mark W. Schwartz, 1998; Hoeksema and Schwartz, 2003; Wyatt et al., 2014; Tasoff et al., 2015; Enyeart et al., 2015; Wyatt et al., 2016). The mutualism between plants and mycorrhizal fungi is often used as an example where the former exchange carbohydrates against phosphorus with the latter. While Schwartz and Hoeksema (Mark W. Schwartz, 1998; Hoeksema and Schwartz, 2003) analyzed when trade is beneficial for both species, Wyatt *et al.* (Wyatt et al., 2014) went further by considering conditions under which the trade is evolutionary stable.
Enyeart *et al.* propose a model to engineer synthetic gene circuits in bacteria. The authors consider two bacterial species that live in an environment with two antibiotics. To be able to grow, both species must produce an antibiotic-resistance protein against each antibiotic. The gene for such a protein is only expressed if the appropriate signaling molecule is present. The authors demonstrate that both species grow better together when they trade the signaling molecules given that each bacterium has a comparative advantage in the production of one signaling molecule.
Tasoff et al. (2015) apply the general equilibrium theory from economics (Varian, 2009) to the mutualistic trade of compounds between microorganisms. The authors assert that a comparative advantage is a necessary condition in order that a trade takes place. The general equilibrium theory models a centralized market of goods that are sold and acquired by agents. The agents enter the market with the goal to sell their goods and buy certain goods. Given the prices for these goods, the market is said to be in equilibrium if the supply equals the demand, that is, there is no good that accumulates or that is undersupplied. The general equilibrium theory can be applied to the trade of several goods between several agents (Tasoff et al., 2015).

5.2.5 Use of minimal sets of precursors and exchanged compounds

In this section, we show how the concepts of minimal precursor sets and minimal sets of exchanged compounds can be integrated into models from game theory and economics. A game is composed of a set of players, a set of actions per player, and an utility (payoff) function that assigns a value to each player that depends on the strategies chosen by each player. The payoff values are usually presented in a matrix form where the matrix entry $[i, j]$ represents the payoff values for player 1 and 2 given that player 1 plays strategy i, and player 2 plays strategy j. The players are assumed to be rational so that they know the components of the game (players, actions, payoffs) and that they choose a strategy to maximize their payoffs. The players in our case are microorganisms that are associated to their metabolic network. Microorganisms are not supposed to be rational, in the sense that they choose a strategy

after reasoning. Nevertheless, game theory can be applied to microorganisms because natural selection leads them on the long term to choose their best action to maximize fitness. We suppose that a game is played on a given minimal precursor set of the joint metabolic model. In the case where such a minimal medium induces commensalism or obligate mutualism, we build the actions of each species based on the minimal sets of exchanged compounds. For each minimal set of exchanged compounds, we infer the action of both species from the compounds that they import and export, *e.g.* if the minimal precursor set consists in the metabolite A that is taken up by both species, and species 1 provides the metabolite B to species 2, which in exchange provides the metabolite C (obligate mutualism), then the action of species 1 can be expressed as $A + C \rightarrow B$ and the action of species 2 is $A + B \rightarrow C$. Usually, only two actions (cooperate and defect) are considered in game theoretical approaches in the context of cross-feeding. In our approach, each player possesses several actions that correspond to the different possibilities to exchange compounds.
The crucial point then is to determine a payoff function. It is notable that no payoff matrix is provided in the literature related to commensalism between microorganisms (Doebeli, 2002; Pfeiffer and Bonhoeffer, 2004). Assigning dummy payoffs to the players to simulate a known outcome is rather unsatisfying (Hummert et al., 2014). In our opinion, the payoff must reflect the cost-benefit ratio of producing and exchanging compounds. The cost should integrate the relative amount of resources that is invested into the production of an exchanged compound. The benefit must take into account whether or not the strategies of the different species are coordinated. Consider that two species can either exchange the compounds A against B, or C against D to sustain growth in co-culture. If species 1 provides A and species 2 provides D, then growth is not possible and there is no benefit for any species. The payoff would be even negative as the cost for producing A and D must be paid. Wintermute and Silver Wintermute and Silver (2010) studied the interactions between *E. coli* auxotrophs. The authors compute how each species values the exchanged compounds. This value corresponds to the marginal change in the objective function in a linear programming problem when a constraint is modified. The maximal biomass yield of the wild type is used as reference. The cost of exporting a compound m to the environment is the reciprocal of the minimum flux of the export reaction of m ensuring the production of at least 90% of the maximal biomass yield. If the minimum flux on the export reaction of m is zero, then the cost of compound m takes the value of infinity meaning that this compound can not be exported et any cost. In a similar way, the benefit to import a compound m from the environment is computed as the reciprocal of the minimum flux on the import reaction of m provided that at least 10% of the maximal biomass yield can be produced. If there is no solution to such an LP formulation, then there is no benefit of importing the compound m. In the case where more than one compound can be exchanged between a pair of auxotrophs, the mean benefit and cost over all exchanged compounds is considered.
In the case where the minimal precursor set of the joint model allows the independent growth of each species (first row of Table 5.1), the precursor set itself provides important information. If it consists of a single compound then we expect a yield versus growth dilemma. Facultative mutualism can be established under some conditions, *e.g.* if there is a comparative advantage. In both cases the cost and benefit must be determined to make a clear assessment.

5.3 Application

Benomar et al. (2015) studied the interaction between the Gram-negative sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough and the fermentative endospore-forming Gram-positive bacterium *Clostridium acetobutylicum*. The authors showed that in a poor medium for *D. vulgaris*, an interspecies cell-cell interaction is established that allows the exchange of

cytoplasmic material, *e.g.* proteins. This interaction is associated with an exchange of metabolites and a higher production of H_2 (Benomar et al., 2015). The authors defined therefore a medium (GY containing glucose, yeast extract, dibasic potassium phosphate, ammonium chloride, and minerals) that enables *Clostridium acetobutylicum* to grow in monoculture. The same medium is very poor for *Desulfovibrio vulgaris* which is reflected by the observation of survival of the bacterium in monoculture; growth however is not possible. Growth of both bacteria is observed in co-culture suggesting a commensal interaction between *Clostridium acetobutylicum* and *Desulfovibrio vulgaris* where the former provides nutrients to the latter. To investigate this hypothesis, the authors grew *C. acetobutylicum* on the GY medium, removed the cells of *C. acetobutylicum* after some time, and added a washed suspension of *Desulfovibrio vulgaris*. No growth of *D. vulgaris* was observed in the GY medium plus the metabolites produced by *C. acetobutylicum*. Another experiment in co-culture was conducted in which one of the bacteria was placed in a dialysis tube that allows the diffusion of small molecules but prevents a physical interspecies interaction (Benomar et al., 2015). Again, no growth of *D. vulgaris* was observed. This suggests that a cell-cell interaction is necessary for *D. vulgaris* to grow in co-culture with *C. acetobutylicum*. The physical interaction and the exchange of calcein, mCherry molecules, and green fluorescent protein was confirmed by fluorescence microscopy (Benomar et al., 2015).

The species interaction between *Desulfovibrio vulgaris* Hildenborough and *Clostridium acetobutylicum* is an exciting and at the same time a challenging case for our methods of minimal precursor sets and minimal sets of exchanged compounds. The metabolic network of *Clostridium acetobutylicum* was obtained from Senger and Papoutsakis (2008) and converted into SBML format (Hucka et al., 2003; Finney and Hucka, 2003). The metabolic network of *Desulfovibrio vulgaris* Hildenborough (version 19.0) was obtained from the Biocyc database (Caspi et al., 2014). The components of the GY medium was kindly provided by the authors of the paper Benomar et al. (2015).

For both bacterial species independently, we enumerated all minimal precursor sets that allow the production of biomass and accumulation of compounds (constraint $(Sv)_{\mathcal{C}\setminus X} \geq 0$). Accepting the accumulation of compounds is due to the fact that at the starting point of the collaboration with the authors of Benomar et al. (2015), only this version of the enumeration of minimal stoichiometric precursor sets was implemented.

We created in both metabolic networks a dummy compound T (the target compound) that was added as a product of the biomass reaction. The components of the GY medium were exclusively considered as source compounds. We find through the enumeration of minimal stoichiometric precursor sets that *Clostridium acetobutylicum* can grow on five minimal media that are each composed of two compounds: riboflavin plus either L-methionine, L-cysteine, thiamine, L-cysteinylglycine, or N-glycyl-L-methionine. No minimal precursor set was found for *Desulfovibrio vulgaris* which is consistent with the experiments conducted by (Benomar et al., 2015). We then built the joint metabolic network of both species (see at the beginning of this chapter) which was not a straightforward task because the metabolic networks were obtained from different sources and the compound identifiers were thus different. We matched the identifiers either through their SMILES codes (Weininger, 1988) or manually based on the compound names. We built exchange reactions for each compound that is found in both individual networks since, according to the observations in (Benomar et al., 2015), all molecules are susceptible to be exchanged. We created a dummy target T and a reaction that consumes the target compounds of each species and produces T. The joint network comprises 3646 reactions and 3197 compounds. The enumeration of all minimal precursor sets yields five solutions each containing a single source, that is either L-methionine, L-cysteine, thiamine, L-cysteinylglycine, or N-glycyl-L-methionine. These are the same sources as in the minimal precursor sets for *Clostridium acetobutylicum*. However, in co-culture, riboflavine can be

class	compounds
1	3-Aminopropionic acid, Phosphopantethein, N-(-R-4-Phosphopantothenoyl)-L-cysteine, Acetyl coenzyme A, Pseudouridine 5'-phosphate, CoenzymeA, Dephospho-CoA, D-4-Phosphopantothenate, Pantothenic acid, Uracil
2	FAD, Guanosine 3'-diphosphate 5'-triphosphate, 6-7-Dimethyl-8–1-D-ribityllumazine, 5-Amino-6–5-phosphoribitylaminouracil, GTP, 2,5-diamino-6-hydroxy-4-(5-phosphoribosylamino)pyrimidine, GDP-mannose, FADH2, 5-Amino-6–5-phosphoribosylaminouracil, GDP-4-keto-6-deoxy-D-mannose, 4–1-D-Ribitylamino-5-aminouracil, Riboflavin, Flavin mononucleotide
3	Carbamoyl phosphate, Dihydrogen carbonate, Citrulline
4	Sulfate, Adenylyl sulfate, Thiosulfate, Sulfite, Sulfide
5	meso-2-6-Diaminoheptanedioate, LL-2-6-Diaminoheptanedioate, Lysine

Table 5.2: The equivalence classes based on the 5850 minimal sets of exchanged compounds between *Desulfovibrio vulgaris* Hildenborough and *Clostridium acetobutylicum* when growing in co-culture on L-cysteinylglycine.

omitted. This result confirms that growth of both species is possible in co-culture. The cell-cell interaction is reflected by the fact that not only compounds that have an exchange reaction in the original networks are exchangeable. The minimal precursor sets are currently examined for growth in the laboratory of the authors of the paper Benomar et al. (2015).

The observation that both bacteria can grow on a minimal precursor set in co-culture but not independently suggests a mutualistic interaction. Notice that *Clostridium acetobutylicum* is able to grow independently in the GY medium but that the minimal precursor sets in co-culture are minimal with respect to the minimal precursor sets in monoculture. To gather a more detailed insight into the species interaction, we enumerate the minimal sets of exchanged compounds. Due to the high number of exchanged compounds (483), the method (recursively solving problem (5.2)) presented in Section 5.1.2 is unfortunately impractical.

For this reason, we solve the alternative problem formulation (5.4) to enumerate all minimal sets of exchanged compounds ignoring which species exports or imports the compounds. The preliminary results for the minimal stoichiometric precursor set that contains the compound L-cysteinylglycine provide 5850 minimal sets of exchanged compounds of size six. Based on these solutions, we computed the equivalence classes of the compounds, where the compounds c_1 and c_2 are in the same equivalence class if c_1 can be replaced by c_2 in each solution and *vice versa*. In the present case, there are five equivalence classes which enable to cluster the solutions into one single solution that contains D-ribulose5-phosphate and one compound from each equivalence class. The compound D-ribulose5-phosphate is thus necessarily exchanged. The equivalence classes are presented in Table 5.2. To verify the correctness of the equivalence classes, one can, in the present case, build the product of the number of compounds per class and check if such product equals the total number of solutions (here: $10 \times 13 \times 3 \times 5 \times 3 = 5850$). It is striking that riboflavin (in equivalence class 2) is exchanged. As we saw above, this compound takes part in every minimal precursor set for *Clostridium acetobutylicum* in monoculture. Further analysis in collaboration with the authors of Benomar et al. (2015) will provide more insight into the role of each exchanged compound. Imaging mass spectrometry at different time points may reveal the exchange of compounds *in vivo*.

5.4 Conclusion and Perspectives

We showed how the methods of minimal precursor sets and minimal sets of exchanged compounds can provide an insight into a species interaction at the metabolic level at the example of the mutualistic relationship between the bacteria *Desulfovibrio vulgaris* Hildenborough and *Clostridium acetobutylicum*. The computed minimal precursor sets that allow growth of both species in co-culture are currently verified in the laboratory. Notice that the bacteria may grow very slowly as we do not maximize the production of biomass during the computation of minimal precursor sets. Here, a minimal precursor set assures, under the assumption that the metabolic network is well curated, that some amount of biomass can be produced. Due to the high number of variables and constraints, we were able to compute only a subset of the minimal sets of exchanged compounds. A convex relaxation approach as it is employed in Julius et al. (2008) may improve the performance of our methods and thus enable to compute all minimal sets of exchanged compounds including the direction of exchange. The provided candidate list of minimal sets of exchanged compounds could be validated experimentally through imaging mass spectrometry at different time points.
We mentioned briefly how the concepts of minimal precursor sets and minimal sets of exchanged compounds can be included into models of game theory and economics. Further investigation is necessary to determine the payoff function. In our opinion, game theory and models from economy are well suited to understand species interactions. It is exciting to see a similarity between markets in economics and trading species with the environment as central market place. Other principles from economics could find their application in species interaction, *e.g.* the avoidance of bad trading partners, the establishment of local business ties, diversification or specialization, monopolization of a market, or elimination of competitors (Werner et al., 2014). The black queen hypothesis (BQH) (Morris et al., 2012) states that many compounds that are exported to the environment are “leaky", that is the export of this public good is unavoidable. A species that does not express the gene of such a leaky function has a selective advantage unless the function is lost from the community. Species that retain the function are called helpers. The BQH is supposed to be responsible for genome reduction. Morris *et al.* postulate that the number of helper species should be small but always present as the public good needs to be produced for the community. The helper species thus corresponds to the definition of a “keystone species" (Morris et al., 2012) or a monopolist.

Conclusion and Perspectives

In this thesis, we presented methods to enumerate exhaustively different objects in metabolic networks: minimal stoichiometric precursor sets, minimal stoichiometric factories, and minimal sets of exchanged compounds. All of them can be applied to provide a deeper insight into species interactions at the metabolic level.

Concerning the minimal stoichiometric precursor sets, we discussed their relationship with the minimal topological precursor sets. Notably, we showed that minimal stoichiometric precursor sets can be obtained from combinations of minimal topological factories in the many-to-one transformed network. Unfortunately this approach is not efficient when applied on genome scale metabolic networks. This brought us to develop an alternative method, called Sasita, which is based on mixed integer linear programming. We applied Sasita on genome-scale metabolic networks of several *Escherichia coli* strains (comprising commensal and both intestinal and extraintestinal pathogens) to enumerate minimal stoichiometric precursor sets that allow growth. On one hand, we confirmed previous results, on the other hand we detected minimal precursor sets that were not found before by previous methods. The obtained solutions for the different strains allow to distinguish them in their ability to catabolise nutrients. Our method can thus be used to determine minimal growth conditions of organisms. Furthermore, we demonstrated that minimal stoichiometric precursor sets enable to distinguish different types of species interactions, namely facultative and obligatory mutualism, competition, and commensalism. This could also be valuable for industrial production in that alternative media may be cheaper or provide higher productivity.

The software Sasita is written in Java and is publicly available at http://sasita.gforge.inria.fr. Currently, Cplex (IBM ILOG AMPL/CPLEX 12.5.1) is used to solve the MILP models. We intend to develop a version that interacts with a publicly available solver such as SCIP. A database that stores minimal precursor sets for a given (joint) metabolic network and a set of sources and targets, may build a new resource of minimal media for mono- and co-cultures. We observed that he computation time to obtain the next minimal precursor set in Sasita increases with the number of already computed minimal solutions. The beforehand identification of source equivalence classes would provide a significant improvement as it implies that less solutions are enumerated.

It was then an obvious objective to enumerate all minimal stoichiometric factories from a given minimal precursor set to the set of targets. We introduced two alternative definitions of stoichiometric factories, one that is constrained to fulfil the steady state condition, and a second one that allows for an accumulation of compounds. The relationship between the former and elementary modes was discussed briefly. To establish a relationship between minimal stoichiometric factories allowing for an accumulation and chemical organizations will be addressed in the future. Although a chemical organization is defined as a set A of compounds, there is an associated set of reactions $\mathcal{R}_A$ that contains all reactions whose substrates are in A. Both stoichiometric factories and the set of reactions that is associated to a chemical organizations, are self-maintaining sets of reactions, that is, all compounds involved in a stoichiometric factory and set of reactions $\mathcal{R}_A$ are produced in a positive amount.

Chemical organizations have the additional constraint to be a closed set, that is, the products of the reactions in $\mathcal{R}_A$ must be part of A. We discussed how the network can be pruned before enumerating all minimal stoichiometric factories from a given minimal precursor set that allow the production of the set of targets. The same two methods that were already used for the enumeration of minimal stoichiometric precursor sets are used for the enumeration of minimal factories. Due to the network pruning and the concept of essential compounds, the bottleneck of the combinatorial approach can be shifted, at least in metabolic networks of medium size, towards the combinations of minimal topological factories in the many-to-one network that must be built to obtain minimal stoichiometric factories. The essential compounds build the basis of another idea, that of computing minimal reaction cut sets for the target production considering only the reactions that either consume or produce a given essential compound. Removing the reactions that are linked to an essential compound but are not part of a minimal cut set, has the advantage of enumerating less solutions in many cases. However, further investigation is required to find a plausible biological motivation to enumerate such a subset of factories. For this purpose, we consider the literature about robustness of metabolic networks as a next step.
Lastly, we demonstrated how minimal precursor sets allow the detection of species interaction inducing media. Computing the minimal sets of exchanged compounds between two species in co-culture provide a deeper insight on the interaction. We applied these methods to the interaction between the bacteria *Desulfovibrio vulgaris* Hildenborough and *Clostridium acetobutylicum*. Five minimal media that induce an obligatory mutualism were determined. The obtained results are currently examined in the laboratory by our collaborators.
We reviewed game theoretical approaches and methods from economics in the context of species interaction at the metabolic level. We think that an interesting game could arise when the three above mentioned concepts are included in such models, *e.g.* as available actions of a species. The next step before playing games is to develop a payoff function. On the other hand, market theory is studied since a long time in economics. We presented two concepts, the one of comparative advantage and the general equilibrium theory, which were already used to study species interactions. In our opinion, these and other concepts from economics are applicable to species interaction at the metabolic level. A longer-term aim is to study species interactions between more than two species.

Bibliography

Acuña, V., Chierichetti, F., Lacroix, V., Marchetti-Spaccamela, A., Sagot, M.-F., and Stougie, L. (2009). Modes and cuts in metabolic networks: Complexity and algorithms. *Biosystems*, 95(1):51 – 60.

Acuña, V., Milreu, P. V., Cottret, L., Marchetti-Spaccamela, A., Stougie, L., and Sagot, M.-F. (2012). Algorithms and complexity of enumerating minimal precursor sets in genome-wide metabolic networks. *Bioinformatics*, 28(19):2474–2483.

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2010). *Essential Cell Biology*. Garland Science, 3 edition.

Andrade, R., Wannagat, M., Klein, C. C., Acuña, V., Marchetti-Spaccamela, A., Milreu, P. V., Stougie, L., and Sagot, M.-F. (2016). Enumeration of minimal stoichiometric precursor sets in metabolic networks. *Algorithms for Molecular Biology*.

Antoniewicz, M. R. (2015). Methods and advances in metabolic flux analysis: a mini-review. *Journal of Industrial Microbiology & Biotechnology*, 42(3):317–325.

Archetti, M. and Scheuring, I. (2013). Trading public goods stabilizes interspecific mutualism. *Journal of Theoretical Biology*, 318:58 – 67.

Archetti, M., Scheuring, I., Hoffman, M., Frederickson, M. E., Pierce, N. E., and Yu, D. W. (2011a). Economic game theory for mutualism and cooperation. *Ecology Letters*, 14(12):1300–1312.

Archetti, M., Úbeda, F., Fudenberg, D., Green, J., Pierce, N. E., and Yu, D. W. (2011b). Let the right one in: A microeconomic approach to partner choice in mutualisms. *The American Naturalist*, 177(1):75–85. PMID: 21091210.

Ausiello, G., Franciosa, P. G., and Frigioni, D. (2001). Directed hypergraphs: Problems, algorithmic results, and a novel decremental approach. In *ICTCS*, pages 312–327.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Bader, J., Mast-Gerlach, E., Popović, M., Bajpai, R., and Stahl, U. (2010). Relevance of microbial coculture fermentations in biotechnology. *Journal of Applied Microbiology*, 109(2):371–387.

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

Bary, A. D. (1879). *Die Erscheinung der Symbiose*. Verlag von Karl J. Trübner.

Benomar, S., Ranava, D., Cárdenas, M. L., Trably, E., Rafrafi, Y., Ducret, A., Hamelin, J., Lojou, E., Steyer, J.-P., and Giudici-Orticoni, M.-T. (2015). Nutritional stress induces exchange of cell material and energetic coupling between bacterial species. *Nat Commun*, 6.

Bever, J. D., Richardson, S. C., Lawrence, B. M., Holmes, J., and Watson, M. (2009). Preferential allocation to beneficial symbiont with spatial structure maintains mycorrhizal mutualism. *Ecology Letters*, 12(1):13–21.

Brandenburger, A. (2007). Cooperative game theory: Characteristic functions, allocations, marginal contribution.

Broom, M. and Rychtar, J. (2013). *Game-Theoretical Models in Biology*. Chapman and Hall/CRC.

Bull, J. J. and Rice, W. R. (1991). Distinguishing mechanisms for the evolution of cooperation. *Journal of Theoretical Biology*, 149(1):63–74.

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

Carbonell, P., Fichera, D., Pandit, S. B., and Faulon, J.-L. (2012). Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. *BMC Systems Biology*, 6(1):1–18.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P., and Karp, P. D. (2014). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 42(D1):D459–D471.

Chung, B. K. S. and Lee, D.-Y. (2009). Flux-sum analysis: a metabolite-centric approach for understanding the metabolic network. *BMC Systems Biology*, 3(1):1–10.

Clark, S. T. and Verwoerd, W. S. (2012). Minimal cut sets and the use of failure modes in metabolic networks. *Metabolites*, 2(3):567–595.

Clarke, B. L. (1980). *Stability of Complex Reaction Networks*, pages 1–215. John Wiley & Sons, Inc.

Cottret, L., Milreu, P., Acuna, V., Marchetti-Spaccamela, A., Viduani Martinez, F., Sagot, M., and Stougie, L. (2007). Enumerating precursor sets of target metabolites in a metabolic network. *Lecture Notes in Bioinformatics*, 5251:233–244.

David, L. and Bockmayr, A. (2014). Computing elementary flux modes involving a set of target reactions. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 11(6):1099–1107.

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

Dejonghe, W., Berteloot, E., Goris, J., Boon, N., Crul, K., Maertens, S., Höfte, M., De Vos, P., Verstraete, W., and Top, E. M. (2003). Synergistic degradation of linuron by a bacterial consortium and isolation of a single linuron-degrading variovorax strain. *Applied and Environmental Microbiology*, 69(3):1532–1541.

Deville, Y., Gilbert, D., van Helden, J., and Wodak, S. J. (2003). An overview of data models for the analysis of biochemical pathways. *Briefings in Bioinformatics*, 4(3):246–259.

Dittrich, P. and di Fenizio, P. S. (2007). Chemical organisation theory. *Bulletin of Mathematical Biology*, 69(4):1199–1231.

Doebeli, M. (2002). A model for the evolutionary dynamics of cross-feeding polymorphisms in microorganisms. *Population Ecology*, 44(2):59–70.

E.coli (2016). E.coli core: http://systemsbiology.ucsd.edu/InSilicoOrganisms/Ecoli/EcoliSBML [accessed: 2016-04-07].

Edwards, D. P., Hassall, M., Sutherland, W. J., and Yu, D. W. (2006). Selection for protection in an ant–plant mutualism: host sanctions, host modularity, and the principal–agent game. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1586):595–602.

Eker, S., Krummenacker, M., Shearer, A., Tiwari, A., Keseler, I., Talcott, C., and Karp, P. (2013). Computing minimal nutrient sets from metabolic networks via linear constraint solving. *BMC Bioinformatics*, 14(1):114.

Enyeart, P. J., Simpson, Z. B., and Ellington, A. D. (2015). A microbial model of economic trading and comparative advantage. *Journal of Theoretical Biology*, 364:326–343.

Estrada, E. and Rodríguez-Velázquez, J. A. (2006). Subgraph centrality and clustering in complex hyper-networks. *Physica A: Statistical Mechanics and its Applications*, 364:581 – 594.

Fabich, A., Jones, S. A., Chowdhury, F. Z., Cernosek, A., Anderson, A., Smalley, D., McHargue, J. W., Hightower, G. A., Smith, J. T., Autieri, S. M., Leatham, M. P., Lins, J. J., Allen, R. L., Laux, D. C., Cohen, P. S., and Conway, T. (2008). Comparison of Carbon Nutrition for Pathogenic and Commensal *Escherichia coli* Strains in the Mouse Intestine. *Infection and Immunity*, 76:1143–1152.

Feist, A. M. and Palsson, B. O. (2010). The biomass objective function. *Current Opinion in Microbiology*, 13(3):344 – 349. Ecology and industrial microbiology * Special section: Systems biology.

Fell, D. A. and Wagner, A. (2000). Structural properties of metabolic networks: Implications for evolution and modeling of metabolism. In *Animating the cellular map*, pages 79 – 85. Stellenbosch University Press.

Finney, A. and Hucka, M. (2003). Systems biology markup language: Level 2 and beyond. *Biochemical Society Transactions*, 31(6):1472–1473.

Gagneur, J. and Klamt, S. (2004). Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, 5(1):1–21.

Gallo, G., Longo, G., Nguyen, S., and Pallottino, S. (1993). Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2-3):177–201.

Gerstl, M. P., Klamt, S., Jungreuthmayer, C., and Zanghellini, J. (2015). Exact quantification of cellular robustness in genome-scale metabolic networks. *Bioinformatics*.

Glick, B. R. (2003). Phytoremediation: synergistic use of plants and bacteria to clean up the environment. *Biotechnology Advances*, 21(5):383–393.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, 144:517–546.

Guinane, C. M. and Cotter, P. D. (2013). Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. *Therapeutic Advances in Gastroenterology*, 6(4):295–308.

Gurvich, V. and Khachiyan, L. (1999). On generating the irredundant conjunctive and disjunctive normal forms of monotone Boolean functions. *Discrete Appl. Math.*, 96:363–373.

Hamburger, H. (1973). N-person prisoner's dilemma. *The Journal of Mathematical Sociology*, 3(1):27–48.

Hamilton, W. (1964). The genetical evolution of social behaviour. i. *Journal of Theoretical Biology*, 7(1):1 – 16.

Handorf, T., Christian, N., Ebenhöh, O., and Kahn, D. (2008). An environmental perspective on metabolism. *J. Theor. Biol.*, 252:530 – 537.

Hauert, C., De Monte, S., Hofbauer, J., and Sigmund, K. (2002). Volunteering as red queen mechanism for cooperation in public goods games. *Science*, 296(5570):1129–1132.

Hoeksema, J. D. and Schwartz, M. W. (2003). Expanding comparative–advantage biological market models: contingency of mutualism on partner's resource requirements and acquisition trade–offs. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1518):913–919.

Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1 edition.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. (2006). Copasi—a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074.

Hopla, C., Durden, L., and Keirans, J. (1994). Ectoparasites and classification. *Rev Sci Tech.*, 13(4):985–1017.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., , the rest of the SBML Forum:, Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novère, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. (2003). The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

Hummert, S., Bohl, K., Basanta, D., Deutsch, A., Werner, S., Theißen, G., Schroeter, A., and Schuster, S. (2014). Evolutionary game theory: cells as players. *Mol. BioSyst.*, 10(12):3044–3065.

Hunt, K. A., Folsom, J. P., Taffs, R. L., and Carlson, R. P. (2014). Complete enumeration of elementary flux modes through scalable demand-based subnetwork definition. *Bioinformatics*, 30(11):1569–1578.

Imieliński, M., Belta, C., Halász, Á., and Rubin, H. (2005). Investigating metabolite essentiality through genome-scale analysis of escherichia coli production capabilities. *Bioinformatics*, 21(9):2008–2016.

Imieliński, M., Belta, C., Rubin, H., and Halász, Á. (2006). Systematic analysis of conservation relations in *Escherichia coli* genome-scale metabolic network reveals novel growth media. *Biophysical Journal*, 90(8):2659–2672.

Jandér, K. C. and Herre, E. A. (2010). Host sanctions and pollinator cheating in the fig tree–fig wasp mutualism. *Proceedings of the Royal Society of London B: Biological Sciences*.

Jevremovic, D., Boley, D., and Sosa, C. (2011). Divide-and-conquer approach to the parallel computation of elementary flux modes in metabolic networks. In *Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium on*, pages 502–511.

Julius, A. A., Imielinski, M., and Pappas, G. J. (2008). Metabolic networks analysis using convex optimization. In *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, pages 762–767.

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

Kelk, S. M., Olivier, B. G., Stougie, L., and Bruggeman, F. J. (2012). Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. *Scientific Reports*, 2(580).

Kiers, E. T., Rousseau, R. A., West, S. A., and Denison, R. F. (2003). Host sanctions and the legume-rhizobium mutualism. *Nature*, 425(6953):78–81.

Kim, P.-J., Lee, D.-Y., Kim, T. Y., Lee, K. H., Jeong, H., Lee, S. Y., and Park, S. (2007). Metabolite essentiality elucidates robustness of escherichia coli metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13638–13642.

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

Klamt, S., Saez-Rodriguez, J., and Gilles, E. D. (2007). Structural and functional analysis of cellular networks with cellnetanalyzer. *BMC Systems Biology*, 1(1):1–13.

Klamt, S. and Stelling, J. (2003). Two approaches for metabolic pathway analysis? *Trends in Biotechnology*, 21(2):64 – 69.

Kleerebezem, R. and van Loosdrecht, M. C. (2007). Mixed culture biotechnology for bioenergy production. *Current Opinion in Biotechnology*, 18(3):207–212.

Klitgord, N. and Segrè, D. (2010). Environments that induce synthetic microbial ecosystems. *PLoS Comput Biol*, 6(11):1–17.

Lacroix, V., Cottret, L., Thébault, P., and Sagot, M.-F. (2008). An introduction to metabolic networks and their structural analysis. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 5(4):594–617.

Larhlimi, A., Blachon, S., Selbig, J., and Nikoloski, Z. (2011). Robustness of metabolic networks: A review of existing definitions. *Biosystems*, 106(1):1–8.

Larhlimi, A. and Bockmayr, A. (2007). Minimal direction cuts in metabolic networks. *AIP Conference Proceedings*, 940(1):73–86.

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

Leatham, M., Banerjee, S., Autieri, S. M., Mercado-Lubo, R., Conway, T., and Cohen, P. S. (2009). Precolonized Human Commensal *Escherichia coli* Strains Serve as a Barrier to *E. coli* O157:H7 Growth in the Streptomycin-Treated Mouse Intestine. *Infection and Immunity*, 77(7):2876–2886.

Lee, S., Phalakornkule, C., Domach, M. M., and Grossmann, I. E. (2000). Recursive {MILP} model for finding all the alternate optima in {LP} models for metabolic networks. *Computers & Chemical Engineering*, 24(2-7):711 – 716.

Lewis, N. E., Nagarajan, H., and Palsson, B. O. (2012). Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat Rev Micro*, 10(4):291–305.

Lin, C., Chang, C.-J., Lu, C.-C., Martel, J., Ojcius, D., Ko, Y.-F., Young, J., and Lai, H.-C. (2014). Impact of the gut microbiota, prebiotics, and probiotics on human health and disease. *Biomedical Journal*, 37(5):259–268.

Ma, H. and Zeng, A.-P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277.

Maltby, R., Leatham-Jensen, M. P., Gibson, T., Cohen, P. S., and Conway, T. (2013). Nutritional Basis for Colonization Resistance by Human Commensal *Escherichia coli* Strains HS and Nissle 1917 against *E. coli* O157:H7 in the Mouse Intestine. *PLoS ONE*, 8(1):e53957.

Mark W. Schwartz, J. D. H. (1998). Specialization and resource trade: Biological markets as a model of mutualisms. *Ecology*, 79(3):1029–1038.

Martin, B. D. and Schwab, E. (2012). Current usage of symbiosis and associated terminology. *International Journal of Biology*, 5(1).

Maynard Smith, J. and Harper, D. (2003). *Animal Signals*. Oxford Series in Ecology and Evolution. Oxford University Press, USA.

Meador, J., Caldwell, M. E., Cohen, P. S., and Conway, T. (2014). *Escherichia coli* pathotypes occupy distinct niches in the mouse intestine. *Infection and Immunity*, 82(5):1931–1938.

Mitchell, R. J., Irwin, R. E., Flanagan, R. J., and Karron, J. D. (2009). Ecology and evolution of plant–pollinator interactions. *Annals of Botany*, 103(9):1355–1363.

Mithani, A., Preston, G. M., and Hein, J. (2009). Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics*, 25(14):1831–1832.

Monk, J., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M., and Palsson, B. Ø. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences*, 110(50):20338–20343.

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012). The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *mBio*, 3(2).

Müller, A. C. and Bockmayr, A. (2013). Flux modules in metabolic networks. *Journal of Mathematical Biology*, 69(5):1151–1179.

Müller, A. C., Bruggeman, F. J., Olivier, B. G., and Stougie, L. (2014). *Research in Computational Molecular Biology: 18th Annual International Conference, RECOMB 2014, Pittsburgh, PA, USA, April 2-5, 2014, Proceedings*, chapter Fast Flux Module Detection Using Matroid Theory, pages 192–206. Springer International Publishing, Cham.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, 54(2):286–295.

Nedosyko, A. M., Young, J. E., Edwards, J. W., and Burke da Silva, K. (2014). Searching for a toxic key to unlock the mystery of anemonefish and anemone symbiosis. *PLoS ONE*, 9(5):e98449.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563.

Nowak, M. A. and May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, 359(6398):826–829.

Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis? *Nat Biotech*, 28(3):245–248.

Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory*. The MIT Press, first edition.

Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A., and Palsson, B. O. (2003). Metabolic pathways in the post-genome era. *Trends in Biochemical Sciences*, 28(5):250–258.

Parmentier, E. and Michel, L. (2013). Boundary lines in symbiosis forms. *Symbiosis*, 60(1):1–5.

Peacock, K. A. (2011). Symbiosis in ecology and evolution. In Brown, K. d. and Peacock, K. A., editors, *Philosophy of Ecology*, volume 11 of *Handbook of the Philosophy of Science*, pages 219 – 250. North-Holland, Amsterdam.

Pearcy, N., Crofts, J. J., and Chuzhanova, N. (2014). Hypergraph models of metabolism. *International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering*, 8(8):19 – 23.

Pelillo, M. (2009). *Encyclopedia of Optimization*, chapter Replicator dynamics in combinatorial optimization, pages 3279–3291. Springer US, Boston, MA.

Pellmyr, O. and Huth, C. J. (1994). Evolutionary stability of mutualism between yuccas and yucca moths. *Nature*, 372(6503):257–260.

Pey, J., Villar, J. A., Tobalina, L., Rezola, A., García, J. M., Beasley, J. E., and Planes, F. J. (2014). Treeefm: Calculating elementary flux modes using linear optimization in a tree-based algorithm. *Bioinformatics*.

Pfeiffer, T. and Bonhoeffer, S. (2004). Evolution of cross-feeding in microbial populations. *The American Naturalist*, 163(6):E126–E135. PMID: 15266392.

Pfeiffer, T. and Schuster, S. (2005). Game-theoretical approaches to studying the evolution of biochemical systems. *Trends in Biochemical Sciences*, 30(1):20 – 25.

Pfeiffer, T., Schuster, S., and Bonhoeffer, S. (2001). Cooperation and competition in the evolution of atp-producing pathways. *Science*, 292(5516):504–507.

Pramanik, J. and Keasling, J. D. (1997). Stoichiometric model of escherichia coli metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering*, 56(4):398–421.

Quek, L.-E. and Nielsen, L. K. (2014). A depth-first search algorithm to compute elementary flux modes by linear programming. *BMC Systems Biology*, 8(1):1–10.

Reeburgh, W. S. (2007). Oceanic methane biogeochemistry. *Chemical Reviews*, 107(2):486–513.

Ricardo, D. (1817). *On the Principles of Political Economy and Taxation*. London: John Murray.

Romero, P. and Karp, P. (2001). Nutrition-related analysis of pathway/genome databases. In *Pacific Symposium on Biocomputing'01*, pages 470–482.

Ronald Noë, P. H. (1994). Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, 35(1):1–11.

Rothschild, M. and Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *The Quarterly Journal of Economics*, 90(4):629–649.

Rozen, D. E. and Lenski, R. E. (2000). Long-term experimental evolution in escherichia coli. viii. dynamics of a balanced polymorphism. *The American Naturalist*, 155(1):24–35.

Ruby, E. G. and McFall-Ngai, M. J. (1999). Oxygen-utilizing reactions and symbiotic colonization of the squid light organ by vibrio fischeri. *Trends in Microbiology*, 7(10):414–420.

Sandholm, W. H. (2010). *Population Games and Evolutionary Dynamics*. MIT Press.

Sapp, J. (2004). The dynamics of symbiosis: an historical overview. *Canadian Journal of Botany*, 82(8):1046–1056.

Schilling, C. H., Letscher, D., and Palsson, B. Ø. (2000). 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):229 – 248.

Schimper, A. (1888). *Die epiphytische Vegetation Amerikas*. Gustav Fischer, Jena.

Schuetz, R., Kuepfer, L., and Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli. *Molecular Systems Biology*, 3(1).

Schuster, S., Dandekar, T., and Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology*, 17(2):53–60.

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

Schuster, S., Hilgetag, C., Woods, J., and Fell, D. (2002a). Reaction routes in biochemical reaction systems: Algebraic properties, validated calculation procedure and example from nucleotide metabolism. *Journal of Mathematical Biology*, 45(2):153–181.

Schuster, S., Klamt, S., Weckwerth, W., Moldenhauer, F., and Pfeiffer, T. (2002b). Use of network analysis of metabolic systems in bioengineering. *Bioprocess and Biosystems Engineering*, 24(6):363–372.

Schwarz, R., Musch, P., von Kamp, A., Engels, B., Schirmer, H., Schuster, S., and Dandekar, T. (2005). Yana – a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC Bioinformatics*, 6(1):1–12.

Schwendener, S. (1868). Ueber die beziehungen zwischen algen und flechtengonidien. *Botanische Zeitung*, 26:289–292.

Segrè, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117.

Senger, R. S. and Papoutsakis, E. T. (2008). Genome-scale model for clostridium acetobutylicum: Part i. metabolic network resolution and analysis. *Biotechnology and Bioengineering*, 101(5):1036–1052.

Sieuwerts, S., de Bok, F. A. M., Hugenholtz, J., and van Hylckama Vlieg, J. E. T. (2008). Unraveling microbial interactions in food fermentations: from classical to genomics approaches. *Applied and Environmental Microbiology*, 74(16):4997–5007.

Sigma-Aldrich (2016). Metabolic pathway map: http://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma/General_Information/metabolic_pathways_poster.pdf [accessed: 2016-03-04].

Simms, E. L., Taylor, D. L., Povich, J., Shefferson, R. P., Sachs, J., Urbina, M., and Tausczik, Y. (2006). An empirical test of partner choice mechanisms in a wild legume–rhizobium interaction. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1582):77–81.

Smith, J. M. and Price, G. R. (1973). The logic of animal conflict. *Nature*, 246(5427):15–18.

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

Stewart, E. J. (2012). Growing unculturable bacteria. *Journal of Bacteriology*, 194(16):4151–4160.

Storey, K. B. (2004). *Functional Metabolism*. John Wiley & Sons, Inc.

Tasoff, J., Mee, M. T., and Wang, H. H. (2015). An economic framework of microbial trade. *PLoS ONE*, 10(7):1–20.

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

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

Toller, W. W., Rowan, R., and Knowlton, N. (2001). Repopulation of zooxanthellae in the caribbean corals montastraea annularis and m. faveolata following experimental and disease-associated bleaching. *The Biological Bulletin*, 201(3):360–373.

Traulsen, A. and Nowak, M. A. (2006). Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences*, 103(29):10952–10955.

Trinh, C. T., Wlaschin, A., and Srienc, F. (2008). Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Applied Microbiology and Biotechnology*, 81(5):813–826.

Urbanczik, R. (2006). Sna – a toolbox for the stoichiometric analysis of metabolic networks. *BMC Bioinformatics*, 7(1):1–4.

Valls, M. and de Lorenzo, V. (2002). Exploiting the genetic and biochemical capacities of bacteria for the remediation of heavy metal pollution. *FEMS Microbiology Reviews*, 26(4):327–338.

Varian, H. R. (2009). *Intermediate Microeconomics: A Modern Approach (Eighth Edition)*. W. W. Norton & Company, eighth edition.

Voet, D. and Voet, J. (2011). *Biochemistry*. John Wiley & Sons.

von Böhm-Bawerk, E. (1891). *The Positive Theory of Capita*. London: Macmillan and Co.

von Kamp, A. and Klamt, S. (2014). Enumeration of smallest intervention strategies in genome-scale metabolic networks. *PLoS Comput Biol*, 10(1):e1003378.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Wagner, A. and Fell, D. A. (2001). The small world inside large metabolic networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1478):1803–1810.

Wagner, C. and Urbanczik, R. (2005). The geometry of the flux cone of a metabolic network. *Biophysical Journal*, 89(6):3837–3845.

Watson, M. R. (1984). Metabolic maps for the apple slowromancapii@. *Biochemical Society Transactions*, 12(6):1093–1094.

Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

Welbaum, G., Sturz, A., Dong, Z., and Nowak, J. (2004). Managing soil microorganisms to improve productivity of agro-ecosystems. *CRITICAL REVIEWS IN PLANT SCIENCES*, 23(2):175–193.

Werner, G. D. A., Strassmann, J. E., Ivens, A. B. F., Engelmoer, D. J. P., Verbruggen, E., Queller, D. C., Noë, R., Johnson, N. C., Hammerstein, P., and Kiers, E. T. (2014). Evolution of microbial markets. *Proceedings of the National Academy of Sciences*, 111(4):1237–1244.

West, S. A., Kiers, E. T., Simms, E. L., and Denison, R. F. (2002). Sanctions and mutualism stability: why do rhizobia fix nitrogen? *Proceedings of the Royal Society B: Biological Sciences*, 269(1492):685–694.

Weyl, E. G., Frederickson, M. E., Yu, D. W., and Pierce, N. E. (2010). Economic contract theory tests models of mutualism. *Proceedings of the National Academy of Sciences*, 107(36):15712–15716.

Wintermute, E. H. and Silver, P. A. (2010). Emergent cooperation in microbial metabolism. *Molecular Systems Biology*, 6(1).

Wyatt, G. A. K., Kiers, E. T., Gardner, A., and West, S. A. (2014). A biological market analysis of the plant-mycorrhizal symbiosis. *Evolution*, 68(9):2603–2618.

Wyatt, G. A. K., Kiers, E. T., Gardner, A., and West, S. A. (2016). Restricting mutualistic partners to enforce trade reliance. *Nat Commun*, 7.

Zarecki, R., Oberhardt, M. A., Reshef, L., Gophna, U., and Ruppin, E. (2014). A novel nutritional predictor links microbial fastidiousness with lowered ubiquity, growth rate, and cooperativeness. *PLoS Comput Biol*, 10(7):1–12.

Zhou, W. and Nakhleh, L. (2011). Properties of metabolic graphs: biological organization or representation artifacts? *BMC Bioinformatics*, 12(1):1–12.