Modelisation mathematique des impacts de l'environnement à l'aide de reseaux metaboliques et de la theorie des jeux

Taneli Pusa

N° d'ordre NNT : xxx

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

École Doctorale ED 341
E2M2

Spécialité de doctorat : Bioinformatique

Soutenue publiquement le 04/02/2019, par :
Taneli Pusa

Mathematical modelling of the impacts of environment using metabolic networks and game theory

---

Devant le jury composé de :

Vinga Susana, DR, Instituto Superior Técnico	Rapporteur
Mishra Bud, Prof., Courant Inst.	Rapporteur
Mouchiroud Dominique, Prof., LBBE	Examinatrice
Mira Nuno, DR, Instituto Superior Técnico	Examinateur
Becchetti Luca, DR, Sapienza Univ. di Roma	Examinateur
Cottret Ludovic, DR, INRA	Examinateur
Sagot Marie-France, DR, LBBE, INRIA Rhône-Alpes	Directrice de thèse
Marchetti-Spaccamela Alberto, Prof., Sapienza Univ. di Roma	Co-directeur de thèse
Mary Arnaud, DR, LBBE,	Co-encadrant

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 no particular order, I'd like to thank: my advisors Marie-France Sagot, Alberto Marchetti-Spaccamela, and Arnaud Mary; all the members of the Erable team, both past and present (and including those that are members of the family if not the research team); a special thanks to our team assistants Florence Bouheddi, Marina da Graça, and Claire Sauer, all of whom have spent countless hours dealing with the administrative complexities related to my employment; all those whose work and code I've built on, in particular Martin Wannagat, Alice Julien-Laferrière, and Louis Duchemin; my collaborators and co-authors Mariana Ferrarini, Ricardo Andrade, Leen Stougie, and Ifigeneia Kyrkou; all the people in MicroWine for making this crazy adventure possible; my advisor, colleague and friend Tadeas Priklopil whose course on game theory was the first step on the path that eventually led to this thesis; my parents for their support as well as for introducing me to academia.

Resumé en français

L'ADN est transcrit en ARN, l'ARN est traduit en protéines, les protéines catalysent des réactions. Ainsi est souvent reformulé la théorie fondamentale de la biologie moléculaire [45]. C'est aussi la façon comme je répondrai si on me pressait d'expliquer le sujet de ma thèse en une phrase. J'ajouterai peut-être un détail supplémentaire : les réactions sont liées entre elles et donnent naturellement naissance à une structure que l'on pourrait décrire comme un réseau.

Cela est essentiellement ce que l'on appelle un *réseau métabolique* ou une *reconstruction métabolique* : un ensemble de réactions chimiques, rassemblées sur la base du génome d'un organisme, et disposées de manière à ce que les réactions correspondantes soient liées entre elles. C'est une structure curieuse dans la mesure où cela correspond à la fois à un modèle et à un objet d'étude. C'est un modèle, car lorsque la cellule est perçue – d'une manière très simplifiée – comme un "sac d'enzymes", l'ensemble des réactions chimiques décrit les molécules que la cellule peut absorber, ce qu'elle peut faire avec ces substrats, et ce qu'elle finira par excréter en tant que produits finis. Ainsi, le réseau métabolique peut être utilisé pour comprendre en grande partie comment un organisme ou une cellule interagit avec son environnement et de quelle façon.

Par contre, le réseau métabolique est intéressant en soi. Ce n'est bien sûr pas un hasard que les conversions chimiques codées dans le génome sont liées entre elles, car chacune représente un élément dans des structures plus grandes, souvent appelées *fonctions métaboliques* ou *voies métaboliques*. Ceci correspondent aux composantes de haut niveau du fonctionnement interne de la cellule, servant les besoins de base de la vie : production d'énergie et de biomasse, élimination des toxines, etc. Il est donc intéressant de se demander comment ces processus sont organisés. Y a-t-il une structure sous-jacente ? Y a-t-il des redondances ? Certaines composantes sont-elles plus cruciales que d'autres ?

Un réseau métabolique est généralement représenté par une sorte de graphe. Il peut s'agir d'un graphe simple, liant entre eux des nœuds correspondant à des métabolites ou à des réactions, d'un graphe bipartite contenant des nœuds pour les métabolites et les réactions, ou d'un hypergraphe liant entre eux des ensembles de nœuds, correspondant à des métabolites, au travers d'hyperarcs correspondant à des réactions. Le graphe peut ne pas être orienté ou ses arêtes peuvent être dotées de directions afin de refléter le fait que de nombreuses réactions métaboliques ont tendance à se dérouler dans une direction.

Une autre représentation couramment utilisée pour le réseau métabolique est la matrice stœchiométrique. Celle-ci contient les coefficients stœchiométriques de toutes les réactions, ce qui permet de penser le réseau en termes de flux de matière : chaque réaction consomme un ensemble de molécules et en produit un autre. De cette manière, le réseau, ou le *métabolisme*, peut être considéré comme un processus, prenant les substrats de l'environnement à une extrémité et excrétant les produits à l'autre.

La prise en compte du métabolisme en termes de flux réactionnels donne lieu à ce qui est sans doute l'application la plus connue des reconstructions métaboliques : l'analyse par contrainte. L'idée sous-jacente est que pour que le métabolisme fonctionne, il doit être équilibré. Une

réaction ne peut pas procéder si on ne lui fournit pas ses substrats. De même, les produits de la réaction doivent pouvoir être utilisés ou excrétés. Cela signifie que le réseau définit les capacités métaboliques d'un organisme non seulement en termes de réactions uniques, mais également en termes d'ensembles de réactions qui doivent toutes fonctionner de manière synchrone.

Les contraintes d'équilibre peuvent être utilisées pour étudier la structure du métabolisme en explorant les différents modes, définis en termes de réactions qui sont actives, selon lesquels il peut opérer. Ces contraintes peuvent également être utilisées afin de prédire le comportement métabolique. Dans l'analyse de l'équilibre des flux, le principe du métabolisme équilibré est associé à une hypothèse d'optimisation afin de rechercher, parmi les comportements métaboliques possibles, celui qu'un organisme pourrait vraisemblablement adopter.

Avec la disponibilité croissante de données dites 'omiques – transcriptomique, protéomique et métabolomique – divers moyens de les intégrer au réseau métabolique ont suscité un intérêt croissant. Lorsque le réseau est représenté par un graphe, les données 'omiques peuvent guider l'extraction des sous-réseaux d'intérêt pour rechercher des voies métaboliques ou des ensembles de gènes liés entre eux. Dans le cadre de la modélisation basée sur les contraintes, les données 'omiques peuvent être utilisées pour améliorer la prédiction du comportement métabolique et pour construire des modèles métaboliques spécifiques au contexte.

Une application intéressante des reconstructions métaboliques en conjonction avec les données 'omiques consiste à utiliser les deux méthodes pour comprendre les changements métaboliques. Lorsqu'un organisme rencontre un changement dans les conditions environnementales, il s'ensuit souvent une réorganisation du métabolisme. Des mesures comparatives de l'expression des gènes et des concentrations de métabolites peuvent être utilisées pour mieux comprendre ces changements, mais ces données sont "sans structure", ce qui signifie qu'elles ne fournissent pas d'informations sur la relation entre les processus métaboliques. Un réseau métabolique, en revanche, contient cette information et peut donc grandement améliorer une telle analyse. Ce sujet est exploré dans le chapitre 2 de cette thèse, où je présente une nouvelle méthode appelée Moomin pour "Mathematical explOration of Omics data on a MetabolIc Network" (exploration mathématique de données 'omiques sur un réseau métabolique). Moomin combine les résultats d'une analyse d'expression différentielle comparant les niveaux d'expression génique dans deux conditions différentes avec un réseau métabolique afin de produire une hypothèse de changement métabolique. L'idée est d'utiliser la structure du réseau pour définir les changements globaux possibles du métabolisme. Ces changements sont ensuite annotés sur la base des données d'expression des gènes dans le but de trouver le changement qui correspond le mieux aux observations. La recherche du meilleur changement de score est formulée en forme d'un problème d'optimisation linéaire.

La théorie des jeux est une branche des mathématiques appliquées qui traite des agents rationnels en interaction ayant des objectifs contradictoires. Lorsque la rationalité est remplacée par la sélection naturelle, la théorie des jeux *évolutionniste* peut être utilisée pour expliquer les "décisions" prises même par des organismes microscopiques. Dans le chapitre 3, je présente l'idée d'un *jeu métabolique*, qui corrspond à un modèle de jeu théorique permettant de prédire le comportement métabolique. Contrairement à l'analyse de l'équilibre des flux, qui permet de prédire l'état métabolique à l'aide d'une optimisation simple, un jeu métabolique prend en compte le fait que l'optimalité est influencée par les membres environnants d'une communauté microbienne. En modifiant la disponibilité des nutriments ou en sécrétant des molécules bénéfiques ou nocives, les microbes créent essentiellement leur propre environnement et adaptent le comportement optimal au contexte.

Il existe une autre façon d'expliquer ce qui m'a employé ces trois dernières années, que la plupart des gens trouvent beaucoup plus "sexy". En effet, techniquement parlant, le sujet de mon projet de ma thèse était le vin. Le projet Microwine, dont je faisais partie, était une

immense entreprise interdisciplinaire visant à faire la lumière sur les différents rôles que jouent les microbes dans la production du vin. Quinze étudiants de doctorat, allant de géologues et biologistes (et même un œnologue) à mathématiciens, ont utilisé le séquençage de nouvelle génération pour caractériser les communautés microbiennes dans les vignobles, étudier la dynamique de la fermentation et rechercher des solutions à diverses maladies menaçant l'avenir de la viticulture.
Cependant, comme souvent dans le cas de projets interdisciplinaires, mes recherches ont fini par ne pas trop impliquer le vin. Au lieu de cela, la majeure partie de cette thèse peut être considérée comme une recherche fondamentale sur les réseaux métaboliques. Cette recherche est bien entendu néanmoins pertinente également dans le contexte du vin, qui est après tout un produit direct du métabolisme microbien. Mon travail avec une autre doctorante du projet, Ifigeneia Kyrkou, avec qui nous avons développé un modèle épidémiologique de l'agent pathogène de la vigne *Xylella fastidiosa* est ainsi plus proche du sujet du vin. Ce travail est décrit au chapitre 4.
Cette thèse est structurée de la manière suivante. Dans le chapitre 1, je présente quelques concepts relatifs aux sujets abordés dans les chapitres suivants. Je présente les objets mathématiques les plus couramment utilisés pour décrire les réseaux métaboliques et donne une brève introduction à la modélisation par contraintes. Dans la dernière partie du chapitre 1, je présente une introduction à la théorie des jeux évolutionnistes.
Dans le chapitre 2, je discute de l'intégration des données omiques dans les réseaux métaboliques. Dans la première partie du chapitre, je passe en revue la littérature sur le sujet, en décrivant les différentes manières dont les données 'omiques ont été utilisées en combinaison avec les réseaux métaboliques et la multitude d'outils informatiques ayant mis en œuvre de telles analyses. La dernière partie du chapitre 2 est dédiée à l'algorithme MOOMIN, une méthode permettant de générer des hypothèses de changements métaboliques à l'aide de données d'expression génique et d'un réseau métabolique. J'introduis le cadre théorique et prouve la complexité du problème d'optimisation résultant. Je décris ensuite deux implémentations, une utilisant "answer set programming" et une autre utilisant l'optimisation linéaire. Dans la dernière partie du chapitre, je présente les résultats obtenus en appliquant la méthode à deux ensembles de données réelles. Un article décrivant la méthode MOOMIN et les résultats obtenus est en préparation. Le logiciel MOOMIN est disponible à l'adresse : `github.com/htpusa/moomin`.
Dans le chapitre 3, je passe en revue la littérature sur les applications de la théorie des jeux à l'étude des microbes, en mettant l'accent sur le métabolisme et en particulier les jeux dérivés de réseaux métaboliques et de la modélisation par contraintes. Dans la dernière partie du chapitre, j'explique plus en détail l'idée d'un jeu métabolique et aborde différents aspects de la définition de tels jeux : le choix des joueurs, les actions et les gains. Ce chapitre correspond à un article de synthèse soumis à *Frontiers in Genetics* au moment de la rédaction.
Dans le dernier chapitre, à savoir le chapitre 4, je présente un modèle épidémiologique de l'agent pathogène de la vigne *Xylella fastidiosa*. Je déduis des expressions pour l'équilibre sans maladie du modèle et le taux de reproduction de base, et présente des simulations numériques explorant les équilibres endémiques. Enfin, je présente les résultats d'une analyse de sensibilité pour le taux de reproduction de base et l'équilibre endémique, et discute des implications pour le contrôle de l'agent pathogène. Le contenu de ce chapitre fait partie d'un article publié dans *Frontiers in Microbiology* [103] dont je suis le deuxième auteur.

Contents

Introduction

DNA is transcribed into RNA, RNA is translated into proteins, proteins catalyse reactions. Thus is often paraphrased the central dogma of molecular biology [45]. It is also how I would respond if pressed to explain the topic of my thesis in one sentence. Perhaps I would add one more detail: the reactions are linked, naturally giving rise to a structure that is best described as a network.

That is essentially what a *metabolic network* or a *metabolic reconstruction* is: a collection of chemical reactions, gathered based on the genome of an organism, and arranged in such a way that related reactions are connected. It is a curious structure in that it is both a model and an object of study in itself. It is a model, because when the cell is seen – in a grossly simplified way – as a "bag of enzymes", the collection of chemical reactions describe what molecules the cell can take in, what it can do with those substrates, and what it will eventually excrete as the end products. Thus the metabolic network can be used to understand much of how an organism or a cell interacts with its environment and how it does it.

On the other hand the metabolic network is interesting in itself. It is of course no accident that the chemical conversions coded into the genome are linked, for each of them serves as a piece in larger structures, often referred to as *metabolic functions* or *pathways*. They are the higher level components of the cell's inner workings, serving the basic needs of biological life: production of energy and biomass, removal of toxins etc. It is thus of interest to ask how such processes are organised. Is there an overlying structure? Are there redundancies? Are some components more crucial than others?

A metabolic network is usually represented by some sort of graph. This can be a simple graph, joining together nodes corresponding to either metabolites or reactions, a bipartite graph that contains nodes for both metabolites and reactions, or a hypergraph that joins together sets of nodes, corresponding to metabolites, using hyperarcs that correspond to reactions. The graph can be undirected, or its edges can be assigned directions to reflect the fact that many metabolic reactions tend to proceed in one direction.

Another commonly used representation for the metabolic network is the stoichiometric matrix. It contains the stoichiometric coefficients of all of the reactions, allowing one to think of the network in terms of a flux of matter: each reaction consumes one set of molecules and produces another. In this way, the network, or *metabolism*, can be seen as one process, taking in substrates from the environment at one end and excreting out products at the other.

Considering the metabolism in terms of reaction fluxes gives rise to what is arguably the most celebrated application of metabolic reconstructions: constraint-based analysis. It is built on the idea that in order for the metabolism to function, it has to be balanced. A reaction cannot proceed if it is not being supplied its substrates. Similarly, the products of the reaction need to be further utilised or excreted. This means that the network defines the metabolic capabilities of an organism not only in terms of single reactions, but in terms of sets of reactions that all need to operate in synchrony.

The balance constraints can be used to study the structure of the metabolism by exploring the different modes, defined in terms of the reactions that are active, it can operate in. They

can also be used to predict metabolic behaviour. In Flux Balance Analysis, the principle of balanced metabolism is combined with an assumption of optimisation to find among the possible metabolic behaviours the one that an organism would likely undertake.

With the increasing availability of so-called 'omics data – transcriptomics, proteomics, and metabolomics – there has been growing interest in various ways of integrating them with the metabolic network. When the network is represented by a graph, 'omics data can guide the extraction of subnetworks of interest to find metabolic pathways or sets of related genes. Within the framework of constraint-based modelling, 'omics data can be used to improve the prediction of metabolic behaviour and to build context-specific metabolic models.

One interesting application of metabolic reconstructions in conjunction with 'omics data is to use the two to understand metabolic shifts. When an organism encounters a change in environmental conditions, often a re-organisation of metabolism follows. Comparative measurements of gene expression and metabolite concentrations can be used to gain insight into these changes but these data are "structureless", meaning they lack the information about how the metabolic components relate to each other. A metabolic network on the other hand contains this information, and can thus greatly benefit such an analysis.

This topic is explored in Chapter 2 of this thesis, where I present a new method called Moomin for "Mathematical explOration of Omics data on a MetabolIc Network". Moomin combines the results of a differential expression analysis comparing the gene expression levels in two different conditions with a metabolic network to produce a hypothesis of a metabolic shift. The idea is to use the network structure to define feasible global changes in metabolism. These changes are then scored based on the gene expression data with the goal of finding the change that best agrees with the observations. Finding the best-scoring change is formulated into an optimisation problem that can be solved using Mixed-Integer Linear Programming.

Game theory is a branch of applied mathematics that deals with interacting rational agents with conflicting goals. When rationality is replaced with natural selection, *evolutionary* game theory can be used to explain the "decisions" taken by even microscopic organisms. In Chapter 3, I present the idea of a *metabolic game*, a game theoretical model for the prediction of metabolic behaviour. In contrast to Flux Balance Analysis, where the metabolic state is predicted using simple optimisation, a metabolic game takes into account the fact that optimality is influenced by the surrounding members of a microbial community. By changing the availability of nutrients, or secreting beneficial or harmful molecules, microbes essentially create their own environment and make optimal behaviour context-specific.

There is another way to explain what has employed me these past three years, one that for some reason most people find much "sexier". Yes, technically the topic of my PhD project was wine. The Microwine project, of which I was part, was an immense interdisciplinary undertaking aimed to shed light on the different roles microbes play in the production of the beverage. Ranging from geologists and biologists (and even an oenologist) to a mathematician (yours truly), 15 PhD students in total used next generation sequencing to characterise the microbial communities in vineyards and wineries, studied the dynamics of fermentations, and sought solutions to various diseases threatening the future of viticulture.

However, as is often the case with interdisciplinary projects, my research ended up not involving much wine. Instead, most of this thesis can be considered basic research on the topic of metabolic networks. It is of course nevertheless relevant also in the context of wine, which is – after all – a direct product of microbial metabolism. Closest to the topic of wine comes my work with another PhD student of the project, Ifigeneia Kyrkou, with whom we developed an epidemiological model of the *Xylella fastidiosa* grapevine pathogen. This work is described in Chapter 4.

This thesis is structured in the following way. In Chapter 1, I present some concepts relevant to the topics discussed in the later chapters. I introduce the mathematical objects most

commonly used to describe metabolic networks and give a brief introduction to constraint-based modelling. In the latter part of Chapter 1, I give an introduction to evolutionary game theory.
In Chapter 2, I discuss the integration of 'omics data into metabolic networks. In the first part of the chapter I review the literature on the subject, describing the various ways in which 'omics data have been used in combination with metabolic networks and the multitude of computational tools that have implemented such analyses. The latter part of Chapter 2 is dedicated to the MOOMIN algorithm, a method to generate hypotheses of metabolic shifts using gene expression data and a metabolic network. I introduce the theoretical framework and prove the complexity of the resulting optimisation problem. I then describe two implementations, one using Answer Set Programming, and another using Mixed-Integer Linear Programming. In the last part of the chapter, I present results obtained by applying the method on two real data sets. An article describing the MOOMIN method and the results obtained is in preparation. The MOOMIN software is available at: `github.com/htpusa/moomin`.
In Chapter 3, I review literature that has applied game theory to the study of microbes, with a focus on metabolism and especially games derived using metabolic networks and constraint-based modelling. In the latter part of the chapter, I further explain the idea behind a metabolic game and discuss different aspects of defining such games: the choice of players, actions, and payoffs. This chapter corresponds to a review article submitted to Frontiers in Genetics at the time of writing.
In the last chapter, Chapter 4, I present an epidemiological model of the *Xylella fastidiosa* grapevine pathogen. I derive expressions for the disease-free equilibrium of the model and the basic reproduction number, and present numerical simulations exploring the endemic equilibria. Lastly, I present the results of a sensitivity analysis for the basic reproduction number and the endemic equilibrium, and discuss the implications for the control of the pathogen. The material in this chapter forms a part of an article published in Frontiers in Microbiology [103] of which I am the second author.

Chapter 1

Preliminaries

Contents

1.1 Metabolic networks

Formally defined, a *metabolic network* is a collection of objects and the relations amongst them [104]. The objects that comprise the network are *metabolites* (or compounds), *biochemical reactions*, *enzymes*, and *genes*.

Metabolites are molecules that are imported into, exported out of, or synthesised or degraded inside a cell. A biochemical reaction converts one set of metabolites, the *substrates*, into another set, the *products*. In principle, the sets are interchangeable. In other words, all chemical reactions can proceed in both directions. In reality, that is, under specific physiological conditions, many reactions are only observed to take place in one direction. We call such reactions *irreversible* as opposed to *reversible* reactions that can be considered to take place in either of the two directions.

Most reactions inside a cell are catalysed by an enzyme. An enzyme is a protein or a complex of proteins that lowers the energy required for the reaction to happen. In other words, the presence of an enzyme increases the probability that a particular biochemical conversion takes place. Some reactions are catalysed by a specific enzyme, others have several distinct *isoenzymes* that are able to facilitate the same reaction. Since a gene is generally speaking associated with a protein, this means that the relation between genes and enzymes – and hence genes and reactions – is a complex one.

Some of the metabolites can be distinguished as *cofactors*, molecules that facilitate the functioning of enzymes and are necessary for their action.

A *pathway* can be loosely defined as a collection of reactions associated with some higher level function. For example, *glycolysis* refers to the process of converting glucose into pyruvate through a series of intermediate steps.

The process of reconstructing a metabolic network starts with the sequencing of an organism's genome. Once genes have been identified, they are assigned functional annotations. While

this step has been automised to a large extent, labour intensive manual refinement is still needed.

1.1.1 A metabolic network as a graph

Metabolites that participate in a reaction are very clearly linked together. Moreover, metabolites can take part in several different reactions, forming further connections. It thus is very natural to represent a metabolic network using a *graph*. A graph is defined as a couple (V, E) corresponding respectively to a set of *nodes* and a set of *edges* which correspond to a subset of V^2.
Several different graphs can be defined based on a metabolic network. I present here the most commonly used ones. In a *compound graph* nodes represent metabolites, and two metabolites are connected by an edge if there is a reaction that has one of them as a substrate and another as a product. At first glance it might seem that in such a graph, an edge denotes the possibility of converting one compound into the other. This is however slightly misleading, and in fact one of the drawbacks of the compound graph representation: there might be several other metabolites participating in the reaction that gives rise to the edge – the compound graph loses this information, sometimes leading to false conclusions if the graph is traversed naively. In contrast, in a *reaction graph*, the nodes are reactions, and edges are drawn if the product of one is the substrate of the other. Similarly to the compound graph, some information is also lost in the reaction graph representation: when two reactions are joined by an edge, it might suggest that the said reactions can proceed in succession, forming a pathway. However, the product set of one need not equal or encompass the substrate set of the other, and hence other reactions might be needed to supply the second reaction. Examples can be seen in figures 1.1b and 1.1c.
A *bipartite graph* is a graph whose node set V can be divided into two disjoint subsets V_1 and V_2 so that in each edge one node is in V_1 and the other in V_2. A metabolic network can be represented by a bipartite graph by defining one subset of nodes for metabolites and the other for reactions. The metabolite nodes are then connected to all the nodes whose reactions they participate in. A *hypergraph* is a generalisation of the simple graph: it joins together sets of nodes. It can be denoted again by (V, E), but this time E is a set of *hyperedges*.
A *directed graph* is a graph whose edges are directed, meaning they have an orientation, and are then called *arcs*. I mentioned earlier that many reactions in a metabolic network can be considered irreversible, that is, to have a direction. It is thus natural to assign directions to the relations in the graph representation so that, for example, two metabolites are connected by an arc that traverses from the substrate to the product of the underlying reactions. A similar logic holds for the other models. A directed hyperedge is a *hyperarc*.
When a network comprises both reversible and irreversible reactions, a mixed graph with both directed and undirected edges can be used. It should be noted however that in a bipartite graph, undirected edges can lead to ambiguities, obscuring the two sets of metabolites at the opposite ends of a reaction. In other words, if the reaction and compound nodes are linked with undirected edges, it is impossible to divide the compounds into substrates and products just based on the graph structure. One solution is to refrain from using undirected edges, and instead use edge labels to indicate reversibility in directed bipartite graphs.

1.1.2 A metabolic network as a stoichiometric matrix

While both the directed bipartite and hypergraph representations preserve the information about the participants of a reaction, they do omit one additional detail: the *stoichiometry* of the reactions. Stoichiometry refers to the relative quantities of the metabolites taking part

in a reaction, in other words, how many molecules of substrate A are needed to produce one molecule of product B. While this information could be included in the aforementioned graphs in the form of edge labels, this seems cumbersome. Instead an alternative representation is often used, the *stoichiometric matrix*.
The stoichiometric matrix $\mathbf{S}$ is an $m \times n$ matrix, with m rows corresponding to the number of metabolites and n columns to the number of reactions in the network. The entry $\mathbf{S}_{ij}$ is the number of molecules of metabolite i participating in the reaction j. If the entry is negative, i is consumed by the reaction, that is, it is a substrate of j, if it is positive, i is produced. An example is shown in Figure 1.1f.
A stoichiometric matrix is usually accompanied by a pair of vectors **ub** and **lb** containing respectively the upper and lower bounds for the rates of the reactions. While the exact entries are often subject to technical details and might not be very informative *per se*, a zero entry in **lb** can be used to recognise reactions that are considered irreversible.
A concept closely related to the stoichiometric matrix is that of *flux vectors*. A flux vector $\mathbf{v}$ expresses the metabolic state of an organism as the rates of all the reactions contained in the network. It is a static concept in the sense that the rates are taken to represent a persistent state of equilibrium. Alternatively, the word *flux distribution* is also used to further underline the idea of allocation of resources: while the network is a collection of the metabolic reactions the organism in question has at its disposal, the flux vector expresses which of those reactions it chooses to utilise.
Two types of reactions can be distinguished: those that consume and produce *internal* metabolites and those that transport them into or out of the cell, appropriately termed *transport reactions*. A transport reaction can be coded into the stoichiometric matrix by omitting the substrate or the products. In other words, a reaction that imports a compound into the cell produces it from "nothing", and one that exports it just consumes it, producing nothing. This type of formulation implies the assumption that in the context of the organism's internal metabolism, an external compound is either present or not, and in the former case it is essentially available *ad libitum*. Similarly, compounds that an organism is able to export, can be disposed of into the environment without limit. However, constraints can be placed for transport reactions in **ub** and **lb** to express relative limits to the extends at which molecules can be moved through the cell membrane. An alternative formulation for denoting the transport reactions is to duplicate the associated metabolites, and to label the clones as external.
Because the time scale of reaction kinetics is much faster than that of the growth of the organism, it is a reasonable assumption to place the metabolism in a *steady state*, namely, the consumption and production of internal metabolites should be balanced. This steady state assumption can be expressed mathematically as

$$\mathbf{S} \cdot \mathbf{v} = 0 \tag{1.1}$$

It defines a concept from linear algebra called the null space [107] that contains all the feasible flux distributions of the system. A set of basis vectors can be used to explicitly define the null space. They are obtained as the columns of the null space matrix $\mathbf{K}$ which satisfies

$$\mathbf{S} \cdot \mathbf{K} = 0$$

However, the basis vectors obtained in this way are not unique. Several approaches for obtaining a unique and biologically relevant basis for the null space using the methods of convex analysis [148] have been proposed.
If all reversible reactions are decomposed into separate forward and backward reactions, all steady state flux vectors lie in the non-negative orthant of the flux space. The solution space now forms a convex polyhedral cone. Because the cone is convex, any vector within it can be

represented by a linear combination of the generating vectors, corresponding to the edges of the cone and termed *extreme currents* [38].
If negative fluxes are allowed for reversible reactions, the flux cone no longer lies in the non-negative orthant. Its edges are now called *elementary flux modes* [156, 158] with the interpretation that they correspond to minimal sets of reactions that are able to operate at a steady state.
Finally, *extreme pathways* were defined in [154]. All internal reversible reactions are split into two, similar to elementary flux modes, but transport reactions are still allowed to have negative fluxes. The extreme pathways again correspond to the edges of the resulting flux cone.

1.1.3 Constraint-based modelling

The matrix representation of a metabolic network is tied to a broader framework known as *constraint-based modelling* [139, 43]. Constraints, such as the steady state condition 1.1, restrict the space of possible fluxes. This space can be explored to discover underlying biochemical structures, as we saw in the previous section, or to find specific flux vectors. The goal can be, for example, to predict if an organism is able to grow in a certain environment or to predict its metabolic state.
In *metabolic flux analysis*, further constraints on the space of feasible fluxes come in the form of additional information, for example, growth rate, substrate uptake, or product formation [12]. These data can be introduced into the **ub** and **lb** vectors to set those entries in **v** that are known to their measured values. The goal is to constrain the flux space sufficiently, so that the remaining unknown fluxes can be calculated. However, a substantial number of fluxes needs to be measured in order for the system to be solvable.
In contrast, in *flux balance analysis* (FBA)[137] information about measured fluxes is not necessarily needed. Instead, FBA relies on optimisation to find a flux vector that is most likely to represent the actual state of an organism's metabolism. First, a pseudo-reaction is defined that corresponds to the formation of biomass. Its substrates are all those metabolites the organism needs in order to grow, in their appropriate quantities. Then, an objective function is formulated. The most common choice is the flux through the biomass reaction [72]. The rationale is that the metabolic state should be – at least approximately – optimised for the production of biomass.
Once the objective function has been defined, the question of finding the optimal flux vector **v** can be formulated as the following *linear programming* (LP) problem [49]:

$$\begin{aligned} \text{max} \quad & \mathbf{c}^T\mathbf{v} && (1.2)\\ \text{subject to} \quad & \mathbf{Sv} = 0 && (1.3)\\ & \mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub} && (1.4) \end{aligned}$$

Equation 1.2 represents the biomass reaction where **c** are the component quantities, Equation 1.3 is the steady state condition, and Equation 1.4 imposes lower and upper bounds on each reaction. While they are not necessarily needed in order to solve the above problem, additional constraints can be added to integrate knowledge about a specific growth environment (restricting which external compounds can be imported by placing upper bounds of zero on appropriate transport reactions) or, for example, to simulate gene knock-outs.
One drawback of FBA is that the optimal flux vector obtained as the solution of the above LP problem is usually not unique. Two variations of the original formulation aim at providing a possible amendment. In flux variability analysis (FVA, [117]), the space of possible optimal flux distributions is explored by establishing the minimum and maximum fluxes for each

reaction. In other words, given that an optimal value of the objective function Z_{opt} has been established as in 1.2-1.4, two additional problems are solved for each reaction:

$$\begin{aligned} \text{max} \quad & \mathbf{v}_i && (1.5)\\ \text{subject to} \quad & \mathbf{Sv} = 0 && (1.6)\\ & \mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub} && (1.7)\\ & \mathbf{c}^T\mathbf{v} = Z_{opt} && (1.8) \end{aligned}$$

and

$$\begin{aligned} \text{min} \quad & \mathbf{v}_i && (1.9)\\ \text{subject to} \quad & \mathbf{Sv} = 0 && (1.10)\\ & \mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub} && (1.11)\\ & \mathbf{c}^T\mathbf{v} = Z_{opt} && (1.12) \end{aligned}$$

The acquired values $\mathbf{v}_{i,max}$ and $\mathbf{v}_{i,min}$ can help better understand the metabolic behaviour and for example the essentiality of individual reactions: if $\mathbf{v}_{i,min} > 0$, it is clear that reaction i is essential for optimal metabolic behaviour.
Parsimonious FBA (pFBA, [110]) relies on the assumption that in an optimal metabolic state, the total required enzyme mass is minimised. This is translated into the language of constraint-based modelling by minimising the sum of all fluxes after having first established the optimal value of the biological objective function:

$$\begin{aligned} \text{min} \quad & \sum \mathbf{v}_i && (1.13)\\ \text{subject to} \quad & \mathbf{Sv} = 0 && (1.14)\\ & \mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub} && (1.15)\\ & \mathbf{c}^T\mathbf{v} = Z_{opt} && (1.16) \end{aligned}$$

1.2 Game theory

Game theory is a branch of applied mathematics originally developed to describe and reason about situations where two or more rational agents, the "*homo economicus*", are faced with choices and have potentially conflicting goals [183]. All participants want to maximise their own well-being, but are doing so taking into account that everyone else is doing the same. Thus paradoxical, suboptimal, outcomes are possible and even common. *Evolutionary* game theory was born out of the realisation that rational choice can be replaced by natural selection: in the course of evolution the strategy (phenotype) that would "win" the game would prevail by simply proliferating more successfully thanks to its success in the "game" [169, 170].
The main concepts that compose a *game* are a set of *players*, a set of *actions* for each player, and a *payoff* function. The players are the participants in the interaction under study. In the simplest case, they are interchangeable, meaning they all have the same set of available actions and the same payoff function. A set of actions defines the choice that each player faces and can correspond for example to the expressed phenotype. The words "action" and "strategy" are often used more or less interchangeably and confusion can ensue. I will adhere to a convention whereby the word "action" refers only to the atomic choices available to a player, that is, the discrete members of the action set, and *strategy* is the "rule" by which the player chooses which choice to enact. A *pure* strategy is a strategy in which one single action is chosen. In this case, the words are thus somewhat interchangeable. In contrast, a *mixed*

strategy is one where two or more different actions are each chosen with some probability. Finally, the payoff function determines the outcome for each player in each scenario, that is, a combination of actions chosen by the players. If mixed strategies are present, the payoff function gives the expected payoff given the probability of each configuration of actions.
The simplest game is the 2-player, 2-strategy matrix game. In it, two players each have (the same) two strategies to choose from. They make their choice simultaneously (or equivalently without information about the other player's choice). The game can be summarized in the following payoff matrix:

	A	**B**
A	a	b
B	c	d

where A and B are the two strategies and a, b, c and d are the payoffs for the row player in each scenario. In such a symmetric game the row and column players are interchangeable, that is the payoff matrix of the column player corresponds to the transpose of the payoff matrix of the row player. If one player chooses action A and the other player B, then the player who has chosen A (B) receives the payoff b (c). It does not matter if a player is considered as row or column player.
Some of such games have become famous and the actions and payoffs can be given generic interpretations, usually denoted by:

	C	**D**
C	R	S
D	T	P

where C stands for "cooperation" and D for "defection", and the payoffs, denoted by their initials, are known as "Temptation", "Reward", "Punishment" and "Sucker's payoff". If $T > R > P > S$, the game is a Prisoner's Dilemma (PD), the "*E. coli* of social psychology" [23]. It corresponds to a situation where the players would both be better off cooperating, but because they will always have the incentive to defect, they end up choosing this inferior outcome, hence the "dilemma".
A common way to analyse games is by using a *solution concept*. A solution is a state of the game (in other words, a configuration of actions/strategies) that can be reasonably assumed to follow from choices made based on some underlying logic. Arguably the two most well-known examples – as well as the ones most often encountered in the context of evolutionary game theory – are the Nash equilibrium [128, 127] and the Evolutionarily Stable Strategy (ESS) [169]. In a Nash equilibrium, all strategies are chosen in such a way that no player has an incentive to unilaterally change theirs. An ESS is a Nash equilibrium, but adds the constraint that a small minority playing a different action cannot invade the population, adding the biological element to the picture. The condition for a strategy A to be an ESS is

$$\forall B \in S, B \neq A : P(A,A) \geq P(B,A) \tag{1.17}$$

$$\text{and if} \quad P(A,A) = P(B,A), \quad P(A,B) > P(B,B), \tag{1.18}$$

where S is the set of available strategies and $P(A,B)$ is the payoff for strategy A against B. A Nash equilibrium can comprise pure or mixed strategies. Similarly, an ESS can be pure or mixed. In the PD, for both players to choose D is the Nash equilibrium of the game: D dominates the other action in all scenarios ($T > R$ and $P > S$) [183]. In this case, it is also the ESS.
If the payoffs of the PD are switched so that $T > R > S > P$, the game is called Hawk-Dove, Snowdrift (SD), or Chicken. In contrast to the PD, in this situation it is still better to

cooperate even if one's partner fails to do so. Here the Nash equilibrium is to choose an action opposite of one's opponent. If mixed strategies are allowed, meaning a player can choose its action probabilistically, we have a mixed Nash equilibrium where both players follow the same strategy of choosing C with some probability (or a portion of the time). This is also an ESS, and can be interpreted as a population of individuals that comprises a mix of C- and D-players.

Another way to analyse matrix games is using the *replicator equation* [173, 79], which models the dynamics of the relative frequencies of a set of strategies in a well-mixed population. It is defined as:

$$\frac{dn_x}{dt} = n_x(A_x - \overline{A}) \tag{1.19}$$

where n_x is the frequency of the strategy x, A_x is the average payoff of an x-player in the population, $\sum_{i \in X} n_i A(x,i)$, X being the set of strategies present, and $\overline{A}$ is the average payoff of the population, $\sum_{i \in X} n_i A_i$. The underlying assumption in the replicator equation is that individuals exist in an essentially infinitely large population, meet others in pairwise encounters in a random fashion, and then increase or decrease in frequency based on how well they do (that is, what is their expected payoff) compared to the population average. In other words, successful strategies will increase in abundance and *vice versa*. However, the replicator equation can also be interpreted to describe a situation where an individual's payoff depends directly on the composition of the population. For example, in a microbial culture, the availability of a certain nutrient might depend on the proportion of the population that is using the said nutrient as their primary resource. If the choice of nutrient is taken to be the strategy, an individual would be expected to obtain higher payoffs when its strategy is rare in terms of frequency.

The simple matrix game can be extended by increasing the number of strategies. Arguably the most well known example is the hand game rock-paper-scissors, captured by the following matrix:

	Rock	**Paper**	**Scissors**
Rock	0	$-b_1$	a_1
Paper	a_2	0	$-b_2$
Scissors	$-b_3$	a_3	0

where all the parameters are positive. Because of the circular nature of the game, there is no pure Nash equilibrium or ESS. The existence of equilibria as well as the replicator dynamics depend on the relationships between the parameters (see [32]).

Increasing the number of strategies available does not in principle change the analysis, but can make things more difficult in practice. Namely, when the number of strategies grows, identifying solutions "by hand" can become infeasible. Computational strategies have been developed for the automated identification of both the Nash equilibria [22, 196] and the ESSs [73].

We can also relax the interchangeability of the players. For example, a game with two players with separate payoff matrices can be represented by (A, B), where

$$A = (a_{ij})_{i=1,\ldots,n;j=1,\ldots,m} \quad \text{and} \quad B = (b_{ij})_{i=1,\ldots,m;j=1,\ldots,n}.$$

The entries of A and B are the payoffs for players in roles 1 and 2 respectively and player 1 has n and player 2 m different actions to choose from. The entry $A[i,j]$ corresponds to the payoff of player 1 when player 1 chooses action i and player 2 chooses j. The replicator equation (1.19) extends readily to the bimatrix game, albeit with one caveat: it does not cover the dynamics *between* the two different types, only the frequencies of individual strategies within them.

Finally, the number of players can also be increased beyond two. In a multiplayer matrix game, the payoff structure is represented by a tensor, the order of which corresponds to the number of players. If the players are interchangeable, only one tensor is needed, otherwise each distinct payoff structure has its own tensor. The analysis of multiplayer matrix games is in general much more difficult than that of simpler games.
The most prominent multiplayer game is the public goods. In its simplest form, it can be seen as an extension of the PD to more than two players [74]. Each player again has a choice between cooperation (C) and defection (D). Cooperation has a cost c (which is paid by the individual) and yields a benefit rc where $r > 1$. All the benefits are summed together and distributed evenly amongst the n players. In other words, the payoffs to a C-player and a D-player respectively are given by

$$P(C,i) = \frac{irc}{n} - c \quad \text{and} \quad P(D,i) = \frac{irc}{n}, \tag{1.20}$$

where i is the number of cooperators in the group. It is easy to see that if $r/n < 1$, defection dominates. This equilibrium is Pareto inefficient, meaning that it is possible to increase the payoff for everyone, and often referred to as the tragedy of the commons [114].
The benefit is not required to be a linear function of the contributions. In the general (non-linear) public goods game the costs and benefits are given by generic functions in the number of cooperators:

$$P(C,i) = b(i) - c(i) \quad \text{and} \quad P(D,i) = b(i)$$

and the expected payoff for an individual in a population with a frequency x of cooperators is

$$E(C,i,x) = \sum_{i=0}^{N-1} p(i,x)(b(i+1) - c(i+1)) \quad \text{and} \quad E(D,i,x) = \sum_{i=0}^{N-1} p(i,x)b(i),$$

where $p(i,x)$ is the probability of finding oneself in a group of i (other) cooperators. The departure from linear costs and benefits allows for more interesting dynamics and the possible coexistence of cooperators and defectors. For examples of Public Goods games with nonlinear benefits, see [125, 75, 19].
A more general game with a set of discrete pure actions can be represented in the *extensive form*: a graph $G = (V, E)$ called a *game tree*, where the set of nodes V corresponds to choices made by the players (or by Nature) and the edges E connect these choices to subsequent choices for other players, and ultimately, the payoffs. An example is shown in fig. 1.2.
The very nature of the extensive form representation seems to suggest a sequential nature for the game, meaning that the players will take turns in choosing their actions. However, this need not be the case, as we can simply assume that at any given level of the game, the player making the choice does not know in which node they are. For example, in fig. 1.2 the female choice could be made without information about the male choice. This is usually represented visually by joining the corresponding nodes with a dashed line.
The Nash equilibrium of a game in extensive form can be found using backwards induction: starting from the terminal nodes and working backwards "up the tree", it is possible to determine the choice of a rational agent in each node. An example of such reasoning is given in fig. 1.2. This of course becomes more difficult if we assume that some choices were made without complete information (knowledge about previous choices). Finding the ESSs of a game in extensive form is in general more complicated (see [44]).
The action set can also be continuous. For example, in a continuous extension of the PD, the player chooses some contribution level $c \in [0, c_{\max}]$ to pay, to yield their opponent a benefit $b(c)$ that is a function of the contribution. The probability of choosing one of the discrete

actions can also be seen as a continuous strategy, for example, the probability to play C in the Hawk-Dove game.
Adaptive dynamics is a framework often used in conjunction with game theory, that combines the ecological and evolutionary time scales to study how strategies will evolve under natural selection [120, 52, 69]. Suppose that x is some continuous strategy and the whole population has adopted x. We assume that x undergoes small mutations so that occasionally a small number of mutants emerge with a strategy y that is close to x but different. We assume that the mutations are rare enough so that the time scales of population and evolutionary dynamics can be separated. If the mutation is beneficial, that is, if it receives a payoff in the reigning population that is higher than that of the *residents*, it will increase in frequency, and may replace x.
Assume that the population dynamics are governed by the replicator equation. Because the mutations are small and the mutant frequency initially low, we can approximate the growth rate of the mutant by the *invasion fitness*:

$$s_x(y) = P(y, x) - P(x, x) \tag{1.21}$$

The rationale is that the mutants will initially only encounter x individuals. If the invasion fitness is positive, the mutants will increase in frequency. Evolution will thus proceed in the direction of the *evolutionary gradient*

$$\left.\frac{\partial s_x(y)}{\partial y}\right|_{y=x} \tag{1.22}$$

until it reaches the neighbourhood of an *evolutionarily singular strategy*: an x^* such that 1.22 is equal to zero (or possibly the boundary of the strategy space).
The nature of the evolutionarily singular strategies can be determined by the second order derivative of the invasion fitness. Most notably, if

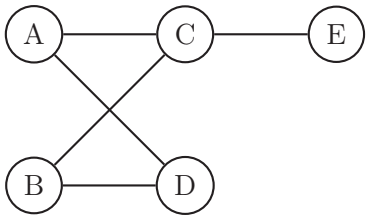
$$\left.\frac{\partial^2 s_x(y)}{\partial y^2}\right|_{y=x=x^*} < 0 \tag{1.23}$$

x^* is an ESS. A complete categorisation of singularities can be found in [69].
Adaptive dynamics can often complement a static analysis of a game. For example, it can turn out that an ESS is unattainable via the sort of gradual mutations considered in the framework. Moreover, through the phenomenon of evolutionary branching [69], adaptive dynamics can explain how the evolution of a monomorphic population towards higher fitness can eventually lead to polymorphisms, thus offering one potential explanation for diversity.

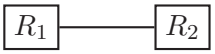
$R_1 : 2A + B \longrightarrow C + D$

$R_2 : 3C \longrightarrow E$

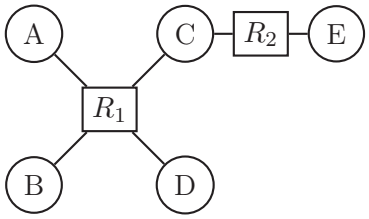
(a)



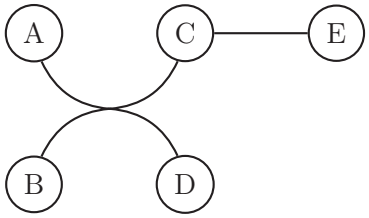
(b)



(c)



(d)



(e)

	R_1	R_2
A	-2	0
B	-1	0
C	1	-3
D	1	0
E	0	1

(f)

Figure 1.1 – A set of chemical reactions – metabolism – lends itself to several different graphical representations. a) An example of a set of reactions behind the different graphs. b) A compound graph. c) A reaction graph. d) A bipartite graph. e) A hypergraph. f) A stoichiometric matrix.

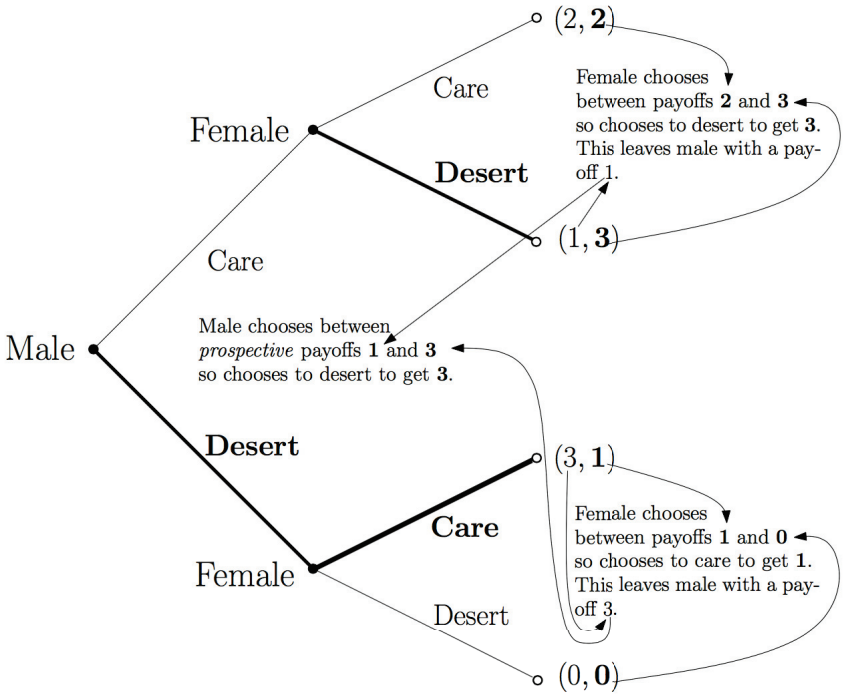


Figure 1.2 – From [32]. Extensive form representation of a brood care game where first the male decides whether or not to care for the offspring, and subsequently the female makes the same choice knowing already the choice made by the male, making the game a sequential one.

Chapter 2

Metabolic shifts

Contents

2.1 Introduction

Cheaper and higher throughput sequencing techniques and other technical advancements have made a wealth of new 'omics-data available in recent years. Transcriptomics measures the RNA levels inside a cell, revealing which genes are being expressed. Metabolomics quantifies the metabolite concentrations, offering further clues into the functioning of the metabolism. Proteomics is also emerging as a new source of information, measuring directly the levels of enzymes, and thus possibly circumventing some of the problems involved in using RNA levels to infer enzyme abundance that are caused by downstream regulation mechanisms.
Often we would like to combine these 'omics data sets with the information offered by a metabolic reconstruction. Since a metabolic network by itself leaves the metabolic state largely undetermined, it is of interest to seek clues about its actual workings by incorporating data gathered from a living cell. A metabolic network can also complement the analysis of 'omics data. Differential expression studies are routinely performed to understand phenotypic variation. However, a mere listing of up- or down-regulated enzymes is often not very informative. A metabolic network on the other hand, represents the relationships between these enzymes, offering a way to connect the observed changes in a meaningful way.
Approaches that incorporate 'omics data into a metabolic network can be roughly divided into two classes based on the representation used. When the network is represented by a graph, the problem usually takes the form of subnetwork extraction [13, 132, 7, 6, 24, 141]. Data about the objects contained in the network such as reactions and metabolites then serve as "edges or nodes of interest" that the method in question tries to join or otherwise include in a subgraph. In approaches that use constraint-based methods, the network is represented by a stoichiometric matrix. If the aim is phenotype prediction, that is, to infer a flux distribution, data such as transcriptomics which can serve as a proxy for enzyme activity can be used to

further constrain the flux space. Here again there are two different approaches, often described as implementing either "switches" or "valves". In methods that use gene expression data as switches, reactions whose genes were not measured at all are blocked [43, 5, 25, 35, 41]. In contrast, in a valve approach expression levels are used as indicators of the activity of reactions [163, 40, 108, 93, 95, 193, 175].

Gene expression data can also be used to build so-called context-specific models [165, 84, 4, 184, 149, 182, 192, 60, 155]. This refers to the process of pruning a network in order to obtain a subset of reactions that are active in a given condition or phenotype.

Constraint-based methods have moreover been adopted to help analyse gene expression data [126, 83, 152, 194, 145]. Methods such as differential expression (DE) analysis are often used to gain information about phenotypic variation. For example, it might be of interest to know how an organism reacts to changes in its environment. While such responses often take the form of a reorganisation of metabolism, direct measurements of internal reaction fluxes remain hard to achieve. In contrast, measuring the expression of genes is now routinely achieved genome wide, thanks to RNA-seq technology. In a DE measurement, gene expression is measured in two or more conditions, and the expression levels are compared. A statistical approach is used to test against the null hypothesis that the expression level of a gene remained the same throughout the different data points. What is obtained is a list of differentially expressed genes, along with associated *fold changes* which tell how much more or less RNA was found for a given gene. Differentially expressed metabolic genes, that is, genes that code for reaction-catalysing enzymes, will then in principle indicate which parts of the metabolism were affected. However, because of different regulatory mechanisms and the intrinsic noisy nature of gene expression data, there is no one-to-one correspondence between differentially expressed genes and changes in reaction activity. Thus DE analysis can greatly benefit from the incorporation of other sources of information, such as the metabolic network.

To this end, we developed a new algorithm, that we called MOOMIN for "Mathematical explOration of Omics data on a MetabolIc Network". It combines the information from a RNA-seq DE analysis with a genome-scale metabolic network to infer which parts of an organism's metabolism increased or decreased in activity. MOOMIN is based on the concept of a *feasible change*: given two different conditions, a change in metabolic activity between them should be such that it preserves a steady state (zero accumulation for all internal metabolites). In other words, assuming a steady state in condition 1, a feasible change from condition 1 to condition 2 guarantees that the metabolism is also in steady state in condition 2. The change in metabolic activity is expressed in qualitative terms as *colours*. Namely, we colour parts of the network green (red) to signify an increase (decrease) in metabolic flux. The gene expression information is leveraged by turning the results of a DE analysis into *weights* for the reactions of the network: evidence for a positive or negative change in expression will result in a positive weight that promotes the inclusion of the said reaction in the solution. Conversely, when there is little or no evidence of change, a negative weight is given to discourage including the reaction. The weights are calculated based on the strength of the evidence for a change in expression. An optimal solution is one that maximises the sum of the weights of the included reactions. The feasibility of the change is ensured using either topological or stoichiometric constraints.

This chapter is composed as follows. In Section 2.2, I review the existing literature on combining 'omics data with a metabolic network. In Section 2.3, I present the MOOMIN algorithm. The general idea and the main theoretical framework are introduced in Section 2.3.1. In this section, I also prove the computational complexity of the optimisation problem. Section 2.3.2 presents the extension of the algorithm to include reaction stoichiometry. In Section 2.3.3, I discuss results obtained by applying MOOMIN to two data sets retrieved from the literature. The work in this section was done in collaboration with Mariana Ferrarini. The chapter ends

with a conclusion.

2.2 State of the art

The first methodological distinction in approaches to combining metabolic reconstructions with 'omics data can be made based on the representation used. In graph-based methods, the metabolic network is represented by a graph whose nodes and edges correspond to the targets of measurement in 'omics investigations (for example, genes or metabolites).
In [13], the authors presented a web-based tool called KEGG SPIDER. KEGG SPIDER first forms a network of metabolic genes by linking together genes whose associated reactions share a metabolite. The input is a list of genes, for example the differentially expressed genes from a DE analysis. Network models are then formed iteratively by first joining together input genes of at most distance one, then distance two etc. The statistical significance of each model is assessed by comparing the number of input genes included to empirical distributions obtained using a simulation procedure with random genes as inputs.
Noirel *et al.* [132] presented a method that uses quantitative proteomics data to infer up- and down-regulated metabolic pathways. Their representation is an enzyme network where nodes correspond to enzymes, and they are joined if the product of one is the substrate of the other. The connections are weighed based on the connectivity of the metabolite(s) that gives rise to the edge following [46]: the weight is equal to the number of occurrences of the metabolite in the network. Based on proteomics data, four kinds of enzymes (nodes) are recognised: up-regulated, mildly up-regulated, down-regulated, and non-quantified. Up-regulated pathways are then extracted by performing a depth-first search, starting from each up-regulated enzyme and selecting paths with the following heuristics: the pathway cannot contain a down-regulated enzyme, the tail of the pathway cannot contain more than one non-quantified enzyme, the pathway must contain more than one up-regulated enzyme, the pathway cannot contain more than two non-quantified enzymes in a row, and the weight of the pathway (the sum of the edge weights) must not exceed a specified maximum. All paths fulfilling these criteria are then joined to form a subgraph. To find down-regulated pathways, the order of measurements can be reversed and the same procedure used.
Alcaraz *et al.* [7] presented KEYPATHWAYMINER, a tool for the extraction and visualisation of interesting subpathways based on a series of gene expression data. The starting point is any biological network that joins together genes based on known physical, regulatory, or genetic interactions. The nodes are categorised into up-regulated, down-regulated or unchanged. The goal is then to find maximal connected subnetworks that contain at most k nodes that remained unchanged in more than l conditions. Because the problem is computationally hard, ant colony optimisation [55] was used to find solutions stochatistically. An update to the software was published in [6] where the authors provided a branch and bound method to find exact solutions.
In [24], KEYPATHWAYMINER was developed further by adding an alternative search problem. Whereas in [7] the goal was to find maximal subgraphs with at most k nodes that were not differentially expressed in more than l conditions, here at most k nodes are allowed to have remained unchanged in at most l conditions *in total*. In other words, the exceptions are counted over all nodes and conditions. The authors reported that this was done to avoid bias towards hub nodes.
Pey *et al.* [141] proposed a method to explain the accumulation or depletion of a given metabolite in a disease phenotype. Their approach can be seen as a mix between the graph-based and contraint-based methods because they use a simple compound graph representation but include additional constraints so that all paths are required to include carbon exchange and to be able to sustain a steady state flux. Enzymes whose malfunction might be responsible

for changes in metabolite concentrations are identified using a *connectivity curve* approach. For an implicated metabolite, the connectivity curve is formed by measuring how many other metabolites can be reached from the focal metabolite by how many steps. The rationale is that if the connectivity curve changes drastically as the result of removal of some enzyme, this could indicate that the said enzyme is responsible for the change in the metabolite's concentration.

In contrast to graph-based approaches, constraint-based methods mainly make use of the stoichiometric matrix to represent a metabolic network. Many of these methods are aimed at improving the flux predictions made by FBA. The standard approach is to consider RNA-levels as a proxy for enzyme activity, and essentially use transcriptomics to further constrain the flux space. The idea of using gene expression data to further constrain the space of feasible fluxes was first presented in [43] by Covert and Palsson. Each reaction in a metabolic reconstruction is dependent on a set of genes. These dependencies are usually expressed in the form of Boolean equations. For example, $r_1 = (g_1 \text{ AND } g_2) \text{ OR } g_3$ means that r_1 needs either the genes g_1 and g_2 or the gene g_3 to be expressed in order to operate. If gene expression is measured and a given gene is not found at all, it can be concluded that said gene is not active, corresponding to an OFF or 0 in the Boolean equation. Given the activity states for all associated genes, if the Boolean value of a reaction is 0, the reaction cannot operate and its flux can be constrained to zero.

In [5], Åkesson *et al.* applied the methodology presented in [43] to gene expression measurements from chemostat and batch cultures of *Saccharomyces cerevisiae*. Flux predictions were obtained by standard FBA using biomass production as the objective. The authors compared the predicted fluxes to fluxes measured using carbon labeling, concluding that incorporating gene expression measurements improved flux predictions in batch cultures.

Chandrasekaran *et al.* presented PROM [35], a method that takes as its input a metabolic network, a regulatory network structure, gene expression data from various different conditions, and additional enzyme regulation information. It then constructs an integrated metabolic-regulatory network. Gene states and gene-TF interactions are represented by probabilities. The probabilities are then used to constrain the fluxes of the associated reactions. Finally, flux distributions are predicted using FBA.

In [41], Collins *et al.* introduced TEAM, an FBA variation based partially on GIMME developed by Becker *et al.* [25] (see later). TEAM takes as its input a series of gene expression measurements, information about the starting composition of growth medium and a series of biomass density measurements. The method then predicts the flux distribution at a series of time points using the input data. At each time step, the medium composition is updated and intake fluxes are constrained accordingly. A flux distribution is then found that both minimises the discrepancy between fluxes and the gene expression data, and satisfies the measured biomass production level. The gene expression data is imposed by assigning penalties to reactions that carry flux when their genes are lowly expressed. Low expression is determined using a threshold inferred from the overall expression data. In case the optimal flux vector is not unique, the sum of fluxes is minimised.

In contrast to the "switch-based" methods presented above, expression levels have also been used to constrain fluxes continuously, assuming that the abundance of transcripts for a gene is proportional to the maximal flux that can pass through its associated reaction. Schwarz *et al.* [163] presented YANA, a toolbox for the analysis of metabolic networks that includes means to incorporate transcriptomics and proteomics data. YANA computes the elementary flux modes (EMs) of the network, with a decomposition step based on metabolite connectivity preventing a combinatorial explosion in the number of EMs. It then finds a flux distribution that minimises the mean squared error between fluxes and enzyme activities. The enzyme activities are inferred based on the expression or proteomics data. This can then be used to

estimate the activity of the EMs.
In [40], Colijn *et al.* introduced E-Flux, an FBA extension that uses gene expression data to derive upper bounds for flux values. The bounds are equal to the normalised value of the expression of the associated gene, and in the case of multiple genes, AND is replaced my a minimum and OR by a sum.
Arguing against the use of biomass yield as the object of optimisation in FBA, Lee *et al.* [108] formulated an alternative objective function using absolute levels of gene expression. Namely, Lee *et al.* minimise the difference between flux values and expression levels, weighted with the accuracy of the expression measurement, for those reactions for which data is available. AND-relationships are again replaced with minimums and ORs with sums.
Kim and Reed [94] presented Relatch, a method that predicts fluxes in a perturbed state using gene expression and flux measurements. First, a reference flux distribution and corresponding enzyme contributions were estimated. Then, fluxes in a perturbed state were predicted by minimising the adjustment to the reference state, consistent with the estimated enzyme contributions.
Another method that predicts fluxes in a perturbed state, Gx-Fba, was presented in [131] by Navid *et al.*. Similarly to [94], a flux distribution is first found for a reference state, after which new constraints are added based on gene expression measurements to predict the flux distribution in a perturbed state.
Kim *et al.* [93] first identified genes whose expression would correlate sufficiently with the activity of their associated reactions. This was done using gene expression and flux measurements from a chemostat culture of *Escherichia coli* with varying dilution rates. After having identified such genes, their expression levels were used to constrain fluxes in single-gene knockout experiments. The authors conclude that this produced on average better flux predictions than simple FBA.
In [95], Kim *et al.* offer an update to the previously introduced E-Flux in the form of two new methods: E-Flux2 and Spot. The goal in both is to improve FBA predictions using gene expression data. E-Flux2 uses a biological objective function and constrains fluxes using the absolute values of gene expression as upper bounds. To find a unique optimum, in a second step the l^2 norm of the flux vector is minimised. If no suitable biological objective is available, Spot can be used. It formulates an objective function that maximises the correlation between flux values and the gene expression levels.
Zhang *et al.* [193] introduced Hpcof to predict flux distributions without the need for a biological objective function. In their formulation, agreement between fluxes and gene expression measurements is ensured using a Huber penalty function. The Huber penalty function is a convex function composed of a quadratic and a linear part that is meant to increase robustness against outliers. The objective function in Hpcof is a combination of the Huber penalty which takes as its input the difference between the flux and the gene expression measurement, and the l^1-norm of the flux vector. To obtain an expression level for a reaction with multiple associated genes, ANDs are replaced with minimums and ORs with maximums.
Tian *et al.* [175] presented LbFba They first estimated linear relationships with the expression levels of genes and the upper and lower bounds of reaction fluxes using a training data set. These bounds were then used to constrain fluxes along with a slack variable that prevents infeasibility. The optimisation procedure simultaneously minimises the slack variables and the l^1-norm of the flux vector. In gene-reaction relations, ANDs were replaced with minimums and ORs with sums.
Improvements to phenotype prediction can also be achieved by forming context-specific metabolic models. Assuming that the metabolic state of an organism varies from one condition to another, or alternatively from tissue to tissue in multicellular organisms, a metabolic network can be pruned to remove reactions that are considered to be inactive in a given context.

In [25], Becker *et al.* introduced GIMME to create context-specific metabolic models. The method takes as its input a metabolic network, a set of gene expression measurements and one or more metabolic functions that the cell is required to achieve. An LP-problem is solved which minimises an inconsistency score formulated as a function of normalised expression levels: if a gene's expression is below a set threshold it contributes to the inconsistency score. The minimisation is done conditional to the required metabolic functions being fulfilled up to some percentage. Finally, reactions that carry no flux are deemed inactive and removed.
Shlomi *et al.* first demonstrated their method for creating context-specific models in [165]: it was later called iMAT and released as software in [197]. Shlomi *et al.* first categorise genes into three different classes: highly, lowly, and normally expressed using thresholds and absolute gene expression levels. This categorisation further induces the same for reactions. A MILP-problem is then solved that produces a flux vector while simultaneously maximising the number of highly expressed reactions and minimising the number of lowly expressed reactions that carry flux. Reactions with zero flux are then removed from the network.
A network pruning method that its authors Jerby *et al.* [84] called MBA has served as the inspiration for many subsequent approaches. MBA is a generic formulation that is able to incorporate data from various sources. In the original publication, Jerby *et al.* used human-curated tissue specific pathways and molecular data to identify core reactions that are strongly believed to be active. The model is then pruned under the condition that all reactions must be able to carry flux, with the aim of removing as many non-core reactions as possible while simultaneously preserving the core reactions.
Agren *et al.* [4] introduced INIT, a method for creating context-specific models similar to iMAT. In addition to integrating gene expression, INIT also imposes the production of metabolites that have been detected in the tissue in question.
In [184], Wang *et al.* presented mCADRE. Their method also uses gene expression data to produce context-specific models. Genes are first categorised into active or inactive across a set of samples based on if they were detected or not. This gives rise to a score function that takes into account in how many samples a given gene was detected. These scores are mapped to reactions using the Boolean functions, replacing AND-relations with minimums and ORs with maximums. Additionally, reaction scores are influenced by network topology: each reaction influences the score of its surrounding reactions, and the strength of the influence is inversely proportional to the connectivity of the reaction. Reactions are also evaluated based on the biological evidence for their presence. The network is then pruned, starting from the lowest scoring reaction. Reactions are removed under the condition that a set of core reactions remain functional and the production of a set of key metabolites is preserved.
Rossell *et al.* [149] expanded on iMAT with EXAMO. The authors started with the MILP-problem presented in [165]. By solving the problem with every reaction forced as active and inactive in turn, they determined reactions that were active or inactive in all optima of the original problem. The network is then pruned under the condition that all the active reactions must carry flux.
FASTCORE, introduced by Vlassis *et al.* [182], is another "generic" method for creating context-specific models. It takes as its input a set of key reactions that must be preserved. It then removes all possible reactions under the condition that the core reactions must be able to carry flux in at least one feasible flux distribution.
Yizhak *et al.* [192] presented PRIME. The method first aims to reduce the space of feasible fluxes by incrementally lowering the upper bounds of fluxes. The bounds are lowered until biomass production is affected. Next the correlation between fluxes and gene expression is evaluated. For reactions that correlated well with the expression of their associated genes, new flux bounds are formulated that are linearly dependent on expression levels.
REGREX, presented by Estévez and Nikoloski in [60], uses the least absolute shrinkage and

selection operator, or LASSO, to simultaneously enforce agreement with gene expression data and remove inactive reactions. Namely, the objective function in the mixed integer quadratic programme formulated by RegrEx consists of a part that minimises the difference between fluxes and expression or protein levels and a part that minimises the l^1-norm of the flux vector. In [155], Schultz and Qutub introduced Corda to produce context-specific models. Corda assesses the dependency of reactions with high experimental evidence on reactions with no evidence by creating pseudo-metabolites that impose a cost on all reactions. "Undesirable" reactions have higher costs. The pseudo-metabolite formulation allows for the use of standard FBA and thus linear programming as opposed to mixed-integer linear programming which is more computationally demanding. The network is then pruned with the goal of preserving desirable reactions.

Finally, a metabolic network can be combined with gene expression data to understand metabolic shifts. Changes in the expression levels of metabolic genes are often used to understand phenotypic changes that occur as the result of a change in conditions, or for example a gene deletion. Combined with a metabolic network, and especially constraint-based methods, rather than simply listing differentially expressed genes, one can infer global shifts in metabolic state.

In [126], Moxley *et al.* investigated changes in yeast metabolism resulting from the removal of a global regulator Gcn4p. The authors sought to predict fluxes in the amino acid biosynthesis following the knock-out using a metabolic network augmented with condition-specific transcription factor binding interactions, protein-protein binding interactions, enzyme-reaction interactions, reaction-metabolite interactions, and metabolite-enzyme interactions. Namely, a flux change was predicted using the equation

$$\Delta\text{flux} = e^{-p_1 d_{interaction}\frac{\Delta\text{mRNA}}{p_2}} \tag{2.1}$$

where p_1 and p_2 are parameters, and $d_{interaction}$ is the "metabolite interaction density", defined as "the ratio of the number of metabolite-enzyme interactions to that of the total reaction enzymes in the pathway". Additionally, the predicted fluxes were required to fulfill the steady state condition. Inclusion of the metabolite interaction density was done to account for feedback inhibition and other enzyme-level regulation. In other words, Moxley *et al.* hypothesised that more correlation between mRNA and flux changes should be observed in a less-connected regulatory network.

Jensen *et al.* ([83]) described Made, a method that takes as its input a time-series of gene expression measurements, and finds a series of binary activity states for all reactions in a metabolic network. When DE analysis is performed for the time series data, one obtains a list of transitions at each two consecutive time points. Namely, $d_{i\to i+1}$ is equal to 1 (-1) if a statistically significant increase (decrease) in expression took place, and to 0 if there was no change. Significance is determined by the associated p-value. Made then tries to find a sequence of binary expression states for the genes that best matches the changes $d_{i\to i+1}$. Each proposed expression state is scored $-\log(p)$, where p is the p-value associated with the change, and the sum of these scores forms an objective function in an MILP problem. For example, if an increase in expression was observed with a p-value 0.01, that is, $d_{i\to i+1} = 1$ and $p = 0.01$, the proposed expression states $x_i = 0$ and $x_{i+1} = 1$ would contribute 2 to the objective function. In contrast, in a case where no change was observed, $d_{i\to i+1} = 0$ and $p = 0.5$, $x_i = 0$ and $x_{i+1} = 1$ would be penalised with -0.3. An optimal series of expression states was found by maximising the afore-described objective function across all time points while simultaneously requiring that the steady condition is satisfied and some fraction of optimal biomass production achieved at each time point. Reactions' dependence on the expression states of genes was implemented by converting the Boolean equations into integer inequalities following [166].

In [62], similarly to [126], Fang *et al.* sought to predict the flux distribution in a perturbed state using the results of a DE analysis. The authors first used FBA and flux variability analysis (FVA, [117]) to establish baseline reference values for fluxes. This was done by calculating the minimum and maximum flux for each reaction under optimal biomass production and then finding a feasible flux distribution close to the mean of the two bounds. Next, a gene expression ratio was assigned to each reaction. The value assigned was obtained by replacing AND relations with the geometric mean and OR relations with the arithmetic mean. If a set of genes was associated with several reactions, the ratio obtained using this procedure was assigned to the overall normalised flux of the said reactions. Fluxes in the perturbed state were predicted using the constraints

$$|x| \leq (R + L)|x_{ref}| \quad \text{if} \quad R < 1 \tag{2.2}$$

$$|x| \geq (R - L)|x_{ref}| \quad \text{if} \quad R > 1 \tag{2.3}$$

where x is the flux in the perturbed condition, x_{ref} is the reference flux, R is the expression ratio, and L is a slack variable. Additionally, alterations in the biomass composition and upper bounds of uptake rates were allowed by introducing another set of slack variables into the biomass reactions coefficients and the bounds for uptake reactions. To obtain the predicted flux distribution, first the sum of the slack variables L was minimised, and after that, the remaining slack variables. Finally, the minimum and maximum flux through each reaction was determined given the optimum obtained, and another problem similar to the first step (establishing the baseline fluxes) was solved to obtain a distribution close to the mean of the minimum and maximum bounds.

Rezola *et al.* [145] searched for EFMs characteristic to specific expression measurements. The authors first used an extension of the algorithm introduced in [51] to compute the K shortest EFMs that differ in at least 5 reactions in the human genome-scale metabolic network. Next, reactions were categorised into highly, moderately, or lowly expressed based on the expression data following [165]. To associate EFMs with physiological conditions, a multivariate hypergeometric test was used to gauge if more highly and less lowly expressed reactions were contained within a given EFM than would be expected by chance.

In [152], Samal *et al.* presented a method that uses sparse group lasso [167] to find extreme currents (ECs) associated with particular phenotypes. First, ECs were calculated, which then gave rise to sets of genes according to the gene-reaction associations. Only unique sets were considered. Because there is considerable overlap within these sets, complicating the analysis, the sets were further clustered using agglomerative hierarchical clustering. Next, a feature matrix was constructed using the expression data, which contains expression measurements for all the genes in multiple samples. The samples are labelled either discretely (for example "healthy"/"sick") or continuously (quantitative response to treatment). Finally, it was assumed that the phenotype can be predicted from the feature matrix using a linear model, and the model was fitted using the sparse group lasso method.

Zhu *et al.* [194] proposed a method for inferring differences in metabolism between two different conditions using gene expression data. First, an expression level was calculated for each reaction by averaging the expression measurements of all its associated genes. Next, to detect reactions that changed their expression levels significantly from one condition to the other, a t-test was first performed to obtain p-values, and then reactions with a Benjamini & Hochberg corrected q-value [26] under 0.05 were deemed as significantly changed. The question of metabolic changes was formulated as an MILP problem where two flux vectors are searched for, one for each condition. Both vectors have to fulfill the standard steady state constraints. Agreement with the expression levels is enforced by an objective function that penalises flux changes that disagree with the observed changes. Namely, if for example a reaction that was up-regulated according to the data has a lower flux value in the second

condition, a penalty is issued proportional to the difference in fluxes. This objective function is then minimised. To account for possible non-unique optima, a modified problem was solved for each reaction to determine if it was predicted to be up- or down-regulated in all optimal solutions.

The literature pertaining to combining 'omics data with metabolic reconstructions has been previously reviewed in several articles. Machado and Herrgård [115] surveyed the various methods for integrating gene expression data into constraint-based methods, categorising them based on the input data (absolute or relative), the treatment of the gene expression levels (discrete or continuous), and the goal of the method (flux prediction, model building or both). The authors also performed a systematic evaluation of seven different methods, GIMME [25], iMAT [165], MADE [83],E-FLUX [40], Lee *et al.* [108], RELATCH [94], and GX-FBA [131], with pFBA, an FBA variant that does not use any data besides the network, used as the baseline. All of the seven methods failed to consistently provide better flux distribution predictions than pFBA. This failure was observed with both transcriptomic and proteomic data. While the methods that utilise relative expression changes rather than absolute levels of expression did not perform better than their counterparts, Machado and Herrgård still argued that relative changes might be better suited for the purposes of improving flux predictions, due to the difficulty of establishing a proper correspondence between absolute transcript levels and metabolic fluxes.

In [96], Kim and Lun grouped methods for the integration of transcriptomic data into metabolic networks according to four criteria: requirement of multiple gene expression measurements, requirement for a threshold for characterising expression levels ("low expression", "high expression" etc.), requirement of a biological objective function, and validation of predictions against measured fluxes. The authors argued that an ideal method would satisfy the following criteria: a single expression data set as input, utilisation of continuous expression values, no need for a biological objective function, and predictions having been validated against measured fluxes. Kim and Lun concluded that at the time of writing, no method satisfied all of the criteria.

Opdam *et al.* [135] performed an evaluation of six different methods for constructing context-specific models, FASTCORE [182], GIMME [25], iMAT [165], INIT [4], MBA [84], and mCADRE [184]. The authors found that the most important parameter influencing model content was the expression threshold which determines when a gene is considered active. While more stringent thresholds lead to more accurate gene-essentiality predictions, they also reduced the number of included metabolic functions. Additional constraints on uptake and secretion fluxes had a smaller effect on gene-essentiality, but influenced more the ability to recover metabolic functions. Opdam *et al.* recommended gene specific thresholds for expression and defining *a priori* known metabolic functions whose inclusion is enforced to improve model construction.

For other reviews on the integration of 'omics data into metabolic networks, see [146, 151, 181].

In the next Section, I present our contribution to this research: the MOOMIN algorithm. While considerable research effort has gone into the integration of metabolic networks with 'omics data, especially transcriptomics, most of it has been dedicated to improving phenotype prediction through methods such as FBA. Only a few methods have been presented for the interpretation of DE data.

MOOMIN adheres to many of the suggested "good practices" for combining 'omics data with a metabolic reconstruction. As the method is aimed at DE data, it naturally uses relative rather than absolute gene expression levels. There is thus no need to assign thresholds for absolute expression levels. Technically, two measurements are needed for the input but this is of course the whole idea behind DE analysis. There is also no need for a biological objective function. Unfortunately, we have not been able to evaluate the changes predicted by MOOMIN against observed fluxes. This is due to the lack of a suitable data set. While studies comprising

both expression and flux measurement data have been published, these data sets are all done with microarray technology. We opted for designing MOOMIN to use DE results obtained using Bayesian statistics, which in turn made the use of microarray data infeasible. However, it seems clear that RNA-Seq will replace the microarray technology completely in the near future, and such data will become available. Furthermore, the use of Bayesian statistics offers the benefit of more robustly quantifying the certainty about genes *not* changing in the level of expression. In contrast to [83], we do not assign binary on/off states to differentially expressed genes, something that might be too restrictive for the model as metabolic re-organisation does not necessarily entail simply activating and de-activating reactions.
The most closely related model to MOOMIN is the one of Zhu *et al.* [194]. The main difference is that Zhu *et al.* largely ignored evidence for no change occurring in expression levels. To my knowledge, an implementation of the method described in [194] has not been made available.

2.3 MOOMIN

The inspiration for MOOMIN were previous methods developed by my team for the analysis of differential metabolomic data. GOBBOLINO & TOUCHÉ [1, 121] used a compound graph, and TOTORO [85] a hypergraph representation to decipher metabolic changes occurring in a transitory state between two equilibria. Both approaches aimed at providing an explanation for changes in metabolite pools in terms of a subgraph. The subgraph consists of reactions that are considered to have taken part in the changes of metabolite concentrations. The problem is then formulated in terms of selecting edges/arcs (reactions) and assigning directions to them: a reaction may have exhibited either an excess or a decreased flux in the transitory state. What is required is that the edge directions *explain* the observed changes in metabolite concentrations. This translates to balance requirements surrounding nodes.
In MOOMIN the methodology is similar, but the question slightly different. MOOMIN was developed to complement DE analysis by leveraging the structural information contained in a metabolic network to further interpret the results. Namely, given the results of a DE analysis pipeline – changes in gene expression – MOOMIN tries to infer a metabolic shift. The underlying logic is that when conditions change, the cell re-organises its metabolism to adapt. While it has been recognised that exactly correlating fluxes with gene expression levels is difficult, we believe that changes in expression can nevertheless serve as clues as to how the metabolism was impacted. Moreover, when the "clues" offered by the transcriptomics data are combined with the network, it is possible to filter out the inherent noise. More specifically, instead of simply projecting the expression changes onto reactions, we require for the metabolic shift to be consistent. In other words, only changes that seem to be supported on the network level – that "fit together" with the changes in surrounding reactions – are accepted, and those that are not are rejected and assumed to be unrelated to the metabolic re-organisation.
I present here two formulations of the problem, a topological one and a stoichiometric one. The former can be seen as a special case of the latter.

2.3.1 Topological formulation

The metabolic network is represented by a hypergraph $G = (V, R)$, where $V = \{v_1, \ldots, v_n\}$ is a set of nodes corresponding to metabolites and $R = \{r_1, \ldots, r_m\}$ is a set of hyperarcs corresponding to reactions. A hyperarc $r_i = (\mathrm{Subs}(r_i), \mathrm{Prod}(r_i))$ is an ordered pair that contains respectively the substrates and products of the corresponding reaction. We will use the words network and hypergraph, metabolite and node, and reaction and hyperarc interchangeably when appropriate.

A *colouring* of a network G is a vector of hyperarc colours $\mathbf{c} = (c_1, \ldots, c_m) \in \{\text{green}, \text{red}, \text{grey}\}^m$. We will call a hyperarc r_i *coloured* if $c_i \in \{\text{green}, \text{red}\}$. The colours "green" and "red" signify respectively an increase and a decrease in flux while "grey" designates no change. The *a priori* colouring $\mathbf{c}_0$ of a hypergraph is a colouring inferred based on a gene expression data set. In it only reactions for which we have direct evidence of change are coloured. In other words, it represents the raw mapping of the gene expression data to the network, based only on gene association relationships of the reactions. It serves as part of the input of the algorithm and imposes constraints on accepted output colourings.
An input hypergraph G also has associated with it a weight vector $\mathbf{w} = \{w_1, \ldots, w_m\}$ that directs the search algorithm and is also derived from the expression data. The purpose is to promote choosing reactions that are deemed most likely to have undergone a change by placing a positive weight on them and to discourage choosing reactions for which we have no evidence of change by assigning a negative weight. Consequently, *a priori* coloured hyperarcs have positive and grey hyperarcs negative weights.
The colours and weights of the hyperarcs are determined by the associated genes and the expression data (or lack thereof). We first assign colours and weights to genes and then use gene association information to map them onto the corresponding hyperarcs.
The colour of a gene is dictated by both its probability to be differentially expressed and the fold change observed. If the gene exceeds the threshold for differential expression (later referred to as the parameter t in the weight function), the gene is considered coloured and the direction of the change, that is, the sign of the fold change, determines if the gene is green or red. If the probability of differential expression is below the threshold or data on the gene is missing due to it containing outliers, the gene is grey. Genes that were not detected at all or were detected at extremely low levels are deemed inactive. For these genes, the viability of the associated reactions is checked through the Boolean functions describing the gene dependency and any reactions deemed inviable are removed from the network. These reactions are thought to carry no flux in either condition and thus there is no sense in trying to infer if they changed or not.
We assign weights to genes according to the function

$$\min\{\beta(-\log(1-p) + \log(1-t)), -\alpha\beta\log(1-t)\}, \tag{2.4}$$

where p is the posterior probability for the gene to be differentially expressed and t, α and β are parameters. The parameter t is the threshold above which we deem a gene to be differentially expressed. Genes that exceed it receive a positive weight and those below a negative one. The parameter α controls the relationship between the positive and negative weights: if, for example, $\alpha = 3$, the highest possible positive weight is three times the (absolute value) of the lowest possible weight. The significance of α is further illustrated in Figure 2.8: in practice, the higher the value, the more coloured hyperarcs will appear in the solution. Finally β is a shape parameter controlling the derivative of the function. The weight function is plotted in Figure 2.1 with the parameters $t = 0.9$, $\alpha = 3$ and $\beta = 2$.
Contrary to determining the viability of a reaction, we do not use the Boolean functions to map the weights onto the reactions. Rather we simply consider the associated genes as a set. The reasoning is that while the Boolean function is well suited for determining the presence or absence of the proteins required for a reaction to take place – an undetected gene corresponding naturally to the truth value false – the dynamics involved in a change in activity are much more complex. For example, in an AND-relationship one of the enzymes might be a bottleneck, meaning that the entire regulation mechanism rests on the expression level of the said enzyme.
For the aforementioned reason, we also wish to be quite liberal in assigning colours. More precisely, regardless of the underlying dependence structure and the number of genes, if one

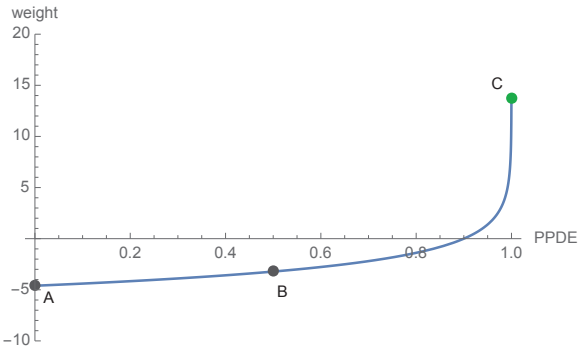
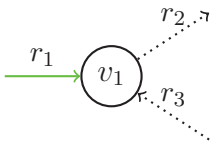


Figure 2.1 – The weight function plotted for parameter values $t = 0.9, \alpha = 3$ and $\beta = 2$. Point A would be the weight of a reaction whose genes did not change their expression or that does not have any known associated genes. B is the weight of a reaction that there is uncertainty about: either a probability of differential expression that does not point in either direction or conflicting evidence from different genes. Finally C is a reaction that has at least one gene showing strong evidence for differential expression.

of the genes is coloured, we consider the reaction coloured as well (excluding of course a situation where two genes indicate opposite colours). The specific rules for assigning colours and weights to reactions are as follows:
1. If at least one of the genes is green (red), the reaction is green (red), and its weight is the maximum of the weights of its genes.
2. If there is at least one gene that is green and one that is red, that is, the colours of the genes contradict each other, the reaction is grey and receives the weight of a gene with 0.5 probability of differential expression.
3. If all the genes are grey, the reaction is grey, and its weight is the maximum of the weights of the genes.
4. If there is no gene association information for the reaction, it is grey and will receive a weight of a gene with 0 probability of differential expression.
Our underlying assumption is that the cell is transforming from one steady state (see Section 1.1.2) to another. Consider the following local situation:



where we have already coloured reaction r_1 green to indicate an increase in flux and we are interested in what this implies for the flux changes in reactions r_2 and r_3. Let $\mathbf{f} = (f_1, \ldots, f_m)$ and $\mathbf{f}'$ be respectively the flux vectors in the control and the changed conditions. The steady state condition gives us the equations

$$a_{11}f_1 - a_{12}f_2 + a_{13}f_3 = 0 \tag{2.5}$$
$$a_{11}f'_1 - a_{12}f'_2 + a_{13}f'_3 = 0 \tag{2.6}$$

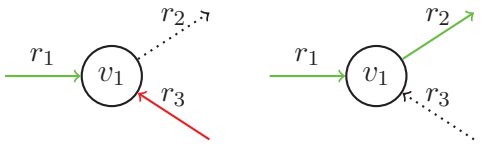
for metabolite v_1, where a_{ij} is the stoichiometric coefficient of metabolite i in reaction j. Let now $\Delta\mathbf{f}$ stand for the difference in fluxes between the two conditions, that is, $\Delta\mathbf{f} = \mathbf{f}' - \mathbf{f}$. The colour green for reaction r_1 corresponds to $\Delta f_1 > 0$. Combining the equations in 2.5, we have

$$a_{11}\Delta f_1 - \Delta a_{12}f_2 + \Delta a_{13}f_3 = 0 \tag{2.7}$$

This implies the following

$$\Delta f_2 \leq 0 \quad \Rightarrow \quad \Delta f_3 < 0 \tag{2.8}$$
$$\Delta f_3 \geq 0 \quad \Rightarrow \quad \Delta f_2 > 0. \tag{2.9}$$

The equalities correspond respectively to the scenarios:



More generally, this dynamic gives rise to the *balance* constraints:

Definition 1. *Given a hypergraph G and a colouring of G $\mathbf{c}$, a node v is* ***balanced*** *if at least one of the following conditions holds:*
1) there exists two coloured hyperarcs r_i and r_j such that $v \in Prod(r_i)$ and $v \in Subs(r_j)$, and $c_i = c_j$
2) there exists two coloured hyperarcs r_i and r_j such that $v \in Subs(r_i)$ and $v \in Subs(r_j)$, and $c_i \neq c_j$
3) there exists two coloured hyperarcs r_i and r_j such that $v \in Prod(r_i)$ and $v \in Prod(r_j)$, and $c_i \neq c_j$.

All solutions must render all nodes balanced, with the exception of sources and sinks of the network: these are metabolites that can be imported into or exported out of the cell and so they need not maintain a steady state. In addition to balancing the nodes, we also require that the solution does not violate the *a priori* colouring. Namely, an accepted solution can leave an *a priori* coloured hyperarc grey, but it cannot change its colour to the opposite one.

Definition 2. *Let G be a hypergraph and $\mathbf{c}_0$ its a priori colouring. A node $v \in V$ is* ***consistent*** *if at least one of the following conditions holds:*
1) there are no hyperarcs $r \in R$ such that r is coloured and $v \in Subs(r)$ or $v \in Prod(r)$
2) v is a source or a sink of G
3) v is balanced.
A colouring $\mathbf{c}$ of G is consistent if every node $v \in V$ is consistent under $\mathbf{c}$ and for all $r_i \in R$ it holds that if $c_{i0} =$ green, $c_i \neq$ red, and if $c_{i0} =$ red, $c_i \neq$ green.

Producing a hypothesis to explain the expression data is now expressed as the optimisation problem

$$\text{find} \quad \arg\max_{\mathbf{c} \in \mathcal{C}_G} \sum_{i \text{ s.t. } c_i \in \{\text{green,red}\}} w_i, \tag{2.10}$$

where $\mathcal{C}_G$ is the set of all consistent colourings of the hypergraph G. We will call this the *colouring problem.*
So far we have not discussed the question of reversible reactions: in theory, every enzymatic reaction can potentially happen in both directions. However, due to thermodynamic constraints, many of them do not. In practice, in a curated metabolic network all reactions are transcribed as directional with one set of compounds as the substrates and another as products. An additional attribute tells if the reaction is reversible or not. Whether or not the reaction does in fact operate in both directions, and if the "primary" direction is in some sense to be favoured, depends on the particular reconstruction.

The aforementioned poses a practical problem that should be reckoned with before any analysis on the network is done. For the purposes of the algorithm, we can assume we know for every reaction whether it is reversible or not. A standard approach then would be to add for every reversible reaction its reverse, usually along with some constraints for excluding cycles composed of these pairs. However, this is not desirable in our case.
Consider the situation in Figure 2.3. Above, a reversible reaction r_2 is still grey, and the nodes B and C need to be balanced. Because r_2 is reversible, both colouring it green with the interpretation that it operates in the $B \rightarrow C$ direction and colouring it red with the opposite direction would balance the two nodes. Two solutions differing only in the colour of this reaction would thus be considered separate. This could potentially cause a combinatorial explosion in the number of distinct solutions.
We solve the above problem by implementing the actual search in the following fashion. First, the network is constructed so that each hyperarc appears only once, that is, in only one direction, all the while conserving the information about potential reversibility. Next, the reverse of *every* hyperarc is added to the network, but as explicitly reversed. In other words, there are two kinds of hyperarcs: the original ones and their reverses. Colouring the network now corresponds to choosing hyperarcs. Choosing a hyperarc, that is, an original hyperarc, corresponds to colouring it green. Choosing the reverse corresponds to the colour red. Finally, simply not choosing a hyperarc, neither its reverse, means to colour it grey.
Obviously, we can only choose a hyperarc or its reverse but not both. Furthermore, the constraint in Definition 2 3) about not choosing colours that contradict the *a priori* colouring translates to the constraint that an *a priori* green hyperarc cannot be chosen in reverse and *vice versa*, except if the reaction is reversible. A reversible reaction that is *a priori* green and chosen in reverse in the search is later interpreted to still be green but operating in the other direction. The same goes for the red hyperarcs.
Unfortunately, this leaves some ambiguity for the reversible, *a priori* uncoloured hyperarcs, as their final colour cannot be deduced from the solution hypergraph. However, this is exactly why we wish to avoid counting the interpretations as separate solutions. Furthermore, we expect that in subsequent analyses, the "true" colour (and direction) can often be inferred from the surrounding hyperarcs. For example in Figure 2.3, if r_2 were a part of a linear pathway, it would be clear that the interpretation on the left is the correct one. We will refer to these undetermined reactions as "yellow" when needed. An example of the hypergraph transformation and the subsequent interpretation of the solutions can be seen in Figure 2.3.1.
I will now give a more formal definition of the transformation.
Let G^* be the transformed hypergraph that is formed from G by adding for every hyperarc $r = (\mathrm{Subs}(r), \mathrm{Prod}(r))$ in R its reverse $r^* = (\mathrm{Prod}(r), \mathrm{Subs}(r))$ (we assume that no hyperarc in G already has a reverse hyperarc). We call the set of reversed hyperarcs R^*, and thus $G^* = (V, R \cup R^*)$. A solution is a subhypergraph of G^*, obtained by choosing hyperarcs, their reverse hyperarcs, or neither. We denote such a selection by $\mathbf{d} \in \{+, 0, -\}^m$, where $d_i = +$, $d_i = 0$, and $d_i = -$ stand respectively for choosing r_i, choosing neither r_i nor r_i^*, and choosing r_i^* to be included in a solution. An *a priori* selection $\mathbf{d}_0$ is formed based on the *a priori* colouring $\mathbf{c}_0$ in the following way:
1) if $c_{i0} =$ green and r_i is not reversible, $d_{i0} = +$.
2) If $c_{i0} =$ red and r_i is not reversible, $d_{i0} = -$.
3) Otherwise $d_{i0} = 0$.

The equivalent of Definition 1 for the transformation is:

Definition 3. *Given a transformed hypergraph G^* and a hyperarc selection $\mathbf{d}$, a node v is **balanced** in the subhypergraph defined by G^* and $\mathbf{d}$ if at least one of the following conditions holds:*

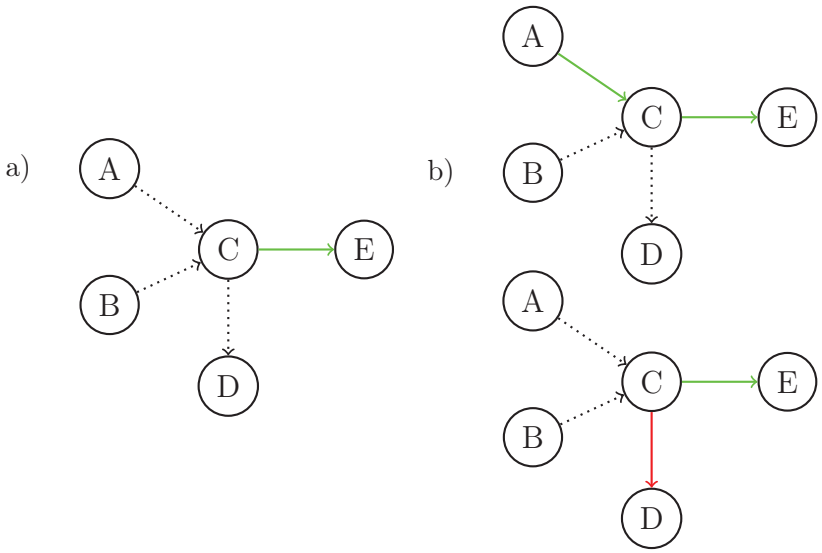


Figure 2.2 – a) Node C is not balanced ($(\{C\},\{E\})$ is green). b) Two possible ways to make C balanced (above: $(\{A\},\{C\})$ is coloured green; below: $(\{A\},\{C\})$ is coloured red). The dotted lines represent grey hyperarcs.

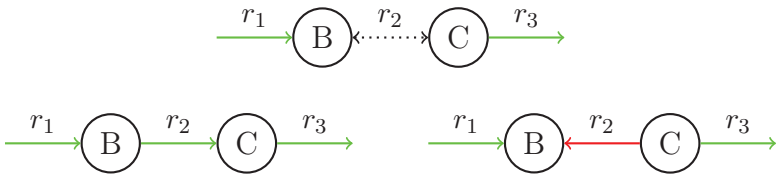


Figure 2.3 – Both B and C can be balanced by colouring r_2 either green or red, depending on the direction the reaction is taken to operate in (r_1 and r_3 are both green).

1) there exists two hyperarcs r_i and r_j such that $v \in Prod(r_i)$ and $v \in Subs(r_j)$, and $d_i = d_j = +$
2) there exists two hyperarcs r_i and r_j such that $v \in Subs(r_i)$ and $v \in Subs(r_j)$, and $d_i = +$ and $d_j = -$
3) there exists two hyperarcs r_i and r_j such that $v \in Prod(r_i)$ and $v \in Prod(r_j)$, and $d_i = +$ and $d_j = -$
4) there exists two hyperarcs r_i and r_j such that $v \in Prod(r_i)$ and $v \in Subs(r_j)$, and $d_i = d_j = -$.

Similarly, we can update Definition 2:

Definition 4. *Let G be a hypergraph, G^* its transformation, and $\mathbf{d}_0$ the a priori selection of G^*. A node $v \in V$ is **consistent** if at least one of the following conditions holds:*
1) there are no hyperarcs $r_i \in R$ such that $d_i \neq 0$ and $v \in Subs(r)$ or $v \in Prod(r)$
2) v is a source or a sink of G
3) v is balanced.
A selection $\mathbf{d}$ of G^* is consistent if every node $v \in V$ is consistent under $\mathbf{d}$ and for all $r_i \in R$ it holds that if $d_{i0} = +$, $d_i \neq -$, and if $d_{i0} = -$, $d_i \neq +$.

The equivalent of the colouring problem 2.10 for the transformation is

$$\text{find} \quad \arg\max_{\mathbf{d}\in\mathcal{D}_G} \sum_{i \text{ s.t. } d_i\in\{+,-\}} w_i, \tag{2.11}$$

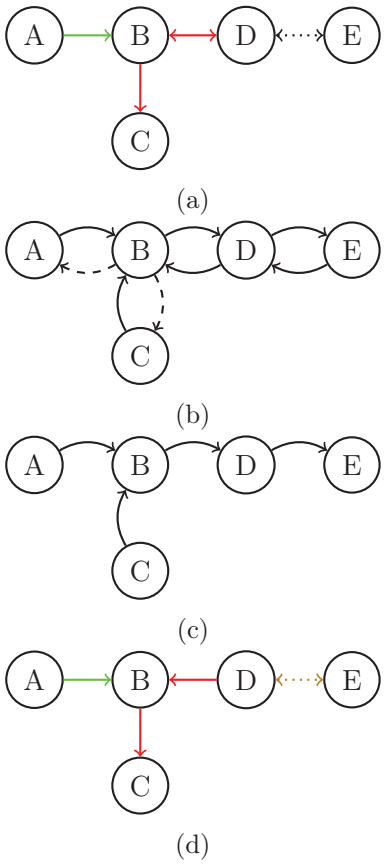


Figure 2.4 – a) The *a priori* colouring of the hypergraph: the hyperarc $(\{A\},\{B\})$ is coloured green, $(\{B\},\{C\})$ and $(\{B\},\{D\})$ are coloured red. Dotted lines indicate a grey colour. b) The transformation into hyperarc directions. Dashed hyperarcs indicate choices that are forbidden based on the *a priori* colouring and reversibility: $(\{A\},\{B\})$ can only be chosen in the actual direction because it is green and irreversible. $(\{B\},\{C\})$ can only be chosen in reverse because it is red, whereas $(\{B\},\{D\})$ can still be chosen in both directions since it is reversible. c) A solution given in terms of hyperarc directions. d) The "interpretation" back into colours: the hyperarc $(\{A\},\{B\})$ is coloured green, $(\{B\},\{C\})$ and $(\{B\},\{D\})$ are coloured red. The hyperarc $(\{D\},\{E\})$ is coloured *yellow* because its direction is not known.

where $\mathcal{D}_G$ is the set of all consistent selections for the hypergraph G. We will call this the *selection problem*.

Theorem 1. *The selection problem is NP-hard.*

Proof. The hardness can be proved by reducing the set cover problem, known to be NP-hard [88], to the selection problem in polynomial time. In the set cover problem the goal is, given a universe of elements and a collection of subsets of the universe, to find a minimum number of subsets in the collection such that their union equals the universe. An illustrative example is shown in Figure 2.5.
Let $U = \{1,\dots,m\}$ be a set of elements, and $S = \{S_1,\dots,S_n\}$ a collection of sets such that $S_i \subseteq U$ for all $i = 1,\dots,n$ and $\cup_{i=1}^{n} S_i = U$. We construct the following hypergraph G: first, a node v_i is created for every element of U and a node v_{S_i} for every set in S. The nodes corresponding to the sets in S are sources, and the nodes corresponding to U are not. Additionally, we create a node T that is a sink. We create a hyperarc $r_T = (\{v_1,\dots,v_m\},\{T\})$

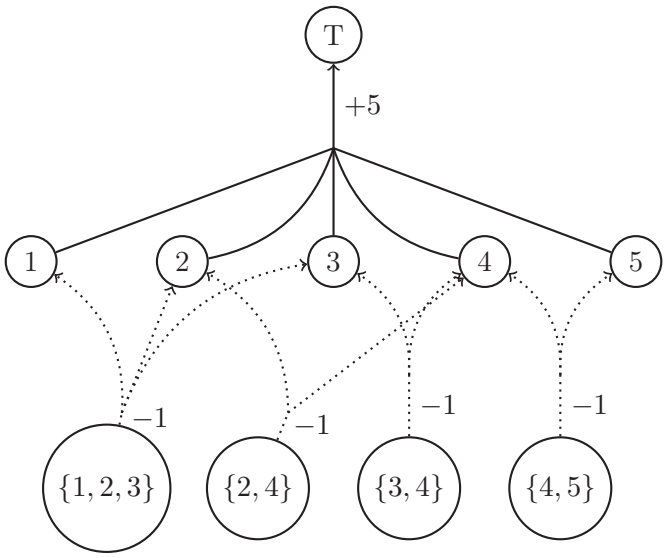


Figure 2.5 – An example of the reduction in the proof. The universe $U = \{1, 2, 3, 4, 5\}$, and the collection $S = \{\{1, 2, 3\}, \{2, 4\}, \{3, 4\}, \{4, 5\}\}$. Finding the minimal number of sets in S to cover U is achieved by finding the highest scoring hypergraph.

with the weight $w_T = n + 1$. We then create a hyperarc r_{S_i} for every node v_{S_i} by connecting it to those nodes v_j whose corresponding element is contained in the set S_i. In other words, for every v_{S_i}, there is a hyperarc $r_{S_i} = (\{v_{S_i}\}, \{v_j : j \in S_i\})$. All of these hyperarcs have a weight $w_{S_i} = -1$.
Let now G^* be the transformation of G and the *a priori* selection $\mathbf{d}_0$ of G^* such that $d_{T0} = +$ and $d_{S_i 0} = 0$ for all $S_i = S$. We observe first that if the hyperarc r_T is selected, that is, $d_T = +$, for every node v_i at least one of the incoming hyperarcs needs to be selected in the default direction to make the node consistent. We further note that r_T should always be selected: if all of the hyperarcs are selected, that is, $d_{S_i} = +$ for all $S_i = S$, all the nodes v_i are made consistent, and thus the hypergraph is consistent. Furthermore, its score will be strictly positive (that is, equal to 1), and so it will be preferred over the empty solution.
Finding the optimal selection $\mathbf{d}$ thus amounts to finding the minimum number of hyperarcs r_{S_i} to select so that the resulting hypergraph is consistent. If we have such a selection, the hyperarcs r_{S_i} that have been selected correspond to the minimum number of sets S_i needed to cover U. □

Answer set programming implementation

Following [85], the first implementation of the MOOMIN-algorithm was done using *Answer Set Programming* (ASP, [30]). ASP is a declarative programming paradigm based on disjunctive logic programming. Oriented towards NP-hard problems, ASP allows to declare problems such as the one formulated in the MOOMIN-algorithm in a fairly concise manner.
The initial input of the pipeline is a metabolic network and the results of a DE analysis. The network is contained in an SBML-file, a representation format commonly used in computational biology [80]. The DE results are stored in a tab-delimited file that contains for every gene the logarithm of the fold change and the associated posterior probability of differential expression (PPDE). An additional file contains a list of inactive genes, obtained in the DE pipeline by filtering out genes that were detected in counts below a certain threshold across

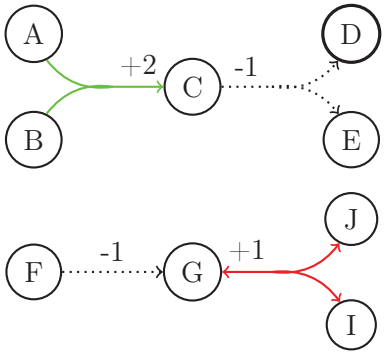


Figure 2.6 – An example hypergraph: the hyperarc $(\{A, B\}, \{C\})$ is coloured green, $(\{G\}, \{I, J\})$ is coloured red. Dotted lines denote grey hyperarcs.

all samples.
The ASP part of the pipeline was implemented using the syntax and solver of the CLINGO suite developed by Postasco [68, 67]. CLINGO combines a grounder tool GRINGO that transforms a user-defined logic programme containing variables into a variable-free format with a solver, CLASP, that computes *answer sets*. This allows for the user to define the problem as fairly human-readable text files that serve as the input for CLINGO. In our case, the answer sets are consistent colourings of the network.
The data input for clingo is produced using a python script. This comprises transforming the network into a hypergraph coded using the CLINGO syntax along with the *a priori* colouring and weights. An example is given in Algorithm 1, where the hypergraph in Figure 2.6 is written using the CLINGO syntax. Lines 1-9 define the nodes of the hypergraph, with node D being specified as an *endnode*, meaning it is either a source or a sink and does not need to be balanced. Lines 10-13 define the hyperarcs, and their *a priori* colours and weights. An additional attribute tells that the reaction represented by the hyperarc $(\{G\}, \{J, I\})$ is reversible.
The rest of the input consists of the problem formulation which GRINGO uses to form the grounded programme. The first part is shown in Algorithm 2. Line 1 defines what can be seen as an auxiliary variable that tells if an hyperarc is selected in a solution in either direction. Line 2 tells that each hyperarc can be chosen either in the original direction or in reverse, but not both. Lines 3 and 4 define the reverse hyperarcs. Lines 5 and 6 impose the consistency with the *a priori* colouring/selection. Lines 7-10 contain the balance constraints in Definition 3. Line 11 tells the solver that we are looking for solutions in terms of hyperarc directions.
Line 12 adds an additional instruction for the enumeration procedure: it tells the solver to "project" the answers, given in terms of hyperarc directions, to hyperarc selections. In other words, when enumerating all optimal solutions, only those that differ in terms of the hyperarcs included in the solution are counted as distinct. The reason for doing this is the observation that sometimes two *a priori* grey hyperarcs can act interchangeably, meaning one can be coloured red and the other green or *vice versa*, with both options producing valid solutions. This can lead to a combinatorial explosion, producing a massive amount of solutions that in actuality differ in trivial ways. For this reason, we opted for the projection, sacrificing in our opinion very little in terms of the information garnered, but gaining considerably in terms of practicality.
Finally, line 13 tells the solver the object of maximisation, in our case the sum of the weights of the selected hyperarcs. In practice, it is often given as a separate file to CLINGO to clarify the input structure.
Based on the fact that ASP as well as the CLINGO solver are specifically targeted at NP-hard

Algorithm 1 The hypergraph in Figure 2.6 expressed in the CLINGO syntax.

```
node(A).
node(B).
node(C).
node(D). endnode(D).
node(E).
node(F).
node(G).
node(J).
node(I).
hyperarc(1). in(A,1). in(B,1). out(1,C). green(1). weight(1,2).
hyperarc(2). in(C,2). out(D,2). out(E,2). grey(2). weight(2,-1).
hyperarc(3). in(F,3). out(G,3). grey(3). weight(3,-1).
hyperarc(4). in(G,4). out(J,4). out(I,4). red(4). weight(4,1). reversible(4).
```

Algorithm 2 The MOOMIN-problem expressed as a logic programme.

```
0<=inanswer(H)<=1:- hyperarc(H).
1<=sel(H);rsel(H)<=1:- inanswer(H).
rin(V,H):-out(H,V).
rout(H,V):-in(V,H).
:- green(H), rsel(H), not reversible(H).
:- red(H), sel(H), not reversible(H).
:- sel(H), in(V,H), not endnode(V), out(H2,V): sel(H2)=0, in(V,H3): rsel(H3)=0.
:- sel(H), out(H,V), not endnode(V), in(V, H2): sel(H2)=0, out(H3,V): rsel(H3)=0.
:- rsel(H), in(V,H), not endnode(V), in(V, H2): sel(H2)=0, out(H3,V): rsel(H3)=0.
:- rsel(H), out(H,V), not endnode(V), out(H2,V): sel(H2)=0, in(V,H3): rsel(H3)=0.
#show sel/1. #show rsel/1.
#project inanswer/1.
#maximizeW,H: hyperarc(H), inanswer(H), weight(H,W).
```

problems, and their successful application in [85] and [86], we expected reasonable running times for the genome-scale metabolic networks of prokaryotes, containing metabolites and reactions in the lower thousands. However, it turned out that the ASP-formulation was not being solved in running times that would enable the algorithm to be used in practice. For example, for a problem instance based on an *E. coli* reconstruction involving 1668 nodes and 2036 hyperarcs, of which 197 were coloured, CLINGO was not able to reach an optimal solution in 1 hour. In contrast, the linear programming formulation presented in the next section was solved in 14 seconds in MATLAB using the CPLEX LP-solver. Both tests were performed on a personal computer with one 2,9 GHz Intel Core i5 processor with two cores. CLINGO was instructed to run two parallel threads.

Linear programming implementation

The MILP-based implementation of MOOMIN was done in MATLAB using the COBRA TOOLBOX [77] for the manipulation of the SBML-files and the CPLEX software suit as the MILP-solver. The COBRA TOOLBOX offers a comprehensive set of utilities for constraint-based modelling. It also offers an interface for solving LP- and MILP-problems that allows the user to choose between several different solvers. This means that our implementation can also be used without the proprietary CPLEX solver.
The inputs and general formulation remain as described in the previous sections. The main function of MOOMIN takes in a COBRA TOOLBOX data structure that contains the metabolic network. Such structure can be created from an SBML-file using functions provided in the Toolbox. The gene expression data is contained in a standard MATLAB data structure and can be read using default MATLAB functions from, for example, a tab-delimited text file.
The MILP-problem can be defined as follows. Let ${x^+}_i$ and ${x^-}_i$ be binary variables that correspond to the selections d_i. Let y_j be a binary variable that represents if node j is included in the solution or not. In other words, y_j tracks if any hyperarc attached to j is chosen. For brevity, let I_i and O_i stand respectively for $\text{Subs}(r_i)$ and $\text{Prod}(r_i)$. The degree of a node j, $|\{i : j \in I_i \cup O_i\}|$, is denoted by d_j. Finally, T is the set of endnodes.
The MILP is defined as

$$\max \quad \sum_{i=1}^{m} ({x^+}_i + {x^-}_i) w_i \tag{2.12}$$

$$\text{s.t.} \quad {x^+}_i + {x^-}_i \leq 1, \forall i \in \{1, \ldots, m\} \tag{2.13}$$

$$\sum_{i|j \in I_i \cup O_i} {x^+}_i + {x^-}_i \leq d_j y_j, \forall j \in \{1, \ldots, n\} \setminus T \tag{2.14}$$

$$\sum_{i|j \in I_i} {x^+}_i + \sum_{i|j \in O_i} {x^-}_i \geq y_j, \forall j \in \{1, \ldots, n\} \setminus T \tag{2.15}$$

$$\sum_{i|j \in O_i} {x^+}_i + \sum_{i|j \in I_i} {x^-}_i \geq y_j, \forall j \in \{1, \ldots, n\} \setminus T \tag{2.16}$$

$${x^-}_i \leq 0, \forall i | d_{0i} = + \tag{2.17}$$

$${x^+}_i \leq 0, \forall i | d_{0i} = - \tag{2.18}$$

$${x^+}_i, {x^-}_i, y_j \in \{0, 1\}, \forall i, j \tag{2.19}$$

Constraint (2.13) guarantees that no hyperarc is chosen twice. Constraint (2.14) requires $y_j = 1$ if j is not an endnode and at least one of the hyperarcs connected to it is chosen. The balance constraints are given in (2.15) and (2.16). Finally, (2.17) and (2.18) enforce the consistency with the *a priori* selection.

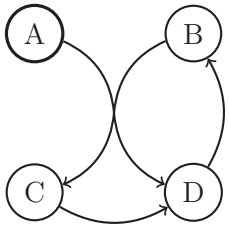


Figure 2.7 – An example of a stoichiometrically infeasible solution. Assume that all the stoichiometric coefficients are equal to one, all reactions are irreversible, and that node A is an endnode. If the figure presents a solution of the transformed problem, it is consistent according to the topological Definition 4. However, it is easy to show that no non-zero flux vector will satisfy the steady state condition.

The above problem can be solved in seconds using the CPLEX solver on a personal computer for a network with 1000+ nodes and 1500+ hyperarcs. The exact reason for the difference in performance between ASP and MILP is unclear. One possibility is that the kind of highly continuous weights used in the MOOMIN-algorithm are not well-suited for the CLINGO solver. While in CLINGO, enumeration of optimal solutions is a built-in feature, in the MILP-formulation this had to be done *ad hoc*. Fortunately, there is a straightforward way to do this. Namely, let $\mathbf{x}_i \in \mathbb{R}^m$ be the ith solution obtained, where $x_{ij} = x_i^+{}_j \wedge x_i^-{}_j$, that is, it is an indicator vector for if a hyperarc is selected in the solution in either direction. We thus follow Section 2.3.1 in only enumerating solutions that differ in terms of the hyperarcs that are selected. To obtain $\mathbf{x}_{i+1}$, we add the following constraint to the problem:

$$2(\mathbf{x}_i \cdot \mathbf{x}_{i+1}) \leq \sum_j^m x_{ij} + \sum_j^m x_{i+1j} - 1 \tag{2.20}$$

and solve it again. Additionally, we obviously need to require that the same value of the objective function is attained.
This procedure in theory guarantees that all optimal solutions are found. However, it needs to be continued until the problem is found to be infeasible which can be difficult for the solver. The switch to MILP allows for a straightforward extension of the algorithm to include reaction stoichiometry. In other words, we can switch from the hypergraph to the stoichiometric matrix representation of the network and formulate the balance constraints to fully account for the conservation of a feasible steady state flux distribution.

2.3.2 Stoichiometric formulation

The balance constraints in Definition 1 guarantee that a solution corresponds to a feasible change from one steady state to another, but only for linear pathways. Namely, in some situations, they allow for stoichiometrically unbalanced cycles (an example is shown in Figure 2.7). To avoid this, stoichiometric constraints can be enforced to guarantee that a solution truly corresponds to a feasible (qualitative) difference in fluxes. To achieve this, we make use of the fact that a difference between two feasible flux vectors is also a feasible flux vector.
To recapitulate, we are assuming that there are two (unknown) flux vectors $\mathbf{f}_1$ and $\mathbf{f}_2$ that both satisfy the steady state condition:

$$S \cdot \mathbf{f}_1 = 0 \quad \text{and} \quad S \cdot \mathbf{f}_2 = 0 \tag{2.21}$$

and that correspond to the flux distributions in the two conditions under study. A trivial calculation shows that any change $\Delta\mathbf{f} = \mathbf{f}_2 - \mathbf{f}_1$ will also satisfy the steady state condition,

that is,

$$S \cdot \Delta\mathbf{f} = 0 \tag{2.22}$$

for any feasible change. A colouring of the hypergraph qualitatively defines a difference vector $\Delta\mathbf{f}$: a green hyperarc implies that $\Delta f_i > 0$, red implies that $\Delta f_i = 0$, and grey that $\Delta f_i = 0$. The consistency of a colouring can thus be ensured *stoichiometrically* by simply requiring that *some* vector $\Delta\mathbf{f}$ complying to the colouring satisfies the steady state condition 2.22. In Figure 2.7 *no* flux vector consistent with the selection displayed will satisfy the condition 2.22.
Observe that we do not wish to infer the magnitude of change, and thus only the qualitative nature of the entries in $\Delta\mathbf{f}$ is of interest. In other words, we will only require a flux distribution that satisfies the steady condition and the actual flux values do not matter (see the implementation details below). In addition, because the flux vector represents a *change*, positive and negative flux values do not have their usual significance. More precisely, where in for example FBA, only reversible reactions can carry a negative flux, meaning that such a reaction is operating in the reverse direction, a negative flux in our situation simply means that a decrease in flux was inferred. In fact, reaction specific flux bounds constraining fluxes to be non-negative or non-positive reflect the *a priori* colours, rather than simply reversibility.
The stoichiometric formulation is also done in terms of the transformation to selection of hyperarc directions. The *a priori* selection $\mathbf{d}_0$ is enforced by setting the flux bounds so that $lb_i = 0, ub_i = C$ whenever $d_{i0} = +$, $lb_i = -C, ub_i = 0$ when $d_{i0} = -$, and $-lb_i = ub_i = C$ when $d_{i0} = 0$. The constant C is just placed to bound the flux cone, and its precise value has no significance. We simply require that $C \gg \varepsilon$, where $\varepsilon > 0$ is another constant that is used to determine what is considered a non-zero flux.
Binary variables ${x^+}_i$ and ${x^-}_i$ are as before indicators of whether a hyperarc is in the solution or not, and in what direction. The difference is that they are now tied to a flux vector $\mathbf{f} \in \mathbb{R}^m$, where $f_i > 0$ implies an increase in flux, and $f_i < 0$ a decrease (we omit the Δ to simplify notation). Thus we will require ${x^+}_i \iff f_i > 0$ and vice versa.
The MILP is defined as

$$\max \quad \sum_{i=1}^{m} ({x^+}_i + {x^-}_i) w_i \tag{2.23}$$

$$\text{s.t.} \quad {x^+}_i + {x^-}_i \leq 1, \forall i \in \{1, \ldots, m\} \tag{2.24}$$

$$S \cdot \mathbf{f} = 0 \tag{2.25}$$

$$f_i + x_i^+(lb_i - \varepsilon) \geq lb_i, \forall i \in \{1, \ldots, m\} \tag{2.26}$$

$$f_i - x_i^+ ub_i \leq 0, \forall i \in \{1, \ldots, m\} \tag{2.27}$$

$$f_i + x_i^-(ub_i + \varepsilon) \leq ub_i, \forall i \in \{1, \ldots, m\} \tag{2.28}$$

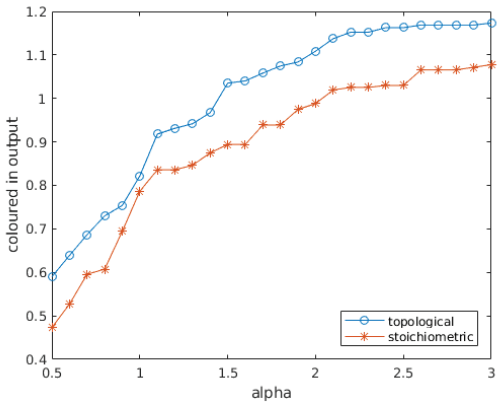
$$f_i + x_i^-(-lb_i) \geq 0, \forall i \in \{1, \ldots, m\} \tag{2.29}$$

$$lb_i \leq f_i \leq ub_i, \forall i \in \{1, \ldots, m\} \tag{2.30}$$

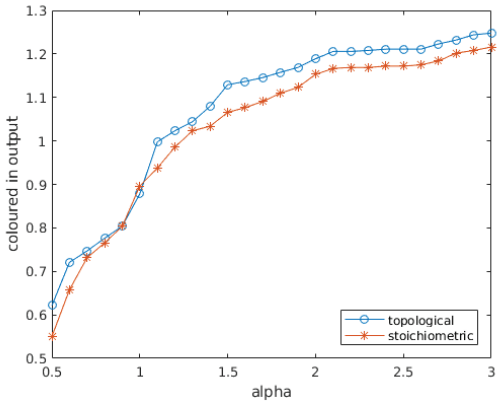
$${x^+}_i, {x^-}_i \in \{0, 1\} \tag{2.31}$$

$$\mathbf{f} \in \mathbb{R}^m \tag{2.32}$$

Constraint (2.24) guarantees that no hyperarc is selected twice. Constraint (2.25) is the steady state condition. Constraints (2.26) and (2.28) guarantee that $x_i^+ = 1 \Rightarrow f_i > 0$ and vice versa, and constraints (2.27) and (2.29) the opposite, $x_i^+ = 0 \Rightarrow f_i \leq 0$. The upper and lower bounds in (2.30) keep the flux cone bounded, and enforce the *a priori* hyperarc selections.
The enumeration of optimal solutions can be done as was described in the previous section.
Based on tests done with real data sets, the stoichiometric problem appears to be harder to solve in practice. Another difference is observed in the number of hyperarcs in a solution:



(a)



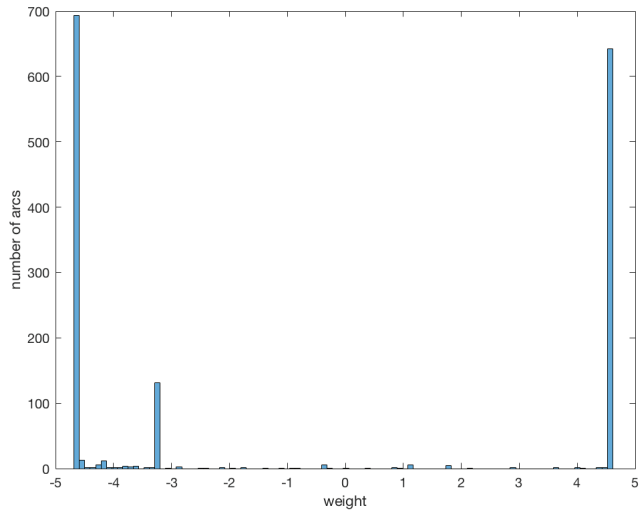
(b)

Figure 2.8 – The plots show the number of coloured hyperarcs in the solution in relation to the number of coloured hyperarcs in the input. a) The *S. cerevisiae* data. b) The *E. coli* data. For both networks, at most the ten first solutions were calculated, and the number shown is the average.

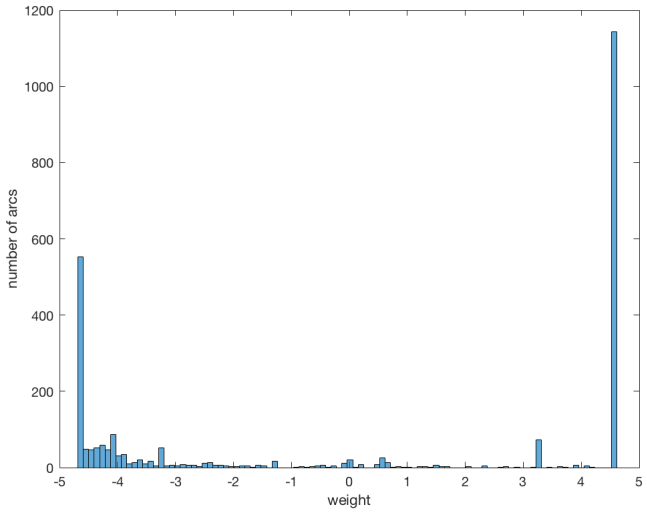
solutions of the topological problem tend to contain more hyperarcs. This might be explained by the fact that the topological constraints are less restrictive, and thus it is "easier" to include hyperarcs with positive weights in the solution (see next section for examples).

2.3.3 Results

The best way to understand the functioning of MOOMIN is through examples. We chose two data sets from the literature for this purpose. The first one, comparing growth of *S. cerevisiae* in a chemostat *versus* a batch culture, is meant as a proof of concept: the resulting shifts in metabolism should be well reflected in the expression of genes and the underlying biology is fairly well understood, and so we expect MOOMIN to easily identify the correct changes. The other set, measuring exposure of *E. coli* to mercury, is meant as a more ambitious test of the method. Our goal here was to see how MOOMIN compares to traditional DE analysis done by the original authors: if it would be able to recover the same results, and to go beyond them, possibly explaining inconsistent findings or inferring changes that were not detected by the gene expression measurements alone. In this section, I will present general observations



(a)



(b)

Figure 2.9 – A histogram of the hyperarc weights in the two networks. a) *S. cerevisiae*. b) *E. coli*.

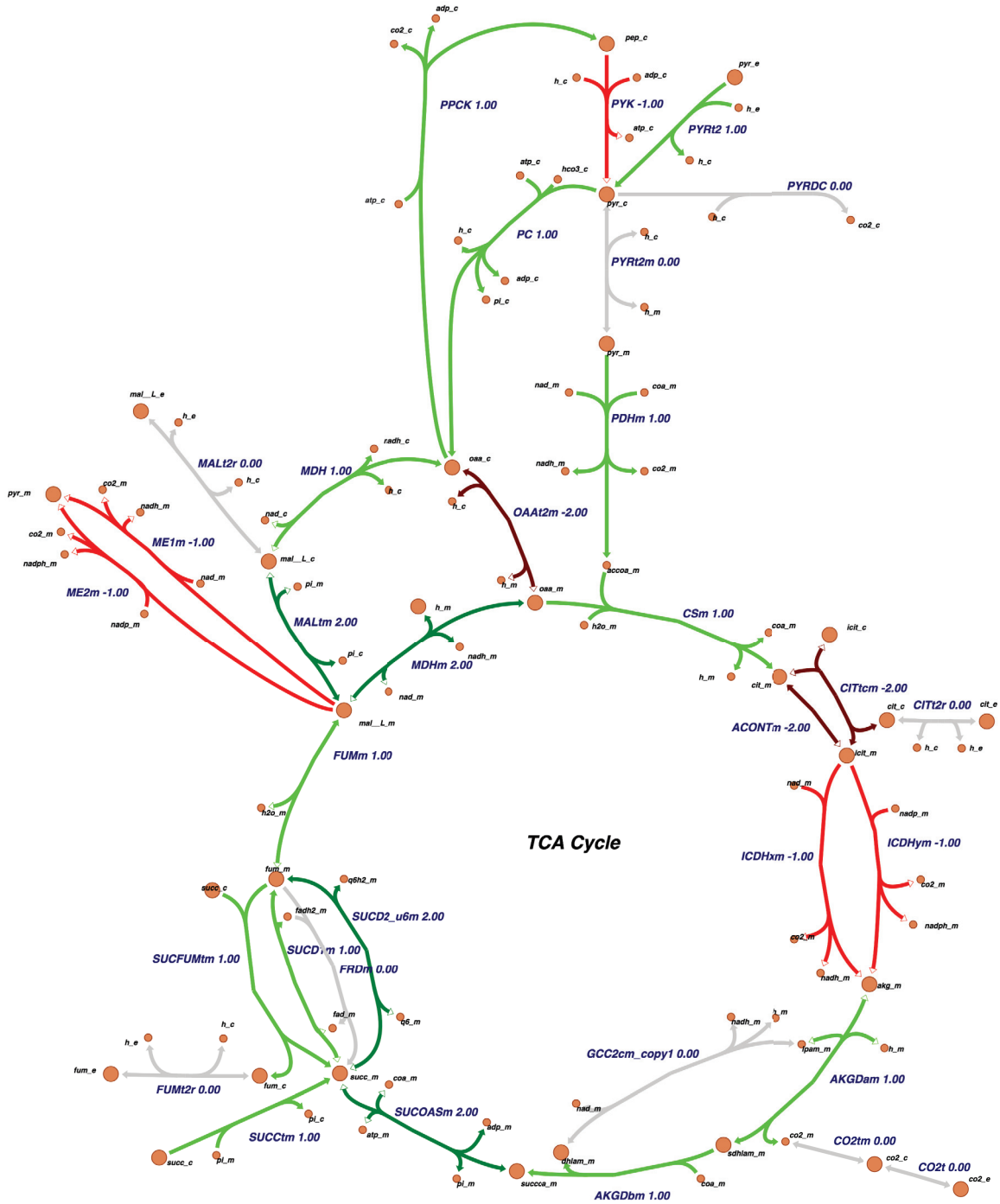


Figure 2.10 – A MOOMIN solution for the *S. cerevisiae* data visualised using ESCHER [97]. The output of the method can be written into a JSON file that can be read and displayed by the ESCHER browser application. In the core metabolism, we can clearly see the ramping up of the TCA cycle and aerobic respiration. Reactions that were inferred to operate in the reverse direction are shown in darker colours. The colours are coded in the JSON file using numbers, these are displayed alongside the reaction names (2: green, reverse; 1: green; 0: grey; -1: red; -2: red, reverse).

about the solutions obtained using MOOMIN as well as biological results.
The *S. cerevisiae* data set was originally published by Nookaew *et al.* [133] and the metabolic network used was the iMM904 published by Mo *et al.* [123]. The *E. coli* data set was originally published by Lavoie and Summers [106] and the network used was iJO1366, published by Orth *et al.* [136]. For both data sets, we performed DE analysis starting from raw counts with the R package EBSEQ [109]. Both networks were downloaded from BIGG. The *S. cerevisiae* network was used as is, two reactions in the *E. coli* network, `R_GLUABUTt7pp` and `R_GTHOr`, were modified from irreversible to reversible based on literature data.

S. cerevisiae growth in batch *versus* chemostat

The *S. cerevisiae* data set contained gene expression measurements obtained using RNA-Seq from two conditions: growth in a batch culture and growth in a chemostat culture. In a batch culture, a small number of cells are inoculated into a nutrient-rich medium, and allowed to grow at maximal rate until the nutrients are exhausted and the growth curve achieves saturation. In contrast, in a chemostat culture, nutrients are added to the culture at a fixed flow rate while the biomass and the products of metabolism are removed from the vessel at the same flow rate to maintain a fixed culture volume. Growth will settle to a steady state value and cells will continue to divide indefinitely. There were three biological replicates for both conditions.
The *S. cerevisiae* network contains 1226 metabolites and 1577 reactions. Gene association information is given for 1043 reactions and in total there are 905 metabolic genes listed in the model. DE analysis of the *S. cerevisiae* data led to a PPDE and a fold change for 6506 genes. These contained 878 of the metabolic genes in the network. We considered a gene inactive if it measured less than 10 counts per million across all samples. There were 620 such genes in the data, 16 of which were metabolic genes in the model. We assessed the feasibility of all reactions associated with these genes, and consequently removed 5 reactions deemed inactive in both conditions. The input hypergraph thus comprised 1225 nodes and 1572 hyperarcs.
Using the threshold of PPDE (the parameter t) 0.9, 1401 genes were up-regulated and 1505 genes down-regulated. The same numbers for the metabolic genes in the network were 284 and 325. Normally, a stricter threshold for differential expression would be used. However, since the weight function is continuous, we use a slightly more lenient value. The pre-processing steps described in Section 2.3.1 resulted in colouring 304 hyperarcs green and 369 hyperarcs red *a priori*. The assigned weights can be seen in the histogram in Figure 2.9 a).
The parameter α controls essentially how lenient or conservative the method is in inferring changes. This is illustrated in Figure 2.8 where the number of coloured arcs appearing in optimal solutions is plotted for different values of α. I used the value $\alpha = 1$ to investigate the general nature of the solutions. For biological results, different values $(1 - 3)$ were tried. The main findings stayed the same.
The topological MILP problem (Equations 2.12 - 2.19) was solved in 18 seconds (including the pre-processing steps) on a 2,9 GHz Intel Core i5 processor with two cores. Of the 673 *a priori* coloured hyperarcs, 430 were also coloured in the solution, and 140 *a priori* grey hyperarcs became coloured.
Using the enumeration procedure described in Section 2.3.1, I enumerated the first 1000 alternative optima. They appear to be quite similar to each other. In total, 620 hyperarcs were coloured in at least one solution and 541 in all of them. On average, a solution contained 570 coloured hyperarcs, and the range was 560-582.
The stoichiometric MILP problem (Equations 2.23 - 2.32) was solved in 446 seconds using the same machine. Remarkably, there was only one optimal solution. It coloured less hyperarcs than the topological formulation (this appears to be the general trend, see Figure 2.8): 370

of the *a priori* coloured reactions were coloured and 164 of the *a priori* grey ones.
The solutions of the topological and the stoichiometric formulations appear to overlap but not completely. Comparing the 541 hyperarcs appearing in all of the topological solutions with the stoichiometric solution, 418 hyperarcs appear in both. This makes sense, since the topological formulation approximates the stoichiometric constraints.
In yeast growth in an aerobic batch culture, the glucose concentration is so high at first that ethanol fermentation takes place even though the cells have enough oxygen to perform only respiration. The ethanol secreted in this first phase is used as a carbon source when glucose is depleted (the so-called diauxic shift). However, if the glucose concentration is kept constant and low (by means of a fedbatch or a chemostat culture), glucose flux into fermentation is extremely low and the yeast growth yield is slightly improved since the respiration is far more efficient in supplying energy and building blocks for biomass production than fermentation.
As a result, as already reported by several studies, up-regulated pathways in a chemostat culture (if we consider batch growth as the control) are usually involved with electron transport, aerobic respiration, and the TCA cycle. In this context, among the results reported by Nookaew *et al.* [133], the authors described an enrichment in Gene Ontology terms related to growth, respiration, the TCA cycle, and fatty acid beta-oxidation.
MOOMIN was able to infer changes within the main pathways associated with growth in a chemostat culture, namely the TCA cycle (11 reactions out of 13), and aerobic respiration (12 reactions out of 16). Figure 2.10 shows a visualisation of this. Increased activity for the TCA cycle is clearly visible (see also Figure S1). We also detected a general down-regulation of nucleotide biosynthesis (Figure S2) and amino-acid metabolism (not shown). Both observations are in accordance with the lower growth rate in chemostat resulting in less need for duplication of DNA, transcription of RNA, and production of proteins, as discussed by Nookaew *et al.*

E. coli exposure to mercury

The *E. coli* data set investigated the bacterium's response to mercury exposure. Gene expression was measured using RNA-Seq in two conditions: an unexposed control and a culture exposed to mercuric chloride ($HgCl_2$) with three biological replicates for both conditions. Mercury is a toxicant that negatively impacts the health of both microscopic and macroscopic organisms, and induces a broad cellular response in *E. coli*.
The *E. coli* network contains 1805 metabolites and 2583 reactions, with 2123 of the reactions having gene association information. There are 1367 metabolic genes listed in the model. PPDE and fold change were obtained for 4326 genes, amongst which 1359 of the metabolic genes. There were 168 inactive genes, including 7 metabolic genes. Based on this, we removed 5 reactions as inactive. Thus the final input hypergraph contained 1805 nodes and 2578 hyperarcs.
In total, 695 genes were up-regulated and 1237 genes down-regulated ($t = 0.9$). For the metabolic genes, the same numbers were 173 and 520 respectively. Based on this, 306 hyperarcs were coloured green and 1026 red *a priori*. A histogram of the assigned weights can be seen in Figure 2.9 b).
The resulting topological MILP problem was solved in 16 seconds. Of the 1332 *a priori* coloured hyperarcs, 894 remained coloured in the solution. On the other hand, 309 *a priori* grey hyperarcs became coloured.
There were again more than 1000 alternative optima. In the first 1000 solutions, 1139 hyperarcs appear in every solution and 1328 in at least one. On average, a solution contained 1163 coloured hyperarcs, ranging from 1151 to 1289.
The stoichiometric MILP problem was solved in 72 seconds. Again, it coloured less hyperarcs

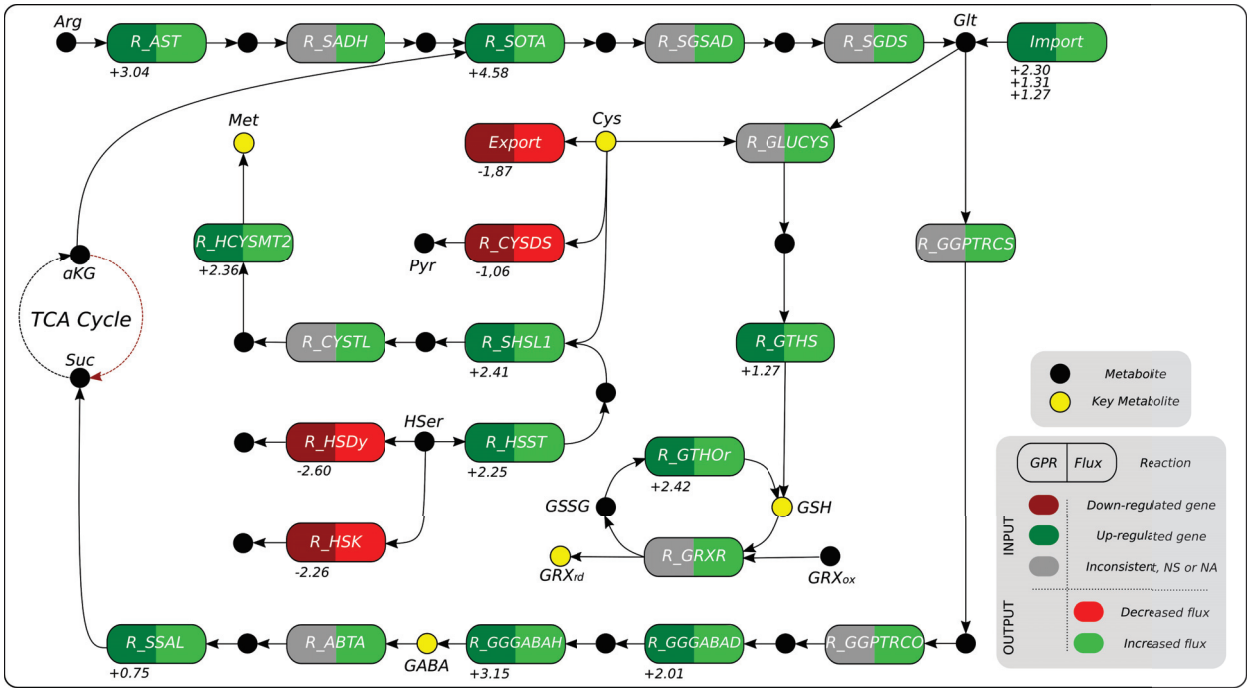


Figure 2.11 – Changes in the production of key metabolites related to the mercury stress response of *E. coli*. Rectangles represent reactions and circles represent metabolites. The reaction rectangles show the *a priori* colour on the left, and the inferred colours, that is, the output of MOOMIN, on the right. Below a reactions is displayed the logarithm of the fold change for the associated gene with the highest PPDE. In the case of import of *Glt*, more than one reaction was considered, therefore, there are three different fold changes. Key metabolites related to the mercury stress response are highlighted in yellow. I thank Mariana Ferrarini for the figure.

than the topological formulation, 1189, however investigating the alternative optima this difference disappeared (see also Figure 2.8 b)). Of the *a priori* coloured hyperarcs, 469 ended up grey in the output. In contrast, 326 *a priori* grey hyperarcs were coloured.
For the *E. coli* data, there were also a multitude of alternative optima for the stoichiometric problem. I again enumerated the first 1000. Here, 1120 hyperarcs were coloured in all of these solutions, and 1240 in at least one. On average, there were 1193 coloured hyperarcs, and the range was 1179-1209.
Comparing the topological and the stoichiometric solutions, 1044 hyperarcs are present in all the first 1000 solutions of both formulations. Thus there is again considerable agreement between the two different formulations. However, it is also clear that they are not equal, and the addition of stoichiometry does alter the space of accepted solutions.
MOOMIN was able to detect changes in all pathways described by Lavoie and Summers [106] as up or down-regulated. As expected, reactions from aminoacid and nucleotide biosynthesis, NADH metabolism, carbohydrate metabolism and glycolysis, whose corresponding genes were downregulated, appeared with predicted reductions in flux in both topological and stoichiometric solutions (not shown). Further in agreement with the expected findings were inferred increases in reactions related to stress response and redox response (Figure 2.11).
Reduced glutathione (GSH), a tripeptide composed of cysteine, glutamate, and glycine, is considered to be one of the most important scavengers of reactive oxygen species (ROS), and is involved in the detoxification of certain xenobiotics and heavy metals. Its ratio with oxidised glutathione (GSSG) can be used as a marker of oxidative stress: in a resting cell, the molar GSH:GSSG ratio exceeds 100:1. However, it is altered under certain stress conditions when GSH is oxidised to GSSG [195]. Following this, Lavoie and Summers expected to find an increase in cysteine and glutathione biosynthesis. However, most genes for biosynthesis of these biothiols appeared to be either down-regulated or showed no significant change.
MOOMIN, on the other hand, was able to detect these features, as showed in Figure 2.11. Even though there was no apparent increase in the production of cysteine (*Cys*), it is possible that more cysteine was available due to a decrease in its turnover to pyruvate (*Pyr*, `R_CYSDS`) and its export to the periplasm. We also predicted an increased production of glutamate (*Glt*) from arginine (*Arg*), which is essential for GSH biosynthesis.
Glutaredoxin (*GRX*) is also a key metabolite for redox homeostasis and stress response. Similarly to GSH, active GRX (*GRXrd*) is oxidized by substrates and it is only recycled to its active form with the concomitant oxidation of GSH (`R_GRXR`). This is why the increased flux towards the production of glutaredoxin (*GRXrd*) is closely linked to the recycling of GSSG into GSH (`R_GTHor`). MOOMIN also connected the down-regulation of genes responsible for the conversion of 2-oxoglutarate (αKG) to succinate (*Suc*) from the TCA cycle (red arrow in Figure 2.11) with the GABA shunt, something not described in [106]. GABA (γ-aminobutyrate) is a metabolite that counteracts a wide range of stresses in several species, and seems to be essential for the acid tolerance in *E. coli* [78, 63]. We were also able to detect a possible increase in the fluxes from cysteine and homoserine (HSer) to another amino acid that has been linked to ROS scavenging in other organisms, namely methionine (*Met*) [33].

2.4 Conclusion

In the first part of this chapter, I presented the state of the art on the integration of 'omics data to metabolic networks. This question has attracted considerable research effort. Most of it has been aimed towards improving phenotype prediction through FBA and similar constraint-based methods by using transcriptomics to further constrain the space of metabolic fluxes, either by using gene expression levels as a proxy for flux magnitude, or by building so-called

context-specific metabolic models. It appears though that this question remains largely unsolved, and no method has provided consistently accurate results.
To a lesser extent, 'omics data and metabolic networks have also been combined to better make use of the former. Comparative analyses of metabolite or transcriptome levels are often performed to understand phenotypic variation or responses to shocks. In this context, a metabolic network offers a new way of forming hypotheses from these data by joining together separate data points using the structure of the network. In the remainder of this chapter, I presented MOOMIN, an algorithm that does this by searching for metabolic shifts that align with the results of a DE analysis.
In the algorithm, the metabolic network is represented by a hypergraph, and metabolic changes, that is, increases and decreases in flux from one condition to another, are coded by colouring the hyperarcs. Each possible colour for a hyperarc/reaction is given a score based on the DE data. Global constraints are placed on accepted colourings based on the assumption that any shift should preserve a steady state of the metabolism. Finding the global shift that best corresponds to the data can the formulated as an optimisation problem. In Section 2.3.1, I proved that this problem is NP-hard.
I first presented an implementation of MOOMIN that used ASP to solve the optimisation problem. ASP has been previously used successfully in [85, 86] for similar problems involving metabolic networks. However, for MOOMIN the performance in genome-scale networks was insufficient. For this reason, a MILP implementation was adopted. This allows the algorithm to run in reasonable times on a personal computer. MILP also admitted the inclusion of stoichiometric constraints for a more accurate preservation of a steady state.
In Section 2.3.3, MOOMIN was applied to real data. Two data sets were chosen: a comparison of *S. cerevisiae* growth in a batch *versus* a chemostat culture, and a study of the response *E. coli* to mercury stress. Overall, the results show that MOOMIN works as intended, upholding for some reactions the change implied by the DE data and rejecting it for others, while additionally inferring changes that were not reflected in the gene expression. The parameter α controls how "conservative" the algorithm is: the lower the value, the fewer coloured reactions appear in a solution. Enumeration of optimal solutions showed that while sometimes the optimum can be unique, it is also possible to have a large number of alternative optima. Based on the results in Section 2.3.3, it appears that these optima do not represent wholly distinct alternative hypotheses but rather share a common core and differ only in a small subset of reactions. Comparing the topological and stoichiometric solutions, they seem to overlap considerably, confirming the assumption that the topological constraints are an approximation of the full stoichiometric steady state condition. It was also observed that the topological constraints are more "lenient", meaning that the resulting solutions contain more coloured hyperarcs.
For the *S. cerevisiae* data, MOOMIN was able to recover the expected results. For the *E. coli* data, not only did we rediscover the findings of the original authors, but MOOMIN also found changes that were expected in the original publication but not found due to inconsistent results of the DE analysis.

Chapter 3

Metabolic games

Contents

3.1 Introduction

One of the main applications of metabolic network models is phenotype prediction. In other words, given the information offered by the reconstruction about the metabolic capabilities of an organism, we wish to know which would be its actual metabolic state in a set of given environmental conditions. The most basic question of this nature is that of viability: is the organism able to survive or grow in a medium? The quantitative version asks how well the organism is able to grow (compared to, for example, some experimentally obtained baseline or to another organism). In biotechnology applications, it can be of interest to predict the production rate of a target compound. (see [119]). Metabolic models can also be used to understand responses to perturbations: what is the cell's response to environmental stress? Finally, the question of what behaviour does a certain set of genes give rise to is of course of general interest by itself in life sciences.

Arguably the most popular tool for predicting phenotypes based on a metabolic reconstrcuction is flux balance analysis (FBA). It is based on the assumption that metabolism is organised to optimise some identifiable target function, most often biomass production (see Section 1.1.3). As a model, FBA is simple and elegant, and does not require extensive computational resources. Despite this, it can provide accurate predictions of metabolic state in a variety of organisms [180, 57, 56, 137, 138], and has been successfully used in many applications, such as metabolic engineering [191, 3], and the identification of drug targets [92].

However, the assumption of growth yield maximisation is not always valid. For example, *S. cerevisiae* uses a mixture of fermentation and respiration even in the presence of oxygen, termed the *Crabtree effect*. This phenomenon is also observed in several other yeasts: *Saccharomyces bayanus*, *Schizosaccharomyces pombe*, *Kluyveromyces thermotolerans*, and *Dekkera bruxelensis* [161]. Lactate production in *Lactobacillus plantarum* could not be predicted by standard FBA because the acetate production pathway has a higher yield [174]. An *E. coli* strain with gene deletions displayed higher biomass yield than the wild type [176].

Another shortfall of FBA are multispecies communities. Microorganisms rarely, if ever, appear in isolation in nature. Thus understanding microbial communities composed of several species is essential for understanding microbial life in general. Furthermore, because these communities are largely shaped by interactions mediated by the intake and secretion of compounds, metabolic modelling is at the centre of this research. Communities are also becoming increasingly important in biotechnology applications [27]. Whereas in a single-species culture viability can be determined solely based on medium composition, for a multispecies community this is not true since organisms may be able to use the metabolic products of other members to grow. It is also difficult to formulate a general optimisation target for a community of organisms because different species can have different, possibly conflicting needs, and it cannot be expected that they would coordinate their behaviour according to a "common goal".
A possible solution is offered by evolutionary game theory. In contrast to "simple" optimisation, such as in the case of traditional FBA, which only takes into account the abiotic factors, game theory incorporates the fact that organisms "create" their own environment through their interactions with it. Thus optimal behaviour cannot be defined simply in terms of an individual on its own, but one also needs to consider the surrounding members of the community whose actions may change the environment and influence what is considered to be optimal. In the context of microbial communities, the availability of nutrients, for example, can depend on the metabolic states of surrounding organisms. In short, game theory shifts the focus from optimal to "best response".
The idea that conflict of interest dilemmas leading to suboptimal outcomes might arise in the choice of pathways in microorganisms was first presented by Pfeiffer *et al.* in [143]. Pfeiffer and Schuster further advocated the use of game-theory in the study of microbes in a later review [160], and finally the concept of considering different pathways in a metabolic network as strategies in a game was mentioned by Schuster *et al.* in [157]. Meanwhile, game theory has been applied to study a wide range of phenomena in the microbial context. Most closely related to metabolism are the question of which pathway to choose for ATP production [143, 64, 116, 161, 8, 157, 87, 15, 16], public goods, that is, the production of costly extracellular products that benefit not only the producer but its neighbours as well [71, 159, 20, 17, 196], and the choice of primary nutrient [54, 91, 76, 196]. Reviews of the use of game theory for the study of microbes can be found in [142, 160, 150, 81].
In this chapter, I review this literature and discuss the idea of a metabolic game. While game theory has been applied to many situations, definitions of the main components vary, and it remains unclear how exactly a game should be defined based on metabolic reconstructions. I discuss the different choices for actions, payoffs, and the game model.
This chapter largely corresponds to a review article titled "Metabolic games", submitted to Frontiers in Genetics at the time of writing. The article version has a slightly different structure and a few more articles are mentioned in the review part of this chapter. It is composed as follows. In Section 3.2, I review the relevant literature. This section considers the three topics mentioned above (yield *versus* rate in ATP production, the production of public goods molecules, and nutrient choice and cross-feeding), as well as some some other applications. In Section 3.3, I describe and discuss the idea of a metabolic game. The chapter ends with a conclusion.

3.2 State of the art

One of the questions that has been extensively studied through applying game theory to metabolism is ATP production. There is a fundamental trade-off between yield and rate of ATP production in heterotrophic organisms: some of the free energy obtained from substrate

degradation is needed to drive the reaction. Increasing the portion of free energy that is used for driving the reaction increases the rate of ATP production but lowers the yield. The choice of pathway thus presents a social dilemma. Choosing the efficient strategy would maximise resource usage and benefit the population as a whole. However, if an individual cell chooses to stray from this cooperative path, its faster growth rate will allow it to increase in numbers and eventually overcome the cooperators at the cost of the interest of the community.

In [143], Pfeiffer *et al.* explored this question in the context of respiration *versus* fermentation. Most organisms can in principle choose to degrade sugar by both the respiration and the fermentation pathways. While fermentation provides ATP faster, it has a significantly lower yield. Thus fermentation can be seen as a wasteful, "selfish" strategy, while respiration is more efficient in terms of nutrient use. By constructing a simple population model, the authors showed that while a population of fermenters will be smaller due to a faster depletion of resources, they can nevertheless take over a population of respirators due to their faster growth rate. This constitutes the famous "tragedy of the commons" [114]. However, if a spatial component is added, respirators can have a chance. This is because at lower nutrient levels, fermenters will deplete their immediate environment of resources and suffer the consequences.

In [64], Frick and Schuster explored this question further. They too constructed a population model for slow but efficient *versus* fast but wasteful resource use. The authors then interpreted the steady state population densities of both strategies in each different scenario as payoffs: in this way, the situation is a Prisoner's Dilemma with pure respiration forming the cooperation strategy. This is important because were the growth rates to be taken as the payoffs, one would conclude that fermentation is the optimal choice in all instances. However, from the point of view of sustaining the highest possible population density, cooperation, that is respiration, is the best choice.

Kreft studied the question in a spatially structured setting [101]. In a simulation of a biofilm, cells are represented by spheres in a continuous space and grow according to Monod kinetics, while metabolites diffuse on a lattice. There are two types of cells: cooperators that use a high yield, low rate growth strategy, and defectors that have the opposite strategy. Simulations performed by Kreft showed that the outcome of the competition depends on the initial conditions, with different starting patterns of the two strategies leading to qualitatively different results.

Experimental evidence for the results described above was provided in [116]. Maclean *et al.* used yeast as their model organism and grew pure respirators and respiro-fermenters together in different culture set-ups. They found that while the "cheaters" win in a chemostat, in serial batch and spatially structured populations, the two strategies can coexist.

In [161], Schuster *et al.* critically examined the assumption made in FBA of maximisation of biomass yield. They argued that in general there is a trade-off between yield and rate, and that it is not *a priori* clear which of these conflicting goals would be selected for. Based on the theoretical results previously put forth by Pfeiffer *et al.* [143] as well as several examples from nature, the authors concluded that maximisation of yield cannot be considered a universal principle.

Aledo *et al.* [8] studied the yield *versus* rate question but this time in glycolysis itself, which can operate under two different regimes: one with a high yield but a slower rate, another with a low yield but a faster rate. Using a simple matrix game model, with payoffs derived as functions of extracellular free energy and in agreement with the Prisoner's Dilemma payoff scheme, the authors showed that in a well-mixed population, cooperation cannot persist. In contrast, if the game is played on a lattice so that players only interact with their neighbours, cooperation is a possible outcome.

Schuster *et al.* returned to the question of yield *versus* rate again in [157]. They presented a toy model representing a simplified version of ATP production to show that whether max-

imising the yield coincides with maximising the rate depends on the particulars of the system. They also further articulated the idea that alternative pathways can be seen as strategies in the game theoretical sense, and that "choosing" which pathway to use can happen not only through changes in genotype, but also through regulatory changes within the life-span of a cell.

Kareva [87] investigated the yield *versus* rate question in the context of cancer cells where the use of the more inefficient glycolysis pathway is observed as one of the hallmarks of cancerous growth and is known as the Warburg effect [187, 186]. However, in contrast to the previous models, the author argued that the use of glycolysis is the cooperative strategy: while recognising the possibility to increase the rate of glucose uptake, she considered the use of glycolysis to remain detrimental to the individual cell due to its low yield. Meanwhile, the associated lactic acid production can benefit the cancer cell population as a whole, if undertaken in sufficient numbers, because it disproportionately harms non-cancerous cells. Thus glycolysis can be considered as public goods production. The contradiction with previous studies is clear. However, in the ODE system used to model a population of cells with varying rates of carbon allocated to glycolysis in [87], it was observed that glycolytic cells do increase in frequency if they have a faster growth rate.

In two successive papers [15, 16], Archetti presented a public goods model of the Warburg effect. He took the same view as [87] and considered glycolysis as the cooperative strategy amongst cancer cells. The benefit accrued by all participants from glycolysis – increased acidity – is modelled by a double sigmoid function: increased acidity yields a benefit over healthy cells if enough cells are producing lactic acid, but too much will start to hamper the growth of even cancer cells. The dynamics of the frequencies of glycolytic and non-glycolytic cells were modelled using the replicator equation. Because an exact solution of the dynamics for a sigmoid shaped benefit is not available, Bernstein polynomials were used to find an approximate solution. Archetti found that if the cost attached to glycolysis is not too high, glycolytic cooperators can persist at intermediate frequencies.

Another possible social dilemma within microbial communities occurs with necessary but costly functions. If a metabolic function is performed at the cell surface or outside the cell, it means that the benefit incurred can be shared by other cells that are possibly not contributing to the undertaking of the said function. Such a situation is best described by a public goods game.

Gore *et al.* [71] studied the invertase production system of *S. cerevisiae*: in order to grow on sucrose, the yeast needs to hydrolyse the sugar molecule. Because invertase is a surface enzyme, much of the resulting monosaccharides leak out. Because producing invertase is costly, it constitutes a public good. The model of Gore *et al.* is a sort of mix between a public goods game and a matrix game: the authors define payoffs in terms of the fraction of invertase-producers in the population but then go on to compare these payoff values to the well-known 2-player games. If the benefits are linear, cooperation cannot persist unless the benefit derived from sucrose degradation by the invertase-producer exceeds the cost, in which case producing the enzyme is not a public good. On the other hand, with non-linear benefits, frequency-dependent selection allows for a fraction of the cooperators to persist. This result was in line with experimental evidence (presented in the same article) which confirmed both the coexistence of producers and non-producers, as well as the non-linear benefit function.

A similar model was presented in [159]. In this paper, Schuster *et al.* studied generic exoenzyme production assuming again that some fraction of the transformed growth product diffuses directly into the producer cell while the rest is available to the surrounding community. This time the benefit from the public good is given by a Monod function modelling the growth rate attained through the available nutrient. The nutrient acquired in turn depends on the fraction of cooperators in the population and cell density, which is a parameter of the

model. The authors conclude that depending on the parameters, the fraction of public good that diffuses away, the cost of enzyme production, and the cell density, the model can be seen as a Prisoner's Dilemma, a Snowdrift or a Harmony game.

Another example of a public good are siderophores, molecules produced by bacteria that bind poorly soluble iron, allowing its transport into the cell. Siderophores constitute a public good because any cell that harbours the appropriate transporters can take in the bound iron without necessarily having to produce the molecule themselves. Cordero *et al.* [42] studied siderophore-production in marine bacteria. Using a large set of isolates of Vibrionaceae, the authors showed that the siderophore-production trait is routinely lost and gained, implying variable selection pressures that can alternatively promote cheating and cooperation. Furthermore, loss of the production trait occurs through loss of the biosynthesis of the siderophore molecule, but not the associated outer-membrane receptors, creating a "cheater" phenotype that no longer contributes to the siderophore-production but enjoys the resulting benefits. The genetic evidence suggests that public goods-type social dilemmas do indeed contribute to the genetic diversity observed in nature.

Allen *et al.* [9] presented a model for public goods production in a population with an explicit spatial structure, taking the invertase production system of *S. cerevisiae* as the inspiration. A 2-dimensional population structure was modelled by placing the cells at the nodes of a weighted graph. Cells are either cooperators or defectors, with cooperators producing a public good with a cost c and yielding a benefit b, of which a fraction is diffused to the neighbouring cells. Frequency of diffusion is proportional to the edge weights. Impressively enough, Allen *et al.* were able to obtain analytical results for their model: if the benefits from public goods production are mostly retained by the producer, cooperation is favoured whenever $b > c$. If the benefits are mostly shared, cooperation is only favoured if the public good is absolutely essential for survival. Between these extremes, the success of cooperation depends on the structure of the population: cooperation decreases with dimensionality, in other words, a 2-dimensional population is more prone to cooperation than a 3-dimensional one.

In [14], Archetti studied growth factor production in cancer cells as a public goods game. Growth factor production is costly but the benefits are available to all surrounding cells. The benefit function was assumed to have a sigmoid shape and population dynamics were modelled by the replicator equation. As in [15, 16], Bernstein polynomials were used to circumvent the problem caused by the sigmoid function. Archetti found that depending on how exactly the fraction of producers influences the benefit from growth factor, different types of dynamics are possible: a globally attracting mixed equilibrium where producers and non-producers coexist, the fixation of one type depending on the initial frequencies, or the fixation of producers regardless of the initial conditions.

The model presented in [14] was expanded on by Archetti in [17] by introducing a spatial component. In this model, cells are placed in the nodes of a Voronoi graph. A Voronoi graph has the average connectivity of 6, with very few nodes beyond degree 4-8. Cells receive benefits from growth factors produced by producer-cells within a neighbourhood defined by a diffusion parameter, discounted with the distance to the focal cell. The benefit itself is given by a normalised logistic function. In other words, benefits are non-linear. Archetti found that similar to well-mixed populations, cooperation declines as the cost of production increases. Stochasticity in the update rules used to model proliferation and a steeper benefit function also decrease cooperation.

Invertase production in *S. cerevisiae* was also modelled in [196] by Zomorrodi and Segrè. The authors constructed a 2-by-2 payoff matrix based on the metabolic reconstruction of the organism. First, the sucrose hydrolysis reaction was modified to account for the cost of invertase production, which was modelled by a reduced ATP yield, and the "leakiness" of the resulting monosaccharides, which was simulated by a forced export of a portion of the reaction

products. Payoffs in the four possible pairwise interactions were obtained as the optimised biomass yields. In the case that a player is facing a producer, intake of the secreted glucose and fructose is allowed to simulate benefits from the public goods production of one's opponent. Three distinct parameter domains were observed: a high cost of enzyme production or high leakiness of sugars leads to a Prisoner's Dilemma, low cost of production and low leakiness of sugars lead to what the authors called a Mutually Beneficial game, where the equilibrium is to be a producer, and intermediate values of the parameters lead to a Snowdrift game where the equilibrium is a mixed one.

A public goods dilemma can also be observed within a cell. This occurs when more than one virus infects a host cell. It was studied in the phage Φ6 using game theory in [177] by Turner and Chao. The viruses generate diffusible intracellular products essential for reproduction. A "defector" strain is one that has lost the associated protein-coding sequences and can thus not replicate in the absence of a complete virus. However, if a "cooperator" virus is present, providing the necessary extracellular products, the defector strain can replicate faster than the cooperator. Thus the situation can be described as a Prisoner's Dilemma.

Chao and Elena [36] studied viruses in a similar system. The authors considered a trade-off in reproduction and the production of public goods. If this trade-off is linear, the resulting game is a Prisoner's Dilemma (according to the authors), and defection dominates. However, with a non-linear trade-off, an adaptive dynamics style simulation revealed a branching of the population into "ultra-cooperators" and "ultra-defectors".

Perhaps the best examples showcasing the usefulness of game theoretic thinking are situations where frequency-dependent selection leads to polymorphisms in nutrient use. It is often the case that in a given environment, there is a preferred choice for the main carbon source. However, in any realistic scenario, nutrient availability is limited, and it can be beneficial for the individual to opt for a carbon source that is slightly less optimal, but abundant due to being the "unpopular" choice. *Cross-feeding* occurs when two organisms depend on each other for the production of some essential metabolite.

In [54], Doebeli considered the evolution of cross-feeding. He constructed a model for a bacterial culture growing in a chemostat, using glucose as its main nutrient. During growth on glucose, acetate is secreted which can also be used as a nutrient, albeit with a lower growth rate. Doebeli assumed that there is a trade-off in using the secondary metabolite: becoming more proficient in using acetate lowers the ability to use glucose efficiently. Furthermore, this trade-off is subject to gradual change through mutations. Bacterial growth and nutrient concentration was modelled using a Michaelis-Menten type model. Using the theory of adaptive dynamics, Doebeli showed that the frequency-dependent selection following from the trade-off can lead to evolutionary branching and the emergence of a stable polymorphism of glucose and acetate specialists. He also found that if the dynamics are changed to model a serial batch culture instead of a chemostat, evolution of cross-feeding becomes much less likely. In a chemostat culture, the concentration of nutrients is kept constant, while in a batch culture nutrients are allowed to be depleted. These results were further expanded and provided experimental confirmation in [65].

Wintermute and Silver [189] studied cross-feeding in *E. coli* auxotroph pairs. The metabolic networks of two different strains were joined to allow for the exchange of metabolic products, and flux distributions were determined by minimising the difference to a wild type strain growing alone. These predictions were then compared to co-culture experiments. It was found that the *in silico* experiments tended to overestimate growth, possibly reflecting the fact that joining the metabolic models may implicitly induce more cooperation between the two strains than is realistic. However, the co-culture experiments nevertheless showed improved growth for a subset of the pairs compared to growth in monoculture, providing evidence for metabolic synergy. The authors also investigated the value of shared metabolites using the concept of

shadow prices. The shadow price of a metabolite, in the constraint-based analysis context, can be understood as a measure of how a change in its availability would affect the value of the objective function. In other words, it can be used to measure how costly a metabolite is for its producer or how much benefit it yields if acquired. The analysis showed that metabolites that tend to be shared are those that are "cheap" for the secreting organism.

Kianercy *et al.* studied the Warburg effect and the *reverse* Warburg effect [91]. The reverse Warburg effect refers to the phenomenon wherein some cells in a tumour use lactate secreted as a by-product of glycolysis as their energy source. The authors' model is a 2-player matrix game with two types of players: hypoxic and oxygenated cells. Both types have the same available strategies: using either glucose or lactate as their nutrient. Lactate is secreted by hypoxic cells using glucose. Similarly to [87] and [15, 16], the authors take yields as payoffs. Thus using glucose gives a lower payoff for hypoxic cells. The authors found that there exist two stable states and conclude that lactate secretion can induce a transition between high and low levels of glucose consumption.

In [76], Healey *et al.* investigated phenotypic *bet-hedging* by experiments and a game theory model. Bet-hedging refers to a hypothesis that microbes may increase their survival in fluctuating environments by implementing a stochastic phenotype. In other words, a genetically homogeneous population might display two (or more) distinct phenotypes. In the language of game theory, this would constitute a mixed strategy. The model system in [76] was *S. cerevisiae* that prefers glucose as its carbon source, but also harbours the GAL network for metabolising galactose. The game theory model used was a simple foraging game, where a population of players must choose between two resources. One of the resources is the preferred one, and so there is an additional cost associated with using the inferior resource. However, if all members of the population have chosen the preferred resource, it is better for an individual to choose the other. This leads to a stable mixed equilibrium of users of both resources. Experiments performed by Healey *et al.* corroborated this theoretical result.

In the article already mentioned above [196], Zomorrodi and Segrè studied amino acid mediated ecological interactions in *Escherichia coli*. Producer strains leak out amino acids which are costly to produce, and can be taken up by mutants lacking the ability to synthesise them. Several different amino acids were investigated, with up to two at a time spanning four possible strategies (genotypes). Like in the case of invertase production in yeast, both the level of leakiness and the cost of production influence the type of equilibria observed. With low enough levels of leakiness, both an equilibrium with a full producer coexisting with a complete auxotroph, as well as cross-feeding are possible. With increasing leakiness, the full producer becomes non-viable. However, it was also observed that due to interdepencies in amino acid production, in some situations cross-feeding is not possible because losing the ability to produce one amino acid leads to the loss of the ability to produce the other. Zomorrodi and Segrè also studied the evolutionary dynamics of these interactions by performing *in silico* invasion experiments. They found that cross-feeding can emerge through the progressive loss of amino acid synthesis capabilities, and that this mutually dependent coalition is often stable against invasion by non-producers, consistent with previous experimental findings [140, 34].

Besides the three topics introduced above, game theory has been applied in a variety of other contexts as well in the study of microorganisms.

In [31], Bremermann and Pickering constructed a game theoretical model of two or more microbial parasites competing inside a host. They assumed that host longevity is influenced by the growth rates of the parasites. Parasites aim to maximise transmission during the lifetime of the host. By exploring different functional forms for the host longevity - parasite reproductive rate dependency, and calculating the resulting Nash equilibria, the authors concluded that parasite reproduction below maximal rates can occur. They also suggest that virulence may increase with an increasing number of parasites competing within a host.

Microbial interactions can also display circular, Rock-Paper-Scissors-type dynamics. An example is bacteoricin production in *E. coli*. Three genotypes are possible: one which possesses the genes for both the production of colicin (a toxin), as well as a protein that provides immunity to the toxin. If the ability to produce colicin is lost, the result is a strain that is still immune to the toxin, but that enjoys a possible growth advantage to the ancestral strain due to forgoing the cost of producing the toxin. Finally, a susceptible strain produces neither the toxin nor the protein. Kerr *at al.* [90] studied this system through both simulations and co-culture experiments. Both investigations showed that when spatial mixing is high, that is, interactions are global rather than local, two of the strains go extinct and the strain that is resistant to the bacteoricin but does not produce it emerges as the winner. However, in spatially structured populations all three strains co-exist, creating patches that "chase" one another according to who can outcompete whom.

Kirkup and Riley [98] provided *in vivo* evidence for a Rock-Paper-Scissors (R-P-S) scenario occurring in *E. coli*. The authors inoculated mice with four different types of the bacterium: two types of colicin producers, a susceptible strain, and a colicin-resistant non-producer. Colonisation and changes in dominant bacteria were detected by monitoring the bacteria present in the fecal pellets of the mice. The results showed that the bacterial competition dynamics indeed conformed to the structure of the R-P-S game.

Eswarappa [61] constructed a 2-player game modelling the conflict between a pathogenic bacterium and its host. Each player has two strategies: the pathogen chooses between residing in intra- or extracellular space inside the host. Correspondingly, the host can choose to activate either intra- or extracellular defences against the infection. By analysing the possible Nash equilibria of the game, Eswarappa concluded that a mixed strategy will be used by both the pathogen and the host.

Martìn and Elena [118] studied mixed viral infections in plants. *Arabidopsis thaliana* was infected with either of two viruses, the Cauliflower mosaic caulimovirus (CaMV) or the turnip mosaic potyvirus (TuMV), or with both, and accumulation of viral load was measured to determine their success. The result can be understood as a 2-by-2 matrix game where the two viruses are the available actions, and the viral loads are the payoffs. It was discovered that TuMV dominates, reaching a higher fitness by itself than CaMV, but also benefiting from a co-infection by CaMV.

Momeni *et al.* [124] studied spatial self-organisation in bacterial cooperator-cheater dynamics. Three different strains of *S. cerivisiae* were cultured together both *in vitro* and *in silico*. One of the strains is an auxotroph for adenine and releases lysine, another is an auxotroph for lysine and releases adenine, and the "cheater" is an auxotroph for lysine but releases nothing. The authors found that cells tended to organise into a non-random pattern where cooperators have more neighbours than cheaters, favouring cooperation.

In game theoretical models of microbial interaction, population size is often implicitly assumes to stay constant. For example, in the commonly used replicator equation only the relative frequencies of strategies are considered, and a globally dominating strategy is hence assumed to completely take over the population. In [112], Li *et al.* showed that this approach might not always be valid. The authors cultivated together two commensal bacteria, *Curvibacter* sp. (AEP1.3) and *Duganella* sp. (C1.2)), and found that which species becomes dominant depends on the initial conditions. However, in either case, both species continue to grow in absolute density, implying that the dominant one does not necessarily displace the other. Thus both frequency and density dynamics might need to be taken into account to accurately describe microbial populations.

Kelsic *et al.* [89] modified the R-P-S-model of bacteriocin production by making resistance to the toxin a public good: instead of simply being immune to an antibiotic, the resistant strain degrades it, rendering its immediate environment safe. Grid-based simulations showed

that in contrast to the previous models, spatial separation of different strains is not needed for stable coexistence. The authors also found that more complicated community structures with several different antibiotics and strains can achieve stability.

Sequential games are rarely applied to the study of microbes. A notable exception can be found in [144]. Pollmächer *et al.* sought to model the dynamics of an invasion of *Aspergillus fumigatus*, a pathogenic fungus, into the lung alveoli of humans. The authors represented the course of an *A. fumigatus* infection in the form of three subsequent matrix games, each representing a distinct phase of the host immune response, played on a graph that serves as a model of the spatial structure of the human lung. The players are fungal cells, and the action space is formed by the different stages of the life-cycle of the pathogen, each with its own susceptibility to different immune responses. Each game is iterated until an equilibrium is reached, with mutation and adaptation taking place at each step. By performing stochastic simulations with different parameter values, Pollmächer *et al.* were able to determine the relative importance of different aspects of the immune response. Namely, for low infection-doses, the phagocytic activity of alveolar macrophages that is present in the first steps of the immune response is sufficient to control infection, while for higher infection-doses the main task of the alveolar macrophages is the recruitment of polymorphonuclear neutrophils.

Wu and Ross [190] adopted game theory to study the human intestinal microbiota. The authors constructed a matrix game with three different actions representing three types of bacteria: antibiotic-sensitive (AS), antibiotic-tolerant (AT), and *Clostridioides difficile*. *C. difficile* is a commonly found human pathogen that can cause infections after antibiotic treatment has disrupted the natural microbiota. Wu and Ross assumed that the interaction between the main components of the intestinal community, AS and AT, is a Snowdrift game, and that the remaining payoffs can be defined by two parameters: the payoff for *C. difficile* against AT, and the payoff for *C. difficile* against AS. The authors explicitly adopted an interpretation of the replicator equation where "the fitness (growth rate) of each phenotype depends on the frequency of each phenotype". Wu and Ross concluded that depending on the values of the two parameters, the system can have one or two stable fixed points, and that this determines how susceptible the host microbial system is to perturbation.

The application of game theory to microbiology has been reviewed in a few articles. Pfeiffer and Schuster [142] advocated the use of game theoretical principles to supplement purely optimisation-based approaches to the study of biochemical systems. Taking as examples the evolution of cross-feeding and choice of pathway in energy production, the authors argued that traditional optimisation may be insufficient to explain the organisation of biochemical systems, because it assumes that organisms evolve in a fixed fitness landscape. In contrast, evolutionary game theory takes into account the fact that the evolving organisms influence their environment and thus the fitness landscape becomes dynamic.

Schuster *et al.* provided similar arguments in [160]. The authors also discussed the choice of payoff: a common choice is *per capita* reproduction rate. However, for example in biofilms, reproduction rate per area covered by the biofilm might be a more suitable choice. Schuster *et al.* pointed out that payoff should be time invariant, and so reproduction rate should be integrated over life span. Steady-state population densities have also been proposed as payoffs, but the authors argued that it is unclear whether this would produce correct predictions in all situations.

Ruppin *et al.* [150] reviewed the literature on constructing and studying metabolic networks, dedicating a part of the article for the use of game theory in this context. The authors posited that the stoichiometric matrix and the associated flux bounds for the reactions of the network should suffice to define a "metabolic game".

In two complementary papers, Bohl, Hummert, *et al.* reviewed the use of game theory in subcellular [29] and cellular [81] biology. In the former article, games describing catalytic RNA,

gene replication, several viruses infecting the same host, and formation of protein complexes were discussed. In the latter, the authors focussed on games where cells are considered as the players. In addition to topics covered in this chapter, they discussed signalling games, relevant in the study of quorum sensing.

3.3 A metabolic game

Most of the studies reviewed in the previous section have evoked game theory as an explanatory device, making use of the established knowledge on famous games such as the Prisoner's Dilemma to qualitatively describe specific observed phenomena, or alternatively used microbiology as a means to provide real life examples of these games. In contrast, our focus here is the idea of using game theory to supplement constraint-based analysis. This is not to dismiss the aforementioned research, which I believe to still be relevant to the topic at hand, as well as of great general importance, but to make clear the difference.

The idea of a metabolic game originates from the articles by Thomas Pfeiffer and Stefan Schuster [143, 142, 161, 160, 157], and was recently given the most concrete realisation to date by Zomorrodi and Segrè [196]. It is meant to answer the need to expand the scope of FBA to cover situations where its assumptions fail.

In constraint-based analysis (see Section 1.1.3), thermodynamic factors and the assumption of a (pseudo-)steady state define a space of what can be considered feasible metabolic behaviours, or more formally, flux distributions. This space is a representation, based on the genome and these few general principles, of the metabolic capabilities of the cell. The question of phenotype prediction then corresponds to the question: which metabolic state will the cell choose. Obviously the cell "choosing" is just rhetoric, for what we are really looking for are choices such that they will best guarantee survival and proliferation, and it is in fact this proliferation that will actually make this "choice". In other words, we are looking for metabolic behaviours that maximise fitness, because it is those behaviours that will by definition persist.

In the language of game theory, the different possible metabolic states can be seen as the actions in a game. A *metabolic strategy* would then correspond to either a specific choice of a metabolic state, or perhaps a mechanism that dictates this choice based on some rule in a more complicated scenario. The players in a metabolic game are surrounding cells that are within the reach of influence through metabolic interactions. The payoff would be any measure of success that can be determined given the metabolic strategies of the players, for example, growth rate or biomass production.

In FBA, the expressed phenotype is predicted based on optimisation. The underlying assumption is that natural selection has configured a cell's metabolism in such a way as to be maximally efficient, further supposing that efficiency would translate to better fitness. In a metabolic game, the expressed phenotype(s) correspond to the solution or equilibrium of the game. The main idea thus remains the same: the phenotype(s) is chosen based on some measure of what is "best" for the organism. However, in contrast to FBA, which can be seen as optimisation "in isolation", the game theoretical perspective takes into account the possible interactions with surrounding cells.

A word or two on the different levels of selection is in order. Namely, natural selection is most often concerned with the propagation of genotypes: successful individuals have more offspring, carrying the same genes. These genes code for the traits that made the ancestors successful, thus spreading the traits along with the genotype. This is also usually the implicit assumption in evolutionary game theory: the strategies are coded in genes, and increase in frequency or "win" when their payoff, proportional to their fitness, is better than that of the alternatives. The usual meaning of "rationality is replaced with natural selection" is this.

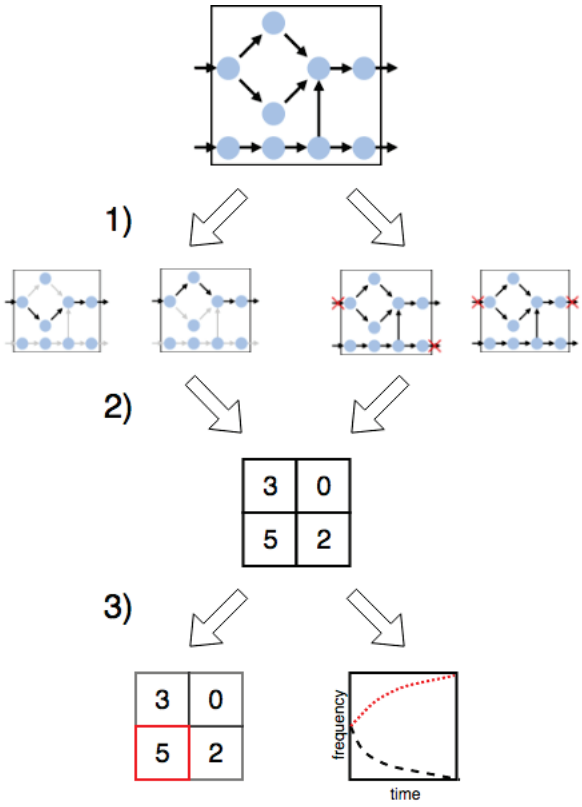


Figure 3.1 – A schematic description of building a metabolic game. 1) One or several metabolic networks are stratified to define an action space in terms of distinct, feasible metabolic behaviours. For example, the different actions can be defined by which pathways are active. Another option is which metabolites are imported into and exported out of the cell. 2) The outcome in each specific interaction scenario spanned by the action space(s) is calculated for the two or more participating cells using FBA. Additional constraints are added to characterise the metabolic behaviours chosen in each scenario. 3) Equilibria are identified in the payoff matrix. Alternatively, a simulation of population dynamics can be performed to determine which metabolic behaviours would triumph in frequency-based competition.

Here our perspective is subtly different, because in a metabolic game the genotype is already given. Thus the choice is made on the level of expression rather than genes. Often this makes very little difference. Knock-outs, the removal of genes, for example, are often considered as genetic mutations. However, blocking a reaction or a pathway in a metabolic network to simulate a gene deletion is no different from blocking it to simulate alternative metabolic behaviour. If it turns out that not carrying out a certain function is better for the organism in a certain situation, the gene does not need be deleted, it will just not be expressed.

The general idea of constructing a metabolic game is illustrated in Figure 3.1. In the remainder of this section, I will discuss the various aspects of systematically defining such as game, namely, how to choose the main components of the model: the players, the action space, and the payoffs.

The first component of a game are the players. They are the participants in the interaction under study. Many of the articles discussed in Section 3.2 used some form of a 2-player matrix game to make their arguments. In principle, this type of game represents a situation where two individuals face each other in a single interaction to obtain a single payoff. With this in

mind, it seems strange to use this model when talking about microbial populations. However, as discussed, when the matrix game is embedded in the replicator dynamics or another kind of frequency dependent selection model, it starts to more closely resemble a microbe culture. In a way, payoffs are obtained according to who one's average neighbour is at any given time, as might be imagined to happen in a well-mixed culture.

Yet the question remains if pairwise encounters are sufficient to capture the interaction dynamics of microbes that mainly influence each other through diffusible molecules. The other type of model often used is the public goods game. At first glance it seems to more accurately describe an interaction through diffusible molecules because it considers several players to take part in the game at the same time. For example, in the case of invertase production, it is intuitive to consider the game to comprise those cells that the released glucose can be assumed to reach. However, there are some problems with using the public goods game as a general model. Firstly, the benefit function must be accurately estimated since its form can greatly influence the type of dynamics it gives rise to (see for example [14], see also [71]). This might be difficult to do without experimental evidence. Secondly, public goods games with nonlinear benefit functions can be difficult to analyse [14], although some progress has been made in this area recently [18].

Explicit consideration of spatial structure could facilitate properly defining interacting agents. Even if the underlying model is a 2-player game, embedding it into a spatial model so that individuals interact with those around them, and the changes resulting in the environment from these actions (depletion of nutrients etc.) happen locally, will be more faithful to nature. The standard way to represent spatial structure in game theory is to assign players to nodes in a graph as was done in [17]. This approach might be most applicable to environments such as biofilms. The other option is to use partial differential equations to include spatial dimensions in the population dynamics. The main problem with both approaches is that usually the only analysis possible is through simulations. Furthermore, parameters such as diffusion coefficients might be needed to specify the model.

Considering all of the above, it seems that if the goal is to specify a systematic framework in which a metabolic game can be defined based mainly on the metabolic reconstructions of the organisms, the simple matrix game should be the model of choice. Indeed, in order to have a computational framework anywhere close to the simplicity of the original FBA formalism, it seems that only the high level ideas from game theory, mainly considering the choice available for one individual in conjunction with the choices available to their opponents, can be included. This is already captured by the matrix game. In addition, authors have arrived at similar conclusions modelling the same situation with various more complicated models [71, 159] and the simpler matrix game [196].

With regard to the choice of action/strategy space, the question is mostly a technical one. In principle, a game constructed on the basis of metabolic networks would consider as available actions the range of feasible metabolic behaviours, in other words, the flux cone ([38, 39], see Section 1.1.3). However, from a practical standpoint it is evident that some abstraction is needed. Firstly, in order for a matrix game to be defined, the action space needs to be discretised. Secondly, with the number of reactions and thus of flux values to be defined routinely reaching to thousands, the game would surely quickly become intractable.

Several approaches to a decomposition of the flux cone have been proposed. Most notably, three related concepts, *elementary flux modes* [156, 158], *extreme currents* [38], and *extreme pathways* [154] (see also Section 1.1.3) all formulate a mathematical definition of a pathway using concepts from linear algebra and convex analysis. Using such concepts, the space of available metabolic phenotypes can be characterised in terms of which reactions are active, each set corresponding roughly to separate biochemical pathways that are able to operate at a steady state. Unfortunately, the number of elements in such a decomposition grows expo-

nentially with the size of the network [99, 162, 2]. It might thus be impossible in practice to define the action space simply using these concepts, at least at the level of genome-scale reconstructions. De Figueiredo *et al.* have offered a possible amendment by proposing an efficient procedure to compute elementary flux modes in order of increasing number of reactions [51].

Other concepts worth exploring are the *phenotypic phase plane* put forth by Edwards *et al.* [58] and the *flux tope* by Gerstl *et al.* [70]. A phenotypic phase plane is defined by the uptake rates of two nutrients. The optimal metabolic behaviour is calculated at each point of the plane using a biomass function. It turns out that such a plane is divided into a finite number of distinct regions with qualitatively different metabolic behaviour. A flux tope is obtained by specifying a direction for all reversible reactions. It corresponds to a maximal "pathway" (as opposed to a minimal one, such as an elementary flux mode). The authors report that the calculation of all flux topes is possible even at a larger scale.

In [196], available metabolic actions were not defined explicitly in terms of flux distributions but rather by excreted compounds. One or several metabolites of interest were first forced to be exported and hence produced (or alternatively to *not* be produced simulating auxotrophy), after which the metabolic state can be determined using standard optimisation principles with the additional constraints. There are compelling arguments for defining actions in metabolic games using extracellular compounds. In general, microbial interactions are often mediated by the exchange of molecules. By focusing on these compounds, the elements of the action space have a clear interpretation in the context of interaction. The set of possible secretions is also much more tractable than the space of all possible metabolic phenotypes.

Interactions based on extracellular metabolites were characterised from a slightly different point of view in [100]. Klitgord and Segrè asked whether it is possible to predict species interactions based on culture media. Using genome-scale stoichiometric models they tested whether growth of two organisms was possible in isolation and in tandem in a given medium. This approach showed examples of both mutualistic and commensal relationship induced by growth media.

The application of shadow prices presented in [189] (see Section 3.2) is also interesting in this regard. To recapitulate, Wintermute and Silver showed how the costs and benefits of extracellular metabolites can be calculated using constraint-based methods. Such an analysis could be very useful for metabolic games since it allows one to compute both the cost of producing a diffusible molecule as well as the benefit derived from it by the organism that is able to receive it.

In a thesis work [185], Wannagat showed how to compute the minimal sets of compounds two organisms need to exchange in order to be able to grow. Here the approach was qualitative and was used to categorise interactions in terms of their type, but such a procedure could be used also to define the action space in a metabolic game.

Finally, in order to construct a game, one needs to define the payoffs. This is arguably the most crucial step since the payoff values will largely determine the predictions of the model. There is a particular importance to not only qualitatively, but also quantitatively establish accurate payoffs here since the hope is for metabolic game theory to match the predictive ability of FBA. One example from the literature discussed in this paper highlights both the importance and the difficulty in defining payoffs.

In several papers [143, 64, 116, 161], fermentation in the presence of oxygen is seen as a classic "cheater" strategy. From an individual's point of view, the inefficiency of fermentation in terms of yield is not "seen": what the cell experiences as the consequence of its choice is a growth rate exceeding that of its conspecifics. The result of a wasteful use of resources is only felt at the population level, resulting in a lower sustainable cell density. This is the (in)famous Prisoner's Dilemma. However, when essentially the same situation has been

discussed in the context of cancer [87, 15, 17], a completely opposite view has been adopted. Here, fermentation was seen as the cooperation strategy. For example, in [15], Archetti described using fermentation as a contribution to a public good, the cost of the action being the loss in yield compared to respiration. While it can be argued that the underlying biology is very different for single-celled microbes and cancerous tissue, the discrepancy is still puzzling. The problem of properly defining payoffs in the yield vs. rate dilemma is related to that of normalisation in FBA [161]. In order to "ground" the flux vector, normalisation is needed. A common choice for a numeraire is the uptake of a primary nutrient. The fact that maximisation of flux through the biomass reaction in FBA leads to a *de facto* maximisation of biomass yield follows from this operation. Consider now the situation in ATP production. If the value of the objective function in a standard FBA approach is taken as the payoff, respiration is a better strategy than fermentation. However, as already discussed, a fermenter can outgrow its respiring neighbour. From the perspective of evolutionary game theory, it is thus clearly the winner, and its payoff should reflect this fact. However, if we simply switch the payoff from yield to actual rate of biomass production, two fermenters would also obtain the highest payoff together. This is because we have assumed in a simplified way that the external resources are infinite, and hence two fermenters are able to sustain the increased uptake of nutrients they achieve in the presence of respirators. In order to arrive at the Prisoner's Dilemma payoff structure, we need to take into account that if everyone uses fermentation, it can no longer provide the benefit it has over respiration because of a depletion of nutrients.
The above example showcases the difficulty in appropriately quantifying the outcomes in a metabolic game. Optimisation of an appropriate objective function can certainly accurately identify "catastrophic" outcomes where growth is not possible, but when conclusions are drawn as to which metabolic strategy would win in intra- or interspecific competition, caution is warranted. One must make sure that the quantity under consideration is apt to decide the winner(s) in an evolutionary sense.
The definition of the action space can also offer a way to quantify the payoffs. For example, if different metabolic phenotypes are characterised by imported and exported metabolites, benefits and costs can be calculated following [189]. This could open the way for a more systematic definition of public goods games using only the knowledge obtained from metabolic models.

3.4 Conclusion

In this section, I reviewed the literature on applying evolutionary game theory to the study of microorganisms, with special attention paid to studies related to metabolic interactions, and discussed the idea of a metabolic game: a game theoretical model of microbial metabolism based on (genome-scale) metabolic reconstructions. Three topics can be distinguished as having received notable attention: choice of pathway in ATP production known as the "yield *versus* rate" question, so-called public goods dilemmas where the production of costly metabolic products yields benefits for individuals other than the producer, and nutrient choice and cross-feeding where frequency-dependent selection dictates the choice of metabolic behaviour. The most often used model is the matrix game, describing either the interaction between two individuals or the interaction between an individual and its population. Frequently, famous examples such as the Prisoner's Dilemma are evoked to either explain observations, or to present a particular biological system as an example of the famous game. Implicit in these treatises is that since the equilibria of these games are known, should the payoff structure follow that of the game, observed behaviour also corresponds to the equilibrium strategies.
Another popular model is the public goods, used to study systems such as the invertase production in *S. cerevisiae* and the production of siderophores in some bacteria. While the simple

model with linear benefits predicts extinction of producers in the absence of remedying factors (for example, spatial structure), nonlinear benefits allow for the coexistence of cooperators and cheaters. The public goods game is an appealing model for the study of microbes because it very naturally incorporates the idea of a group of individuals interacting simultaneously through diffusible molecules. However, quantifying benefit functions can be difficult, and the analysis of public goods games with nonlinear benefits can be difficult.
Game theory has also been proposed as a possible way to improve the applicability of FBA. This leads to the idea of a metabolic game. By constructing a game where actions are defined as alternative metabolic behaviours, phenotypes can be predicted as the equilibria of the game. It appears the matrix game, possible with more than two participants, is the most promising model for the game. Two approaches have been proposed for defining the action space: alternative pathways and excreted compounds. While considering pathways as the actions seems conceptually very promising, current systematic definitions of a pathway such as the concept of elementary modes can lead to a computationally intractable number of distinct choices. In contrast, excreted compounds appears to be a more feasible way to systematically discretising the space of possible metabolic behaviours.
Quantifying payoffs is another remaining challenge. For a straightforward application of the framework, it would desirable to be able to determine payoffs in a manner independent of the context. In other words, calculating payoffs should not require specific knowledge about the particular interactions or, for example, modifications of the metabolic reconstructions. However, knowledge about the costs and benefits of specific molecules should be considered as an option, since this information can be derived from stoichiometric models of metabolism as shown in [189]. The difficulty of accurately defining payoffs is showcased by the yield *versus* rate question of ATP production, where different authors have considered respiration to be alternatively a cooperative or a defective strategy.
It is my belief that suitable model systems are needed for the further development of metabolic games. While the standard workhorses – *E. coli* and *S. cerevisiae* – are useful because they are well understood and very high quality reconstructions are available, there is a risk of developing the model to fit what is already known. Moreover, a true test for the metabolic game model would be a system composed of two or more different species.
While I have focussed here on the application of game theory to specifically predict metabolic behaviour, research involving microbes and game theory in general remains an interesting topic. As has been suggested [81], it can be argued that the assumption of strict rationality often underlying game theoretic analyses applies *better* to organisms below the level of complex cognition. Furthermore, whereas in macroscopic animals accurately quantifying payoffs can be next to impossible, in microorganisms it cappears much more feasible [134]. Microbes thus present a very appealing source of model organisms for the study of evolutionary game theory.
Finally, besides games, other models from economics have generated interest in the field of microbiology. The concept of comparative advantage [147] was thus applied to gene circuits in [59]. The authors showed that when two bacterial species trade signalling molecules necessary for survival, they both enjoy improved growth, as predicted by the theory of comparative advantage. In [172], Tasoff *et al.* used general equilibrium theory [179] to understand the mutualistic exchange of compounds between micro-organisms. The authors argued that comparative advantage is a necessary condition for the exchange to take place. This theory can be further extended to several organisms exchanging multiple compounds. Other concepts that have been suggested for applications in the microbial context include avoidance of bad trading partners, establishment of local business ties, diversification or specialisation, monopolisation of a market, and elimination of competitors [188].

Chapter 4

Xylella fastidiosa epidemiological model

Contents

4.1 Introduction

Xylella fastidiosa is a plant pathogenic bacterium, capable of infecting a wide variety of commercially important crops such as almond, mulberry, peach, olive, citrus, and plum [171]. In grapevine (*Vitis vinifera*), it causes the so-called *Pierce's disease* (PD) which threatens the future of viticulture, especially in highly impacted areas such as Southern California where it has caused massive decline in vine acreage [66]. *X. fastidiosa* impedes the water transmission inside the vine, causing leaves to wither and shoots to die, and significantly decreasing the longevity of the plant.
X. fastidiosa is spread by specialised insect vectors [82]. The main vector responsible for its transmission in grapevine is the glassy-winged sharpshooter (GWSS, *Homalodisca vitripennis*). Adult GWSSs harbour the pathogen in their foregut and infect vines when feeding on their xylem-sap.
So far, no cure has been found for a *X. fastidiosa* infection. However, several approaches have been proposed for controlling the disease. These include control of the insect vectors, control of non-vine host plants, alteration of cropping techniques, breeding or engineering resistance to PD, control through avirulent strains of *X. fastidiosa*, control via other beneficial microbes, bacteriophages, antagonistic bacteria, antibacterial substances, and other defence-stimulating compounds.

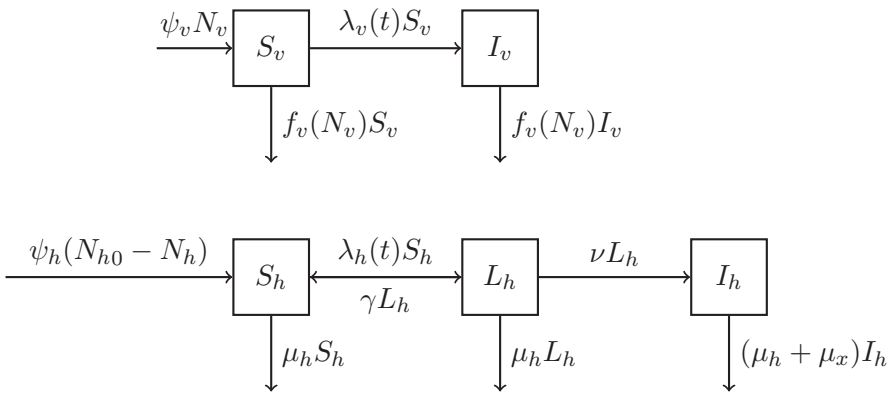


Figure 4.1 – A schematic description of the model.

Given the vast variety of different control strategies, and the threat posed by the PD, it is of great interest to better understand the dynamics of the disease and the relevance of different approaches to its control. To this end, we developed an epidemiological model of the spread of *X. fastidiosa* in a grapevine population (*i.e.*, a vineyard), and used it to estimate the potential of different control strategies. The model is defined by a system of ordinary differential equations (ODEs), and it is based on previously published models of malaria in humans. We calculated analytically the disease-free equilibrium of the system and the basic reproduction number and performed simulations with parameters derived from the literature to study the endemic equilibria and their stability. To gauge the significance of different parameter values and thus evaluate the potential of different control strategies, we performed sensitivity analysis for the basic reproduction number and the endemic equilibrium.
This chapter is composed as follows. In Section 4.2, I first describe the model and present the baseline parameter values chosen. I derive expressions for the disease-free equilibrium and the basic reproduction number, and investigate the endemic equilibria through numerical simulations. Lastly, I present the results of a sensitivity analysis performed for the basic reproduction number and the endemic equilibrium, and discuss some of the implications for the control of PD. The chapter ends with a conclusion.
The work presented here forms a part of a journal article titled "Pierce's Disease of Grapevines: A Review of Control Strategies and an Outline of an Epidemiological Model" with a much larger scope that was published in Frontiers in Microbiology [103]. The idea for the model came from Ifigeneia Kyrkou. I developed the model with her and performed all the analyses. The biological interpretations of the results are due to Kyrkou.

4.2 Epidemiological model

4.2.1 Model description

The model was based on epidemiological models of malaria, most notably one put forth by Chitnis [37]. It describes a vineyard under high PD pressure, with Southern California used as the reference. A schematic description of the model is given in Figure 4.1.
The vine population is described by three state variables, S_h, L_h, and I_h, corresponding respectively to healthy, latent, and symptomatic. Latent vines are ones that harbour the pathogen and can transmit it to GWSSs, but that can still recover from the disease and do not show any external symptoms. Symptomatic vines cannot recover, and due to external symptoms will be removed by vineyard managers.

Since the vine population is an artificial one, we consider it to have a maximum size N_{h0} that the manager aims to maintain. This is modelled by missing vines, that is, the difference between N_{h0} and the total number of vines, denoted by N_h, being replaced at a constant per capita rate ψ_h. We assume that all new vines added to the population are free of the pathogen.
Vines contract the disease from infected GWSSs probing on them. The per capita rate of inoculation is a function of the number of probes per vine the current GWSS population is performing per unit of time, the proportion of GWSSs that are infected, and the probability of transmission per probe. It should be noted that this describes a fairly local situation, where it is assumed that the total number of probes the GWSS population performs is distributed evenly across all vines. In other words, we do not account for GWSS aggregation and swarm behaviour, a factor that can significantly influence the disease spread.
A latent vine will recover at a constant rate γ and progress to symptomatic at a constant rate ν.
Vines are being removed because of age and other diseases at a constant rate. There is an additional removal rate for symptomatic vines, reflecting the fact that viticulture guidelines dictate managers to remove vines that show symptoms of PD. Consequently, in the absence of PD, vines have some stable population level $N_h < N_{h0}$.
The GWSS population is described by two state variables, S_h, and I_v, which correspond respectively to susceptible and infected. New GWSSs are born through a constant per capita birth rate. We consider only the adult stage of the GWSS life cycle, since the nymphal stages have minor or no effect on disease spread. Thus GWSSs are born as adults, and we adjust the birth rate to account for egg and nymphal survival.
GWSSs contract the pathogen by feeding on infected vines, and harbour it for the rest of their lives. We assume they probe plants at a constant rate σ. There is a preference parameter p that controls which proportion of these probes are on grapevine. We assume the GWSS does not contract the pathogen from other sources. Thus the per capita rate at which susceptible GWSSs contract the pathogen is a function of the probing rate, GWSS preference to vine, the probability of transmission per probe, and the proportion of vines that harbour the pathogen.
GWSSs die at a *per capita* death rate that consists of a constant, density-independent part and a density-dependent part. The density-dependent death rate μ_{v2} is used to control the equilibrium population size of the GWSS, which we set so that it amounts to approximately 2 insects per vine in the absence of disease for most of the analysis (values given in Table 4.3). However, we also vary μ_{v2} to simulate situations with different vector densities to explore how it affects the disease dynamics (see Figures 4.2 and 4.4). There is no pathogen-induced death for GWSSs.
The model is given by the following set of differential equations:

$$\frac{dS_v}{dt} = \psi_v N_v - \lambda_v(t) S_v - f_v(N_v) S_v \tag{4.1}$$

$$\frac{dI_v}{dt} = \lambda_v(t) S_v - f_v(N_v) I_v \tag{4.2}$$

$$\frac{dS_h}{dt} = \psi_h(N_{h0} - N_h) - \lambda_h(t) S_h + \gamma L_h - \mu_h S_h \tag{4.3}$$

$$\frac{dL_h}{dt} = \lambda_h(t) S_h - \gamma L_h - \nu L_h - \mu_h L_h \tag{4.4}$$

$$\frac{dI_h}{dt} = \nu L_h - (\mu_h + \mu_x) I_h \tag{4.5}$$

Explanations for the state variables are given in Table 4.1 and for the parameters in Table 4.2.

N_h:	Total number of vines
S_h:	Number of healthy vines
L_h:	Number of latent vines
I_h:	Number of symptomatic vines
N_v:	Total number of GWSSs
S_v:	Number of healthy GWSSs
I_v:	Number of infected GWSSs

Table 4.1 – Variables

ψ_v:	Per capita birth rate of GWSSs, time^{-1}
ψ_h:	Per capita replacement rate of (missing) vines, time^{-1}
N_{h0}:	The maximum number of vines, unit
μ_{v1}:	Density-independent part of GWSS death rate, time^{-1}
μ_{v2}:	Density-dependent part of GWSS death rate, unit $\cdot$ time^{-1}
$f_v(N_v) = \mu_{v1} + \mu_{v2}N_v$:	Per capita density-dependent death rate of GWSSs, time^{-1}
μ_h:	PD-independent death and removal rate of vines, time^{-1}
μ_x:	PD-induced removal rate of vines, time^{-1}
$\lambda_v(t) = \frac{\beta_{hv} p \sigma (L_h + I_h)}{N_h}$:	Per capita inoculation rate for GWSSs, time^{-1}
$\lambda_h(t) = \frac{\beta_{vh} p \sigma I_v}{N_h}$:	Per capita inoculation rate for vines time^{-1}
β_{hv}:	Probability of transmission from vine to GWSS during a probe, dimensionless
β_{vh}:	Probability of transmission from GWSS to vine during a probe, dimensionless
σ:	Number of probes a GWSS performs on vines per unit of time, dimensionless
ν:	Rate of progression from latent to symptomatic vine, time^{-1}
γ:	Rate of recovery for latent vines, time^{-1}

Table 4.2 – Parameters

ψ_v:	0.32, 2.1 eggs per female per day, 30% of which survive
ψ_h:	1/365, giving an average replacement time of 365 days
N_{h0}:	10000
μ_{v1}:	0.01, giving an average lifetime of 100 days
μ_{v2}:	$1.55 \cdot 10^{-5}$, which leads to $N_v^*/N_{h0} = 2$
μ_h:	$1.1 \cdot 10^{-4}$, giving an average lifetime of 25 years
μ_x:	1/180, giving an expected time to be removed once symptomatic of 180 days
β_{hv}:	0.2
β_{vh}:	0.35
σ:	1.5, vector performs 5 probes, 30% of which are on vines
ν:	1/120, average time of progressing 120 days
γ:	0.0033, giving a 28% chance of recovery once a vine has become latent

Table 4.3 – Parameter values

To facilitate analysis, we set

$$i_v = \frac{I_v}{N_v}; \quad s_v = 1 - i_v; \quad s_h = \frac{S_h}{N_{h0}}; \quad l_h = \frac{L_h}{N_{h0}}; \quad i_h = \frac{I_h}{N_{h0}}$$

to arrive at an equivalent system in terms of fractional quantities:

$$\frac{di_v}{dt} = \lambda_v(t)(1 - i_v) - \psi_v i_v \tag{4.6}$$

$$\frac{dN_v}{dt} = \psi_v N_v - f_v(N_v)N_v \tag{4.7}$$

$$\frac{ds_h}{dt} = \psi_h(1 - s_h - l_h - i_h) - \lambda_h(t)s_h + \gamma l_h - \mu_h s_h \tag{4.8}$$

$$\frac{dl_h}{dt} = \lambda_h(t)s_h - \gamma l_h - \nu l_h - \mu_h l_h \tag{4.9}$$

$$\frac{di_h}{dt} = \nu l_h - (\mu_h + \mu_x)i_h \tag{4.10}$$

4.2.2 Baseline parameter values

Baseline parameter values were either taken directly from or derived based on the literature, and chosen to reflect the conditions in Southern California. The values are given in Table 4.3. Because we only consider the adult phase of the life cycle of GWSS, the birth rate was adjusted to account for survival through the earlier life stages. During active oviposition, a female lays on average 2.1 eggs per day [164] and 50% of the females lay eggs every day [168]. Of these eggs, 30% survive [105]. This gives a *per capita* birth rate $\psi_v = 0.32$ per day for GWSSs.
The density-independent part of the GWSS death rate was set to $\mu_{v1} = 1/100$ per day to reflect the longevity reported in [105]. The equilibrium size of the GWSS population is controlled through the density-dependent part of the death rate. The baseline value was set to obtain a stable population that gives a vector density of approximately 2 GWSSs per vine (in the absence of PD). This was based on the GWSS densities reported in [102, 129, 130].
The pathogen transmission probabilities during a probe were set to $\beta_{hv} = 0.2$ for transmission from vine to GWSS, and to $\beta_{vh} = 0.35$ for transmission from GWSS to vine. Both probabilities were obtained from [10].
We combined information on the number of probes from [153], GWSS feeding preferences from [47, 122], and the impact of environmental conditions from [28] to set the number of probes a GWSS performs on vines per unit of time to $\sigma = 1.5$.

The PD-independent death and removal rate of vines was set to $\mu_h = 1.1 \cdot 10^{-4}$ per day to give an expected lifespan of 25 years, reflecting Californian viticulture practices [11]. Growers are advised to rogue symptomatic vines within a year from the appearance of symptoms [178], and so we set the PD-induced removal rate to $\mu_x = 1/180$ per day to give an expected removal time of 6 months.
The per capita replacement rate of (missing) vines was set to $\psi_h = 1/365$ per day so that on average it takes one year to replace a missing vine, in accordance with the literature [11, 48]. It takes on average 120 days for a latent vine to become symptomatic [111], and we set the rate of progression to $\nu = 1/120$ per day. The chance of recovery for latent vines was estimated to be 28% [113, 50], and so the rate of recovery was set to $\gamma = 0.0033$ per day accordingly.

4.2.3 Disease-free equilibrium

The total number of vectors N_v is not influenced by the disease dynamics. It can be calculated by

$$\frac{dN_v}{dt} = 0 \iff N_v = 0 \quad \text{or} \quad N_v^* := \frac{\psi_v - \mu_{v1}}{\mu_{v2}}.$$

Clearly the non-trivial equilibrium N_v^* exists if and only if $\psi_v > \mu_{v1}$, and is asymptotically stable.
For the vine, in the absence of the disease the only dynamics are

$$\frac{ds_h}{dt} = \psi_h(1 - s_h) - \mu_h s_h,$$

that is, the non-PD-related death and removal, and subsequent replacement. This gives the disease-free equilibrium for vine

$$s_h^* := \frac{\psi_h}{\psi_h + \mu_h},$$

which equals approximately 0.96 for the parameter values in Table 4.3.
Jointly, N_v^* and s_h^* define the disease-free equilibrium of the system

$$x_{df} = (0, N_v^*, s_h^*, 0, 0).$$

4.2.4 The basic reproduction number

The basic reproduction number R_0 is a metric to determine whether an infection will spread in a healthy population. There are different definitions and ways to calculate it, but in general it tries to approximate the number of secondary infections one infected individual will cause when it enters a fully susceptible population. Thus $R < 1$ would mean the disease will not establish itself.
We follow [53], and calculate the basic reproduction number using

$$R_0 = \sqrt{K_{vh} K_{hv}},$$

where K_{vh} and K_{hv} are respectively the number of vines one infected GWSS is expected to infect and the number of GWSSs one infected vine is expected to transmit the pathogen to, assuming a completely healthy population. The functions K_{vh} and K_{hv} are given by

$$K_{vh} = \frac{\sigma \beta_{vh}}{\mu_{v1} + \mu_{v2} N_v^*}$$
$$K_{hv} = \frac{\sigma \beta_{hv} N_v^*}{s_h^* N_{h0}} \frac{1}{\gamma + \nu + \mu_h} + \frac{\nu}{\gamma + \nu + \mu_h} \frac{\sigma \beta_{hv} N_v^*}{s_h^* N_{h0}} \frac{1}{\mu_h + \mu_x},$$

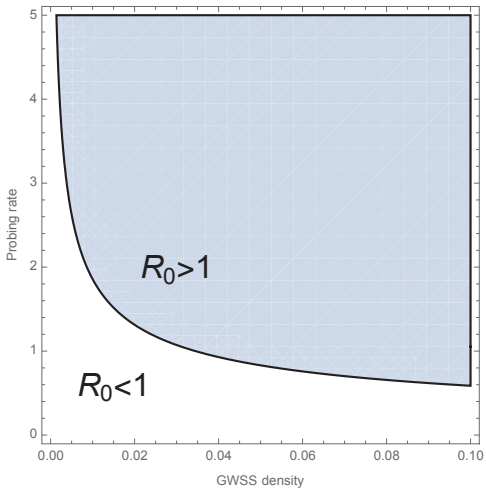


Figure 4.2 – The basic reproduction number R_0 indicates whether the disease can establish itself in a healthy population through an initial infection. Namely, when $R_0 < 1$, the disease-free equilibrium x_{df} is stable, and a small number of infected individuals will not cause endemic disease. The plot shows how the vector density N_v^*/N_{h0} and the vector probing rate affect R_0: the line is the isocline $R_0 = 1$. The rest of the parameters are as given in Table 4.3.

that is, for the GWSS it is the product of its expected lifetime and the rate at which it successfully inoculates a vine. For the vine, we first have the expected time spent as latent, times the number of successful transmissions to a GWSS. The second summand is the same for the symptomatic period, multiplied by the probability to transition from the latent stage to the symptomatic.
Figure 4.2 shows how the basic reproduction number depends on the vector density N_v^*/N_{h0} and the probing rate σ.

4.2.5 The endemic equilibria

An endemic equilibrium is a steady state solution in which the disease will persist in the system. Unfortunately we cannot solve these equilibria analytically. Both numerical simulations and previous results (see [37]) suggest the existence of a unique, asymptotically stable endemic equilibrium when $R_0 > 1$. We will not pursue a rigorous proof here.
The results of a numerical simulation of the system in Equations 4.1-4.5, using the baseline parameter values, are shown in Figure 4.3. We see that starting from a healthy vine population and a small number of infected GWSSs ($S_h = 10000, L_h = 0, I_v = 0, S_v = 18000, I_v = 2000$) the system converges to an endemic equilibrium where a significant proportion of both populations is infected.
Figure 4.4 shows how the disease prevalence depends on the vector density N_v^*/N_{h0}. We ran numerical simulations of the fractional system (Equations 4.6-4.10) at different vector densities, starting each run from a low initial disease prevalence ($i_v = 0.1, s_h = 1$) and until the system had converged to an equilibrium. We can see that PD starts to become endemic when the vector density exceeds 0.01.

4.2.6 Sensitivity analysis

In order to evaluate the significance of the model parameters, and thus estimate the potential of different control strategies, we performed a sensitivity analysis following the approach

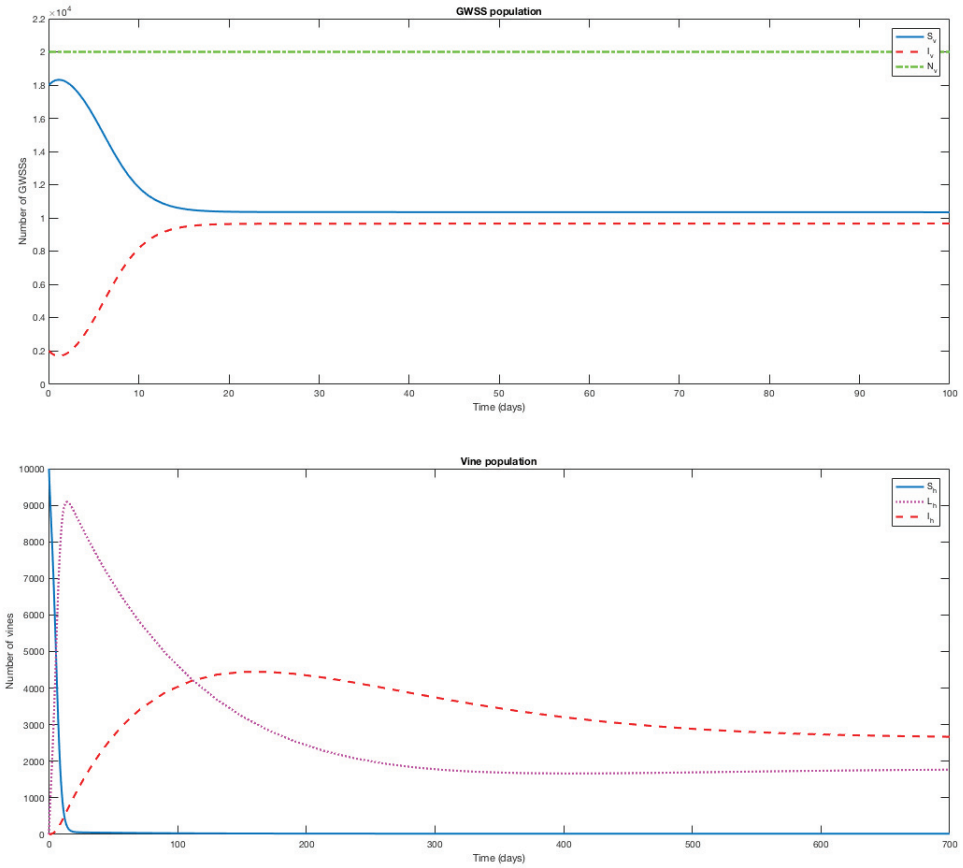


Figure 4.3 – A numerical simulation of the model with parameter values given in Table 4.3. The initial condition was $S_h = 10000, L_h = 0, I_v = 0, S_v = 18000, I_v = 2000$. Note the difference in time scales.

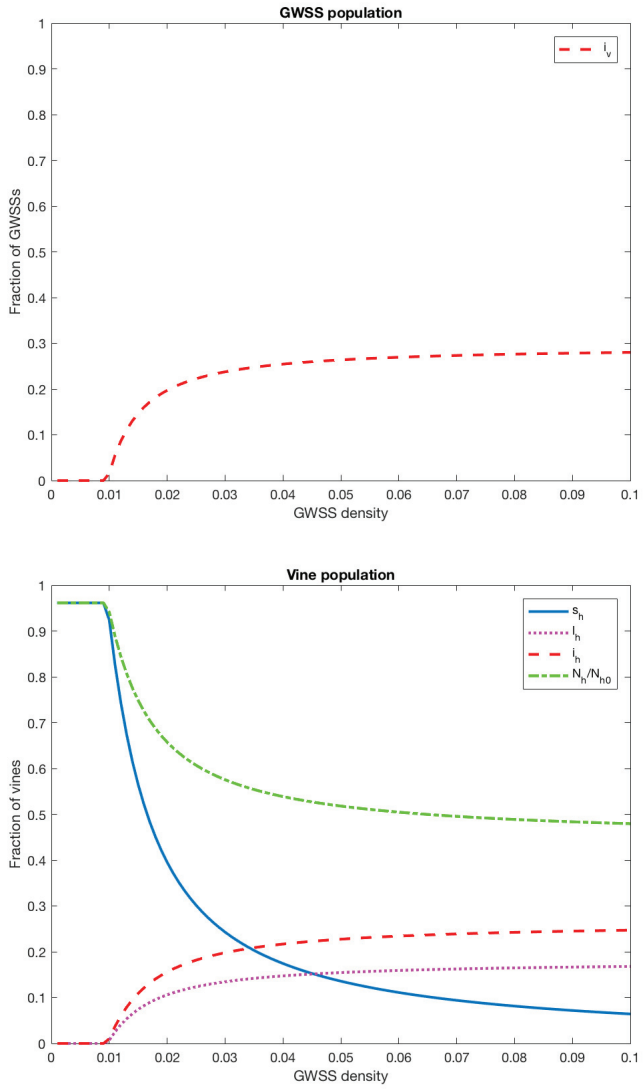


Figure 4.4 – Equilibrium prevalence of PD, as a function of the vector density N_v^*/N_{h0}.

	R_0
ψ_v	0.016
ψ_h	-0.019
N_{h0}	-0.50
μ_{v1}	-0.016
μ_{v2}	-0.50
μ_h	0.0088
μ_x	-0.29
β_{hv}	0.5
β_{vh}	0.5
σ	1
ν	-0.057
γ	-0.14

Table 4.4 – Sensitivity indices of the basic reproduction number R_0 to the parameters, evaluated using the baseline parameter values in Table 4.3.

	i_v	N_v	s_h	l_h	i_h
ψ_v	-0.77	1.0	-0.27	0.00079	0.0023
ψ_h	0.70	0.0	0.76	0.55	1.6
N_{h0}	-0.0038	0.0	1.0	-0.0030	-0.0088
μ_{v1}	-0.00012	-0.032	0.032	-0.000096	-0.00028
μ_{v2}	-0.0038	-1.0	1.0	-0.0030	-0.0088
μ_h	-0.012	0.0	0.00025	-0.0056	-0.036
μ_x	-0.014	0.0	0.048	0.088	-0.72
β_{hv}	0.77	0.0	-0.77	0.0023	0.0067
β_{vh}	0.0038	0.0	-1.0	0.0030	0.0088
σ	0.77	0.0	-1.8	0.0052	0.016
ν	-0.27	0.0	-0.18	-0.81	0.56
γ	-0.0011	0.0	0.28	-0.00083	-0.0025

Table 4.5 – Sensitivity indices of the endemic equilibrium to the model parameters, evaluated using the baseline parameter values in Table 4.3.

detailed in [21]. The normalised sensitivity index of an output u to a parameter p is defined as

$$S_{u_p} := \frac{p}{u}\frac{\partial u}{\partial p}.$$

It measures how the output u changes when small perturbations are made to the value of p. We calculated the sensitivity index for the basic reproductive number R_0 and the endemic equilibrium. The results are given in Tables 4.4 and 4.5. For R_0, since we have an analytical expression for it, the sensitivity indices can be calculated directly. For the endemic equilibrium, we first solved a linear system for each parameter to find an expression for S_{u_p}, and then evaluated it at the numerically solved equilibrium values of the state variables.
The parameters with the most positive correlation with the basic reproduction number are the number of probes a GWSS performs on a vine per unit of time (σ), and the two transmission probabilities (β_{hv} and β_{vh}), meaning that a decrease in any of them would also significantly decrease the value of R_0. Based on this, strategies related to these parameters could be important for curbing the emergence of the disease. Strategies related to σ are those that restrict

GWSS access to vines (for example, screen barriers protecting vineyards) and strategies that make grapevine a less desirable food source for the insect (for example, coating the vines with products that deter the GWSS). Strategies related to the transmission probabilities include releasing major GWSS predators (this reduces the time an insect will spend feeding on a vine) and prophylactic strategies that aim at increasing resistance to the *X. fastidiosa* pathogen. The most negatively correlated parameters, the vineyard size (N_{h0}) and the density-dependent part of the GWSS death rate (μ_{v2}) are both related to simply reducing the GWSS density in the vineyard.
With regards to the endemic equilibrium, it should be pointed out that due to the way the vine population was modelled, some of the results may appear counter-intuitive: because the fractional quantities s_h, l_h, and i_h are all defined with respect to the maximum number of vines N_{h0}, they do not sum to one. For this reason, it is possible for the sensitivity index of a parameter to have the same sign for all three variables. This is the case for the *per capita* replacement rate of vines (ψ_h) which correlates positively with all the three state variables. The simple explanation is that increasing the rate at which vines are being replaced increases the overall occupation level of the vineyard.
Overall, it appears that the parameters with the most relevance for preventing and controlling PD are σ, β_{hv}, β_{vh}, and μ_x. This suggests that in areas where PD has not become endemic, efforts should be concentrated at reducing GWSS preference for vine and the insect access to vines. Where the disease has already been established, regular inspection and removal of infected vines may help control the levels of disease.

4.3 Conclusion

In this chapter, I presented an epidemiological model of Pierce's disease, a grapevine disease caused by the bacterium *X. fastidiosa*. The model is to my knowledge the first fully realised one to be published. Baseline parameter values were extracted from the literature to describe conditions of high disease pressure in Southern California. I provided an analytical expression for the disease-free equilibrium and the basic reproduction number R_0. Endemic equilibria were studied through numerical simulations, leading to the conclusion that when $R_0 < 1$, the disease-free equilibrium is asymptotically stable, and when $R_0 > 1$, it becomes unstable and a unique, asymptotically stable endemic equilibrium emerges.
Sensitivity analysis of the basic reproduction number and the endemic equilibrium were performed for the model parameters to evaluate the potential of different disease control strategies. This led to the conclusion that the most promising approaches to disease prevention and control are those targeted at reducing GWSS preference for and access to vines, and increasing detection and removal of infected plants.

Conclusion and Perspectives

In the Introduction, I offered two ways to summarise my PhD project: metabolic networks and wine. In the end, perhaps the best characterisation would be mathematical modelling. In this thesis, I have covered three different topics, two more closely related and one that appears – at least at first glance – more distant. The underlying principle, however, is the same: capturing nature into equations.

In Chapter 2, I presented the MOOMIN algorithm. It is a computational method that produces a hypothesis of a metabolic shift using a metabolic network and the results of a differential expression analysis. MOOMIN was implemented in MATLAB and the code is available at `github.com/htpusa/moomin`. The algorithm relies on the assumption of a metabolic steady state: zero net production of all internal metabolites. I presented two different ways to enforce this assumption: one based solely on the topology of the hypergraph representation of the network, and one that takes fully into account the reaction stoichiometry. The former can be seen as an approximation of the latter. In the centre of the MOOMIN method is an optimisation problem that aims to find the most likely metabolic shift given the gene expression data. I proved that this problem is NP-hard and showed how it can be solved using Mixed-Integer Linear Programming (MILP). The finalised MOOMIN method was applied to two real data sets and the results were compared to those presented in the original accompanying publications. We found that not only was MOOMIN able to replicate the previous findings but it also inferred changes that were expected but not found in the gene expression data alone.

In addition to finding *an* optimum, MILP can be used to enumerate all optimal solutions. Based on the examples presented in Chapter 2, it appears that there can be a multitude of alternative optima (more than 1000 were found in all but one of the four cases). This poses a potential problem: while a single optimum can be obtained in a reasonable time on a desktop computer (at worst in a matter of minutes), finding each additional alternative optimum usually takes at least as long. Thus if there are possibly thousands of different optima, an exhaustive enumeration can be a lengthy procedure. The examples in Chapter 2 seem to indicate that the alternative optima do not present qualitatively distinct solutions, but rather a small subset of included reactions are responsible for their number. Hence in practice it might not be necessary to attempt full enumeration. Indeed, all of the biological results in Chapter 2 were obtained by simply looking at the first solution found. Nevertheless, the possibility of enumerating all solutions is important since the existence of biologically relevant alternative solutions cannot be discarded. To make this procedure more efficient, it would be beneficial to develop a way to group alternative optima into equivalence classes to avoid searching for trivially differing solutions. This might also have implications for other similar methods.

The inspiration for MOOMIN came from previous methods developed in the team that deal with metabolomics data. Similarly to comparing the gene expression levels as is done in differential expression analysis, metabolite concentrations can also be compared to gain knowledge about changes in metabolism. These two different 'omics data are obviously closely related, and so an interesting question is if they can be combined in a framework similar to MOOMIN.

Ideally, we would have used data from the Microwine project to test MOOMIN. Unfortunately none was available in time for this thesis. We are, however, currently working on two data sets obtained through the Microwine network: one with Ahmad Zeidan at the industrial partner Chr. Hansen, another with Witold Kot at the Aarhus University.

In Chapter 3, I explored the application of evolutionary game theory to microbial metabolism. I reviewed the literature on the topic, with a focus on studies directly related to metabolic networks. I then presented the idea of a metabolic game: a game theoretical model defined based on one or several metabolic reconstructions. Such a model can be used similarly to methods such as Flux Balance Analysis (FBA) to predict metabolic behaviour. Predicted phenotypes are found among the solutions of the game, rather than by simple optimisation as in FBA. With regard to how the game is actually defined, several approaches have been proposed and adopted. In examples put forth by Schuster and Pfeiffer [143, 142, 161, 160, 157], the different available actions were taken to be alternative pathways. In contrast, in a more recent realisation of the idea, Zomorrodi and Segrè's [196] actions were defined in terms of secreted molecules. While both approaches seem to be valid options, the latter might be more suitable for use in a computational framework. Ideally, payoffs would be obtained using constraint-based methods. However, when measures such as the flux through the biomass reaction are used, attention should be paid to distinguishing between growth yield and growth rate.

It remains to be seen whether the metabolic game can be formulated into a method of phenotype prediction that would match the simplicity of FBA. I believe that this development would greatly benefit from suitable model organisms.

In Chapter 4, I presented an epidemiological model of the *Xylella fastidiosa* grapevine pathogen. This was based on previously described models of malaria and defined as a system of ordinary differential equations. I derived expressions for the disease-free equilibrium and the basic reproduction number, and presented the results of numerical simulations exploring the endemic equilibria. The main motivation for developing the model was to assess the potential of different control strategies. To this end, I performed a sensitivity analysis of the model parameters for the basic reproduction number and the endemic equilibrium. This led us to conclude that the most promising approaches to disease control are those related to reducing how much the insect vector feeds on grapevine and to regular inspection and removal of infected vines.

There are several potential ways to extend the model presented in Chapter 4. The first is the inclusion of seasonality: for the sake of simplicity, we modelled most processes with constant rates. However, many factors of the system are affected by the yearly cycle, and this can potentially influence the disease dynamics. Secondly, we did not consider spatial structure. In reality, factors such as insect swarming behaviour and the topology of the vineyard can influence the spread of the infection. Including the spatial component in the model could be especially helpful in assessing and designing control strategies related to physical blocking, such as screen barriers. Lastly, we based our model on the conditions in Southern California and used baseline parameter values obtained from there. It would thus be interesting to see how accurate the model is for other locales.

Bibliography

[1] Vicente Acuña, Etienne Birmelé, Ludovic Cottret, Pierluigi Crescenzi, Fabien Jourdan, Vincent Lacroix, Alberto Marchetti-Spaccamela, Andrea Marino, Paulo Vieira Milreu, Marie-France Sagot, et al. Telling stories: Enumerating maximal directed acyclic graphs with a constrained set of sources and targets. *Theoretical Computer Science*, 457:1–9, 2012.

[2] Vicente Acuna, Flavio Chierichetti, Vincent Lacroix, Alberto Marchetti-Spaccamela, Marie-France Sagot, and Leen Stougie. Modes and cuts in metabolic networks: Complexity and algorithms. *Biosystems*, 95(1):51–60, 2009.

[3] J. Adkins, S. Pugh, R. McKenna, and D. R. Nielsen. Engineering microbial chemical factories to produce renewable "biomonomers". *Front Microbiol*, 3:313, 2012.

[4] Rasmus Agren, Sergio Bordel, Adil Mardinoglu, Natapol Pornputtapong, Intawat Nookaew, and Jens Nielsen. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. *PLoS computational biology*, 8(5):e1002518, 2012.

[5] Mats Åkesson, Jochen Förster, and Jens Nielsen. Integration of gene expression data into genome-scale metabolic models. *Metabolic engineering*, 6(4):285–293, 2004.

[6] Nicolas Alcaraz, Tobias Friedrich, Timo Kötzing, Anton Krohmer, Joachim Müller, Josch Pauling, and Jan Baumbach. Efficient key pathway mining: combining networks and omics data. *Integrative Biology*, 4(7):756–764, 2012.

[7] Nicolas Alcaraz, Hande Kücük, Jochen Weile, Anil Wipat, and Jan Baumbach. Keypathwayminer: detecting case-specific biological pathways using expression data. *Internet Mathematics*, 7(4):299–313, 2011.

[8] Juan Carlos Aledo, Juan A Pérez-Claros, and Alicia Esteban Del Valle. Switching between cooperation and competition in the use of extracellular glucose. *Journal of molecular evolution*, 65(3):328–339, 2007.

[9] B. Allen, J. Gore, and M. A. Nowak. Spatial dilemmas of diffusible public goods. *Elife*, 2:e01169, Dec 2013.

[10] R. P. Almeida and A. H. Purcell. Transmission of xylella fastidiosa to grapevines by homalodisca coagulata (hemiptera: Cicadellidae). *Journal of economic entomology*, 96(2):264–271, 2003.

[11] J. M. Alston, K. B. Fuller, J. D. Kaplan, and K. P. Tumber. Economic consequences of pierce's disease and related policy in the california winegrape industry. *Journal of Agricultural and Resource Economics*, 2013.

[12] Maciek R Antoniewicz. Methods and advances in metabolic flux analysis: a mini-review. *Journal of industrial microbiology & biotechnology*, 42(3):317–325, 2015.

[13] Alexey V Antonov, Hans W Mewes, and Sabine Dietmann. Kegg spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome biology*, 9(12):R179, 2008.

[14] Marco Archetti. Evolutionary game theory of growth factor production: implications for tumour heterogeneity and resistance to therapies. *British journal of cancer*, 109(4):1056–1062, 2013.

[15] Marco Archetti. Evolutionary dynamics of the warburg effect: glycolysis as a collective action problem among cancer cells. *Journal of theoretical biology*, 341:1–8, 2014.

[16] Marco Archetti. Heterogeneity and proliferation of invasive cancer subclones in game theory models of the warburg effect. *Cell proliferation*, 48(2):259–269, 2015.

[17] Marco Archetti. Cooperation among cancer cells as public goods games on voronoi networks. *Journal of Theoretical Biology*, 396:191 – 203, 2016.

[18] Marco Archetti. How to analyze models of nonlinear public goods. *Games*, 9(2):17, 2018.

[19] Marco Archetti and Istvan Scheuring. Review: Game theory of public goods in one-shot social dilemmas without assortment. *Journal of Theoretical Biology*, 299:9–20, 2012.

[20] Marco Archetti and Istvan Scheuring. Trading public goods stabilizes interspecific mutualism. *Journal of theoretical biology*, 318:58–67, 2013.

[21] L. Arriola and J. M. Hyman. Sensitivity analysis for uncertainty quantification in mathematical models. *In Mathematical and Statistical Estimation Approaches in Epidemiology*, 2009.

[22] David Avis, Gabriel D Rosenberg, Rahul Savani, and Bernhard Von Stengel. Enumeration of nash equilibria for two-player games. *Economic Theory*, 42(1):9–37, 2010.

[23] Robert Axelrod. Effective choice in the prisoner's dilemma. *Journal of conflict resolution*, 24(1):3–25, 1980.

[24] Jan Baumbach, Tobias Friedrich, Timo Kötzing, Anton Krohmer, Joachim Müller, and Josch Pauling. Efficient algorithms for extracting biological key pathways with global constraints. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*, pages 169–176. ACM, 2012.

[25] Scott A Becker and Bernhard O Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS computational biology*, 4(5):e1000082, 2008.

[26] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[27] Hans C Bernstein and Ross P Carlson. Microbial consortia engineering for cellular factories: in vitro to in silico systems. *Computational and structural biotechnology journal*, 3(4):e201210017, 2012.

[28] M.J. Blua and D.J.W. Morgan. Dispersion of homalodisca coagulata (hemiptera: Cicadellidae), a vector of xylella fastidiosa, into vineyards in southern california. *Journal of economic entomology*, 96(5):1369–1374, 2003.

[29] K. Bohl, S. Hummert, S. Werner, D. Basanta, A. Deutsch, S. Schuster, G. Theissen, and A. Schroeter. Evolutionary game theory: molecules as players. *Mol Biosyst*, 10(12):3066–3074, Dec 2014.

[30] Piero Bonatti, Francesco Calimeri, Nicola Leone, and Francesco Ricca. Answer set programming. In *A 25-year perspective on logic programming*, pages 159–182. Springer-Verlag, 2010.

[31] Hans J Bremermann and John Pickering. A game-theoretical model of parasite virulence. *Journal of Theoretical Biology*, 100(3):411–426, 1983.

[32] Mark Broom and Jan Rychtář. *Game-theoretical models in biology*. CRC Press, 2013.

[33] K. Campbell, J. Vowinckel, M. A. Keller, and M. Ralser. Methionine Metabolism Alters Oxidative Stress Resistance via the Pentose Phosphate Pathway. *Antioxid. Redox Signal.*, 24(10):543–547, Apr 2016.

[34] Kate Campbell, Jakob Vowinckel, Michael Mülleder, Silke Malmsheimer, Nicola Lawrence, Enrica Calvani, Leonor Miller-Fleming, Mohammad T Alam, Stefan Christen, Markus A Keller, et al. Self-establishing communities enable cooperative metabolite exchange in a eukaryote. *Elife*, 4:e09943, 2015.

[35] Sriram Chandrasekaran and Nathan D Price. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, 107(41):17845–17850, 2010.

[36] L. Chao and S. F. Elena. Nonlinear trade-offs allow the cooperation game to evolve from Prisoner's Dilemma to Snowdrift. *Proc. Biol. Sci.*, 284(1854), May 2017.

[37] N. R. Chitnis. Using mathematical models in controlling the spread of malaria. *[dissertation]: University of Arizona*, 2005.

[38] Bruce L Clarke. Stability of complex reaction networks. *Advances in chemical physics*, pages 1–215, 1980.

[39] Bruce L Clarke. Stoichiometric network analysis. *Cell biophysics*, 12(1):237–253, 1988.

[40] Caroline Colijn, Aaron Brandes, Jeremy Zucker, Desmond S Lun, Brian Weiner, Maha R Farhat, Tan-Yun Cheng, D Branch Moody, Megan Murray, and James E Galagan. Interpreting expression data with metabolic flux models: predicting mycobacterium tuberculosis mycolic acid production. *PLoS computational biology*, 5(8):e1000489, 2009.

[41] Sara B Collins, Ed Reznik, and Daniel Segrè. Temporal expression-based analysis of metabolism. *PLoS computational biology*, 8(11):e1002781, 2012.

[42] Otto X Cordero, Laure-Anne Ventouras, Edward F DeLong, and Martin F Polz. Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proceedings of the National Academy of Sciences*, 109(49):20059–20064, 2012.

[43] Markus W Covert and Bernhard O Palsson. Constraints-based models: regulation of gene expression reduces the steady-state solution space. *Journal of theoretical biology*, 221(3):309–325, 2003.

[44] Ross Cressman. *Evolutionary dynamics and extensive form games*, volume 5. MIT Press, 2003.

[45] Francis HC Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[46] Didier Croes, Fabian Couche, Shoshana J Wodak, and Jacques Van Helden. Metabolic pathfinding: inferring relevant pathways in biochemical networks. *Nucleic acids research*, 33(suppl_2):W326–W330, 2005.

[47] K. M. Daane, M. W. Johnson, K. A. Center, and G. Yokota. Biology and ecology of the glassy-winged sharpshooter in the san joaquin valley. *In Proceedings*, Pierce's Disease Research Symposium:M. Athar Tariq, 2003.

[48] C. H. Daniels, L. Sosnoskie, T. Miller, D. Walsh, G. Hoheise, and I. Zasada. 2018 pest management guide for grapes in washington. *WSU Extension*, Washington State University:Pullman, 2017.

[49] George B Dantzig and Mukund N Thapa. *Linear programming 1: introduction*. Springer Science & Business Media, 2006.

[50] M. P. Daugherty, R. Almeida, R. J. Smith, E. D. Weber, and A. H. Purcell. Severe pruning of infected grapevines has limited efficacy for managing pierce's disease. *American Journal of Enology and Viticulture*, 2018.

[51] Luis F De Figueiredo, Adam Podhorski, Angel Rubio, Christoph Kaleta, John E Beasley, Stefan Schuster, and Francisco J Planes. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, 25(23):3158–3165, 2009.

[52] Ulf Dieckmann and Richard Law. The dynamical theory of coevolution: a derivation from stochastic ecological processes. *Journal of mathematical biology*, 34(5-6):579–612, 1996.

[53] O. Diekmann, J. Heesterbeek, and M. Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, 2009.

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

[55] Marco Dorigo, Mauro Birattari, Christian Blum, Maurice Clerc, Thomas Stützle, and Alan Winfield. *Ant Colony Optimization and Swarm Intelligence: 6th International Conference, ANTS 2008, Brussels, Belgium, September 22-24, 2008, Proceedings*, volume 5217. Springer, 2008.

[56] Jeremy S Edwards, Rafael U Ibarra, and Bernhard O Palsson. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125–130, 2001.

[57] Jeremy S Edwards and Bernhard O Palsson. Metabolic flux balance analysis and the in silico analysis of escherichia coli k-12 gene deletions. *BMC bioinformatics*, 1(1):1, 2000.

[58] Jeremy S Edwards, Ramprasad Ramakrishna, and Bernhard O Palsson. Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnology and bioengineering*, 77(1):27–36, 2002.

[59] Peter J. Enyeart, Zachary B. Simpson, and Andrew D. Ellington. A microbial model of economic trading and comparative advantage. *Journal of Theoretical Biology*, 364:326–343, 1 2015.

[60] Semidán Robaina Estévez and Zoran Nikoloski. Context-specific metabolic model extraction based on regularized least squares optimization. *PloS one*, 10(7):e0131875, 2015.

[61] Sandeepa M Eswarappa. Location of pathogenic bacteria during persistent infections: insights from an analysis using game theory. *PLoS One*, 4(4):e5383, 2009.

[62] Xin Fang, Anders Wallqvist, and Jaques Reifman. Modeling phenotypic metabolic adaptations of mycobacterium tuberculosis h37rv under hypoxia. *PLoS computational biology*, 8(9):e1002688, 2012.

[63] J. W. Foster. Escherichia coli acid resistance: tales of an amateur acidophile. *Nat. Rev. Microbiol.*, 2(11):898–907, Nov 2004.

[64] Tobias Frick and Stefan Schuster. An example of the prisoner's dilemma in biochemistry. *Naturwissenschaften*, 90(7):327–331, 2003.

[65] Maren L Friesen, Gerda Saxer, Michael Travisano, and Michael Doebeli. Experimental evidence for sympatric ecological diversification due to frequency-dependent competition in escherichia coli. *Evolution*, 58(2):245–260, 2004.

[66] LC Galvez, K Korus, J Fernandez, JL Behn, and N Banjara. The threat of pierce's disease to midwest wine and table grapes. *Online. APSnet Features. doi*, 10, 2010.

[67] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, Max Ostrowski, Torsten Schaub, and Philipp Wanko. Theory solving made easy with clingo 5. In *OASIcs-OpenAccess Series in Informatics*, volume 52. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[68] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Clingo= asp+ control: Preliminary report. *arXiv preprint arXiv:1405.3694*, 2014.

[69] S. A. H. Geritz, E. Kisdi, G. Meszena, and J. A. J. Metz. Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *EVOL. ECOL*, 12:35–57, 1998.

[70] Matthias P Gerstl, Stefan Müller, Georg Regensburger, and Jürgen Zanghellini. Flux tope analysis: studying the coordination of reaction directions in metabolic networks. *Bioinformatics*, 2018.

[71] Jeff Gore, Hyun Youk, and Alexander Van Oudenaarden. Snowdrift game dynamics and facultative cheating in yeast. *Nature*, 459(7244):253–256, 2009.

[72] Willi Gottstein, Brett G Olivier, Frank J Bruggeman, and Bas Teusink. Constraint-based stoichiometric modelling from single organisms to microbial communities. *Journal of The Royal Society Interface*, 13(124):20160627, 2016.

[73] John Haigh. Game theory and evolution. *Advances in applied probability*, 7(1):8–11, 1975.

[74] Henry Hamburger. N-person prisoner's dilemma†. *Journal of Mathematical Sociology*, 3(1):27–48, 1973.

[75] Christoph Hauert, Miranda Holmes, and Michael Doebeli. Evolutionary games and population dynamics: maintenance of cooperation in public goods games. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1600):2565–2571, 2006.

[76] David Healey, Kevin Axelrod, and Jeff Gore. Negative frequency-dependent interactions can underlie phenotypic heterogeneity in a clonal microbial population. *Molecular systems biology*, 12(8):877, 2016.

[77] Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastian N Mendoza, Anne Richelle, Almut Heinken, Hulda S Haraldsdottir, Sarah M Keating, Vanja Vlasov, Jacek Wachowiak, et al. Creation and analysis of biochemical constraint-based models: the cobra toolbox v3. 0. *arXiv preprint arXiv:1710.04038*, 2017.

[78] Bradley Hersh, F.T. Farooq, D.N. Barstad, D.L. Blankenhorn, and J.L. Slonczewski. A glutamate-dependent acid resistance gene in escherichia coli. *Journal of bacteriology*, 178:3978–81, 07 1996.

[79] Josef Hofbauer and Karl Sigmund. *Evolutionary games and population dynamics*. Cambridge university press, 1998.

[80] Michael Hucka, Andrew Finney, Herbert M Sauro, Hamid Bolouri, John C Doyle, Hiroaki Kitano, Adam P Arkin, Benjamin J Bornstein, Dennis Bray, Athel Cornish-Bowden, et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[81] Sabine Hummert, Katrin Bohl, David Basanta, Andreas Deutsch, Sarah Werner, Günter Theißen, Anja Schroeter, and Stefan Schuster. Evolutionary game theory: cells as players. *Molecular BioSystems*, 10(12):3044–3065, 2014.

[82] J. D. Janse and A. Obradovic. Xylella fastidiosa: its biology, diagnosis, control and risks. *Journal of Plant Pathology*, 2010.

[83] Paul A Jensen and Jason A Papin. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics*, 27(4):541–547, 2010.

[84] Livnat Jerby, Tomer Shlomi, and Eytan Ruppin. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Molecular systems biology*, 6(1):401, 2010.

[85] Alice Julien-Laferriere. *Models and algorithms applied to metabolism: From revealing the responses to perturbations towards the design of microbial consortia*. PhD thesis, Université de Lyon, 2016.

[86] Alice Julien-Laferrière, Laurent Bulteau, Delphine Parrot, Alberto Marchetti-Spaccamela, Leen Stougie, Susana Vinga, Arnaud Mary, and Marie-France Sagot. A combinatorial algorithm for microbial consortia synthetic design. *Scientific Reports*, 6:29182, 2016.

[87] Irina Kareva. Prisoner's dilemma in cancer metabolism. *PloS one*, 6(12):e28576, 2011.

[88] Richard M. Karp. *Reducibility among Combinatorial Problems*, pages 85–103. Springer US, Boston, MA, 1972.

[89] E. D. Kelsic, J. Zhao, K. Vetsigian, and R. Kishony. Counteraction of antibiotic production and degradation stabilizes microbial communities. *Nature*, 521(7553):516–519, May 2015.

[90] B. Kerr, M. A. Riley, M. W. Feldman, and B. J. Bohannan. Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature*, 418(6894):171–174, Jul 2002.

[91] Ardeshir Kianercy, Robert Veltri, and Kenneth J Pienta. Critical transitions in a game theoretic model of tumour metabolism. *Interface focus*, 4(4):20140014, 2014.

[92] H. U. Kim, S. Y. Kim, H. Jeong, T. Y. Kim, J. J. Kim, H. E. Choy, K. Y. Yi, J. H. Rhee, and S. Y. Lee. Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. *Mol. Syst. Biol.*, 7:460, Jan 2011.

[93] Hyun Uk Kim, Won Jun Kim, and Sang Yup Lee. Flux-coupled genes and their use in metabolic flux analysis. *Biotechnology journal*, 8(9):1035–1042, 2013.

[94] Joonhoon Kim and Jennifer L Reed. Relatch: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. *Genome biology*, 13(9):R78, 2012.

[95] Min Kyung Kim, Anatoliy Lane, James J Kelley, and Desmond S Lun. E-flux2 and spot: validated methods for inferring intracellular metabolic flux distributions from transcriptomic data. *PloS one*, 11(6):e0157101, 2016.

[96] Min Kyung Kim and Desmond S Lun. Methods for integration of transcriptomic data in genome-scale metabolic models. *Computational and structural biotechnology journal*, 11(18):59–65, 2014.

[97] Z. A. King, A. Drager, A. Ebrahim, N. Sonnenschein, N. E. Lewis, and B. O. Palsson. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput. Biol.*, 11(8):e1004321, Aug 2015.

[98] Benjamin C Kirkup and Margaret A Riley. Antibiotic-mediated antagonism leads to a bacterial game of rock–paper–scissors in vivo. *Nature*, 428(6981):412–414, 2004.

[99] Steffen Klamt and Jörg Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Molecular biology reports*, 29(1-2):233–236, 2002.

[100] Niels Klitgord and Daniel Segrè. Environments that induce synthetic microbial ecosystems. *PLoS computational biology*, 6(11):e1001002, 2010.

[101] J. U. Kreft. Biofilms promote altruism. *Microbiology (Reading, Engl.)*, 150(Pt 8):2751–2760, Aug 2004.

[102] Rodrigo Krugner, James R Hagler, Russell L Groves, Mark S Sisterson, Joseph G Morse, and Marshall W Johnson. Plant water stress effects on the net dispersal rate of the insect vector homalodisca vitripennis (hemiptera: Cicadellidae) and movement of its egg parasitoid, gonatocerus ashmeadi (hymenoptera: Mymaridae). *Environmental entomology*, 41(6):1279–1289, 2012.

[103] Ifigeneia Kyrkou, Taneli Pusa, Lea Ellegaard-Jensen, Marie-France Sagot, and Lars Hestbjerg Hansen. Pierce's disease of grapevines: A review of control strategies and an outline of an epidemiological model. *Frontiers in microbiology*, 9, 2018.

[104] Vincent Lacroix, Ludovic Cottret, Patricia Thébault, and Marie-France Sagot. An introduction to metabolic networks and their structural analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 5(4):594–617, 2008.

[105] I. Lauzière and M. Sétamou. Life history studies of homalodisca vitripennis (hemiptera: Cicadellidae), a vector of pierce's disease of grapevine. *Annals of the Entomological Society of America*, 103(1):57–65, 2010.

[106] Stephen P LaVoie and Anne O Summers. Transcriptional responses of escherichia coli during recovery from inorganic or organic mercury exposure. *BMC genomics*, 19(1):52, 2018.

[107] David C Lay. Linear algebra and its applications, 1997, 1997.

[108] Dave Lee, Kieran Smallbone, Warwick B Dunn, Ettore Murabito, Catherine L Winder, Douglas B Kell, Pedro Mendes, and Neil Swainston. Improving metabolic flux predictions using absolute gene expression data. *BMC systems biology*, 6(1):73, 2012.

[109] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29(8):1035–1043, Apr 2013.

[110] Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, et al. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, 6(1):390, 2010.

[111] W.B. Li, C.H. Zhou, W.D. Pria Jr, D.C. Teixeira, V.S. Miranda, E.O. Pereira, A.J. Ayres, C.X. He, P.I. Costa, and J.S. Hartung. Citrus and coffee strains of xylella fastidiosa induce pierce's disease in grapevine. *Plant disease*, 86(11):pp.1206–1210, 2002.

[112] X. Y. Li, C. Pietschke, S. Fraune, P. M. Altrock, T. C. Bosch, and A. Traulsen. Which games are growing bacterial populations playing? *J R Soc Interface*, 12(108):20150121, Jul 2015.

[113] J. H. Lieth, M. M. Meyer, K. H. Yeo, and B. C. Kirkpatrick. Modeling cold curing of pierce's disease in vitis vinifera 'pinot noir' and 'cabernet sauvignon' grapevines in california. *Phytopathology*, 101(12):1492–1500, 2011.

[114] William Forster Lloyd. *Two Lectures on the Checks to Population: Delivered Before the University of Oxford, in Michaelmas Term 1832*. JH Parker, 1833.

[115] Daniel Machado and Markus Herrgård. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS computational biology*, 10(4):e1003580, 2014.

[116] R Craig MacLean and Ivana Gudelj. Resource competition and social conflict in experimental populations of yeast. *Nature*, 441(7092):498, 2006.

[117] R Mahadevan and CH Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264–276, 2003.

[118] Susana Martín and Santiago F Elena. Application of game theory to the interaction between plant viruses during mixed infections. *Journal of general virology*, 90(11):2815–2820, 2009.

[119] Douglas McCloskey, Bernhard Ø Palsson, and Adam M Feist. Basic and applied uses of genome-scale metabolic network reconstructions of escherichia coli. *Molecular systems biology*, 9(1):661, 2013.

[120] Johan AJ Metz, Stefan AH Geritz, Géza Meszéna, Frans JA Jacobs, Joost S Van Heerwaarden, et al. Adaptive dynamics, a geometrical study of the consequences of nearly faithful reproduction. *Stochastic and spatial structures of dynamical systems*, 45:183–231, 1996.

[121] Paulo Vieira Milreu, Cecilia Coimbra Klein, Ludovic Cottret, Vicente Acuña, Etienne Birmelé, Michele Borassi, Christophe Junot, Alberto Marchetti-Spaccamela, Andrea Marino, Leen Stougie, et al. Telling metabolic stories to explore metabolomics data: a case study on the yeast response to cadmium exposure. *Bioinformatics*, 30(1):61–70, 2014.

[122] RF Mizell III, C Tipping, PC Andersen, BV Brodbeck, WB Hunter, and T Northfield. Behavioral model for homalodisca vitripennis (hemiptera: Cicadellidae): optimization of host plant utilization and management implications. *Environmental entomology*, 37(5):1049–1062, 2008.

[123] M. L. Mo, B. O. Palsson, and M. J. Herrgard. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol*, 3:37, Mar 2009.

[124] Babak Momeni, Adam James Waite, and Wenying Shou. Spatial self-organization favors heterotypic cooperation over cheating. *Elife*, 2:e00960, 2013.

[125] Uzi Motro. Co-operation and defection: playing the field and the ess. *Journal of Theoretical Biology*, 151(2):145–154, 1991.

[126] Joel F Moxley, Michael C Jewett, Maciek R Antoniewicz, Silas G Villas-Boas, Hal Alper, Robert T Wheeler, Lily Tong, Alan G Hinnebusch, Trey Ideker, Jens Nielsen, et al. Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator gcn4p. *Proceedings of the National Academy of Sciences*, 106(16):6477–6482, 2009.

[127] John Nash. Non-cooperative games. *Annals of Mathematics*, pages 286–295, 1951.

[128] John F Nash et al. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

[129] NASS/CDFA. 2016 citrus acreage report, 2017.

[130] NASS/CDFA. 2016 grape acreage report, 2017.

[131] Ali Navid and Eivind Almaas. Genome-level transcription data of yersinia pestis analyzed with a new metabolic constraint-based approach. *BMC systems biology*, 6(1):150, 2012.

[132] Josselin Noirel, Saw Yen Ow, Guido Sanguinetti, Alfonso Jaramillo, and Phillip C Wright. Automated extraction of meaningful pathways from quantitative proteomics data. *Briefings in Functional Genomics and Proteomics*, 7(2):136–146, 2008.

[133] I. Nookaew, M. Papini, N. Pornputtapong, G. Scalcinati, L. Fagerberg, M. Uhlen, and J. Nielsen. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res.*, 40(20):10084–10097, Nov 2012.

[134] Martin A Nowak and Karl Sigmund. Phage-lift for game theory. *Nature*, 398(6726):367–368, 1999.

[135] Sjoerd Opdam, Anne Richelle, Benjamin Kellman, Shanzhong Li, Daniel C Zielinski, and Nathan E Lewis. A systematic evaluation of methods for tailoring genome-scale metabolic models. *Cell systems*, 4(3):318–329, 2017.

[136] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. ?. Palsson. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Mol. Syst. Biol.*, 7:535, Oct 2011.

[137] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245, 2010.

[138] Edward J O'Brien, Jonathan M Monk, and Bernhard O Palsson. Using genome-scale models to predict biological capabilities. *Cell*, 161(5):971–987, 2015.

[139] Bernhard Palsson. The challenges of in silico biology. *Nature biotechnology*, 18(11):1147, 2000.

[140] Samay Pande, Holger Merker, Katrin Bohl, Michael Reichelt, Stefan Schuster, Luís F De Figueiredo, Christoph Kaleta, and Christian Kost. Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *The ISME journal*, 8(5):953, 2014.

[141] Jon Pey, Luis Tobalina, Joaquín Prada J De Cisneros, and Francisco J Planes. A network-based approach for predicting key enzymes explaining metabolite abundance alterations in a disease phenotype. *BMC systems biology*, 7(1):62, 2013.

[142] Thomas Pfeiffer and Stefan Schuster. Game-theoretical approaches to studying the evolution of biochemical systems. *Trends in biochemical sciences*, 30(1):20–25, 2005.

[143] Thomas Pfeiffer, Stefan Schuster, and Sebastian Bonhoeffer. Cooperation and competition in the evolution of atp-producing pathways. *Science*, 292(5516):504–507, 2001.

[144] Johannes Pollmächer, Sandra Timme, Stefan Schuster, Axel A Brakhage, Peter F Zipfel, and Marc Thilo Figge. Deciphering the counterplay of aspergillus fumigatus infection and host inflammation by evolutionary games on graphs. *Scientific reports*, 6, 2016.

[145] Alberto Rezola, Jon Pey, Luis F de Figueiredo, Adam Podhorski, Stefan Schuster, Angel Rubio, and Francisco J Planes. Selection of human tissue-specific elementary flux modes using gene expression data. *Bioinformatics*, 29(16):2009–2016, 2013.

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

[147] David Ricardo. *On the Principles of Political Economy and Taxation*. London: John Murray, 1817.

[148] R Tyrrell Rockafellar. Convex analysis. princeton landmarks in mathematics, 1997.

[149] Sergio Rossell, Martijn A Huynen, and Richard A Notebaart. Inferring metabolic states in uncharacterized environments using gene-expression measurements. *PLoS computational biology*, 9(3):e1002988, 2013.

[150] E. Ruppin, J. A. Papin, L. F. de Figueiredo, and S. Schuster. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr. Opin. Biotechnol.*, 21(4):502–510, Aug 2010.

[151] Rajib Saha, Anupam Chowdhury, and Costas D Maranas. Recent advances in the reconstruction of metabolic models and integration of omics data. *Current opinion in biotechnology*, 29:39–45, 2014.

[152] Satya Swarup Samal, Ovidiu Radulescu, Andreas Weber, and Holger Fröhlich. Linking metabolic network features to phenotypes using sparse group lasso. *Bioinformatics*, 33(21):3445–3453, 2017.

[153] WRM Sandanayaka and EA Backus. Quantitative comparison of stylet penetration behaviors of glassy-winged sharpshooter on selected hosts. *Journal of economic entomology*, 101(4):1183–1197, 2008.

[154] Christophe H Schilling, Jeremy S Edwards, David Letscher, Bernhard Ø Palsson, et al. Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnology and bioengineering*, 71(4):286–306, 2000.

[155] André Schultz and Amina A Qutub. Reconstruction of tissue-specific metabolic networks using corda. *PLoS computational biology*, 12(3):e1004808, 2016.

[156] R Schuster and Stefan Schuster. Refined algorithm and computer program for calculating all non–negative fluxes admissible in steady states of biochemical reaction systems with or without some flux rates fixed. *Bioinformatics*, 9(1):79–85, 1993.

[157] Stefan Schuster, Luis F de Figueiredo, Anja Schroeter, and Christoph Kaleta. Combining metabolic pathway analysis with evolutionary game theory. explaining the occurrence of low-yield pathways by an analytic optimization approach. *Biosystems*, 105(2):147–153, 2011.

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

[159] Stefan Schuster, Jan-Ulrich Kreft, Naama Brenner, Frank Wessely, Günter Theißen, Eytan Ruppin, and Anja Schroeter. Cooperation and cheating in microbial exoenzyme production–theoretical analysis for biotechnological applications. *Biotechnology journal*, 5(7):751–758, 2010.

[160] Stefan Schuster, Jan-Ulrich Kreft, Anja Schroeter, and Thomas Pfeiffer. Use of game-theoretical methods in biochemistry and biophysics. *Journal of biological physics*, 34(1-2):1–17, 2008.

[161] Stefan Schuster, Thomas Pfeiffer, and David A Fell. Is maximization of molar yield in metabolic networks favoured by evolution? *Journal of theoretical biology*, 252(3):497–504, 2008.

[162] Stefan Schuster, Thomas Pfeiffer, Ferdinand Moldenhauer, Ina Koch, and Thomas Dandekar. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to mycoplasma pneumoniae. *Bioinformatics*, 18(2):351–361, 2002.

[163] Roland Schwarz, Patrick Musch, Axel von Kamp, Bernd Engels, Heiner Schirmer, Stefan Schuster, and Thomas Dandekar. Yana–a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC bioinformatics*, 6(1):135, 2005.

[164] M. Sétamou and W. A. Jones. Biology and biometry of sharpshooter homalodisca coagulata (homoptera: Cicadellidae) reared on cowpea. *Annals of the Entomological Society of America*, 98(3):322–328, 2005.

[165] Tomer Shlomi, Moran N Cabili, Markus J Herrgård, Bernhard Ø Palsson, and Eytan Ruppin. Network-based prediction of human tissue-specific metabolism. *Nature biotechnology*, 26(9):1003, 2008.

[166] Tomer Shlomi, Yariv Eisenberg, Roded Sharan, and Eytan Ruppin. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular systems biology*, 3(1):101, 2007.

[167] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[168] Mark S Sisterson. Egg load dynamics of homalodisca vitripennis. *Environmental entomology*, 37(5):1200–1207, 2008.

[169] J Maynard Smith and GR Price. The logic of animal conflict. *Nature*, 246:15–18, 1973.

[170] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge university press, 1982.

[171] Giuseppe Stancanelli, Rodrigo Almeida, Domenico Bosco, David Caffier, Ewelina Czwienczek, Jean-Claude Gregoire, Gabor Hollo, Olaf Mosbach-Schulz, Stephen Parnell, and Claude Bragard. Assessing the risk posed to plant health by xylella fastidiosa in the european union. *CIHEAM International Centre for Advanced Mediterranean Agronomic Studies, Watch Letter*, 33:1–8, 2015.

[172] Joshua Tasoff, Michael T. Mee, and Harris H. Wang. An economic framework of microbial trade. *PLoS ONE*, 10(7):1–20, 07 2015.

[173] Peter D Taylor and Leo B Jonker. Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, 40(1-2):145–156, 1978.

[174] Bas Teusink, Anne Wiersma, Douwe Molenaar, Christof Francke, Willem M de Vos, Roland J Siezen, and Eddy J Smid. Analysis of growth of lactobacillus plantarum wcfs1 on a complex medium using a genome-scale metabolic model. *Journal of Biological Chemistry*, 2006.

[175] Mingyuan Tian and Jennifer L Reed. Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. *Bioinformatics*, 2018.

[176] Cong T Trinh, Ross Carlson, Aaron Wlaschin, and Friedrich Srienc. Design, construction and performance of the most efficient biomass producing e. coli bacterium. *Metabolic engineering*, 8(6):628–638, 2006.

[177] Paul E Turner and Lin Chao. Prisoner's dilemma in an rna virus. *Nature*, 398(6726):441–443, 1999.

[178] L. G. Varela, D. R. Haviland, F. G. Zalom, L. J. Bettiga, R. J. Smith, and K. M. Daane. Uc ipm pest management guidelines: Grape. *UC ANR Pub*, 3448, 2015.

[179] Hal R. Varian. *Intermediate Microeconomics: A Modern Approach (Eighth Edition)*. W. W. Norton & Company, eighth edition, December 2009.

[180] Amlt Varma and Bemhard O Palsson. Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/technology*, 12, 1994.

[181] RP Vivek-Ananth and Areejit Samal. Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. *Biosystems*, 147:1–10, 2016.

[182] Nikos Vlassis, Maria Pires Pacheco, and Thomas Sauter. Fast reconstruction of compact context-specific metabolic network models. *PLoS computational biology*, 10(1):e1003424, 2014.

[183] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton university press, 2007.

[184] Yuliang Wang, James A Eddy, and Nathan D Price. Reconstruction of genome-scale metabolic models for 126 human tissues using mcadre. *BMC systems biology*, 6(1):153, 2012.

[185] Martin Wannagat. *Study of the evolution of symbiosis at the metabolic level using models from game theory and economics*. PhD thesis, Université de Lyon, 2016.

[186] Otto Warburg. On the origin of cancer cells. *Science*, 123(3191):309–314, 1956.

[187] Otto Warburg, Franz Wind, and Erwin Negelein. Ueber den stoffwechsel von tumoren im körper. *Klinische Wochenschrift*, 5(19):829–832, 1926.

[188] Gijsbert D. A. Werner, Joan E. Strassmann, Aniek B. F. Ivens, Daniel J. P. Engelmoer, Erik Verbruggen, David C. Queller, Ronald Noë, Nancy Collins Johnson, Peter Hammerstein, and E. Toby Kiers. Evolution of microbial markets. *Proceedings of the National Academy of Sciences*, 111(4):1237–1244, 2014.

[189] Edwin H Wintermute and Pamela A Silver. Emergent cooperation in microbial metabolism. *Molecular systems biology*, 6(1):407, 2010.

[190] Amy Wu and David Ross. Evolutionary game between commensal and pathogenic microbes in intestinal microbiota. *Games*, 7(3):26, 2016.

[191] H. Yim, R. Haselbeck, W. Niu, C. Pujol-Baxley, A. Burgard, J. Boldt, J. Khandurina, J. D. Trawick, R. E. Osterhout, R. Stephen, J. Estadilla, S. Teisan, H. B. Schreyer, S. Andrae, T. H. Yang, S. Y. Lee, M. J. Burk, and S. Van Dien. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat. Chem. Biol.*, 7(7):445–452, May 2011.

[192] Keren Yizhak, Edoardo Gaude, Sylvia Le Dévédec, Yedael Y Waldman, Gideon Y Stein, Bob van de Water, Christian Frezza, and Eytan Ruppin. Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *Elife*, 3:e03641, 2014.

[193] Shao-Wu Zhang, Wang-Long Gou, and Yan Li. Prediction of metabolic fluxes from gene expression data with huber penalty convex optimization function. *Molecular BioSystems*, 13(5):901–909, 2017.

[194] Lvxing Zhu, Haoran Zheng, Xinying Hu, and Yang Xu. A computational method using differential gene expression to predict altered metabolism of multicellular organisms. *Molecular BioSystems*, 13(11):2418–2427, 2017.

[195] O. Zitka, S. Skalickova, J. Gumulec, M. Masarik, V. Adam, J. Hubalek, L. Trnkova, J. Kruseova, T. Eckschlager, and R. Kizek. Redox status expressed as GSH:GSSG ratio as a marker for oxidative stress in paediatric tumour patients. *Oncol Lett*, 4(6):1247–1253, Dec 2012.

[196] Ali R Zomorrodi and Daniel Segrè. Genome-driven evolutionary game theory helps understand the rise of metabolic interdependencies in microbial communities. *Nature Communications*, 8(1):1563, 2017.

[197] Hadas Zur, Eytan Ruppin, and Tomer Shlomi. imat: an integrative metabolic analysis tool. *Bioinformatics*, 26(24):3140–3142, 2010.

Supplementary figures

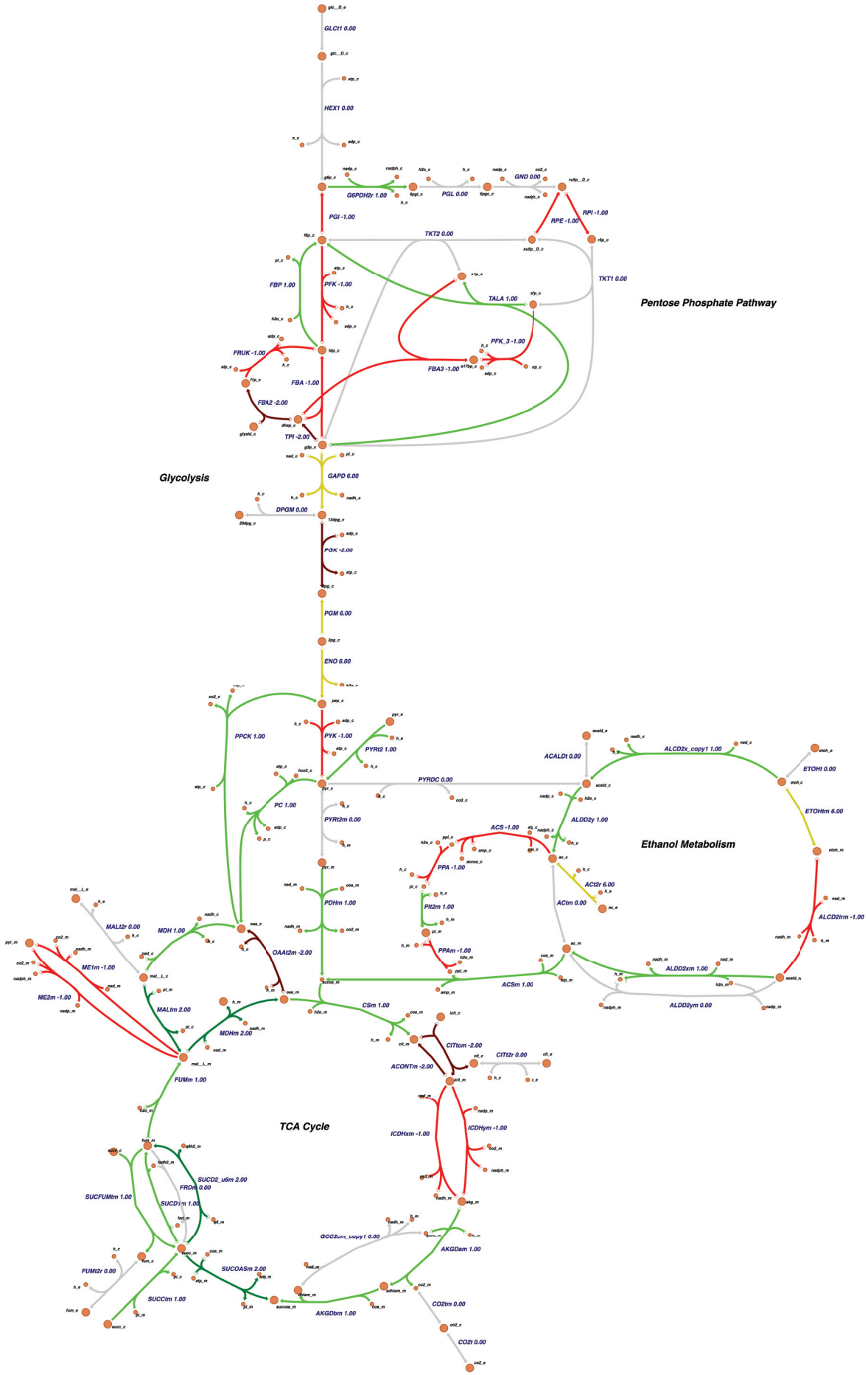

GLCt1 0.00
HEX1 0.00
G6PDH2r 1.00
PGL 0.00
GND 0.00
RPE -1.00
RPI -1.00
PGI -1.00
TKT2 0.00
TKT1 0.00
TALA 1.00
FBP 1.00
PFK -1.00
PFK_3 -1.00
FBA3 -1.00
FRUK -1.00
FBA -1.00
FBA2 -2.00
TPI -2.00
Pentose Phosphate Pathway
Glycolysis
GAPD 6.00
DPGM 0.00
PGK -2.00
PGM 6.00
ENO 6.00
PPCK 1.00
PYK -1.00
PYRt2 1.00
PYRDC 0.00
PC 1.00
PYRt2m 0.00
ACALDt 0.00
ALCD2x_copy1 1.00
ETOHt 0.00
ETOHtm 6.00
ALDD2y 1.00
ACS -1.00
PPA -1.00
ACt2r 6.00
ACtm 0.00
Plt2m 1.00
PPAm -1.00
ACSm 1.00
Ethanol Metabolism
ALCD2irm -1.00
ALDD2xm 1.00
ALDD2ym 0.00
PDHm 1.00
MALt2r 0.00
MDH 1.00
OAAt2m -2.00
ME1m -1.00
ME2m -1.00
MALtm 2.00
MDHm 2.00
CSm 1.00
CITtcm -2.00
CITt2r 0.00
ACONTm -2.00
FUMm 1.00
ICDHxm -1.00
ICDHym -1.00
TCA Cycle
SUCD2_u6m 2.00
FRDm 0.00
SUCD1m 1.00
SUCFUMtm 1.00
AKGDam 1.00
SUCOASm 2.00
FUMt2r 0.00
SUCCtm 1.00
AKGDbm 1.00
CO2tm 0.00
CO2t 0.00

Figure S1

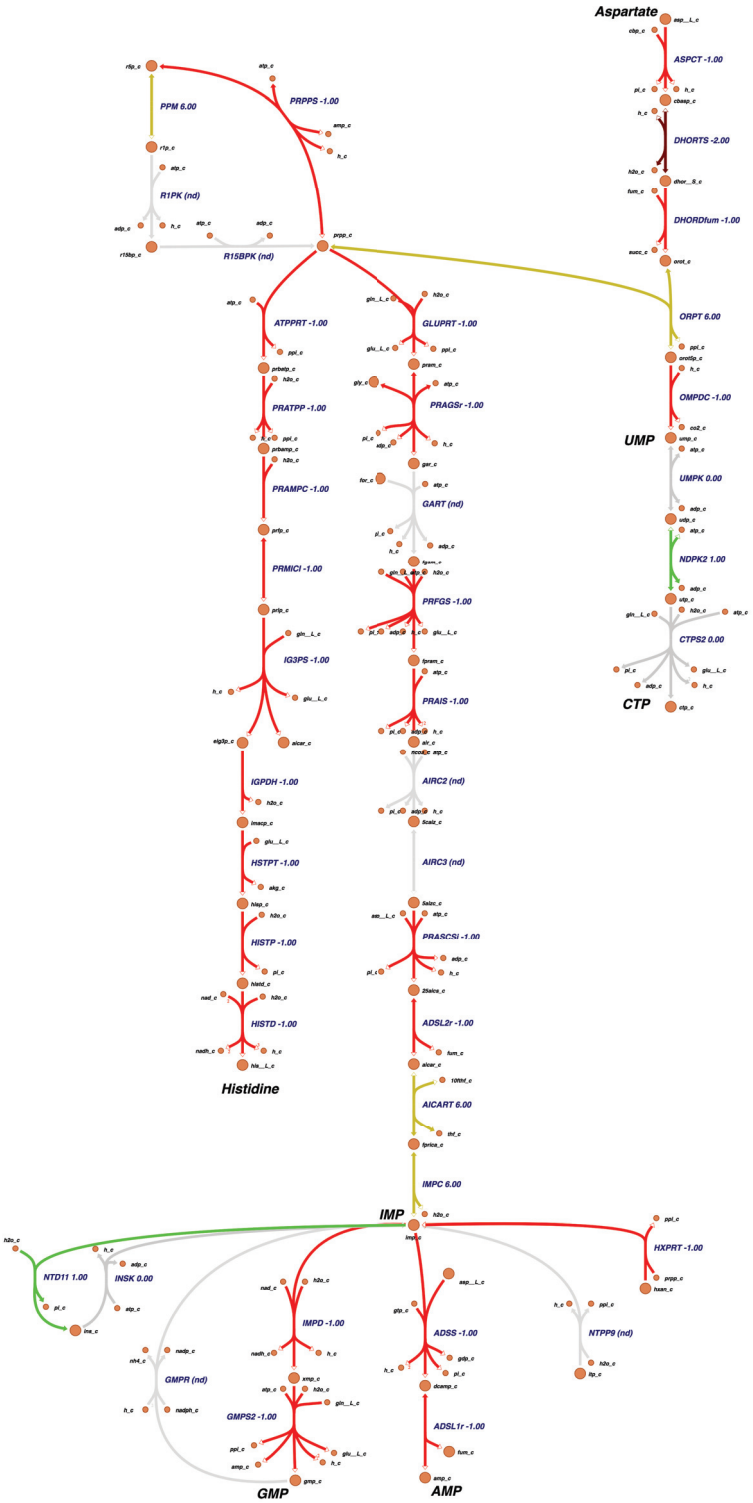

Aspartate
ASPCT -1.00
DHORTS -2.00
DHORDfum -1.00
ORPT 6.00
OMPDC -1.00
UMP
UMPK 0.00
NDPK2 1.00
CTPS2 0.00
CTP
PPM 6.00
PRPPS -1.00
R1PK (nd)
R15BPK (nd)
ATPPRT -1.00
PRATPP -1.00
PRAMPC -1.00
PRMICI -1.00
IG3PS -1.00
IGPDH -1.00
HSTPT -1.00
HISTP -1.00
HISTD -1.00
Histidine
GLUPRT -1.00
PRAGSr -1.00
GART (nd)
PRFGS -1.00
PRAIS -1.00
AIRC2 (nd)
AIRC3 (nd)
PRASCSI -1.00
ADSL2r -1.00
AICART 6.00
IMPC 6.00
IMP
NTD11 1.00
INSK 0.00
IMPD -1.00
GMPR (nd)
GMPS2 -1.00
GMP
ADSS -1.00
ADSL1r -1.00
AMP
HXPRT -1.00
NTPP9 (nd)

Figure S2